UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO
UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Comparación de medidas de similitud para desambiguación del sentido de las palabras utilizando rankeo de grafos"

Tesis

Que para obtener el grado de
Maestra en Ciencias de la Computación

Que Presenta
Ing. Selene Itzel Vargas Flores

Tutor Académico:
Dra. Yulia Nikolaevna Ledeneva

Tutores Adjuntos:
Dr. René Arnulfo García Hernández
Dr. Grigori Sidorov

TIANGUISTENCO, MÉXICO. ENERO 2016

Dedicatoria

A mis amados padres, quienes en todo momento estuvieron presentes, brindándome su apoyo incondicional y su amor.

A mi hermano que durante mi maestría me dio el mayor regalo que he recibido en la vida, el cual abrió muchas posibilidades en mi vida para realizar mis sueños.

A Isael que sin lugar a duda es una personita maravillosa, que confió en mí y que estuvo en todo momento, apoyándome y sacando cada palabra para saber que si se podía obtener un logro más en mi vida, lo amo.

Agradecimientos

A mí por ser un extraordinario ser humano que la vida me ha permitido ser día a día, me aprecio por ser una hija responsable, amorosa y comprometida con lo que realiza.

Agradezco a Dios por la oportunidad de dejar realizarme en un logro más en mi vida, por la sabiduría, paciencia y dedicación que supo guiarme en los momentos difíciles.

A mis padres por ser mi ejemplo a seguir, por el apoyo incondicional que me brindaron durante mi estudio, el amor puro y sincero, los aprecio por forjar al gran ser humano que soy y poder concluir con éxito.

A mi hermano que amo, por ser un extraordinario ser humano, por el apoyo, los ánimos y buenos consejos que me dio.

A mis asesores de tesis la Dra. Yulia Ledeneva, el Dr. René Arnulfo García Hernández y al Maestro Rafael Cruz Reyes, por creer en mí, por su paciencia, sabiduría, comprensión y por el gran apoyo brindado para la realización de esta tesis y por ser parte fundamental de mi formación académica.

A mi compañero Miguel Ángel Calderón por el apoyo y tiempo brindado.

También expreso mi gratitud al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico otorgado durante mis estudios de maestría en el programa Maestría en Ciencias de la Computación a través de la beca con número (CVU/Becario): 559604/296810.

Resumen

La desambiguación del sentido de las palabras es uno de los problemas más importantes del área del procesamiento del lenguaje natural. Es indispensable en la aplicación de diferentes tarea tales como recuperación de información, traducción automática, búsqueda de respuestas y generación automática de resúmenes, entre otras. Resolver el problema de la desambiguación del sentido de las palabras consiste en seleccionar el sentido correcto de una palabra en un contexto específico de un conjunto determinado de sentidos posibles.

En esta tesis, se utiliza la tarea léxica completa *english-all-words* en el idioma inglés del foro de Senseval-2 y el diccionario de sentidos llamado *WordNet* en la versión 2.1. Para todas las palabras a desambiguar se extraen todos los sentidos del dicho diccionario y se selecciona un sentido correcto.

Uno de los mejores métodos que se aplica para esta tarea es el método de rankeo basado en grafos, llamado *TextRank* [Mihalcea, 2006]. Se propone comparar diferentes medidas de similitud en este método de rankeo de grafos, que son la medida de coseno, la medida de edición y la subsecuencia común más larga (del inglés, *Longest Common Subsequence* (LCS)). La evaluación se realiza utilizando las medidas de Precisión y Recuerdo.

Contenido

Página

Lista de figuras

Lista de tablas

CAPÍTULO 1.

Introducción

El lenguaje es el medio de comunicación más eficaz de que disponen los seres humanos, utilizado de diversas maneras, sirve para expresar todo tipo de sentimientos, emociones, dar a entender un concepto o idea y empezar alguna comunicación. El lenguaje es parte fundamental de los seres humanos [Gelbukh & Sidorov, 2006].

El lenguaje es el recurso más importante que poseemos los seres humanos. Gran parte de esta información se comunica, almacena y maneja en forma de lenguaje natural (español, inglés, ruso, etc.). En la actualidad, podemos obtener grandes volúmenes de información en forma escrita, ya sea de manera impresa o electrónica. Las palabras comúnmente tienen múltiples sentidos. Los seres humanos hacen intuitivamente la selección del sentido correcto de una palabra en un contexto en particular [Torres, 2009].

La ciencia encargada de estudiar el lenguaje humano y establecer diferentes modelos que permitan las máquinas comprender el lenguaje humano, es la tarea de la lingüística computacional.

El uso de la computadora es cada vez más frecuente en el desarrollo de las actividades cotidianas de los seres humanos. Siendo una herramientas necesarias para el procesamiento de la información formada en los textos, ya que son capaces de manejar grandes volúmenes de datos y múltiples sentidos. Pero también una computadora no puede hacer todo lo que las personas normalmente hacemos con el texto, por ejemplo, responder preguntas basándose en la información proporcionada, o, hacer deducciones de algún contenido, o elaborar un resumen de esta información [Torres, 2009].

Por lo anterior, el Procesamiento de Lenguaje Natural (PLN) tiene por objetivo habilitar a las computadoras para que entiendan el texto, procesándolo por su sentido. Para llevar a cabo esta tarea, un sistema de PLN necesita conocer sobre la estructura del lenguaje, la cual se analiza normalmente en 4 niveles: morfológico, sintáctico, semántico y pragmático. En el nivel morfológico se trabaja cómo se construyen las palabras; en el sintáctico, cómo combinar las palabras para formar oraciones; en el semántico, el significado de las palabras, y por último, en el pragmático se trabaja cómo el contexto afecta a la interpretación de las oraciones. Nuestra investigación se centra en el nivel semántico. Los 4 niveles antes mencionados suelen tener un problema; la ambigüedad, es el proceso de decidir (seleccionar) el sentido de una palabra en su contexto, decimos seleccionar porque cada palabra tiene un conjunto determinado de sentidos posibles. Este problema surge debido a que las palabras pueden asumir diferentes significados dependiendo texto, oración o palabra en el que se usan [Gelbukh & Sidorov, 2002].

Existen varios tipos de ambigüedad: sintáctica (estructural), léxica y semántica. En el presente trabajo nos centramos en el problema de la

ambigüedad semántica, que se presenta cuando una palabra tiene múltiples significados [Gelbukh & Sidorov, 2002].

Se dice que una estructura gramáticas (un texto, una oración, una palabra) es ambigua, cuando ésta puede ser entendida de dos o más formas, es decir que expresa más de un significado. Si la ambigüedad corresponde a la interpretación de una oración o un fragmento es llamada estructural o sintáctico, y si esta se presenta en un vocablo es llamada léxica. Por ejemplo, la oración *the man saw the girl with the telescope,* presenta ambigüedad estructural, ya que podría ser interpretada de dos maneras: el hombre vio a una niña que tenía un telescopio o el hombre uso el telescopio para ver a una niña. Ahora, la oración *the man saw the girl with the telescope*, expresa un significado no ambiguo para un ser humano, ya que éste sabe que un sombrero no es utilizado para ver; sin embargo, sigue siendo ambigua para una computadora, porque ésta no sabe que un sombrero no se una para ver [Tejeda, 2006].

La ambigüedad léxica, ocurre cuando un vocablo expresa múltiples sentidos, como por ejemplo *church*, el cual posee cuatro sentido diferentes si se toma como referencia el diccionario de *WordNet* 2.1, tal como se muestra en la figura 1.

CHURCH

1: one of the groups of Christians who have their own beliefs and form of worship
2: a place for public (especially Christian) worship
3: a service conducted in a church
4: the body of people who attend or belong to a particular local church

Figura 1. Definición de "church" según *WordNet* 2.1.

1.1 Planteamiento del problema

¿Cómo encontrar el sentido correcto de una palabra, utilizando diferentes medidas de similitud en el método de rankeo de grafos para la tarea léxica completa *english-all-words,* en el idioma inglés, del foro Senseval-2?

1.2 Objetivo general

Comparar diferentes medidas de similitud en el método de rankeo de grafos para la tarea léxica completa *english-all-words,* en el idioma inglés, del foro Senseval-2.

1.3 Objetivos específicos

- Probar diferentes medidas de similitud para conocer el desempeño que se pueda obtener.
- Definir las medidas de similitud del método del rankeo basado en grafos para seleccionar el sentido correcto de las palabras.
- Evaluar los resultados obtenidos utilizando Precisión y Recuerdo.
- Extraer los sentidos del diccionario de *WordNet* versión 2.1.

1.4 Hipótesis

Si se utilizan diferentes medias de similitud en el método de rankeo de grafos, entonces es posible seleccionar el sentido correcto, conociendo el desempeño del método de rankeo de grafos para la tarea léxica competa del Senseval-2 de desambiguación del sentido de las palabras.

1.5 Delimitación del problema

- El foro Senseval-2 que es utilizado ya no será elaborado, ya que está desarrollado por el foro de Senseval para datos de prueba.
- Se prueba el método propuesto con el foro Senseval-2, en el idioma inglés en la tarea léxica (english-all-words)
- Se utiliza el diccionario de *WordNet* en la versión 2.1 que se descargó de la página https://wordnet.princeton.edu/.

1.6 Justificación

La desambiguación del sentido de las palabras *(WSD)* es considerada como uno de los problemas más importantes de investigación en el PLN. Es esencial para las aplicaciones que requieren la comprensión del lenguaje, ya que utiliza la información para estudiar y tratar el lenguaje humano, con ellos se reconoce a la computación lingüística. Pero la tarea no es pensada como un fin en sí misma, sino como una mejora para otras tareas y aplicaciones de la computación lingüística y del PLN, donde puede ser aplicada en: traducción automática, recuperación de información, búsqueda de respuestas, resumes de textos y la desambiguación del sentido de las palabras.

1.7 Estructura de la tesis

En el capítulo 1, se ha descrito una breve introducción de lo que será el presente trabajo, así como el planteamiento del problema, el objetivo general y los objetivos específicos, también la hipótesis y delimitaciones.

En el capítulo 2, se describen los conceptos teóricos esenciales que van a ser utilizados durante el presente trabajo y que nos ayudarán en los posteriores capítulos.

En el capítulo 3, se presenta el estado del arte donde se describen los trabajos previos, tomado como base para el desarrollo de la propuesta.

En el capítulo 4, se describe la construcción de la metodología propuesta así como los conceptos que la integran, describiendo cada uno de los pasos de la metodología planteada.

En el capítulo 5, se describe la experimentación de las pruebas del método propuesto y se expone los resultados generados durante el proceso.

En el capítulo 6, se dan conclusiones y se comentan los resultados más significativos que se han obtenido en la experimentación. Se apuntan las ideas del trabajo futuro para continuar el esfuerzo de esta tesis.

CAPÍTULO 2.

Marco Teórico

En este capítulo, se introducen los conceptos fundamentales que permiten comprender el entorno y desarrollo del presente trabajo de tesis.

2.1 Lenguaje

El lenguaje natural se entiende como el lenguaje hablado y escrito con el propósito que exista comunicación entre una o varias personas, es más directo para expresar lo que se quiere comunicar [Gelbukh & Sidorov, 2006]. Por ello el ser humano ha utilizado el lenguaje para transmitir un conocimiento, sentimientos, con el fin de comunicarse con el resto de los seres humanos ya sea de manera gráfica, oral o escrita.

2.1.1 Niveles del lenguaje

Existen diferentes niveles principales para el análisis de la estructura del lenguaje dentro de la lingüística [Bolshakov & Gelbukh, 2004] que son:

- Nivel fonológico: trata de los sonidos que componen el habla, permitiendo formar y distinguir palabras.
- Nivel morfológico: trata sobre la estructura de las palabras y las leyes para formar nuevas palabras a partir de unidades de significado más pequeñas llamadas morfemas.
- Nivel sintáctico: trata sobre cómo las palabras pueden unirse para construir oraciones y cuál es la función que cada palabra realiza en esa oración.
- Nivel semántico: trata del significado de las palabras y de cómo se unen para dar significado a una oración.
- Nivel pragmático: estudia la intención del hablante al producir oraciones específicas o textos en una situación específica.

Por lo anterior, conlleva a involucrar a la Lingüística Computacional, es el campo multidisciplinario (lingüística y la informática) que utiliza la información para estudiar y tratar el lenguaje humano, así como intentar modelar de forma lógica el lenguaje natural desde un punto de vista computacional [Torres, 2009]. Por ello tiene como objetivo la realización de aplicaciones informáticas que imiten la capacidad humana de hablar y entender [Tello, 2010].

2.2 Ambigüedad

La ambigüedad surge en el lenguaje natural cuando una estructura gramatical puede ser interpretada de varias maneras [Gómez Montes, y otros]. Es una situación en la que la información se puede entender o

interpretar de más de una manera. La desambiguación de sentido de las palabras, que en inglés se denomina *Word Sense Desambiguation (WSD),* consiste en identificar el sentido de un vocablo ambiguo en un determinado contexto usando un conjunto de candidatos establecido [Tejeda, 2006].

2.2.1 Tipos de ambigüedad

Existen tres tipos de ambigüedad como son ambigüedad: léxica, sintáctica (estructural) y semántica.

- Ambigüedad léxica: se presenta cuando las palabras pueden pertenecer a diferentes categorías gramaticales, por ejemplo bajo puede ser una preposición, un sustantivo, un adjetivo o una conjugación del verbo bajar [Sidorov, 2005].
- Ambigüedad sintáctica o estructural: este tipo de ambigüedad se presenta en la estructura de la oración, uno de los ejemplos típicos es la oración: "veo al gato con el telescopio", lo que no es claro es ¿quién está en el telescopio? [García, 2003].
- Ambigüedad semántica: este tipo de ambigüedad se encarga de procesar aquellos vocablos que tienen múltiples significados [Diccionario de la Lengua Española, 2002]. Por ejemplo, banco puede significar banco de peces, banco para tomar asiento o institución financiera [Tejeda, 2006].

Cabe recordar, que la ambigüedad puede tener más de una interpretación, es decir puede existir ambigüedad léxica, sintáctica o estructural, semántica, contextual y referencial, de acuerdo al nivel de análisis [Molina, 1999].

2.3 Desambiguación del sentido de las palabras

La desambiguación del sentido de las palaras consiste en asignar el sentido más apropiado a una palabra dentro de un contexto dado. Existen diferentes aplicaciones prácticas tales como la traducción automática, y la adquisición de conocimientos, entre otras, requieren de conocimientos sobre el significado de palabras y desambiguación del sentido de las palabras se considera esencial para todas estas aplicaciones [Torres, 2009].

En los últimos años se han incrementado las investigaciones para crear métodos de *WSD*. A continuación se describe la clasificación para métodos de *WSD* de acuerdo a los recursos que utilizan, ver figura 2 [Torres, 2009].

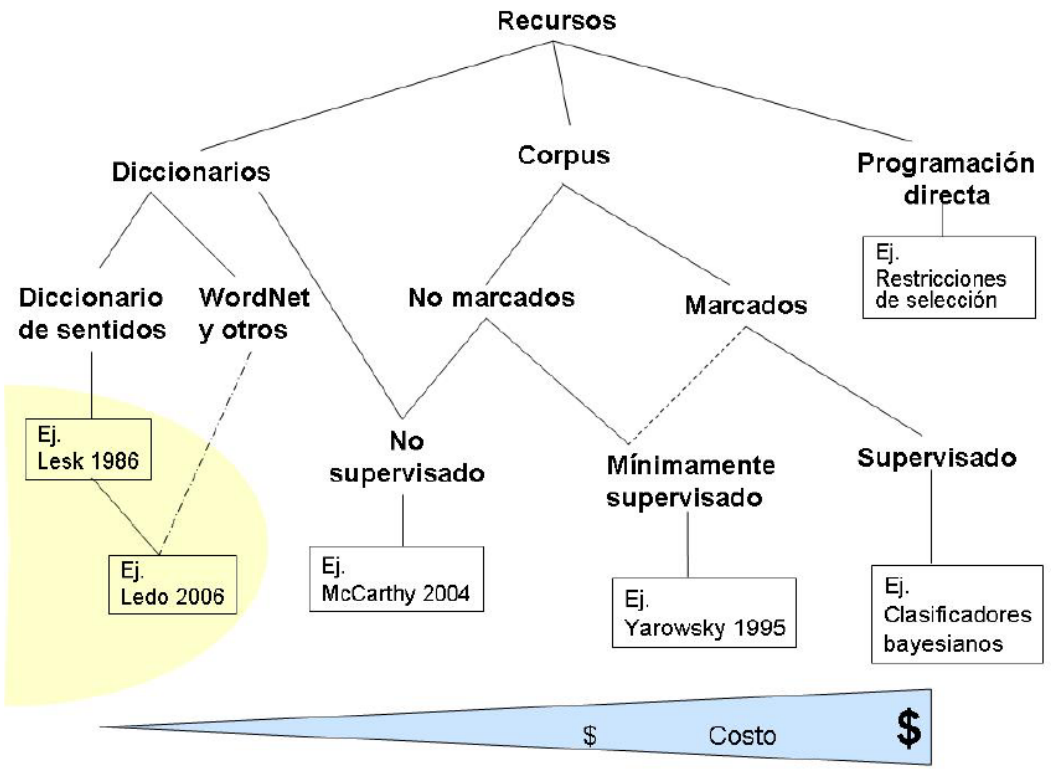


Figura 2. Clasificación de los métodos para *WSD* de acuerdo a los recursos que utilizan [Torres, 2009].

El uso de un diccionario en la desambiguación del sentido de las palabras es de suma importancia, ya que ésta consiste en asignar a cada palabra en un texto dado, un sentido que se relaciona con una lista de sentidos en un diccionario [Tejada, 2006].

WordNet es un sistema de referencia léxica, el cual fue desarrollado en la universidad de Princeton bajo la dirección del profesor George A. Miller [Tejada, 2006]. Este recurso combina muchas características usadas para desambiguación del sentido de las palabras un solo sistema. *WordNet* incluye definiciones de sentidos de palabras como un diccionario, define *synsets* o conjuntos de sinónimos, los cuales representan un concepto léxico, y además proporciona las relaciones existentes entre palabras. Está conformado por cuatro bases de datos correspondientes a sustantivos, verbos, adjetivos y adverbios. Cada palabra es asociada con un conjunto de sentidos (ver capítulo 4) [Torres, 2006].

