UNIVERSIDAD DE MURCIA

Facultad de Comunicación y Documentación
Departamento de Información y Documentación

TESIS DOCTORAL

Metodología para la población automática de ontologías. Aplicación en los dominios de medicina y turismo

Juana María Ruiz Martínez

Dirigida por:

Dr. D. Rafael Valencia García
Dr. D. Rodrigo Martínez Béjar

Octubre 2011

RESUMEN

La Web Semántica tiene como meta la ampliación de los estándares y tecnologías utilizados en la web actual, de manera que, el contenido semántico de los documentos sea inteligible tanto para el ser humano como para las aplicaciones software. El desarrollo de esta potente infraestructura requiere, así mismo, de la creación de un conjunto de herramientas que le puedan dar soporte. Entre estas herramientas se encuentran las ontologías, utilizadas en el ámbito de la Web Semántica como estructuras capaces de organizar, inferir o compartir el conocimiento. La creación de ontologías y su mantenimiento de la forma más automatizada posible es uno de los principales retos a los que se ha enfrentado desde su nacimiento la Web Semántica.

La construcción automática de ontologías, o como se denomina en inglés *Ontology Learning*, consiste en la creación (semi-) automática de conceptos, relaciones, instancias y atributos ontológicos utilizando diversos métodos computacionales y tomando como fuente de conocimiento, en la mayoría de los casos, textos en lenguaje natural. Una subtarea dentro de la construcción automática de ontologías, es la población o instanciación (semi-)automática, también conocida como *Ontology Population*. Esta tarea, que es el elemento vertebrador del trabajo que aquí se presenta, consiste en enriquecer mediante instancias de las clases y/o de las relaciones una ontología ya existente. La importancia de la población de ontologías reside fundamentalmente en el hecho de que (1) las ontologías existentes necesitan de actualizaciones periódicas (2) el tejido ontológico es lo suficientemente amplio como para dar cobertura a numerosos dominios especializados, la instanciación de dichas ontologías con datos obtenidos de la web supone un salto cualitativo en las tareas de recuperación de información.

Por otra parte, el contenido web es predominantemente textual, siendo la web la fuente a partir de la que se obtiene conocimiento relevante para la instanciación de ontologías mediante el uso de técnicas de procesamiento de lenguaje natural.

En este trabajo se proponen dos metodologías para la instanciación automática de ontologías abordando el desarrollo de la mismas desde una perspectiva lingüística y computacional.

La primera metodología se basa en la distancia cotextual y la ganancia de conocimiento. Se parte de dos corpora, relativos al dominio del turismo, que se analizan desde un punto de vista discursivo. Los datos obtenidos se interpretan, de modo que la información que arroja el análisis se pueda utilizar como elemento de partida para la extracción de información textual de forma automática.

En primer lugar se desarrollan una serie de patrones lingüísticos y listas de entidades nombradas, que se integran en la herramienta GATE. Una vez procesados los textos con esta herramienta, el resultado es un conjunto de anotaciones clasificadas como entidades nombradas. A continuación, la metodología propuesta calcula cuáles de esas entidades podrían ser individuos de la ontología, basándose en la distancia que las separa en el texto, por un lado, y en la cantidad de conocimiento que aportan a la ontología, por otro. La consistencia de la ontología se comprueba en la fase final mediante un razonador.

La segunda metodología propuesta se basa en la instanciación de ontologías a partir de roles semánticos. La integración de diversos recursos ontológicos y lingüísticos es la base de esta metodología, en la que se combinan ontologías de alto nivel del dominio biomédico con *frames* semánticos extraídos de FrameNet. El resultado es un modelo ontológico, que permite la extracción de relaciones entre entidades en textos de carácter biomédico. Las entidades implicadas en dichas relaciones se convierten en candidatas a instancias de la ontología. Finalmente, un razonador comprueba la consistencia e infiere nuevas instancias, en función de los axiomas definidos. La validación de la metodología se lleva cabo mediante el mapeo del modelo ontológico con una ontología de dominio biomédico y la instanciación de la misma.

AGRADECIMIENTOS

A lo largo de mi trayectoria investigadora son muchas las personas que me han apoyado, ayudado y animado tanto a nivel profesional como personal, y a todas ellas me gustaría darles mi más sincero agradecimiento.

Muy especialmente, quiero dar las gracias a mis directores, Rafael Valencia García, por su atenta labor de dirección, apoyo y amabilidad; y Rodrigo Martínez Béjar, por confiar en mí y brindarme la oportunidad de comenzar esta aventura multidisciplinar.

A todos los compañeros y profesores, sin cuya inestimable colaboración no hubiera sido posible esta tesis, y en especial a Dago por sus valiosos consejos, a José Antonio y Laura por su ayuda en distintas etapas de este trabajo, a Paco y Yesi por sus correcciones, y a Piedad por su apoyo y sugerencias en cuestiones lingüísticas.

El día a día no hubiera sido lo mismo sin mis compañeros-amigos de grupo Cati, Rafa, Nieves, Mari Carmen, Consuelo, Lizeth, Astrid, Edu, Lucía, Isidro, Manolo, Tedy, Cristina, Lilian, Eneko, Ángel, María… con los que tan buenos momentos he compartido.

A aquellos profesores que me han brindado la oportunidad de unirme temporalmente a sus grupos de investigación en distintos países, Diego Mollà, Paola Velardi, Roberto Navigli y Roxana Girju.

Una mención especial merecen mis amigos Mari Sol, Narci, Cristina, Toñi, Roberto y Lucia por todas las charlas sobre ontologías y procesamiento de lenguaje natural que han soportado y por los ánimos que me han dado siempre. Y Ale, por su apoyo y, sobre todo, por la paciencia.

A mi familia, y en especial a mis padres, en los que siempre he encontrado apoyo incondicional.

SUMARIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

BLOQUE I

INTRODUCCIÓN, ESTADO DEL ARTE Y OBJETIVOS

CAPÍTULO 1

INTRODUCCIÓN, OBJETIVOS Y METODOLOGÍA

Resumen. La web semántica y el procesamiento del lenguaje natural son áreas de investigación en las que convergen, entre otras ramas del conocimiento, la ingeniería informática, la lingüística y la documentación. Esta tesis, fruto de la colaboración multidisciplinar, trata de afrontar desde diversas perspectivas, el problema de la instanciación automática de ontologías a partir de texto en lenguaje natural. En este capítulo se introducen las nociones básicas que se desarrollarán profusamente a lo largo de esta tesis doctoral, y que están relacionadas fundamentalmente con la denominada Web Semántica y con el Procesamiento de Lenguaje Natural. Finalmente, en este capítulo se perfilan los objetivos que se persiguen, así como la metodología que se va a seguir a lo largo de este trabajo de investigación.

1.1 Introducción

La World *Wide Web* se ha convertido en el mayor repositorio digital del que disponemos en la actualidad. Imágenes, video, audio y texto conforman este entramado multimedia que crece exponencialmente. Desde sus inicios, la gestión y recuperación de los de datos contenidos en la Web se ha convertido en uno de los principales retos al que se han enfrentado, desde puntos de vista diversos, usuarios, investigadores y entidades gestoras de contenidos on-line.

Como indica (Stumme et al., 2006), el origen del problema es que los datos de la Web no están estructurados, de manera que sólo pueden ser entendidos por el ser humano, pero la cantidad de datos es tan grande que sólo puede ser procesada eficientemente por máquinas.

La organización de la información contenida en la web se ha abordado así mismo, desde distintas perspectivas. Por ejemplo, se han aplicado métodos de organización ya conocidos, como los índices o los tesauros, se han reinventado otros ya existentes, como las ontologías, llegadas de la mano de la Web Semántica

o incluso se han creado nuevos estructuras de organización como las folksonomías.

La realidad es que no existe una metodología única y definitiva que permita gestionar el contenido web de forma completamente eficiente. No obstante, de entre las propuestas que se han realizado, es la Web Semántica la que se perfila como aquella que permite explotar de forma más eficaz los contenidos web, siendo al mismo tiempo la más ambiciosa.

Por otro lado, el lenguaje natural en su forma escrita es el más utilizado en los motores de búsqueda, el correo electrónico o la producción científica que circula por la red. Esta gran cantidad y diversidad de información textual, sobrepasa las capacidades de los buscadores actuales que todavía no pueden dar respuesta a todas de las consultas o preguntas directas que un usuario pueda realizar.

El campo de investigación de la Web Semántica surgió con el objetivo de dotar tanto a los contenidos textuales como no textuales, accesibles desde la Web, de una estructura y un significado bien definidos de manera que no sólo sean inteligibles para el ser humano, sino también para las aplicaciones software (Berners-Lee et al., 2001)

Para obtener una adecuada definición de los datos, la Web Semántica utiliza RDF y OWL, dos estándares que ayudan a convertir la Web en una infraestructura global en la que es posible compartir y reutilizar conocimiento.

La tecnología empleada para la representación del conocimiento en la Web Semántica son las ontologías, que se pueden definir como una especificación formal de la conceptualización de un domino (Studer et al., 1998). Las ontologías, no sólo proporcionan una estructura formal para la representación del conocimiento de un dominio, sino que pueden reutilizarse y compartirse (Studer et al., 1998; Valencia- García et al. 2004).

A pesar de las ventajas que ofrecen, la construcción y desarrollo de ontologías de manera manual, sigue siendo, aún hoy, una tarea lenta y costosa. La necesidad de superar esta limitación (Shamsfard & Bargoroush, 2004) ha dado lugar a la proliferación de estudios e investigaciones cuya finalidad es obtener métodos para

contribuir al proceso de construcción automática o semiautomática de ontologías (*Ontology Learning*), proceso del cual la instanciación automática de ontologías (*Ontology Population*) forma parte.

La mayoría de métodos de aprendizaje de ontologías (*Ontology Learning*) utilizan técnicas de *machine learning* (aprendizaje automático), que no sólo permiten descubrir conocimiento ontológico a gran escala y de forma más rápida que manualmente, sino que se minimizan los efectos negativos de factores como la subjetividad introducida por el análisis humano y las posibles inconsistencias (Zhou, 2007).

Mientras que la finalidad del *Ontology Learning* es la adquisición de nuevos conceptos y relaciones con el consecuente cambio de la ontología en sí misma, la meta del *Ontology Population* o población de ontologías[1] es la extracción y clasificación de instancias de los conceptos y relaciones definidas en la ontología (Tanev & Magnini, 2006).

Es en este punto, en la creación de ontologías, y en concreto en la instanciación automática de ontologías para el caso que nos ocupa, en donde la ingeniería ontológica y las técnicas de extracción de información y procesamiento de lenguaje natural convergen.

La web, como depósito de contenidos textuales, y los nuevos sistemas de computación, han generado una nueva inquietud entre la comunidad científica, sobre todo en los campos de la ingeniería informática, pero también en la lingüística, la sociología o la psicología, y es el intento de sistematizar, mediante el desarrollo de recursos y herramientas, el lenguaje humano. Algunas de las principales disciplinas lingüísticas, como la morfología, la sintaxis, la semántica y la pragmática han sido los puntos desde los que se ha abordado la informatización del lenguaje. Surge así una nueva disciplina, la Lingüística Computacional, e importantes recursos como WordNet o FrameNet, además, aparecen nuevos

1 En esta memoria, se utilizan de forma intercambiable los términos población de ontologías e instanciación de ontologías.

conceptos, como el de las Entidades Nombradas y toda una serie de corpora anotados como muestra representativa del lenguaje que se quiere sistematizar. Herramientas como GATE (Cunningham , 2002) contribuyen a la integración de todos estos recursos para generar aplicaciones de minería de textos.

La instanciación o población de ontologías de forma (semi)automática a partir de texto estructurado o no estructurado[2] hace uso de técnicas de extracción de información. De este modo, el conocimiento contenido en los textos pasa a formar parte de la ontología en forma de instancias.

Aunque en un principio se pensó en la Web Semántica como una infraestructura global que diera cobertura a la totalidad de la Web, y de hecho existen ontologías que pretenden abarcar un amplio espectro de ámbitos del conocimiento como SUMO (IEEE, 2011), esta idea se ha ido diluyendo, quedando el uso de ontologías restringido a ámbitos del conocimiento específicos y delimitables. La representación del conocimiento mediante ontologías se ha consolidado sobre todo en dominios científicos especializados, como es el caso de la biomedicina, en donde son cada vez más las ontologías disponibles.

La instanciación de ontologías tiene dos funcionalidades principalmente, por un lado, la adición de nuevo conocimiento a las ontologías existentes y por otro el mantenimiento de ontologías ya instanciadas. Las ontologías de dominios no estáticos, como por ejemplo, la biomedicina o el turismo, necesitan de una actualización casi continua.

El nuevo conocimiento se incorpora a la ontología de forma "inteligente" gracias a los axiomas de las propiedades y de las clases y a la expresividad de OWL 2 que permite no sólo comprobar inconsistencias, si no también inferir nuevo conocimiento.

---

[2] La distinción entre texto estructurado y no estructurado se utiliza en este trabajo en su acepción computacional. Con el término texto estructurado nos referimos a un texto que posee una estructura creada artificialmente, en la que se incluye una cierta categorización de los elementos, ya sea mediante el uso de metadatos o mediante la división artificial del texto. En el ámbito lingüístico, una producción textual bien formada posee una estructura discursiva.

1.2 Objetivos

Con las dos metodologías propuestas en esta tesis, se aborda la instanciación automática de ontologías desde un punto de vista que combina el análisis lingüístico tradicional y tecnologías para la extracción de conocimiento textual. Así mismo, se recurre a la combinación de recursos lingüísticos y ontológicos ya desarrollados que permiten llevar a cabo el proceso de instanciación.

Los objetivos de la tesis son:

- Analizar desde un punto de vista lingüístico las características de un lenguaje de especialidad. La lingüística como ciencia del lenguaje se ha desarrollado durante siglos. Ciertamente, los métodos estadísticos y computacionales son necesarios para el desarrollo de herramientas de procesamiento de lenguaje natural, pero estos métodos serán más eficaces si consideran las características lingüísticas del tipo de textos objeto del procesamiento.
- Diseñar una metodología para la instanciación automática de ontologías basada en la distancia cotextual y en la ganancia de conocimiento. La distancia cotextual se refiere a la distancia física que existe entre dos unidades lingüísticas en el texto. En cuanto a la ganancia de conocimiento es una medida que hace referencia al conocimiento cuantitativo adquirido por el sistema, es decir, a mayor conocimiento adquirido mayor ganancia de conocimiento.
- Diseñar una metodología para la instanciación de ontologías basada en roles semánticos. Los roles semánticos son los papeles que desempeñan los actuantes de una oración con respecto al verbo. El marco de trabajo de esta metodología implica, así mismo, el diseño de un modelo ontológico en el que se combinan ontologías de alto nivel con recursos y herramientas lingüísticas y ontológicas.
- Validar la metodología basada en la distancia cotextual y la ganancia de conocimiento en el dominio del turismo. La validación consiste en la

extracción de instancias a partir de textos no anotados relativos a los dominios de la hostelería y restauración con el objetivo de instanciar una ontología.

- Validar la metodología basada en roles semánticos. La validación consiste en el mapeo con el modelo ontológico diseñado de una ontología de dominio biomédico. A partir de un corpus formado por textos del dominio de la biomedicina, se extraen las relaciones entre instancias y las instancias en sí.

Como se ha dicho previamente, las investigaciones que se presentan en esta tesis son fruto de la colaboración interdisciplinar entre lingüistas, documentalistas e informáticos. El punto de vista desde el que se abordan las metodologías propuestas es fundamentalmente un punto de vista lingüístico y documental, en consecuencia están fuera de los límites de esta tesis la descripción detallada de la parte más técnica referida a la implementación de los prototipos.

1.3 Metodología

La metodología propuesta para llevar a cabo los objetivos descritos previamente es la siguiente:

- Se ha llevado a cabo el análisis del estado del arte referido a las distintas áreas implicadas en las dos metodologías de instanciación que se proponen.
    - Web Semántica. Se ha realizado un análisis relativo a metodologías y lenguajes Web que permiten anotación semántica de contenidos, haciendo especial hincapié en OWL y OWL2. Así mismo se ha realizado un estudio de las metodologías y tecnologías de representación del conocimiento, entre las que se encuentran las ontologías pero también otros sistemas de representación como las folksonomías, mapas

conceptuales o tesauros. Conscientes de las limitaciones de la Web Semántica actual, y de lo inviable que resulta su aplicación a nivel global a corto o incluso medio plazo, se han individualizado aquellas áreas de conocimiento en donde las ontologías se han utilizado con mayor profusión y en donde, en consecuencia, la instanciación de ontologías cobra un mayor sentido.

- Procesamiento de Lenguaje Natural (PLN). Se han analizado diferentes metodologías y tecnologías de procesamiento de lenguaje natural. Las herramientas y recursos de PLN están en auge, ante el reto de la organización de la web semántica, la comunidad científica, ha apostado por el desarrollo de recursos que sistematizan en la medida de lo posible el lenguaje humano. Se estudian los distintos niveles de un sistema para el PLN y los proyectos de mayor relevancia en el ámbito de las tecnologías lingüísticas como por ejemplo WordNet. Se estudian así mismo a las metodologías propuestas para la anotación de roles semánticos, ya sea en dominios generales, como en el dominio de la biomedicina.
- Extracción de información. Se ha procedido al análisis de las metodologías y tecnologías para la extracción de conocimiento a partir de textos en lenguaje natural, prestando especial atención a al reconocimiento e identificación de entidades nombradas y las relaciones entre ellas.

- Diseño e implementación de una metodología para la instanciación automática de ontologías basada en la distancia cotextual.
  - Se ha llevado a cabo un análisis de las características lingüísticas de un conjunto representativo de textos del dominio siguiendo la metodología propuesta por Bhatia (1993). Un conocimiento

exhaustivo de los textos de los que se extraen las instancias de la ontología, así como de su contexto, facilita la elaboración de recursos lingüísticos que se ajustan a las necesidades del dominio, optimizando, de este modo, la precisión y exhaustividad en la fase de extracción de entidades nombradas.

- Validación de la metodología mediante la instanciación de una ontología del dominio turístico. Se desarrollado una ontología de dominio reutilizando otras ya existentes. Las instancias para la población de la ontología se extraen de dos corpora seleccionados relativos a las actividades de hostelería y restauración.

- Diseño e implementación de una metodología para la instanciación de ontologías basada en roles semánticos.
  - Dada la disponibilidad de distintos recursos y herramientas para el PLN cuyo uso está extendido entre la comunidad científica, aunque generalmente no de manera combinada, se ha procedido al diseño de un marco de trabajo en el que se integran dicho recursos que son de carácter ontológico, lingüístico así como herramientas de PLN.
  - Diseño de un modelo ontológico en que se asocian *frames* semánticos a las relaciones de una ontología de alto nivel. Las relaciones ontológicas y los *frames* tienen en común que son generalizaciones de situaciones del mundo real cuya expresión lingüística se realiza fundamentalmente mediante una forma verbal o una nominalización de la misma.
  - Validación de la metodología mediante la instanciación de una ontología de dominio biomédico. El modelo ontológico que se ha desarrollado prevé que se mapeen con una ontología de dominio el máximo de relaciones posible.

1.4 Estructura

La estructura del documento consiste en tres bloques principales. El bloque I consta de 4 capítulos y en él se incluye la introducción y el estado del arte, mientras que en el bloque II que consta de 3 capítulos, se describen las metodologías de población de ontologías, se validan los resultados y se ponen de manifiesto algunas conclusiones y el trabajo futuro. Finalmente, el bloque III consta de un solo Capítulo en el que se realiza un resumen en inglés de la memoria de tesis. En detalle, la estructura del documento es la siguiente.

Capítulo 1 Introducción. En este capítulo, en el que nos encontramos, se enmarca el trabajo realizado dentro del ámbito de la web semántica, se presentan los objetivos perseguidos durante la investigación, así como la metodología que se ha utilizado para llevar a cabo el cumplimiento de los mismos.

Capítulo 2 Web Semántica. En este capítulo se hace un recorrido por los principales elementos de la web semántica, se describen sus objetivos, y se introduce el concepto de ontología del que se aportan algunas de las definiciones que se han dado a lo largo de la historia. Además, se ponen de manifiesto las principales clasificaciones de ontologías que encontramos en la literatura. Así mismo, se describe OWL y OWL2 y finalmente se hace alusión a otras estructuras de clasificación como los tesauros o folksonomías.

Capítulo 3 Procesamiento de Lenguaje Natural. Se describen los distintos niveles de un sistema de Procesamiento de Lenguaje Natural que coinciden casi en su totalidad con los distintos niveles de análisis de la lingüística tradicional. En primer lugar se describe en qué consiste el preprocesamiento textual, necesario para que la información sea procesable computacionalmente, a continuación se describen el procesamiento morfológico, sintáctico, semántico y finalmente pragmático. En cada uno de estos niveles se describen varios recursos que se utilizan en el desarrollo de las metodologías. Esto es, la base de datos léxica

WordNet y etiquetado de roles semánticos. Se hace un recorrido por los distintos proyectos para el etiquetado de roles semánticos tanto de carácter general, como VerbNet, FrameNet y PropBank, como de carácter especifico aplicados al dominio de la biomedicina. Finalmente se describen dos herramientas que permiten llevar a cabo el procesamiento, GATE y Freeling.

Capítulo 4 Instanciación automática de ontologías. Este capítulo comienza con la descripción de los sistemas de Extracción de Información, ya que como se indica, la población de ontologías posee numerosos elementos en común con esta tarea. Se habla así mismo del reconocimiento y extracción de entidades nombradas, especialmente relevante para los sistemas de instanciación de ontologías. Se describen algunas de las metodologías desde las que se puede abordar tanto la extracción de entidades como la extracción de relaciones entre ellas. Finalmente se analizan los principales sistemas de instanciación de ontologías propuestos en la literatura y se realiza una comparación entre ellos.

Capítulo 5 Instanciación de Ontologías Basada en la distancia contextual y la ganancia de conocimiento. En este capítulo en primer lugar se realiza una aproximación lingüística a un corpus de dominio turístico. Se extraen las distintas características discursivas y se seleccionan aquellos datos relevantes para el desarrollo de la metodología de instanciación.

En segundo lugar se desarrolla la metodología para la instanciación de ontologías aplicada al dominio del turismo. La validación se lleva a cabo en dos corpora de dominio y finalmente se muestran los resultados obtenidos.

Capítulo 6 Instanciación de Ontologías basada en los roles semánticos. En este capítulo se describe en primer lugar el proceso de integración entre los distintos recursos y herramientas que permiten el desarrollo de la metodología propuesta. Una vez descrito el marco de trabajo, y obtenido un modelo ontológico, se lleva a cabo la validación de la metodología. Para ello muestra como una

ontología de dominio se puede mapear con el modelo ontológico propuesto. Finalmente se ponen de manifiesto los resultados obtenidos.

Capítulo 7 Conclusiones, Trabajo Futuro y Contribuciones. En este capítulo se exponen las conclusiones generales de la tesis y se comentan cuales podrían ser las futuras líneas de trabajo.

Capítulo 8 English Summary. En este capítulo se realiza un resumen en inglés de lo expuesto a lo largo de la memoria, junto con las conclusiones. Este capítulo se ajusta a lo exigido por la normativa para la obtención de la mención de doctor europeo.

CAPÍTULO 2

WEB SEMÁNTICA Y ONTOLOGÍAS

Resumen En este capítulo se describe la Web Semántica y uno de sus principales componentes, las ontologías. Se realiza un breve recorrido histórico por el concepto de ontología a través de algunas de las principales definiciones presentes en la literatura. Así mismo se describen cada uno de los componentes ontológicos y el lenguaje OWL, un estándar del W3C para representación de ontologías. Se presentan varias de las clasificaciones de ontología que se encuentran en la literatura y que responden a diversos criterios clasificatorios. Finalmente, se comparan las ontologías con otros sistemas de representación del conocimiento como los tesauros, folksonomías o mapas conceptuales

2.1. Introducción

La era digital, sobre todo Internet y los nuevos soportes electrónicos, han generado nuevas áreas de investigación, disciplinas y herramientas que tienen como objetivo la sistematización y organización del conocimiento al que ahora se pude acceder computacionalmente. El contenido documental, ya sea textual, visual o auditivo es objeto de incursiones automatizadas para, fundamentalmente, identificar las partes, separar lo relevante de lo irrelevante y finalmente generar un nuevo producto procesado que puede ir desde un documento anotado semánticamente a un producto totalmente nuevo, normalizado y estandarizado, como puede ser una ontología. El porqué de tan arduo trabajo es que un sistema informático a pesar de su capacidad de procesamiento y almacenamiento, no es capaz de "comprender" la información que contiene a no ser que posea "instrucciones" de cómo entenderla.

La Web Semántica nace como una solución prometedora para el acceso y organización a los contenidos digitales, siendo las ontologías la piedra angular sobre la que se sustenta.

Más concretamente, esta tesis se encuadra dentro del aprendizaje automático de ontologías, que engloba técnicas provenientes de ámbitos diversos, como la adquisición de conocimiento, el aprendizaje automático, el Procesamiento de Lenguaje Natural, la inteligencia artificial o la gestión de bases de datos (Gacitua et al., 2007).

En este capítulo se describen las principales características de la Web semántica junto con uno de sus componentes fundamentales, las ontologías.

2.2. Web Semántica

El gran éxito de la Web ha traído consigo nuevos retos, ya que cada día se genera y acumula una gran cantidad de información inteligible únicamente para personas, y no para el computador. La Web Semántica se fundamenta en una innovadora visión de Tim Berners Lee acerca de la Web, dónde se pretende dotar a los contenidos de información semántica inteligible para las máquinas, de manera que éstas "entiendan" el significado de los contenidos con los que operan. De este modo, la información se podría compartir, procesar y transferir de forma más eficiente. Según Beerners Lee, la Web Semántica es una extensión de la Web actual donde la información viene dotada de significado bien definido, y permite a computadoras y personas trabajar en cooperación (Berners-Lee et al., 2001).

Como señalaban Antoniou y Harmelem (2004) la meta de la Web Semántica es, entre otras, permitir el avance de los sistemas de gestión del conocimiento.

El proyecto de la Web Semántica tal y como lo concibieron Berners Lee y Lassila, ha avanzado no sólo en el ámbito científico, sino que existen en la actualidad aplicaciones reales que están llegando al mercado como indica Jorge Cardoso en su estudio sobre el estado de la Web Semántica (Cardoso, 2007), aunque su avance es más lento de lo que se preveía en un principio, ya que la Web Semántica requiere el uso y desarrollo de una apreciable cantidad de tecnologías. Entre los objetivos del desarrollo de la Web Semántica, se pueden mencionar los siguientes:

- **Datos comprensibles por las máquinas**. La Web Semántica es una visión, a saber, la idea de que los datos en la WWW sean definidos y enlazados de forma que la aplicaciones puedan utilizarlos con fines no únicamente de visualización.

- **Agentes inteligentes**. La Web Semántica pretende conseguir una WWW más legible para las aplicaciones con el fin de facilitar que los agentes inteligentes puedan recuperar y manipular la información de forma autónoma.

- **Bases de datos distribuidas**. La Web Semántica pretender conseguir con los datos lo mismo que consiguió el HTML con los sistemas de información textuales, esto es, proporcionar suficiente flexibilidad para representar las bases de datos y las reglas lógicas para enlazarlas. Un logro de la Web Semántica sería transformar la actual estructura de libro con hipervínculos de la WWW en una gran base de datos distribuida.

- **Infraestructura automática**. La Web Semántica es una infraestructura y no una aplicación. Por lo tanto, el gran problema para el desarrollo de la Web Semántica es la falta de un marco simple de automatización en la WWW actual.

- **Servicio a los humanos**. La visión de la Web Semántica es permitir que las aplicaciones localicen recursos relevantes para nosotros en la Web de manera automática y sin mucha intervención de los usuarios.

- **Anotación mejorada**. La idea de la Web Semántica proporciona, a la Web informal, unas anotaciones entendibles por las aplicaciones informáticas.

- **Búsqueda mejorada**. La meta principal de la Web Semántica es construir un índice estructurado de la Web, que permitirá acceder a los recursos Web por contenido en vez de por palabras clave.

- **Servicios Web**. La Web Semántica no sólo proporcionará acceso a documentos estáticos, sino también a servicios capaces de ofrecer comportamientos útiles, permitiendo a los agentes software automatizar procedimientos actualmente manuales.

Para obtener una adecuada definición de los datos, la arquitectura de la Web Semántica está formada por una serie de capas, según definió Berners Lee en el siguiente esquema (figura 2.1).

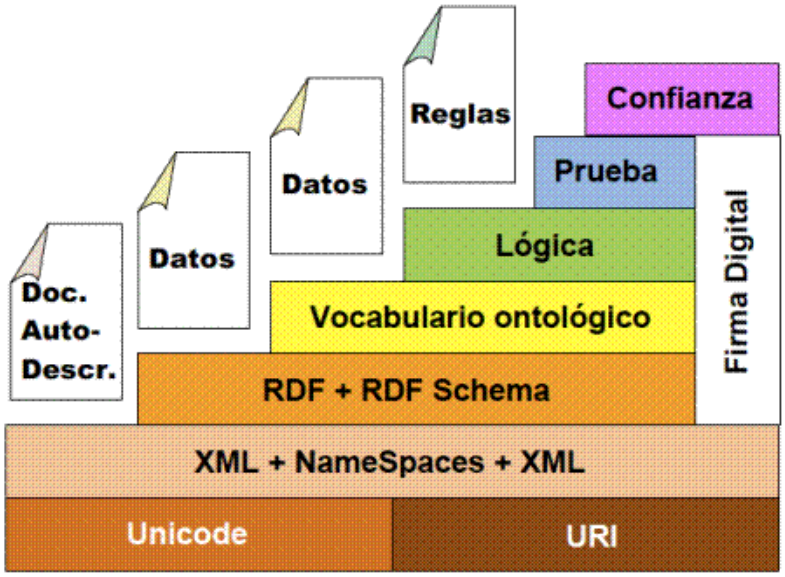


Figura 2.1 Capas de la Web Semántica.

Los dos elementos de la primera capa proporcionan una sintaxis común. *Uniform resource identifiers* (URIs) proporciona un estándar para referirse a las entidades, a los recursos, mientras que Unicode es un estándar universal para la codificación de caracteres y símbolos.

El metalenguaje XML (*Extensible Markup Language*) establece una notación para estructurar los datos de los documentos en forma de árboles de etiquetas con atributos, mientras que XML *Schema* permite la definición de gramáticas válidas para documentos XML, En la actualidad, la normalización de estas dos capas está ampliamente aceptada e implantada. XML es un primer paso para la representación explícita de los datos y la estructura de los contenidos de la web separada de su presentación en HTML. No obstante, este lenguaje ofrece una capacidad limitada para expresar la semántica, ya que el modelo de datos de XML consiste en un árbol que no distingue entre objetos y relaciones, ni tiene noción de jerarquía de clases (Castells, 2003).

Resource Description Framework (RDF) es el primer nivel en donde la información es, en cierto modo, inteligible para la máquina. Según el W3C, RDF es un lenguaje que permite describir metadatos y proporciona interoperabilidad entre aplicaciones que intercambian información inteligible para las máquinas en la Web (Lassila & Swick, 1998).

Los documentos RDF están formados por declaraciones de recursos en expresiones con la forma sujeto-predicado-objeto. Los recursos pueden ser páginas Web, partes o colecciones de páginas Web, o cualquier objeto que no sea directamente parte de la WWW (por ejemplo un libro, una persona o un concepto). En RDF los recursos se designan siempre mediante URIs. Las propiedades son atributos específicos, características o relaciones que describen recursos. Cada propiedad tiene un significado específico, en ellas se definen sus valores permitidos, los tipos de recursos que puede describir y sus relaciones con otras propiedades.

Un recurso específico junto con una propiedad denominada, más el valor de dicha propiedad para ese recurso es una sentencia RDF. Es decir, una sentencia está formada por una tripleta que consiste en dos nodos (sujeto y objeto) unidos por un arco (predicado), donde los nodos representan recursos, y los arcos propiedades.

En http://www.w3.org/TR/REC-rdf-syntax/ encontramos en el ejemplo que se muestra en la figura 2.2:

```
<?xml version="1.0"?>
<rdf:RDF   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"

xmlns:contact="http://www.w3.org/2000/10/swap/pim/conta
ct#">

  <contact:Person
rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>

</rdf:RDF>
```

Figura 2.2 Ejemplo de documento rdf

Que significa que existe una persona (contact:Person) identificada con la URI "http://www.w3.org/People/EM/contact#me", cuyo nombre (contact_fullName) es "Eric millar", que tiene la dirección de email (contact:mailbox) "em@w3.org", y cuyo título (conctact:personalTitle) es "Dr."

Con *RDF Schema* (RDFS) se pueden definir jerarquías de clases de recursos, especificando las propiedades y relaciones que se admiten entre ellas.

La siguiente capa es *Ontology Vocabulary*. En la actualidad, el lenguaje estándar para el desarrollo de ontologías es *Ontology Web Language* (OWL) que se considera una extensión de RDF, incluyendo toda su capacidad expresiva y permitiendo utilizar expresiones lógicas. Una descripción más detallada de estos lenguajes de ontologías se presenta en la sección 2.4.

A continuación, según la arquitectura que estamos analizando, se encuentra la capa Lógica. Se suele considerar que el nivel lógico y el ontológico están integrados, ya que la mayoría de las ontologías permiten la definición de axiomas

lógicos. Mediante la aplicación de la deducción lógica (reglas de inferencia) es posible inferir nuevo conocimiento de la información que ya hay explícita.

La capa Pruebas (*Proof*) se refiere a la capacidad de realizar demostraciones y de probar porqué se ha tomado o aconsejado tomar una decisión y no otra, en base a los axiomas de la capa lógica (Codina & Rovira, 2006).

Por último, la capa superior Confianza (*Trust*) es la que garantiza la fiabilidad de las transacciones comerciales realizadas a través de la Web en las que intervienen no sólo personas sino también agentes. No obstante en la actualidad, estas dos últimas capas no se tienen en cuenta.

2.3. El concepto de ontología

En los últimos años el proyecto de la Web Semántica ha servido para asentar lo que podríamos denominar el uso actual del término ontología (Pedraza-Jiménez et al., 2007). No obstante, a pesar de su apogeo, este concepto no es novedoso y ya había sido desarrollado (con otros sentidos) en el ámbito filosófico.

El concepto de ontología ha cambiado a lo largo de la historia, añadiéndose a su sentido original nuevos significados procedentes de distintos dominios. Nacido en el seno de la filosofía, el término ontología fue utilizado por primera vez por J.C. Wolf a principios del siglo XVIII, que entendió por ontología la ciencia del ser como tal. En su doctrina, ontología es equivalente a la metafísica general, y se planteaba el problema sobre cómo estaba organizado el ser. Aunque el concepto ha evolucionado a lo largo de los siglos, se ha mantenido dentro de un ámbito filosófico. Autores como Leibniz (Leibniz en Couturat, 1903), Kant (Kant, 2001) o Sowa (Sowa, 20009), también han aportado sus propias definiciones de ontología. Algunas de las definiciones propuestas son:

Ontología es la ciencia de algo y de nada, del ser y del no ser, de la cosa y del modo de la cosa, de la sustancia y el accidente (Leibniz, en Couturat 1903)

La filosofía trascendental es el sistema de todas nuestras cogniciones puras a priori, que podemos llamar ontología. Así, ontología trata con cosas en general, desde abstractas hasta particulares. Abarca todos los conceptos puros de la comprensión y todos los principios de la razón. Las ciencias principales que pertenecen a la metafísica son: ontología, cosmología, y teología. Ontología es una pura doctrina de elemento de toda nuestra cognición al completo, o: contiene la suma de todos nuestros conceptos puros que podemos tener a priori sobre la cosas **(Kant, 2001)**

Estas definiciones de carácter filosófico han dado paso en las últimas décadas a otras definiciones del concepto de ontología que se enmarcan dentro de un ámbito muy distinto: la Inteligencia Artificial (IA) y más recientemente en el ámbito de la Web Semántica. En este sentido, una de las primeras y más citadas definiciones de ontología en este nuevo marco es la de Gruber:

Una ontología es una especificación explícita de una conceptualización. El término proviene de la filosofía, donde una ontología es un recuento sistemático de la existencia.
En sistemas de Inteligencia Artificial, lo que existe es lo que puede ser representado.
Cuando el conocimiento de un dominio se representa mediante un formalismo declarativo, el conjunto de objetos que puede ser representado se llama universo del discurso. Esos conjuntos de objetos, y las relaciones que se establecen entre ellos, son reflejados en un vocabulario con el cual representamos el conocimiento en un sistema basado en conocimiento. Así, en el contexto de IA, podemos describir la ontología de un programa como un conjunto de términos. En tal ontología, las definiciones asocian nombres de entidades del universo del discurso con textos comprensibles por los humanos que describen el significado de los nombres, y axiomas formales que limitan la interpretación y buen uso de

dichos términos. Formalmente, una ontología es una teoría lógica (Gruber, 1993).

Esta primera aproximación al concepto de ontología en el ámbito de la IA ha suscitado diversas críticas como por ejemplo en (Guarino, 1995), en parte motivadas por su carácter ambiguo. De igual modo, otros autores como Borst (Borst, 1997) la han perfilado, dando lugar a una definición más precisa. *Una ontología es una especificación formal de una conceptualización compartida*.

Como se indica en (Pedraza-Jiménez et al., 2007), una "conceptualización" es un modelo abstracto de algún fenómeno del mundo construido mediante la identificación de los conceptos relevantes de ese fenómeno (normalmente un dominio del conocimiento). "Explícito" significa que los conceptos utilizados en la ontología, y las restricciones para su uso, están claramente definidos. "Formal" se refiere al hecho de que debe ser comprensible para las máquinas, es decir, estar expresada mediante sintaxis (como el lenguaje OWL) que permita a un ordenador operar sobre ella. Por último, "compartida" refleja la noción que contendría conocimiento consensuado en algún grado (en el caso de un dominio de conocimiento, se supone que estará consensuado por los expertos en él).

En la actualidad, no existe una definición consensuada de ontología, aunque, como se ha mencionado, la más extendida es la de Gruber.

2.3.1 Componentes de una ontología

El contenido de una ontología se formaliza usando cinco tipos de componentes: clases, relaciones, propiedades, axiomas e instancias (Gruber, 1993).

Clases o tipos: una clase es un conjunto de objetos (físicos, tareas, funciones, etc.). Cada objeto en una clase es una instancia de esa clase. Desde el punto de vista de la lógica los objetos de una clase, se pueden describir especificando las propiedades que éstos deben satisfacer para pertenecer a esa clase. Las clases son

la base de la representación del conocimiento en las ontologías, ya que describen los conceptos del dominio. Una clase se puede dividir en subclases, las cuales representarán conceptos más específicos que la clase a la que pertenecen. Una clase cuyos componentes son clases, se denomina superclase o metaclase.

Relaciones: se establecen entre conceptos de una ontología para representar las interacciones entre éstos. Están definidas, por lo general, como el producto cartesiano de n conjuntos: R: C1 x C2 x …x Cn. Algunas de las relaciones más utilizadas son:

- Instancia de: asocian objetos a clases.
- Relaciones temporales: implican precedencia en el tiempo.
- Relaciones topológicas: establecen conexiones espaciales entre conceptos.
- Relaciones taxonómicas.
- Relaciones partonómicas.

Axiomas: elementos que permiten el modelado de verdades que se cumplen siempre en el dominio. Los axiomas pueden ser estructurales o no estructurales. Un axioma estructural establece condiciones relacionadas con la jerarquía de la ontología, conceptos y atributos definidos. Un axioma no estructural establece relaciones entre atributos de un concepto y son específicos de un dominio.

Instancias o individuos: son objetos, miembros de una clase, que no pueden ser divididos sin perder su estructura y características funcionales. Pueden ser agrupados en clases.

Propiedades o slots: los objetos se describen por medio de un conjunto de características o atributos que son almacenados en los slots. Éstos almacenan diferentes clases de valores. Las especificaciones, rangos y restricciones sobre estos valores se denominan características o facetas. Para una clase dada, los slots y las restricciones sobre ellos son heredados por las subclases y las instancias de la clase.

2.3.2 Clasificación de los tipos de ontologías

En la literatura se encuentran diferentes clasificaciones de tipos de ontologías. Principalmente, se siguen dos criterios para dichas clasificaciones: el tipo de conocimiento que contienen y la motivación de la ontología.

2.3.2.1 Clasificación por el conocimiento que contienen

Este es el criterio donde existe mayor diversidad, como se puede ver en las dos siguientes clasificaciones de ontologías. La primera de ellas propuesta por (van Heijst et al, 1997), distingue tres tipos de ontologías:

- **Ontologías terminológicas y lingüísticas**. Especifican los términos usados para representar conocimiento en un dominio determinado. Un ejemplo de ontologías terminológicas es la red semántica UMLS (Unified Medical Language System) (Lindberg et al, 1993). En cuanto a las ontologías lingüísticas, una de la más utilizadas es WordNet (Miller, 1995; Fellbaum, 1998) (ver apartado 3.2.1.4.1), que es una gran base de datos léxica en la que se contemplan distintos tipo de relaciones.
- **Ontologías de información**. Especifican la estructura de los registros de la base de datos. Los esquemas de bases de datos son un ejemplo.
- **Ontologías para modelar conocimiento**. Especifican conceptualizaciones de dominios. Estas ontologías tienen una estructura interna mucho más rica que los anteriores tipos y son las ontologías que interesan a los desarrolladores de sistemas basados en conocimiento.

Una clasificación alternativa es la que se puede encontrar en (Mizoguchi et al, 1995), donde también se proponen tres categorías:

- **Ontologías de tarea**. Establecen la forma en la cual se puede usar el conocimiento del dominio para realizar tareas específicas. De esta forma,

una aplicación podría realizar búsquedas de información, mientras otra podría gestionar la asignación de bloques libre de memoria.

- **Ontologías de dominio**. Contienen todos los conceptos asociados a un dominio particular.
- **Ontologías generales**. Contienen descripciones generales sobre objetos, eventos, relaciones temporales, relaciones causales, modelos de comportamiento y funcionalidades.

2.3.2.2 Clasificaciones por motivación

A continuación se presentan dos clasificaciones distintas atendiendo al criterio de la motivación. Según la primera de ellas, se distinguen cuatro tipos de ontologías:

- **Ontologías para la representación de conocimiento**. Permiten explicar las conceptualizaciones que subyacen en los formalismos de representación de conocimiento (Davis et al, 1993).
- **Ontologías genéricas**. Definen conceptos considerados genéricos en diferentes áreas. Ejemplos de tales conceptos son componente, subclase, proceso, estado, etc. Estas ontologías son reutilizables en diferentes dominios. Se llaman también ontologías abstractas o superteorías porque permiten definir conceptos abstractos.
- **Ontologías del dominio**. Definen conceptualizaciones específicas del dominio. Las metodologías actuales de adquisición de conocimiento distinguen entre ontologías y conocimiento del dominio que describe situaciones factuales del dominio, mientras que las ontologías imponen descripciones sobre la estructura y contenido del conocimiento del dominio.
- **Ontologías de aplicación**. Están ligadas al desarrollo de una aplicación concreta. Normalmente estas ontologías toman conceptos de ontologías del dominio y genéricas, así como métodos específicos para realizar la tarea, por lo que no suelen ser reutilizables

Una clasificación alternativa es la propuesta por Poli (Poli, 2000). En dicha clasificación, se identifican los siguientes tipos de ontologías:

- **Ontologías generales**. Tienen que ver con las categorías fundamentales y sus conexiones de dependencia. Existen categorías fundamentales que se aplican a todos los niveles ontológicos. Sin embargo, muchas de las categorías de alto nivel pueden tener diferentes valores en niveles diferentes de la ontología, aunque deben tener algo en común.
- **Ontologías categóricas**. Estudian las diversas formas en las que una categoría da cuenta de los diversos niveles ontológicos, determinando la posible presencia de una teoría general que subsume sus concretizaciones. Mientras que la ontología general está más relacionada con la arquitectura de la teoría, la ontología categórica es más sensible a los detalles de las categorías individuales. Sin embargo, es obvio que ambas son necesarias.
- **Ontologías del dominio**. Se refieren a la estructuración detallada de un contexto de análisis con respecto a los subdominios que lo componen.
- **Ontologías genéricas**. Aparecen ligadas a corpus lingüísticos y léxicos conceptuales. De hecho, los términos se pueden clasificar en varios niveles. Esto significa que cada término debería ser accesible por defecto únicamente en su sentido genérico, mientras que sus significados especializados quedan para cuando se active una ontología del dominio específica. Por otro lado, la ontología del dominio contiene términos que no tienen correspondencias analíticas en ontologías genéricas. El conocimiento del dominio "satura" el conocimiento genérico.
- **Ontología regional.** Analiza las categorías y sus conexiones de interdependencia para cada nivel ontológico (estrato o capa).
- **Ontología aplicada**. Estas ontologías son la aplicación concreta del entorno ontológico a un objeto específico (por ejemplo, un hospital).

2.3.2.3 Clasificación por el grado de formalidad

Las ontologías, también se pueden clasificar basándose en el grado de formalidad de la ontología. Según este criterio Poli (2003) distingue entre:

- **Ontología descriptiva,** relacionada con la recolección de información sobre los ítems del dominio analizado. La unidad y variedad del mundo es la salida de las conexiones de dependencia y formas de independencia entre los ítems. Cosas materiales, plantas y animales, así como los productos de los talentos y actividades de animales y humanos, son ítems del mundo. En otras palabras, el mundo no sólo contiene cosas, animadas o no, sino también actividades y procesos, así como los productos derivados de los mismos. Es difícil negar que existen pensamientos, sensaciones y decisiones, así como el completo espectro de actividades mentales, y estamos obligados a admitir la existencia de reglas, lenguajes, sociedades y costumbres (Poli, 2001).
- **Ontología formal,** que destila, filtra, codifica y organiza los resultados de una ontología descriptiva. Según esta interpretación, la ontología formal lo es en el sentido de Husserl en sus *Logical Investigations*. Ser formal en este sentido implica tratar con categorías como cosa, proceso, materia, forma, todo, parte, etc. Estas categorías caracterizan aspectos y tipos de realidad que todavía no han sido utilizados bajo ningún formalismo. La ontología formal se ha desarrollado de dos maneras principales (Albertrazzi, 1996). El primer enfoque consiste en estudiar la ontología formal como parte de la ontología, y analizarla usando las herramientas de modo que se aproxima a la lógica formal. Desde este punto de vista, la ontología formal examina las características lógicas de predicación, así como aquellas de las diferentes teorías de universales. El uso del paradigma específico de la teoría de conjuntos aplicada a predicación condiciona su interpretación. El segundo enfoque vuelve a los orígenes Husserlianos y analiza las categorías fundamentales de objeto, estado,

parte, todo, etc., así como las relaciones entre partes y todos y sus leyes de dependencia, una vez que los conceptos materiales han sido sustituidos por sus conceptos formales correlativos al "algo" puro. Este tipo de análisis no trata con el problema de relaciones entre ontología formal y ontología material.

2.4. Lenguajes ontológicos

Los lenguajes ontológicos son el vehículo para expresar ontologías de forma comprensible por las máquinas. Desde los años 90, se han presentado diversos propuestas, aunque ha sido OWL el que se ha impuesto sobre todos los demás en parte por ser la recomendación del Consorcio World Wide Web (W3C) (http://www.w3.org/) y porque de los lenguajes desarrollados es el que presenta una mayor expresividad y adecuación a la Web Semántica, como se verá a continuación.

Uno los predecesores de OWL es RDF (Resource Description Framework) (Lassila & Webick, 1998). Fue desarrollado por el W3C como un lenguaje general para representar información de la web, ya que es capaz de especificar contenido semántico de un modo estandarizado, interoperable y basado en XML. Esta especificación muestra tres representaciones del modelo de datos, como tripletas, como grafo y en XML.

