INSTITUTO POLITÉCNICO NACIONAL CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Similitud léxico-semántica en un corpus de textos grande coherente

TESIS QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN PRESENTA

Reyna Elia Melara Abarca

DIRECTORES DE TESIS:

DR. ALEXANDER GELBUKH ***DRA. SOFÍA N. GALICIA HARO***

MÉXICO D.F DICIEMBRE DE 2010

Índice

Capítulo 2

Trabajo relacionado

2.1 Introducción: La Wikipedia.

Es un proyecto sin fines de lucro de la Fundación Wikimedia, que consiste en una enciclopedia en línea, de libre edición y consulta.

Algunas de sus características que la han convertido en un recurso en línea relevante son:

- su gran tamaño, con un índice de crecimiento constante,
- su contenido semi-estructurado, que se reconoce de alta calidad,
- es independiente del dominio,
- está disponible en varios idiomas (multilingüe)
- es de acceso libre, disponible para edición y uso incluso fuera de línea.

Su interfaz de usuario es una aplicación de software basada en Web, que se ejecuta en el nivel más alto en una arquitectura LAMP. Se edita en texto plano por un lenguaje de marcas para estructurar documentos Wiki, que permiten la creación de manera simple y adhoc de documentos de contenido colaborativo. Es independiente del dominio, se actualiza constantemente y es multilingüe. Se gestiona a través del software libre y de código abierto MediaWiki que permite mantener, crear, configurar y usar los Wiki.

El recurso básico de Wikipedia es un *artículo* (o página), que define y describe una entidad o evento, y consiste de un documento de hipertexto con hipervínculos a otras páginas internas o externas de Wikipedia (Mihalcea, 2007). Los artículos se relacionan con otros artículos a través de hipervínculos que son palabras o frases dentro del contenido de cada artículo, de modo que la información se complementa formando una red de artículos y vínculos, los cuales no necesariamente son creados por el mismo autor.

Estos artículos se van agregando bajo el concepto de colaboración de los usuarios, quienes pueden estar geográficamente distribuidos en diferentes países. La mayoría de estos artículos pueden editarse libremente.

Una de las ventajas de Wikipedia es que la "libertad de contribución" tiene un impacto positivo, tanto en aspectos cualitativos, como es el aumento en el número de artículos y cuantitativos, como la corrección rápida de errores, que se propicia por tener esta característica de ambiente colaborativo (Mihalcea, 2007).

Sin embargo, también tiene desventajas. Esta misma característica de "libertad de contribución" genera vicios como la saturación en algunos artículos, artículos con vínculos que no conducen a información relevante, vínculos rotos y falta de seriedad en las fuentes de información.

Cifras publicadas en Wikipedia indican que para enero de 2010, se habían registrado alrededor de 68 millones de visitas a las Wikipedias (que es como se le denomina al conjunto de las Wikipedias de diferentes lenguaje), y que cerca de 91,000 usuarios participan activamente, contribuyendo a la edición de aproximadamente 15,000,000 artículos en más de 270 lenguajes. Durante el 2008, las Wikipedias de los lenguajes inglés, alemán, francés, polaco y japones, en ese orden, fueron las que reflejaron el mayor número de ediciones de artículos.

En algunos estudios comparativos sobre las Wikipedias, las versiones en inglés y en alemán indican ser las más robustas y mejor estructuradas, mientras que la Wikipedia en español es menos confiable, ya que incurre en un sin número de errores y carece de fuentes fidedignas de información, esto debido a que "cualquiera puede verla o editarla, sin la intervención de moderador alguno, ni la sujeción a ningún filtro..." (Maldonado, 2010).

En el caso de la Wikipedia en inglés[1], algunas cifras estadísticas a septiembre de 2011, indican lo siguiente (Cuadro 1):

Cuadro 1: Datos estadísticos de la Wikipedia en Inglés.
[Warning: Image ignored]

El propio fundador de Wikipedia Jimmy Wales señala: "A una década de distancia, cerca de 400 millones de personas utilizan Wikipedia y sus sitios hermanos cada mes – alrededor de la tercera parte del mundo conectado por Internet[2]".

Como puede observarse, Wikipedia tiene características interesantes, como lo son su vasto volumen de contenidos, el número de usuarios que la editan y la consultan, lo cual deriva en un comportamiento y crecimiento dinámico.

Wikipedia constituye también un recurso que ha sido ampliamente utilizado para fines académicos y de investigación[3], tal es el caso del Procesamiento de Lenguaje Natural (PLN), lo cual se abordará en el resto de este capítulo.

2.2 Estructura de Wikipedia

Wikipedia está conformada por millones de artículos, organizados en categorías. Los artículos se relacionan por medio de hipervínculos. La mayoría de estos artículos pueden ser editados libremente.

Los artículos son páginas Web. Además de estas páginas existen otras que no son de contenidos que tengan entradas bibliográficas, sino las que se conocen como páginas administrativas.

Las páginas, dependiendo del tipo tienen un formato específico, de manera concreta, todo lo que se incorpore a la Wikipedia debe de realizarse de acuerdo a las plantillas que la plataforma determina y en base a las políticas que al respecto tengan señaladas.

El contenido de la Wikipedia puede ser descargado de los sitios que oficialmente tiene disponibles, lo cual permite que sea un recurso útil para fines académicos.

De manera formal se puede decir que es un recurso semiestructurado, constituido por un lado, por componentes bien definidos, como son la estructura de hipervínculos y la jerarquía de temas o categorías, y por otro lado, recursos no estructurados como lo son las colecciones de texto que de ella se generan.

---

[1] Statistics, http://en.wikipedia.org/wiki/Special:Statistics

[2] An appeal from Wikipedia founder Jimmy Wales, http://wikimediafoundation.org/wiki/Special:LandingCheck?landing_page=WMI

[3] Wikipedia:Academic studies of Wikipedia, http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies

2.2.1 Recursos estructurados Se puede considerar que su estructura está altamente organizada en relación a las siguientes características:

1 Cuenta con un sistema de categorías, cuya finalidad es organizar los artículos, en el que cada artículo debe de pertenecer al menos a una categoría. Este sistema se puede pensar como un grafo o como tesauro.

2 Artículos, que conforman una red con características de grafo, los cuales deben cumplir con un formato establecido (MediaWiki), dividido en secciones, párrafos y las correspondientes relaciones entre páginas:

- 2.1 *páginas de redireccionamiento* (redirect pages), que establecen relaciones de sinónimos,
- 2.2 *páginas de desambiguación*, (disambiguation pages), que establecen relaciones de homónimos,
- 2.3 Conjunto de hipervínculos o vínculos internos (internal links), que establecen relaciones de referencias cruzadas y que también se puede considerar un grafo similar a la red WWW.

3 Información contenida en estructuras en forma de listas y tablas:

- 3.1 contenido de las páginas,
- 3.2 listas de vínculos página-a-página (tablas pagelinks, categorylinks, imagelinks),
- 3.3 metadatos de imágenes (tablas image y oldimage),
- 3.4 misceláneas (tablas interwiki y site_stats).

2.2.2 Recursos no estructurados

1. Texto en formato wiki (wikitext) y metadatos embebidos en XML disponibles para descarga, en forma de respaldos (dump) de la base de datos de Wikipedia.
2. Archivos estáticos en hipertexto, Wikipedia Wikis en formato HTML, disponibles para descarga, en forma de respaldos (dump).

Los archivos en texto plano Wikipedia, suelen contener varios gigabytes de información, que requieren de recursos adecuados tanto de almacenamiento como de procesamiento.

2.2.3 Artículos Como ya se mencionó, la unidad representativa de la Wikipedia es el *artículo*. Los artículos tienen una forma y un contenido. La forma la determina el software de MediaWiki y el contenido se conforma en co-autoría por una comunidad de usuarios.

2.2.3.1 Formato de los artículos: Páginas wiki Los artículos se construyen a través del navegador de MediaWiki utilizando un patrón de diseño conocido como *página wiki*, por medio de un lenguaje propio conocido como *wikitexto*.

Usualmente las páginas wiki tienen tres representaciones[4] (Cuadro 2):

[4] Wiki, http://es.wikipedia.org/wiki/Wiki

Código fuente	Que puede editarse por la comunidad de usuarios. Formato que se almacena localmente en el servidor. Texto plano que puede visualizarse unicamente por medio de la operación "Editar".
Código HTML	El utilizado por el servidor para mostrar la información a partir del código fuente en tiempo real ("al vuelo").
Plantilla o patrón de diseño	Contiene la disposición y elementos comunes a todas las páginas

Cuadro 2: Páginas y edición.

Los usuarios a través de la interfaz de MediaWiki tienen opciones de creación y edición de artículos.

Wikipedia tiene registro de los datos de creación y modificación de cada artículo.

2.2.3.2 Contenido de las Páginas wiki: Artículos

Cada artículo tiene una identificador que lo referencia de manera unívoca, el cual consiste en una o más palabras separadas por espacios o guiones bajos y ocasionalmente una explicación entre paréntesis (Cuadro 3).

Título del artículo	**Identificador único del artículo**
Pentagonal number theorem	Pentagonal_number_theorem
Partition (number theory)	Partition_function_(number_theory)

Cuadro 3: Identificador único de un artículo de Wikipedia.

Los artículos utilizan uno de los recursos más importantes de la World Wide Web, los *hipervínculos*. Los hipervínculos pueden ser *internos* o *externos*. Los *hipervínculos internos* o *vínculos internos,* son artículos que hacen alusión a otras "entradas enciclopédicas" a las que se apunta con hipervínculos, lo cual da origen a un modelo de *referencias cruzadas* de los artículos de Wikipedia.

Los hipervínculos internos se forman con los identificadores de los artículos y forma externa del hipervínculo o texto anclado (anchor text). En el lenguaje de marcas de Wikipedia, los hipervínculos se forman colocando el texto de la forma externa del hipervínculo entre doble paréntesis cuadrados. Se utiliza el símbolo | (barra vertical) cuando la forma externa se vincula a un artículo con un identificador único diferente (Cuadro 4).

Lenguaje de marcas de Wikipedia	**Anchor text**	**URL**
[[Partition function (number theory)|unrestricted partition functions]]	unrestricted partition functions	http://en.wikipedia.org/wiki/Partition_function_(nu

[[pentagonal numbers]]	pentagonal numbers	http://en.wikipedia.org/wiki/Pentagonal_numbers

Cuadro 4: Formato de los hipervínculos internos.

Los hipervínculos externos, son vínculos a páginas fuera de la Wikipedia y en su lenguaje de marcas re forman utilizando paréntesis cuadrados y un espacio que separa la URL del nombre del vínculo (Cuadro 5).

Lenguaje de marcas de Wikipedia	Forma externa del hipervínculo externo
[http://www.mathpages.com/home/kmath623/kmath623.htm On Euler's Pentagonal Theorem] at MathPages	On Euler's Pentagonal Theorem at MathPages
[http://front.math.ucdavis.edu/math.HO/0510054 Euler and the pentagonal number theorem]	Euler and the pentagonal number theorem

Cuadro 5: Formato de los hipervínculos externos.

