UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

Modalidad Clásica

Escuela de Ciencias de La Computación

"Modelo Multilingüe para la Extracción de Información a nivel de Frases usando Técnicas de Procesamiento de Lenguaje Natural y Recursos Multilingües"

Trabajo de fin de carrera previo a la obtención del Título de Ingeniero en Sistemas Informáticos y Computación.

AUTORA:

Nero Ortega Elizabeth Margarita

DIRECTOR:

Ing. Riofrío Calderón Guido Eduardo

CODIRECTOR:

Ing. Sucunuta España Manuel Eduardo

Loja, 2010

CERTIFICACIÓN

Ing.

Guido Riofrío

DIRECTOR DEL PROYECTO DE FIN DE CARRERA

C E R T I F I C O:

Que el presente trabajo de fin de carrera previo a la obtención del título de Ingeniero en Sistemas Informáticos y Computación, titulado **"MODELO MULTILINGÜE PARA LA EXTRACCIÓN DE INFORMACIÓN A NIVEL DE FRASES USANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL Y RECURSOS MULTILINGÜES"** realizado por la profesional en formación ELIZABETH MARGARITA NERO ORTEGA ha sido orientado, revisado y corregido bajo mi dirección por lo que autorizo su presentación.

Loja, 11 de Noviembre del 2010.

f) ----------

CERTIFICACIÓN

Ing.

Manuel Sucunuta

CODIRECTOR DEL PROYECTO DE FIN DE CARRERA

C E R T I F I C O:

Que el presente trabajo de fin de carrera previo a la obtención del título de Ingeniero en Sistemas Informáticos y Computación, titulado **"MODELO MULTILINGÜE PARA LA EXTRACCIÓN DE INFORMACIÓN A NIVEL DE FRASES USANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL Y RECURSOS MULTILINGÜES"** realizado por la profesional en formación ELIZABETH MARGARITA NERO ORTEGA ha sido orientado, revisado y corregido bajo mi dirección por lo que autorizo su presentación.

Loja, 11 de Noviembre del 2010.

f) ----------

DECLARACIÓN Y CESIÓN DE DERECHOS

"Yo Elizabeth Margarita Nero Ortega" declaro ser autora del presente trabajo y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales.

Adicionalmente declaro conocer y aceptar la disposición del Art. 67 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que en su parte pertinente textualmente dice: "Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se realicen a través, o con el apoyo financiero, académico o institucional (operativo) de la Universidad"

f)----------

Autora

AUTORÍA

Las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo, son de exclusiva responsabilidad del autor.

f) ----------

Elizabeth Margarita Nero O

AGRADECIMIENTO

Agradezco a Dios por haberme permitido culminar mis estudios superiores, a mi familia por haberme brindado su apoyo y comprensión y a mis profesores quienes contribuyeron en mi formación profesional, en especial al Ing. Guido Riofrío Director de esta Tesis, quien con su conocimiento y motivación ha guiado el desarrollo de la misma.

También quisiera mencionar a mis compañeros por los ánimos de aliento para culminar este trabajo de investigación.

Loja, 2010

Elizabeth Margarita Nero Ortega

DEDICATORIA

El presente trabajo de investigación representa la culminación de mi carrera universitaria y el inicio de mi vida profesional, por tal motivo está dedicado de manera especial a Dios quien me ha brindado sabiduría, salud y vida, a las personas que más quiero en la vida, mis padres Sr. José Nero y Sra. Olga Ortega, quienes con su ejemplo me han inculcado buenos valores y me han dado un ejemplo de vida que me ha permitido cumplir con mis metas propuestas y a mis hermanos quienes siempre me han brindado su apoyo incondicional.

Me parece importante también dedicar este trabajo a todos mis profesores y amigos con los cuales he compartido momentos de tristeza y alegría.

Loja, 2010

Elizabeth Margarita Nero Ortega

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Luego de haber aparecido el computador y el internet en la vida de los seres humanos surgieron nuevas áreas de investigación como: el Procesamiento de Lenguaje Natural, el estudio del multilingüismo entre otras. Las áreas de investigaciones mencionadas son el centro de atención en la actualidad ya que aún no se ha podido eliminar por completo el problema de poder recuperar con mayor precisión lo que el usuario necesita, sumado a esto, la dificultad para recuperar información que está en diferentes lenguajes, por esta razón se desarrolla la presente tesis la misma que cubre el estudio del PLN orientado al multilingüismo y los recursos multilingües como BalKanet y EuroWordNet que existen para poder reutilizarlos, terminando en el planteamiento de un modelo que trata de extraer las mejores características lingüísticas con significado adecuado según el contexto de la frase, con técnicas del PLN, para luego pasar dichas características a un lenguaje universal con lo cual la frase puede ser fácilmente pasada a cualquier lenguaje.

Dentro del PLN se hace un estudio de los modelos, técnicas y herramientas que existen para poder extraer características lingüísticas adecuadas al contexto de la frase y se usa Word Sense Desambiguation para asignar un sentido apropiado a las características antes mencionadas.

Existen varios recursos lingüísticos los mismos que para su construcción se ha necesitado mucho tiempo y un equipo de trabajo amplio, por lo cual se hace un estudio de los mismos para poder agruparlos y así aprovechar su existencia en el modelo, que se plantea como resultado de la presente tesis.

CAPÍTULO 1

1. INTRODUCCIÓN

El auge de las computadoras en la vida de los seres humanos fue el primer paso en el desarrollo de la tecnología, luego hubo otro gran avance, el Internet, con lo cual aparecieron nuevas áreas de investigación como: **Procesamiento del Lenguaje Natural (PLN)** área de estudio de la Inteligencia Artificial con el objetivo de poder hacer que el computador responda a lo que desea realizar el ser humano, estudios de los recursos lingüísticos, la multilingüalidad entre otras.

El PLN es una disciplina que tiene diversas áreas de aplicación una de ellas es la recuperación de información, la cual se está utilizando hoy en día con los buscadores en la web. Pero al problema de cómo saber que información es la que el usuario necesita (uso de buscadores) se suma la diversidad de lenguajes en los cuales está disponible dicha información, por cuanto ha surgido el tema de multilingüalidad, que se presenta actualmente como barrera para encontrar y compartir información relevante en la web.

El volumen de datos inaccesible a causa de las barreras idiomáticas crece cada año porcentualmente. Para darnos cuenta de cuan diversos son los idiomas que se utilizan en la web podemos observar la Figura 1.1.

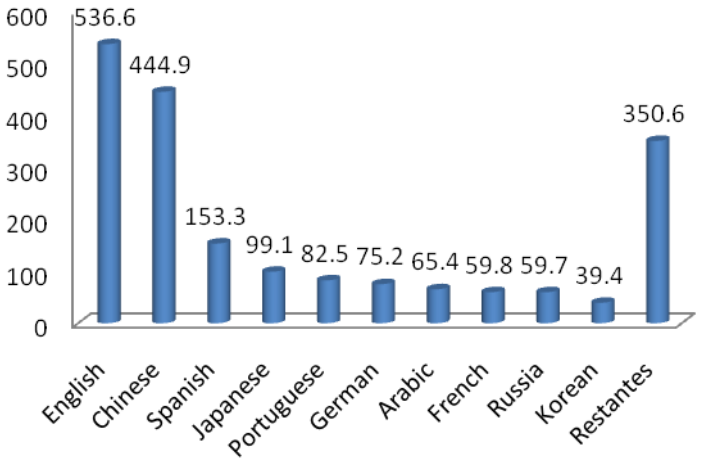


Figura 1.1: Top 10 de lenguajes usados en el Internet por millones de usuarios [13]

El PLN para solventar en algo la diversidad de lenguajes en la disponibilidad de información, ha utilizado la traducción automática entre pares de idiomas, lo cual no resultó tan eficiente debido a la complejidad de cada lenguaje y a que muchas veces no hay una correspondencia adecuada entre los mismos. Por cuanto surge el tema de utilizar el PLN para la implementación de recursos multilingües más no para la traducción de pares de idiomas, investigación que se estará tratando a lo largo del desarrollo de la presente tesis.

El PLN como ya lo mencione tiene muchas áreas de aplicación por cuanto es conveniente delimitar en que ámbito se lo va ha utilizar, en este caso es la Recuperación de Información, dentro de lo que corresponde al análisis de la consulta de los usuarios en Lenguaje Natural.

La combinación del tratamiento de los recursos multilingües (barrera actual) con el procesamiento de lenguaje natural está dentro de la disciplina Ingeniería Lingüística denominada también tecnología del Lenguaje, ya que esta aprovecha el marco de conocimiento del PLN y lingüístico como la traducción, terminología y lingüística Computacional.

Los Sistemas de PLN en sus inicios no tenían apoyo de la lingüística, este conocimiento se incorporo a partir de los años sesenta, y se convierte en uno de sus componentes más importantes. A partir de entonces se definió una área de conocimiento llamada **Lingüística Computacional (CL Computational Linguistics**) apoyada por la **Asociación para la Lingüística Computacional (ACL Association for Computational Linguistics)[1].**

Procesamiento del lenguaje Natural (PLN)

El Procesamiento de Lenguaje Natural fue uno de los primeros conceptos de la Inteligencia Artificial, el PLN ha desempeñado múltiples papeles en el contexto de la IA, y su importancia dentro de este campo ha crecido y decrecido a consecuencia de cambios tecnológicos y científicos. Los primeros intentos de traducir texto por ordenador utilizando el PLN fueron a finales de los cuarenta, pero a finales de los 50 esto fracasó debido a la baja potencia de los ordenadores y la escasa sofisticación

[1] The Association for Computational Linguistics, http://www.aclweb.org/

Lingüística. Sin embargo en la década de los sesenta y setenta se intentó realizar interfaces en Lenguaje Natural para acceder a base de datos y otras aplicaciones teniendo éxito en esto. A partir de ello es que en la década de los 80 y 90 se dio más interés a la traducción automática basándose en el Procesamiento de Lenguaje Natural de tal manera que también empezó el estudio del PLN con diferentes áreas de aplicación.

Definición

"El Procesamiento del Lenguaje Natural consiste en el estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos". Un sencillo ejemplo de PLN es un corrector ortográfico de un procesador de textos que todos hemos empleado alguna vez [5].

Componentes

- **Análisis Morfológico:** Este análisis es el más simple de todos pero importante ya que de aquí se derivan las estructuras que se deben tener para generar palabras para nuestro lenguaje que estemos tratando.

- **Análisis Sintáctico:** Estructura de una oración, se utiliza varios métodos para este análisis:
  - **Ascendente[2].-** construcción de un árbol de análisis sintáctico de una cadena de texto, desde las hojas del árbol hasta la raíz.

  - **Descendente[3].-** se caracterizan porque se analizan la cadena de componentes léxicos de izquierda a derecha, se obtiene la derivación más a la izquierda y el árbol de derivación se construye desde la raíz hasta las hojas.

[2] Departamento de Ingeniería de la Información y las Comunicaciones Universidad de Murcia Análisis Sintáctico Ascendente, http://ants.dif.um.es/staff/juanbot/traductores/files/20022003/tema5.pdf

[3] Análisis Sintáctico Descendente, http://informatica.uv.es/docencia/iiguia/asignatu/2000/PL/2007/tema4.pdf

- **Redes de transición.-** utilizado más para procesamiento de lenguaje natural que nos permita acceder a bases de datos. Por ejemplo escribir en el idioma español "Recuperar datos de personas matriculadas y la base de datos" me devuelva dichos datos sin tener que utilizar un lenguaje técnico.

El Análisis Morfológico y Sintáctico no depende mucho del Lenguaje Natural que se intente procesar. El resultado del análisis sintáctico de una frase es una estructura de árbol.

- **Análisis Semántico:** Este análisis sí depende del lenguaje natural que se esté procesando, por ejemplo no es lo mismo el tratamiento semántico para acceso a una base de datos que para traducciones de texto. Este análisis se refiere al significado de las frases del lenguaje procesado. Este análisis consta básicamente de dos partes:

  - Representar en forma lógica las frases pero fuera de un contexto
  - Eliminar la ambigüedad en el significado de la frase
  - Inferir conocimiento a partir del significado

  Para este análisis se pueden utilizar diversos recursos por ejemplo diccionarios semánticos etc.

- **Análisis Pragmático:** Analizar las oraciones del lenguaje natural dentro de un contexto, entendiendo mucha información que está sobreentendida en una oración que pertenece a un contexto determinado; de esta forma la pragmática estudia "las acciones del discurso" y las situaciones en las cuales el lenguaje es usado.

Los recursos multilingües son recursos de un sistema en diferentes Lenguajes, para su implementación se puede utilizar diccionarios, tesauros o corpus multilingües etc. Un ejemplo se lo puede encontrar en el tesauro EuroVoc de la Comunidad Europea, que abarca 9 idiomas y se utiliza en la actualidad para la recuperación de documentos europeos [3].

Hasta la fecha se han realizado varias investigaciones referentes al PLN utilizado para temas como traducción automática, recuperación de información, extracción de

información para resúmenes, tutores inteligentes (enseñanza asistida por computador), corrección de faltas ortografías, reconocimiento del habla, síntesis de voz y respuesta a preguntas [1]. Es necesario tener presente que dichas investigaciones no son de corta duración al contrario necesitan de mucho tiempo es decir varios años.

El tema de investigación de recuperación de información es muy analizado ya que la tendencia de los buscadores en la web es que estos sean mucho más precisos a consultas que entiendan casi en su totalidad el Lenguaje Natural que maneja el usuario dando como resultado mejores respuestas a las búsquedas y no simplemente las más relevantes de acuerdo a análisis de palabras específicas dentro de la consulta. Dentro de esto también existe actualmente la limitante de buscar información sólo en el idioma nativo, convirtiéndose esto en una barrera de lenguaje en el proceso de búsqueda de información.

En la actualidad existen entidades dedicadas por completo a la investigación sobre Procesamiento de Lenguaje Natural una de ellas es la **Sociedad Española para el Procesamiento de Lenguaje Natural (SEPLN)** que fue creada en 1982, sus objetivos son: actividades referentes a las tecnologías del Lenguaje Humano, foros de encuentros de proyectos, agrupar equipos de investigación de universidades, instituciones y empresas [26].

A nivel internacional la investigación del uso del Procesamiento de Lenguaje Natural para los recursos multilingües tiene un gran apoyo por ejemplo existe el Natural **Language Group at the Spanish National Distance University (*UNED*)**[4] de Madrid creado en 1993 por Prof. Dra. Felisa Verdejo tiene 20 investigadores, sus líneas de investigación son: acceso Inteligente a la Información y adquisición, representación de conocimiento léxico y gramatical, este grupo ha realizado diversos proyectos y publicaciones referentes al tema. Uno de los proyectos es ***"Recuperación de Información Textual en un Entorno Multilingüe con Técnicas de Lenguaje Natural"***[5] el cual tuvo una duración de 3 años. También es necesario nombrar el proyecto Evaluación de la mejor práctica y colaboración para acceso a información multilingüe[6]. Entre otros de los grupos que se pueden mencionar son: **Grupos de investigación de la**

[4] Natural Language Group at the Spanish National Distance University (UNED), http://nlp.uned.es/

[5] Recuperación de Información Textual en un Entorno Multilíngüe con Técnicas de Lenguaje Natural, http://nlp.uned.es/item/

[6] Evaluación de la mejor práctica y colaboración para acceso a información multilingüe, http://www.trebleclef.eu/

UPM[7] (tiene departamentos como: Ingeniería de Sistemas Telemáticos, de Matemática Aplicada, de Inteligencia Artificial, Ingeniería y Arq. Telemáticas. Las líneas de investigación son: creación de recursos lingüísticos, desarrollo de herramientas eficientes, recuperación de Información y búsqueda de Respuestas) y **Grupo de investigación de la UAM: NLP@UAM** (tiene departamentos de: lingüística e Ingeniería Informática. Las líneas de investigación son: creación de recursos lingüísticos multilingües, compilación de corpus y bases de datos, anotación morfosintáctica en español y árabe, recuperación de Información, búsquedas de respuestas). Uno de los departamentos de la UAM que es necesario mencionar es el **Laboratorio de Lingüística Informática (LLI-UAM)**[8], el mismo que fue creado en 1988, tiene varias líneas de investigación entre ellas: Compilación de corpus orales y escritos, multilingües y multimodales, Anotación lingüística en todos los niveles: fonológico, morfológico, sintáctico, semántico y pragmático, Treebanks, Bases de datos acústicas, Diccionarios electrónicos, Extracción de información, Gramáticas computacionales, Herramientas para manejo de corpus lingüísticos (orales y escritos) (actuales y diacrónicos), Herramientas informáticas para estudios lingüísticos y/o filológicos, Terminología y Traducción automática.

Otro aporte a los recursos multilingües son los tesauro multilingüe que ofrecen la ventaja de correspondencias entre conceptos idénticos expresados en diferentes lenguas lo cual permiten interrogar un sistema documental en la propia lengua del usuario y encontrar los documentos indizados en cualquiera de las lenguas del tesauro [3].

La web semántica un tema muy en auge también está considerando los recursos multilingües para lo cual utiliza ontologías multilingües, un proyecto acerca del tema es **"Recuperación de Información multilingüe en la Web semántica"**[9]**.** Todo esto con el objetivo de disminuir las limitantes de la utilización de la web sólo en lenguaje nativo del usuario.

Para las búsquedas poco precisas se está ya utilizando ontologías las mismas que realizar todo el procesamiento del lenguaje natural.

[7] Grupos de Investigación, Universidad Politécnica de Madrid, http://www.fi.upm.es/?id=investigacion/grupos

[8] Laboratorio de Lingüística Informática (LLI-UAM), http://www.lllf.uam.es/

[9] Instituto de Computación Facultad de Ingeniería – UDELAR 2006 – 2007, Recuperación de Información Bilingüe en la Web Semántica, http://www.fing.edu.uy/inco/grupos/pln/prygrado/InformeRecBilWS.pdf

FlareNet[10] ***"Es una red temática cuyo objetivo es la elaboración de estrategias y recomendaciones para la promoción y el desarrollo de las tecnologías lingüísticas y los recursos lingüísticos asociados por su importancia, para minimizar el impacto de la diversidad lingüística en Europa digital multilingüe"*** [8].

"Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano"[11], el objetivo del proyecto es analizar, experimentar y desarrollar tecnologías inteligentes, interactivas y multilingües de minería de textos, como pieza clave de la próxima generación de motores de búsqueda y análisis textual, sistemas capaces de encontrar "the need behind the query" (la necesidad que subyace a la consulta).

A nivel de las Universidades de Ecuador no se puede encontrar mucha información ni proyectos acerca de Procesamiento de Lenguaje Natural y recursos multilingües.
Por cuanto este tema de investigación **"Modelo Multilingüe para la Extracción de Información a nivel de Frases usando técnicas de Procesamiento de Lenguaje Natural y Recursos Multilingües"** es algo nuevo en nuestro ámbito Nacional pero que tiene mucho futuro en diferentes ámbitos relacionados con la informática como la construcción de buscadores que eliminen la brecha de la diferencia de idiomas de tal manera que presten una mejor funcionalidad y servicio a los usuarios de la web de cualquier parte del mundo.