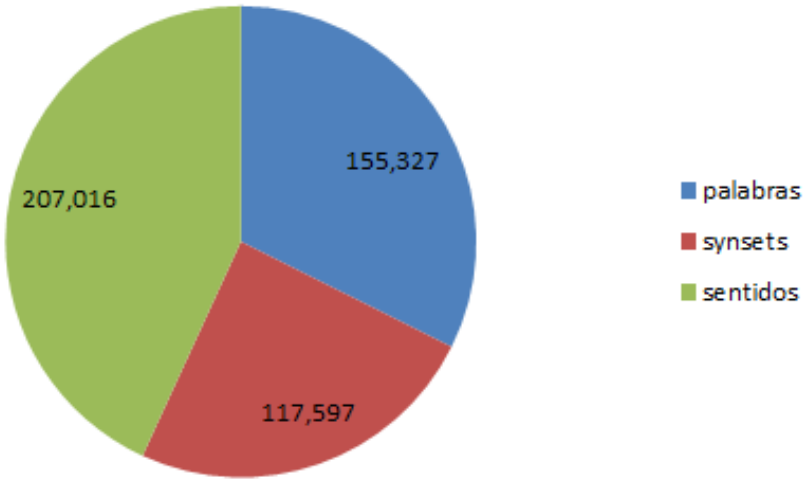


Figura 3. Estructura de *WordNet* versión 2.1.

2.4 Métodos de la desambiguación del sentido de las palabras

Actualmente existen dos categorías principales para la clasificación de los métodos empleados para la ***WSD***: métodos basados en conocimiento y métodos basados en corpus.

2.4.1 Métodos basados en diccionarios

Los diccionarios pueden ser de sentidos y otros como ***WordNet***. Los diccionarios proporcionan una lista de glosas (definición de sentido) para las palabras. Los métodos que utilizan sólo diccionarios de sentidos, buscan elegir un sentido (de esta lista) para cada palabra en un texto dado, tomando en cuenta el contexto en el que aparece. Como ejemplo, ***Lesk*** (1986) propone utilizar la coherencia global del texto, es decir, el total de sentidos de palabras relacionadas en el texto: mientras más relacionadas estén las palabras entre sí, más coherente será el texto. Además existen variantes del algoritmo de ***Lesk*** que utilizan no sólo diccionarios de sentidos, sino también otro tipo de diccionarios como ***WordNet.***

2.4.2 Métodos basados en corpus

2.4.2.1 Métodos supervisados

Los métodos que utilizan corpus marcados son los métodos supervisados. Éstos reducen la desambiguación de sentidos de palabras a un problema de clasificación, donde a una palabra dada se le asigna el sentido más apropiado de acuerdo a un conjunto de posibilidades, basadas en el contexto en el que ocurre. Hay muchos algoritmos de aprendizaje supervisado utilizados para ***WSD***, como ejemplo tenemos los clasificadores bayesianos, máquinas de soporte vectorial, árboles y listas de decisión, etc.

Hay métodos que utilizan una gran cantidad de corpus no marcados y muy pocos marcados llamados mínimamente supervisados. Como ejemplo de éstos tenemos el método de *Yarowsky* (1995), el cual identifica todas las ocurrencias de una palabra a desambiguar en un corpus no marcado. Después identifica un número pequeño de colocaciones semilla representativos de cada sentido de la palabra y etiqueta todos los ejemplos que contienen la colocación semilla con la palabra de dicha colocación (así tenemos los conjuntos etiquetados con cada sentido representativo y el conjunto residuo).

El algoritmo utiliza los conjuntos etiquetados para entrenar una lista de decisión y encontrar nuevas colocaciones, para después etiquetar sobre el conjunto residuo. El algoritmo termina cuando el conjunto residuo se estabiliza.

2.4.2.2 Métodos no supervisados

Los métodos que utilizan corpus no marcados son los no supervisados, estos métodos también utilizan otros recursos como *WordNet* para poder asignar un sentido a cada palabra que aparece en los textos no marcados. Como ejemplo de éstos tenemos el método de [McCarthy, 2004], el cual elige de un diccionario (tesauro) las palabras relacionadas con la palabra a desambiguar. Cada palabra relacionada tiene un peso, éstas y la palabra a desambiguar tienen sentidos en un diccionario. Para elegir el sentido correcto, las palabras relacionadas votan por un sentido de la palabra a desambiguar con cierto peso. Se elige el sentido con más peso.

2.4.3 Métodos de rankeo

El rankeo basado en grafos es básicamente una manera de decidir la importancia de un vértice dentro de un grafo, con base en la información generada por la estructura del grafo.

El *PageRank* es uno de los más populares algoritmos de rankeo y fue diseñado como un método de análisis de los enlaces en la *Web*.

La idea básica utilizada en el modelo de rankeo basado en grafos es el voto. Cuando un vértice esta enlazado a otro, este es un voto para el otro vértice. El más alto número de votos asignado a un vértice es la importancia del vértice [Mihalcea, 2006].

2.4.5 Métodos basados en tesauros/ontologías

Las categorías semánticas de las palabras del contexto determinan la categoría de todo el contexto, y a su vez el sentido de las palabras que lo conforman.

Algoritmo de *Walter* (1987).

- Cada palabra tiene uno o varias categorías semánticas que corresponden a sus diferentes sentidos.
- Para cada categoría semántica se cuentan las palabras que la disparan.
- Se selecciona la categoría semántica con mayor conteo. Ésta establece el sentido de todas las palabras.

Una ontología de relaciones léxicas, se organiza en conjunto de sinónimos:

- Una palabra que tiene varios sentidos pertenece a varios *synsets* (conjunto de sinónimos)
- Considera varias relaciones de *synsets* (hiponimia, hiperonimia, meronimia, etc).
- Incluye información sobre sustantivos, verbos, adjetivos y adverbios.

2.4.6 Métodos basados traducción

La desambiguación de las palabras puede realizarse considerando sus traducciones en otros lenguajes. Existe el algoritmo de [Dangan-Itai, 1991], el cual tiene que identificar en un corpus en un segundo lenguaje todas las traducciones de las palabras que se desea desambiguar. Así como contar las veces que cada traducción ocurre junto a las traducciones de las palabras del contexto y por último seleccionar el sentido con la mayor cuenta.

2.5 Stopwords

Las palabras que son demasiado frecuentes en los documentos de una determinada colección no aportan información relevante. De hecho, se considera que una palabra que aparece en al menos el 80% de los documentos de una determinada colección carece de utilidad, pues generalmente se trata de preposiciones, conjunciones, artículos, etc. Estas palabras se consideran vacías y normalmente se eliminan para evitar que puedan ser consideradas como potenciales [Moreiro, 2002].

Existe para cada idioma un conjunto de palabras vacías, comunes a todos los dominios, fácilmente identificable: artículos, preposiciones, conjunciones, etc., aunque también puede haber verbos, adverbios y adjetivos [Moreiro, 2002]. Una herramienta de pre-procesamiento consiste en eliminar las stopwords o palabras vacías del texto.

2.6 Lematización

La lematización es hallar el lema (raíz) de las palabras, el cual no tiene que tener significado. Un algoritmo de lematización es un proceso que trata de minimizar la información lingüística. En este algoritmo, las diferentes formas que puede adoptar una palabra son reducidas a una única forma común, a la cual se denomina *stem* o lema. *Stem* o lema es la porción de la palabra después de eliminar sus afijos. Por ejemplo, perr para las palabras perro, perros, perrito, perrote, etc. [Kryscia, 2007].

A diferencia de la lematización, el *stemming* es un proceso de truncamiento utilizado para gestionar de manera automática las diferentes formas de una palabra. Así, un algoritmo de este tipo extrae los pseudo-sufijos, las terminaciones de una palabra, y crea una pseudo-raíz de la misma [Valder, 2004].

Los primeros algoritmos de *stemming* se desarrollaron para el idioma inglés, pero esta técnica ha sido adaptada para el español. El algoritmo Porter [Peinado, 2003] es el más utilizado para el idioma inglés. También existen algoritmos para otros idiomas tales como el francés, el español, el holandés, el griego y el latín. En general, estos algoritmos se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común [Deco, 2007].

El algoritmo de Porter es un lematizador línea secuencial, considerado uno de los mejores y más conocidos. Este remueve en cinco pasos controlados más de 60 terminaciones, removiendo las terminaciones cortas sin excepciones.

Cada paso resulta en la remoción de una terminación o la transformación de la raíz [Peinado, 2003].

2.7 Secuencias Frecuentes Maximales

Una secuencia se considera que es frecuente (*SF*) si aparece al menos β veces en el documento, donde β es el umbral de frecuencia dado. El número de secuencias frecuentes existentes en un documento puede ser muy grande, de manera que una forma de reducir el conjunto de las *SF's* encontradas es tomar en cuenta solo aquellas que no son subsecuencia de alguna otra *SF*; es decir, las *SF's* que son maximales [Hernández, 2007].

2.8 PageRank

La idea básica utilizada en el modelo de rankeo basado en grafos es el voto. Cuando un vértice está enlazado a otro, este es un voto para el otro vértice. El más alto número de votos asignado a un vértice es la importancia del vértice.

El rankeo basado en grafos es básicamente una manera de decidir la importancia de un vértice dentro de un grafo, con base en la información generada por la estructura del grafo.

El *PageRank* es uno de los más populares algoritmos de rankeo y fue diseñado como un método de análisis de los enlaces en la Web.

El *PageRank* en el algoritmo de rankeo, para saber que vértice (sentido) tuvo mayor valor (importancia) y con ello poder saber qué sentido es el correcto con respecto a la palabra en la oración. Para cada uno de los nodos del grafo deben ser numerados comenzando por cero, el archivo que se va a generar debe inicia con la primera línea que indica el número de nodos, seguido por el número de aristas [Mihalcea, 2006].

$$PR(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|}$$

d **=** 0.85 va de (0-1)
|| **=** valor absoluto
Out **=** grado de salida
In **=** enlaces
V_i **=** vértice entrante
V_j **=** vértice saliente
PR **=** Page Rank

Figura 4. Formula del *PageRank* [Mihalcea, 2006].

2.9 Ranking

Los sistemas requieren que el usuario proporcione suficientes restricciones sintácticas en su consulta para limitar el número de documentos recuperados, y esos documentos recuperados no se clasifican por orden de cualquier relación con la consulta del usuario.

El enfoque *ranking* en la recuperación está orientado para los usuarios finales. Este enfoque permite que el usuario introduzca una consulta simple como una oración o una frase y de esta manera recuperar una lista de documentos clasificados por orden de probabilidad en relevancia.

La razón principal del enfoque *ranking* es porque es más eficaz ya que todos los términos de la consulta se utilizan para la recuperación y los resultaos se clasifican con base en la concurrencia o similaridad de los términos de la consulta. Sin embargo, algunas métricas utilizan la estadística término ponderación para hacer el cálculo de similaridad [Baeza-Yates, 1992].

2.10 Métricas de similitud en conjuntos

En esta sección, se describen métricas de similitud, que consideran un texto como un conjunto de cadenas y son utilizadas para desambiguar un término.

2.10.1 Similitud Coseno

El algoritmo de similitud coseno es utilizado como un método de desambiguación y tienen bastante éxito en diferir nombres de entidades, también este algoritmo puede ser visto como un clasificador de documentos dada una consulta. La similitud coseno se emplea en la búsqueda y recuperación y está representado en el modelo estándar de espacio vectorial donde cada componente corresponde a un término del vocabulario [Bunescu, 2006]. El cálculo de la similitud se obtiene entre dos vectores en un espacio que posee un producto interior con el que se evalúa el valor del coseno del ángulo comprendido entre ellos. Cuando la función de coseno es de 0° el valor arrojado es 1, los valores se utilizan sobre todo en el espacio positivo, donde el resultado está limitado entre [0,1] [Baeza, 1999].

En la representación de un documento o una consulta como un vector, un peso debe asignarse a cada término que representa el valor del componente en el vector correspondiente. La ponderación de los términos que se encuentran en el modelo de espacio vectorial se calcula con *TF*IDF* que es mucho más preciso que la limitada modalidad de peso binario. Pero al igual que en los pesos binarios solo se calcula *TF*IDF* a aquellos términos que tiene coincidencia en el conjunto de documentos y la consulta [Buttcher, 2010]. Dado dos vectores de términos *X* & *Y*, la similitud coseno, *cos* (θ), se representa usando un producto de punto y magnitud como:

$$similaridad = \cos(\theta) = \frac{\vec{x}}{|x|} \cdot \frac{\vec{y}}{|y|} = \frac{\sum_{i=1}^{|v|} x_i \, y_i}{\left(\sqrt{\sum_{i=1}^{|v|} (x_i)^2}\right)\left(\sqrt{\sum_{i=1}^{|v|} (y_i)^2}\right)}$$

Figura 5. Formula de similaridad de coseno.

2.10.2 Similitud de Dice

El cálculo de coeficiente Dice fue desarrollado por Lee Raymond Dice [Dice, 1945] y se centra en usos estadístico, pero también es utilizada en la recuperación de información para obtener la similitud entre conjuntos. La similitud de Dice puede definirse como el tamaño de la intersección colocada en el numerador dividido por el tamaño de la unión de la muestra.

El coeficiente de Dice es utilizado en el procesamiento del lenguaje natural almacenado en índices, permitiendo tener una búsqueda más completa. El uso de este algoritmo de similitud puede tener valores en los resultados entre 0 y 1, en muchos casos los archivos tienen una similitud cercana o igual a cero. Por esta razón, los umbrales son generalmente aplicados en el proceso de búsqueda. Los umbrales son un valor en que la medida de similitud debe ser igual o superior o también se puede considerar como un número que limita los elementos obtenidos de acuerdo a su puntuación [Kowalski, 1997]

Cuando el algoritmo se utiliza como una medida de similitud de caracteres, el coeficiente puede calcularse por dos cadenas *C1* y *C2* usando bigramas de la siguiente manera [Rao, 2013].

Donde n_t es el número de bigramas encontrados en ambas cadenas n_x es el número de bigramas en la cadena x, n_y es el número de bigramas en la cadena *y*.

2.10.3 Similitud de Jaro-Winkler

Jaro desarrolló una función de similitud que define la transposición de dos caracteres como la única operación de edición permitida. Winkler ayudó a normalizar esta similitud por eso se conoce como similitud Jaro-Winkler. En esta similitud, los caracteres no necesitan ser adyacentes, sino que puede estar alejada cierta distancia que depende de la longitud de ambas cadenas. La similitud de Jaro tiene un orden de complejidad de *0(n)*.

Cuanto mayor sea la distancia Jaro-Winkler en dos cadenas, más similares son las cadenas, la métrica de distancia Jaro-Winkler está diseñada y es más adecuada para cadenas cortas tales como nombres de personas. La puntuación se normaliza tal que *0* equivale a ninguna similitud y *1* es una coincidencia exacta [Uribe, 2010].

La medida de Jaro [Gelbukh, 2009] busca la similitud entre dos cadenas de longitud de caracteres $|a|$ y $|b|$, tiene un espacio y tiempo de complejidad de solamente $0(|a| + |b|)$. Se considera que el número de caracteres en C común en las dos cadenas y el número de transposiciones *t* como:

$$sim_{jaro}\ (a,b) = \frac{1}{3}\ (\frac{c}{|a|} + \frac{c}{|b|} + \frac{c-t}{c})$$

Figura 6. Formula de la medida de Jaro

2.10.4 Distancia de Edición

Es una modificación del modelo original conocido como distancia de Levenstein [Levenshtein, 1966]. La distancia de edición es el número mínimo de operaciones (inserción, eliminación y sustitución) obligatoria para

modificar una cadena *C1* en una cadena *C2*. Estas oraciones son a nivel de carácter y si el resultado es más cercano a 0 las cadenas son más similares [Manning, 2009]. Las operaciones son:

- *Remplazar un carácter de A por otro de B*
- *Eliminar un carácter de A o B*
- *Insertar un carácter de B en A [Uribe, 2010]*

Las medidas de distancia de edición puede ser normalizado en el intervalo *[0,1]* dividiendo el número total de operaciones por el número de caracteres de la cadena más larga. Una vez normalizado, la distancia de edición se puede convertir a la similitud restando el valor de la distancia con el número 1.

2.10.5 Distancia de Hamming

El algoritmo de Hamming toma las cadenas como un vector de caracteres y cada vector es una palabra representada en código binario de acuerdo a la ausencia o presencia de un carácter. El código se obtiene mediante la adición de una comprobación de paridad, si hay un error, se puede detectar por la suma impar de los dígitos. Entre mayor sea la suma de los dígitos impares mayor es la dificultad de convertir un código valido en otro. A esta diferencia se le llama distancia de Hamming [Rao, 2013].

Por ejemplo: *El código binario 11011011010110 = tiene una distancia de 5.*

La palabra "banco central" y las palabras "banco mundial"= tiene una distancia de 7.

2.11 Medidas de evaluación

Las medidas de evaluación son calculadas por el promedio de los valores obtenidos para cada solicitud con respecto al número de consultas que se han hecho. Estas medidas son usadas para satisfacer las necesidades del usuario en cuanto a relevancia de los documentos recuperados, cada métrica utiliza parámetros distintos para medir la efectividad de los sistemas de recuperación de información [Pérez, 2012].

Dos medidas de evaluación frecuentemente utilizadas en el área de la recuperación de información son la precisión y el recuerdo, las cuales permiten conocer los objetos relevantes dentro de una colección. El recuerdo está definido como la probabilidad de detectar un objeto dado que es relevante, mientras que la precisión se define como la probabilidad de que un objeto es relevante dando que fue detectado [Zhu, 2004].

Se define como correctas al número de oraciones extraídas por el sistema y por el humano; incorrectas como el número de oraciones extraídas por el sistema pero no por el humano; y olvidadas como el número de oraciones extraídas por el humano pero no por el sistema. Así entonces tenemos [Villat, 2006]. De esta forma, se dice que la precisión refleja cuantas de las oraciones extraídas por el sistema fueron buenas, y el recuerdo refleja cuantas de las oraciones buenas olvido el sistema.