2.2.4 Páginas de redireccionamiento Como resultado del ambiente colaborativo de Wikipedia, se tienen algunas implicaciones. Una de estas es la falta de consistencia respecto al uso del identificador único para una determinada entidad, lo que da como resultado las *páginas de redireccionamiento* (Cuadro 6), que no tienen contenido, pero redirigen al lector a otro artículo, sección de un artículo u otra página, usualmente con un título alternativo[5].

Tipo de página	Ejemplo
Página de redireccionamiento: Se utilizan para ayudar a los usuarios a encontrar información y mantener organizados los wikis, de modo que múltiples nombres, abreviaciones, errores de ortografía o temas relacionados se dirijan hacia la misma página.	Mathematics "Maths" and "Math" redirect here. For other uses of "Mathematics" or "Math", see Mathematics (disambiguation) and Math (disambiguation). Mathematics is the body of knowledge justified by deductive reasoning about abstract structures, starting from axioms and definitions. **Mathematics** may also refer to • *Mathematics* (producer), a hip-hop producer • *Mathematics* (album), an album by the band The Servant • "*Mathematics*" (song), a song by Mos Def • "*Mathematics*" (Little Boots song), a song by Little Boots • *Mathematics Magazine*, a publication of the Mathematical Association of America

Cuadro 6: Páginas de redireccionamiento.

[5] Wikipedia:Redirect, http://en.wikipedia.org/wiki/Wikipedia:Redirect.

2.2.5 Páginas de desambiguación Las *páginas de desambiguación* (Cuadro 7), no son artículos, tienen por objetivo proporcionar al usuario que realiza una búsqueda con un término ambiguo, una lista de artículos que pudieran ser lo que esta buscando.

La desambiguación de artículos de Wikipedia se realiza cuando dos o más artículos hacen referencia a temas diferentes pero el título de la página tiene el mismo nombre[6], es decir, que pueden ser referenciados por el mismo término de búsqueda.

Una página de desambiguación tiene los vínculos a distintos artículos que corresponden a términos susceptibles de crear confusión o generar ambigüedad.

Las páginas de desambiguación, incluyen los diferentes significados de un término y el vínculo correspondiente a cada artículo de Wikipedia por término[7]. Para evitar la ambigüedad, existen convenciones para títulos de los artículos que recomiendan que los títulos de las páginas sean con un nombre natural seguido de un dato relevante que ayude a distinguir el significado de la palabra.

Tipo de página	Ejemplo
Página de desambiguación: típicamente su identificador único consiste de una explicación entre paréntesis (desambiguación) junto al nombre de la entidad ambigua.	• *Number*(game), a number-guessing computer game • *Number*(magazine), a Japanese sports magazine • *Number*(manga), a manga by Tsubaki Kawori • *Number*(music), a self-contained piece of music

Cuadro 7: Páginas de desambiguación.

2.2.6 Categorías de la Wikipedia Las *categorías* se utilizan para organizar la Wikipedia. Representan temas principales y los artículos o páginas se clasifican de acuerdo al tema, es decir, se asocian al menos a una categoría.

El software de MediaWiki, permite que las páginas se agreguen a listados automáticos, que ayudan al proyecto a tener una estructura que agrupa páginas de temas similares[8].

Wikipedia destaca dos categorías principales:

(1) Categorías de temas. La categoría contiene artículos relacionados a un tema y comparten el nombre: **Category:Mathematics** , organiza todo el contenido relacionado a Matemáticas, se puede leer en la página principal de Mathematics:

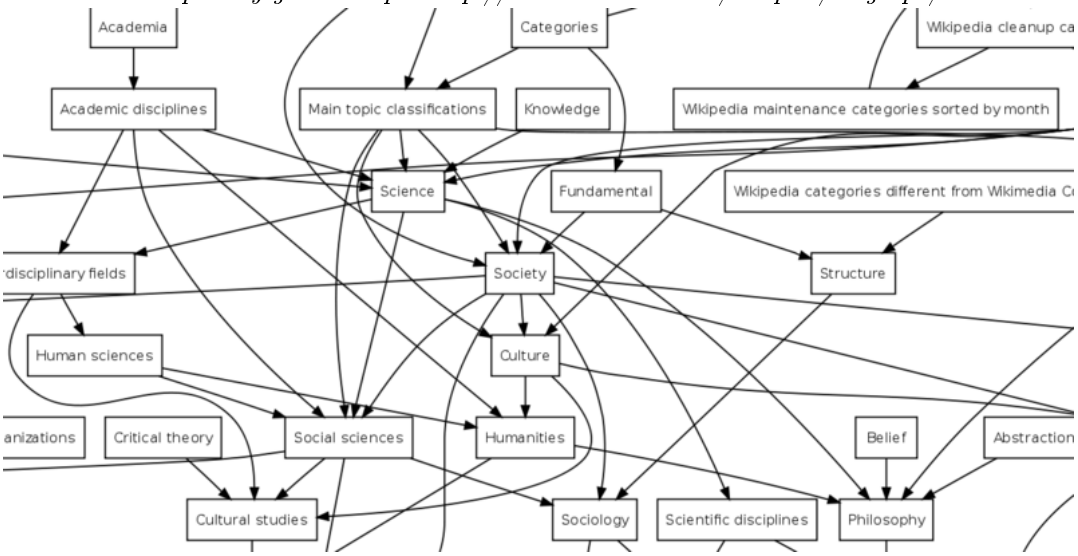
The main article for this category is Mathematics.
(El artículo principal de esta categoría es Matemáticas).

[6]Category:Disambiguation pages, http://en.wikipedia.org/wiki/Category:Disambiguation_pages
[7]Wikipedia:Disambiguation, http://en.wikipedia.org/wiki/Wikipedia:Disambiguation.
[8]Wikipedia: Help:Category, http://en.wikipedia.org/wiki/Help:Category

(2) Categorías conjunto. Definen una clase usualmente en plural: **Category:Mathematical theorems**, que debe incluir todos los teoremas matemáticos documentados en la Wikipedia.

De acuerdo a la propia definición de Wikipedia, las categorías (Figura 1) pueden pensarse como *árboles* que se traslapan.

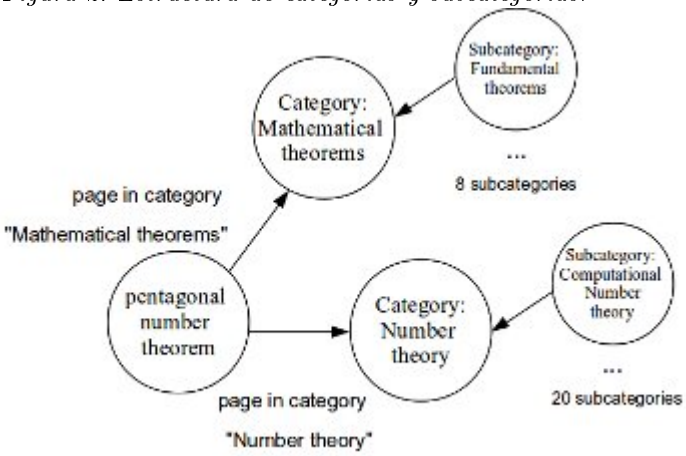
Figura 1: Pequeña porción del Grafo de Categorías de Wikipedia, imagen extraída de Wikipedia y generada por http://tools.wikimedia.de/~dapete/catgraph/.



Cualquier categoría puede tener *subcategorías* (Figura 2) y es posible que una categoría sea subcategoría de más de un *padre* y que matemáticamente, el sistema de categorías se aproxima a un *grafo acíclico dirigido*[9]*:*

$$Se\ dice\ que\ A\ es\ una\ categoría\ padre\ de\ B,\ cuando\ B\ es\ subcategoría\ de\ A.$$

Figura 2: Estructura de categorías y subcategorías.



[9] Wikipedia:Categorization, http://en.wikipedia.org/wiki/Wikipedia:Categorization

Existen guías del uso recomendado del software de MediaWiki y de edición de artículos, que sugieren evitar la creación de ciclos, sin embargo, debido al volumen de la Wikipedia y la libertad de los usuarios para elegir libremente los criterios para generar, nombrar los artículos y clasificarlos (Folksonomía[10]) en las que no se establecen explícitamente relaciones semánticas subyacentes, la calidad con la que se organizan los artículos en categorías es altamente variable, existen algunas muy detalladas y bien organizadas y en otros casos la categorización se ha dado de manera ad hoc o pobremente.

Tampoco existe una restricción sobre el tipo de categoría de mayor nivel al que debe vincularse una categoría hija, por tanto la estructura de categorías no se considera estrictamente como un árbol o un grafo acíclico dirigido, ya existen casos paradójicos en los que una categoría puede ser su propio padre (Kittur et al., 2009).

De acuerdo a (Zesch & Gurevych, 2007), la organización de las categorías también es como un grafo con una estructura taxonómica, en el que cada categoría puede tener un número arbitrario de subcategorías en el que se establecen relaciones de hiponimia o meronimia, y tampoco se considera del todo estricta, debido a la existencia de ciclos y categorías disconexas[11].

Por otro lado, para (Ponzetto & Strube, 2007b) el sistema de categorías es como un tesauro organizado temáticamente, en el que las categorías no forman una estructura arbórea, sino un grafo dirigido (Ponzetto & Strube, 2006b), en el que un artículo puede aparecer en más de una categoría, y cada categoría puede tener más de una categoría padre.

Todas estas percepciones son coincidentes al mencionar que la estructura de categorías no puede considerarse de manera incuestionable como un grafo o una taxonomía, ya que la libertad del ambiente colaborativo genera que sea desordenada, mal formada y difícil de encontrarle sentido (Kittur et al., 2009). De aquí que para el PLN, su tratamiento sea diverso como se documenta en el resto del capítulo.

2.2.7 Wikipedia como recurso para descarga

Al ser un proyecto de software libre, Wikimedia permite descargar[12] el contenido wiki de la Wikipedia, en forma de archivos de respaldo (*dumps*), el sitio de Wikipedia para descarga menciona los posibles usos de este recurso:

- con propósitos de respaldos,
- pasa uso fuera de línea,
- con fines académicos,
- para publicar bajo los términos de las licencias pertinentes,
- por diversión.

Los archivos para descarga están protegidos bajo los términos de las licencias Creative Commons Attribution-ShareAlike License 3.0 (CC-BY-SA) y GNU Free Documentation License (GFDL).

Los *dumps* están en SQL o en XML en el sitio http://dumps.wikimedia.org/. Las imágenes y otros archivos con o sin derechos de autor, se descargan por separado en base a términos de uso diferentes de los que tienen los archivos de texto.

