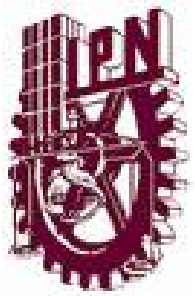


Instituto Politécnico Nacional
Centro de Investigación en Computación

Desambiguación de los sentidos de las palabras en español usando textos paralelos

Que para obtener el grado de:

Doctorado en Ciencias de la Computación

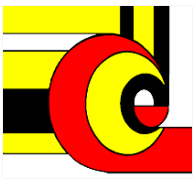
Presenta:

Grettel Barceló Alonso

Directores:

Dr. Alexander Gelbukh

Dr. Grigori Sidorov



México, D.F. Junio 2010

Índice general

Resumen

El carácter polisémico en las palabras se ha definido como la integración en un símbolo de diversas significaciones. Así, el mismo símbolo puede ser utilizado en varios contextos para hacer referencia a conceptos disímiles. En el campo de la lingüística computacional, al proceso de adjudicación del correcto significado se denomina *Desambiguación de los Sentidos de las Palabras* (WSD por sus siglas en inglés - Word Sense Disambiguation).

La WSD es una tarea ampliamente necesaria en aplicaciones relacionadas con procesamiento de lenguaje natural. Destaca su extensamente usanza en ejercicios de traducción artificial y recuperación de información. Por tanto, constituye uno de los pilares en el aseguramiento de la calidad de dichas aplicaciones.

Muchos métodos han sido propuestos para llevar a cabo la WSD y entre ellos, los supervisados han sido de los más sobresalientes; con la presente dificultad de que requieren datos etiquetados de sentidos, adquiridos de forma manual en la gran mayoría de las ocasiones. En este trabajo, se presenta un algoritmo de desambiguación que combina las ventajas de las aproximaciones supervisadas y no supervisadas. Como resultado, se produce un sistema que toma un corpus paralelo bilingüe o multilingüe, se alinea a nivel de palabras y genera como salida las etiquetas de sentidos para todos los idiomas implicados.

Aunque el método puede ser aplicado sobre cualquier par de idiomas, se han utilizado textos de origen en español durante los experimentos realizados. Se emplearon, como textos meta, traducciones a los idiomas italiano, inglés y hebreo. La selección de estas lenguas estuvo basada en la diferenciación de las mismas según el grupo y familia de correspondencia con referencia al español.

Abstract

The polysemous character of words has been defined as the integration in a symbol of several meanings. Thus, the same symbol can be used in different contexts to refer to dissimilar concepts. In the computer linguistic field, the process of adjudication of the correct meaning is called *Word Sense Disambiguation* (WSD).

WSD is a widely necessary task in applications related to natural language processes. Something to bring out is that it is a very common custom to be used in artificial translation and information retrieval exercises. That is why WSD constitutes one of the main pillars in guarantying the quality of such applications.

Many methods have been proposed to carry out WSD and, among them, the supervised ones have been the most outstanding, taking into account the current difficulty that labeling the data of senses implies, which are acquired manually most of the time. In this project, a disambiguation algorithm is presented that combines the advantages of the supervised and unsupervised approximations. The result is a system that takes a bilingual or multilingual parallel corpus, it is aligned at word level and it generates as output labels of senses for all the implied languages.

Although the method can be applied to any set of given languages, Spanish texts have been the ones to be tested. As target texts, translations in to Italian, English and Hebrew have been employed. The reason for selecting these languages was based on the differences among them according to the correspondence of group and family as reference to Spanish.

Capítulo 1

Introducción

Idealmente, debería existir una relación unívoca en la que cada término, en un lenguaje dado, representara únicamente un concepto. Sin embargo, es frecuente que una sola palabra sea empleada para expresar distintos significados, claramente diferentes y muy probablemente sin relación entre uno y otro. A esta pluralidad de significados de cualquier signo lingüístico, se denomina *polisemia* y conlleva a que el significado o sentido de una palabra sea un principio infinitamente variable y sensible al contexto.

La ambigüedad en el sentido de las palabras es una característica intrínseca del lenguaje natural. Por ejemplo, la palabra «sierra» puede tener varios sentidos y hacer referencia a una herramienta para cortar madera, una cordillera de montes con peñascos cortados o inclusive a un pez, según el diccionario de la lengua española. El sentido específico que se debe asociar a la palabra ambigua tendrá que ser determinado por el contexto textual en el cual la instancia aparezca. En el campo de la lingüística computacional este problema es denominado *Desambiguación de los Sentidos de las Palabras* (WSD, por sus siglas en inglés).

La ambigüedad también es un problema en la notación morfológica de un corpus[1]. Puede radicar en una palabra que pertenece a más de una categoría, como «joven» (sustantivo o adjetivo), o en palabras con un rasgo morfológico ambiguo, como «cólera» (género masculino o femenino).

Existen varios mecanismos lingüísticos para definir el sentido correcto de una palabra, basándose en el contexto donde ésta sea empleada y en función de sus posibles sentidos semánticos. Estos acercamientos son clasificados tomando en cuenta la fuente principal de conocimiento usada en la diferenciación de los sentidos. Los principales son: basados en conocimiento y basados en corpus (supervisados y no supervisados), aunque también han surgido combinaciones de los anteriores.

[1]Colecciones muy grandes de textos, normalmente con alguna información lingüística adicional, como las marcas morfológicas, sintácticas, referenciales, etc.

La tarea de desambiguación de sentidos está relacionada con casi todas las aplicaciones que requieren compresión del lenguaje en el área de procesamiento del lenguaje natural. La original y más obvia es la traducción automática, aunque se ha suscitado un reciente crecimiento en áreas como bioinformática y la web semántica [1].

- Traducción automática: Se requiere WSD para elecciones léxicas en palabras que tiene diferentes traducciones para diferentes sentidos y que son potencialmente ambiguas si desconocemos el dominio dado.
- Recuperación de información: Deben ser resueltos problemas de ambigüedad en algunas consultas, aunque en muchas ocasiones el usuario no proporciona suficiente información para determinar el contexto y sólo recuperar documentos relevantes al sentido entendido.
- Extracción de información y procesamiento de textos: Se requiere de WSD para la calidad en el análisis del texto en muchas aplicaciones. Por ejemplo, tener la capacidad de crear el resumen de un documento sobre la base de los datos proporcionados, con un análisis detallado del contenido, sin truncar las primeras líneas de los párrafos.
- Lexicografía: La lexicografía moderna está basada en corpus, con WSD se suministran agrupaciones empíricas e indicadores contextuales de sentidos a los lexicógrafos.

1.1. Planteamiento del problema

El objetivo de la desambiguación de los sentidos de las palabras es determinar el significado correcto, o el sentido de una palabra dada en el contexto. Éste se considera como uno de los problemas más difíciles en el nivel léxico del procesamiento del lenguaje natural [2], [3] y es fundamental en aplicaciones de traducción automática, recuperación de información, etc.

Las dificultades provienen de algunos orígenes, incluyendo la falta de medios para formalizar las propiedades o parámetros del contexto que caracterizan el uso de una palabra ambigua en un sentido en particular y la falta de un inventario de sentidos usuales.

Entre los enfoques que existen para WSD, el de aprendizaje *supervisado* es el más exitoso hasta la fecha [4]. En este se emplea un corpus en el que cada ocurrencia de una palabra ambigua w ha sido manualmente anotada con el sentido correcto, de acuerdo con la existencia de sentidos en un diccionario. De esta forma, este corpus sirve como material de entrenamiento para el algoritmo de aprendizaje, cuyo resultado es un modelo automáticamente aprendido y que puede ser usado para atribuir el sentido

correcto a cualquier ocurrencia de w en un nuevo contexto.

Los corpus empleados en la tarea de clasificación supervisada, requieren una gran cantidad ejemplos para cada sentido de la palabra. Este requisito hace al enfoque supervisado impracticable al tratar de desambiguar todas las palabras en los textos. Este problema ha sido llamado: *obstáculo de adquisición de conocimientos* [5].

Ha sido reconocido que la mejor manera adquirir etiquetas de sentidos para las palabras en un corpus es a mano [6]. Sin embargo, la formulación manual de inventarios de sentidos ha estado basada en la intuición de los lexicógrafos y se ha visto afectada de problemas que incluyen el alto costo, la asignación arbitraria del significado a palabras y el emparejamiento equivocado a dominios de aplicación [7].

Debido a los anteriores inconvenientes, los acercamientos *no supervisados* han recibido mucha atención en los últimos años, pues tienen el potencial para superar el cuello de botella de adquisición de sentidos. Esto lo han conseguido con la introducción del sentido de la palabra directamente del corpus. Sin embargo, lo que parece a simple vista una ventaja, constituye también su principal desventaja: la desambiguación se lleva a cabo partiendo de un conjunto no muy bien definido de sentidos, lo que conlleva en muchas ocasiones a la obtención de resultados pobres durante el proceso de WSD.

A modo de resumen se pueden plantear los siguientes aspectos:

- Los algoritmos supervisados tienen sus desventajas:
  - alto costo del desarrollo del corpus etiquetado de entrenamiento,
  - subjetividad y falta de consistencia en la asignación de las etiquetas,
  - dependencia del dominio y lenguaje: para cada caso se hace un corpus específico.
- Los algoritmos existentes no supervisados tienen otras desventajas:
  - curvas de aprendizaje muy lentas, con lo cual requieren de corpus muy grandes,
  - resultados malos, como consecuencia.

Probablemente, la calidad que proporcionan los algoritmos supervisados se debe a la información adicional que usan, aunque la obtención de esta información conlleva un proceso costoso y muchas veces arbitrario.

Característica	Supervisados	No supervisados	Propuesto
Uso de información adicional para el entrenamiento	√	×	√
Bajo costo de obtención de esta información	×	√	√/×
Objetividad y consistencia de los datos usados	×	√	√
Independencia del dominio y lenguaje	×	√	√
Curva de aprendizaje rápida	√	×	√
Calidad de los resultados	√	×	√

Cuadro 1.1: Métodos supervisados vs. no supervisados

1.2. Solución propuesta

En esta tesis, se propone un método de WSD que combina las ventajas de las dos aproximaciones sin incurrir en sus dificultades. El método usa información adicional, (como lo hacen los métodos supervisados) pero esta información es más fácil de obtener y más objetiva; a saber, se usa una traducción del texto de entrada a otro(s) idioma(s). Con esto, se esperan alcanzar resultados semejantes a los de los métodos supervisados, pero al costo de los no supervisados.

Básicamente la idea consiste en "*utilizar un lenguaje para desambiguar otro*". La ventaja de contar con una traducción del texto facilita el proceso de etiquetado, pues se pueden encontrar intersecciones en los sentidos de las palabras para el conjunto de lenguas. En este punto existe una principal desventaja y es que muchas ambigüedades están preservadas a través de los idiomas. Para aliviar este problema se realiza un filtrado tomando en cuenta los dominios semánticos a los que pertenecen los sentidos comunes y se incorporan cuatro medidas de similitud semántica: Leacock and Chodorow [8], Hirst and St-Onge [9], edge [10] y Random.

El algoritmo de desambiguación que se propone está basado en dos recursos principales: (1) MultiWordNet como léxico especializado para cada uno de los idiomas involucrados y (2) textos paralelos como información adicional para proporcionar diversas lexicalizaciones de las palabras polisémicas. Utiliza como fundamento principal el hecho de que las redes de palabras que conforman MultiWordNet están alineadas.

Se ha considerado en el análisis, la desambiguación de palabras de tres categorías gramaticales (sustantivos, adjetivos y verbos).

El sistema desarrollado toma como entrada un corpus paralelo bilingüe o multilingüe, alineado o NO a nivel de palabra y produce como salida las etiquetas de sentidos para todos los idiomas implicados. La arquitectura general del sistema, organizada por módulos, se muestra en la Figura 1.1.

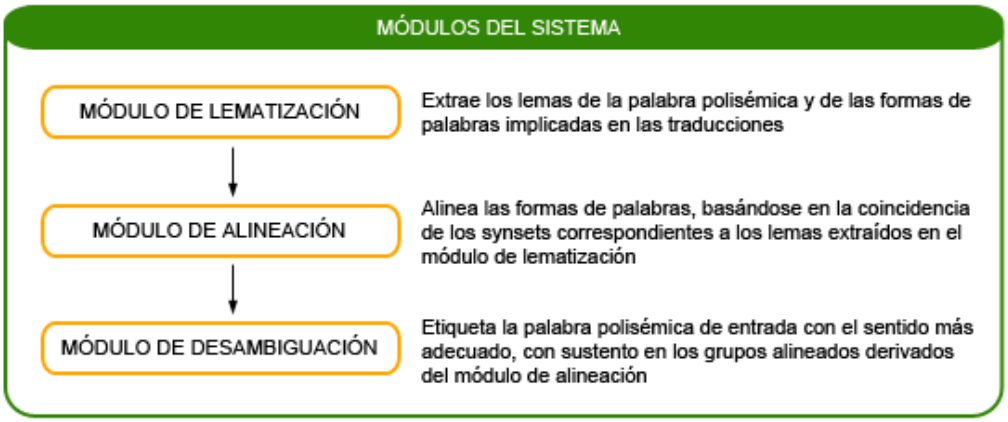


Figura 1.1: Módulos principales del sistema

1.3. Justificación

El uso tan frecuente de las computadoras en la actualidad ha conllevado al desarrollo de múltiples aplicaciones tecnológicas, cuya base es el lenguaje natural. Aunado a esto, el surgimiento de Internet y con ello la gran reserva de recursos de información digitales, incita el desarrollo de algoritmos de búsqueda eficientes, que ponen a nuestra disposición todo un mundo de lenguaje. Con esta plataforma es común el uso de aplicaciones importantes de la lingüística computacional, como: la traducción automática, el manejo y gestión de documentos, la generación de texto y resúmenes, las bibliotecas digitales, las interfaces en lenguaje natural, etc.; cuya efectividad depende en gran medida de la eficacia de los algoritmos que han sido propuestos y que en todas estas tareas consideran la desambiguación de los sentidos de las palabras.

La investigación y mejora de algoritmos de WSD conduce, por consiguiente, a la mejora en las aplicaciones antes mencionadas, lo que justifica el área de investigación seleccionada para esta tesis.

Muchos métodos han sido propuestos para llevar a cabo la desambiguación y entre ellos los supervisados han sido de los más sobresalientes, con la presente dificultad de que requieren datos etiquetados de sentidos, adquiridos en la gran mayoría de las ocasiones, de forma manual.

La disponibilidad de los conjuntos adecuados de datos etiquetados es por tanto un factor significativo y crucial para el entrenamiento de algoritmos basados en aprendizaje. Por desgracia, estos recursos son difíciles de obtener y de crear, pues requieren que cada una de las palabras haya sido etiquetada con su sentido. Este procedimiento resulta muy costoso y por ello solamente algunos artículos de vocabulario terminan siendo etiquetados.

Mientras el cuello de botella léxico parece estar resuelto para el idioma inglés, no existe un amplio rango de recursos lingüísticos para el procesamiento del lenguaje natural en otros idiomas [11]. El recurso de mayor cobertura de palabras es Princeton WordNet, un lexicón de concordancia semántica para el idioma inglés[1], que agrupa los sinónimos en conjuntos denominados synsets y proporciona a las palabras su definición. El propósito de éste es producir una combinación de diccionario y tesauro que sea intuitivamente utilizable y respalde el análisis de texto automático y las aplicaciones de inteligencia artificial. Sin embargo, su utilidad como recurso de entrenamiento y evaluación para anotadores de sentidos supervisados está actualmente algo limitada por su metodología secuencial. Esta notación obliga a los intérpretes a familiarizarse con los inventarios de sentidos de cada palabra [2].

A pesar de sus inconvenientes, este recurso representa una contribución muy importante al campo, proporcionando el primer conjunto de datos balanceados de larga escala para el estudio de las propiedades de polisemia en el idioma inglés.

Desafortunadamente, no existe para el idioma español un recurso léxico de tal magnitud con el cual se pueda proveer de sentido a las palabras, aunque se han están llevando a cabo esfuerzos en esta dirección. Por lo tanto, la búsqueda de recursos disponibles que permitan asociar etiquetas de sentidos y más aún, esto lo realice de forma automática, es una tarea intermedia indispensable en el procesamiento de lenguaje natural.

1.4. Hipótesis

Un recurso que se hace cada vez más disponible en los medios electrónicos compartidos a través de Internet son los textos paralelos. Éstos poseen el mismo contenido semántico, pero expresado en lenguajes diferentes [12] y reciben el nombre de corpus paralelos bilingües. Actualmente existe un creciente interés en la alineación de dichos corpus, que consiste en el establecimiento de correspondencias entre los elementos de un texto y sus traducciones en el otro [13]. La alineación puede ser realizada tomando con base a alguno de los elementos estructurales: párrafos, oraciones o palabras.

Como hipótesis de la investigación se planteó la posibilidad de utilizar los textos paralelos como fuente potencial en el proceso de etiquetado para la desambiguación de los sentidos de las palabras del idioma español; dado que en un corpus paralelo alineado por palabras, las diferentes traducciones en una lengua de meta sirven de etiquetas de sentido de una palabra ambigua en la lengua de origen. Es decir, las distinciones en las traducciones proveen una correlación práctica para la distinción de sentidos.

[1] Según Wikipedia, enciclopedia libre escrita por voluntarios de todo el mundo. http://es.wikipedia.org

Para ello, se partió de los siguientes supuestos:

- Los corpus muy grandes, hasta de texto plano no preparado, son difíciles de obtener para un lenguaje, género y dominio específico.
- Información adicional disponible para un corpus de entrenamiento hace que la curva de aprendizaje se haga más rápida, y por lo tanto, con corpus de un tamaño no muy grande se logran mejores resultados.
- Es más barato encontrar un corpus paralelo alineado (traducido) para un lenguaje, género y dominio específico, que etiquetar un corpus del tamaño adecuado como se hace para el WSD supervisado tradicional.
- La traducción paralela proporciona información adicional sobre el corpus, útil para el aprendizaje de WSD.
- Esta información adicional es suficiente para que el balance de costo/calidad para un método basado en el uso de tal traducción sea mejor que para los métodos tradicionales supervisados y no supervisados.
- La información aprendida de un corpus con traducción (paralelo) será útil para desambiguar nuevos textos (monolingües), de tamaño corto.

1.5. Aportaciones

Las siguientes son las aportaciones a la ciencia más importantes obtenidas durante el trabajo de tesis:

1. Método de alineación a nivel de palabras basado en la correspondencia entre WordNets.
2. Procedimiento de filtrado por dominios semánticos para reducir la polisemia.
3. Técnica para la aplicación de medidas de similitud en pares de palabras con idiomas diferentes.
4. Traducción artificial en sustitución de textos paralelos.
5. Algoritmo para la adquisición automática de etiquetas de sentidos extraídas de la alineación de textos.

Contribuciones técnicas para la evaluación de los métodos anteriores:

1. Estructura de diccionarios de lemas para el español, italiano e inglés y procedimientos de extracción de formas de palabras para cada uno de dichos idiomas.
2. Corpus de evaluación (gold standard) en los pares de idiomas: español / inglés y español / italiano.
3. Sistema de WSD y un kit de herramientas basados en dicho algoritmo.

1.6. Posibles aplicaciones

Con los métodos anteriores implementados se pueden desarrollar las siguientes aplicaciones:

- Desambiguación de corpus monolingües cortos del mismo lenguaje, género y dominio que alguno de los textos etiquetados (usados como datos de entrenamiento).
- Recuperación de información previa desambiguación.
- Incorporación del método propuesto en algoritmos de alineación a nivel de palabras.

1.7. Objetivos

Para dar solución a las necesidades planteadas y poder cumplir satisfactoriamente con cada etapa de investigación y desarrollo de la tesis, se definieron los objetivos que se relacionan a continuación.

1.7.1. Objetivo general

Desarrollar un método que permita desambiguar los sentidos de las palabras para el idioma español y el resto de las lenguas involucradas en corpus paralelos multilingües, alineados a nivel de palabra; explotando así las diferencias polisémicas que prevalecen entre los idiomas.

1.7.2. Objetivos específicos

- Investigar las características de familias de lenguajes para identificar diferencias léxicas importantes, que delimiten los sentidos de las palabras, con el idioma español.
- Encontrar los recursos lingüísticos requeridos (corpus paralelos bi- o multilingües) que se utilizarán en la tarea de desambiguación, en función de los lenguajes seleccionados en la primera fase.
- En caso de que los corpus localizados no estuvieran alineados, seleccionar una herramienta efectiva para su alineación a nivel de palabra o diseñar un método para ello.
- Desarrollar un procedimiento que permita la adquisición automática de etiquetas de sentidos a partir del corpus.
- Comparar las distinciones de sentidos derivadas del algoritmo propuesto con las correspondientes anotaciones hechas por humanos.

1.8. Metodología

Para desarrollar el trabajo propuesto se siguió un conjunto de pasos que para asegurar el cumplimiento de cada uno de los objetivos presentados. A continuación se enumeran las necesidades superadas en el desarrollo de la investigación:

1. Recopilación bibliográfica y análisis detallado de los acercamientos de desambiguación existentes.
2. Caracterización de las familias de lenguajes y su relación con el lenguaje español.
3. Selección del idioma(s) que se empleará como lenguaje meta en los textos paralelos.
4. Búsqueda de corpus paralelos con el par de lenguas: español / meta(s).
5. Comparación y aplicación de diversas herramientas de alineación a nivel de palabras sobre el corpus elegido.
6. Análisis de diccionarios monolingües y bilingües disponibles.
7. Diseño de un algoritmo para la adquisición de etiquetas de sentidos extraídas de la alineación resultante.
8. Definición e implementación computacional de los procedimientos requeridos para la desambiguación de los sentidos de las palabras.
9. Aplicación de los procedimientos implementados sobre el corpus de prueba.
10. Incorporación de un módulo de estadísticas al sistema.
11. Selección de conjunto de pruebas, siguiendo criterios estudiados para la realización de pruebas y refinamiento del método definido.
12. Desarrollo de artículos relacionados con los resultados obtenidos.

1.9. Estructura del documento

Este documento de tesis se encuentra dividido en 6 capítulos. En el segundo capítulo se presentan aspectos relacionados con los antecedentes de la investigación. Consta de dos partes fundamentales. La primera tiene que ver con la desambiguación de los sentidos de las palabras. Al respecto se mencionan: los modos de representación de los sentidos, las etapas principales de la desambiguación, los recursos empleados y los métodos desarrollados para este fin. La segunda parte está relacionada con el concepto de texto paralelo, los niveles y métodos de alineación para este tipo de contenido y su empleo en la desambiguación.

El capítulo tres establece el marco teórico. Se analizan las características principales del español (texto origen) y se determinan los idiomas que fueron usados como texto meta, de los cuales se resumen los rasgos lingüísticos. Además, se profundiza en la estructura de MultiWordNet y WordNet Domains (recursos elegidos para ser utilizados en la desambiguación). Se propone la traducción artificial como fuente de distinción de sentidos y finalmente se establecen los preliminares para el uso de MultiWordNet en la alineación.

En el capítulo cuatro se presenta el método de desambiguación propuesto y los módulos de preprocesamiento requeridos previos a la asignación de sentidos: lematización y alineación. En la descripción de la lematización se especifican los recursos morfológicos empleados para cada uno de los idiomas implicados y la composición de los mismos. En cuanto a la alineación, se describe el algoritmo y la forma en que los dominios y las medidas de similitud semántica son aplicados para la determinación de enlaces no triviales. Por último, en este capítulo se definen el idioma de soporte y el método de back-off cuando no es posible asignar una etiqueta de sentido a una palabra por ausencia de evidencias.

El capítulo cinco inicia con el análisis de la composición de las redes de palabras que conforman MultiWordNet y estadísticas con relación a los sentidos contenidos, a partir de las cuales se determinan los niveles de polisemia de los idiomas empleados. El capítulo continúa con la presentación del corpus de prueba y los resultados obtenidos de la aplicación de los métodos de alineación y desambiguación propuestos, haciendo énfasis en la reducción de ambigüedad categorial y semántica.

La discusión de los resultados obtenidos y las conclusiones se exponen en el capítulo seis. Además, se identifican algunas necesidades no resueltas y se presentan elementos que pueden ser incorporados para dar continuidad al sistema como trabajo futuro.

Los apéndices contienen aspectos particulares relacionados con la algunas peculiaridades de WordNetDomains, gráficos de precisión y recall para los resultados de la alineación de cada oración del corpus de prueba y las palabras polisémicas en la tarea

de resolución de ambigüedad de muestra léxica para el español.

Capítulo 2

Antecedentes

El problema de la desambiguación de los sentidos de las palabras fue formulado por primera vez a principios de la década de los cuarentas, en los orígenes de la traducción automática, haciéndolo uno de los problemas más viejos en la lingüística computacional [1].

En términos generales, la desambiguación involucra la asociación de una palabra ambigua dada en un texto o discurso, con una definición o significado (sentido), que es distinguible del resto de los significados potencialmente atribuibles a dicha palabra. Existen tres modos fundamentales de representar los sentidos de las palabras [14]:

- *Con respecto a un diccionario*:
  dedo = Cada uno de los cinco apéndices articulados en que terminan la mano y el pie del hombre y, en el mismo o menor número, de muchos animales.
  dedo = Medida de longitud, duodécima parte del palmo, que equivale a unos 18 mm

- *Con respecto a su traducción en un segundo lenguaje*:
  dedo = finger
  dedo = toe

- *Con respecto al contexto donde la palabra ocurre (discriminación)*:
  "Me apuntó con el dedo"
  "El zapato me lastima el dedo"

La desambiguación de los sentidos de las palabras consta de dos etapas fundamentales [15]:

1) La definición del conjunto de sentidos para la palabra ambigua o la extracción de los mismos de un diccionario. Por ejemplo, para la palabra ambigua «planta» el diccionario de la lengua española define, entre otros, los siguientes significados:

- planta_1 = Parte inferior del pie.

- planta$_2$ = Árbol u hortaliza que, sembrada y nacida en alguna parte, está dispuesta para trasplantarse en otra.

- planta$_3$ = Fábrica central de energía, instalación industrial.

2) El desarrollo de un algoritmo que asigne el sentido correcto a la palabra para un determinado contexto.

Así, si se tiene la siguiente sentencia:
"*Científicamente se conoció la planta de hierba mate en Europa desde principios del siglo XIX.*"
Se puede determinar que el sentido correcto se corresponde a la definición 2.

Existen varios recursos que permiten obtener los sentidos predefinidos de las palabras durante la primera etapa, como diccionarios, tesauros, corpus y más recientemente, textos paralelos bilingües, que incluyen las traducciones de una entrada en otro lenguaje.

En la segunda etapa, la asignación correcta de los sentidos requiere del análisis del contexto en el cual la palabra ambigua está siendo empleada a partir de fuentes de conocimiento externas, información sobre los contextos de casos previamente desambiguados derivados de un corpus o ejemplos de relaciones entre artículos léxicos de diferentes lenguajes obtenidas de textos paralelos. En cualquier caso, se utiliza algún método de asociación para determinar la mejor correspondencia entre el contexto actual y una de las fuentes de información mencionadas.

2.1. Métodos de desambiguación de los sentidos

Los métodos de asociación para la desambiguación son a menudo clasificados según la fuente principal de conocimiento empleada en la diferenciación del sentido. Los acercamientos que utilizan principalmente diccionarios, tesauros y bases de conocimiento léxicas, son conocidos como métodos basados en diccionario o basados en conocimiento [16].

Las aproximaciones que evitan (casi) completamente información externa y trabajan con corpus que no contienen ninguna información lingüística adicional se denominan métodos no supervisados (adoptando la terminología de aprendizaje de máquina). En esta categoría son incluidas las técnicas que utilizan textos paralelos alineados por palabras. Finalmente, los métodos supervisados y semi-supervisados hacen uso de corpus anotados [1].

Investigaciones recientes estudian los beneficios de combinar las técnicas existentes [17], [18], [19], [11], [20].