La presente investigación se desarrollará con el objetivo de plantear un modelo multilingüe que permita realizar la extracción de información a nivel de frases usando técnicas del PLN y Recursos Multilingües. Para llegar a esto se realizará el estudio de:

- PLN
- Herramientas del PLN
- Recursos Lingüísticos

Luego de haber desarrollado todos los temas antes mencionados se planteará el modelo con las respectivas descripciones referentes a sus módulos.

[10] FLARANET, www.flarenet.eu
[11] Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano, intime.dlsi.ua.es/text-mess/doku.php

CAPÍTULO 2

2. PROCESAMIENTO DEL LENGUAJE NATURAL

Durante la historia de la humanidad, la mayor parte del conocimiento se ha difundido, guardado y manejado en la forma de lenguaje natural, en la actualidad esto se mantiene con la diferencia de que este conocimiento es almacenado en digital por lo cual las computadoras deben ser capaces de procesarlo. Pero lo que es conocimiento para los seres humanos no lo es para las computadoras, para ellas son solo un conjunto de datos que pueden ser almacenados o borrados, de esos datos no se pueden hacer inferencias lógicas, no se pueden obtener resúmenes es decir el computador no puede hacer todo lo que una persona podría realizar con dichos datos porque no puede entenderlos como conocimiento [14].

Para resolver dicha situación todos los países sobre todo los más desarrollados del mundo, se están dedicando al estudio de la ciencia que se encarga de habilitar a las computadoras para que entiendan el texto en lenguaje natural. Esta ciencia se denomina Procesamiento del Lenguaje Natural, Tecnología del Lenguaje, o Lingüística Computacional [14].

"Los inicios del PLN se pueden situar después de la segunda guerra mundial. En esa época, sólo EE.UU. estaba capacitada para llevar a cabo investigaciones de este nivel. Desde entonces hasta ahora, las investigaciones se han ido trasladando también a Europa, y se ha evolucionado desde sistemas muy toscos de traducción hasta algoritmos y software con gran potencia de resolución" [15].

El avance en el PLN ha sido básicamente afectado por el avance en el procesamiento del computador y el conocimiento de la lingüística, actualmente no se posee un conocimiento bastante preciso de cómo trabaja el lenguaje humano por lo cual se puede decir que el segundo factor aún es un problema para el PLN.

Para que la computadora pueda verdaderamente ayudar en el procesamiento de texto, se necesita pasar un largo camino de aprendizaje de la estructura del texto en lenguaje natural, durante este camino existen diversos problemas que deben ser resueltos con técnicas adecuadas, dichas técnica en la actualidad las proporciona el PLN.

El PLN considera varios conceptos que son parte de la lingüística entre ellos los guiones, la ortografía, la gramática, el estilo, hechos y coherencia lógica [14].

2.1.Modelos del PLN

Existen varios modelos del PLN que dependiendo de dominio del lenguaje que se esté tratando de procesar se pueden combinar para hacer uso de sus ventajas, por ejemplo

Modelo Simbólico

"Los sistemas simbólicos se basan en la manipulación de símbolos, ellos fueron concebidos por los matemáticos para captar de manera rigurosa y sistemática la demostración de teoremas matemáticos y lógicos" [6]. Este modelo tiene como base la aplicación de gramáticas. Dichas gramáticas pueden ser regulares o redes de transición, gramáticas independientes del contexto, gramáticas de unificación y rasgos

Modelo Estadístico

"La aplicación de la probabilidad y la estadística al estudio del lenguaje tiene una tradición al menos tan antigua como la de los modelos formales. La idea general es inferir conocimiento directamente de los datos, buscando regularidades significativas" [6]. Uno de los conceptos fundamentales en este tipo de modelos es la Teoría de la Información la misma que trata de encontrar modelos matemáticos que gobiernan los sistemas diseñados para comunicar y manipular información.

Modelo Biológico

"Cualquier sistema PLN está organizado en diferentes módulos (reconocimiento léxico y morfológico, análisis sintáctico, interpretación semántica y pragmática). La mayoría de estos sistemas usan una estrategia lineal, que no se corresponde con el procesamiento simultáneo y en paralelo que realiza nuestro cerebro" [6], por tal motivo surgen los modelos biológicos aplicados al PLN.
Estos modelos se dividen en dos grandes grupos: los inspirados en el cerebro, el conexionismo (redes neuronales) y los inspirados en la vida y la evolución, la computación evolutiva (algoritmo genético).

2.1.1. Comparación de los Modelos

Características	Simbólico	Estadístico	Biológico
Idea fundamental	Consiste en sistemas con una representación formal, posee reglas y símbolos	Se basa en el estudio de probabilidades para determinar la correspondencia más cercana y adecuada de un texto en un cierto contexto	Utilizan lo que es la evolución en el PLN considerando reglas, entrenamiento y tests
Técnicas usadas	Gramáticas: generativas, de estados finitos, independientes del Contexto, Unificación y Rasgos	N-gramas(consiste en la identificación de aquellas palabras que suelen aparecer juntas con el fin de tratarlas como una sola unidad conceptual) Cadenas de Markov Árbol de Decisión Gramáticas probabilísticas	Algoritmos Evolutivos Redes Neuronales
Ventajas	Están definidos formalmente Fáciles de entender Se tiene un nivel bajo de la resolución de ambigüedades	Resolución de ambigüedades con mayor precisión Mayor eficiencia para dominios específicos	Adaptación a nuevos dominios Procesamiento paralelo Tolera errores humanos en la comunicación lingüística
Limitaciones	Procesamiento	Dependen	Complejo

	lineal No consideran la ambigüedad Gran esfuerzo para cambios de dominios	demasiado del corpus de entrenamiento	Existe poco desarrollo documental y técnico

Tabla 2.1: Comparación de Modelos

2.1.2. Aplicaciones de los Modelos

Hasta la actualidad los modelos mencionados han sido utilizados para desarrollar diferentes aplicaciones dentro del PLN

Modelo Simbólico: Correctores ortográficos, correctores sintácticos y de estilo

Modelo Estadístico: Etiquetadores estocásticos, desambiguación léxica y sintáctica, reconocimiento del habla y traducción automática

Modelo Biológico: Reconocimiento del habla, simulación y reconstrucción de lenguas, modelos de evolución de lenguas y planificación lingüística

2.2.Problemas Generales

Al hacer que el computador procese el lenguaje natural que los seres humanos usamos, surgen algunos problemas, los mismos que en parte son resueltos hoy en día gracias a las técnicas y herramientas que ofrece el PLN. Dichos problemas son: ambigüedad léxica y variación terminológica (variaciones morfosintácticas, variaciones semánticas (varios significados, sinónimos)).

La ambigüedad léxica se refiere a que una misma palabra puede pertenecer a diferentes categorías gramaticales. Por ejemplo: La palabra “para” puede ser: preposición o forma del verbo parar.

El problema de variaciones morfosintácticas se lo puede entender fácilmente como uno de los problemas comunes en la recuperación de Información por ejemplo la consulta realizada por un usuario, la cual es “recursos de educación” no puede

recuperar un documento que contengan la expresión “recursos educativos” a pesar de que son expresiones equivalentes ya que existe una variación morfosintáctica denominada permutación. Otro tipo de variaciones morfosintácticas son inserciones (recursos audiovisuales de educación es una variación de recursos de educación), o variaciones por coordinación (recursos culturales y educativos es una variación de recursos educativos).

Las variaciones semánticas se refieren a que una palabra puede tener varios significados por ejemplo recurso puede referirse a recurso educativo o recurso humano.

A medida que se fueron estudiando los problemas mencionados surgieron técnicas para resolverlos, entre ellas se tiene: tratamiento de la ambigüedad léxica de las palabras, resolución de la anáfora y la elipsis, tratamiento de formas lógicas y roles semánticos y reconocimiento de entidades con nombre.

Muchas investigaciones en el campo del PLN han estudiado métodos para resolver las ambigüedades léxicas mediante diccionarios, gramáticas, bases de conocimiento y correlaciones estadísticas.

Un problema que es de gran impacto en el PLN es la ambigüedad en el sentido de las palabras correspondiente a la ambigüedad semántica dentro de las variaciones semánticas, una palabra pueda ser interpretada de diferentes formas, es decir, posea más de un significado o sentido (fenómeno lingüístico conocido como polisemia). Para esto existe una tarea llamada **Desambiguación del Sentido de las Palabras** (WSD: Word Sense Disambiguation), la misma que persigue la asignación automática de sentidos a las palabras de un texto o frase [17], se puede decir que WSD dentro del PLN se considera como una tarea intermedia ya que no proporciona datos finales para el usuario, así como la traducción automática o extracción de información.

Como se mencionó anteriormente, para esta investigación se ha escogido el PLN aplicado a la recuperación de información. Uno de los mayores problemas dentro de esta área de estudio es que dicha recuperación solo se la realiza basándose en las coincidencias de palabras claves mas no considerando el contexto de las peticiones del usuario, por cuanto la búsqueda de información es bastante imprecisa.

A todos los problemas mencionados se agrega uno más el de contar con diversas lenguas en el mundo para poder compartir información, a lo cual se denomina multilingüismo de lo cual se hablará en próximos temas.

En la actualidad existen diversos recursos lingüísticos que son de gran ayuda para el PLN, esto también se abordará a lo largo de esta investigación en el capítulo referente a Recursos Lingüísticos.

2.3.Niveles de Estudio

Para que un texto en Lenguaje Natural sea procesado como tal por un computador se necesita que dicha cadena de caracteres pase por diferentes fases de procesamiento, a nivel general puede ser [15]:

- Separar oraciones, palabras y signos de puntuación de forma que sea posible su posterior tratamiento con herramientas como el analizador morfológico y el etiquetador de categorías gramaticales (tokenizing).

- Identificar las raíces de las palabras, contrastándolas con lemas y afijos (stemming).

- Analizar la sintaxis de los conjuntos de términos que conforman la frase.

- Tratar de desambiguar el significado de los términos (en inglés, Word Sense Disambiguation, o WSD).

En cada una de estas fases es necesario disponer de una información empírica que posteriormente se contrasta contra la información introducida, y se procesa en base a las reglas del idioma procesado.

Para entender mejor las fases mencionadas es necesario considerar los niveles de estudio que implica el PLN es decir su arquitectura, la misma que es mostrada en la Figura 2.1, y descrita en los siguientes apartados.

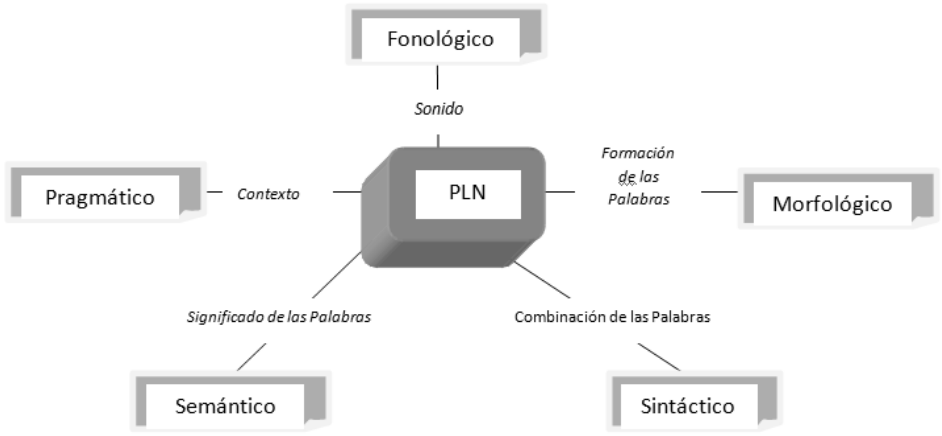


Figura 2.1: Representación de los Niveles del PLN

2.3.1. Nivel Fonológico

Este nivel es considerado para entender el sonido del lenguaje hablado como estructuras abstractas denominadas fonemas. Cabe recalcar que este nivel sólo es aplicable al PLN en el caso del reconocimiento de la voz y de la síntesis de voz. Tema que no es muy relevante en la presente investigación por cuanto no se tiene mayor descripción [15].

2.3.2. Nivel Morfológico

Se trata de los mecanismos de formación de la adaptación de los lemas al contexto de uso, y de las unidades mínimas de modificación de forma (morfemas). El PLN realiza un análisis morfológico a nivel de términos, del análisis se extraen lexemas (también llamados monemas independientes) y morfemas (o monemas dependientes) que pueden ser flexivos o derivativos.

Los morfemas adaptan el lexema al contexto de uso, ya sea añadiendo matices de significado (en el caso de los morfemas derivativos) o marcando relaciones gramaticales (en el caso de los flexivos) con el resto de términos de una oración [15]. En este nivel al igual que en otros existe un proceso de desambiguación.

"Los métodos de la morfología computacional —la rama del procesamiento de lenguaje natural que se encarga del modelado de las formas morfológicas de palabras— varían, y van desde el uso de diccionarios que especifican las formas para cada palabra, hasta las heurísticas que ayudan a adivinarlas" [14] .

La información lingüística utilizada en este nivel incluye los morfemas del lenguaje analizado y las reglas de formación de las palabras. Las reglas de formación morfológicas pueden expresarse mediante gramáticas formales dependiendo del lenguaje que se esté analizando. [18]

En la actualidad, ya se han alcanzado cotas de efectividad importantes, en cuanto a analizadores morfológicos alcanzando el 95% de efectividad [22].

2.3.3. Nivel Sintáctico

Noam Chomsky (1957) fue el primero en hablar sobre este tema en el libro **"Estructuras Sintácticas",** el introdujo las **"Gramáticas Generativas"** en donde se construyen las oraciones a partir de las reglas mencionadas en dichas gramáticas. Este nivel trata de como se relacionan los conjuntos de palabras en los subconjuntos de una frase (denominados sintagmas) o en la frase en general, centra su estudio en la función que cada palabra presenta y las relaciones gramaticales entre las mismas, se identifica el rol de los términos dentro de la frase y las dependencias con los demás términos [15]. Antes de seguir con el tema es necesario aclarar el término sintagma, este se refiere a la unidad de función, es decir, una palabra o conjunto de palabras que realiza una función sintáctica determinada (por ejemplo sujeto) dentro de la oración o dentro de otro sintagma mayor. Cada vez que se distinga un sintagma, hay que especificar qué función realiza, además todo sintagma contiene un núcleo, el cual es la palabra más importante de su estructura, de esta manera el nombre del sintagma va acorde con su núcleo y su función. Para entender mejor lo que es un núcleo y una función se plantea el siguiente ejemplo: la unión de un SN (Sintagma Nominal, núcleo sustantivo) con función de sujeto, con un SV (Sintagma Verbal, núcleo verbo) con función de predicado dará lugar a una unidad superior la oración. Existen diversos tipos de sintagmas estos son: sintagma nominal (SN, grupo de palabras que se articulan alrededor del sustantivo, la estructura de este sintagma es Determinante+ Núcleo +Adyacente), sintagma preposicional (SPREP, se llama así porque va

introducido por una preposición, su estructura es Enlace + Término), sintagma adjetival (SADJ, su estructura es Cuantificador + Núcleo + Complemento del adjetivo), sintagma adverbial (SADV, su estructura es Cuantificador + Núcleo + Complemento del adverbio) y sintagma verbal (SV, su función es la de predicado porque indica lo que se dice o se predica del sujeto, con el cual concuerda en número y persona, la estructura de este sintagma es Núcleo (verbo copulativo o semicopulativo) + Atributo + Complementos). Todos excepto el Sintagma preposicional, reciben su nombre del núcleo que contienen.

En este nivel se puede tener ambigüedad, para reducir esto se utiliza en un nivel posterior, lo que es la interpretación del texto, para lo cual es necesario tener una representación del conocimiento del mundo. Este conocimiento se ha logrado expresar en la actualidad con técnicas de representación del conocimiento, tales como modelos, redes semánticas (constituyen representaciones del conocimiento estructurado, las mismas que están organizadas como grafos con enlaces etiquetados entre nodos. Los nodos son sentidos de las palabras o clases abstractas de sentidos, mientras que los enlaces son relaciones semánticas entre los sentidos por ejemplo la relación de sinonimia) y ontologías.
Un analizador sintáctico robusto es aquel que considera la corrección de errores al procesar frases agramaticales obteniendo árboles sintácticos completos. Al hacer referencia a frases agramaticales se está diciendo que dichas frases no son cubiertas por la gramática propuesta y como resultado no se pueden obtener árboles sintácticos completos. Una técnica para poder tener un análisis sintáctico con corrección de errores es realizar análisis aproximados obtenidos antes de llegar al análisis completo, mediante ítems aproximados, luego de obtener dichos ítems se compara con elementos de la gramática que sean similares obteniendo así análisis aproximados, la similaridad se la mide mediante una función de distancia. Para tener un analizador sintáctico mediante un esquema de análisis con corrección de errores, se necesita decidir primero qué función de distancia se va utilizar para definir el conjunto de ítems aproximados. Uno de los algoritmos para reconocedor de errores es el de Lyon, se pueden tener varios analizadores con corrección de errores pero los aconsejables para cadenas de texto grandes son los analizadores regionales [39].

2.3.4. Nivel Semántico

Trata el significado de los términos, establece relaciones entre significados y significantes. Esto implica el tratamiento de la ambigüedad, tanto en casos de sinonimia y antonimia, como de polisemia. En la recuperación de información actualmente se está agregando este nivel para llevar a cabo lo que es la RI basada en contexto que sólo puede lograrse por medio de la creación y mantenimiento de bases de conocimientos como los tesauros conceptuales y las **"Redes Semánticas".**

Para abordar lo que se refiere a desambiguación semántica se debe considerar todos los sentidos de una palabra y determinar el sentido adecuado de la palabra.

Para este nivel es necesario considerar la tarea de WSD, la misma que trata de solucionar el problemas de asignación de un sentido adecuado a una palabra que fue mencionado anteriormente, dicha tarea se la puede denominar como de clasificación, en ella los sentidos son las clases y el contexto es la evidencia, de esta forma lo que se pretende en WSD es asignar una palabra a una clase (sentido) según una evidencia. Interpretar el sentido correcto para una palabra en un cierto texto o en una conversación es una tarea bastante fácil y usual que la realiza el ser humano. Por el contrario cuando se usa un computador para dicha tarea se convierte en una actividad de mucha dificultad, esto debido a que el computador procesa el texto sin un significado de tal manera que para lograr la comprensión e interpretación adecuada de la información, se necesita analizar exhaustivamente cada una de las palabras y así obtener la interpretación o sentido más apropiado. El contexto de la palabra es el conjunto de las palabras que la acompañan a dicha palabra, junto con las relaciones sintácticas y categorías semánticas. Por ello el contexto ha sido tomado precisamente como el medio más eficaz para identificar el sentido de una palabra polisémica [17].