RDF se basa en grafos que se pueden expresar como sentencias en las que el sujeto es el recurso a describir, el predicado es una propiedad o característica propia de dicho recurso y el objeto es un valor concreto que tiene dicha característica.

Este lenguaje, de expresividad suficiente para representar contenido en la Web, no posee sin embargo la capacidad de expresión necesaria para la Web Semántica.

Con el objetivo de aprovechar al máximo la semántica integrada en la nueva web se propuso el lenguaje OWL.

OWL (del inglés *Ontology Web Language*), es una propuesta de estandarización de lenguaje ontológico, especificado por un grupo de trabajo Web Ontology del consorcio W3C (McGuinness & Harmelen, 2007), que ayudaría a solucionar los impedimentos actuales para la construcción cooperativa de ontologías entre diferentes plataformas de construcción ontológica, además de potenciar a la Web Semántica.

Los lenguajes ontológicos anteriores a OWL han sido utilizados para desarrollar herramientas y ontologías destinadas a comunidades específicas (especialmente para ciencias y aplicaciones específicas de comercio electrónico). Estos lenguajes no fueron definidos para ser compatibles con la arquitectura de la World Wide Web en general, y la Web Semántica en particular. OWL da solución a este problema proporcionando un lenguaje, que utiliza la conexión proporcionada por RDF para añadir las siguientes capacidades a las ontologías:

- Capacidad de ser distribuidas a través de varios sistemas.
- Escalable a las necesidades de la Web.
- Compatible con los estándares Web de accesibilidad e internacionalización.
- Abierto y extensible.

El primer borrador de la especificación de este lenguaje (OWL 1.0) apareció en julio de 2002 y se presentó de manera formal por el W3C en febrero de 2004 (McGuinness & Harmelen, 2007). En octubre de 2009 apareció la ultima versión hasta el momento, la 2.0 (W3C, 2009). En OWL 1.0 se definieron los tres siguientes sublenguajes de expresividad creciente:

- **OWL Lite**: es el sublenguaje menos expresivo. Comparado con RDFS, el lenguaje anterior a OWL, añade restricciones de rango local, restricciones existenciales, restricciones de cardinalidad simple y varios tipos de propiedades (inversa, transitiva y simétrica). Está destinado a usuarios que necesiten sobre todo una jerarquía de clasificación y restricciones sencillas.

- **OWL DL**: proporciona una expresividad máxima con garantías de cómputo incluyendo todas las construcciones de su lenguaje con el inconveniente de que estas construcciones sólo pueden ser utilizadas bajo ciertas restricciones. Su nombre proviene de la correspondencia que mantiene con la lógica descriptiva y es el más usado en la actualidad, por el hecho de asegurar la completitud y finitud de los razonamientos.

- **OWL Full**: está destinado a usuarios que desean una expresividad máxima y la libertad sintáctica ofrecida por RDF aunque sin garantías de cómputo. Permite que una ontología aumente el significado, RDF u OWL, del vocabulario predefinido.

En la versión 2.0 de OWL, además de otras mejoras, la nomenclatura de sublenguajes ha sido modificada al concepto de perfil. Esta nomenclatura, además de mantener los tres lenguajes anteriores, está orientada a las características particulares interesantes para cierto tipo de aplicaciones.

- **OWL 2 EL**: perfil cercano a la lógica descriptiva EL++, que asegura un tiempo polinomial para la resolución de problemas de razonamiento. Es particularmente útil en aplicaciones que emplean ontologías que contienen un gran número de propiedades y/o clases. Es sencillo de implementar y permite una gran escalabilidad para expresiones complejas, aunque la expresividad que proporciona es bastante limitada. Por ejemplo, no permite el uso de cuantificadores universales.

- **OWL 2 QL**: destinado a aplicaciones que utilizan un gran número de instancias de datos y en los que la consulta es la tarea de razonamiento más importante. Es una variante de OWL-Lite, muy habitual en tareas de integración en bases de datos. Resulta muy sencillo extender los

habituales lenguajes relacionales (como SQL), incorporando consultas con los axiomas definidos por el subconjunto. Finalmente, facilita el mapeo entre UML y los diagramas Entidad-Relación, con lo que la representación de esquemas de datos es bastante inmediata. En cuanto a expresividad sigue siendo bastante limitada, por ejemplo no permite cuantificadores existenciales ni propiedades encadenadas.

- **OWL 2 RL**: indicado para aplicaciones que requieren un razonamiento escalable sin sacrificar demasiada expresividad. Está orientado fundamentalmente a ser utilizado en otras tecnologías basadas en reglas, facilitando el razonamiento.

En la siguiente figura (Figura 2.3) se pueden ver representados de forma esquemática los lenguajes descritos previamente.

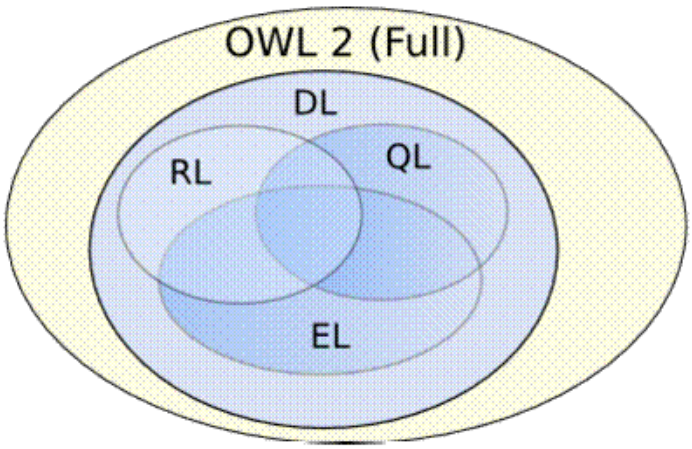


Figura 2.3 Diagrama de Venn de los perfiles de OWL 2.0 (W3C, 2009).

2.5. Las Ontologías en contraposición a otros sistemas de organización del conocimiento.

Las estructuras que se utilizan para organizar la información presentan distintos grados de complejidad, pudiendo ir desde una lista de términos como puede ser un vocabulario de un dominio o una lista de palabras clave, pasando por el establecimiento de relaciones entre los términos, como, por ejemplo, ocurre en un tesauro, hasta la posibilidad de realizar inferencias lógicas, como sucede con las ontologías. Por su parte, las relaciones que se establecen entre términos (conceptos en el caso de las ontologías, y descriptores y no-descriptores en el caso de los tesauros), presentan a su vez distintos grados de complejidad, siendo las más extendidas las taxonómicas o jerárquicas, probablemente seguidas por las partonómicas. A partir aquí, se abre un amplio abanico de relaciones.

A continuación se presenta la clasificación que realiza Moreiro (2007) de los principales sistemas existentes para la representación[3] de contenidos digitales:

- **Taxonomía**: Red semántica de conceptos interrelacionados. Generan sus estructuras jerárquicas de acuerdo con un contexto y unos usuarios determinados.

- **Tesauro documental**: Lista de descriptores formada con las posibilidades de representación de los conceptos generales en los documentos de un dominio concreto. Es un vocabulario destinado exclusivamente a la indización y la recuperación de la información. El tesauro debe partir de una categorización y por tanto de una taxonomía del conocimiento temático.

[3] Representación se utiliza aquí en un sentido documental, como el conjunto de términos u otros símbolos que permitan identificar, clasificar y localizar un documento o una parte del mismo.

- **Ontologías**: Tienen como misión representar el conocimiento a partir de la organización taxonómica en cuanto al modo de clasificación o categorización jerárquica de los conceptos pertenecientes a un conjunto temático.

- **Tesauros automáticos o conceptuales**: Red semántica en la que cada nodo contiene un único concepto semántico que puede tener una serie de descriptores asociados que se identificarían según las relaciones habituales en los tesauros. Puede verse como una red semántica conceptual por la que se navega desde los términos más genéricos de una faceta hacia los más específicos. Una de las propuestas de mejora de los tesauros es la inclusión de verbos.

- **Mapas conceptuales:** Grafos o redes conceptuales constituidas por conceptos y relaciones entre conceptos. Son colecciones ordenadas de nodos conectados por arcos que se usan para representar documentos.

- **Topic Map**: Un documento o conjunto de documentos SGML o XML interrelacionados en un espacio multidimensional en el que las localizaciones se denominan *Topics*. No tienen inferencia, ni permiten describir reglas ni axiomas.

- **Folksonomías:** Conjuntos de palabras clave incorporadas y asignadas por cualquier internauta para colaborar en la indización de todo tipo de contenidos en el espacio Web compartido y abierto.

A excepción de los Topic Maps, lo que tienen en común todos estos sistemas de representación de conocimiento es que se estructuran en torno al lenguaje humano con mayor o menor grado de formalización. Un mayor grado de normalización y estandarización implica un mayor control sobre las unidades lingüísticas empleadas para representar el conocimiento. Estableciendo de nuevo

una gradación que parta de las representaciones más informales a las más formales, la escala resultante sería: palabra- término- descriptor- concepto.

Según la Recomendación Estándar Británica para la selección, formación y definición de términos técnicos (en Sager & Nkwenti-Azeh, 1990), los conceptos son constructos mentales (abstracciones) que se pueden emplear para clasificar los distintos objetos del mundo exterior e interior, mientras que los términos son las unidades léxicas concretas que se emplean para referirse a un concepto.

Spasic (2004) puntualiza que los términos son representaciones lingüísticas de los conceptos específicos de un domino, siendo ésta una definición ampliamente aceptada.

No siempre existe un único término para referirse a un concepto. Por este motivo, con el objetivo de facilitar la comunicación científico-técnica, ha habido intentos de imponer una relación biunívoca entre término-concepto como en la norma ISO 704 (Thomsen & Madsen, 2009), atribuyendo un término a cada concepto y representando cada concepto por un solo término. En una ontología, las clases están representadas por un concepto seleccionado de entre el conjunto de términos posibles con los que se puede designar una clase, es decir, un concepto engloba todos los términos con los que se puede nombrar una clase.

Ontologías, terminologías y el lenguaje natural tienen propósitos diversos, como se señala en (Baud et al., 2007). Más concretamente, las ontologías permiten la comunicación semántica hombre-máquina, el lenguaje natural permite la comunicación entre humanos, y las terminologías deberían ser el puente de unión entre ambos.

De la clasificación de los sistemas de representación del conocimiento mencionada previamente, se desprende que las ontologías aportan información semántica que otros sistemas no aportan. Además, son entendibles por el computador y por los seres humanos. Las folksonomías, por su parte, pueden aportar una información más rica semánticamente. Además, son creadas por los usuarios en entornos colaborativos, por lo que su coste es bajo. No obstante, la falta de normalización las convierten en sistemas clasificatorios poco fiables, ya

que, junto con las ventajas, presentan todos los inconvenientes del uso de lenguaje natural para organización de la información. En las folksonomías no existe un control sobre el uso de términos sinónimos, ni el uso del singular y plural. Además, las categorizaciones, en el caso de existir, dependen del usuario por lo que responden más a las necesidades personales que a un afán clasificatorio universal. Por otro lado no se establecen relaciones entre los términos más allá de la taxonomía que en ocasiones puede crear el usuario.

De la complejidad del lenguaje natural y de los principales problemas que presenta para la representación de la información, se trata en el siguiente capítulo de esta memoria.

El grado de normalización que presentan las ontologías puede ser equivalente al que aporta un tesauro. Sin embargo, este último no es entendible por un ordenador, no soporta procesos de inferencia automáticos y además la riqueza potencial de las relaciones es menor que en una ontología. Aunque un tesauro no es equivalente a una ontología, sí que puede constituir la base para la construcción de una ontología. Así, uso de un tesauro existente puede aportar un buen número de conceptos a la ontología de dominio, así como relaciones jerárquicas que aparecen explícitas y las relaciones de equivalencia (sinonímicas). Existen algunas propuestas para la adaptación de los tesauros a la web semántica como se puede ver por ejemplo en (Pastor Sánchez, 2009).

Como se indica en (Sánchez-Jiménez & Gil-Urdiciain, 2007), una de las principales diferencias entre los lenguajes documentales y las ontologías es la capacidad para representar instancias de los conceptos:

En una ontología, las clases se definen por un conjunto de propiedades o atributos que tendrán valores distintos en diferentes instancias de dichas clases y se declararán de forma explícita. En los lenguajes documentales las instancias no se explicitan en ninguna parte, y no se define el conjunto de atributos que hacen que esa instancia se a miembro de una clase, por lo que no se codifican de forma

que resulten explícitas para una máquina. (Sánchez-Jimenez & Gil-Urdiciain, 2007)

Por otro lado, los sistemas de clasificación como las ontologías, en su acepción informática, y las folksonomías nacen para la web, que es un sistema descentralizado y heterogéneo (Pedraza-Jiménez et al., 2007), mientras que los sistemas de clasificación tradicionales nacen para entornos cerrados, en donde el flujo de información entrante está controlado.

Independientemente de ser la estructura de conocimiento elegida dentro de la Web Semántica, las ontologías, como sistema de representación del conocimiento, presentan un mayor potencial que el resto de sistemas. Sin embargo, uno de los principales hándicap que impiden su proliferación es, como se ha dicho, el coste y dificultad de su construcción y mantenimiento. De ahí la importancia del desarrollo de metodologías que permitan automatizar el máximo posible el proceso.

El lenguaje es multiforme y heteróclito; a caballo en diferentes dominios, a la vez físico, fisiológico y psíquico, pertenece además al dominio individual y al dominio social; no se deja clasificar en ninguna de las categorías de los hechos humanos, porque no se sabe cómo desembrollar su unidad.
La lengua, por el contrario, es una totalidad en sí y un principio de clasificación.
(Saussure, Curso de Linguística General, p.33)

CAPÍTULO 3

EL PROCESAMIENTO DE LENGUAJE NATURAL

Resumen En este capítulo se describe el Procesamiento de Lenguaje Natural (PLN) y los distintos niveles de los que consta, esto es, procesamiento fonético, preprocesamiento textual, procesamiento morfológico, sintáctico, semántico y finalmente pragmático. Al hilo de cada uno de los niveles se plantean, desde un punto de vista lingüístico, las dificultades que encontramos en las diferentes fases para llevar a cabo el procesamiento. De igual modo, se analizan algunos de los principales recursos existentes tales como WordNet. Se hace especial hincapié en los sistemas para el etiquetado de roles semánticos, profundizando en algunos de ellos como FrameNet o VerbNet para dominios generales y PasBio o BioProp para el dominio de la biomedicina. Finalmente se describen GATE y Freeling.

3.1 Introducción

El lenguaje natural es el vehículo de comunicación más profusamente utilizado por el ser humano, siendo las lenguas manifestaciones concretas del lenguaje. La lengua como sistema semiótico de naturaleza lingüística, se manifiesta en cada una de las realizaciones individuales de los miembros de una comunidad, que la utilizan como código común y compartido sobre el que tienen un conocimiento parcialmente idéntico (Hernández Terrés & Escavy Zamora, 1999). La lengua está

presente tanto en los actos de habla como en la escritura, y ésta última, a lo largo de los siglos, ha sido el medio principal para la transmisión de conocimiento.

La llamada era digital ha traído consigo nuevos soportes documentales, generando novedosas perspectivas para el acceso y uso de los datos contenidos en los documentos que el soporte en papel no ofrecía. Pero, a su vez, se ha hecho patente la necesidad de invertir en el desarrollo de herramientas que permitan aprovechar la potencialidad ofrecida por los soportes digitales. La ambiciosa tarea de compartir a todos los niveles el lenguaje natural con un ente inanimado, como puede ser un ordenador, es lo que da origen al denominado Procesamiento de Lenguaje Natural (PLN).

En este capítulo se realiza una aproximación general al PLN, describiendo las distintas fases de las que generalmente consta un sistema encargado de dicho procesamiento y que se corresponden con los diferentes niveles de análisis lingüístico. Al hilo de cada una de estas fases, se introduce la descripción de herramientas relevantes en cada uno de los ámbitos de análisis, por ejemplo lexicones computacionales como WordNet, proyectos para la anotación de roles semánticos de carácter general como VerbNet, FrameNet y PropBank y otros desarrollados para el dominio de la biomedicina como BioFrameNet, PasBio o BioProp. Además, se describe la herramienta GATE que es una infraestructura para el desarrollo de e implementación de herramientas para el PLN. Por otro lado, se mencionan las características de Freeling, un conjunto de herramientas que permite llevar a cabo tareas relativas al análisis morfo-sintáctico e identificación de entidades nombradas en castellano. Finalmente, se relaciona PLN y Ontologías, tópico que está presente en el resto de capítulos de esta tesis.

3.2 El Procesamiento de Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN) se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio del lenguaje natural.

El PLN puede verse como un intento de simular el comportamiento lingüístico humano, de manera que el sistema de signos, que constituye la lengua, sea adquirido y procesado por el computador, siendo éste capaz de reconocer, comprender, interpretar y generar lenguaje humano, ya sea escrito o hablado.

En el intento por la automatización de los procesos lingüísticos han confluido distintas disciplinas, tanto informáticas como lingüísticas, dando lugar a denominaciones que sintetizan esa hibridad. Este es el caso de la lingüística computacional, que puede verse como un conjunto heterogéneo de teorías, métodos, herramientas, aplicaciones y productos que tienen en común la consideración de la lengua como un objeto susceptible de ser tratado mediante procedimientos informáticos (Llisterri, 2003).

En sus inicios, el PLN se centró principalmente en tres áreas: la traducción automática, el reconocimiento del habla y el acceso a bases de datos (Jackson & Schilder, 2006). Aunque estos tres elementos siguen siendo objeto de las investigaciones en PLN, a lo largo de las décadas y con el desarrollo de nuevas tecnologías como Internet, se han incorporado nuevos usos, entre ellos la recuperación de información, los sistemas de diálogo, la búsqueda de respuestas, el resumen automático o la extracción de información que desempeña un papel importante en la construcción automática de ontologías.

La arquitectura de los sistemas y metodologías que existen para el PLN se han articulado en torno a los niveles de análisis lingüístico tradicionales, esto es, fonética (en los sistemas de reconocimiento del habla), morfología, sintaxis, semántica y pragmática.

En la siguiente figura (Figura 3.1), se pueden visualizar las diferentes fases o niveles de procesamiento.

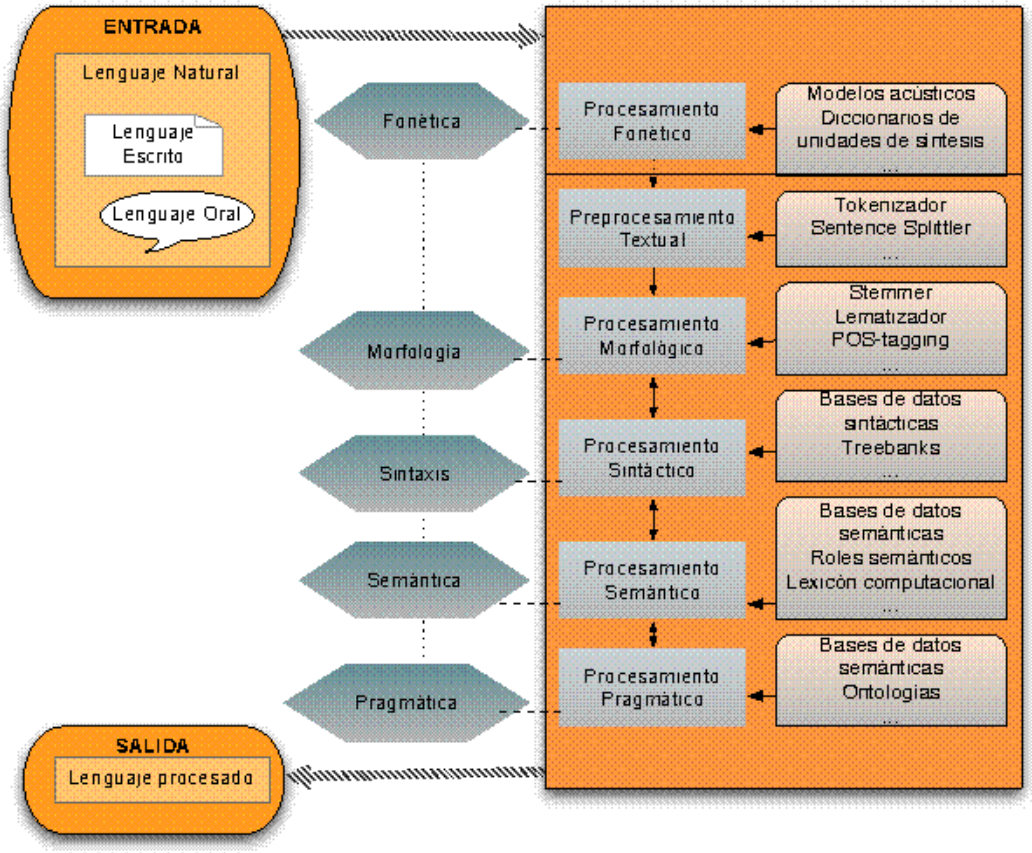


Figura 3.1 Fases del Procesamiento de Lenguaje Natural.

A continuación, se describen brevemente cada una de estas fases.

- El **Procesamiento fonético.** Sólo es aplicable a aquellos sistemas que incluyen reconocimiento de voz, ya sea como entrada y/o como salida. La fonética se encarga de la descripción de las dimensiones físico-acústicas, articulatorias y auditivas de los sonidos del lenguaje. Esta parte es necesaria para el desarrollo de tecnologías del habla. El procesamiento fonético está fuera del alcance de esta tesis.

- **Preprocesamiento textual**. En primer lugar se realiza una segmentación del texto para identificar tokens y otras unidades relevantes para el análisis, tales como oraciones y párrafos, se trata de

una tarea de preprocesado, a partir de la cual se puede llevar a cabo el resto del procesamiento lingüístico.

- **Procesamiento léxico-morfológico**. La morfología estudia la estructura de la forma de las palabras, básicamente a través del uso de morfemas (flexivos y derivativos). Basándose en esos morfemas y en la posición que ocupan con respecto al lexema, se pueden determinar aspectos como el tiempo, el género, el número, el grado, etc. El procesamiento morfológico identifica y clasifica las unidades lingüísticas en las distintas categorías gramaticales: sustantivo, verbo, adjetivo, adverbio, etc.

  El **procesamiento sintáctico**. La sintaxis estudia las relaciones que se establecen entre las palabras dentro de la oración y las reglas que rigen estas relaciones. Por lo tanto, el procesamiento sintáctico obtiene la estructura de constituyentes de una oración, basándose, por lo general, en una gramática previamente definida. Uno de los formalismos sintácticos más habituales basado en gramáticas es el de las gramáticas de estructuras sintagmáticas (*Phrase Structure Grammar*, PSG).

- El **procesamiento semántico.** Consiste en el establecimiento de una serie de relaciones semánticas entre las estructuras textuales con el objetivo de determinar el sentido en el que se está utilizando cada unidad lingüística en el discurso. La adquisición de conocimiento semántico es crucial para mejorar las aplicaciones en lenguaje natural.

- El **procesamiento pragmático.** Analiza el texto en su conjunto, teniendo en cuenta el contexto comunicativo en el que se inserta. Los principales elementos de este análisis son la resolución de la anáfora, la catáfora, la correferencialidad y otros elementos deícticos. Para sustentar el análisis pragmático se suelen utilizar ontologías de dominio.

Las fases o niveles de procesamiento no son independientes unas de otras, sino que están interrelacionadas. A medida que se va avanzando en el análisis, será necesario recurrir al conocimiento extraído en niveles anteriores, e incluso posteriores. Por ejemplo, para la desambiguación de las categorías morfológicas es necesario recurrir a la sintaxis para determinar la función de un término en la oración.

Por otro lado aunque el procesamiento en los niveles inferiores como el análisis morfológico y la desambiguación han alcanzado cotas aceptables de eficacia, no se puede decir lo mismo de niveles superiores como la semántica y la pragmática debido a su complejidad.

Cada una de estas fases se describe en más profundidad a continuación.

3.2.1 Preprocesamiento textual.

Esta fase no forma parte del análisis lingüístico como tal, pero es necesaria para la obtención de unidades procesables computacionalmente.

Una secuencia textual está formada por caracteres alfanuméricos, símbolos y espacios en blanco. A su vez, las unidades discursivas son la palabra, la oración, el párrafo y finalmente el discurso. El preprocesamiento textual identifica estos elementos dentro del texto. Entre los procesos que se realizan en esta fase podemos destacar la tokenización y la identificación de oraciones. A continuación, se explican en detalle estos procesos.

3.2.1.1 Tokenización

Un token es una cadena de caracteres alfabéticos, numéricos y/o simbólicos separada del resto por un espacio en blanco. La tokenización consiste en la segmentación de los caracteres del texto de entrada en unidades procesables lingüísticamente, es decir, computacionalmente.

Se suele distinguir entre:

- **Palabras**: Cadenas de caracteres alfabéticos y alfanuméricos, en mayúscula o minúsculas, constituyen generalmente las unidades léxicas del texto.
- **Números**: Cadenas de caracteres numéricas.
- **Símbolos**: Símbolos permitidos por el código empleado. Por ejemplo @ o $.
- **Puntuación**: Identifica los distintos signos de puntuación del texto.

Una estrategia sencilla para llevar a cabo la tokenización de un texto es la división del mismo mediante la localización de espacios en blanco y signos de puntuación, aunque dicha estrategia no contempla una serie de dificultades (Ananiadou & McNaught, 2006) tales como:

- **Abreviaturas**: Un signo de puntuación no siempre indica el final de una palabra o de una oración. Por ejemplo *ej.*, *Vol.*
- **Apóstrofes**: Aunque en castellano no son frecuentes las apóstrofes, sí lo son en otros idiomas como el inglés, por ejemplo en *Bloom's syndrome protein.*
- **Separación por guiones**: No siempre está claro dónde debe un tokenizador establecer la división en uno o varios términos separados por guiones. Por ejemplo *teórico-práctico* o *co-operate.*
- **Múltiples formatos**: Los números y otras entidades como direcciones pueden aparecer en distintos formatos. Por ejemplo *4.578* o *4578*

Tradicionalmente, las soluciones empleadas para la resolución de estos problemas en la tokenización se basan en el uso de expresiones regulares, lexicones o una mezcla de ambos.

En el dominio biomédico, la tokenización presenta problemas adicionales como se señala en (Ananiadou & McNaught, 2006), donde un término que se refiere a una misma entidad extralingüística puede expresarse de diferente modo. Por ejemplo, NF-KappaB, NF-Kappa B o NF-Kappa-B.

3.2.1.2 Identificación de oraciones o *Sentence Splitter*

La identificación de oraciones y párrafos se lleva a cabo obteniendo la lista de palabras contenida entre dos signos de puntuación que indican el fin de la oración. Así, el sistema identificará que un punto y seguido indica final de oración y que un punto y aparte indica, además, final de párrafo.

No obstante, la aparición de un punto no siempre significa el fina de una oración. Por ejemplo *El rojo, el amarillo, el verde, etc. son ejemplos de colores* en donde el punto de la abreviatura *etc.* no indica el final de la oración.

3.2.2 Procesamiento Léxico-Morfológico.

Como se ha indicado, el procesamiento-morfosintáctico consiste en la asignación de una categoría gramatical a las distintas unidades léxicas. El análisis morfológico se puede descomponer a su vez en distintas fases que son, la reducción al lexema, la lematización y finalmente la etiquetación morfológica. A continuación, se explican cada uno de estos procesos.

3.2.2.1 Reducción al lexema o *Stemmer*

El *stemming* es una técnica de procesamiento lingüístico que consiste en reducir un término a su raíz, es decir, en separar el lexema de los morfemas que forman una determinada unidad lingüística. Con el *stemming* se obtiene el lexema común que subyace al conjunto de formas que son variantes de una misma unidad.

De este modo, todos los términos que compartan un mismo lexema serán considerados términos similares. Esta técnica se utiliza de manera automática en la recuperación de información, dando lugar al truncamiento. Por ejemplo la mayoría motores de búsqueda actuales utilizan esta técnica para recuperar el mismo término en singular y en plural, en masculino y en femenino.

Sin embargo, para lenguas altamente flexivas como el español y el francés, el stemming no es suficiente, puesto que en una sola unidad lingüística pueden

aglutinarse pronombre, verbo y objeto (por ejemplo en *dímelo*). Además, cuentan con gran número de términos cuyo lexema es irregular, como por ejemplo los verbos irregulares pensar (pienso- pensamos) o morir (moría – muero). Por tanto, en estos casos, es necesario recurrir a la lematización con el fin de que todas estas formas sean clasificadas adecuadamente.

Por otro lado, para la etiquetación gramatical la reducción a la raíz tampoco es suficiente, ya que, un conjunto de palabras que compartan el mismo lexema, no tienen porqué pertenecer a la misma categoría gramatical. Por ejemplo, cantar es un verbo, cantante es un sustantivo y el lexema de ambos es "cant-".

3.2.2.2 Lematización

La lematización consiste en la obtención de la forma canónica de una unidad léxica, esto es, en la identificación del lexema y los posibles morfemas derivativos y flexivos para reducir la palabra al lema. Se considera lema la forma en la que un término aparece representado en un diccionario o enciclopedia. Por ejemplo, en el caso de los verbos el lema es el infinitivo.

Las técnicas utilizadas para la lematización pueden prever el funcionamiento de los morfemas flexivos, pero sólo en algunos casos, el de los morfemas derivativos. Los morfemas flexivos son aquellos que tienen un significado gramatical preestablecido. Con ellos no se crean nuevas unidades léxicas y constituyen un inventario cerrado. Son los utilizados para indicar el género y número en los sustantivos y adjetivos, y el aspecto, diátesis, modo, persona y tiempo en los verbos.

Por ejemplo, sabemos que el morfema de plural en español es –s, -es, o Ø según la terminación del singular sea vocal o –s. De este modo si en el sistema se inserta un término en singular y el mismo término en plural, con la aplicación de estas reglas debería ser capaz de distinguir uno y otro.

Los morfemas derivativos son aquellos utilizados para la creación de nuevas unidades léxicas (con un nuevo significado léxico) mediante el proceso lingüístico

de la derivación, que consiste en añadir un sufijo, prefijo, infijo o interfijo a la raíz de la palabra. Por ejemplo –able puede formar palabras como urbanizable, irritable, adorable a partir de los verbos urbanizar, irritar y adorar.

Aunque existen ciertas reglas en la formación de palabras, no es posible sistematizar el comportamiento de los morfemas derivativos, ya que estos se comportan de forma más irregular que los flexivos.

Así, estos sistemas pueden identificar el término *pequeño* como masculino singular, y a partir de ahí, pueden inferir que el término *pequeños* es masculino plural. Si se han incluido reglas básicas sobre morfología derivativa, el sistema podría identificar que el término *pequeñito* es un diminutivo; pero sin embargo difícilmente podrá inferir que a partir de ellos se puedan formar nuevos vocablos como *empequeñecer*.

3.2.2.3 Análisis Morfológico o *Part-of-Speech-Tagging*

El pos-tagging (*part of speech tagging*) o etiquetación morfológica consiste en la asignación de una etiqueta morfológica a cada una de las unidades léxicas de una secuencia textual dada. El objetivo es determinar cuál es la categoría gramatical a la que una unidad léxica pertenece dentro del conjunto de categorías gramaticales que el sistema de la lengua contempla para dicha unidad.

Una vez identificadas las categorías, se les asigna de forma automática una etiqueta gramatical (nombre, adjetivo, adverbio, conjunción, etc.). Fundamentalmente, se pueden distinguir dos metodologías: las que están basadas en reglas desarrolladas manualmente y las metodologías basadas en aprendizaje automático, por ejemplo, los etiquetadores basados en Modelos Markovianos (Jurafsky et al., 2008).

La ambigüedad es el principal problema que afecta a la etiquetación gramatical. Un término tomado de manera aislada es susceptible de pertenecer a distintas categorías morfológicas. Sin embargo, teniendo en cuenta su entorno

discursivo, habrá una única categoría gramatical válida para cada unidad lingüística.

A continuación, se muestra un ejemplo de ambigüedad en inglés y en español:

- *The light is on*. Aquí podemos observar que *light* es un nombre (luz).
- *This is a light case*. Sin embargo, en este caso, light se puede clasificar como un adjetivo (ligero).
- *Light the fire*. Por último, en este caso esta palabra representa un verbo (encender).
- *Hoy como fuera de casa*. Aquí 'como' actúa como verbo.
- *La vida es como una caja de bombones*. Y en cambio aquí 'como' se comporta con un adverbio.

Es interesante para entender el alcance de la ambigüedad, aludir a la distinción que realiza Coseriu (Coseriu,1987) entre significado léxico y significado categorial. El significado léxico se refiere a lo que está organizado por el lenguaje, y el categorial al modo de organizarlo. Así, el significado léxico se corresponde a qué significa una palabra y el categorial al cómo de la significación. Mientras que los significados categoriales (las categorías gramaticales) no pertenecen a las lenguas como tales, sino al lenguaje en general, los significados léxicos, son diferentes en las distintas lenguas, puesto que cada lengua delimita y estructura de una manera peculiar la realidad conocida.

Siguiendo el ejemplo propuesto por Coseriu, y aplicándolo a la etiquetación morfológica, podemos ver que por ejemplo el término *amo* tiene al menos dos significados léxicos, que son querer y dueño, y a su vez tienen dos significados categoriales, que son verbo y sustantivo.

Sin embargo entre *verde*, en el contexto *el árbol verde* (adjetivo) y *verde* en *el verde es un color* (sustantivo), la diferencia es sólo de significado categorial. (Almela Pérez, 2002).

La clasificación del significado categorial en estos casos se realiza a través del análisis sintáctico, ya que se debe recurrir al contexto para determinar con qué

patrón estructural se corresponde una determinada estructura lingüística. Por ejemplo, si el adjetivo va precedido por un artículo, se produce una transcategorización o cambio de categoría, y, en ese caso, el adjetivo se convierte en un sustantivo.

En cuanto al significado léxico, éste puede recogerse en los lexicones computacionales. Para poder llegar a entender el texto, el sistema necesita un conocimiento amplio del vocabulario de una lengua. En un lexicón computacional se recopila y representa la terminología de uno o varios dominios. Normalmente, los lexicones contienen una lista con las raíces de los términos y de los afijos, junto con un conjunto de reglas morfosintácticas que indican cuáles son las combinaciones válidas entre los distintos elementos, además de las etiquetas morfológicas que se pueden asignar a cada término.

Por ejemplo, dentro del dominio de la biomedicina se ha desarrollado el UMLS *specialist lexicón*[4], en el que se han incluido muchos de los términos biomédicos del inglés. Para cada entrada, en el léxico se ha recogido la información sintáctica, morfológica y ortográfica, de modo que pueda ser utilizada por los sistemas de PLN.

3.2.3 Procesamiento Sintáctico o *Syntactic Parsing*

Mediante el análisis sintáctico se obtiene la estructura de constituyentes de una oración. El análisis sintáctico está estrechamente relacionado con el análisis morfológico ya que la ambigüedad existente entre la adjudicación de una u otra categoría gramatical queda generalmente resuelta con la determinación de las funciones gramaticales de esa unidad dentro de la oración. Por otro lado, los analizadores sintácticos identifican en primer lugar los tipos de sintagma:

[4] http://www.nlm.nih.gov/research/umls/knowledge_sources/index.html#specialist

nominal, preposicional, verbal, adverbial y adjetival, por lo que necesitan las categorías de cada palabra o conjunto de palabras etiquetadas por el pos-tagging.

Existen dos aproximaciones básicas a la hora de abordar el análisis sintáctico:

- **Análisis de dependencias**: La oración se divide en elementos léxicos individuales entre los cuales existe una relación de dependencia binaria, el resultado sería un árbol. Un ejemplo se puede ver en la figura 3.2.

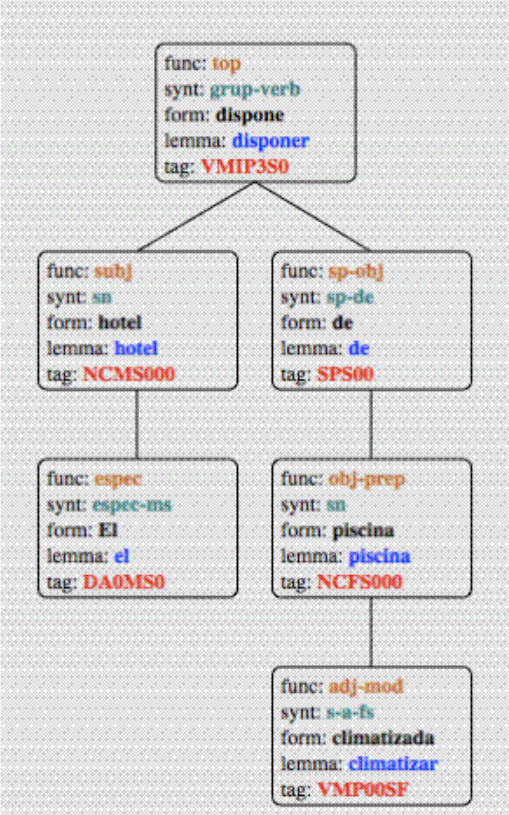


Figura 3.2 Ejemplo de análisis de dependencias con Freeling.

- **Análisis de Constituyentes**: Las relaciones entre los elementos léxicos es de inclusión, se hacen divisiones superiores en las que se van incluyendo otras inferiores. Para llegar de una relación inferior a otra superior habrá

que pasar por todas las relaciones de inclusión intermedias. Se muestra un ejemplo en la figura 3.3.

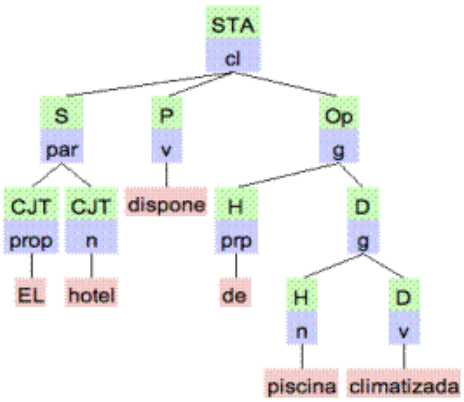


Figura 3.3 Ejemplo de análisis de constituyentes con HISPAL[5] en VISL.

Uno de los principales problemas del análisis sintáctico es la presuposición del conocimiento no explícito como la elipsis. En español es frecuente elidir u omitir el sujeto de la oración por ejemplo en *(Él) Llegó muy tarde* **o el verbo en** *María no estaba esperando, ni Juan tampoco (estaba esperando)* **o** *¿Cuándo te vas? (Me voy) mañana*

Un buen análisis sintáctico no debe ceñirse a los límites de cada oración, sino que para resolver ambigüedades deberá recurrir al análisis del discurso completo.

[5] http://beta.visl.sdu.dk/visl/es/parsing/automatic/trees.php

3.2.4 Procesamiento semántico

El procesamiento semántico consiste en establecimiento de una serie de relaciones de semánticas entre las estructuras textuales con la finalidad de determinar el sentido en el que se está utilizando cada unidad lingüística en el discurso.

No hay que olvidar que uno de los objetivos esenciales en el PLN es que el sistema sea capaz de "comprender" el contenido de los enunciados textuales. Por esto, el procesamiento semántico es una tarea fundamental y compleja.

Dentro del procesamiento semántico se suele llevar a cabo el reconocimiento y clasificación de entidades nombradas. No obstante, este tópico se trata junto con la extracción de información y la instanciación de ontologías con las que está estrechamente relacionado (ver apartado 4.3).

Una vez que se han identificado las unidades lingüísticas y su función gramatical en el discurso, es necesario discernir qué sentido, de entre todos los posibles, es al que se está refiriendo esa secuencia textual. El sentido de un término viene dado por el contexto y por el cotexto. El cotexto sintagmático de una entidad verbal está constituido por las palabras concomitantes con dicha entidad verbal, y por lo tanto, aparece explícito en el texto. Sin embargo, el contexto es la situación socio-cultural que envuelve al acto discursivo y no está explícita en el texto (Ver apartado 5. de esta tesis).

Las técnicas para la desambiguación semántica de las palabras (*Word Sense Desambiguation*) consisten en identificar el significado de una palabra en un determinado contexto dentro de un conjunto de candidatos posibles.

Para llevar a cabo la desambiguación, se ha recurrido al uso de herramientas como corpus anotados, lexicones computacionales o anotación de roles semánticos, junto con la utilización de otras herramientas para la representación del conocimiento.

La anotación semántica de los términos dentro de un corpus se está convirtiendo en una de las principales soluciones para el procesamiento semántico, ya que en el corpus se recogen la mayor cantidad de sentidos posibles,

cuyo patrón oracional podrá ser comparado con los patrones extraídos del texto a analizar. Una vez identificadas las coincidencias, se etiquetan también los nuevos patrones. Este tipo de métodos se han denominado métodos basados en corpus (*corpus-based methods*).

Por otro lado, los métodos basados en el conocimiento (*Knowledge-based methods*) hacen uso del conocimiento previamente recogido en tesauros, ontologías, lexicones, y terminologías. Un ejemplo de este tipo de herramientas es WordNet, en donde, a través del concepto de *synset,* se pretende recoger el máximo de sentidos posibles para un término dado. Con el uso de estas herramientas, se intentan solucionar problemas como la sinonimia y la polisemia.

3.2.4.1 WordNet

Uno de los principales componentes de multitud de aplicaciones en las que se incluye un análisis semántico es WordNet (Miller, 1995; Fellbaum, 1998).

WordNet es una gran base de datos léxica en inglés desarrollada por un grupo de lingüistas y psicólogos de la Universidad de Princenton (Miller, 1995; Fellbaum, 1998). En ella, nombres, verbos, adjetivos y adverbios son agrupados dentro de conjuntos de sinónimos o *synsets*. Cada uno de los synsets, que representa un concepto distinto, está enlazado con otros synsets por medio de relaciones léxicas y semánticas. El resultado es una red interconectada de significados representados en cada synset por las correspondientes formas léxicas.

La fundamentación teórica del sistema tiene su origen en la idea de matriz de vocabulario (*vocabulary matrix*) (Miller, 1986). Con el término forma léxica (*word form*), Miller se refiere a la expresión física que se escribe o se pronuncia, mientras que significado léxico (*word meaning*) se refiere al concepto lexicalizado que se expresa por medio de una forma léxica.

Las columnas de la matriz de vocabulario contienen todas las palabras (formas léxicas) de un idioma, mientras que las filas contendrían todos los significados. Una entrada de una celda de la matriz implica que la forma léxica de una columna

puede usarse, en el contexto apropiado, para expresar el significado de esa fila (ver tabla 3.1).(Moreno Ortiz, 1997).

Tabla 3.1 Matriz de vocabulario de WordNet.

Significados Léxicos	Formas Léxicas				
	F_1	F_2	F_3	.....	F_n
M_1	$E_{1.1}$	$E_{1.2}$			
M_2					
M_3		$E_{2.2}$			
:					
:			$E_{3.3}$		
M_n					
				..	
				.....	$E_{m.n}$

En la tabla 3.1, la entrada E1.1 implica que la forma léxica F1 puede usarse para expresar el significado M1. En el caso de que haya dos entradas en la misma columna, la forma léxica es polisémica; si hay dos entradas en la misma fila, las dos formas léxicas son sinónimas.

Aparte de la relación de sinonimia, que es la que con mayor frecuencia encontramos entre los elementos del mismo synset, existen otras relaciones como la antonimia, la hiponimia y, en consecuencia, la hiperonimia, la meronimia y su inversa, la holonimia. Además cabe mencionar relaciones como la troponimia, que equivale a la relación de hiponimia (la hiponimia es una relación exclusiva de los sustantivos) pero aplicada a los verbos, y por último la implicación.

Al hilo de WordNet surgieron otras iniciativas como **EuroWordNet**[6], que se trata de un proyecto europeo en el que se está desarrollando una base de datos multilingüe para el holandés, el italiano, el español, el catalán, el alemán, el francés, el checo y el estonio siguiendo la estructura propuesta por WordNet. No obstante, la accesibilidad y visibilidad del mismo no es todavía comparable a la de WordNet.

[6] http://www.illc.uva.nl/EuroWordNet/

Dentro del procesamiento semántico, el etiquetado de roles semánticos, también conocidos como papeles temáticos o *semantic roles* en inglés, se ha convertido en un elemento fundamental para tareas como la localización de relaciones entre entidades, la generación de lenguaje, la traducción automática, la extracción de información o los sistemas de búsqueda de respuesta.

Un rol semántico describe la relación semántica que se establece entre el predicado, normalmente un verbo, y sus argumentos o constituyentes sintácticos.

Un rol identifica el papel de un argumento del verbo en el evento que dicho verbo expresa, por ejemplo, un agente, un paciente, un beneficiario, etc., o también adjuntos como causa, manera o temporal (Moreda Pozo, 2008).

Aunque su identificación está estrechamente relacionada con la sintaxis, van más allá de la misma, ya que no es la función gramatical de los actuantes de la oración la que determina el tipo de rol, sino la función semántica de los componentes con respecto al predicado. Esto permite que en dos oraciones con el mismo sentido pero con distinta diátesis se puedan identificar los mismos roles. Es decir, los roles semánticos permiten generalizar las distintas realizaciones de los argumentos de los predicados.

Por ejemplo, en la figura 3.4 la frase A y la frase B tienen el mismo significado a pesar de la variación de la diátesis del verbo que en el primer caso es activa, mientras que en el segundo es pasiva. La diferencia sintáctica de ambas no altera la asignación de roles semánticos, en este caso Causa y Efecto.

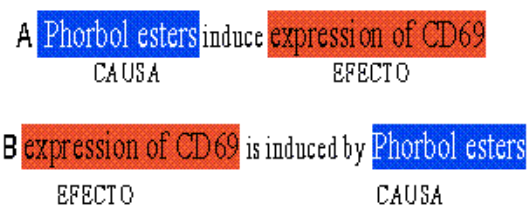


Figura 3.4 Ejemplo de anotación de roles semánticos.

Desde un punto de vista lingüístico, el etiquetado de roles semánticos consiste en localizar los constituyentes que son argumentos de un predicado dado y asignarles etiquetas semánticas que describan cuál es su relación con el predicado (Márquez, 2008).

El etiquetado de roles semánticos se ha llevado a cabo fundamentalmente en corpus de referencia que constituyen la base para el entrenamiento de sistemas que identifican los roles semánticos.

Los principales recursos dedicados al etiquetado de roles son FrameNet (Baker et al., 1998), PropBank (Kingsbury & Palmer, 2002) y VerbNet (Kipper-Schuler, 2005). Además, han surgido propuestas para dominios específicos, como el de la biomedicina, para el que se han desarrollado sistemas como PasBio (Wattarujeekrit, 2004) o BioFrameNet (Dolbey, 2006).

A continuación, se describen los proyectos mencionados previamente.

3.2.4.2 VerbNet

El proyecto VerbNet[7] (Kipper- Schuler, 2005) se encarga del desarrollo de un extenso lexicón verbal organizado jerárquicamente e independiente del dominio. El idioma para el que se ha desarrollado es el inglés y la principal diferencia que presenta con respecto a proyectos similares es que está enlazado a otros lexicones como WordNet (Fellbaum, 1998; Miller, 1995), XTag (Doran, 1994)[8] y FrameNet (Baker et al., 1998).

VerbNet se basa en la clasificación verbal de Levin (Levin, 1993) a la que se han añadido otras subclases para mejorar la coherencia semántica y sintáctica (Kipper et al., 2008).

Cada clase verbal en VerbNet consta de uno o más verbos (miembros) que comparten uno o más significados y que tienen un comportamiento sintáctico

[7] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[8] http://www.cis.upenn.edu/~xtag/tech-report/

similar. En dichas clases, se indican los roles temáticos para la estructura de argumentos del predicado de esos miembros, donde los argumentos están predefinidos según el tipo de verbo.

Las clases están descritas mediante tres elementos:

- Roles semánticos, como por ejemplo actor, agente, beneficiario, causa, etc., donde cada clase verbal tiene asociado un conjunto de roles que desempeñan los argumentos de un verbo de esa clase.