Precisión *P* = ejemplos clasificados correctamente/ejemplos clasificados
Recuerdo *R* = ejemplos clasificados correctamente/total de ejemplos

$$\mathrm{P} = \frac{clasificados\ correctamente}{correctas + incorrectas} \qquad R = \frac{clasificados\ correctamente}{correctas + olvidadas}$$

Figura 7. Fórmula de Precisión y Recuerdo [Villat, 2006].

CAPÍTULO 3

Estado del Arte

En este capítulo, se habla sobre los trabajos relacionados de los métodos basados en la *WSD* que han sido utilizados en el problema de la desambiguación del sentido de las palabras.

Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization.
Rada Mihalcea; 2004 [Mihalcea, 2004]

Mihalcea propone un método innovador para la extracción de oraciones, llamado *TextRank*, el cual consiste en un algoritmo de *ranking* basado en grafos. Básicamente, un algoritmo de *ranking* basado en grafos es una forma de decidir acerca de la importancia de un vértice dentro de un grafo, considerando la información global del mismo.

De acuerdo con Pajares [Pajares, 2006], un grafo es un conjunto de nodos y arcos (vértices) que unen esos nodos. Cada nodo representa un elemento que equivale a una situación valida. Dado lo anterior, Mihalcea [Mihalcea, 2004] construye un grafo para representar el texto, de manera que los nodos son palabras (u otras entidades de texto) interconectadas mediante vértices con relaciones significativas. Para la tarea de extracción de oraciones, el objetivo es calificar oraciones enteras y ordenarlas de mayor a menor calificación. Por lo tanto, se agrega un vértice al grafo por cada oración en el texto.

Para establecer las conexiones (ciclos) entre oraciones, se define una relación de similitud, donde la relación entre dos oraciones puede ser vista como un proceso de "recomendación", una oración que señala a cierto concepto en el texto da al lector una "recomendación" para referirse a otras oraciones en el texto que señalan a los mismos conceptos y por tanto, un vínculo puede establecerse entre dos oraciones cualesquiera que compartan un contenido común.

Random Walks on Text Structures.
Rada Mihalcea; 2006 [Mihalcea, 2006]

Mihalcea propone un algoritmo de rankeo basados en grafo, que es básicamente una manera de decidir la importancia de un vértice dentro de un grafo, en base a la información generada por la estructura del grafo. La idea básica utilizada en el modelo de rankeo basado en grafos es el voto o recomendación. Cuando un vértice esta enlazado a otro, este es un voto para el otro vértice. El más alto número de votos asignado a un vértice es la importancia del vértice.

Los algoritmos de rankeo de grafos están basados en el modelo del camino aleatorio, donde se dan pasos aleatorios en un grafo *G*, comenzando el

modelado del camino, llamado proceso de Marcov. Bajo ciertas condiciones este modelo converge a la distribución estacionaria de probabilidades, asociada a los vértices del grafo.

Sea *G= (V,* E) un grafo direccionado con un conjunto de vértices *V* y un conjunto de aristas *E*, para cada vértice *Vi*, *In(Vi)* es el conjunto de vértices que apuntan a él y *Out(Vi)* es el conjunto de vértices que son apuntados desde el. Se describen 2 algoritmos de rankeo basado en grafos: *HITS* y *PageRank*.

El Algoritmo *HITS (Hyperlinkinked Introduced Topic Search*) es un algoritmo iterativo que fue diseñado para la clasificación de páginas Web de acuerdo a su grado de autoridad. Los algoritmos *HITS* hace una distinción entre autoridades (páginas con un gran número de enlaces de entrada) y repartidores (*hubs*) (páginas con un gran número de enlaces de salida).

El Algoritmo *PageRank*, es quizá uno de los más populares algoritmos de rankeo y fue diseñado como un método de análisis de los enlaces en la Web. El rankeo de páginas integra los enlaces de entrada y los de salida en un solo modelo, y genera un solo conjunto de grados.

Se analiza los algoritmos de rankeo basado en grafos para el procesamiento de texto en dos niveles diferentes: a nivel de oraciones y a nivel de documento. El procesamiento de texto a nivel de oraciones, detecta ambigüedad: esta tarea consiste en asignar el significado más apropiado a una palabra polisémica, dentro de un contexto. Para habilitar la aplicación de algoritmos de rankeo basada en grafos en la desambigüedad de todas las palabras en un texto restringido, tenemos que construir un grafo que represente el texto y conectar las palabras en relación a su significado.

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

Grettel Barceló Alonso; 2010 [Alonso, 2010]

Dentro el problema de la desambiguación de los sentidos de las palabras es determinar el significado correcto, o el sentido de una palabra dada en el contexto. Este se considera como uno de los problemas más difíciles en el nivel léxico del procesamiento del lenguaje natural.

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.

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

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, los módulos principales del sistema se muestran el la figura 8.

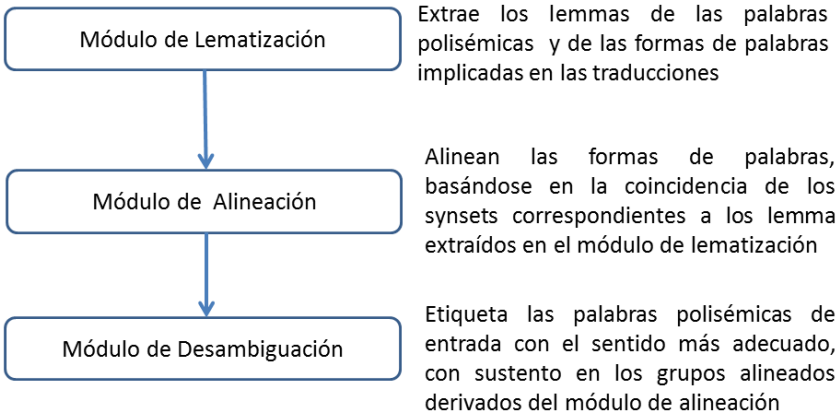


Figura 8. Módulos principales del sistema.

La desambiguación involucra la asociación de una palabra ambigua dada en un texto, con una definición o significado (sentido), que es distinguible del resto de los significados potencialmente atribuibles a dicha palabra.

Existen tres métodos fundamentales de representar los sentidos de las palabras:

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 = nger

dedo = toe

Con respecto al contexto donde la palabra ocurre (discriminación):

\Me apunto con el dedo"

\El zapato me lastima el dedo"

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

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:

planta1 = Parte inferior del pie.
planta2 = Árbol u hortaliza que, sembrada y nacida en alguna parte, está dispuesta para trasplantarse en otra.
planta3 = Fabrica central de energía, instalación industrial.

2) El desarrollo de un algoritmo que asigne el sentido correcto a la palabra para un determinado contexto, como se muestra a continuación.

\Científicamente se conoció la planta de hierba mate en Europa desde principios del siglo XIX.

El algoritmo que se propone está basado en dos recursos principales:
1) WordNet como léxico especializado para cada uno de los idiomas involucrados.
2) Textos paralelos como información adicional para proporcionar diversa lexicalizaciones de las palabras polisémicas.

La arquitectura del sistema de desambiguación implementado, ha sido concebida como la integración, en una herramienta, de tres módulos principales como es la *Lematización, Alineación y Desambiguación.*

Optimización global de coherencia en la desambiguación del sentido de las palabras

Sulema Torres Ramos; 2009 [Torres, 2009]

Los métodos para desambiguación de sentidos de palabras se clasifican en: los que utilizan diccionarios, los que utilizan corpus y los que no utilizan ningún recurso léxico.

Los que utilizan diccionarios: Los diccionarios pueden ser de sentidos y otros como *WordNet*. Los diccionarios proporcionan una lista de glosas (definición de sentido) para las palabras. Los métodos que utilizan solo diccionarios de sentidos, buscan elegir un sentido (de esta lista) para cada palabra en un texto dado, tomando en cuenta el contexto en el que aparece.

Además existen variantes del algoritmo de *Lesk* que utilizan no solo diccionarios de sentidos, sino también otro tipo de diccionarios como *WordNet.*
Los que utilizan corpus: Los corpus pueden ser no marcados y marcados.

El algoritmo de *Lesk* (1986) es uno de los primeros algoritmos exitosos usados en la desambiguación de sentidos de palabras. El primero es acerca de desambiguar palabras, considerando la optimización global del texto, esto es, encontrar la combinación de los sentidos que maximice la relación total entre los sentidos de todas las palabras.

Algoritmo de *Lesk* Simplificado: Para reducir el espacio de búsqueda del algoritmo de *Lesk Si*mple o *Lesk* Simplificado, es donde los sentidos de las palabras en el texto son determinados uno a uno encontrando el mayor traslape entre los sentidos de las definiciones de cada palabra con el contexto actual.

En lugar de buscar, asignar, simultáneamente, el significado de todas la palabras en un texto dado, este enfoque determina el sentido de las

palabras uno a uno, por lo que se evita la explosión combinatoria de sentidos.

Para medir el desempeño de un sistema de desambiguación de sentidos de palabras, se utilizan medidas de evaluación que son: Coverage (cobertura), Precisión (precisión) y Recall (recuerdo).

- Coverage determina el porcentaje de casos cubiertos por el sistema. Esta dado por el número de veces que el sistema asigno un sentido entre el total de casos.
- Precisión es usado para medir la exactitud o fidelidad del algoritmo. En el caso de desambiguación del sentido de las palabras está definido por, el total de sentidos correctos entre el total de casos cubiertos por el sistema.
- Recall está definida por el total de sentidos correctos sobre el total de casos.

En este trabajo, se utilizan como algoritmos para desambiguación de sentidos de palabras *Lesk* Completo y *Lesk* Simple, además del uso de dos estrategias de back-off; sentido aleatorio y sentido más frecuente.

Para llevar a cabo la desambiguación, los métodos tipo *Lesk* requieren el uso de un diccionario de sentidos llamado *WordNet*.

La medida de similitud utilizada por ambos algoritmos es la medida original de *Lesk*. La ventana de contexto para ambos algoritmos es la oración. Para la optimización del método de *Lesk* Completo se utiliza un método de optimización conocido como Algoritmos con Estimación de Distribuciones.

Desambiguación de sentidos de palabras usando sinónimos

Miguel Ángel Gaona Ríos; 2007 [Gaona, 2007]

El problema que surge, es que en diferentes contextos, la misma palabra puede tener diferentes significados: Juan trabaja en un banco, Juan está sentado en un banco.

En la tarea de *WSD*, la computadora debe determinar cuál de las acepciones que tiene la palabra en el diccionario, es la que el autor tenía en mente. La *WSD* es en esencia una tarea de clasificación: Los sentidos de la palabra son las clases, el contexto provee la evidencia y cada ocurrencia de una palabra es asignada a una o más de las posibles clases basado en la evidencia.

Se implementan dos métodos no supervisados para la solución de la *WSD* y una aplicación de uno de estos métodos. A saber, las aportaciones principales son:

Un método no supervisado para la *WSD*, que se basa en el método de Mihalcea y Moldovan (1999). El método hace uso de la Web y de un diccionario de las palabras semánticamente cercanas a la palabra ambigua.

Un método no supervisado para la *WSD*, que modifica el algoritmo de *Lesk* (1986). El método se basa en el uso de las estadísticas de las coocurrencias de las palabras. Estas estadísticas se colectan a través de la Web. El método modificado proporciona mejor precisión que el método original.

Dentro de la evaluación en la calidad de un sistema de *WSD* se suele cuantificar mediante dos medias de precisión, la precisión absoluta y la precisión relativa. La precisión absoluta (*recall*, *R*) es la métrica básica para decidir la calidad de la tarea de *WSD* porque muestrea el número de casos

correctamente desambiguados sobre todos los casos de prueba. La precisión relativa *(precisión, P)* se define como el porcentaje de respuestas correctas sobre todas los casos tratados por el sistema de *WSD.* Esta medida favorece a los sistemas que obtienen una alta fiabilidad en la asignación de sentidos. Las métricas son realmente informativas si se interpretan conjuntamente. Un sistema puede tener una precisión relativa muy buena, pero una precisión absoluta muy baja, lo que significa que ha identificado la respuesta correctamente para muy pocos caos, pero con alta fiabilidad [Ted, 2000].

Tabla 1. Características de autores con respecto a diferentes métodos.

AUTORES CARACTERÍSTICAS	[Alonso, 2010]	[Torres, 2009]	[Som, 2008]	[Gaona, 2007]	[Pancardo, 2006]	[Mihalcea, 2006]
Tipo de foro/corpus	SENSEVAL-3	SENSEVAL-2	SENSEVAL-2 y SENSEVAL-3	SENSEVAL-2	SENSEVAL-3	SENSEVAL-2
Tipo de diccionario	MiniDir-2.1 /MultiWordNet	WordNet	WordNet	WordNet 2.1	WordNet 2.0	WordNet 2.1
Tamaño del corpus	12625 ejemplos etiquetados de fragmentos de la novela Don Quijote de la Mancha	2436 palabras ambiguas (de las cuales 1136 son sustantivos, 544 verbos, 457 adjetivos y 299 adverbios)	Tiene sustantivos, verbos, adjetivos y adverbios	La Web	215 sustantivos	3 archivos con un total de 238 oraciones y 2436 palabras para desambiguar (1136 sustantivos, 544 verbos, 457 adjetivos y 299 adverbios)
Idioma del corpus	Español	Inglés	Inglés	Inglés	Inglés	Inglés
Etiquetado del corpus	Corpus semánticamente etiquetado, del repositorio de sentido de MiniDir-2.1	Corpus etiquetado sintética y semánticamente		EuroWordNet para traducir cada palabra		
Tipo de tarea	Tarea de muestra léxica	Tarea de todas las palabras	Tarea de todas las palabras	Traducción palabra por palabra	Tarea léxica completa	Tarea léxica completa
Medidas de evaluación	Precisión (P) y Recuerdo (R)	Coverage (C) , Precisión (P) y Recuerdo (R)		Precisión (P) y Recuerdo (R)	Coverage (C) , Precisión (P) y Recuerdo (R)	Precisión (P) y Recuerdo (R)
Medidas de similitud		Lesk, traslape	Lesk y jcn			Lesk Sample
Método	supervisado y no supervisado	Lesk Completo y Lesk Simple	Lesk y Lesk Simple	Lesk Simple utilizando un back–off (escoger un sentido de manera aleatoria)	Lesk y Lesk Simple	Lesk Completo y Lesk Simple
Algoritmo		Algoritmos con Estimación de Distribuciones				Algoritmos de rankeo basados en grafos (HITS y PageRank)

CAPÍTULO 4.

Metodología de Trabajo

En este capítulo, se describe la metodología de trabajo para llevar a cabo el proceso para la desambiguación del sentido de las palabras. Se describen los pasos correspondientes a la metodología propuesta. Para llevar a cabo la desambiguación del sentido de las palabras, se requiere el uso de un diccionario de sentidos. En esta tesis, el diccionario utilizado es *WordNet* en la versión 2.1. La evaluación se hizo sobre el corpus etiquetado en el idioma inglés sobre la tarea léxica completa *english-all-words* del foro Senseval-2.

4.1 Metodología de trabajo

En la figura 7, se presenta el diagrama general de la metodología propuesta.

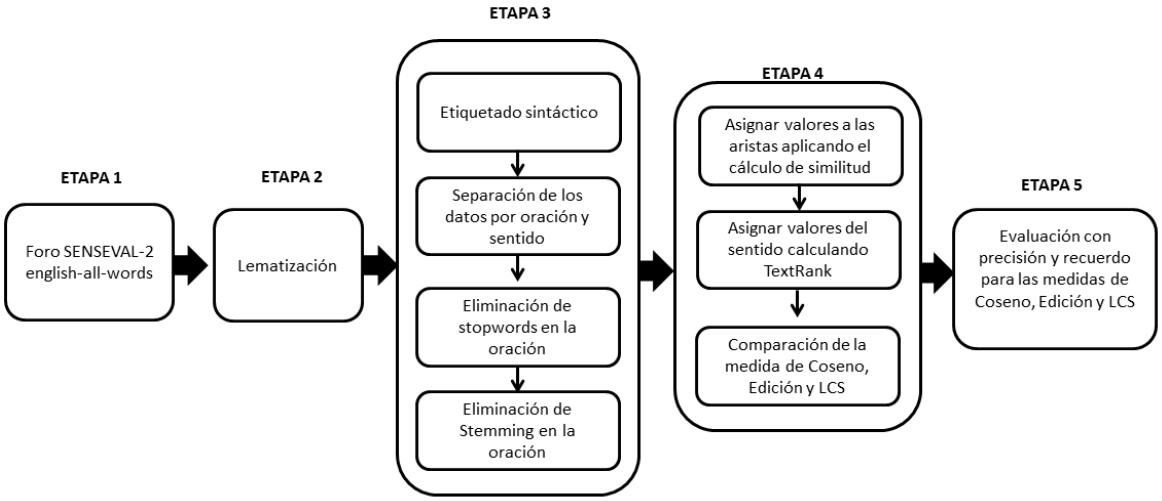


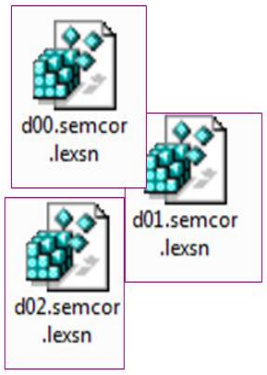
Figura 9. El esquema general de la metodología propuesta.

A continuación se presenta cada uno de las etapas de la metodología de trabajo propuesta.

Etapa 1. Datos de entrada

Los datos de la entrada son del foro SENSEVAL-2 de la tarea léxica completa que lleva por nombre *English-all-words*. Se constituye por los archivos d00.semcor.lexsn.key (D00), d01.semcor.lexsn.key (D01), d02.semcor.lexsn.key (D02).