2.3 Representaciones de Wikipedia

[10] Folksonomía es la traducción al español de la palabra folksonomy que surge de la combinación de los morfemas en inglés Folk (gente) y taxonomy (taxonomía), que es el resultado del etiquetado colaborativo, clasificación social, clasificación colectiva o indexado social para crear y administrar contenidos digitales.

[11] De acuerdo a los autores, para mayo 15 del 2006, en la Wikipedia en alemán, el componente con más conexiones del grafo de categorías de Wikipedia contenia 99.8% de todos los nodos de categorías y 7 ciclos.

[12] Página de información de los archivos de descarga de la Wikipedia http://en.wikipedia.org/wiki/Wikipedia:Database_download

En este sentido, para las tareas de PLN se tiene noción de que la Wikipedia ha sido considerada como:

- Corpus, por su una amplia colección de textos.
- Tesauro, con relaciones jerárquicas y de equivalencia entre términos y relaciones como la sinonimia, entre otras.
- Base de conocimiento, los artículos en Wikipedia se encuentran fuertemente interconectados y esta estructura se enriquece con su sistema de categorías (Zesch et al., 2008a).
- Base de datos, extracción y codificación de estructuras y relaciones.
- Ontología, la expresión formal de las relaciones en la Web Semántica y construcciones lógicas.
- Estructura de red, análisis de las relaciones y extracción, sobre una representación tipo grafo de red (network graph).

FALTA AMPLIAR LA LISTA

2.4 Wikipedia en el Procesamiento de Lenguaje Natural Wikipedia constituye base amplia de conocimiento léxico-semántica, que por sus características ha permitido que sea utilizada en el Procesamiento de Lenguaje Natural en tareas tales como (Zesch et al., 2008a):

- Recuperación de información,
- Categorización de textos,
- Wikificación de textos,
- Búsqueda de respuestas,
- Resúmenes automáticos,
- Cálculo de proximidad semántica,
- Desambiguación del sentido de las palabras.

Muchas de las investigaciones al respecto, han sido motivadas para que el PLN apoye tareas relacionadas a la educación.

2.4.1 Recuperación de Información (RI) La recuperación de información (RI) debe interpretar los contenidos de los documentos y hacer un ranking de las respuestas. La consulta no es estructurada y puede ser ambigua. La relevancia es el principal punto de interés (Baeza-Yates & Berthier, 1999).

La base documental más grande del mundo, es sin duda la Web, sin embargo presenta algunos problemas como:

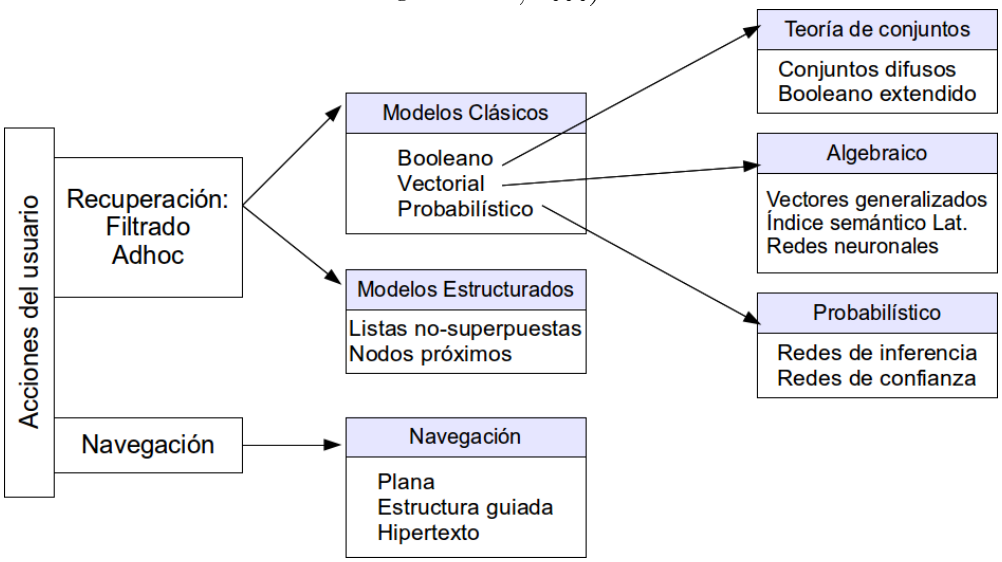
- No hay responsables de todos los contenidos,
- No es fácil buscar ni indexar,

- No se usa un lenguaje útil para las máquinas.

La RI se basa en la utilización de términos para indexar y recuperar documentos. Indexar un documento consiste en representar su contenido por un conjunto de términos índices. Recuperar significa especificar un conjunto de términos que deben hallarse entre los índices del documento, estableciendo un ranking de relevancia.

El problema de recuperación de la información es la manera de predecir la relevancia de los documentos y su ranking. Las distintas premisas utilizadas en el cálculo de la relevancia darán lugar a distintos modelos de trabajo (Figura 3) o de recuperación de información.

Figura 3: Taxonomía de los Modelos de Recuperación de Información (Baeza-Yates & Berthier, 1999).



El modelo de la técnica RI se define como una cuádrupla $[D, Q, F, R(q_i, d_j)]$, con:

(1) D es un conjunto de representaciones de documentos

(2) Q es un conjunto de representaciones de necesidades de información de los usuarios

(3) F es un marco de modelado de documentos, consultas y sus relaciones

(4) $R(q_i, d_j)$ es una función de ranking que asocia un número real con una consulta y un documento. El ranking define el orden en el que el documento satisface la consulta.

Las premisas que forman la base para los algoritmos de ranking determinan el modelo de IR a seguir.

2.5 Categorización de textos La categorización de textos consiste en asignar documentos a dos o más subcategorías, que ya han sido definidas como resultado de procesos de Recuperación de Información al entrenar corpus de documentos que previamente fueron clasificados. (Manning & Hinrich, 1999).

Mayormente, los clasificadores de texto utilizan técnicas de aprendizaje de máquina y representan los textos como *bag of words (BOW)*.

La información contenida en Wikipedia, ha sido utilizada también para crear un clasificador auxiliar de texto, que permite relacionar documentos con artículos relevantes de Wikipedia, esto es, con los conceptos de los artículos se aumenta espacio de palabras (o BOW), método que los autores enmarcan en el constructivismo inductivo[13] (Gabrilovich & Evgeniy, 2006).

El procesamiento utiliza texto plano, que permite aplicar los algoritmos de similitud para identificar automáticamente los artículos relevantes para cada documento. Se realiza un generador de características, que trabaja en cada documento con contextos individuales de diferentes niveles: por palabras , sentencias, después párrafos y finalmente con el documento completo. Al trabajar con contextos individuales, implícitamente se realiza desambiguación del sentido de las palabras y se controla la polisemia, gracias al contexto y texto circundante.

En este caso, de acuerdo a lo expuesto por los autores, si se quiere información sobre "jaguar car models" los resultados tendrán únicamente relación con autos y modelos de autos, mientras que si se solicita información de "jaguar Panthera onca" los resultados serán relacionados a animales.

Grafos de Wikipedia

Como se ha mencionado, Wikipedia tiene una estructura organizada, que puede apreciarse en sus categorías y los vínculos entre artículos. Otra forma de organización en la que es concebida la Wikipedia es como un grafo dirigido, en el que los nodos son temas y las aristas son los hipervínculos entre estos. Este tipo de estructura comparte características topológicas similares a las de la WWW (World Wide Web) (Capocci2006),

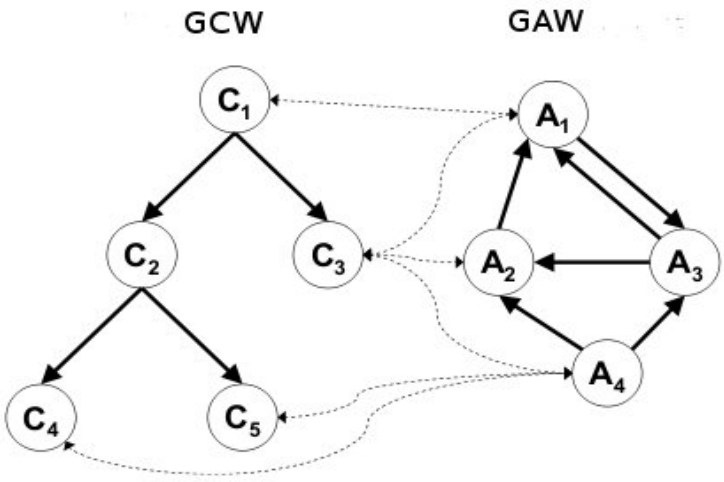
Suponiendo que cada artículo de Wikipedia puede vincularse a un número arbitrario de categorías, en la que cada categoría es un tipo de etiqueta semántica para ese artículo. Una categoría regresa los vínculos, a todos los artículos de esa categoría, además de visualizarse como un solo grafo, se le ha tratado como grafos específicos:

- el *grafo de categorías GCW (Wikipedia Category Graph)* y
- el *grafo de artículos GAW (Wikipedia Article Graph)*.

Los grafos de artículos y categorías están fuertemente vinculados (Figura 4) pero son tratados como estructuras diferentes, cada una con sus propias características específicas.

[13] Estudio de métodos que dotan al estudiante con la habilidad de modificar o mejorar la representación del lenguaje.

Figura 4: Relación entre los grafos de categorías (GCW) y de artículos (GAW) de Wikipedia.



Esta separación obedece a que, los vínculos entre artículos son las relaciones producto del trabajo colaborativo, mientras que los vínculos entre categorías son establecidos por relaciones de hiponimia o meronimia (18).

En el PLN, Wikipedia en su forma de grafo, mayormente como GCW, ha sido explotado como:

- Mapa de categorías basado en la co-ocurrencia de categorías (Holloway).
- Tesauro que combina etiquetado colaborativo e indexado jerárquico (Voss2006).
- Fuente de conocimiento léxico-semántica (Zesch2007).

Grafo de categorías de Wikipedia -GCW

El grafo de categorías de Wikipedia - GCW es equiparable a los grafos bien tipificados de las redes léxico-semánticas. Es el resultado del ambiente colaborativo en el que se ha construido Wikipedia.

Como se ha mencionado el GCW tiene características y propiedades de las redes léxico-semánticas, como WordNet y ha sido utilizado como un recurso en el PLN, por ejemplo, para calcular la proximidad semántica (Zesch2007).

Para el cálculo de proximidad semántica del GCW se ha probado por medio de un análisis formal, basado en la teoría de grafos y orientado particularmente para redes semánticas de palabras (wordnets).

Para este análisis el grafo dirigido de categorías G, se trata como un grafo no dirigido, dado que las relaciones que conectan categorías son reversibles:

$$\text{Learning to link with Wikipedia. } G = (V, E)$$

donde,
V representa el conjunto de nodos o vértices y
E es el conjunto no ordenado de pares de distintos vértices, conocidos como arcos o aristas.

