El Procesamiento del Lenguaje Natural para extraer conocimiento de la Web, Documentos y Redes sociales

Dr. José Luis Ochoa Hernández

Departamento de Ingeniería Industrial

Universidad de Sonora

Hermosillo, Sonora, México

29 y 30 DE OCTUBRE 2012 CHIHUAHUA

ÍNDICE GENERAL

Contenido

Índice de Figuras

Índice de Tablas

1 Inteligencia Artificial

1.1 Introducción[1]

Actualmente, una gran cantidad de personas en todo mundo, conoce lo que es la inteligencia artificial, otra gran cantidad de gente ha trabajado y otra gran cantidad se ha beneficiado de sus servicios y en muchas ocasiones sin saberlo, un ejemplo muy básico es el procesador de Textos Word, con su corrector ortográfico automático de palabras o Google, el cual corrige automáticamente las palabras que escribimos mal en la caja de texto.

La inteligencia artificial hoy en día se encuentra presente en una gran cantidad de aplicaciones que en ocasiones ni imaginamos, por ejemplo se pueden mencionar los siguientes casos: *las **redes neuronales*** se están usando para detectar fraudes en instituciones bancarias y para el análisis de riesgos financieros; ***los sistemas expertos*** se emplean en el ámbito comercial, para tomar decisiones con base en la inormación detallada de los clientes; la ***lógica difusa*** se aplica tanto en electrodomésticos (lavadoras, refrigeradores, sistemas de aire acondicionado, etc.) como en el control de sistemas tan complejos como el transporte Metro; los ***sistemas multiagente*** intervienen en la gestión de cadenas de suministro en procesos industriales y en actividades de mercadotecnia como subastas e inversiones; se están utilizando ***algoritmos genéticos*** para la optimización del corte en líneas de producción de papel y para pronósticos en mercados bursátiles; la ***visión artificial*** es de gran importancia en robótica, control de calidad y aplicaciones de diagnóstico médico; también se están aplicando técnicas de inteligencia artificial para el **reconocimiento de voz** en servicios de telefonía.

Sin embargo, a pesar de las ventajas de los sistemas inteligentes, estos no se aplican masivamente en las empresas debido entre otras causas a: la **falta de personal especializado**, **la desconfianza en los beneficios de la inteligencia artificial** y **la incapacidad de detectar las áreas de oportunidad** y definir un diseño adecuado a los requerimientos y restricciones de la situación a resolver. De ahí la necesidad de contar con gente especializada, en nuestro país, que sea capaz de diseñar sistemas basados en técnicas de inteligencia artificial, para resolver problemas, en los diferentes campos de trabajo, de una manera eficiente e innovadora.

Es muy importante impulsar su estudio y difundir sus bondades para incentivar el acercamiento de los jóvenes a esta rama de la ciencia, ya que sin lugar a duda, los sistemas inteligentes son y serán una parte primordial de las tecnologías empleadas en la vida cotidiana, el trabajo y la industria.

1.2 Características de la IA

Una de las características principales es que incluye varios campos de desarrollo, como la robótica, la comprensión y traducción de lenguajes, el reconocimiento y aprendizaje de palabras de máquinas o los variados sistemas computacionales expertos, que son los encargados de reproducir el comportamiento humano en una sección del conocimiento.

Tales tareas reducen costos y riesgos en la manipulación humana en áreas peligrosas, mejoran el desempeño del personal inexperto y el control de calidad en el área comercial.

[1] http://www.cnnexpansion.com/tecnologia/2010/04/07/inteligencia-artificial-robot-expansion

1.3 Ramas de la Inteligencia Artificial[2]

La inteligencia artificial (IA) debido a su naturaleza cuenta con muchos campos de estudio, es decir, ramas en las que se divide y se investiga con especialización; cada una de estas ramas surge por medio de ideas innovadoras y el surgimiento de nuevos paradigmas de cómputo en el ámbito de la investigación computacional que permitieron nuevas técnicas de programación, éstas técnicas incluyen el concepto heurístico[3] y se apoyan en nuevos sistemas de hardware que se derivan del desarrollo de la tecnología a través de las generaciones de computadora.

Cada una de las ramas de la IA no son más que una metodología diferente para tratar la resolución de problemas aplicando el principio de inteligencia a los sistemas. Estas ramas se pueden dividir en áreas clásicas y áreas de vanguardia de acuerdo a la época en que surgieron, pero esta clasificación varía debido a la diversidad de metodologías de IA existentes, por tal motivo se presentarán cada una de las ramas y sus características tratando de seguir el orden de surgimiento de cada una.

1.3.1 Reconocimiento del Habla

El ***"Reconocimiento del Habla"*** es una rama que utiliza el método interactivo de comunicación primaria al igual que el ser humano que es el habla, lo que le permite "escuchar" a una persona hablar, decodificar el significado de las palabras, interpretarlas, y transmitir una respuesta. Actualmente se utiliza mucho para personas con capacidades diferentes que no son capaces de utilizar el teclado y necesitan comunicarse por medio del habla para poder utilizar diferentes dispositivos.

1.3.2 Procesamiento del Lenguaje Natural

Otra rama se conoce como ***"Procesamiento del Lenguaje Natural",*** que es un intento de comunicación cada vez más clara entre humano-máquina y máquina-humano, dejando el uso de lenguajes de programación o de un conjunto de comandos, para procesar el lenguaje humano natural. Para procesar dicho lenguaje humano natural se necesita dividirlo, primero se obtiene la comprensión del lenguaje natural, que investiga métodos para que la computadora permita comprender instrucciones dadas en este tipo de lenguaje, para que así la computadora nos pueda comprender más fácilmente; como segundo paso es la generación de lenguaje natural, en donde la computadora es capaz de expresarse en el lenguaje humano natural, de tal forma que podamos entenderla de manera más sencilla.

1.4 Importancia de la Inteligencia Artificial

Las computadoras son fundamentales hoy día en nuestras vidas, afectando todos los aspectos de esta. La Inteligencia Artificial se crea con la implementación en las computadoras para realizar mecanismo de computación que utiliza *programas fijos* y contiene una seria de *reglas* que lo hacen funcionar. Esto permite a las computadoras a ser creadas en máquinas artificiales que desempeñan tareas *monótonas, repetitivas y simples* más eficiente y efectivas que un ser

[2] http://icopcion.wordpress.com/2011/02/08/ramas-de-la-inteligencia-artificial/

1.- Inteligencia Artificial e Ingeniería del Conocimiento. Gonzalo Pajares Martinsanz y Matilde Santos Peñas. AlfaOmega. México, D.F. Mayo 2007.

2.- A Fondo: Inteligencia Artificial. Henry C. Mishkoff. Ediciones Anaya Multimedia S. A. Madrid 1988.

[3] Heurística: Idea Basada en la experiencia que ayuda a determinar cómo se debe proceder.

humano. Estudios sobre trabajos o tareas repetitivas han demostrado que el ser humano no le agrada este tipo de trabajo y al pasar el tiempo son más susceptibles a cometer errores en el mismo.

Para situaciones complejas, el objetivo se hace más complejo, debido a que la inteligente artificial dotada a las computadoras no es suficiente, estas tienen dificultad en entender ciertas situaciones o problemas específicos, por lo tanto no saben cómo reaccionar a estas. También ocurre que dentro de un problema, tienen la variabilidad del mismo y no pueden adaptarse a un cambio que pueda ocurrir. Estos problemas son de suma importancia para la Inteligencia Artificial que busca mejorar, aprender y entender el razonamiento, para que en estas situaciones las computadoras puedan dar una solución.

El campo de la ciencia de la Inteligencia Artificial está todavía en etapas de crecimiento, comparadas con otras ramas de la computación, pero poco a poco el estudio del comportamiento humano dará paso a aplicar estos conocimientos a las computadoras, y por lo tanto estas lograr de manera primitiva razonar sobre diferentes situaciones. La complejidad en aplicarle conocimientos del ser humano a las computadoras, es la capacidad de nosotros mismos de ser impredecible, ya que cada persona puede reaccionar diferente a una situación específica por esa razón no se puede implementar un patrón dentro de la memoria de una computadora. Hasta ahora, no existe la posibilidad de predecir o almacenar todo tipo de comportamiento de un ser humano a todas las situaciones que se enfrenta durante su existencia.

Lo que si tiene que quedar claro es que debemos usar nuestros recursos materiales y humanos con más eficiencia, y para lograrlo, es necesaria la ayuda que nos ofrecen las computadoras.

Existe la falsa impresión de que uno de los objetivo del IA es sustituir a los trabajadores humanos y ahorrar dinero. Pero en el mundo de los negocios, la mayoría de personas está más entusiasmada ante las nuevas oportunidades que ante el abatimiento de costos. Además, la tarea de reemplazar totalmente a un trabajador humano abarca de lo difícil a lo imposible, ya que no se sabe cómo dotar a los sistemas de IA de toda esa capacidad de percibir, razonar y actuar que tienen las personas. Sin embargo, debido a que los humanos y los sistemas inteligentes tienen habilidades que se complementan, podrían apoyarse y ejecutar acciones conjuntas:

- En la agricultura, controlar plagas y manejar cultivos en forma más eficiente.
- En las fábricas, realizar montajes peligrosos y actividades tediosas (labores de inspección y mantenimiento).
- En la medicina, ayudar a los médicos a hacer diagnósticos, supervisar la condición de los pacientes, administrar tratamientos y preparar estudios estadísticos.
- En el trabajo doméstico, brindar asesoría acerca de dietas, compras, supervisión y gestión de consumo energético y seguridad del hogar.
- En las escuelas, apoyar la formación de los estudiantes, especialmente en aquellas materias consideradas complejas.
- Ayudar a los expertos a resolver difíciles problemas de análisis o a diseñar nuevos dispositivos.
- Aprender de los ejemplos para explorar bases de datos en busca de regularidades explotables.
- Proporcionar respuestas a preguntas en lenguaje natural usando datos estructurados y texto libre.

La IA aplicada es la contraparte de ingeniería de la ciencia cognoscitiva y complementa sus perspectivas tradicionales. La ciencia cognoscitiva es una mezcla de psicología, lingüística y filosofía.

La metodología y terminología de la IA está todavía en vías de desarrollo. La IA se está dividiendo y encontrando otros campos relacionados: lógica, redes neuronales, programación orientada a objetos, lenguajes formales, robótica, etc. Esto explica por qué el estudio de IA no está confinado a la matemática, ciencias de la computación, ingeniería (particularmente la electrónica y la mecánica), o a la ciencia cognoscitiva, sino que cada una de estas disciplinas es un potencial contribuyente. Por ejemplo, la robótica es considerada como un campo interdisciplinario que combina conceptos y técnicas de IA, con ingeniería óptica, electrónica y mecánica.

1.5 Aplicaciones de la Inteligencia Artificial[4]

Hay muchos usos y aplicaciones de la Inteligencia Artificial, tanto en el mercado comercial como en el servicio militar. Para propósito de esta investigación, se limitó al uso comercial debido a que mucho de los usos de esta metería en el servicio militar, es confidencial.

Algunos de los usos de la Inteligencia Artificial son:

1.5.1 Reconocimiento de Voz

Los sistemas que contienen su memoria los datos para reconocer voz, comenzaron su auge en la década de los ’90, cuando se logró implementar el programa en los sistemas de las aerolíneas, aunque con capacidades limitadas. El sistema consiste en indicarle las instrucciones mediante los comandos de voz para que este realizara sus tareas. Un ejemplo de este tipo de sistema es “Via Voice” el cual fue el primer sistema credo para las aerolíneas.

1.5.2 Entendimiento del lenguaje natural

Los sistemas tienen data para el reconocer el significado de la palabra estando esta sola pero cuando se tienen en contexto y en una oración hay problemas en que el sistema reconozca el significado de cada una de las palabras. En este campo se esta muy primitivo por que el objetivo final es que el sistema reconozca lo que se le indica y responda a este.

1.5.3 Visión computarizada

Existen ciertos sistemas que están diseñados con programas que reconocen en solo dos dimensiones pero para lograr eficientemente que el sistema funcione a capacidad total se tiene que desarrollar información que se pueda ser leída por el sistema en tercera dimensión porque esta la que se requiere. Este sistema está en etapas

[4] http://es-es.facebook.com/pages/AIS-Aplicaciones-de-Inteligencia-Artificial/123835344323009

primitiva y su uso mayor es en el área verificación y/o lectura de la retina para propósitos de seguridad.

2 Procesamiento del Lenguaje Natural

2.1 Introducción[5],[6]

El *Procesamiento del Lenguaje Natural* es una subrama de la Inteligencia Artificial y de la Lingüística. También se suele referir a esta rama de la informática de forma abreviada como PLN o NLP, del inglés Natural Language Processing.

El *Procesamiento del Lenguaje Natural* es una disciplina con una larga trayectoria. Nace en la década de 1960, con el objeto de estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural, originalmente desarrollado a comienzos de la Guerra Fría como el mecanismo que usaban los físicos Soviéticos para la traducción de documentos [Locke y Booth].

El fin del PLN es construir sistemas y mecanismos que permitan la comunicación entre personas y máquinas por medio de lenguajes naturales ya sea por medio de la voz o del texto. Además, trata de que los mecanismos que permitan esa comunicación sean lo más eficaces posibles, computacionalmente hablando. En definitiva, se busca poder crear programas que puedan analizar, entender y generar lenguajes que los humanos utilizan habitualmente, de manera que el usuario pueda llegar a comunicarse con la computadora de la misma forma que lo haría con un humano.

Veamos una definición [Covington]

> "*El Procesamiento de lenguaje natural (PLN) es el uso de computadoras para entender lenguajes (naturales) humanos tales como inglés, francés o japonés. Por entender no se quiere decir que el computador tenga pensamientos, sentimientos y conocimientos humanizados, sino que el computador pueda reconocer y usar información expresada en lenguaje humano*".

Otra definición [Manaris y Slator]

"*Un sistema de PLN es aquel que encapsula un modelo del lenguaje natural en algoritmos apropiados y eficientes. En donde las técnicas de modelado están ampliamente relacionadas con eventos en muchos otros campos*", incluyendo:

- *Ciencia de la computación*, la cual provee métodos para representar modelos, diseñar e implementar algoritmos para herramientas de software.
- *Lingüística*, la cual contribuye con nuevos modelos lingüísticos y procesos.
- *Matemática*, la cual identifica modelos formales y métodos.
- *Neurociencia*, la cual explora los mecanismos mentales y otro tipo de actividades físicas.

Durante toda la historia de humanidad el conocimiento, en su mayor parte se comunica, se guarda y se maneja en la forma de lenguaje natural –griego, latín, inglés, español, etc. La época actual no es ninguna excepción: el conocimiento sigue existiendo y creándose en la forma de

[5] http://www.cicling.org/ampln/NLP.htm#Recuperaci%C3%B3n_de_Informaci%C3%B3n
[6] http://procesamientolenguajerecuperacion.50webs.org/index.html

documentos, libros, artículos, aunque éstos se guardan en forma electrónica (digital). El gran avance que tenemos hoy en día, es que las computadoras ya pueden ser una ayuda enorme al momento de procesar todo este conocimiento.

En la época actual de información, del manejo eficiente de este conocimiento depende el uso de todos los demás recursos, sean estos: naturales, industriales y humanos.

Sin embargo, lo que es conocimiento para nosotros –los seres humanos– no lo es para las computadoras. Para ellas son solos archivos, una secuencia de caracteres, y nada más. Una computadora puede copiar tal archivo, respaldarlo, transmitirlo, borrarlo –como un funcionario–, pero no puede buscar las respuestas a las preguntas en este texto, hacer las inferencias lógicas sobre su contenido, generalizar y resumirlo –es decir, hacer todo lo que las personas normalmente hacemos con el texto–, porque no lo puede entender.

Para combatir esta situación, se dedica mucho esfuerzo, sobre todo en los países más desarrollados, al desarrollo de la *ciencia que se encarga de habilitar a las computadoras a entender el texto,* en función del enfoque práctico versus teórico, del grado en el cual se espera lograr la comprensión y de otros aspectos tiene varios nombres: *procesamiento de lenguaje natural, procesamiento de texto, tecnologías de lenguaje, lingüística computacional*. En todo caso, se trata de procesar el texto por su sentido y no como un archivo binario.

El gran avance que tenemos hoy en día, es que las computadoras ya pueden ser una ayuda enorme al momento de procesar todo este conocimiento, sobre todo si se enfoca al manejo eficiente de la información, aplicado a los recursos naturales, industriales y humanos.

No solo basta con procesar el conocimiento, sino también hay que almacenarlo, para ello, existe una forma de representación del conocimiento que permite utilizarlo de forma útil y eficiente, a este medio de almacenamiento se le conoce como '*ontologías*'. El investigador Tomas Gruber [Gruber] define ontología como "*a formal explicit specification of a shared conceptualization*". Sin embargo, esta definición, no es la única ni la última, pero si, es la más utilizada en este dominio, por su sencillez y completitud. Dicho de otra forma, una ontología, es una jerarquía de conceptos con atributos y relaciones, que define una terminología consensuada para definir redes semánticas de unidades de información interrelacionadas. También, una ontología, proporciona un vocabulario de clases y relaciones para describir un dominio, poniendo el acento en la compartición del conocimiento y el consenso en la representación de éste.

Las ontologías son actualmente aplicadas en varios sectores, como el financiero, educacional, etc, etc, sin embargo, en donde serán más utilizadas, será en la denominada *Web Semántica, la cual* está incluida dentro de la Web a la que se le conoce como 3.0, esta famosa Web destaca por la forma en que la información es utilizada, ya que su estructura permite obtener resultados más exactos, como por ejemplo, en esta Web, a la consulta realizada en un buscador, *Quiero un vuelo a Cancún por la mañana, para el día 15 de Agosto y que no me cueste más de 1500 pesos*, los resultados serían exactamente los que estamos pidiendo, ya que el navegador detectaría de forma automática nuestro origen y nos mostraría las empresas que ofrecen estos servicios, en cambio, actualmente, no sucede esto.

2.2 Descripción del problema

Como hemos dicho anteriormente, la tendencia se encuentra en mejorar de forma notoria la Web, sin embargo, la Web no es lo único que necesita actualización, el conocimiento que

existe actualmente principalmente en las empresas o en los centros de investigación, sean estos las universidades o laboratorios tecnológicos, que se está dejando de explotar por falta de recursos, sean estos de personal, económicos o tecnológicos, por esa razón se presenta en este artículo una herramienta que extrae el conocimiento de forma automática, en el dominio que sea, principalmente de textos escritos en lenguaje natural y en español, refiriéndonos a esto como cualquier texto que se pueda ser almacenado en un solo o una colección de archivos de texto , como por ejemplo, Leyes, reglamentos, manuales, informes, reportes, quejas, libros, etc, etc.

2.3 Medios de existencia del conocimiento[7]

Philip G. Armour distingue cinco formas conocidas de almacenar el conocimiento –DNA, el cerebro/memoria, herramientas/aparatos, libros y software– y analiza las características, ventajas y desventajas de cada uno de ellos [Philip G. Armour].

1. **DNA:** Es el primer método de almacenamiento del conocimiento. EL DNA existe para almacenar el conocimiento de cómo crear vida, como una máquina de Turing. El conocimiento está profundamente empotrado, pasar de grado es obligatorio para la supervivencia de las especies. El conocimiento es persistente, pero se actualiza muy lentamente. No tenemos la capacidad de cambiar el conocimiento –todavía, o sí…– de forma intencionada. El DNA puede hacer crecer un objeto físico que interactúa y modifica el entorno.
2. **Cerebro:** Es un "experimento" casi exclusivo de la raza humano: almacenar más conocimiento en el cerebro que lo que se hereda en el DNA. Usamos nuestro cerebro para almacenar el conocimiento que adquirimos, fue el segundo método de almacenar el conocimiento que conocimos. El conocimiento es muy volátil, pero podemos cambiarlo rápida e intencionalmente. Podemos aplicar ese conocimiento para afectar y modificar el mundo.
3. **Máquinas y herramientas:** El valor más importante de una herramienta no es ella en sí misma, sino como ha sido creada y modificada. El conocimiento del creador de esas herramientas es lo que marca las diferencias. Se las suele llamar también "conocimiento sólido" y fue la tercera forma de almacenar el conocimiento. El conocimiento es bastante persistente, pero no es fácil de actualizar. Es intencional y existe para afectar el mundo exterior.
4. **Libros:** Ha permitido nuevas formas de depositar y acceder al conocimiento que hasta ese momento estaban confinados al cerebro. Hizo al conocimiento portable en el tiempo y en el espacio. El conocimiento es muy persistente, pero de actualización lenta. Aunque los libros son intencionales no tienen capacidad para cambiar al mundo.
5. **Software:** Es la última forma conocida –de hace sólo unos 50 años– para almacenar el conocimiento. Después de unos inicios dubitativos, está creciendo a una velocidad vertiginosa. Multitud de personas están trabajando para obtener información de las fuentes más diversas, comprenderla, clasificarla y trasladarla a este medio, y entonces intentan validar todo ese conocimiento. Hay una razón para que se invierta tanto esfuerzo, este medio tiene las características que deseamos y que no tienen los otros medios: es intencional, persistente, de actualización sencilla y rápida, y sobre todo es activo.

Sin embargo, existen otras formas no contempladas por el autor, como lo son las Páginas web y las ontologías.

[7] http://mnm.uib.es/gallir/posts/2005/10/06/455/

6. **Páginas Web:** Una página web es el nombre de un documento o información electrónica adaptada para la World Wide Web y que puede ser accedida mediante un navegador para mostrarse en un monitor de computadora o dispositivo móvil. Esta información se encuentra generalmente en formato HTML o XHTML, y puede proporcionar navegación a otras páginas web mediante enlaces de hipertexto. Las páginas web frecuentemente incluyen otros recursos como hojas de estilo en cascada, guiones (scripts) e imágenes digitales, entre otros.

   Las páginas web pueden estar almacenadas en un equipo local o un servidor web remoto. El servidor web puede restringir el acceso únicamente para redes privadas, p. ej., en una intranet corporativa, o puede publicar las páginas en la World Wide Web. El acceso a las páginas web es realizado mediante su transferencia desde servidores utilizando el protocolo de transferencia de hipertexto (HTTP).

   Una página web está compuesta principalmente por información (sólo texto y/o módulos multimedia) así como por hiperenlaces; además puede contener o asociar datos de estilo para especificar cómo debe visualizarse, y también aplicaciones embebidas para así hacerla interactiva.