En la figura 8, se presenta un ejemplo de una oración completa. Se marcan en verde los lemmas que se extraen por oración.



```
<context filename=d00 source=senseval2>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done id=d00.s00.t01 pos=NN lemma=art wnsn=3 lexsn=1:09:00::>art</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done id=d00.s00.t03 pos=NN lemma=change_ringing wnsn=1 lexsn=1:04:00::>change-ringing</wf>
<wf cmd=done id=d00.s00.t04 pos=VBZ lemma=be wnsn=1 lexsn=2:42:03::>is</wf>
<wf cmd=done id=d00.s00.t05 pos=JJ lemma=peculiar wnsn=2;4
lexsn=5:00:00:specific:00;5:00:00:characteristic:00>peculiar</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t08 pos=NNS lemma=english wnsn=2 lexsn=1:18:00::>English</wf>
<punc>,</punc>
<wf cmd=ignore pos=CC>and</wf>
<punc>,</punc>
<wf cmd=ignore pos=IN>like</wf>
<wf cmd=done id=d00.s00.t13 pos=JJS lemma=most wnsn=1 lexsn=3:00:02::>most</wf>
<wf cmd=done id=d00.s00.t14 pos=JJ lemma=english wnsn=1 lexsn=3:01:00::>English</wf>
<wf cmd=done id=d00.s00.t15 pos=NNS lemma=peculiarity wnsn=1;2 lexsn=1:09:00::;1:07:02::>peculiarities</wf>
<punc>,</punc>
<wf cmd=done id=d00.s00.t17 pos=JJ lemma=unintelligible wnsn=2
lexsn=5:00:00:incomprehensible:00>unintelligible</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t20 pos=NN lemma=rest wnsn=1 lexsn=1:24:00::>rest</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t23 pos=NN lemma=world wnsn=7 lexsn=1:14:02::>world</wf>
<punc>.</punc>
</s>
```

Figura 10. Ejemplo de los archivos English-all-words.

Etapa 2. Pre-procesamiento

- **Paso 1. Lematización**

Se extraen los lemas por oración (ver figura 9) de cada uno de los archivos d00.semcor.lexsn.key (D00), d01.semcor.lexsn.key (D01), d02.semcor.lexsn.key (D02). La oración original es tomada del archivo D00, se muestra de color naranja a continuación:

The art of change_ringing be peculiar to the english, and, like most english peculiarity, unintelligible to the rest of the world

art, change_ringing, be, peculiar, english, most, english, peculiarity, unintelligible, rest, world

Figura 11. Extracción de lemas.

Etapa 3. Preparación de los datos

- **Paso 1. Etiquetado Sintáctico**

Se etiquetan sintácticamente cada uno de los lemas dependiendo si es un sustantivo (n), verbo (v), adjetivo (a), adverbio (r), como se muestra en la figura 10, ver anexo 6 para el etiquetado.

art#n change_ringing#n be#v peculiar#a english#n most#a english#a peculiarity#n unintelligible#a rest#n world#n

Figura 12. Etiquetado sintáctico por oración.

- **Paso 2. Preparación de las oraciones por cada sentido**

Se generan las oraciones por cada uno de los sentidos correspondientes a cada uno de los lemas, de los archivos d00.semcor.lexsn.key (D00), d01.semcor.lexsn.key (D01), d02.semcor.lexsn.key (D02), como se muestra en la figura 11.

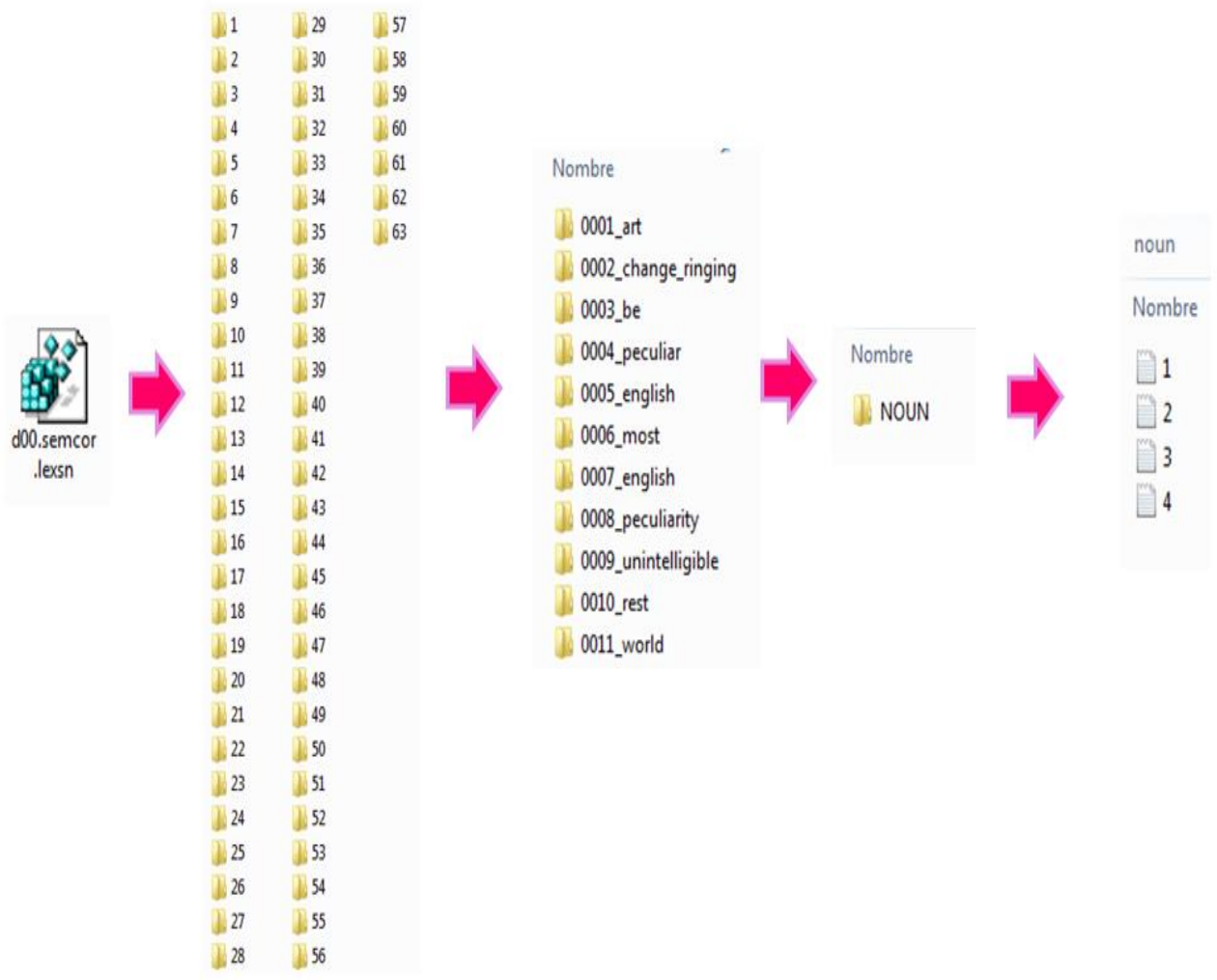


Figura 13. Sentidos correspondientes al lema y clasificación por etiquetas sintácticas.

- **Paso 3. Eliminación de stopwords en la oración**

Se eliminan las *stopwords* (palabras vacías) de cada de la oración.

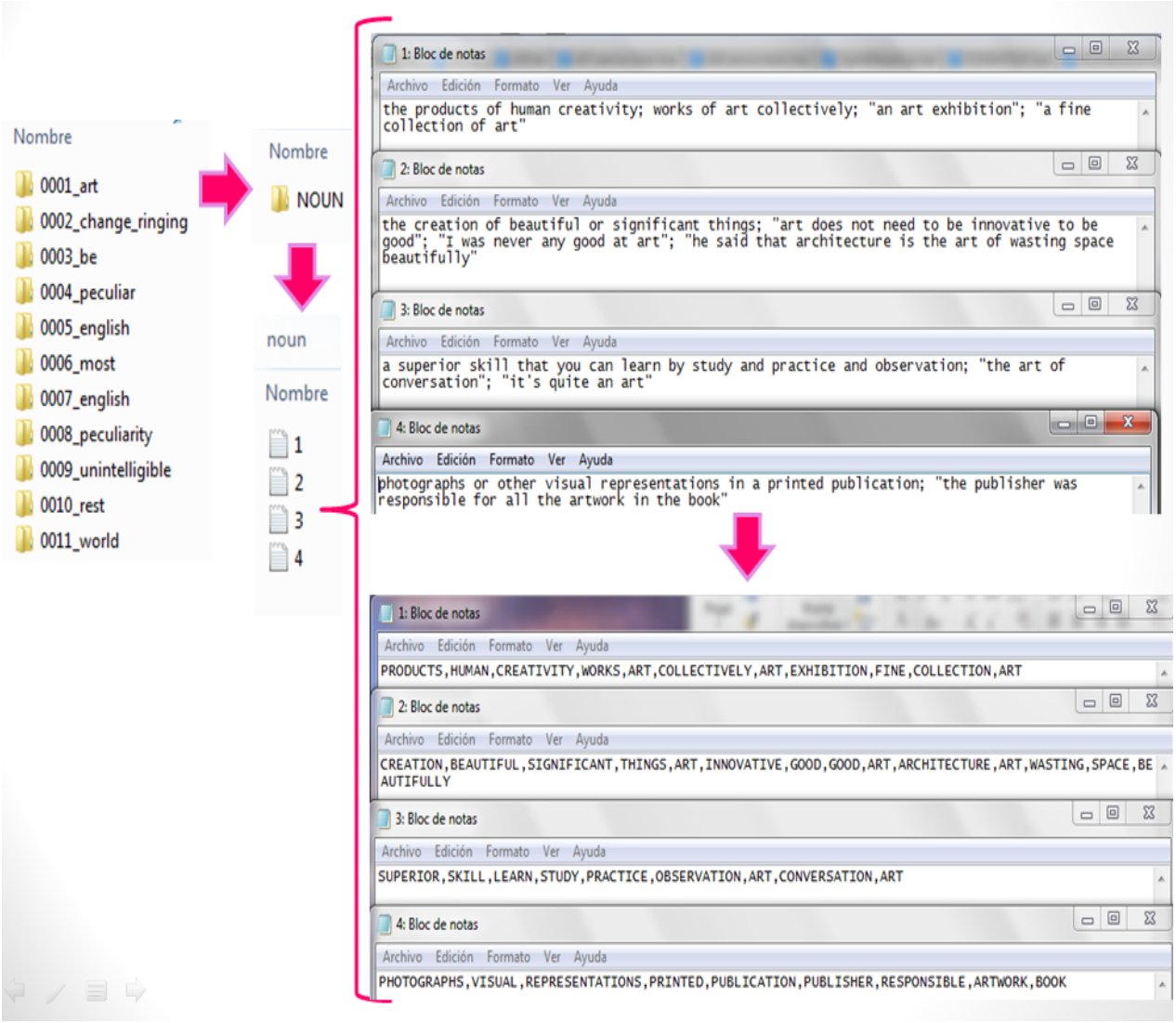


Figura 14. Eliminación de stopwords por archivo.

- **Paso 4. Eliminación de *stemming* en la oración**

Se aplica *stemming* (raíz de la palabra) a cada uno de los sentidos.

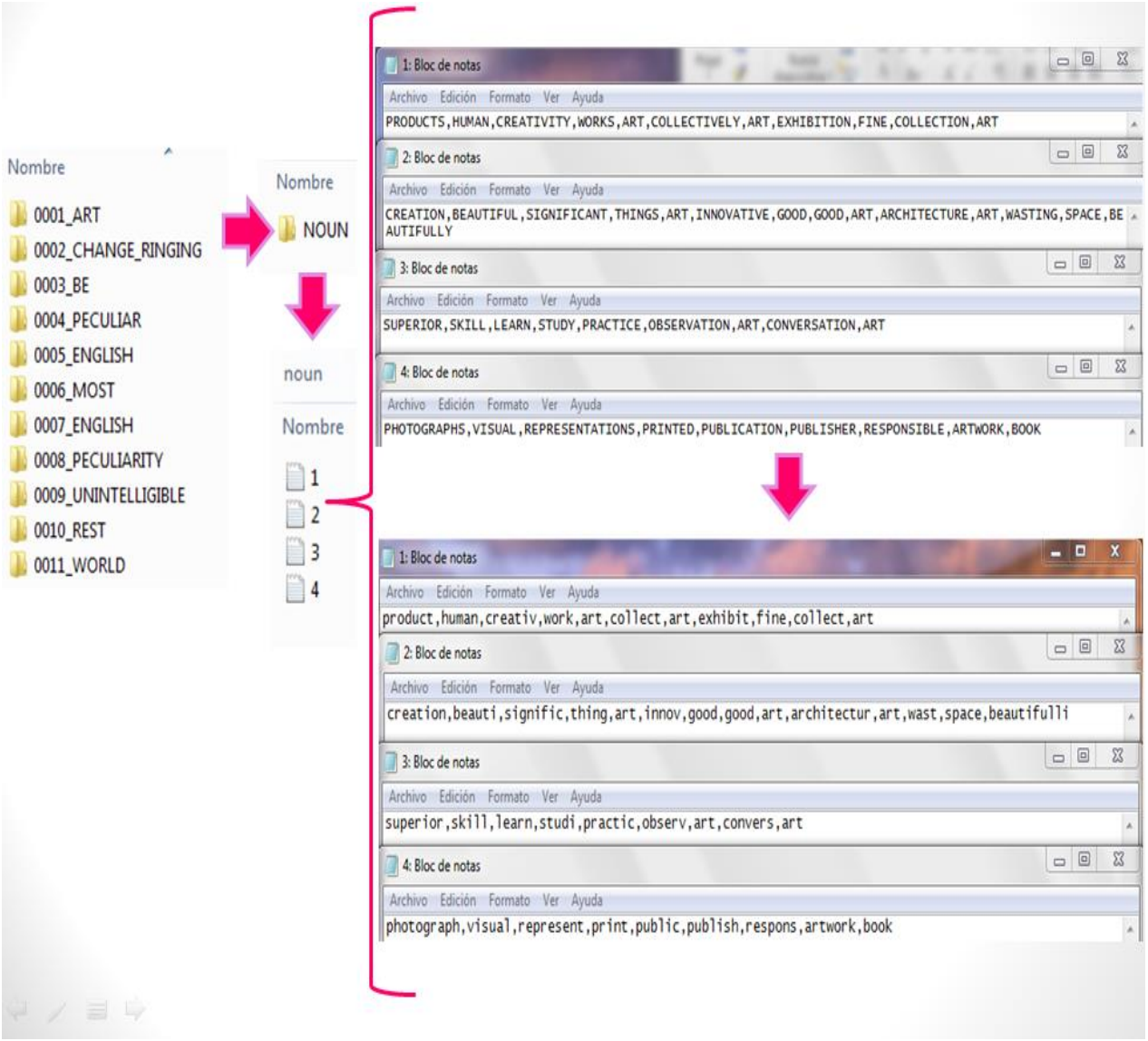


Figura 15. Eliminación de *stemming* por archivo.

Etapa 4. Similitud de sentidos

- **Paso 1. Asignar valores a las aristas aplicando cálculo de similitud**

Se aplica la medida de coseno, edición y LCS para sacar el valor de la similitud entre cada uno de los sentidos de acuerdo a *WordNet* (ver figura 14).

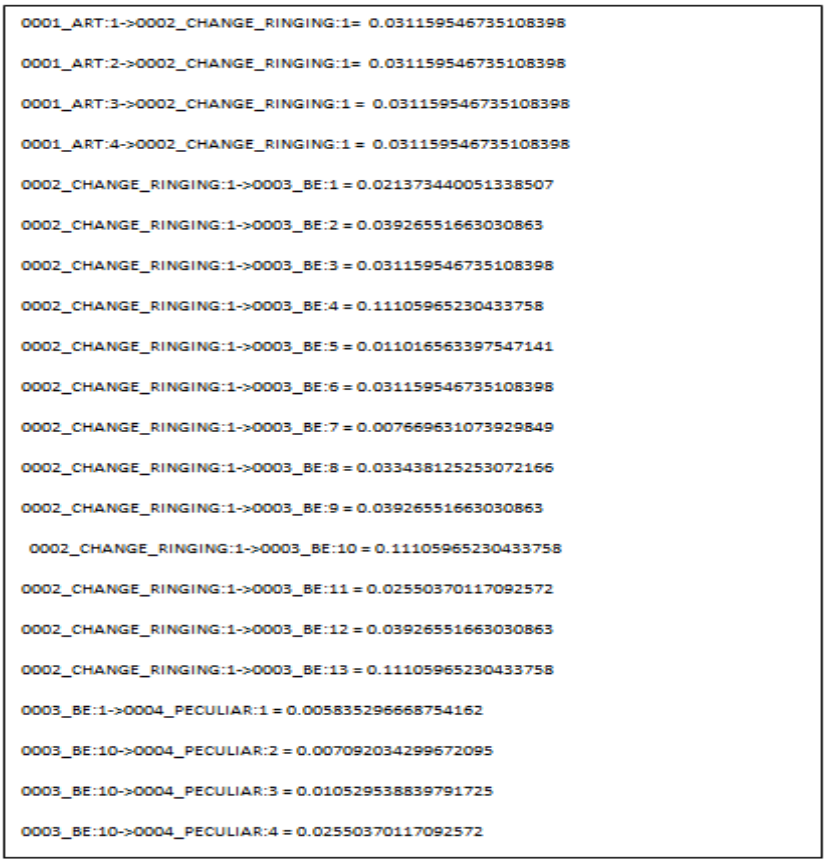
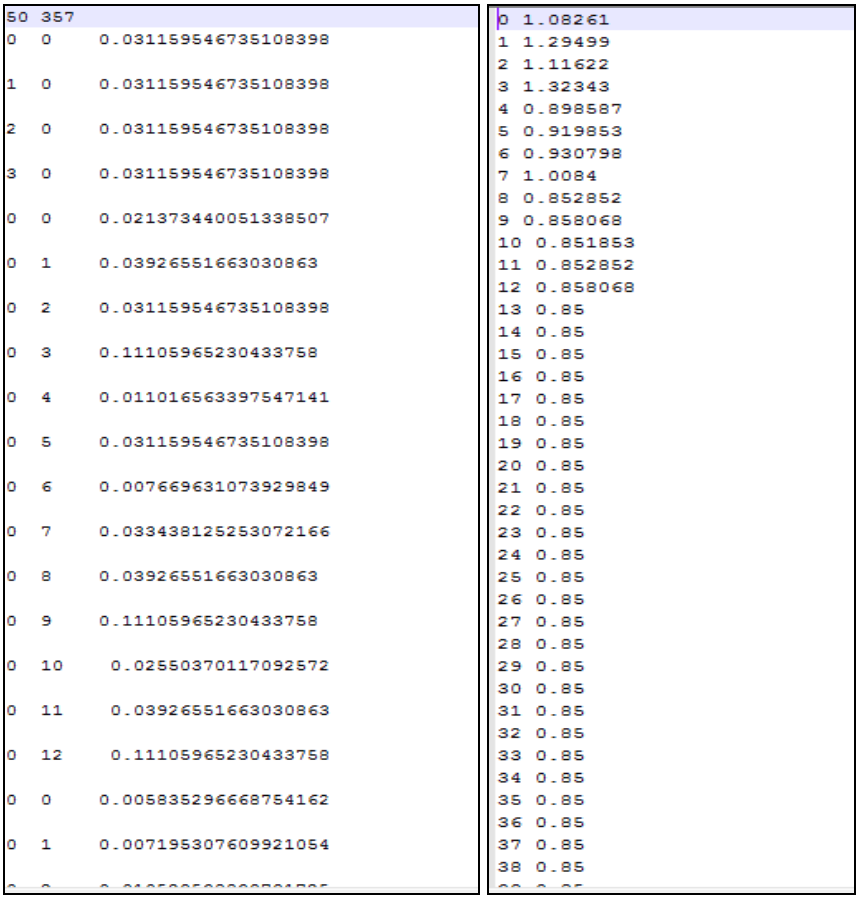


Figura 16. Archivo .txt con los valores de similitud de cada sentido de acuerdo a la palabra.