2.1.1. Basados en IA

La utilización de los métodos basados en Inteligencia Artificial (IA) para tratar los problemas de la comprensión del lenguaje surge en la década de los sesentas. A menudo involucraron el uso de conocimiento detallado sobre la sintaxis y semántica para alcanzar su objetivo, lo cual fue explotado más tarde para la desambiguación de los sentidos de las palabras.

Métodos simbólicos

Las redes semánticas fueron inmediatamente aplicadas al problema de la representación de los significados de las palabras. Masterman en 1961, trabajando en el área de la traducción automática, utilizó una red semántica para la representación de oraciones en los conceptos fundamentales del lenguaje; la distinción de los sentidos fue hecha implícitamente por la elección de dichas representaciones que reflejaban grupos de nodos estrechamente relacionados en la red [15].

En este sentido, se construyó una red que incluía las relaciones entre palabras y conceptos (o tipos). Estas relaciones fueron etiquetadas con dependencias semánticas o simplemente indicando la asociación entre las palabras. La red fue creada partiendo de un diccionario de definiciones, pero es reforzada por el conocimiento humano que fue codificado manualmente [21], [22].

Las siguientes aproximaciones basadas en IA explotaron el uso de marcos, los cuales contenían información acerca de las palabras y sus roles y relaciones con el resto de las palabras en la oración [23].

Métodos conexionistas

Los trabajos en psico-lingüística en los años sesentas y setentas introdujeron el concepto de la desambiguación hecha por los humanos. Esta idea fue utilizada en los modelos de activación extendida donde los conceptos en una red semántica eran activados sobre su uso y la activación separaba los nodos conectados. Al modelo se agregó la noción de inhibición de los nodos, donde la activación de uno puede suprimir la de sus vecinos [15].

Aplicado a la desambiguación léxica, este acercamiento asumía que la activación de un nodo se correspondía a la activación de alguno de los sentidos del concepto, por ejemplo el físico, y que tal activación inhibiría la activación del resto de los sentidos.

2.1.2. Basados en conocimiento

Las propuestas basadas en conocimiento de los años setentas y ochentas todavía son objeto de investigación actual. Las técnicas principales aplican selección de restric-

ciones, uso de ejemplos para cada significado, traslape en el texto de las definiciones, medidas de similitud semántica y heurísticas. Finalmente, su meta es hacer inferencia semántica general usando el conocimiento almacenado en las bases.

Como recurso léxico en este tipo de métodos se pueden emplear:

- *Diccionarios legibles por las computadoras* [24], [11], [25].- Entre los más utilizados se encuentran: Longman Dictionary of Contemporary English, Collins English Dictionary, Oxford English Dictionary
- *Tesauros.*- Como el Roget's Thesaurus [26]
- *Lexicones computacionales* [27], [28], [29].- Como Princeton WordNet.

Diccionarios legibles por las computadoras

Durante los años ochentas comenzaron los intentos de extracción automática de bases de conocimiento léxicas y semánticas de los diccionarios digitales. A pesar de las limitaciones, estos diccionarios constituyen una fuente de información pre-elaborada sobre los sentidos de las palabras y por consiguiente rápidamente se convirtieron en un elemento fundamental en la investigación de la desambiguación [15].

Todos estos métodos confían en la noción de que el sentido que debe ser asignado a palabras con múltiples co-ocurrencias es aquel que maximiza la relación entre los sentidos elegidos. El algoritmo más importante creado sobre la base de este razonamiento, es el algoritmo de Lesk.

En 1986 Lesk creó una base de conocimiento en la que asociaba con cada sentido en un diccionario una "firma" compuesta de la lista de palabras que aparecen en la definición de ese sentido. La desambiguación es acometida seleccionando aquel sentido de la palabra a etiquetar cuya firma contenga el mayor número de traslapes con las firmas de las palabras vecinas en su contexto [30].

Tesauros

Los tesauros proporcionan información acerca de las relaciones que se establecen entre las palabras, más comúnmente la sinonimia. El Tesauro de Roget ha sido uno de los más empleados en el campo.

Se han desarrollado técnicas para la discriminación de los sentidos de los verbos [31] y la determinación de clases de palabras en categorías comunes [26], de esta forma las clases resultantes se usan para desambiguar las nuevas ocurrencias de una palabra ambigua.

Al igual que los diccionarios, un tesauro es un recurso creado para los humanos y no es por consiguiente una fuente de información perfecta sobre las relaciones entre palabras. Se reconoce ampliamente que los niveles superiores de su jerarquía de concepto están abiertos a la discordancia. No obstante, los tesauros proporcionan una red rica de asociaciones y un conjunto de categorías semánticas potencialmente valioso para el lenguaje; sin embargo, Roget's y otros tesauros no se han usado extensivamente para WSD [15].

Lexicones computacionales

A mediados de los ochentas, se comenzaron a construir manualmente bases de conocimiento de gran escala. En la actualidad el recurso más conocido y empleado en estos métodos es Princeton WordNet(PWN) [32] para la desambiguación de palabras en el idioma inglés.

PWN combina los rasgos de muchos de los otros recursos comúnmente usados en la tarea de desambiguación: incluye las definiciones para los sentidos individuales de palabras, las cuales son agrupadas en conjuntos de sinónimos (synsets) para representar un solo concepto léxico, organizados en una jerarquía. Incluye otros enlaces entre palabras según varias relaciones semánticas, incluyendo: hiperonimia[1] / hiponimia[2], antonimia, meronimia[3], etc. Otra posible razón para el uso extendido de WordNet es su condición de recurso léxico libre y ampliamente disponible.

La mayoría de los trabajos de desambiguación basados en medidas de similitud semántica, utilizan PWN como recurso lingüístico. Se asignan pesos en sus relaciones y se define una métrica que toma en cuenta el número de arcos del mismo tipo que salen de un nodo y la profundidad del árbol [33], [34].

2.1.3. Basados en corpus

En los métodos basados en corpus el contexto de la instancia de una palabra se trata de emparejar con información acerca de los contextos de instancias previamente desambiguadas de la palabra, derivados del corpus [16]. Un corpus es un conjunto estructurado de textos que proporcionan una colección de ejemplos que validan el desarrollo de los modelos de lenguaje.

[1]Cuando un término general puede ser utilizado para referirse a la realidad nombrada por un término más particular.

[2]Cuando una palabra posee todos los rasgos semánticos de otra más general (su hiperónimo), pero que añade en su definición otros rasgos semánticos que la diferencian de la segunda.

[3]Cuando una la palabra cuyo significado constituye una parte del significado total de otra palabra.

Métodos supervisados

Se basan en conjuntos de datos de entrenamiento previamente etiquetados. El sistema de aprendizaje debe poseer entonces una colección de entrenamiento y sus respectivas etiquetas de sentido (categoría), manualmente adquiridas.

Su metodología consta de los siguientes pasos:

- La creación de una muestra de datos de entrenamiento donde una palabra dada es previamente anotada con un sentido de un conjunto predeterminado de posibilidades,
- La conversión de las instancias etiquetadas de sentidos en vectores de características,
- La utilización de un algoritmo de aprendizaje para entrenar al clasificador y finalmente,
- La aplicación del clasificador a las instancias de prueba para asignar la etiqueta de sentido correcta.

En la desambiguación de los sentidos se han aplicado una gran cantidad de algoritmos supervisados con buenos resultados, entre ellos: árboles y listas de decisión, clasificadores bayesianos, perceptrones, redes neuronales, modelos gráficos, etc [14].

Métodos no supervisados

Se basan en corpus no etiquetados. Al igual que los métodos supervisados, el sistema de aprendizaje debe poseer un conjunto de entrenamiento, pero sin sus correspondientes etiquetas de sentidos.

Los algoritmos de aprendizaje no supervisados tienen el potencial para superar el cuello de botella de adquisición de sentidos. Esto lo han conseguido con la introducción del sentido de la palabra directamente del corpus.

En la desambiguación se han empleado técnicas de clustering [35], [36], matrices de co-ocurrencia [37], [38], [39] y textos paralelos [40], [41], [42]. Para lograr la tarea de desambiguación, los algoritmos no supervisados de clustering agrupan las palabras basándose en la similitud del contexto, por tanto no requieren herramientas de soporte, como diccionarios, tesauros o anotaciones manuales [14]. Están soportados por la noción de que "palabras con significados similares tienden a ocurrir en contextos similares [43]."

Recientemente se han propuesto aplicaciones basadas en grafos de co-ocurrencia para todos los pares de palabras que co-ocurren en el contexto de la palabra a ser

etiquetada. Se ha demostrado que esta co-ocurrencia puede utilizada para identificar el sentido correcto de las palabras.

Los textos paralelos requieren estar alineados a nivel de palabra para identificar las traducciones entre dos lenguajes. Esta idea parece proveer una respuesta, pues intuitivamente se supone que si otra lengua lexicaliza una palabra en dos o más maneras, debe haber una motivación conceptual. Si miramos suficientes lenguas, seríamos capaces de encontrar las diferencias léxicas importantes que delimitan los diferentes sentidos de una palabra [44].

2.1.4. Conclusiones

El estudio anterior permite un análisis de las desventajas que presenta cada uno de los métodos existentes para la desambiguación de los sentidos de las palabras.

Métodos basados en:	Desventaja
IA	La dificultad de elaborar manualmente las fuentes conocimiento requeridas para los sistemas basados en IA los restringió a ser implementaciones "juguete" manejando solamente una fracción minúscula del lenguaje. Por lo tanto, los procedimientos de desambiguación embebidos en tales sistemas son probados usualmente en pequeños conjuntos de prueba en un contexto limitado (comúnmente, una sola oración), haciendo imposible determinar su eficacia en los textos reales.
Conocimiento	Utiliza orígenes de conocimientos diferentes hechos a mano, muchos de los sistemas de WSD tempranos usaron este enfoque.
Corpus (*supervisados*)	Los textos etiquetados con los sentidos de las palabras son prácticamente inexistentes por la cantidad de tiempo que se consume y la dificultad para producirlos manualmente.
Corpus (*no supervisados*)	Este acercamiento ha recibido mucha atención en los últimos años, aunque tiene la desventaja que la desambiguación se lleva a cabo partiendo de un conjunto no muy bien definido de sentidos.

Cuadro 2.1: Desventajas de los métodos de desambiguación

2.2. Fundamentos para la alineación de textos

En la década de los ochentas se introdujo la idea de almacenar electrónicamente traducciones pasadas en un formato bilingüe. El concepto consistía en la construcción de una concordancia bilingüe teniendo un operador de datos que manualmente introdujera el texto y su traducción [45], creando así una herramienta de referencia valiosa

para los traductores. Como consecuencia, se originó un gran interés en la construcción automática de tales bases de datos a mayor escala.

En años recientes, se ha suscitado un progreso considerable en el campo de la alineación paralela de textos. El creciente interés en éstos viene dado básicamente por dos motivos: por una parte, la popularización de Internet ha hecho de la red una enorme colección documental bilingüe, y por otra, en la sociedad de la globalización, las organizaciones multinacionales generan enormes cantidades de documentos, escritos usualmente en los idiomas que son nativos en las diversas regiones donde la organización está presente. La expresión "texto paralelo" por sí misma es ahora bien establecida dentro de la comunidad de la lingüística computacional.

Un *texto paralelo* es la unión de dos o más textos que poseen el mismo contenido semántico, pero expresados en lenguajes diferentes [12]. El término paralelo no implica que los textos tengan una correspondencia exacta entre palabras, oraciones y/o párrafos; es decir, dos textos pueden estar completamente desalineados sin dejar de ser textos paralelos [46].

Las herramientas de alineación permiten cotejar textos paralelos, presentándolos en columnas que utilizan marcas textuales para equipararlos. Los algoritmos reforman los textos, estableciendo una relación entre las estructuras de los implicados. Dependiendo del nivel de alineación que se requiera, esta reestructuración puede ser realizada entre párrafos, sentencias o palabras, del contenido expresado en lenguaje *original* y su *traducción* (usualmente denominada meta).

Actualmente se reconocen cuatro[1] niveles de alineación [47]:

- *1er orden*: Este tipo de alineación se emplea cuando los textos que conforman el corpus son muy cortos. Se denomina alineación a nivel de texto.
- *2do orden*: Cuando las correspondencias entre los textos se realizan por párrafos.
- *3er orden*: Los segmentos por alinear son oraciones.
- *4to orden*: Se refiere a la alineación donde los segmentos por alinear son palabras, se le denomina también alineación léxica.

Los textos alineados han demostrado ser una fuente inestimable de datos de traducción para los bancos de terminología y los diccionarios bilingües. Actualmente, la alineación de traducciones está proporcionando la base para el desarrollo de una nueva generación de herramientas de asistencia para los traductores humanos, que permitan mejorar la calidad y productividad de su trabajo.

[1] Aunque algunos estudios han expuesto métodos de alineación a nivel de caracteres [45]

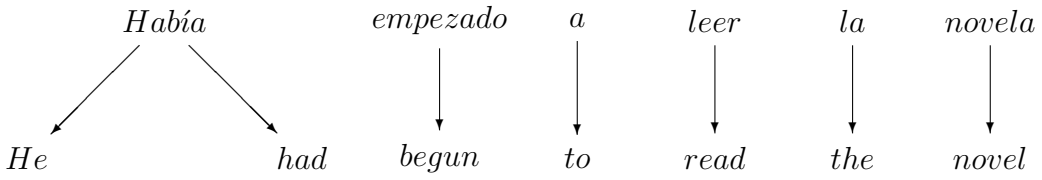


Figura 2.1: Ejemplo de alineación a nivel de palabra

2.2.1. Métodos de alineación

Los métodos propuestos para encontrar alineaciones en textos paralelos, se han clasificado generalmente en dos tipos de aproximaciones: los estadísticos y los lingüísticos, mismos que pueden ser utilizados en cualquiera de los niveles de emparejamiento.

Métodos estadísticos

Los métodos *estadísticos* clásicos utilizan información no-léxica, como la correlación esperada de longitud y posición de las unidades de texto (párrafos u oraciones), la frecuencia de co-ocurrencia, la proporción del tamaño de las oraciones en los diferentes idiomas, etc. [48]. Intentan establecer la correspondencia entre las unidades del tamaño esperado. El tamaño puede medirse en el número de palabras o caracteres[49].

Entre los métodos más destacados que han sido desarrollados en esta categoría y que están basados en longitud se encuentran:

- El propuesto por Gale-Church, que toma como base a la longitud en caracteres de las frases; y su mejora, a la que se denominó método de los costes, donde se asume que cada carácter en un lenguaje produce de forma aleatoria un número de caracteres en otro lenguaje [50].
- Una adaptación del método de Gale-Church propuesta por Brown en el cual se mide la distancia en palabras en lugar de caracteres [51].
- La ampliación con pistas léxicas del método de Gale-Church hecha por Wu, en donde demuestra que la asunciones de Gale-Church no son válidas para lenguajes dispares, como por ejemplo: inglés y chino [52].

Otras contribuciones importantes son los alineamientos basados en cognados [53] y en distancias entre las apariciones de una misma palabra.

Lo atractivo de estos acercamientos es el contraste que se establece entre la utilización mínima de recursos y los resultados satisfactorios que consiguen. Además, como no requieren ninguna fuente externa de información, se consideran independientes del idioma.

Métodos lingüísticos

Por otro lado, los métodos lingüísticos se apoyan en recursos léxicos existentes, como diccionarios bilingües de gran escala y glosarios [48], para establecer la correspondencia entre las unidades estructurales.

Algunos métodos propuestos en esta clase son:

- El formulado por Kay-Roscheisen, en el que asume que las primeras y últimas frases del bitexto están alineadas (serán las primeras anclas) y construye un conjunto de posibles alineaciones según un criterio de distancia. En base a correspondencias léxicas detecta cuáles de las posibles alineaciones son más probables y las fija como nuevas anclas [13].
- Un ajuste al método de los costes de Gale-Church, planteado por Chen. En éste los costes se calculan a través de un modelo de traducción basado en relaciones palabra - palabra [54].
- Una variante del método de Kay-Rocheisen pensada para lenguajes muy dispares (como inglés y japonés), propuesta por Haruno-Yamazaki, donde considera que en lenguajes dispares resulta muy difícil alinear las palabras funcionales (preposiciones, artículos) y por tanto, sólo intenta alinear palabras con contenido (verbos, nombres, adjetivos) [55].

Por la disponibilidad cada vez mayor de recursos bilingües, se invierte más esfuerzo en la investigación de la efectividad de los acercamientos basados en léxicos.

2.2.2. Alineación a nivel de párrafos

El alineado de los textos paralelos, en una primera fase a nivel de párrafo, es insuficiente. Muchas veces se crean paralelizaciones ruidosas y con ambigüedad sintáctica y/o semántica.

El trabajo de detección de párrafos es mucho más simple que el trabajo de detección de oraciones, pues casi siempre están en una relación 1:1, lo cual resulta fortuito ya que la alineación a nivel de oraciones mejora mucho si se realiza primero la alineación a nivel de párrafos [46].

La segmentación y alineado a nivel de párrafo están favorecidos por el marcado DocBook de los documentos.

2.2.3. Alineación a nivel de oraciones

El trabajo inicial en este nivel estuvo enfocado en la reconstrucción de las ligas entre el texto de origen y el texto meta. Este nivel de resolución presentó un desafío

mayor a la alineación automática de unidades textuales más bastas (como párrafo o sección), descubriendo que en él las correspondencias uno - a - muchos y muchos - a - uno, no son raras.

Los primeros algoritmos publicados para alinear oraciones en textos paralelos estuvieron basados únicamente en la observación de la correlación entre la longitud de un texto y la de su traducción, ellos calculan las correspondencias de la oración más probable como una función de la longitud relativa de los candidatos [50], [51]. La desventaja mayor de este método es que una vez el algoritmo tiene accidentalmente un par de oraciones desalineadas, tiende a ser incapaz de corregirse. Por tanto, los algoritmos de alineación basados en las longitudes no son muy robustos ni fiables.

Es por ello, que se introduce el método estadístico de alineamiento basado en cognados, que junta el criterio de la longitud con la noción de cognación [53] para mejorar la eficiencia de los algoritmos antes propuestos.

Otro método de refinamiento, considerado en la categoría de los métodos lingüísticos, que más se ha empleado en este nivel de alineación, es el propuesto por Kay y Roscheisen, en el cual la alineación preferida de la oración es aquella que maximiza el número de las correspondencias sistemáticas entre las palabras [13]. Debili y Sammouda se basaron en este método usando un diccionario bilingüe para dirigir las hipótesis iniciales en las correspondencias de la palabra, de tal modo redujeron el espacio de la búsqueda del algoritmo [56].

Aunque los resultados obtenidos por algunos de los métodos antedichos han sido absolutamente exactos cuando están probados en recopilaciones relativamente limpias y extensas, siguen siendo alineaciones "parciales", pues ocultan el grado más fino de resolución debajo del nivel de oración: el nivel de palabra.

2.2.4. Alineación a nivel de palabras

Este nivel de alineación tiene mayor dificultad que el de oración, puesto que la relación 1:1 entre los elementos que son alineados llega a ser cada vez más rara.

La alineación a nivel de palabras, que puede ser definida como la indicación de la correspondencia entre las palabras en un texto paralelo, fue introducida por primera vez como un resultado intermedio de modelos de traducción estadística [57]. En la traducción automática la alineación de palabras juega un papel crucial, pues los corpus alineados de esta forma han sido considerados como excelentes fuentes de conocimiento relacionadas con la traducción.

Se han propuesto varios métodos para encontrar las alineaciones entre palabras en textos paralelos. Algunas técnicas dependen de un conjunto de parámetros que son

aprendidos mediante un proceso de entrenamiento de datos [57], [58]. Otros obtienen las alineaciones de palabras usando diversas funciones de similitud entre los tipos de los dos idiomas [59], [60], [61].

A pesar de la existencia de diversos métodos a nivel de palabra, todavía la tarea de alineación está lejos de ser un trabajo trivial debido a la diversidad de idiomas naturales. Por ejemplo, la alineación de palabras dentro de las expresiones idiomáticas y traducciones libres son problemáticas. Además, cuando dos idiomas difieren ampliamente en el orden de las palabras, resulta muy difícil encontrar las correspondencias. Por consiguiente, es necesario incorporar información lingüística útil para aliviar estos problemas.

2.3. Textos paralelos en la WSD

Como se mencionó en la sección de métodos de desambiguación, los textos paralelos han sido empleados, dentro de las aproximaciones no supervisadas, como base para la adquisición automática de los sentidos de las palabras [2]. Para ello, el corpus paralelo debe estar alineado a nivel de palabra; de esta forma las diversas traducciones en una lengua meta sirven como las etiquetas de sentido de una palabra ambigua en la lengua original.

El uso de corpus alineados y de bases léxicas multilingües para adquirir automáticamente datos etiquetados de sentido, explotan las diferencias polisémicas entre dos (o más) idiomas. El punto de partida básico para aprovechar esta peculiaridad, es la admisión de que si dos textos son uno la traducción del otro y aluden a los mismos hechos, entonces las palabras contenidas en ellos se deben referir a los mismos conceptos. Con eficacia, la traducción captura el contexto como el traductor lo concibió.

En consecuencia, si un artículo léxico en una lengua es ambiguo, y cada significado corresponde a un artículo léxico distinto en otra lengua, podemos utilizar corpus paralelos para extraer una gran cantidad de "ejemplos desambiguados" en la primera lengua.

En el Cuadro 2.2 se muestran ejemplos de palabras que podrían estar alineadas en textos paralelos. En la primera columna aparece la palabra ambigua en su lenguaje original, la segunda columna muestra dos de los sentidos que puede llegar a tomar la palabra indicada y finalmente su traducción en el idioma meta.

Así, si se encuentra por ejemplo la palabra "bank" en un texto en inglés, se podría conocer el sentido, (sin tomar en cuenta el contexto) al buscar su correspondiente palabra alineada en su contraparte francesa.

Recientes investigaciones plantean nociones diversas para la desambiguación de los sentidos utilizando textos paralelos, aunque pocos reportan resultados con el idioma

Palabra ambigua (lengua original)	Sentidos	Traducciones (lengua meta)
chanel	A path over which electrical signals can pass	频道 (chino)
(inglés)	A relatively narrow body of water	海峡 (chino)
dedo	Cada uno de los cinco apéndices articulados en que termina la mano del hombre y, en el mismo o menor número, de muchos animales	finger (inglés)
(español)	Cada uno de los cinco apéndices articulados en que termina el pie del hombre y, en el mismo o menor número, de muchos animales	toe (inglés)
bank	A building in which the business of banking transacted	banque (francés)
(inglés)	Sloping land (especially the slope beside a body of water)	rive (francés)

Cuadro 2.2: Ejemplos de palabras alineadas en textos paralelos

español. Hay estudios que ocupan textos con traducciones de los idiomas: inglés, rumano, esloveno, checo, búlgaro, estonio y húngaro, extraídos del corpus paralelo George Orwell's Nineteen Eighty-Four [42], [44], [3].

Para el par de lenguas chino e inglés, algunos investigadores utilizan textos paralelos alineados a nivel de palabra con la herramienta GIZA++ [4]. Entre los corpus que son tomados como entrada de este sistema están: Hong Kong Hansards, Hong Kong News, Hong Kong Laws, Sinorama , Xinhua News y la traducción al inglés del Chinese Treebank [62]. Otros acercamientos se apoyan de los diccionarios bilingües inglés / chino y chino / inglés [41].

Del italiano al inglés se han usado como recursos MultiWordNet y MultiSemCor [5] y para inglés - japonés los corpus Wall Street Journal y Nihon Keizai Shimbun [63], [7].

2.3.1. Ventajas y desventajas en el uso de textos paralelos

Después del análisis anterior se pueden puntualizar los beneficios e inconvenientes a los que conduce la utilización de textos paralelos en la desambiguación de los sentidos de las palabras.

Ventajas:

- Disminuye el impacto de la necesidad de corpus anotados manualmente, los cuales provocan problemas de altos costos, división arbitraria de los sentidos de las palabras y otros.
- Utiliza las traducciones en lugar de un inventario de sentidos de palabras.
- Constituye una fuente para buscar datos potenciales de prueba y entrenamiento.

Desventajas:

- Primero, los corpus paralelos, especialmente los corpus paralelos exactamente alineados, son muy escasos.
- El empleo de herramientas de alineación de texto ineficaces a nivel de palabras conlleva a una mala calidad en la desambiguación de los sentidos.
- A menudo no posible distinguir todos los sentidos de una palabra en la lengua original, simplemente confiando en corpus paralelos, especialmente cuando los corpus son relativamente pequeños.
- Muchas ambigüedades están preservadas a través de los idiomas, principalmente cuando los idiomas que se relacionan son de la misma familia.

Capítulo 3

Marco Teórico

Aunque el método de desambiguación propuesto es independiente del lenguaje, el objetivo del trabajo consiste en la asignación correcta de sentidos a las palabras que conforman textos en español. Por tanto, es este idioma considerado como origen en los textos de entrada del algoritmo presentado.

3.1. El lenguaje origen

El *español* es una continuación moderna del latín hablado (denominado latín vulgar), que tras el desmembramiento del Imperio Romano fue divergiendo de las otras variantes del latín que se hablaban en las distintas provincias del antiguo Imperio, dando lugar mediante una lenta evolución a las distintas lenguas neolatinas [64]. Debido a su propagación por América, el español es, con diferencia, la lengua neolatina que ha logrado mayor difusión, siendo el idioma oficial en la mayoría de los países de Latinoamérica. Es conocido también como "castellano", por ser el dialecto peninsular a partir del cual evolucionó el idioma estatal.

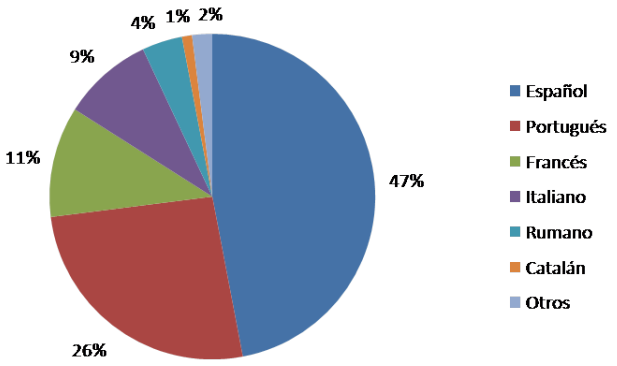


Figura 3.1: Distribución de las lenguas romances

Desde el punto de vista estrictamente lingüístico, el español es un conjunto de cincuenta y ocho variedades [65]. Pertenece a la familia indoeuropea y específicamente al grupo de las lenguas romances, en donde es por mucho el idioma más hablado. Un gráfico de distribución del grupo romance se muestra en la Figura 3.1.

Es uno de los seis idiomas oficiales de la ONU y, tras el chino mandarín, es la lengua más hablada del mundo por el número de hablantes que la tienen como lengua materna [66]. Por otro lado, el español es el segundo idioma más estudiado tras el inglés.

Las siguientes son algunas propiedades lingüísticas del español:

- Según la mayoría de los autores, se distinguen por lo general 24 fonemas en el español.
- Su alfabeto está formado por 27 letras (5 vocales y 22 consonantes) y 2 dígrafos.
- La estructura silábica más frecuente es consonante más vocal.
- Es una lengua flexiva de tipo fusional.
- El acento es de intensidad y estadísticamente dominan las palabras llanas, o acentuadas en la penúltima sílaba, después las agudas y por último las esdrújulas.

3.2. El lenguaje meta

Como se trató en los capítulos anteriores, los textos paralelos pueden ser utilizados como información adicional en el proceso de desambiguación, pues estos proporcionan diversas lexicalizaciones de la palabra en cuestión. Después de esta base debemos plantearnos una pregunta fundamental: ¿Qué lenguaje meta utilizar para desambiguar el español?