- Restricciones de los argumentos, tales como tiempo, localización, idea, animado que se utilizan para acotar el significado del verbo

- *Frames* o esquemas sintácticos, en los que se incluye una descripción sintáctica del verbo que indica las posibles realizaciones lingüísticas de los argumentos de un predicado. Por ejemplo, transitivo, intransitivo, sintagma preposicional, alternancias de diátesis, etc.

VerbNet cuenta en su versión extendida con 274 clases de primer nivel junto con 23 roles semánticos. El número total de verbos que incluye la base de datos es de 5257.

3.2.4.3 FrameNet

Otro de los proyectos dedicado a la anotación de roles semánticos en inglés es FrameNet (Baker et al., 1998). Según la definición de Fillmore (1976) un *frame* o marco semántico (en inglés, *semantic frame*) es una representación esquemática de una situación en la que se ven involucrados varios participantes. Es decir, un *frame* es una representación esquemática de una situación del mundo real en base a la cual se organiza la información.

Cada marco semántico tiene asociadas unas unidades léxicas determinadas, unos roles y varios ejemplos anotados manualmente que representan ese *frame* y

que han sido extraídos de un corpus. FrameNet cuenta con unos 17.0000 ejemplos anotados en total.

Las unidades léxicas que conforman un marco semántico son aquellas que pueden aparecer como predicados del *frame*, y que, por tanto evocan dicho marco. En el caso del *frame* definido como *shopping,* las unidades léxicas que lo invocan son el verbo *shop* y el sustantivo *shopping*. Por otro lado, las unidades léxicas que se consideran predicativas no son sólo verbos, sino que los elementos del frame pueden ser argumentos de cualquier predicado, como adjetivos y nombres.

Además, cabe distinguir entre *core roles* o roles centrales, que son aquellos que necesariamente van unidos a un determinado *frame* y los *non-core roles* o roles no centrales que son aquellos que pueden o no manifestarse en un determinado *frame* y que delimitan el significado del *frame*. Por ejemplo, el *frame Ingestion* tiene asociados dos roles centrales que son *Ingestor* e *Ingestibles* y nueve roles no básicos como por ejemplo *Duration*, *Instrument*, *Place*, etc.

Dado que los roles se definen para un *frame* concreto, éstos son más específicos que aquellos roles asignados a un verbo en general (como por ejemplo agente o paciente). La ventaja su adaptación a cada situación particular, mientras que el problema es que para tener una cobertura amplia es necesaria la definición de una gran cantidad de *frames*.

3.2.4.4 PropBank

En el proyecto *Proposition Bank* (PropBank) (Kingsbury & Palmer, 2002), las anotaciones semánticas se llevan a cabo en las estructuras sintácticas del corpus Penn Treebank[9]. Cada verbo tiene asociados un conjunto de *frames* o marcos sintácticos, denominado frameset o roleset, en donde se especifican los roles

[9] Corpus de un millón de palabras, extraido del Wall Street Journal y anotado sintáctica semánticamente por el LINC laboratory en la Universidad de Pensilvania. http://www.cis.upenn.edu/~treebank/

admisibles por cada verbo. Los argumentos asignados a cada verbo son etiquetados con una etiqueta numérica que va dese Arg0 hasta Arg5.

Los verbos que comparten significado comparten también el frameset en el que se insertan. Por ejemplo, *buy* y *purchase* pertenecen al mismo conjunto de roles.

Además, PropBank cuenta con un conjunto general de roles etiquetados como ArgM y una etiqueta que especifica su función, por ejemplo PNC propósito o LOC lugar.

En los verbos polisémicos, sentidos diferentes tienen asociados framesets distintos. Por ejemplo, para el verbo *run* se definen tres esquemas distintos en función de su semántica.

Aunque el proyecto principal se ha desarrollado en inglés, se está desarrollando también una versión para el árabe.

El número de verbos que contiene hasta el momento es de unos 5.500 que se distribuyen en un total de 7.268 frameset y que tienen asignados unos 6.000 predicados en total.

A pesar de que los proyectos descritos comparten la finalidad común de anotar el conocimiento semántico generalmente en torno a los verbos, presentan algunas diferencias como se aprecia en la tabla 3.2. En dicha tabla, se puede ver cómo los distintos sistemas de etiquetado de estructuras sintáctico-semánticas han gestionado el etiquetado de las unidades léxicas referidas a la acción de comprar (*purchase*).

Tabla 3.2 Comparativa entre distintos sistemas de etiquetado de roles semánticos.

VerbNet	**FrameNet**	**PropBank**
Verbo: OBTAIN **Miembros:** Aquire Obtain Purchase **Roles:** Asset [+CURRENCY] **Frames:** NP.asset V NP **Ejemplo** "$50 won't even purchase a dress." **Sintaxis** NP V NP PP.asset **Ejemplo** "Carmen purchased a dress for $50." **Sintaxis** Agent V Theme {for} Asset **Semántica** has_possession(start(E), ?Source, Theme) transfer(during(E), Theme) has_possession(end(E), Agent, Theme) cause(Agent, E) cost(E, Asset) **Sintaxis** Asset V Theme **Semántica** has_possession(start(E), ?Source, Theme) tralllInsfer(during(E), Theme) has_possession(end(E), ?Agent, Theme) cause(?Agent, E) cost(E, Asset)	**Frame:** COMMERCE_BUY **Unidades léxicas:** Buy Purchase_(act)*s.* Purchase *v.* **Core Roles:** Buyer Goods **Non-Core Roles:** Manner Means Money Period_of_iterat Place Purpose Propose_of_goods Rate Reason Recipient Seller Time Unit **Ejemplo:** The Buyer wants the Goods and offers Money to a Seller in exchange for them. Jess bought a coat. Lee bought a textbook from Abby.	**Verbo:** PURCHASE **Miembros:** Buy Purchase **Argumentos:** Arg0: purchaser Arg1: thing purchased Arg2: seller Arg3: price paid Arg4: benefactive **Ejemplo:** **Arg0**: They **Rel:** purchased **Arg1**: $2.4 billion in Fannie Mae bonds

Como se puede ver en la tabla anterior, en VerbNet el verbo *Obtain* da nombre a la clase en la que se incluye *Purchase*. Los tres miembros del grupo comparten características sintáctico-semánticas así como un conjunto de roles preestablecidos que se le asigna a cada grupo, en este caso Asset.

En FrameNet se parte del *frame*, de la situación comunicativa, a la que se asocian determinadas unidades léxicas, en este caso *Purchase* como sustantivo y

como verbo. Para el *frame* se han descrito una serie de roles específicos, entre los que se encuentran los roles esenciales o *core roles* que son *Buyer* y *Goods*, y los roles no-esenciales que pueden o no aparecer entre los argumentos del predicado.

Finalmente, PropBank comparte la estructura de VerbNet en cuanto a la existencia de una clase verbal con distintos miembros, aunque en este caso los miembros difieren a los contenidos en la clase de VerbNet. Por otro lado, los roles descritos para cada grupo verbal son específicos, como sucede en FrameNet. Además, en PropBank existe un número máximo de argumentos para un verbo.

Los proyectos de propósito general descritos anteriormente, han servido de base para la creación de otros de carácter específico, es decir, asociados a un dominio. En concreto, para el dominio de la bioinformática se han desarrollado algunas propuestas basadas en las estructuras de etiquetado de predicado y argumentos.

3.2.4.5 Propuesta de Korhonen et al.

En esta propuesta, se extraen los verbos más frecuentes de un corpus de artículos biomédicos y mediante diversos métodos de distribución de frecuencias y estadísticos, se clasifican los verbos extraídos en alguna de las clases léxicas de VerbNet, obteniendo, de este modo, sus roles.

Para llevar a cabo la tarea han creado un "gold standard", recogido en la Tabla 3.3, en donde, además, se muestran algunos ejemplos para cada verbo.

Tabla 3.3 Clasificación verbal propuesta por Korhonen et al. (2006).

1 Have an effect on activity (BIO/29)	8 Physical Relation Between Molecules (BIO/20)
1.1 Activate / Inactivate 1.1.1 Change activity: *activate, inhibit* 1.1.2 Suppress: *suppress, repress* 1.1.3 Stimulate: *stimulate* 1.1.4 Inactivate: *delay, diminish* **1.2 Affect** 1.2.1 Modulate: *stabilize, modulate* 1.2.2 Regulate: *control, support* **1.3 Increase / decrease:** *increase, decrease* **1.4 Modify:** *modify, catalyze*	**8.1 Binding:** *bind, attach* **8.2 Translocate and Segregate** 8.2.1 Translocate: *shift, switch* 8.2.2 Segregate: *segregate, export* **8.3 Transmit** 8.3.1 Transport: *deliver, transmit* 8.3.2 Link: *connect, map*
	9 Report (GEN/30)
2 Biochemical events (BIO/12)	**9.1 Investigate** 9.1.1 Examine: *evaluate, analyze* 9.1.2 Establish: *test, investigate* 9.1.3 Confirm: *verify, determine* **9.2 Suggest** 9.2.1 Presentational: *hypothesize, conclude* 9.2.2 Cognitive: **9.3 Indicate:** *demonstrate, imply*
2.1 Express: *express, overexpress* **2.2 Modification** *dephosphorylate, phosphorylate* 2.2.1 Biochemical modification: *dephosphorylate, phosphorylate* 2.2.2 Cleave: *cleave* **2.3 Interact:** *react, interfere*	10 Perform (GEN/10)
	10.1 Quantify 10.1.1 Quantitate: *quantify, measure* 10.1.2 Calculate: *calculate, record* 10.1.3 Conduct: *perform, conduct* **10.2 Score:** *score, count*
3 Removal (BIO/6) *consider, believe*	
3.1 Omit: *displace, deplete* **3.2 Subtract:** *draw, dissect*	11 Release (BIO/4): *detach, dissociate*
4 Experimental Procedures (BIO/30)	12 Use (GEN/4): *utilize, employ*
4.1 Prepare 4.1.1 Wash: *wash, rinse* 4.1.2 Mix: *mix* 4.1.3 Label: *stain, immunoblot* 4.1.4 Incubate: *preincubate, incubate* 4.1.5 Elute: *elute* **4.2 Precipitate:** *coprecipitate coimmunoprecipitate* **4.3 Solubilize:** *solubilize,lyse* **4.4 Dissolve:** *homogenize, dissolve* **4.5 Place:** *load, mount*	13 Include (GEN/11) **13.1 Encompass:** *encompass, span* **13.2 Include:** *contain, carry*
5 Process (BIO/5): *linearize, overlap*	14 Call (GEN/3): *name, designate*
6 Transfect (BIO/4): *inject, microinject progress, proceed*	15 Move (GEN/12)
7 Collect (BIO/6)	**15.1 Proceed:** **15.2 Emerge:**
7.1 Collect: *harvest, select arise, emerge* **7.2 Process:** *centrifuge, recover*	16 Appear (GEN/6): *appear, occur*

Su propuesta se fundamenta en la creación de 16 clases verbales en las que un verbo actúa como concepto general que representa una clase. En cada clase verbal están incluidos aquellos verbos que responden específicamente al sentido que expresa dicha clase. Algunas clases están organizadas jerárquicamente, pero entre el conjunto de las 16 clases no se establece una relación. Por otro lado, aunque algunos verbos son propios de la terminología del domino como *dephosphorylate* o *coprecipitate*, otros como *evaluate* o *employ* son de carácter general, es decir, sólo en determinados contextos, esto es, en contextos biomédico, se podrán clasificar como pertenecientes a un dominio.

3.2.4.6 PasBio

PasBio (Wattarujeekri et al., 2004) toma como punto departida los *framesets* descritos en PropBank para buscar en artículos científicos oraciones que contienen verbos relevantes en el dominio de la biología molecular.

El proyecto estuvo activo desde 2002 hasta 2005, y en la actualidad no es posible acceder a las estructuras de lo argumentos[10].

En PasBio se han analizado 30 verbos en distintas oraciones y con distintos sentidos, siempre dentro de la biología molecular. Dichos verbos se han obtenido de PubMed.

En el trabajo mencionado, se introduce la noción de análisis sistemático de los argumentos en textos biológicos y se propone la construcción de estructuras de Argumento-Predicado (PAS). Las PAS son estructuras de conocimiento que representan las relaciones entre un verbo y sus argumentos. En PasBio los predicados describen los roles de los genes y lo que de ellos resulta mediante sus funciones biológicas.

[10] Consulta realizada en Junio de 2011 en http://sites.google.com/site/nhcollier/projects/pasbio

Los 30 nuevos *frames* propuestos siguen el formato de roleset descrito en PropBank, aunque están ejemplificados con una frase del corpus seleccionado como ocurre en FrameNet, es decir, para cada *frame* existe un ejemplo seleccionado manualmente.

Para llevar a cabo la selección de verbos relevantes del dominio, en primer lugar se han identificado los sentidos de cada verbo utilizando WordNet. Los argumentos de cada verbo se han dividido en *core arguments*, aquellos que son fundamentales para completar el significado del evento, y en *adjuntos* aquellos que completan y modifican el significado de los argumentos y que pueden ser de tres tipos, adverbiales, de negación y modales.

En la siguiente tabla (tabla 3.4) se pueden ver algunos de las estructuras verbales creadas con PasBio.

Tabla 3.4 Etiquetado de roles semánticos en PasBio

Transform Arg2 transformation_of Arg1 Arg0:agent/causer of transformation Arg1:entity undergoing transformation Arg2 :effect of transformation/ end state
Modify Arg3 transformation_of Arg1 Arg0:agent, cause of transformation Arg1:thing changing Arg2:method Arg3:consequence // secondary predication
Develop Arg3 transformation_of Arg2 Arg1:non-intentional theme Arg2:thing developed Arg3:end result
Alter Arg1 Arg2 transformation_of Arg3 Arg0:causer of transformation // mutation, protein Arg1:thing changing // codon, exon, premRNA Arg2:end state Arg3:start state Arg4:location refering to tissue, organelle, person, gene

Generate Arg0:agent, causer //gene, process// Arg1:thing created

Los verbos incluidos en la tabla indican relaciones que no son exclusivas de los textos biomédicos. Sin embargo, es la asignación de los argumentos propios del dominio lo que los diferencia de otras propuestas más generales. Por ejemplo, *Generate* puede tener un argumento que sea un gen (gene).

No obstante, en Cohen & Hunter (2006) se presentan una serie de cuestiones en torno a la validez de PasBio para la representación de los eventos en biología molecular. Como los autores indican, las representaciones de los predicados están normalmente asociadas con un verbo. Sin embargo, en los textos médicos las formas nominalizadas de los verbos juegan un importante papel (por ejemplo, *inhibits* frente a *inhibition*), siendo ésta y el escaso número de predicados descritos una importante limitación a la hora de su aplicación real.

3.2.4.7 BioFrameNet

BioFrameNet (Dolbey, 2006) es el resultado de una tesis doctoral y en ella se propone una metodología para la extensión de FrameNet aplicada al dominio de la biología molecular.

La estructura, en consecuencia, es la misma que en FrameNet, estando algunos de los frames de ambos relacionados entre sí. No obstante, dada la complejidad del lenguaje biomédico, se ponen de relieve nuevas características semánticas y sintácticas que no se ajustan a los verbos de dominio general descritos en FrameNet.

Los datos en los que se ha basado el estudio son una colección de textos sobre los distintos tipos de eventos de trasporte entre proteínas intracelulares.

Además, los frames han sido implementados en el lenguaje ontológico OWL DL para facilitar el mapeo entre ontologías de dominio biomédico como GO[11] o Entrez Gene[12].

3.2.4.8 BioProp

BioProp (Chou et al., 2006) es un repositorio de estructuras de predicado-argumento (*Predicate Argument Structures* PAS) del dominio biomédico, anotadas mediante el etiquetador automático de roles semánticos BIOSMILE (Tsai et al., 2007).

BIOSMILE utiliza técnicas de *machine learning* o aprendizaje automático, en concreto un modelo de máxima entropía, que permite extraer relaciones biomédicas.

El sistema ha sido entrenado en 500 resúmenes etiquetados sintácticamente pertenecientes al corpus GENIA.

BioProp contiene 30 verbos que han considerado relevantes para el dominio de la biología molecular, y, aunque para la estructura de anotaciones utilizan el modelo de PropBank, como PasBio, las anotaciones se basan en las clases definidas en VerbNet, en donde los argumentos de cada verbo son representados a un nivel semántico y tienen asociados roles semánticos. Este sistema incorpora formas lematizadas junto con etiquetas gramaticales y entidades nombradas.

Dadas las similitudes entre BioPop y PasBio, se ha realizado un mapeo entre ambos (Tsai et al., 2008).

---

[11] http://www.geneontology.org/

[12] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

3.2.5 Procesamiento Pragmático

La pragmática pone en relación un determinado enunciado con el contexto comunicativo.

El significado atiende a las reglas del sistema lingüístico, que asignan un valor a un conjunto de signos fónicos o gráficos organizados en unas estructuras determinadas, ya sean palabras, frases u oraciones, sin tener en cuenta el contexto concreto en el que se producen, es decir, sin considerar ningún factor extralingüístico. Sin embargo, el sentido – o significado pragmático-discursivo – resulta de la interdependencia de los factores contextuales y las formas lingüísticas (Casalmiglia Blancafort, 1999).

Es decir, durante procesamiento pragmático se analiza el texto en su conjunto, siendo una de las principales dificultades de este análisis la resolución de la anáfora, la catáfora y la deixis.

La anáfora es el término empleado para referirse a una forma lingüística (con frecuencia un pronombre) que remite al significado de otra forma que ya ha sido expresada en el texto. Por el contrario la catáfora es utilizada para referirse a una forma lingüística que aún no ha aparecido en el texto. Por ejemplo: *Me dijo que no lo había visto*. **Donde** *lo* **es una anáfora o** *Me lo dijo claramente: no vuelvas* **donde** *lo* **es una catáfora.**

Si el referente al que hace alusión la anáfora o la catáfora, está contenido en el discurso, es decir, si es conocimiento explícito, podría identificarse. Sin embargo, el conocimiento implícito, aquel que no aparece ex profeso en el texto es mucho más difícil de sistematizar para ser identificado por el sistema.

La deixis es un fenómeno lingüístico por el cual todo acto de habla queda anclado en unas coordenadas espacio-temporales, es el llamado yo-aquí-ahora. Existen en las lenguas numerosos términos que sirven para ubicar tanto espacial como temporalmente la enunciación. Por ejemplo, los adverbios allí, ahora, antes, etc.

Otro elemento para el que el contexto es imprescindible es la implicación. Es decir, las deducciones que se pueden realizar partiendo de un enunciado en base a una situación determinada. Por ejemplo: A: *¿Habrá llegado ya?* **B:** *La luz de su*

habitación está encendida. El interlocutor B puede sugerir que, puesto que la luz está encendida, lo más probable es que haya llegado.

Finalmente, entre los problemas que afectan al procesamiento pragmático y al semántico cabría mencionar el uso de algunas figuras retóricas entre las que destaca la metáfora, que no sólo se utiliza en el discurso literario, sino que se suele utilizar con cierta frecuencia en todos los discursos, incluso en el científico. Si se trata de metáforas lexicalizadas, pueden estar incluidas en el lexicón, pero existen infinitas combinaciones que son imposibles de recopilar. Por ejemplo, *El Amazonas es el pulmón del mundo* o *The Bottleneck of ontology building*.

El desarrollo de sistemas que permitan analizar la pragmática textual automáticamente es una tarea difícil, ya que en el texto se pone de manifiesto la intención comunicativa del hablante, que no sólo estará determinada por las estructuras oracionales, sino que vendrá marcada también por el contexto socio-cultural en el que se integra así como la finalidad que persigue la producción textual, más allá de la mera comunicación.

En la lingüística tradicional, el análisis del discurso es la aproximación más metódica a las características supraoracionales de un texto. El análisis del discurso nos permite indagar sobre porqué se han elegido unas construcciones sintagmáticas y no otras, cuál es el léxico predominante y porqué, o si el texto se ajusta a los patrones preestablecidos para el género discursivo al que pertenece

En relación con este enfoque teórico del análisis del discurso aplicado al PLN, encontramos un enfoque más técnico, la DRT (*Discourse Representation Theory*). Se trata de un formalismo lógico que plantea una representación e interpretación más allá del nivel de una oración, es decir, a nivel del discurso, para poder tratar fenómenos que necesitan un contexto como la anáfora y el tiempo (Verdejo, 2007).

La DRT está relacionada con la *Discourse Representation Structure* (DRS) que consta en dos partes: un universo de referentes del discurso, que representa los objetos mencionados, y un conjunto de condiciones que representa la información acumulada sobre esos objetos hasta el momento.

La ambigüedad que genera la pragmática afecta, sobre todo, a situaciones comunicativas reales o a dominios generales. Sin embargo, en dominios específicos como el de la biomedicina, por ejemplo, la ambigüedad generada por el contexto discursivo se reduce notablemente, ya que el marco en el que se insertan las producciones científicas biomédicas está bien definido. En consecuencia, la resolución pragmática en este tipo de textos afecta a fenómenos como la anáfora y catáfora y a procesos de razonamiento como la implicación textual[13] o *textual entailment*.

El procesamiento pragmático queda fuera del ámbito de esta tesis y los trabajos realizados se quedan en el nivel semántico del PLN.

3.3 Herramientas para el PLN

En el apartado anterior se han descrito cuáles son los niveles para el procesamiento automático del lenguaje natural y cuáles son los principales problemas en cada una de las fases. Para cada uno de los niveles, se han desarrollado numerosas herramientas, generalmente integradas en otros sistemas y fundamentalmente para el inglés.

Aquí nos centraremos en la descripción de uno de los principales y más completos entornos para el PLN, GATE. Y, por otra parte, se describirá Freeling, que aunque está limitado al análisis morfo-sintáctico y al reconocimiento de entidades nombradas, es una de las principales herramientas que existen para el PLN en español.

[13] La implicación textual o *textual entailment* es un relación direccional entre fragmentos de texto. La relación se refiere a si la verdad de un enunciado implica la verdad de otro enunciado denominado también hipótesis. http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Portal

3.3.1 GATE

Gate[14](A General Architecture for Text Engineering) (Cunningham, 2002) es una infraestructura para el desarrollo e implementación de componentes software capaces de procesar lenguaje humano. GATE ayuda a investigadores y desarrolladores de tres modos diferentes:

1- Especificando una arquitectura o estructura organizativa para el software de procesamiento de lenguaje.
2- Mediante la provisión de un marco de trabajo o librería de clases que implementa la arquitectura y que puede ser utilizada para agregar las capacidades de procesamiento lingüístico de diversas aplicaciones
3- Proveyendo un entorno de desarrollo construido sobre el marco de trabajo con herramientas gráficas útiles para el desarrollo de componentes.

La arquitectura explota el desarrollo de software basado en componentes, la orientación a objetos y el código portable. El marco de trabajo y el entorno de desarrollo están desarrollados en Java y disponibles como software libre bajo licencia de la librería GNU. En la siguiente figura (figura 3.5), se muestra un ejemplo de esta interfaz.

[14] http://gate.ac.uk/

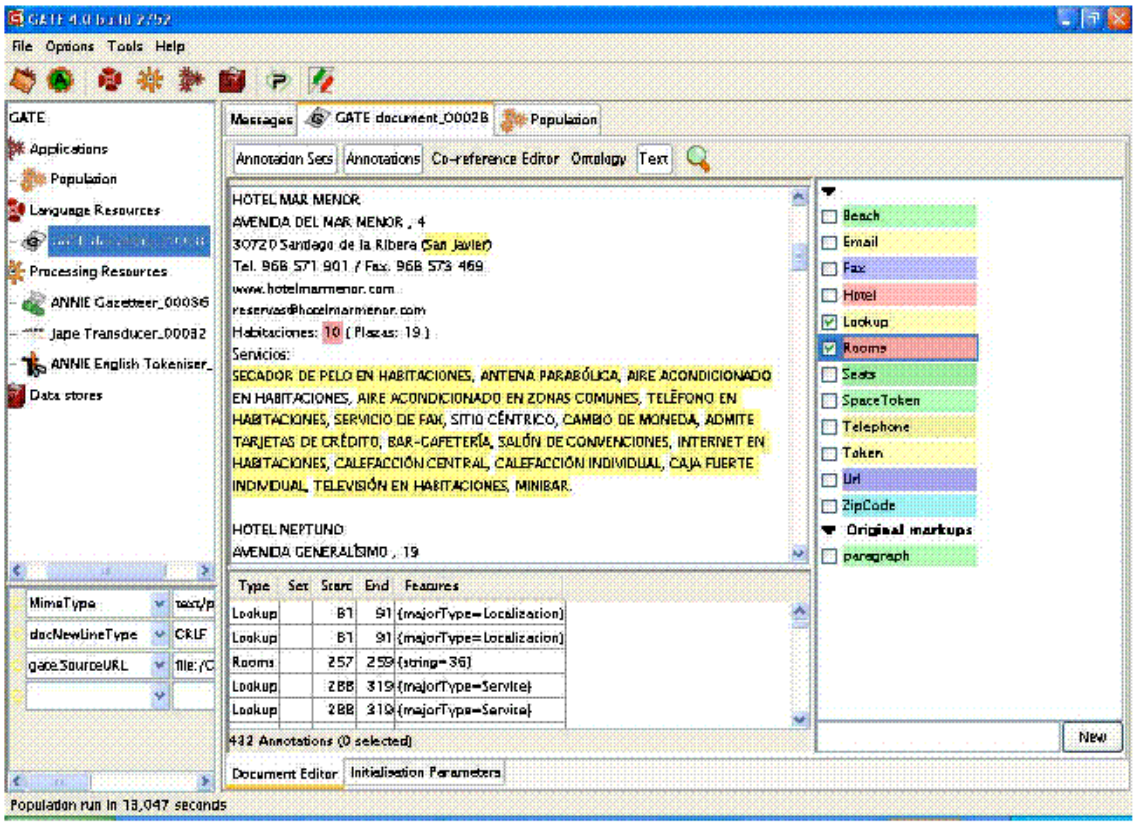


Figura 3.5 Elementos identificados con GATE.

GATE, como arquitectura, sugiere que los elementos de los sistemas de software que procesan lenguaje natural se pueden dividir en varios tipos denominados recursos. Estos recursos son partes reusables con interfaces bien definidas y son una forma de arquitectura popular utilizada por ejemplo en Java Beans y Microsoft .Net. Los componentes de GATE son tipos específicos de Java Bean y se puede distinguir entre:

- *Recursos del Lenguaje (LR):* representan entidades como léxicos, recopiladores u ontologías.

- *Recursos de Procesamiento (PR):* representan entidades principalmente algorítmicas, como parsers, etiquetadores, lematizadores, traductores, reconocedores de voz, etc. En general, los PR incluyen a los LR, por ejemplo, un etiquetador usualmente incluye un léxico.

- *Recursos Visuales:* representan componentes de visualización y edición que participan en la GUI.

El conjunto de recursos integrados en GATE es conocido como CREOLE (*Collection of Reusable Objects for Language Engineering*). Todos los recursos que se utilizan en GATE se representan en paquetes en Java con archivos JAR, además de algunos datos de configuración en archivos XML. Cuando se ha desarrollado un conjunto apropiado de recursos, estos pueden ser embebidos en la aplicación cliente objetivo utilizando el marco de trabajo de GATE.

Los documentos, recopilaciones y anotaciones en GATE se almacenan en bases de datos de diferentes clases, visualizados en el entorno de desarrollo, y accedidos a nivel de código vía el marco de trabajo. Los recursos pueden ser cargados en GATE, y guardados en conjunto en un Data Store. Además, los recursos de procesamiento se pueden "unir" por medio de una tubería (pipeline), igual que en las tuberías de Unix donde la salida de una aplicación es la entrada a la siguiente.

GATE soporta el procesamiento de documentos en una gran variedad de formatos, incluyendo XML, RTF, email, HTML, SGML y texto plano. En todos los casos, el formato se analiza y se convierte en un modelo unificado de anotaciones. Una anotación es un metadato unido a una sección particular del contenido del documento. La conexión entre una anotación y el contenido al que se refiere se hace por medio de dos indicadores que representen las localizaciones de inicio y del final del contenido del texto cubierto por el análisis. Una anotación debe también tener un tipo (o un nombre) que se utiliza para crear clases de anotaciones similares, ligado generalmente junto con su semántica.

Un documento en GATE puede tener una o más capas de anotaciones, una anónima (también llamada *default*), y tantas *conocidas* como se requiera. Una capa de anotación se organiza como un gráfico dirigido acíclico (Figura 3.6), en el cual los nodos tienen una localización determinada en el documento y los arcos representan las anotaciones que alcanzan la localización indicada por el nodo del

inicio al de fin. Gracias a esta estructura, un conjunto de anotaciones (*Annotations Set*) mantiene un número de anotaciones y una serie de índices que proporcionan un acceso rápido a las anotaciones contenidas.

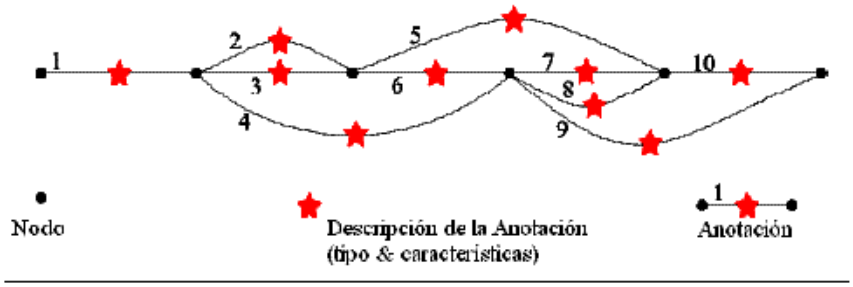


Figura 3.6 Modelo de Grafo de Anotaciones.

Una de las principales herramientas integradas en GATE es ANNIE (A Nearly-New Information Extraction System). Originalmente desarrollada en el contexto de extracción de información, se trata de un sistema altamente modular cuyos componentes se organizan en una arquitectura de estilo tubería. Algunos de sus componentes más importantes son:

- *Tokenizer:* Separa el texto en tokens simples como pueden ser números, símbolos de puntuación, y palabras.
- *Sentence Splitter:* Es una cascada de transductores de estado finito que separan el texto en oraciones.
- *Part of Speech Tagger:* Identifica y anota las categorías gramaticales de un texto. Este tagger utiliza TreeTagger[15] para distintos idiomas.
- *Gazetteer:* Utiliza una serie de listas de palabras. Cada lista representa un conjunto de nombres y se utiliza un archivo índice para acceder a estas listas.
- *Semantic Tagger:* Basado en el lenguaje JAPE, contiene reglas que actúan en anotaciones asignadas a frases previamente, y produce como salida entidades de anotaciones.

[15] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

- *Orthographic Coreference (NameMatcher):* agrega identidad a los nombres de entidades encontrados por el etiquetador semántico.

A continuación se describen detalladamente los gazetteers y las reglas JAPE que conforman el núcleo de GATE.

3.3.1.1 Gazetteers Lists

En el ámbito de GATE, los *gazetteers* consisten en un conjunto de listas que contienen nombres de entidades tales como días de la semana, nombres de persona, ciudades, etc. que no tienen porqué estar ordenados alfabéticamente. Su principal uso es ayudar en tareas de reconocimiento de entidades nombradas, aunque pueden ser usadas para otros propósitos.

Un *gazetteer* es una lista de términos (unipalabra o multipalabra), un inventario, que hace referencia a las entidades de un dominio concreto (ver apartado 4.3.1). Estas listas pueden tener carácter abierto, como los nombres de personas, que pueden ir aumentando al aparecer nuevos nombres, o pueden ser un inventario cerrado, como los meses del año.

Los *gazetteers* se pueden combinar con reglas JAPE, permitiendo la elaboración de expresiones en las que existan elementos fijos, ya sea un elemento de un gazetteer o la lista completa.

En GATE, cuando el *gazetteer* se ejecuta en un documento, se crean anotaciones del tipo *Lookup* para cada secuencia coincidente en el texto. El *gazetteer* es independiente de cualquier Token u otra anotación. Es decir, una entrada podría involucrar a más de un token (términos multipalabra). Una anotación *Lookup* se crea solamente cuando exista coincidencia con la entrada completa, las entradas parciales no serán anotadas.

Estas listas se almacenan en ficheros de texto plano en las cuales cada línea corresponde a un término. Es importante, sobre todo cuando lo que se quiere

procesar es lenguaje español, con tildes y eñes, estar seguro de que en el editor que se esté utilizando para crear estas listas se guarde el fichero en UTF-8.

Para facilitar este proceso, el framework de GATE proporciona un editor Unicode. Así mismo, el módulo ANNIE Gazetteers facilita la creación y modificación tanto del fichero de índice como de los ficheros de gazetteers.

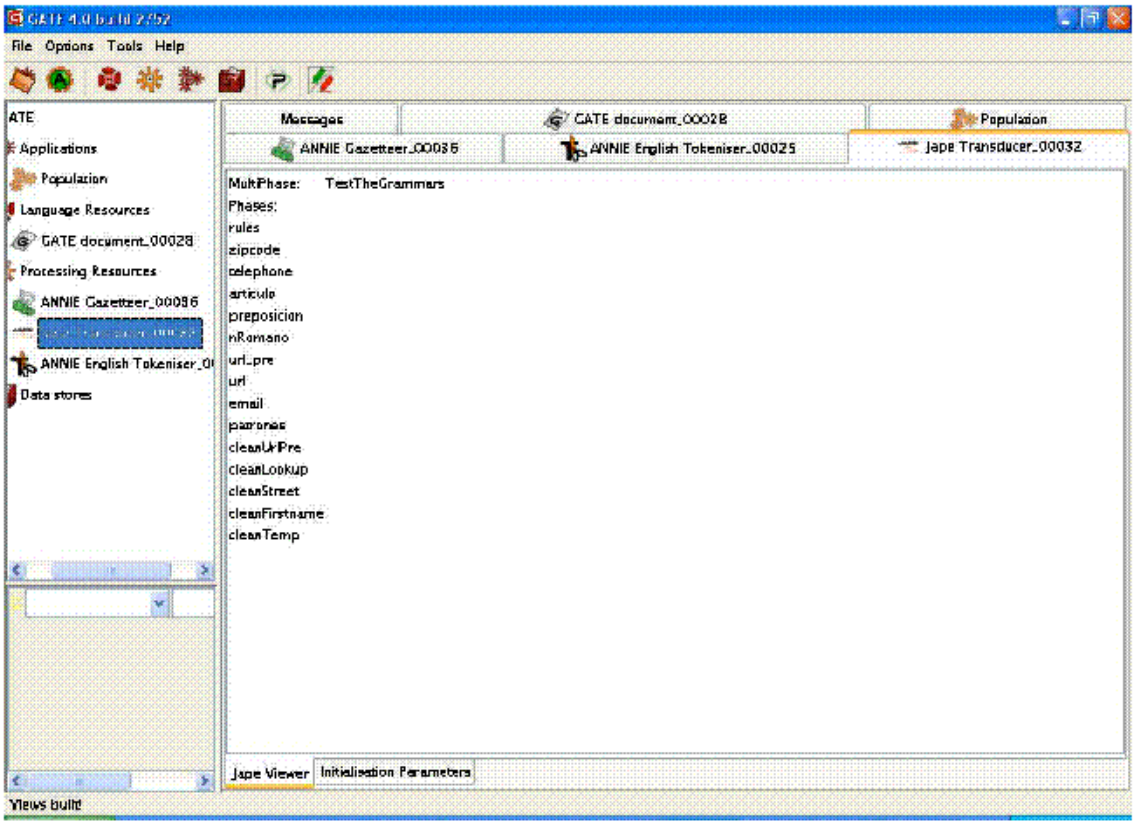


Figura 3.7 Listado de reglas Jape aplicadas en GATE.

Para acceder a estas listas, debe crearse un fichero que haga la tarea de índice, normalmente suele llamarse lists.def. Cada gazetteer debe estar localizado en el mismo directorio que su índice. Además, este fichero de índice contiene una especificación de las listas a las que indexa. Para cada lista se debe especificar un "major type" (tipo de mayor importancia o principal) y un "minor type" (tipo de menor importancia), éste último opcional. Estas listas se compilan después en máquinas de estado finitas que permiten que cualquier secuencia de texto

coincidente sea anotada con las características que especifican los tipos principales y de menor importancia.

3.3.1.2 Reglas JAPE

JAPE (*Java Annotation Patterns Engine*) permite reconocer expresiones regulares de anotaciones en documentos. El modelo de anotaciones de GATE está basado en grafos. El resultado es que, en ciertos casos, el proceso que ajusta es no determinístico (esto es, los resultados son dependientes de factores aleatorios como las direcciones en que los datos se guardan en la maquina virtual). Cuando hay estructuras en el grafo que necesitan algo más que una automatización regular para ser reconocidas, JAPE elige una alternativa arbitrariamente. En la práctica, en muchos casos, los datos guardados en los grafos de anotaciones de GATE son secuencias simples, y pueden ajustarse determinísticamente mediante expresiones regulares.

Una gramática de JAPE consiste en un conjunto de sentencias, cada una de las cuales consiste en un conjunto de reglas patrón / acción. La parte izquierda (LHS) de las reglas es un patrón de anotación que puede contener operadores de expresiones regulares (*, ?, +). La parte derecha (RHS) la conforman sentencias de manipulación de anotaciones. Las anotaciones que se ajustan a la parte izquierda de una regla pueden ser referenciadas en la parte derecha por medio de etiquetas que se adjuntan a los patrones.

Las reglas de gramática pueden ser esencialmente de dos tipos. El primer tipo de reglas no involucra búsqueda en diccionarios, pero éstas pueden ser definidas utilizando un conjunto reducido de formatos posibles. En general, son directas y ofrecen poco potencial para la ambigüedad.

El segundo tipo de reglas tiene un uso más pesado de diccionarios, y cubre un mayor rango de posibilidades. Esto no sólo significa que se van a necesitar muchas reglas para describir todas las situaciones, sino que la ambigüedad va a ser frecuente. Esto lleva a la necesidad de ordenar y priorizar las reglas.

Por ejemplo, una única regla es suficiente para identificar direcciones IP como se puede ver en la siguiente tabla (tabla 3.5):

Tabla 3.5 Ejemplo de regla JAPE para identificar una dirección.

```
Rule: IPAddress (
{Token.kind == number}
{Token.string == "."}
{Token.kind == number}
{Token.string == "."}
{Token.kind == number}
{Token.string == "."}
{Token.kind == number} ) :ipAddress -->
:ipAddress.Address = {kind = "ipAddress"}
```

En cambio, para identificar una fecha, puede haber muchas variaciones posibles, por lo que se necesitan varias reglas. Por ejemplo, la palabra "Domingo" puede ser el nombre de una persona o el día de la semana, con lo cual será necesario mantener información sobre el contexto para eliminar la ambigüedad.

El contexto se utiliza para definir bajo qué situaciones se aplica un patrón. Por ejemplo, en el reconocimiento de una fecha, puede definirse un contexto en el que la fecha se reconozca sólo si el año está precedido por las palabras "en" o "de". La parte derecha de una regla en JAPE puede contener código en lenguaje Java. Esto es útil para borrar anotaciones temporales y para manipular características de anotaciones previas.

Las gramáticas JAPE se escriben en archivos con extensión ".jape" que son parseados y compilados en tiempo de ejecución para luego ser ejecutados en el documento de GATE.

GATE provee una herramienta denominada "Jape Debugger" que permite localizar errores en programas JAPE de manera que el usuario puede ver en detalle cómo trabaja una regla en particular cuando se aplica a un rango del texto. La herramienta permite ver qué reglas fueron aplicadas y cuáles no, y además, por qué algunas reglas fueron o no aplicadas. Así mismo, es posible colocar puntos de

corte para reglas particulares permitiendo al usuario ver cómo se aplicó una regla y qué anotaciones se crearon.

3.3.2 Freeling

Freeling (http://www.lsi.upc.edu/~nlp/freeling/) es una suit de herramientas para el PLN de código abierto desarrollada por el grupo de investigación TALP de la Universidad Politécnica de Cataluña. Las tareas de PLN que permiten llevar a cabo sus componentes son:

- Tokenización y división de oraciones.
- Análisis morfológico junto con identificación de sufijos, reconocimiento de términos multipalabra y predicción probabilística de categorías de palabras desconocidas.
- Detección y clasificación de Entidades Nombradas, entre las que se incluyen el reconocimiento de números, monedas o magnitudes físicas. Además es capaz de resolver la correferencia nominal o menciones nominales
- PoS Tagging
- Análisis sintáctico de dependencias y de constituyentes
- Anotación de sentidos mediante WordNet.

Los idiomas para los que está disponible actualmente son: castellano, catalán, gallego, italiano, inglés, galés, portugués, asturiano y recientemente se ha incluido un PoS Tagger para el ruso. Para integrar la herramienta Freeling en GATE, se ha hecho uso de un plugin desarrollado por nuestro grupo de investigación. En la figura 3.8 se puede ver un ejemplo de pos-tagging realizado directamente en la página web en la versión de demostración.

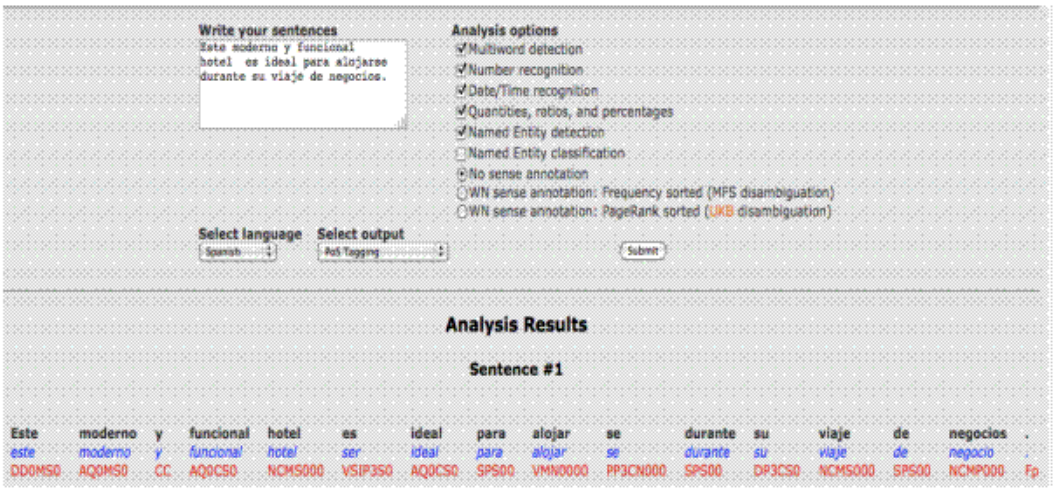


Figura 3.8 Ejemplo de PoS Tagger con Freeling.

3.4 Las ontologías y el PLN

El uso y desarrollo de herramientas y técnicas de PLN no suele ser un fin en sí mismo, sino que estas técnicas se integran en sistemas que requieran para su funcionamiento conocimiento obtenido a partir de los textos, como, por ejemplo, la traducción automática, el resumen automático o la creación e instanciación de ontologías.

Dado que la mayoría de los datos contenidos en la web se encuentran en lenguaje natural, la técnicas de PLN se convierten en la opción lógica para obtener conocimiento de los textos.

Dentro de los sistemas para el aprendizaje automático de ontologías, el objetivo principal del PLN es el de generar un conjunto de datos pre-procesados usables por los componentes del sistema (Maedche & Staab 2001). Las técnicas empleadas varían en función de si se trata de documentos semi-estructurados como por ejemplo, los diccionarios o los documentos HTML, o si por el contrario se trata de textos no estructurados.

La relación que se establece entre las ontologías y el PLN es bidireccional:

- Las ontologías pueden ser un elemento dentro de un sistema de PLN que permita llevar a cabo algunas de las tareas como por ejemplo la identificación de entidades nombradas, la extracción de relaciones o la desambiguación. Es decir, en este caso la ontología no sería el fin, sino el medio, un componente dentro del sistema de PLN que facilita la estructuración conceptual de un campo de conocimiento permitiendo llevar a cabo tareas de razonamiento.

- Por otro lado, las técnicas de PLN pueden ser utilizadas para la construcción e instanciación de ontologías. En este caso, las ontologías serían un fin en sí mismas y PLN se convertiría en un vehículo para la obtención de conocimiento de forma automática (En el apartado 4.8 se describen varios sistemas que utilizan el PLN para el enriquecimiento de ontologías).

En algunos sistemas, como los descritos en los capítulos 5 y 6 de esta memoria, PLN y ontologías están estrechamente relacionados, y aunque, como se verá más adelante, el objetivo es el enriquecimiento de la ontología añadiendo nuevas instancias, las reglas y axiomas de la ontología también se utilizan como herramienta de razonamiento para la desambiguación y obtención de conocimiento textual.

CAPÍTULO 4

INSTANCIACIÓN AUTOMÁTICA DE ONTOLOGÍAS

Resumen La Extracción de Información (EI) es un componente esencial en la instanciación automática de ontologías. En este capítulo se describe en qué consiste en proceso de EI del que forma parte el reconocimiento y clasificación de Entidades Nombradas. Se hace un recorrido por los distintos métodos existentes para la extracción de entidades nombradas y las relaciones entre ellas. Finalmente se analizan y clasifican los principales sistemas de instanciación de ontologías descritos en la literatura.

4.1 Introducción

La instanciación automática de ontologías, u *Ontology Population*, es un elemento crucial en la tarea de relacionar texto con ontologías. Por un lado, la instanciación proporciona una ontología adaptada a los datos y al dominio con el que se relaciona y, por otro lado, la ontología resultante, más rica que la primera, se puede utilizar para tareas relacionadas con la web semántica, tales como la gestión de conocimiento, la recuperación de información, los sistemas de pregunta-respuesta, aplicaciones semánticas de escritorio, etc. (Maynard, 2009).

En la literatura referente a la construcción de ontologías, la mayoría de los autores coinciden en lo lento y costoso que es el proceso de crear una ontología de forma manual e insisten en la necesidad de desarrollar métodos semiautomáticos o automáticos que permitan generar, o al menos enriquecer, ontologías a partir de la información contenida en la web.

El objetivo de este capítulo es, en primer lugar, realizar una aproximación a dos de las herramientas fundamentales para la instanciación de ontologías como son la extracción de Información y, dentro de ésta, el reconocimiento y clasificación de entidades nombradas junto con la extracción de relaciones.

En segundo lugar, se ha realizado una revisión bibliográfica de los principales sistemas de instanciación e ontologías escritos en la literatura.

4.2 Extracción de Información

Instanciación de ontologías y Extracción de información (EI) están estrechamente relacionadas, ya que el conocimiento contenido en los textos, que pasará a formar parte de la ontología, se obtiene por lo general con el uso de técnicas de EI. La EI se usa, por ejemplo, para la obtención de términos a partir de texto en lenguaje natural, tales como entidades nombradas y términos técnicos, así como para relacionar dichos términos con conceptos de la ontología (Maynard, 2009).

La Extracción de Información se considera una tecnología que puede ayudar al ontólogo durante la construcción y el mantenimiento de la ontología. Desde esta perspectiva, la extracción de información puede verse como la tarea de extraer a partir del texto entidades predefinidas, tales como nombres, localizaciones, fechas, etc. (Celjuska & Vargas-Vera, 2004), que son las denominadas Entidades Nombradas. Algunos autores han considerado esta tarea como la tarea de rellenar plantillas (Ramshaw & Weischedel, 2005), es decir, a partir de una colección de documentos se obtienen los datos relevantes sobre uno o más tipos predefinidos de hechos. Cada hecho se representa como una plantilla cuyos slots se completan con la información que se ha obtenido del texto (Ananiadou & McNaught, 2006).

Por ejemplo, con los datos obtenidos de la siguiente noticia se puede completar la tabla 4.1:

> ***Huelva, 30 de Mayo. EuropaPress.*** *La XXXVII edición de la Feria del Libro, que organiza la Concejalía de Cultura del Ayuntamiento de Huelva junto con la Asociación Provincial de Libreros, celebra este martes una nueva edición donde la poesía onubense se configura como protagonista.*

Tabla 4.1. Ejemplo de plantilla para la extracción de información.