- **Paso 2. Asignar valores del sentido calculando *TextRank***

Se aplica la fórmula del *PageRank* en el algoritmo de rankeo, para saber que vértice (sentido) tiene mayor valor (importancia) y con ello poder saber qué sentido es el correcto con respecto a la palabra en la oración. Para cada uno de los nodos del grafo deben ser numerados comenzando por cero, el archivo que se va a generar debe inicia con la primera línea que indica el número de nodos, seguido por el número de aristas [Mihalcea, 2006].



```
50 357
0  0    0.031159546735108398
1  0    0.031159546735108398
2  0    0.031159546735108398
3  0    0.031159546735108398
0  0    0.021373440051338507
0  1    0.03926551663030863
0  2    0.031159546735108398
0  3    0.11105965230433758
0  4    0.011016563397547141
0  5    0.031159546735108398
0  6    0.007669631073929849
0  7    0.033438125253072166
0  8    0.03926551663030863
0  9    0.11105965230433758
0  10    0.02550370117092572
0  11    0.03926551663030863
0  12    0.11105965230433758
0  0    0.005835296668754162
0  1    0.007195307609921054
```

```
0 1.08261
1 1.29499
2 1.11622
3 1.32343
4 0.898587
5 0.919853
6 0.930798
7 1.0084
8 0.852852
9 0.858068
10 0.851853
11 0.852852
12 0.858068
13 0.85
14 0.85
15 0.85
16 0.85
17 0.85
18 0.85
19 0.85
20 0.85
21 0.85
22 0.85
23 0.85
24 0.85
25 0.85
26 0.85
27 0.85
28 0.85
29 0.85
30 0.85
31 0.85
32 0.85
33 0.85
34 0.85
35 0.85
36 0.85
37 0.85
38 0.85
```

Figura 17. Archivo de salida .txt con respecto al *PageRank*.

- **Paso 3. Comparación de las medias de Coseno, Edición y LCS**

	Coseno	**Edición**	**LCS**
Precisión	56.96%	43.29%	39.99
Recuerdo	56.56%	35.65	36.03

Etapa 5. Evaluación

En esta etapa se evalúa el desempeño del método con diferentes medidas de similitud. La precisión se calcula dividiendo cuantas de las oraciones de cada sentido son extraídas correctamente entre las oraciones correctas de los sentidos y el recuerdo se calcula cuantas de las oraciones buenas olvidó el sistema.

CAPÍTULO 5.

Experimentación

Como ya se había mencionado, los archivos d00.semcor.lexsn.key (D00), d01.semcor.lexsn.key (D01), d02.semcor.lexsn.key (D02), están considerados para la desambiguación del sentido de las palabras, por lo cual para determinar el sentido correcto de cada una de las palabras de acuerdo a la oración, se consideró hacer un pre-procesamiento con cada archivo para su posterior uso, generando así el valor de cada uno de los sentidos (vértices) y la similitud que existe entre ellos (aristas). Por esta razón se realizaron experimentos para determinar el sentido correcto de la oración de acuerdo al contexto.

5.1 Corpus

Se define un corpus como una colección de textos en lenguaje natural, elegido para caracterizar un estado o variedad de un lenguaje. Sin embargo, en la actualidad existen diversos tipos de corpus, entre los que destacan: corpus del lenguaje escrito y corpus del lenguaje hablado. En cualquier caso, un corpus actúa como repositorio de información la cual puede ser manipulada para extraer conocimiento [Orta, 2007].

5.1.1 SENSEVAL-2

Se celebra desde 1998 y cuyo objetivo principal es la evaluación objetiva de sistemas, métodos y técnicas que determinan automáticamente el significado de una palabra en un contexto [Guzmán, 2005].
Sirve como un foro que reúne a investigadores de *WSD* y dominios que utilizan *WSD* para diversas tareas. Esto permite a los investigadores discutir modificaciones que mejoran el rendimiento de los sistemas y analizar las combinaciones que son óptimas [Taulé & Martí, 2003].
A continuación se describe las versiones de Senseval, el año y el lenguaje utilizando en cada foro ver tabla 2.

Tabla 2. Versiones de Senseval.

SENSEVAL		
SENSEVAL-1	1988	Constituidos por tres lenguas inglés, francés y el italiano.
SENSEVAL-2	**2001**	Las lenguas se incrementaron hasta un total de doce (**inglés**, francés, italiano, español, vasco, koreano, sueco, holandés, estonio, checo, chino y japonés).
SENSEVAL-3	2004	Surgió en un taller realizado en Barcelona en colaboración de ACL, tiene catorce tareas diferentes para la desambiguación semántica.

A continuación se describe las distintas lenguas de la versión de Senseval-2 ver tabla 3.

Tabla 3. Tareas que han participado en las distintas lenguas de Senseval-2.

TAREAS	
Tarea léxica completa	Checo (1), Estonio (2), Holandés , **Ingles** (12)
Tarea basada en muestra léxica	Español (5), Vasco (2), Inglés (15), Italiano (2), Japonés (3), Coreano (2), Sueco (5) .
Tarea de traducción	Japonés (8)

La categoría gramatical se observa para los datos de Senseval-2 constan de **3 archivos** con un total de **238 oraciones** y **2436 palabras** para **desambiguar**, de las cuales, en la siguiente tabla se observa la distribución de estos datos [Torres, 2009]. A continuación se muestra en la tabla4 las diferentes categorías de acuerdo al foro Senseval-2.

Tabla 4. Categoría gramatical de Senseval-2.

	Sustantivos	**Verbos**	**Adjetivos**	**Adverbios**
Archivo 1 D00	331	162	99	86
Archivo 2 D01	495	242	170	107
Archivo 3 D02	310	140	188	106
Total	**1136**	**544**	**457**	**299**

5.1.2 SemCor

Creado por la Universidad de Princeton, es un corpus etiquetado sintáctica y semánticamente. En la figura 16 podemos observar un ejemplo del formato y etiquetado que usa *SemCor* para la representación de oraciones, la información de cada una de las oraciones se encuentra etiquetada por <wf> y </wf>.

```
<context filename=d00 source=senseval2>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done id=d00.s00.t01 pos=NN lemma=art wnsn=3 lexsn=1:09:00::>art</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done id=d00.s00.t03 pos=NN lemma=change_ringing wnsn=1 lexsn=1:04:00::>change-
ringing</wf>
<wf cmd=done id=d00.s00.t04 pos=VBZ lemma=be wnsn=1 lexsn=2:42:03::>is</wf>
<wf cmd=done id=d00.s00.t05 pos=JJ lemma=peculiar wnsn=2;4
lexsn=5:00:00:specific:00;5:00:00:characteristic:00>peculiar</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t08 pos=NNS lemma=english wnsn=2 lexsn=1:18:00::>English</wf>
<punc>,</punc>
<wf cmd=ignore pos=CC>and</wf>
<punc>,</punc>
<wf cmd=ignore pos=IN>like</wf>
<wf cmd=done id=d00.s00.t13 pos=JJS lemma=most wnsn=1 lexsn=3:00:02::>most</wf>
<wf cmd=done id=d00.s00.t14 pos=JJ lemma=english wnsn=1 lexsn=3:01:00::>English</wf>
<wf cmd=done id=d00.s00.t15 pos=NNS lemma=peculiarity wnsn=1;2
lexsn=1:09:00::;1:07:02::>peculiarities</wf>
<punc>,</punc>
<wf cmd=done id=d00.s00.t17 pos=JJ lemma=unintelligible wnsn=2
lexsn=5:00:00:incomprehensible:00>unintelligible</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t20 pos=NN lemma=rest wnsn=1 lexsn=1:24:00::>rest</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t23 pos=NN lemma=world wnsn=7 lexsn=1:14:02::>world</wf>
<punc>.</punc>
</s>
```

Figura 18. Formato y etiquetado usado por *SemCor*.

5.1.3 Resultados

A continuación se muestra una oración aplicando la metodología propuesta, la siguiente oración fue extraída del foro Senseval-2 del archivo D00 en el idioma inglés.

Etapa 1. Datos de entrada

Primera oración

The art of change_ringing be peculiar to the english, and, like most english peculiarity, unintelligible to the rest of the world

Etapa 2. Pre-procesamiento

- **Paso 1. Lematización**

art change_ringing be peculiar english most english peculiarity unintelligible rest world

Etapa 3. Preparación de los datos

- **Paso 1. Etiquetado Sintáctico**

art#n change_ringing#n be#v peculiar#a english#n most#a english#a peculiarity#n unintelligible#a rest#n world#n

- **Paso 2. Preparación de las oraciones por cada sentido**
-

PALABRA/ETIQUETA	sustantivo (**n**)	verbo (**v**)	adjetivo (**a**)	adverbio (**r**)	TOTAL DE SENTIDOS

art	**X**				**4**
change_ringing	**X**				**1**
be		**X**			**13**
peculiar			**X**		**4**
english	**X**				**4**
most			**X**		**2**
english			**X**		**4**
peculiarity	**X**				**3**
Unintelligible			**X**		**2**
Rest	**X**				**7**
World	**X**				**8**

Paso 3. Eliminación de stopwords en la oración

Se tomó como ejemplo la palabra ART para el pre-procesamiento en stopwords.

Palabra	Sustantivo (**n**)	Total de sentidos	Sentidos
art	**X**	**4**	**1. PRODUCTS,HUMAN,CREATIVITY,WORKS,ART,COLLECTIVELY,ART, EXHIBITION,FINE,COLLECTION,ART** **2. CREATION,BEAUTIFUL,SIGNIFICANT,THINGS,ART,INNOVATIVE, GOOD,GOOD,ART,ARCHITECTURE,ART,WASTING,SPACE, BEAUTIFULLY** **3. SUPERIOR,SKILL,LEARN,STUDY,PRACTICE,OBSERVATION,ART, CONVERSATION,ART** **4. PHOTOGRAPHS,VISUAL,REPRESENTATIONS,PRINTED, PUBLICATION, PUBLISHER,RESPONSIBLE,ARTWORK,BOOK**

Paso 4. Eliminación de *stemming* en la oración

Se tomó como ejemplo la palabra ART para el pre-procesamiento en *stemming.*

Palabra	Sustantivo (**n**)	Total de sentidos	Sentidos
art	**X**	**4**	**1. product,human,creativ,work,art,collect,art,exhibit ,fine,collect,art** **2. creation,beauti,signific,thing,art,innov,good,good, art,architectur,** **art,wast,space,beautifully** **3. superior,skill,learn,studi,practic,observ,art,convers ,art** **4. photograph,visual,represent,print,public,publish,r espons,artwork,** **book**

Etapa 4. Similitud de sentidos

- **Paso 1. Asignar valores a las aristas aplicando cálculo de similitud**

SIMILARIDAD ENTRE PALABRA	
0001_ART:1->0002_CHANGE_RINGING:1= 0.031159546735108398	0002_CHANGE_RINGING:1->0003_BE:6 = 0.031159546735108398
0001_ART:2->0002_CHANGE_RINGING:1= 0.031159546735108398	0002_CHANGE_RINGING:1->0003_BE:7 = 0.007669631073929849
0001_ART:3->0002_CHANGE_RINGING:1 = 0.031159546735108398	0002_CHANGE_RINGING:1->0003_BE:8 = 0.033438125253072166
0001_ART:4->0002_CHANGE_RINGING:1 = 0.031159546735108398	0002_CHANGE_RINGING:1->0003_BE:9 = 0.03926551663030863
0002_CHANGE_RINGING:1->0003_BE:1 = 0.021373440051338507	0002_CHANGE_RINGING:1->0003_BE:10 = 0.11105965230433758
0002_CHANGE_RINGING:1->0003_BE:2 = 0.03926551663030863	0002_CHANGE_RINGING:1->0003_BE:11 = 0.02550370117092572
0002_CHANGE_RINGING:1->0003_BE:3 = 0.031159546735108398	0002_CHANGE_RINGING:1->0003_BE:12 = 0.03926551663030863
0002_CHANGE_RINGING:1->0003_BE:4 = 0.11105965230433758	0002_CHANGE_RINGING:1->0003_BE:13 = 0.11105965230433758
0002_CHANGE_RINGING:1->0003_BE:5 = 0.011016563397547141	

- **Paso 2. Asignar valores del sentido calculando *TextRank***

VÉRTICE (SENTIDO) CON MAYOR VALOR
0 1.08261
1 1.29499
2 1.11622
3 1.32343
4 0.898587
5 0.919853
6 0.930798
7 1.0084
8 0.852852
9 0.858068
10 0.851853
11 0.852852
12 0.858068
13 0.85
14 0.85
15 0.85

Etapa 5. Evaluación

Aplicando precisión y recuerdo para la mediad de Coseno, Edición y LCS

Los resultados obtenidos en las siguientes tablas 5, 6 y 7, son usados los datos de las 2473 palabras. La evaluación se hizo sobre foro Senseval-2 en la tarea *english-all-words* del idioma inglés. Para llevar a cabo la desambiguación, se hiso uso de un diccionario de sentidos, en nuestro caso fue utilizado *WordNet* en la versión 2.1. Se siguió con la metodología experimental para llevar a cabo la evaluación de los métodos de desambiguación del sentido de las palabras basados grafos. A continuación se muestran los resultados obtenidos de cada una de las medias de evaluación:

Tabla 5. Resultados generados de los archivos D00, D01 y D02 del foro Senseval-2 en el idioma inglés aplicando medida de Coseno.

ARCHIVO	TOTAL DE PALABRAS	COINCIDENCIAS	PRECISIÓN	RECUERDO
D00	684	268	39.18	40.86
D01	1032	604	58.52	79.86
D02	757	554	73.18	48.96
PROMEDIO	2473	1426	56.96%	56.56%

Tabla 6. Resultados generados de los archivos D00, D01 y D02 del foro Senseval-2 en el idioma inglés aplicando medida de Edición.

ARCHIVO	TOTAL DE PALABRAS	COINCIDENCIAS	PRECISIÓN	RECUERDO
D00	684	400	58.47	50.20
D01	1032	438	42.44	30.25
D02	757	203	28.81	26.50
PROMEDIO	2473	1041	43.24%	35.65%

Tabla 7. Resultados generados de los archivos D00, D01 y D02 del foro Senseval-2 en el idioma inglés aplicando medida de LCS.

ARCHIVO	TOTAL DE PALABRAS	COINCIDENCIAS	PRECISIÓN	RECUERDO
D00	684	184	27.04	26.30
D01	1032	503	48.74	43.60
D02	757	312	41.21	38.20
PROMEDIO	2473	1000	38.99%	36.03%

CAPÍTULO 6.

Conclusiones y Trabajo Futuro

En este capítulo, se presentan las conclusiones generales del trabajo de tesis, así como las principales aportaciones del presente trabajo y trabajos futuros.

6.1. Conclusiones

En la presente tesis, se compararon diferentes medidas de similitud del método de rankeo de grafos para la tarea léxica completa *english-all-words,* en el idioma inglés, del foro Senseval-2.

- Se extrajeron los sentidos del diccionario de *WordNet* versión 2.1.
- Se definieron las medidas de similitud del método del rankeo basado en grafos para seleccionar el sentido correcto de las palabras.
- Se probaron las diferentes medidas de similitud y se conoció el desempeño.
- Se evaluaron los resultados obtenidos utilizando Precisión y Recuerdo.

- Se determinó la importancia del pre-procesamiento para cada uno de los sentidos.
- En esta tesis, se realizaron diversos experimentos con diferentes medidas de similitud, como, por ejemplo, Coseno, Edición y LCS. Los resultados que se obtuvieron son: para Coseno con una precisión de 56.96% y recuerdo de 56.56%, para Edición con una precisión de 43.24% y recuerdo de 35.65% y para LCS con una precisión de 38.99% y recuerdo de 36.03%.
- Se realizó el análisis de cada medida ya que el foro cuenta con tres documentos, los cuales vienen siendo pequeños, por ellos los valores en las tres medidas varían del 39 hasta 72%.

6.2. Aportaciones

Algunas aportaciones en el área de la desambiguación del sentido de las palabras son:

- Una de las principales aportaciones de este trabajo es la realización de un conjunto de módulos que ayudaron al desarrollo del método.
- Se compararon tres medias de similitud como Coseno, Edición y LCS, obteniendo mejor resultado la similitud de Coseno con una precisión de 56.96%.
- Se propuso una metodología para la desambiguación del sentido de las palabras.
- Se implementó una metodología que consta de cinco etapas.
- Se utilizó un diccionario de sentidos como lo fue *WordNet* en la versión 2.1, arrojando cada sentido automáticamente.