Para realizar el estudio se han seguido tres aproximaciones, tomando en consideración la familia y el grupo de pertenencia del idioma respecto al español:

1. Tomar un lenguaje que pertenezca a la misma familia y al mismo grupo.
2. Tomar un lenguaje que pertenezca a la misma familia, pero a un grupo diferente.
3. Tomar un lenguaje que pertenezca a una familia diferente.

Para ello, se seleccionaron como idiomas meta el italiano, el inglés y el hebreo respectivamente. La Figura 3.2 muestra la clasificación de estas lenguas por familia y grupo.

La existencia de los recursos disponibles para los lenguajes fue también un factor de influencia en la selección.

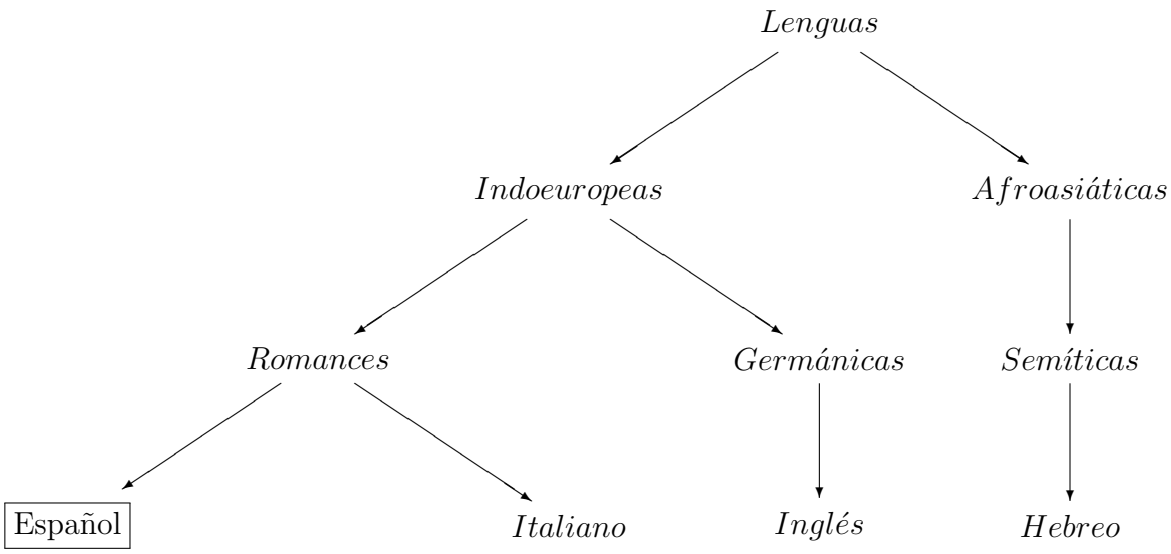


Figura 3.2: Clasificación de lenguajes por familia y grupo

3.3. Familia de lenguas indoeuropeas

Por el número de hablantes, la familia de *lenguas indoeuropeas* es considerada la mayor del mundo. Incluye 150 lenguas habladas por alrededor de 3000 millones de personas. Los siguientes son reconocidos como los grupos más importantes que conforman esta familia:

Grupo Báltico
Grupo Celta
Grupo Eslavo
Grupo Germánico
Grupo Griego
Grupo Romance

Se hará hincapié en los grupos de lenguas más conocidos en América: el Romance y el Germánico, que son además los grupos a los cuales pertenecen los idiomas elegidos italiano e inglés, respectivamente.

3.3.1. Grupo de lenguas romances

Las *lenguas romances*, también conocidas como neolatinas, son una rama indoeuropea de lenguas estrechamente relacionadas entre sí y que históricamente aparecieron como evolución del latín vulgar. Estas lenguas se clasifican a su vez en varios subgrupos, cada uno de los cuales puede comprender diversos dialectos [67]. Las lenguas romances poseen alrededor de 700 millones de hablantes.

Clasificación

El grupo romance se divide en cinco ramas:

1. Ibero-románicas: español, portugués y gallego
2. Italiano: con más de 200 dialectos
3. Lenguas galo-románicas: francés
4. Lenguas occitano - románicas: catalán
5. Rumano

Muchas veces las diferencias entre estos lenguajes son mínimas, llegando a ser inteligibles entre sí en la mayoría de casos. Por ejemplo un hispanohablante puede comprender, tanto de manera escrita como oral, los idiomas: gallego, portugués, catalán e italiano. De manera escrita puede entender: francés y rumano.

Características lingüísticas

- Presencia de al menos dos posibilidades para el género gramatical: masculino y femenino, y dos posibilidades para el número gramatical: singular y plural.
- Para eliminar la ambigüedad en el uso de flexiones se emplean preposiciones.
- Poseen un sistema de artículos.
- Los determinantes generalmente preceden al sustantivo (aunque en rumano el artículo es postpuesto).
- Presencia de un sistema de flexión verbal con numerosas formas.
- Muchas conjugaciones de los verbos poseen carácter irregular.
- El verbo incluye las categorías de persona, número, tiempo y modo gramatical.
- Existen tiempos compuestos.
- Se emplean verbos auxiliares.
- Utilizan adverbios para marcar grados de intensidad.
- Presencia de concordancias gramaticales de género entre el sustantivo y el adjetivo, y entre el número del sujeto y el número expresado en el verbo.
- El orden de los términos no es relevante para el sentido, sino principalmente para el estilo y el énfasis.

3.3.2. Grupo de lenguas germánicas

Las *lenguas germánicas* proceden del territorio ubicado en la parte que hoy ocupan Alemania, Holanda, Bélgica, Dinamarca y el sur de los actuales territorios de Noruega y Suecia. En la actualidad comprenden el inglés, el alemán, el sueco, el islandés, el danés, el holandés, el noruego, entre otros. En conjunto, tienen más hablantes que ninguna otra rama indoeuropea [68]. Este grupo tiene alrededor de 560 millones de hablantes.

Clasificación

El grupo germánico se divide en tres ramas:

1. Oriental: rama extinta. A ella pertenecía la lengua hablada por los godos.
2. Nórdica: sueco, danés, noruego e islandés
3. Occidental: inglés y alemán

Características lingüísticas

- Presencia de dos tipos distintos de conjugación verbal: la débil y la fuerte.
- Doble declinación del adjetivo: fuerte y débil; en dependencia de la precedencia o no de artículo u otros determinantes.
- Mutación consonántica conocida como la Ley de Grimm.
- La migración del acento hacia la raíz de la palabra. Las palabras de origen germánico tienen siempre un acento fijo independientemente de lo que se les añada.
- Pérdida de varias categorías formales del verbo en el plano de tiempos y modos.
- Es típica la adición de -s o -es en la formación del singular genitivo.
- La comparación de adjetivos sigue un patrón regular.
- El orden común en estas lenguas es (Sujeto) - Verbo - Objeto.

3.4. Familia de lenguas afroasiáticas

Las *lenguas afroasiáticas* son una macrofamilia de lenguas que consta de unas 240 lenguas habladas por unos 285 millones de personas distribuidas por el norte y el este de África, el Sahel y el sudoeste asiático. Los siguientes son reconocidos como los grupos más importantes que conforman esta familia:

Grupo Semítico
Grupo Egipcio
Grupo Bereber
Grupo Chádico
Grupo Cusítico
Grupo Omótico

Se enfatiza en el grupo Semítico que es al cual pertenece el idioma hebreo.

3.4.1. Grupo de lenguas semíticas

Las *lenguas semíticas* se desarrollaron sobre todo por Cercano Oriente y el norte y este de África. Toman su nombre del personaje bíblico Sem, cuyos descendientes serían, según la tradición, los pueblos semitas. Las lenguas semíticas más habladas en el presente son el árabe, el amárico, el hebreo y el tigriña. Este grupo tiene alrededor de 350 millones de hablantes.

Clasificación

Las lenguas semíticas suelen dividirse en tres grandes grupos: orientales (Mesopotamia), noroccidentales (Cercano Oriente) y suroccidentales (Península Arábiga y Cuerno de África) [69]. Este agrupamiento está basado en principios geográficos y culturales.

1. Lenguas semíticas orientales: Todas las lenguas de este grupo están actualmente extintas. El lenguaje más conocido de este grupo fue el acadio, que se hablaba en zonas del actual Iraq.

2. Lenguas semíticas noroccidentales: La mayor parte de estas lenguas están también extintas. Sobreviven únicamente dos: el hebreo y el arameo.

3. Lenguas semíticas suroccidentales: Incluye las lenguas del sur de Arabia, el árabe y las lenguas etiópicas.

Características lingüísticas

- Incluyen el género gramatical dual indicado generalmente mediante sufijos especiales.

- Los lexemas de las palabras suelen estar constituidos por esqueletos consonánticos, teniendo el esquema vocálico entre las consonantes información gramatical.

- Entre las raíces consonánticas o lexemas la gran mayoría son de tres consonantes.

- Tienen un sistema de infijos muy desarrollado (los patrones vocálicos son una forma de infijación).

- El infinitivo semítico desempeña la doble función de acción verbal y de sustantivación de esa acción.

- Si los pronombres personales van sufijados a un nombre, equivalen a pronombres posesivos; si van sufijados a un verbo, tienen valor de complemento directo.

- Las lenguas semíticas tienden a la parataxis, es decir, a la agrupación de las oraciones mediante la conjunción copulativa.

3.5. Idiomas elegidos vs. español

3.5.1. Italiano

El *italiano* es una lengua romance. El italiano moderno es, como toda lengua nacional, un dialecto que ha conseguido imponerse como lengua propia de una región mucho más vasta que su región dialectal. En este caso se trata del dialecto toscano de Florencia, Pisa y Siena, que se ha impuesto no por razones políticas, económicas o militares como suele ocurrir, sino debido al prestigio cultural que llevaba consigo al ser el idioma en el que se escribió La Divina Comedia, que se considera la primera obra literaria escrita en la "lingua moderna" [70].

El italiano utiliza 21 letras. Las letras k, j, w, x e y se emplean únicamente en palabras de origen extranjero o variantes gráficas de escritura. Las vocales pueden llevar acentos gráficos, generalmente al final de la palabra. El acento puede ser llano o agudo, pudiendo indicar no sólo la sílaba acentuada sino incluso la apertura de la vocal abierta o cerrada.

El italiano, al igual que el español, tiene una ortografía altamente fonética, es decir, existe una correspondencia considerable entre la lengua escrita y la oral. El acento tónico se encuentra normalmente en la penúltima sílaba, pero también puede estar en la última o en la antepenúltima.

Algunas de las características lingüísticas del italiano son [71]:

- Presencia de dos géneros gramaticales: masculino y femenino.
- Dos números: singular y plural.
- Hay sustantivos cuyo plural es irregular.
- Los artículos y adjetivos son una categoría variable, por tanto tienen variación en género y en número, a fin de concordar con el núcleo del sintagma nominal.
- Existen dos tipos de artículos: indeterminado y determinado.
- Los verbos se dividen en tres categorías o conjugaciones: los verbos terminados en *-are*, en *-ere* y en *-ire*.
- Algunos verbos son irregulares.

- Modos de conjugación: indicativo, subjuntivo, condicional e imperativo.
- Tiempos verbales:
  - Presente: simple, perfecto
  - Pasado: simple, perfecto anterior, perfecto remoto
  - Futuro: simple, perfecto
- Tres formas impersonales: infinitivo, gerundio, participio.
- Los pronombres personales sujetos se sobreentienden, a menos que se quiera insistir en la persona que realiza la acción.

3.5.2. Inglés

El *inglés* pertenece a la familia germánica. Es un idioma originario del norte de Europa, de raíz germánica, que se desarrolló en Inglaterra. Difundido desde su origen por todas las Islas Británicas y en muchas de sus antiguas colonias de ultramar. El inglés, al extender Inglaterra su lengua por todo el mundo (Imperio Británico), y al convertirse los Estados Unidos de América en la mayor potencia económica y militar, se ha convertido de facto en la lingua franca de nuestros días.

El inglés utiliza 26 letras, correspondientes al alfabeto latino sin ninguna adición, salvo en las palabras tomadas directamente de otros idiomas con abecedarios diferentes. No existe el acento gráfico (su acentuación es de tipo prosódica).

Algunas de las características lingüísticas del inglés son [72], [73]:

- Presencia de tres géneros gramaticales: masculino, femenino y neutro.
- La distinción de género sólo existe en los pronombres y adjetivos posesivos de la tercera persona de singular.
- Dos números: singular y plural.
- Hay sustantivos cuyo plural es irregular.
- Tanto el artículo como el adjetivo son invariables de número (excepto los adjetivos demostrativos).
- Existen dos tipos de artículos: indeterminado y determinado.
- Presencia de verbos irregulares.
- Conjugación simple de verbos. Los verbos ingleses tienen una misma forma de escritura (excepto 3ra. persona del singular).

- Modos de conjugación: indicativo, subjuntivo, condicional e imperativo.
- Tiempos verbales:
  - Presente: simple, continuo, perfecto, perfecto continuo
  - Pasado: simple, continuo, perfecto, perfecto continuo
  - Futuro: simple, continuo, perfecto, perfecto continuo
- Tres formas impersonales: infinitivo, gerundio, participio.
- Hay obligatoriedad de expresar el sujeto pronominal.

3.5.3. Hebreo

El *hebreo* es una lengua semítica hablada en su mayoría en Israel y en comunidades judías repartidas por el mundo. En Israel es, junto con el árabe, una de las dos lenguas oficiales del país y es hablada por la mayoría de su población.

La lengua hebrea, como todas las lenguas semíticas, se escribe de derecha a izquierda. Está compuesta por 22 caracteres, de los cuales cinco tienen una grafía distinta en final de palabra.

Algunas de las características lingüísticas del hebreo son [74], [75], [76]:

- Presencia de dos géneros gramaticales: masculino y femenino.
- Tres números: singular, plural y dual. La forma dual para se usa en expresiones de tiempo, cantidades y ciertas formas de sustantivos que inherentemente incluyen parejas (por ejemplo: ojos, orejas, pies)
- El adjetivo sucede al sustantivo y poseen coincidencia en género y número.
- Posee un sistema verbal en el que el uso de ciertas vocales y consonantes denota diferencias en el significado (por ejemplo «katab» “él escribió”; «niktab» “eso fue escrito”; «hiktîb» “él hizo escribir”)
- No tiene el verbo “ser” o “estar” para expresar la identidad o el estado, y en su lugar se emplean frases nominales.
- La conjugación refleja el número, la persona y el género.
- Modos de conjugación: indicativo, imperativo y condicional. No existe el modo subjuntivo.
- Tiempos verbales:
  - Presente perfecto

  - Pasado perfecto
  - Futuro simple
- Tres formas impersonales: infinitivo, gerundio, participio.
- Relativa libertad del orden sintáctico en las frases.

3.5.4. Resumen de las principales características lingüísticas

El estudio anterior permite establecer una comparación entre los idiomas metas y el español, tomando en consideración algunos de los aspectos lingüísticos tratados anteriormente.

Los siguientes son los parámetros que se eligieron para establecer la comparación en el Cuadro 3.1:

1. Géneros gramaticales
   a) M: masculino
   b) F: femenino
   c) N: neutro
2. Número
   a) S: singular
   b) P: plural
   c) D: dual
3. Existe concordancia entre sustantivo / adjetivo?
4. La conjugación del verbo refleja el número, la persona y el género?
5. Modos de conjugación
   a) IND: indicativo
   b) SUBJ: subjuntivo
   c) IMP: imperativo
   d) COND: Condicional
6. Tiempos verbales
   a) Presente: simple (PreS), continuo (PreC), perfecto (PreP), perfecto continuo (PrePC)
   b) Pasado: simple (PasS), continuo (PasC), perfecto (PasP), perfecto continuo (PasPC), perfecto anterior (PasPA), perfecto remoto (PasPR)

   c) Futuro: simple (FutS), continuo (FutC), perfecto (FutP), perfecto continuo (FutPC)

7. Formas impersonales

   a) INF: infinitivo

   b) GER: gerundio

   c) PAR: participio

8. Orden sintáctico

#	Español	Italiano	Inglés	Hebreo
1	M,F,N	M,F	M,F, N	M,F
2	S,P	S,P	S,P	S,P,D
3	SI	SI	NO	SI
4	SI	SI	NO	SI
5	IND,SUBJ,IMP,COND	IND,SUBJ,IMP,COND	IND,SUBJ,IMP,COND	IND,IMP,COND
6a	PreS,PreP	PreS,PreP	PreS,PreC,PreP,PrePC	PreP
6b	PasS,PasC,PasP,PasPA	PasS,PasPA,PasPR	PasS,PasC,PasP,PasPC	PasP
6c	FutS,FutP	FutS,FutP	FutS,FutC,FutP,FutPC	FutS
7	INF,GER,PAR	INF,GER,PAR	INF,GER,PAR	INF,GER,PAR
8	Libre	Libre	SVO	Libertad relativa

Cuadro 3.1: Características lingüísticas de los idiomas implicados

3.6. WordNet

Ya que se han elegido los lenguajes implicados para los textos paralelos, el siguiente paso consiste en la búsqueda de los recursos que se emplearán en la desambiguación. Es necesario entonces contar con un léxico especializado para cada uno de los idiomas involucrados, de donde se puedan extraer los significados de las palabras polisémicas que se desean desambiguar.

Uno de los recursos más importantes empleados en la actualidad para aplicaciones que requieren procesamiento de lenguaje natural son las redes de palabras. Éstas describen las relaciones léxicas y semánticas existentes entre las palabras y recopilan sus sentidos al igual que un diccionario monolingüe.

Princeton WordNet (PWN) es considerada como la base de datos léxica más significativa. Su diseño está en consonancia con teorías psicolingüísticas relativas a la organización de la información en la mente del hablante [32]. WordNet constituye un intento de reflejar el modelo de memoria basado en redes semánticas propuesto por

Collins y Quillian en un modelo lexicográfico.

El recurso empleado en el método propuesto basa su estructura en la implementada por PWN, por lo que es necesario estudiar los conceptos básicos y la organización que posee este sistema electrónico de referencia léxica.

3.6.1. Matriz de vocabulario

La fundamentación teórica del sistema tiene su origen en la idea de la "matriz de vocabulario" [77]; donde el término *forma léxica* se corresponde a la expresión física que se escribe o se pronuncia y *significado léxico* se refiere al concepto que se expresa por medio de una forma.

En la matriz de vocabulario, las columnas contienen todas las formas léxicas del idioma (inglés para el caso de PWN) y las filas contienen todos los significados. Una entrada de una celda de la matriz implica que la forma léxica de una columna puede usarse (en el contexto apropiado) para expresar el significado de esa fila. En el Cuadro 3.2, la entrada E_{11} muestra que la forma F_1 puede usarse para expresar el significado S_1.

Significados	Formas de palabras				
	F_1	F_2	F_3	$\cdots$	F_N
S_1	E_{11}	E_{12}			
S_2		E_{22}			
S_3			E_{33}		
$\vdots$				$\ddots$	
S_M					E_{MN}

Cuadro 3.2: Matriz de vocabulario de PWN

Para identificar una forma polisémica en la matriz, basta con encontrar dos o más entradas en la misma columna, puesto que esto manifiesta que la palabra puede ser empleada para expresar sentidos diferentes. Si hay dos entradas en la misma fila, las dos formas son sinónimas ya que se pueden utilizar indistintamente para expresar el mismo significado.

Ciertamente esta estructura nos permite recuperar información en dos direcciones:

1. Obtener todos los posibles sentidos para una forma en cuestión y
2. Obtener todas las formas posibles de expresar un significado determinado.

Sin embargo, en el método de WSD que se propone, sólo se requiere el primer modo de acceso a la información. Es decir, el punto de partida consiste en la obtención de la forma léxica que se utilizará como entrada en la matriz.

3.6.2. Synsets

Para la representación de los conceptos, PWN se ha estructurado mediante conjuntos de sinónimos. Un conjunto de sinónimos o *synset* (por la abreviación de la denominación inglesa "synonym set") consiste en la unión de todas las formas de palabra de una fila, de modo que esta agrupación representa un concepto abstracto o significado. Por tanto, la sinonimia es la relación léxica primordial en PWN.

Las formas en un synset están agrupadas de tal manera que son intercambiables para cierto contexto. Asimismo, una forma puede aparecer en más de un synset -ambigüedad semántica- y pertenecer a más de una clase gramatical ("part of speech" - POS) -ambigüedad categorial-.

Los synsets se distinguen por medio de un identificador único, compuesto por la clase gramatical y un número arbitrario asignado: *pos#offset*.

Además del conjunto de sinónimos y el identificador, los synsets poseen una descripción del concepto abstracto que representan. A esta descripción o significado se le denomina *glosa*.

El Cuadro 3.3 muestra ejemplos de synsets que contienen la palabra «triangle».

Id	Forma	Glosa
n#03538338	triangle	a percussion instrument consisting of a metal bar bent in the shape of an open triangle
n#03538480	triangle	any of various triangular drafting instruments used to draw straight lines at specified angles
n#09999242	triangle trigon trilateral	a three-sided polygon

Cuadro 3.3: Estructura de los synsets

3.6.3. Relaciones

PWN divide el lexicón en cuatro categorías principales: sustantivos, verbos, adjetivos y adverbios. Éstas se encuentran organizadas en jerarquías basadas en relaciones entre los synsets. Dos tipos de relaciones son representadas por punteros: léxicas y semánticas. Las relaciones *léxicas* tienen lugar entre las formas léxicas semánticamente relacionadas y las relaciones *semánticas* entre los significados. La lista completa de las relaciones es mostrada en la Figura 3.3.

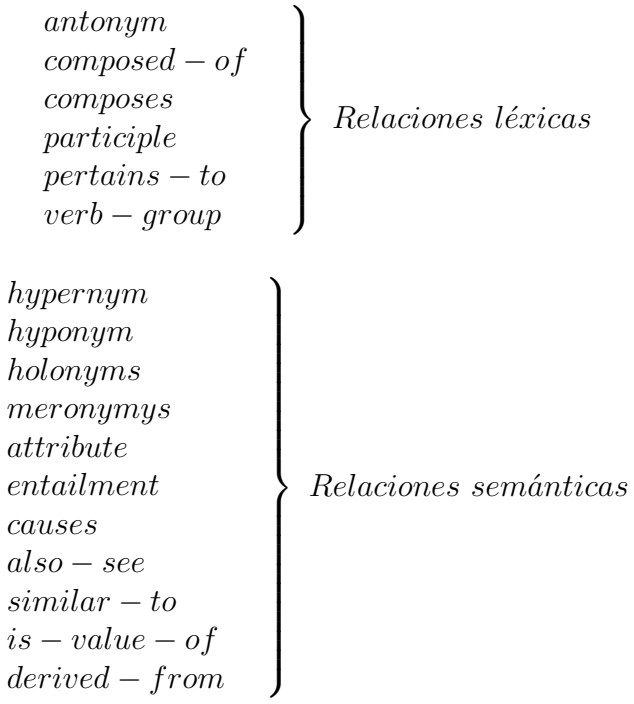


Figura 3.3: Relaciones en PWN

3.7. MultiWordNet

Se han utilizado al menos dos metodologías de construcción de redes de palabras multilingües. La primera consiste en la construcción independiente de wordnets específicas del lenguaje, con una fase posterior de búsqueda de correspondencias entre ellas. Este enfoque es el utilizado por EuroWordNet [78], proyecto que emplea un índice interlingua (ILI) para materializar las relaciones entre idiomas. La segunda metodología consiste en la construcción de las de wordnets específicas del lenguaje manteniendo tanto como sean posibles las relaciones semánticas disponibles en PWN. Este acercamiento se empleó para el desarrollo de MultiWordNet [79].

MultiWordNet (MWN) es una base de datos léxica multilingüe, en la cual se ha realizado una alineación estricta entre Princeton WordNet y redes de palabras para el español, el italiano y el hebreo, entre otros lenguajes.

Los synsets, para cada uno de los idiomas alineados, fueron creados en correspondencia con los synsets de PWN en la medida de las posibilidades. Las relaciones semánticas también fueron importadas de los synsets ingleses correspondientes. Es decir, se asume que si existen dos synsets relacionados en PWN, la misma relación existe en los synsets pertinentes en los otros idiomas [80].

Mientras el proyecto hace hincapié en la utilidad de una alineación precisa entre

redes de palabras de lenguas diferentes, la jerarquía plurilingüe implementada puede representar las verdaderas rarezas léxicas entre idiomas, como brechas léxicas y diferencias de denotación.

El procedimiento de desambiguación que se presenta, se vale de MWN como recurso lingüístico para todos los lenguajes partícipes.

3.8. Dominios en WordNet

Los *dominios semánticos* proveen una manera natural de establecer relaciones semánticas entre los sentidos de palabra. Los dominios son áreas humanas de debate, como POLÍTICA, ECONOMÍA, DEPORTE, que presentan su propia terminología y coherencia léxica. Han sido usados para distinguir los usos técnicos de las palabras y para describir textos de acuerdo con temas generales caracterizados por un ámbito léxico específico [81].

WordNet Domains (WND) fue creado para incluir etiquetas de dominio en PWN. Los synsets fueron anotados con al menos una etiqueta, seleccionada de un conjunto de 169 etiquetas jerárquicamente ordenadas. Un dominio puede incluir synsets de categorías sintácticas y jerarquías diferentes.

Una ventaja adicional que poseen las etiquetas de dominio es el efecto secundario de reducir la polisemia de las palabras en PWN.

3.8.1. Etiquetas de dominio

Las etiquetas empleadas en WND fueron tomadas del sistema de clasificación decimal Dewey[1]. Tienen una estructura de árbol. Cada synset de PWN fue etiquetado con una o más etiquetas [82]. La etiqueta FACTOTUM fue asignada si no existe otra adecuada.

La Figura 3.4 muestra un ejemplo de la jerarquía de dominios empleada.

3.9. Textos paralelos

Los grupos de investigación en procesamiento de lenguaje natural se encuentran en búsqueda constante de nuevos recursos que puedan ser empleados en tareas de desambiguación. Por los inconvenientes que han sido tratados en capítulos anteriores, en términos del coste en la generación de recursos apropiados, se han creado aplicaciones que dependen esencialmente de la existencia de corpus paralelos; es decir, traducciones

[1]Sistema de clasificación de bibliotecas, desarrollado por Melvil Dewey, bibliotecario del Amherst College en Massachusetts, EUA en 1876.

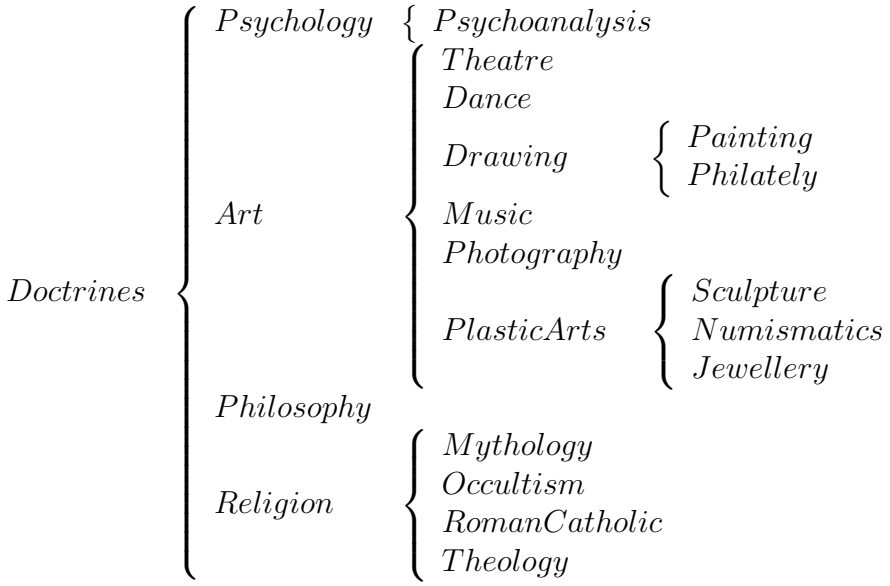


Figura 3.4: Fragmento de la jerarquía de dominios

de textos en grandes cantidades. Sin embargo, la creación de estas traducciones aún requiere de la intervención humana, por lo que se han propuesto algunos métodos para la compilación de grandes cantidades de textos paralelos de la WWW [83].