Evento	XXXVII edición de la Feria del Libro
Organizador	Concejalía de Cultura del Ayuntamiento de Huelva Asociación Provincial de Libreros
Fecha	30-5-2011
Lugar	Huelva
Tema	Poesía
Fuente	EuropaPress

En este caso, la plantilla incluye los campos Evento, Organizador, Fecha, Lugar, Tema y Fuente que se completan con la información obtenida del texto del ejemplo.

Estos datos pueden utilizarse para mostrárselos directamente a un usuario o pueden ser almacenados en una base de datos para su posterior análisis. En otras ocasiones, se utilizan para indexar documentos en aplicaciones de Recuperación de Información (RI) (Maynard et al., 2003).

Los métodos de EI comenzaron a desarrollarse a finales de los años 80 y comienzos de los noventa propiciados por el aumento de la capacidad de cómputo y de almacenamiento de información, junto con la existencia de grandes cantidades de datos textuales disponibles en formato electrónico. Por otra parte, en 1987 comenzaron a tener lugar las conferencias denominadas *Message Understanding Conferences* (MUC), cuya principal aportación al área fue la de proporcionar un marco de referencia en el que poder evaluar diferentes propuestas para abordar la EI utilizando los mismos conjuntos de datos y métricas de evaluación (Alfonseca, 2008).

En la última conferencia MUC7, celebrada en 1998, se evaluaron las siguientes tareas:

- Reconocimiento y Clasificación de Entidades Nombradas, consistente en la identificación en el texto de nombres de personas, organizaciones, lugares, unidades, monedas, etc.
- Plantillas para Elementos, consistente en rellenar información de plantillas acerca de las entidades.

- Plantillas para las Relaciones, consistente en rellenar plantillas con información acerca de relaciones entre las entidades (por ejemplo, situado en, trabaja para, producto de, etc.)
- Plantillas para Escenarios, consistente en rellenar plantillas con información acerca de eventos mencionados en los textos.
- Resolución de la correferencia, consistente en la identificación de qué menciones del texto se refieren a la misma entidad en el mundo real.

Aunque la extracción de información y la recuperación de información están relacionadas, no deben confundirse entre sí. A continuación, se presentan las principales diferencias entre ambas.

4.2.1 Extracción de Información vs. Recuperación de Información

La Recuperación de Información (RI), es una disciplina que se ha desarrollado profusamente en el ámbito de las ciencias documentales a lo largo de los últimos 30 años. No obstante, ya en 1951, Mooers se refería a ella como el *proceso o el método por el cual un usuario es capaz de convertir su necesidad informativa en una lista de citas de documentos almacenados, que contienen la información útil para él* (Gómez Díaz, 2005).

Esta definición, aún hoy, refleja la esencia de la RI, esto es, un conjunto de técnicas (procedimientos o métodos) mediante los cuales el usuario recupera información relevante para él y que se suele presentar en forma de lista de documentos. Esto implica que el usuario deberá revisar dichos documentos para localizar la información que considere relevante para su consulta.

Por otro lado, las técnicas de PLN se aplican tanto a EI como RI con diferentes grados de complejidad. Mientras la RI localiza textos/documentos y se los presenta al usuario, la EI analiza dichos textos y presenta sólo partes específicas en las que el usuario puede estar interesado (Maynard et al., 2003).

Con el nacimiento de Internet, el desarrollo de herramientas para la RI transcendió del ámbito de las bibliotecas, centros de documentación y archivos, siendo el vasto dominio de la Web el principal objeto de las investigaciones en RI.

Según la distinción tradicional de Baeza-Yates y Ribeiro Neto (Baeza-Yates et al. 1999), en la evolución de los sistemas de RI en las bibliotecas, se pueden distinguir tres momentos fundamentales:

- **Desarrollos iniciales: al principio las búsquedas se hacían a través de las tablas de contenido de los libros.**
- **Recuperación de la Información en bibliotecas: en principio, se utilizaron los propios sistemas de cada institución, evolucionando posteriormente a sistemas comerciales, con interfaces gráficas mejoradas, sistemas que permiten las búsquedas hipertextuales, etc.**
- **La Web y las bibliotecas digitales: el abaratamiento de los costes permite el acceso a múltiples puntos de información. Se puede acceder a gran cantidad de fuentes desde casi cualquier punto del planeta.**

En la siguiente tabla (tabla 4.2), extraída de Ananiadou &McNaught (2006), se resumen las principales diferencias entre RI y EI.

Tabla 4.2 Diferencias entre RI y EI traducida y adaptada de Ananiadou & McNaught (2006).

Recuperación de Información (RI)	**Extracción de Información (EI)**
Devuelve documentos	Devuelve hechos
Es una tarea de clasificación (cada documento es relevante/no relevante para una consulta)	Es una aplicación de PLN, lo que implica análisis del texto y transformación en una representación estructurada.
Se puede llevar a cabo sin realizar una análisis sintáctico de la consulta, es decir, tratando dicha consulta como una "bolsa de palabras" que será lo que se busque en los documentos.	Está basada en el análisis sintáctico y semántico

4.3 Detección y clasificación de Entidades Nombradas

El primer paso en la mayoría de las tareas de EI es detectar y clasificar todos los nombres propios mencionados en un texto, a esta tarea se le denomina Reconocimiento de Entidades Nombradas también conocida como *Named Entity Recognition* (NER) en inglés.

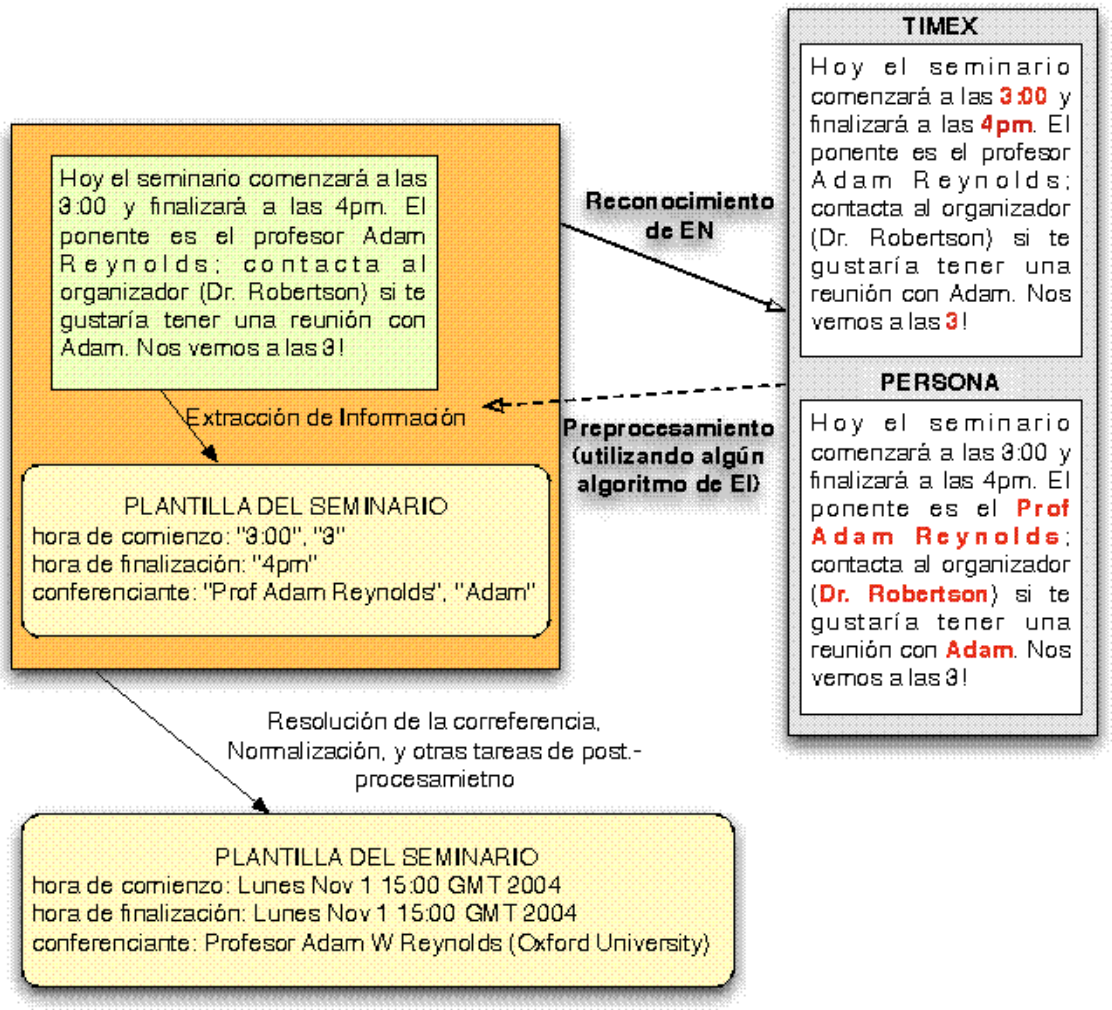


Figura 4.1 Extracción de Información y Entidades Nombradas. Traducido y adaptado de Alberto Lavelli et al. (2008).

En la figura 4.1, los elementos del texto a considerar para cumplimentar una plantilla de EI son entidades nombradas del tipo Persona y Tiempo. En un proceso posterior las entidades nombradas serán normalizadas y desambiguadas en el caso

de que sea necesario. En primer lugar, se identifican las entidades nombradas del tipo Persona y Tiempo en este caso que se utilizan para completar los campos de la plantilla. Cada mención diferente de una entidad nombrada se incluye en la plantilla tantas veces como sea localizada en el texto. Por ejemplo como *Hora de comienzo* se incluyen tanto *3* como *3:00*, en una fase posterior se procederá a normalizar y desambiguar las entradas en el caso de que sea necesario.

Las entidades nombradas se han convertido en un importante componente tecnológico para muchas aplicaciones de PLN, entre las que se incluye la EI, pero también los sistemas de Pregunta-Respuesta, el Resumen Automático y la Recuperación de Información. (Sekine et al., 2002).

La extracción de entidades del dominio que pueden representar instancias de los conceptos de la ontología, es el primer paso en la mayoría de los sistemas de instanciación automática de ontologías.

Una entidad nombrada es un término uni-palabra o multipalabra que se refiere a una entidad que tiene nombre y que posee algún tipo de interés (Alfonseca, 2008). Aunque en muchas ocasiones coinciden con los nombres propios, el fenómenos de las entidades nombradas es más abarcador y no todas las entidades nombradas son nombres propios.

Inicialmente, únicamente se consideraron entidades nombradas los términos utilizados para referirse a Personas, Organizaciones y Localizaciones (POL), y también los utilizados para referirse a fechas, expresiones temporales, expresiones monetarias y porcentajes (Grishman, 2006), aunque en la actualidad se puede considerar una entidad nombrada cualquier término que haga referencia a un elemento representativo del dominio. Por ejemplo, una entidad nombrada sería "José López", "J. López", "Estados Unidos", "Pizza Hut" y en determinados dominios pueden serlo "Hotel Príncipe Felipe", "Hidruro de aluminio" o "Renault Clio". En la jerarquía de entidades nombradas propuesta por Sekine (2004), se pueden identificar más de 200 tipos de entidades nombradas.

La tarea que se encarga de la identificación y clasificación de Entidades Nombradas en un texto se denomina como Reconocimiento y Clasificación de

Entidades Nombradas (*Named Entity Recognition and Classifcation*), que consiste en detectar una secuencia de palabras que describen una entidad relevante en un dominio y determinar el tipo al que pertenece. Esta tarea es altamente dependiente de la aplicación para la que haya sido desarrollada, distintos dominios implican distintas entidades. Es decir, que una misma producción lingüística puede ser considerada como Entidad Nombrada en un sistema y en otros no (Borrega et al., 2007). Los distintos tipos de textos contienen una buena parte de nombres que no pueden ser adecuadamente analizados a menos que esos nombres sean identificados como entidades nombradas (Grishman, 2006).

En el siguiente ejemplo, se pueden ver identificadas algunas de las entidades nombradas clásicas.

[PERSONA Jorge Francisco Isidoro Luis Borges]([LUGAR Buenos Aires] , [FECHA 24 de agosto de 1899] – [LUGAR Ginebra], [FECHA 14 de junio de 1986]) fue un escritor argentino, uno de los autores más destacados de la literatura del [FECHA siglo XX]. Publicó ensayos breves, cuentos y poemas.

En (Doddington, 2004), se distingue entre Entidades Nombradas y Menciones de Entidades Nombradas. Las menciones, según esta distinción, se refieren a las expresiones de referencia, mientras que las entidades se refieren al referente. Es decir, las menciones son fragmentos de texto que se pueden clasificar como entidades y que tienen el mismo referente en determinados contextos (Magnini, 2006).

En términos lingüísticos, una mención de una entidad nombrada sería el significante (la cadena de caracteres) utilizada para hacer alusión a un mismo referente del mundo extralingüístico, que en este caso estaría restringido a aquellas entidades que previamente se han considerado entidades nombradas.

Adaptando el ejemplo que propone Magnini (2006), tendríamos que, en un contexto particular, tanto "el presidente de España" como "José Luis Rodríguez Zapatero" se refieren a la misma entidad, es decir, son instancias particulares del

tipo Persona cuyo nombre es "José Luis", los apellidos son "Rodríguez Zapatero" y su rol es "Presidente de España". En este caso, además, hay que tener en cuenta el momento en el que se ha producido en mensaje, ya que "Presidente de España" no siempre se refiere a la misma persona.

Se puede distinguir entre tres tipos de menciones:

- Menciones nombradas, cuando se utiliza el nombre propio para referirse a la entidad: Federico García Lorca, Hotel los Narejos.
- Menciones nominales, cuando se utiliza otro nombre diferente del nombre propio para referirse a la entidad: el poeta, el establecimiento.
- Menciones pronominales, cuando se utiliza un pronombre para referirse a la entidad: *él*.

Por otro lado, la identificación de entidades nombradas implica la resolución de dos fenómenos textuales, esto es, la correferencia y la coocurrencia.

- La **Correferencia** se refiere a la identificación de entidades que tengan el mismo referente en el mundo real aunque la expresión lingüística utilizada para referirse a ella varíe. Es a lo que se ha hecho alusión previamente como menciones nombradas, nominales y pronominales. En Bontcheva et al. (2002), se presenta un algoritmo para la resolución de la correferencia tanto de nombres propios como de pronombres mediante el desarrollo de un nuevo módulo de ANNIE (ver apartado 3.3.1). Para ello, establecen relaciones entre elementos con el mismo referente dependiendo de su posición en el texto tal que determinados pronombres y siglas se pueden relacionar con una misma entidad nombrada.

- La **Coocurrencia** con otras entidades consiste en que una entidad implica la aparición de otras entidades en un mismo contexto. Por ejemplo, Federico García Lorca puede implicar la aparición de otras entidades como *La casa de Bernarda Alba* o *1936*.

Como se ha dicho, la extracción de instancias para la población de ontologías se asemeja, en gran medida, al reconocimiento y clasificación de entidades nombradas. Sin embargo, una de las diferencias fundamentales es que en el reconocimiento y clasificación de entidades nombradas, cada ocurrencia de un término reconocido se clasifica de forma independiente, mientras que en la población de ontología, es el término, independientemente del contexto en el que aparece, el que tiene que ser clasificado (Tanev & Magnini, 2006).

Es decir, en un texto pueden aparecer entidades del tipo Hotel, el reconocedor de entidades nombradas identifica todas y cada una de las menciones de este tipo de entidad que aparecen en el texto. Por el contrario, en la instanciación de ontologías, las entidades nombradas extraídas y normalizadas son las que tienen que ser clasificadas en la ontología de dominio, y será solamente una mención de la misma entidad la que se incluya en la ontología.

La detección e inclusión de las entidades nombradas como instancias de las clases de la ontología, es una técnica que ha sido empleada por autores como (Tanev & Magnini, 2006; Magnini et al., 2006; de Boer, 2007).

Las entidades nombradas pueden ser obtenidas mediante listas o gazetteers, mediante reglas o mediante otros métodos automáticos (p.ej. las palabras semilla), aunque generalmente se utiliza una combinación de estos tres métodos.

4.3.1 Métodos basados en listas o Gazetteers

Las listas utilizadas para el reconocimiento de entidades nombradas se denomina Gazetteers y aunque originariamente hacía alusión a los diccionarios geográficos, las listas pueden contener cualquier tipo de nombres propios, por ejemplo nombres de personas y apellidos, nombres de compuestos químicos, obras de arte, etc.

Como se indica en (Nadeau et al., 2006), con la identificación mediante listas de entidades nombradas, surgen tres problemas fundamentales:

- Ambigüedad entidad-nombre, que se produce cuando una palabra o frase que comienza por mayúscula es una entidad, a menos que:
  - Aparezca en otros lugares del documento con otra inicial.
  - Sólo aparezca al principio de una oración o de una cita.
  - Sólo aparezca en una oración en la cual todas las palabras contienen tres caracteres que comienzan por mayúscula.

- Detección de los límites de la entidad, cuando la entidad está compuesta por dos o más palabras. Generalmente, se agrupan todas las entidades consecutivas del mismo tipo y todas las entidades con cualquier palabra consecutiva que empiece por mayúscula. (Esto evidentemente no funciona si las entidades de las que estamos hablando no comienzan por mayúscula).

- Ambigüedad entidad-entidad, tiene lugar cuando una cadena de caracteres pertenece a una entidad nombrada de más de un tipo (homonimia). Nadeau et al. (2006) establecen la hipótesis de que al menos una de las entidades debe aparecer en el contexto que la identifica unívocamente, y para reconocerlas habría que definir otro tipo de reglas como, por ejemplo, si van precedidas de algún otro término, como se verá en el siguiente apartado.

Considerando los inconvenientes arriba descritos, un sistema de reconocimiento de entidades nombradas basado solamente en listas de entidades es poco fiable. No obstante, suele ser un componente en los sistemas de aprendizaje automático. La extracción automática de gazetteers a partir de la web o de colecciones documentales no etiquetadas para sistemas de reconocimiento de entidades nombradas se ha llevado a cabo en trabajos como los de (Riloff & Jones, 1999), en donde se parte de un conjunto de patrones léxicos y entidades

que van aumentando paulatinamente a medida que el sistema descubre nuevas entidades. De igual manera, Etzioni et al., (2005) proponen un método semiautomático de palabras semilla para la identificación de términos afines. La metodología propuesta por (Lin & Pantel, 2001) se basa también en la similitud entre contextos terminológicos, donde se crean clusters de palabras relacionadas semánticamente en función de las dependencias sintácticas que comparten. Más recientemente (Toral & Muñoz, 2006; Kazama & Torisawa, 2007; Ratinov & Roth, 2009) han extraído listas de gazetteers a partir de la Wikipedia. Otro sistema para la extracción automática de gazetteer es el propuesto por (Nadeau et al., 2006), en el que se extraen listas de ciudades y marcas de coche de manera no supervisada partiendo de palabras semilla.

4.3.2 Métodos basados en patrones lingüísticos o reglas

Las reglas para la extracción de entidades nombradas se elaboran en base a la formalización de las características textuales recurrentes que existen en el texto. Este tipo de reglas o patrones pueden expresarse con distintos grados de abstracción lingüística y van desde estructuras sintácticas sencillas a la indicación de la coocurrencia de dos o más términos al análisis morfológico.

Por ejemplo, para encontrar una dirección podríamos utilizar la siguiente combinación:

Calle + (det) + Palabra en Mayúscula

Encontramos una palabra que es el disparador o *trigger Calle*, que facultativamente puede ir seguida por un determinante y a continuación una palabra en mayúscula.

La implementación de dichas reglas se puede realizar mediante *transducers* o lenguajes de expresiones regulares, como por ejemplo JAPE, descrito en el apartado 3.3.1.2. Un *transducer* permite asociar una etiqueta concreta con una

palabra en el texto, por ejemplo, puede asignarle una categoría sintáctica a una palabra. En la figura 4.2 se puede ver el *transducer* de Persona.

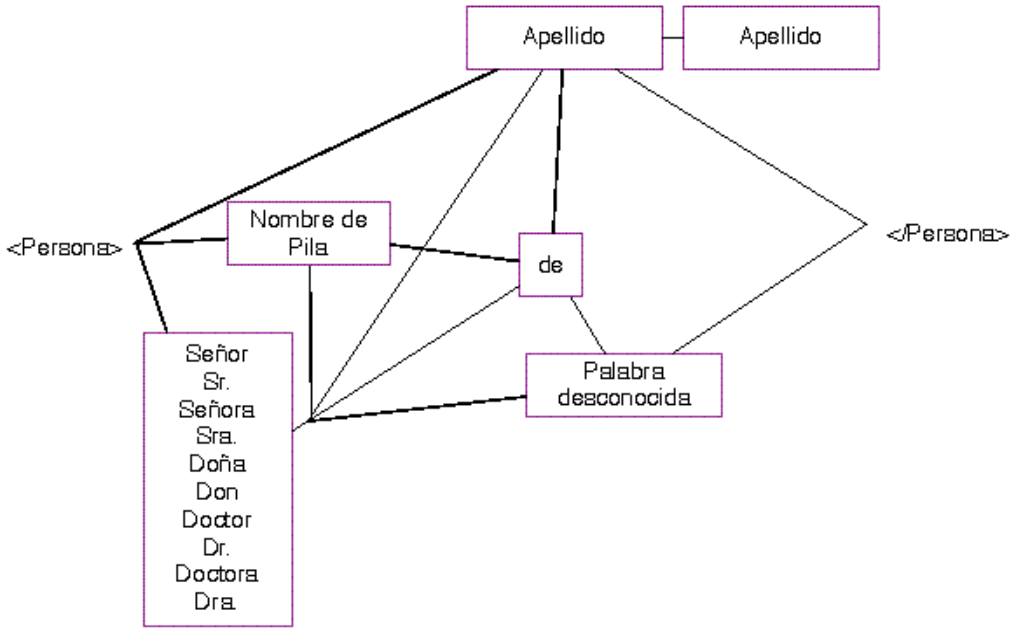


Figura 4.2 Ejemplo de transducer.

Por ejemplo, la expresión "Sr. Tomás de Soto", representa una persona, como también lo es "Sr. De Soto", "Sr. Soto", "Tomás", "Tomás Soto", etc.

Hay *transducer* que son estándar pero hay otros que vienen dados por el dominio. Por ejemplo, la clase semántica *bombardear puede* implicar la aparición del sustantivo y preposición *explosión de* seguida de otro determinante y un elemento perteneciente a la clase semántica de los explosivos como *bomba, granada, obús,* etc. Esto quedaría reflejado en el siguiente patrón:

<explosión> <de> <DET> <Explosivo>

En el trabajo de (Faure & Poibeau, 2000) se extraen entidades nombradas mediante este tipo de patrones léxico-sintácticos.

4.3.3 Métodos basados en aprendizaje automático

El aprendizaje automático de entidades nombradas es el más frecuente, y en él convergen tanto el uso de gazetteers como el uso de patrones léxico-sintácticos.

Como señala Alfonseca (2008), los sistemas que utilizan técnicas de aprendizaje automático incluyen el aprendizaje basado en memoria, los modelos de entropía máxima y modelos de Markov ocultos (McCallum et al., 2000; Klein et al., 2003; Florian et al., 2003; Kozareva et al. 2005). Otros sistemas utilizan para el aprendizaje listas de transformación (Black & Vasilakopoulos, 2002), algoritmos de boosting (Carreras et al. 2003), y Support Vector Machines (SVM) (Isozaki & Kazawa, 2002, Mayfield et al., 2003; Li et al., 2005).

4.4 Entidades nombradas en el dominio de la biomedicina

El reconocimiento de entidades nombradas en biomedicina forma parte de muchos del los sistemas para la minería de texto y EI en dominios biomédicos.

En biomedicina, las entidades nombradas, también llamadas bio-entidades, suelen incluir genes, proteínas o enfermedades.

Aunque existen varias nomenclaturas biomédicas que facilitan la identificación de genes y proteínas, tales como HUGO[16] o Swiss_Prot[17], la identificación de entidades nombradas en biomedicina presenta problemas adicionales. Siguiendo la clasificación que se realiza en Ananiadou el al. 2006 y (Zhou et al., 2005), los principales problemas son los siguientes:

- **Nombres Ambiguos**
  - Algunos nombres denotan distintos genes y proteínas, como por ejemplo ARF.

[16] http://www.genenames.org/

[17] http://web.expasy.org/docs/swiss-prot_guideline.html

- Hay genes y proteínas que se pueden confundir con palabras inglesas tales como *can, for, zip,* etc.
- Algunos nombres de genes no se pueden desambiguar si no se tienen en cuenta la especie.
- Hay nombres que pueden denotar entidades biomédicas de clases diferentes, por ejemplo *myc-c gene* y *myc-c-protein*.

- **Longitud**
  - Una entidad nombrada puede estar expresada por varios términos como *47kDa sterol regulatory element binding factor*.

- **Sinónimos**
  - Una entidad puede ser denotada por múltiples nombres.
  - Algunos nombres de genes y proteínas denotan la misma proteína que es idéntica a su proteína homóloga en diferentes especies.

- **Variaciones**
  - Las variaciones gráficas en la expresión de genes y nombres de proteínas que denotan las mismas entidades es frecuente en la literatura. Entre ellas se incluyen las variaciones de caracteres, variaciones en las mayúsculas, en el orden de las palabras, sintácticas y en las abreviaturas. Por ejemplo, *D(2)* o *D2*, *RNase P protein* o *RNase P*, cuyo referente es la misma entidad en cada uno de los pares de casos.
- **Variación en el rango de los nombres a identificar.**
  - Dependiendo del objetivo del sistema, se reconocen diferentes tipos de entidades nombradas. Un sistema no debe reconocer todas las entidades nombradas biomédicas que aparezcan en un texto, sino sólo aquellas que sean de interés para la aplicación. Por ejemplo, si se está elaborando una lista de proteínas no será necesario recopilar nombres genéricos que hacen referencia a las proteínas, como "la proteína" o "dichas proteínas", sino solamente nombres específicos.

En la literatura, existen numerosas propuestas para la identificación de entidades nombradas en biomedicina. Aunque algunas de ellas se basan en el uso de diccionarios para la identificación de entidades nombradas (Ono et al., 2001), la mayoría de los sistemas están basados en reglas o en técnicas de aprendizaje automático (Saquete et al., 2008). Fukuda (1998) desarrolló uno de los primeros sistemas para el reconocimiento de entidades nombradas aplicado a las proteínas en el que las reglas estaban definidas manualmente.

Con los sistemas de aprendizaje automático, tales como Modelos de Markov, *Conditional Random Field* o Modelos de máxima entropía, se trata de automatizar el proceso. El trabajo presentado en (Settles, 2004) extrae entidades nombradas utilizando *Conditional Random Fields*. Este método toma en consideración una serie de características ortográficas y semánticas para entrenar el sistema. Otros trabajos como (Shen et al., 2003), utilizan Modelos de Markov para el reconocimiento de entidades nombradas. Finalmente, una propuesta basada en Support Vector Machine es la que se presenta en (Lee et al., 2004).

Otros sistemas como GENIA (Kim et al., 2003) parten de corpus anotados para la extracción de entidades nombradas. El reconocedor de entidades nombradas de GENIA es capaz de extraer 6 tipos de entidades nombradas (genes, proteínas, DNA, RDN, tipo de célula y línea celular) a partir del corpus anotado de GENIA. El sistema, que ha sido entrenado siguiendo varias técnicas de aprendizaje automático, ha demostrado una mayor eficiencia cuando se aplican *Support Vector Machines* (Kim et al., 2004).

La ventaja de las técnicas de aprendizaje automático es que pueden identificar entidades biomédicas potenciales que no han sido previamente incluidas en recursos léxicos estándar. Este sistema y el corpus al que está asociado se describen más en profundidad en el capítulo 6 de esta memoria.

4.5 Extracción de relaciones entre Entidades Nombradas

Una vez que las entidades han sido identificadas en el texto, el siguiente paso es la extracción de las relaciones entre ellas. Por lo tanto, la extracción de relaciones es la tarea del detectar y caracterizar una relación entre dos entidades nombradas (Lavelli et al., 2008).

Una relación conecta dos entidades de acuerdo con algún criterio (Alfonseca, 2008). Por ejemplo, la relación *is located in* o *está localizado en* puede conectar una entidad del tipo HOTEL con una entidad del tipo LOCALIZACIÓN como se puede ver en la figura 4.3, en donde un hotel que ha sido anotado como una entidad nombrada se relaciona mediante una relación de localización con una entidad nombrada anotada como localización.

> *Un nuevo concepto en hotel boutique,* [HOTEL *Racó de Buenos Aires*] ***localizado en*** *el corazón de* [LOCALIZACIÓN *Buenos Aires*]*, ofrece una nueva visión en materia de lujo y confort a la hora de viajar y hospedarse en nuestra ciudad.*

Figura 4.3 Relación entre entidades nombradas.

Las relaciones que se extraen entre entidades nombradas no son necesariamente binarias. No obstante, la extracción de relaciones no-binarias implica un problema añadido a la hora de representarlas, como, por ejemplo, en una ontología. Eso se debe a que las relaciones se establecen de manera directa entre dos clases, aunque haya relaciones como las taxonómicas, que por la propiedad transitiva, relacionan más de dos elementos ontológicos.

Identificar relaciones a partir de texto en lenguaje natural no es una tarea sencilla, ya que las relaciones pueden expresarse de diversos modos.

En la literatura, generalmente se establece una distinción entre la extracción de relaciones taxonómicas y el resto de relaciones (Alfonseca, 2008). Las relaciones

taxonómicas se refieren a relaciones de jerarquía también denominadas relaciones de hiponimia o relaciones is_a/es_un.

Curse (1986) se refiere a la relación léxica de hiponimia como la inclusión de una clase en otra. La clase más general es el hiperónimo y la clase más específica es el hipónimo. Un hipónimo posee todos los rasgos semánticos o semas de otra más general, su hiperónimo, pero que añade en su definición otros rasgos semánticos que lo diferencian del segundo.

Curse (1986) representa la relación mediante el esquema que se muestra en la figura 4.4.

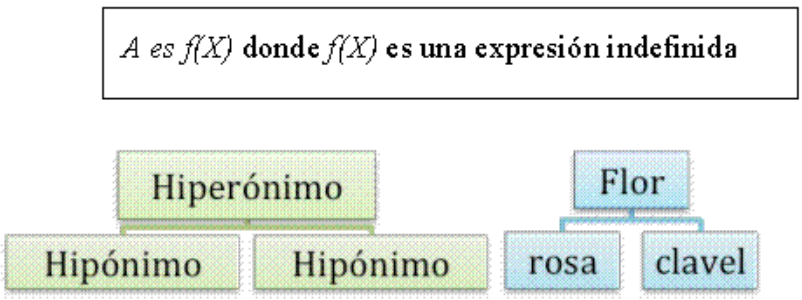


Figura 4.4 Representación gráfica de la relación de hiponimia.

La relación de hiponimia o taxonómica es importante, ya que las entidades nombradas a menudo se organizan en taxonomías, es decir, en subtipos de entidades. Como señala Alfonseca (2008), se puede considerar que el aprendizaje de las relaciones taxonómicas entre los conceptos pertenecientes a una ontología y las instancias encontradas en los textos está estrechamente relacionada con la extracción de relaciones entre entidades nombradas.

Alfonseca clasifica la metodologías de extracción de relaciones entre entidades en cuatro grupos que se describen a continuación.

4.5.1 Métodos basados en diccionarios

Las definiciones de diccionarios y glosarios tienen una estructura predefinida en la que la información se presenta de forma concisa enumerando a todos sus participantes. Como se indica en (Navigli et al., 2010), los participantes se pueden clasificar en:

- *Definendum*: El término que se va a definir.
- *Defininiens*: El significado del término.
- *Definitor*: El término o términos que ligan el *Definendum* y el *Definiens*, con frecuencia se trata de un verbo.

Este tipo de estructuras favorecen la extracción de relaciones de hiponimia entre los conceptos a partir de un corpus anotado (Navigli & Velardi, 2010).

4.5.2 Métodos basados en propiedades distribucionales de las palabras

Según la semántica distribucional, los términos que son semánticamente similares comparten contextos similares, y por tanto se pueden utilizar para distribuciones de coocurrencia de términos para calcular una medida de similitud semántica entre ellos (Firth, 1961).

Aunque estos métodos se suelen utilizar para el aprendizaje de relaciones taxonómicas, también se puede llevar a acabo el aprendizaje de otro tipo de relaciones.

Por ejemplo en la figura 4.5, en el fragmento de texto A "Rafael Nadal" se puede etiquetar como una entidad de tipo Persona- Deportista- Tenista, según el grado de especificidad. A su vez, "Copa Davis" se puede etiquetar como un Evento deportivo o como un Torneo, mientras que la relación que se establece entre Deportista y Evento Deportivo en este caso es NoParticiparEn.

Entonces utilizando un corpus donde se han anotado estas entidades nombradas, se pueden anotar en diferentes corpus entidades semánticamente

similares que aparezca en el mismo contexto, pudiéndose inferir el mismo tipo de relación entre ellas. Este es el caso de los fragmentos de texto B, C y D.

A
El tenista español [DEPORTISTA Rafael Nadal] ha anunciado que no participará en la eliminatoria de cuartos de final de la [EVENTO DEPORTIVO Copa Davis] entre España y Estados.

B
El tenista español [DEPORTISTA Tommy Robredo] ha anunciado que no participará en el [EVENTO DEPORTIVO Roland Garros].

C
Vencedor de dos etapas y subiendo puertos como nunca en Suiza, [DEPORTISTA Peter Sagan] no participará en el proximo [EVENTO DEPORTIVO Tour de Francia].

D
El delantero [DEPORTISTA Carlos Vela] no tomará parte de la [EVENTO DEPORTIVO Copa América Argentina 2011] con la selección

Figura 4.5 Ejemplo de anotación de relaciones entre entidades nombradas.

Este método se utiliza en Hasegawa et al. (2004), en donde la metodología propuesta se basa en la hipótesis de que entidades similares aparecen en contextos similares y que, además, si se trata de relaciones significativas se repiten con frecuencia a lo largo de un corpus. Los autores descubren las relaciones entre entidades mediante un proceso de *clustering*. Para llevar a cabo este proceso, en primer lugar etiquetan las entidades en el corpus y obtienen aquellas entidades que coocurren en un contexto similar, agrupándolas en un *cluster*. Cada *cluster* se etiqueta con un tipo de relación de modo que, cuando el sistema encuentre dos entidades en un contexto similar las considerará como pertenecientes al mismo tipo de *cluster*.

En (Agirre, 2000) extraen relaciones no taxonómicas entre conceptos de una ontología a partir de la web. Elaboran unos patrones lingüísticos básicos basados en sintagmas verbales que, generalmente, son los encargados de indicar las

relaciones entre dos entidades. Utilizan estos patrones para realizar consultas en la web. Entonces, aquellas entidades extraídas de la web que coocurren en contextos similares a los de las entidades semilla utilizadas para la creación de los patrones, mantienen entre sí una relación equivalente a la de las primeras.

4.5.3 Métodos basados en patrones lingüísticos

Se trata de patrones léxico-sintácticos en los que se refleja la relación entre dos entidades. Como la mayoría de los métodos basados en patrones, presentan el problema de la creación de los mismos, ya que si se hace de forma manual, a parte del elevado costo, la cobertura es limitada. Por el contrario, los patrones elaborados automáticamente tienen un coste menor, pero su precisión es también más baja.

Uno de los métodos más comunes utilizado para el aprendizaje automático de patrones está basado en las palabras semilla. Las palabras semilla son un conjunto de términos que están relacionadas entre sí y que servirá como base para extraer frases, normalmente de la web, y a partir de ellas extrapolar las relaciones que existen entre los términos.

Por ejemplo, para aprender un patrón relativo a la fecha de nacimiento, se puede comenzar buscando los términos Picasso-1881

Algunos de los resultados obtenidos y los patrones que se pueden obtener de ellos son:

Pablo Picasso (1881–1973)	*<palabra1><palabra2> <paréntesis> <número>*
Pablo Ruiz nació en 1881	*<palabra1><palabra2> nació en <número>*

Una vez obtenidos estos patrones, se pueden aplicar a otros términos que no se encontraban entre las palabras semilla, obteniendo la relación "fecha de nacimiento" entre otras entidades.

Uno de los primeros sistemas descritos en la literatura que utilizan este método para obtener información de la Web es el de Brin (1999), en el que se extrae la relación autor- título. Este mismo método se utiliza en (Agichtein & Gravano, 2000; Ravichandran & Hovy, 2002; Shinyama & Sekine, 2006; Nakamura-Delloye, 2011).

En (Bollegala et al., 2011) se utiliza esta metodología para extraer relaciones a gran escala de la web, cuyo objetivo es la adaptación de sistemas ya existentes a la extracción de nuevas relaciones.

4.5.4 Métodos basados en aprendizaje automático a partir de corpus anotados

Los corpus pueden estar anotados sintácticamente y, a partir de la creación de árboles de dependencias, se puede inferir la relación que existe entre pares de entidades (Zelenko et al., 2003). En (Culotta & Sorensen, 2004), se extraen relaciones de un corpus anotado mediante Support Vector Machine, donde las relaciones extraídas las representan en un árbol de dependencias gramaticales.

Los métodos de aprendizaje automatizado se han utilizado también en los trabajos de Zhang et al. (2005), Chen et al. (2006) y He et al. (2006).

4.6 Extracción de relaciones entre entidades nombradas en el dominio de la biomedicina

El reconocimiento automático de relaciones entre entidades nombradas en dominios biomédicos es una práctica importante en el campo de la bioinformática, en donde la producción textual, sobre todo en forma de artículos científicos, ha aumentado de forma tan significativa. A continuación, se mencionan los principales sistemas para la extracción de relaciones en biomedicina recogidos en la literatura.

El sistema de Ray Craven (2001), uno de los primeros que se desarrolló para la obtención de relaciones biomédicas, utiliza Modelos de Markov para la obtención de dos tipos de relaciones entre entidades. Por un lado, obtienen la relación subcellular-location (subcelular-localización) entre las entidades *Proteins* (Proteínas) y *Locations* (Localización) y, por otro lado, la relación *disorder-association* (trastorno-asociación) que existe entre las entidades *Gene* (Gen) y *Disorder* (Trastorno). Esta metodología es la que se usa también en (Rosario & Hearst, 2004), que identifican la relación semántica entre *Treatment* (Tratamiento) y *Disease* (Enfermedad) en textos biocientíficos mediante modelos gráficos y redes neuronales.

En (Chun et al., 2006), se identifica la relación *gene- disease* (gen-enfermedad) a partir de los resúmenes del MEDLINE estudiando la coocurrencia de términos. Las relaciones se clasifican en función de seis tópicos. Para la identificación de entidades nombradas y relaciones, utilizan técnicas de aprendizaje automático, además de diccionarios de genes y enfermedades.

En Bundschus (Bundschus et al., 2008), llevan a cabo dos procesos diferentes. En primer lugar, extraen las relaciones entre enfermedades y tratamientos, y, en segundo lugar, identifican las relaciones entre genes y enfermedades a partir de un conjunto predefinido de relaciones. Para la detección de relaciones, utilizan Conditional Random Files (CRFs), que son modelos de gráficos probabilísticos utilizados para etiquetar y segmentar secuencias.

Otras aproximaciones, como la de (He et al., 2006) utilizan técnicas de aprendizaje automático (*machine learning*), junto con técnicas de análisis del discurso, para la extracción de las interacciones entre *proteins-proteins* (proteínas-proteínas). Sin embargo, las únicas relaciones que extraen son taxonómicas y partonómicas.

El algoritmo que se presenta en (Sharma et al., 2010) extrae cinco tipos de entidades y las relaciones existentes entre ellas. Como elemento que indica una relación, utilizan los verbos extraídos de UMLS, y las relaciones se extraen en base a la sintaxis de las oraciones.

Otra de las metodologías basadas en patrones léxico-semánticos es la de (Sahay et al., 2008). Los patrones creados son del tipo "d is caused by e" y se utilizan para realizar búsquedas en la Web. De entre los resultados obtenidos, se seleccionan aquellos que aparecen en los fragmentos de la web presentados por un buscador. Con la información obtenida, enriquecen algunos recursos existentes, tales como UMLS.

4.7 Aprendizaje automático de ontologías (*Ontology Learning*)

En el ámbito de la Web Semántica la construcción de ontologías de forma (semi-) automática (*ontology learning* en inglés) a partir de texto en lenguaje natural, se plantea como uno de los principales retos. Las ontologías se consideran una piedra angular en el desarrollo de la Web Semántica, y la construcción manual de las mismas es una tarea lenta y costosa que ralentiza el desarrollo de dicha web (Cimiano, 2005). Por este motivo, ya en la década de los 90 encontramos las primeras propuestas que asentaron las bases para metodologías posteriores (Gruninger & Fox, 1995; Uschold & King, 1995; Gruber, 1993).

Ontology Learning o Aprendizaje automático de ontologías hace referencia al proceso de adquirir (construir o integrar) una ontología (semi) automáticamente (Gómez-Pérez & Manzano-Macho, 2003).

A diferencia de la construcción manual de ontologías, la construcción automática de ontologías no sólo es capaz de descubrir conocimiento ontológico a gran escala y de forma rápida, sino que es capaz de mitigar los posibles elementos subjetivos y las inconsistencias derivadas de las construcción manual (Zhou, 2007).

En un proceso ideal de creación de una ontología de forma manual, encontramos al menos tres recursos fundamentales, el ontólogo, el experto del dominio y los *corpora* (ver figura 4.6). El ontólogo es el encargado de modelar en forma de ontología el conocimiento del dominio que viene dado tanto por el

experto del dominio como por los corpora especializados. La colaboración ontólogo-experto dará las claves sobre los conceptos fundamentales a modelar en la ontología, siendo el resultado una ontología semilla que se enriquecerá paulatinamente. Por otro lado, el ontólogo recurre a los corpora especializados para la obtención de conocimiento específico, como por ejemplo, las instancias de las clases o sus relaciones.

Por su parte, el experto del dominio supervisará y evaluará la ontología resultante, enriqueciéndola y actualizándola. El coste de una ontología producida de este modo es tan alto que resulta inviable para el desarrollo de la Web Semántica.

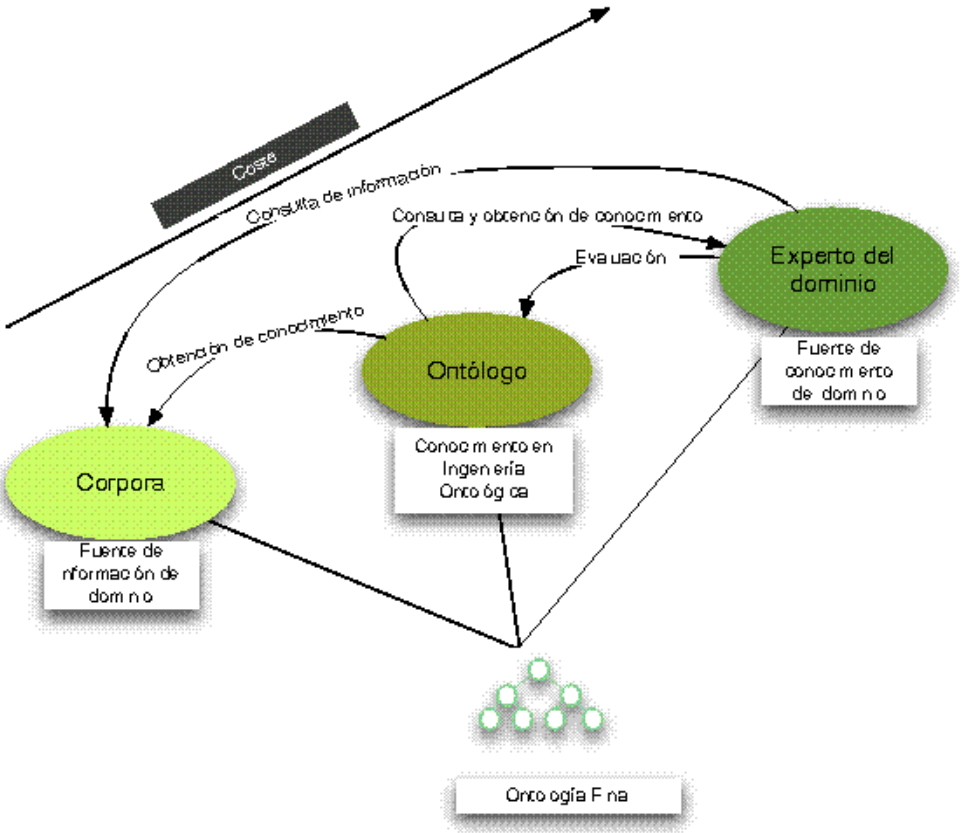


Figura 4.6 Desarrollo manual de una ontología.

Es aquí donde entran en juego las técnicas previamente descritas tanto de PLN como de EI y reconocimiento de entidades nombradas.

El conocimiento contenido en la web en general, o en corpora especializados en particular, puede ser la base para la obtención de los conceptos o de las instancias para enriquecer una ontología de dominio, como es el caso de las metodologías que se describen en esta memoria.

En el proceso de construcción de ontologías de forma automática, el coste de los recursos humanos necesarios se reduce significativamente. La reutilización de ontologías existentes, la selección de *corpora* los suficientemente abarcadores con información relevante y correcta o la inclusión de razonadores para comprobar la consistencia de la ontología, son elementos que convierten al proceso de construcción automática en el único modo viable del desarrollo de ontologías para la Web Semántica.

Como se indica en Petasis et al. (2007), el aprendizaje automático de ontologías puede dividirse en tres aspectos fundamentalmente:

- Integración de ontologías existentes
- Construcción de una ontología desde cero o ampliando una ya existente
- Adaptación de una ontología genérica a un dominio específico.

Además desde un punto de vista procedimental, el aprendizaje automático de ontologías se puede dividir en 4 tareas principales (Petasis et al., 2007):

- *Ontology Population* o Instanciación automática de ontologías: Consiste en insertar en la ontología nuevas instancias de los conceptos y las relaciones.
- *Ontology enrichment* o Enriquecimiento de una ontología: Consiste en añadir nuevos conceptos, relaciones y reglas.
- Resolución de inconsistencias: Mediante esta tarea, se trata de mantener la consistencia de una ontología mediante la eliminación de información contradictoria.
- Evaluación: Se trata de medir la calidad de una ontología mediante ciertos parámetros preestablecidos.

La mayoría de los sistemas desarrollados combinan técnicas de análisis lingüístico con algoritmos de aprendizaje automático con el objetivo de localizar los conceptos y relaciones potencialmente interesantes y las relaciones entre ellos.

Mientras que la meta del *Ontology Learning* es la adquisición de nuevos conceptos y relaciones con el consecuente cambio de la ontología en sí misma, la meta del Ontology Population es la extracción y clasificación de instancias de los conceptos y relaciones definidas en la ontología (Tanev & Magnini, 2006).

La instanciación automática de ontologías es fundamental para la provisión de servicios de conocimiento basados en ontologías que sean de calidad.

4.8 Instanciación automática de ontologías (*Ontology Population*)

El proceso de extender una ontología mediante la inserción de nuevas instancias de los conceptos y/o de las relaciones se denomina Ontology Population o Instanciación automática de Ontologías y se puede considerar una subtarea dentro del proceso de *Ontology Learning* (Petasis et al., 2007). En la siguiente figura, se expone de forma esquemática cuál sería el proceso de población de una ontología. En primer lugar, se parte de uno o más corpus y de una ontología inicial. De los corpora, se obtienen las posibles instancias de los conceptos y relaciones mediante el uso de herramientas de EI. De entre el conjunto de instancias candidatas, aquellas que sean consideradas válidas por el sistema pasan a formar parte de la ontología durante el proceso de instanciación. La salida es una ontología poblada.