Se puede decir que la semántica de las palabras tiene dos niveles, la semántica oracional y la semántica del discurso (relación entre oraciones). Tanto la oracional como la de discurso pueden ser representadas de una manera lógica y usan la representación del conocimiento lingüístico que se encuentra en los Recursos Lingüísticos de los cuales se hablará más adelante. La semántica puede ser tomada de una manera más general que

el nivel lingüístico, es decir, puede ser de cierta manera general para diversos idiomas. Con esta idea se dio la creación de un Gramática de estilos. Estas gramáticas son usadas para sacar resúmenes de documentos, fueron propuestas desde el punto de vista de un tópico que, se refiere a tener, una oración representativa como resumen de un párrafo. El estilo se refiere a como debe estar organizada una oración, en donde se deben encontrar las oraciones que representan la idea principal de un párrafo etc. La Gramática de estilos pretende crear reglas de estilo para la escritura propuesta por Williams (1990)[37].

Existen diversos tipos de conocimiento útiles para la WSD, entre estos:

- Categoría sintáctica PoS, se usa para organizar los sentidos de las palabras. La categoría asignada en el PoS se usa para asignar un sentido adecuado, por ejemplo en WordNet handle tiene 5 sentidos como verbo y solo 1 como sustantivo, de esta manera si antes la palabra fue etiquetada con la categoría sustantivo entonces el sentido que se asigna es el correspondiente a cuando se considera dicha palabra como sustantivo.

- Morfología, especialmente la relación entre las palabras derivadas y sus raíces. En el proceso de derivación morfológica, algunos sentidos no se trasmiten. Por ejemplo, en WN agreement tiene 6 sentidos y su raíz agree tiene 7 sentidos de esta manera se puede descartar un sentido para agreement.

- Colocaciones, parque natural, zona verde. Por ejemplo, el sustantivo partido tiene 9 sentidos pero un solo sentido es asignado cuando funciona con el concepto partido de futbol.

- Asociaciones semánticas de las palabras:

  - ✓ **Organización taxonómica**: Relación básica hipo/hiperonímia, por ejemplo, perro es-un animal.

  - ✓ **De dominio**: Se hace referencia a áreas concretas de conocimiento, por ejemplo, en el dominio de los deportes racket es simplemente una raqueta mientras que en el deporte tenis racket es una raqueta de tenis.

- ✓ **Frecuencia de los sentidos**: Sobre los cuatro sentidos de people (gente), el sentido general corresponde a 90% de las ocurrencias de SemCor.

Existen varios métodos de WSD, a continuación se presenta una clasificación de los mismos [17]:

- **Métodos Basados en Conocimiento**: Estos métodos utilizan un conocimiento lingüístico previamente adquirido. Consiste en utilizar recursos externos para desambiguar las palabras, tales como diccionarios, tesauros (vocabularios controlados que representan las relaciones semánticas con otras palabras y sus significados), textos sin ningún tipo de etiquetado e incluso recursos de la Web. Aquí se pueden mencionar los siguientes métodos:

  - ✓ **Método de Lesk** (1986): Aquí se usa el algoritmo de Lesk, el cual fue uno de los primeros algoritmos desarrollados para la desambiguación semántica de todas las palabras en cualquier texto, requiere de un único recurso un diccionario, usa un conjunto de entradas del mismo, una por cada posible sentido y conocimiento sobre el contexto inmediato donde se desarrolla la desambiguación. La principal idea de este algoritmo es desambiguar palabras encontrando el solapamiento entre las definiciones de sus sentidos (palabras similares en sus sentidos) [71]. Dicho de otra manera, dadas dos palabras, W1 y W2, cada una con sus respectivos sentidos Nw1 y Nw2 definidos en un diccionario, para cada par de combinaciones de posibles sentidos, en un inicio se determina el solapamiento con las correspondientes definiciones contando el número de palabras que tienen en común. Luego el par de sentidos con el mayor solapamiento es seleccionado, asignándose así a cada palabra un sentido apropiado, es necesario mencionar que existen variaciones de este algoritmo, en la Figura 2.2 se muestra los principales pasos de este algoritmo.

(1) Para cada uno de los sentidos i de W_1
(2) Para cada uno de los sentidos j de W_2
(3) Calcular el solapamiento (i,j), el número de palabras en común entre las definiciones del sentido i y el sentido j
(4) Encontrar i y j tales que el solapamiento (i, j) sea el máximo
(5) Asignar el sentido i a W_1 y el sentido j a W_2

Figura 2.2: Pasos del Algoritmo Lesk [71]

Un ejemplo de aplicar el algoritmo Lesk se muestra a continuación: Si se considera que queremos desambiguar las palabras **"pine"** y **"cone",** mediante el par de palabras **"pine cone"**. El diccionario Oxford Advanced Learner's define cuatro sentidos para **"pine"** y tres sentidos para **"cone"**, tal y como muestra la Figura 2.3.

Pine
1* seven kinds of evergreen tree with needle-shaped leaves
2* pine
3 waste away through sorrow or illness
4 pine for something, pine to do something cone
1 solid body which narrows to a point
2 something of this shape, whether solid or hollow
3* fruit of certain evergreen trees (fir,pine)

Figura 2.3: Sentidos para las palabras **"pine" y "cone"** [71]

En la Figura 2.4 se puede ver el solapamiento existente entre cada sentido de **"pine"** y cada sentido de **"cone"**.

Pine#1 ∩ Cone#1 = 0
Pine#2 ∩ Cone#1 = 0
Pine#3 ∩ Cone#1 = 0
Pine#4 ∩ Cone#1 = 0

Pine#1 ∩ Cone#2 = 0
Pine#2 ∩ Cone#2 = 0
Pine#3 ∩ Cone#2 = 1
Pine#4 ∩ Cone#2 = 0

Pine#1 ∩ Cone#3 = 0
Pine#2 ∩ Cone#3 = 1
Pine#3 ∩ Cone#3 = 0
Pine#4 ∩ Cone#3 = 1

Figura 2.4: Solapamiento entre conceptos [71]

El primer sentido de **"*pine*"** y el tercero de **"*cone*"** tienen el máximo solapamiento entre todas las posibles combinaciones de sentidos, con dos palabras en común: **"evergreen"** y **"tree",** por lo tanto, estos son los sentidos seleccionados por el algoritmo de Lesk para dichas palabras.

- ✓ **Similitud Annealing**: Este método trata de solucionar el problema principal del algoritmo inicial de Lesk (se tienen que hacer demasiadas combinaciones entre conceptos cuando se trata de desambiguar más de dos palabras con un número elevado de sentidos), para saber como funciona este algoritmo se puede revisar en **"simulated annealing"** por Cowie (1992) [72].

- ✓ **Similitud semántica**: Las medidas de similitud semántica son extraídas a través de redes semánticas. Aquí se incluye métodos que tratan de encontrar la distancia semántica existente entre diferentes conceptos, se puede ampliar este tema en Rada (1989) [62].

- ✓ **Preferencias de selección** (Selectional preferences): Adquiridas de forma automática o semi-automática, como

una forma de restringir los posibles sentidos de una palabra, dichas preferencias están basados en la relación que las palabras tiene con otras palabras en un cierto contexto. Un ejemplo de restricción sería Comer-Comida, Beber-Bebida. Analizando la frase "Mary tiene una gata como mascota", el sentido para "gata" como instrumento de mecánica, no tiene cabida en este contexto porque mascota requiere un objeto animal (gato) de esta manera gata tiene un sentido de animal.

- **Métodos Basados en Corpus**: Se basan en el uso de técnicas estadísticas y de aprendizaje automático para inducir modelos del lenguaje a partir de grandes conjuntos de ejemplos textuales. El propósito de un corpus es servir de fuente de datos, proporcionando ejemplos de oraciones y ejemplos de uso de varias palabras para ser utilizados en algoritmos de aprendizaje automático. Entonces en este caso la desambiguación se realiza mediante un algoritmo que no usa información explícita de una fuente léxica, sino que adquiere conocimientos sobre los sentidos de las palabras a partir de un corpus. Dentro de estos métodos se tienen:

  - ✓ **Métodos supervisados**: Utilizan la clasificación para asigna a una palabra objetivo el sentido más apropiado dado un conjunto de posibles combinaciones de las palabras de su contexto, identifican características o rasgos, estos métodos utilizan clasificadores o técnicas basadas en de aprendizaje automático (Machine learning systems)[12] para llevar a cabo la desambiguación, hacen el uso de textos anotados para su entrenamiento. En inicios de las tareas de PLN las técnicas basadas en aprendizaje automático fueron usadas para resolución de ambigüedades léxicas.

  - ✓ **No supervisados**: Estos métodos encuentran similitud entre contextos para lo cual identifican patrones en los conjuntos de datos sin el beneficio de los datos etiquetados, estos patrones se utilizan para dividir los datos en grupos, donde cada uno de los miembros del grupo posee más características en común con los miembros de un cierto grupo

[12] Mitchell, Tom M. (1997), Machine learning meets natural language

que con otros miembros del resto de grupos, dentro de estos métodos se tienen los métodos distribucionales**.**

Los métodos supervisados por lo general dan mejor resultado aunque su desventaja es que requieren de la labor humana para el entrenamiento del clasificador, lo cual es poco deseable, los métodos no supervisados no sufren de la desventaja mencionada para los supervisados ya que la adquisición de conocimiento no la realizan de corpus anotados. Además debido a que los métodos no supervisados basados en corpus, no se basan en ningún diccionario, repositorio de sentidos o tesauro, no se tienen la restricción de la interpretación de sentidos que el autor del diccionario haya impuesto, concluyendo así que al evitar hacer uso de estos recursos, se garantiza la adaptabilidad de estos sistemas a diferentes campos. Otra ventaja no menos importante, de los métodos no supervisados es que son independientes del lenguaje es decir son fáciles de adaptarlos a cualquier idioma que disponga de un corpus del cual obtener información.

Otra clasificación más general de los métodos para WSD son: sistemas supervisados y sistemas no supervisados. Esta clasificación es la utilizada en la competición Senseval para la evaluación de los distintos sistemas de WSD presentados. Como ya se ha comentado anteriormente, cuando se habla de sistemas supervisados se hace referencia a aquellos sistemas que necesitan de corpus de entrenamiento anotados semánticamente. Mientras que los sistemas no supervisados son aquellos que no necesitan esa anotación para poder funcionar correctamente.

Para explorar los aspectos científicos y técnicos de la WSD y así poder establecer unas bases objetivas para la evaluación sistemas WSD en 1997, se sentaron las bases de una competición libre y voluntaria, denominada **Senseval**[13]. El primer certamen **Senseval** tuvo lugar en 1998 y las lenguas participantes fueron el inglés, el francés y el italiano. La metodología **Senseval** desarrollada permite evaluar los sistemas de desambiguación automáticamente determinando el sentido de una única palabra en un contexto determinado es decir, exclusivamente en función de una muestra léxica. Con **Senseval** -2, las lenguas participantes se incrementaron hasta un

[13] Sens Eval, Evaluation Exercises for the Semantic Analysis of Text, http://www.senseval.org

total de 12 idiomas (inglés, francés, italiano, español, vasco, danés, sueco, holandés, estonio, checo, chino y japonés).

2.3.5. Nivel Pragmático

Un sistema automático que incorpora información pragmática del análisis lingüístico es capaz de procesar textos completos y extraer tópicos generales comprendidos, el ejemplo más ilustrativo lo constituyen las redes neuronales artificiales o sistemas expertos [16].

2.4.Ejemplos de Características de cada Nivel del PLN

- **Nivel Morfológico**
  - Raíces de todas las palabras (stemming)
  - Lemas de todas las palabras
  - Lemas de los nombres y verbos
- **Nivel Sintácticas**
  - Parte de las Oraciones (Part Of Sentences) del las palabras
  - La longitud de la oración
- **Características Semánticas**
  - El foco de la oración
  - Hiperónimos, sinónimos del foco de la oración

2.5.PLN y Multilingüismo

El multilingüismo nace con la necesidad de comunicarse entre personas que poseen diferentes lenguajes. Una forma de cubrir esta necesidad fue la traducción automática, pero esta solución no fue la más adecuada ya que había mucha perdida

de información al pasar de un lenguaje a otro, otro problema fue la consideración de la ambigüedad de las palabras y su posible correspondencia entre las diferentes lenguas y algo aún más problemático fue la falta de correspondencia entre palabras de diversos idiomas (algunas palabras existen en algunos idiomas pero su equivalente en los demás idiomas no).

Considerando que lograr una traducción exacta de una lengua a otra es muy complicado se ha visto la posibilidad de llegar a una representación universal o común a todos los lenguajes para luego si hacer el pase de dicha representación a cualquier lenguaje, a eso se denomina interlingua, una representación semántica común. Una aplicación de este concepto se lo puede observar en EuroWordNet.

Dentro del marco de PLN están áreas como la extracción de información multilingüe y búsqueda de respuestas multilingües, por lo cual el Multilinguismo y el PLN tienen una relación de apoyo. Los sistemas de **Extracción de Información MultiLingüe (Cross Lingual Information Retrieval, CLIR)** y **Búsqueda de Respuestas MultiLingüe (Cross Lingual Question Answering CLQA)** tratan de recuperar información aunque las preguntas (o consultas) se formulen en un idioma y las respuestas se localicen en documentos escritos en otro idioma distinto. La CLQA utiliza muy regularmente un clasificador de preguntas basado en aprendizaje automático, este clasificador se apoya de lo que son las características léxicas, sintácticas y semánticas para poder reconocer partes importantes de las preguntas para luego poderlas clasificar y posteriormente utilizarlas en un sistema completo de búsqueda de respuestas.

La importancia de las aplicaciones multilingües se aumentó mucho por las siguientes circunstancias:

- El uso de la web ya es mundial por lo cual implica el uso de diversos lenguajes para la búsqueda de información.
- Con la formación de la Unión Europea, las oficinas europeas manejan los documentos en 12 lenguajes oficiales, y este número va a crecer más con la expansión de la unión a otros países Europeos.
- Con el crecimiento de democracia en los países multilingües, se fortalece la posición de los lenguajes aunque no oficiales pero muy usados –como es el español en los EE.UU.

- Con el desarrollo técnico de los países de tercer mundo, la revolución informática empieza a llegar a los mismos, donde en muchos casos hay decenas de lenguajes usados y en muchos casos ellos son oficiales.

Para manejar el tema del multilingüismo se han creado varios recursos lingüísticos como son los tesauros multilingües, las ontologías multilingües, además existe un aporte muy importante en la actualidad esto es el **UNL Universal Networking Language**. [24]

Los tesauros multilingües sobre un dominio determinado permiten traducción de términos específicos de un dominio que quizá no se puede encontrar en un diccionario bilingüe, esto ayuda por ejemplo cuando las palabras de un idioma no tienen equivalencia en otro idioma. Un ejemplo de tesauro multilingüe sobre el dominio médico es el metatesauro de UMLS [42], otro ejemplo es el tesauro EuroVoc de la Comunidad Europea, abarca 9 idiomas y se utiliza en la actualidad para la recuperación de documentos europeos. Este tipo de tesauros fueron los primeros tipos de recursos lingüísticos diseñados para la recuperación de información translingüe.

En la actualidad además de hablar de multilingüismo ha surgido otra definición esto es el término translingüe, lo cual se refiere a que se hace correspondencias entre lenguajes completamente diferentes por ejemplo una búsqueda de información translingüe sería una búsqueda en español y como resultado su correspondencia en el idioma chino.

Después de todo lo mencionado anteriormente en el tema de PLN y Multilingüismo se puede deducir que el multilingüismo implica correspondencia más exacta entre diferentes idiomas para hacer búsquedas en diferentes lenguas, lo que conlleva a utilizar conocimiento lingüístico, el mismo que puede ser procesado por las máquinas usando el PLN.

2.6.Aplicaciones del PLN

El PLN puede ser aplicado en diferentes ámbitos, según Moreno Sandoval autor del Libro "Lingüística Computacional: Introducción a los modelos simbólicos, estadísticos y biológicos"[37] clasifica las aplicaciones del PLN en:

- Sistemas que tratan de emular la capacidad humana para procesar lenguas naturales. Ejemplos: traducción automática, recuperación y extracción de información, interfaces hombre-máquina.

- Sistemas que ayudan en las tareas lingüísticas. Entre las aplicaciones de este tipo se tiene: herramientas de análisis textual, herramientas de manejo de corpus, bases de datos lexicográficas.

- Programas de ayuda a la escritura y composición textual. Este tipo de aplicaciones han sido ampliamente desarrolladas, algunos ejemplos de estas son: correctores ortográficos, correctores sintácticos y de estilo.

- Enseñanza asistida por computador. Este es un campo de aplicación en continua expansión y que tiene varias vertientes. La más importante es la de los programas educativos para la enseñanza de las lenguas extranjeras.

2.7.Herramientas

Las herramientas para el PLN consisten en corpus, lexicones y analizadores en los diferentes niveles del PLN. Muchas veces se piensa que, herramientas son sólo los analizadores, pero esto no es así, las herramientas también son los diferentes algoritmos y sistemas implementados para resolver las ambigüedades.

Se ha realizado un estudio de las herramientas más conocidas para el PLN y que ya han sido probados en diferentes sistemas, además se ha podido encontrar implementación de algoritmos que interactúan con recursos lingüísticos que actualmente están siendo muy usados en tareas del PLN, como resultado de esto se puede mencionar herramientas como: MPRO programa para el análisis morfológico y sintáctico de textos en español, TreeTagger para obtener el POS y el lema de las palabras, Lingpipe es un Kit de herramientas libre, Porter Stemmer, GATE, MACO, RELAX, Le*ff*e (Léxico de formas flexionadas del Español), 3LB-SAT (3LB-Semantic Annotation Tool), Interfaz para implementación del Método de Marcas de Especificación para la desambiguación léxica[14], APOLN Analizador Parcial de Oraciones en Lenguaje Natural y sistemas para WSD. Para poder tener una descripción de cada una de las herramienta ya mencionadas referirse al ANEXO 1.

[14] WSD USING SPECIFICATION MARKS METHOD, http://gplsi.dlsi.ua.es/wsd

Es necesario mencionar que GATE no es solo una herramienta sino que es una arquitectura completa para PLN, algo también importante es la Interfaz para implementación del Método de Marcas de Especificación para la desambiguación léxica ya que esta interactúa con WordNet Figura 2.5, recurso lingüístico muy usado por tener valiosas características.

Figura 2.5: Resultado de usar la implementación de WSD usando el método de marcas de especificación

2.8. Discusión

Identificar los niveles del PLN y dentro de cada uno de ellos las unidades más importantes del lenguaje ha servido para poder etiquetar y relacionar conceptos en los recursos lingüísticos de tal manera que se pueda tener información mejor representada logrando con ello obtener características adecuadas de cada lenguaje para poder hacer una correspondencia adecuada en otros idiomas, tratando con ello también de resolver el problema de pérdida de información al pasar de un lenguaje a otro.

Al realizar el estudio del PLN en el nivel semántico se ha podido deducir que existen recursos lingüísticos que brinden la información necesaria para tratar la ambigüedad semántica de las palabras usando Word Sense Desambiguation.

Gracias a que se ha tenido el estudio del PLN también se ha podido ir relacionando en los recursos lingüísticos las diferentes anotaciones de los niveles de tal manera que esto sirva para la tarea de desambiguación.