No obstante, los problemas persisten en lenguajes que están menos representados en formatos electrónicos. El inglés continúa siendo la lengua más difundida en nuestros días. De hecho, pueden encontrarse múltiples contenidos que han sido transcritos del inglés al español y viceversa. Por otra parte, la dupla español / italiano, puede llegarse a encontrar en páginas donde se publican noticias (relacionadas mayormente con la religión [84], [85]), si se realiza una exploración exhaustiva. Examinando el Cuadro 3.4, se podría conjeturar que la menor divulgación de textos paralelos relacionados con esta copla puede deberse a la cantidad de parlantes actualmente del italiano, en comparación con el inglés y el español. Finalmente, el problema se hace más evidente cuando se intentan localizar documentos paralelos para el par español / hebreo.

Lengua	Hablantes como lengua materna (en millones)	Hablantes como 1ra y 2da lengua (en millones)
Español	358	417
Inglés	341	508
Italiano	62	-
Hebreo	8	-

Cuadro 3.4: Número de hablantes de los idiomas implicados

Por la escasa existencia de los recursos requeridos y para aliviar el esfuerzo humano en la creación de las traducciones, parece natural considerar recursos de datos alternativos como pseudo-traducciones, es decir versiones creadas por sistemas de traducción automática.

3.9.1. Traducción artificial

Lo trascendental de la investigación es la observación de que las traducciones, aún siendo automáticas, pueden servir como fuentes de distinción de sentidos. Esto permite evaluar un mismo corpus en los tres lenguajes que han sido elegidos como metas y por ende, obtener resultados más concluyentes. El empleo de corpus diferentes impediría asegurar qué lengua desambigua mejor al español, pues los resultados podrían depender del género del texto e incluso de las características morfológicas de las palabras seleccionadas para desambiguar.

La necesidad de contar con un objetivo de comparación, es otra justificación para que la traducción artificial constituyera la mejor opción en la obtención de los corpus paralelos de entrada en el sistema propuesto. Los textos traducidos manualmente mantienen el inconveniente de que no están etiquetados de sentidos. Por tanto, el acercamiento propuesto parte de un conjunto de prueba estándar, que es transcrito a la lengua meta mediante una tecnología de traducción de máquina comercial, produciendo un corpus paralelo artificial. Los resultados obtenidos pueden ser entonces comparados con los sentidos que han sido asignados por anotadores humanos a cada aparición de las palabras ambiguas.

Por supuesto, la traducción artificial también posee una desventaja: El desempeño de los sistemas comerciales es menor que el de los traductores humanos, en términos de la calidad de la traducción en el corpus obtenido. Sin embargo, un corpus producido artificialmente es más fácil de crear y permite la evaluación de experimentos controlados sin restricciones en el lenguaje meta [40].

3.10. Alineación

Se ha partido del principio de que diversas lexicalizaciones de una palabra en un idioma meta proporcionan la evidencia para un sentido distinto de la palabra en el idioma origen. En este supuesto se necesita contar con la pareja de palabras $< F_k^i(O), F_k^i(T) >$, donde F_k^i representa la k-ésima ocurrencia de la palabra i en la lengua origen y en la traducción respectivamente. Para contar con estas duplas es preciso que los textos estén alineados a nivel de palabras.

El problema en este punto consiste en que los corpus paralelos exactamente alineados son muy escasos. Considerando además que se requieren corpus alineados para un conjunto bien definido de idiomas y que el etiquetado de estos corpus requiere de una

evaluación objetiva, es necesario entonces encontrar métodos de alineación para los corpus paralelos artificiales, obtenidos por medio del sistema comercial de traducción.

3.10.1. Alineación con WordNets

La idea consiste básicamente en el aprovechamiento de los mismos recursos empleados en la desambiguación para su uso en el proceso de alineado.

MultiWordNet fue elegido como léxico especializado, porque contiene redes de palabras para cada uno de los idiomas involucrados. Sin embargo, ofrece también una ventaja adicional a nivel estructural: su concepción ha estado guiada por la exacta alineación con Princeton WordNet. Esta característica constituye la clave en el método que se plantea y permite que tanto el proceso de alineado, como el de desambiguación, se lleven a cabo de forma simultánea.

El hecho de que los WordNets están alineados implica que para expresar un determinado significado, todos utilizan el mismo synset, independiente del idioma. La constitución de los synsets, permite incorporar un grupo de formas sinónimas, por tanto, todos los WordNets en dicho synset almacenan las formas de palabras que pueden representar el significado específico.

Haciendo énfasis en el proceso de alineación, lo que se hace a grandes rasgos es buscar todos los synsets de la palabra a desambiguar, es decir, todos sus posibles sentidos. Este conjunto se compara entonces con los conjuntos extraídos del contexto paralelo (texto en el idioma meta), y aquella palabra que posea mayor coincidencia o intercepción de synsets, será la palabra que se alinea con la palabra polisémica que se desea etiquetar.

En el siguiente capítulo se presenta una descripción formal de este método de alineación con redes de palabras.

Capítulo 4

Algoritmo de Desambiguación

El algoritmo de desambiguación que se propone está basado en dos recursos principales: (1) MultiWordNet como léxico especializado para cada uno de los idiomas involucrados y (2) textos paralelos como información adicional para proporcionar diversas lexicalizaciones de las palabras polisémicas. Utiliza como fundamento principal el hecho de que las redes de palabras que conforman MultiWordNet están alineadas.

Antes de efectuar la desambiguación se requieren dos módulos de preprocesamiento de los textos paralelos: un módulo de lematización y uno de alineación.

4.1. Módulo de lematización

Las correspondencias que deben establecerse en el proceso de alineado, previo a la desambiguación, pueden producirse entre dos palabras como entradas básicas, dos lemas obtenidos de la aplicación de reglas morfológicas sobre las palabras, o la combinación de éstas. Por ejemplo, el establecimiento de la equivalencia entre: *clothes* (en inglés) y *ropas* (en español) se representa por medio de un enlace entre una palabra y su traducción en plural, es decir, entre una entrada léxica básica y otra a la que se le ha aplicado la regla morfológica de formación del plural. Por tanto, todas las palabras implicadas en ambos contextos deben ser lematizadas.

4.1.1. Extracción de lemas para el español

En el caso del español se utilizó un lexicón con información morfológica. Éste fue desarrollado por el Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC. Actualmente cuenta con 1007454 entradas. Cada registro contiene la forma de palabra, su lema y características morfológicas.

Los siguientes son ejemplos de los resultados que se obtienen para la palabra «abrazo»:

abrazo *V0IP1S0* *abrazar*
abrazo *NCMS000* *abrazo*

La primera columna contiene la forma de palabra que estamos consultando y constituye la entrada en el archivo. La segunda muestra información de sus propiedades morfológicas (varían según el POS) y finalmente, la tercera columna presenta todos los lemas posibles. En este análisis no se resuelve la homonimia, es decir, se pueden extraer varios lemas para una misma forma.

A partir de este lexicón se diseñó una base de datos morfológica con 1000799 entradas (sólo sustantivos, adjetivos y verbos), correspondientes a 928211 formas de palabra y 21878 lemas. El Cuadro 4.1 resume las cantidades por clase gramatical.

Clase gramatical	Entradas	Formas	Lemas
Sustantivos	29128	28446	15543
Adjetivos	16482	16456	4947
Verbos	955189	893789	3457

Cuadro 4.1: Composición de la base de lemas para el español

4.1.2. Extracción de lemas para el italiano

Para el italiano se utilizó el mismo acercamiento que para el español, un recurso morfológico denominado Morph-it! [86].

Morph-it! es un diccionario de formas declinadas con su lema y rasgos morfológicos. Puede ser empleado como fuente de información para un lematizador, analizador o generador morfológico del italiano. Los datos para Morph-it! fueron preparados usando una mezcla de métodos de recopilación basados en corpus, reglas basadas en expresiones regulares y chequeo manual.

La versión empleada (0.4.7) cuenta con 504906 entradas. Cada registro contiene la forma de palabra, su lema y características morfológicas.

Los siguientes resultados son obtenidos para un ejemplo de forma en cada una de las clases gramaticales:

gattini *gattino* *NOUN-M:p*
andarono *andare* *VER:ind+past+3+p*
fastidiosetto *fastidioso* *ADJ:dim+m+s*

En este caso, la primera columna contiene la forma de palabra, la segunda almacena su lema y la tercera exhibe sus propiedades morfológicas. Al igual que para el español, las características varían según la clase gramatical.

Se diseñó también una base de datos morfológica para el idioma italiano partiendo del diccionario Morph-it! La BD creada contiene 499052 entradas, correspondientes a 930155 formas de palabra y 23682 lemas. El Cuadro 4.2 resume las cantidades por clase gramatical.

Clase gramatical	Entradas	Formas	Lemas
Sustantivos	35396	32551	17322
Adjetivos	72371	64795	9404
Verbos	391285	313395	6124

Cuadro 4.2: Composición de la base de lemas para el italiano

4.1.3. Extracción de lemas para el inglés

Aunque en PWN se almacenan lemas generalmente, se pueden realizar búsquedas en formas declinadas. Morphy es un conjunto de funciones morfológicas que es aplicado para generar formas que estén presentes en WordNet [87]. Morphy usa dos tipos de procesos para transformar la palabra de entrada en una forma que pueda ser encontrada en la base de datos de PWN: listas de excepciones y reglas de separación.

Esta misma aproximación se ha reproducido para realizar la lematización en el módulo propuesto. Las entradas de las listas de excepciones y las formas regulares fueron almacenadas en una base de datos que contiene información de los lemas y el POS. Así, para cada ingreso particular, se aplican las reglas de división en función de la categoría gramatical. Luego, se ejecuta una búsqueda en PWN para la forma resultante y la categoría específica.

El Cuadro 4.3 muestra las reglas de separación utilizadas por Morphy y que han sido implementadas. Si una palabra termina con uno de los sufijos indicados, éste se elimina de la palabra y se añade la terminación correspondiente.

La base de datos morfológica para el idioma inglés se llenó con las entradas de los archivos de excepciones y de formas regulares. La BD creada contiene 153328 entradas, correspondientes a 145275 formas de palabra y 141116 lemas. El Cuadro 4.4 resume las cantidades por clase gramatical.

POS	Sufijo	Terminación
Sustantivos	-s-	- -
	-ses-	-s-
	-xes-	-x-
	-zes-	-z-
	-ches-	-ch-
	-shes-	-sh-
	-men-	-man-
	-ies-	-y-
Adjetivos	-er-	- -
	-est-	- -
	-er-	-e-
	-est-	-e-
Verbo Verbos	-s-	- -
	-ies-	-y-
	-es-	-e-
	-es-	- -
	-ed-	-e-
	-ed-	- -
	-ing-	-e-
	-ing-	- -

Cuadro 4.3: Reglas de separación del inglés

Clase gramatical	Entradas	Formas	Lemas
Sustantivos	116691	116600	115253
Adjetivos	22928	22910	21666
Verbos	13709	13684	11604

Cuadro 4.4: Composición de la base de lemas para el inglés

4.2. Módulo de alineación

La alineación de las redes de palabras que conforman MultiWordNet constituye el punto de partida en este módulo, pues todas las formas que se usan para representar un significado específico poseen el mismo identificador de synset, independientemente del idioma.

En MWN se manejan elementos léxicos simples, por tanto las expresiones multipalabras no son encontradas y consideraremos sólo equivalencias 1:1, durante la fase de

alineación. De este modo, en el texto quedan descartadas todas aquellas frases indivisibles con un sentido específico.

Por otra parte, el mecanismo desarrollado está enfocado en la extracción de equivalencias con limitación uno a uno [61]. Es decir, todas las palabras participan en una única equivalencia, con excepción del nulo. Esto para evitar correspondencias no deseadas, como por ejemplo tener muchas palabras en español alineadas con la misma palabra en inglés. Así, en cada fase del algoritmo, si una traducción potencial pasa a ser parte de un par de equivalencia, la lista de traducciones potenciales se va reduciendo.

Además, en el algoritmo no existe restricción alguna de la clase gramatical (POS) de las palabras en un par, pues es muy frecuente que durante la traducción se cambie el POS para expresar la misma idea. Por ejemplo, en los fragmentos: ...*che non voglio ricordare come si chiami*...(italiano) y ...*de cuyo nombre no quiero acordarme*...(español), *nombre* puede ser alineado con *si chiami*.

Sea $F_k^i(O)$ la k-ésima ocurrencia de la palabra i en el idioma origen, el módulo de alineación busca determinar las palabras $F_k^i(T1)$, $F_k^i(T2)$, ..., $F_k^i(TN)$ que se corresponden con las traducciones de $F_k^i(O)$ para formar un grupo alineado $g = (l^O, l^{T1}, l^{T2}, ..., l^{TN})$.

Este grupo contendrá todos los lemas de $F_k^i(O)$ en el idioma origen (l^O) y todos los lemas de $F_k^i(T1)$, $F_k^i(T2)$,..., $F_k^i(TN)$ en los idiomas metas ($l^{T1}, l^{T2}, ..., l^{TN}$), que permitirán efectuar el proceso de desambiguado en la fase posterior.

Descripción del flujo en el proceso de alineación

Entradas

- Un corpus paralelo multilingüe ($CP = O \cup T1 \cup T2 \cup ... \cup TN$), donde O representa el texto origen y Tj la traducción o texto meta en la lengua j
- Palabra marcada para desambiguar en el texto origen $F_k^i(O)$

Salidas

- Traducciones de la palabra polisémica en los textos metas $F_k^i(T1), F_k^i(T2), ..., F_k^i(TN)$
- Los lemas de la palabra marcada y de las traducciones anteriores ($l^O, l^{T1}, l^{T2}, ..., l^{TN}$)

Procesamiento

- Extracción del contexto en los textos paralelos.
    - Basados en el supuesto de que los textos origen y metas se encuentran alineados a nivel de oración, el contexto del origen será aquella oración que contenga la palabra que ha sido marcada para desambiguar.
    - Los contextos metas serán las oraciones que se correspondan en numeración con el contexto del origen.
- Lematización de $F_k^i(O)$
    - Siguiendo el proceso descrito en el módulo de lematización, se extraen los lemas correspondientes a $F_k^i(O)$
    - En esta fase ya se cuenta entonces con el primer elemento del grupo (l^O) de alineación.
    - Para cada lema l_x en l^O:
        - Obtención del conjunto de synsets $s(l_x)$
- Lematización de todos los contextos de traducción
    - Para cada texto meta Tj y cada forma de palabra F_x en él:
        - Extracción de sus lemas asociados (posibles l^{Tj})
        - Por cada lema l_x en el l^{Tj} potencial:
        - Obtención del conjunto de synsets $s(l_x)$
        - Se determina la cardinalidad del conjunto de intersección entre el total de synsets de $F_k^i(O)$ y la forma de palabra F_x
- Alineación de las formas de palabra
    - A partir de las cardinalidades obtenidas, puede presentarse alguna de las siguientes situaciones:

1. Que una de las formas de palabra en el contexto paralelo pueda ser directamente asignada como equivalencia léxica de $F_k^i(O)$ por ser partícipe del conjunto de intersección de mayor cardinalidad.
2. Que varias de las formas de palabra en el contexto paralelo posean la misma cardinalidad.
3. Que ninguna de las formas de palabra en el contexto paralelo pueda ser considerada como el par de alineación de $F_k^i(O)$.

En el primer caso el proceso de alineación queda concluido para la palabra $F_k^i(O)$. Cuando no es posible seleccionar una de las formas de palabra por haber muchas que poseen el mismo valor de cardinalidad, se utiliza el sentido más cercano; y en el último caso, donde todas las palabras en el contexto paralelo tienen intersección nula con el conjunto de synstes de $F_k^i(O)$, se aplica una medida de similitud semántica o un método de descartación por concordancia y discordancia de dominios para efectuar el alineado.

4.2.1. Sentido más cercano

Cuando varios pares $< F_k^i(O), \textit{traducción potencial} >$ poseen la mayor cardinalidad de los conjuntos de intersección de sus synsets correspondientes, se toma en cuenta la posición del primer synset común.

Por ejemplo, suponga que la palabra *quiero* ha tenido un synset en común con *have* y uno con *desire*. Para determinar cuál de las dos traducciones será elegida como equivalente, se comparan las posiciones que ocupan en cada una, el synset coincidente. Aquella palabra cuya posición del synset coincidente sea menor, será la ganadora. A continuación se describe el procedimiento para el ejemplo citado.

$< SO_i/SM_j > = <$*"En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor." / "In a village of La Mancha, the name of which I have no desire to call to mind, there lived not long since one of those gentlemen that keep a lance in the lance-rack, an old buckler, a lean hack, and a greyhound for coursing."*$>$

$< F_1^{10}(O), F_1^{11}(T) > = <$ *quiero*, *have* $>$

$g = (l^O, l^T)$ donde:

- $l^o =$ { *querer* [Verb] }

- $l^T =$ { *have* [Sust], *have* [Verb] }

$s(l^O)$	$s(l^T)$		
- v#00472243	- n#07533214	- v#00786286	- v#00523422
- v#00479719	- v#01508689	- v#01620370	- v#01530096
- v#00479841	- v#01794357	- v#01185771	- v#01513366
- v#00808096	- v#01443215	- v#01509557	- v#00045966
- v#01211759	- v#01509295	- v#01876679	- v#01608899
- v#01245362	- v#01857688	- v#00080645	- v#00039991
- v#01530096	- v#00080395	- v#00045715	- v#00978092

$\Rightarrow I(g) = s(l^O) \cap s(l^T) = \{v\#01530096\}$

$< F_1^{10}(O), F_1^{13}(T) > = <$ *quiero*, *desire* $>$

$g = (l^O, l^T)$ donde:

- $l^o =$ { *querer* [Verb] }

- $l^T =$ { *desire* [Sust], *desire* [Verb] }

$s(l^O)$	$s(l^T)$
- v#00472243	- n#05564026
- v#00479719	- n#03868099
- v#00479841	- n#10112569
- v#00808096	- v#01245362
- v#01211759	- v#01246466
- v#01245362	- v#01246175
- v#01530096	

$\Rightarrow I(g) = s(l^O) \cap s(l^T) = \{v\#01245362\}$

Para la forma de palabra *have*, el synset coincidente (v#01530096) ocupa la posición 16, mientras que para *desire*, el synset coincidente (v#01245362) está en la posición 4. Lo anterior se resumen en que es mucho más común utilizar *desire* en el sentido de *querer*, que *have*.

4.2.2. Similitud semántica

La necesidad de determinar la relación semántica entre dos conceptos léxicos es un problema que prevalece en muchas de las tareas de procesamiento del lenguaje. Se han usado medidas de similitud en aplicaciones tales como la desambiguación de los sentidos de las palabras, la determinación de la estructura de los textos, resúmenes y anotación, extracción y recuperación de información, corrección automática de errores, entre otras [88].

En el último caso, donde todas las palabras en el contexto paralelo tienen intersección nula con el conjunto de synstes de $F_k^i(O)$, se han aplicado cuatro medidas de similitud semántica: Leacock and Chodorow [8], Hirst and St-Onge [9], edge [10] y random. Estas medidas han sido implementadas en WordNet::Similarity package [89] y todas ellas se basan de algún modo en la estructura de WordNet. A continuación se describe cada una de estas cuatro medidas:

- *Leacock and Chodorow* (LCH): Esta medida se basa en la longitud $len(s_1, s_2)$ de la ruta más corta entre dos synsets en la jerarquía de hiperónimos y la profundidad máxima D de la taxonomía:

$$LCH(s_1, s_2) = -log\frac{len(s_1, s_2)}{2D}$$

  El hecho de que la medida LCH tome en cuenta la profundidad de la taxonomía en la que se encuentran los synsets significa que el comportamiento de la medida se ve profundamente afectado por la presencia o ausencia de un nodo raíz único. Si hay un nodo raíz único, entonces sólo hay dos taxonomías: una para los sustantivos y otra para los verbos. Todos los sustantivos, entonces, estarán en la misma taxonomía y los verbos en la misma taxonomía. Hay once nodos de partida (primer nivel en la jerarquía) que estructuran la taxonomía de sustantivos en once jerarquías y la estructura o nivel más profundo tiene quince niveles. Para los verbos hay 573 topes. Si el nodo raíz no está siendo utilizado, entonces es posible que los synsets pertenezcan a más de una taxonomía.

  Por ejemplo, el synset que contiene *turtledove#n#2* pertenece a dos taxonomías: una con raíces en *group#n#1* y una con raíces en *entity#n#1*. En tal caso, la relación se calcula hallando el LCS que representa la ruta más corta entre los synsets. Entonces, el valor de D es la máxima profundidad de la taxonomía en la que se encuentra el LCS. Si el LCS pertenece a más de una taxonomía, la taxonomía con la profundidad máxima es seleccionada (es decir, el valor más grande para D).

  El valor máximo de esta medida dependerá de la profundidad de la taxonomía y se obtiene al comparar un synset consigo mismo.

- *Hirst-St-Onge* (HSO): La idea en esta medida de similitud es que dos conceptos lexicalizados son semánticamente cercanos si sus synsets en WordNet están conectados por una ruta que no es tan larga y que no cambia de dirección frecuentemente. Las direcciones de los enlaces en la misma ruta pueden variar (entre: hacia arriba -hiperonimia y meronimia-, hacia abajo -hiponimia y holonimia- y horizontal -antonimia-). La cercanía de la relación está dada por:

$$HSO(s_1, s_2) = C - len(s_1, s_2) - k * d$$

  donde C y k son constantes (en la práctica se usa $C = 8$ y $k = 1$) y d es el número de cambios de dirección en la ruta. Si no existe dicha ruta, entonces HSO es cero y los synsets se consideran independientes.

- *Edge*: Realiza un conteo de la ruta más corta con respecto a la distancia entre synsets. El número de nodos en la jerarquía de WordNet es la medida de la distancia conceptual entre los synsets.

- *Random*: Asigna un número aleatorio entre 0 y 1 como medida de similitud, donde 0 indica que no existe similitud entre los conceptos y 1 que poseen el mismo sentido.

La utilidad *similarity.pl* permite al usuario introducir pares de conceptos en la forma *word#pos#sense* para medir qué tan parecidos son semánticamente. Por ejemplo, en el Cuadro 4.5 aparecen los sentidos de la palabra *breed* como sustantivo.

Entrada en similarity.pl	Synset correspondiente
breed#n#0	*n#06037479*
breed#n#1	*n#06037015*
breed#n#2	*n#07308064*
breed#n#3	*n#03852666*

Cuadro 4.5: Synsets asociados a cada sentido de la palabra *breed* como sustantivo

Esta forma de especificar las entradas presenta una desventaja, pues sólo se aceptan palabras inglesas y las posiciones que ocupan los sentidos definidos de dichas palabras en WordNet. Es decir, *similarity.pl* no se puede aplicar directamente si se están comparando palabras de idiomas diferentes al inglés.

Para solucionar este problema se obtiene la traducción inglesa para cada uno de los synset de la palabra española o italiana implicada en el par de comparación, aprovechando nuevamente la ventaja de que las redes de palabras en MultiWordNet están alineadas.

Así por ejemplo, si se tiene la palabra generación, cuyos synstes como sustantivo son: *n*#06196326, *n*#06195881, *n*#10955750 y *n*#00546392; las posibles traducciones para dichos synstes serían las palabras inglesas que poseen dichos offsets. Sin embargo, como las posibles traducciones para cada synset se emplean para hacer referencia al mismo concepto (mismo sentido), se puede elegir la primera traducción para cada synset como se muestra en el Cuadro 4.6.

Synset	Posibles traducciones	Traducción elegida
n#06196326	*coevals contemporaries generation*	*coevals*
n#06195881	*generation*	*generation*
n#10955750	*generation*	*generation*
n#00546392	*generation multiplication propagation*	*generation*

Cuadro 4.6: Selección de la traducción para cada synset de la palabra *generación* como sustantivo

Ahora sólo bastaría determinar la posición que ocupa el sentido buscado sobre la lista de sentidos de la traducción y se podrían realizar comparaciones entre las palabras *breed* y *generación* con el formato aceptado por el paquete *similarity.pl*, por ejemplo usando las entradas: *breed#n#0 / generation#n#1*. Los valores de similitud[1] de todas las posibles combinaciones se presentan en el Cuadro 4.7.

	T_1	T_2	*breed*	...	T_n
W_1	-1	-1	-1	...	-1
generación	2	-1	4	...	0
W_3	0	-1	3	...	2
...	...	...	...	...	...
W_m	3	-1	2	...	0

Figura 4.1: Matriz de similitud

Una vez que se tienen los valores de similitud de todas las posibles parejas de synstes que se pueden formar con las dos palabras que se comparan, se toma como similitud del par, el mayor valor obtenido.

Este valor se coloca en una matriz de similitud como se muestra en la Figura 4.1. Tal matriz posee el valor -1 en todas las palabras que fueron directamente asignadas por poseer la mayor cardinalidad absoluta, con el objetivo de que no sean tomadas nuevamente en cuenta en la actual etapa de extracción. Así, de la matriz se advierte

[1]Obtenidos empleando el método *Hirst-St-Onge*

coevals#*n*#0 - *breed*#*n*#0 = 0
coevals#*n*#0 - *breed*#*n*#1 = 0
coevals#*n*#0 - *breed*#*n*#2 = 2
coevals#*n*#0 - *breed*#*n*#3 = 2
generation#*n*#1 - *breed*#*n*#0 = 0
generation#*n*#1 - *breed*#*n*#1 = 0
generation#*n*#1 - *breed*#*n*#2 = 2
generation#*n*#1 - *breed*#*n*#3 = 2
generation#*n*#2 - *breed*#*n*#0 = 0
generation#*n*#2 - *breed*#*n*#1 = 0
***generation*#*n*#2 - *breed*#*n*#2 = 4**
generation#*n*#2 - *breed*#*n*#3 = 0
generation#*n*#3 - *breed*#*n*#0 = 0
generation#*n*#3 - *breed*#*n*#1 = 0
generation#*n*#3 - *breed*#*n*#2 = 0
generation#*n*#3 - *breed*#*n*#3 = 0

Cuadro 4.7: Similitud de *breed* y las traducciones de *generación* (*coevals* y *generation*)

que W_1 y T_2 fueron previamente relacionadas en un par de equivalencia.

Finalmente, los pares de equivalencia se forman comenzando por la mayor similitud y así sucesivamente. En el caso de la matriz anterior los pares que se obtendrían con los valores definidos son:

$$< generación, breed >$$

$$< W_m, T_1 >$$

$$< W_3, T_n >$$

4.2.3. Dominios en la alineación

En los casos donde la alineación no puede ser determinada por la cardinalidad del conjunto de intersección de synsets, la correspondencia se realiza analizando los dominios semánticos que posee la palabra polisémica en el origen, y las palabras del contexto de traducción.

De manera general el método consiste en rechazar todas aquellas traducciones que posean dominios diferentes a los de la palabra origen y dar mayor peso a aquellas que

posean dominios coincidentes.

Descripción del flujo en el proceso de alineación

Entradas

- Palabra marcada para desambiguar en el texto origen $F_k^i(O)$

- Contexto de las traducciones

Salidas

- Traducciones de la palabra polisémica en los textos metas $F_k^i(T1), F_k^i(T2), ..., F_k^i(TN)$

- Los lemas de la palabra marcada y de las traducciones anteriores $(l^O, l^{T1}, l^{T2}, ..., l^{TN})$

Procesamiento

- Descartar todas las traducciones que posean al menos una etiqueta de dominio que no está en los dominios de $F_k^i(O)$ (excluyendo FACTOTUM). Es decir, si $dom(F_k^i(T)) \setminus dom(F_k^i(O)) > 0$, se rechaza la traducción, donde *dom* es la función de extracción de categorías semánticas.