Cada página se considera un *nodo* n , cada vínculo entre páginas es un arco o una arista e . Cómo los grafos de estructuras semánticas, el GCW se puede definir por el conjunto de parámetros propios de los grafos, como lo son:

- *grado de los nodos* k , que es el número de arcos conectados con ese nodo.
- *grado promedio* $\bar{k}$, el promedio sobre todos los nodos.
- *rutas o caminos entre nodos* $p_{i,j}$, es la secuencia de los arcos que conectan a un nodo n_i con un nodo n_j .
- *longitud de ruta* $l(p_{i,j})$, es el número de arcos en la ruta. Puede haber más de una ruta entre los nodos, de modo que se puede calcular la longitud de la ruta más corta $L_i, j = minl(p_{i,j})$.
- *promedio de la ruta más corta* $\bar{L}$ sobre todos los nodos.
- *diámetro* D , es la máxima longitud de la ruta más corta entre todos los pares de nodos del grafo.
- *coeficiente de cluster* de un nodo n_i puede calcularse de la siguiente manera:

$$C_i = \frac{T_i}{\frac{k_i(k_i-1)}{2}} = \frac{2T_i}{k_i(k_i-1)}$$

donde T_i se refiere al número de arcos entre vecinos del nodo n_i y $k_i(k_i-1)/2$ es el número máximo de arcos que pueden existir entre los k_i vecinos del nodo n_i .

El coeficiente de cluster C para todo el grafo es el promedio de todos los C_i . En un grafo conexo, el coeficiente de cluster es 1.

Las medidas de proximidad semántica aplicables a redes semánticas de palabras (i.e. WordNet, tesauro Roget, entre otras (Zesch2007)) se aplican sobre el GCW:

- Todos los grafos analizados son grafos de mundos pequeños, los cuales contienen clusters que están conectados por vínculos de rangos amplios que conducen a valores pequeños de $\bar{L}$ y D , esto es, los grafos de mundos pequeños están caracterizados por tener (i) valores pequeños de $\bar{L}$ y (ii) valores altos de C .
- Todas las redes semánticas son grafos de escala libre (scale-free graphs), ya que su grado de distribución sigue una ley exponencial.

De modo particular los resultados demostraron que tanto WordNet como el GCW son (i) grafos de libre escala y grafos de mundos pequeños y (ii) tienen un conjunto de parámetros muy similar.

Medidas de proximidad semántica aplicadas a un grafo

Existen muchas medidas para el cálculo de proximidad semántica en redes semánticas (Zesch2007) (Tabla 2):

Tabla 2. Medidas basadas en WordNet y aplicables a Wikipedia

Autor	Descripción	Cálculo
Rada et al. (1989)	Longitud de la ruta en arcos (Path Lenght) entre dos nodos.	$dist_{PL} = l(n_1, n_2)$

Leacock y Chodorow (1998)	Normaliza la longitud de la ruta con la profundidad del grafo	$\mathrm{sim}_{LC}(n_1, n_2) = -\log(\frac{l(n_1), n_2}{2 \times depth})$
Wu y Palmer (1994)	Introduce una medida para saber que tan similares son los sentidos de dos palabras basandose en la profundidad de los dos sentidos en la taxonomía y en el Least Common Subsumer de dos nodos (nodo antecesor más específico).	$\mathrm{sim}_{\mathrm{WP}} = \frac{2depth(lcs)}{l(n_1, lcs) + l(n_2, lcs) + 2depth(lcs)}$
Resnik (1995)	Define la similitud semántica entre dos nodos como el valor de contenidos de información (IC) de su lcs. Utiliza la frecuencia relativa en un corpus para estimar el valor de contenido de la información.	Res
Jiang y Conrath (1997)	Adicionalmente usa el IC de los nodos. El resultado que devuelve es una distancia en lugar de un valor de similitud.	$dist_{JC}(n_1, n_2) = IC(n_1) + IC(n_2) - 2IC(lcs)$
Lin (1998)	Define la similitud semántica usando una formula derivada de la teoría de la información.	$\mathrm{sim}_{Lin}(n_1, n_2) = \frac{2 \times IC(lcs)}{IC(n_1) + IC(n_2)}$

Como las palabras polisémicas pueden tener más de un nodo correspondiente en una red semántica de palabras, la proximidad semántica entre dos palabras w_1 y w_2 puede calcularse de la siguiente manera:

$$SR = \begin{cases} \min\limits_{n_1 \in s(w_1), n_2 \in s(w_2)} dist(n_1, n_2) & \text{path} \\ \max\limits_{n_1 \in s(w_1), n_2 \in s(w_2)} sim(n_1, n_2) & \text{IC} \end{cases}$$

donde $s(w_i)$ es el conjunto de nodos que representa sentidos de la palabra w_i, esto significa, que la proximidad de dos palabras es igual al par de nodos más relacionados.

A diferencia de otras redes léxico-semánticas, los nodos del GCW no son synsets o términos únicos, sino un concepto generalizado o una categoría, de modo que es necesario hacer modificaciones para adaptar las medidas de proximidad semántica al GCW. La proximidad semántica entre artículos se mide por las categorías asignadas a los artículos.

Se definen C_1 y C_2 como un conjunto de categorías asignadas a los artículos a_i y a_j respectivamente. Se determina la proximidad semántica para cada par de categorías (c_k, c_l) con $c_k \in C_1$ y $c_l \in C_2$. Se selecciona el valor más apropiado de entre todos los pares (c_k, c_l) , como por ejemplo, el valor mínimo para basados en ruta y el máximo para las medidas basadas en contenidos de información.

$$SR_{best} = \begin{cases} \min\limits_{c_k \in C_1, c_l \in C_2} (sr(c_k, c_l)) & \text{path based} \\ \max\limits_{c_k \in C_1, c_l \in C_2} (sr(c_k, c_l)) & \text{IIC based} \end{cases}$$

Si se sustituye el contenido de información de Resnik con información de contenido intrínseco IIC del grafo subyacente se obtienen mejores resultados y además es independiente de corpus, el IIC de un nodo n_i :

$$IIC(n) = 1 - \frac{\log(hypo(n_i) + 1}{\log(|C|)}$$

donde:

$hypo(n_i)$ es el número de hipónimos de n_i y
$|C|$ es el número de nodos de la taxonomía.

Para contar adecuadamente el número de hipónimos se debe de romper ciclos del GCW pero sin desconectar nodos de un componente conexo.

Una vez que se realizaron los cálculos y adecuaciones correspondientes de acuerdo a la información que se ha descrito, los resultados se comparan con un estándar de oro, en este caso con tres conjuntos de datos en idioma alemán y posteriormente con el momento de producto de correlación de Pearson *r* para comparar con las valoraciones que se hacen de manera manual. Los resultados obtenidos demuestran que las medidas de proximidad semántica fueron exitosas al ser aplicadas al GCW.

Grafo de artículos de Wikipedia – GAW (Wikipedia Article Graph)

En Wikipedia los artículos están estrechamente relacionados por medio de la estructura de vínculos, dado que los vínculos pueden ser insertado mientras se edita un artículo. El grafo de artículos se concibe como un grafo dirigido, como el que se muestra en la parte derecha de la Figura 4. Cada artículo es un nodo y cada vínculo entre artículos un arco que va de un nodo hacia otro (Zesch2007). Los vínculos entre artículos se establecen con cualquier tipo de relación entre estos.

Los hipervínculos que apuntan de un artículo hacia otro pueden ser tratados como vínculos dirigidos, mientras que los artículos representan los nodos de una red (Zlatic2006), un grafo resultado de la estructura hiper vinculada de los artículos de Wikipedia (Buriol2006) que es un tipo de grafo de Web.

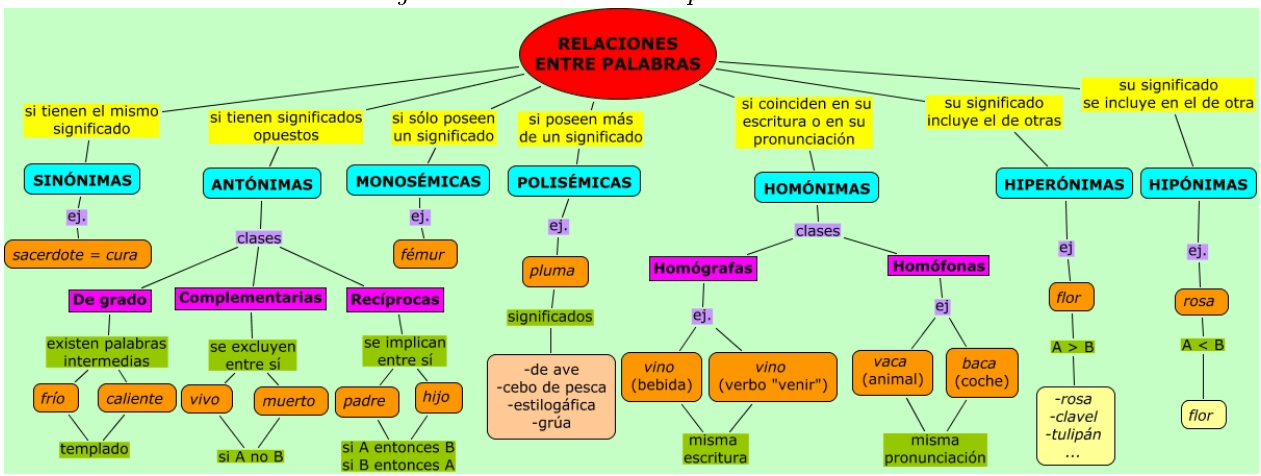
Los grafos de Web se han estudiado en base a sus propiedades topológicas. El estado del arte respecto al GAW parece indicar que se ha realizado poca investigación sobre su evolución estadística y propiedades topológicas. Justamente es esta una de las características que representa un área de oportunidad para el presente trabajo de tesis.

Relaciones entre palabras

Una manera de entender los recursos léxico-semánticos es estudiar como se relacionan los significados de las palabras. Las palabras pueden organizarse en una jerárquía léxica, como en el caso de WordNet, en la que las palabras se organizan jerárquicamente. Cada nodo consiste de un synset de palabras con idénticos (o casi idénticos) significados (ManningFoundations). Existen además otro tipo de relaciones entre palabras como la meronimia o relaciones *parte-todo,* en la Figura 5 se pueden ver otro tipo de estas relaciones.

La hiponimia y hiperonimiason también relaciones entre palabras. La hiperonima es una palabra con un sentido más general. La hiponima es una palabra con un significado más especializado. En general, si w_1 es una hiperonima de w_2, entonces w_2 es un hiponima de w_1.

Figura 5: Relaciones entre palabras.



WordNet

WordNet es un sistema de referencia léxica, en forma de diccionario electrónico en inglés que fue desarrollado en la universidad de Princeton. Este recurso combina muchas características usadas para desambiguación del sentido de las palabras en un solo sistema.