CAPÍTULO 3

3. RECURSOS LINGÜÍSTICOS

En el proceso de realizar el tratamiento de un lenguaje natural con la ayuda de un computador surgen varias necesidades entre ellas la de obtener unos elementos que le permitan entender las unidades mínimas de tratamiento del lenguaje, en base a la morfología del lenguaje (lemas y flexiones), la estructura y relaciones entre términos (sintaxis) y los significados implícitos (semántica). Por esta razón existen los recursos Lingüísticos.

En cada nivel de estudio del PLN se necesita tener una información empírica que se la puede obtener con los recursos lingüísticos además de las reglas propias del lenguaje que se esté procesando.

Una desventaja que existe con los recursos lingüísticos es que algunos de ellos están restringidos por licencias.

3.1. Recursos Lingüísticos más usados

3.1.1. Lexicones

Un Lexicón es una colección de términos utilizados en un idioma, sobre los que se puede incluir o no una descripción. Unos de los lexicones más familiares son los diccionarios, los lexicones también pueden incluir información sobre prefijos y sufijos, raíces y otras formas o variaciones [15].

Según el grado de especialización, un lexicón puede ser general de la lengua o más bien especializado. También se puede hablar de lexicones enciclopédicos (Wikipedia podría considerarse como tal) cuando la base de lemas no sólo se refiere a términos del lenguaje, sino a personalidades, aspectos históricos y demás [15].

El lexicón aporta principalmente información semántica, lo que permite tratar la sinonimia y la antonimia (la relación que establecen dos un significantes en relación a un significado) y la polisemia (la relación entre un significante y dos o más significados) [15].

Para trasladar la estructura del lexicón a la informática, es frecuente que aparezcan abstracciones típicas de las estructuras de datos, como son árboles y grafos en general. También se puede hablar de ontologías (estructuras de representación del conocimiento que aplican el criterio de relaciones entre términos en su sentido más amplio) [15].

Debido al gran esfuerzo que presenta desarrollar y mantener un Lexicón, se han planteado diferentes especificaciones de estructuración entre las cuales se pueden destacar MULTILEX, GENELEX, COMLEX [19].

Dentro de esta clase de recursos existen otros recursos que son las bases de conocimiento, estas fueron creadas con el propósito de disminuir el esfuerzo de desarrollo de los sistemas de lA mediante la reutilización de información ya almacenada. Una de estas bases de conocimiento que ha sido de gran ayuda para el PLN es el WordNet[18] la misma que posee una base de datos que agrupa las palabras en conjuntos de sinónimos llamados synsets y provee definiciones, comentarios y ejemplos de uso de estas palabras y sus sentidos. De esta manera, combina los elementos de un diccionario (definiciones y algunos ejemplos) y los de un tesauro (sinónimos). Actualmente, el tesauro WordNet contiene alrededor de 155000 palabras organizadas en más de 117000 synsets formando un total de más de 206000 definiciones y sentidos. WordNet maneja 4 categorías léxicas (o tipos de partes de la oración) en sus synsets: sustantivos, verbos, adjetivos y adverbios [25], este recurso será ampliado en los siguientes temas de este capítulo.

Aquí también se puede mencionar a EuroWordnet [20] como recurso, este es una base de datos multilingüe con WordNets para varios idiomas europeos (Holandés, Italiano, Español, Alemán, Francés, Checo y Estonio) presenta restricciones de copyright, ofreciendo sólo descargas de muestras. EuroWordNet permite conocer la traducción de un término a otros idiomas, para cada sentido WordNet del término original, esto es útil para el PLN ya que permite obtener según el sentido apropiado la traducción correcta, los principales propósitos de este recurso son: servir de esqueleto de los léxicos semánticos para desarrollar sistemas de reconocimiento y comprensión automática del lenguaje, servir de punto de partida para la elaboración de un léxico para la Traducción Automática, servir de herramienta de aprendizaje de diferentes lenguas y servir a los correctores gramaticales y ortográficos para lograr reglas más precisas haciendo uso de la información semántica, de este recurso se tendrá un mayor detalle en temas posteriores en este mismo capítulo.

3.1.2. Gramáticas Computacionales

Las gramáticas computacionales pueden entenderse como una descripción formalizada del conocimiento lingüístico, en el PLN pueden ser empleadas por las herramientas de análisis en el nivel sintáctico.

Describen la estructura de una lengua en diversos niveles: palabra (gramática morfológica), frase, oración etc. Por ejemplo: Las gramáticas pueden tratar la estructura en términos de significado (semántica y discurso) o en términos de sintaxis.

La relación entre la gramática y los recursos lingüísticos es la plasmación de sus reglas y estructuras en un sistema informático que permita su posterior tratamiento [15] [21].

- Gramáticas de unificación: Tienen como principal característica la codificación de la máxima información posible en el léxico.

- Gramáticas de restricciones: Parten de la anotación de la posibles funciones (funciones identificadas en los sintagmas, tema mencionado anteriormente) sintácticas de una palabra, para luego realizar una desambiguación y seleccionar la función adecuada.

3.1.3. Corpus

Según Leech[15] el término **corpus** se aplica a ***"Un conjunto de material lingüístico que existe en forma electrónica y que puede ser procesado por una computadora con distintos fines como la investigación lingüística y la ingeniería del lenguaje"*** [74].
De la definición anterior se puede decir también, que consiste en un gran conjunto de textos en lenguaje natural que incluye información extra tales como etiquetas (anotaciones del corpus) para cada palabra indicando los constituyentes gramaticales, en ellos se pueden encontrar oraciones, párrafos o fragmentos más extensos que proporcionan ejemplos correctos

[15] Professor Geoffrey Leech, http://www.ling.lancs.ac.uk/profiles/296/

de uso del Lenguaje. Las etiquetas permiten que los corpus sean útiles computacionalmente.

Los corpus sirven para el modelo estadístico del PLN ya que en base a la información lingüística que estos almacenan se puede calcular las frecuencias de las palabras que aparecen en un conjunto de textos, y deducir todas las probabilidades medias y condicionadas, de esta manera lo que se quiere es predecir acontecimientos a partir de cierta información incompleta, por ejemplo el análisis más probable de una oración en un texto a partir de análisis anteriores. Con todo ello se puede deducir por ejemplo si "el" corresponde a un artículo o se refiere a un pronombre personal utilizando las probabilidades.

Los corpus no incorporan necesariamente documentos completos, ya que pueden ser frases sueltas que exponen la estructura sintáctica de un lenguaje de forma empírica.
Dado que el corpus es la base a partir del cual se va a extraer información para el tratamiento posterior, es importante que su selección sea estricta, de acuerdo con los objetivos del sistema para el cual va a estar destinado.
En esencia una función del corpus es ofrecer lo que en Inteligencia Artificial se denomina información de aprendizaje o entrenamiento, es decir, casos a partir de los cuales el sistema puede detectar patrones (aspectos comunes) y discriminantes (aspectos diferenciadores), de modo que establece criterios para tratar las posteriores entradas de texto.
Las anotaciones en un corpus pueden realizarse de dos formas, de manera manual y automática. Cuando se realiza de la primera forma por lo general los primeros datos son marcados por especialistas que definan las ambigüedades, y luego se usan estos datos para el entrenamiento de anotadores automáticos. Cuando se realiza de manera automática se usan *gramáticas computacionales***.** Con las anotaciones que se realizan en los corpus se logra solucionar en un cierto porcentaje el problema de la ambigüedad a nivel morfológico. También se pueden encontrar que la anotación puede ser realizada usando una mezcla de las dos formas mencionadas. Uno de los procesos más usados para la anotación de Corpus es el de Penn Treebank, el mismo que posee dos fases: la de POS tags (asignación automática de etiquetas) mas corrección manual y la información sintáctica (generación automática de subárboles). Algo que se debe considerar dentro de la anotación es el conjunto de POS tags (conjunto de etiquetas para la anotación) a utilizar ya que reducir el

tamaño del conjunto de etiquetas reduce las posibilidades de inconsistencias en el etiquetado.

Luego de realizar la anotación se debe calcular la probabilidad asociada a cada palabra con una determinada categoría sintáctica. Para ello se aplican alguna técnica estadística **(***probabilidad condicionada, Ley de Bayes, n-gramas, árboles de decisión***).**

Los corpus para ser usados como recurso en el PLN tendrían que tener un alto grado de entrenamiento, lo cual implica un coste bastante elevado en recursos computacionales, ya que se necesita considerar un corpus que contenga millones de millones de palabras etiquetadas, por todo esto puedo decir que los corpus pueden ser usados para dominios específicos de aplicación. Por ejemplo sería muy beneficioso un corpus para la construcción de un buscador solo de ciertas áreas de medicina, se lograría una buena precisión el lo que es PLN en este ámbito. Dentro de un dominio específico se debe decidir el tamaño de la muestra para el corpus de tal manera que se pueda obtener estimaciones fiables con un margen de error aceptable.

Un corpus bastante usado y que utiliza wordNet es SemCor, consiste es un subconjunto del Brown Corpus, donde cada término está manualmente etiquetado con su correspondiente sentido en WodNet. Este recurso puede ser usado para la desambiguación del sentido de las palabras, contiene aproximadamente 700000 palabras. En SemCor todas las palabras están etiquetas gramaticalmente y más de 200000 están también lematizadas y etiquetadas por su sentido de acuerdo a WordNet 1.6, está compuesto por 352 textos, de los cuales 186 textos contienen palabras de clase abierta (sustantivos, verbos, adjetivos y adverbios) y están anotados gramaticalmente, lematizados y con su sentido, en los textos restantes solo los verbos han sido anotados con lemas y sentidos [43].

3.1.3.1. Recomendaciones para la anotación en la Lingüística de Corpus

En Leech, McEnery[16] y Wilson presentan un conjunto de pautas, estándares o recomendaciones de buenas prácticas para la anotación de textos. Existen diversos proyectos dedicados la unificación y

[16] Tony McEnery, http://www.ling.lancs.ac.uk/profiles/Tony-McEnery/

estandarización de esquemas de anotación, un ejemplo de este tipo de proyectos es la iniciativa EAGLES[17] (Expert Advisory Groups on Language Engineering Standards), en la Tabla 3.1 y 3.2 se puede ver lo que define EAGLES, uno de los resultados de dicho proyecto es el Estándar de Codificación de Corpus –Corpus Encoding Standard o CES[18] (1999), este estándar incluye algunos criterios generales que deben considerarse cuando se elabora un esquema de anotación. Las anotaciones deben contener algunos de los distintos niveles del PLN mencionados en el Capítulo 2 de esto se tiene más explicación en las siguiente tema [56].

- **EAGLES**

ADJETIVOS			
Pos	**Atributo**	**Valor**	**Código**
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
3	Grado	Apreciativo	Q
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Caso	-	O
7	Función	Participio	P

Forma	Lema	Etiqueta
Alegres	alegre	AQ0CP00
Alegre	alegre	AQ0CS00
Bonitas	bonito	AQ0FP00
Bonita	bonito	AQ0FS00
bonitos	bonito	AQ0MP00
Bonito	bonito	AQ0MS00
Quemada	quemado	AQ0FS0P

Tabla 3.1: Estandarización de anotación EAGLES

[17] The essentials of EAGLES, http://www.ilc.cnr.it/EAGLES/intro.html
[18] Corpus Encoding Standard, http://www.cs.vassar.edu/CES/

- **EAGLES**

NOMBRES			
Pos	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	-	O
6	Género semántico	-	O
7	Grado	Apreciativo	A

Forma	Lema	Etiqueta
chico	chico	NCMS00
chicos	chico	NCMP00
chica	chico	NCFS00
chicas	chico	NCFP00
tesis	tesis	NCFN000
Antonio	antonio	NP00000

Tabla 3.2: Estandarización de anotación EAGLES

3.1.3.2. Niveles de anotación lingüística

Los corpus no contemplan las anotaciones en los 4 niveles del PLN, algunos consideran sólo dos o tres de los mismos, a continuación se menciona las anotaciones en los niveles más comunes.

- **Anotación de lemas**

La anotación de lemas supone acompañar cada token léxico con su lema, el lema corresponde a la palabra que existe en un diccionario. Este tipo de anotación es usada para la recuperación de información, aquí a los tokens correspondientes de la consulta del usuario se les asigna una palabra que existe en el diccionario del lenguaje en el cual está la consulta [56].

- **Anotación morfosintáctica**

  Éste tipo de anotación es una de las más extendido en la Lingüística de Corpus, también se la llama etiquetación POS (Part-of-speech) o etiquetación gramatical (anotación de la clase gramatical por ejemplo, nombre, verbo, etc.) de cada token léxico en un texto, la información que se obtiene de esta anotación es esencial para el análisis sintáctico y el semántico [56].

- **Anotación sintáctica**

  Esta anotación al igual que la anterior es una de las más extendidas en la Lingüística de Corpus, se trata de asignar anotaciones de las relaciones sintácticas a las anotaciones morfosintácticas [56].

- **Anotación semántica**

  La anotación semántica puede ser vista desde dos perspectivas, desde las relaciones semánticas entre elementos de un texto y la anotación del significado, este nivel de anotación es poco usado aunque si existen estudios de criterios para este tipo de anotación [56].

3.1.3.3. Tipos de Corpus

- **Corpus Paralelo**
  Siguiendo la terminología de Baker (1995) y McEnery (1996) [38] que parece ser la aceptada comúnmente en la actualidad, la diferencia principal de este tipo de corpus con el ya mencionado es que, se tiene un conjunto de documentos, los mismos que son traducidos a diversos lenguajes. Utilizando un corpus paralelo se puede devolver documentos en más de un idioma, aunque se trate de los mismos documentos traducidos. El corpus paralelo se usa para la construcción automática de léxicos y para la investigación sobre la traducción. Un ejemplo de Corpus Paralelo bastante interesante es el

corpus compuesto de manuales técnicos de IBM escritos en francés y en inglés.

- **Corpus Multilingües**
  Su característica principal es que el conjunto de documentos que se tienen en el corpus están en diversos idiomas pero están agrupados por un criterio común. No se trata de documentos traducidos a diferentes idiomas.

3.1.4. Ontologías

Una ontología es una entidad computacional es decir es creada como un recurso artificial, los investigadores de la inteligencia artificial han encontrado que las ontologías son un modelo del conocimiento bastante eficiente para describir formalmente los recursos web, su vocabulario y para hacer explícito de alguna forma el significado de los términos incluidos en las páginas web. De esta definición nace la Semántica Ontológica la misma que es la teoría que estudia el significado del lenguaje humano o lenguaje natural y la aproximación al Procesamiento del Lenguaje Natural (PLN) utilizando las ontologías como recurso central para extraer y representar el significado de textos en lenguaje natural, brindando así la posibilidad de razonar con el conocimiento que se deriva a partir de estos textos representados [56]. De esto se puede decir que las ontologías pueden servir como representación del conocimiento, por ejemplo la estructuración de los lemas puede estar representada en las ontologías.

Las ontologías como representación del conocimiento tienen varias características por ejemplo: son independientes del sistema de procesamiento que se sutilice, los conceptos se organizan mediante relaciones de Hiponimia (relación entre un genérico y sus específicos (flor y rosa), relación es-un) y Meronimia (relación entre un todo y sus partes (coche y chasis, motor), relación parte-de).

Además dentro del PLN las ontologías se están empleando para construir representaciones independientes de la lengua que puedan servir de punto de encuentro entre dos o más lenguas naturales.

Otra relación entre el PLN y las ontologías es que actualmente se tiene la creación de sistemas de mapeo-alineación de términos gracias a las representaciones independientes de la lengua anteriormente mencionadas y su agrupación en ontologías, también se usa ontologías en extracción y análisis de términos de glosarios on-line.

Una aplicación bastante demostrativa del uso de las ontologías con el PLN para la solución del problema de multilingüismo es EuroWordNet, este recurso está compuesto de diversos tipos de ontologías por ejemplo de dominio y de alto nivel.

Un ejemplo de ontología considerada para el PLN es CIRCA Technology Ontology la cual contiene millones de palabras, significados y relaciones entre las palabras. Fue creada por un equipo de lexicógrafos y lingüísticos computacionales.

Como conclusión, las ontologías están siendo utilizadas para el problema de multilingüismo lo cual implica también el PLN para diversas lenguas, actualmente existen 149 ontología multilingües con algún tipo de información lingüística.

3.2.WordNet

3.2.1. Descripción

Sistema con información léxica extraída de forma semiautomática de diccionarios, este sistema ha sido desarrollado en el Cognitive Science Laboratory de la Universidad de Princeton, el contenido de este recurso se organiza mediante una base de datos léxica donde se agrupan conjuntos de palabras (nombres, verbos, adjetivos y adverbios) en grupos de sinónimos llamados synsets: un synset se codifica como un número único de ocho dígitos.

Cada synset representa un concepto distinto y entre cada uno de ellos existen conexiones que expresan relaciones semánticas, conceptuales o léxicas. Como resultado del conjunto de conexiones se tiene una extensa red navegable que con un gran número de inter-relaciones entre palabras.

Entre los conjuntos de relaciones se tienen:

- Sinonimia: Relación léxica entre palabras, son sinónimas dentro de una misma categoría sintáctica (nombre, verbo, adjetivo o adverbio), aquellas palabras que al sustituirse dentro de un contexto determinado el significado no cambia. Por ejemplo a continuación se presenta los sinónimos en Wordnet para bank.

Sinónimos para "bank#1"	Depository financial institution #1 Baking concern#1 Banking company#1 Financial institution#1 Financial Organization#1 Financial organisation#1

- Antonimia: Relación léxica entre palabras , palabras con significados opuestos, por ejemplo:

Antónimos para "clean#1"	dirty #1 soil#1 begrime#1 grime#1 colly#1 bemire#1

- Hiponimía: Relación entre significados de las palabras, este tipo de relaciones se dan únicamente para los nombres. Por ejemplo: "arce" es un hipónimo de "árbol" y "árbol" es un hipónimo de "planta".

Hipónimo para “cat#1”
- domestic#1
- house#1
- Felis domesticus#1
- Felis catus#1
- wildcat#1

- Hiperonimia: Se define como la relación inversa de la hiponimia.

Hiperónimos para “cat#1”
- feline#1
- felid#1
- carnivore#1
- placental#1
- placental mammal#1
- eutherian#1
- eutherian mammal#1
- mammal#1
- mammalian#1
- vertebrate#1

- Meronimía: Relación semántica, identificada como “*un tipo de*” dicho de otra manera una palabra X es merónima de Y si “ *X es una parte de Y* “. Por ejemplo dedos, uñas son merónimos de mano.

Merónimos para “body#1”
- Articulatory system #1
- Digestive system#1
- Gastrointestinal system#1
- Endocrine system#1
- Lymphatic system#1
- Musculoskeletal system#1
- Sensory system#2
- Trunk#3

- Holonimia: Relación inversa de meronimia.

Holónimos para “eye#1”
- Visual system #1
- face#1
- human face1#1

- Troponimía: Relaciona verbos y es el equivalente de la relación de hiponimia para los nombres.

Tropónimos para “eat#1”
- Wash down #1
- gluttonize#1
- gluttonise#1
- fress#1
- wolf#1
- slurp#1
- fare#2

- Entailment: Un término implica al otro. Por ejemplo divorcio implica matrimonio.