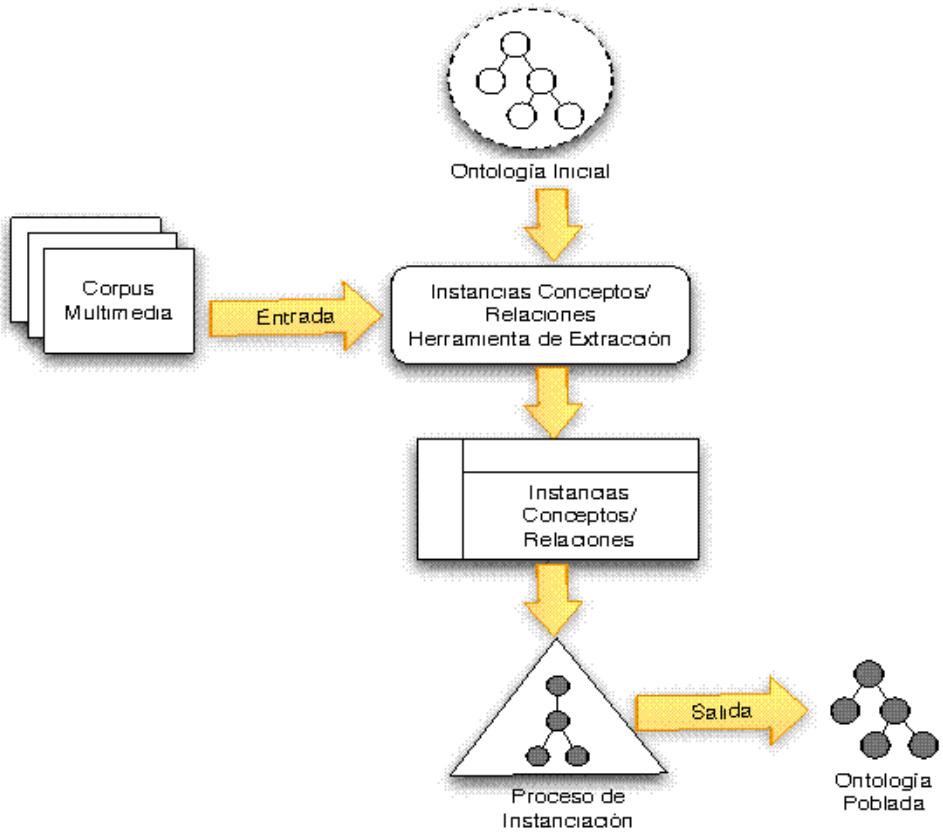


Figura 4.7 Proceso de Instanciación de una ontología. Traducido y adaptado de (Petasis et al., 2007).

El proceso de Ontology Population se puede dividir en tres tareas principales, extracción de instancias de las clases, extracción de valores de los atributos y extracción de instancias de las relaciones.

Extracción de Instancias de las clases: La mayoría de los sistemas cuentan con un modulo para la identificación de entidades nombradas que se convierten en candidatos a instancias de la ontología. Los términos que se identifican hacen referencia en el texto a instancias concretas de una clase o concepto. Los principales problemas que se presentan es decidir, por un lado, si se trata de una instancia o de una clase y, por otro lado, cómo proceder a la desambiguación para la inclusión en uno u otro concepto de la ontología.

Por ejemplo, supongamos el siguiente texto en lenguaje natural:

El [HOTEL Hotel Dos Mares] cuenta con piscina climatizada y amplios jardines en los que celebrar todo tipo de eventos. El [HOTEL hotel] además dispone de una terraza desde la que se puede disfrutar de magníficas vistas.

En este caso, ambas menciones hacen referencia a la misma entidad del tipo hotel. Por lo tanto, se convertirían en candidatos a instancias de la ontología. No obstante, en el segundo caso, la entidad se refiere anafóricamente a un elemento que ya se ha mencionado previamente en el texto y que coincide con el nombre de la clase "Hotel".

Extracción de valores de los atributos: Esta tarea es la que más se asemeja a la EI en el sentido descrito anteriormente como tarea que consiste en rellenar plantillas preestablecidas. Los atributos se refieren a datos concretos asociados a una instancia, y pueden ser desde números de teléfono, cantidades monetarias pasando por nombres de proteínas o cualquier otro elemento que aporte información relevante sobre una instancia incluida en la ontología.

Extracción de instancias de las relaciones: La extracción de relaciones entre instancias se refiere al proceso de conectar dos instancias a través de la relación que existe entre las clases a las que pertenecen y están descritas en la ontología.

A continuación se presentan las aproximaciones más importantes para la instanciación automática de ontologías.

4.9 Sistemas para la instanciación automática de ontologías

En los últimos años, la instanciación automática de ontologías a partir de texto en lenguaje natural se ha abordado desde diferentes perspectivas. La mayoría de ellas combinan técnicas de Procesamiento de Lenguaje Natural, tales como extracción

y reconocimiento de patrones lingüísticos, POS-tagger y análisis sintáctico, junto con otras técnicas de aprendizaje automático.

A continuación, se describen y comparan los principales sistemas que encontramos en la literatura:

- **Ontoshopie** (Celjuska & Vargas-Vera 2004). Es un sistema basado en aprendizaje automático en donde en primer lugar se identifican las entidades para decidir posteriormente si formarán parte de la ontología como instancias. El punto de partida es un corpus anotado con etiquetas XML donde cada entidad está asociada con la clase correspondiente en la ontología. Entonces, mediante un diccionario conceptual, se entrena el sistema y el resultado es una serie de reglas para la extracción de instancias y sus relaciones. Una vez validado por el usuario, los elementos aprendidos se insertan en la ontología.

- **PANKOW** (Cimiano et al. 2004). La metodología que se propone este sistema, se basa en patrones no-supervisados que son capaces de categorizar las etiquetas HTML de la Web como si se tratara de instancias de una ontología. Para aquellas etiquetas que contienen nombres propios, se crean una serie de expresiones regulares, basadas en los patrones propuestos por Hearst (1992). Con dichos patrones, se realizan búsquedas a través de Google para localizar nombres propios que se convierten en posibles candidatos a instancias de la ontología. El sistema categoriza los resultados y los nombres propios se asocian con la clase correspondiente en la ontología.

- **OntoPop** (Amardeilh et al. 2006)**.** Se basa en textos con anotaciones semánticas, es decir, en corpora anotados. Para cada etiqueta semántica, se crea un árbol conceptual, representado como un documento XML. Cada nodo del árbol se corresponde con una etiqueta semántica. Las

etiquetas semánticas pueden ser mapeadas en la ontología mediante un conjunto de reglas desarrolladas manualmente.

- **Navigli & Velardi** (2006). En esta propuesta, se expone una metodología que permite la anotación automática de las glosas de un tesauro con el objetivo de enriquecer una ontología del mismo dominio. Cada glosa se analiza morfológicamente, así mismo se le aplica un reconocedor de entidades nombradas. A continuación se define un conjunto de expresiones regulares para anotar ciertos fragmentos de la glosa con propiedades ontológicas (relaciones conceptuales). El resultado es un fragmento anotado en donde un cada par de términos se asocia mediante una relación ontológica. A estos términos los consideran el dominio y el rango de la relación ontológica. Tras un proceso de desambiguación, los términos elegidos se insertan en la ontología como individuos de los conceptos definidos por las glosas anotadas.

- **Magnini et al.** (2006). Este sistema, identifica las Menciones de Entidades Nombradas y les asigna los conceptos y las relaciones previamente definidas en la ontología. Para cada entidad se define una lista de atributos que se irán completando con las menciones extraídas. La clasificación se apoya en las características sintácticas de los términos.

- **Class-Example** (Tanev & Magnini, 2008). El punto de partida de este sistema semi-automático es un grupo de instancias de entidades nombradas pertenecientes a cada una de las clases que se pretenden poblar en la ontología y que sirven como conjunto de entrenamiento para el sistema. A continuación, un algoritmo basado en un vector de similitud se aplica a un corpus en el que aparece al menos mencionada

dos veces cada entidad de entrenamiento. Se obtienen las características sintácticas de cada ocurrencia en el corpus, de modo que, dichas características se utilizan para la creación de un nuevo vector de características. Éste se compara con los ya existentes, y entonces la nueva instancia se inserta en la clase más similar, es decir la clase cuyo vector de características es más similar.

- **OntoSyphon** (McDowell & Cafarella, 2008)**.** Dada una ontología con una clase raíz y algunos patrones lingüísticos simples, esta plataforma extrae de la web tantas instancias de los conceptos seleccionados y de las relaciones taxonómicas como es posible. La ontología, además de ser poblada, se usa también para verificar las instancias extraídas.

- **Giuliano & Gliozzo** (2008). Proponen una metodología no supervisada basada en sustitución de Entidades Nombradas. A partir de una ontología semi-poblada, buscan en la Web entidades que aparecen en contextos similares a las instancias de la ontología, apoyándose en la premisa de que las entidades que aparecen en contextos similares pertenecen a los mismos conceptos. Entonces, se seleccionan aquellas sustituciones que generan un elevado número de oraciones recuperadas de la web.

- **Danger & Berlanga** (2009). Esta metodología, que se basa en las estructura del documento (título, subtítulo, párrafos), extrae entidades que se consideran relevantes, apoyándose, para ello, en un lexicón predefinido que, a su vez, ha sido mapeado con las clases y relaciones de la ontología. Una vez completada esta fase, proponen un método heurístico que permite combinar las instancias. El resultado es un conjunto de instancias complejas, es decir, instancias que poseen más de una relación en la ontología.

- **SPRAT** (Maynard et al. 2009). Los autores usan tres tipos de patrones lingüísticos implementados en GATE con el fin de localizar Entidades Nombradas de los textos. Estas entidades se convierten en candidatos a instancias de la ontología.

- **AllRight** (Jannach et al., 2009). Este sistema extrae instancias a partir de datos semi-estructurados. Mediante un motor de búsqueda, localizan en la Web descripciones de productos cuyas características están expresadas de forma tabular. Para ello, utilizan un conjunto de palabras clave. Entonces, las páginas Web que describen el mismo producto, se agrupan y la información de las características de cada producto se utilizan para instanciar atributos ontológicos. Por lo tanto, este sistema solamente extrae instancias de los atributos de la ontología y no relaciones entre instancias.

- **BOEMIE** (Petasis et al., 2007). La meta de este sistema es la instanciación de una ontología a partir de fuentes multimedia mediante el uso de patrones. Una de las herramientas desarrollada en el marco de este proyecto es HMatch(I), con la que se pretenden identificar, por un lado, diferentes instancias que denotan el mismo objeto en el mundo real y, por otro, expresiones que hacen referencia a la misma instancia, solventando, de este modo, algunos de los problemas que ocurren durante el proceso de mapeo de instancias en la ontología.

4.10 Clasificación de los sistemas para la instanciación de ontologías

Siguiendo los parámetros que proponen en Petasis et al. (2007), se ha establecido una comparación entre los sistemas de población de ontologías descritos más arriba:

- **Requerimientos Iniciales.** En relación con los requerimientos previos, es decir los recursos o el conocimiento que son necesarios para que el sistema pueda funcionar, la mayoría de los sistemas descritos utilizan módulos para la Clasificación y Reconocimiento de Entidades Nombradas (Named Entity Recognition and Classification). Por ejemplo, en (Tanev & Magnini, 2006), el punto de partida es un conjunto de entrenamiento de instancias para cada clase de la ontología. SPRAT (Maynard et al. 2009) considera las Entidades Nombradas identificadas por GATE como candidatos a instancias de la Ontología. En (Magnini et al., 2006), el sistema identifica las menciones de entidades nombradas y las relaciona con los conceptos y relaciones ya definidos en la ontología. La metodología descrita en (Giuliano & Gliozzo, 2008) también se basa en sustitución de entidades nombradas.

  Otros sistemas además necesitan un corpus anotado de entrenamiento. Por ejemplo, el punto de partida de Ontoshopie (Celjuska & Vargas-Vera 2004) es un corpus con anotaciones XML donde cada entidad es asociada con la clase correspondiente en la ontología. En (Tanev & Magnini, 2008), un corpus analizado sintácticamente contiene las entidades de entrenamiento necesarias. (Navigli & Velardi, 2006) hacen uso de las definiciones de un glosario anotadas gramaticalmente y un conjunto de patrones lingüísticos definidos manualmente, y OntoPop (Amardeilh et al., 2006) utiliza anotaciones semánticas de textos. Finalmente, (McDowell & Cafarella, 2008) requiere de una ontología con una clase raíz y algunos patrones lingüísticos simples para extraer instancias de los conceptos y relaciones taxonómicas de la web.

- **Método de aprendizaje.** La mayoría de las metodologías utilizan, en mayor o menor medida, técnicas de aprendizaje automático para poblar la ontología. Por ejemplo, Ontosophie (Celjuska & Vargas-Vera 2004) se basa en un diccionario conceptual que genera reglas de extracción. Estas

reglas se utilizan para entrenar el sistema. En (Tanev & Magnini, 2008) se emplea un algoritmo no supervisado basado en un vector de similitudes. El algoritmo se aplica a un corpus anotado sintácticamente que contiene las entidades de entrenamiento. (Giuliano & Gliozzo, 2008) utilizan también un método de aprendizaje semisupervisado para extraer instancias de la Web.

Algunos sistemas hacen uso de patrones construidos manualmente, como es el caso de (Maynard et al. 2009; Amardeilh et al., 2006; McDowell & Cafarella, 2008), así como en (Navigli & Velardi, 2006), donde las glosas son anotadas con las propiedades de la ontología mediante un conjunto de expresiones regulares.

Finalmente (Magnini et al., 2006), establece un conjunto de referencia creado manualmente en el que las menciones de entidades nombradas son asignadas a conceptos y relaciones ya definidos en la ontología.

- **Grado de automatización.** Otro parámetro utilizado para clasificar los diversos métodos de población de ontologías es el grado de automatización. Algunos sistemas como (Maynard et al. 2009; Magnini et al., 2006; Giuliano & Gliozzo, 2008; Navigli & Velardi, 2006; McDowell & Cafarella, 2008), son no supervisados o débilmente supervisados (Tanev & Magnini, 2008), mientras que otros, tales como (Celjuska & Vargas-Vera 2004; Navigli & Velardi, 2006; Amardeilh et al., 2006), necesitan ser guiados por un experto.
- **Portabilidad a otros dominios.** Muchos de los sistemas han sido testados en colecciones de documentos de uno o varios dominios específicos. Por ejemplo, la metodología descrita en (Navigli & Velardi, 2006)se ha testeado ene el dominio del patrimonio cultural y su portabilidad requiere desarrollar nuevos patrones lingüísticos de acuerdo con el dominio y el lenguaje. Este también es el caso de (Maynard et al. 2009; Celjuska & Vargas-Vera 2004). Otros sistemas extraen entidades nombradas de tipo

genérico, como por ejemplo, personas o localizaciones (Tanev & Magnini, 2008; Magnini et al., 2006; Giuliano & Gliozzo, 2008; Amardeilh et al., 2006). Finalmente, Ontoshypon (McDowell & Cafarella, 2008) es una metodología independiente del dominio.

- **Mantenimiento de la consistencia.** No son muchos los sistemas que proveen información acerca de si se comprueba la consistencia de la ontología durante o al final del proceso de instanciación (Tanev & Magnini, 2008; Magnini et al., 2006; Giuliano & Gliozzo, 2008; Celjuska & Vargas-Vera 2004; Navigli & Velardi, 2006; McDowell & Cafarella, 2008). En (Amardeilh et al. 2006) se requiere un mantenimiento manual de las reglas de adquisición, aunque no utilizan ningún razonador para comprobar la consistencia de la ontología. Finalmente en (Maynard et al. 2009), el plugin de GATE utilizado para insertar las instancias en la ontología comprueba la consistencia de las mismas durante su inserción.

- **Desambiguación de Entidades.** Algunos sistemas llevan a cambo ciertas tareas de desambiguación durante el proceso de población de la ontología. Por ejemplo en (Celjuska & Vargas-Vera 2004), los valores de fiabilidad son asignados a las entidades extraída, en el caso de que sean ambiguas, seleccionan los valores con una mayor fiabilidad. Otros sistemas como (Giuliano & Gliozzo, 2008; Amardeilh et al. 2006), utilizan características contextuales para desambiguar. En (Maynard et al. 2009), a pesar de que el sistema puede detectar y avisar acerca de posibles ambigüedades, el proceso de desambiguación depende en gran mediad del usuario final. En (Tanev & Magnini, 2008), las entidades nombradas ambiguas se permiten dentro del corpus de entrenamiento. Sin embargo, si se encuentra una entidad nombrada ambigua durante la ejecución de proceso de instanciación, ésta no se incluye en la ontología. En (McDowell & Cafarella, 2008) se proponen algunas estrategias de desambiguación

basadas en la inclusión de información adicional durante la búsqueda de instancias en la Web.

Finalmente, otros sistemas aplican métodos de desambiguación durante el proceso, aunque no relacionados necesariamente con las entidades nombradas. Por ejemplo (Navigli & Velardi, 2006) aplica un algoritmo de desambiguación semántica basado en patrones estructurales de las glosas anotadas, y (Magnini et al., 2006) utiliza diferentes medidas de correferencia para gestionar el problemas de la ambigüedad de las menciones.

- **Dependencia del Lenguaje.** Casi todas las metodologías examinadas sólo dan soporte a recursos en lengua inglesa (Maynard et al. 2009; Giuliano & Gliozzo, 2008; Celjuska & Vargas-Vera 2004; Navigli & Velardi, 2006; McDowell & Cafarella, 2008). No obstante, el grado de dependencia del lenguaje es variable de acuerdo con la portabilidad de sus componentes lingüísticos.

  Es posibles distinguir entre sistemas fuertemente dependientes del lenguaje (tales como (Maynard et al. 2009; Celjuska & Vargas-Vera 2004; Navigli & Velardi, 2006) y sistemas que, aunque dependientes, la adaptación a otras lenguas no supondría un elevado coste de recursos, por ejemplo (Giuliano & Gliozzo, 2008; McDowell & Cafarella, 2008).

  Solamente (Magnini et al., 2006) y (Amardeilh et al. 2006) toman en consideración otras lenguas tales como el italiano y el francés respectivamente.

BLOQUE II

DESARROLLO DE LAS METODOLOGÍAS Y VALIDACIÓN

CAPÍTULO 5

INSTANCIACIÓN DE ONTOLOGÍAS BASADA EN LA DISTANCIA CO-TEXTUAL Y LA GANANCIA DE CONOCIMIENTO

APLICACIÓN AL DOMINIO DEL TURISMO.

Resumen. El estudio riguroso y sistemático de las características lingüísticas de los textos utilizados para (1) la extracción de las instancias y (2) el desarrollo de los componentes lingüísticos la metodología de instanciación, es el eje sobre el que se articula la primera parte de este capítulo. En primer lugar, se describen dos corpora de dominio turístico, para, en segundo lugar, proceder a su análisis desde un punto de vista de análisis del discurso. En la segunda parte del capítulo, se describe la metodología de instanciación de la ontología, basada en dos elementos, la distancia cotextual y la ganancia de conocimiento. Finalmente, se realiza una validación de la metodología propuesta mediante la instanciación de una ontología de dominio turístico. Los resultados de la evaluación son alentadores, corroborando la hipótesis de que un conocimiento exhaustivo de los textos de los que se extraen las instancias favorece la obtención de buenos resultados en cuanto a exhaustividad y precisión se refiere.

5.1 Introducción

El significado, la semántica, de una buena parte de los documentos que circulan por a red es desconocida para las máquinas, hecho que empobrece los resultados que le llegan a usuario final.

En el contexto de la Web Semántica, las ontologías constituyen una potente herramienta para el desarrollo de aplicaciones capaces de incorporar y representar

el conocimiento semántico. Por este motivo, el enriquecimiento de ontologías mediante la adquisición automática de conocimiento textual es una tarea fundamental que permite avanzar en la infraestructura necesaria para dar soporte a la web semántica.

En este capítulo, se presenta una metodología para la instanciación automática de ontologías a partir de textos en lenguaje natural. La validación se ha realizado en textos de dominio turístico.

La metodología se basa en dos aspectos fundamentales, la distancia cotextual, como elemento lingüístico, y la ganancia de conocimiento, como elemento ontológico.

El cotexto se refiere a qué elementos lingüísticos circundan un determinado fragmento de texto, mientras que la ganancia de conocimiento se utiliza durante el proceso de instanciación para determinar qué candidatos a instancias enriquecen en un grado mayor la ontología. Ambos conceptos se desarrollarán profusamente en los siguientes apartados.

Este capítulo se organiza como sigue. En primer lugar, se ha realizado un estudio lingüístico de dos corpora de dominio turístico siguiendo los parámetros de Bhatia (1993). De este modo, se ha obtenido la información pertinente tanto para el desarrollo de los recursos de PLN que forman parte de la metodología, como para la ontología que se va a instanciar.

En segundo lugar, se describe la metodología propuesta, que se divide en tres fases principales:

- Procesamiento de Lenguaje Natural y Procesamiento del Corpus.
- Reconocimiento de entidades nombradas.
- Instanciación de la Ontología y comprobación de la consistencia.

Finalmente, se valida la metodología, aplicándola a dos *corpora* de dominio turístico, y se presentan los resultados de evaluación.

5.2 Desarrollo de los recursos

Cualquier sistema para la población de ontologías prevé, al menos, la existencia de dos elementos básicos para su desarrollo: el corpus o documentos de los que se extraerán las instancias, y la ontología a instanciar.

En este apartado se presentan tanto los recursos lingüísticos como ontológicos necesarios para el desarrollo del sistema. Esto recursos consisten en:

- Dos *Corpora* representativos del dominio del turismo y que se han utilizado, por un lado, para el desarrollo de la ontología y de los componentes lingüísticos del sistema, y por otro para la extracción de instancias.
- La ontología de dominio que se pretende instanciar.

El texto no es sólo la fuente de datos de la que extraer las instancias de las ontología, sino que, en fases previas, el conocimiento de las características del mismo a todos los niveles lingüísticos, esto es, las características lexicográficas, morfológicas, sintácticas, semánticas y pragmáticas, nos dará una visión general de cómo se representa la información en un dominio concreto y permitirá adaptar las herramientas de PLN a las características textuales inferidas. En este caso, los textos seleccionados pertenecen al dominio del turismo.

5.2.1 Selección y análisis lingüístico de los *corpora*

En este apartado, se presenta el análisis discursivo de los textos , análisis que sustenta, por un lado, el desarrollo de la ontología, y por otro, el desarrollo de los componentes lingüísticos del sistema, a los que se hace mención más adelante.

La Web, como repositorio universal de información textual, se ha convertido en una atractiva posibilidad para la adquisición de *corpora* textuales y el análisis léxico a gran escala (Castano et al., 2008).

Un corpus es una colección de elementos lingüísticos, seleccionados y ordenados de acuerdo con criterios lingüísticos explícitos, con la finalidad de ser utilizada como muestra de la lengua (Santmaría et al., 2003).

Para la extracción de conocimiento a partir de texto en lenguaje natural es necesario disponer de un corpus lo suficientemente representativo del dominio. De este modo, los recursos lingüísticos, cuyo desarrollo se fundamenta en dicho corpus, se adecúan a las producciones textuales del dominio en cuestión. Además, la metodología empleada para la extracción de instancias se puede aplicar a otras producciones textuales pertenecientes al mismo dominio. Es decir, el corpus adquiere una doble funcionalidad. Por un lado, como punto de referencia, del que inferir cómo está expresada la información de interés de un dominio y por otra parte es la fuente de información a partir de la cual extraer las instancias de la ontología.

Como señala Alcántara Plá (2007), el corpus es el único medio con el que contamos en la actualidad para acercarnos a nuestro objeto de estudio de forma imparcial. Este mismo autor añade que el corpus ayuda sólo como referencia relativa, pero es fundamental para saber qué fenómenos son frecuentes y cuáles marginales, para así focalizar nuestros esfuerzos en los problemas adecuados.

El uso de un corpus es siempre una tarea de simplificación, no tenemos que considerar todos los datos siempre que consideremos los datos correctos y esenciales. Dado que no se puede considerar todo el contenido de la web, hay que seleccionar un conjunto de datos suficiente y fiable para llevar a cabo las tareas de creación de los recursos y, posteriormente, instanciación de la ontología.

La metodología se ha validado en el dominio del turismo y, aunque como se verá más adelante, la modularidad del sistema garantiza cierta independencia de la metodología con respecto al ámbito de aplicación, la ontología y los recursos

lingüísticos son específicos del dominio en cuestión. Los corpora textuales que se han compilado son el corpus *Hoteles* y el corpus *Restaurantes*.

El Corpus *Hoteles* está formado por la descripción en español de 112 hoteles (unas 13.000 palabras), extraídas de la página oficial de turismo de la Región de Murcia en el año 2008 (www.murciaturistica.com). De estas descripciones, 20 (el 18% del corpus aproximadamente) han sido empleadas para el análisis lingüístico, mientras que el resto se han utilizado para la evaluación del sistema.

Por otro lado, el Corpus *Restaurantes* está constituido por la descripción en español de 848 restaurantes (aproximadamente 67.500 palabras) extraídos de la página oficial de turismo de Murcia en el año 2009. De estas descripciones, 100 (el 13% del corpus aproximadamente) han sido empleadas para el análisis lingüístico y el resto para la evaluación.

El método desarrollado para la instanciación automática de ontologías necesita, además, de un conjunto de recursos lingüísticos. El conocimiento exhaustivo de los textos de los que se obtienen las instancias de la ontología, así como de su contexto, facilita la elaboración de dichos recursos, optimizando, de este modo, el proceso de extracción de información, no sólo para los corpora analizados, sino también para las producciones textuales pertenecientes al mismo género discursivo, incluso en distintos idiomas, como se verá más adelante.

Con el objetivo de determinar qué elementos lingüísticos y entidades de conocimiento son los más representativos del dominio y cuáles son las estructuras léxicas en las que se insertan, se ha llevado a cabo un análisis lingüístico de los dos corpora turísticos comentados anteriormente.

Los documentos escritos que están relacionados de manera directa con el turismo constituyen un material muy heterogéneo: guías turísticas, revistas de viajes, catálogos, anuncios publicitarios, folletos y materiales promocionales, documentos de viaje, diccionarios, etc. y el soporte utilizado para su difusión puede ser impreso o/y digital. En este último, entran en juego elementos propios

de la web, como por ejemplo la hipertextualidad, en donde los límites entre un documento y otro aparecen difuminados. En una misma página web, podemos reservar un viaje, consultar una guía sobre las zonas de interés o visualizar anuncios publicitarios. Junto a elementos puramente textuales encontramos elementos visuales que definen como rasgos paralingüísticos. Este tipo de información no textual puede dificultar el análisis.

Ambos *corpora* están constituidos por textos que se sitúan dentro del discurso de la promoción turística, al mismo tiempo que ofrecen información de interés para el viajero. El objetivo que persiguen es describir las características de los hoteles y restaurantes de forma positiva para que los clientes potenciales decidan si se ajustan o no a sus necesidades. Poseen, por tanto, un componente evaluativo que aparece, predominantemente, de forma implícita.

En nuestro caso, y siguiendo la clasificación que realiza Alcántara Plá (2007) sintetizando otras clasificaciones anteriores, se trata de *corpora* especializados, ya que los textos han sido elegidos porque poseen unas características específicas. Como se ha indicado son textos representativos del dominio del turismo, en concreto de la descripción de hoteles y restaurantes.

El grado de objetividad varía de unos a otros, y puede ir desde lo meramente informativo, mediante listas o iconos que enumeran los servicios de los que dispone el hotel, a descripciones subjetivas en donde se realzan distintas cualidades del establecimiento (y donde el componente evaluativo es más explícito).

Para el estudio, en cuestión se han seleccionado los textos de la página www.murciaturistica.com los cuales ofrecen producciones textuales más allá de lo puramente enumerativo (listas de servicios), incluyendo partes de discurso elaborado en donde las propiedades del hotel son descritas en lenguaje natural. Por otro lado, a pesar de su carácter publicitario, en muchos de ellos se intenta mantener una apariencia de objetividad que haga más creíble el mensaje.

En el corpus restaurantes, el número de descripciones es menor, mientras que se pueden encontrar más elementos en forma de lista, aunque los términos no están normalizados, es decir, para referirse a un mismo servicio no existe una terminología estándar.

La extracción de conocimiento de textos publicitarios es una tarea compleja, debido a que una de las principales características de dichos textos radica en la originalidad, y en general, para su interpretación es necesario un conocimiento del mundo muy amplio, difícil de recopilar en una ontología.

Como señalan (Bosch Abarca et al., 2005) en el lenguaje publicitario lo más interesante no es tanto ofrecer información sobre los atributos del producto o servicio, sino crear un mensaje sugerente y atractivo, cargado de recursos retóricos y emociones. Con respecto a los textos turísticos que encontramos en la web, estas autoras añaden que existe una dicotomía en lo referente a los registros, ya que hay un registro que se acerca a la oralidad, como por ejemplo las opiniones de los usuarios, y otros textos con un mayor grado de formalidad, muy próximos a géneros tradicionales.

En el caso que nos ocupa, los textos recopilados están más cercanos al registro formal que al informal.

En cuanto a los protocolos comunicativos utilizados en la web en el dominio publicitario del turismo, cabe distinguir entre protocolos comunicativos visuales y protocolos comunicativos textuales.

Los protocolos comunicativos visuales en los documentos seleccionados están constituidos en su mayoría por representaciones icónicas, como por ejemplo el icono de una silla de ruedas para indicar que es accesible para discapacitados físicos o el icono de un sobre en lugar de una dirección de correo electrónico. Aunque los iconos que aparecen junto a las descripciones de hoteles han sido recopilados en el corpus, no son tenidos en cuenta a la hora de realizar el análisis debido, por un lado, a la dificultad que supone procesarlos automáticamente

(aplicación de técnicas de reconocimiento de imagen) y, por otro, a que en este caso no aportan información relevante, puesto que esos mismos servicios ya están descritos textualmente.

Por otro lado, los rasgos textuales pretenden atraer la atención sobre la oferta, generar interés e incentivar y mover a la acción, son similares al discurso persuasivo. La diferencia es, que están regidos por las leyes de la eficacia y la economía informativa (textos breves) y suelen ir dirigidos a la satisfacción inmediata de los deseos de compra de los consumidores.

En el caso de los hoteles, al tratarse de descripciones completas (incluye tanto listas de servicios como descripciones textuales), el proceso utilizado para la extracción de conocimiento es aplicable a gran parte de las descripciones de hoteles que encontramos en la web.

Por tanto, las delimitaciones que se han establecido para el corpus seleccionado son de dos tipos, por un lado, se analizan solamente elementos textuales y no las imágenes o listas de iconos, y por otro, no se analizan las opiniones de usuarios o incluso de profesionales evaluadores del hotel (opiniones como las que pueden aparecer en las guías de viaje).

En este caso, el corpus será el modelo empleado para la elaboración de reglas que permitan extraer conocimiento del texto, es decir, es un corpus con carácter modélico del que se presupone que tiene unas características estructurales y léxicas similares a otros textos del mismo tipo y dominio. Pero, por otro lado, la información obtenida de dicho corpus, las instancias, varían con respecto a la información que se pueda extraer de otros corpus, ya que, como es lógico, los distintos textos sobre turismo aportan información diferente. El corpus se usa a dos niveles: en un primer nivel, es utilizado como modelo o representación de todos los documentos similares del dominio, y a otro nivel, el de la instanciación de la ontología, se utiliza como fuente para la extracción de instancias.

La motivación a la hora de seleccionar la temática de los hoteles deriva del hecho de que el alojamiento es una de las necesidades básicas del turista, y en consecuencia la información hotelera será con frecuencia solicitada por el viajero, como ya se ha señalado. Por otro lado, a nivel lingüístico suponen un reto interesante, ya que el hotel está relacionado con otras entidades turísticas como los monumentos, los sitios de interés o la restauración, lo que puede dar lugar a un enriquecimiento cualitativo de la ontología.

En cuanto al corpus Restaurantes, la restauración es un elemento que a menudo aparece ligado al alojamiento y aunque el número de recursos web dedicados a la descripción de restaurantes es menor que los dedicados a los hoteles y la descripción de los mismos es más heterogénea, existen elementos clave que se pueden sistematizar y, en consecuencia, se pueden añadir a la ontología.

Cabría hacer un pequeño apunte en cuanto a la redacción de estos textos, y es que el autor de los mismos es desconocido, siendo probablemente el propio personal del hotel en algunos casos, de manera que la veracidad de la información no está garantizada, por lo que en este caso se cuenta con la veracidad de la fuente, al tratarse de un organismo oficial. Por otro lado, este hecho no afecta a las técnicas empleadas para la extracción de conocimiento, objeto fundamental de este trabajo, a pesar de que la validez de las instancias en la ontología sí pueda quedar afectada. En cualquier caso, para la creación automática de ontologías o para su instanciación, es necesario asumir que la información textual puede estar sesgada, ser parcial o incluso no ser verdadera. La fuente de la que se obtiene la información debe ser fiable si finalmente se pretende que la ontología se ajuste lo más posible a la realidad a la que responde.

Por último, hay que añadir que en los textos electrónicos la presencia de errores ortográficos, gramaticales y erratas es superior, en general, a los errores de este tipo que encontramos en textos impresos. De nuevo, la calidad de la fuente es determinante para que este fenómeno no se convierta en un obstáculo para el

análisis. Al tratarse de un corpus de entrenamiento, se ha creído conveniente modificar las erratas y errores gramaticales que el corpus contenía.

5.2.2 El lenguaje del turismo como lenguaje de especialidad o lenguaje para fines específicos

Una visión global del discurso turístico se ofrece en *Lengua y comunicación en el español del turismo* (Calvi, 2005), en donde, desde una perspectiva más abarcadora que en el presente trabajo, en el sentido de que engloban todas las producciones textuales del domino del turismo, se realiza un análisis de las mismas, tomando en consideración guías turísticas, catálogos de reservas, folletos, así como los distintos soportes en los que se pueden presentar, es decir, digital e impreso. El trabajo mencionado se centra en los mecanismos de formación de palabras y algunos rasgos propios de la deixis espacio-temporal y personal, junto con el análisis de las formas de cortesía.

El lenguaje del turismo ha sido considerado por varios autores como un lenguaje para fines específicos (LFE) entre ellos Calvi (2005). Aun teniendo en cuenta la heterogeneidad de este tipo de lenguaje, en él se mezclan componentes de diversas áreas temáticas como economía, geografía, historia del arte, etc. y que el léxico propiamente turístico no es muy abundante, se podría clasificar como un lenguaje para fines específicos

Según esta autora, existen dos dimensiones de los LFE:

- Una dimensión horizontal, que se relaciona con un componente temático, es decir, con el conocimiento compartido por los expertos en el dominio, lo que incide en el componente léxico dando lugar a la monosemia y, en algunos casos, a la monoreferencialidad.
- Y una dimensión vertical, que añade un componente sociológico y pragmático que se puede clasificar según los distintos niveles comunicativos. El grado de densidad conceptual y la terminología

específica va descendiendo en función de si la comunicación se produce entre un especialista y otro especialista; un especialista y especialistas en formación, y finalmente cuando la comunicación es divulgativa.

El lenguaje del turismo también es considerado un lenguaje de especialidad por Alcalá Varó et al. (2006). No obstante, realizan varias consideraciones en cuanto a lenguas de especialidad se refiere. En este sentido, los autores consideran que las lenguas de especialidad son un sublenguaje de la lengua general, en donde un lenguaje especializado es una variedad producto de una selección y combinación concreta de determinados medios lingüísticos.

Esta consideración implica que, a pesar de la dificultad en la obtención de conocimiento de un texto en el que confluyen rasgos propios de la publicidad, como metáforas evocadoras, elementos subjetivos con apariencia de objetividad, ambigüedades léxicas introducidas por cuantificadores de límites difusos o subjetivos como *cerca*, *lejos*, o *un poco*, es posible identificar elementos y características que sistemáticamente se van a repetir en los distintos textos y que van a posibilitar la aplicación de reglas generales para el procesamiento del discurso.

Los corpora analizados se enmarcan dentro de la comunicación publicitaria, en concreto en la modalidad de actividad promocional, que como señala Calvi (2006) tiene el objetivo de configurar la imagen turística de un lugar, destacando sus atractivos en función de las distintas tipologías de turistas y sus necesidades.

El léxico que encontramos contiene términos específicos del lenguaje turismo junto con otros términos, la mayoría, del lenguaje común.

Aunque el análisis desde el punto de vista publicitario es interesante, en este caso no se hace hincapié en este carácter promocional, es decir, se parte del supuesto de que en los documentos aparecen una serie de datos que son susceptibles de clasificación, obviando el entorno publicitario en el que se encuentran. Por ejemplo, si el hotel dispone de vistas pintorescas es posible que

esto se destaque en la descripción, mientras que si no dispone de algún servicio esto no se sitúa en el lugar destacado, pero al fin y al cabo todo son datos con los que reconstruir la descripción del hotel en la ontología.

5.2.3 Análisis del discurso según los parámetros de Bhatia

En nuestro caso, para la realización del análisis lingüístico se seguirán fundamentalmente las directrices indicadas por Bhatia en su trabajo *From description to explanation in discourse analysis* (1993).

Para el estudio del corpus, se han numerado las descripciones de los hoteles como H1, H2, H3, etc. y las de los restaurantes como R1, R2, R3, etc. En este apartado y los siguientes nos referimos a los hoteles y restaurantes en estos términos para ejemplificar el análisis llevado a cabo. La oración o fragmento citado en cada ejemplo se ha incluido en el momento en el que se hace alusión a dicho texto, ya que la publicación del corpus en su totalidad no es posible en este trabajo debido a los derechos de autor.

Como se ha mencionado previamente, el objetivo de este estudio es determinar cuáles son las características lingüísticas relevantes que serán utilizadas en el desarrollo tanto de los componentes lingüísticos del sistema como de la propia ontología.

Según Bhatia (1993), *el texto puede ser analizado cuantitativamente mediante el estudio de las características específicas de cada lengua que están predominantemente usadas según la variedad a la que el texto pertenezca.*

Este tipo de análisis de carácter estadístico aporta información sobre el predominio de determinados elementos léxicos (repetición de términos o de estructuras, y concordancia de los términos). Por otro lado, si se ha llevado a cabo un análisis morfosintáctico, el análisis estadístico puede aportar información sobre qué tiempos verbales son los más utilizados, si la frecuencia de aparición de

sustantivos es alta o si las oraciones subordinadas son más abundantes que las coordinadas, por ejemplo.

En el análisis del discurso, todos los elementos están imbricados. Por ejemplo, la superestructura se determina mediante la observación de los constituyentes temáticos y organizativos del corpus a nivel global. En este caso, se puede observar que, generalmente, en el primer párrafo se procede a la descripción del hotel indicando la ubicación del mismo, pero para esto se ha observado que existen elementos léxicos, como verbos y preposiciones que indican situación. Entonces, la observación de elementos morfoléxicos, sintácticos, semánticos y pragmáticos va a indicar la presencia o no de fenómenos textuales característicos del dominio.

El análisis estadístico puede aportar información relevante sobre el texto. De hecho, las primeras aproximaciones a los textos en soporte digital que se llevaron a cabo, uniendo tecnologías informáticas y lingüísticas, fueron en este sentido, aunque, como se precisó en su momento, los datos numéricos sin interpretación y contextualización arrojan poca información. Por este motivo, es necesario interpretarlos adoptando un enfoque concreto. En este caso, como se ha indicado previamente, el punto de vista desde el que se analiza el texto es lingüístico, y poniendo de manifiesto aspectos que se consideran interesantes para la extracción de conocimiento.

Por otro lado, la estadística puede confirmar o desmentir algunas de las apreciaciones que se han obtenido *a priori* sobre las características del corpus, aportando datos cuantificables al respecto.

En la siguiente tabla (tabla 5.1), se muestran algunos de los datos numéricos relacionados con el tamaño y diversidad los corpora. Para la obtención de estos datos, no han sido contabilizados elementos textuales repetitivos de tipo estructural, es decir, que se encuentran de manera predeterminada en la ficha de descripción y que afectan fundamentalmente al corpus Restaurantes, por ejemplo

"jefe de cocina" "tipo de cocina" "recomendaciones" etc. Son elementos que se repiten en cada una de las descripciones.

Tabla 5.1 Datos numéricos del corpus Hoteles.

	Corpus Restaurantes	Corpus Hoteles
Número de palabras	55123	13980
Número de palabras distintas	7578	2540
Media de palabras por oración	10.36	13.42

La primera aproximación a un corpus de estudio es durante la etapa de recopilación. A partir de ese momento, se van observando una serie de fenómenos, de rasgos, propios del dominio y de la tipología textual correspondiente. Durante este proceso, si el corpus posee una estructura más o menos fija, como es el caso, y un núcleo léxico poco variado, se ponen de manifiesto ciertas características que se repiten y que en un análisis más riguroso podrán corroborarse, así como explicar qué las origina.

Estas características que se han observado *a priori*, fruto del proceso de recogida del corpus y de su lectura, las llamaremos **características superficiales**. Algunas son datos empíricos fácilmente comprobables tras un análisis cuantitativo o superficial, otras requieren un mayor grado de abstracción y un análisis exhaustivo.

Las **características superficiales** del corpus se pueden dividir en dos partes, las relacionadas con la macroestructura textual y las relacionadas con los elementos gramaticales. A continuación, se explican cada una de estas características para cada corpus.

Características relacionadas con la macroestructura en el Corpus Hoteles.

1. El corpus está dividido en subtextos constituidos por la descripción de cada hotel. Esto le confiere unas limitaciones claras a la hora de obtener

información puesto las características de un hotel comienzan en el momento que se menciona y acaban cuando se nombra un hotel distinto.

2. Cada subtexto está formado por entre uno y tres párrafos.
3. La descripción del hotel comienza siempre con el nombre del hotel y la dirección de donde se encuentra, seguida del teléfono y/o fax y la página web en el caso de tenerla o estar recogida en la base de datos.
4. El primer párrafo está dedicado a la descripción del hotel ensalzando o simplemente aportando datos sobre su ubicación y servicios destacados.
5. El hotel es ubicado en la descripción con respecto a otros edificios o lugares de interés, por este motivo es frecuente encontrar en el primer párrafo expresiones del tipo: *localizado en, ubicado en, cerca de, a poca distancia de*, etc.
6. Si el hotel ha sufrido alguna reforma o mejora, se menciona en al principio.
7. Si el hotel cuenta con SPA o con Balneario, es también puesto de manifiesto al inicio de la descripción.
8. Los servicios de los que dispone son enumerados en una lista en el último párrafo, la terminología que se utiliza para referirse a un mismo servicio es en la mayoría de los casos la misma.
9. La descripción de las habitaciones es un elemento destacable dentro del corpus, pudiendo mencionarse aspectos como el estilo arquitectónico y decorativo, o los servicio de los que dispone.

Características relacionadas con los elementos gramaticales en el Corpus Hoteles.

1. El uso de verbos es escaso y predominan los participios con función de adjetivo.
2. El uso de sustantivos y adjetivos es superior al uso de verbos. (Estilo nominal)

3. Es frecuente el uso de adjetivos valorativos positivos.
4. La persona verbal que aparece suele ser la tercera del singular referida al hotel.
5. En ocasiones encontramos la segunda persona del plural cuando el sujeto es el personal del hotel.

Características relacionadas con la macroestructura en el Corpus Restaurantes

1. Cada restaurante constituye un registro independiente del corpus, por lo tanto, los límites entre una entrada y otra están claros.
2. La descripción sigue una estructura preestablecida constituida por una serie de apartados fijos y otros variables en los que se incluyen los datos pertinentes.
3. En el primer párrafo se incluye en nombre del restaurante, la dirección, el teléfono y fax, la página web y la dirección de correo electrónico.
4. A continuación, la descripción se estructura en una serie de apartados preestablecidos que en el caso de aparecer lo hacen siempre en el mismo orden: Jefe de cocina, Jefe de sala, Tipo de cocina, Descripción, Recomendaciones, Postres, Especialidades culinarias, Otras recomendaciones, Tarjetas de crédito, Cierre semanal, Cierre vacacional, Precio menú, Plazas restaurante, Recetas disponibles y Aparcamiento.
5. Finalmente, en algunos casos se incluye una descripción adicional que no se corresponde con ninguno de los apartados.
6. En el apartado Descripción, como su nombre indica, se incluye una descripción en lenguaje natural de elementos que se han considerado relevantes.
7. Las enumeraciones de alimentos y platos cocinados son frecuentes, y no existe una normalización terminológica al respecto.

Características relacionadas con los elementos gramaticales en el Corpus Restaurantes.

1. La marcada estructura de las descripciones condiciona el lenguaje utilizado en cada apartado, siendo predominante el estilo nominal dado que la presencia de enumeraciones es muy alta.
2. En las descripciones que no son meras listas, se utiliza un estilo directo y conciso en donde la subordinación es muy escasa, siendo habitual la yuxtaposición.
3. Es frecuente el uso de adjetivos valorativos positivos.

Estas características superficiales observadas en los corpora y que son consideradas de interés para el procesamiento automático, extracción de conocimiento y población de la ontología, son estudiadas más en profundidad a continuación.

Con este objetivo, se han establecido tres niveles de análisis que siguen el modelo de Bhatia (1993) y que van desde las características léxico gramaticales, pasando por los patrones textuales para terminar con la interpretación estructural del género. Se ha hecho un mayor hincapié en el corpus Hoteles, que no tiene un carácter tan estructurado como el corpus de restaurantes. A continuación se describen los distintos niveles de análisis.

Nivel 1. Análisis de las características léxico gramaticales.

En esta fase, se procedió al análisis morfológico y estadístico del corpus, para ello se utilizaron el software de libre distribución Freeling 2.0 (Freeling, 2008) que ha sido incluido en GATE y la librería nltk[18] integrada en Phyton[19].

[18] http://www.nltk.org/

[19] http://www.python.org/

Para el análisis estadístico del corpus restaurantes, se han eliminado los elementos estructurales repetitivos a los que se ha hecho alusión previamente.

La frecuencia de los términos de un texto puede arrojar algunos datos significativos para el estudio, aunque sin otro tipo de análisis como por ejemplo morfológico, de "colocaciones", etc. puede aportar información que induzca a error.

Tabla 5.2 Las 10 palabras más frecuentes del corpus sin incluir stopwords.

Corpus Restaurantes		**Corpus Hoteles**	
restaurante	943	habitaciones	572
Murcia	253	968	200
euros	252	hotel	189
caseros	221	zonas	181
Cartagena	212	comunes	177
carnes	209	aire	145
pescados	205	acondicionado	144
mediterránea	189	plazas	130
murciana	183	ascensor	123
arroces	171	individual	113

Por ejemplo, en la tabla 5.2 se muestran las 10 palabras más frecuentes, sin incluir palabras vacías, aunque sí han incluido los números. De ella se puede inferir cierta información contextualizada por el conocimiento de ambos corpora que ya se tiene. Sin este conocimiento previo los datos podrían interpretarse de forma errónea, ya que no necesariamente las palabras más abundantes son las más representativas o las de más utilidad para la extracción de conocimiento, ni por el contrario, tampoco es correcto suponer que una palabra muy frecuente en el corpus no aporta información de interés, siempre y cuando no sea una palabra vacía considerada de forma aislada.

Las palabras más representativas de un documento son aquellas que tienen frecuencias intermedias. Sin embargo, en este caso, los términos con una frecuencia alta son indicadores de los principales aspectos que se quieren destacar.

Las palabras más frecuentes generalmente están distribuidas en el corpus de manera uniforme, lo que puede significar que constituyen núcleos temáticos recurrentes que llevan asociados un campo léxico del que son modificadoras o modificadas.

Las palabras más frecuentes son "restaurante" y "habitaciones", respectivamente. En el primer caso, no aporta demasiada información. Sin embargo, en el segundo puede ser un indicativo de que la descripción del hotel se va a centrar en las habitaciones y en los servicios que se ofrecen. Por tanto, en torno al término habitaciones aparecerán de forma recurrente una serie de palabras en su cotexto más inmediato.