WordNet, incluye definiciones de sentidos de palabras como un diccionario, define synsets o conjuntos de sinónimos, los cuales representan un concepto léxico, y además proporciona las relaciones jerárquicas existentes entre palabras.

Los sentidos de WordNet comprenden un conjunto de sinónimos y definiciones al igual que un diccionario, las cuales son llamadas glosas. El número que se encuentra al inicio de la definición de algunos sentidos, es la frecuencia de los valores obtenidos del corpus SemCor. A diferencia de un diccionario, WordNet contiene un conjunto de relaciones léxicas entre sysnsets o lemas, los cuales aparecen al inicio de la glosa (Torres_Ramos2009).

WordNet puede descargarse libremente de Internet y ha sido ampliamente utilizada en aplicaciones de PLN.

Wikipedia como red

Wikipedia comparte características de las redes léxico-semánticas, las cuales estructuran el vocabulario en función de las relaciones semánticas entre palabras, como sinonimia, antonimia, hiponimia, hiperonimia o meronimia (ie. WordNet)[14].

Cada Wikipedia (por cada idioma se le considera una diferente Wikipedia) puede considerarse como una red en la que los artículos son los nodos y los hipervínculos entre artículos son enlaces directos entre ellos (Zlatic2006).

Ambigüedad

[14] Tecnologías lingüísticas: los recursos lingüísticos http://liceu.uab.cat/~joaquim/language_technology/HLT/tecnol_ling_recursos.html.

Una de las tareas del PLN es la resolución de la ambigüedad sobre como deben ser interpretadas las palabras, esto debido a que una palabra puede tener más de un significado o sentido (polisemia).

La ambigüedad, en el proceso lingüístico, se presenta cuando pueden admitirse distintas interpretaciones a partir de una representación dada o cuando existe confusión al tener diversas estructuras y no tener los elementos necesarios para eliminar las eventualmente incorrectas (Torres_Ramos2006). Para desambiguar, es decir, para seleccionar los significados o las estructuras más adecuados de un conjunto distinto de posibilidades, se requieren de diversas estrategias de solución en cada caso (Galicia_Haro2007).

Se distinguen tres tipos principales de ambigüedad: léxica, semántica y sintáctica o estructural.

- *Ambigüedad léxica*. Se presenta cuando las palabras pueden pertenecer a diferentes categorías gramaticales, por ejemplo *bajo* puede ser una preposición, un sustantivo, un adjetivo o una conjugación del verbo bajar.
- *Ambigüedad semántica*. Se presenta cuando las palabras tienen múltiples significados, por ejemplo la palabra *banco* puede significar banco de peces, banco para tomar asiento o institución financiera.
- A*mbigüedad sintáctica*. También conocida como ambigüedad estructural se presenta cuando una oración puede tener más de una estructura sintáctica. Por ejemplo, hablando de una pintura de arte en la oración "Este trabajo no tiene título" (ManningFoundations) se pueden entender dos cosas diferentes: a) el trabajo no recibió por parte de su autor un nombre con el cuál pueda denotarse, o bien, b) la pintura no tiene una placa en la cuál pueda leerse el título de la obra.

Desambiguación del sentido de las palabras

La desambiguación del sentido de las palabras es una fase necesaria para la consecución de tareas de PLN como lo son el análisis sintáctico o la interpretación semántica. La desambiguación consiste en asignar automáticamente el sentido adecuado de una palabra en polisémica en un texto con relación al contexto en el que se utiliza, las palabras tienen un número finito discreto de sentidos, normalmente escritos en un diccionario, tesauro u alguna otra fuente de referencia (ManningFoundations) (Mihalcea2007).

La desambiguación es considerada una tarea de clasificación: los sentidos de la palabra son las clases, el contexto provee la evidencia y cada ocurrencia de una palabra es asignada a una o más de las posibles clases en base a la evidencia. Si las decisiones de un sistema dependen del significado del texto, entonces la desambiguación es necesaria.

En un sistema de recuperación de información, cuando se realiza una consulta (query) sobre la palabra *carácter* deberá devolver como resultado documentos que traten sobre alguna de las acepciones de esta palabra, como pueden ser:

- Señal o marca que se imprime, pinta o esculpe en algo.
- Conjunto de cualidades o circunstancias propias de una cosa, de una persona o de una colectividad, que las distingue, por su modo de ser u obrar, de las demás. *El carácter español. El carácter insufrible de Fulano.*
- Condición dada a alguien o a algo por la dignidad que sustenta o la función que desempeña. *El carácter de juez, de padre. Medidas de carácter transitorio.*
- Fuerza y elevación de ánimo natural de alguien, firmeza, energía. *Un hombre de carácter.*

- Modo de decir, o estilo.

La desambiguación de sentidos de las palabras se ha estudiado con métodos estadísticos, métodos basados en conocimiento (o basados en diccionarios) y con métodos mixtos (Galicia_Haro2007).

Métodos basados en conocimiento

Los diccionarios, tesauros y bases léxicas de conocimiento, textos sin ningún tipo de etiquetado e incluso recursos de la Web que no utilizan un corpus como evidencia, son conocidos como métodos basados en conocimiento (WSD book) y están basados en heurísticas y análisis del contexto en donde se encuentra una ambigüedad.

Estos métodos utilizan el conocimiento lingüístico previamente adquirido. La idea básica consiste en utilizar recursos lingüísticos externos para desambiguar las palabras. Los recursos que comúnmente son utilizados por estos métodos son los diccionarios MRD (Machine Readable Dictionaries), como son:

- Longman Dictionary of Contemporary English (LDOCE)
- Collins English Dictionary (CED) (http://www.collinslanguage.com/)

Desambiguación supervisada

Utiliza datos de entrenamiento, anotaciones manuales, extracción de características para ser utilizada por los clasificadores.

Redes Semánticas

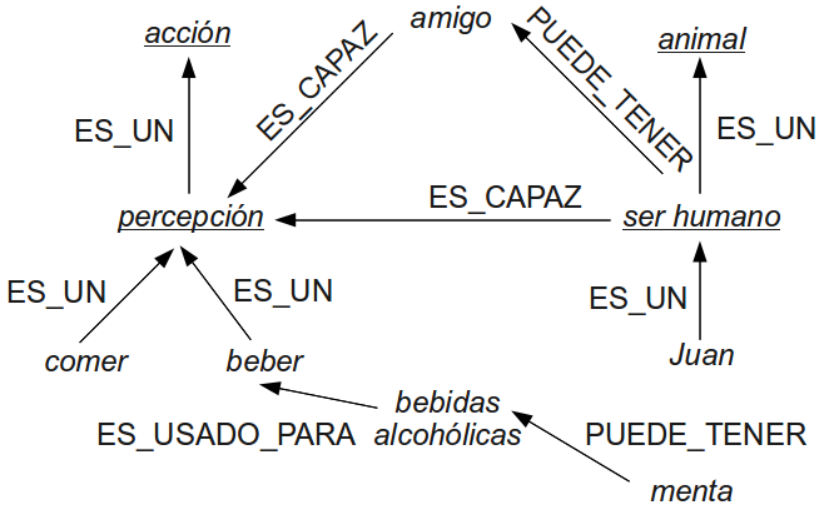
Se hace uso también de estructuras como las redes semánticas, en las que la clase semántica ayuda a resolver una ambigüedad, es decir, se analiza con que parte de la frase están enlazadas otras frases. La red semántica surge bajo la idea de que los conceptos están entrelazados formando una red, cada concepto constituye un nodo de la red que se conecta con otros nodos mediante enlaces de distinta naturaleza. Los enlaces establecen un tipo de relación:

- enlaces de pertenencia a una clase ("es un tipo de"),
- enlaces de meronimia ("es una parte de"),
- enlaces de sinomimia ("es igual que"),
- enlaces de función ("tiene la función de"),
- enlaces de contención ("contiene un"), por mencionar los más comunes.

La red semántica es un conjunto de relaciones entre pares de palabras, o una combinación de palabras, que se refieren a una cosa específica o idea.

Una red semántica es un grafo, que representa cadenas de relaciones. Los elementos semánticos se representan por nodos. En este grafo se representan trayectorias que se trazas siguiendo las relaciones de una palabra a otra, los elementos con alguna relación semántica se dibujan por medio de líneas o flechas conocidas como aristas. De esta forma se pueden medir que tan cercanos o lejanos se encuentra los pares de palabras en la red.

Figura 6: Red semántica para la frase Juan bebe bebidas alcohólicas con sus amigos.



Desambiguación del sentido de las palabras utilizando la estructura de vínculos de Wikipedia (Mihalcea2007)

Utilizando los hipervínculos de Wikipedia se puede crear un corpora de sentidos etiquetado que puede ser utilizado para construir clasificadores de sentidos exactos y robustos. Por medio de experimentos de desambiguación de sentidos de las palabras para el corpus Wikipedia de sentidos etiquetado generado para un subconjunto de palabras ambiguas de SENSEVAL[15], se demuestra que las anotaciones de Wikipedia son fiables y la calidad del clasificador de etiquetación de sentidos construido sobre estos datos excede en mucho la precisión de una base inicial que selecciona el sentido de la palabra más frecuente por default.

Wikificación de Textos, Wikify!

Su propósito es añadir información adicional a un texto, mediante la obtención de sus palabras relevantes, para asociarles enlaces o hipervínculos a artículos de Wikipedia (Mihalcea2007).

La Wikificación de textos se realiza en base a dos tareas:

1. Extracción de texto.
2. Desambiguación del sentido de las palabras.

De los textos se extraen las palabras más relevantes, como pueden términos técnicos, entidades con nombre, nueva terminología y aquellos que tienen relación estrecha con el texto.

[15] El propósito de SENSEVAL es realizar ejercicios de evaluación para el análisis semántico de textos de las fortalezas y debilidades de los programas útiles para determinar automáticamente el sentido de las palabras en un contexto.

La extracción de textos utilizó un vocabulario controlado de frases clave de los títulos de los artículos de Wikipedia y de las "formas superficie" de los artículos de Wikipedia. Se utilizaron tres algoritmos no supervisados para la extracción de palabras clave: *tf.idf*, $\chi 2$ y el método de Clasificación (Ranking) *keyphraseness*.

Posteriormente para elegir la página que se ha de asociar a las palabras relevantes y a una página Wikipedia, se requiere de un proceso de desambiguación, el cual se debe dar de acuerdo al contexto en el que se encuentra cada palabra.

Agregar la imagen de la arquitectura del sistema Wikify!

Para la evaluación de los resultados obtenidos de los algoritmos de extracción, se creo un estándar de oro con un conjunto de 85 documentos y 7,286 conceptos vinculados anotados de forma manualmente, contra el cual se compararon los términos automáticamente extraídos. El desempeño se evaluó en términos de precisión, memoria (recall) y medida-F (F-measure).