7. **Ontologías:** El término ontología en informática hace referencia a la formulación de un exhaustivo y riguroso esquema conceptual dentro de uno o varios dominios dados; con la finalidad de facilitar la comunicación y el intercambio de información entre diferentes sistemas y entidades.

   Un uso común tecnológico actual del concepto de ontología, en este sentido semántico, lo encontramos en la inteligencia artificial y la representación del conocimiento. En algunas aplicaciones, se combinan varios esquemas en una estructura de facto completa de datos, que contiene todas las entidades relevantes y sus relaciones dentro del dominio.

2.4 Tipos de Información

2.4.1 Información Estructurada

Cuando hablamos de Información estructurada, nos estamos refiriendo a documentos cuya estructura es declarada explícitamente de algún modo, ya sea asociando etiquetas a elementos de estructura o mediante la sintaxis con la que se escribe el documento, como hacen los lenguajes de programación. No se pueden entender como documentos estructurados aquellos escritos utilizando cualquier procesador de textos, ya que en ellos la estructura (por ejemplo, el título de un capítulo) se denota a través de la forma que adopta el contenido.

La información estructurada es la que estamos acostumbrados a administrar y a procesar, ya sea para el soporte a la toma de decisiones o, simplemente, para consultar información como los diccionarios. Este tipo de información, cuando se extrae a partir de documentos escritos en lenguaje natural, representa una gran desventaja para una organización, puesto que perdemos de vista información muy valiosa que se encuentra no-estructurada.

2.4.2 Información Semiestructurada

Internet constituye hoy en día la fuente de información para la inteligencia competitiva más potente al alcance de todo tipo de organizaciones y personas. La paradoja radica en que la mayoría de la información disponible en Internet no es visible para las herramientas tradicionalmente utilizadas en la búsqueda y explotación de dicha información. Nos estamos refiriendo a la llamada "Web Oculta" y que los buscadores por regla general, no son capaces de mostrar.

La importancia de la Web Oculta no deriva, no obstante, de la cantidad de información que contiene, sino principalmente en la calidad de la misma. Todo lo que está detrás de un formulario tendrá una calidad mucho mayor de lo que podamos encontrar en él. La información más valiosa se encuentra almacenada en bases de datos y es accesible por Internet mediante formularios de consulta, tal y como pone de manifiesto Francois Libmann.

La propiedad más importante que caracteriza a la Web Oculta es que su información está semiestructurada (es decir, débilmente estructurada). O sea, la principal ventaja de la información residente en la Web Oculta consiste en que ésta puede ser procesada de forma automática una vez que ha sido estructurada, por lo que puede consultarse con potentes lenguajes de consulta similares a los que nos ofrecen las bases de datos (por ejemplo, SQL o XQuery).

2.4.3 Información no estructurada expresada en lenguaje natural

La información no estructurada la encontramos en fuentes tales como documentos, la Web y cualquier otro medio que contenga información expresada principalmente en lenguaje natural y en formatos muy diversos como texto, videos, audio o imágenes. Este tipo de información ocupa más del 94 % del universo digital.

En los últimos años, los esfuerzos para la gestión de la información en las empresas se han centrado, principalmente, en los datos estructurados y semiestructurados, con el único objetivo de sacar el máximo partido posible a las costosas inversiones en bases de datos y sistemas transaccionales.

Sin embargo, con la famosa "explosión de los datos" vivida en los últimos años, y, gracias a la llegada de Internet y al uso que a dicha red se le ha dado por todo tipo de organizaciones y personas, la información no estructurada, como el correo electrónico, los informes, contratos, facturas, formularios, hojas de cálculo, imágenes, presentaciones, etc., ha llegado a ser muy importante. De hecho, el 80% de la información de una empresa reside hoy en estos contenidos no estructurados, vitales para las operaciones diarias de los departamentos más estratégicos de la empresa, desde marketing e I+D[8] a recursos humanos y finanzas.
A pesar de tanto éxito, la Web ha introducido nuevos problemas.

- lo *tedioso* o *aburrido* que es; y
- lo *difícil* que es encontrar la *información útil* en ella.

Las tecnologías actuales de IR (representadas en el contexto Web por "buscadores" o "portales", tales como Google, Yahoo, MSN Search, Altavista, Lycos, Infoseek, etc.) hacen que las necesidades de los usuarios sean satisfechas únicamente de forma *parcial* o *primitiva*. El

[8] **I + D**, proveniente de las siglas que hacen referencia a ***I****nvestigación* + ***D****esarrollo.*

gran número de documentos irrelevantes que usualmente recuperan estos buscadores, y el gran número de documentos relevantes que estos sistemas no recuperan, son un obstáculo substancial para una mejor utilización de los recursos de información disponibles en la Web.

2.5 Problemas en el uso del Lenguaje Natural[9]

El lenguaje natural, entendido como la herramienta que utilizan las personas para expresarse, posee propiedades que merman la efectividad de los sistemas que emplean el PLN, sean estos los de recuperación de información, reconocimiento de voz, traducción de documentos, etc. Estas propiedades son la *variación* y la *ambigüedad lingüística*. Cuando hablamos de la ***variación lingüística*** nos referimos a la posibilidad de utilizar diferentes palabras o expresiones para comunicar una misma idea. En cambio, la ***ambigüedad lingüística*** se produce cuando una palabra o frase permite más de una interpretación.

Ambos fenómenos inciden en el este proceso aunque de forma distinta. La *variación lingüística* provoca el silencio documental, es decir la omisión de documentos relevantes para cubrir la necesidad de información, ya que no se han utilizado los mismos términos que aparecen en el documento. En cambio, la *ambigüedad implica* el ruido documental, es decir, la inclusión de documentos que no son significativos, ya que se recuperan también documentos que utilizan el término pero con significado diferente al requerido. Estas dos características dificultan considerablemente el tratamiento automatizado del lenguaje. A continuación se muestran algunos ejemplos que ilustran la repercusión de estos fenómenos en este proceso.

A nivel morfológico una misma palabra puede adoptar diferentes roles morfo-sintácticos en función del contexto en el que aparece, ocasionando problemas de ambigüedad.

> *Ejemplo 1. Deja la comida que sobre sobre la mesa de la cocina, dijo llevando el sobre en la mano.*

La palabra sobre es ambigua morfológicamente ya que puede ser un sustantivo masculino singular, una preposición, y también la primera o tercera persona del presente de subjuntivo del verbo sobrar.

A nivel sintáctico, centrado en el estudio de las relaciones establecidas entre las palabras para formar unidades superiores, sintagmas y frases, se produce ambigüedad a consecuencia de la posibilidad de asociar a una frase más de una estructura sintáctica. Por otro lado, esta variación supone la posibilidad de expresar lo mismo pero cambiando el orden de la estructura sintáctica de la frase.

> *Ejemplo 2. María vio a un niño con un telescopio en la ventana.*

La interpretación de la dependencia de los dos sintagmas preposicionales, con un telescopio y en la ventana, otorga diferentes significados a la frase: (i) María vio a un niño que estaba en la ventana y que tenía un telescopio, (ii) María estaba en la ventana, desde donde vio a un niño que tenía un telescopio, y (iii) María estaba en la ventana, desde donde miraba con un telescopio, y vio a un niño.

[9] http://www.hipertext.net/web/pag277.htm

A nivel semántico, donde se estudia el significado de una palabra y el de una frase a partir de los significados de cada una de las palabras que la componen. La ambigüedad se produce porque una palabra puede tener uno o varios sentidos, es el caso conocido como polisemia.

Ejemplo 3. Luis dejó el periódico en el banco.

El término banco puede tener dos significados en esta frase, (i) entidad bancaria y (ii) asiento. La interpretación de esa frase va más allá del análisis de los componentes que forman la frase, se realiza a partir del contexto en que es formulada.

Y también hay que tener en cuenta la variación léxica que hace referencia a la posibilidad de utilizar términos distintos a la hora de representar un mismo significado, es decir el fenómeno conocido como sinonimia:

Ejemplo 4: Coche / Vehículo / Automóvil.

A nivel pragmático, basado en la relación del lenguaje con el contexto en que es utilizado, en muchos casos no puede realizarse una interpretación literal y automatizada de los términos utilizados. En determinadas circunstancias, el sentido de las palabras que forman una frase tiene que interpretarse a un nivel superior recurriendo al contexto en que es formulada la frase.

Ejemplo 5. Se moría de risa.

En esta frase no puede interpretarse literalmente el verbo morirse si no que debe entenderse en un sentido figurado.

Otra cuestión de gran importancia es la ambigüedad provocada por la anáfora, es decir, por la presencia en la oración de pronombres y adverbios que hacen referencia a algo mencionado con anterioridad.

Ejemplo 6. Ella le dijo que los pusiera debajo

La interpretación de esta frase tiene diferentes incógnitas ocasionadas por la utilización de pronombres y adverbio: ¿quién habló?, ¿a quién?, ¿qué pusiera qué?, ¿debajo de dónde?. Por tanto, para otorgar un significado a esta frase debe recurrirse nuevamente al contexto en que es formulada.

Con todos los ejemplos expuestos queda patente la complejidad del lenguaje y que su tratamiento automático no resulta fácil ni obvio.

3 Web Semántica (Semantic Web)

3.1 Introducción[10]

En poco más de una década desde su aparición, la *World Wide Web* se ha convertido en un instrumento de uso cotidiano en nuestra sociedad, comparable a otros medios tan importantes como la radio, la televisión o el teléfono, a los que aventaja en muchos aspectos. La web es hoy un medio extraordinariamente flexible y económico para la comunicación, el comercio y los negocios, ocio y entretenimiento, acceso a información y servicios, difusión de cultura, etc. Paralelamente al crecimiento espectacular de la web, las tecnologías que la hacen posible han experimentado una rápida evolución. Desde las primeras tecnologías básicas: HTML y HTTP, hasta nuestros días, han emergido tecnologías como CGI, Java, JavaScript, ASP, JSP, PHP, Flash, J2EE, XML, por citar algunas de las más conocidas, que permiten una web mejor, más amplia, más potente, más flexible, o más fácil de mantener. Estos cambios influyen y son al tiempo influidos por la propia transformación de lo que entendemos por WWW. La generación dinámica de páginas, el acoplamiento con bases de datos, la mayor interactividad con el usuario, la concepción de la web como plataforma universal para el despliegue de aplicaciones, la adaptación al usuario, son algunas de las tendencias evolutivas más marcadas de los últimos años.

La evolución de la web no termina aquí ni mucho menos. Son diversos los aspectos susceptibles de mejorar. Entre las últimas tendencias que pueden repercutir en el futuro de la web a medio plazo, a finales de los 90 surge la visión de lo que se ha dado en llamar la *web semántica*. Se trata de una corriente, promovida por el propio inventor de la web y presidente del consorcio W3C, cuyo *último fin es lograr que las máquinas puedan entender*, y por tanto utilizar, lo que la web contiene. Esta nueva web estaría poblada por agentes o representantes software capaces de navegar y realizar operaciones por nosotros para ahorrarnos trabajo y optimizar los resultados.

Para conseguir esta meta, la web semántica propone describir los recursos de la web con representaciones procesables (es decir, entendibles) no sólo por personas, sino por programas que puedan asistir, representar, o reemplazar a las personas en tareas rutinarias o inabarcables para un humano. Las tecnologías de la web semántica buscan desarrollar una web más cohesionada, donde sea aún más fácil localizar, compartir e integrar información y servicios, para sacar un partido todavía mayor de los recursos disponibles en la web.

[10] http://blogs.enap.unam.mx/asignatura/francisco_alarcon/wp-content/uploads/2012/01/web_semantica.pdf

3.2 Un poco de historia

Ilustración 3. 1. Tim Berners-Lee creador de la World Wide Web.

La aparición de la WWW se puede situar en 1989, cuando Tim Berners-Lee presentó su proyecto de "World Wide Web" en el CERN (Suiza), con las características esenciales que perduran en nuestros días. El propio Berners-Lee completó en 1990 el primer servidor web y el primer cliente, y un año más tarde publicó el primer borrador de las especificaciones de HTML y HTTP. El lanzamiento en 1993 de Mosaic, el primer navegador de dominio público, compatible con Unix, Windows, y Macintosh, por el National Center for Supercomputing Applications (NCSA), marca el momento en que la WWW se da a conocer al mundo, extendiéndose primero en universidades y laboratorios, y en cuestión de meses al público en general, iniciando el que sería su vertiginoso crecimiento. Los primeros usuarios acogieron con entusiasmo la facilidad con que se podían integrar texto y gráficos y saltar de un punto a otro del mundo en una en una misma interfaz, y la extrema sencillez para contribuir contenidos a una web mundial.

Aunque es sumamente difícil medir el tamaño de la web, se estima que hoy día unos 10^9 usuarios utilizan la web, y que ésta contiene del orden de 4×10^9 documentos, un volumen de información equivalente a entre 14 y 28 millones de libros. Como dato comparativo, la asociación *American Research Libraries*, que agrupa unas 100 bibliotecas en EE.UU., tiene catalogados unos 3.7 millones de libros. La biblioteca de la Universidad de Harvard, la mayor de EE.UU., contiene en torno a 15 millones de libros. Estas cifras incluyen sólo lo que se ha dado en denominar la *web superficial*, formada por los documentos estáticos accesibles en la web.

Hoy casi todo está representado de una u otra forma en la web, y con la ayuda de un buen buscador, podemos encontrar información sobre casi cualquier cosa que necesitemos. La web está cerca de convertirse en una enciclopedia universal del conocimiento humano. No obstante, en este panorama tan favorable hay espacio para mejoras. Por ejemplo, el enorme tamaño que ha alcanzado la web, a la vez que es una de las claves de su éxito, hace que algunas tareas (por ejemplo encontrar la planificación óptima con transporte, alojamiento, etc., entre todas las posibles para un viaje bajo ciertas condiciones), requieran un tiempo excesivo para una persona o resulten sencillamente inabarcables.

Esto se debe principalmente a que la web actual está enfocada totalmente al consumo humano, por ejemplo: ahora vemos Ilustración 3.2 A) *La web vista por una persona y B) la web vista por el ordenador.*

A)

Ilustración 3. 2. Representación de una página web.

B) Mostrar todos <script type="text/javascript">addDOMLoadEvent(function(){jQuery('#jcarousel-cartelera-4f47f0d024173').jcarousel({scroll:1});});</script> <!-- BILLBOARD ZONE #ap-100 --> 61% de actividad social <h2 style="line-height: 31px;" class="title-32 bold">De Guindos: "El problema ya no es la austeridad fiscal, sino volver al crecimiento económico" 59</h2> "La austeridad fiscal es necesaria, pero se tiene que complementar con políticas que fomenten el crecimiento", dice el ministro.
El ministro traslada a Bernanke el compromiso de corregir el déficit.
Guindos mantendrá un encuentro con el 'think tank' Brooking Institution y este sábado otro con empresarios españoles en México. 71% de actividad social <h2 class="title-18">Portugal rechaza que las parejas homosexuales puedan adoptar 58</h2

3.3 La web semántica

La web semántica propone superar las limitaciones de la web actual mediante la introducción de descripciones explícitas del significado, la estructura interna y la estructura global de los contenidos y servicios disponibles en la WWW. Frente a la semántica implícita, el crecimiento caótico de recursos, y la ausencia de una organización clara de la web actual, la web semántica aboga *por clasificar, dotar de estructura y anotar los recursos con semántica explícita procesable por máquinas*. La ilustración 3.3 muestra esta propuesta. Actualmente la web se asemeja a un grafo formado por nodos del mismo tipo, y arcos (hiperenlaces) igualmente indiferenciados. Por ejemplo, no se hace distinción entre la página personal de un profesor y el portal de una tienda on-line, como tampoco se distinguen explícitamente los enlaces a las asignaturas que imparte un profesor de los enlaces a sus publicaciones. Por el contrario en la web semántica cada nodo (recurso) tiene un tipo (profesor, tienda, pintor, libro), y los arcos representan relaciones explícitamente diferenciadas (pintor – obra, profesor – departamento, libro – editorial).

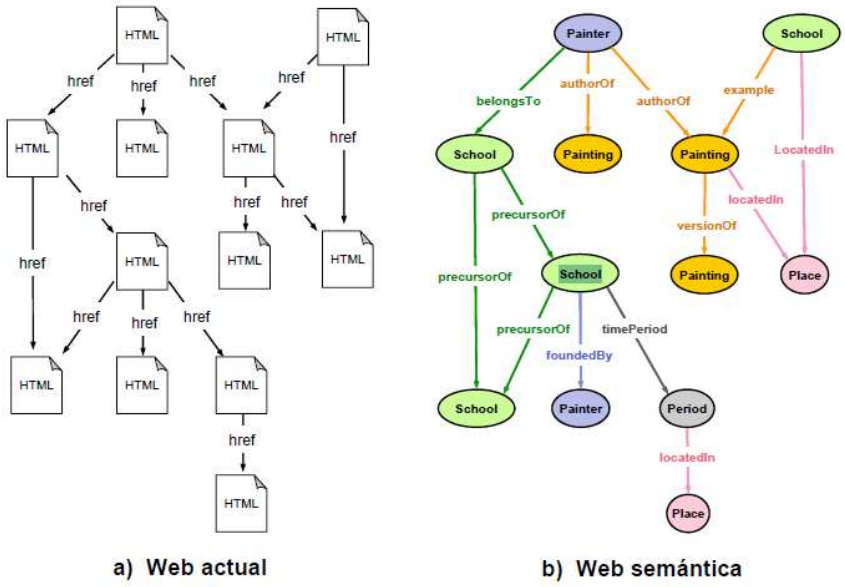


Ilustración 3. 3. Comparación de la Web Actual Vs la Web semántica (ejemplo).

La web semántica mantiene los principios que han hecho un éxito de la web actual, como son los principios de descentralización, compartición, compatibilidad, máxima facilidad de acceso y contribución, o la apertura al crecimiento y uso no previstos de antemano. En este contexto un problema clave es alcanzar un entendimiento entre las partes que han de intervenir en la construcción y explotación de la web: usuarios, desarrolladores y programas de muy diverso perfil. La web semántica rescata la noción de ontología del campo de la Inteligencia Artificial como vehículo para cumplir este objetivo.

3.4 ¿Qué es la Web semántica?

Definición 1. La Web semántica o Web 3.0 (o semantic web), sin más palabras es la "*Web de los datos*". Se basa en la idea de añadir metadatos semánticos y ontológicos a la *World Wide*

Web. El objetivo es mejorar Internet ampliando la interoperabilidad entre los sistemas informáticos usando "agentes inteligentes"[11].

Definición 2. La Web Semántica es la nueva generación de la Web, que intenta realizar un filtrado automático preciso de la información. Para ello, es necesario hacer que la información que reside en la Web sea entendible por las propias máquinas. Especialmente su contenido, más allá de su simple estructura sintáctica.

Con lo cual, podemos determinar que la Web Semántica trata sobre diferentes ámbitos, por un lado es un conjunto de lenguajes y procedimientos para poder añadir esa semántica a la información para que sea entendible por los agentes encargados de procesarla. Y por el otro lado trata, el desarrollo y la construcción de los agentes encargados de procesar esa información y filtrar la que es útil para los usuarios o para agentes que tienen que realizar una determinada función.

[11] Agentes inteligentes son programas en las computadoras que buscan información sin operadores humanos.

4 Metodología base de la implementación

La metodología que fue seguida para lograr el desarrollo de la aplicación se basa en la metodología de Ontology Learning en español diseñada por el Dr. Ochoa en [Ochoa, J.L. et al., 2011]. La cual esta formada por cuatro etapas principales, a las cuales se les conocerá desde ahora como módulos. Estos módulos se identifican como 1) módulo de extracción de conceptos, 2) módulo de aprendizaje automático de patrones, 3) módulo de extracción de relaciones y 4) módulo de creación de la ontología.

4.1 Módulo de Extracción de conceptos

Este módulo esta formado por 4 etapas principales, a saber:

- Fase de PLN
- Fase de patrones lingüísticos
- Fase de extracción de conceptos compuestos
- Fase de extracción de conceptos simples

4.2 Fase de PLN

El objetivo principal de esta fase, es el análisis del texto de forma lingüística. Para ello, se ha empleado una herramienta de análisis del lenguaje natural llamada *FreeLing*[12] *2.2* [(Atserias et al., 2006); (Padró et al., 2010)]. Cuyas funcionalidades básicas que realiza esta herramienta son las siguientes:

- ***Tokenizador (Tokeneizer):*** Subproceso que se aplica a un texto sin formato (lenguaje natural) y devuelve una lista de objetos separados en palabras individuales.

- ***Divisor de oraciones (Sentence Splitter):*** Subproceso que emplea una lista de objetos palabra y devuelve una lista de objetos oraciones.

- ***Etiquetado gramatical (PosTagger):*** Subproceso que etiqueta cada una de las palabras con su categoría gramatical. Este proceso se puede realizar en base a la definición de la palabra o el contexto en que aparece.

- ***Lematizador (Lemmatizer):*** Subproceso que permite el reconocimiento, la generación y la manipulación de las relaciones morfoléxicas a partir de cualquier palabra, incluyendo la recuperación de toda su información lexicogenética hasta llegar a una palabra primitiva.

- ***Reconocedor de Entidades Nombradas (Named Entity Recognition):*** Subproceso que permite identificar y clasificar en un documento de texto, expresiones que identifican instancias de conceptos relevantes para algún dominio de aplicación.

[12] **FreeLing**, Consultado en (Junio 2012): http://nlp.lsi.upc.edu/freeling/.

Algo que hay que tomar en cuenta, es que la aplicación de cada uno de estos procesos se realiza por cada una de las oraciones identificadas en el corpus. El proceso es el siguiente (ver ilustración 4.1).