- De las traducciones con conjunto diferencia vacío, se toman aquellas que posean más etiquetas a favor (excluyendo FACTOTUM). Es decir, se toma el conjunto $dom(F_k^i(T)) \cap dom(F_k^i(O))$ de mayor cardinalidad.

- Si en el paso anterior resultan varias traducciones, entonces se determina la ganadora por la que tenga similar cantidad de etiquetas FACTOTUM a $F_k^i(O)$. O sea, se elige aquella con $\mid card(dom(F_k^i(T)) = FACTOTUM) - card(dom(F_k^i(O)) = FACTOTUM) \mid$ menor

La extracción de los dominios se lleva a cabo por medio de WND, léxico en el que cada synsets de PWN ha sido asociado con una o varias etiquetas de dominio. Por tanto, si existen synsets en las redes de palabras de español e italiano que no existen en PWN, este análisis no puede ser efectuado.

4.3. Módulo de desambiguación

Una vez que se tienen los textos alineados a nivel de palabra, se puede llevar a cabo el proceso de desambiguación.

Del módulo de alineación se obtiene como salida el grupo $g = (l^O, l^{T1}, l^{T2}, ..., l^{TN})$, tal que que l^O y l^{Tj} se corresponden con los lemas de las formas de palabra $F_k^i(O)$ y $F_k^i(Tj)$ respectivamente.

Entonces, sea $s(l^O)$ los synsets pertinentes a l^O y $s(l^{Tj})$ los pertinentes a l^{Tj}, este módulo busca etiquetar la palabra $F_k^i(O)$, con el synset cuyo significado describa más adecuadamente la intención que se le da a la palabra en el contexto textual en que ésta aparece.

En otros acercamientos, el contexto sería clave en el procedimiento de desambiguación porque de él depende el sentido con el que se interpreta la palabra. Sin embargo, el método propuesto sólo se vale de la traducción alineada, que ha sido determinada en fase anterior por el módulo de alineación.

Descripción del flujo en el proceso de desambiguación

Entrada

- El grupo alineado $g = (l^O, l^{T1}, l^{T2}, ..., l^{TN})$ para la palabra a desambiguar.

Salida

- La etiqueta de sentido para la palabra polisémica.

Procesamiento

- Intersección del conjunto de synsets

  - Se realiza un filtrado por POS, tomando sólo aquellos lemas que coincidan en categoría gramatical.
  - Si para cualesquiera lemas $l_i \in l^O, l_j \in l^{T1}, l_k \in l^{T2}, ..., l_m \in l^{TN}$, se tiene que $POS(l_i) = POS(l_j) = POS(l_k) = ... = POS(l_m)$ entonces se realiza un filtrado por offset. Es decir, $I(g) = s(l_i) \cap s(l_j) \cap s(l_k) \cap ... \cap s(l_m)$, siendo $I(g)$ el conjunto de intersección del grupo alineado g y $s(l_x)$ una función de extracción del identificador para los synsets asociados al lema l_x

- Asociación de la etiqueta de sentido:

  - Si el conjunto de intersección $I(g) = 1$, las palabras pueden ser desambiguadas completamente, pues se toma el sentido del empalme como el correcto.
  - Si el conjunto de intersección $I(g) = \emptyset$, se realiza un filtrado por dominios, quedando como posibles etiquetas sólo aquellos sentidos de la traducción que posean dominios comunes con los sentidos del texto origen.
  - Si el conjunto de intersección $I(g) > 1$, se aplica el método de back-off indicado en la configuración del sistema.

4.3.1. Dominios en la desambiguación

En la sección 3.8 se mencionó la ventaja adicional que ofrecen los dominios al reducir la polisemia de palabras en WordNets. Esta cualidad es la que ahora se utiliza para disminuir el número de etiquetas potenciales cuando el conjunto de intersección $I(g)$ es vacío.

La ausencia de synsets comunes en la primera etapa de filtrado puede deberse a una o varias de las siguientes razones:

- Errores en la traducción automática
- Alineación incorrecta
- Brechas léxicas (gaps)
- Incompletitud de las redes de palabras que conforman MWN

Ante la presencia de los dos primeros problemas es muy improbable que se obtenga un etiquetado de sentido correcto para la palabra polisémica. Sin embargo, en los dos últimos se puede conseguir un marcado congruente tras realizar el filtrado por dominios.

De cualquier modo, independientemente del problema por el cual no existan synsets comunes entre los lemas del texto origen y las traducciones, se procede con la supresión de sentidos mediante el filtrado por dominios; pues éstos, en su función de categorías semánticas, establecen relaciones entre los sentidos.

Descripción del flujo en el proceso de filtrado por dominios

Entrada

- Los synsets (no comunes): $s(l^O), s(l^{T1}), s(l^{T2}), ..., s(l^{TN})$ para el grupo alineado g.

Salida

- La etiqueta de sentido para la palabra polisémica.

Procesamiento

- Intersección del conjunto de synsets

  - Se realiza un filtrado por POS, tomando sólo aquellos lemas que coincidan en categoría gramatical.
  - Si para cualesquiera lemas $l_i \in l^O, l_j \in l^{T1}, l_k \in l^{T2}, ..., l_m \in l^{TN}$, se tiene que $POS(l_i) = POS(l_j) = POS(l_k) = ... = POS(l_m)$ entonces se realiza un filtrado por dominio. Es decir, $I(g) = d(l_i) \cap d(l_j) \cap d(l_k) \cap ... \cap d(l_m)$, siendo $I(g)$ el conjunto de intersección del grupo alineado g y $d(l_x)$ una función de extracción del dominio para los synsets asociados al lema l_x

- Asociación de la etiqueta de sentido:

  - Si el conjunto de intersección $I(g) = 1$, se puede etiquetar la palabra ambigua con cualquiera de los $N + 1$ sentidos -1 por idioma- cuyos dominios coincidieron, donde N representa la cantidad de traducciones del texto origen. La elección de uno u otro sentido se indica en la configuración del sistema con la elección de un idioma de soporte.
  - Si el conjunto de intersección $I(g) = \emptyset$, se aplica directamente el método de back-off sobre todos los synsets del idioma de soporte seleccionado.
  - Si el conjunto de intersección $I(g) > 1$, se aplica el método de back-off sobre los synsets del idioma de soporte seleccionado que pertenecen a dicho conjunto.

4.3.2. Idioma de soporte

En teoría, se puede elegir cualquier red de palabra para el etiquetado de sentidos porque éstas están alineadas en MWN.

La elección del WordNet resulta indiferente cuando hay synsets comunes [y no es necesario realizar el filtrado por dominios], pues independientemente del idioma que se elija durante el proceso de marcado, la identificación del synset ganador (ya sea por unicidad o por aplicación del back-off) es la misma.

Sin embargo, cuando no hay synsets comunes (independientemente que hayan o no dominios comunes) hay que decidir sobre qué conjunto de synsets aplicar el back-off. Para un corpus paralelo con N traducciones de un texto original, existen $N + 1$ posibles idiomas y por tanto $N + 1$ bases de datos léxicas que se pueden emplear para el marcado. Así, la designación de un synset, sí depende del idioma de marcado que se haya preferido.

A este lenguaje se le ha denominado idioma de soporte y su elección se puede hacer en la configuración del sistema.

4.3.3. Back-off

Cuando no es posible asignar una etiqueta de sentido a una palabra por ausencia de evidencias, es necesario utilizar un método de respaldo o back-off. Al presente, el sistema que se ha implementado permite elegir entre dos métodos de back-off:

1. Primer sentido
2. Sentido aleatorio

4.4. Arquitectura del sistema

El sistema de desambiguación implementado, ha sido concebido como la integración, en una herramienta, de tres módulos principales:

1. Lematización
2. Alineación
3. Desambiguación

Los dos primeros forman parte de la etapa de preprocesamiento de los textos paralelos.

Cada módulo cumple una función específica dentro del sistema y las salidas en uno son tomadas como entradas en el siguiente nivel. La Figura 4.2 muestra la dependencia y la forma en que deben ser acoplados los módulos referidos.

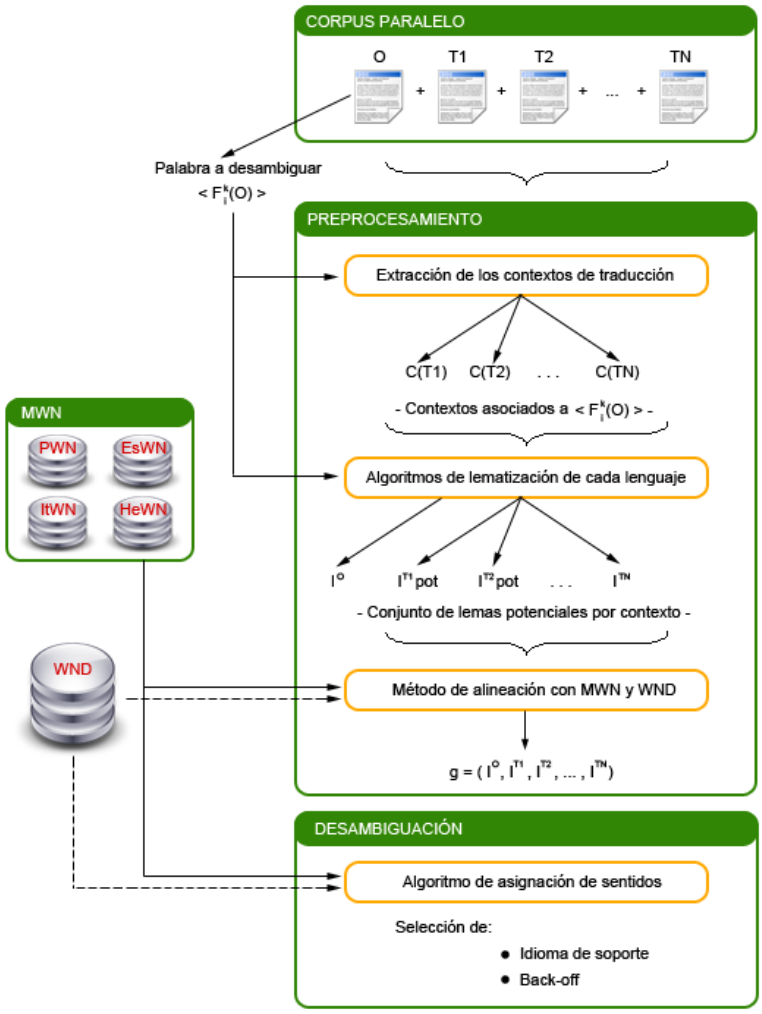


Figura 4.2: Arquitectura del sistema de desambiguación

4.4.1. Interfaz del sistema

La Figura 4.3 muestra la interfaz del sistema de desambiguación. En un inicio se cargan los textos paralelos y se especifican los lenguajes de los mismos. Las formas que pueden ser desambiguadas por su categoría gramatical (sustantivos, adjetivos y verbos) se indican mediante enlaces en el texto origen.

Al presionar sobre un enlace, correspondiente en el algoritmo con $F_k^i(O)$, se ilumina en el texto meta la palabra con la cual $F_k^i(O)$ ha sido alineada, es decir $F_k^i(T)$. Al mismo tiempo se despliega la información de los lemas del grupo alineado $g = (l^O, l^T)$.

En la segunda etapa se realiza el filtrado de sentidos por intersección de synsets o en caso de ser necesario, por dominios. Como se describió en los algoritmos, el conjunto

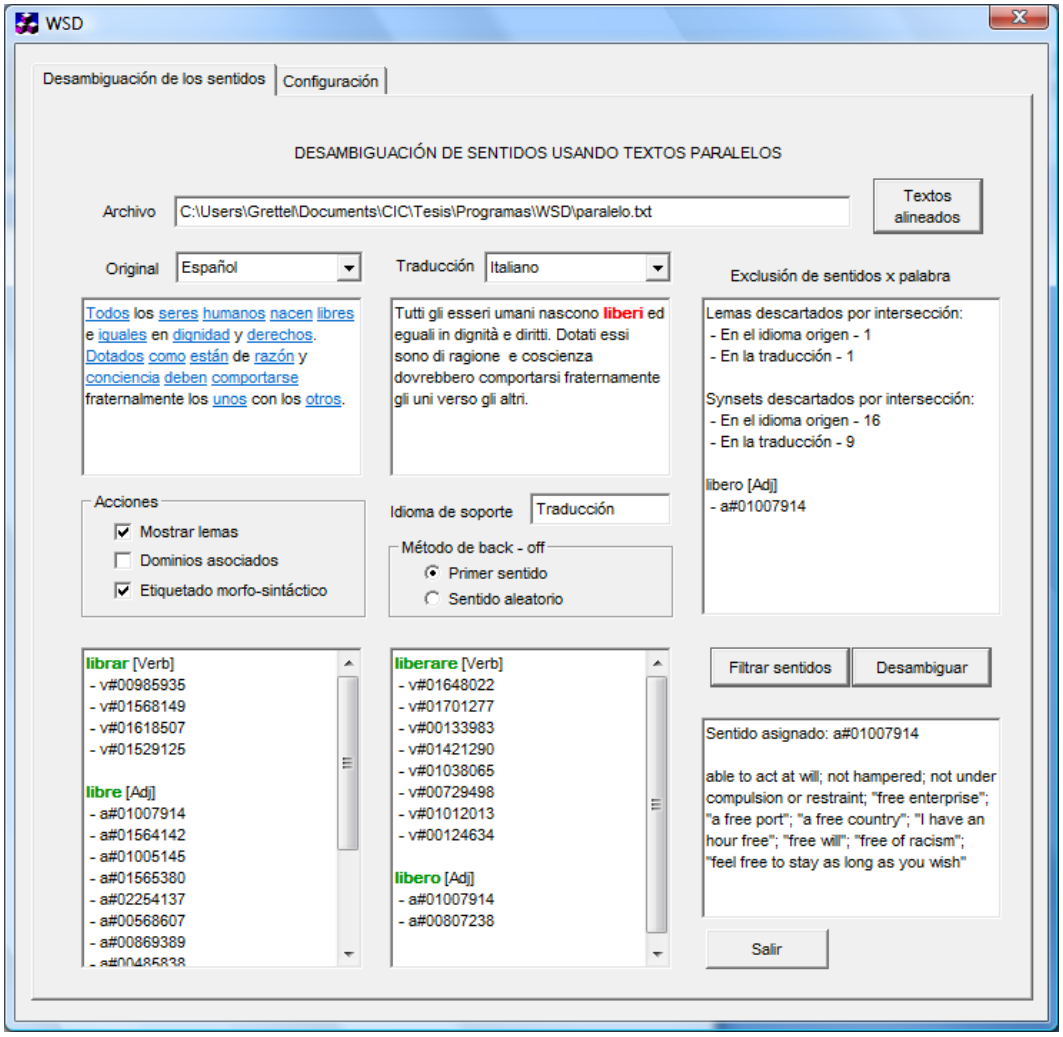


Figura 4.3: Interfaz del sistema de desambiguación

de intersección resultante puede contener:

- Un único elemento. En cuyo caso se desambigua totalmente la palabra.
- Ninguno o varios elementos. En tales casos se procede a ejecutar el método de back-off indicado.

4.5. Aplicación del algoritmo de desambiguación

Se utiliza un ejemplo con el par de idiomas español / italiano para ilustrar el caso en el que la intersección de synsets resulta en un único elemento y cuando resulta en varios elementos.

"*Todos los seres humanos nacen libres e iguales en dignidad y derecho. Dotados como están de razón y conciencia deben comportarse fraternalmente los unos con los otros.*"

"*Tutti gli esseri umani nascono liberi ed eguali in dignità e diritti. Dotati essi sono di ragione e coscienza dovrebbero comportarsi fraternamente gli uni verso gli altri.*"

- Caso en el que la intersección de synsets resulta en un elemento

$< F_1^{12}(O), F_1^{12}(T) > = <$ *derecho*, *diritti* $>$

$g = (l^O, l^T)$ donde:

- $l^o =$ { *derecho* [Adj], *derecho* [Sust] }

- $l^T =$ { *diritto* [Adj], *diritto* [Sust] }

$s(l^O)$	$s(l^T)$
$-a\#01959343$	$-a\#02366741$
$-a\#01182936$	$-a\#01184641$
$-a\#01186048$	$-n\#04030305$
$-a\#01959702$	$-n\#03051767$
$-a\#01956633$	$-n\#06243906$
$-a\#01188689$	$-n\#04680638$
$-a\#02203298$	$-n\#00370892$
$-n\#04030305$	$-n\#02894644$
$-n\#00403152$	$-n\#00371093$

$\Rightarrow I(g) = s(l^O) \cap s(l^T) = \{n\#04030305\}$

Como el conjunto de intersección está formado por un único synset, la etiqueta de sentido que se le asigna a la palabra $F_1^{12}(O) =$ *derecho* es la glosa correspondiente a dicho synset.

Sentido asignado: $n\#04030305$

Glosa: *an abstract idea of that which is due to a person or governmental body by law or tradition or nature: "they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness"; "Certain rights can never be granted to the government but must be kept in the hands of the people"- Eleanor Roosevelt; "it is his right to say what he pleases"*

- Caso en el que la intersección de synsets es de varios elementos

$< F_1^{19}(O), F_1^{19}(T) > = <$ *deben*, *dovrebbero* $>$

$g = (l^O, l^T)$ donde:

- $l^o =$ { *deben* [Verb] }

- $l^T =$ { *dovrebbero* [Verb] }

$s(l^O)$	$s(l^T)$
$-\ v\#01857688$	$-v\#01857688$
$-\ v\#01542552$	$-v\#01857799$
$-\ v\#01869087$	$-v\#01858275$
$-\ v\#01542388$	$-v\#01858069$
	$-v\#01862782$
	$-v\#01542388$

$\Rightarrow I(g) = s(l^O) \cap s(l^T) = \{v\#01857688, v\#01542388\}$

Como el conjunto de intersección está formado por varios synsets, se utiliza el método de back-off para elegir la etiqueta de sentido. En este caso se utilizó back-off primer sentido. Por tanto, la glosa que se le asigna a la palabra $F_1^{19}(O) =$ *deben* es:

Sentido asignado: $v\#01857688$

Glosa: *be obliged, required, or forced to*

- Caso en el que la intersección de synsets es vacía

Para el último caso, donde la intersección de synsets resulta vacía, se utiliza el mismo ejemplo con el par de idiomas español / inglés.

"*Todos los seres humanos nacen libres e iguales en dignidad y derecho. Dotados como están de razón y conciencia deben comportarse fraternalmente los unos con los otros.*"

"*All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.*"

$< F_1^{20}(O), F_1^{18}(T) > = <$ *comportarse*, *act* $>$

$g = (l^O, l^T)$ donde:

- $l^o =$ { *comportar* [Verb] }

- $l^T =$ { *act* [Sust], *act* [Verb] }

$s(l^O)$	$s(l^T)$
$-$ v#01798530	$-v$#01612822
$-$ v#01798202	$-v$#00007023
	$-v$#01177316
	$-v$#00741672
	$-v$#00008958
	$-v$#01879916
	$-v$#01721982
	$-v$#01178748
	$-v$#01177668

$\Rightarrow I(g) = s(l^O) \cap s(l^T) = \{\emptyset\}$

Como el conjunto de intersección es vacío, la etiqueta de sentido que se le asigna a la palabra $F_1^{20}(O) =$ *comportarse* dependerá del idioma de soporte y back-off. Considerando este último como primer sentido y el inglés como idioma de soporte:

Sentido asignado: v#01612822

Glosa: *perform an action; "think before you act"; "We must move quickly"*

Capítulo 5

Resultados

5.1. Composición de las redes de palabras

Todas las redes de palabras que forman parte de MultiWordNet, poseen una estructura similar. En ellas se almacena información de los lemas que constituyen el lenguaje y los synsets asociados a los mismos. Además, la naturaleza del lexicón permite determinar las brechas léxicas entre los idiomas.

El Cuadro 5.1 muestra algunas cifras relacionadas con la composición de cada una de las WordNets empleadas en el trabajo de tesis.

Propiedad	Español	Italiano	Inglés	Hebreo
Lemas	60037	44416	123781	15041
Sustantivos	47732	34224	96324	10991
Adjetivos	9053	4905	20171	2273
Verbos	5297	4896	10329	1614
Synsets	105516	36270	102105	5922
Glosas	7717	2732	99649	4103
Brechas léxicas	-	986	465	329

Cuadro 5.1: Estructura de los WordNets en MWN

La estructura de la red para el español no permite extraer información de las brechas léxicas.

La misma información de la cantidad de lemas por categoría gramatical se muestra en la Figura 5.1. Con ello se tiene una perspectiva más clara de la completitud de las redes de palabras, comparando los cuatro idiomas involucrados en el estudio.

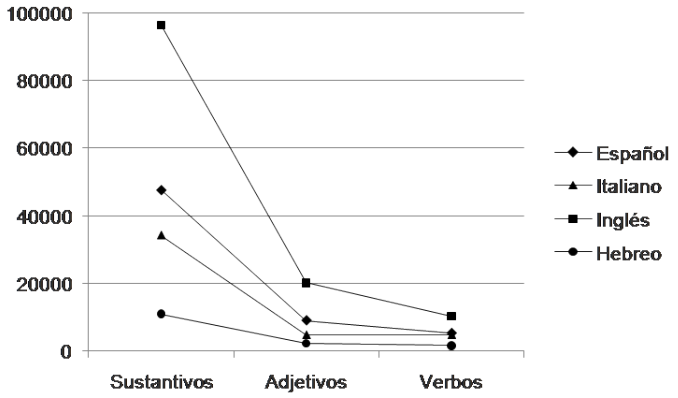


Figura 5.1: Distribución de los lemas según la categoría gramatical en MWN

5.1.1. Estadísticas de sentidos en MultiWordNet

Se ha efectuado un análisis con objeto de determinar cuál de los idiomas involucrados en el estudio, es más ambiguo. Para ello se han calculado los promedios de sentidos por palabra almacenada en su red correspondiente. El análisis comprende la observación por categoría gramatical.

	Español	Italiano	Inglés	Hebreo
Sentidos asignados	93425	64384	171018	18667
Sentidos sustantivos	63028	48376	119050	13792
Sentidos adjetivos	17999	6228	29883	2675
Sentidos verbos	12398	9780	22085	2200
Promedio de sentidos	**1.55612**	1.50009	1.42733	1.25484
Promedio sustantivos	1.32046	1.41351	1.23591	1.35217
Promedio adjetivos	1.98818	1.26972	1.48148	1.87094
Promedio verbos	2.34057	1.99755	2.13815	2.49632

Cuadro 5.2: Estadísticas de sentidos en MWN

La primera fila en el Cuadro 5.2 muestra la cantidad de sentidos que han sido atribuidos a los lemas de entrada en las tablas de índice. De esta forma, si un mismo sentido ha sido asignado a dos palabras diferentes, se toma dos veces en cuenta en este conteo. Por tanto, no se está haciendo referencia al total de synsets del idioma, sino a la cantidad de veces que éstos han sido asignados. Las siguientes tres filas, se corresponden con la división de esta cantidad por POS.

Con la medida de asignación y el número de lemas, se puede determinar el promedio de sentidos. Mientras mayor sea el resultado obtenido, mayor será el grado de polisemia

en el lenguaje. Por tanto el español es, de los cuatro idiomas estudiados, el más ambiguo.

Tomando únicamente en cuenta los valores obtenidos en el cuadro anterior, el hebreo será el mejor desambiguador, puesto que posee el menor promedio. Así, sus formas de palabra deberán ser las más adecuadas sirviendo como fuentes de distinción de sentidos en traducciones del español. Sin embargo, en este punto es necesario considerar que aspectos como la incompletidud de los WordNets influyen en los resultados.

Si se analizan los promedios por categoría gramatical, el italiano es el más ambiguo en lo que a sustantivos se refiere, el español en adjetivos y el hebreo en verbos. Esta información es útil, cuando se conoce el POS de las palabras polisémicas, pues en base a ello se puede elegir el idioma más conveniente en la tarea de desambiguación.

En la Figura 5.2 se presentan los valores del cuadro anterior con formato de gráfica de barras.

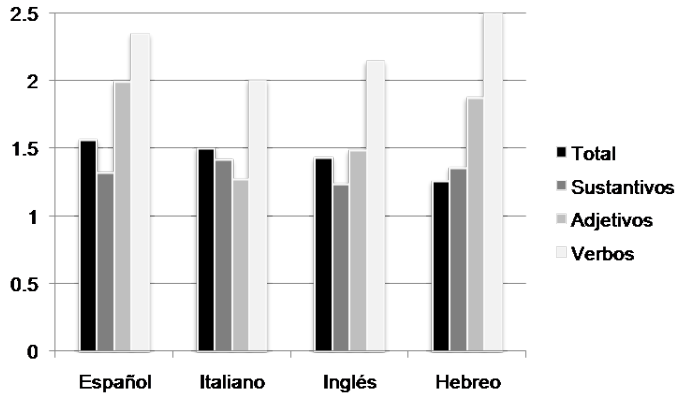


Figura 5.2: Promedio de sentidos asociados en MWN

5.1.2. Estadísticas de dominios en MultiWordNet

Un análisis similar se ha efectuado, pero ahora tomando como referencia los dominios semánticos.

La primera fila en el Cuadro 5.3 muestra la cantidad de lemas que no tienen asociado ningún dominio semántico. Esto se debe a que existen lemas cuyos sentidos han sido definidos mediante synsets que no existen en PWN. Por ejemplo, la red de palabras del español incluye synsets cuyos offset tienen la forma $5000XXXX$, como: v#50000002 - *cariarse*, v#50000003 - *achaflanar*, v#50000009 - *humar*, entro otros. En PWN no hay equivalentes, de ninguna categoría gramatical, que posean esta estructura.

	Español	Italiano	Inglés	Hebreo
Lemas sin dominio asociado	7556	0	0	925
Promedio de etiquetas de dominio por lema	1.47785	1.32216	1.48059	1.25942
Lemas con dominio FACTOTUM	16365	31612	13426	5415
Dominio más ambiguo (cantidad de lemas)	*Factotum* *Biology* (10621)	*Factotum* *Biology* (5040)	*Biology* (31926)	*Factotum* *Biology* (682)
Dominio menos ambiguo -con 1 lema asociado-	*Sub*	*Veterinary* *Sub*	*Veterinary*	8 dominios: *Dance, Plastic_Arts, Betting, Table_Tennis,...*
Dominio(s) con ningún lema asociado	*Environments* *Veterinary*	*Environments*	*Environments*	22 dominios: *Environments, Veterinary, Philately, Psychoanalysis,...*

Cuadro 5.3: Estadísticas de dominios en MWN

Cada lema en MWN está relacionado con uno o varios sentidos representados por medio de los synsets. Cada sentido a su vez puede tener asociado uno o más dominios. Por ejemplo, *a*#00009940 - *moderado* tiene a *Psychological_Features* y *Quality* como dominios. En consecuencia, es posible determinar el promedio de etiquetas por lema, lo cual se indica en la segunda fila del cuadro.

En la sección 3.8 se mencionó que FACTOTUM se asignaba como etiqueta de dominio en aquellos casos que no hubiera otra más adecuada. Como implicación, se obtiene que esta etiqueta supera al resto en número. Por tanto, en la mayoría de los lenguajes, excepto para el inglés, es precisamente FACTOTUM el dominio más ambiguo. En dichos casos se ha incluido el segundo dominio de mayor nivel de ambigüedad para esta entrada del cuadro.

Finalmente, las últimas dos hileras contienen los dominios con uno o ningún lema asociado respectivamente. Los casos que poseen dos o más rótulos, coinciden en cantidad. La lista completa para el idioma hebreo puede consultarse en el Apéndice A.