Sin embargo, el siguiente token más frecuente es 968, es decir, una cifra, que sólo adquiere sentido cuando se contextualiza. Por tratarse de un corpus de hoteles de la Región de Murcia, podemos saber que 968 es el prefijo telefónico de dicha comunidad.

Este dato es importante para la extracción de conocimiento, puesto que es indicativo de que en el corpus van a aparecer números de teléfono y fax, y que esta información estará relacionada con el hotel. Por lo tanto, habrá que tenerlo en cuenta en la metodología para la extracción de conocimiento.

Entre las palabras más frecuentes, encontramos, por un lado, "zonas" y "comunes" y, por otro lado, "aire" y "acondicionado". Para un humano, no es difícil suponer que estas palabras son constituyentes de un término multipalabra, es decir, que en el texto lo que encontraremos será "zonas comunes" y "aire acondicionado". La identificación de los términos multipalabra y de las Entidades Nombradas es uno de los puntos clave para el procesamiento de lenguaje natural, existiendo diversas herramientas que pueden ayudar a su identificación.

Con el objetivo de combinar el análisis morfológico y el estadístico se ha desarrollado un plug-in de GATE denominado STEX, el cual contabiliza la frecuencia de aparición de las palabras en el corpus, en combinación con el

análisis morfológico realizado por Freeling. De este modo, es posible saber si por ejemplo la palabra "juego" es un sustantivo o es un verbo y qué número de veces aparece en el texto como perteneciente a una u otra categoría. Por otro lado, mediante STEX se determinan las colocaciones, es decir la tendencia que tienen algunas palabras a coocurrir más frecuentemente con otras.

Esta aproximación cuantitativa se integra con la visión cualitativa que aporta el análisis del discurso, obteniendo así una perspectiva global a la vez que precisa de las características textuales. En la tabla 5.3 se puede ver qué tiempos verbales son los de mayor frecuencia en el corpus hoteles.

Tabla 5.3 Análisis de los tiempos verbales y la presencia de verbos en el corpus hoteles. (Realizado con STEX).

Uso de los verbos en el corpus	
Porcentaje de verbos en el texto	3,2%
Verbos en presente de indicativo	30%
Participios verbales	24%
Verbos en infinitivo	18%
Verbos en 3ª persona	42,55%

La presencia de verbos en el texto no es muy elevada, se opta por el uso de frases nominales en donde en ocasiones encontramos formas no personales del verbo, como el participio (se entiende que el participio se inserta dentro de una construcción pasiva en tercera persona en la que el verbo "estar" se ha eludido). Por ejemplo, "Situado en la principal avenida de la ciudad" H8, "Diseñado especialmente para las familias" H12. La presencia de este tipo de formas no personales y de verbos y construcciones impersonales como "hay", "se puede disfrutar", "se pueden practicar" le confieren objetividad al texto, y son utilizadas a la hora de hablar sobre las características del hotel, y dado que este es el objetivo fundamental del texto, estas formas son las más abundantes.

En el corpus Restaurantes, la presencia de formas verbales es mucho menor, y entre las que encontramos, las formas en tercera persona haciendo referencia a restaurante son las más frecuentes "ofrece una variada carta de vinos" R29, "dispone así mismo de pescados" R34. Encontramos también verbos impersonales "se realizan platos de comida regional" R37 y formas no personales "Situado en primera línea de mar" R6.

El presente simple de indicativo en forma activa es el tiempo verbal más utilizado en el corpus Hoteles, se trata de un rasgo propio de la descripción, ya que presenta las características de un elemento de forma "atemporal" y como un hecho de verdad. Por ejemplo, "Cuenta con" H10, "le ofrece" H11 "dispone de" H9. Las variaciones hacia tiempos de pasado se suelen utilizar casi exclusivamente para referirse a las reformas que ha sufrido el hotel, el número de veces que aparece es tan bajo que realmente es irrelevante.

Cabe destacar también el número de construcciones impersonales con "se", que como se ha señalado confieren objetividad al texto y suelen ser utilizadas o bien porque el hotel no se considera un sujeto agente, o bien porque el sujeto serían los posibles clientes y se elude la personalización. También podría establecerse una relación entre la presencia de estas formas y la veracidad que pretende transmitir el texto. Por ejemplo, "Se realizan actividades" H7, "se puede disfrutar del almuerzo"H17.

En cuanto al tipo de oraciones que encontramos, son en su mayoría oraciones simples. Por ejemplo, "En temporada alta se puede disfrutar del Almuerzo temático una vez por semana sin suplemento adicional" H17, "El hotel goza de una ubicación privilegiada en el principio de La Manga, en primera línea de playa del Mar Mediterráneo" H11. El corpus Restaurantes se limita casi exclusivamente a las oraciones simples que le confieren un carácter telegráfico y enumerativo. "Excelentes materias primas y cocina creativa bien elaborada, con cambios frecuentes en la carta. Sugerencias diarias según mercado." R51

Las oraciones subordinadas más comunes son las de relativo. Por ejemplo, "En el corazón de La Manga Club, un resort deportivo y de ocio único, con una extensión de 6 km cuadrados, que incluyen tres de los mejores campos de golf europeos" H1, "El NH Cartagena está situado en el centro neurálgico, comercial y de negocios de Cartagena, junto al ayuntamiento, con un magnífica ubicación que le permitirá disfrutar de toda la personalidad de esta bella ciudad" H16.

La baja presencia de oraciones subordinadas se debe a la brevedad y concisión del mensaje que se quiere transmitir, y a que el contenido textual está muy cercano a la enumeración de características, motivo por el cual son frecuentes las oraciones yuxtapuestas. "Situado sobre la playa del Mediterráneo denominada *Playa Galúa* y a 300 m del centro comercial de la Manga. A 100 m de la Plaza Bohemia. Existen habitaciones y zonas del restaurante buffet para no-fumadores" H17.

Nivel 2. Análisis de los patrones textuales o textualización.

Como señala Bhatia (1993), existe un tipo de anuncios publicitarios en donde abundan los sintagmas nominales. Eso es debido a que en los anuncios se trata de describir previamente el producto, y la clase morfológica que se emplea habitualmente para ello es el adjetivo. Y puesto que el adjetivo suele estar inserto en un Sintagma Nominal, estas son las formas más abundantes en el texto. Por ejemplo, en H16 podemos detectar los siguientes sintagmas nominales:

NH ha desarrollado una gama de hoteles con una personalidad extraordinariamente singular: los hoteles Collection. Combinan la magia de su entorno con la belleza de su emblemático edificio. La historia y la modernidad conviven en armonía para ofrecer el mejor recuerdo al cliente. El NH Cartagena está situado en el centro neurálgico, comercial y de negocios de Cartagena, junto al ayuntamiento, con una magnífica ubicación que le permitirá disfrutar de toda la personalidad de esta bella ciudad. Combinando con perfecta armonía el diseño con el equipamiento más actual y completo, en el NH Cartagena está todo preparado para garantizarle una confortable estancia. Unas cuidadas instalaciones y un servicio de exquisita calidad son su mejor tarjeta de presentación.

En el texto seleccionado, aparecen una serie de sintagmas nominales (resaltados en gris) en los que se han incluido adjetivos que, en su mayoría, poseen connotaciones semánticas positivas, frente al escaso número de verbos. En realidad, la información que aportan estas valoraciones es escasa, pero contribuyen a formar una idea positiva del hotel en el cliente.

No obstante, como se ha señalado anteriormente, a pesar de que el número de verbos no es elevado, son el elemento vertebrador del texto, ya que no sólo serán un indicativo de las relaciones que se establecen entre las distintas entidades identificadas, sino que, debido a la concordancia sujeto-verbo, la persona gramatical va a facilitar la identificación de las distintas partes que conforman el texto.

Por un lado, los verbos en 3ª persona del singular se refieren habitualmente a las propiedades del hotel en general. Por ejemplo, "Cuenta con", "Dispone de". Y los verbos en tercera persona del plural se suelen referir a las características de los elementos del hotel "todos nuestros salones y terrazas exteriores ofrecen buenas vistas panorámicas" H13, "Todas las habitaciones disponen de aire acondicionado" H1.

Por otro lado, vemos en algunos textos un cambio hacia la primera persona del plural, refiriéndose también a las características del hotel, pero en este caso el sujeto del verbo sería el personal del hotel. Mediante este uso personal, se pretende un acercamiento al cliente. Por ejemplo, "En él reflejamos nuestro compromiso" H9

El interlocutor o la persona a la que va dirigido el mensaje puede ser evocado de forma neutra, como por ejemplo, "viajero", "visitante", permitiendo que el interlocutor se identifique con esas figuras ("ofrece a sus visitantes habitaciones climatizadas (…)" H15).

Es interesante también observar cómo la semántica de los verbos nos permite distinguir a qué partes del hotel se está refiriendo la descripción. Por ejemplo,

"decorar", "diseñar" y "renovar" se suelen referir a las características de las habitaciones ("Sus habitaciones han sido totalmente renovadas" H20)

Los verbos como "servir" o "abrir" se refieren a las características del bar-restaurante.

Un elemento importante en la descripción hotelera es la situación o localización del hotel con respecto al entorno que lo rodea.

La localización en los *corpora* analizados viene indicada, por un lado, por la dirección oficial y, por otro por, la descripción que se hace destacando aspectos considerados de interés turístico. Es lo que se puede llamar "localización turística". En la localización turística, se realzan o mencionan aspectos que pueden ser un reclamo turístico y posicionan al elemento descrito, en este caso el hotel o el restaurante, con respecto a otras construcciones, paisajes, entornos que generalmente se encuentran próximos. La mayoría de estos elementos se refieren a playas, lugares de ocio o monumentos que, en fases posteriores, se clasifica aquí como entidades nombradas.

En el texto, existe un amplio abanico de marcadores que indican localización y que incluyen verbos de situación, preposiciones, locuciones prepositivas y un número restringido de sustantivos y adjetivos generalmente derivados de los verbos de localización.

El uso de estas formas locativas, es un factor importante en el caso que aquí se analiza, ya que determina la ubicación del hotel y del restaurante con respecto a otras infraestructuras o lugares de interés para el viajero, siendo la ubicación de los establecimientos hoteleros y de restauración un valor añadido para el visitante.

En la siguiente tabla (tabla 5.4), se pueden ver los principales verbos y locuciones verbales que indican localización en el texto.

Tabla 5.4 Verbos de localización en el corpus Hoteles

VERBOS DE LOCALIZACIÓN DEL CORPUS "HOTELES"
Localizar : se localiza, está localizado
Hallar: se halla
Enclavar: está enclavado
Encontrar: se encuentra
Asentar: se asienta, está asentado
Estar: Está en
Rodear: Rodeado por, rodeado de

Por otro lado, en la tabla 5.5 se muestran las variantes extraídas del corpus para la indicación de la localización de los hoteles y restaurantes. Estas expresiones pueden aparecer precedidas por un verbo o perífrasis verbal que indique localización seguido de una preposición o locución preposicional. En algunos casos, se prescinde del verbo para indicar la localización. Los elementos entre paréntesis son opcionales y los elementos entre corchetes indican datos variables.

Tabla 5.5 La indicación de la localización en el corpus.

VERBO DE LOCALIZACIÓN	PREPOSICIONES/LOCUCIONES PREPOSICIONALES CERCA	PREPOSICIONES/LOCUCIONES PREPOSICIONALES LEJOS
(en)	frente a (el)	Lejos de
	Junto a (el)	(más) alejado de
	A menos de + [distancia |tiempo]	
	Antes de	
	Al lado de	
	A la entrada de	
	a ((tan) sólo)(unos) [distancia | tiempo] de	
	a escasos [distancia|tiempo] de	
	A orillas de	
	Ser el más próximo a	

Las expresiones que indican cercanía, como "frente a" "junto a" "al lado de", etc., se suelen utilizar en sistemas basados en patrones para la extracción de información textual. Por ejemplo, se puede traducir en la relación ontológica "CercaDe" contenida en la ontología de hoteles y restaurantes que se ha desarrollado para la validación del sistema

Existen otro tipo de expresiones, que pueden aparecer en torno a las mostradas en la tabla anterior y que las modifican. Los modificadores más frecuentes se muestran en la siguiente tabla (Tabla 5.6).

Tabla 5.6 Modificadores de localización.

Modificador1	**Modificador2**
(pleno) (corazón de |(centro (neurálgico)de))	cercanías de principal acceso directo por proximidades de alrededores de orillas de con vistas a

Uno de los modificadores más comunes en el corpus es "en el corazón de" o "en el centro neurálgico". Este tipo de estructuras lingüísticas se puede traducir en el atributo ontológico "céntrico". No obstante, la sistematización de las mismas es compleja, ya que implican, al menos, dos entidades: por un lado, la entidad de la que se indica la situación, y, por otro, la entidad con respecto a la que se sitúa. Por ejemplo, *X está a orillas de Y.*

Nivel 3. Interpretación estructural del género textual.

La macroestructura discursiva de ambos corpora se repite recursivamente en cada una de las descripciones, posicionando las partes del discurso en lugares más o menos fijos, sobre todo en el corpus Restaurantes. Esta estructura permite al consumidor identificar rápidamente las distintas partes y es una mezcla entre la información que supuestamente el usuario desea conocer de un determinado

establecimiento, y la información que al establecimiento le interesa resaltar en su favor. Es decir, que, en realidad, no se van a ofrecer todos los datos relevantes para el usuario, sino sólo aquellos que benefician la imagen del hotel o del restaurante. De este modo, en algunas descripciones, se hace hincapié por ejemplo en que el hotel cuenta con balneario, en detrimento de otros aspectos valorados de forma positiva en otros casos como la localización.

Las descripciones de los hoteles del corpus están divididas en tres secciones que se comentan a continuación.

En primer lugar, aparece el nombre del hotel junto con la localización y la información de contacto (dirección, teléfono, fax, y página web). Se trata de información práctica a la que el usuario puede acceder de forma directa.

En segundo lugar, las características del hotel se describen en lenguaje natural, poniendo de relieve aquellos aspectos que pueden resultar atractivos para el usuario, tales como la ubicación del hotel ("en el centro de la ciudad", "cerca de la playa"), si cuenta con campo de golf o SPA, si ha sido recientemente reformado o, incluso, cuál es el estilo arquitectónico o decorativo de las habitaciones. Es en esta parte donde la extracción de las instancias es más compleja, ya que, al tratarse de textos promocionales, y a pesar del intento por mantener una apariencia de objetividad, se ensalzan las características positivas de los hoteles, poniendo de relieve aquellos elementos que despiertan el interés de los consumidores potenciales.

Algunos de los aspectos más recurrentes en esta parte del texto son:

- Localización con respecto al entorno: "El complejo está enclavado entre limoneros, buganvillas y palmeras convirtiendo este verde paisaje en uno de los más espléndidos de Europa" H1.
- Estilo arquitectónico o decoración de las habitaciones. "Las habitaciones destacan por su amplitud y cuidada decoración que combina el estilo mediterráneo con detalles muy actuales" H2.

- **Actividades que ofrece: "Dentro del balneario se realizan actividades al margen de las del aspecto meramente Balneoterápico como Aquagym, Senderismo, Actuaciones en directo o excursiones programadas a lugares de interés de la Región". H6**
- **Servicios más destacados: "Otros complementos del hotel son boutique y tienda de prensa y souvenirs, salón social, cibercorner" H17**

Finalmente, se presenta una lista con los servicios que el hotel ofrece. Esta es la parte más estructurada y objetiva del texto, y en consecuencia de donde se han extraído una mayor cantidad de instancias. Por ejemplo, en H3 encontramos la siguiente lista:

SECADOR DE PELO EN HABITACIONES, ANTENA PARABÓLICA, JARDÍN-TERRAZA, AIRE ACONDICIONADO EN HABITACIONES, RESTAURANTE ACCESIBLE SIN ESCALONES, AIRE ACONDICIONADO EN ZONAS COMUNES, TELÉFONO EN HABITACIONES, SALA(S) DE REUNIONES, SERVICIO DE FAX, HABITACIONES ESPECIALMENTE ADAPTADAS, INSTALACIONES Y ZONAS COMUNES ACCESIBLES, WC ADAPTADO A MINUSVÁLIDOS EN ZONAS COMUNES, SITIO CÉNTRICO, ASCENSOR, ADMITE TARJETAS DE CRÉDITO, GARAJE DE PAGO, BAR-CAFETERÍA, PISCINA, INTERNET EN HABITACIONES, SERVICIO DE SECRETARÍA, ASCENSOR ACCESIBLE ENTRE PLANTAS, CALEFACCIÓN CENTRAL, CAJA FUERTE INDIVIDUAL, HABITACIONES CON SALÓN-SUITES, TELEVISIÓN EN HABITACIONES, MINIBAR.

Como se puede ver, la segmentación del texto, es decir, la distribución de los enunciados está en relación con la distribución de los temas, los subtemas y los cambios de tema. En este caso, la unidad básica es el párrafo, unidad significativa supraoracional, que está constituida por un conjunto de enunciados relacionados entre sí por el contenido.

Esta macroestructura cuenta con elementos estándar que se repiten en otras producciones textuales pertenecientes al mismo dominio. Por ejemplo, en la

descripción del hotel Ritz Carlton Chicago[20] (Figura 5.1), se puede observar una distribución similar de los elementos discursivos:

Figura 5.1 Descripción del Hotel Ritz Carlton Chicago.

La misma estructura se presenta en textos escritos en inglés, como por ejemplo en el siguiente texto (Figura 5.2).

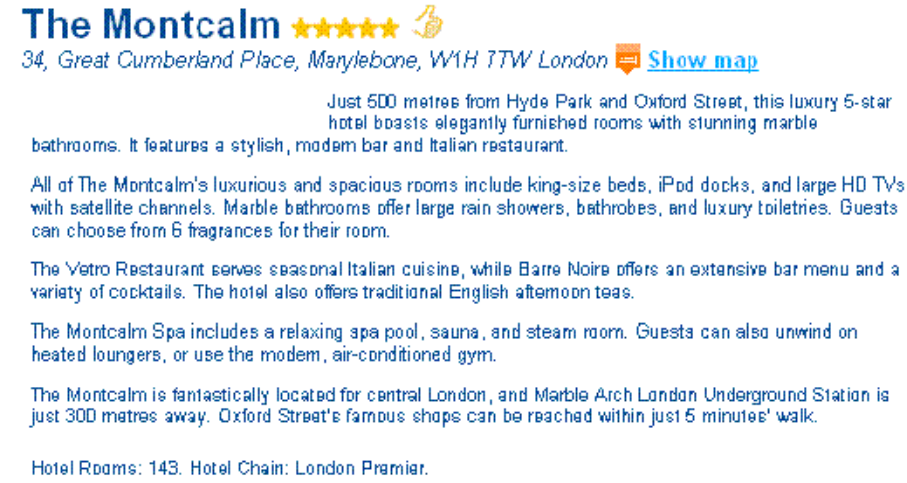


Figura 5.2 Descripción en inglés de un hotel.

20 Descripción extraída de http://www.booking.com

5.3 Uso del la información adquirida durante el análisis

La información lingüística obtenida tras el análisis del discurso es útil, por un lado, para la determinación de los parámetros cotextuales del sistema y, por otro, sirve para determinar cuáles son las entidades relevantes del dominio.

El cotexto es el conjunto lingüístico que rodea un elemento dado de un texto (Aznar et al., 1991). Es decir, aquella información incluida en segmentos anteriores y posteriores del texto (Bustos Gisbert, 1996).

Como indica Bustos Gisbert (1996), con el cotexto se hace alusión a dos tipos de información explícita. Por un lado, la que se incluye en el texto propiamente dicho y, por otro, la que aparece en textos paralelos al que es objeto de análisis. El cotexto, a diferencia del contexto, es siempre un elemento explícito.

En el caso de un corpus textual, el cotexto, es, por una parte, lo que circunda a un elemento textual concreto, y por otra parte, el cotexto lo constituyen aquellos documentos que aparecen junto con un documento dado.

El cotexto en la metodología propuesta es un elemento parametrizable, su configuración dependerá de las características textuales de los documentos de los que se va a extraer la información. Así, por ejemplo, en textos descriptivos en los que se enumeran una serie de elementos con sus características, el cotexto relevante para la extracción de información aparecerá siempre a continuación del elemento que se va a describir, mientras que la información de elementos anteriores puede ser poco o nada relevante.

Una vez descrita la macroestructura textual y las características lingüísticas de los *corpora* Hoteles y Restaurantes, los límites cotextuales de las entidades principales se pueden establecer con claridad.

Así, por ejemplo, en la figura 5.3 se puede ver que el cotexto relevante para la extracción de elementos informativos de la descripción de "Hotel 2" es aquel que aparece a continuación de la descripción y no el que aparece previamente. De

igual manera, no es relevante el cotexto que aparece en la descripción del "Hotel 3".

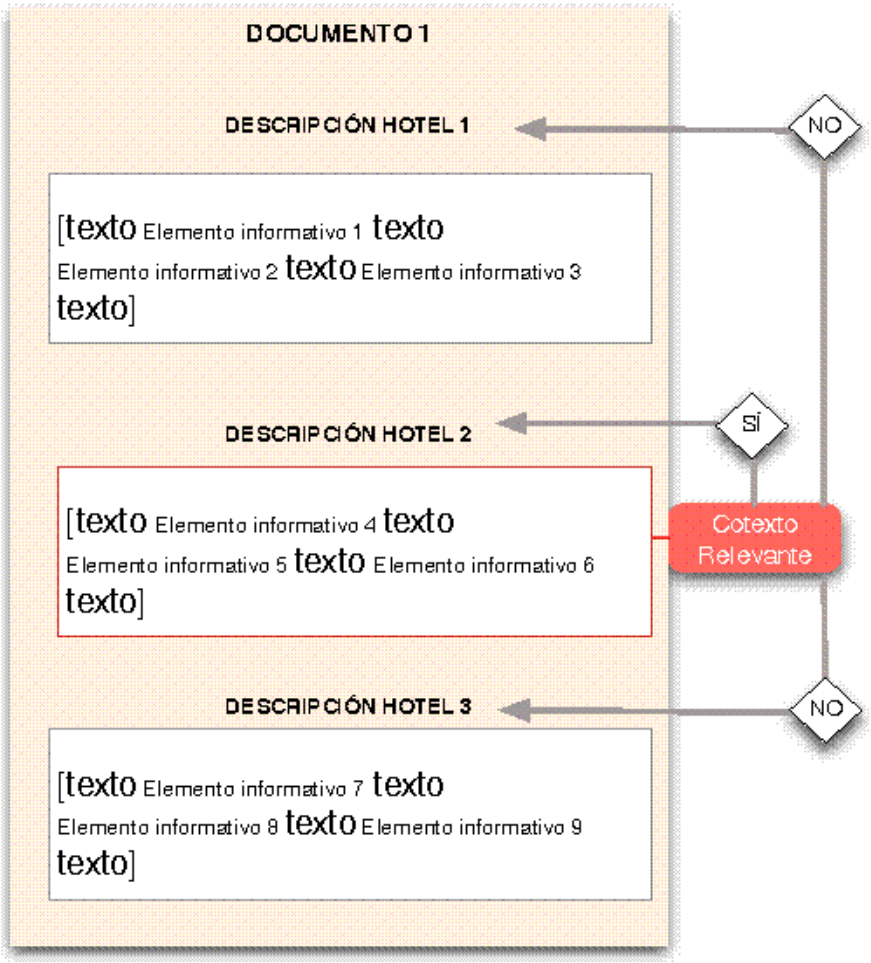


Figura 5.3 Cotexto en textos semi-estructurados.

En otro tipo de textos, como por ejemplo, en un artículo científico, la permeabilidad entre los distintos apartados es mayor y es más difícil realizar compartimentos estancos en la influencia del cotexto. Si bien es cierto, que los elementos más influyentes en un fragmento de texto son aquellos más cercanos al mismo, y cuanto mayor es el radio, menor es la influencia del cotexto.

En este tipo de textos en donde los límites son más difusos, el cotexto relevante se puede representar en forma de círculos concéntricos (figura 5.4), en donde los

círculos más cercanos al elemento textual informativo son los que mayor información aportan con respecto al mismo, y los más periféricos son irrelevantes.

En el caso de los corpora descritos, la información potencial que se puede obtener de un círculo concéntrico alejado de un elemento o entidad relevante es errónea, ya que, como se ha visto, se trata de información asociada a la entidad precedente o posterior.

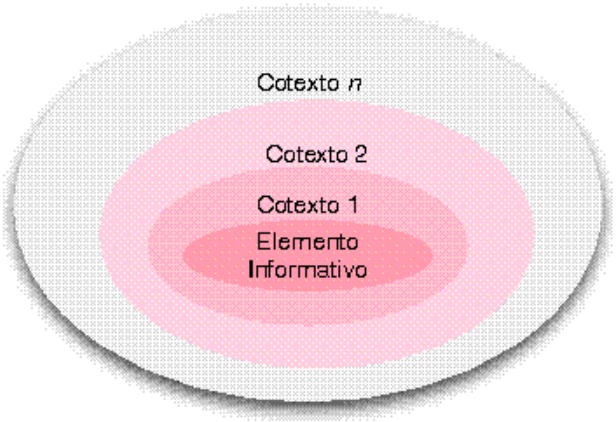


Figura 5.4 Cotexto en textos narrativos.

Además del cotexto, la metodología de instanciación automática de ontologías presentada en este capítulo, se fundamenta en la combinación con mayor ganancia de conocimiento.

Por ejemplo, tras el análisis discursivo realizado, se han obtenido cuáles son los elementos más frecuentes que aparecen junto a una entidad de la clase Hotel. Estos elementos aportan información relevante para la ontología y además se utilizan para la desambiguación, como se pone de manifiesto más adelante. Si una instancia se puede clasificar en varias clases de la ontología, se clasifica en la que aporte una mayor ganancia de conocimiento, es decir, en aquella con el mayor número de relaciones y propiedades.

Por otro lado, con la información obtenida a partir del análisis lingüístico, se ha determinado qué entidades son relevantes para el dominio. En la siguiente tabla (tabla 5.7), se pueden ver cuáles son los tipos de entidades relevantes que se han

obtenido del corpus Hoteles, con qué tipo de entidad ontológica se corresponden y cuál es su denominación en la ontología. En consecuencia, se han elaborado los recursos lingüísticos necesarios para las siguientes fases, entre las que se encuentra la identificación de dichos elementos como entidades.

Tabla 5.7 Tipos de entidades relevantes en el corpus Hoteles.

Entidad nombrada	**Tipo de entidad ontológica**	**Denominación en la ontología**
Hotel	Clase	Hotel
Dirección	ObjectProperty	hasAdress
Código Postal	DataProperty	zipCode
Ciudad	Class	Location
Teléfono	DataProperty	phone
Fax	DataProperty	fax
País	Clase	Country
URL	DataProperty	url
e-mail	DataProperty	e-mail
Estilo arquitectónico	Clase	ArchitectonicStyle
Monumentos	Clase	Monuments
Número de habitaciones	DataPropery	numberOfRooms
Aparcamiento	DataPropery	parking
Tarjetas	DataPropery	creditCard
Servicios	Clase	Services
	Clase	Activity

De igual manera, se han obtenido los tipos de entidades relevantes de corpus Restaurantes, recogidas en la tabla 5.8.

En la ontología, se han creado las clases, relaciones y propiedades pertinentes para cada una de ellas. En la siguiente tabla (Tabla 5.8), se puede ver una correspondencia entre las entidades relevantes en el texto y su representación en la ontología.

Tabla 5.8 Tipos de entidades relevantes en el corpus Restaurantes

Entidad nombrada	**Tipo de entidad ontológica**	**Denominación en la ontología**
Restaurante	Clase	Restaurant
Dirección	ObjectProperty	hasAdress
Ciudad	Class	Location
Código Postal	DataProperty	zipCode
Teléfono	DataProperty	phone
Fax	DataProperty	fax
País	Clase	Country
URL	DataProperty	url
e-mail	DataProperty	e-mail
Jefe de cocina	DataPropery	chef
Jefe de sala	DataPropery	headWaiter
Estilo arquitectónico	Clase	ArchitectonicStyle
Menú	Clase	Menu
Cierre Semanal	DataPropery	weeklyClosing
Cierre Vacacional	DataPropery	hollidayClosing
Número de Comensales	DataPropery	numberOfSeats
Aparcamiento	DataPropery	parking
Tarjetas	DataPropery	creditCard
Especialidades	ObjectProperty	hasSpecialty

Dado que algunas de ellas presentan un contenido cerrado, como por ejemplo la entidad "Municipio", que sólo se puede corresponder con los municipios de la Región de Murcia, o "cierre semanal", que sólo se puede corresponder con los días de la semana, se ha elaborado una serie de Gazetteers para en GATE.

Como se ha comentado previamente, un gazetteer es una lista de entidades relevantes de un domino, es procesable por GATE y se pueden combinar con reglas JAPE. Un ejemplo de Gazetteer pueden ser los días de la semana o los servicios ofertados por un hotel. En la figura 5.5, se muestra un ejemplo de Gazetteer integrado en GATE.

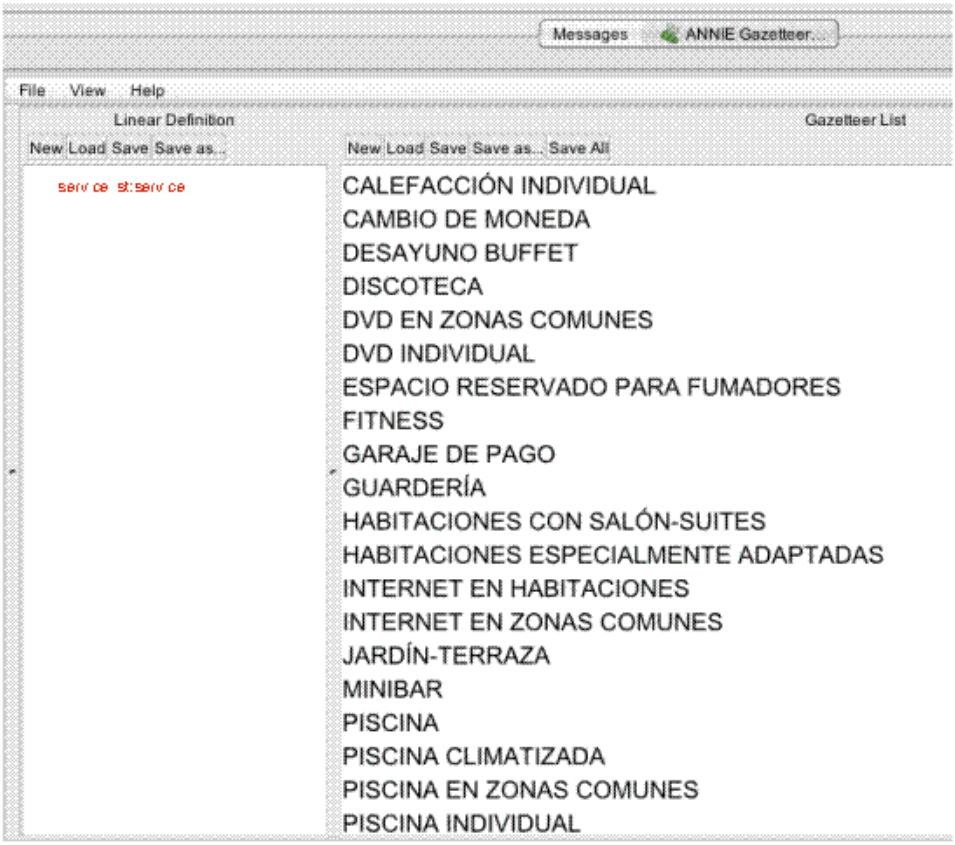


Figura 5.5 Gazetteer de servicios en GATE.

No obstante, el análisis del discurso contribuye especialmente a la elaboración de patrones generales para el reconocimiento de entidades que no pertenecen a listas cerradas. Este es el caso de los patrones de localización que se han especificado en el apartado 5.2.3. Es el uso de estos patrones, junto con los Gazetteers, lo que permite la creación de reglas JAPE utilizadas para la anotación y extracción de las entidades en el texto.

Una regla JAPE es un conjunto de expresiones regulares que permite la extracción de información textual. La siguiente regla (tabla 5.9) se utiliza para localizar información referente al cierre semanal de un restaurante.

Tabla 5.9 Regla JAPE cierre semanal.

```
  Rule:CreaCierreSemanal
  priority:20
  (
   (
    (CIERRESEMANAL)
    (
    (SPACE) |
    {Token.kind ==word} |
    {Token.kind == punctuation} |
    (SPACE) (CONECTOR) |
    (SPACE) {Token.kind == number}
   )+
  ):CierreSemanal
  )
  --
>:CierreSemanal.CierreSemanal={rule="CierreSemanal"}
```

En cuanto a la macroestructura textual, la existencia de una estructura fija, no sólo se encuentra en el corpus que se ha analizado, sino que, con algunas variaciones, se trata de una estructura estándar para la descripción de establecimientos hoteleros, lo que favorece la portabilidad del sistema.

Por otro lado, la estructura de la ontología, descrita más adelante (ver apartado 5.4.1), recoge, en forma de clases y propiedades, las entidades nombradas relevantes identificadas durante el análisis lingüístico.

5.3.1 Adaptación de los recursos a otros idiomas

Aunque los corpora principales utilizados para el entrenamiento y validación del sistema son los descritos anteriormente, se han realizado también experimentos en textos escritos en lengua inglesa (Ruiz-Martínez et al., 2010). Para ello, las listas de entidades han sido traducidas al inglés. De igual modo se han extraído los patrones pertinentes para el desarrollo de reglas que permiten

identificarlas, con la consecuente modificación del módulo del sistema encargado del reconocimiento de entidades nombradas.

La modularidad de la metodología propuesta, permite que se pueda adaptar a otros idiomas.

5.4 Desarrollo de la ontología de dominio turístico

La gran cantidad de servicios turísticos ofertados a través de la web, ha auspiciado el desarrollo de proyectos que pretenden sistematizar la información que contienen, lo que a su vez ha propiciado el desarrollo de ontologías relacionadas con el dominio del turismo. Por ejemplo, Hi-Touch es un proyecto europeo relacionado con el turismo sostenible en el que se ha utilizado una ontología desarrollada por la compañía Mondeca (Delahousse, 2003). El objetivo del proyecto es la creación de un conjunto de herramientas que permitan tanto a operadores turísticos como al usuario final acceder a los productos turísticos de su zona de forma personalizada. La ontología Hi-Touch cuenta con 6 clases principales que abarcan de manera muy genérica los distintos ámbitos turísticos. Se complementa con un tesauro multilingüe que es una adaptación del *Thesaurus on Tourism and Leisure Activities* desarrollado por la Organización Mundial de Turismo (OMT) (WTO, 2001). Los descriptores son utilizados para indizar los productos turísticos y las consultas de los usuarios. A su vez, la compañía Mondeca cuenta con su propia ontología turística basada en el tesauro de OMT.

Por su parte, el grupo de trabajo e-Tourism en DERI (Digital Enterprise Research Institute) ha creado la ontología OnTour (Pratner, 2004), la cual se centra, fundamentalmente, en la descripción de alojamiento y actividades y es uno de los componentes de un proyecto para la búsqueda de paquetes vacacionales, alojamiento y actividades.

Dentro del proyecto europeo Harmonise, se desarrolló la ontología IMHO (Interoperable Minimum Harmonisation Ontology) (Missikof, 2003), que, en un principio, se centró en el alojamiento, eventos y actividades. En la actualidad,

abarca muchos más subdominios y forma parte del proyecto HarmoNET, en el que participan numerosas organizaciones internacionales, y cuyo objetivo es el intercambio de información turística a nivel global.

La ontología QALL-ME (Izquierdo, et al., 2007), también dentro del marco de un proyecto financiado por la UE, ha sido aplicada a un sistema de pregunta-respuesta (*question answering*) y sus clases principales son destinos turísticos, sitios turísticos, eventos turísticos y transportes. Además, esta ontología ha sido alineada con WordNet (Fellbaum, 1998; Miller, 1995) y SUMO (IEEE, 2011).

Otras ontologías, son la desarrollada por el grupo de investigación SEED (SEmantic E-tourism Dynamic packaging) para sistemas de información turística (OTIS) (Cardoso, 2005); la ontología AuSTo (Australian Sustainable Tourism) (Jakkilinki, 2007) o LA_DMS (Kanellopoulos et al., 2008), a saber, una ontología general pensada para adaptarse a las necesidades del usuario sobre información turística de distintos destinos.

Uno de los proyectos más recientes es la ontología cDott (Barta et al., 2009) construida a partir de otras ontologías turísticas como Harmonise y que posee una estructura modular que permite la integración de distintas ontologías con un grado variable de especificidad.

Finalmente, existen ontologías muy específicas desarrolladas para ámbitos locales, como la Ontología Cruzar (Gutiérrez Losada, 2010), cuyo objetivo es la construcción de una aplicación para el cálculo de rutas turísticas personalizadas en la ciudad de Zaragoza.

5.4.1 Ontología Turismo

Como se ha indicado, la metodología desarrollada ha sido testeada en el dominio del turismo. Para ello, ha sido necesaria la creación de una ontología turística centrada, fundamentalmente, en alojamiento y restauración.

Teniendo en cuenta la existencia de numerosas ontologías de dominio turístico, como las mencionadas anteriormente, se ha creído conveniente la reutilización y adaptación de algunas de ellas para el propósito de esta investigación.

El punto de partida de la Ontología Turismo ha sido la ontología travel.owl desarrollada en Protégé (Knublauch, 2004). Se trata de una ontología básica que se ha modificado considerablemente añadiendo elementos ontológicos de la ontología OnTour, descriptores del tesauro de la OMT y aquellas clases y propiedades específicas que se han considerado relevantes después de estudiar los corpora y que se refieren al sector de la hostelería y restauración en España.

El resultado es una ontología implementada en OWL2 que contiene toda la información turística que requiere el escenario de estudio que nos ocupa.

A continuación, se muestran las clases principales de la misma con algunas de sus subclases (ver figura 5.6).

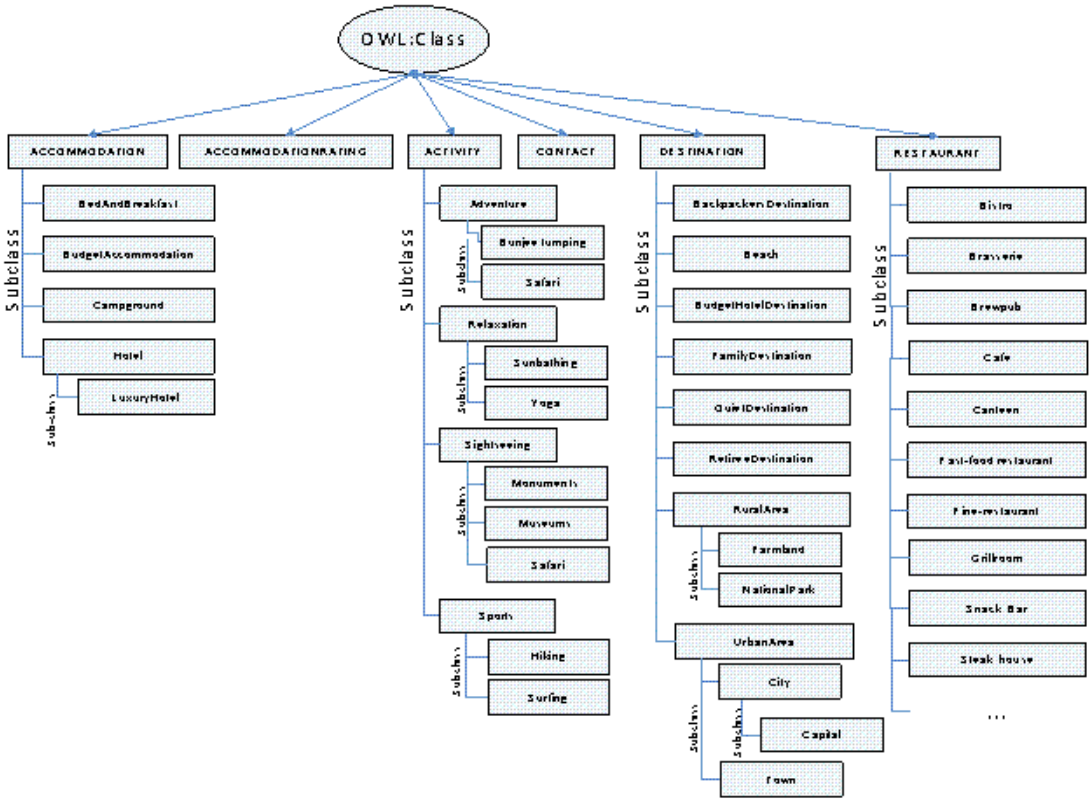


Figura 5.6 Extracto de la ontología Turismo.

La ontología se articula en torno a dos clases principales, la clase *Hotel*, que es una subclase de *Accomodation*, y la clase *Restaurant* (Restaurante), que, a su vez, cuenta con varias subclases. La ontología se ha proyectado para dar cobertura a un amplio sector del ámbito turístico. Los hoteles y restaurantes son las clases que mayor ganancia de conocimiento obtienen durante el experimento, debido a que los *corpora* utilizados son de este ámbito, pero también se pueden incluir instancias de la clase *Monuments* (Monumentos) o *Sports* (Deportes).

La ontología tiene 6 clases principales, que son *Accomodation* (Alojamiento), *AccomodationRating* (Clasificación del Alojamiento), *Activity* (Actividad), *Contact* (Contacto), *Destination* (Destino), *Restaurant* (Restaurante), *Facility* (Instalaciones), *Location* (Localización), *Room* (Habitación), *Service* (Servicio).

En cuanto a las *ObjectProperties*, existen algunas de carácter general, como *hasLocation* (tiene localización) o *hasContact* (tiene contacto), aplicables tanto a hoteles como a restaurantes, y otras de carácter más específico como *hasRoom* (tiene habitación) o *hasActivity* (tiene actividad) para los hoteles y *hasMenu* (tiene menú) o *hasHeadWaiter* (tiene jefe de sala) para los restaurantes.

Las *DatatypeProperties* se han definido para datos específicos como el número de habitaciones (nuberRooms), el teléfono (hasTelephone), las plazas de aparcamiento (hasParking), etc.

Además, se han definido algunas expresiones lógicas que permiten inferir nuevo conocimiento a partir del conocimiento disponible en la ontología. Por ejemplo, un destino familiar o *FamilyDestination* (Figura 5.7) es aquel que tiene habitaciones (*hasRoom*) de tipo familiar o que tiene servicios (*hasServices*) del tipo guardería, servicio_de_canguro, video_juegos, área_de_juegos_para_niños o piscina_infantil.

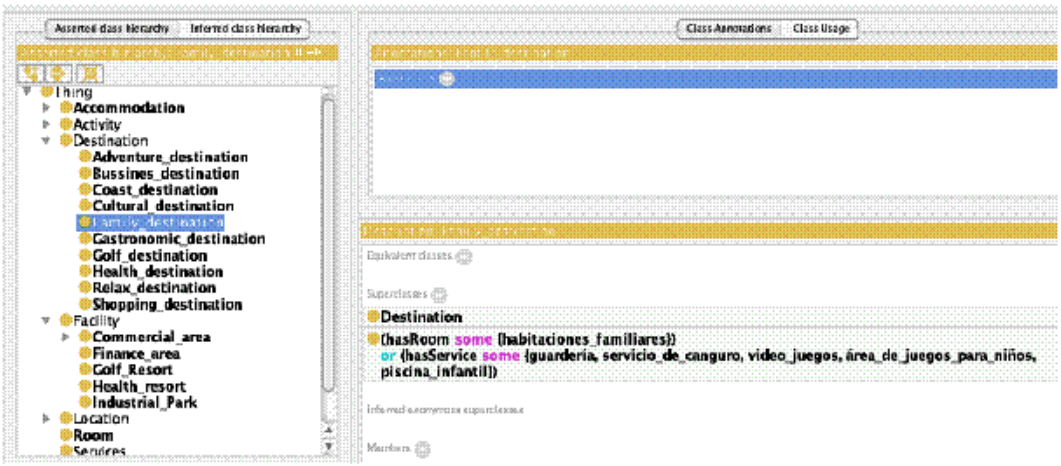


Figura 5.7 Ejemplo de clases en la ontología de turismo.

Para el desarrollo de esta ontología, se ha utilizado el editor de ontologías Protégé (SMI, 2007), que es el editor de ontologías más utilizado en la actualidad (Cardoso, 2007) y que ha sido desarrollado por la Universidad de Stanford.

5.5 Metodología para la instanciación automática de ontologías

En esta sección, se explica la metodología desarrollada para la instanciación automática de ontologías. El proceso de población de la ontología se lleva a cabo en cuatro fases secuenciales ilustradas en la figura 5.8:

1. Fase de Procesamiento de Lenguaje Natural y de Procesamiento del Corpus.
2. Fase de Reconocimiento e identificación de las Entidades Nombradas.
3. Fase de Población de la Ontología.
4. Verificación de la consistencia de la ontología.

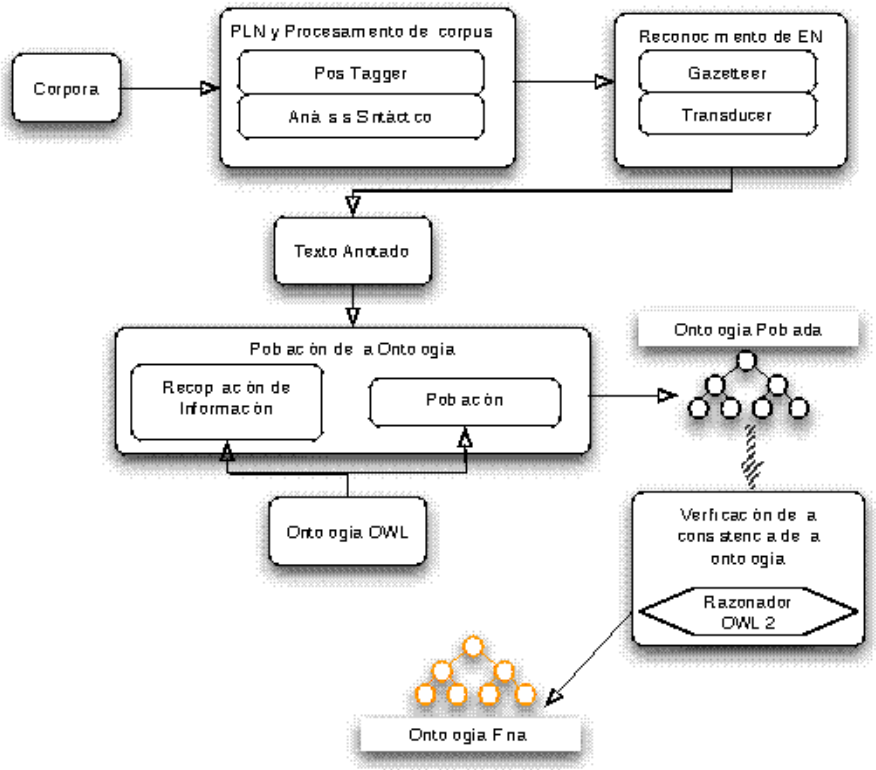


Figura 5.8 Arquitectura general del sistema.