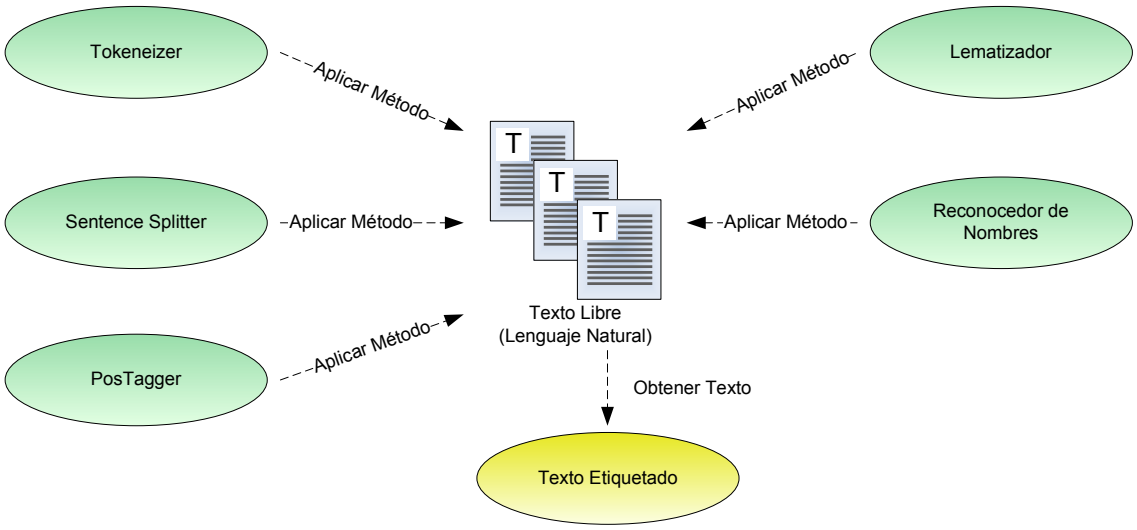


Ilustración 4. 1. Muestra de procesos de la herramienta de etiquetado del lenguaje FreeLing.

Por ejemplo, supongamos que el sistema se encuentra con la siguiente frase:

El procedimiento para designar al nuevo Rector es establecido en el presente estatuto.

Tabla 4. 1 Texto de ejemplo en Lenguaje Natural.

Una vez ejecutada esta fase obtendríamos el siguiente resultado:

el·el·DA0MS0	es·ser·VSIP3S0
procedimiento·procedimiento·NCMS000	establecido·establecer·VMP00SM
para·para·SPS00	en·en·SPS00
designar·designar·VMN0000	el·el·DA0MS0
a·a·SPS00	presente·presente·AQ0CS0
el·el·DA0MS0	estatuto·statuto·NCMS000
nuevo·nuevo·AQ0MS0	
rector·rector·NCMS000	

Tabla 4. 2. Texto etiquetado por FreeLing.

Donde: el primer elemento es la palabra en su forma normal, el segundo elemento es la palabra en su forma lematizada y el tercer elemento es la etiqueta morfológica.

4.3 Fase de patrones lingüísticos

Una de las formas de extraer conocimiento es la basada en la formación de patrones lingüísticos. Estos patrones nos ayudarán a extraer los términos que emplearemos después cono conceptos al terminar este primer módulo. Para ello, existen básicamente tres métodos: la extracción lingüística, la extracción estadística y los métodos híbridos. En esta investigación se ha implementado una metodología híbrida que combina tanto métodos de información

lingüística como estadística. Esta metodología extrae aquellos términos formados por una sola palabra, como los formados por más de una palabra '*términos compuestos*'. Podemos definir patrón lingüístico como:

"la combinación de uno o varios elementos gramaticales, que son representados por su estructura morfosintáctica, para definir uno o varios términos válidos, los cuales pueden pertenecer a un dominio o ser genéricos".

Lo patrones lingüísticos pueden ser obtenidos de diversas formas como: 1) A partir de términos válidos ya definidos, 2) que sean identificados a partir de un texto previamente etiquetado, o como los empleados en esta metodología, 3) empleando métodos informáticos basados en la estadística y/o heurística o bien 4) que sean proporcionados por un experto lingüista dependiendo de nuestras necesidades.

4.4 Módulo de Aprendizaje automático de Patrones

Para que el sistema adquiera los patrones de forma automática, es necesario introducir un listado de patrones guía con los cuales el sistema identificará los mejores patrones existentes en el corpus, un ejemplo de los patrones guía se puede ver en la tabla 4.3.

Longitud del Término	Estructura Morfosintáctica
2	xx·xx
3	xx·xx·xxx
4	xxxx·xx·xxx·xx

Tabla 4. 3. Longitud del término especificada por patrones morfosintácticos.

Los patrones guía representan la longitud de las términos que deseamos recuperar y la especialización de cada palabra incluida en el término, es decir, en la tabla 4.3, se pueden ver tres filas, donde se representa la longitud del término, es decir, los patrones guía recuperarán patrones que contengan términos formados por 2, 3 y 4 palabras. También se puede ver que cada patrón guía está formado por un conjunto de "xx" separadas por un punto " · ", cada 'x' representa un nivel de especialización que se encuentra definido en las etiquetas Eagles[13]. Donde cada etiqueta representa hasta 8 niveles de especialización para algunas categorías morfosintácticas, como en el caso de los pronombres, y un nivel de especialización de 7 para el caso de los sustantivos y los verbos.

Dependiendo de la especificación que requiramos para cada dominio o aplicación, se pueden ajustar el número de 'x', es decir, si deseamos recuperar conceptos muy específicos, tenemos que definir, 3, 4 ó más 'x', para conceptos más genéricos o simples, con 1 ó 2 'x' es suficiente.

El algoritmo emplea una serie de análisis estadísticos y heurísticos para definir los mejores patrones que recuperarán los conceptos, ver [Ochoa et al., 2011b] para obtener mas detalle. Este algoritmo tiene la peculiaridad de que puede ser configurable, es decir, podemos elegir la longitud y especialización, además tiene la opción de filtrar algunos elementos morfosintácticos que no sean necesarios, ya sean al principio, al medio o al final de cada patrón, esta característica permite realizar una mejor selección, ya que desde un principio

[13] **Etiquetas EAGLES**, Consultado en (Junio 2012): http://www.lsi.upc.es/~nlp/tools/parole-sp.html.

eliminamos todos aquellos que no sean de importancia para la investigación. Un ejemplo de los patrones extraídos de forma automática con algunos ejemplos de los conceptos que extrae se presenta a continuación (ver tabla 4.4).

Linguistic Pattern	**Descripción**	***Terms***
NC + A + A	Nombre Común + Adjetivo + Adjetivo	*servicio social universitario, secretario general académico*
NC + AQ	N. Común + Adjetivo Calificativo	*colégio académico, contrato colectivo*
NC + SP + NC + A	N. Común + Preposición + N. Común + Adjetivo	*institución de educación superior, profesor de tiempo completo*
NC + SP + A + NC	N. Común + Preposición + Adjetivo + N. Común	*alumno de nuevo ingreso, tutor de práctica profesional*
NC + SP + NP	N. Común + Preposición + N. Propio	*comisión de gobernación, programa de doctorado*
NC+SP+NC+SP+DA+NC	N. Común + Preposición + N. Común + Preposición + Determinante + N. Común	*órgano de gobierno de la universidad*
NC+A+SP+NC+SP+NC	N. Común + Adjetivo + Preposición + N. Común + Preposición + N. Común	*título profesional a nivel de licenciatura*

Tabla 4. 4. Patrones lingüísticos

4.5 Fase de extracción de conceptos compuestos[14,15]

4.5.1 C-Value / NC-Value[16]

Uno de los métodos más extendidos para el reconocimiento de términos compuestos, es el que se presenta en [Frantzi et al., 2000]. Este método se basa en dos fases: la primera es el método *C-value*, que tiene como objetivo, realizar una extracción de términos compuestos a partir de un conjunto de patrones lingüísticos definidos, y la segunda, es el método *NC-value,* que incorpora información de contexto al método C-value, con el objetivo de mejorar aun más la extracción de los términos compuestos. Su ecuación es la siguiente:

$$C-Value = \begin{cases} \log_2|a| * f(a) & si\ a\ no\ está\ incluido\ en\ b \\ \log_2|a| * \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & en\ otro\ caso \end{cases} \quad (1)$$

Donde:

- a es el término candidato
- $|a|$ es la longitud del término candidato
- $f(a)$ es la frecuencia del término candidato a en el corpus
- T_a es el conjunto de candidatos de mayor longitud que contienen a a.
- $P(T_a)$ es el número de los candidatos de mayor longitud que contienen a a.
- $\sum f(b)$ es la ocurrencia total de a como sub término de cualquier término candidato b tal que $|a| < |b|$.

[14] http://user.phil-fak.uni-duesseldorf.de/~rumpf/SS2005/Informationsextraktion/Pub/KagUmi96.pdf
[15] http://people.kmi.open.ac.uk/petr/papers/atr_znalosti_2009.pdf
[16] http://www.springerlink.com/content/kwefkd55ludr8vu4/fulltext.pdf

La primera parte de esta ecuación, se utiliza para aquellos términos que no están incluidos en algún término de mayor longitud. Por ejemplo, si tenemos el término *resonancia nuclear* y también tenemos el término *resonancia nuclear magnética*, esta ecuación ya no sería aplicable, y tendríamos que aplicar la segunda parte de la ecuación.

Tabla 4.5 mustra algunos de los valores C-Value Obtenidos.

C-Value	*Term*
162.00	*dirección de servicio escolar* (school service address)
156.91	*plan de estudio* (study plan)
134.00	*personal académico* (academic staff)
112.53	*jefe de departamento* (department head)

Tabla 4. 5. *Ejemplo del valor C-value obtenidos para algunos terminos.*

La siguiente parte de este algoritmo es el método *NC-value.* Este método, necesita palabras de contexto y es necesario obtener un peso, el "*Factor de ponderación de contexto*", que mide la probabilidad de que una palabra de contexto aparezca con un término específico, y se calcula utilizando la siguiente ecuación:

$$weight(w) = \frac{t(w)}{n} \qquad (2)$$

Donde:

- w es la palabra de contexto.
- t(w) es el número de veces que aparece la palabra de contexto con el término.
- *n* es el número total de veces que se considero.
- *weight (w)* es el factor de ponderación de contexto.

El objetivo principal del método NC-value, es refinar la lista de términos candidatos que obtuvimos en la anterior fase (los candidatos del método C-value). Para ello, es necesario incorporar información de contexto, esto es, las palabras que están alrededor (tanto en la parte izquierda como en la derecha) de los términos candidatos. En [Grefenstette, 1994], se dice que existe una mayor probabilidad de encontrar verbos, adjetivos y nombres, alrededor de un término, por lo tanto, siguiendo estas pautas, el sistema identifica a todas aquellas palabras con esas categorías gramaticales alrededor de cada uno de los términos candidatos, para obtener las palabras de contexto.

Por ejemplo, para el candidato a término *plan de estudio* en la tabla 4.6 se muestran algunas palabras de contexto con su probabilidad y el número de veces que aparecen con el término:

t(w)	weight (w)	palabra de contexto
1.0	0.003623	docencia
2.0	0.007246	máximo
11.0	0.039855	correspondiente
12.0	0.043478	programa
15.0	0.054347	asignatura
5.0	0.018115	nivel
2.0	0.007246	técnico

1.0	0.003623	artículo
2.0	0.007246	conclusión
4.0	0.014492	posgrado

Tabla 4. 6. Palabras de contexto de un término.

Una vez obtenido este factor de ponderación para cada palabra de contexto, se tienen que unificar esos valores en uno solo, para ello, se emplea la ecuación 3.

$$cw_1 \cdot w_1 + cw_2 \cdot w_2 + \ldots + cw_n \cdot w_n \qquad (3)$$

Dónde:

- cw_x es el número de veces que aparece la palabra de contexto *x* con el término.
- w_x es el factor de ponderación obtenido para la palabra de contexto *x*.

Como hemos visto en la anterior tabla (ver tabla 4.6) cada candidato a término aparece con sus palabras de contexto, entonces, al aplicar la ecuación (3), el resultado será el valor de ponderación final de esta parte del método NC-value.

Una vez obtenido este valor, se reordena la lista de candidatos C-value, dando lugar al valor final buscado NC-value, mediante el cálculo de la ecuación (4).

$$NC-value(a) = 0.8 * C-value(a) + 0.2 * \sum_{b \in C_a} f_a(b) weight(b) \qquad (4)$$

Dónde:

- *a* es el término candidato
- C_a es el conjunto de palabras de contexto de *a*
- *B* es una palabra de C_a
- $f_a(b)$ es la frecuencia de *b* como palabra de contexto de *a.*
- weight (b) es el peso de *b* como palabra de contexto

En la ilustración 4.2, se presenta de forma gráfica, el proceso seguido para obtener estos conceptos compuestos.

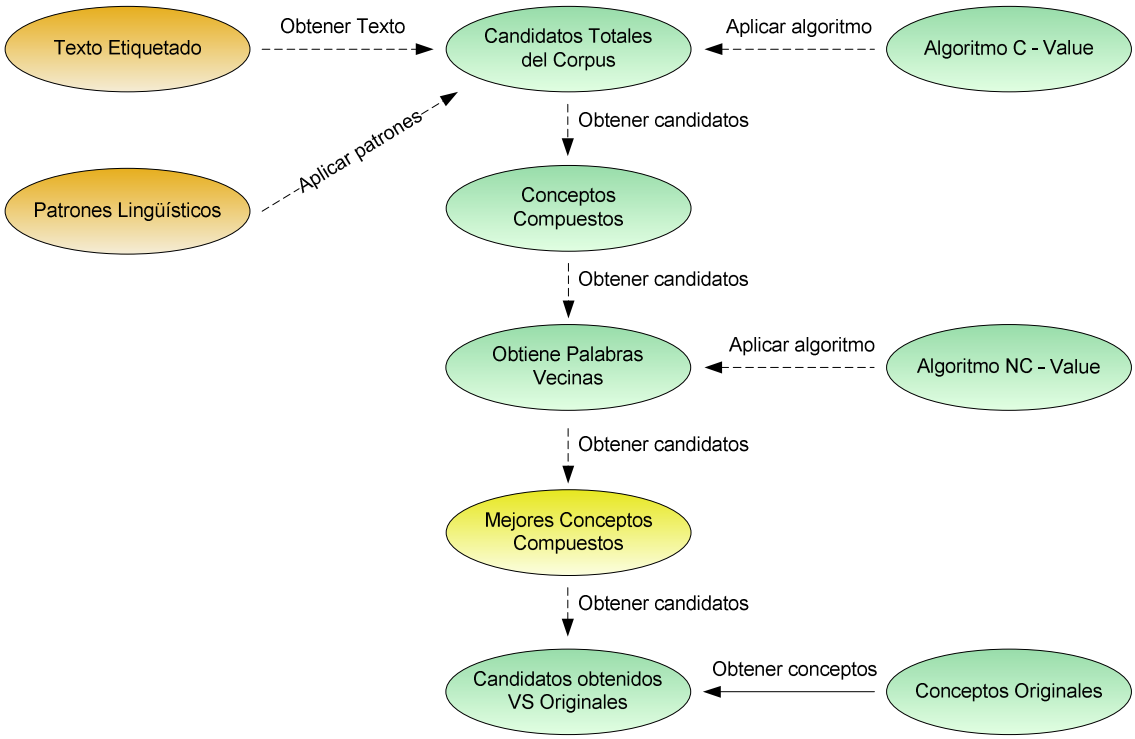


Ilustración 4. 2.Extracción de conceptos compuestos.

4.6 Fase de Extracción de conceptos simples

4.6.1 TF-IDF

El método TF-IDF, fue presentado por Gerard Salton en [Salton, 1991]. En este trabajo, se ha desarrollado este método para obtener los mejores candidatos a términos formados por una sola palabra. Esta medida TF-IDF, es un peso de uso frecuente en los procesos de recuperación de la información y minería de textos. Representa una medida estadística para evaluar la importancia de una palabra, en un documento dentro de un corpus. La importancia se incrementa proporcionalmente debido al número de veces que una palabra aparece en el documento, pero se compensa con la frecuencia de la palabra contenida en la totalidad del corpus.

Esta medida se define por la siguiente ecuación:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \qquad (5)$$

Dónde: $tf_{i,j}$ representa la *frecuencia del término* (dando importancia al término t_i en el documento d_j) e idf_i, que representa la *frecuencia del documento inversa*, (lo importante que es la palabra, en todo el corpus). A continuación, se muestran las fórmulas para calcular el valor $tf_{i,j}$ e idf_i.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (6)$$

Dónde:

- El ***numerador*** $n_{i,j}$**:** es el número de ocurrencias del término considerado (t_i) en el documento d_j.
- El ***denominador*** es la suma del número de términos del documento d_j, es decir, el tamaño del documento $| d_j |$.

Y para la *frecuencia del documento inversa* tenemos:

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \qquad (7)$$

Dónde:

- $| D |$: es el número total de documentos en el corpus.
- $|\{d : t_i \in d\}|$ es el número de documentos en los que aparece el término t_i.
- En el supuesto de que el término no este en el corpus, se producirá una división por cero. Por lo tanto es común utilizar $1+|\{d : t_i \in d\}|$.

En el trabajo presentado por [Knoth et al., 2009], se expone que la medida TF-IDF es una buena medida para evaluar la importancia de una palabra dentro de un documento contenido en un corpus. Por ejemplo, en [Ramos, 2003], se implementa el método TF-IDF para determinar, que palabras dentro de un corpus pueden ser más favorables de usar en una consulta de base de datos.

El proceso para extraer los mejores términos formados por una sola palabra de forma gráfica (ver ilustración 4.3), se muestra a continuación.

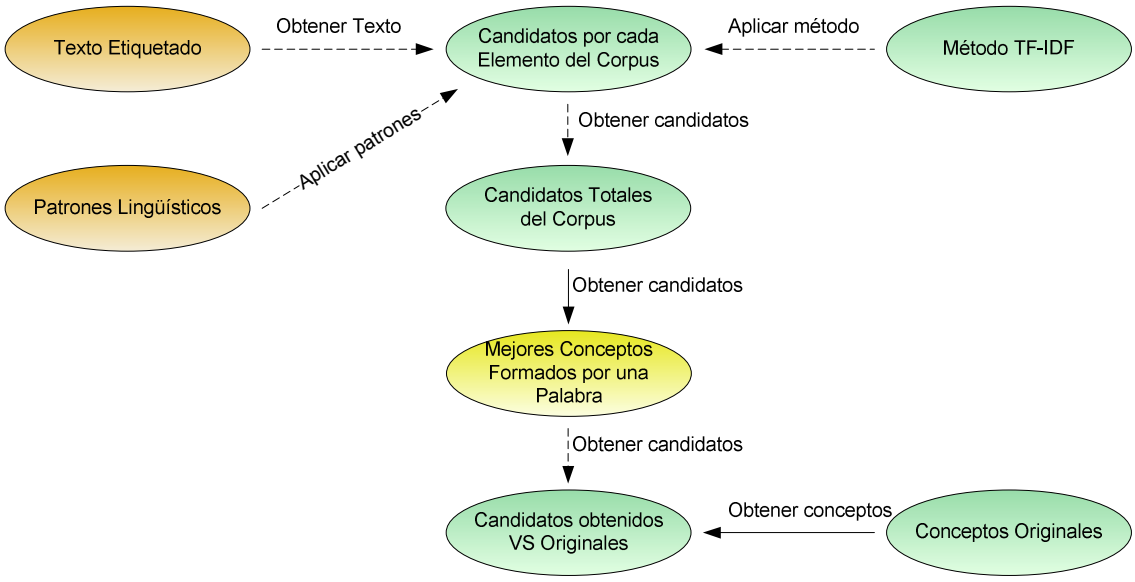


Ilustración 4. 3. Extracción de Conceptos formados por una sola palabra.

Ese conjunto de términos seleccionados, serán identificados como los conceptos superiores y estos, serán los que emplearemos en futuros procesos.

4.7 Módulo de Extracción de Relaciones

En los textos escritos en lenguaje natural, las relaciones entre conceptos suelen estar asociadas a los verbos [Valencia-García et al., 2008]. Por esa razón, una vez que los

conceptos han sido identificados en el corpus, es necesario identificar las relaciones semánticas entre estos conceptos. Nuestro trabajo se basa en el uso de roles semánticos para identificarlas.

Un rol semántico se define como la relación entre un componente sintáctico y su predicado [(Ochoa et al., 2011b); (Moreda et al., 2011); (García-Miguel et al., 2010)]. Actualmente el conjunto de roles semánticos mayormente utilizados en la literatura, son los desarrollados en el proyecto "*the Proposition Bank"* (PropBank) proyect [Palmer et al., 2005] el cual se ha desarrollado sólo para el inglés.

En nuestro caso, es necesario emplear un conjunto de roles semánticos en español, por esa razón, se ha decidido emplear la Base de Datos Sintáctica en Español ADESSE [(García-Miguel et al., 2010); (Albertuz-Carneiro, 2007); (García-Miguel et al., 2005)]. Esta base de datos, contiene información sintáctico-semántica sobre los verbos.

En la ilustración 4.4 se muestra un ejemplo del rol semántico SUSTITUCIÓN [Albertuz-Carneiro, 2007] que se puede expresar a través de los verbos *sustituir*, *reemplazar*, *relevar* o *suplir* que son integrantes del rol semántico.

```
<predicate lemma="sustituir">
<roleset id="sustitución">
<roles>
  <role n="0" descr="holder"/>
  <role n="1" descr="sustituir"/>
</roles>
</roleset>
```

El **[$_0$rector]** es sustituido por el **[$_1$Secretario General Académico]**.

Ilustración 4. 4. Un ejemplo del verbo *disminuir* en ADESSE.

Este módulo se compone de un conjunto de fases secuenciales (ver ilustración 4.5) Fase de Identificación de Verbos y la Fase de Identificación de Relaciones. A continuación, se explica cada una de estas fases.