Entailment para “eat#1”
- chew #1
- masticate#1
- manducate#1
- jaw#1
- swallow#1
- get down#1

WordNet además de distinguir los significados de cada término mediante synsets establece una relación de orden entre los diferentes sentidos de las palabras, de acuerdo a su frecuencia de aparición es decir da un orden de prioridad a los sentidos de una palabra. Un ejemplo de esto se muestra a continuación [71]:

1. {03912097} *plant#1, Works#1, industrial plant#1 – (building for carrying on industrial labor; "they built a large plant to manufacture automobiles")*
2. {00016858} *plant#2, flora#2, plan lifet#1 – (a living organism lacking the power of locomotion)*
3. {05831211} *plant#3 - (something planted secretly for discovery by another; "the police used a plan to trick the thieves"; "he claimed that the evidence against him was a plant")*
4. {10282477} *plant#4- (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)*

El significado plant#1 es el más frecuente para dicha palabra.

Además de lo ya mencionada wordnet presenta para cada sentido de una palabra una descripción y un ejemplo de uso.

Wornet organiza los conceptos a nivel jerarquías principales de las cuales se deriva todos los demás conceptos, esta jerarquía es: entity, psychological feature, abstraction, state, event, act, human action, human activity, group, grouping, possession, phenomenon.

Para entender como usa WordNet el concepto de red semántica para cada sentido de una palabra dada, se presenta en la Figura 3.1 la red semántica de la palabra "airplane" con el sentido uno que es el más frecuente para dicha palabra, dicha red semántica contiene las relaciones mencionadas anteriormente.

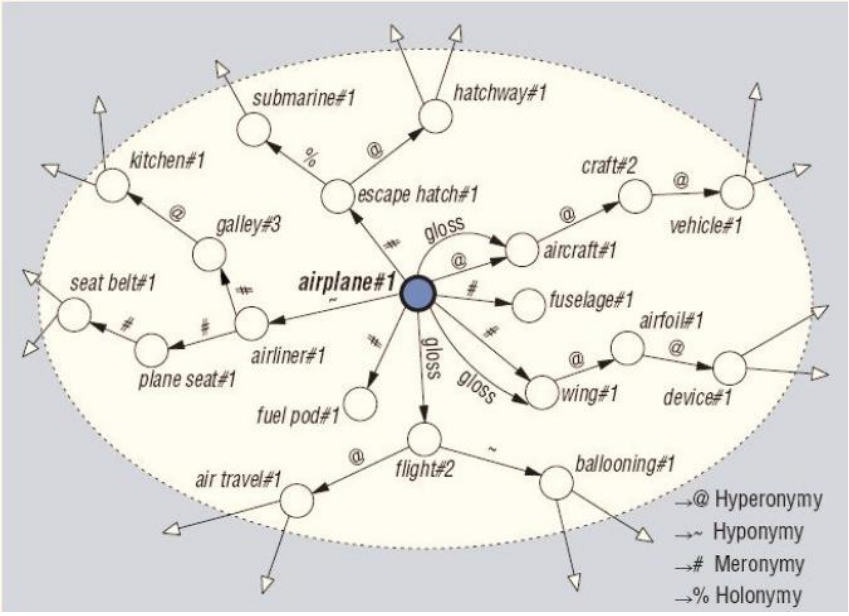


Figura 3.1: Red Semántica para la palabra airplane con sentido#1 según Wordnet [76]

Otro ejemplo de las relaciones semánticas existentes en Word-Net se lo puede ver en la Figura 3.2.

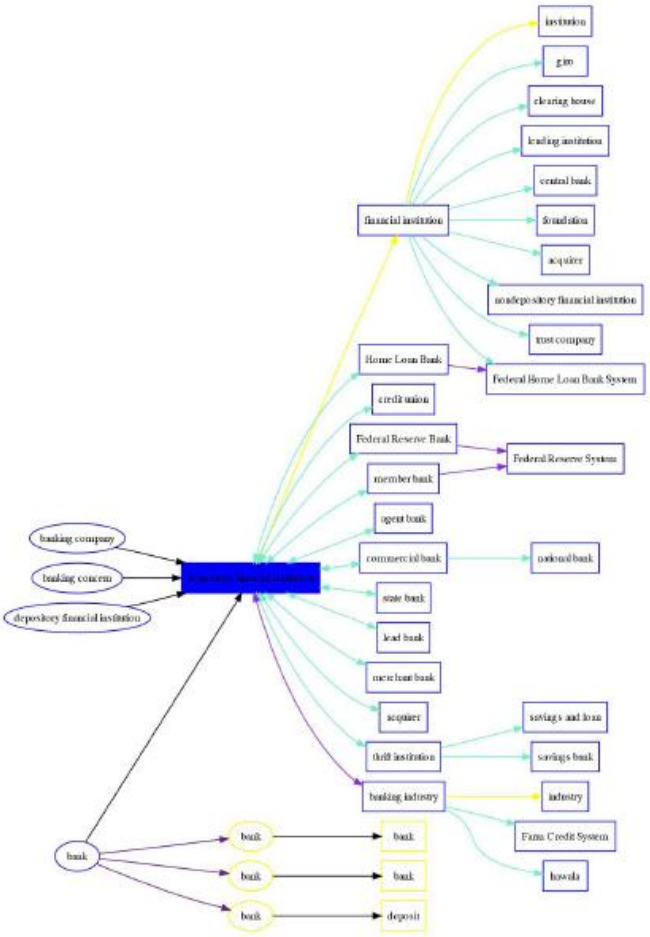


Figura 3.2: Representación simplificada de las relaciones de "bank#1" (sentido más frecuente de bank) [76]

En la Figura 3.2 la relación de sinónimos está representada por flechas de color negro (sinónimos: banking company, banking concern y depository finantial institution), los hipónimos se representan mediante flechas de color verde, los merónimos son repre sentados mediante flechas de color amarillo y los holónimos están representados mediante flechas de color morado.

WordNet tiene diferentes dominios, los mismos que están estructurados desde dos puntos de vista: jerárquicamente Figura 3.3 y semánticamente.

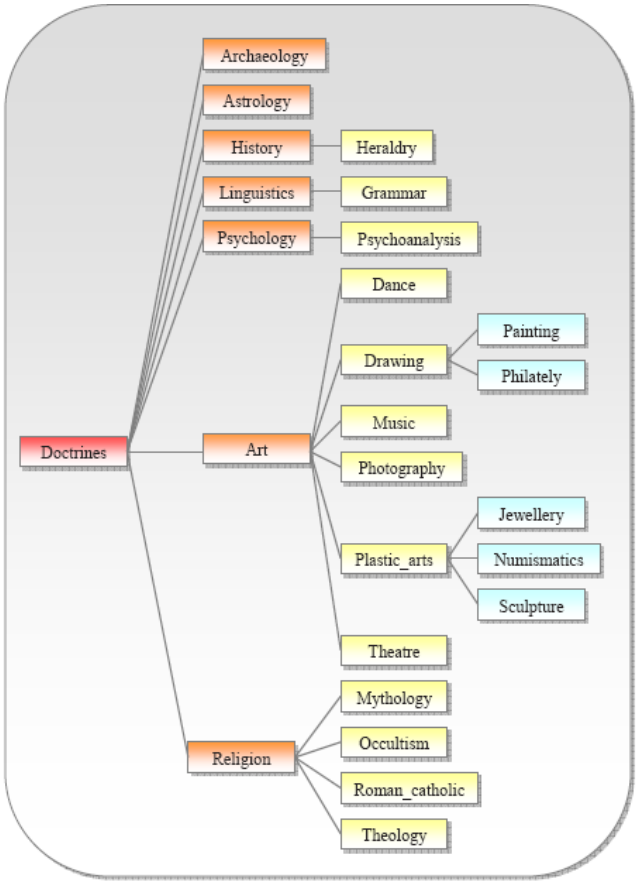


Figura 3.3: Uno de los cinco dominios principales de Wordnet Domains representados Jerárquicamente [78]

La representación Semántica de los dominios se organiza en 22 familias. La organización jerárquica es fija, mientras que la organización por familias se puede reorganizar permitiendo así la creación de nuevas relaciones interdisciplinarias.

3.2.2. WordNet Domains

WordNet Domains[19] es un recurso léxico creado de una manera semiautomática en el cual se aumentado etiquetas de dominio a WordNet, dichas etiquetas se denominan *"Subject Field Codes"* (SFC) con esto se

[19] WordNet Domains, http://wndomains.itc.it/

trata de agrupar conjuntos de palabras relevantes para un dominio específico.
En este recurso las synsets de WordNet han sido anotadas con varias etiquetas de dominio de entre un conjunto de 200 etiquetas organizadas jerárquicamente.

La creación de WordNet Domains fue motivada por las siguientes razones:

- ✓ Crear nuevas relaciones entre palabras: Mediante las etiquetas de dominio se puede establecer relaciones entre palabras que pertenecen a distintas categorías.

- ✓ Anotar a nivel semántico: Se tiene una anotación a nivel de conceptos no a nivel de palabras ya que la anotación de dominios se realiza a nivel de synsets.

- ✓ Obtener recursos multilingües: Los *SFC* son independientes del lenguaje, por lo que se pueden incluir en recursos multilingües tales como EuroWordNet.

El proyecto WordNet Domains es sus inicios tuvo algunos problemas, las etiquetas propuestas para la jerarquía no tuvieron un significado semántico significativo, por dicha razón se implementó un sistema de código para representación de la jerarquía de dominios cada código tiene una descripción y una relación especificada con las otras jerarquías [78].

3.2.2.1. Uso de Jerarquía de Dominios

La jerarquía de dominios ha tenido varios usos, entre ellos los siguientes [78]:

- ✓ EuroWordNet dominio-ontología, una jerarquía de dominios independientes del idioma para la cual los conceptos de interlingua (ILI-records) pueden ser asignados.

- ✓ Grandes jerarquías de dominio también están disponibles en Internet, principalmente destinado a la clasificación web

de los documentos, esto se puede ver por ejemplo en el directorio de Google y Yahoo.

- ✓ Dentro de la jerarquía de dominios WordNet Domains ha sido usado por la comunidad de PLN para desambiguar el sentido de las palabras en varios idiomas.

- ✓ Dado que la jerarquía de dominios es independiente del lenguaje la misma jerarquía es usada para construir corpus de referencia para distintos idiomas

- ✓ Dos de los proyectos que se pueden mencionar y que han usado el WordNet Domains son el MEANING[20] y BALKANET[21]

3.2.3. WordNet y PLN

WordNet es el recurso que se ha usado hasta el momento en varios proyectos ya que es bastante completo para poder realizar lo que es la desambiguación semántica, en él se busca el significado correcto para las palabras que en una fase anterior fueron procesadas por un analizador morfológico, obteniéndose así las palabras procesadas con su sentido correspondiente. Una manera de realizar esta desambiguación usando WordNet es primero adquirir desde WordNet los conjuntos de sinónimos de las palabras a desambiguar, luego determinar la coincidencia entre el contexto de las palabras a desambiguar y el contexto del conjuntos de sinónimos para finalmente escoger el significado correcto de las palabras en un texto. De esta manera WordNet es de gran ayuda para el nivel semántico del PLN.

A continuación se mencionan varios proyectos que han usado WordNet conjuntamente con el PLN:

[20] The free dictionary, http://idioms.thefreedictionary.com/

[21] A Multilingual Semantic Network for the Balkan, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.5679

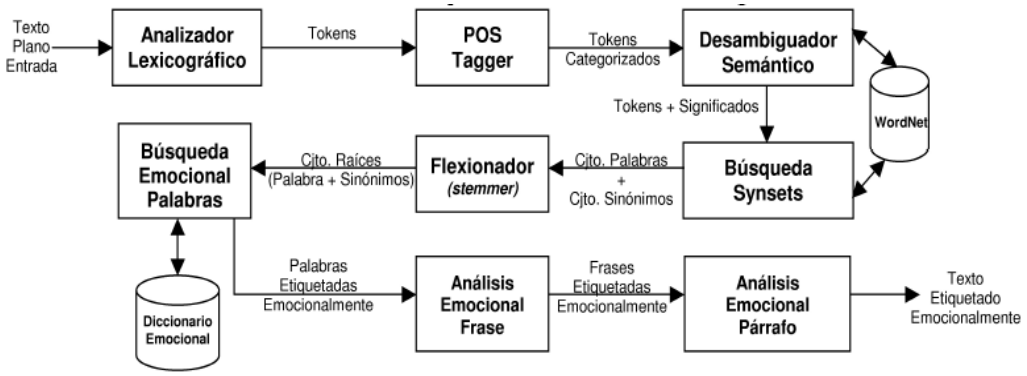


Figura 3.4: ***"Identificación de emociones a partir de texto usando desambiguación semántica"*** [45]

En la Figura 3.4 se muestra el ejemplo de un sistema que usa WordNet para la desambiguación semántica, las palabras que pasan por el POS son luego pasadas a la desambiguación semántica, en dicho paso usan WordNet para asignarles un significado correspondiente de acuerdo a los tokens recibidos. Por ejemplo a la palabra ratón se le puede asignar concepto de dispositivo electrónico o animal, con la ayuda de WordNet se realizará una asignación correcta de significado. El algoritmo aplicado para la desambiguación semántica usado en el sistema fue una modificación del algoritmo 54Aximum Related-ness Disambiguation propuesto por Pedersen, Banejee, y Patwardhan (2005), con una ventana de desambiguación que abarca toda la frase [45].

Otra investigación del uso del PLN en conjunto con WordNet es ***"Uso del recurso lingüístico WordNet en la expansión de consultas con un modelo del usuario de recuperación de información"*** [51], en dicho proyecto se usa wordNet para conseguir un conjunto de sinónimos que se usan para expandir la consulta del usuario además se utiliza la información del uso de dichos sinónimos en diferentes contextos para escoger los sinónimos adecuados para la consulta de tal manera que se va expandiendo las consultas de acuerdo al contexto que el usuario poco a poco va especificando.

La investigación titulada ***"Utilizando WordNet para Complementar la Información de Entrenamiento en la Identificación del Significado de las Palabras"*** [52] trata de combinar un recurso de colección de entrenamiento (SemCor) con el uso de una base léxica (WorNet) para

mejorar la efectividad del proceso de desambiguación del sentido de las palabra (WSD), la forma de lograr esto se lo representa en la Figura 3.5. Una colección de entrenamiento es un conjunto de documentos con los significados etiquetados manualmente, que permite al sistema asignar los significados a nuevos documentos, de acuerdo con su similitud a otros documentos de la colección de entrenamiento.

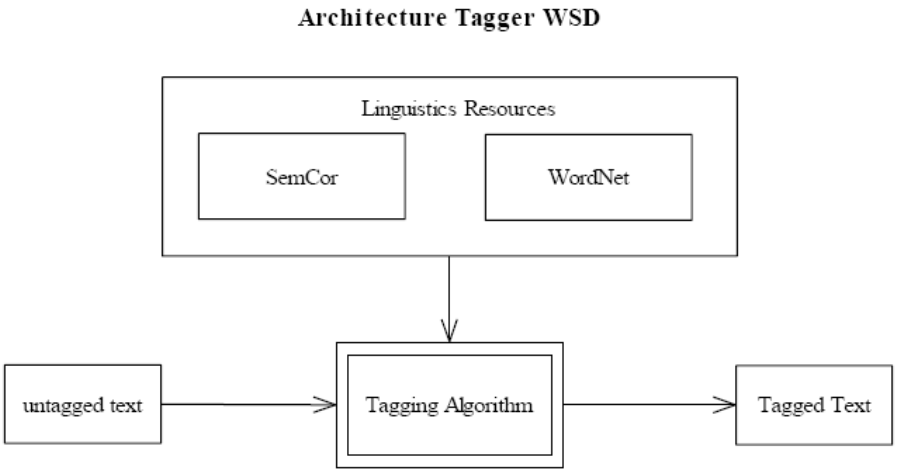


Figura 3.5: Modelo WSD usado para ***"Utilizando WordNet para Complementar la Información de Entrenamiento en la Identificación del Significado de las Palabras"*** [52]

3.2.4. Aplicaciones WordNet

Con el pasar de los años se ha visto la construcción de wordNets para diferentes idiomas ya que estos han sido usados para varias aplicaciones, entre dichas aplicaciones se tienen: Word Sense Desambiguation (WSD), Information Retrieval (IR), Question Answering (QA), Information Extraction (IE), Resolución de Anáfora entre otras [80].

3.3.EuroWordNet

Este proyecto se inició en 1996, el objetivo de este trabajo fue crear una base de datos léxico-semántica, la misma que se puede considerar como un recurso multilingüe con wordnets de algunos lenguajes europeos, cada wordNet

representa un único lenguaje, dichos wordnets están relacionados a través de un Inter-Lingual-Index, por medio del índice los lenguajes son interconectados siendo posible ir de una palabra de un lenguaje a su correspondiente palabra en otros lenguajes.

3.4. BalkaNet

BalkaNet[22] es un proyecto empezó en Septiembre del 2001 y terminó en Agosto del 2004 contempla lenguajes como: Bulgarian, Greek, Romanian, Serbian, Turkish and to extend the Czech wordnet desarrollado previamente en EuroWordNet. El objetivo principal de este proyecto fue alinear wordnets para los 6 lenguajes antes mencionados y demostrar su utilidad en aplicaciones modernas reales. La construcción de wordnets para los lenguajes que abarca BalkaNet fue apoyado por muchos recursos monolingües y bilingües. Un ejemplo de la aplicación de este recurso es "Utilizing BalkaNet's Conceptual Taxonomies to IndexWeb Documents", aquí los dominios conceptuales ILI son usados como clusters dentro de los cuales pueden ser clasificados los documentos de la web, para realizar dicha clasificación se realiza un mapeo entre los términos de los documentos y los conceptos de ILI calculando luego la similaridad semántica. Balkanet usa los dominios de SUMO lo cual permite que este recurso pueda ser fácilmente extendible para otros idiomas y también permite que cualquier aplicación pueda utilizar el mismo sin requerir hacer modificación alguna [80].

3.5. MultiWordNet

El proyecto MultiWordNet[23] se inició con el propósito de crear un WordNet para Italiano alineado con el WordNet de la Universidad de Pricenton. MultiWorNet es una base de datos lexica multilingue, incluye información acerca de palabras en Italiano e Inglés, además incluye información léxica como: relación léxica (sinonimia) entre palabras, relaciones semánticas (has_hypernym, has_hyponym, has_part) entre conceptos léxicos, correspondencia entre conceptos en Inglés e Italiano y existencia de dominios.

[22] BalkaNet, http://is.dblab.upatras.gr
[23] MultiWordNet, http://multiwordnet.fbk.eu/english/home.php

Este recurso fue construido tomando como referencia las relaciones existentes en el WordNet, de esta forma el nuevo wordNet para Italiano fue construido con alineación al wordNet de Princeton.