En cuanto a la desambiguación de palabras se utilizaron dos métodos, uno del tipo del algoritmo de Lesk, que identifica el significado más parecido de una palabra en un determinado contexto basándose en una medida de traslape contextual entre las definiciones del diccionario de la palabra ambigua. El segundo método es de datos dirigidos, que incorpora características locales y de interés actual en un clasificador Naive Bayes.

Para evaluar esta fase se realizó un estándar de oro de un conjunto de datos de la la misma colección de 85 páginas de Wikipedia y 7,286 conceptos vinculados, con anotaciones manuales de sentidos correspondientes a los vínculos.

Identificación de temas en base a un algoritmo de centralidad de un grafo ensayado

Utiliza el método no supervisado Wikify! (MihalceaCsomai2007) para encontrar temas o categorías relevantes en un documento utilizando un algoritmo de centralidad en el grafo ensayado construido de Wikipedia.

Se toma como premisa que el conocimiento enciclopédico externo puede utilizarse para identificar temas relevantes de un documento (Mihalcea2009). Se realizan dos tareas:

(1) Se construye un grafo que incluye toda la información de la Wikipedia en inglés, los nodos equivalen a las categorías y artículos, los arcos representan las relaciones de proximidad entre artículos (5.8 milliones de nodos y 65.5 millones de arcos).

(2) Por cada documento, se identifica el concepto enciclopédico en el texto y se crea un vínculo entre el contenido del artículo y el grafo enciclopédico externo. Se hace entonces una clasificación utilizando una variación del algoritmo de centralidad para grafos ensayados PageRank[16] que

> Next, we run a biased graph centrality algorithm on the entire graph, so that all the nodes in the external knowledge repository are ranked based on their relevance to the input document. We use a variation of the PageRank (Brin and Page, 1998) algorithm, which accounts for both the relation between the nodes in the document and the encyclopedic graph, as well as the relation between the nodes in the encyclopedic graph itself.

2.3.1 Wikipedia como base de conocimiento

Desde el punto de vista computacional, Wikipedia constituye una base de conocimiento. Las bases de conocimiento para PLN deben reunir algunas características como: ser independientes del dominio, actualizadas frecuentemente, multilingües, las cuáles posee la Wikipedia, lo que la convierte en un recurso importante en aplicaciones de PLN.

Como ya se ha mencionado, el sistema de categorías de Wikipedia es considerado una taxonomía al igual que la de WordNet, esto ha permitido que ambas sean explotadas para calcular medidas sobre proximidad semántica y similitud semántica, con el propósito de permitir a las computadoras razonar sobre un texto escrito [10] y derivar conocimiento.

Extracción de conocimiento léxico-semántico

Así como Wikipedia, también Wiktionary[17] es considerada una base de conocimiento, ambas se han construido por medio de la colaboración de usuarios. Se les denomina Bases de Conocimiento Colaborativo, a diferencia de lo que se conoce como Bases de Conocimiento Lingüístico (WordNet es un ejemplo de este tipo).

Wiktionary es la parte léxica de Wikipedia, y las entradas añadidas incluyen información léxico-semántico como parte-del-discurso, sentido de las palabras, glosas, etimología, pronunciación, ejemplos, traducción, colocación, términos derivados. También se incluye sinónimos, antónimos, hiperónimos e hipónimos. A diferencia de las bases de conocimiento lingüístico incluye información como abreviaciones, acrónimos, pronunciación correcta, contracciones, proverbios, onomatopeyas, jerga coloquial, entre otras.

[16] Algoritmo de análisis de vínculos que asigna un valor numérico (peso) a cada elemento de un conjunto de documentos vinculados (hipervínculos) y de este modo determinar la relevancia relativa de cada documento del conjunto. Google tiene la marca registrada "PageRank" aunque la patente del algoritmo es propiedad de la Universidad de Stanford.

[17] http://www.wiktionary.org/

Relaciones Semánticas en Wikipedia

Como se ha explicado, Wikipedia se compone de artículos, vinculados a otros artículos y organizados mayormente por medio de Categorías. De estas estructuras se ha observado que, además de las relaciones que se establecen solo por estas estructuras, se pueden encontrar relaciones semánticas.

En algunos trabajos se buscan relaciones semánticas fuertes (Chernov et al., 2006), en la que por medio de un esquema de base de datos se establecen relaciones que se consideran importantes entre las categorías.

Proximidad semántica

Este modelo está relacionado con el conocimiento semántico. Se requiere para desambiguar oraciones completas, porque sus diversas estructuras sintácticas son perfectamente posibles, o para enlazar frases circunstanciales que al no estar directamente enlazados con el sentido del lexema rector requieren un método conectado con la semántica de contexto.

El cálculo de proximidad semántica determina que tan relacionados están dos conceptos en una taxonomía por medio de las relaciones que existan entre ellos. Es una tarea útil en aplicaciones del PLN como recuperación de información, corrección ortográfica y desambiguación del sentido de las palabras.

Para realizar estos cálculos se analizan las relaciones entre palabras, las más comunes son:

- antonimia
- hiponimia/hiperonimia
- meronimia
- relaciones funcionales como parte-de, hecho-de, es-un-atributo-de, etc..

Es usual que las medidas de proximidad semántica sean evaluadas, esto es, se compara los resultados obtenidos con un "estándar de oro" producto del juicio humano, es decir, personas que en base a su juicio determinan la proximidad de pares de palabras.

Estándares de oro para evaluación de Proximidad Semántica y Similitud Semántica

Uno de las métodos de evaluar las medidas de proximidad semántica , es comparándolas con juicios humanos, es decir, personas que realizan un juicio sobre el grado de relación entre pares de palabras, de lo cuál se obtiene un estándar de oro . Sobre las investigaciones de proximidad semántica y similitud semántica algunos de los estándares de oro utilizados son:

Colección WordSimilarity-353

Contiene dos conjuntos de pares de palabras con las anotaciones humanas sobre similitud. Uno de los conjuntos contiene 153 pares de palabras con su valor de similitud asignado para 13 materias. El segundo conjunto contiene 200 pares de palabras, con la similitud resultado de la valoración de 16 materias. Un tercer conjunto se conforma por los dos conjuntos anteriores, el cual se integra por 353 palabras con los valores de similitud.

Los conjuntos pueden descargarse libremente[18] y están disponibles en dos tipos de formatos:

[18] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

- Valores separados por comas (Comma-separated values- CSV).
- Delimitados por tabuladores (Tab-delimited – TAB).

RG (Rubenstein & Goodenough)

Contiene 65 pares de sustantivos, los cuales fueron recopilados en 1965, escritos en tarjetas de papel que tenían que ser ordenadas de acuerdo a la similitud de su significado, es decir, basado en sinonimia en vez de proximidad y calificados en una escala de 0 a 4 por 51 personas [24].

MC (Miller & Charles)

Subconjunto de 30 pares de RG (Rubenstein & Goodenough) con 38 temas de prueba.

R (Resnick)

Replicó las clasificaciones del conjunto MC utilizando a 10 personas para la calificación.

Wikirelate!

Considera el árbol de categorías de Wikipedia como una Folksonomía resultante de la clasificación colaborativa, en la que los usuarios pueden categorizar el contenido de las entradas enciclopédicas. Las categorías constituyen una red semántica en la que los conceptos se relacionan. Esta red de categorías se utiliza para calcular la proximidad semántica.

Aplica a la Wikipedia las medidas para cálculo de proximidad semántica desarrolladas originalmente para WordNet:

1. Medidas basadas en recorridos.
2. Medidas basadas en contenido de información.
3. Medidas basadas en traslape de textos.

El proceso de minería de este modelo consiste en tomar pares de palabras i, j, recuperar las páginas de Wikipedia a las que estas se refieren, posteriormente se asocian al árbol de categorías extrayendo las categorías a las que las páginas recuperadas pertenecen para finalmente calcular la proximidad semántica de acuerdo a las páginas extraídas y las rutas encontradas sobre la taxonomía de categorías.

Modelo de Vector de Vínculos de Wikipedia (Wikipedia Link Vector Model – WLVM)

Al identificar la proximidad semántica de términos y conceptos en Wikipedia se puede extraer un tesauro como mapa de relaciones semánticas entre palabras y frases para recuperación de información.

El modelo WLVM utiliza la estructura de hipervínculos de Wikipedia en vez de utilizar contenidos textuales para el cálculo de medidas de proximidad semántica. El proceso de extracción en este modelo consiste en tomar pares de términos de la estructura de hipervínculos de Wikipedia. El procedimiento es extraer todos los artículos relacionados a cada término, para lo cuál se realiza el listado de todas las páginas cuyos títulos coincidan con el término y procesarlos de modo que:

- Los artículos se usen directamente.

- Se siguen los vínculos de redirección, de modo que se utilizan los artículos correspondientes.
- Se procesan las páginas de desambiguación, de modo que cada artículo al que vinculan sea utilizado.

Lo siguiente es obtener la similitud entre los términos juzgando la similitud de las páginas obtenidas previamente. La similitud semántica entre dos artículos de Wikipedia se define por el ángulo formado por los vectores de los vínculos encontrados entre ellos. Este enfoque es similar al modelo de espacio de vectoresutilizado en recuperación de información. Particularmente en el modelo WLVM los vectores no se construyen por medidas de probabilidad como TF-IDF sino usando el peso de los valores de los vínculos que se obtiene por la probabilidad de ocurrencia de cada vínculo definida por el número total de vínculos al artículo sobre el número total de artículos.

La proximidad semántica más alta para los artículos está dada por el ángulo más pequeño entre vectores, que va desde los 0° si los artículos contienen una lista idéntica de vínculos, a 90° si no existe traslape entre ellos. Así se da la desambiguación de artículos, de modo que solo los dos artículos que están más relacionados son los utilizados como medida de similitud. El método para evaluar estas medidas fue comparándolas con WordSimilarity-353[19].

Otra medida de proximidad semántica evaluada con este modelo se obtuvo al sumar los pesos de los vínculos compartidos con mejores resultados sobre proximidad semántica entre términos, mientras que el método WLVM aplicado tal cual, resultó tener mejores resultados de proximidad semántica entre artículos (Milne2007).

Para desarrollar este modelo se utilizó la herramienta Wikipedia Miner desarrollada en específico para la exploración de la estructura de hipervínculos de Wikipedia.
Modelo de Medidas basadas en Vínculos de Wikipedia (Wikipedia Link-based Measure - WLM)

Cálculo de medidas de proximidad entre artículos (Milne_2008):

1. basada en los vínculos que se extienden hacia afuera de los artículos, esto se hace calculando el ángulo entre vectores de los vínculos encontrados en dos artículos, esto es similar a los que se hace con los vectores TF-IDF utilizados en los modelos de RI. La diferencia radica en que se utiliza el valor del cómputo de pesos de vínculos, en vez del cómputo de los pesos de la ocurrencia de términos. Esta probabilidad se define por el número total de vínculos al artículo objetivo sobre el total del número de artículos. Esto es, si s y t son el artículo fuente y el artículo objetivo respectivamente, entonces el peso w del vínculo s $\rightarrow$ t está definido por:

$$w(s \rightarrow t) = \log(\frac{W}{T}) sis \in T, 0 deotromodo$$

donde T es el conjunto de todos los artículos que se vinculan a t y W es el conjunto de todos los artículos de Wikipedia. Los pesos de los vínculos se utilizan para generar los vectores correspondientes, la similitud de los artículos es el ángulo entre vectores (similitud por cosenos). El rango va de los 0° si el artículo contiene una lista idéntica de vínculos a 90° si no existe traslape entre ellos.