5.2. Evaluación del algoritmo de alineación

Para realizar los experimentos se emplearon fragmentos elegidos aleatoriamente de la novela *Don Quijote de la Mancha* en sus versiones paralelas *español / inglés* y *español / italiano*. Los fragmentos usados en ambos pares de idiomas fueron los mismos. Los conjuntos de prueba estuvieron formados por 23 sentencias alineadas. El texto en

español está constituido por 828 palabras.

El Cuadro 5.4 muestra la composición de cada uno de los fragmentos de los textos meta empleados como corpus de prueba.

	Italiano	Inglés
# de palabras	866	796
Promedio de sentidos	4.4276	6.8653

Cuadro 5.4: Composición de los textos meta

Para producir el gold standard de los pares de alineación, dos anotadores fueron instruidos para asignar un equivalente a todas las palabras no funcionales, con procedimientos específicos de cuándo asignar un equivalente nulo. No se incluyeron etiquetas de probabilidades. En caso de que hubiera un desacuerdo para un par específico, un tercer anotador definía el correcto.

El Cuadro 5.5 muestra la cantidad de pares de alineación determinados por los anotadores (gold standard). El topline indica el número máximo de equivalencias que podrían ser extraídas por el sistema, considerando relaciones 1:1 exclusivamente, la incompletitud de las redes de palabras que conforman MWN y las brechas léxicas (gaps) de los lenguajes implicados.

Versión	Gold standard	Topline
español / italiano	333	277
español / inglés	389	329

Cuadro 5.5: Pares de alineación sugeridos por los anotadores y topline del sistema

5.2.1. Medidas de evaluación

Realizamos la evaluación respecto a tres diferentes medidas: precisión, recall y F-measure. La *precisión* es calculada como el número de equivalencias extraídas correctamente entre el número de equivalencias sugeridas por el sistema. El *recall* se corresponde al número de equivalencias extraídas correctamente entre el número de equivalencias sugeridas por los anotadores.

Sin embargo, ni la precisión ni el recall pueden, de manera independiente, determinar la calidad del emparejamiento. Por lo general, la maximización del recall compromete la precisión y viceversa [90]. Por tanto, se requiere una medida que combine ambos parámetros. La *F-measure* posee esta característica y se determina como:

$$F - measure = \frac{2 * precisión * recall}{precisión + recall}$$

En este caso, la precisión y el recall poseen el mismo peso, pero la fórmula anterior puede ajustarse si se desea otorgar mayor peso a algunas de estas dos medidas.

5.2.2. Resultados

Los resultados fueron obtenidos con una iteración del algoritmo. Además, fue necesario establecer un umbral de relación para cada una de las medidas de similitud. De esta forma se evitan despropósitos lingüísticos a través de un juicio booleano (relacionados / desvinculados), pues las medidas devuelven un valor numérico de semejanza, pero no establecen aquellos valores que cuentan como cercanos. Los umbrales en cada método fueron los siguientes: $LCH \rightarrow 2$, $HSO \rightarrow 2$, $Edge \rightarrow 0{,}2$ y $Random \rightarrow 0{,}2$. Así, un valor de 0.166667 para el par <*astillero, villaje*> con el método *Edge* es descartado y se asigna similitud 0, o sea, sólo se toman valores de similitud mayores a 0.2 como fue señalado en el umbral. Esto evita la extracción de equivalencias erróneas y por tanto, tiene influencia en la precisión.

El Cuadro 5.6 muestra la cantidad de pares extraídos por el sistema (total), la cantidad de éstos que son correctos y los valores de las tres medidas de evaluación empleadas para cada tipo de alineación[1]. La entrada *Sólo coincidentes*, se refiere al caso donde no se ha aplicado ninguna medida de similitud semántica para efectuar el alineado cuando no hay synsets comunes entre la palabra $F_k^i(O)$ y sus traducciones potenciales.

Los mismos valores han sido graficados en la Figura 5.3.

Gold standard	Método similitud	Total	Correctos	Precisión	Recall	F-measure
389	*Sólo Coincidentes*	228	206	90.35 %	52.96 %	66.78 %
	LCH	289	220	76.12 %	56.56 %	64.88 %
	HSO	321	226	70.40 %	58.10 %	63.66 %
	Edge	307	221	71.99 %	56.81 %	63.51 %
	Random	377	203	53.85 %	52.19 %	53.01 %

Cuadro 5.6: Evaluación de los resultados en el corpus de prueba

Sin embargo, si se toman los valores del topline en vez de los del gold standard (descartando las equivalencias establecidas por pertenencia a frases, incompletitud de

[1]En los apéndices B, C y D se muestran relaciones detalladas por cada una de las oraciones que conforman el corpus

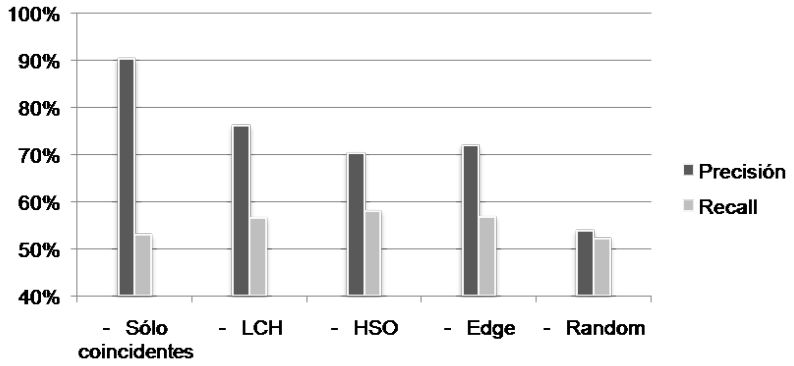


Figura 5.3: Valores de precisión y recall por método de similitud

MWN y gaps) los resultados de las tres medidas de evaluación sufren algunas variaciones, como se observa en el Cuadro 5.7 y las Figuras 5.4 y 5.5.

Topline	Método similitud	Total	Correctos	Precisión	Recall	F-measure
329	*Sólo coincidentes*	221	206	93.21 %	62.61 %	74.91 %
	LCH	271	220	81.18 %	66.87 %	73.33 %
	HSO	299	226	75.59 %	68.69 %	71.98 %
	Edge	288	221	76.74 %	67.17 %	71.64 %
	Random	346	203	58.67 %	61.70 %	60.15 %

Cuadro 5.7: Evaluación de los resultados considerando el topline del sistema

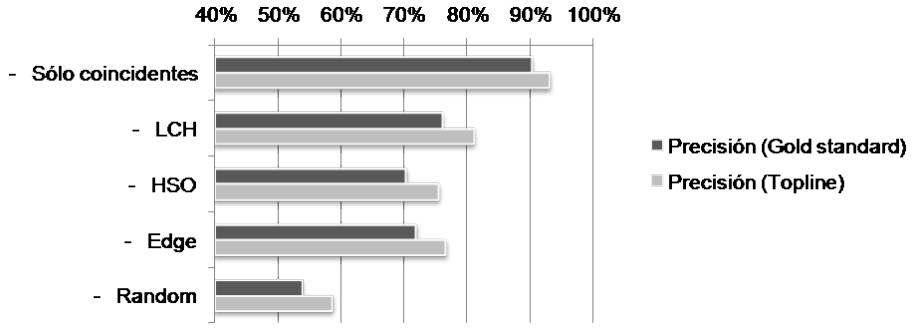


Figura 5.4: Comparación de los valores de precisión con base en el gold standard y el topline

La precisión del método de *sólo coincidentes* tiene que ver con la asertividad de las alineaciones de las redes de palabras que conforman MWN. Sin embargo, los valores de recall obtenidos para este método son pobres en ambos casos (usando el gold

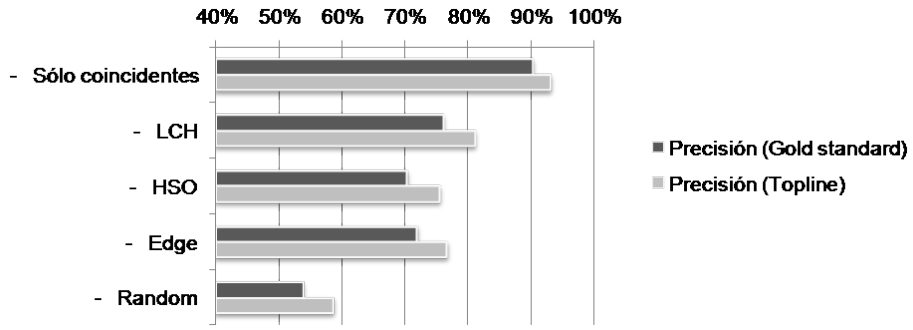


Figura 5.5: Comparación de los valores de recall con base en el gold standard y el topline

standard y el topline). Esto se debe a dos razones fundamentales: (1) el hecho de que en MWN se almacenan elementos léxicos simples y por tanto es imposible asignar un sentido específico a las expresiones multipalabra y (2) la incompletitud de las redes (véase Cuadro y Figura 5.1 para observar la desproporción entre el inglés y el resto de los idiomas).

Por otra parte, la medida LCH es comparable en términos de F-measure con el método de *sólo coincidencia*, a pesar de la diferencia significativa si se analizan los valores de precisión. Los métodos de similitud HSO y $Edge$, aumentan el recall, aunque no de manera significativa (entre 4 y 6 % aproximadamente), pero lo hacen a costa de una disminución considerable de la precisión (entre 16 y 18 %). En este sentido, es lógico que el método HSO posea el mayor valor de recall, puesto que su definición toma en cuenta cuatro tipos de relaciones semánticas (hiperonimia, meronimia, hiponimia y holonimia) y una léxica (antonimia) de WordNet, en tanto que LCH sólo se basa en la hiperonimia.

La baja precisión del método *Random*, con respecto al resto de las medidas, está relacionada con el simple marcaje basado en la aleatoriedad de la asignación del valor de similitud.

Si se colocan en el mismo gráfico la precisión y el recall, como se muestra en la Figura 5.6, se puede advertir que la precisión disminuye a medida que aumenta el recall, independientemente de la base (gold standard o topline). Para cuantificar la relación que existe entre estas medidas se ha determinado el coeficiente de correlación producto o momento de Pearson, r, un índice adimensional acotado entre -1 y 1 que refleja el grado de dependencia lineal entre dos conjuntos de datos.

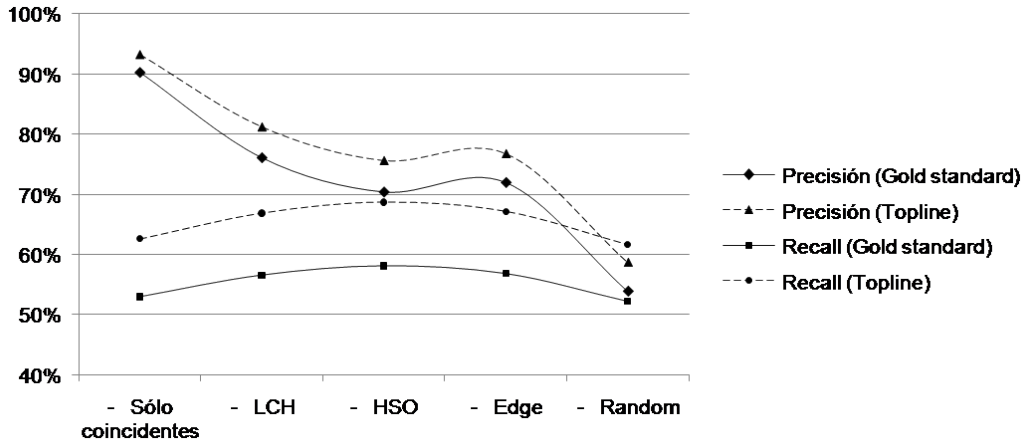


Figura 5.6: Comparación de los valores de precisión y recall con base en el gold standard y el topline

La fórmula determinar el coeficiente de correlación es:

$$r = \frac{\sum_{i=1}^{n}(x-\overline{x})^2(y-\overline{y})^2}{\sqrt{\sum_{i=1}^{n}(x-\overline{x})^2 \sum_{i=1}^{n}(y-\overline{y})^2}}$$

donde:

- n es el número de métodos de similitud empleados
- x el valor de precisión para el método i (conjunto de valores independientes)
- y el valor de recall para el método i (conjunto de valores dependientes)
- $\overline{x}$ e $\overline{y}$ son las medias de muestra promedio para el conjunto de valores independientes y dependientes respectivamente.

El resultado de aplicar el coeficiente de correlación a los valores de precisión y recall obtenidos en los Cuadros 5.6 y 5.7 son r = -0.9838 y r = -0.9787 respectivamente, lo que indica que estas medidas poseen un alto vínculo y son inversamente dependientes.

5.3. Evaluación del algoritmo de desambiguación

Para la evaluación del sistema de desambiguación propuesto se requiere de un corpus que haya sido previamente anotado de sentidos en forma manual y que sirva como gold-standard. Entre los corpus más difundidos para entrenamiento y evaluación, están

los creados para el concurso SENSEVAL (acrónimo de *Sense Evaluation*).

El primer certamen SENSEVAL tuvo lugar en 1998 y las lenguas participantes fueron el inglés, el francés y el italiano. La metodología desarrollada permitía evaluar los sistemas de desambiguación exclusivamente en la determinación automática del sentido de una única palabra en un contexto determinado.

Con SENSEVAL-2, las lenguas participantes se incrementaron hasta un total de 12 (inglés, francés, italiano, español, vasco, danés, sueco, holandés, estonio, checo, chino y japonés) al igual que el tipo de tareas posibles en las que podían concursar los competidores.

En este sentido, se diseñaron tres tipos de tareas que básicamente se diferenciaban entre sí por el número de palabras que los programas tenían que desambiguar [91].

1. Tarea basada en una muestra léxica (lexical sample task)
2. Tarea léxica completa (all-words task)
3. Tarea de traducción (translation task)

La *tarea de muestra léxica* tiene como objetivo evaluar únicamente una palabra por frase. En la *tarea léxica completa*, los sistemas deben etiquetar todas las palabras con contenido semántico de un texto, exceptuando las funcionales (conjunciones, preposiciones, artículos, etc.). Finalmente, la *tarea de traducción*, es de hecho un subtipo de muestra léxica porque sólo hay que desambiguar una única palabra; la diferencia es que el sentido de la palabra se define de acuerdo a su traducción. En esta última sólo ha participado la lengua japonesa.

Para poder llevar a cabo dichas tareas es necesario disponer de dos tipos básicos de datos:

1. Un diccionario (o léxico)
2. Un corpus etiquetado manualmente

Todos los sistemas que concursan tienen que analizar semánticamente el mismo corpus y, en consecuencia, es necesario que cada lengua participante disponga de este corpus y de un diccionario en el que se incluyan los distintos significados de las palabras que se deben desambiguar.

El tercer y último certamen, SENSEVAL-3, se realizó en el 2004 e incluyó las siguientes tareas: tarea léxica completa (inglés e italiano), tarea de muestra léxica (vasco, catalán, chino, inglés, italiano, rumano, español y sueco), adquisición automática de subcategorización para verbos ingleses, muestra léxica multilingüe (inglés-francés e inglés-hindi), WSD de las glosas de WordNet, identificación de roles semánticos (inglés y sueco) e identificación de formas lógicas en inglés.

5.3.1. Muestra léxica para español en SENSEVAL-3

Como en todas las tareas en SENSEVAL, para el español fueron necesarios dos recursos lingüísticos: el diccionario (MiniDir-2.1) y el corpus anotado (MiniCors).

MiniDir-2.1 contiene 46 palabras de tres categorías sintácticas: 21 sustantivos, 7 adjetivos y 18 verbos. La selección de éstas se hizo tratando de mantener el núcleo de palabras de Senseval-2 y con el fin de compartir alrededor de 10 palabras de la meta con otras tareas de muestra léxica (vasco, catalán, inglés, italiano y rumano) [92]. El promedio de sentidos por palabra es de 5.33, siendo de 4.52 sentidos para sustantivos, 6.78 para verbos y 4 para adjetivos.

La Figura 5.7 muestra un ejemplo de una entrada léxica en MiniDir-2.1.

```
<lexelt item="actuar" pos="VM">

<sense id="actuar.1" definition="Realizar actos, ejercer funciones propias de su cargo o
naturaleza" used="yes">
<example text="hay que actuar"/>
<example text="la fiscalía actuó contra el terrorista"/>
<example text="desde la muerte de su padre, el hermano mayor actuó como cabeza de
familia"/>
<example text="actuar legalmente"/>
<example text="los jugadores suramericanos que actúan en España"/>
<collocation text="actuar en defensa propia"/>
<synset wordnet="1.5" id="01341700v"/>
<synset wordnet="1.5" id="00618376v"/>
</sense>
```

Figura 5.7: Entrada léxica en MiniDir-2.1

Cada entrada contiene información general y una relación de sentidos. Cada sentido contiene un identificador, su definición, una lista de ejemplos, de sinónimos (posiblemente vacía), de posibles colocaciones (posiblemente vacía) y el conjunto de synstes equivalentes en PWN. Este conjunto también puede estar vacío porque MiniDir-2.1 es de grano más grueso que PWN. Por lo tanto, típicamente un sentido en MiniDir-2.1 corresponde a varios sentidos en PWN. Sin embargo, podría haber casos en los que un sentido en MiniDir-2.1 no tenga equivalentes en PWN. Sentidos descartados por baja frecuencia tienen un atributo: *used* = "no".

El Apéndice E muestra el conjunto de las 46 palabras polisémicas seleccionadas. También se agrega información de su POS, número de sentidos y cantidad de ejemplos de entrenamiento, sin etiquetar y de prueba.

Por otra parte, *MiniCors* es un corpus semánticamente etiquetado de acuerdo a los parámetros de las tareas de muestra léxica en SENSEVAL. Las etiquetas son tomadas

del repositorio de sentidos MiniDir-2.1.

El corpus MiniCors está formado por 12625 ejemplos etiquetados (35875 oraciones y 1506233 palabras). El contexto considerado para cada ejemplo incluye la oración meta, además de una anterior y otra siguiente. Todos los ejemplos fueron extraídos de un corpus del año 2000 de la agencia de noticias española EFE, que incluye 289066 noticias (2814291 de oraciones y 95344946 palabras).

Cada participante dispone de un conjunto relativamente pequeño de ejemplos etiquetados por palabra (2 tercios de 75 + 15 * # sentidos) y un conjunto comparativamente amplio de ejemplos sin etiquetar. El conjunto de prueba se compone de un tercio de los 75 + 15 * # sentidos [93]. En cada caso, se facilitan dos versiones de los archivos: una con inclusión de los lemas y etiquetas de categoría gramatical y la otra sin ninguna información morfológica.

5.3.2. Preparación de los textos paralelos

Como se mencionó en la subsección 3.9.1, se utilizó traducción artificial para obtener los textos paralelos de pruebas, por tres razones principalmente:

- La facilidad para su creación
- La necesidad de contar con un objetivo de comparación
- La evaluación de experimentos controlados sin restricciones en el lenguaje meta

Para la traducción del corpus de muestra léxica del español en SENSEVAL-3, se empleó el traductor comercial *LEC Power Translator 11*. Con ello se produjeron los textos en italiano e inglés.

Todos los resultados que se presentan en las siguientes secciones se obtuvieron a partir del archivo de ejemplos de entrenamiento y sus traducciones, en la versión que no incluye información de los lemas ni etiquetas de POS. El contexto de cada ejemplo consta de tres oraciones, que han sido marcadas con las etiquetas $< previous >$, $< target >$ y $< following >$. La Figura 5.8 muestra el formato.

<instance id="arte.n.217" docsrc="efe_17115_2000/09/21">
<answer instance="arte.n.217" senseid="arte.1"/>
<context>
<previous>
Más de cuarenta obras de grandes artistas españoles del siglo XX, desde Picasso o Miró a Julio González, se han reunido en el Museo Provincial de Teruel para la exposición "Viaje a la semilla", un intento por mostrar la relación de las vanguardias con la arqueología.
</previous>
<target>
La exposición, que se inaugura hoy y permanecerá abierta hasta el próximo 19 de noviembre, es un recorrido de las relaciones del <head>arte</head> contemporáneo, desde las vanguardias históricas hasta los artistas contemporáneos, con el pasado remoto.
</target>
<following>
El título de la muestra, "Viaje a la semilla", tomado de un cuento del escritor cubano Alejo Carpentier, quiere ser una metáfora de los referentes que las vanguardias encontraron en el arte de los pueblos primitivos africanos, en la Grecia arcaica o preclásica y, en el caso de España, en los íberos.
</following>
</context>
</instance>

Figura 5.8: Entrada léxica en el corpus de entrenamiento

El archivo contiene ejemplos manualmente anotados de sentidos para las 46 palabras polisémicas. Sin embargo, hay sentidos que no tienen equivalentes en WordNet, por tanto, las oraciones que contienen dichos significados han sido descartadas del corpus. El Cuadro 5.8 contiene los sentidos en cuestión. El número que acompaña a la palabra ambigua representa la posición del significado en el MiniDir-2.1.

Tras desechar los contextos con los sentidos especificados, el corpus resultante queda compuesto por 7079 entradas. La distribución por clase gramatical de dichas entradas se muestra en la Figura 5.9.

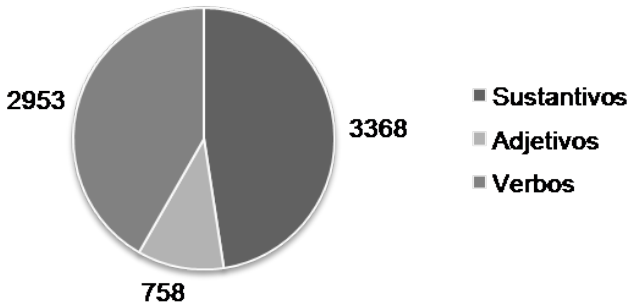


Figura 5.9: Distribución de las entradas según la categoría gramatical en el corpus

Sentidos descartados	
arte.3	pasaje.4
arte.4	perder.1
autoridad.4	perder.2
bajar.2	perder.7
bajar.5	perder.8
banda.1	perder.9
canal.6	perder.10
circuito.5	perder.11
columna.2	popular.3
columna.5	subir.2
corazón.6	tabla.4
coronar.6	tabla.6
gracia.2	tocar.8
grano.4	tratar.8
jugar.5	tratar.12
mina.4	usar.3
natural.2	verde.5

Cuadro 5.8: Sentidos en MiniDir-2.1 sin equivalentes en PWN

5.3.3. Resultados de la alineación

Durante el proceso de alineación los contextos fueron reducidos a la oración $< target >$, pues sólo interesa determinar la palabra, en el idioma destino, que se corresponde con la palabra ambigua. Las sentencias $< previous >$, $< following >$ continúan descartadas en el método de desambiguación propuesto, ya que éste se basa en el conjunto de alineación obtenido, de esta forma el contexto resulta indiferente.

El Cuadro 5.9 presenta los resultados obtenidos después de aplicar el método de alineación sobre las 7079 apariciones de las palabras polisémicas. La primera entrada del cuadro indica que las palabras del texto en español se relacionaron en su totalidad con una forma destino, tanto para el italiano como para el inglés.

Las siguientes filas señalan la manera en que se produjo la alineación: por intersección de synsets o por aplicación de los pesos de dominios. Este mismo análisis se dividió luego por clase gramatical, donde el primer valor se corresponde a la aproximación de intersección y el segundo a la de pesos por categoría semántica.

La última hilera del cuadro muestra la precisión de la alineación. Los resultados obtenidos en este concepto indican que en el inglés se logró mayor número de correspondencias correctas. Esto se explica con la relación entre las aproximaciones empleadas para ejecutar la alineación de ambos idiomas. Por supuesto el enfoque de intersección es más exacto y por tanto, las correspondencias realizadas bajo el mismo tienen mayor

	Italiano (%)	Inglés (%)
Recall	100	100
Por intersección de synsets	88.668	93.728
Por aplicación de pesos de dominio	11.332	6.272
Sustantivos	97.312 / 2.688	97.922 / 2.078
Adjetivos	92.421 / 7.579	98.813 / 1.187
Verbos	78.799 / 21.201	87.409 / 15.591
Precisión	88.563	93.403

Cuadro 5.9: Porcentajes de los resultados de alineación obtenidos

probabilidad de ser las apropiadas. Para el inglés se utilizó dicho enfoque 5.06 % más que para el italiano. Es decir, hubo mayor cantidad de sentidos de las palabras de traducción del inglés que coincidieron con los sentidos de las palabras polisémicas en español y por tanto, las cardinalidades de los grupos de intersección son superiores.

Con esto se deduce que la preservación de los niveles de ambigüedad en el idioma destino, favorece la correctitud de la alineación. Sin embargo, las cifras que se presentaron en el Cuadro 5.2 revelan que el italiano es más ambiguo que el inglés, por lo que los resultados pudieran parecer contradictorios. El problema en este punto consiste en que dichos valores fueron determinados para la totalidad de los lemas que se encuentran en las wordnets de los idiomas correspondientes y la alineación se está ejecutando sobre un grupo reducido de palabras (46). Entonces, para medir con exactitud el grado de polisemia, se ha determinado el promedio de sentidos de todas las posibles traducciones de estos 46 lemas. Los resultados se presentan en el Cuadro 5.10.

	Italiano	Inglés
Total	5.2692	8.6696
Sustantivos	5.7073	6.3589
Adjetivos	5.1539	10.625
Verbos	10.066	9.3768

Cuadro 5.10: Promedio de sentidos en las traducciones

Ahora, es posible advertir que para las traducciones que devolvió el LEC Power Translator, el inglés es más polisémico que el italiano, con excepción en los verbos.

5.3.4. Reducción de la polisemia

Después de ejecutar la alineación sobre el corpus, se tienen los grupos de correspondencias entre idiomas. Estas asociaciones constituyen la única entrada en el módulo de desambiguación, pues el contexto se descarta.

Para no afectar los resultados de la desambiguación, los casos de alineación incorrecta fueron eliminados. El Cuadro 5.11 muestra estadísticas con relación a los lemas, deducidas a partir de las salidas generadas durante el proceso de asignación de etiquetas de sentidos.

		Español - Italiano		Español - Inglés	
Total	Promedio de lemas	1.2099	1.3426	1.2495	1.8390
	Lemas descartados	0.1539	0.2838	0.1106	0.6954
	Lemas en la intersección	1.0410		1.1294	
Sustantivos	Promedio de lemas	1.1771	1.2410	1.2497	1.7829
	Lemas descartados	0.1014	0.1638	0.0845	0.6171
	Lemas en la intersección	1.0729		1.1640	
Adjetivos	Promedio de lemas	1.8984	2.0181	1.7563	1.7124
	Lemas descartados	0.7585	0.8894	0.3462	0.3049
	Lemas en la intersección	1.1151		1.4048	
Verbos	Promedio de lemas	1.0668	1.2944	1.1021	1.9487
	Lemas descartados	0.0588	0.2771	0.0755	0.9098
	Lemas en la intersección	0.9808		1.0048	

Cuadro 5.11: Desambiguación de los lemas

A partir de los valores anteriores se desprenden los porcentajes en que se reduce la ambigüedad categorial. Hay que recordar que el apoyo con la traducción carece de direccionalidad, por tanto se presentan las disminuciones en ambos sentidos. Por ejemplo, en el Cuadro 5.12 se observa que para el total de palabras, el español reduce su polisemia en pertenencia del POS 12.72 % auxiliado por el italiano y el italiano a su vez 21.138 % con el español.