El modelo de construcción de este recurso consta de dos procedimientos automáticos el primero se llama *Assign-procedure*, en el cual teniendo en cuenta un sentido de la palabra italiana, el procedimiento selecciona una lista ponderada de los synsets correspondientes más probables del wordNet de Princeton. Esta lista es utilizada por los lexicógrafos para construir los synsets para el wordNet en litaliano. El segundo procedimiento apoya la detección de *Lexical Gaps* (LG-procedimiento), que son los casos en que un concepto léxico de una lengua se expresa a través de una combinación libre de palabras en otro idioma [79].

3.5.1. Modelo de Datos de MultiWordNet

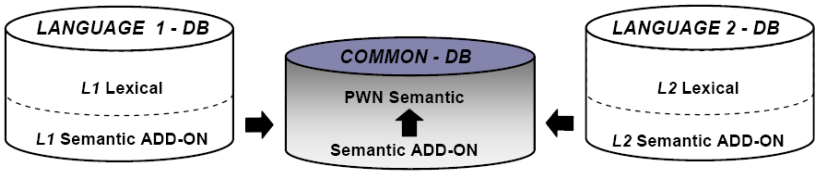


Figura 3.6: Modelo de Datos MultiWordNet [79]

En la Figura 3.6 se muestra el modelo de datos usado para multiWordNet, el cual tiene 2 módulos básicos, el COMMON-DB y LANGUAJE-DB, en el primero están almacenadas las relaciones semánticas comunes a los diversos lenguajes mientras que en LANGUAJE DB están almacenadas las relaciones léxicas específicas de cada lenguaje por ejemplo has-synonym. Dicho de otra manera la información acerca de qué palabras pertenecen a cada synsets están en el módulo LANGUAGE-DB y la información sobre las relaciones entre synsets, comunes a todos los lenguajes, están en COMMON-DB.

3.6.EuroWordNet vs MultiWordNet

La diferencia básica entre estos dos recursos multilingües es el modelo de construcción que se ha usado para cada uno de ellos, el proyecto EuroWordNet consistió en construir wordNets específicos de cada lenguaje independientes

uno del otro, luego se estableció la correspondencia entre dichos wordNets de diferentes idiomas, por el contrario MultiWordNet se construyó usando como base las relaciones presente en el wordNet de Princeton es decir en la construcción de dicho recurso no hubo independencia entre los wordNets [79].

El modelo de construcción de MultiWordNet es menos complejo y garantiza el más alto grado de compatibilidad entre diferentes wordnets, la desventaja de este modelo es que se puede forzar a una dependencia excesiva del léxico y estructura conceptual de una de lenguas implicadas, lo cual traería como consecuencia demasiada atención a un solo lenguaje, por esta razón sería mejor el modelo usado para EuroWordNet.

Otra ventaja importante del modelo de construcción de MultiWordNet es que se usa procedimientos automáticos que fueron mencionados anteriormente, lo cual acelera tanto la construcción de synsets correspondientes y la detección de las divergencias entre el wordNet de Princeton y el WordNet en construcción.

EuroWordNet abarca más lenguajes que MultiWordNet.

3.7. Acoplamiento de Recursos Semánticos

La construcción de diversos recursos semánticos ha sido una tarea bastante compleja, con una duración de muchos años y colaboración de varios grupos de investigación. El uso de estos recursos se ha convertido en una práctica frecuente para los sistemas actuales de PLN. Los recursos se usan por separado lo cual, no permite aprovechar al máximo a los mismos, por esto es necesario buscar como se complementan estos entre sí. Entre los recursos más usados se pueden mencionar a WordNet, SemCor y Topic Signatures para cada synset adquiridas de la web.

Con el propósito de tener compatibilidad de los recursos mencionados se ha creado un repositorios que los agrupe, este es Multilingual Central Repository (MCR) el cual sigue el mismo principio de EuroWordnet que tiene una agrupación de WordNets en diferentes idiomas con lo cual se ha podido aprovechar mejor el conjunto de wordnets, obteniendo así un recurso multilingüe de gran tamaño útil para un gran número de procesos semánticos

que necesitan de una gran cantidad de conocimiento multilingüe. El MCR también integra WordNet Domains, nuevas versiones de los Base Concepts y la Top Concept Ontology y la ontología SUMO. La versión que se tiene hasta el 2008 contiene 934.771 relaciones semánticas entre synsets. MCR representa un volumen casi cuatro veces más grande que el de Princeton WordNet.

Tener en conjunto las bases de conocimiento más grandes y probadas nos permite realizar WSD con mayor facilidad sin la necesidad de aplicar algoritmos tan complejos.

3.8. Integración de herramientas del PLN

Actualmente existen varios recursos que se han utilizado en el PLN, todos ellos han sido producto de varios años de estudio por lo cual no es muy adecuado seguir creando más recurso, si se puede reutilizar los que existen, de esta forma se estaría logrando más investigación en el ámbito del PLN. Pero integrar dichos recursos no es una tarea fácil aunque existe un proyecto que trabajó en una plataforma de integración de recursos de PLN InTiMe[46] (INtegration of Tools and corpora In the text- MEss project), el proyecto fue desarrollado por el Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante para más información sobre este proyecto se puede revisar el ANEXO 2.

A parte de tener una plataforma que integre recursos que sean más accesibles también es necesario reutilizar dichos recursos agrupándolos para crear recursos consistentes y completos de tal manera que se reutilice conocimiento lingüístico formalizado lo que garantizará obtener mejores recursos lingüísticos, un ejemplo de esto es Leff (*Léxico de formas flexionadas del Español*), el cual es un léxico morfológico y sintáctico, en este recurso se ha reutilizado Multext, el léxico de la USC, SRG y ADESSE los dos primeros contienen solamente información morfológica, mientras que los restantes incluyen información sintáctica, la forma de reutilizar los recursos mencionados fue construir un léxico morfológico inicia tomando como base Multext y transformándolo en el formato Alexina (es un modelo que permite describir información morfológica y sintáctica de manera legible, completa y eficiente) luego se convirtió el léxico de la USC al formato de Alexina y se lo fusionó con el léxico inicial extraído de Multext, obteniendo así el léxico que contiene la información morfológica de L*eff*e, seguidamente se convirtió la información sintáctica de ADESSE y del léxico SRGen al formato de Alexina para finalmente fusionar el L*eff*e morfológico con los léxicos sintácticos obtenidos [47].

3.9. Discusión

Luego de haber realizado la investigación de varios recursos lingüísticos puedo decir que lo mejor no es construir recursos desde cero sino aprovechar los ya existentes, para así poder obtener recursos que cubran un mayor número de lenguajes y poder usarlos en tareas del PLN, además al momento de agrupar dichos recursos se deben tener presente que las anotaciones lingüísticas que se usan deben ser estandarizadas para no tener problemas cuando se haga el uso de los nuevos recursos lingüísticos.

Los recursos más adecuados para propósitos de esta tesis es EuroWorNet y BalKaNet, pero en este capítulo se describe más a detalle a WordNet ya que EuroWorNet y BalKanet es un conjunto de WorNets.

Gracias a que se realiza una revisión de MultiWordNet se puede saber el modelo de construcción de este recurso y así se puede tomar esto como referencia para poder alinear EuroWordNet y BalKaNet.

CAPÍTULO 4

4. MODELO

El modelo que se presenta a continuación tiene dos partes relevantes: lo que es el PLN y una técnica de Representación Universal del Lenguaje. Lo novedoso de este modelo es la reutilización de herramientas del PLN y agrupación de recursos multilingües ya existentes, para extraer características universales de una frase y luego representarlas en un lenguaje universal de tal manera que la conversión sea posible a cualquier lenguaje sin perder información. Este modelo está planteada solo para el procesamiento de frases simples, esto implica que no procesa párrafos completos, ni textos completos; esta limitante es debido a que en un futuro se pretende que dicho modelo sea útil para el área de Recuperación de Información en la web (las búsquedas que se realizan por lo general son mediante textos cortos).

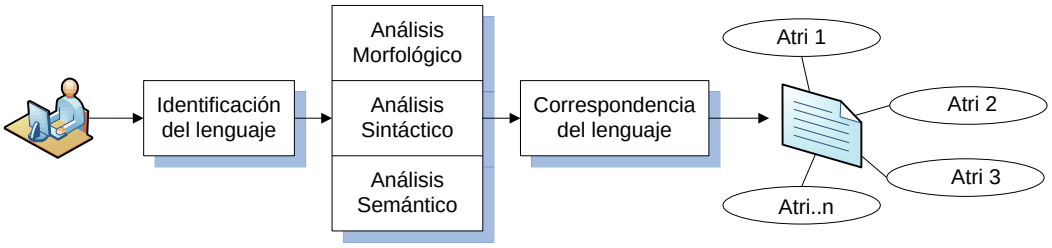


Figura 4.1: Representación General del Modelo Multilingüe

En la Figura 4.2 se muestra el modelo a detalle, la descripción de la misma se hará más adelante.

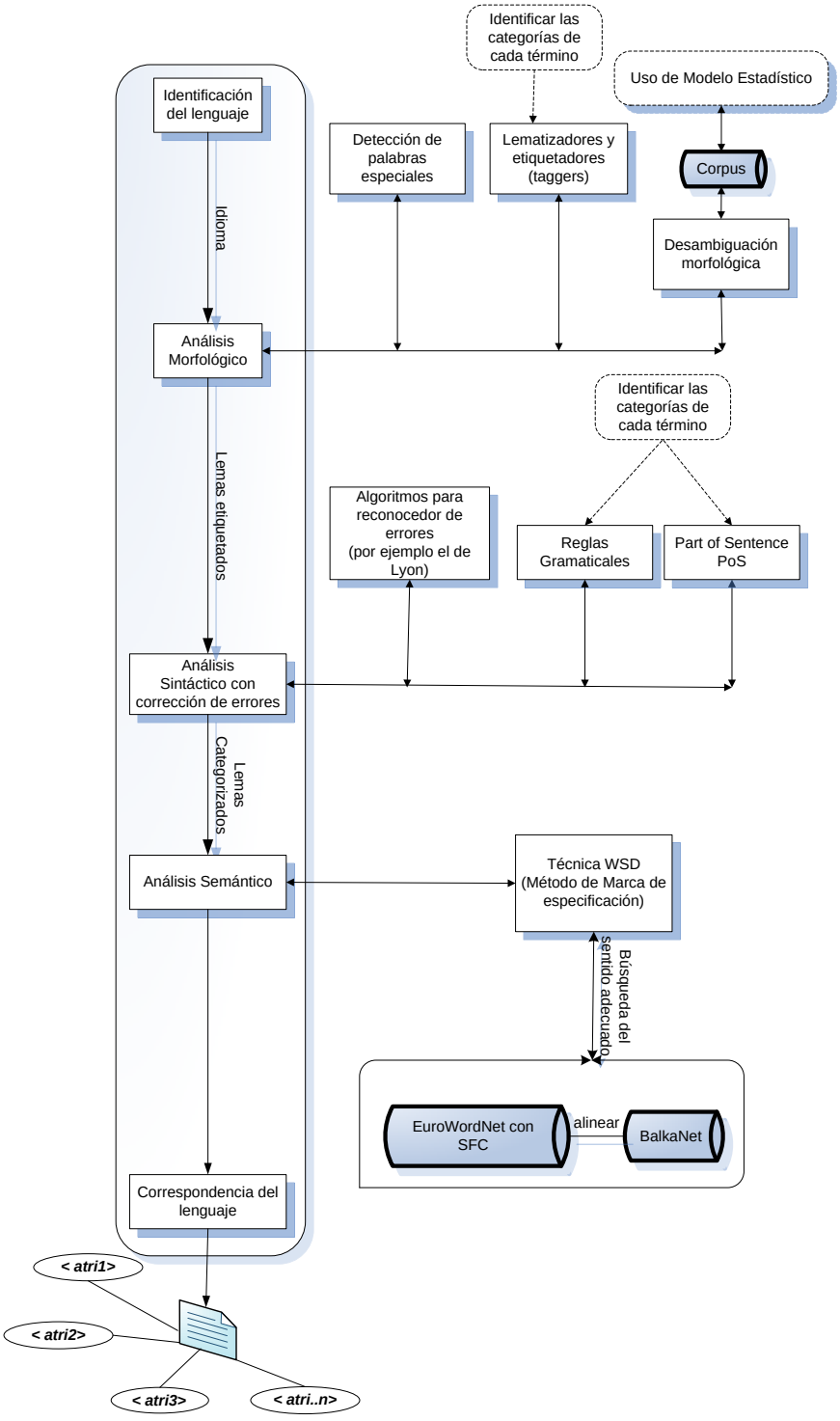


Figura 4.2: Modelo

4.1. Identificación del Lenguaje

Este módulo tiene como principal objetivo deducir en que lenguaje está escrita la palabra o frase ingresada, mediante diccionarios, ontologías o haciendo un estudio de la morfología de la palabra. Este dato lo necesitamos para poder realizar el PLN adecuado ya que dependiendo del idioma este puede tener ciertas variantes. En la actualidad existen herramientas para poder realizar la identificación del idioma en el que está escrito un texto por ejemplo TEXTCAT [27], hay una versión disponible que no es comercial y sirve para identificación de 76 lenguajes.

4.2. Análisis Morfológico

Luego de haber identificado el lenguaje se realiza el análisis morfológico de la frase ingresada al inicio, este proceso conlleva a varias actividades que se citan a continuación [28][29][30]:

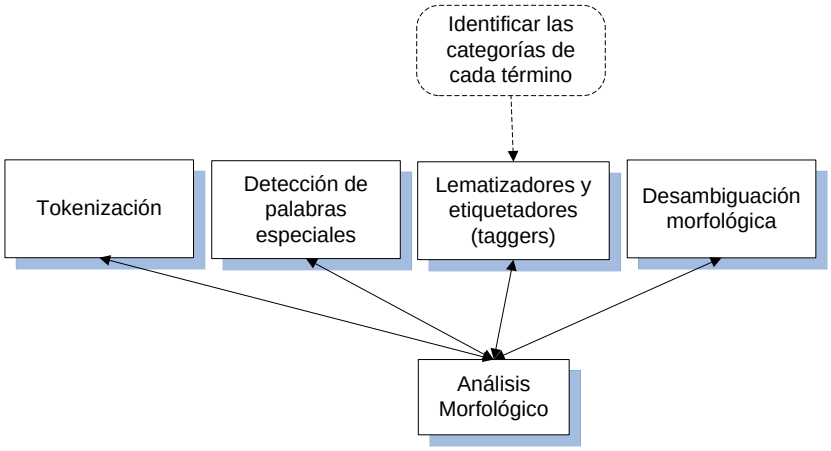


Figura 4.3: Análisis Morfológico

- **Tokenización:** El objetivo de este proceso es separar la frase tratable para procesos posteriores. Se debe realizar un proceso de tokenización a la frase de entrada, este resultado será la entrada del proceso de etiquetación.
- **Detección de Palabras Especiales**: Se debe hacer un reconocimiento de las entidades con nombre por ejemplo los nombres de los países de las ciudades

etc. Todo lo que tenga un nombre propio. Por ejemplo María llega hoy a Ecuador, las palabras especiales serían María y Ecuador. Esto se lo puede realizar usando una lista de palabras especiales como lo propone el modulo de GATE o se puede tener un conjunto de reglas para identificar dichas palabras especiales. Este proceso de identificación es necesario porque las palabras especiales no pueden ser tratadas en el proceso anterior.

- **Lematizadores y Etiquetadores**: Luego de haber realizado la *"Detección de palabras especiales"* y la tokenización se extraen los morfemas (unidad más pequeña a la que se le puede asignar significado por ejemplo *de, no, yo, le, el*) y lexemas (unidad mínima con significado léxico que no presenta morfemas gramaticales por ejemplo *sol*) de las palabras , los morfemas adaptan el lexema al contexto de uso, añadiendo matices de significado tales como: si la palabra es un *verbo, sustantivo, adjetivo, etc* o marcando relaciones gramaticales con el resto de términos de una oración de tal forma que se identifican propiedades de inflexión como: género (masculino, femenino, neutro), número (plural o singular), caso (nominativo, acusativo, etc.) [31]. El resultado de esta actividad es el *"Etiquetado de palabras"*, esta metodología hace uso de modelos estadísticos que toman en cuenta la probabilidad de que una palabra pueda ser etiquetada como una de las posibles formas dentro del lenguaje nativo, esto ayuda para que en base a probabilidades se pueda determinar cual es la mejor etiqueta (por ejemplo sustantivo o adjetivo) que se puede asignar, según análisis exhaustivos que se han realizado sobre la precisión de dicha metodología es de un 90%; para mejorar esta precisión se plantea básicamente que se tenga en cuenta la probabilidad que hay de etiquetar una palabra de un determinado modo si la que la precede tiene otra etiqueta dada, por ejemplo lo más normal es que después de un determinante se encuentre un sustantivo o un adjetivo, siendo la probabilidad de haber un sustantivo mucho mayor, este tipo de comprobaciones se llevan a cabo mediante el uso de *"Modelos Ocultos de Markov"* (HMM) (Brants, 2001).

  Existe otra técnica que supera un poco a la ya mencionada, es la de Transformaciones Progresivas, se basa en tener reglas generadas en base a un contexto de entrenamiento para luego en base a dichas reglas realizar el

etiquetado, su principal diferencia con los Modelos Ocultos de Markov es la rapidez con la que cuenta para etiquetar las palabras.

Una técnica que también es considerada para asignar etiquetas es el "*Etiquetado con listas de transformaciones"* (Brill, 1995) [41], una lista de transformaciones es una lista de reglas de la forma *Si-entonces* dichas reglas se las obtiene en base a un entrenamiento con texto anotado, esto funciona aplicando a cada palabra las reglas, obteniendo como resultado etiquetas apropiadas. Por ejemplo:

Etiquetado Inicial:

Todos-NN los-NN niños-NN y-NN las-NN niñas-NN comen-NN

Regla: If the word is currently tagged as NN and it ends with an s, then retag it as NNS.

Siguiente Etiquetado:

Todos-NN los-NN niños-NNS y-NN las-NN niñas-NNS comen-NN

El etiquetado con listas de transformaciones tiene varias ventajas y desventajas entre ellas: es más expresivo que las listas de decisión y los arboles de decisión, una regla puede deshacer lo que ha hecho otra regla anterior en un caso particular, el aprendizaje es muy lento, pues hay que evaluar a cada paso muchas posibilidades, no da varias posibles etiquetaciones en casos dudosos en conclusión tiene una precisión 95-96%.

Otras técnicas que se pueden usar en el etiquetado son: Aprendizaje basado en memoria (k nearestneighbor), Redes neuronales, Bootstrapping, Combinación de varios métodos y EngCG: Reglas definidas a mano por expertos.

Las etiquetas más usadas son las de PoS del Brown Corpus y las del Penn Treebank. En la Figura 4.4 se muestra un ejemplo de dichas etiquetas:

Part-of-speech	Morphological variation	tag
noun	Singular	NN
	Plural	NNS
	Proper, singular	NNP
	Proper, plural	NNPS
adjective	Normal	JJ
	Comparative	JJR
	Superlative	JJS
verb	Base	VB
	Non-3rd, present tense	VBP
	3rd person, present	VBZ
	Past tense	VBD
	Past participle	VBN
	Gerund	VBG

Figura 4.4: Etiquetado PoS [41]

El conjunto de etiquetas que se usa debe estar bajo un estándar por ejemplo el estándar europeo ya mencionado EAGLES[24].