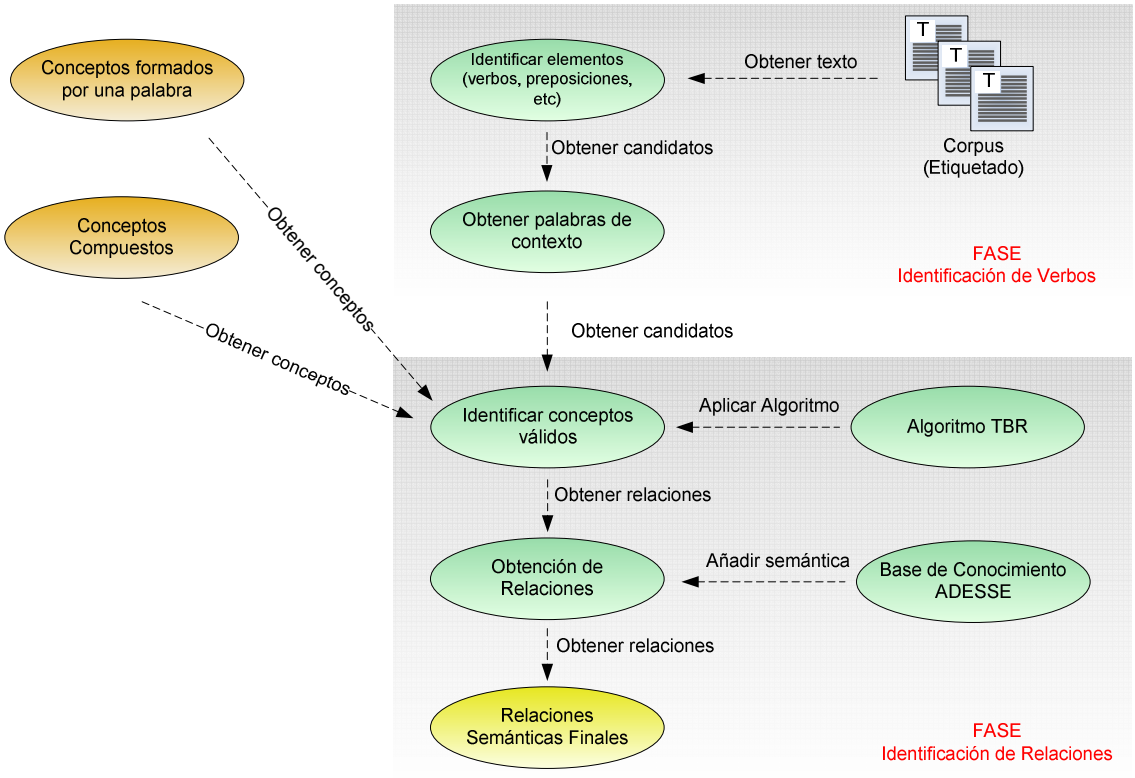


Ilustración 4. 5. Representación gráfica del proceso de Extracción de Relaciones.

4.7.1 Fase de Identificación de Verbos

Se comienza con la localización de todos aquellos elementos principales de las relaciones, para las relaciones no taxonómicas, se buscan los verbos principales, para las relaciones taxonómicas, se busca la unión del verbo "ser" en tercera persona "es" con un determinante "un ó una", para las relaciones partonómicas, se busca la unión de preposiciones y artículos "de la ó de el" etc, que se encuentran en todo el corpus. Después, el sistema obtiene un número configurable de palabras a la izquierda y a la derecha del verbo, con el objetivo de identificar algún concepto extraído en la fase anterior. Por ejemplo, de la siguiente frase se obtiene la expresión verbal *proporcionadas* cuya forma lematizada es *proporcionar*.

*…y difundir las convocatorias de movilidad estudiantil·** proporcionadas proporcionar **·por la dirección de movilidad, intercambio y…*

4.7.2 Fase de Identificación de Relaciones

Una vez identificados los conceptos válidos a la izquierda y a la derecha del verbo, se crea una relación candidata. Dependiendo de las palabras y, de la longitud de los patrones que se hayan elegido, se obtendrán pocas o una gran cantidad de relaciones candidatas. Para los casos, en los que exista más de un concepto tanto a la izquierda como a la derecha del verbo, se obtendrán los conceptos más cercanos al verbo. Además, los conceptos compuestos, tendrán prioridad sobre los conceptos formados por una sola palabra. Del ejemplo anterior, se obtienen las relaciones mostradas en la tabla 4.7.

Conceptos identificados a la izquierda	verbo	Conceptos identificados a la derecha
movilidad estudiantil	proporcionar	dirección
convocatorias		movilidad
movilidad		intercambio
estudiantil		dirección de movilidad

Tabla 4. 7. Conceptos candidatos obtenidos a partir de una relación candidata.

Para identificar una relación válida pueden existir dos casos: 1) que solo exista un concepto sencillo tanto a la derecha y a la izquierda del verbo, o bien 2) que exista una combinación de conceptos (sencillos y compuestos) en cada uno de los lados. Para ello, seguiremos el algoritmo creado e implementado para este fin, llamado TBR[17] (ver ilustración 4.6):

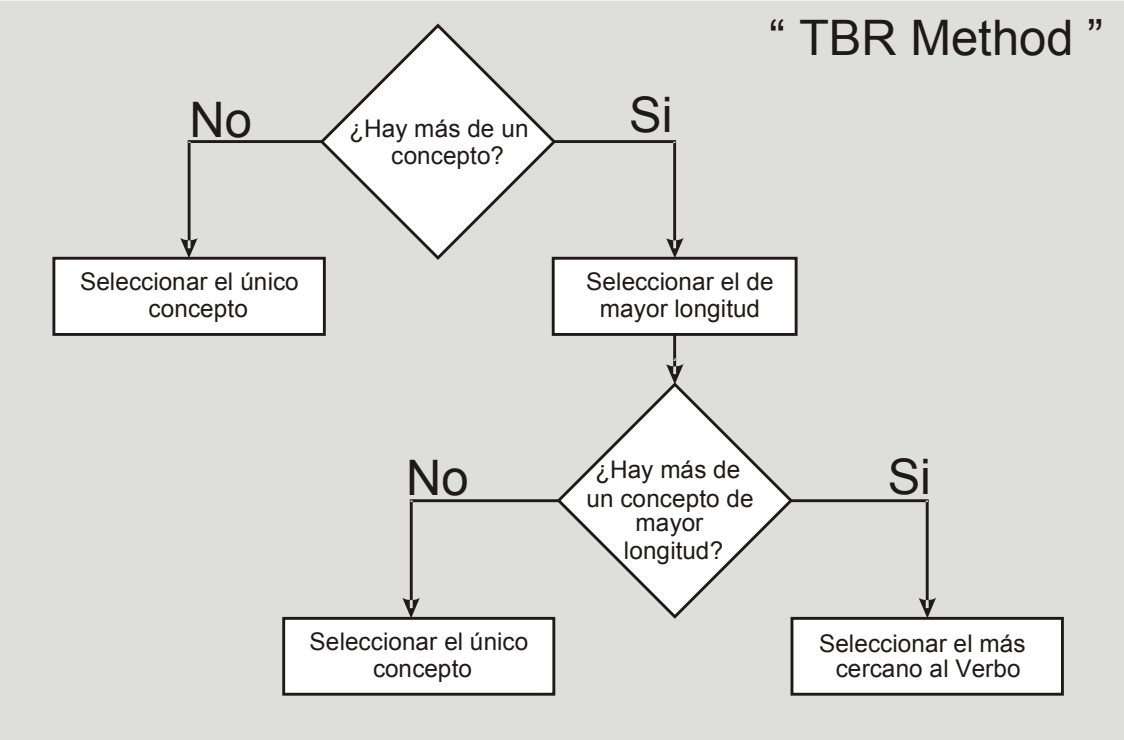


Ilustración 4. 6. Algoritmo de selección de relaciones TBR.

Para detectar las relaciones taxonòmicas, la unión de los elementos (verbo + determinante) '*es–un*' o '*es–una*' es identificado en las oraciones, en particular, ste tipo de relaciones es muy simple de obtener, ya que solo es necesario identificar la unión en la oración y obtener los conceptos validos alrededor de ésta. Por ejemplo Ver tabla 4.8.

[17] **TBR**, proveniente de las siglas en inglés de ***T**he **B**est **R**elation* o La Mejor Relación, desarrollado en esta tesis.

Oración	Relación Taxonómica	Explicación
"*La Junta Directiva es un órgano colegiado con facultades de nombrar al Rector.*"	junta directiva **es un** órgano colegiado	Con esto, el conocimiento que se extrae es que la *junta directiva* es *un tipo de órgano colegiado,* nos está dando una simple y pequeña definición de lo que se conoce como *junta directiva.*
La Universidad de Sonora es una institución autónoma de servicio público.	Universidad de sonora **es una** Institución autónoma	Esta relación nos indica que la Universidad de Sonora es un tipo de institución independiente.

Tabla 4. 8. Ejemplos de relaciones taxonómicas extraídas.

Las relaciones partonómicas han existido desde hace tiempo, lo novedoso es la técnica de extracción que se presenta en este artículo, como es de saberse, estas se caracterizan por ser del tipo "parte de" o "tiene un". En nuestro caso, nosotros las formamos de forma implícita (es decir, no explícitamente) a partir de la unión de preposiciones y artículos, generando el tipo de relaciones partonómicas deseadas, lo cual fue descubierto al analizar una gran cantidad de relaciones extraídas que contenían esta peculiaridad, en la tabla 4.9 se puede ver unos ejemplos.

Oración	Relación	Relación Semántica	Explicación
"Nos fue turnada para su estudio y dictamen la Iniciativa de Ley Orgánica de la Universidad de Sonora."	Ley Orgánica **de la** Universidad de Sonora	Universidad de Sonora **tiene una** Ley Orgánica.	De la cual obtenemos que el rol semántico asociado a la conjunción '*de la*', puede ser *"tener" o "tiene una.*
La obligatoriedad del Estado de sostener el "Fondo Universidad de Sonora", administrado por el Instituto de Crédito Educativo del Estado de Sonora…."	a) obligatoriedad **del** Estado = obligatoriedad **de el** Estado b) Instituto de crédito educativo **del** Estado de Sonora = Instituto de crédito educativo **de el** Estado de Sonora	a) el estado **tiene una** obligatoriedad, es decir, obligaciones b) el estado de Sonora **tiene un** Instituto de Crédito Educativo	Esta relación nos indica que la Universidad de Sonora es un tipo de institución independiente.

Tabla 4. 9. Ejemplos de relaciones Partonómicas extraídas.

Con el conjunto de estos cuatro tipos de relaciones estamos aportando nuevas formas de adquirir conocimiento para el proceso actual de Ontology Learning en español, el cual es escaso, en todos los dominios, pero particularmente aún más, en este dominio (Universitario) el cual es muy difícil de procesar.

Una vez identificadas estas relaciones a partir de los elementos requeridos, se hace una búsqueda a partir de la forma lematizada del verbo, sobre la base de datos ADESSE, para encontrar su significado. Una vez encontrado, se asigna a la relación.

5 Herramienta de Extracción y generación de Conocimiento

En este apartado, se explica el desarrollo de una herramienta de software que pueda servir como entorno para la extracción de conocimiento y creación automática de ontologías a partir de textos escritos en lenguaje natural, utilizando la metodología de Ontology Learning diseñada por [Ochoa, J.L. et al., 2011]. Esta aplicación, toma un corpus (colección de archivos) de texto como entrada, lo procesa y como resultado final, proporciona un archivo en lenguaje OWL con la ontología del dominio obtenida.

En este apartado, se explica el desarrollo de una herramienta de software que pueda servir como entorno para la creación automática de ontologías a partir de texto en lenguaje natural, utilizando la metodología de Ontology Learning diseñada por [Ochoa, J.L. et al., 2011]. Esta aplicación, toma un corpus de texto como entrada, lo procesa y como resultado final, proporciona un archivo en lenguaje OWL con la ontología del dominio obtenida. Adicionalmente, proporciona una serie de archivos intermedios, en los que se pueden consultar los resultados de cada una de las etapas de la arquitectura (ver ilustración 5.12, Pág. 106).

La interfaz de esta herramienta software, se ha diseñado para que pueda ser usada tanto por usuarios expertos en procesamiento del lenguaje natural, interesados en representar el conocimiento contenido en textos por medio de ontologías, como por expertos del dominio que no tengan muchos conocimientos en las tecnologías desarrolladas en el trabajo de tesis del profesor [Ochoa, J.L., 2011].

5.1.1 Interfaz Inicial

En la ilustración 5.1, se muestra la interfaz principal del sistema, que permite el acceso a todas las opciones disponibles a través de sus cuatro menús:1) *Archivo*, 2) *Selección del proceso,* 3) *Idioma* y 4) *Ayuda*.

Ilustración 5. 1. Interfaz principal del sistema

Ilustración 5. 2. Gestión de Proyectos en el sistema (menú Archivo).

En el menú *Selección del Proceso*, se pueden configurar las opciones necesarias para realizar el proceso de construcción de ontologías a partir de texto. En este menú, aparecen cuatro opciones distintas (ver ilustración 5.3):

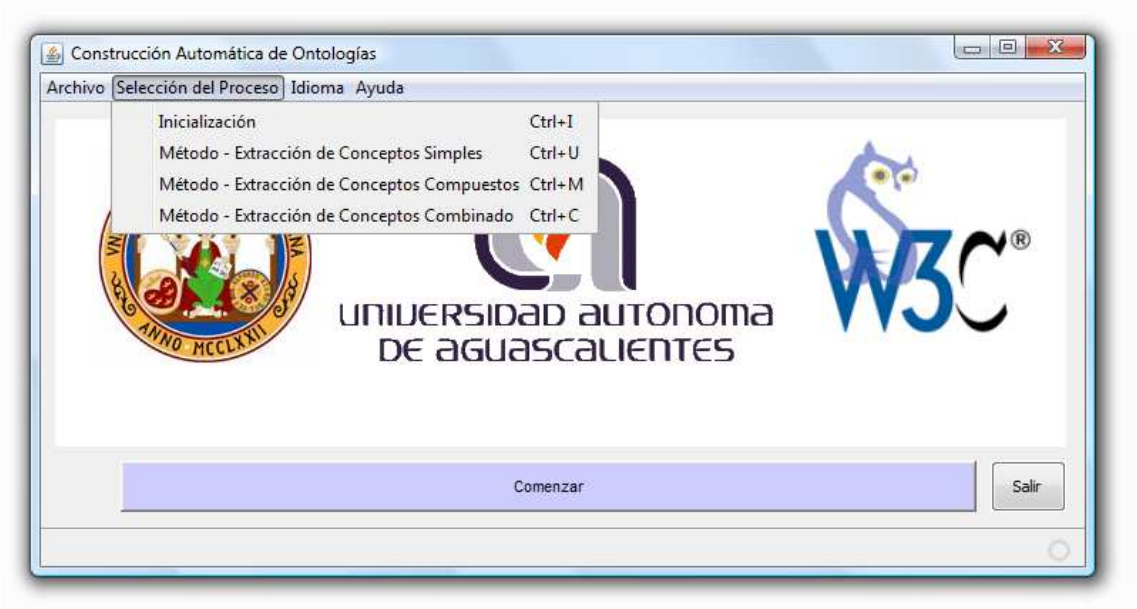


Ilustración 5. 3. Selección de procesos en el sistema (menú Selección del Proceso).

En el menú *Idioma*, se puede seleccionar el idioma en el que se desean manipular las interfaces del sistema. En este menú, aparecen dos opciones (ver ilustración 5.4 y 5.5):

Ilustración 5. 4. Selección del idioma de las interfaces del sistema (Español).

Ilustración 5. 5. Selección del idioma de las interfaces del sistema (Inglés).

En el menú *Ayuda*, es posible visualizar la información general, incluida la ayuda del sistema, al hacer clic en este menú, se muestran tres opciones (ver ilustración 5.6):

Ilustración 5. 6. Información General del sistema (menú Ayuda).

5.1.2 Interfaces de configuración

- **Inicialización**

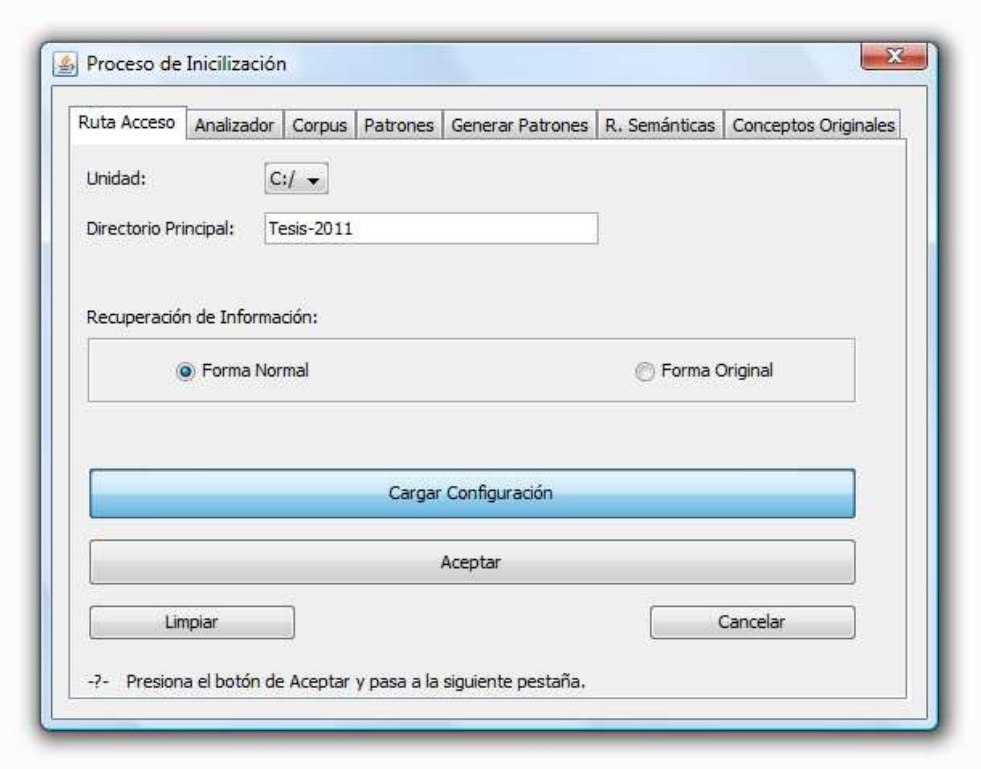


Ilustración 5. 7. Proceso de Inicialización General del sistema.

Dentro de la pestaña *Analizador* (ver ilustración 5.8), se debe indicar la ruta donde se encuentra el software *FreeLing*.

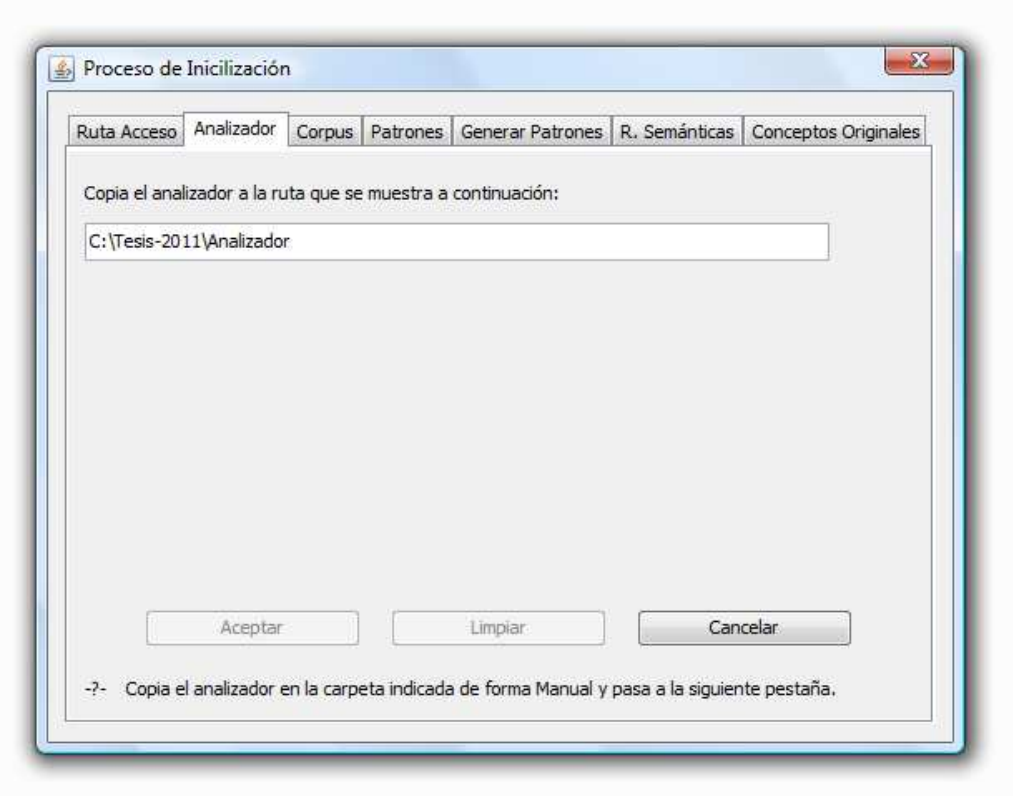


Ilustración 5. 8. Ruta del Analizador de Textos FreeLing en el sistema.

Seguidamente, en la pestaña *Corpus* (ver ilustración 5.9), el usuario tiene que aportar los documentos de texto que formaran parte del corpus, para ello, disponemos de dos formas principales de hacerlo:

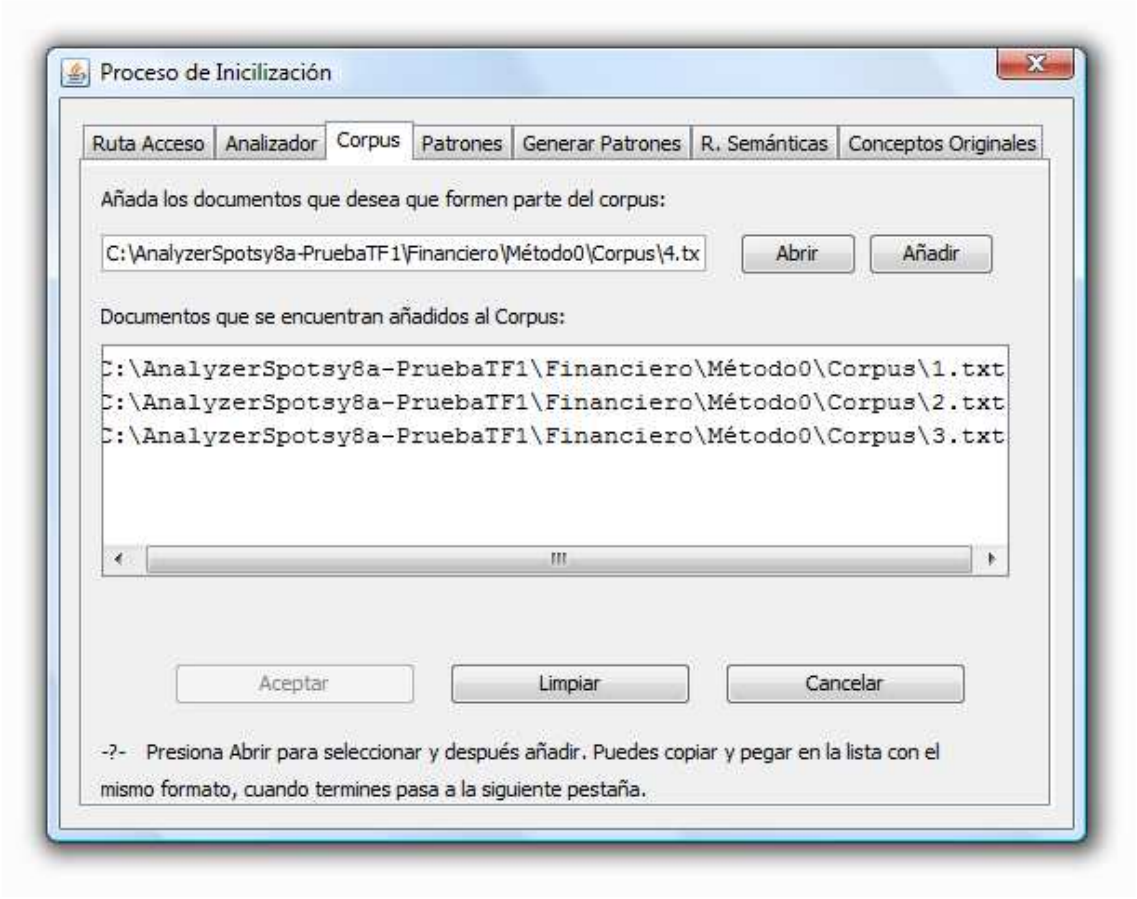


Ilustración 5. 9. Pestaña Corpus, documentos de texto que serán incluidos al sistema.