	Español - Italiano		Español - Inglés	
Total	12.720 %	21.138 %	8.852 %	37.814 %
Sustantivos	8.614 %	13.199 %	6.762 %	34.612 %
Adjetivos	39.955 %	44.071 %	19.712 %	17.805 %
Verbos	5.512 %	21.408 %	6.851 %	46.688 %

Cuadro 5.12: Porcentajes de reducción de ambigüedad categorial

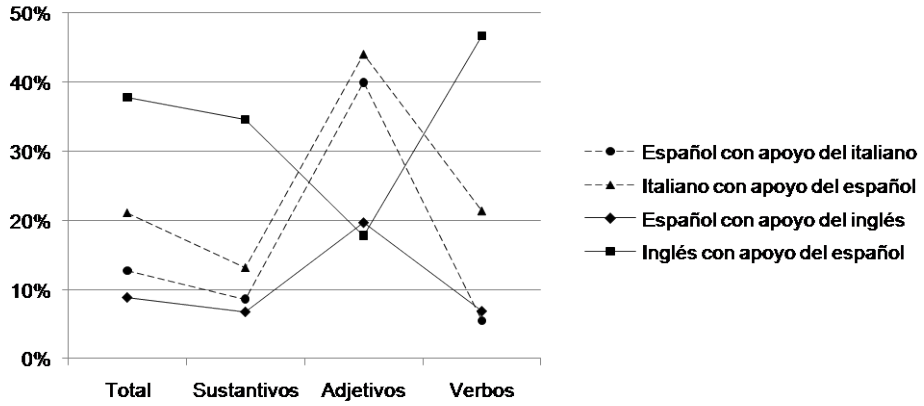


Figura 5.10: Gráficos de reducción de ambigüedad categorial

La misma exploración se ha efectuado para obtener estadísticas respecto a los sentidos. El Cuadro 5.13 muestra los resultados obtenidos.

		Español - Italiano		Español - Inglés	
Total	Promedio de synstes	7.1963	7.5890	7.4084	11.6978
	Synstes descartados	4.3958	4.8416	3.3046	7.6047
	Synstes en la intersección	2.7019		4.0301	
Sustantivos	Promedio de synstes	6.4229	7.9576	6.8512	8.3540
	Synstes descartados	3.4948	5.0243	2.9558	4.4634
	Synstes en la intersección	2.9086		3.8875	
Adjetivos	Promedio de synstes	10.9323	7.8014	10.8322	11.6498
	Synstes descartados	7.4718	4.3747	4.0200	4.8469
	Synstes en la intersección	3.3837		6.7909	
Verbos	Promedio de synstes	7.1905	7.0192	7.1288	15.4777
	Synstes descartados	4.7260	4.6871	3.5719	11.9400
	Synstes en la intersección	2.2610		3.3800	

Cuadro 5.13: Desambiguación de los synsets

De manera análoga, los porcentajes de reducción de polisemia semántica calculados con los valores precedentes, se presentan en el Cuadro 5.14.

	Español - Italiano		Español - Inglés	
Total	61.084 %	63.798 %	44.606 %	65.010 %
Sustantivos	54.412 %	63.138 %	43.143 %	53.428 %
Adjetivos	68.346 %	56.076 %	37.112 %	41.605 %
Verbos	65.726 %	66.775 %	50.105 %	77.143 %

Cuadro 5.14: Porcentajes de reducción de ambigüedad semántica

Si se analizan los resultados por categoría gramatical, se puede distinguir que en todos los casos la reducción de la polisemia está en correspondencia con el nivel de ambigüedad: mientras menor es el grado de ambigüedad del idioma de traducción, mayor será el porcentaje de reducción de la polisemia en el idioma origen (referirse al grado de ambigüedad en el Cuadro 5.10).

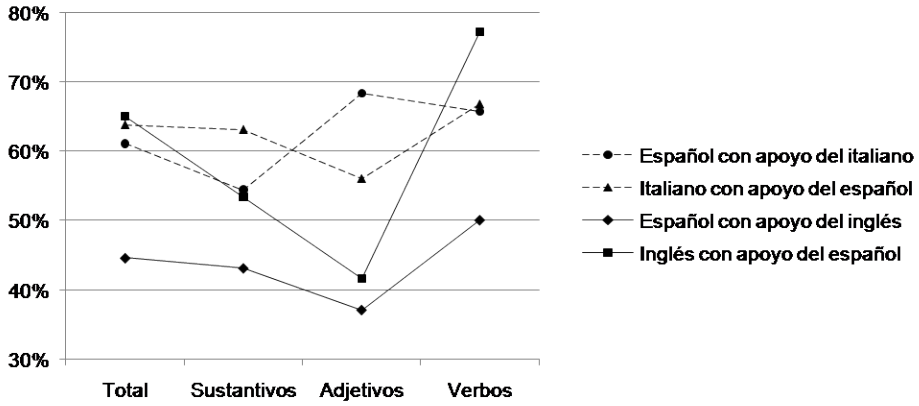


Figura 5.11: Gráficos de reducción de ambigüedad semántica

Capítulo 6

Conclusiones

Muchas de las aplicaciones relacionadas con procesamiento de lenguaje natural requieren previa desambiguación para conseguir resultados satisfactorios en las labores que desempeñan. Es por ello que la diferenciación de los sentidos ha sido considerada, desde hace muchos años, objeto de estudio como tarea independiente.

Desde sus inicios, se han utilizado diversas aproximaciones para la diferenciación de sentidos en palabras polisémicas. El método de desambiguación propuesto combina el enfoque supervisado y no supervisado, empleando textos paralelos alineados a nivel de palabra como corpus de entrada. En los ensayos realizados se usaron textos en español como idioma origen y traducciones de los mismos al italiano e inglés.

Como diccionario léxico multilingüe se usó MultiWordNet, aprovechando la ventaja que su concepción ofrece: las redes de palabras que lo conforman se encuentran alineadas.

El sistema implementado está dividido en tres módulos principales: (1) lematización, (2) alineación y (3) desambiguación. En la primera fase se obtienen los lemas de la palabra polisémica y de todas las formas que aparecen en el contexto del texto meta, pues es posible establecer relaciones con los lemas obtenidos de la aplicación de reglas morfológicas sobre las entradas. Durante esta etapa, no se utiliza heurística sintáctica alguna para resolver homonimia morfológica. En la fase de alineación se toma ventaja de la correspondencia que existe entre las wordnets de MWN y se incluye un análisis de pesos por dominio semántico o medidas de similitud, bajo el supuesto de que los textos se encuentran alineados a nivel de sentencias. Como resultado de este módulo se obtienen grupos de alineación que constituyen la entrada en el proceso de desambiguación. Finalmente, cada palabra polisémica en el texto origen es etiquetada con el sentido determinado por el desambiguador, basándose en intersecciones de synstes y categorías semánticas.

Para la evaluación de los resultados obtenidos se efectuaron experimentos en dos direcciones. La primera estuvo enfocada en el proceso de alineación de palabras no fun-

cionales, para la cual se emplearon fragmentos de la novela Don Quijote de la Mancha, sin etiquetado ni preprocesamiento previo. Los resultados fueron comparados con un gold standard, producido por anotadores humanos, a través de tres medidas de evaluación: precisión, recall y F-measure.

Para contar con un corpus de prueba estándar que incluyera anotación manual de sentidos, en el segundo acercamiento de los experimentos se hizo uso de traducción artificial para obtener los textos meta. Se partió del supuesto que aún las traducciones conseguidas mediante un traductor automático proporcionarían un alto porcentaje de lexicalizaciones que permitiesen descartar sentidos de las palabras polisémicas en el origen. Para ello se empleó el traductor comercial *LEC Power Translator 11* sobre el corpus de muestra léxica del español en SENSEVAL-3.

6.1. Discusión

De los estudios realizados en los corpus de prueba descritos anteriormente, se derivan dos resultados significativos:

1. ***Mientras mayor sea el grado de polisemia en el texto meta, mejores resultados se obtendrán en la alineación***. Esto está relacionado con la cardinalidad de las intersecciones de los synstes: Si ambos idiomas son altamente ambiguos, la preservación de la polisemia en la traducción, hará que el número de sentidos coincidentes sea mayor. Así, la seguridad de que la alineación es correcta, aumenta.

   La Figura 6.1 muestra la correspondencia descrita entre el promedio de sentidos y la precisión de la alineación. Los valores del gráfico para el corpus de *Don Quijote de la Mancha* se tomaron de los Cuadros X y 5.4[1], del italiano e inglés respectivamente; para el corpus de muestra léxica del español en SENSEVAL-3 de los Cuadros 5.9 y 5.10.

[1]Obtenidos empleando el método *Sólo coincidentes*

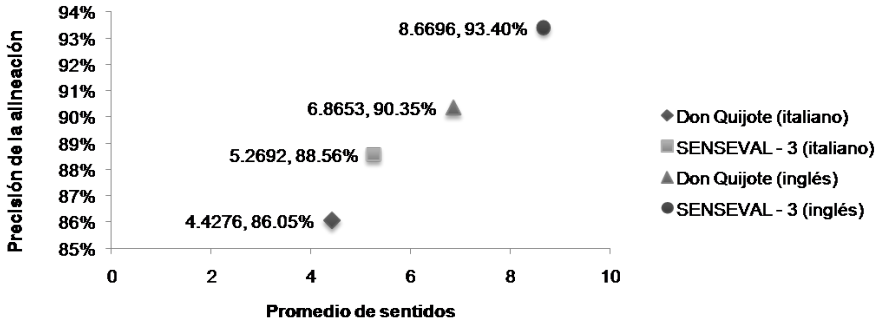


Figura 6.1: Relación entre el nivel de polisemia y la precisión de la alineación

2. En concordancia con la hipótesis inicial del estudio, se ha llegado a la conclusión de que ***mientras menor es el grado de polisemia en el texto destino, mayor información de diferenciación otorgará a la palabra en el idioma origen***.

   Las Figuras 6.2 y 6.3 muestran los resultados que se obtuvieron para el idioma español utilizando como traducción el italiano y el inglés. De ellos se puede observar la reducción en el promedio de sentidos para el total del corpus y por categoría gramatical. Los valores han sido obtenidos a partir de los datos del Cuadro 5.13.

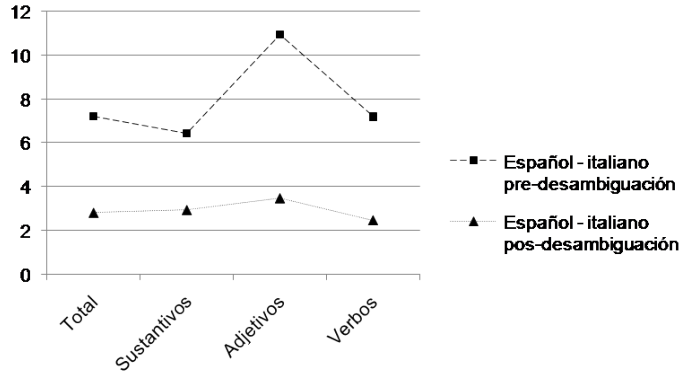


Figura 6.2: Reducción de la polisemia semántica del español con apoyo del italiano

Si se comparan las distancias entre marcadores (por categoría) en los gráficos de reducción, es posible distinguir que el italiano ha contribuido en mayor medida con la disminución de la ambigüedad del español, lo cual está en correspondencia con los niveles de polisemia en el cuadro 5.13 $\rightarrow$ $polisemia$(italiano) $<$ $polisemia$(inglés).

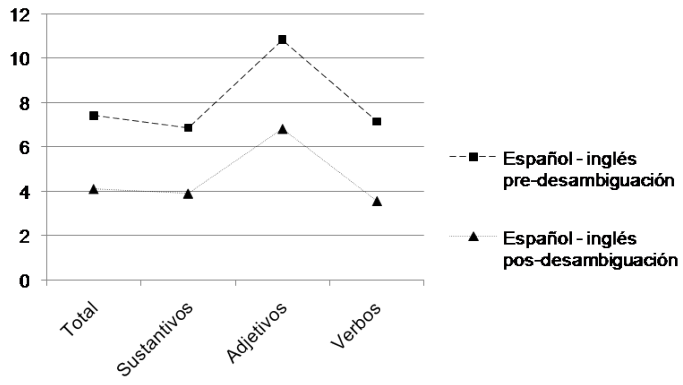


Figura 6.3: Reducción de la polisemia semántica del español con apoyo del inglés

6.2. Aportaciones

Método de alineación a nivel de palabras basado en la correspondencia entre redes de palabras

Los synsets de las redes de palabras que constituyen MultiWordNet han sido ubicados según la posición que ocupan los synsets de Princeton WordNet y respetando sus relaciones semánticas. De este modo, se dice que las redes se encuentran alineadas. Esta característica permite entonces que MWN pueda ser utilizado como recurso léxico en tareas de alineación de textos y específicamente de palabras con relación 1:1. El algoritmo propuesto no posee restricción de clase gramatical y con excepción del nulo, está enfocado en la extracción de equivalencias con limitación uno a uno. Todas las palabras de los contextos que participan del procesamiento, requieren ser lematizadas, pues los enlaces pueden producirse entre dos palabras como entradas básicas, dos lemas obtenidos de la aplicación de reglas morfológicas sobre las palabras, o la combinación de éstas.

El algoritmo se basa en el cálculo de la intersección de los conjuntos de synsets para todas las combinaciones de pares de palabras del contexto origen y el contexto meta (producto cartesiano de cada palabra $x \in$ *oración origen* y cada palabra $y \in$ *oración meta*). Una vez que se han determinado todas las cardinalidades.

Procedimiento de filtrado por dominios semánticos para reducir la polisemia

Se utiliza en aquellas palabras a las que no se les puedan asignar directamente sus pares de alineación con la intersección de los conjuntos de synsets. Los dominios son categorías semánticas que permiten distinguir los usos técnicos de las palabras y describir textos de acuerdo con temas generales caracterizados por un ámbito específico.

En el procedimiento propuesto se utiliza WordNet Domains como recurso léxico. Éste fue creado para incluir etiquetas de dominio en PWN. En resumen, el filtrado por dominios rechaza las traducciones potenciales con dominios diferentes a los de la palabra origen y brinda mayor peso a las que poseen dominios en común, con excepción del dominio FACTOTUM, pues la frecuencia de esta etiqueta impide su uso en la correcta determinación de relaciones semánticas.

Técnica para la aplicación de medidas de similitud en de pares palabras con idiomas diferentes

Otro recuso para la determinación de pares de alineación cuando no se cuenta con información concluyente de los grupos de synsets, son las medidas de similitud semántica. En este trabajo se emplearon cuatro de las medidas implementadas en el paquete *similarity* de WordNet: Leacock and Chodorow, Hirst and St-Onge, edge y random. Sin embargo, el formato de entrada de esta utilidad no permite formas de palabra en idiomas diferentes, pues fue creada para ser usada en PWN únicamente.

Se planteó entonces una técnica de obtención de traducciones basada nuevamente en la propiedad de alineación de las redes de MWN. El algoritmo consiste en la utilización de las primeras traducciones inglesas para cada synset de la palabra origen. Estos pares serán las entradas de *similarity.pl* y el mayor de los valores resultantes se incorpora a la matriz de alineación, de la cual se obtienen las parejas de equivalencia.

Traducción artificial en sustitución de textos paralelos

La necesidad de un objetivo de comparación para los idiomas implicados en la desambiguación, provocó la búsqueda de una alternativa que brindase información adicional para proporcionar diversas lexicalizaciones de las palabras polisémicas y obtuviera resultados comparables con textos paralelos. Fue por ello que se realizaron pruebas con traducciones artificiales, consiguiendo reducciones de hasta el 46.69 % para la ambigüedad categorial y del 77.14 % para la semántica.

Algoritmo para la adquisición automática de etiquetas de sentidos extraídas de la alineación de textos

La desambiguación se lleva a cabo a partir de los pares de alineación obtenidos en la fase de preprocesamiento. En el acercamiento planteado, el contexto para determinar el sentido se limita a la traducción alineada y específicamente a los synsets comunes. Si sólo existe un synset en común, éste se asigna como sentido correcto. Si no hay synsets compartidos, se utilizan nuevamente los dominios semánticos para disminuir el número de etiquetas potenciales. Por último si existen más de un synset en común se utiliza el método de back-off definido en la configuración del sistema para el idioma de soporte especificado.

6.3. Trabajos futuros

Como consecuencia del desarrollo y aplicación del sistema de desambiguación se han identificado ciertas necesidades que brindarían ventajas sobre los resultados alcanzados y servirán para darle continuidad a los mismos.

1. Utilizar alineadores completos como GIZA++ para obtener las correspondencias entre las palabras de los textos paralelos.
2. Incorporar la hipótesis de un dominio por discurso en el análisis de pesos por categoría semántica.
3. Implicar el contexto y utilizar alguna medida de similitud en el método de desambiguación en el caso donde el conjunto de intersección de synstes para los idiomas involucrados tenga cardinalidad mayor que uno.
4. Extender las nociones intuitivas del algoritmo, introduciendo un marco formal y conceptual para el uso de índices de interlingua (ILI). De esta forma se podrían realizar estudios con redes de palabras construidas bajo el enfoque de EuroWordNet.
5. Realizar experimentos con textos paralelos manuales y analizar el desempeño de las traducciones no automáticas en la reducción de la polisemia.
6. Determinar coeficientes de correlación para estudiar el impacto de diversos factores lingüísticos en el idioma meta.

6.4. Publicaciones

Relacionadas con el trabajo de tesis

- "Ambigüedad en Nombres Hispanos". *Revista Signos*, ISSN 0035-045, Vol. 42, No. 70 (2009) pp. 153-169
- "Medidas de Complejidad Cuantitativas para Sistemas Expertos Basados en Reglas". *Revista Iberoamericana de Inteligencia Artificial*, ISSN 1988-3064, Vol. 13, No. 43 (2009) pp. 16-31
- "Formal Grammar for Hispanic Named Entities Analysis". In *Proceedings of 10th International Conference, CICLing-2009*, Mexico City, Mexico. Lecture Notes in Computer Science N 5449, ISSN 0302-9743, ISBN 978-3-642-00381-3 (2009) pp. 183-194
- "Hybrid Algorithm for Word Alignment in Parallel Texts". In *Proceedings of 14th International Conference on Applications of Natural Language to Information Systems, NLDB-2009*, Saarbrücken, Germany.

- "Incorporating Linguistic Information to Statistical Word-Level Alignment". In *Proceedings of 14th Iberoamerican Congress on Pattern Recognition, CIARP 2009*, Jalisco, Mexico. Lecture Notes in Computer Science N 5856, ISBN 978-3-642-10267-7 (2009) pp. 387-394

Otras publicaciones

- "VBCH Herramienta para la verificación de Bases de Conocimiento". *En Memorias del IEEE 3er Congreso Internacional sobre Innovación y Desarrollo Tecnológico, CIINDET-2005*, Morelos, México (2005)

- "An Artificial Intelligence Based Model for Algebra Education". In *Proceedings of 5th International Conference on Digital Content, Multimedia Technology and its Applications, IDC-2009*, Seoul, Korea.

- "Pronóstico para la inyección de tenso-activos en pozos de petróleo a partir de una metodología que integra técnicas de Inteligencia Artificial y Minería de Datos". *Revista Interciencia*, ISSN 0378-1844, Vol. 34, No. 10 (2009) pp. 703-709

Bibliografía

[1] E. Agirre and P. Edmonds. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2006.

[2] P. Resnik and D. Yarowsky. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL-SIGLEX, Workshop Tagging Text with Lexical Semantics: Why, What, and How?*, pages 79–86, EUA, 1997.

[3] D. Tufiş, R. Ion, and N. Ide. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1312–1318, Geneva, 2004.

[4] H. Tou, B. Wang, and Y. Seng. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 455–462, Japón, 2003.

[5] A. Massimiliano, M. Ranieri, and C. Strapparava. Crossing parallel corpora and multilingual lexical databases for wsd. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 242–245, México, 2005.

[6] M. Diab. An unsupervised method for multifingual word sense tagging using parallel corpora: A preliminary investigation. In *Proceedings of the ACL-2000, Workshop on Word senses and multi-linguality*, pages 1–9, Hong Kong, 2000.

[7] H. Kaji. Word sense acquisition from bilingual comparable corpora. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 32–39, Canada, 2003.

[8] C. Leacock and M. Chodorow. Combining local context andwordnet similarity for word sense identification. *WordNet: An electronic Lexical Database*, pages 265–283, 1998.

[9] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic Lexical Database*, pages 305–332, 1998.

[10] R. Rada, H. Mili, E. Bicknell, and M. Bletner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[11] G. Rigau, J. Atserias, and E. Agirre. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting on Association for Computational Linguistics*, pages 48–55, España, 1997.

[12] C. Nevill and T. Bell. Compression of parallel texts. *Information Processing and Management: an International Journal*, 28(6):781–793, 1992.

[13] M. Kay and M. Roscheisen. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, 1993.

[14] R. Mihalcea and T. Pedersen. Advances in word sense disambiguation. Tutorial at IX Ibero-American Conference on Artificial Intelligence, 2004.

[15] N. Ide and J. Véronis. Word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1), 1998.

[16] A. Nikunen. Different approaches to word sense disambiguation. *Language technology and applications*, 2007.

[17] R. Mihalcea and D.I. Moldovan. An iterative approach to word sense disambiguation. In *Proceedings of the 13th International Florida Artificial Intelligence Research Society Conference*, pages 219–223, EUA, 2000.

[18] R. Mihalcea and D.I. Moldovan. Word sense disambiguation based on semantic density. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Canada, 1998.

[19] A. Molina, F. Pla, E. Segarra, and L. Moreno. Word sense disambiguation using statistical models and wordnet. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, España, 2002.

[20] D. Buscaldi, P. Rosso, and F. Masulli. Integrating conceptual density with wordnet domains and cald glosses for noun sense disambiguation. *Lecture Notes in Computer Science*, pages 183–194, 2004.

[21] M.R. Quillian. Semantic memory. *Semantic Information Processing*, pages 227–270, 1968.

[22] M.R. Quillian. The teachable language comprehender: a simulation program and theory of language. *Communications of the ACM*, 12(8):459–476, 1969.

[23] P.J. Hayes. On semantic nets, frames and associations. In *Proceedings of the 5th International. Joint Conference on Artificial Intelligence*, pages 99–107, EUA, 1977.

[24] J. Véronis and N Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 389–394, Finlandia, 1990.

[25] N. Rattanawongchaiya, K. Naruedomkul, N. Cercone, and B. Sirinaovakul. Multidictionary with word sense disambiguation system architecture. In *Proceedings of the National UniServe Conference*, pages 107–112, Australia, 2006.

[26] D. Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, pages 454–460, Francia, 1992.

[27] R. Navigli. Online word sense disambiguation with structural semantic interconnections. In *Proceedings of the 11th Conference of the European Association for Computational Linguistics*, pages 107–110, Italia, 2006.

[28] M. Sinha, M. Kumar, P. Pande, L. Kashyap, and P. Bhattacharyya. Hindi word sense disambiguation. In *Proceedings of the International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems*, India, 2004.

[29] R. Navigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(27):1075–1086, 2005.

[30] M. Lesk. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference*, pages 24–26, Canada, 1986.

[31] P. Archibald. An exploration of abstract thesaurus instantiation. Master thesis, University of Kansas, Lawrence, Kansas, 1985.

[32] G.A. Miller. Wordnet: An on-line lexical database. *International Journal of lexicography*, 3(4), 1990.

[33] M. Warin, H. Oxhammar, and M. Volk. Enriching an ontology with wordnet based on similarity measures. In *Proceedings of the MEANING-2005 Workshop*, Italia, 2005.

[34] S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–257, México, 2003.

[35] A. Purandare and T. Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the 8th Conference on Natural Language Learning*, pages 41–48, EUA, 2004.

[36] P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Canada, 2002.

[37] S. Brin and L. Page. The anatomy of a large - scale hypertextual web search engine. In *Proceedings of the 7th International Conference on World Wide Web 7*, pages 107–117, Australia, 1998.

[38] J. Véronis. Hyperlex: lexical cartography for information retrieval. *Computer Speech and Language*, 18(3):223–252, 2004.

[39] E. Agirre, D. Martínez, O. López, and A. Soroa. Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 585–593, Australia, 2006.

[40] M. Diab and P. Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262, EUA, 2001.

[41] X. Wang and J. Carroll. Word sense disambiguation using sense examples automatically acquired from a second language. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 547–554, Canada, 2005.

[42] N. Ide, T. Erjavec, and D. Tufi? Automatic sense tagging using parallel corpora. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, pages 83–89, Japón, 2001.

[43] G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1991.

[44] N. Ide, T. Erjavec, and D. Tufis. Sense discrimination with parallel corpora. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense*, pages 54–60, Filadelfia, 2002.

[45] E. Macklovitch and M.L. Hannan. Line 'em up: Advances in alignment technology and their impact on translation support tools. *Machine Translation*, 13(1):41–57, 1998.

[46] J.A. Vera and G. Sidorov. Proyecto de preparación del corpus paralelo alineado español-inglés. In *Memorias del Encuentro Internacional de la Ciencias de la Computación*, Mexico, 2004.

[47] M. Mikhailov. Parallel corpus aligning: Illusions and perspectives. *The Austrian Academy Corpus*, 2002.

[48] C. Kit, J. Webster, H. Pan K. Sin, and H. Li. Clause alignment for bilingual hong kong legal texts with available lexical resources. In *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages*, pages 286–292, China, 2003.

[49] A. Gelbukh, G. Sidorov, and J.A. Vera. A bilingual corpus of novels aligned at paragraph level. In *Proceedings of the 5th International Conference on NLP*, pages 16–23, Finlandia, 2006.

[50] W. Gale and K. Church. Program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993.

[51] P. Brown, J. Lai, and R. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, EUA, 1991.

[52] D. Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 80–87, EUA, 1994.

[53] M. Simard, G. Foster, and P. Isabelle. Using cognates to align sentences in parallel corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Canada, 1992.

[54] S. Chen. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 9–16, EUA, 1993.

[55] M. Haruno and T. Yamazaki. High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of Annual Conference of the Association for Computational Linguistics*, pages 131–138, EUA, 1996.

[56] F. Debili and E. Sammouda. Appariement des phrases de textes bilingues français-anglais et français-arabe. In *Proceedings of the 14th Conference on Computational Linguistics*, Francia, 1992.

[57] P. Brown, S. Della, V. Della, and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[58] S. Vogel, H. Ney, and C. Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Denmark, 1996.

[59] F. Smadja, V. Hatzivassiloglou, and K. McKeown. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[60] S. Ker and J. Chang. A class-based approach to word alignment. *Computational Linguistics*, 23(2):313–343, 1997.

[61] D. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

[62] Y. Seng and H. Tou. Word sense disambiguation with distribution estimation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1010–1015, Escocia, 2005.

[63] H. Kaji and Y. Morimoto. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Taiwan, 2002.

[64] Idioma español. http://es.wikipedia.org/wiki/Castellano, 2008. Consultado 09/12/08.

[65] J.C. Moreno. *El universo de las lenguas. Clasificación, denominación, situación, tipología, historia y bibliografía de las lenguas*. Castalia, S.A., 2003.

[66] Ethnologue, languages of the word. http://www.ethnologue.com/, 1999. An encyclopedic reference work cataloging all of the world's 6,912 known living languages.

[67] J.E. Gargallo and M. Reina. *Manual de lingüística románica*. Ariel, 2007.

[68] H. Krahe. *Lingüística germánica*. Ediciones Cátedra, 1977.

[69] R. Hetzron. *The Semitic Languages*. Taylor & Francis, 1997.

[70] Idioma italiano. http://es.wikipedia.org/wiki/Idioma_italiano, 2008. Consultado 18/12/08.

[71] Italian grammar. http://en.wikipedia.org/wiki/Italian_grammar, 2008. Consultado 23/12/08.

[72] C. Valero. Inglés y español mano a mano: dos lenguas y dos formas de ver el mundo. *Cuadernos Cervantes de la lengua española (Especial lingüística contrastiva)*, (29), 2001.

[73] English verbs. http://en.wikipedia.org/wiki/English_verbs, 2008. Consultado 23/12/08.

[74] D. Levin. El español en ruso y en hebreo. In *XIII Congreso Internacional de ASELE: El español, lengua del mestizaje y la interculturalidad*, pages 515–521, España, 2002.

[75] Hebrew grammar. http://en.wikipedia.org/wiki/Hebrew_grammar, 2008. Consultado 23/12/08.