- **Desambiguación Morfológica**: En el análisis morfológico ya empieza a presentarse cierto nivel de ambigüedad así pueden existir diversas etiquetas para una sola palabra por cuanto se necesita desambiguar a nivel morfológico esto haciendo uso de cualquiera de las técnicas anteriormente mencionadas. En una de las técnicas anteriores es necesario tener un conjunto de reglas dicho conjunto puede ser obtenido de un corpus, dependiendo del corpus se puede tener menor o mayor precisión de las mismas, al estar usando un corpus se está usando un modelo estadístico en el nivel morfológico del PLN.[25]

[24] Descripción de EAGLES en la sección 3.1.3.1

[25] Es posible obtener un conjunto de reglas del corpus ya que la constitución básica de este son ejemplos de uso del lenguaje en el cual está el corpus.

4.3. Análisis Sintáctico con Corrección de Errores

A este nivel se identifica la estructura interna de una oración, dicha estructura a su vez se desglosa en sintagmas, los cuales vienen constituidos por sujeto, predicado, objetos directo-indirecto, etc. Aquí se establece la relación entre las categorías de los lemas identificados y etiquetados en la actividad anterior (Ver Figura 4.5) [32]. Para llevar a cabo este nivel del PLN es necesario:

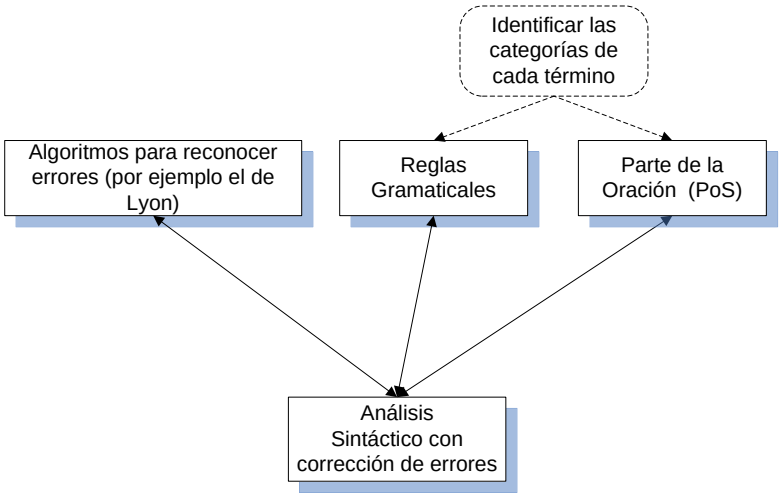


Figura 4.5: Análisis Sintáctico

- **Algoritmo para Reconocer Errores**: Lo primero que se toma en consideración para este análisis es la corrección de errores, muchas veces las frases no corresponde a la gramática del lenguaje por ello es necesario aplicar un algoritmo para reconocer los errores y así poder continuar con el análisis de la oración uno de estos algoritmos es el de Lyon.

- **Reglas Gramaticales**: Esto hace referencia a las gramáticas y a las reglas sintácticas de cada lenguaje, por ejemplo se hace el uso de lo que son grupos nominales (el gato, pescado), verbales(come), para definir la estructura de una oración, por ejemplo en el lenguaje español que se define una estructura como: Sujeto Predicado y dentro de ello muchas otras estructuras como objeto, especificadores(el) etc.

- **"Parte de la Oración" (PoS)**: Se identifica el rol de cada término dentro de una frase haciendo uso de las reglas gramaticales.

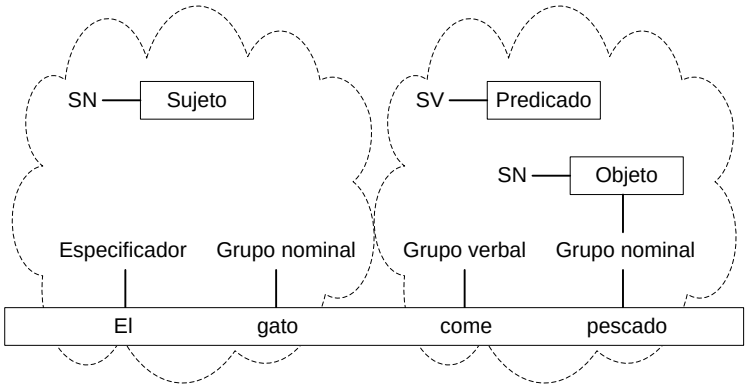


Figura 4.6: Resultado del Análisis Sintáctico

4.4. Análisis Semántico

Para realizar el análisis semántico se puede hacer uso de los recursos lingüísticos por ejemplo WordNet, pero como el objetivo de este modelo es el ámbito multilingüe se plantea usar EuroWordNet y BalkaNet. Se recibe el resultado de los análisis anteriores y se aplica WSD (Word Sense Desambiguation o desambiguación del sentido de las palabras) usando la red semántica que se encuentra en cada wordnet del EuroWordNet y BalkaNet. Aquí se puede hacer referencia a otras investigaciones que ya se han realizado por ejemplo una interfaz desarrollada para desambiguar el sentido de las palabras usando Wordnet y el Método de especificación de Marcas (Specification Marks Method, el cual está dentro de la clasificación de Método de Similaridad semántica mencionada anteriormente) [34] este método se basa en el emparejamiento del contexto de la palabra a ser desambiguada con información de un recurso de conocimiento léxico externo WordNet, este algoritmo utiliza la taxonomía de nombres de Word-Net, sus relaciones de hiponimia e hiperonimia, para desambiguar palabras dentro de un contexto local (oración). La hipótesis en la que se basa este algoritmo es que las palabras que aparecen en un mismo contexto tienen sus sentidos relacionados entre sí, por tanto, puede existir un concepto dentro de la red semántica que relacione todas las palabras del contexto. Este concepto superior es la denominada *Marca de Especifidad* (ME). El proceso de la realización de desambiguación aplicando este algoritmo es el siguiente: a través de la jerarquía de WordNet y de las palabras del contexto, se obtiene el conjunto de hiperónimos/hipónimos que comparten las palabras, usando dicha información se

trata de determinar el concepto superior (ME) que engloba el mayor número de palabras del contexto con sus respectivos sentidos. Si como resultado para la ME inicial aún existen palabras ambiguas en el contexto, se va descendiendo por la jerarquía obteniendo nuevas ME, de forma que se seleccionará aquella ME que maximice el número de palabras del contexto no ambiguas. En la Figura 4.7

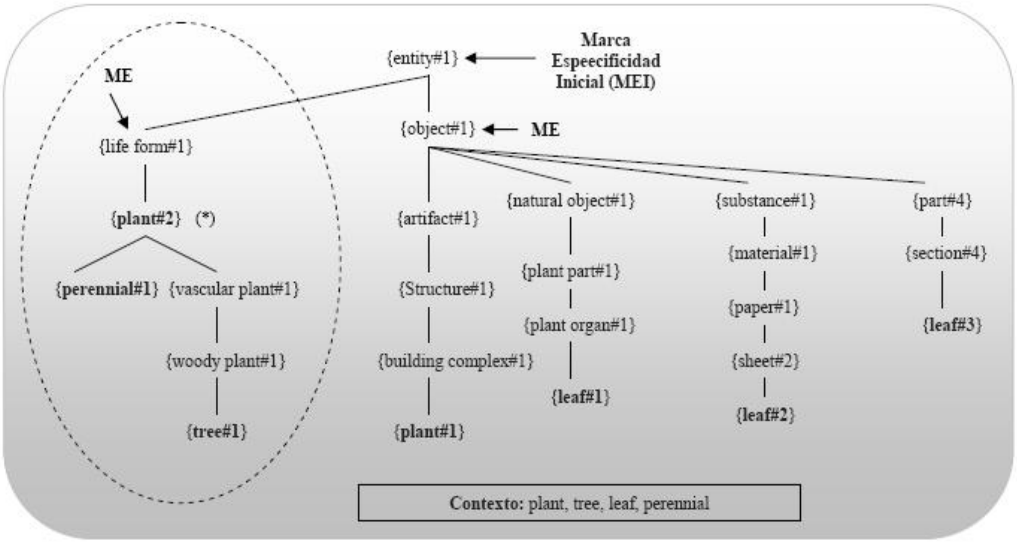


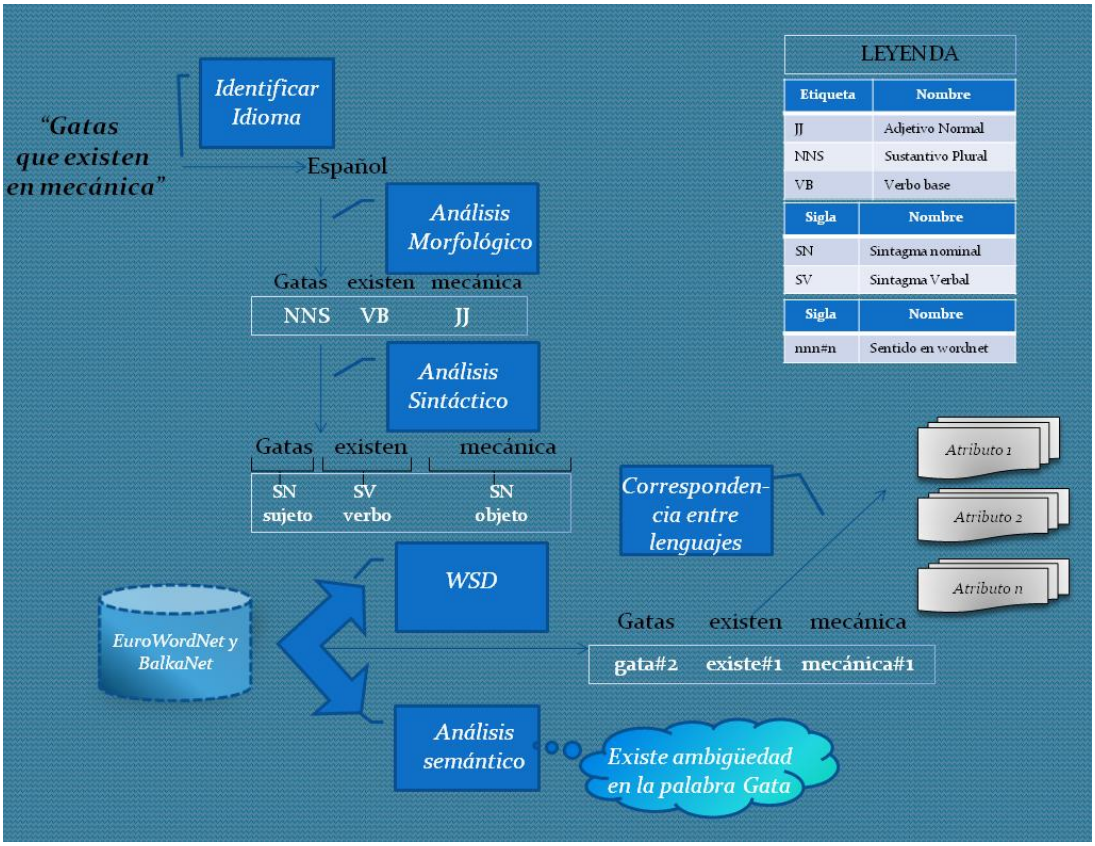
Figura 4.7: Funcionamiento del Specification Marks Method [76]

4.5. Correspondencia del Lenguaje

En este módulo se busca la correspondencia de las palabras al idioma inglés de tal forma que después dichas palabras se las convierte en palabras universales obteniendo así una representación que es muy fácil para pasarla a diversas lenguas.

El objetivo principal de alinear Balkanet y Eurowordnet es abarcar más lenguajes, EurowordNet considera 7 lenguajes y Balkanet 6. La alineación de este recurso es posible ya que ambas usan la misma jerarquía de dominios y la intelingua ILI.

4.6.Ejemplo



Los procesos que se deben llevar a cabo se muestran en los cuadros azules, cada proceso debe usar algoritmos mencionados en la descripción anterior del modelo, teniendo así un resultado de palabras con las características lingüísticas apropiadas, pasando luego a representarse de una forma universal, con dicha representación es fácil pasar la frase a cualquier lenguaje.

CAPÍTULO 5

5. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

5.1.Conclusiones

- Existen varios modelos para el PLN los cuales pueden ser usados de una manera conjunta como se propone en el modelo presentado en esta tesis.
- Al usar un sistema de WSD en el modelo se obtiene una correspondencia idónea entre palabras de diversos lenguajes, solucionando así el problema básico de la traducción, como lo es la pérdida de información.
- Actualmente existen varios recursos lingüísticos producto de varias investigaciones, lo necesario es agrupar dichos recursos para poder obtener mayores beneficios.
- Para realizar la desambiguación a nivel semántico existe la técnica de WSD con varios métodos y algoritmos por lo cual es necesario escoger los apropiados.
- Para alinear EuroWordNet y Balkanet se necesitan hacer el estudio de alineación de WordNets, se puede usar el modelo de MultiWordNet.
- El PLN está presente en el tratamiento del texto y en información almacenada en los recursos lingüísticos por ejemplo en WordNet se tienen etiquetas para dichas etiquetas se necesitan algoritmos pertenecientes a PLN.

5.2. Recomendaciones

- Para la implementación del modelo se recomienda considerar un etiquetado estándar tanto para el PLN como para los recursos lingüísticos que se utilicen.
- Para implementar el modelo se recomienda realizar un estudio de las herramientas adecuadas que se puedan acoplar de tal manera que haya reutilización logrando así optimizar tiempo y recursos.
- El corpus para la desambiguación morfológica debe ser elegido de acuerdo a un dominio de aplicación, esto con la finalidad de mejorar los resultados de los proyectos inherentes.
- Si se va a dar continuidad al proyecto, mismo que representa la base para recursos multilingües, se recomienda adquirir la licencia de EuroWordNet.
- Para realizar este tipo de investigaciones se recomienda contar con un equipo de ingenieros en sistemas y lingüistas, considerando también un tiempo mayor a 3 años para su desarrollo.

5.3.Trabajos Futuros

- Adaptar el modelo para utilizarla en la construcción de un buscador multilingüe en un área específica
- Adaptar el modelo propuesto para hacer posible su aplicación en la conversación de dos personas en el internet sin importar su idioma.
- Realizar una estandarización en cuanto al etiquetado a usar, para la implementación del modelo.
- Estudio de los diferentes métodos para WSD de tal manera que se elija el mejor, cuando se vaya a implementar el modelo propuesto.
- Estudio comparativo de EuroWordNet y BalkaNet.
- Reutilización de recursos para crear nuevos recursos léxicos y sintácticos para el idioma español según un dominio o de una forma generalizada de tal manera que se puedan aprovechar los recursos ya creados y no gastar esfuerzo innecesariamente
- Uso de WordNet para la extensión consultas en dominios específicos.
- Investigación sobre la extensión de consultas a nivel multilingüe usando EuroWordNet y PLN.

6.ANEXOS

Anexo 1

Herramientas para PLN

- MPRO[40], programa para el análisis morfológico y sintáctico de textos en español, los resultados de este programa pueden ser usados para diferentes fines entre ellas traducción automática, indexación y elaboración de corpora. MPRO consiste en una serie de subprogramas, diccionarios, léxicos y gramáticas que interactúan entre sí. Los componentes de esta herramienta son LESEN y PARSER los mismos que pueden ser adaptados a las necesidades del usuario y al tipo de textos con los que se desee trabajar, sean estos textos especializados o generales. LESEN abarca todo lo que es el análisis morfológico mientras que PARSER cubre lo que es el análisis sintáctico

- TreeTagger[48] para obtener el POS y el lema de las palabras. Se ha utilizado con éxito para taggear textos en alemán, inglés, francés, italiano, español, griego, y francés antiguo, además esta herramienta es fácilmente adaptable a otros lenguajes si se dispone de un lexicón y corpus marcado manualmente. Usa un árbol de decisión para poder asignar una cierta categoría a una palabra y para estimar la probabilidad que existe de que una palabra pertenezca a una u otra categoría. La probabilidad de que una palabra sea asignada a un a categoría se la determina haciendo varias pruebas desde los niveles altos de nodos a las hojas del árbol de decisión [49]. Está disponible para sistemas Windows, Linux y Mac.

- Lingpipe es un Kit de herramientas libre, desarrollado por Alias-I. Sirve para procesar texto usando lingüística computacional, es usado en tareas como: reconocimiento de entidades (personas, organizaciones y locaciones), sugerencia en ortografía correcta para las consultas entre otras actividades.

Existen tutoriales que sirven de guía para saber utilizar la herramienta en las tareas mencionadas, por ejemplo se tienen tutoriales para: Topic Classification, Named Entity Recognition, Clustering, Part-of-Speech Tagging, Sentence Detection, Language Identification, Word Sense Disambiguation etc [50].

- Porter Stemmer [53] es un algoritmo que está implementado en diversos lenguajes de programación, este fue originalmente descrito en Porter 1980, ha sido usado y adaptado en los últimos 20 años, sirve para quitar las inflexiones de las terminaciones de las palabras en inglés. Su aplicación principal es como parte de un proceso de normalización de términos que es muy usado en los sistemas de recuperación, existen implementaciones en línea a continuación se muestra un ejemplo de uso online de este algoritmo:

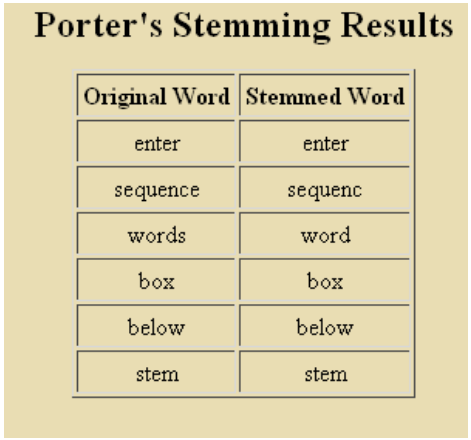


Figura 6.1: Resultados de haber usado la implementación online del Porter Stemmer[54]

- GATE [44] es una herramienta open source, diseñada por la Universidad de Sheffield, se la considera como una arquitectura general para la ingeniería del texto, es capaz de resolver muchos de los problemas que surgen al procesar texto, tiene una gran comunidad de desarrolladores para procesamiento de lenguaje, el análisis morfosintáctico de esta herramienta utiliza los procesos Sentence Splitter, Tokenizer y POS tagger. Un ejemplo de uso de esta herramienta se puede observar en la Figura 2.2, en dicha arquitectura GATE se usa para identificar tokens cuya categoría morfosintáctica es nombre (común, propio o plural) y luego de identificarlos se hace uso de wordNet para descartar aquellos nombres que no son específicos del dominio biomédico.