Los ficheros que son añadidos al corpus, no tienen algún formato en especial, son ficheros de texto cualquiera, un ejemplo de estos se puede ver en la ilustración 5.10.

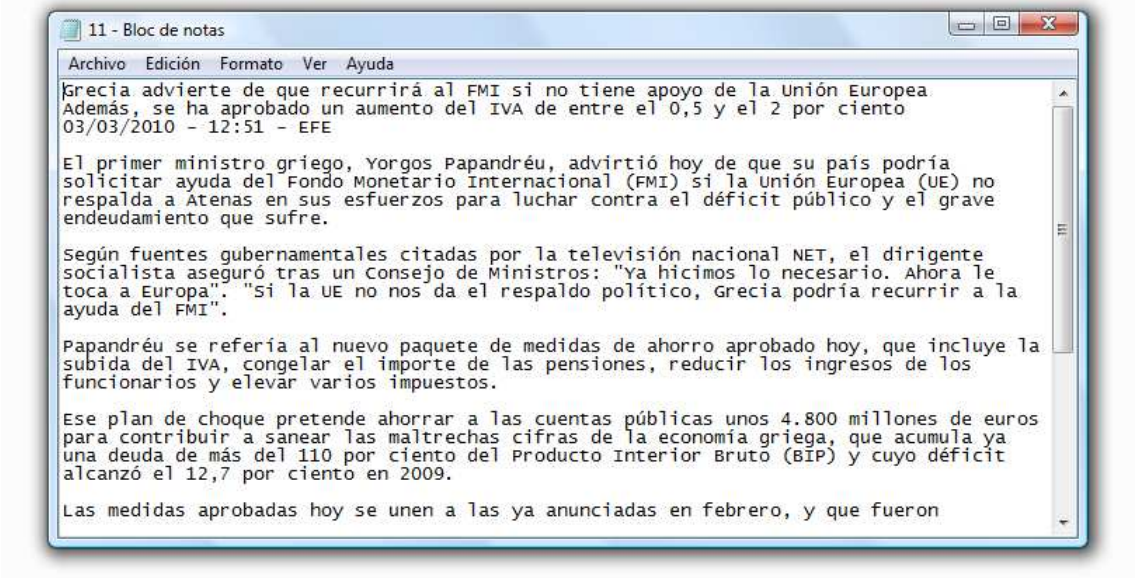


Ilustración 5. 10. Ejemplo de un documento perteneciente al corpus.

La pestaña *Patrones* (ver ilustración 5.11), tiene un objetivo primordial en el sistema, ya que es en esta, donde introduciremos los patrones lingüísticos que empleará el sistema para identificar los conceptos que serán utilizados en el resto del proceso de extracción de

conceptos. Es necesario que los patrones se encuentren catalogados Morfológicamente para poder introducirlos al sistema, para ello se tienen tres elementos principales:

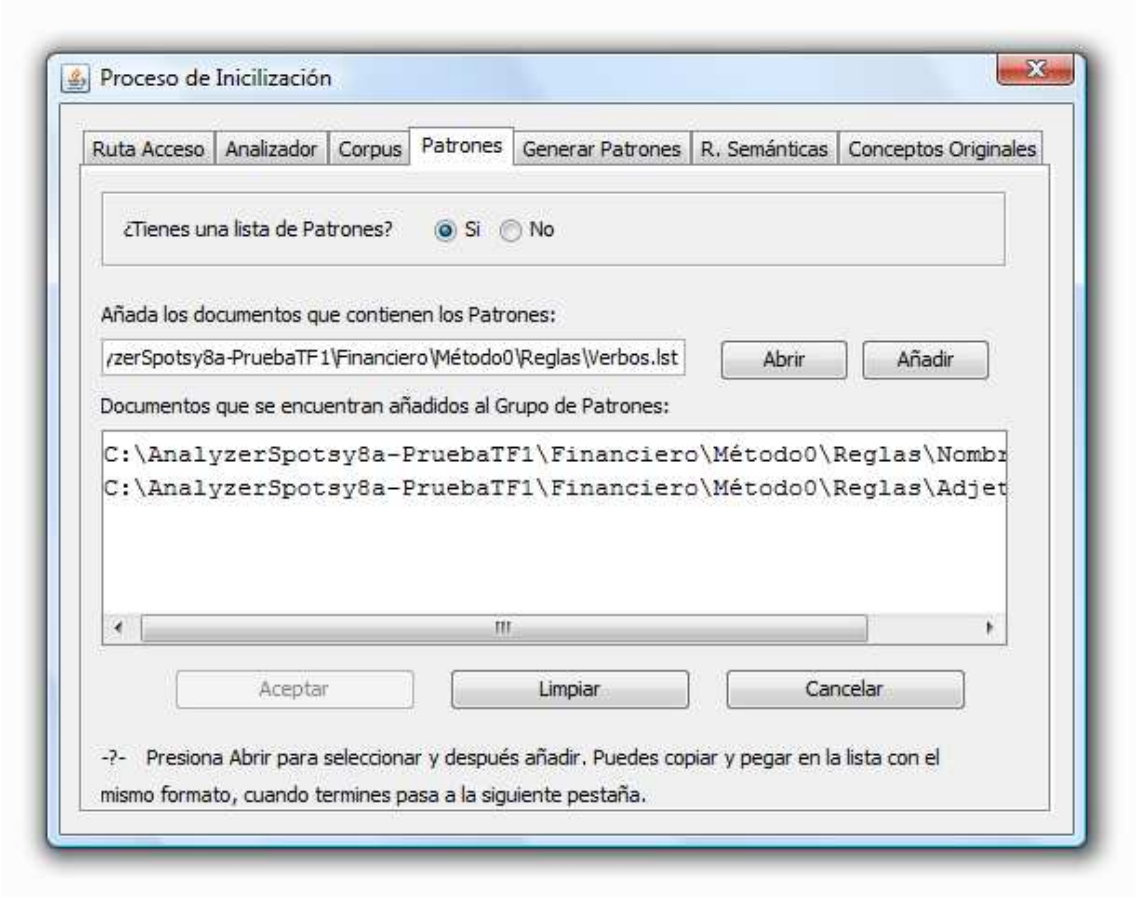


Ilustración 5. 11. Pestaña Patrones, documentos de texto con patrones que serán incluidos al sistema.

El formato que debe contener cada fichero es simple, un patrón en cada línea y cada elemento gramatical de cada patrón, se separa por un espacio, al final de cada patrón, no deben existir espacios en blanco, ya que esto podría causar resultados erróneos al sistema. Un ejemplo de estos ficheros puede verse en las ilustración 5.12 y 5.13.

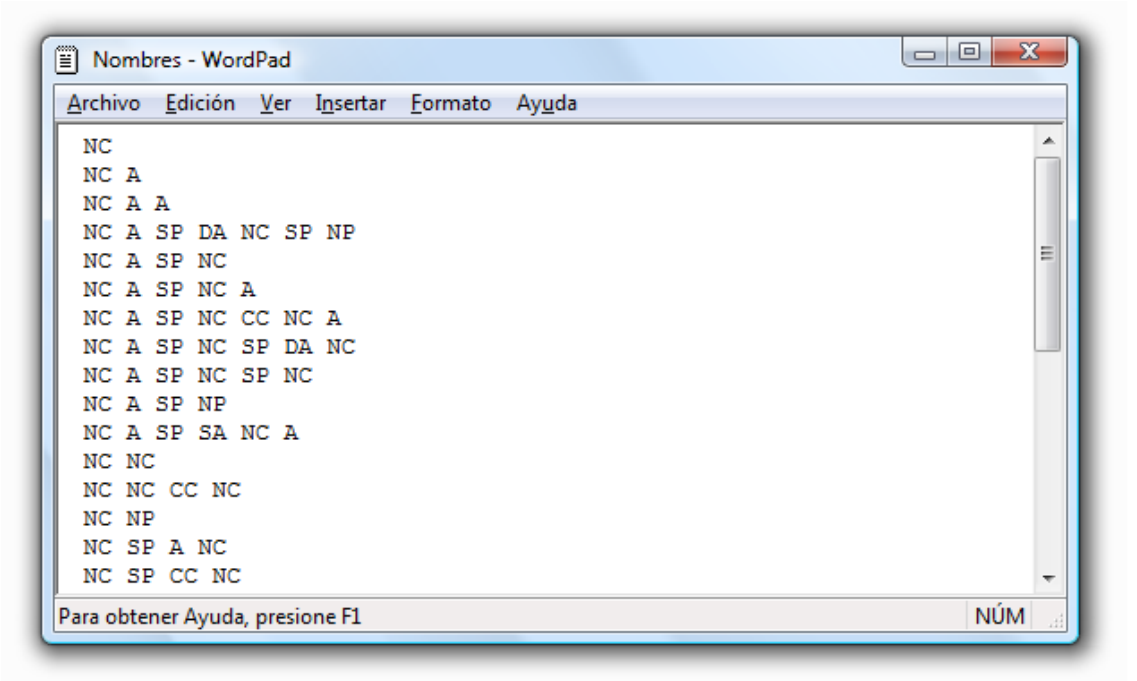


Ilustración 5. 12. Fichero de Patrones Lingüísticos (Nombres.lst).

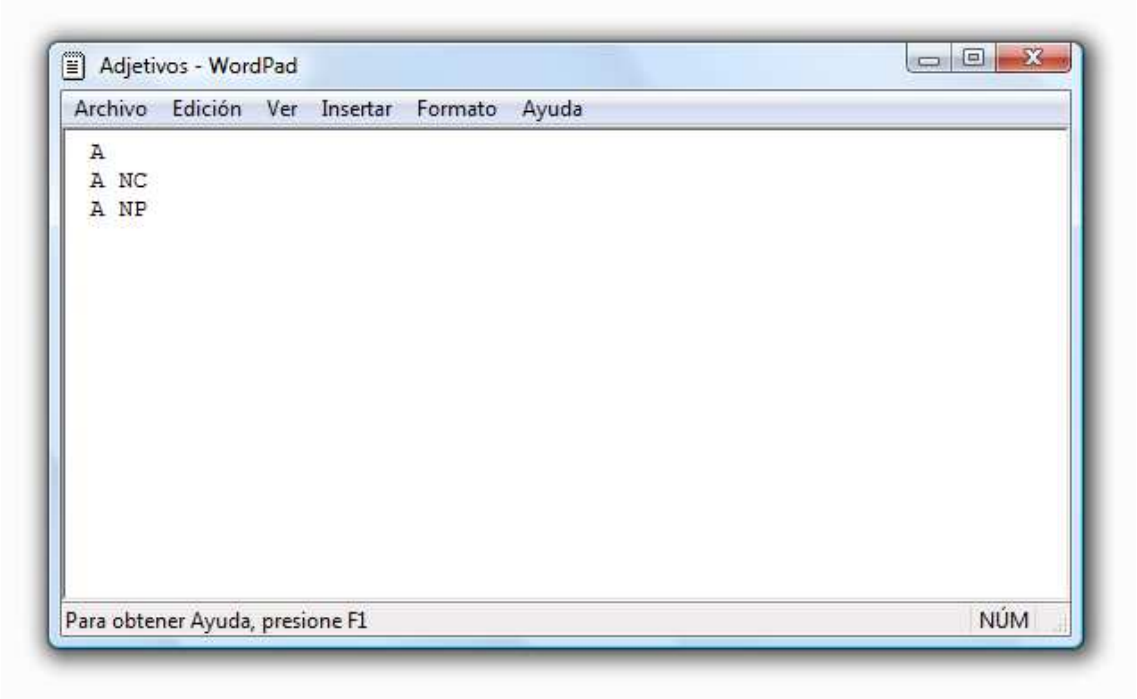


Ilustración 5. 13. Fichero de Patrones Lingüísticos (Adjetivos.lst).

Si no se han definido los patrones lingüísticos para este dominio, se podrán generar patrones a partir de la pestaña *Generar Patrones* (ver Ilustración 5.14). En esta pestaña, se define el conjunto de patrones guía que se utilizan para generar los patrones definitivos que serán utilizados por el sistema para recuperar los conceptos. A continuación, se definen los elementos de esta ventana:

Ilustración 5. 14. Pestaña Generar Patrones, definición de patrones guía.

Una vez indicados los patrones, el sistema generará los patrones automáticamente (ver ilustración 5.15).

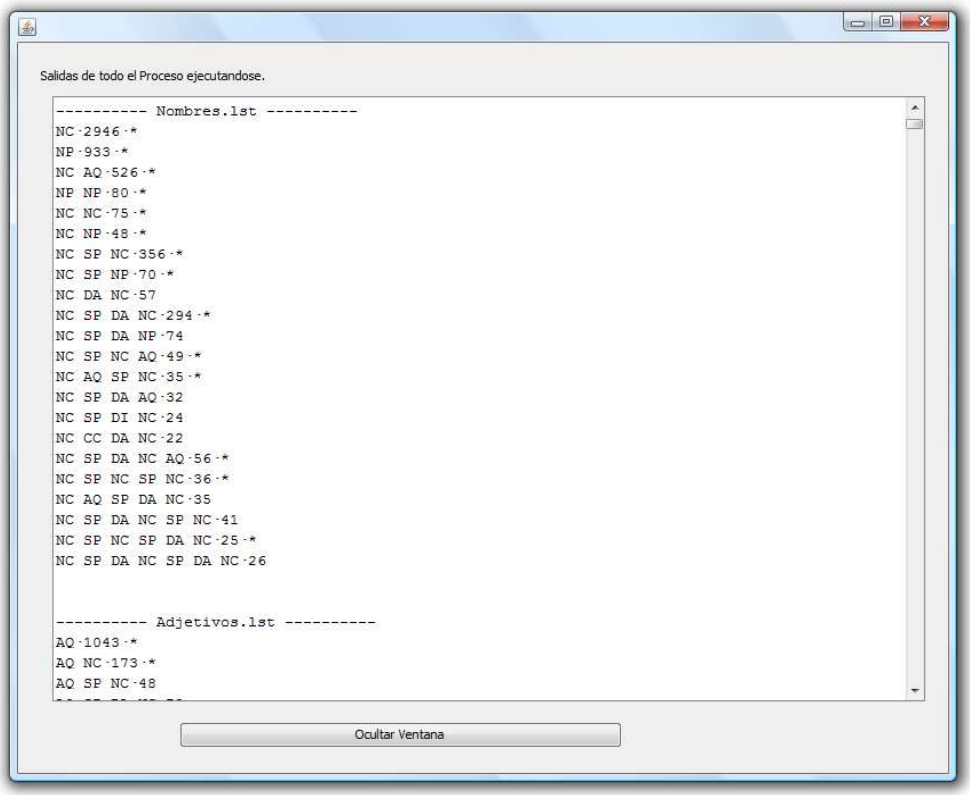


Ilustración 5. 15. Resultado obtenido al Generar Patrones con los patrones guía.

En la pestaña *Relaciones Semánticas,* se puede configurar la base de datos que contiene los roles semánticos que serán asignados a las relaciones entre conceptos. Por defecto, la base de datos que se ha utilizado es ADESSE para el idioma español, pero fácilmente podría configurarse otra base de datos como VerbNet para el idioma inglés (ver ilustración 5.16).

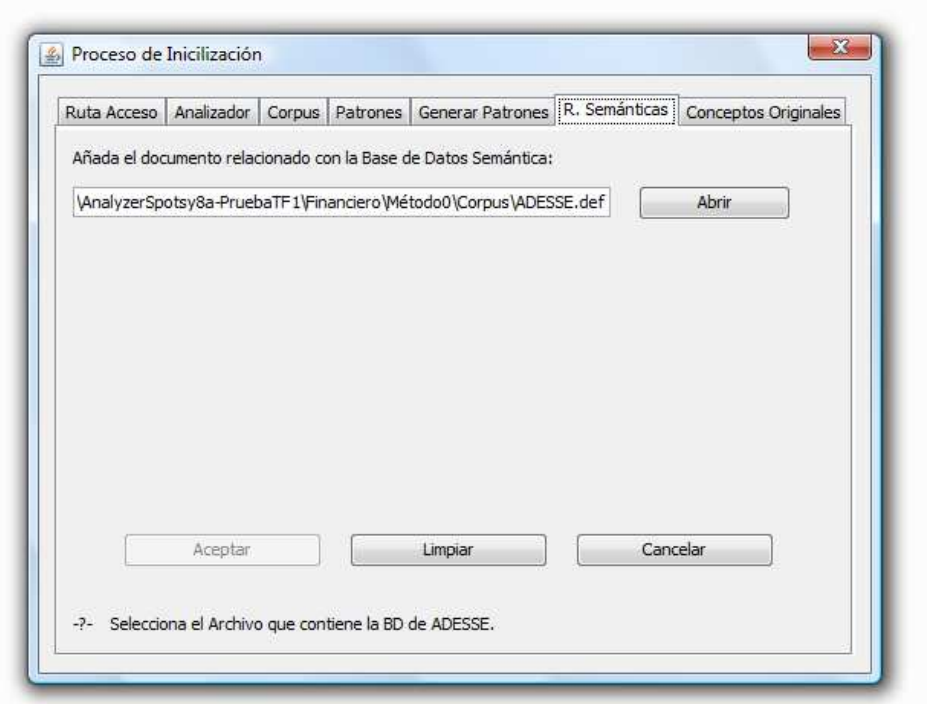


Ilustración 5. 16. **Pestaña Relaciones Semánticas, configuración de la Base de Datos semántica al sistema.**

El documento que contiene la base de datos semántica, no requiere de un formato especialmente complejo, es decir, únicamente son necesarios: el verbo en formato lematizado, y el rol semántico asociado por línea. Un ejemplo de una línea del documento de la base de datos ADESSE puede ser: "deplorar·1·8·sensación", en donde, el verbo en formato lematizado es "*deplorar*" y el rol semántico asignado es "*sensación*". Los números intermedios, representan un valor de acepción y un Número de ejemplos, respectivamente, los cuales, no son empleados en este trabajo. La división de cada uno de estos valores, es representado por un punto central "·" o un espacio " ". A continuación, se presentan varios ejemplos de esta base de datos (ver ilustración 5.17).

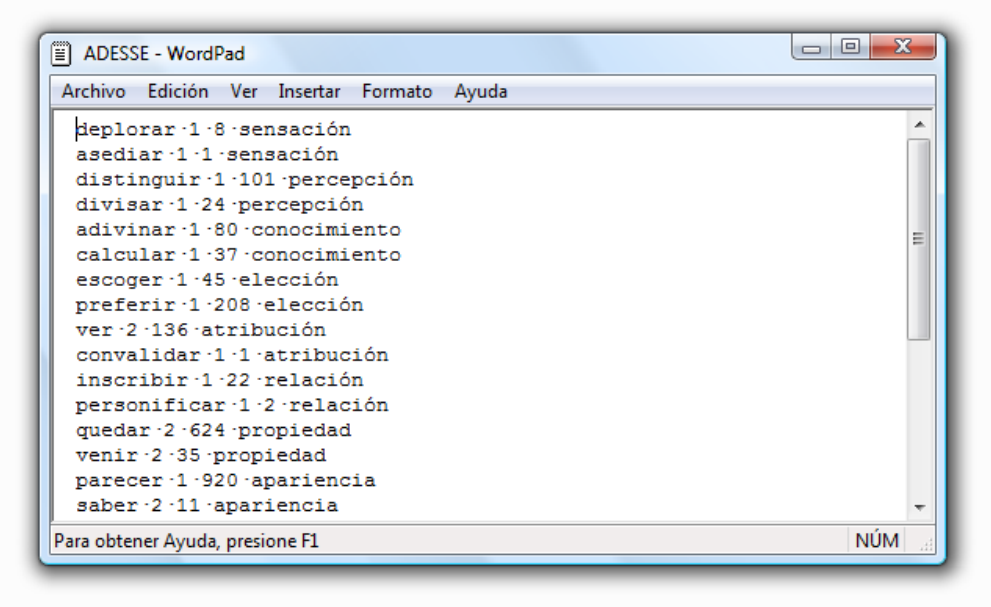


Ilustración 5. 17. Ejemplo del formato de base de datos ADESSE.

En la última pestaña *Conceptos Originales,* se pueden configurar un par de ficheros de forma totalmente opcional, con los cuales, se comprobará la veracidad del sistema, es decir, que se compararán los conceptos formados por una sola palabra y los conceptos compuestos, obtenidos por el sistema, con los conceptos originales proporcionados por el usuario, los cuales, han sido obtenidos a partir de estándares terminológicos, mas, los que el usuario ha identificado de forma manual en el corpus. En el supuesto de no desear obtener estadísticas del sistema, las opciones de esta pestaña pueden quedar en blanco. A continuación, se definen los elementos de esta ventana (ver ilustración 5.18).