6.3 Trabajo futuro

En la tabla 1, se puede ver una lista de las características de los diferentes métodos. En este trabajo solo se tomaron las siguientes características, la clasificación que fuera por sustantivo, verbo, adjetivo y adverbio, se utilizó en específico la tarea léxica completa en el idioma inglés y utilizando a *WordNet* versión 2.1 como el diccionario de sentidos, sin embargo, como trabajo futuro quedan las siguientes preguntas:

- ¿Se podrá utilizar otras tareas en el área de la desambiguación del sentido de las palabras y otros idiomas, sin ayuda de un diccionario?
- En este trabajo solo se hicieron pruebas con la tarea léxica completa, pero como trabajo futuro habría que probar con otras tareas en diferentes idiomas.
- Habría que desarrollar y probar otras medidas para desambiguación del sentido de las palabras como la medida de Dice, Hamming, Jaro-Winkler, entre otras y así se podría encontrar una mejor medida para obtener mejores resultados.
- Habría que ajustar los valores de las diferentes medidas de similitud para ver si se podría mejorar los resultados.

Referencias

[Alonso, 2010] Barcelo Alonso Grettel. Desambiguación de los sentidos de las palabras en español usando textos paralelos, Tesis del Doctorado, Instituto Politécnico Nacional, 2010.

[Baeza-Yates, 1992] Baeza, Y. (1992). Information Retrieval: data structures & Algorithms (First Edition). New Yersey: Prentice Hall Sitio web: http://www.amazon.com/Information-Retrieval-Data-Structures-Algorithms/dp/0134638379.

[Baeza Yates, 1999] Baeza, Y. (1999). Modern Information retrieval (First Printed). Octubre 20, 2015, de New York ACM Press: Addison Wesley Sitio web: http://people.ischool.berkeley.edu/~hearst/irbook/

[Bolshakov & Gelbukh, 2004] Bolshakov, Igor, Gelbukh, Alexander. Computational Linguistics. Models, Resources, Applications. Ciencia de la Computación.

[Bunescu, 2006] Bunescu, R. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. Agosto 20, 2015, de

Presented at the 11th Conference of the European Chapter of the Association for Computational Linguistics, Italy Sitio web: http://mys.yoursearch.me/web?q=Using+Encyclopedic+Knowledge+for+Named+Entity+Disambiguation

[Buttcher, 2010] Buttcher, S. (2010). Implementing and Evaluating Search Engines. Septiembre 19, 2015, de Information Retrieval, The MIT Press Cambridge, Massachusetts London, England Sitio web: http://mys.yoursearch.me/web?q=A+simple+tokenization+of+english+text%2C+in+information+retrieval+implementing+and+evalating+search+ngines

[Deco, 2007] Claudia Deco, Cristina Bender, Mario Chiari; Problemas de la Traducción de la Consulta en la Búsqueda de Información Multilingüe; Departamento de Investigación Institucional; Facultad de Química e Ingeniería; Universidad Católica Argentina; Argentina; 2007. Sitio web: http://www.infosurrevista.com.ar/biblioteca/INFOSUR-Nro1-2007- DecoBenderChiari.pdf

[Dice, 1945] Dice.L.R. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3).

[Gaona, 2007] Ríos Gaona Miguel Ángel. Desambiguación de sentidos de palabras usando sinónimos, Tesis de Licenciatura, Instituto Politécnico Nacional, 2007.

[Gelbukh y Sidorov, 2002] Gelbukh, A. y G. Sidorov. Recuperación de documentos con comparación semántica suave. In: Proc. TAINA-2002, Workshop on Soft Computing at MICAI'2002: 2nd Mexican International Conference on Artificial Intelligence, Merida, Mexico, April 2002, pp 253–261.

[Gelbukh & Sidorov, 2006] Gelbukh, A. & Sidorov G. (2006). Procesamiento Automático del Español con Enfoque en Recursos Léxicos Grandes. México: Tresguerras 27, 06040, DF. [Agrawal 1994] Agrawal Rakesh, Srikant Ramakrishnan, "*Fast Algorithms for Mining Association Rules*", Proceedings of 20th International Conference on Very Large Data Bases (VLDB), Morgan Kaufmann, ISBN 1-55860-153-8, 1994, pp. 487–499.

[Gelbukh, 2009] Gelbukh, A. (2009). Generalized Mongue Elkan Method for Approximate Text String Comparison. Springer-Verlag, pp. 559-570.

[Guzmán, 2005] Guzmán (2005). Descubrimiento de patrones léxicos en la Web para su integración en

sistemas de desambiguación del sentido de las palabras. Diciembre 14, 2007, de Departamento de Sistemas Informáticos y Computación Sitio web: http://users.dsic.upv.es/~prosso/resources/GuzmanDEA.pdf.

[Hernández, 2007] Edith Hernández Reyes; Agrupamiento de documentos basado en Secuencias Frecuentes Maximales; Tesis de Maestría; Instituto Nacional de Astrofísica, Óptica y Electrónica; Puebla, México; 2007.

[Kowalski, 1997] Kowalski, G. (1997). Information Retrieval Systems theory and implementation. Boston: Kluwer Academic.

[Kryscia, 2007] Kryscia Daviana Ramirez Benavides; *Stemming* - Lematización; Escuela de Ciencias de la Computación e Informática; Universidad de Costa Rica; Costa Rica; 2007. Sitio web: http://www.ecci.ucr.ac.cr/~kramirez/RI/Material/Presentaciones/*Stemming*.pdf

[Levenshtein, 1966] Levenshtein, V.I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, 10(8), pp. 707-710.

[Manning, 2009] Manning, C.D., Prabhakal Raghavan, P., & Schutze, H.

(2009). An Introduction to information retrieval. Cambridge, England: Cambridge University Press.

[Mihalcea, 2004] Rada Mihalcea; Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization; Department of Computer Science; University of North Texas; Texas; EUA; 2004.

[Mihalcea, 2006] Rada Mihalcea. Random Walks on Text Structures, 2006. Sitio web: http://link.springer.com/chapter/10.1007%2F11671299_27

[Moreiro, 2002] José Antonio Moreiro González; Aplicaciones al análisis automático del contenido provenientes de la teoría matemática de la información; Departamento de Biblioteconomía y Documentación; Universidad Carlos III de Madrid; España; 2002. Sitio web: http://www.um.es/fccd/anales/ad05/ad0515.pdf

[Orta, 2007] Rosa María Ortega Mendoza. (2007). Descubrimiento automático de hipónimos a partir de texto no estructurado. Enero 23, 2015, Tesis de INAOE Sitio web: http://ccc.inaoep.mx/~villasen/index_archivos/tesis/TesisMaestria-RosaOrtega.pdf

[Pajares, 2006] Gonzalo Pajares Martin Sanz, Matilde Santos Penas; Inteligencia Artificial e Ingeniería del Conocimiento; Facultad de Informática; Universidad Complutense de Madrid; Madrid; España; 2006; Editorial Alfaomega Ra-Ma

[Pancardo, 2006] Rodríguez Pancardo Aarón. La Web como Recurso Lingüístico para la Desambiguación Semántica, 2006.

[Peinado, 2003] Jesús Peinado Rodríguez; Lematización para palabras médicas complejas: Implementación de un algoritmo en LISP; Facultad de Salud Pública y Administración Carlos Vidal Layseca; Universidad Peruana Cayetano Heredia; Peru; 2003. Sitio web: http://www.scielo.org.pe/pdf/rmh/v14n4/v14n4cc02.pdf

[Pérez, 2012] Pérez, M. (2012). Construcción de un árbol de términos latentes y su uso en el cálculo de la semejanza de documentos. Diciembre 20, 2015, Tesis de Instituto Politécnico Nacional Sitio web: https://www.google.com.mx/?gws_rd=cr#q=Construcci%C3%B3n+de+un+%C3%A1rbol+de+t%C3%A9rminos+latentes+y+su+uso+en+el+c%C3%A1lculo+de+la+semejanza+de+documentos.+Instituto+Polit%C3%A9cnico+Nacio

nal%2C+M%C3%A9xic%o2C+D.F%2C+2012

[Penn, 2004] Gerald Penn, Word Sense Disambiguation. CSC401, Spring 2004. University of Toronto.

[Rao, 2013] Rao, D., McNamee, P., & Dredze, M. (2013).Entity linking: Finding extracted entities in a knowledge base. In Multi-source, Multilingual Information Extraction and Summarization, pp. 93-115. Springer.Retreved from. Sitio web: http://link.springer.com/chapter/10.1007/978-3-642-28569_5

[Sidorov, 2005] Sidorov G. Etiquetador Morfológico y Desambiguador Manual: Dos Aplicaciones del Analizador Morfológico Automático para el Español. En: Memorias del VI encuentro internacional de computación ENC-2005, México, Puebla, 2005, pp. 147–149.

[Som 2008] Som Sinha Ravi. Graph-based Centrality Algorithms for Unsupervised Word Sense Disambiguation, Tesis, UNT, 2006.

[Sinha & Mihalcea, 2007] Ravi Sinha and Rada Mihalcea, Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity, in Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007), Irvine, CA, September 2007.

[Taulé & Martí, 2003] Taulé & Martí. (2003). SENSEVAL, una aproximación computacional al significado. Noviembre 18, 2013, de SENSEVAL Sitio web: http://www.uoc.edu/humfil/articles/esp/taule0303/taule0303.html

[Ted, 2000] Ted Pederse. 2000. A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. Proceedings of the first conference on North American chapter of the Association for Computational Linguistics. Seattle, Washington, pp. 63–69.

[Tejeda, 2006] Tejeda Cárcamo Javier. Desambiguación de sentidos de palabras usando relaciones sintácticas como contexto local. Mayo de 2006. Octubre 13, 2013 Sitio web: http://www.gelbukh.com/thesis/Javier%20Leandro%20Tejada%20Carcamo%20-%20MSc.pdf.

[Torres, 2009] Torres Ramos Sulema. Optimización global de coherencia en la desambiguación del sentido de las palabras, Tesis, IPN, 2009.

[Uribe, 2010] Uribe, I.A. (2010). Guía Metodológica para la Selección de Técnicas de Depuración de Datos (Tesis de Maestría) Universidad Nacional de Colombia, Medellín.

[Villat, 2006] Esau Villatoro Tello; Generación automática de resúmenes de múltiples documentos; Tesis de Maestría; Instituto Nacional de Astrofísica, Óptica y Electrónica; Puebla, México; 2006.

[Zhu, 20004] Mu Zhu; Recall, Precision and Average Precision; Working Paper 2004-09; Department of Statistics & Actuarial Science; University of Waterloo; Canada, 2004.

Anexo 1. Significado de las etiquetas sintácticas utilizadas para la anotación del corpus SENSEVAL

La siguiente tabla muestra las etiquetas asignadas por el etiquetador estocástico de categoría Gramatical de Eric Brill.

Etiqueta sintáctica	**Significado**
CC	Conjunción coordinada
CD	Cardinalidad
CT	Determinante
EX	Existencial "there"
FW	Palabra extranjera
IN	Preposición o conjunción subordinada
JJ	Adjetivo
JJR	Adjetivo, comparativo
JJS	Adjetivo, superaltivo
LS	Marcador de elemento de lista
MD	Modal
NN	Sustantivo, singular o no contable
NNP	Nombre propio, singular
NNPS	Nombre propio, plural
NNS	Sustantivo, plural
NP	Nombre propio, singular
NPS	Nombre propio, plural

POS	Terminación posesiva
PP	Pronombre personal
PR	Pronombre
PRP	Pronombre
PRP$	Pronombre, plural
RB	Adverbio
RBR	Adverbio, comparativo
RBS	Adverbio, superlativo
RP	Partícula
SYM	Símbolo
TO	"to"
UH	Intersección
VB	Verbo, forma básica
VBD	Verbo, tiempo pasado
VBG	Verbo, gerundio o presente participio
VBN	Verbo, pasado participio
VBP	Verbo, presente singular sin 3era persona
VBZ	Verbo, 3era persona presente singular
WDT	Wh-determinante
WP	Wh-pronombre
WP$	Wh-pronombre posesivo
WRB	Wh-adverbio
#	muestra de la libra
$	signo de dólar
.	puntuación final de oración
,	coma
;	punto y coma
(	corchete izquierdo
)	corchete derecho
“	comillas dobles

El conjunto de etiquetas sintáctica Penn Treebank [Santorini, y otros, 1990] [Marcus, y otros, 1993]

Etiqueta sintáctica	**Significado**
ADJP	frase adjetivo
ADVP	frase adverbio
NP	sintagma nominal
PP	sintagma preposicional
S	cláusula declarativa simple
SBAR	cláusula introducida por conjunción subordinante
SBARQ	pregunta directa introducido por wh-palabra o wh-frase
SINV	oración declarativa con el tema de inversión
SQ	Subconstituyente de SBARQ excluyendo wh-palabra o wh-frase
VP	sintagma verbal
WHADVP	wh- frase adverbio
WHNP	wh-sintagma nominal
WHPP	Wh- sintagma preposicional
X	constituyente de categoría desconocido o incierto

Ejemplo

Las etiquetas sintácticas o mejor dicho y para hablar con propiedad morfosintácticas son aquellas que identifican cada palabra como perteneciente a una clase o categoría gramatical.

1.- El coche rojo necesita un cambio de aceite

El	Articulo
coche	Sustantivo
rojo	Adjetivo
necesita	Verbo
un	Articulo
cambio	Sustantivo
de	Preposición
aceite	Sustantivo

2.- Le cambio el aceite al coche rojo

Le	Pronombre
cambio	Verbo
el	Articulo
aceite	Sustantivo
al	Preposición + articulo
coche	Sustantivo
rojo	Adjetivo

Anexo 2. Significado de las etiquetas del formato SemCor

Ejemplo de un archivo en formato *SemCor*:

```
<context filename=d00 source=senseval2>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done id=d00.s00.t01 pos=NN lemma=art wnsn=3 lexsn=1:09:00::>art</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done id=d00.s00.t03 pos=NN lemma=change_ringing wnsn=1 lexsn=1:04:00::>change-
ringing</wf>
<wf cmd=done id=d00.s00.t04 pos=VBZ lemma=be wnsn=1 lexsn=2:42:03::>is</wf>
<wf cmd=done id=d00.s00.t05 pos=JJ lemma=peculiar wnsn=2;4
lexsn=5:00:00:specific:00;5:00:00:characteristic:00>peculiar</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t08 pos=NNS lemma=english wnsn=2 lexsn=1:18:00::>English</wf>
<punc>,</punc>
<wf cmd=ignore pos=CC>and</wf>
<punc>,</punc>
<wf cmd=ignore pos=IN>like</wf>
<wf cmd=done id=d00.s00.t13 pos=JJS lemma=most wnsn=1 lexsn=3:00:02::>most</wf>
<wf cmd=done id=d00.s00.t14 pos=JJ lemma=english wnsn=1 lexsn=3:01:00::>English</wf>
<wf cmd=done id=d00.s00.t15 pos=NNS lemma=peculiarity wnsn=1;2
lexsn=1:09:00::;1:07:02::>peculiarities</wf>
<punc>,</punc>
<wf cmd=done id=d00.s00.t17 pos=JJ lemma=unintelligible wnsn=2
lexsn=5:00:00:incomprehensible:00>unintelligible</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t20 pos=NN lemma=rest wnsn=1 lexsn=1:24:00::>rest</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t23 pos=NN lemma=world wnsn=7 lexsn=1:14:02::>world</wf>
<punc>.</punc>
</s>
```

La Organización Internacional de Estándares (ISO) ha normalizado este lenguaje en 1986. El lenguaje SGML sirve para especificar las reglas de etiquetado de documentos y no impone en si ningún conjunto de etiquetas en especial. Los archivos que se utilizaron fue Senseval_2 en archivos:

d00.semcor.lexsn, d01.semcor.lexsn y d02.semcor.lexsn, los cuales tienen diferentes etiquetas las cuales se describen a continuación:

Este elemento indica el comienzo de un contexto. filename es el nombre del archivo del corpus original del cual se extrae el contexto. d00 indica que este documento contiene delimitadores de párrafo.

<context filename=d00 source=senseval2>

Inicio de una nueva oración. sentence_number es un entero. La primera oración en cada contexto es numerada 1, y los números de oración son incrementados secuencialmente en todo el contexto. Los números de oración no reinician en uno en cada párrafo.

<s snum=1>

Este elemento representa una palabra. word es la forma ortográfica tal y como aparece en el documento original. Toda la información sintáctica y semántica es almacenada en forma de pares de atributo/valor descrito abajo.

<wf cmd=ignore pos=DT>The</wf>

cmd= cmd. Indica el estatus de un elemento **wf**.

cmd	Significado
tag	Palabra que debe ser etiquetada
done	Palabra etiquetada semáticamente
ignore	Palabra que no debe ser etiquetada
update	Utilizada solo durante el desarrollo de concordancia semántica
retag	Utilizada solo durante el desarrollo de concordancia semántica

pos= pos. Es la etiqueta sintáctica asignada por el etiquetador estocástico de categoría gramatical de Eric Brill.

lema= lema. La forma básica de la palabra o colocación que pertenece a los otros pares de atributo/valor en su wf. Esta es la forma de la cadena utilizada para buscar en la base de datos de *WordNet.* Si rdf está presente, lema es la forma básica de la redefinición. Cuando pn está presente, redefinición, lema y category tienen el mismo valor.

wnsn= sense_number. Es el número de sentido (entero) correspondiente a la salida de pantalla de *WordNet.*

lexsn= lex_sense. Cuando la encontramos concatenada con lema usando el carácter de concatenación "%", se crea una sense_key que indica a cual sentido de *WordNet* debemos ligar la palabra (word). Esta es la etiqueta semántica de una palabra (word).

<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done id=d00.s00.t01 pos=NN lemma=art wnsn=3
lexsn=1:09:00::>art</wf>
<wf cmd=ignore pos=IN>of</wf>

Los sentidos de *WordNet* comprenden un conjunto de sinónimos y definiciones al igual que un diccionario, las cuales son llamadas glosas.
El número que se encuentra al inicio de la definición de algunos sentidos, es la frecuencia de los valores obtenidos del corpus SemCor.
A diferencia de un diccionario, *WordNet* contiene un conjunto de relaciones léxicas entre sysnsets o lemas, los cuales aparecen al inicio de la glosa.

Anexo 3. Ejemplo de la estructura de un archivo en formato SemCor

En este ejemplo podemos ver una muestra del corpus del archive d00, para un parrafo que contiene varias oraciones, como por ejemplo: The art of change_ringing be peculiar to the English, and, like most English peculiarity, unintelligible to the rest of the world.