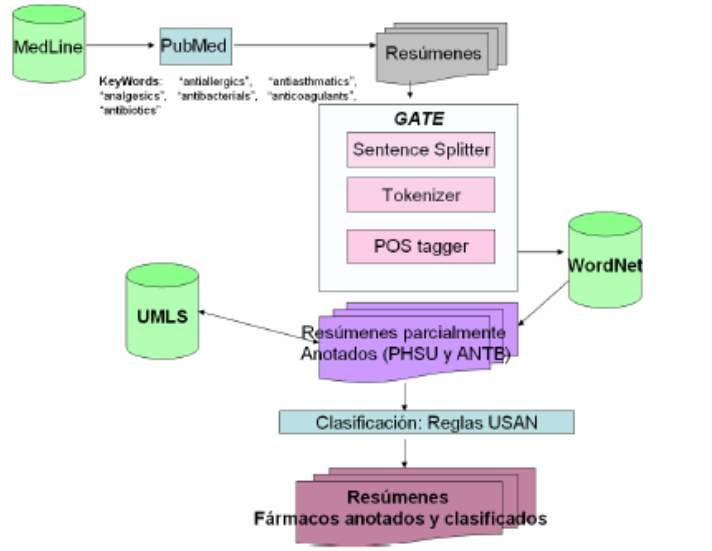


Figura 6.2: Arquitectura para la Detección de fármacos genéricos en textos biomédicos [45]

En el congreso internacional sobre Web Semántica Osaka 2005, la presencia de aplicaciones centradas en PLN para la web semántica fueron las que predominaron, las mismas que usaron en su mayoría GATE [55].

- MACO [61] es un analizador morfológico en 1998, desarrollado en la Universidad de Barcelona y la Universidad Politécnica de Cataluña. El sistema tokeniza el texto y devuelve, para cada token, todos los posibles pares POS-lema. Ofrece la posibilidad de reconocer fechas, números, nombre propios, signos de puntuación, abreviaturas, términos compuestos, algunos de los módulos de este analizador aún requieren mejora. Su plataforma es UNIX/LINUX. En la Figura 2.3 se muestra la salida que produce el analizador morfológico después de procesar un texto de entrada.

Análisis morfosintáctico de cada palabra

```
Recursos
        recursos NP00000
        recurso NCMP000
educativos
        educativo AQ0MP00
para
        para SPS00
        para NCMS000
        parar VMMP2S0
        parar VMIP3S0
        parir VMSP1S0
        parir VMSP3S0
        parir VMMP3S0
enseñanza
        enseñanza NCFS000
media
        media NCFS000
        mediar VMMP2S0
        mediar VMIP3S0
        medio NCFS000
        medio AQ0FS00
. . Fp
```

Figura 6.3: Resultado del analizador MACO al procesar la frase: Recursos educativos para enseñanza media

- RELAX [61] es un POS etiquetador, fue desarrollado por la Universidad Politécnica de Cataluña, este realiza la desambiguación morfosintáctica sobre la salida de MACO asignando a cada una de las palabras una sola etiqueta y su lema correspondiente entre todas las etiquetas. La precisión actual sobre textos en castellano es superior al 97%. Su plataforma es UNIX/LINUX. En la Figura 2.4 se puede observar el resultado de la desambiguación morfológica de la frase Recursos educativos para enseñanza media.

```
Recursos
        recurso NCMP000
educativos
        educativo AQ0MP00
para
        para SPS00
enseñanza
        enseñanza NCFS000
media
        medio AQ0FS00
. . Fp
```

Figura 6.4: Resultado de usar MACO y RELAX

- Le*ff*e [47] (*Léxico de formas flexionadas del Español*), un léxico morfológico y sintáctico de amplia cobertura y libre, puede ser usado directamente en aplicaciones de PLN de alto nivel, especialmente en aquellas que requieren un análisis sintáctico profundo.

- 3LB-SAT (3LB-Semantic Annotation Tool) [58], está orientado a la palabra (o token), permite introducir el corpus en diferentes formatos (TBF y XML) y que usa EuroWordNet para consultar el sentido de las palabras en cuatro lenguas (español, ca-talán, euskara e inglés). En la Figura 2.5 se puede observar un esquema de dicha herramienta. Sirve para el etiquetado semántico de corpus multilingüe.

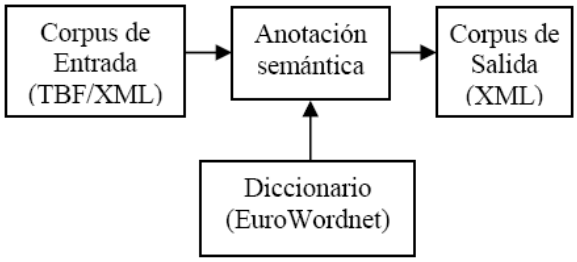


Figura 6.5: Estructura de la herramienta [58]

- Interfaz para implementación del Método de Marcas de Especificación para la desambiguación léxica[26]

- APOLN [75] Analizador Parcial de Oraciones en Lenguaje Natural, este ha sido usado en varias aplicaciones: sistemas de extracción de datos, traducción, resoluciones de ambigüedad, resumen de textos entre otras. Algo muy interesante de esta herramienta es que es capaz de manejar cualquier frase es decir a pesar de que esta tenga errores léxicos o no aceptados en la gramática del lenguaje. El análisis parcial es una alternativa a la definición de las gramáticas de amplia cobertura.

- Algunos ejemplos de sistemas para la WSD son:

  - ✓ Duluth Senseval-2 systems [27].- basado en árboles de decisión

  - ✓ SyntaLex.- versión mejorada del Duluth Senseval-2 con el uso de rasgos sintácticos [28]

  - ✓ SenseTools[29]

  - ✓ SenseRelate TargetWord[30] .- identifica los sentidos de una palabra basado en la similitud semántica de sus vecinos

  - ✓ SenseRelate-TargetWord [31].- usa WordNet Similarity, mide la similitud en Word-Net

  - ✓ InfoMa[32].- Representa los significados de las palabras en un vector (no supervisado)

  - ✓ SenseCluster[33].- encuentra clusters de palabras que ocurren en contextos similares

[26] WSD USING SPECIFICATION MARKS METHOD, http://gplsi.dlsi.ua.es/wsd
[27] Duluth Senseval-2 systems, www.d.umn.edu/~tpederse/senseval2.html
[28] www.d.umn.edu/~tpederse/syntalex.html
[29] SenseTools, www.d.umn.edu/~tpederse/sensetools.html
[30] search.cpan.org/dist/WordNet
[31] http://lit.csci.unt.edu/~senselear
[32] InfoMap, http://infomap-PLN.sourceforge.net
[33] SenseClusters, http://senseclusters.sourceforge.net

Anexo 2

Proyecto InTiMe

InTiMe es una plataforma que permite conocer, acceder, usar y compartir herramientas y corpus. Tiene una arquitectura cliente/servidor y distribuida realizada con servicios web que permite, por una parte, integrar en los servidores cualquier recurso y, por otra, tener acceso a las herramientas que trabajan con esos corpus remotamente y ejecutarla como si dicha herramienta estuviera ejecutándose localmente. El desarrollo de esta plataforma se la realizó usando JAVA lo cual trae varias ventajas. La arquitectura InTiMe puede ser adaptada a las necesidades del grupo de investigación de PLN. El la Figura 3.7 se muestra la arquitectura de dicha plataforma.

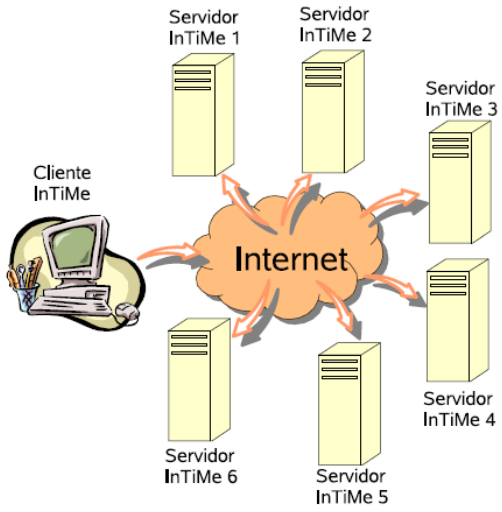


Figura 6.6: Arquitectura general de la plataforma InTiMe [46]

Cada servidor InTiMe almacena un subconjunto de las herramientas y los corpus que se pretendan integrar en la plataforma pero, al mismo tiempo, estos servidores conocerán que recursos hay instalados en los otros servidores. Esta plataforma hace más fácil el acceso a los recursos de tal manera que no tengamos que tener instaladas todas ellas en la máquina del grupo de trabajo sino acceder a ellas remotamente.

REFERENCIAS:

[1] Vallez, M., & Rafael Pedraza Jimenez, R. (s.f.). *El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines*. Obtenido de http://eprints.rclis.org/10700/1/El_Procesamiento_del_Lenguaje_Natural_en_la_Recuperaci%C3%B3n_de_Informaci%C3%B3n_Textual_y_%C3%A1reas_afines.pdf

[3] *Euro Voc*. (s.f.). Obtenido de http://www.bizkaia.net/kultura/eurovoc/presentacion.asp?Tem_Codigo=2862&idioma=CA

4] UED, N. L. (s.f.). Obtenido de http://nlp.uned.es/projects

[5] Gil Leiva, I., & Rodriguesz Muños, J. (s.f.). *Departamento de Información y Documentación de la Universidad de Murcia. .* Obtenido de revistas.ucm.es/byd/11321873/articulos/RGID9696220205A.PDF

[6] Moreno Sandoval, A. (1998). *Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos.* Madrid: Sintesis.

[8] Bel, N., & Calzolari, N. (s.f.). *SEPLN.* Obtenido de www.sepln.org/revistaSEPLN/revista/43/articulos/art51.pdf

[13] M., G. M. (s.f.). *Internet World Stats*. Recuperado el 2010, de http://www.internetworldstats.com/stats7.htm

[14] Gelbukh, A., & Sidorov, G. (2006). Obtenido de http://www.gelbukh.com/libro-procesamiento/LibroPLN.pdf

[15] Alberich, M. (2007). *Procesamiento del Lenguaje Natural.* Recuperado el 2010, de http://www.sopadebits.com/extranet/gallery/download/4492/pln-1.0-20070630.pdf

[16] Rodríguez Perojo, K., & León, R. R. (s.f.). Un enfoque desde la perspectiva de la automatización . *Organización y recuperación de la información* .

[17] Tello Leal, E. (s.f.). *Universidad Autónoma de Tamaulipas en México*. Obtenido de La Desambiguación del Sentido de las Palabras: revisión metodológica : http://www.nosolousabilidad.com/articulos/desambiguacion.htm

[18] Burnaga Rodriguez, M. (s.f.). *Universidad Complutense de Madrid*. Obtenido de Integración de Técnicas de Procesamiento del Lenguaje Natural para la Recuperación de información en Bibliotecas de Componentes Software: http://eprints.ucm.es/tesis/19911996/X/1/X1023301.pdf

[19] Tiedemann, J. (s.f.). *Lexicon Architecture Norms*. Obtenido de Guidelines and Recommendations: http://stp.ling.uu.se/~joerg/diplom/node5.html#SECTION00510000000000000000 0

[20] Vossen, P., Diez Orzas, P., Peters, W., & A. X. (s.f.). *Multilingual design of EuroWordNet.* Obtenido de http://www.aclweb.org/anthology/W/W97/W97-0801.pdf

[21] *OSI*. (s.f.). Recuperado el 2010, de Tecnología Lingüística: http://oesi.cervantes.es/oesi/tls.jsp

[22] InTiMe Plataforma de Integración de Recursos de PLN. (2008). *Procesamiento del Lenguaje Natural*

[24] Universal Networking Digital Language Fundation. (s.f.). Obtenido de http://www.undl.org/

[25] University Princeton. (s.f.). *WordNet A Lexical DataBase for English.* Obtenido de http://wordnet.princeton.edu/

[26] *Sociedad Española para el Procesamiento de Lenguaje Natural.* (s.f.). Obtenido de www.sepln.org

[27] Trenkle, W. (Abril de 1994). *TextCat Language Guesser Demo.* Obtenido de http://odur.let.rug.nl/~vannoord/TextCat/Demo/textcat.html

[28] *Fredy Andrés, R. R. (Julio de 2006). Recuperado el 2010, de http://www.cesarcabrera.info/proyectoGrado/Proyecto%20AIWSLN.pdf*

[29] *Alemany, L. A. (2005). Herramientas Libres para Procesamiento del Lenguaje Natural. Obtenido de Facultad de Matemática, Astronomía y Física UNC, Córdoba (Argentina): http://cs.famaf.unc.edu.ar/~laura/freeNLP*

[30] Alberich, M. (2007). *Procesamiento del Lenguaje Natural.* Recuperado el 2010, de http://www.sopadebits.com/extranet/gallery/download/4492/pln-1.0-20070630.pdf

[31] *CALA. (2010). Epistemowikia. Recuperado el 2010, de Procesamiento de lenguajes naturales/Principales dificultades del PLN: http://campusvirtual.unex.es/cala/epistemowikia/index.php?title=Procesamiento_de_lenguajes_naturales/Principales_dificultades_del_PLN*

[32] Sociedad Española para el Procesamiento de Lenguaje Natural (SPLN). (s.f.). *Procesamiento de Lenguajes Naturales.* Obtenido de http://procesamientolenguajenatural.50webs.com/pdf/Procesamiento%20de%20Lenguajes%20Naturales.pdf

[33]*Alemany, L. A. (2005). Herramientas Libres para Procesamiento del Lenguaje Natural. Obtenido de Facultad de Matemática, Astronomía y Física UNC, Córdoba (Argentina): http://cs.famaf.unc.edu.ar/~laura/freeNLP*

[34] *Montoyo, A. and Palomar, M. (2001). Specification Marks for Word Sense Disambiguation: New Development. 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001). México D.F. (México). Published in Lecture Notes in Computer Science, VOL 2004, Springer-Verlag, 2001. http://gplsi.dlsi.ua.es/wsd/*

[35] *Monográfico, Red Temática. Tratamiento de la información Multilingüe y Multimodal. Procesamiento del Lenguaje Natural, Revista nº 38, Abril de 2007*

[37] *Williams, Joseph. M. "Style Toward Clarity and Grace". The University of Chicago Press. Chicago and London*

[38] *Baker, Mona (1995): "Corpora in translation studies. An overview and some suggestions for future research"*

[39] (Revista de Procesamiento del Lenguaje Natural)Aplicación de técnicas basadas en PLN al tratamiento de preguntas médicas en Búsqueda de Respuestas.

[40] Haller, J., Donoso, A., & Ramírez, Y. (s.f.). *MPRO .* Obtenido de Un programa para el análisis morfológico y sintáctico de textos en español: http://www.sepln.org/revistaSEPLN/revista/29/29-Pag307.pdf

[41] Alfonseca, E., & Pilar, R. (s.f.). *Procesamiento Natural del Leguaje,.* Obtenido de Universidad Autónoma de Madrid: http://alfonseca.org/nlp/tema2.pdf

[42] *National Library of Medicine (1997). Unid Medical Language System (UMLS) Knowledge Sources, 6th experimental editio*

[43] *Miller, 1994, The MultiSemCor corpus, http://multisemcor.itc.it/*

[44] *General Architecture for text engineering, http://www.gate.ac.uk/*

[45] Isabel, S. B., Paloma, M., & Samy, D. (2008). Detección de fármacos genéricos en textos biomédicos. *Procesamiento del Lenguaje Natural* , 27-34.

[46] Gómez, J. M. (s.f.). *Departamento de Lenguajes y Sistemas Informáticos Universidad de Alicante.* Obtenido de Plataforma de Integración de Recursos de PLN: http://rua.ua.es/dspace/bitstream/10045/5039/1/PLN_40_10.pdf

[47] Molinero, M. A., Sagot, B., & Lionel, N. (2009). Construcción y extensión de un léxico morfológico y sintáctico para el español: el Leffe. *Revista de Procesamiento del Lenguaje Natural*

[48] *Schmid, 1994, TreeTagger, Institute for Computational Linguistics of the University of Stuttgart, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/*

[49] *Schmid, Probabilistic Part-of-Speech Tagging Using Decision Tree,http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf*

[50] *Alias-I, LingPipe, http://alias-i.com/lingpipe/*

[51] *João Pinto, F. (s.f.). Departamento de Ciencias de Computación Universidad de la Coruña. Obtenido de http://www.sistedes.es/sistedes/pdf/2007/Pinto.pdf*

[52] *Ureña López, A., & Buenaga Rodríguez, M. (s.f.). Departamento de Informática. Obtenido de Universidad de Jaén: http://erevista.aepia.org/index.php/ia/article/viewFile/248/235*

[53] *(Ureña López & Buenaga Rodríguez) (Porter)* http://tartarus.org/~martin/PorterStemmer/index-old.html

[54] *Porter's Stemming Algorithm Online,* http://maya.cs.depaul.edu/classes/ds575/cgi-bin/porter-online-stop.pl

[55] Pérez Agüera, J. R. (s.f.). *Recuperación de información.* Obtenido de Depto. De Sistemas Informáticos y Programación, Facultad de Informática, Universidad Complutense de Madrid

[56] Aguado de Cea, G., Álvarez de Mon y Rego, I., & Pareja Lora, A. (s.f.). *Facultad de Informática, UPM, Madrid-España.* Obtenido de Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: OntoTag: http://www.fi.upm.es/dlacyt/sites/www.fi.upm.es.dlacyt/files/onto_tag.pdf

[58] Bisbal, E., Antonio, M., Moreno, L., Pla, F., & Saiz Noeda, M. (s.f.). *Universidad de Alicante y Politécnica de Valencia, 3LB-SAT.* Obtenido de http://hdl.handle.net/10045/1510

[61]MACO, Analizador Morfológico, RELAX, Etiquetador POS, http://nlp.uned.es/~anselmo/catalogo_rile.html#MACO

[62] Rada, R., H. Mili, E. Bicknell y M. Blettner (1989), *Development an Application of a Metric on Semantic Nets, IEEE Transactions on Systems, Man and Cybernetics*

[71]Lesk, Michael (1986), *Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone, en Proceedings of the 1986 SIGDOC Conference, Association for Computing Machinery*

[72] Cowie, Jim, Joe Guthrie y Louise Guthrie (1992), *Lexical disambiguation using simulated annealing, en Proceedings of the 14th International Conference on Computational Linguistic*

[74] Leech, G. 1997 *Introducing corpus annotation. Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman*

[75] *APOLN: A Partial Parser Of Unrestricted Text (1999), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.8160*

[76] Vázquez Pérez, S. (s.f.). *Departamento de Lenguajes y Sistemas Informáticos Universidad de Alicante.* Obtenido de http://rua.ua.es/dspace/bitstream/10045/11456/1/Tesis_vazquez.pdf

[78] Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Workshop on, Multilingual Linguistic Resources,* http://wndomains.fbk.eu/publications/Coling-04-ws-WDH.pdf

[79] Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet Developing an aligned multilingual database. *1st International WordNet Conference*, (págs. pag. 293-302). Mysore, India http://multiwordnet.fbk.eu/paper/MWN-India-published.pdf

[80] Tufis. D, Cristea D., Stamou S., Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania Faculty of Computer Science, "Al. I. Cuza" University of Iasi Romania, Research Academic Computer Technology Institute, Patras, Greece, BalkaNet: Aims, Methods, Results and Perspectives, Romanian Journal of Information Science And Technology Tufis-CS-ROMJIST2004.pdf