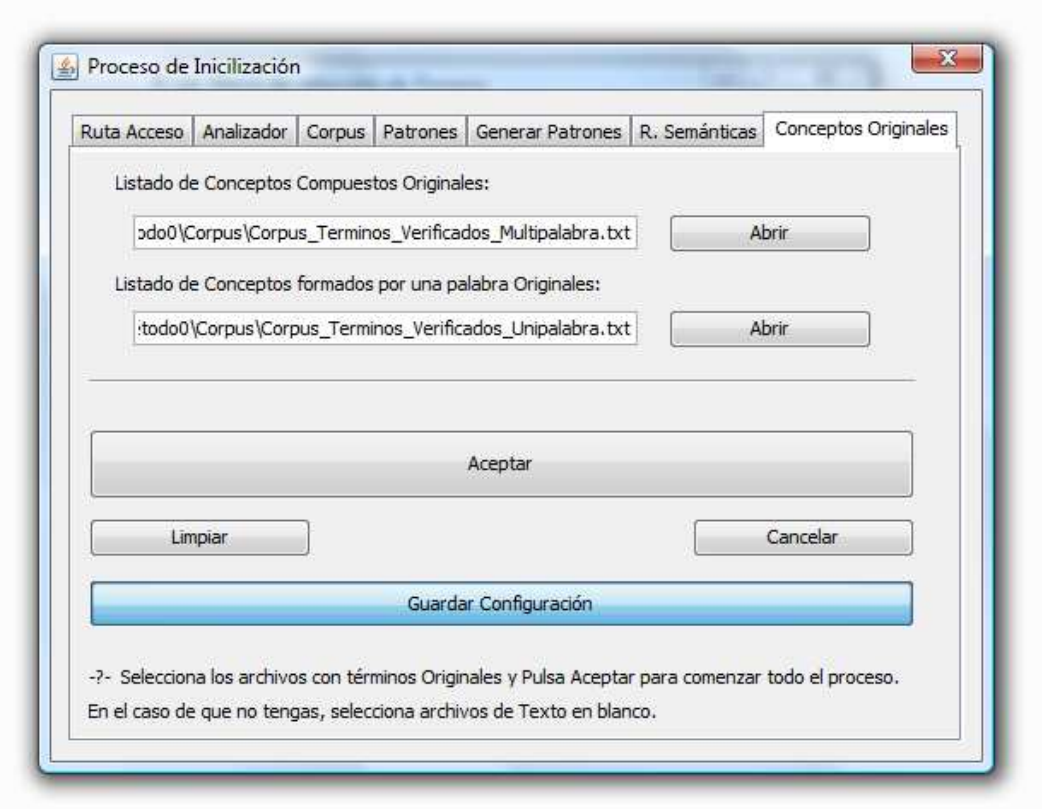


Ilustración 5. 18. Pestaña Conceptos Originales, introducción de la terminología valida para comprobar la veracidad del sistema.

❖ Selección del Proceso a ejecutar

Una vez definido y configurado el proyecto, se muestra una nueva ventana en la que se puede elegir el método que deseamos utilizar para generar la ontología final. Aquí, se puede definir si la ontología va a estar formada por conceptos formados por una sola palabra, por conceptos compuestos o por ambos conceptos. A continuación, se definen los elementos de esta ventana (ver ilustración 5.19).

Ilustración 5. 19. Selección del método para Generar la Ontología.

❖ Proceso de Configuración – El método Conceptos Simples

Si el usuario selecciona el método Conceptos Simples, se le mostrará una ventana que le permita introducir los valores correspondientes a este método (ver ilustración 5.20).

En la pestaña TF-IDF, se definen principalmente los patrones que identificaran los conceptos formados por una palabra en el corpus, junto con otros elementos. A continuación, se definen los elementos de esta ventana (Ver ilustración 5.20).

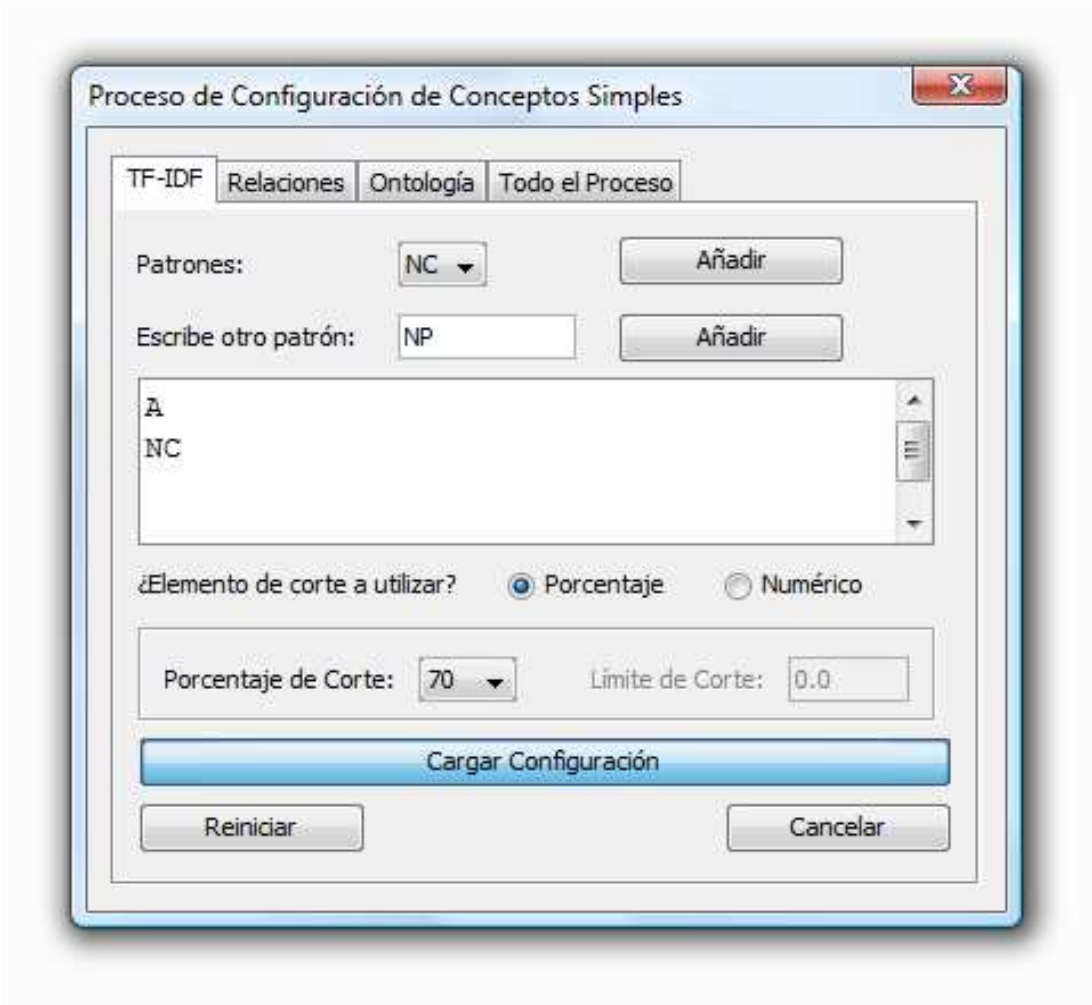


Ilustración 5. 20. **Configuración del Método TF-IDF para el proceso identificación de conceptos simples.**

En el supuesto de que el usuario cuente con alguna configuración guardada anteriormente, presionando el botón *Cargar configuración*, podrá elegir la configuración deseada en el cuadro de diálogo que le sea presentado, cuando termine, todas sus configuraciones quedarán reflejadas en las pestañas de configuración de este proceso.

Seguidamente, en la pestaña Relaciones, se ajustarán los parámetros necesarios para el módulo de extracción de relaciones semánticas. Los elementos principales son: las palabras a la izquierda y a la derecha, ya que en estas, es donde se identificarán a los conceptos válidos que generarán las relaciones, las cuales formarán parte de la ontología. A continuación, se definen los elementos de esta pestaña (ver ilustración 5.21).

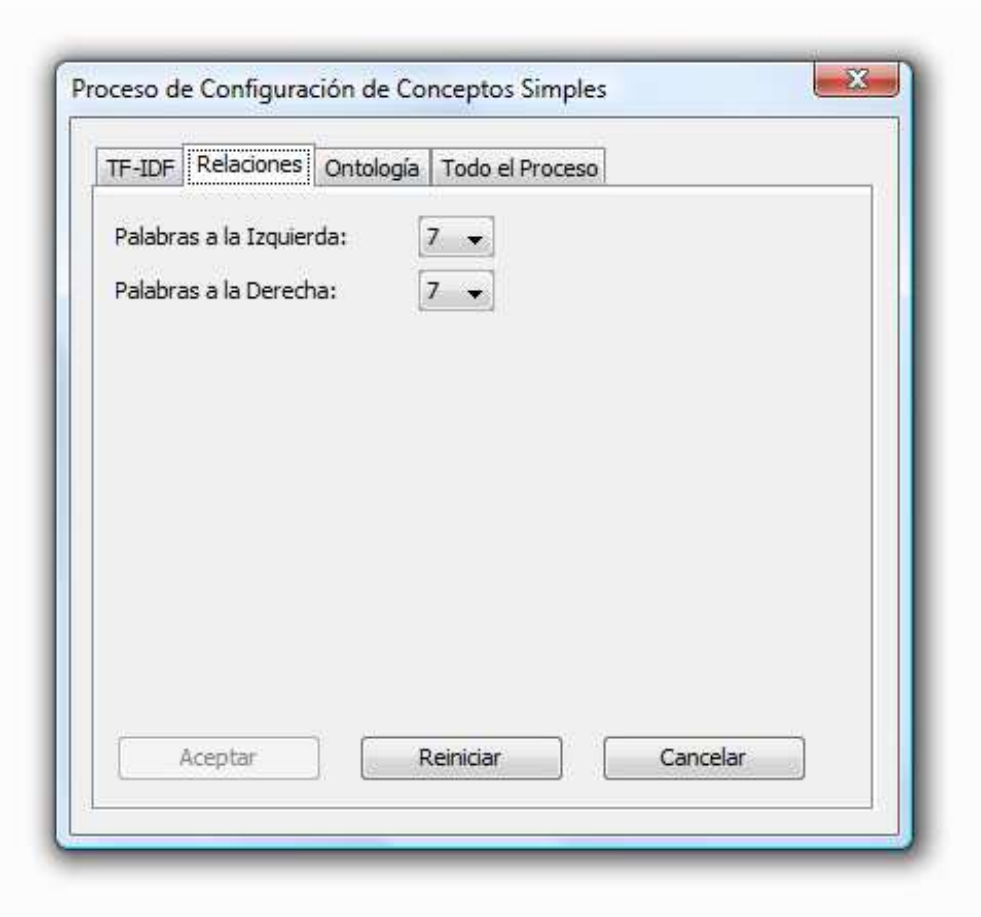


Ilustración 5. 21. Configuración de la identificación de conceptos para el proceso de identificación de relaciones en el método conceptos simples.

La siguiente pestaña "*Ontología*", permite ajustar y aportar los parámetros necesarios como el nombre y la URI de la ontología y la ruta donde se guardará dicha ontología. A continuación, se definen los elementos de esta ventana (ver ilustración 5.22).

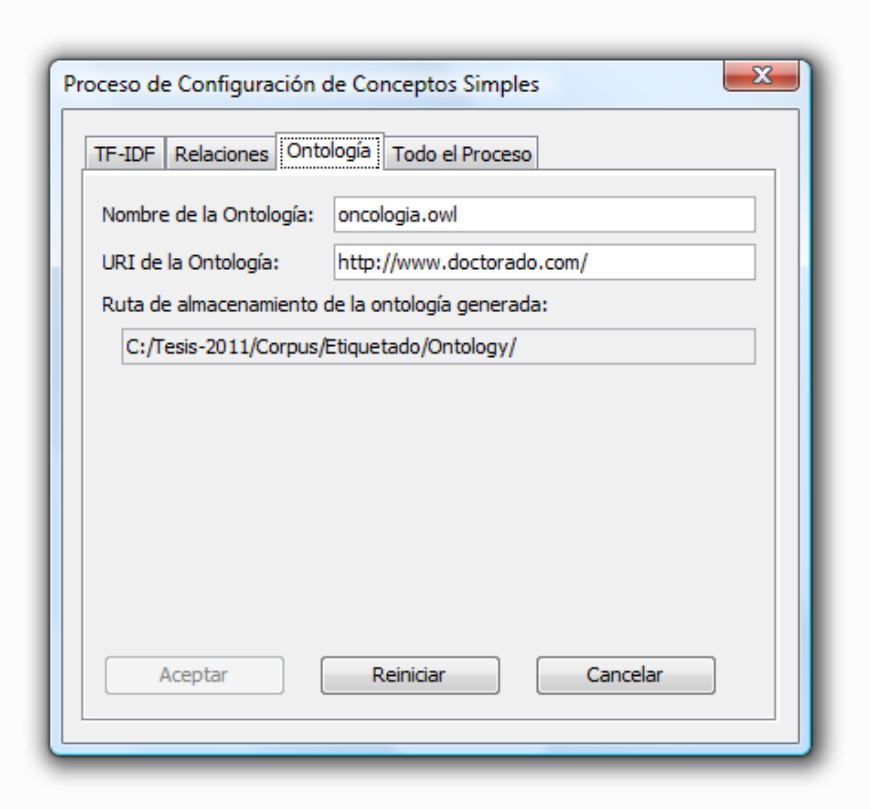


Ilustración 5. 22. Configuración de los elementos de la Ontología para el método de conceptos Simples.

En la última pestaña llamada *Todo el Proceso*, se seleccionarán las fases del proceso que se quieren ejecutar: la Búsqueda de Conceptos, Extracción de Relaciones y / o Creación de la Ontología. A continuación, se definen los elementos que hay que configurar en esta pestaña (ver ilustración 5.23).

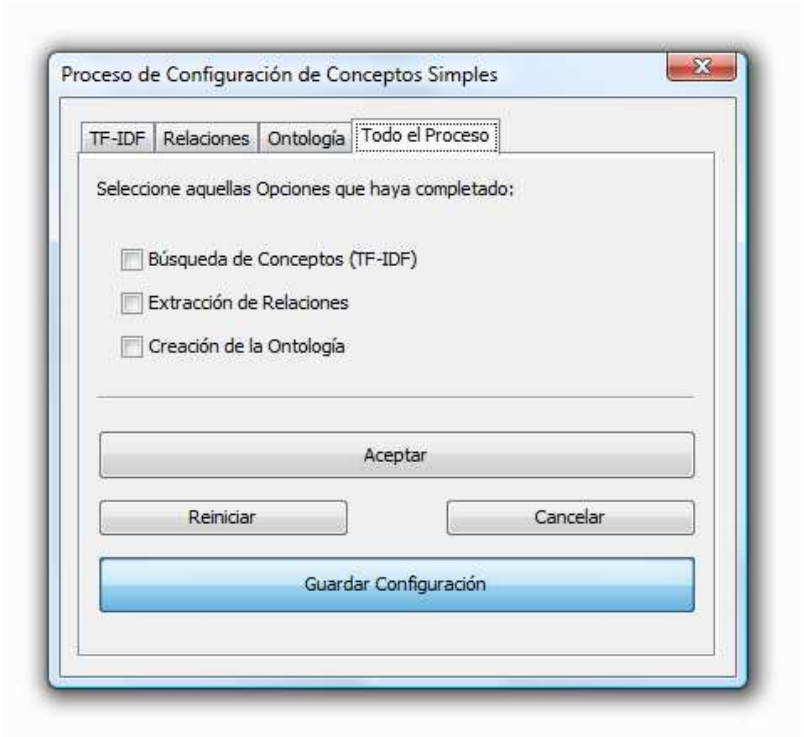


Ilustración 5. 23. Selección de los procesos a ejecutar por el sistema para el método de conceptos simples.

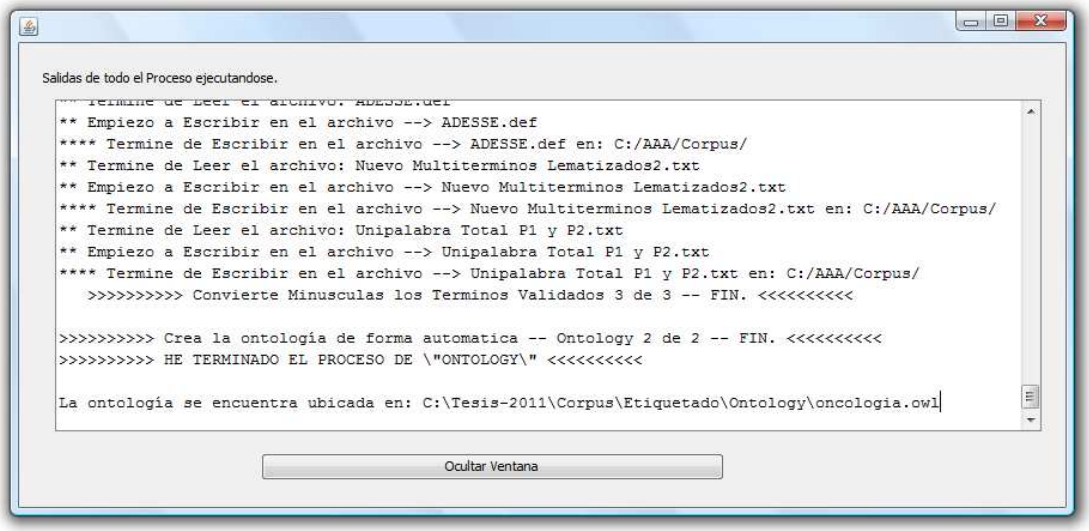


Ilustración 5. 24. Resultado final del proceso de creación de la Ontología para el método de conceptos simples.

Por ejemplo, en la ilustración 5.25, se muestra una ontología resultante con este método.

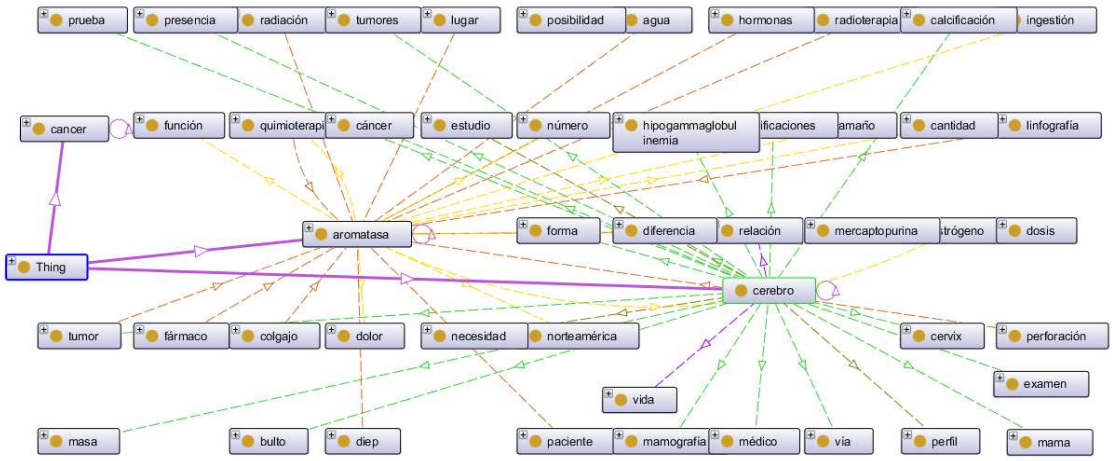


Ilustración 5. 25. Ejemplo de una Ontología creada solo con conceptos formados por una palabra.

❖ Proceso de Configuración – El método Conceptos Compuestos

Si el usuario selecciona el Método Conceptos Compuestos (ver ilustración 5.19), se le mostrará una ventana, la cual, le permitirá introducir los valores correspondientes a este método (ver ilustración 5.26).

En la pestaña C-value, se define principalmente el porcentaje de corte del método C-value. A continuación, se definen los elementos de esta ventana (ver ilustración 5.26).

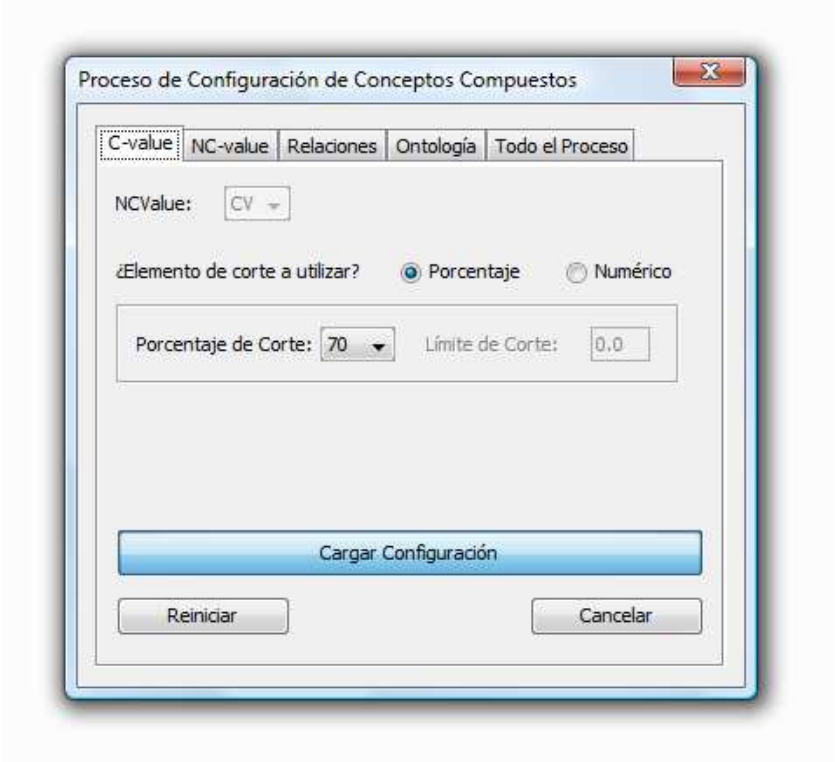


Ilustración 5. 26. Configuración del Método C-value para el proceso de Búsqueda de Conceptos Compuestos.

Si el usuario cuenta con alguna configuración guardada anteriormente, presionando el botón *Cargar configuración*, podrá elegir la configuración guardada en el cuadro de diálogo que le sea presentado, cuando termine, todas sus configuraciones quedarán reflejadas en las pestañas de configuración de este proceso.

Seguidamente, en la Pestaña NC-value, se ajustaran los parámetros necesarios para que el método funcione a las necesidades que se tengan, para ello, se configuran los parámetros principales que son: las palabras a la izquierda, a la derecha y el porcentaje de corte, ya que de ellos depende la identificación de los mejores conceptos compuestos. A continuación, se definen los elementos de esta ventana (ver ilustración 5.27).