2. La segunda métrica que utiliza el modelo está basada en las ocurrencias de un término en páginas web, tal como sería el resultado de búsquedas en el motor de Google que obtiene páginas donde ocurren los términos, pero en vez de utilizar los resultados de búsqueda de Google, esta métrica se basa en la estructura de vínculos de Wikipedia.

$$sr(a,b) = \frac{\log(max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(min(|A|,|B|))}$$

[19] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html

donde a y b son dos artículos, A y B son conjuntos de todos los artículos vinculados a a y b respectivamente y W es la totalidad de Wikipedia.

Análisis Semántico Explícito (Explicit Semantic Analysis – ESA)

Propone un método que representa el significado de cualquier texto en términos de conceptos de Wikipedia. A diferencia de los métodos de proximidad semántica basados en WordNet que se limitan a proximidad de palabras, este método calcula la proximidad de textos arbitrarios.

Utiliza técnicas de máquinas de aprendizaje (machine learning) para representar el significado de los términos(textos) como vectores de los conceptos de Wikipedia, llamados vectores de interpretación, los cuales tienen asignados pesos utilizando TF-IDF.

Se evalúan los resultados calculando automáticamente el grado de proximidad semántica entre fragmentos de textos de lenguaje natural, comparando los vectores usando la métrica de cosenos.

En correlación con métodos de proximidad semántica humanos se observan mejoras de r =0.56 a 0.75 para palabras, y de r =0.60 a 0.72 para textos.

Wikipedia como red semántica

Para calcular la proximidad semántica entre palabras, se puede utilizar el sistema de categorización de Wikipedia como una red semántica (Ponzetto&Strube).
Wikipedia como *corpus* de entrenamiento

Wikipedia ha sido utilizada como *corpus* de entrenamiento para categorización de temas de recursos de aprendizaje (Meyer2007), se utiliza el método de comparación de proximidad de vecinos, (k-Nearest-Neighbors)para categorizar los recursos de aprendizaje con artículos de Wikipedia.

Los artículos se transforman a un vector de palabras, los recursos son mapeados a vectores de palabras para compararlos. Los vectores similares se considera que cubren temas similares. La categorización de los recursos de aprendizaje, en este caso, de objetos de aprendizaje, se hace tomando el sistema de categorías de Wikipedia como base.

El procesamiento de todas las categorías de Wikipedia, como muchos de los procesos que explotan su información, consume muchos recursos de cómputo, como lo son la capacidad de memoria y la complejidad de calculo, lo que se convierte en factores limitantes.

Wikipedia como corpus XML [20]

Wikipedia también tiene una estructura de archivos XML. Se tiene noción de la utilización de los archivos XML de la Wikipedia como corpus, integrado por documentos XML de ocho colecciones en idiomas diferentes: inglés, francés, alemán, español, chino, árabe y japonés.
Los textos están codificados en UTF-8 resultado de los textos originales de Wikipedia, del corpus se han eliminado textos de tipo plantillas y de comentarios (Denoyer2006). Este recurso se ha empleado para tareas de recuperación de información, como son recuperación ad-hoc, categorización, clustering y tareas de mapeo de estructuras.

[20] http://www-connex.lip6.fr/~denoyer/wikipediaXML/

Crear un corpus de texto plano de Wikipedia

Las colecciones de texto de Wikipedia pueden convertirse en corpus[21][22], como en el caso de los artículos Wikipedia, archivos con formato y etiquetado MediaWiki[23] y que pueden obtenerse por idioma, descargando el respaldo de la base de datos (dump databases)[24] correspondiente.

De manera general el procedimiento consiste en descargar el respaldo conveniente, los cuales son documentos de tamaño considerable y por estas características están en formato comprimido. Los respaldos se descomprimen y el contenido debe ser analizados sintácticamente (*parser)* para cambiar la sintaxis propia de marcas de los artículos Wiki y obtener otro tipo de archivo (i.e. XML o texto plano).

Ejemplos del procedimiento para obtener un corpus de este tipo pueden encontrarse en internet, el respaldo de la base de datos (Dump) del idioma inglés es el más explotado, pero pueden encontrarse inclusive en idioma noruego[25].

La tarea más complicada de este proceso es la del análisis sintáctico, sin embargo, existen varios proyectos de parsers[26] alternativos al de MediaWiki además del propio.

Wikicorpus (Reese2010)

Se trata de otro proyecto de corpus producto de porciones de Wikipedia y enriquecido con información lingüística, trilingüe (catalán, español, inglés) con alrededor de 750 millones de palabras. Es un corpora anotado, con lemas y categorías gramaticales (part-of-speech) usando la libreria de código abierto FreeLing[27] que es una librería orientada al desarrollo que proporciona servicios de análisis del lenguaje. También está anotado con sentidos por medio del Estado del Arte del algoritmo de Desambiguación de Sentidos de las Palabras UKB que asigna anotaciones de WordNet. El algoritmo UKB[28] se basa en una red de relaciones semánticas para eliminar la ambigüedad de los sentidos más probables de las palabras en un texto utilizando el algoritmo de Clasificación de Páginas (PageRank).

El proyecto Wikicorpus tiene dos propósitos principales:

1. Poner a disposición un corpus enriquecido con información lingüística,
2. Explorar la inducción de recursos léxico semánticos multilingües.

Como parte del proceso de obtención del corpus se desarrollaron dos productos disponibles para PLN: un parser de páginas de Wikipedia desarrollado en Java y la integración de un algoritmo de desambiguación semántica al FreeLing.

El corpus se obtuvo mediante tres subprocesos:

1. Filtrado, mediante el cual se excluyeron páginas sin categoría, como son las páginas de redirección.

[21] Extracting Text from Wikipedia, http://evanjones.ca/software/wikipedia2text.html.
[22] Generating a Plain Text Corpus from Wikipedia, http://blog.afterthedeadline.com/2009/12/04/generating-a-plain-text-corpus-from-wikipedia/.
[23] http://en.wikipedia.org/wiki/MediaWiki
[24] http://download.wikimedia.org/backup-index.html
[25] https://www.hf.ntnu.no/hf/isk/Ansatte/petter.haugereid/cl/wiki-corpus.html.
[26] http://www.mediawiki.org/wiki/Alternative_parsers
[27] FreeLing: An Open Source Suite of Language Analyzers, http://nlp.lsi.upc.edu/freeling/
[28] UKB: Graph Based Word Sense Disambiguation and Similarity, http://ixa2.si.ehu.es/ukb/

2. Extracción de texto, utilizando el parser JavaCC que se desarrolló específicamente para la obtención del corpus.

3. Procesamiento lingüístico de los textos utilizando FreeLing para lo cual fueron tokenizados, lematizados y etiquetados con partes del discurso. Además se integró el algoritmo para desambiguación semántica basado en grafos UKB.

Wikitology: Wikipedia como una ontología.

El proyecto Wikitology (Syed2008) propone un sistema para identificar temas y conceptos asociados a un conjunto de documentos, esto es, predecir conceptos comunes a ese conjunto utilizando la Wikipedia (versión en inglés) como una ontología, en la que sus artículos y categorías representan conceptos.

Explota la información semántica contenida en sus páginas:

1. vínculos desde y hacia otros artículos de Wikipedia,
2. vínculos a páginas de desambiguación,
3. vínculos de redireccionamiento,
4. vínculos desde y hacía páginas web externas,
5. valores calculados PageRank calculado por motores de búsqueda como Google,
6. historial de páginas que indican cuando y qué tan frecuente han sido editadas.

También es una fortaleza el que sus páginas están ligadas a ontologías formales como DBPedia y Semantic MediaWiki, así como en el sistema comercial Freebase.

Este sistema utiliza los artículos de Wikipedia y los grafos de categorías y de vínculos a los artículos para la predicción de conceptos. El grafo de categorías se utiliza para la predicción de conceptos generalizados y el grafo de vínculos a los artículos para predicción de conceptos específicos que no se encuentran en la jerarquía de categoría.

APIs para Wikipedia

Explorar el árbol de categorías

La herramienta Extension:CategoryTree[29]integrada al software de MediaWiki, permite explorar dinámicamente el árbol de categorías, utiliza AJAX para cargar las porciones del árbol que se están explorando.

[29] http://www.mediawiki.org/wiki/Extension:CategoryTree

Calcular la proximidad de palabras

Para calcular proximidad semántica se transforman recursos léxicos en una red o grafo y se calcula la proximidad utilizando rutas o caminos o la longitud resultante del número de nodos entre los términos de una jerarquía.

Esta API calcula la proximidad semántica de la siguiente manera:

1. Se parte de un par de palabras como entrada,
2. Se recuperan los artículos a los que hacen referencia (utilizando estrategias de desambiguación basadas en la estructura de vínculos de los artículos)
3. Se calculan las rutas o caminos en el grafo de categorización entre las categorías a las que están asignados los artículos,
4. Se entregan los resultados de las rutas o caminos encontrados,
5. Se califica de acuerdo a determinadas medidas definidas.

La implementación toma como base teórica cálculo de longitud de caminos o rutas (Syed2008), información de contenidos (information-content) (Zesch2008) y medidas de superposición de texto (text-overlap) (Milne_2008).

Esta API implementa las clases para realizar consultas a la Wikipedia para recuperar entradas de esta enciclopedia así como determinar el resultado de pares de palabras.

Sistema para indexar textos Wiki

El concepto de Wiki se refiere a un sitio web colaborativo el cual permite la edición páginas web ligadas utilizando un lenguaje simplificado de marcas o un editor de tipo **WYSIWYG** (***W****hat* ***Y****ou* ***S****ee* ***I****s* ***W****hat* ***Y****ou* ***G****et)*. Es erróneo referirse a la Wikipedia como Wiki, ya que Wikipedia es un proyecto de textos o páginas Wiki. Los datos contenidos en Wikipedia están divididos en textos y vínculos, los cuales pueden ser[7]:

- internos, que son vínculos que ligan páginas dentro de un mismo sitio;
- externos, que son vínculos hacia sitios externos;
- inter wikis, especifícan el artículo que describe un término de la enciclopedia pero en otro lenguaje;
- categorías, clasifican los artículos temáticamente.