Brevemente, la metodología funciona de la siguiente manera. En primer lugar, el corpus es analizado morfo-sintácticamente con el objetivo de extraer la información lingüística necesaria para las siguientes fases. Para ello, se ha utilizado el framework GATE. Durante la segunda fase, se extraen un conjunto de menciones de entidades, de modo que, cuanto mayor sea el número de menciones en esta fase, mayor será el número potencial de instancias de la ontología, ya que, en la siguiente fase, dichas menciones serán candidatas a instancias o a valores de propiedades de la ontología. A continuación, se desambiguan aquellas ocurrencias de entidades identificadas en la fase previa y se instancia la ontología. Durante esta tercera fase, cada entidad nombrada se inserta en la ontología como un individuo de una o más clases. Así mismo, se identifican los valores de sus atributos y relaciones. Si una instancia no había sido insertada previamente, entonces el sistema instancia la ontología con la información extraída, creando así un nuevo individuo. Si, por el contrario, una instancia ya existía, es enriquecida con nuevas relaciones y atributos, en el caso de que hayan sido identificados. Finalmente, en la última fase, se comprueba la consistencia de la ontología mediante un razonador OWL-DL para detectar individuos erróneos y que no cumplan las restricciones de la ontología

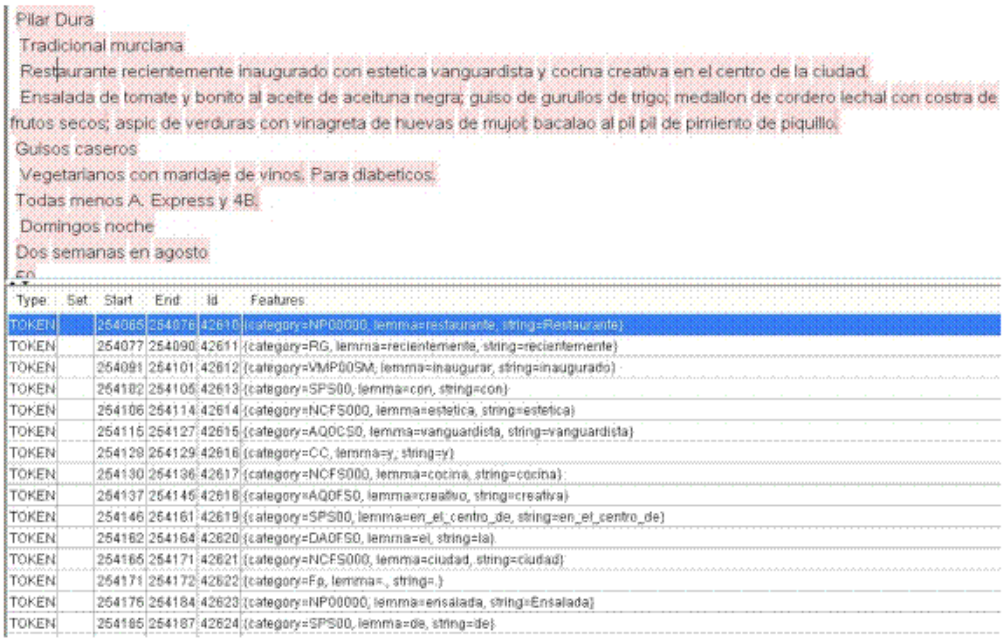
A continuación, se explica detalladamente cada una de las fases de este proceso.

5.5.1 Fase de procesamiento del corpus

El principal objetivo de esta fase es obtener la estructura morfológica y sintáctica de cada oración en el corpus. Con este fin, se han utilizado el conjunto de herramientas de PLN previamente descritas, que han sido desarrolladas en GATE.

Haciendo uso de dichas herramientas, se procede a la tokenización, lematización y etiquetado morfológico y gramatical de aquellas partes del corpus

que no habían sido previamente utilizadas para el análisis lingüístico y desarrollo de los recursos.


Pilar Dura
Tradicional murciana
Restaurante recientemente inaugurado con estetica vanguardista y cocina creativa en el centro de la ciudad.
Ensalada de tomate y bonito al aceite de aceituna negra; guiso de gurullos de trigo; medallon de cordero lechal con costra de frutos secos; aspic de verduras con vinagreta de huevas de mujol; bacalao al pil pil de pimiento de piquillo.
Guisos caseros
Vegetarianos con maridaje de vinos. Para diabeticos.
Todas menos A. Express y 4B.
Domingos noche
Dos semanas en agosto
50

Type	Set	Start	End	Id	Features
TOKEN		254065	254076	42610	{category=NP00000, lemma=restaurante, string=Restaurante}
TOKEN		254077	254090	42611	{category=RG, lemma=recientemente, string=recientemente}
TOKEN		254091	254101	42612	{category=VMP00SM, lemma=inaugurar, string=inaugurado}
TOKEN		254102	254105	42613	{category=SPS00, lemma=con, string=con}
TOKEN		254106	254114	42614	{category=NCFS000, lemma=estetica, string=estetica}
TOKEN		254115	254127	42615	{category=AQ0CS0, lemma=vanguardista, string=vanguardista}
TOKEN		254129	254129	42616	{category=CC, lemma=y, string=y}
TOKEN		254130	254136	42617	{category=NCFS000, lemma=cocina, string=cocina}
TOKEN		254137	254145	42618	{category=AQ0FS0, lemma=creativo, string=creativa}
TOKEN		254146	254161	42619	{category=SPS00, lemma=en_el_centro_de, string=en_el_centro_de}
TOKEN		254162	254164	42620	{category=DA0FS0, lemma=el, string=la}
TOKEN		254165	254171	42621	{category=NCFS000, lemma=ciudad, string=ciudad}
TOKEN		254171	254172	42622	{category=Fp, lemma=., string=.}
TOKEN		254176	254184	42623	{category=NP00000, lemma=ensalada, string=Ensalada}
TOKEN		254185	254187	42624	{category=SPS00, lemma=de, string=de}

Figura 5.9 Análisis morfológico con Freeling a través de GATE.

En la Figura 5.9, se puede ver un pequeño fragmento del corpus analizado mediante la integración de Freeling en GATE. En él se muestra la separación en tokens, la posición que ocupa dentro del corpus, la categoría gramatical a la que pertenece la palabra, indicada mediante una etiqueta, el lema y el string, es decir, el término que encontramos en el texto.

En cuanto a las etiquetas morfológicas que utiliza el sistema, se trata del estándar EAGLES, definido por Sinclair (1996). En este estándar, a cada elemento morfológico se le asigna una etiqueta alfanumérica, en la que se indica la categoría gramatical, junto con elementos como el género, el número y, en el caso de los verbos, la persona, el tiempo, el aspecto, la diátesis, etc.

5.5.2 Fase de reconocimiento de Entidades Nombradas

Durante esta segunda etapa, se procede a la identificación de los elementos del texto que son candidatos a entidades. Para ello, se ha utilizado de nuevo el Framework GATE y los plugins desarrollados para tal efecto.

Como se ha indicado anteriormente, en esta tesis doctoral, el Reconocimiento e Identificación de entidades nombradas es una subtarea de la extracción de información que trata de localizar y clasificar elementos concretos del texto según una serie de categorías predefinidas tales como nombres de persona, organizaciones, localizaciones, expresiones de tiempo, cantidades monetarias o porcentajes.

La lista de entidades empleada varía en función del dominio objeto de análisis. Como se ha dicho, la validación de la metodología presentada aquí se ha realizado en el dominio turístico y, para ello, se han desarrollado los recursos específicos a los que se hace alusión en el apartado 5.3.

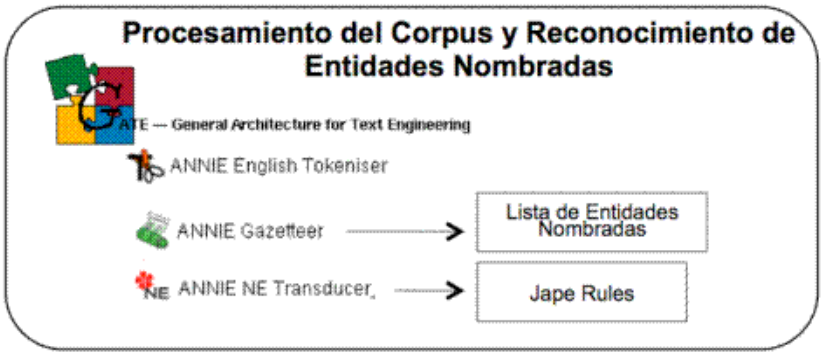


Figura 5.10 Procesamiento del corpus y Reconocimiento de entidades nombradas.

Como se puede observar en la figura 5.10, la fase de reconocimiento de entidades consta de dos componentes principales: un Gazetteer o lista de entidades y el transducer o aplicación que interpreta un conjunto de reglas JAPE. La salida de cada componente de GATE es un conjunto de anotaciones, es decir, metadatos asociados con una sección concreta del contenido del documento. La función de

los gazetteers es identificar entidades nombradas en el texto basándose en listas predefinidas como ciudades, nombres de persona, etc. Así, el sistema obtiene anotaciones para cada palabra que aparece en la lista de entidades que son relevantes para el dominio en cuestión.

Las listas desarrolladas para el dominio del turismo son de dos tipos, de carácter general, tales como, localizaciones, código postal, nombres, apellidos e identificadores de direcciones, y de carácter específico, tales como servicios de los restaurantes, comidas o categorías profesionales.

JAPE *transducer* es un módulo para la ejecución de las gramáticas JAPE que están basadas en expresiones regulares. Las reglas desarrolladas en JAPE son capaces de obtener códigos postales, teléfonos, urls, direcciones de correo electrónico, direcciones postales, restaurantes, nombres de persona o cantidades monetarias, entre otras.

En las siguientes tablas (tabla 5.10 y tabla 5.11), se muestran dos ejemplos de reglas JAPE, una para la identificación de cantidades monetarias y otra para la identificación de servicios de un hotel.

Tabla 5.10 Regla JAPE para la identificación de cantidades monetarias

```
Rule:Money
// $30
// $40,4
(
({Token.string == "$"}
({Token.kind == number}
 (({Token.string == ","}|
  {Token.string == "."}
  )
  {Token.kind == number}
 )?
)
)
:number
--> :number.Money = {kind = "money", rule = "Money"}
```

Tabla 5.11 Regla JAPE para la identificación de servicios hoteleros.

```
Phase:     Service
Input: Token Lookup SpaceToken
Options: control = appelt
Rule:ServiceRule
Priority: 50
(
     {Lookup.majorType == service}

)
:service -->
 :service.Service=   {kind   =   "Service",   rule   =
"Service1"}
```

Cada anotación obtenida mediante el JAPE *transducer* es considerada una mención de una entidad nombrada y, a su vez, todas la entidades nombradas identificadas en el texto son candidatas a ser instancias o valores de los atributos de una instancia en la ontología. Por ejemplo, una entidad nombrada que representa un Hotel es considerada como candidato de una instancia de la clase Hotel, y las entidades nombradas que representan números de teléfono o direcciones de mail son consideradas como candidatos a posibles valores de instancias de atributos de la ontología (por ejemplo, el número de teléfono de un hotel).

GATE anota las menciones de cada entidad nombrada. Por ejemplo, si se ha definido una entidad nombrada del tipo Hotel, GATE anota en el documento todas las veces que aparece un miembro de ese tipo de entidad nombrada como puede ser "Hotel las Palmeras", "Hotel la Manga", "Hotel don Juan", etc. Cada una de las instancias de la clase Hotel pasa a ser una mención de una entidad nombrada. Por otra parte, si la misma entidad nombrada aparece más de una vez en el documento, solamente se inserta una instancia con dicho nombre en la ontología. Por ejemplo, si la entidad "Hotel don Juan" ha sido anotada tres veces en distintas partes del corpus, se contabiliza como tres menciones distintas de la misma entidad, pero sólo se inserta una instancia con dicho nombre.

De igual modo, GATE puede anotar "Av. Ramón y Cajal" y "Avenida Ramón y Cajal", que son exactamente la misma entidad nombrada expresada de distintos modos, con lo cual sólo una de las opciones pasa a ser un valor de un atributo de un individuo de la ontología.

En la siguiente figura (figura 5.11), se muestra un ejemplo de la lista de las entidades nombradas y las menciones de éstas resaltadas en el texto.

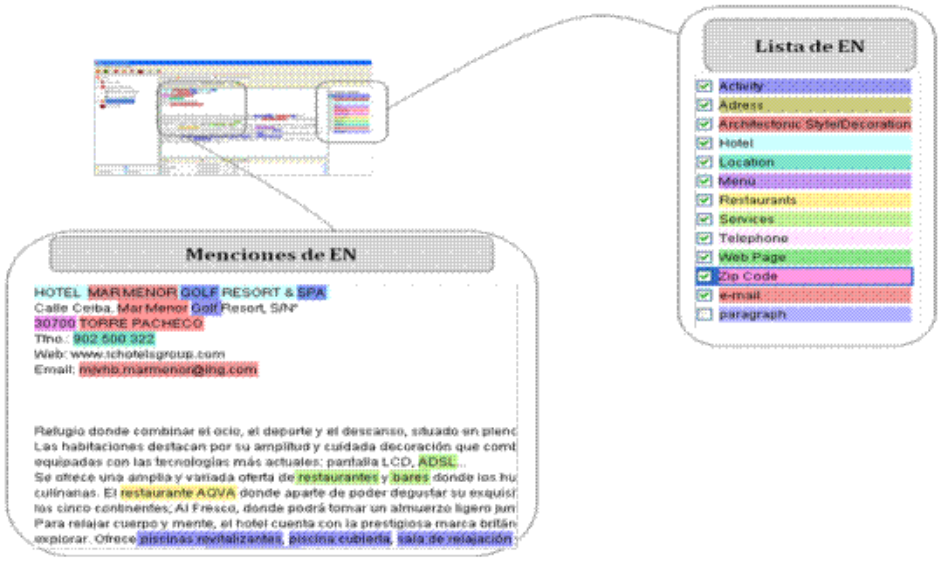


Figura 5.11 Ejemplo de anotación de entidades nombradas.

Una vez anotadas las entidades nombradas, el sistema agrupa las anotaciones y las clasifica como pertenecientes a uno o más tipos de entidad nombrada (por ejemplo Activity, Hotel, Restaurant, etc.). Cada grupo puede estar formado por una o varias anotaciones que se solapan en el texto. Por ejemplo, el grupo "ADSL" está formado por una sola anotación del tipo "Services", mientras que el grupo "piscina cubierta" está formado por dos anotaciones, ambas pertenecientes al tipo "Services". A su vez, las anotaciones que conforman un grupo se pueden clasificar como pertenecientes a uno o más tipos de entidad. En este último caso, el grupo presenta una ambigüedad. Por ejemplo, en la Figura 5.11 es posible observar que hay tres tipos de anotaciones resaltadas en la primera frase. Por un lado, "Hotel Mar Menor Golf Resort & Spa" es anotado como una entidad del tipo

Hotel. Por otro lado, "Mar Menor" se anota como una entidad del tipo *Location,* y, finalmente, "Golf" y "Spa" son anotadas como entidades del tipo *Activity*. Este tipo de ambigüedad se resuelve en la siguiente fase, donde el sistema debe decidir si se corresponden con alguna instancia, relación o atributo de la ontología.

5.5.3 Fase de instanciación de la ontología

En la fase anterior, se han obtenido un conjunto de grupos de anotaciones de entidades nombradas gracias a los recursos lingüísticos implementados. Durante la fase de instanciación de la ontología, el sistema determina si dichas anotaciones se corresponden con instancias, atributos o relaciones en el modelo ontológico y, además, permite resolver las posibles ambigüedades surgidas en la fase anterior y generadas por la clasificación de un grupo de anotaciones en más de un tipo de entidad nombrada. De este modo, el sistema debe asociar cada anotación o grupo de anotaciones a un tipo concreto de entidad ontológica. En las ontologías desarrolladas en OWL, los principales elementos ontológicos son Clases, Relaciones taxonómicas (Subclase de), *DatatypeProperties* (atributos), *ObjectProperties* (relaciones) e *Individuals* (Instancias).

El proceso para llevar a cabo la fase de instanciación de la ontología consta de 4 subfases principales:

1. Recopilación de las entidades nombradas identificadas en la fase anterior.
2. Creación de un árbol con todas las combinaciones posibles de ENs ambiguas.
3. Cálculo de los valores de todas las combinaciones.
4. Inserción de los individuos en el modelo ontológico.

A continuación, se explica cada una de estas subfases en detalle.

1. Recopilación de las entidades nombradas identificadas en la fase anterior.

Este primer paso toma como entrada todos los grupos de anotaciones que han sido identificadas en la fase de Reconocimiento de entidades nombradas.

La lista de grupos contiene todas las anotaciones que se han realizado en el texto y que se han relacionado con una o más entidad nombrada, es decir, contiene tanto entidades nombradas ambiguas como no-ambiguas.

Las ambigüedades son, fundamentalmente, de carácter lingüístico y se pueden agrupar en tres tipos:

- **Tipo A:** Una anotación o grupo puede estar relacionado con más de una entidad nombrada. Este tipo de ambigüedad se produce cuando la posición de comienzo y fin de la anotación de una entidad es exactamente la misma que la posición de comienzo y fin de una entidad diferente. Cuando se produce este tipo de ambigüedad sólo una de las anotaciones es correcta en ese escenario concreto.

  Por ejemplo, "Salzillo" puede ser una Persona, un Museo o una Obra de Arte. Aunque durante el proceso de instanciación podrían ser insertadas en la ontología cada una de las instancias en la clase correspondiente de forma independiente, para un determinado escenario sólo una de las instancias que se pueden extraer es la verdadera, es decir, aunque todas pueden ser instancias de la ontología, las relaciones con otros componentes ontológicos son diferentes. Por ejemplo, si en el texto se está haciendo alusión a Salzillo como Museo, la anotación identificada como entidad nombrada estará relacionada con otros elementos o propiedades asociadas con la clase Museo, tales como número de teléfono, dirección, horario, precio de la entrada, etc. Estas relaciones son diferentes si la instancia hace alusión a una persona u obra de arte.
- **Tipo B**: Varias entidades nombradas pueden estar solapadas en el texto. Este tipo de ambigüedad se produce cuando la posición de comienzo o fin de una entidad se solapa con la posición de comienzo o fin de una entidad diferente.

Por ejemplo, "Hotel Jardines de Lorca" es un grupo de anotaciones en el que se solapan dos tipos de entidades "Lorca" como Localización, por un lado, y "Hotel Jardines de Lorca" como Hotel.

- **Tipo C**: Una mención de una entidad nombrada puede estar relacionada con varios recursos de la ontología. Este tipo de ambigüedad aparece gráficamente representada en el texto como una ambigüedad de tipo A.
  Por ejemplo, el número "968232426" puede ser tanto un número de teléfono como un número de fax de un hotel o un restaurante.

2. Creación de un árbol con todas las combinaciones posibles de ENs ambiguas.

Las ambigüedades de Tipo A se resuelven mediante la definición de todos los posibles grupos de anotaciones no-ambiguas. Dos anotaciones pertenecen al mimo grupo si sus posiciones de comienzo y fin se solapan. Los grupos de anotaciones se representan en forma de árbol, de forma que cada grupo se corresponda con un grupo de anotaciones no ambiguo. Es decir, cada nivel del árbol, desde la raíz hasta los nodos-hoja, representa una entidad nombrada y sus hermanas, a un nivel en el que las anotaciones ambiguas relacionadas con una entidad son incompatibles. (Ver figura 5.14)

Por otro lado, la resolución de las ambigüedades del Tipo B y C se lleva a cabo relacionando cada entidad con la más cercana. Este hecho se basa en cuántas entidades nombradas pueden ser interrelacionadas y cuál es la distancia textual que separa unas de otras. Por lo tanto, cuanto mayor sea el número de anotaciones que se adapten al modelo ontológico y cuanto más cercanas estén, mejor será la puntuación que cada grupo puede lograr.

3. Cálculo de los valores de todas las combinaciones.

Para obtener las anotaciones más adecuadas a partir de todas las combinaciones, a cada rama del árbol se le asigna una puntuación basada en tratar

de completar la mayor información posible de individuos detectados en la ontología. Es lo que se denomina ganancia de conocimiento

Para ello, se ha desarrollado un algoritmo que permite evaluar todos los posibles grupos de entidades nombradas y que se describen en la tabla 5.12. El parámetro de entrada del algoritmo es la *NE_list*. Esta lista contiene todas las anotaciones de las entidades nombradas identificadas. El árbol, que representa todos los grupos de entidades permitidos, se genera mediante la función *combinatorial_tree_of_NE(NE_list)*. Esta función crea un árbol donde los casos ambiguos se representan en un mismo nivel, creándose un nodo para cada anotación que se considera incompatible. La profundidad del árbol será igual al número de entidades.

Tabla 5.12 Cálculo de combinaciones posibles entre las anotaciones.

```
Procedure get_the_best_combination(NE_list) BEGIN
  root_of_NE_tree = combinatorial_tree_of_NE(NE_list);
  NE_to_be_visited = stack_of_NE();
  NE_to_be_visited.push(root_of_NE_tree);
  score_Best = 0;
  solution_Best = list_of_NE();
  WHILE NE_to_be_visited.has_elements() DO
    NE_node = NE_to_be_visited.pop();
    IF NE_node.has_children() THEN
      NE_to_be_visited.pushAll(NE_node.getChildren());
    ELSE
      solution_Current = NE_node.list_path_nodes();
      score_Current = calculate_score(NE_node);
      IF score_Current > score_Best THEN
        score_Best = score_Current;
        solution_Best = solution_Current;
      END IF
    END IF
  END WHILE
  return solution_Best;
END PROCEDURE
```

El número total de grupos de anotaciones no-ambiguas se calcula con la siguiente función:

$$\prod_{i=1}^{n} NE_i$$

donde NE_i es el número de anotaciones permitidas de una entidad ambigua "*i*" en el texto, es decir, determina el número de niveles en el árbol.

Cada vez que se encuentra una anotación ambigua, se crea un nuevo nivel, mientras las anotaciones no ambiguas se agrupan en un mismo nivel.

El algoritmo recorre todos los nodos desde la raíz hasta las hojas del árbol. Cuando el algoritmo llega a un nodo-hoja, calcula la puntación del grupo de anotaciones mediante la función *calculate_score(NE_node)*. Finalmente, cuando todos los grupos han sido generados y ha sido establecida su puntuación, sólo permanece el nodo del árbol que posea la máxima puntuación. La valoración de cada grupo se basa en el número de anotaciones potenciales que pueden ser mapeadas al modelo ontológico y el número de relaciones que se pueden crear entre ellas.

Las anotaciones están representadas en la ontología como clases o propiedades. Por el contrario, en el texto, las anotaciones aparecen aisladas aunque, normalmente, las anotaciones más cercanas están relacionadas entre sí, estableciéndose una red de relaciones textuales entre las anotaciones. Cuando una anotación puede ser relacionada con otras anotaciones, sólo se selecciona aquella más cercana, es decir, aquella cuya distancia cotextual es menor. En la tabla 5.13 se puede ver el pseudo-código que define el cálculo de esta función.

Tabla 5.13 calculate_score function.

```
Function calculate_score(list_of_Annotations) BEGIN
  score_total = 0;

  FOR EACH annotationA IN list_of_Annotations DO
    score_A = 0;
    IF annotationA.is_a_class THEN
      score_A += Wclass;
      FOR EACH annotationB IN list_of_Annotations DO
```

```
        IF  (annotationA.can_be_related_to(annotationB)  AND
            annotationA.is_the_closest_to(annotationB)) THEN
          distance_AB      =      annotationA.position      -
          annotationB.position
          score_A += (Wrelationship / distance_AB)
        END IF
      END FOR
    ELSIF annotationA.is_a_property THEN
      score_A += score_A + Wproperty;
    END IF
    score_total += score_A;
  END FOR
  return score_total;
END FUNCTION
```

En la tabla 5.13, la variable $score_{total}$ representa el valor total obtenido por la lista de anotaciones, mientras que $score_A$ representa el valor que se le asigna a cada anotación en la lista. W_{class} se refiere al peso que se asigna a una anotación que es un individuo de una clase de la ontología, $W_{property}$ es el peso asignado a una anotación cuando es mapeada con una *datatype property* en la ontología, y $W_{relationship}$ es el peso que se le da a un individuo cada vez que es relacionado con otro individuo o *data property* en la ontología. No obstante, $W_{relationship}$ puede ser reajustado más adelante en función de la distancia cotextual entre anotaciones. La distancia se mide calculando el número de palabras que separan dichas anotaciones en el texto. De este modo, cuanto más cercanas se encuentren dos anotaciones en el texto, mayor será la puntación asignada.

Es el usuario final quien establece los valores de las constantes W_{class}, $W_{property}$ y $W_{relationship}$ y, en última instancia, el resultado final del proceso de población de la ontología depende de los valores asignados a estos pesos.

El ámbito de influencia de las relaciones de las anotaciones es también un parámetro que el usuario puede definir, pudiendo penalizar la ocurrencia de éstas en función del lugar en donde se encuentren en el texto. Generalmente, el texto se divide en párrafos y el ámbito de influencia se expresa en número de párrafos. En

el caso del dominio turístico, como se ha señalado, la estructura del corpus prevé que las instancias relacionadas con un hotel o restaurante se encuentren a continuación del mismo, y a una distancia de dos o tres párrafos. Por lo tanto, en este caso, se penalizan los elementos ontológicos que se pueden relacionar con una instancia del tipo hotel y que aparecen en párrafos precedentes, ya que lo más probable es que sean instancias de una entidad diferente.

4. La combinación con mayor ganancia de conocimiento.

La combinación con mayor ganancia de conocimiento se refiere a aquella combinación de elementos ontológicos que va a enriquecer, en mayor medida, la ontología, es decir, que va a aportar mayor conocimiento.

La ganancia de conocimiento se basa en que una instancia no aparece aislada en el texto, sino que, en el texto circundante, es decir, en el cotexto, se pueden localizar otros elementos que aporten información sobre la misma. Dicha información es susceptible tanto de ser incluida en la ontología como de ser utilizada para la desambiguación de una instancia.

El objetivo perseguido con la ganancia de conocimiento es maximizar el número de propiedades y relaciones asociadas a cada entidad, de manera que el proceso se reduce a un problema de optimización cuyo objetivo es el de maximizar la cantidad de información delimitada por el modelo ontológico.

Por ejemplo, supongamos que en el texto que se está analizando se ha identificado una anotación que puede pertenecer tanto al tipo de entidad nombrada *Hotel* como al de *Activity*. Si la entidad se clasifica como del tipo *Hotel,* aportaría a la ontología una nueva instancia de tipo hotel con 4 valores asignados a dicha instancia (por ejemplo, dirección, nº de teléfono, servicios y URL), mientras que si se clasifica como *Activity* se trataría de una instancia con sólo 3 valores asignados (por ejemplo, dirección, nº de teléfono y URL). Dado que si la instancia se anota como Hotel la ganancia de conocimiento es mayor, se selecciona esta posibilidad y no la otra.

5. Inserción de individuos en la ontología y comprobación de la consistencia.

Finalmente, una vez que se han resuelto los problemas de ambigüedad y que se ha identificado el grupo de anotaciones con la puntuación más alta, es posible iniciar el proceso de población de la ontología. Las entidades identificadas pueden ser insertadas en el modelo ontológico como individuos de propiedades o como individuos de una clase, según corresponda.

Es fundamental comprobar la consistencia de la ontología, para ello se utiliza el razonador Hermit en Protégé[21]. Cada vez que se inserta una nueva instancia, propiedad o relación el sistema comprueba que no se ha infringido ninguna regla ontológica que garantice la consistencia.

En la siguiente sección, se explica en detalle mediante un ejemplo cómo funciona el algoritmo de instanciación automática de ontologías.

5.6 Validación de la metodología en el dominio del turismo

Como se ha comentado previamente, en esta fase, a partir de las entidades reconocidas con anterioridad, se instancia la ontología. A continuación, se muestra el ejemplo para cada una de las subfases.

1. Lista de Entidades Nombradas identificadas mediante GATE

En la figura 5.12, se puede ver la descripción de un hotel ya anotado después de la fase del Procesamiento del Corpus y Reconocimiento de entidades nombradas. A cada grupo de anotaciones GATE, les asigna un color, pudiendo superponerse distintos colores. Si a una entidad nombrada se le asigna más de una anotación, entonces se producirá una ambigüedad, que gráficamente aparece

[21] http://protege.stanford.edu

representada mediante la superposición de distintos colores (los colores de las diferentes entidades a las que puede pertenecer una anotación).

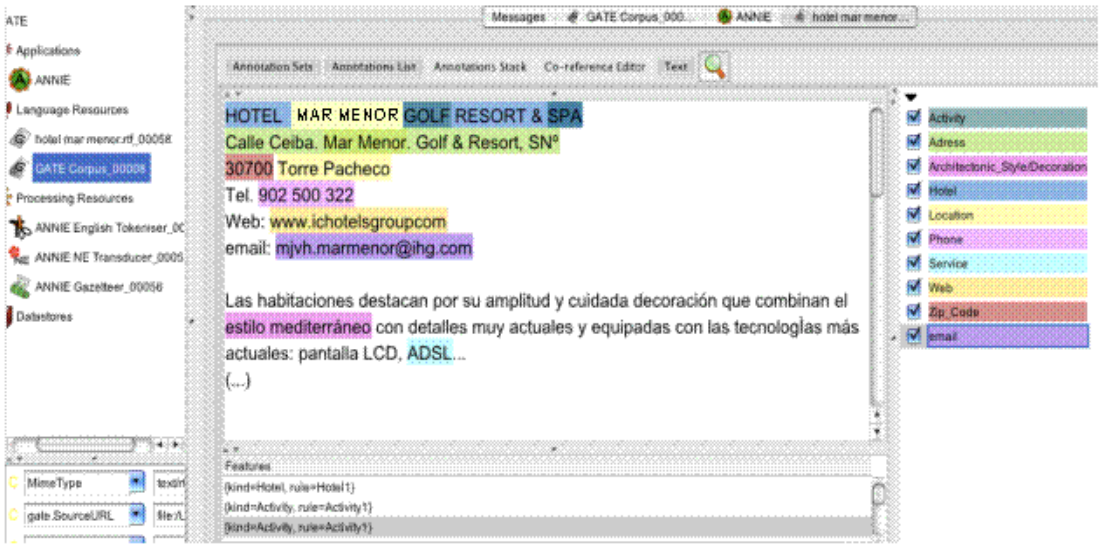


Figura 5.12 Ejemplo de anotaciones en GATE.

Por ejemplo, en la figura 5.12 se muestran las anotaciones para un fragmento del corpus Hoteles en GATE. En la primera línea, se superponen tres anotaciones diferentes, *Activity*, *Location* y *Hotel*. El resto de entidades identificadas sólo tienen una anotación posible. Por ejemplo, estilo mediterráneo es una entidad del tipo *Architectonic_StyleDecoration* y *ADSL* es una entidad del tipo *Service*.

La identificación por colores es sólo un elemento visual enfocado al usuario. Para el sistema, una anotación superpuesta es aquella cuyo token de inicio se superpone con uno o más tokens pertenecientes a otra anotación.

2. **Creación de un árbol con todas las combinaciones posibles de entidades nombradas ambiguas.**

Cuando se detecta una ambigüedad, se crean varios grupos de anotaciones. Por ejemplo, en la figura 5.13 se muestran tres grupos de anotaciones ambiguas.

Figura 5.13 Ejemplo de anotaciones ambiguas.

Por un lado, la expresión lingüística "Hotel Mar Menor Golf Resort & Spa" ha sido anotada como una entidad del tipo *Hotel*; "Mar Menor" ha sido anotada con una entidad del tipo *Location*; mientras que "Golf" y "Spa" han sido anotadas como expresiones del tipo *Activity*.

La ambigüedad surge por la superposición de las anotaciones *Hotel* con las anotaciones *Activity* y *Location*. El modo de resolver dicha ambigüedad es la creación de dos grupos separados de anotaciones, cada uno con las anotaciones de la entidad correspondiente.

Una vez que se han creado los grupos, se procede a la construcción de un árbol en el que se representa la combinación de los grupos. La función "combinatorial_tree_of_NE (NElist)" es la responsable de este paso en el algoritmo. La figura 5.14 muestra el árbol con los grupos correspondientes que se han creado en este caso, es decir, con los grupos que se muestran en la Fig 5.13. El algoritmo utiliza el árbol para recorrer todas las anotaciones desambiguadas desde la raíz hasta los nodos hoja.

En concreto, el número total de rutas que el algoritmo necesita visitar en este ejemplo es: 2 * 2* 1* =4, donde el número de la primera parte de la ecuación representa la cantidad de hermanos en cada nivel del árbol. De este modo, el algoritmo generará cuatro combinaciones diferentes de anotaciones no-ambiguas.

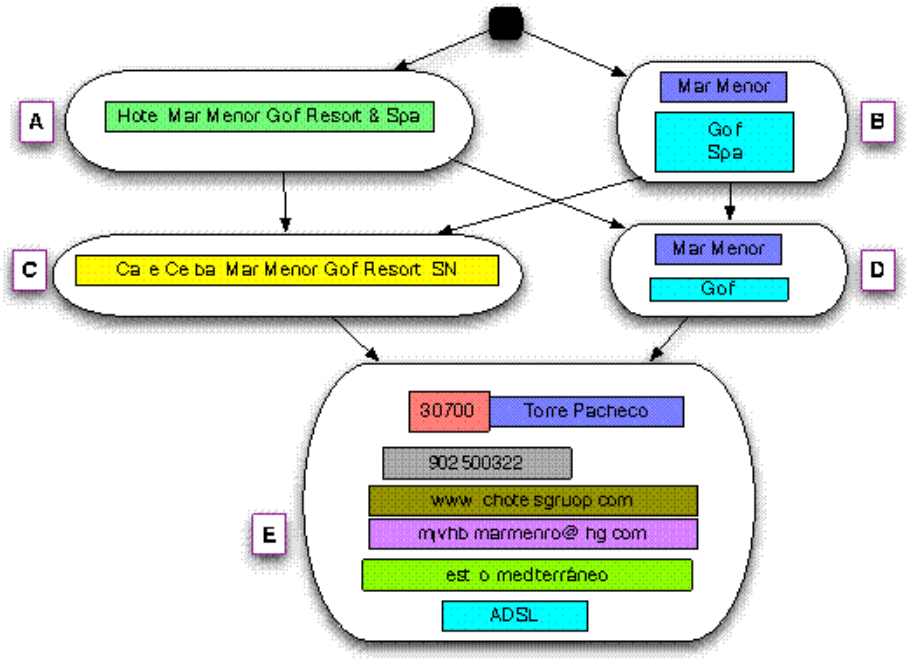


Figura 5.14 Árbol de desambiguación de entidades nombradas.

3. Cálculo de la medida (puntuación) para cada grupo de anotaciones

Esta fase, está relacionada con la función "calculate_score(NE node)" en el algoritmo. La función "calculate_score (NE node)" toma como entrada las anotaciones no-ambiguas y las mapea en la clase o propiedades correspondientes de la ontología. Una vez que todas las anotaciones han sido mapeadas con los correspondientes elementos ontológicos, la función trata de combinar cada recurso con aquellos más cercanos en el texto. Cuando dos recursos pueden combinarse, se crea una nueva relación o se le asigna un valor a un atributo. Además, la puntuación de cada anotación depende de la distancia de las entidades que se pueden relacionar en el texto.

Para calcular la puntuación de cada grupo, es necesario, en primer lugar, determinar cuáles son los pesos que están relacionados con la creación de una clase, propiedad y relación. En el ejemplo indicado más arriba, a cada parámetro se le han asignado los siguientes valores:

Wclases = 0.1

Wproperty = 0.1

Wrealtionship = 1

El peso asignado a $W_{relationship}$ debe ser mayor que los otros pesos ya que su valor se reduce por la distancia entre anotaciones en el texto. Es decir, cuanto mayor es la distancia cotextual entre dos anotaciones en el texto, menor es la probabilidad de que exista una relación entre ellas.

Una anotación dada se relaciona con aquellas anotaciones que son candidatas a ser instancias de un concepto de la ontología, con el que la primera anotación tuviera relación. Por ejemplo, si en el texto aparece una instancia de localización y la primera instancia que se encuentra con la que puede tener relación es una instancia de hotel, entonces se relacionará con dicha anotación. El peso de esta relación sería el asignado a $W_{relationship}$, 1 en este caso, dividido por el número de palabras que separan una anotación de otra. En la siguiente figura (figura 5.15), se puede ver una representación gráfica del ejemplo descrito. Entre las entidades de tipo *Location* y de tipo *Hotel,* se ha presupuesto una distancia de 4 palabras. Así mismo, las clases de la ontología con las que se relaciona la clase *Location* son *Hotel*, *Restaurant* y *Monument*.

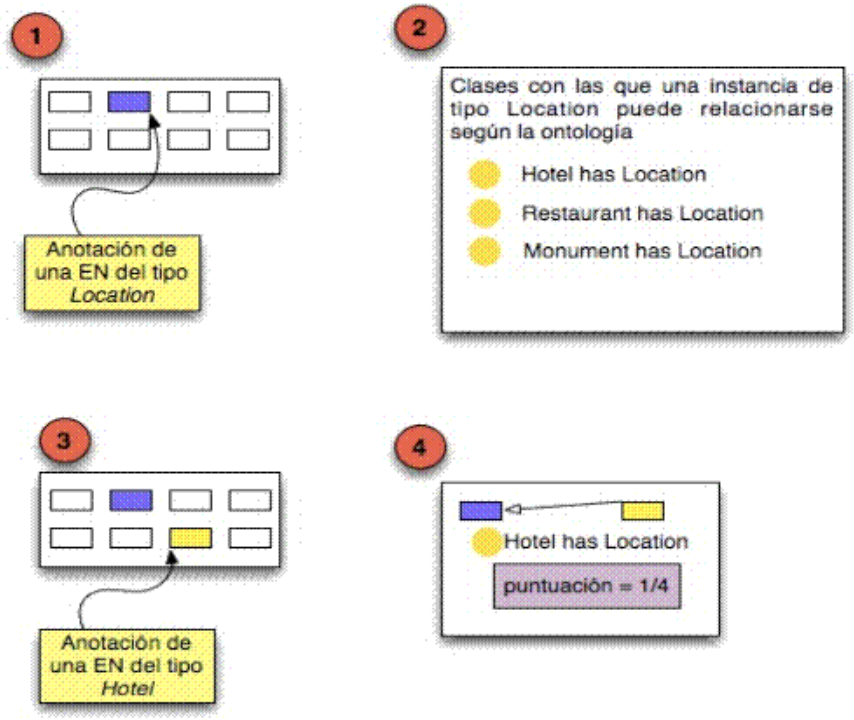


Figura 5.15 Asignación de relaciones en el texto.

Normalmente, el valor que se le suele dar a $W_{property}$ y W_{clases} es más bajo, porque la creación de una nueva instancia en la ontología, ya sea de una clase o de una propiedad, sólo aporta información si dicha instancia está relacionada con otras. Es decir, cuanto mayor es el número de relaciones que puede crear una instancia, mayor es la información que aporte.

Finalmente, en la tabla 5.14 se indica la puntuación para cada grupo de anotaciones desambiguadas que aparece en el ejemplo citado anteriormente.

Tabla 5.14 Puntuación para cada grupo.

Secuencia de Anotaciones	Puntuación Total
A, C, E	**1,31799498**
A, D, E	1,2
B, C, E	1,2
B, D, E	1,2

A continuación, se detalla cómo ha sido calculada la puntuación de cada grupo, es decir, cómo se ha asignado a cada grupo la puntuación más alta de aquellas analizadas.

Las anotaciones con las puntuaciones más altas son A,C,E. Para obtener estas puntuaciones, se invoca la función "calculate_score".

La primera anotación es Hotel "Hotel Mar Menor Golf Resort & SPA", la cual se localiza en la posición cero en el texto y está relacionada con la clase Hotel en la ontología. La puntuación de esta anotación es: 0.51799498. De dicha puntuación, 0.1 pertenece W_{class}, y el resto de la puntuación pertenece al peso de las relaciones: la relación con *Adress* **"***Calle Ceiba. Mar Menor Golf Resort, s/nº***" está separada por 5 palabras, de manera que su valor es 1/5; la relación con** *Zip Code* **"30700" está separada por 12 palabras, de forma que el valor es 1/12; la relación con** *Location* **"Torre Pacheco", que está separada por 13 palabras, tiene un valor de 1/13; la relación con** *Architectonic Style/Decoration* **"Estilo Mediterráneo" está separada por 29 palabras, entonces 1/29; y, finalmente, la relación con** *Service ADSL* **está separada por 43 palabras, por lo tanto 1/43. Además, a las $W_{property}$ y W_{Class} correspondientes, se le ha asignado una puntuación de 0.1 a cada una, y en el ejemplo propuesto se corresponden con:** *Adress* **"***Calle Ceiba. Mar Menor Golf Resort, s/nº", Zip Code "30700", Location "Torre Pacheco", Telephone "902500322", Web-page "***www.ichotelsgroup.com***",** *e-mail "***mjvhb.marmenor@ihg.com***",** *Architectonic Style/Decoration* **"estilo mediterráneo"** *y Services "ADSL".*

4. La mejor combinación

Llegados a este punto, el sistema ha evaluado qué grupo de anotaciones es el mejor mapeado en la ontología y cuál es el que provee la mayor cantidad de información relacionada extraída del texto. En la figura 5.13, se muestran las mejores combinaciones para el ejemplo mostrado en la figura 5.14.

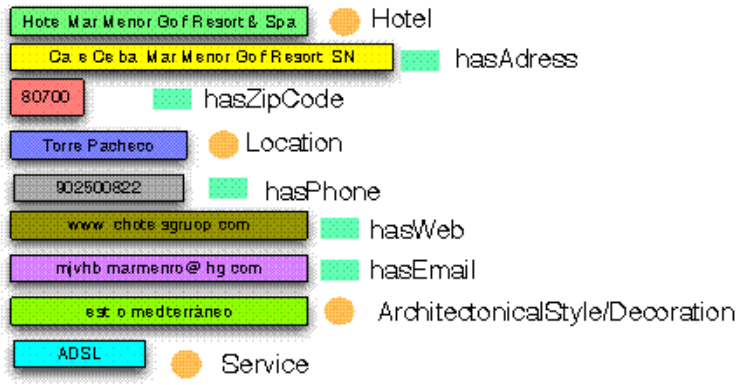


Figura 5.16 Ejemplo de la mejor combinación de entidades nombradas.

La figura 5.16 muestra cómo se han mapeado en la ontología las anotaciones que se han considerado mejores en el texto. Por ejemplo, "Hotel Mar Menor Golf Resort & Spa" se ha incluido como un individuo de la clase *Hotel*, a su vez la dirección "Calle Ceiba, Mar Menor Golf Resort SN" y la localización "Torre Pacheco" han sido relacionadas con el mismo.

5. Inserción de los individuos en la Ontología de Turismo y comprobación de la consistencia.

Una vez que se ha seleccionado el mejor grupo de anotaciones, los elementos correspondientes son insertados en la ontología. Un ejemplo de los resultados del proceso de instanciación se muestra en la Figura 5.17 mediante una captura de pantalla del editor de ontologías Protégé, en donde aparecen parte de las instancias insertadas.

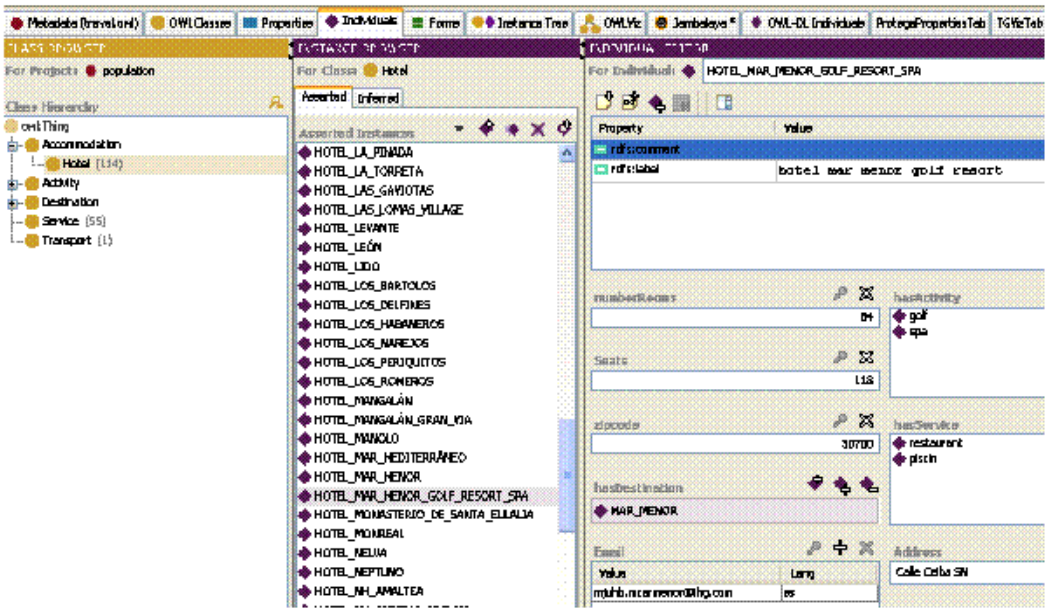


Figura 5.17 Resultado del proceso de instanciación de la ontología.

La instancia de la clase hotel que se ha insertado es HOTEL_MAR_MENOR_GOLF_RESORT_SPA, y lleva asociadas una serie de object y datatype properties:

- numberRooms: 64
- seats: 118
- zipCode 30700
- hasDestination: Mar Menor
- email: mjvh.marmenor@ihg.com
- hasActivity: golf, spa
- hasService: restaurante, piscina
- Adress: Calle Ceiba s/n

Finalmente, se ha comprobado la consistencia de la ontología para verificar que todas las instancias cumplen con las propiedades ontológicas que se han definido previamente.

5.7 Evaluación de la metodología en el dominio del turismo

A continuación, se expone cómo se ha evaluado la metodología presentada y cuáles son los resultados obtenidos.

Los experimentos se han realizado sobre los *corpora* Hoteles y Restaurantes descritos previamente. Cada corpora ha sido dividido en dos partes: una se ha utilizado para la creación de los recursos para el PLN y la otra se ha utilizado para instanciar la ontología y medir la exhaustividad y precisión del sistema. En concreto, para esta última parte, han sido utilizados el 87% del corpus Restaurantes y el 78% del corpus Hoteles.

Para la evaluación de la metodología, se han llevado a cabo dos experimentos independientes. Por un lado, se ha evaluado la capacidad del sistema en cuanto a identificación, extracción y clasificación de entidades nombradas se refiere. Y, por otro lado, se ha evaluado la capacidad del sistema para incluir instancias en la ontología, es decir, para llevar a cabo el proceso de instanciación.

El éxito de la tarea de instanciación de la ontología está directamente relacionado con dos factores:

1. La cantidad de entidades nombradas extraídas mediante GATE durante la primera parte del proceso.
2. La capacidad del sistema en la resolución de ambigüedades, es decir, su habilidad para clasificar candidatos a entidades nombradas.

En un primer experimento, se han evaluado los resultados obtenidos durante la fase de extracción de entidades nombradas, valorando tres aspectos diferentes que se muestran en la tabla 5.15.

Esta tabla incluye aquellas menciones de entidades nombradas que GATE identificó como ambiguas, es decir, aquellas anotaciones que se pueden mapear con más de una entidad ontológica y/o aquellas que están solapadas; aquellas anotaciones que el sistema finalmente no pudo desambiguar con la ayuda de la

ontología, es decir aquellas entidades todavía ambiguas después del proceso de instanciación; finalmente, se muestran las anotaciones de entidades nombradas correctas.

Tabla 5.15 Entidades Nombradas extraídas

	Restaurantes	**Hoteles**
Anotaciones de EN ambiguas	123	73
Anotaciones de EN ambiguas después del proceso de instanciación	59	8
EN correctamente anotadas	13072	2958

Del corpus Restaurantes, se han extraído correctamente 13.072 anotaciones de entidades nombradas, de las cuales 123 fueron ambiguas y solamente 59 no pudieron ser desambiguadas después del proceso de población de la ontología, mientras que del corpus hoteles se extrajeron 73 entidades ambiguas, de las cuales 8 no se pudieron desambiguar. Las entidades correctas extraídas han sido 2958.

Por un lado, el número de anotaciones de entidades nombradas correctamente identificado por GATE representa el 99% del total de entidades nombradas anotadas. Esta información se refiere a los grupos que contienen una sola anotación de una entidad nombrada dada y, en consecuencia, no presentan ambigüedades, lo que representa la mayor parte de los casos. Por el contrario, las anotaciones que presentan alguna ambigüedad, es decir, aquellas anotaciones que se pueden mapear con más de un entidad nombrada se agrupan para su desambiguación.