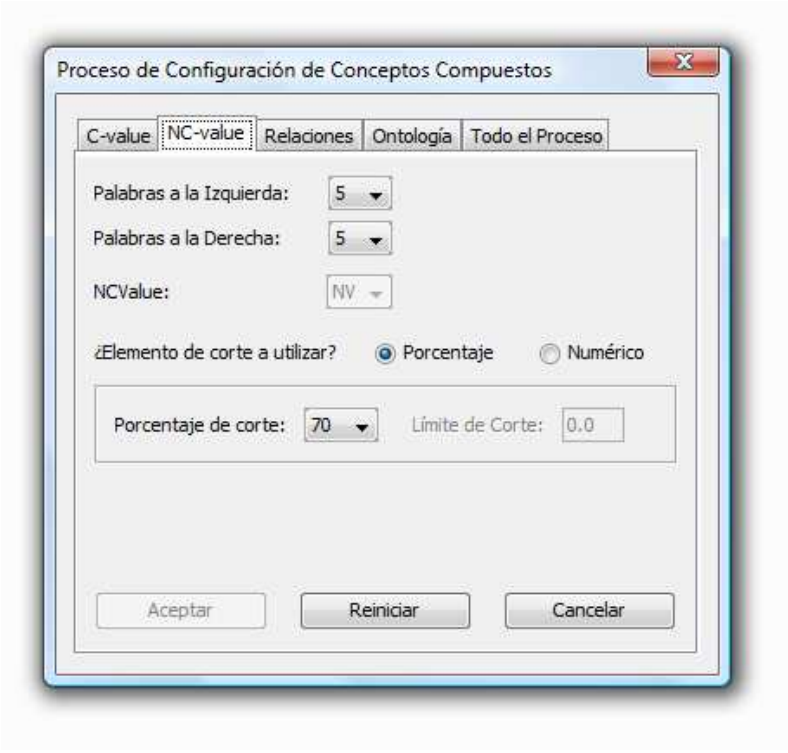


Ilustración 5. 27. Configuración del Método NC-value para la identificación de los mejores conceptos compuestos.

En la pestaña Relaciones Semánticas, se ajustarán los parámetros necesarios para el módulo de extracción de relaciones semánticas. Al igual que en el método de conceptos Simples, los elementos principales nuevamente son: las palabras a la izquierda y a la derecha, ya que en estas es donde se identificarán a los conceptos válidos que generarán las relaciones, que formarán parte de la ontología. A continuación, es mostrada la ilustración 5.28.

Ilustración 5. 28. Configuración de la identificación de conceptos para el proceso de identificación de relaciones en el método Conceptos Compuestos.

La siguiente pestaña, es la Pestaña *Ontología,* la cual, permite ajustar y aportar los parámetros necesarios como el nombre y la URI de la ontología y la ruta en donde se va a guardar dicha ontología. Los elementos son idénticos a los empleados en el método de Conceptos Simples. A continuación, es mostrada la ilustración 5.29, que representa esta pestaña.

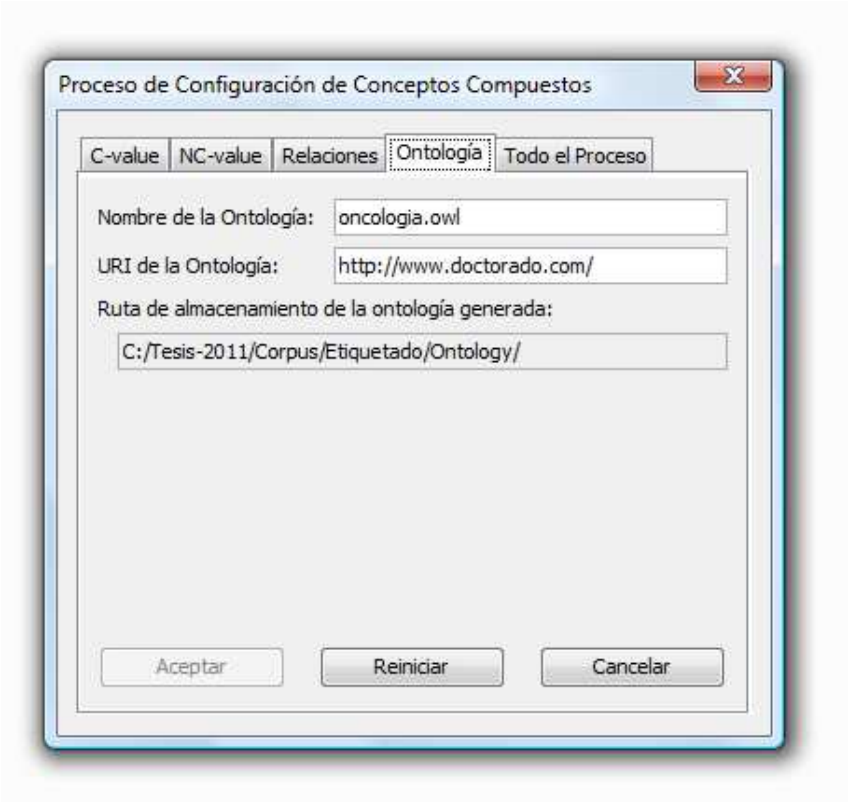


Ilustración 5. 29. Configuración de los elementos de la Ontología para el método Conceptos Compuestos.

En la última pestaña llamada Todo el Proceso, son presentadas todas las fases de este método, es decir, las fases que se desean ejecutar, estas son: la Búsqueda de Conceptos para C-value y NC-value, Extracción de Relaciones y Creación de la Ontología. A continuación, se definen los elementos que hay que configurar en esta pestaña (ver ilustración 5.30).

Si se desea guardar esta configuración, es necesario presionar en el botón *Guardar Configuración,* para almacenarla con un nombre específico y esta pueda ser recuperada en un futuro.

Al igual que en el método de configuración de Conceptos Simples, los procesos pueden ejecutarse de forma independiente, es decir, se puede ejecutar uno a uno en periodos de tiempo no consecutivos, lo único que se necesita, es cargar el proyecto guardado cada vez que se desee continuar con un mismo proyecto y seleccionar el método que se desea ejecutar.

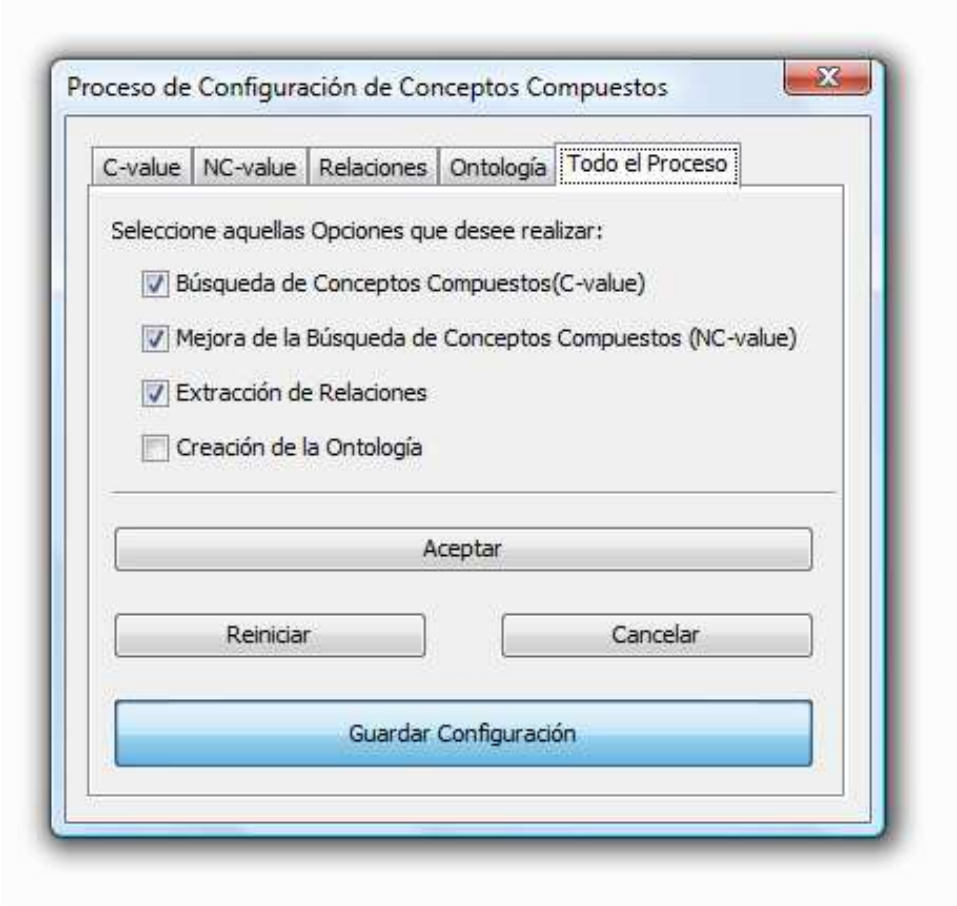


Ilustración 5. 30. Selección de los procesos a ejecutar por el sistema para el método conceptos compuestos.

Como parte final de este proceso, es necesario pulsar al botón Aceptar, para que se ejecuten todos los procesos marcados. En una ventana, se irán mostrando los logs del sistema, y cuando se haya finalizado todo el proceso, se obtendrá un mensaje de salida como el presentado en la ilustración 5.31, donde se podrá observar, la ruta del fichero en lenguaje OWL creado.

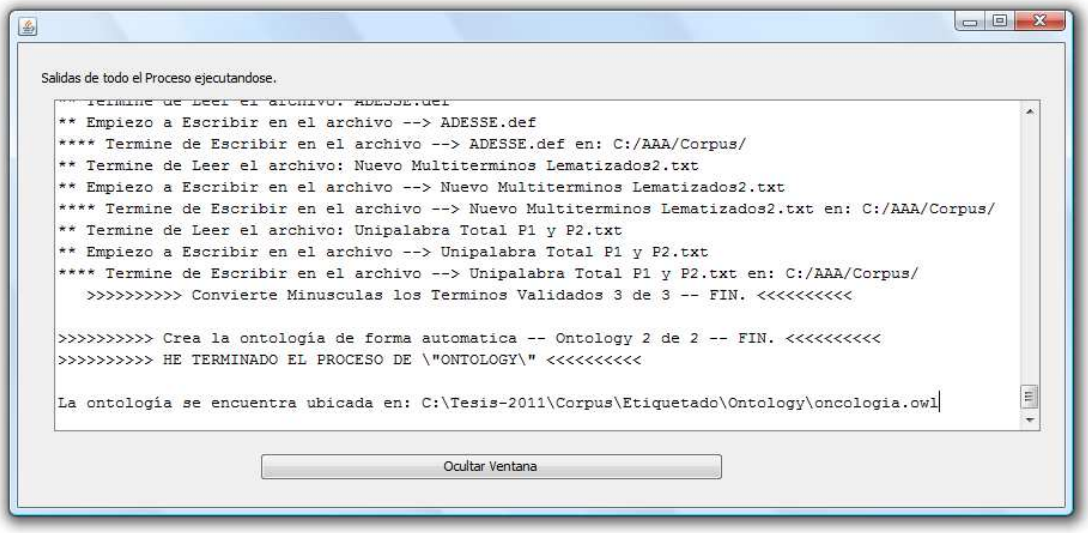


Ilustración 5. 31. Resultado final del proceso de creación de la ontología para el método conceptos compuestos.

Por ejemplo, en la ilustración 5.32, se muestra un ejemplo de la ontología resultante con este método.

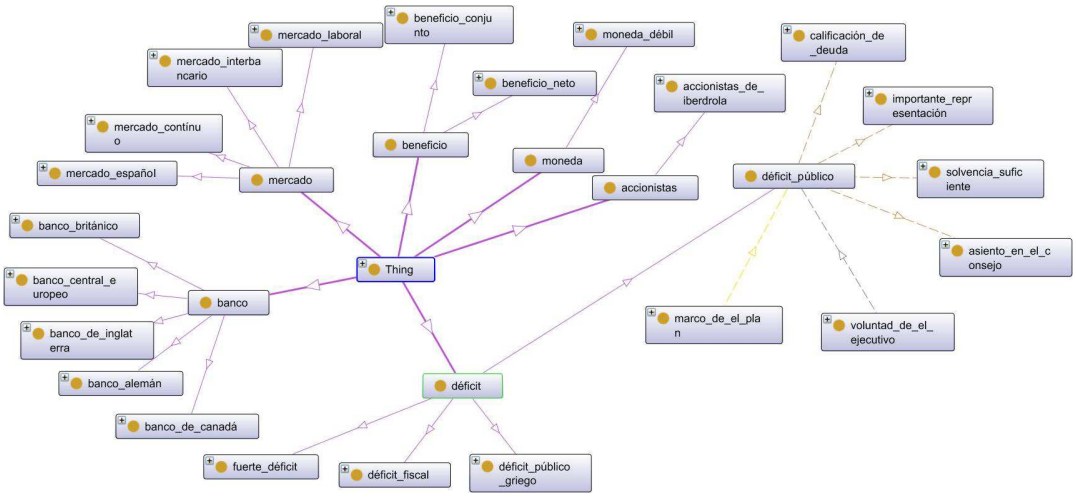


Ilustración 5. 32. Ejemplo de una Ontología creada solo con conceptos compuestos

❖ Proceso de Configuración – El método Conceptos Combinados

Si el usuario selecciona el Método Conceptos compuestos (ver ilustración 5.19), se le mostrará una ventana, en la cual, podrá introducir los valores correspondientes a este método.

En la pestaña C-value + TF-IDF, los elementos principales que el usuario tiene que aportar son: los patrones formados por una palabra (en el área de texto debe existir al menos un patrón) y el Porcentaje de corte (ya que este define el límite de corte de los mejores candidatos). La definición de los elementos de esta pestaña, se puede apreciar en la ilustración 5.20 y en la ilustración 5.26, por lo que a continuación, se muestra la figura representativa de esta pestaña (ver ilustración 5.33).

Ilustración 5. 33. Asignación de valores para la obtención de los Mejores conceptos formados por una palabra y los conceptos compuestos para el método Conceptos Combinados.

Seguidamente, en la pestaña NC-value, los parámetros principales que deben ser configurados son: las palabras a la izquierda, a la derecha y el porcentaje de corte, ya que de ellos depende la identificación de los mejores conceptos compuestos. La definición de los elementos de esta pestaña, se puede apreciar en la descripción de la pestaña NC-value del método conceptos compuestos, (ver ilustración 5.27), por lo que a continuación, se muestra la figura representativa de esta pestaña (ver Ilustración 5.34).

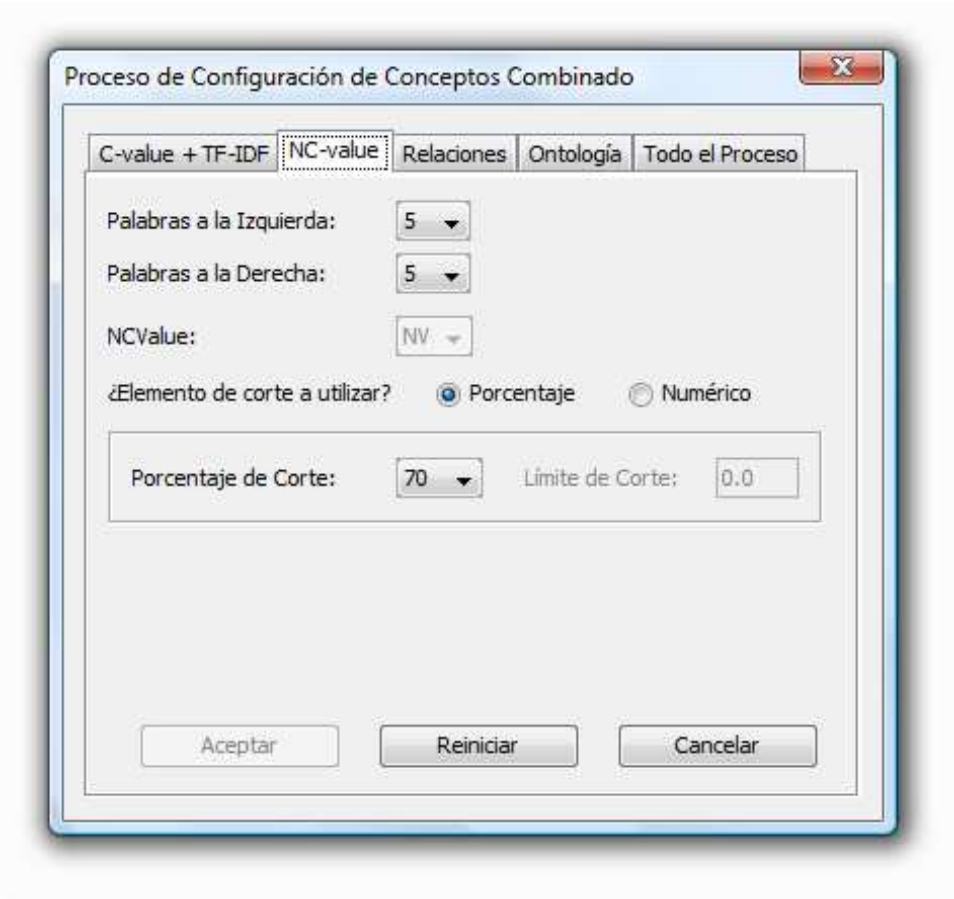


Ilustración 5. 34. Configuración del Método NC-value para la identificación de los mejores conceptos compuestos.

En la pestaña *Relaciones Semánticas*, se ajustarán los parámetros necesarios para el módulo de extracción de relaciones semánticas. Al igual que en el método de conceptos Compuestos, los elementos principales son: las palabras a la izquierda y a la derecha de un concepto, ya que en estas, es donde se identificarán a los conceptos válidos que generarán las relaciones que formarán parte de la ontología.

La definición de los elementos de esta pestaña se puede apreciar en la descripción de la pestaña NC-value del método conceptos compuestos, (ver ilustración 5.28), por lo que a continuación, se muestra la figura representativa de esta pestaña (ver Ilustración 5.35).

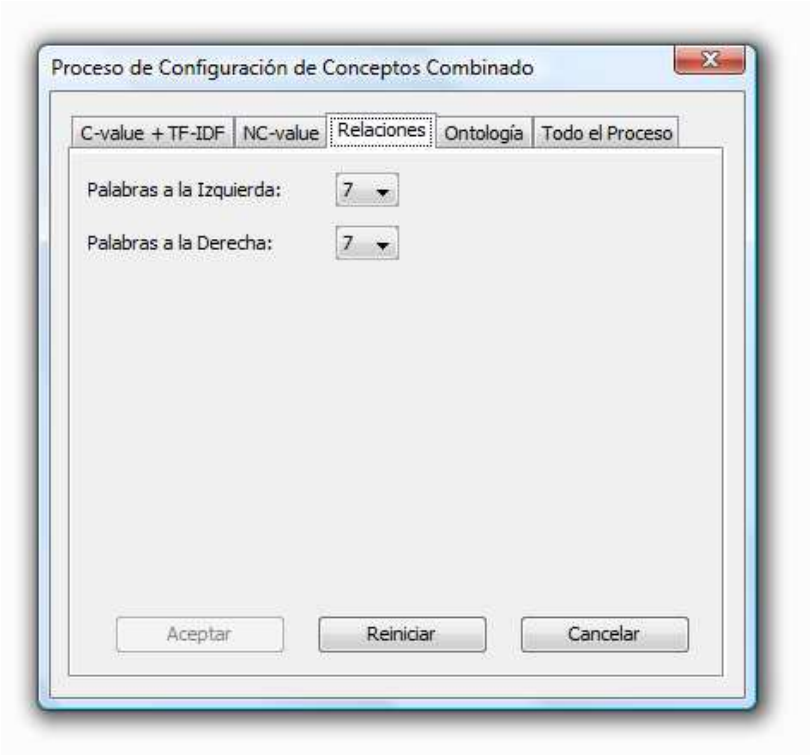


Ilustración 5. 35. Configuración de la identificación de conceptos para el proceso de identificación de relaciones en el método Conceptos Combinados.

La penúltima pestaña, es la Pestaña *Ontología,* la cual permite ajustar y aportar los parámetros necesarios como el nombre, la URI y la ruta de donde se va a guardar dicha ontología. Los elementos, al ser idénticos a la pestaña Ontología del método Conceptos Compuestos, fueron descritos anteriormente. A continuación, es mostrada la ilustración 5.36, que representa esta pestaña.

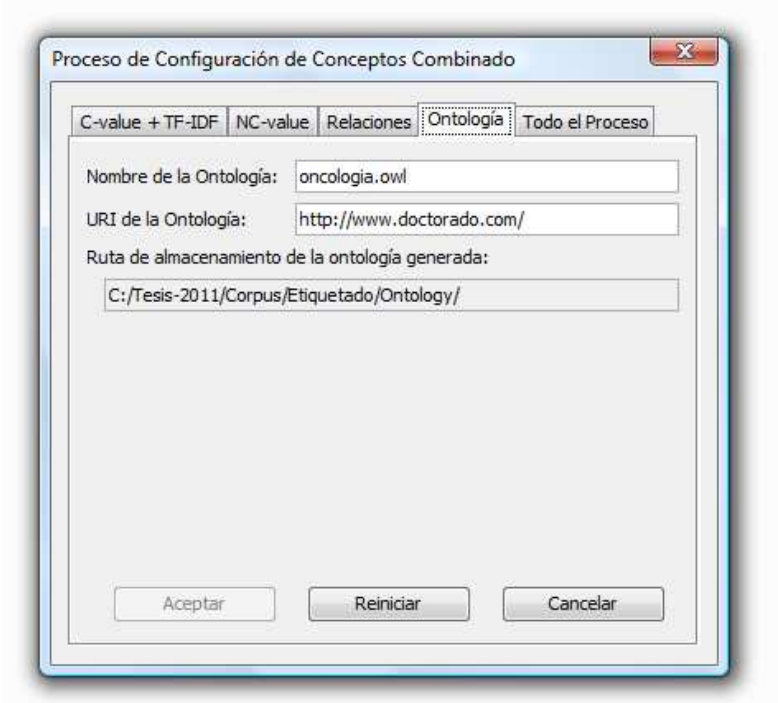


Ilustración 5. 36. Configuración de los elementos de la Ontología para el método Conceptos Combinados.

En la última pestaña llamada *Todo el Proceso*, son presentadas todas las fases de este método, para ser ejecutadas de forma continuada o por separado. A continuación, se definen los elementos que hay que configurar en esta pestaña (ver ilustración 5.37).

Si se desea guardar esta configuración, es necesario presionar en el botón *Guardar Configuración,* para almacenarla con un nombre específico y esta pueda ser recuperada en un futuro.