[76] Hebrew verb conjugation. http://en.wikipedia.org/wiki/Hebrew_verb_conjugation, 2008. Consultado 23/12/08.

[77] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. Five papers on wordnet. *CSL Report 43*, 1993.

[78] Eurowordnet. http://www.illc.uva.nl/EuroWordNet/, 2001. Consultado 29/12/08.

[79] Multiwordnet. http://multiwordnet.itc.it/english/home.php, 2004. Consultado 29/12/08.

[80] E. Pianta, L. Bentivogli, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 21–25, India, 2002.

[81] B. Magnini and G. Cavaglia. Integrating subject field codes into wordnet. In *Proceedings of the LREC-2000*, pages 1413–1418, Grecia, 2000.

[82] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources*, pages 101–108, Suiza, 2004.

[83] P. Resnik. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, EUA, 1999.

[84] Agenzia fides. http://www.fides.org, 2009. Consultado 5/01/09.

[85] Al-bushra. http://www.al-bushra.org/, 2009. Consultado 5/01/09.

[86] E. Zanchetta and M. Baroni. Morph-it! a free corpus-based morphological resource for the italian language. In *Proceedings of Corpus Linguistics 2005*, Reino Unido, 2005.

[87] Wordnet's morphological processing. http://wordnet.princeton.edu/man/morphy, 2006. Consultado 25/09/08.

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

[89] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 144–152, EUA, 2004.

[90] H.H. Do, S. Melnik, and E. Rahm. . comparison of schema matching evaluations. In *Proceedings of the GI-Workshop: Web and Databases*, pages 221–237, Alemania, 2002.

[91] M. Taulé and M.A. Martí. Senseval, una aproximación computacional al significado. *Digithum. Revista digital d'humanitats*, (5), 2003.

[92] L. Márquez, M. Taulé, M.A. Martí, N. Artigas, M. García-Vega, F. Real, and D. Ferres. Senseval-3: The spanish lexical sample task. In *Proceedings of the SENSEVAL International Workshop On The Evaluation Of Systems For The Semantic Analysis Of Text*, pages 21–24, España, 2004.

[93] Senseval-3 evaluation exercises for word sense disambiguation. http://www.senseval.org/senseval3, 2004. Consultado 7/01/09.

Apéndice A

Dominios sin lema asociado en el hebreo

El Cuadro A.1 muestra aquellos sominios del hebreo que no tienen ningún lema asociado.

Dominios
Environments
Veterinary
Philately
Psychoanalysis
Paranormal
Badminton
Rugby
Cycling
Skating
Hockey
Mountaineering
Rowing
Sub
Bowling
Applied_Science
Radiology
Biochemistry
Genetics
Paleontology
Statistics
Topography
Telegraphy

Cuadro A.1: Dominios con ningún lema asociado en el hebreo

Apéndice B

Resultados de la alineación por oración

Los Cuadros B.1 y B.2 muestra los valores, por oración, de las alineaciones obtenidas por el sistema utilizando las medidas de similitud semántica: *LeacockandChodorow* (LCH), $Hirst - St - Onge$ (HSO), *Edge* y *Random* en el corpus de 23 oraciones extraídas de fragmentos de *Don Quijote de la Mancha*.

Las entradas indican *cantidad_pares_correctos*/*cantidad_pares_sugeridos*. El topline indica el número máximo de equivalencias que podrían ser extraídas por el sistema, considerando relaciones 1:1 exclusivamente, la incompletitud de las redes de palabras que conforman MWN y las brechas léxicas (gaps) de los lenguajes implicados.

#P	#O	Gold standard	Topline	Sólo coincidentes	LCH	HSO	Edge	Random
1	1	17	13	10/11	11/14	12/17	11/14	10/16
1	2	19	18	9/9	11/12	12/15	11/15	10/18
1	3	11	6	3/4	5/9	5/10	6/10	5/11
1	4	15	14	10/10	12/14	12/14	12/15	10/16
1	5	14	9	8/8	8/8	8/11	8/9	6/11
1	6	17	12	10/11	9/12	9/11	9/12	10/15
1	7	8	7	4/4	5/6	6/6	6/7	4/10
2	1	21	16	10/12	10/13	11/18	11/15	9/20
2	2	32	31	21/22	20/26	22/28	20/27	18/33
2	3	20	18	14/16	15/18	15/19	15/18	15/21
2	4	15	10	6/6	6/8	6/9	6/9	8/12
2	5	15	12	8/9	8/10	9/12	8/11	6/13
2	6	20	17	9/10	10/12	9/13	10/15	9/18
2	7	9	6	5/6	5/7	5/7	6/8	5/8

continúa...

#P	#O	Gold standard	Topline	Sólo coincidentes	LCH	HSO	Edge	Random
3	1	15	13	8/8	10/13	11/14	10/13	8/14
3	2	32	31	18/24	19/27	18/29	18/27	18/33
3	3	25	24	12/12	13/18	13/21	13/19	10/26
3	4	6	5	2/2	3/4	2/5	2/4	4/5
4	1	9	8	4/4	3/5	3/6	3/5	2/8
4	2	18	17	8/8	9/12	8/13	8/13	8/19
4	3	12	9	6/8	7/10	7/11	7/10	6/12
4	4	23	19	14/16	13/18	14/18	13/18	15/23
4	5	16	14	7/8	8/13	9/14	8/13	7/15
Totales		389	329	206/228	220/289	226/321	221/307	203/377

Cuadro B.1: Pares extraídos por oración usando el *inglés* como lenguaje meta

#P	#O	Gold standard	Topline	Sólo coincidentes	LCH
1	1	17	14	7/8	8/13
1	2	18	16	8/8	9/14
1	3	13	8	5/5	3/10
1	4	12	9	3/3	4/10
1	5	13	12	5/6	5/8
1	6	15	10	4/6	5/9
1	7	8	7	2/2	2/6
2	1	14	9	8/8	10/15
2	2	25	22	15/19	17/25
2	3	20	19	10/11	13/16
2	4	12	11	5/5	9/10
2	5	11	10	3/3	5/8
2	6	13	12	7/9	7/11
2	7	8	7	0/1	2/6
3	1	12	9	6/6	6/10
3	2	27	25	19/24	19/27
3	3	21	17	9/10	11/17
3	4	4	2	0/0	1/2

continúa...

#P	#O	Gold standard	Topline	Sólo coincidentes	LCH
4	1	6	4	3/4	4/5
4	2	18	16	9/9	9/12
4	3	11	9	5/7	5/10
4	4	21	17	10/13	10/17
4	5	14	12	5/5	5/8
Totales		333	277	148/172	220/289

Cuadro B.2: Pares extraídos por oración usando el *italiano* como lenguaje meta

Apéndice C

Recall de la alineación por oración

De la Figura C.3 a la C.14 representan la cantidad de pares correctos propuestos por el sistema versus el gold standard establecido por los anotadotes y el topline, por lo que tienen relación con el recall. Se presentan los gráficos para ambos idiomas meta (italiano e inglés)

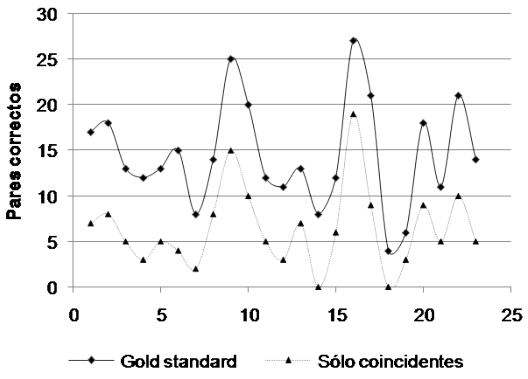


Figura C.1: Sólo coincidentes vs. gold standard (*italiano*)

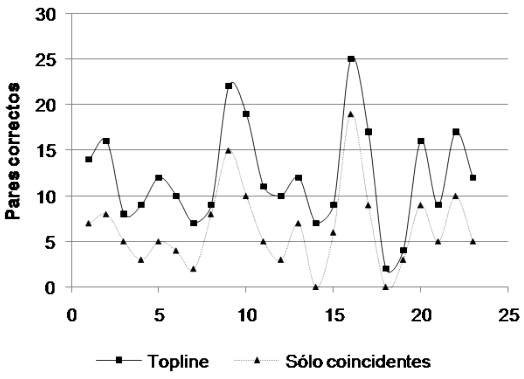


Figura C.2: Sólo coincidentes vs. topline (*italiano*)

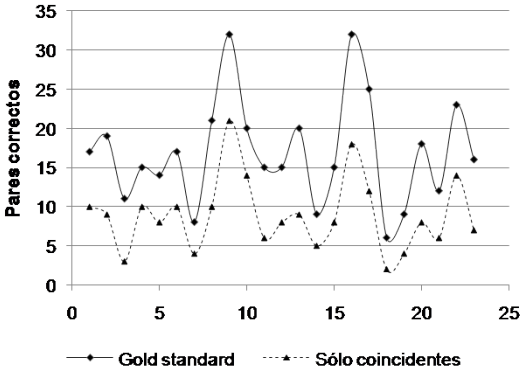


Figura C.3: Sólo coincidentes vs. gold standard (*inglés*)

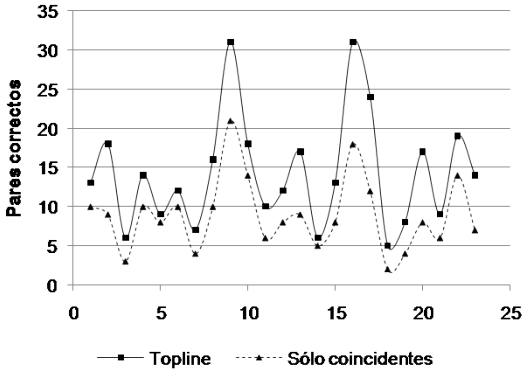


Figura C.4: Sólo coincidentes vs. topline (*inglés*)

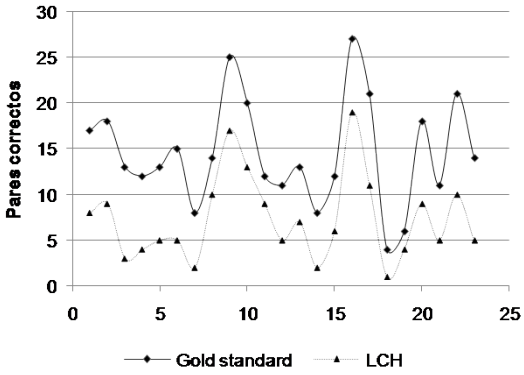


Figura C.5: LCH vs. gold standard (*italiano*)

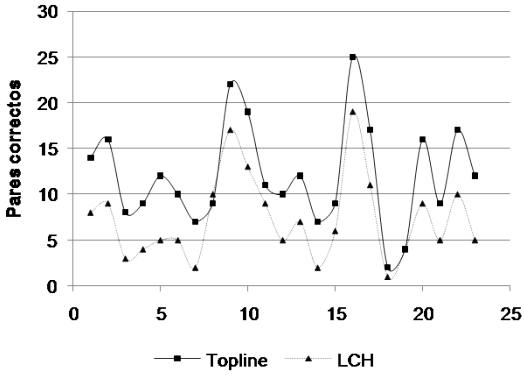


Figura C.6: LCH vs. topline (*italiano*)

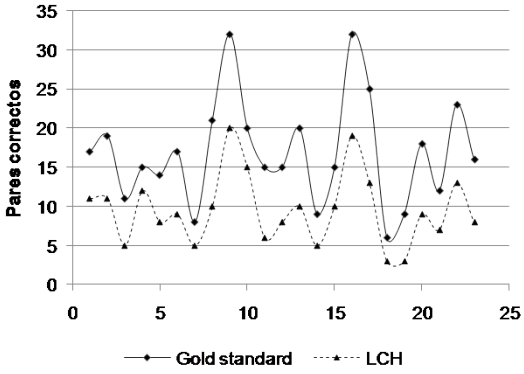


Figura C.7: LCH vs. gold standard (*inglés*)

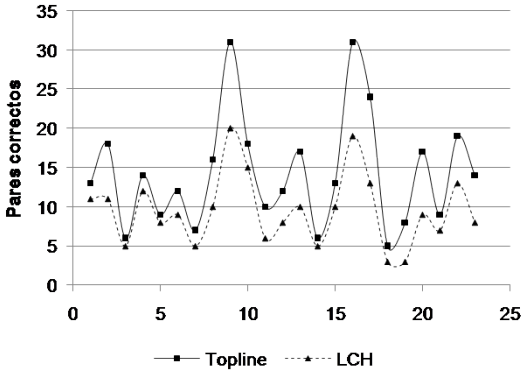


Figura C.8: LCH vs. topline (*inglés*)

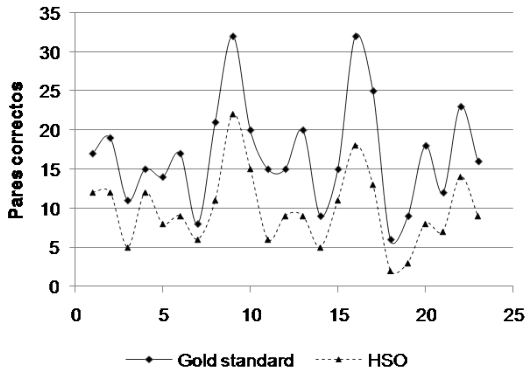


Figura C.9: HSO vs. gold standard (*inglés*)

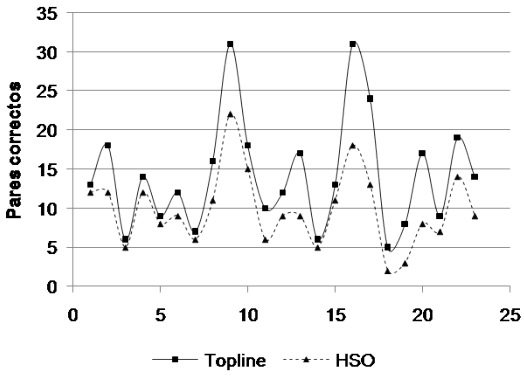


Figura C.10: HSO vs. topline (*inglés*)

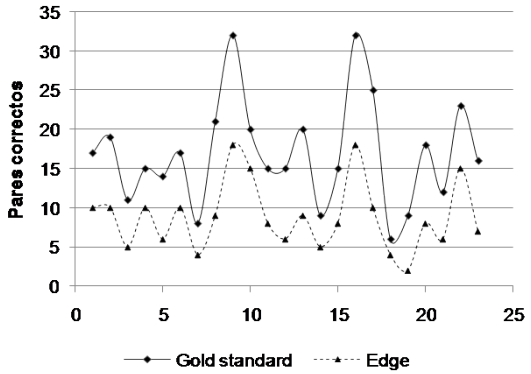


Figura C.11: Edge vs. gold standard (*inglés*)

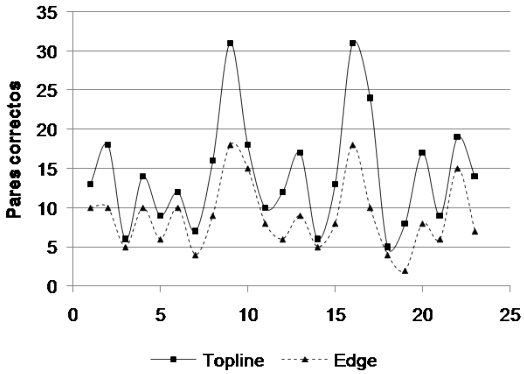


Figura C.12: Edge vs. topline (*inglés*)

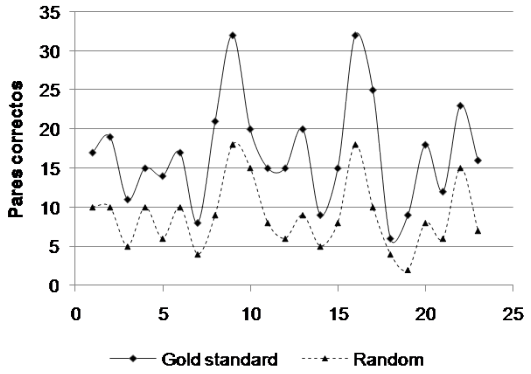


Figura C.13: Random vs. gold standard (*inglés*)

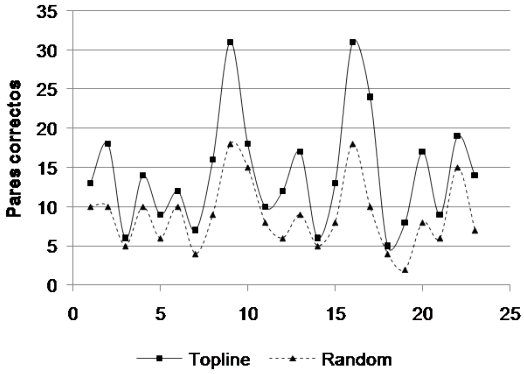


Figura C.14: Random vs. topline (*inglés*)

Apéndice D

Precisión de la alineación por oración

De la Figura D.1 a la D.7 representan la cantidad de pares correctos propuestos por el sistema versus la cantidad de pares sugeridos por el sistema, por lo que tienen relación con la precisión.

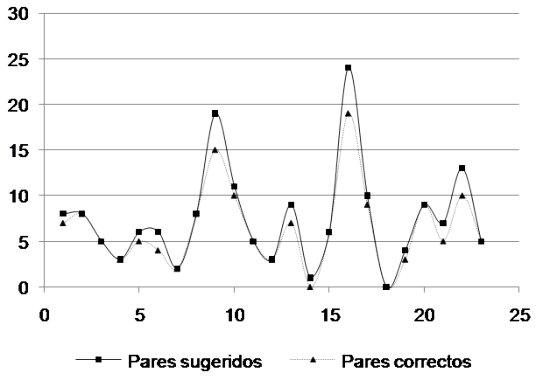


Figura D.1: Precisión sólo coincidentes (*italiano*)

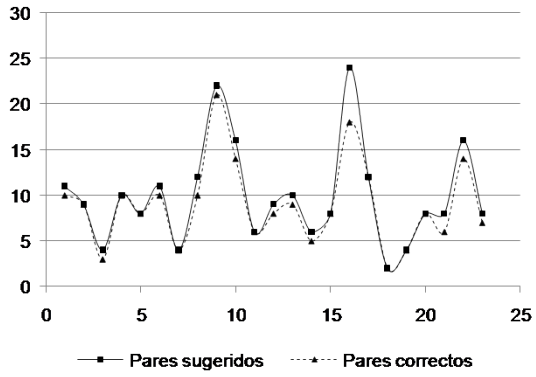


Figura D.2: Precisión sólo coincidentes (*inglés*)

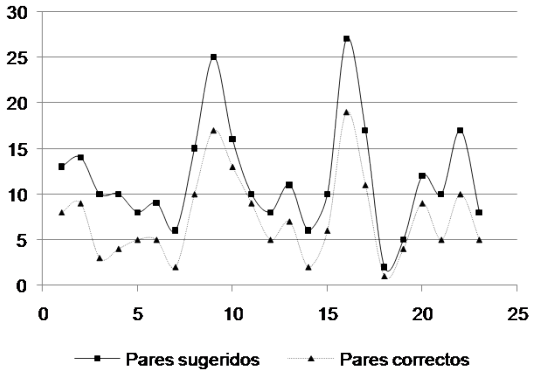


Figura D.3: Precisión LCH (*italiano*)

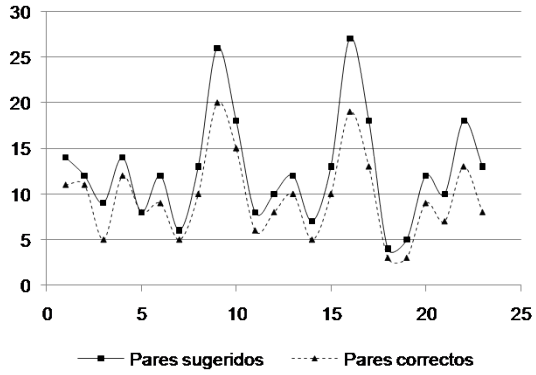


Figura D.4: Precisión LCH (*inglés*)

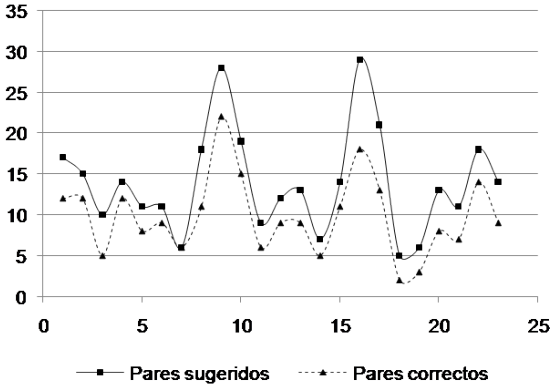


Figura D.5: Precisión HSO (*inglés*)

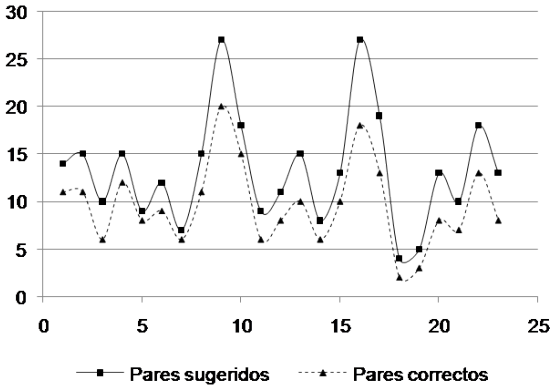


Figura D.6: Precisión Edge (*inglés*)

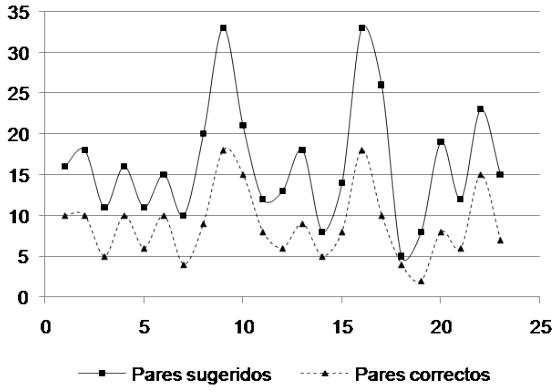


Figura D.7: Precisión Random (*inglés*)

Apéndice E

Palabras polisémicas

El Cuadro E.1 muestra el conjunto de las 46 palabras polisémicas seleccionadas en la evaluación del algoritmo de desambiguación. También se agrega información del POS, número de sentidos y cantidad de ejemplos de entrenamiento, sin etiquetar y de prueba.

			# de ejemplos		
Palabra	POS	# sentidos	de entrenamiento	de prueba	no etiquetados
actuar	v	3	133	67	1500
apoyar	v	3	259	128	1500
apuntar	v	4	213	106	1500
arte	n	3	251	121	1500
autoridad	n	2	268	132	1500
bajar	v	3	235	115	1500
banda	n	4	230	114	1500
brillante	a	2	126	63	1369
canal	n	4	262	131	1500
canalizar	v	2	253	126	700
ciego	a	3	102	52	390
circuito	n	4	261	132	1500
columna	n	7	129	64	1258
conducir	v	4	134	66	1094
corazón	n	3	123	62	1500
corona	n	3	124	64	916
duplicar	v	2	254	126	1500
explotar	v	5	212	103	1500
ganar	v	3	237	118	1500
gracia	n	3	72	38	1209
grano	n	3	117	61	524
hermano	n	2	128	66	1500
jugar	v	3	236	117	1500
letra	n	5	226	114	1251

continúa...

Palabra	POS	# sentidos	# de ejemplos de entrenamiento	# de ejemplos de prueba	# de ejemplos no etiquetados
masa	n	3	172	85	1151
mina	n	2	134	66	1458
natural	a	5	215	107	1500
naturaleza	n	3	258	128	1500
operación	n	3	134	66	1500
órgano	n	2	263	131	1500
partido	n	2	133	66	1500
pasaje	n	4	220	111	375
perder	v	4	218	106	1500
popular	a	3	133	67	1500
programa	n	3	267	133	1500
saltar	v	8	200	101	1117
simple	a	3	117	61	1500
subir	v	3	231	114	1500
tabla	n	3	130	64	1500
tocar	v	6	158	78	1500
tratar	v	3	143	72	1235
usar	v	2	263	130	1500
vencer	v	3	134	65	1500
verde	a	2	69	33	1500
vital	a	2	131	65	1500
volar	v	3	122	60	705
TOTAL			8430	4195	61252

Cuadro E.1: Palabras polisémicas en tarea de muestra léxica para el español

Apéndice F

Diccionarios de optimización

Las siguientes son entradas representativas del los diccionarios de optimización empleados para agilizar la determinación de los valores de similitud semántica.

Medida de similitud LCH

n#06381267	n#06107243	1.519826	lugar	village
n#06381267	n#07533214	1.386294	lugar	have
n#00014887	n#00049044	1.268511	lugar	call
n#06381267	n#04444887	1.268511	lugar	mind
n#06381267	n#06262675	2.079442	lugar	there
n#06381267	n#10951384	1.067841	lugar	long
n#06381267	n#07660871	1.163151	lugar	gentleman
n#06381267	n#02823427	1.268511	lugar	keep
n#06381267	n#03316835	1.268511	lugar	buckler
n#06352837	n#00287536	1.163151	lugar	coursing
v#00518436	v#01508689	1.856298	acordar	have
v#00518436	v#00696267	1.856298	acordar	call
v#00336188	v#00491303	2.079442	acordar	mind
v#00518436	v#01231785	1.673976	acordar	long
v#00518436	v#01832078	1.856298	acordar	keep
v#00518436	v#01414286	1.673976	acordar	course

Figura F.1: Similitud *LCH* de los lemas *lugar* y *acordar* con las palabras del contexto meta

Medida de similitud HSO

n#05369177	n#03194375	5.000000	pendencia	quarrel
n#05369177	n#05372686	0.000000	pendencia	wooing
n#05369177	n#05572363	0.000000	pendencia	agony
n#05369177	n#04496504	0.000000	pendencia	sort
n#05369177	n#04345975	0.000000	pendencia	mind
n#05369177	n#02656657	0.000000	pendencia	fabric
n#05369177	n#04531761	0.000000	pendencia	fancy
n#05369177	n#03963063	0.000000	pendencia	true
n#05369177	n#04476879	0.000000	pendencia	reality
n#00266805	n#10195710	2.000000	camino	sore
n#00266805	n#06846731	4.000000	camino	want
n#06348591	n#00112808	5.000000	camino	road
n#06348591	n#02835429	2.000000	camino	inn
n#00266805	n#04974979	2.000000	camino	welcome

Figura F.2: Similitud *HSO* de los lemas *pendencia* y *camino* con las palabras del contexto meta

Medida de similitud Edge

n#01900545	n#10746382	0.076923	vaca	olla
n#01403710	n#05707187	0.333333	vaca	beef
n#01900545	n#09869910	0.076923	vaca	mutton
n#01900545	n#05839780	0.076923	vaca	salad
n#01900545	n#02638930	0.100000	vaca	scrap
n#07189690	n#01329440	0.125000	vaca	pigeon
n#01900545	n#02655039	0.111111	vaca	extra
n#01900545	n#03412153	0.166667	vaca	quarter
n#01900545	n#09538207	0.071429	vaca	income
n#05840548	n#10746382	0.100000	salpicón	olla
n#05840548	n#05386686	0.125000	salpicón	beef
n#05840548	n#09869910	0.125000	salpicón	mutton
n#05840548	n#05839780	0.500000	salpicón	salad
n#05840548	n#10656657	0.111111	salpicón	scraps
n#05840548	n#01329440	0.062500	salpicón	pigeon
n#05840548	n#07597116	0.100000	salpicón	extra
n#05840548	n#09891778	0.090909	salpicón	quarter
n#05840548	n#09538207	0.066667	salpicón	income

Figura F.3: Similitud *Edge* de los lemas *vaca* y *salpicón* con las palabras del contexto meta