<context filename=d00 source=senseval2>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done id=d00.s00.t01 pos=NN lemma=art wnsn=3
lexsn=1:09:00::>art</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done id=d00.s00.t03 pos=NN lemma=change_ringing wnsn=1
lexsn=1:04:00::>change-ringing</wf>
<wf cmd=done id=d00.s00.t04 pos=VBZ lemma=be wnsn=1
lexsn=2:42:03::>is</wf>
<wf cmd=done id=d00.s00.t05 pos=JJ lemma=peculiar wnsn=2;4
lexsn=5:00:00:specific:00;5:00:00:characteristic:00>peculiar</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t08 pos=NNS lemma=english wnsn=2
lexsn=1:18:00::>English</wf>
<punc>,</punc>
<wf cmd=ignore pos=CC>and</wf>
<punc>,</punc>

<wf cmd=ignore pos=IN>like</wf>
<wf cmd=done id=d00.s00.t13 pos=JJS lemma=most wnsn=1
lexsn=3:00:02::>most</wf>
<wf cmd=done id=d00.s00.t14 pos=JJ lemma=english wnsn=1
lexsn=3:01:00::>English</wf>
<wf cmd=done id=d00.s00.t15 pos=NNS lemma=peculiarity wnsn=1;2
lexsn=1:09:00::;1:07:02::>peculiarities</wf>
<punc>,</punc>
<wf cmd=done id=d00.s00.t17 pos=JJ lemma=unintelligible wnsn=2
lexsn=5:00:00:incomprehensible:00>unintelligible</wf>
<wf cmd=ignore pos=TO>to</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t20 pos=NN lemma=rest wnsn=1
lexsn=1:24:00::>rest</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=ignore pos=DT>the</wf>
<wf cmd=done id=d00.s00.t23 pos=NN lemma=world wnsn=7
lexsn=1:14:02::>world</wf>
<punc>.</punc>
</s>

Anexo 4. Lista de palabras vacías (StopWords)

A continuación se presenta la lista de palabras vacías en inglés, utilizada en la etapa de pre procesamiento del método propuesto en esta tesis.

A

ABLE

ABOUT

ABOVE

ACCORDING

ACCORDINGLY

ACROSS

ACTUALLY

AFTER

AFTERWARDS

AGAIN

AGAINST

AIN, T

ALL

ALLOW

ALLOWS

ALMOST

ALONE

ALONG

ALREADY

ALSO

ALTHOUGH

ALWAYS

AM

AMONG

AMONGST

AN

AND

ANOTHER

ANY

ANYBODY

ANYHOW

ANYONE

ANYTHING

ANYWAY

ANYWAYS

ANYWHERE

APART

APPEAR

APPRECIATE

APPROPRIATE

ARE

AREN ,T

AROUND

AS

ASIDE

ASK

ASKING

ASSOCIATED

AT

AVAILABLE

AWAY

AWFULLY

B

BE

BECAME

BECAUSE

BECOME

BECOMES

BECOMING

BEEN

BEFORE

BEFOREHAND

BEHIND

BEING

BELIEVE

BELOW

BESIDE

BESIDES

BEST

BETTER

BETWEEN

BEYOND

BOTHBRIEF

BUT

BY

C

C, MON

C, S

CAME

CAN

CAN, T

CANNOT

CANT

CAUSE

CAUSES

CERTAIN

CERTAINLY

CHANGES

CLEARLY

CO

COM

COME

COMES

CONCERNING

CONSEQUENTLY

CONSIDER

CONSIDERING

CONTAIN

CONTAINING

CONTAINS

CORRESPONDING

COULD

COULDN, T

COURSE

CURRENTLY

D

DEFINITELY

DESCRIBED

DESPITE

DID

DIDN T

DIFFERENT

DO

DOES

DOESN T

DOING

DON T

DONE

DOWN

DOWNWARDS

DURING

E

EACH

EDU

EG

EIGHT

EITHER

ELSE

ELSEWHERE

ENOUGH

ENTIRELY

ESPECIALLY

ET

ETC

EVEN

EVER

EVERY

EVERYBODY

EVERYONE

EVERYTHING

EVERYWHERE

EX

EXACTLY

EXAMPLE

EXCEPT

F

FAR

FEW

FIFTH

FIRST

FIVE

FOLLOWED

FOLLOWING

FOLLOWS

FOR

FORMER

FORMERLY

FORTH

FOUR

FROM

FURTHER

FURTHERMORE

G

GET

GETS

GETTING

GIVEN

GIVES

GO

GOES

GOING

GONE

GOT

GOTTEN

GREETINGS

H

HAD

HADN ,T

HAPPENS

HARDLY

HAS

HASN T

HAVE

HAVEN ,T

HAVING

HE

HE ,S

HELLO

HELP

HENCE

HER

HERE

HERE ,S

HEREAFTER

HEREBY

HEREIN

HEREUPON

HERS

HERSELF

HI

HIM

HIMSELF

HIS

HITHER

HOPEFULLY

HOW

HOWBEIT

HOWEVER

I

I ,D

I,LL

I, M

I ,VE

IE

IF

IGNORED

IMMEDIATE

IN

INASMUCH

INC

INC

INDEED

INDICATE

INDICATED

INDICATES

INNER

INSOFAR

INSTEAD

INTO

INWARD

IS

ISN ,T

IT

IT ,D

IT ,LL

IT ,S

ITS

ITSELF

J

JUST

K

KEEP

KEEPS

KEPT

KNOW

KNOWS

KNOWN

L

LAST

LATELY

LATER

LATTER

LATTERLY

LEAST

LESS

LEST

LET

LET S

LIKE

LIKED

LIKELY

LITTLE

LOOK

LOOKING

LOOKS

LTD

M

MAINLY

MANY

MAY

MAYBE

ME

MEAN

MEANWHILE

MERELY

MIGHT

MORE

MOREOVER

MOST

MOSTLY

MUCH

MUST

MY

MYSELF

N

NAME

NAMELY

ND

NEAR

NEARLY

NECESSARY

NEED

NEEDS

NEITHER

NEVER

NEVERTHELESS

NEW

NEXT

NINE

NO

NOBODY

NON

NONE

NOONE

NOR

NORMALLY

NOT

NOTHING

NOVEL

NOW

NOWHERE

O

OBVIOUSLY

OF

OFF

OFTEN

OH

OK

OKAY

OLD

ON

ONCE

ONE

ONES

ONLY

ONTO

OR

OTHER

OTHERS

OTHERWISE

OUGHT

OUR

OURS

OURSELVES

OUT

OUTSIDE

OVER

OVERALL

OWN

P

PARTICULAR

PARTICULARLY

PER

PERHAPS

PLACED

PLEASE

PLUS

POSSIBLE

PRESUMABLY

PROBABLY

PROVIDES

Q

QUE

QUITE

QV

R

RATHER

RD

RE

REALLY

REASONABLY

REGARDING

REGARDLESS

REGARDS

RELATIVELY

RESPECTIVELY

RIGHT

S

SAID

SAME

SAW

SAY

SAYING

SAYS

SECOND

SECONDLY

SEE

SEEING

SEEM

SEEMED

SEEMING

SEEMS

SEEN

SELF

SELVES

SENSIBLE

SENT

SERIOUS

SERIOUSLY

SEVEN

SEVERAL

SHALL

SHE

SHOULD

SHOULDN T

SINCE

SIX

SO

SOME

SOMEBODY

SOMEHOW

SOMEONE

SOMETHING

SOMETIME

SOMETIMES

SOMEWHAT

SOMEWHERE

SOON

SORRY

SPECIFIED

SPECIFY

SPECIFYING

STILL

SUB

SUCH

SUP

SURE

T

T S

TAKE

TAKEN

TELL

TENDS

TH

THAN

THANK

THANKS

THANX

THAT

THAT S

THATS

THE

THEIR

THEIRS

THEM

THEMSELVES

THEN

THENCE

THERE

THERE S

THEREAFTER

THEREBY

THEREFORE

THEREIN

THERES

THEREUPON

THESE

THEY

THEY D

THEY, LL

THEY, RE

THEY, VE

THINK

THIRD

THIS

THOROUGH

THOROUGHLY

THOSE

THOUGH

THREE

THROUGH

THROUGHOUT

THRU

THUS

TO

TOGETHER

TOO

TOOK

TOWARD

TOWARDS

TRIED

TRIES

TRULY

TRY

TRYING

TWICE

TWO

U

UN

UNDER

UNFORTUNATELY

UNLESS

UNLIKELY

UNTIL

UNTO

UP

UPON

US

USE

USED

USEFUL

USES

USING

USUALLY

UUCP

V

VALUE

VARIOUS

VERY

VIA

VIZ

VS

W

WANT

WANTS

WAS

WASN T

WAY

WE

WE D

WE LL

WE RE

WE VE

WELCOME

WELL

WENT

WERE

WEREN T

WHAT

WHAT S

WHATEVER

WHEN

WHENCE

WHENEVER

WHERE

WHERE S

WHEREAFTER

WHEREAS

WHEREBY

WHEREIN

WHEREUPON

WHEREVER

WHETHER

WHICH

WHILE

WHITHER

WHO

WHO, S

WHOEVER

WHOLE

WHOM

WHOSE

WHY

WILL

WILLING

WISH

WITH

WITHIN

WITHOUT

WON T

WONDER

WOULD

WOULDN T

X

Y

YES

YET

YOU

YOU, D

YOU, LL

YOU, RE

YOU, VE

YOUR

YOURS

YOURSELF

YOURSELVES

Z
ZERO

Anexo 5. Lemas

Se extrajeron los lemas de cada uno de los archivos, d00.semcor.lexsn.key, d01.semcor.lexsn.key, d02.semcor.lexsn.key.

art
change_ringing
be
peculiar
english
most
english
peculiarity
unintelligible
rest
world
tailor
england
scene
evoke
rural
england
lovely
ancient
stone
church
stand
field
sound
bell
cascade
tower
call
faithful
evensong
parishioner
stop
chat
church
door
member
here
always
tower
man
woman
pull
rhythmically
rope
attach
same
bell
first
sound
here
also
discordant
church
britain
set
bell
once
own
band
ringer
herald
sunday
morning
evening
service
now
only
local
ringer
remain
other
here
today
live
elsewhere
belong_to
group
ringer

many
belfry
even
service
tower
scrape
say
retired
worker
try
train
youngster
disco
dance
just
drift
away
worry
old_age
flightiness
youth
diminish
rank
group
keep
bell
peal
history
side
nationwide
survey
modern
note
n't
churchgoer
enjoy
peal
bell
cool
autumn
evening
most
other
bell
no
longer
ring
sunday
ring
easy
see
why
ancient
art
less
complicated
version
play
tune
bell
carillon
continental
include
octogenarian
youngster
training
drive
sunday
church
church
effort
keep
bell
sound
foreigner
change_ringing
mind-boggling
exercise
english
invent
year
ago
require
physical
dexterity
bell
weigh
more
ton
combine
intense
mental
concentration

take
year
ago
nearly
third
england
church
low
simple
descend
scale
use
large
church
as
many
bell
then
signal
ringer
begin
vary
order
bell
sound
alter
steady
rhythm
striking
variation
change

europe
consider
english
childish
fit
only
change
peal
take
about
hour
look
thursday
night
practice
church
district
give
idea
work
involve
ringer
stand
circle
foot
ahead
other
prizefighter
stance
pull
rope

proper
english
bell
start
round
high-pitched
bell
bell
muffle
ceiling
totally
absorb
ringer
stare
straight
ahead
use
peripheral_vision
call
watch
other
rope
thus
time
pull
far
above
belfry
huge
bronze
bell

occur
only
once
rule
state
ringer
memorize
pattern
change
know
method
name
series
so
wrist
advance
retard
next
swing
so
bell
swap
place
following
change
well-known
detective_story
involve
church
bell
english
small
hole
high
ceiling
ringing
chamber
speak
snake
rope
seem
make
much
sound
bit
obsession
admit
master
band
england
best
female
ringer
passion
usually
stay
tower
however
more
often
not
ringer
mount
wheel
swing
madly
full
degree
start
end
surprisingly
inverted
position
skilled
ringer
use
not
well
cleric
membership
church_of_england
steadily
dwindle
strong-willed
vicar
press
equally
strong-willed
often
ringer
attend
service
year

novelist
describe
passion
find
satisfaction
mathematical
completeness
mechanical
perfection
ringer
add
fill
solemn
intoxication
come
intricate
ritual
faultlessly
perform
ringing
become
say
new
band
today
several
member
congregation
still
n't
enough

think_of
church
something
stick_on
bottom
belfry
when
change
complete
work_up
sweat
ringer
often
skip
local
pub
leave
worship
other
below
college
joy
bell_ringing
shortly
publish
booklet
vicar
country
entitle
bell
care

ago
vicar
get
so
ringer
attend
service
sack
entire
band
ringer
promptly
set_up
picket_line
protest
club
treat
tower
sort
separate
premises
n't
always
live
bell
need
ring
small
rural
parish
inner_city

ringer
ring
more
bell
bell
fall
silent
follow
dustup
church
attendance
vicar
refuse
talk_about
say
wound
here
come
ringer
above
make
vicar
say
president
hope
speak
student
theological
third
ringer
today

say
mr
recognize
no
longer
high
priority
church
life
experience
mr
also
attack
greater
problem
lack
ringer
survey
say
train
bell_ringer
england
today
only
still
also
ringer
newspaper
ringer
writer
sign

church
council
program
attract
train
ringer
only
partly
successful
say
mr
right
now
lucky
year
keep
new
ringer
add
bright
sign
grow
number
woman
enter
once
field
more
large
frequently
unwashed

woman
n't
accept
everywhere
however
oldest
bell_ringing
group
country
found
remain
fact
particularly
galling
woman
group
sole
source
ringer
britain
most
prestigious
church
britain
woman
file
equal_opportunity
suit
extent
problem
surface

letter
red-blooded
balanced
male
remark
frequency
woman
faint
peal
suggest
settle
back
traditional
role
make
tea
meeting
torrent
reply
follow
woman
ringer
observe
average
male
ringer
leave
quite
lot
desire
badly

unbearably
flatulent
peal
woman
write
say
year
never
know
lady
faint
belfry
see
man
die
bless
art
change_ringing
be
peculiar
english
most
english
peculiarity
unintelligible
rest
world
tailor
england
scene
evoke

summer
series
letter
weekly
stone
church
stand
field
sound
bell
cascade
tower
call
faithful
evensong
parishioner
stop
chat
church
door
member
here
always
tower
man
woman
pull
rhythmically
rope
attach
same
dress
decorate
acne
peal
bell
cool
autumn
evening
most
other
church
britain
set
bell
once
own
band
ringer
herald
sunday
morning
evening
service
now
only
local
ringer
remain
other
here
today
rural
england
lovely
ancient
year
ago
require
physical
dexterity
bell
weigh
more
ton
combine
intense
mental
concentration
proper
english
bell
start
round
high-pitched
bell
low
simple
descend
scale
use
large
church

bell
first
sound
here
also
discordant
modern
note
n't
churchgoer
enjoy
take
year
ago
nearly
third
england
church
bell
no
longer
ring
sunday
ring
easy
see
why
ancient
art
less
complicated

live
elsewhere
belong_to
group
ringer
include
octogenarian
youngster
training
drive
sunday
church
church
effort
keep
bell
sound
many
belfry
even
service
tower
scrape
say
retired
worker
try
train
youngster
disco
dance

as
many
bell
then
signal
ringer
begin
vary
order
bell
sound
alter
steady
rhythm
striking
variation
change
occur
only
once
rule
state
ringer
memorize
pattern
change
know
method
name
series
so

version
play
tune
bell
carillon
continental
europe
consider
english
childish
fit
only
foreigner
change_ringing
mind-boggling
exercise
english
invent
foot
ahead
other
prizefighter
stance
pull
rope
small
hole
high
ceiling
ringing
chamber
just
drift
away
worry
old_age
flightiness
youth
diminish
rank
group
keep
bell
peal
history
side
nationwide
survey
year
ago
require
physical
dexterity
bell
weigh
more
ton
combine
intense
mental
concentration
proper
change
peal
take
about
hour
look
thursday
night
practice
church
district
give
idea
work
involve
ringer
stand
circle
intoxication
come
intricate
ritual
faultlessly
perform
ringing
become
bit
obsession
admit
master
band

speak
snake
rope
seem
make
much
sound
very
obvious
exit
congregation
prayer
novelist
describe
passion
find
satisfaction
mathematical
completeness
mechanical
perfection
ringer
add
fill
solemn
often
skip
local
pub
leave
worship

english
bell
start
round
high-pitched
bell
low
simple
descend
scale
use
large
church
as
many
bell
then
signal
ringer
begin
vary
order
bell
sound

picket_line
protest
club
treat
tower
sort

england
best
female
ringer
passion
usually
stay
tower
however
more
often
not
ringer
think_of
church
something
stick_on
bottom
belfry
when
change
complete
work_up
sweat
ringer
fault
stairs
bell_tower
locate
next
altar

other
below
not
well
cleric
membership
church_of_england
steadily
dwindle
strong-willed
vicar
press
equally
strong-willed
often
ringer
attend
service
year
ago
vicar
get
so
ringer
attend
service
sack
entire
band
ringer
promptly

separate
premises
say
new
band
today
several
member
congregation
still
n't
enough
ringer
ring
more
bell
bell
fall
silent
follow
dustup
church
attendance
vicar
refuse
talk_about
say
wound
vicar
nearby
church

so
crunch
crunch
crunch
bang
bang
bang
here
come
ringer
above
make
very
obvious
exit
congregation
prayer
say
admit
mixed
feelings
issue
vicar
active
bell_ringer
sound
bell
net
draw
people
church

set_up
live
hope
ringer
draw
full
life
sort
parliament
ringing
group
aim
improve
relations
vicar
say
president
hope
speak
student
theological
college
joy
bell_ringing
shortly
publish
booklet
vicar
country
entitle
bell

feel
church
life
experience
mr
also
attack
greater
problem
lack
ringer
survey
say
train
bell_ringer
england
today
only
still
also
ringer
n't
always
live
bell
need
ring
small
rural
parish
inner_city

say
successful
say
mr
right
now
lucky
year
keep
new
ringer
add
bright
sign
grow
number
woman
enter
once
field
more
third
ringer
today
woman
n't
accept
everywhere
however
oldest
bell_ringing

care
say
mr
recognize
no
longer
high
priority
group
sole
source
ringer
britain
most
prestigious
church
britain
woman
file
equal_opportunity
suit
extent
problem
surface
summer
series
letter
weekly
newspaper
ringer
writer

church
council
program
attract
train
ringer
only
partly
make
tea
meeting
torrent
reply
follow
woman
ringer
observe
average
male
ringer
leave
quite
lot
desire
badly
dress
decorate
acne
large
frequently
unwashed

group
country
found
remain
fact
particularly
galling
woman
bless
wrist
advance
retard
next
swing
so
bell
swap
place
following
change
well-known
detective_story
involve
church
bell
english
draw
people
church
say
live

sign
letter
red-blooded
balanced
male
remark
frequency
woman
faint
peal
suggest
settle
back
traditional
role
far
above
belfry
huge
bronze
bell
mount
wheel
swing
madly
full
degree
start
end
surprisingly
inverted

unbearably
flatulent
peal
woman
write
say
year
never
know
lady
faint
belfry
see
man
die
feel
fault
stairs
bell_tower
locate
next
altar
so
crunch
crunch
crunch
bang
bang
bang
say
admit

hope
ringer
draw
full
life
sort
parliament
ringing
group
aim
improve
relations
thus
time
pull
bell
net
sound
totally
absorb
ringer
stare
straight
ahead
use
peripheral_vision
call
watch
other
rope
vicar

position
skilled
ringer
use
bell

mixed
feelings
issue
vicar
ceiling

nearby
church
ctive
bell_ringer
muffle

Anexo 6. Etiquetado Sintáctico

Se generó por cada uno de los archivos d00.semcor.lexsn.key, d01.semcor.lexsn.key, d02.semcor.lexsn.key, la clasificación de sustantivo (**n**), verbo (**v**), adjetivo (**a**), adverbio (**r**).