Al igual que en el método de configuración de Conceptos Compuestos, los procesos pueden ejecutarse de forma independiente, es decir, se puede ejecutar uno a uno en periodos de tiempo no consecutivos, lo único que se necesita, es cargar el proyecto guardado cada vez que se desee continuar con un mismo proyecto y seleccionar el método que se desea ejecutar.

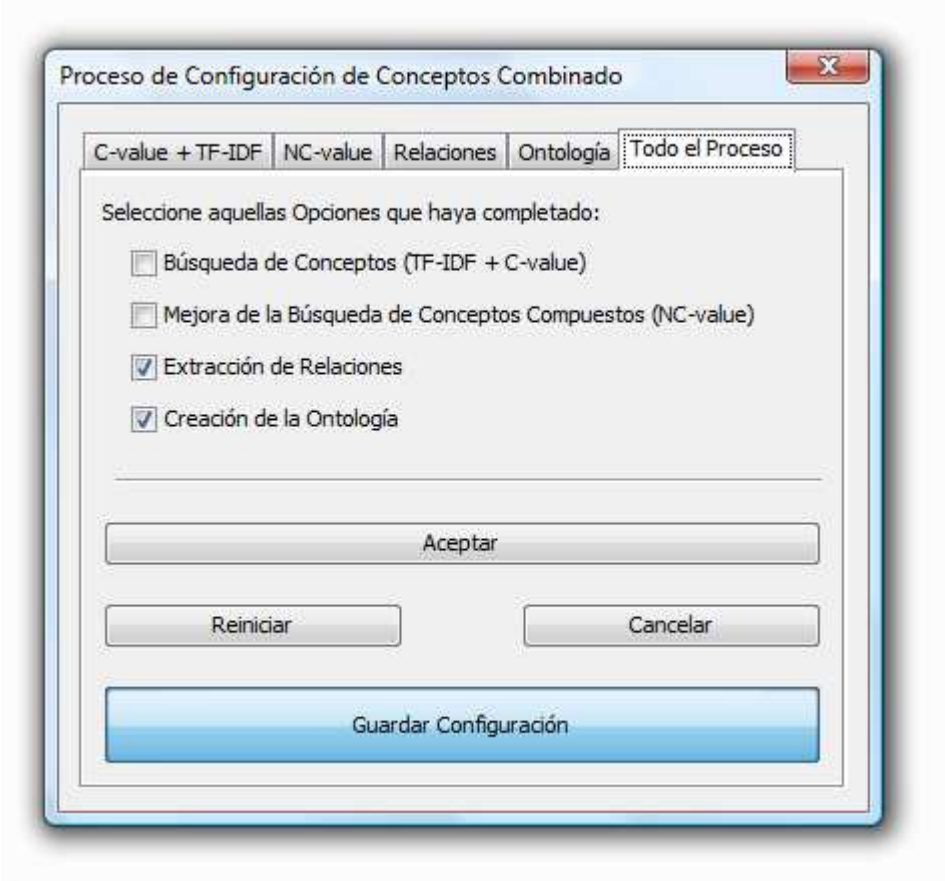


Ilustración 5. 37. Selección de los procesos a ejecutar por el sistema para el método Combinado.

Como parte final de este proceso, es necesario pulsar al botón Aceptar, para que se ejecuten todos los procesos marcados. En una ventana, se irán mostrando los logs del sistema, y cuando se haya finalizado todo el proceso, se obtendrá un mensaje de salida como el presentado en la ilustración 5.38, donde se podrá observar la ruta del fichero en lenguaje OWL creado.

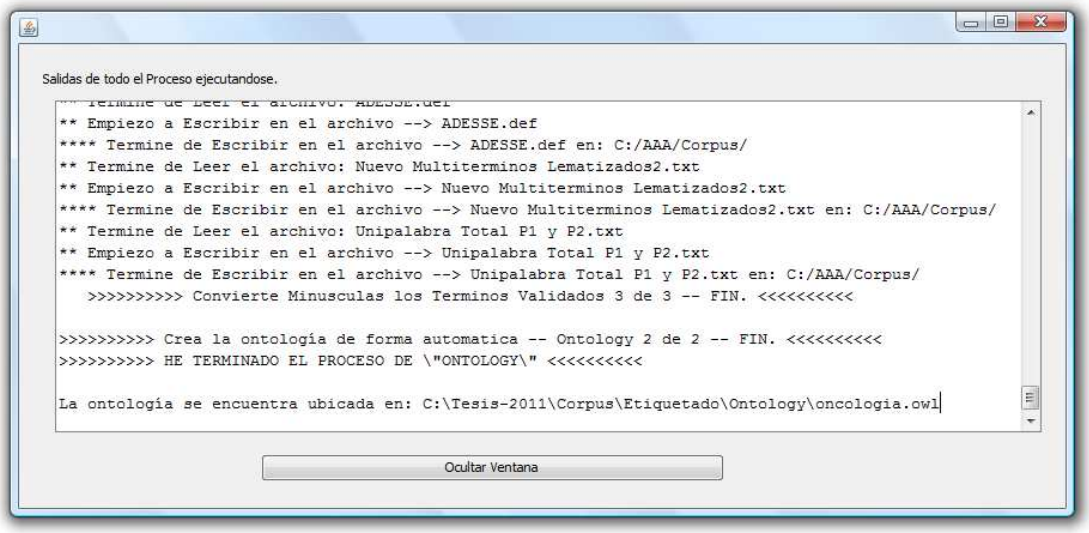


Ilustración 5. 38. Resultado final del Proceso de creación de la Ontología para el método Conceptos Combinados.

Por ejemplo, en la ilustración 5.39, se muestra un ejemplo de la ontología resultante con este método.

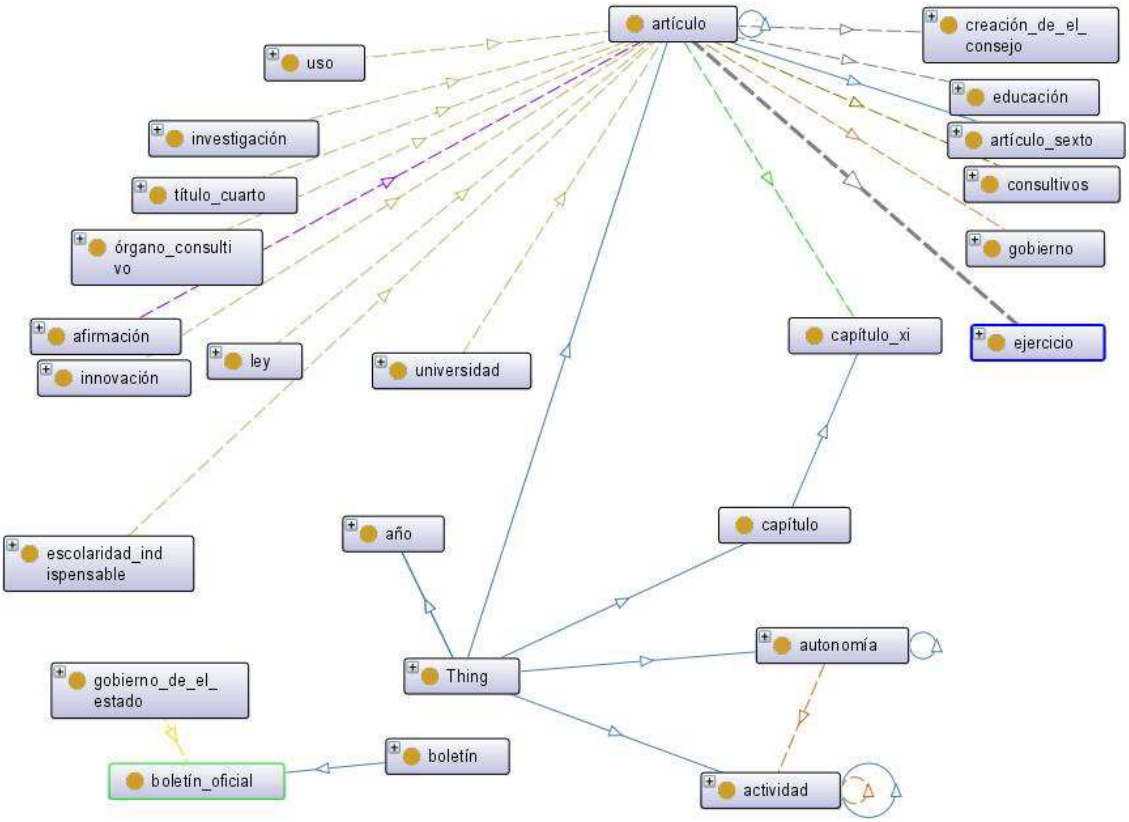


Ilustración 5. 39. Ejemplo de una Ontología creada con conceptos formados por una palabra y compuestos.

6 Validación de la metodología

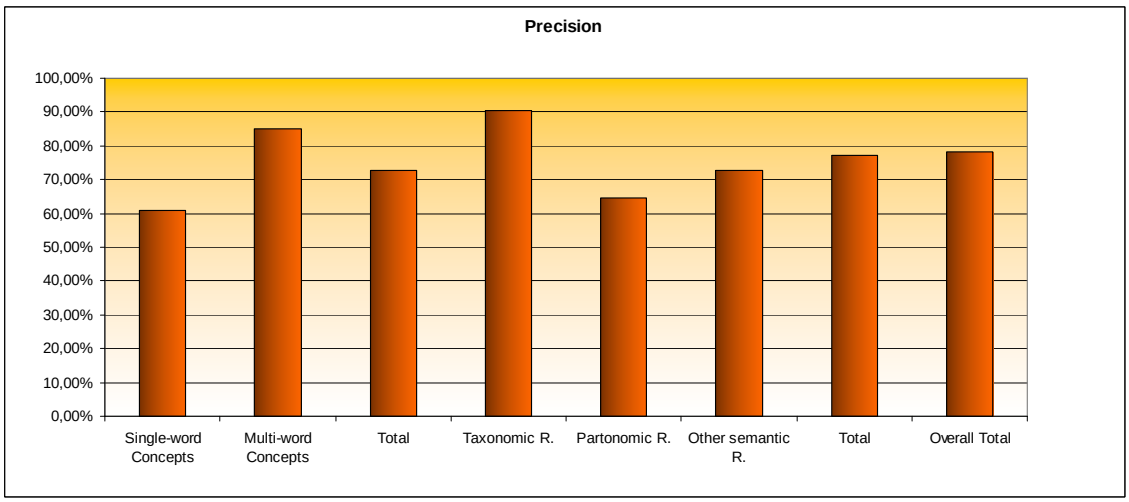
Para validar los resultados obtenidos con esta metodología, se presenta una de las métricas de evaluación más reconocidas y estándar empleadas en el proceso de recuperación de información [Subramaniam et al., 2010]. Su nombre es la medida – F o F1 y se define como "*La media armónica de los valores de Precisión y Recall"* (ver ecuación 8), Esta media F1, se basa en otras dos métricas para obtener sus resultados, estas son la Precisión que representa una medida de exactitud o fidelidad y el Recall que representa una medida de integridad, sus ecuaciones son 9 y 10 respectivamente. Los resultados aquí presentados fueron obtenidos al evaluar la herramienta en el dominio universitario.

$$Medida - F = 2 * precision * recall \Big/ precision + recall \quad \ldots(8)$$

$$\text{Pr}ecisión = conocimiento\,correcto\ \ encontrado\ \ por\,\text{el}\,sistema \Big/ conocimiento\ total\ sugerido\ por\ el\ sistema \quad \ldots.(9)$$

$$\text{Re}call = conocimiento\ \ correcto\ encontrado\ \ por\,\text{el}\,sistema \Big/ conocimiento\ total\ existente\,en\ el\,corpus \quad \ldots..(10)$$

A continuación, se muestra un gráfica donde se puede apreciar el conocimiento extraído por cada una de las fases principales de la herramienta, donde resalta el 81.73% de las relaciones subclassOf y el 73.57% de las relaciones taxonómicas, y el 78.25% de las relaciones partonómicas. También se destaca el 86.35% de los conceptos multiplabra extraidos, estos valores son vistos en la figura de F-Measure. En cuanto a la figura de Precisión, tanto los conceptos multipalabra, las relaciones taxonómicas, presentan muy buenos resultados. Y en la figura de Recall, lo mas destacado es que casi todos los elementos están arriba el 80%, lo que indoca que se extrajo arriba del 80% del conocimeinto total encontrado en los documentos analizados.



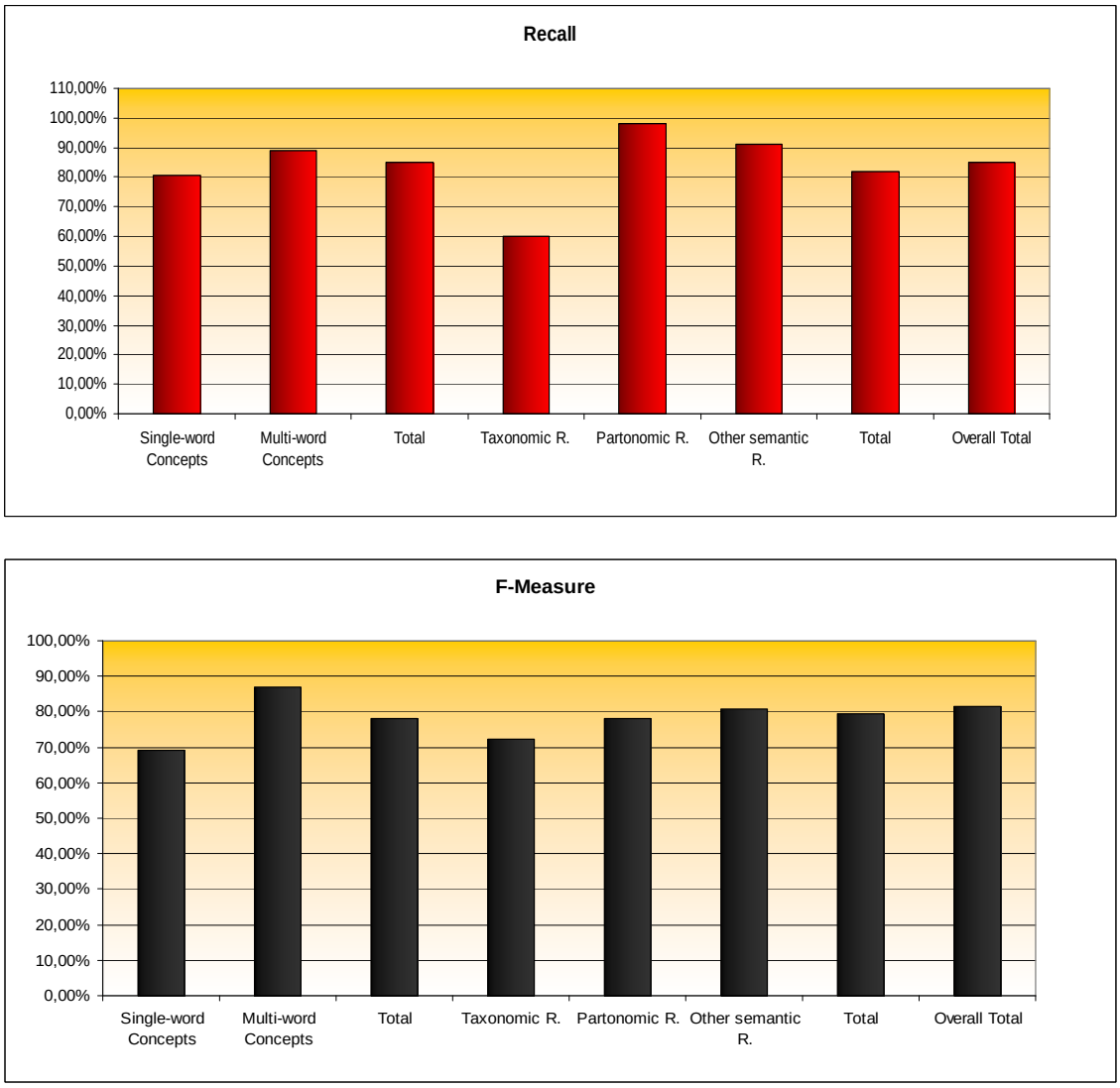


Ilustración 6. 1. Resultados de la evaluación del análisis.

7 Conclusiones

Se ha presentado en este artículo una herramienta que emplea la metodología de ontology learning para adquirir conocimiento a partir de textos escritos en lenguaje natural y en español, para ello se han utilizando algunas de las técnicas más representativas del procesamiento de lenguaje natural adaptadas al español. Esta herramienta se centra en recuperar conocimiento implícito empleando para ello la identificación de 4 tipos de relaciones entre conceptos: las relaciones taxonómicas, no taxonómicas, partonómicas y del tipo subClassOf, las tres primeras se basan en identificar los roles semánticos definidos en la base de datos semántica ADDESE y la última en la jerarquía de clases y propiedades de la ontología.

Una importante novedad es que en esta metodología se incorpora la aplicación de una metodología de aprendizaje de patrones de forma automática, para adquirir conocimiento la cual, da la diversidad de aplicarse a cualquier dominio y a cualquier lenguaje, ya que si reemplazamos las herramientas específicas para el lenguaje español, con las de cualquier otro lenguaje, podemos obtener los mismos resultados que los obtenidos en el español y dado que la metodología de aprendizaje de patrones es configurable, se pueden obtener infinitas combinaciones de patrones para el lenguaje en el que sea implementado.

La construcción de ontologías a partir de textos no representa una novedad, ya que a la fecha se han publicado diversas metodologías, sin embargo, para el idioma español, no existe aún una gran cantidad de estas, por esa razón, presentar esta metodología para este lenguaje representa un gran avance. En lo que respecta al exterior, en la comunidad de la ingeniería del conocimiento, la construcción de ontologías de forma automática es considerada como una importante actividad, ya que desde hace años las ontologías están llegando a ser la base de múltiples actividades, entre ellas tenemos principalmente a la Web semántica [Shamsfard and Barforoush, 2004], y entre otras aplicaciones tenemos la gestión del conocimiento en bibliotecas, los tutores online en diversas áreas del conocimiento, en traducciones automáticas, en la detección de plagios editoriales o en una de las aplicaciones más recientes de las ontologías que son los fraudes bancarios.

8 Referencias

[Locke y Booth] Locke W.N. y Booth A.D., "Machine Translation of Languages", Technology Press of MIT ans Wiley, Cambridge, Mass., 1955.

[Covington, M.] Covington, Michael A. "Natural Language Processing for Prolog Programmers". Artificial Intelligence Programs The University of Georgia Athens, Georgia. PRENTICE HALL, Englewood Cliffs. NewJersey 07632.

[Manaris y Slator] Manaris, B. Z. y Slator, B. M. 1996. Interactive Natural Language Processing: Building on Success. Computer, IEEE.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5, pp. 199-220.

Phillip G. Armour. *The five orders of ignorance*. Communications of the ACM Volumen 43, número 10. October 2000.

[Ochoa, J.L. et al., 2011] Ochoa, J.L., Hernández-Alcaraz, M.L., Valencia-García, R. and Martínez-Béjar, R. (2011). A semantic role based Ontology Learning approach for Spanish texts. *In DCAI 2011*, Salamanca, Spain, 91/2011, pp. 273-280. Doi: 10.1007/978-3-642-19934-9_35.

[Atserias et al., 2006] Jordi Atserias and Bernardino Casas and Elisabet Comelles and Meritxell González and Lluís Padró and Muntsa Padró. (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genoa, Italy. May, 2006.

[Padró et al., 2010] Lluís Padró and Miquel Collado and Samuel Reese and Marina Lloberes and Irene Castellón. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), ELRA, La Valletta, Malta. May, 2010.

[Ochoa et al., 2011b] Ochoa J.L, Almela A. and Valencia-García R. (2011) Identifying patterns for unsupervised learning of multiword terms. Educational Research and Reviews vol 6(9) pp. 645-656.

[Frantzi et al., 2000] K.T. Frantzi, S. Ananiadou and M. Hideki. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries, 3(2), pp. 115-130.

[Grefenstette G., 1994] Grefenstette G (1994). Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers Norwell, MA, USA.

[Salton, 1991] Salton, G. (1991). Developments in Automatic Text Retrieval, Science, August 30, 253, pp. 974–80.

[Knoth et al., 2009] Knoth, P., Schmidt, M., Smrz, P. & Zdráhal, Z. (2009). Towards a framework for comparing automatic term recognition methods. Paper given at Znalosti (Knowledge) 2009, Brno, Prague, pp. 12.

[Ramos, 2003] J. Ramos. (2003). Using tf-idf to determine word relevance in document queries. In First International Conference on Machine Learning, New Brunswick: NJ, USA, 2003. Rutgers University.

[Valencia-García et al., 2008] Valencia-García R, Fernández-Breis JT, Ruiz-Martínez JM, García-Sánchez F, Martínez-Béjar R (2008). A knowledge acquisition methodology to ontology construction for information retrieval from medical documents. Expert Systems: The Knowledge Engineering Journal. 25(3):314-334.

[Moreda et al., 2011] P. Moreda, H. Llorens, E. Saquete, and M. Palomar, (2011). Combining semantic information in question answering, Information Processing and Management.

[García-Miguel et al., 2010] García-Miguel, José M.; Fita González Domínguez y Gael Vaamonde. (2010). ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanish, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), Valletta (Malta), 17-23 de mayo.

[Palmer et al., 2005] Palmer M, Gildea D, Kingsbury P (2005). The proposition bank: an annotated corpus of semantic roles. Computational Lingustics. 1(31): 71-106.

[Albertuz-Carneiro, 2007] Albertuz Carneiro, Francisco, (2007). Sintaxis, semántica y clases de verbos: Clasificación verbal en el proyecto ADESSE, Cano López, Pablo (coord): Actas del VI Congreso de Lingüística General, Santiago de Compostela, Las lenguas y su estructura (IIb), 2(2), pp. 2015-2030. ISBN 84-7635-672-2.

[García-Miguel et al., 2005] García-Miguel, José M.; Lourdes Costas y Susana Martínez (2005). Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. En Wotjak, Gerd, & Juan Cuartero Otal, eds. Entre semántica léxica, teoría del léxico y sintaxis. Frankfurt am Main: Peter Lang, pp. 373-384.

[Subramaniam et al., 2010] Subramaniam T, Jalab HA, Taga AY (2010) Overview of textual anti-spam filtering techniques. Int. J. Phys. Sci. 5(12): 1869-1882.

[Shamsfard & Barforoush, 2004] Shamsfard M. and Barforoush A. (2004). Learning ontologies from natural language texts, International Journal of Human-Computer Studies, 60(1), pp. 17-63.