Esta organización permite distinguir tres tipos de algoritmos de búsqueda:
Convertir textos Wiki en textos de Lenguaje Natural mediante el uso de expresiones regulares realizando dos tareas: eliminación y transformación del texto:

Paso 1, eliminación de etiquetas
Paso 2, transformación de etiquetas Wiki

Con las Wikipedias rusa y en ingés se implementó un sistema para indexar textos Wiki (Krizhanovsky2009)conocida como WikIDF. Para indexar la lista de lemas y la frecuencia de su ocurrencia se utilizó el sistema GATE, también se utilizó el analizador morfológico Lemmatizer, y el módulo Russian-POStagger. Al utilizar WikIDF, se diseñaron los índices DB para ambas Wikipedias.

Herramienta de minería de Wikipedia (Wikipedia Miner Toolkit)

Herramienta de código abierto[30]y de acceso orientado a objetos que permite consultar y desplegar información de la estructura y contenido de Wikipedia. Con esta herramienta es posible comparar semánticamente términos y conceptos y detectar temas cuando son mencionados en documentos.

Por medio del procesamiento de los respaldos (dumps) de la Wikipedia se pueden extraer sumarios del gráfico de hipervínculos y la jerarquía de categorías.

La herramienta cuenta con scripts en PERL para la extracción de sumarios, se puede comunicar con la base de datos MySQL para indexar de manera persistente la información que se extrae, también incluye una API desarrollada en Java que simplifica los datos para su acceso.

Permite el cálculo de proximidad semántica utilizando la estructura de hipervínculos de Wikipedia (Wikipedia Link-based Measure - WLM) , con la que se obtiene resultados más precisos y menos costosos que ESA, lo cuál atribuye a que la estructura de hipervínculos es la conexión definida manualmente de dos conceptos desambiguados manualmente [10], es decir, el método está más cercano a la semántica definida manualmente en Wikipedia.

Librería Wikipedia basada en Java (JWPL - Java-based Wikipedia Library) y Librería Wiktionary basada en Java (JWKLT – Java-based Wiktionary Library)

Con respaldos (dumps) de Wikipedia (Zesch2008)se importa a una base de datos, para explotar el indexado dela base de datos que garantiza casi tiempo constante de recuperación por cada artículo. Se utilizó el framework objeto-relacional de Hibernate. La interface de programación orientada a objetos está centrada en los objetos: Wikipedia, Page y Category:

Objeto	Operaciones
Wikipedia	• Establecer la conexión con la base de datos. • Iterar sobre los artículos, categorías, páginas de redirección y desambiguación.
Page	• Representar artículos normales, páginas de redirección y desambiguación. • Proporcionar acceso al texto de un artículo (con información de marcas o texto plano), las categorías asignadas, los vínculos internos y externos del artículo, vínculos de redirección.

[30]http://wikipedia-miner.sourceforge.net/

Category	• Acceder a los artículos de una categoría. • Las categorías al estar conformadas como un tesauro, este objeto proporciona los métodos para recuperar categorías padre e hijas. • Objeto CategoryGraph que entre otras operaciones, permite encontrar la ruta más corta entre dos categorías.

Con respaldos (dumps) de Wiktionary en formato XML en diferentes lenguajes se hace el parseo utilizando la librería de base de datos Berkley DB. Para cada entrada del Wiktionary, la API regresa un objeto Java Wiktionary para hacer consultas a la base de datos sobre información sobre alguna palabra utilizando el grafema como argumento de la consulta. Adicionalmente puede especificarse parte-del-discurso o el lenguaje de la palabra.

Estas librerías han sido utilizadas en investigaciones de PLN como son:

- Analizar y acceder a las estructura de grafo de categorías de Wikipedia (Zesch2007).
- Cálculo de proximidad semántica entre palabras .
- Recuperación de información semántica.

WikiLibros

Este proyecto (WikiProject) tiene como objetivo que los usuarios escriban *Libros-Wikipedia* (*Wikipedia-Books*), en español el proyecto se conoce como *WikiLibros*. La idea del proyecto es permitir a los uaurios crear libros de acuerdo a sus especificaciones utilizando el amplio banco de contenidos libres de Wikipedia y proporcionar las reglas para utilizar la página Wikipedia:Books, sus subpáginas[31] y las herramientas (Wikipedia-Books tools) que permiten crear un WikiLibro.

Un Libro-Wikipedia es una colección de artículos Wikipedia que pueden ser salvados, presentados vía electrónica en formato portable PDF (portable document format) o en formato abierto ODF (Open-Document format) u ordenado como un libro impreso.

La empresa encargada de las ediciones impresas es PediaPress (http://pediapress.com/) que se pueden obtener dentro del mismo sitio web de Wikipedia en la sección de imprimir/exportar. El servicio tiene un costo que varia de acuerdo al número de páginas.

Herramienta para generar Libros (Book Tool)[32]

Esta herramienta permite a los usuarios organizar la selección de páginas que realicen en forma de libro que puede ser:

- editado y estructurado en capítulos,
- compartido y cargado,
- presentado como documento ODF,

[31] Wikipedia:WikiProject Wikipedia-Books, http://en.wikipedia.org/wiki/Wikipedia:WBOOKS
[32] Book tool, http://meta.wikimedia.org/wiki/Book_tool

- exportado como documento ODF,
- ser solicitado como un libro impreso a PediaPress.

El núcleo del proyecto es todas las páginas en el espacio de nombres *Book* (i.e. Páginas que comienzan con *Book:...),* y cualquier categoría o plantilla que este relacionada con estas páginas.
El proceso de creación para usuarios normales se desagrega en cuatro pasos[33]:

1. En la página de Wikipedia, en el menú lateral izquierdo se puede ver el submenu de opciones Imprimir/exportar (Print/export) como se muestra en la Figura ??, en el cual se encuentran las opciones para generar un WikiLibro, por medio de la opción Create a book se habilita la herramienta *Book creator*, que se permanece activa en la parte superior de las páginas de Wikipedia.

2. Las páginas pueden irse agregando al libro con la opción Agregar esta página al libro (Add this page to your book) como se muestra en la Figura ??:

   Las categorías pueden también agregarse y cada vez que se agrega una categoría todas sus páginas pertenecientes se anexan al libro, esto se hace con la opción Agregar esta categoría al libro (Add this category to your book) como se aprecia en la Figura 7.

Figura 7: Agregar categorías a un WikiLibro.

   Una vez que se han agregado páginas o categorías se puede revisar el material seleccionado con la opción Show book que se puede observar en la Figura 8 y se puede agregar título y subtítulos, también se puede reorganizar el orden de los libros.

Figura 8: Opción Show Book para visualizar el contenido del libro.

   El libro ya terminado puede ser descargado en formato PDF u ODF como se ha mencionado o bien puede ordenarse una copia impresa a PediaPress, la Figura ?? muestra como puede seleccionarse el formato en el cual obtener el libro.

   La creación de un WikiLibro depende de la selección de artículos y categorías de Wikipedia que seleccionen los usuarios y no puede exceder a 500 artículos en tanto que los libros impresos no pueden exceder de 800 páginas impresas. Se estructuran en base a una plantilla con la sintaxis Wiki.

[33] Book tool/Help/Books, http://meta.wikimedia.org/wiki/Book_tool/Help/Books

Bibliografía [Buriol2006] Buriol Luciana S, Castillo Carlos, Donato Debora, Leonardi Stefano & Millozzi Stefano. *Temporal Analysis of the Wikigraph. Web Intelligence, Hong Kong* (2006) : .

[Capocci2006] Capocci A, Servedio V D P, Colaiori F, Buriol L S, Donato D, Leonardi S & Caldarelli G. *Preferential attachment in the growth of social networks: the case of Wikipedia* . (2006) : .

[Milne_2008] David Milne & Ian H Witten. *An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links.* (2008) : .

[Denoyer2006] Denoyer Ludovic & Gallinari Patrick. *The Wikipedia XML Corpus.* (2006) : .

[Holloway] Holloway. *Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its authors. ArXiv Computer Science e-prints* () : .

[Krizhanovsky2009] Krizhanovsky AA & Smirnov AV. *On the Problem of Wiki Texts Indexing.. Journal of Computer and Systems Sciences International* (2009) **48**: p. 616-624..

[Maldonado_Arias2010] Maldonado Arias Manuel. *Wikipedia: un estudio comparado* . (2010) : .

[ManningFoundations] Manning Christopher D & Schiitze Hinrich. Foundations of Statistical Natural Language Processing
. MIT (Ed.). , 1999.

[Meyer2007] Meyer Marek, Rensing Christoph & Steinmetz Ralf. *Categorizing Learning Objects Based On Wikipedia as Substitute Corpus.* (2007) : .

[Mihalcea2007] Mihalcea Rada. *Using Wikipedia for Automatic Word Sense Disambiguation. NAACL* (2007) : .

[Milne2007] Milne David. *Computing Semantic Relatedness using Wikipedia Link Structure.* (2007) : .

[Ponzetto&Strube] Ponzetto Simone Paolo & Strube Michael. An API for Measuring the Relatedness of Words in Wikipedia.. In *Proceedings of the 45th annual meeting of the association of computational linguistics.* .

[Ponzetto2006] Ponzetto Simone Paolo & Strube Michael. *Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution.* (2006) : .

[Ponzetto2007] Ponzetto Simone Paolo & Strube Michael. *Deriving a Large Scale Taxonomy from Wikipedia.* (2007) : .

[Reese2010] Reese Samuel, Boleda Gemma, Cuadros Montse, Padró Lluís & Rigau German. Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In *7th language resources and evaluation conference (lrec'10)..* 2010.

[Galicia_Haro2007] Sofía N Galicia Haro & Alexander Gelbukh. *Investigaciones en Análisis Sintáctico para el Español.* (2007) : p. 324.

[Syed2008] Syed Zareen Saba, Finin Tim & Joshi Anupam. Wikitology: Wikipedia as an ontology. In *Proceedings of the grace hopper celebration of women in computing conference.* 2008.

[Torres_Ramos2006] Torres Ramos Sulema. **Aprendizaje supervisado de colocaciones para la resolución de la ambigüedad sintáctica**. CIC. 2006.

[Torres_Ramos2009] Torres Ramos Sulema. **Optimización global de coherencia en la desambiguación del sentido de las palabras**. CIC.2009.

[Voss2006] Voss Jakob. *Collaborative thesaurus tagging the Wikipedia way.* (2006) : .

[18] Zesch Torsten & Gurevych Iryna. Analysis of the Wikipedia Category Graph for NLP Applications.. In *Proceedings of the second workshop on textgraphs: graph-based algorithms for natural language processing (2007), pp. 1-8..* 2007.

[Zesch2007] Zesch Torsten & Gurevych Iryna. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the second workshop on textgraphs: graph-based algorithms for natural language processing (2007), pp. 1-8..* 2007.

[Zesch2008] Zesch Torsten, Müller Christof & Gurevych Iryna. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the conference on language resources and evaluation (lrec)..* 2008.

[Zlatic2006] Zlatic V, Bozicevic M, Stefancic H & Domazet M. *Wikipedias: Collaborative web-based encyclopedias as complex networks*

. (2006) : .