art#n
change_ringing#n
be#v
peculiar#a
english#n
most#a
english#a
peculiarity#n
unintelligible#a
rest#n
world#n
tailor#n
england#n
scene#n
evoke#v
rural#a
england#n
lovely#a
ancient#a
stone#n
church#n
stand#v
field#n
sound#n
vicar#n
say#v
president#n
hope#v
speak#v
student#n
theological#a
college#n
joy#n
bell_ringing#n
shortly#r
publish#v
booklet#n
vicar#n
country#n
entitle#v
bell#n
care#n
say#v
mr#n
recognize#v
no#r
longer#r
high#a
novelist#n
describe#v
passion#n
find#v
satisfaction#n
mathematical#a
completeness#n
mechanical#a
perfection#n
ringer#n
add#v
fill#v
solemn#a
intoxication#n
come#v
intricate#a
ritual#n
faultlessly#r
perform#v
ringing#n
become#v
bit#n
obsession#n
admit#v

bell#n
cascade#v
tower#n
call#v
faithful#n
evensong#n
parishioner#n
stop#v
chat#v
church#n
door#n
member#n
here#r
always#r
tower#n
man#n
woman#n
pull#v
rhythmically#r
rope#n
attach#v
same#a
bell#n
first#r
sound#v
here#r
also#r
discordant#a
modern#a
note#n
n't#r

priority#n
church#n
life#n
experience#n
mr#n
also#r
attack#v
greater#a
problem#n
lack#n
ringer#n
survey#n
say#v
train#v
bell_ringer#n
england#n
today#n
only#r
still#r
also#r
ringer#n
n't#r
always#r
live#v
bell#n
need#v
ring#v
small#a
rural#a
parish#n
inner_city#n

master#n
band#n
england#n
best#a
female#n
ringer#n
passion#n
usually#r
stay#v
tower#n
however#r
more#r
often#r
not#r
ringer#n
think_of#v
church#n
something#n
stick_on#v
bottom#n
belfry#n
when#r
change#n
complete#v
work_up#v
sweat#n
ringer#n
often#r
skip#v
local#a
pub#n

churchgoer#n
enjoy#v
peal#n
bell#n
cool#a
autumn#n
evening#n
most#a
other#a
church#n
britain#n
set#n
bell#n
once#r
own#a
band#n
ringer#n
herald#v
sunday#n
morning#n
evening#n
service#n
now#r
only#r
local#a
ringer#n
remain#v
other#n
here#r
today#n
live#v
church#n
council#n
program#n
attract#v
train#v
ringer#n
only#r
partly#r
successful#a
say#v
mr#n
right#r
now#r
lucky#a
year#n
keep#v
new#a
ringer#n
add#v
bright#a
sign#n
grow#v
number#n
woman#n
enter#v
once#r
field#n
more#a
third#a
ringer#n
today#n
leave#v
worship#n
other#n
below#r
not#r
well#r
cleric#n
membership#n
church_of_england#n
steadily#r
dwindle#v
strong-willed#a
vicar#n
press#v
equally#r
strong-willed#a
often#r
ringer#n
attend#v
service#n
year#n
ago#r
vicar#n
get#v
so#r
ringer#n
attend#v
service#n
sack#v
entire#a
band#n

elsewhere#r
belong_to#v
group#n
ringer#n
include#v
octogenarian#n
youngster#n
training#n
drive#v
sunday#n
church#n
church#n
effort#n
keep#v
bell#n
sound#v
many#a
belfry#n
even#r
service#n
tower#n
scrape#v
say#v
retired#v
worker#n
try#v
train#v
youngster#n
disco#n
dance#n
just#r
woman#n
n't#r
accept#v
everywhere#r
however#r
oldest#a
bell_ringing#a
group#n
country#n
found#v
remain#v
fact#n
particularly#r
galling#a
woman#n
group#n
sole#a
source#n
ringer#n
britain#n
most#a
prestigious#a
church#n
britain#n
woman#n
file#v
equal_opportunity#n
suit#n
extent#n
problem#n
surface#v
ringer#n
promptly#r
set_up#v
picket_line#n
protest#n
club#n
treat#v
tower#n
sort#n
separate#a
premises#n
say#v
new#a
band#n
today#n
several#a
member#n
congregation#n
still#r
n't#r
enough#a
ringer#n
ring#v
more#a
bell#n
bell#n
fall#v
silent#a
follow#v
dustup#n
church#n

drift#v
away#r
worry#v
old_age#a
flightiness#n
youth#n
diminish#v
rank#n
group#n
keep#v
bell#n
peal#v
history#n
side#n
nationwide#a
survey#n
take#v
year#n
ago#r
nearly#r
third#a
england#n
church#n
bell#n
no#r
longer#r
ring#v
sunday#n
ring#v
easy#a
see#v
summer#n
series#n
letter#n
weekly#a
newspaper#n
ringer#n
writer#n
sign#v
letter#n
red-blooded#a
balanced#a
male#n
remark#v
frequency#n
woman#n
faint#v
peal#n
suggest#v
settle#v
back#r
traditional#a
role#n
make#v
tea#n
meeting#n
torrent#n
reply#n
follow#v
woman#n
ringer#n
observe#v
attendance#n
vicar#n
refuse#v
talk_about#v
say#v
wound#n
vicar#n
nearby#a
church#n
feel#v
fault#n
stairs#n
bell_tower#n
locate#v
next#a
altar#n
so#r
crunch#n
crunch#n
crunch#n
bang#n
bang#n
bang#n
here#r
come#v
ringer#n
above#r
make#v
very#r
obvious#a
exit#n

why#r
ancient#a
art#n
less#r
complicated#a
version#n
play#v
tune#n
bell#n
carillon#n
continental#a
europe#n
consider#v
english#n
childish#a
fit#a
only#r
foreigner#n
change_ringing#n
mind-boggling#a
exercise#n
english#n
invent#v
year#n
ago#r
require#v
physical#a
dexterity#n
bell#n
weigh#v
more#a

average#a
male#a
ringer#n
leave#v
quite#r
lot#n
desire#v
badly#r
dress#v
decorate#v
acne#n
large#a
frequently#r
unwashed#a
unbearably#r
flatulent#a
peal#n
woman#n
write#v
say#v
year#n
never#r
know#v
lady#n
faint#v
belfry#n
see#v
man#n
die#v
bless#v
art#n

congregation#n
prayer#n
say#v
admit#v
mixed#v
feelings#n
issue#n
vicar#n
active#a
bell_ringer#n
sound#n
bell#n
net#n
draw#v
people#n
church#n
say#v
live#v
hope#n
ringer#n
draw#v
full#a
life#n
sort#n
parliament#n
ringing#n
group#n
aim#v
improve#v
relations#n
vicar#n

ton#n
combine#v
intense#a
mental#a
concentration#n
proper#a
english#a
bell#n
start#v
round#n
high-pitched#a
bell#n
low#a
simple#a
descend#v
scale#n
use#v
large#a
church#n
as#r
many#a
bell#n
then#r
signal#n
ringer#n
begin#v
vary#v
order#n
bell#n
sound#v
alter#v
change_ringing#n
be#v
peculiar#a
english#n
most#a
english#a
peculiarity#n
unintelligible#a
rest#n
world#n
tailor#n
england#n
scene#n
evoke#v
rural#a
england#n
lovely#a
ancient#a
stone#n
church#n
stand#v
field#n
sound#n
bell#n
cascade#v
tower#n
call#v
faithful#n
evensong#n
parishioner#n
stop#v
say#v
president#n
hope#v
speak#v
student#n
theological#a
college#n
joy#n
bell_ringing#n
shortly#r
publish#v
booklet#n
vicar#n
country#n
entitle#v
bell#n
care#n
say#v
mr#n
recognize#v
no#r
longer#r
high#a
priority#n
church#n
life#n
experience#n
mr#n
also#r
attack#v
greater#a

steady#a
rhythm#n
striking#a
variation#n
change#n
occur#v
only#r
once#r
rule#n
state#v
ringer#n
memorize#v
pattern#n
change#n
know#v
method#n
name#n
series#n
so#r
change#n
peal#n
take#v
about#r
hour#n
look#n
thursday#n
night#n
practice#n
church#n
district#n
give#v
chat#v
church#n
door#n
member#n
here#r
always#r
tower#n
man#n
woman#n
pull#v
rhythmically#r
rope#n
attach#v
same#a
bell#n
first#r
sound#v
here#r
also#r
discordant#a
modern#a
note#n
n't#r
churchgoer#n
enjoy#v
peal#n
bell#n
cool#a
autumn#n
evening#n
most#a
problem#n
lack#n
ringer#n
survey#n
say#v
train#v
bell_ringer#n
england#n
today#n
only#r
still#r
also#r
ringer#n
n't#r
always#r
live#v
bell#n
need#v
ring#v
small#a
rural#a
parish#n
inner_city#n
church#n
council#n
program#n
attract#v
train#v
ringer#n
only#r
partly#r

idea#n
work#n
involve#v
ringer#n
stand#v
circle#n
foot#n
ahead#r
other#a
prizefighter#n
stance#n
pull#v
rope#n
small#a
hole#n
high#a
ceiling#n
ringing#n
chamber#n
speak#v
snake#v
rope#n
seem#v
make#v
much#a
sound#n
bell#n
muffle#v
ceiling#n
totally#r
absorb#v
other#a
church#n
britain#n
set#n
bell#n
once#r
own#a
band#n
ringer#n
herald#v
sunday#n
morning#n
evening#n
service#n
now#r
only#r
local#a
ringer#n
remain#v
other#n
here#r
today#n
live#v
elsewhere#r
belong_to#v
group#n
ringer#n
include#v
octogenarian#n
youngster#n
training#n
successful#a
say#v
mr#n
right#r
now#r
lucky#a
year#n
keep#v
new#a
ringer#n
add#v
bright#a
sign#n
grow#v
number#n
woman#n
enter#v
once#r
field#n
more#a
third#a
ringer#n
today#n
woman#n
n't#r
accept#v
everywhere#r
however#r
oldest#a
bell_ringing#a
group#n

ringer#n
stare#v
straight#r
ahead#r
use#v
peripheral_vision#a
call#v
watch#v
other#a
rope#n
thus#r
time#v
pull#n
far#r
above#r
belfry#n
huge#a
bronze#n
bell#n
mount#v
wheel#n
swing#v
madly#r
full#a
degree#n
start#v
end#v
surprisingly#r
inverted#a
position#n
skilled#a
drive#v
sunday#n
church#n
church#n
effort#n
keep#v
bell#n
sound#v
many#a
belfry#n
even#r
service#n
tower#n
scrape#v
say#v
retired#v
worker#n
try#v
train#v
youngster#n
disco#n
dance#n
just#r
drift#v
away#r
worry#v
old_age#a
flightiness#n
youth#n
diminish#v
rank#n
country#n
found#v
remain#v
fact#n
particularly#r
galling#a
woman#n
group#n
sole#a
source#n
ringer#n
britain#n
most#a
prestigious#a
church#n
britain#n
woman#n
file#v
equal_opportunity#n
suit#n
extent#n
problem#n
surface#v
summer#n
series#n
letter#n
weekly#a
newspaper#n
ringer#n
writer#n
sign#v

ringer#n
use#v
wrist#n
advance#v
retard#v
next#a
swing#n
so#r
bell#n
swap#v
place#n
following#a
change#n
well-known#a
detective_story#n
involve#v
church#n
bell#n
english#a
novelist#n
describe#v
passion#n
find#v
satisfaction#n
mathematical#a
completeness#n
mechanical#a
perfection#n
ringer#n
add#v
fill#v
group#n
keep#v
bell#n
peal#v
history#n
side#n
nationwide#a
survey#n
take#v
year#n
ago#r
nearly#r
third#a
england#n
church#n
bell#n
no#r
longer#r
ring#v
sunday#n
ring#v
easy#a
see#v
why#r
ancient#a
art#n
less#r
complicated#a
version#n
play#v
tune#n
letter#n
red-blooded#a
balanced#a
male#n
remark#v
frequency#n
woman#n
faint#v
peal#n
suggest#v
settle#v
back#r
traditional#a
role#n
make#v
tea#n
meeting#n
torrent#n
reply#n
follow#v
woman#n
ringer#n
observe#v
average#a
male#a
ringer#n
leave#v
quite#r
lot#n
desire#v
badly#r

solemn#a
intoxication#n
come#v
intricate#a
ritual#n
faultlessly#r
perform#v
ringing#n
become#v
bit#n
obsession#n
admit#v
master#n
band#n
england#n
best#a
female#n
ringer#n
passion#n
usually#r
stay#v
tower#n
however#r
more#r
often#r
not#r
ringer#n
think_of#v
church#n
something#n
stick_on#v
bell#n
carillon#n
continental#a
europe#n
consider#v
english#n
childish#a
fit#a
only#r
foreigner#n
change_ringing#n
mind-boggling#a
exercise#n
english#n
invent#v
year#n
ago#r
require#v
physical#a
dexterity#n
bell#n
weigh#v
more#a
ton#n
combine#v
intense#a
mental#a
concentration#n
proper#a
english#a
bell#n
dress#v
decorate#v
acne#n
large#a
frequently#r
unwashed#a
unbearably#r
flatulent#a
peal#n
woman#n
write#v
say#v
year#n
never#r
know#v
lady#n
faint#v
belfry#n
see#v
man#n
die#v
bless#v
wrist#n
advance#v
retard#v
next#a
swing#n
so#r
bell#n
swap#v
place#n

bottom#n
belfry#n
when#r
change#n
complete#v
work_up#v
sweat#n
ringer#n
often#r
skip#v
local#a
pub#n
leave#v
worship#n
other#n
below#r
not#r
well#r
cleric#n
membership#n
church_of_england#n
steadily#r
dwindle#v
strong-willed#a
vicar#n
press#v
equally#r
strong-willed#a
often#r
ringer#n
attend#v
start#v
round#n
high-pitched#a
bell#n
low#a
simple#a
descend#v
scale#n
use#v
large#a
church#n
as#r
many#a
bell#n
then#r
signal#n
ringer#n
begin#v
vary#v
order#n
bell#n
sound#v
alter#v
steady#a
rhythm#n
striking#a
variation#n
change#n
occur#v
only#r
once#r
following#a
change#n
well-known#a
detective_story#n
involve#v
church#n
bell#n
english#a
draw#v
people#n
church#n
say#v
live#v
hope#n
ringer#n
draw#v
full#a
life#n
sort#n
parliament#n
ringing#n
group#n
aim#v
improve#v
relations#n
thus#r
time#v
pull#n
far#r
above#r
belfry#n

service#n
year#n
ago#r
vicar#n
get#v
so#r
ringer#n
attend#v
service#n
sack#v
entire#a
band#n
ringer#n
promptly#r
set_up#v
picket_line#n
protest#n
club#n
treat#v
tower#n
sort#n
separate#a
premises#n
say#v
new#a
band#n
today#n
several#a
member#n
congregation#n
still#r
rule#n
state#v
ringer#n
memorize#v
pattern#n
change#n
know#v
method#n
name#n
series#n
so#r
change#n
peal#n
take#v
about#r
hour#n
look#n
thursday#n
night#n
practice#n
church#n
district#n
give#v
idea#n
work#n
involve#v
ringer#n
stand#v
circle#n
foot#n
ahead#r
huge#a
bronze#n
bell#n
mount#v
wheel#n
swing#v
madly#r
full#a
degree#n
start#v
end#v
surprisingly#r
inverted#a
position#n
skilled#a
ringer#n
use#v
bell#n
muffle#v
ceiling#n
totally#r
absorb#v
ringer#n
stare#v
straight#r
ahead#r
use#v
peripheral_vision#a
call#v
watch#v
other#a

n't#r
enough#a
ringer#n
ring#v
more#a
bell#n
bell#n
fall#v
silent#a
follow#v
dustup#n
church#n
attendance#n
vicar#n
refuse#v
talk_about#v
say#v
wound#n
here#r
come#v
ringer#n
above#r
make#v
sound#n
net#n
other#a
prizefighter#n
stance#n
pull#v
rope#n
small#a
hole#n
high#a
ceiling#n
ringing#n
chamber#n
speak#v
snake#v
rope#n
seem#v
make#v
much#a
sound#n
very#r
obvious#a
exit#n
congregation#n
prayer#n
bell_ringer#n
bell#n
rope#n
vicar#n
nearby#a
church#n
feel#v
fault#n
stairs#n
bell_tower#n
locate#v
next#a
altar#n
so#r
crunch#n
crunch#n
crunch#n
bang#n
bang#n
bang#n
say#v
admit#v
mixed#v
feelings#n
issue#n
vicar#n
active#a