Proporcionalmente, el número de ambigüedades es mayor en el corpus Hoteles. Esto se debe principalmente a que en muchos casos los nombres de los hoteles incluyen elementos que se pueden clasificar en diferentes tipos de entidades nombradas , como por ejemplo Localización o Actividad. La mayoría de estas

ambigüedades se pueden resolver con la metodología propuesta en este trabajo. Sin embargo, existen algunas anotaciones de entidad nombrada cuya ambigüedad no se resuelve durante el proceso de población de la ontología. El número de ambigüedades no resueltas es mayor en el corpus Restaurantes. La razón es que los descriptores son más cortos y el número de posibles relaciones entre ellos es menor. En consecuencia, en algunos casos no hay una diferencia entre la ganancia de conocimiento si el sistema elige una entidad u otra, de manera que el sistema no puede determinar con certeza a qué clase pertenece una entidad dada.

Se ha llevado a cabo un segundo experimento para evaluar los resultados obtenidos después del proceso de instanciación de la ontología.

En primer lugar, se han extraído las entidades ontológicas relevantes (individuos, *object properties* y *data properties*) que aparecen en los corpora, de manera manual.

En segundo lugar, se ha obtenido información a cerca del número de entidades de conocimiento recuperados y las entidades correctamente recuperadas por la metodología propuesta. Por otro lado, se ha comprobado qué instancias de las propiedades han sido correctamente creadas y qué instancias de las clases se han insertado correctamente en la ontología. Finalmente, se muestra la suma de todas las propiedades correctamente instanciadas en el corpus hoteles y en el corpus Restaurantes. Los resultados obtenidos se muestran en la tabla 5.16.

Tabla 5.16 Resultados del proceso de instanciación.

Evaluación	**Entidades de Conocimiento relevantes en los corpora**	**Entidades Recuperadas**	**Entidades de Conocimiento correctamente recuperadas**
Instancias de Restaurantes	1320	1195	1179
Instancias de Hoteles	217	221	206
Datatype Proerties Restaurantes	1907	1665	1554
Datatype Properties Hoteles	421	347	335
Object Porperties Restaurantes	5959	5267	4898
Object Porperties Hoteles	2251	1992	1804
Total propiedades Restaurantes	7866	6932	6452
Total propiedades Hoteles	2672	2339	2139

La diferencia entre el número de instancias de ambos corpora es significativo. Así, mientras el corpus restaurantes contiene 1320 instancias de las que se han recuperado 1179, el corpus Hoteles contiene 217 de las que se han recuperado 206. Esta diferencia se mantiene en el resto de categorías, como se puede observar gráficamente en la figura 5.18.

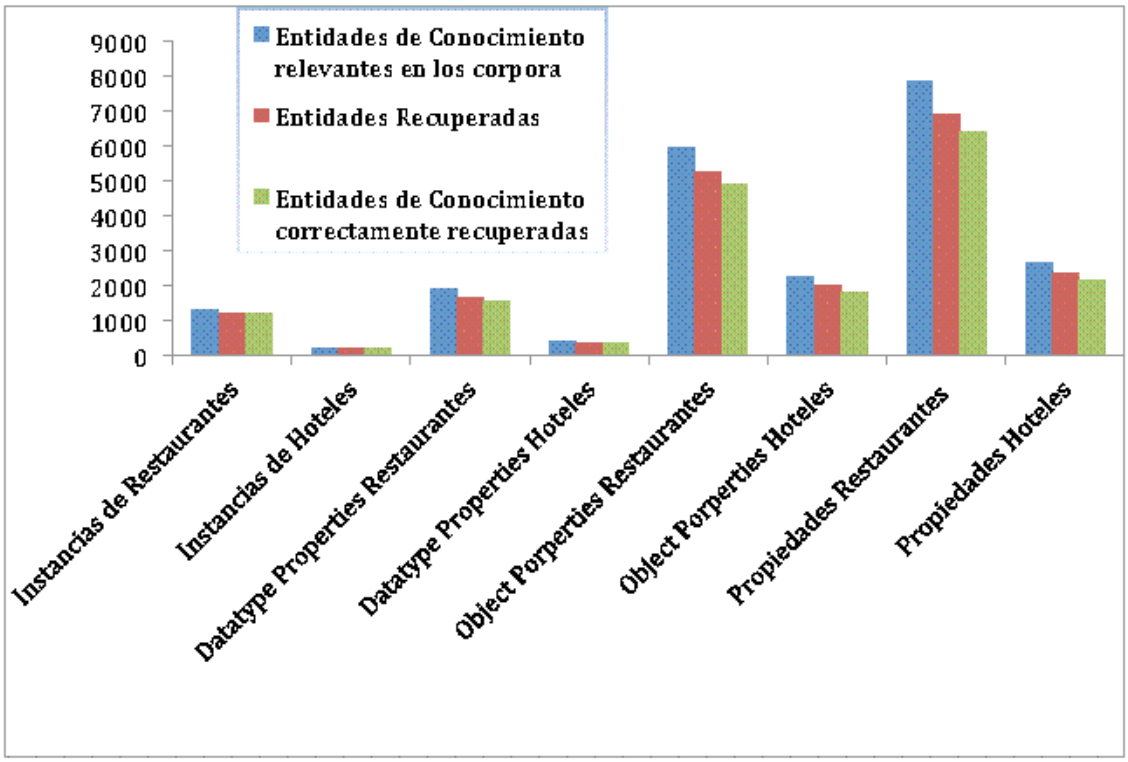


Figura 5.18 Valores y medidas obtenidos durante la evaluación.

Las medidas estándar que se han utilizado para la evaluación del sistema son la exhaustividad (1) y la precisión (2). El número total de entidades de conocimiento se ha comparado con aquellas que fueron relevantes. Por otro lado, la precisión se ha calculado utilizando la entidades relevantes recuperadas y aquellas que fueron irrelevantes. Finalmente, la medida-F (3) es el promedio ponderado de la exhaustividad y la precisión. El resultado de estas medidas de evaluación se indica en la tabla 5.17.

$$(1)\ \text{Exhaustividad} = \frac{\text{Entidades de Conocimiento Correctamente Recuperadas}}{\text{Entidades de Conocimiento en el Corpus}}$$

$$(2)\ \text{Precisión} = \frac{\text{Entidades de Conocimiento Correctamente Recuperadas}}{\text{N° total de Entidades de Conocimiento en el Corpus}}$$

$$(3)\ \text{Medida de F} = \frac{2\ (\text{Exhaustividad} \cdot \text{Precisión})}{\text{Exhaustividad} + \text{Precisión}}$$

Tabla 5.17 Exhaustividad, Precisión y Medida de F.

	Corpus Restaurantes			Corpus Hoteles		
	Exhaustividad (%)	Precisión (%)	Medida de F (%)	Exhaustividad (%)	Precisión (%)	Medida de F (%)
Instancias	89.32	98.66	93.76	94.93	93.21	94.06
Object Properties	82.19	92	87.26	80.14	90.56	85.03
Datatype Properties	81.49	93.33	87.01	79.57	96.54	87.24
Total	84.10	95.66	89.51	81.19	92.35	86.85

Los elevados valores de exhaustividad y precisión obtenidos en ambos corpora se deben a dos razones, fundamentalmente. Por un lado, ambos corpora son muy específicos y, en consecuencia, las entidades de conocimiento que los representan

son acotables. Por otro lado, el análisis lingüístico que se ha llevado a cabo ha permitido la creación de recursos lingüísticos para el PLN adaptados a dichos corpora.

El mejor valor obtenido ha sido la precisión en las instancias que se han extraído del corpus Restaurantes, con un 98,66%. Como se ha dicho, el carácter estructurado del corpus, la repetición de patrones lingüísticos y el hecho de que las instancias no aparezcan aisladas en el texto, ha favorecido el aumento de la precisión. Por el contrario, el carácter publicitario de los textos prevé que, en ocasiones, se incluyan estructuras más originales para la presentación de entidades como, por ejemplo, del tipo *Specialities*. Esto ha supuesto un descenso en la exhaustividad, con un valor del 89,32%, ya que el sistema no ha sido capaz de identificar este tipo de instancias por no disponer del modelo lingüístico adecuado.

En cuanto al corpus Hoteles, el valor más bajo se ha obtenido para las *Datatype Properties*, con un valor de 79,57%. Dado que las instancias de la mayoría de estas relaciones son del tipo teléfono, dirección, email, etc., que suponen el grueso de la información contenida en el corpus, es normal que este valor descienda, ya que, sobre todo las instancias de la relación *hasAdress* son complejas de desambiguar.

En cuanto al valor más alto en el corpus Hoteles, es de nuevo la exhaustividad de las instancias de las clases. Los motivos son los mismos que se han mencionado previamente para el corpus Restaurantes.

De igual modo, el valor de la medida F de los individuos y las relaciones muestran que la ontología fue correctamente instanciada, con una media del 89% en el corpus Restaurantes y del un 86% en el corpus Hoteles.

Con un valor medio de la medida F de 89,51% , el rendimiento del sistema en el dominio de los restaurantes es un poco mejor que en el dominio de los hoteles.

Esta ligera variación en el rendimiento puede ser debida al grado diferente de especificidad de sus contextos textuales.

CAPÍTULO 6

INSTANCIACIÓN DE ONTOLOGÍAS BASADA EN ROLES SEMÁNTICOS

APLICACIÓN AL DOMINIO DE LA BIOMEDICINA

Resumen. En este capítulo, se presenta una metodología para la instanciación automática de ontologías de dominio haciendo uso de roles semánticos. Esta metodología, que se ha validado en el domino de la biomedicina, se basa en la integración de distintos recursos ontológicos y léxicos. Dos ontologías de alto nivel, que definen relaciones semánticas básicas, se han mapeado con marcos semánticos provenientes de FrameNet. El resultado es un modelo ontológico denominado BioOntoVerb OM, que contiene relaciones de alto nivel y *frames* asociados a las mismas. El proceso de instanciación se lleva a cabo, por un lado, mediante la identificación de entidades nombradas, que se convierten en candidatas a instancias de la ontología, y por otro lado mediante la identificación de relaciones entre ellas. Cuando dos entidades nombradas se relacionan a través de las unidades léxicas incluidas en BioOntoVerb OM, dicha relación se incluye provisionalmente como instancia de una *ObjectProperty,* y las instancias implicadas en la misma se convierten en candidatas a instancias de la ontología. Finalmente un razonador comprueba la consistencia de la ontología, de modo que si la ontología es consistente, las clases y propiedades correspondientes son instanciadas.

6.1 Introducción

El crecimiento exponencial de la literatura científica en el dominio biomédico genera una dificultad cada vez mayor a la hora de acceder selectivamente a la información de interés para sus usuarios (principalmente investigadores) y de obtener datos actualizados relacionados con este campo.

La Bioinformática es la disciplina que se encarga de la gestión de datos biomédicos asistida por el ordenador. Algunos de sus objetivos son recuperar, analizar y representar la información de manera que facilite el entendimiento, por

parte de los científicos, de los procesos vitales así como la investigación en nuevos y mejores medicamentos (Persidis, 1999), favoreciendo así mismo el avance en campos como el de la genómica.

En consecuencia, uno de los principales retos es la creación de herramientas de minería de datos en general y textos en particular que sean capaces de recuperar la información relevante en los cada vez más nutridos repositorios de información biomédicos.

Desde su inicio, las ontologías han ido captando el interés de la comunidad biomédica, ya que poseen una estructura jerárquica que permite, tanto describir conceptos de alto nivel como conceptos muy específicos. Su estructura organizacional permite además almacenar el conocimiento y compartirlo, facilitando el acceso y recuperación al mismo. De hecho, se han desarrollado diversos sistemas como el descrito en (Wächter et al., 2008), basados en ontologías para la búsqueda en la literatura científica. De entre las numerosas ontologías que se han desarrollado en esta área, es la *Genome Ontology* (GO)[22] (Ashburner, et al., 2000) la que está teniendo una mayor repercusión en la comunidad científica.

GO describe los procesos biológicos, las funciones moleculares y los componentes celulares de los productos génicos.

El *Gene Ontolgy Consortium* tiene como objetivo principal producir un vocabulario controlado y dinámico que pueda aplicarse a todas las eucariotas, incluyendo el conocimiento cambiante de los genes y proteínas. Con este fin, se han creado tres ontologías independientes, una sobre procesos biológicos, otra sobre funciones moleculares y otra sobre componentes celulares. Las ontologías pueden accederse en la URL http://www.geneontology.org.

No obstante, el mantenimiento y actualización de todos los vocabularios y ontologías es inabarcable con una metodología completamente manual. Por este

[22] http://www.geneontology.org/

motivo han surgido, de forma paralela al desarrollo de las ontologías, proyectos que permiten el enriquecimiento automático de las mismas.

El objetivo de este capítulo es presentar una metodología para obtener y clasificar automáticamente instancias biomédicas extraídas de textos en lenguaje natural, aprovechando las propiedades semánticas de un modelo ontológico definido en el lenguaje OWL2.

Para la definición del modelo ontológico, se han integrado una serie de recursos. Estos recursos son, por un lado, la ontología de relaciones de OBO (Smith et al., 2005) y la ontología BioTop (Beisswanger et al., 2008) y, por otro, los *frames* de FrameNet. OBO y BioTop son dos ontologías de alto nivel o *top level ontologies* en las que se expresan las relaciones genéricas del domino biomédico. Para cada una de las *Object Properties,* se han definido un conjunto de axiomas.

La ontología resultante se ha mapeado con *frames* o marcos semánticos extraídos de FrameNet (Baker et al., 1998, Miller, 1995), es decir, las relaciones de la ontología tienen asociado un conjunto de marcos semánticos. Los *frames* o marcos semánticos son representaciones esquematizadas de situaciones del mundo real en base a las cuales se organiza la información. Aquellas relaciones ontológicas que tienen asociado un *frame,* contienen tanto expresiones lingüísticas, que representan dichas relaciones a un nivel textual, como unos roles semánticos que describen cada relación. El modelo ontológico resultante es BioOntoVerb OM. A partir del mismo, y junto con el reconocedor de entidades nombradas de GENIA (Tsuruoka et al., 2005), se extraen las instancias de la ontología.

El capítulo se estructura de la siguiente manera. En primer lugar, se describen las ontologías de alto nivel y de dominio biomédico que forman parte de la metodología propuesta, esto es, OBO *relation ontology* y BioTop, por un lado, y la ontología GENIA y XGENIA, por otro. A continuación, se describen algunos

copora biomédicos, entre ellos el corpus GENIA, un corpus anotado ampliamente utilizado en la comunidad científica y que en nuestro caso se utiliza para la validación de la metodología. La integración de estos recursos en el marco de trabajo propuesto da como resultado el modelo ontológico BioOntoVerb.

Finalmente, se describe el proceso de instanciación de la ontología, que consta de tres fases principales y la validación de la metodología. En la última parte del capítulo se ponen de relieve algunas conclusiones.

6.2 Ontologías biomédicas o bio-ontologías

Las Bio-Ontologías formalmente representan las relaciones entre conceptos biológicos definidos. Son recursos compartidos cuyo continuo desarrollo y la incorporación en sistemas bioinformáticos y biocomputacionales permite la integración de información científica y facilita el descubrimiento de conocimiento (Blake, 2004).

En este trabajo, se ha tomado como base la clasificación de ontologías realizada por Guarino (1998), en la que se distingue entre Ontologías de alto nivel (Top-level ontologies) , Ontologías de domino (*Domain ontologies*), Ontología de Tarea (*Task ontology*) y Ontología de aplicación (*Application ontology*). Términos a los que ya se ha hecho referencia en el apartado 2.3.2.

La figura 6.1 es una representación gráfica de la clasificación realizada por Guarino (1998) y de cómo las ontologías se relacionan entre ellas. La metodología presentada en esta memoria, hace uso de ontologías de alto nivel, ontologías de dominio y ontologías de aplicación, como se explica en los siguientes apartados.

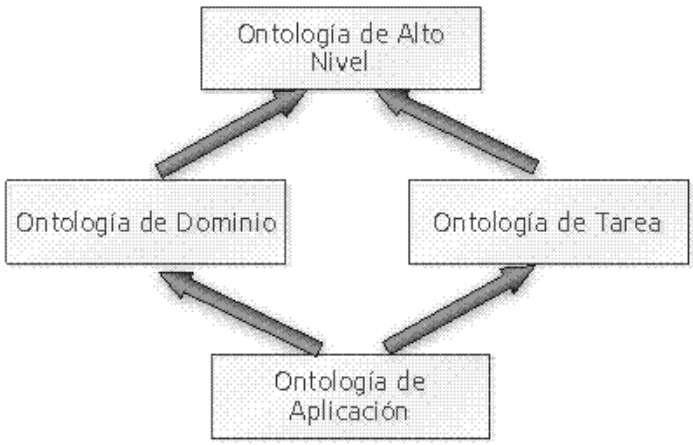


Figura 6.1 Tipos de Ontologías.Traducido y adaptado de (Guarino, 1998)

En las ontologías de alto nivel, se describen conceptos generales tales como localización, espacio o tiempo. Para el desarrollo de la metodología, se parte de dos ontologías de alto nivel muy extendidas dentro del dominio biomédico, la ontología de relaciones OBO (Smith et al., 2005) y la ontología BioTop (Beisswanger et al., 2008).

Por otro lado, una ontología de dominio describe el vocabulario de un domino específico y se considera una especialización de una ontología de alto nivel. Como se verá más adelante, para la validación de la metodología se ha utilizado la ontología de domino xGENIA (Rak et al., 2007).

Finalmente, el resultado de la combinación del modelo ontológico creado, con la ontología xGENIA es una ontología de aplicación, esto es, una ontología especialmente adaptada para llevar a cabo una tarea que, en este caso, es la instanciación automática a partir de textos en lenguaje natural.

A continuación, se describen las ontologías relevantes para la metodología. OBO *relations Ontology* (Smith et al., 2005), BioTop (Beisswanger et al., 2008), GENIA *ontology* (Kim et al., 2006a) y xGENIA (Rak et al., 2007).

6.2.1 Ontologías de alto nivel o metaontologías en biomedicina

Con la proliferación de ontologías y bases de datos en el dominio de la biología, es necesario clasificar las ontologías, y esto es lo que pretenden las ontologías de alto nivel, también llamadas *metaontologías*.

Las ontologías de alto nivel sirven de anclaje para las ontologías de dominio, de manera que se pueda reutilizar y compartir el conocimiento contenido en las mismas. De este modo, se evita que las ontologías de dominio permanezcan incomunicadas entre sí y que el mismo conocimiento se defina repetidamente (Pasquier, 2008). En este trabajo, nos centramos en la ontología de relaciones OBO *relations ontology* (Smith et al., 2005) y en las relaciones expresadas en BioTop (Beisswanger et al., 2008).

6.2.1.1 OBO (Open Biological Ontologies)

La librería de ontologías OBO es un repositorio de vocabularios controlados desarrollados para un uso compartido en los dominios biológico y médico.

El proyecto OBO (Open Biomedical Ontologies)[23] se centra en integrar ontologías relativas a los genes y las proteínas. Entre los criterios de inclusión, se encuentra que la ontología sea abierta, que esté editada mediante sintaxis GO u OWL, que tenga definiciones, identificadores únicos y que se complemente con otras ontologías de OBO (Baclawski & Niu, 2005).

No obstante, como señalan Smith et. al (2005) existen algunas limitaciones en las relaciones definidas tanto en OBO como en GO. Por ejemplo, en GO la relación "part of" representa simultáneamente tres tipos de relaciones, a saber, la relación de inclusión, la relación de una posible relación parte-todo entre

[23] http://www.obofoundry.org/

entidades biológicas y una relación parte-todo necesaria entre entidades biológicas.

Dadas estas limitaciones, en Smith et al. (2005) se propone un conjunto de relaciones, conocido como *OBO Relations Ontology* (http://www.obofoundry.org/ro/) en el que se describen formalmente 10 relaciones generales, incluyendo taxonómicas y partonómicas, aplicables a las ontologías de dominio biomédico.

En primer lugar, estos autores distinguen tres tipos de relaciones binarias:

- **Clase-Clase**: Se trata de la relación taxonómica que se establece entre las clases de la ontología expresada como is_a
- **Instancia- Clase**: Es la relación entre una instancia y una clase, un ejemplo es *instance_of*
- **Instancia-Instancia**: Es la relación que se establece entre las instancias de una clase, por ejemplo, una relación de carácter partonómico como *part-of*

Las relaciones propuestas en OBO *relations ontology* incluyen los distintos tipos de relaciones binarias mencionadas que, a su vez, se dividen en cuatro grupos: *Foundational relations* (relaciones fundacionales), *Spatial relations* (relaciones espaciales), *Temporal relations* (relaciones temporales) y *Participation relations* (relaciones de participación).

A continuación, se describe brevemente cada una de ellas.

- **Foundational relations.** Se trata de las relaciones básicas ontológicas:
  - ***is_a***
  - ***part_of***

Las relaciones de taxonomía y partonomía son el eje sobre el que se articulan la mayoría de las ontologías. La extracción de conceptos relacionados taxonómicamente ha sido un tema ampliamente tratado en la literatura, como por

ejemplo en (Cimiano et al., 2005; Welty, 2001; Ceusters et al., 2003; Hahn et al., 1999), por lo que existen sistemas capaces de llevar a cabo esta tarea con relativo éxito.

La metodología que se propone en esta tesis no se adecúa a la extracción de relaciones taxonómicas y sólo parcialmente a las partonómicas, ya que el conjunto de roles semánticos que se podría asociar a este tipo de relaciones es demasiado amplio y las expresiones lingüísticas ligadas a dichos roles presentaría un alto grado de ambigüedad por se demasiado generales.

La extracción de relaciones partonómicas sólo se ha llevado a cabo cuando las expresiones lingüísticas asociadas a roles semánticos poseen un grado de ambigüedad bajo. Un ejemplo de este tipo de relación es "contained_in", que se puede ver como una relación de localización pero también de partonomía.

La metodología aquí presentada se centra en relaciones más específicas, y, por otro lado, menos tratadas en la literatura. No obstante, tanto las relaciones taxonómicas como partonómicas son tenidas en cuenta por el razonador, siendo relevantes para la consistencia de la ontología.

- **Spatial relations**: Conectan una entidad con otra en términos de las relaciones entre regiones espaciales que ocupan. Las relaciones son las siguientes:
  - ***located_in,*** en donde cada entidad está asociada en todo momento con una región espacial determinada, es decir, esta relación indica la situación exacta de una entidad.
  - ***contained_in,*** en la que una entidad forma parte de otra que actúa como la entidad contenedora.
  - ***adjacent_to***, en la cual una entidad mantiene una relación espacial de proximidad con otra.

- **Temporal relations**: Conectan entidades existentes en diferentes momentos, y son las siguientes:
  - ***transformation_of*** En donde una entidad es la transformación de otra. La entidad resultante mantiene las mismas propiedades que la primera, pero la existencia de ambas se da en momentos diferentes. Por ejemplo un embrión y un feto.
  - ***derives_from*** En donde una entidad deriva de otra y la entidad resultante posee nuevas características que la primera no poseía.
  - ***preceded_by*** En donde una entidad precede a otra en el tiempo.

- **Participation relations:** Conectan procesos con las entidades que los llevan a cabo. Las relaciones son las siguientes:
  - ***has_agent,*** es una relación que se produce entre una instancia y un proceso cuando la instancia toma parte activamente en el proceso
  - ***has_participant,*** es una relación que se produce entre una instancia y un proceso cuando la instancia toma parte en un proceso en un momento determinado.

La representación de este tipo de relaciones a un nivel textual es difícil de sistematizar, ya que, por ejemplo, las relaciones fundacionales son demasiado generales y es difícil encontrarlas en textos en lenguaje natural. Por lo tanto, aunque no se han excluido del modelo ontológico, no se han asociado roles a las mismas, ya que los individuos que en ellas participan no se obtienen directamente del texto, sino mediante la inferencia lógica llevada a cabo por el razonador.

En consecuencia, los dos grupos de relaciones de la ontología OBO que se han considerado son las relaciones espaciales y temporales. Estas relaciones

representan un total de 6 a las que se ha añadido otra relación significativa a partir de la ontología BioTop (Beisswanger et al., 2008), que se comenta a continuación.

6.2.1.2 BioTop

Otra aproximación en la que se representan formalmente relaciones entre ontologías es BioTop (Beisswanger et al., 2008), que es una ontología de alto nivel aplicada al dominio de la ciencia de la vida en la que se trata de unir e integrar varias ontologías de dominio específico. Aunque originalmente fue una extensión de la ontología GENIA (ver apartado 6.2.2.1), en la actualidad posee nuevas categorías relativas a la medicina y a la salud.

De entre las relaciones descritas en BioTop, algunas son equivalentes a las descritas en la ontología de relaciones de OBO. Sin embargo, otras son más específicas.

Dentro del marco de BioTop, se llevó a cabo un mapeo entre las ontologías OBO *relations ontology*, BioTop y Basic Formal Ontology[24] (BFO). La ontología resultante está disponible en http://www.imbi.uni-freiburg.de/ontology/BioTop/ y el mapeo entre las *Object Properties* se puede ver en la figura 6.2.

No obstante, en la ontologías de relaciones de OBO no están presentes todas las relaciones de la ontología BioTop. En esta última, el número de relaciones es mayor y su grado de abstracción es más bajo, siendo menos generales.

Para este trabajo, además de seleccionar aquellas relaciones de OBO mapeadas con BioTop, se ha añadido la relación de causalidad, es decir, la relación *caused_by* presente en la ontología BioTop.

La ventaja de este mapeo es que el modelo ontológico resultante será válido tanto para aquellas ontologías de dominio que se enlacen con OBO como para aquellas que se enlacen con BioTop.

[24] http://www.ifomis.org/bfo

Figura 6.2 Mapeo entre las relaciones de OBO, BioTop y BFO.

6.2.2 Ontologías de domino en biomedicina

Existe un gran número de ontologías de domino en el área de la biomedicina, ya que su desarrollo va unido a proyectos, metodologías o subcorpus particulares. Solamente en OBO se recogen decenas de ontologías de dominio biológico. OBO se divide en dos apartados principales, uno en donde se puede acceder a las denominadas OBO *foundry ontologies*, es decir, aquellas ontologías que han sido desarrolladas o mapeadas siguiendo unos parámetros definidos y que presentan un alto grado de normalización. El otro apartado está dedicado a las ontologías aportadas por la comunidad científica que aún no forman parte de las ontologías principales de OBO pero que pueden ser de interés. En la siguiente figura, se pueden ver algunas de las ontologías contenidas en el repositorio de OBO dedicado a los aportes de la comunidad y en donde se indica el título de la

ontología, el dominito específico al que pertenecen, el prefijo con el que se las denomina, desde dónde se puede acceder a ellas y la fecha de la última actualización.

OBO Foundry candidate ontologies and other ontologies of interest

Title	Domain	Prefix	File	Last changed
Adverse Event Reporting Ontology	health	AERO		
Amphibian gross anatomy	anatomy	AAO	AAO_v2_edit.obo	
Amphibian taxonomy	anatomy	ATO	amphibian_taxonomy.obo	
Anatomical Entity Ontology	anatomy	AEO	aeo.obo	2011/01/17
Ascomycete phenotype ontology	phenotype	APO	ascomycete_phenotype.obo	2011/03/28
Basic Formal Ontology	upper	BFO	1.1	
Bilateria anatomy	anatomy	BILA	bilateria_nmca.obo	
Biological imaging methods	experiments	FBbi	image.obo	2011/05/24
BRENDA tissue / enzyme source	anatomy	BTO	BrendaTissueOBO	
C. elegans development	anatomy	WBls	worm_development.obo	2008/01/31
C. elegans gross anatomy	anatomy	WBbt	WBbt.obo	
C. elegans phenotype	phenotype	WBPhenotype	worm_phenotype.obo	2011/09/19
Cell type	anatomy	CL	cell.obo	2011/08/24
Chemical Information Ontology	biochemistry	CHEMINF	cheminf.owl	
Common Anatomy Reference Ontology	anatomy	CARO	caro.obo	2011/09/12

Figura 6.3 Repositorio de ontologías de dominio biomédico en OBO.

En este trabajo, nos centraremos en la descripción de dos ontologías de dominio, a saber, GENIA y xGENIA. Ésta última, que está basada en GENIA, se ha utilizado para la evaluación de la metodología propuesta en esta tesis.

6.2.2.1 GENIA Ontology

La ontología GENIA (Kim et al., 2006a) es una taxonomía diseñada como soporte para las anotaciones semánticas del corpus GENIA (ver apartado 6.3.1). Esta ontología, cuya versión más actual es la 2.0 se divide a su vez en dos ontologías: la ontología terminológica y la terminología de eventos. La ontología terminológica, GENIA *term ontology*, está compuesta por tres subontologías que cubren las anotaciones terminológicas referentes a las reacciones químicas, organismos y partes anatómicas de los organismos.

Dado que se puede establecer una relación casi directa entre los conceptos de la ontología terminológica y las categorías de MeSH (NLM, 2011), se ha llevado a

cabo un mapeo entre ambos. Dicho mapeo ha permitido la extracción de listas de términos que se han utilizado para mejorar la exhaustividad del reconocedor de entidades nombradas de GENIA y que forma parte de la metodología propuesta. En la tabla 6.7 del apartado 6.4.3, en donde se describe la extracción de entidades nombradas, se puede ver el mapeo entre las clases de la ontología GENIA y las categorías del MeSH (NLM, 2011).

Por otro lado, la ontología de eventos da soporte a la anotación de eventos en el corpus GENIA, estando conectada con la ontología GO (Kim et al., 2008).

6.2.2.2 xGENIA Ontology

La ontología de GENIA es en realidad una taxonomía, de manera que la riqueza de las relaciones semánticas es escasa. Con el objetivo de superar esta limitación, en (Rak et al., 2007) proponen la ontología xGENIA, que se basa en el corpus GENIA y está desarrollada en OWL.

Las clases de la ontología xGENIA están constituidas por la taxonomía original de GENIA, es decir, por las 47 clases que se definen en la ontología GENIA original. Sin embargo, añaden nuevas relaciones utilizando para ello verbos, ya sea en forma nominalizada o en infinitivo, expresados mediante *Object Properties*.

En cuanto a los individuos, éstos se han obtenido a partir de las entidades biológicas anotadas en el corpus GENIA. A cada entidad biológica o individuo se le ha asignado un identificador único, garantizando que cada uno se inserta sólo una vez en la ontología, aunque en el texto aparezca expresado de distintas formas. El nombre original de la entidad en el corpus se recoge en las etiquetas *rdfs:label*.

Aunque OWL permite la asignación de más de una clase a un individuo, esto no ocurre en el corpus GENIA, en el que cada instancia está asociada a una única clase.

En xGENIA, cada relación binaria relaciona dos individuos a través de un verbo o de la forma nominalizada de un verbo, esto es, cada predicado se representa en la ontología como un *owl:ObjectProperty* y tiene su propio dominio *rdfs:domain* y rango *rdfs:range*. Además, las propiedades se han organizado en una jerarquía utilizando *rdfs:subPropertyOf* para indicar que una propiedad es la variante de otra.

Para facilitar la identificación de algunas entidades, se les ha asignado un Identificador Único de Concepto (*Concept Unique Identifiers (CUI)),* extraído del metatesauro de UMLS (NLM, 2008). Las entidades se han relacionado con los identificadores por medio de *datatype properties*.

De este modo, se ha conseguido una ontología con una riqueza semántica mayor que la que tiene la ontología GENIA original. La ontología además está instanciada, lo que permite realizar la evaluación de sistemas dedicados a la instanciación automática de ontologías sin la validación de un experto.

En la siguiente tabla (tabla 6.1), se puede ver información técnica de la ontología.

Tabla 6.1 Estadística de xGENIA. Adaptado y traducido de (Rak. et al. 2008)

Clases	47
Object Properties	142
Individuos (entidades biológicas)	34.842
Relaciones entre individuos	7.174
Taxonomía léxica (StemsFrom)	10.386
Individuos unidos con CUI (directamente/indirectamente)	14.700 (6.851/7.849)

En la ontología de Rak et al. (2008), se han incluido las etiquetas anidadas presentes en el corpus GENIA, en las que se indica la raíz léxica (el *stem*) de las

entidades anotadas. Se ha creado una relación, denominada *StemsFrom,* que une la entidad con su raíz, aumentando de este modo el número de individuos de la ontología.

Por otro lado, las instancias de la ontología se han mapeado con los identificadores únicos de UMLS (CUIs). El mapeo, tanto de los individuos como de los componentes de la ontología léxica, ha generado un mayor número de relaciones, de forma que los individuos que directamente han sido mapeados con una entrada de UMLS son 6.851 e, indirectamente, son 7.849.

6.3 Corpora biomédicos

Así como existen numerosas ontologías de dominio biomédico, no ocurre lo mismo con los corpora[25] anotados disponibles. Las bases de datos médicas como PubMed o BioMed, son la fuente utilizada generalmente para la extracción de resúmenes o textos completos de la literatura científica biomédica.

BioMed (BioMed, 2011), es un corpus formado por más de 96.000 artículos. Está disponible para su descarga en formato XML. Por otro lado, **PubMed** (NCBI, 2011), desarrollado por la Biblioteca Nacional de Medicina de EEUU, permite realizar búsquedas en las más de 20 millones de referencias bibliográficas recogidas en MEDLINE[26], revistas científicas y libros online. Los textos de PubMed/MEDLINE se han utilizado para la creación de corpora como PennBioIE (Mandel, 2006), dedicado al dominio oncológico, GENIA (Kim et al., 2003) o Biotext (Schwartz & Hearst, 2003).

[25] El concepto de corpus especializado se ha descrito en el apartado 5.1.1 de este trabajo.

[26] http://www.nlm.nih.gov/bsd/pmresources.html

Otro de los recursos que se han creado en los últimos años es **BioInfer** (Pyysalo et al., 2007). Se trata de un corpus en el que se anotaron manualmente entidades nombradas y sus relaciones a partir de un análisis sintáctico de dependencias. El corpus cuenta con casi 1.100 oraciones extraídas de resúmenes de artículos científicos.

Por otro lado, **BioText** (Schwartz & Hearst, 2003) es una colección de 1.000 resúmenes de MEDLINE, seleccionados aleatoriamente y que han sido etiquetados con el objetivo de relacionar abreviaturas biomédicas con su forma desarrollada. En consecuencia, las anotaciones no son relevantes para muchas de las tareas de PLN.

6.3.1 El corpus GENIA

Uno de los corpus de domino cuyo uso está más extendido en la comunidad científica es el corpus GENIA. Esto se debe a que posee diversos niveles de anotación referentes tanto a la forma como a la semántica, y a que es un corpus de libre distribución que, además, tiene asociados otros recursos, como la ontología previamente descrita y otras herramientas para la anotación morfosintáctica.

El **corpus GENIA**[27] (Kim et al., 2003, Kim et al., 2006b), desarrollado dentro del *GENIA Project* por la universidad de Tokio, es un corpus formado por 2.000 títulos y resúmenes de artículos extraídos de la base de datos MEDLINE. Se centra en un subdomino muy específico, a saber, las reacciones biológicas en los factores de transcripción en las células de la sangre humana.

Las anotaciones que contiene son tanto sobre el conocimiento biomédico, es decir, anotaciones semánticas y de eventos, como sobre las estructuras lingüísticas en las que se insertan, es decir anotaciones morfosintácticas (Kim et al. 2006b).

[27] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi

Para llevar a cabo las anotaciones, en primer lugar los autores dividieron el corpus en oraciones y procedieron a la tokenización del corpus. Las oraciones y los tokens están divididos entre sí por etiquetas XML.

Como se ha dicho, el corpus está anotado con etiquetas morfológicas y sintácticas. El resultado es un árbol de frases y cláusulas en donde el elemento raíz cubre el total de la oración, los elementos internos se corresponden con la frases y cláusulas, y las hojas del árbol representan las palabras.

Esta estructura de árbol es posible porque, para la etiquetación sintáctica, han utilizado de nuevo etiquetas XML en las que se especifica la categoría sintáctica de cada elemento. Por ejemplo, *Prepositional phrase* (Sintagma preposicional), *Adjective phrase* (Sintagma adjetival), *Particle* (participio), *Reduced Relative clause* (Oración de relativo reducida), etc.

A la información gramatical que no tiene cabida en las etiquetas de los constituyentes sintácticos, del tipo *Dative* (Dativo), *Vocative* (Vocativo), *Locative complement*, (Complemento locativo), etc., se le ha asignado etiquetas que hacen referencia a los roles gramaticales.

Por otro lado, la anotación semántica consiste en la asignación de etiquetas a los términos del dominio, es decir, a aquellas expresiones nominales referentes a entidades biológicas, con un significado específico que es compartido y repetido a lo largo del corpus.

Por ejemplo, en la figura 6.4 se puede ver la anotación de términos en el término *IL-2 gene transcripiton in T cells*.

```
<term><term sem="DNA_domain_or_region">IL-2 gene</term>
transcription</term> in <term sem="Cell_type">T cells</term>
```

Figura 6.4 Anotación de un término en GENIA.

En la figura, se han identificado los términos "IL-2 gene", "IL-2 gene transcription" y "T-cells". Las etiquetas utilizadas para la anotación de los términos se corresponden con conceptos de la ontología.

Finalmente, el corpus contiene anotaciones referentes a eventos. Un evento biológico se define como una ocurrencia temporal que sucede en una o más entidades biológicas (Kim et al., 2006b). Para la anotación de eventos, se creó una ontología específica en la que se definieron eventos que causan algún cambio en genes o productos génicos (como las proteínas).

El corpus GENIA se ha utilizado en la metodología que se describe en esta memoria tanto para instanciar la ontología como para evaluar los resultados obtenidos, como se observa en los siguientes apartados.

6.4 El framework BioOntoVerb

La integración de algunos de los recursos lingüísticos y ontológicos previamente descritos constituye el núcleo de la metodología de la propuesta para la instanciación automática de ontologías mediante roles semánticos.

En anteriores apartados, se ha hecho alusión a varios recursos, tanto léxicos como ontológicos, que fueron desarrollados o están siendo desarrollados, con el objetivo de organizar y sistematizar el conocimiento de manera que pueda ser, por un lado, computable y, por otro, inteligible para el ser humano. Es mediante la combinación e integración de estos recursos de diversa naturaleza, como se ha desarrollado la metodología propuesta aquí, que reutiliza, en la medida de lo posible, el contenido puesto a disposición por y para la comunidad científica.

En la siguiente tabla (Tabla 6.2), se han clasificado los recursos a los que nos referimos, estableciendo una distinción entre recursos léxicos y ontológicos. Así mismo, se han añadido otras herramientas como razonadores y herramientas de PLN.

Tabla 6.2 Recursos ontológicos y lingüísticos.

Recursos Léxicos		Recursos Ontológicos		Corpora	Herramientas	
Generales	**Domino Biomédico**	**Alto Nivel**	**Domino**	**Domino Biomédico**	**Generales**	**Dominio Biomédico**
FrameNet	PasBio	OBO relations ontology	xGENIA	Corpus GENIA	GATE	GENIA tagger e identificador de EN
VerbNet	BioProp	BioTop	GENIA	BioText	Protègè	
WordNet	UMLS			PubMed/ MEDLINE	Hermit Pellet2	

Estos recursos y herramientas poseen un alto grado de estandarización y se utilizan ampliamente dentro de la comunidad científica.

A continuación, se describe cómo estos recursos se han integrado para ser parte de la metodología de instanciación de ontologías de dominio biomédico denominada BiOntoVerb. En líneas generales, relaciones seleccionadas de ontologías biomédicas de alto nivel se han mapeado con los roles semánticos. De este modo, se ha creado un modelo ontológico integrable en ontologías de dominio, y que permite la extracción de instancias a través de las relaciones semánticas textuales.

Los recursos lingüísticos y ontológicos se han integrado en tres capas diferentes, como se muestra en la figura 6.5. A continuación, se introducen brevemente cada una de las capas, que se describirán más en profundidad en los siguientes apartados.

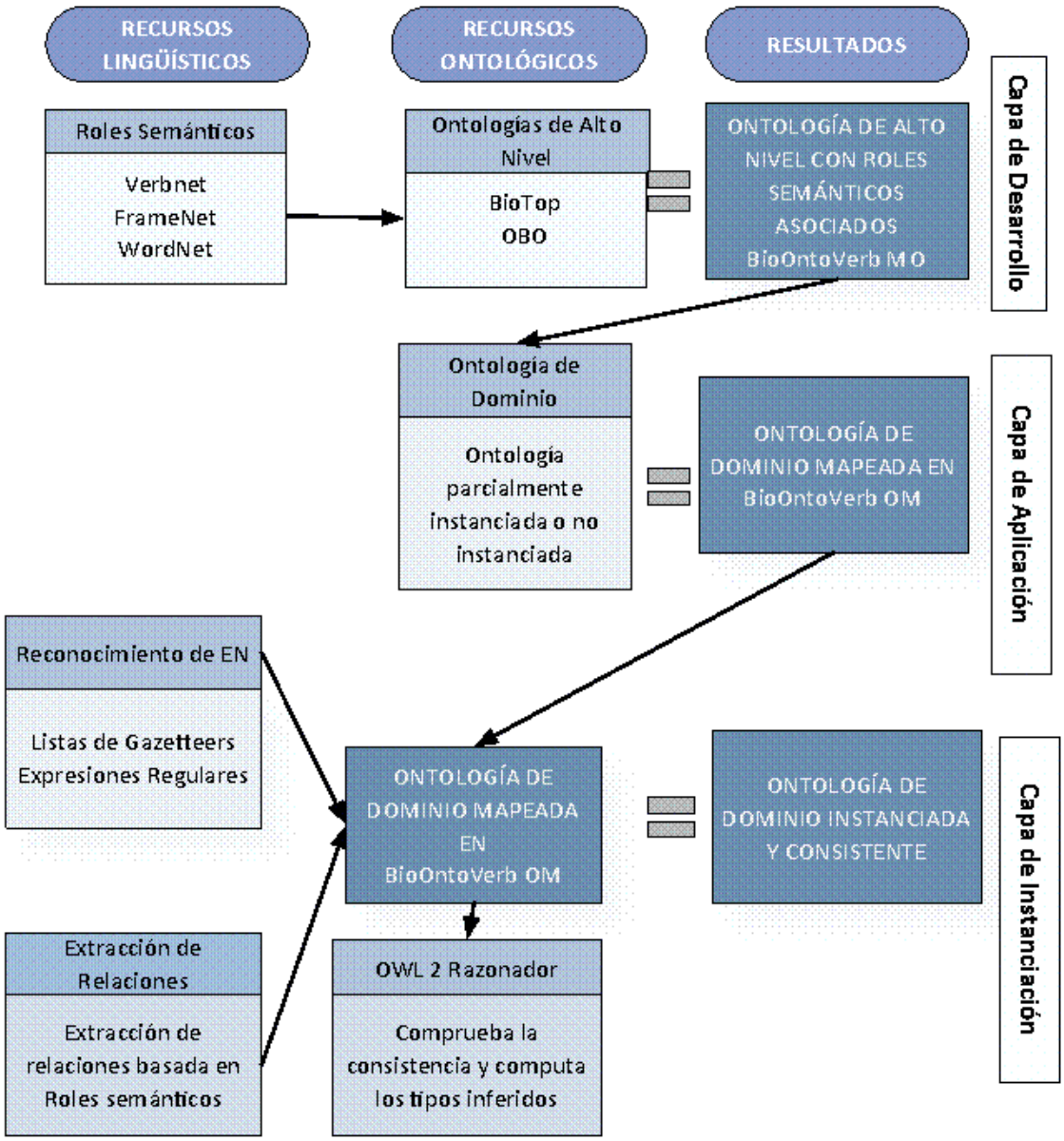


Figura 6.5 Integración de los recursos en BioOntoVerb.

Capa de desarrollo. En esta capa, se ha definido un modelo ontológico basado en ontologías de alto nivel en el que se integran las diferentes relaciones descritas en OBO *relations ontology* y en BioTop. Este modelo ontológico, llamado BioOntoVerb MO, está descrito en OWL 2 y permite definir ontologías de dominio basándose en las relaciones semánticas de alto nivel y que comúnmente se usan en las ontologías de dominio. En el modelo ontológico, las relaciones se

expresan por medio de *Object Properties* , a las que se les han asignado distintos axiomas, entre los que se incluyen la transitividad o la asimetría. Así mismo, a cada una de las relaciones u *Object Properties* que conforman el modelo, se les ha asignado un conjunto de frames semánticos extraídos de FrameNet. Cada frame semántico, está asociado, a su vez, a un conjunto de expresiones verbales u otras unidades léxicas que representan dicha relación a un nivel textual.

Las unidades léxicas incluidas en el modelo han sido extraídas de FrameNet principalmente, pero también de VerbNet y WordNet, y mapeadas con roles existentes en FrameNet. En la tabla 6.3, se puede ver un ejemplo del mapeo de estas relaciones, que más adelante se describe en profundidad.

Por ejemplo, a la relación ontológica *PhisicalContainedIn* de la ontología BioTop que se corresponde con la relación *ContainedIn* de la ontología OBO, le han sido asignados los axiomas Transitivo, Irreflexivo y Asimétrico. Así mismo, se le ha asignado el *frame Contain* cuyas unidades léxicas asociadas son *contain, include, have as a component, comprise, hold inside* e *incorporate*.

Tabla 6.3 Mapeo entre relaciones ontológicas y esquemas semánticos.

Relaciones Ontológicas	**Axiomas de las Propiedades**	**Roles semánticos**	**Unidades Léxicas**
PhysicallyContainedIn (BioTop) Contained_in (OBO)	Transitive Irreflexive Asymmetric	**Containing:** Container holds within its physical boundaries the Contents.	Contain Include Have as a component Comprise Hold inside Incorporate

Capa de Aplicación. En esta capa, se define la ontología que va a ser instanciada, es decir, la ontología de dominio. Para ello se utiliza el modelo ontológico previamente descrito. Las relaciones de la ontología de domino se

mapean con las relaciones definidas en la ontología de alto nivel, es decir, en el modelo ontológico BioOntoVerb OM. Al mapear las relaciones de la ontología de alto nivel con la ontología de dominio, todas la propiedades definidas, así como los roles asociados, pasan a formar parte de la ontología de dominio.

Capa de instanciación de la ontología. En esta capa se lleva a cabo el proceso de instanciación de la ontología dividido en dos fases en las que se han integrado los recursos pertinentes. No se describe aquí el proceso de instanciación en sí mismo, sino los recursos que forman parte de la metodología.

En primer lugar, los candidatos de entidades nombradas se identifican utilizando el framework de GATE. Para ello, se ha utilizado una combinación de reglas JAPE y listas de Gazetteers junto con el identificador de entidades nombradas de GENIA. Las listas que se han utilizado contienen términos biológicos extraídos de UMLS.

Una vez que se han extraído las entidades nombradas, se identifican en el texto los roles semánticos con el objetivo de extraer las posibles relaciones entre las entidades reconocidas.

Finalmente, un razonador, como por ejemplo Pellet2, se ejecuta con el objetivo de comprobar, por un lado, la consistencia de la ontología, y por otro, de añadir los tipos inferidos.

Si la ontología es inconsistente, la última relación incluida en la ontología se elimina. En el caso de que la ontología sea consistente, y el razonador haya inferido que uno de los individuos que pertenece a la relación se puede clasificar en una nueva clase, dicha clasificación se lleva a cabo.

A continuación, se explica en detalle cómo se han desarrollado los recursos en cada una de las capas de las que consta el sistema, a saber, el modelo ontológico, la asignación de roles semánticos a las relaciones de la ontología y el desarrollo de los recursos para el reconocimiento de entidades.

6.4.1 Modelo ontológico BioOntoVerb

Como se ha comentado anteriormente, en este trabajo se propone un modelo ontológico basado en distintos tipos de relaciones definidas en las ontologías OBO *relation ontology* y BioTop. Este modelo ontológico se ha implementado utilizando la nueva versión del OWL, es decir, OWL 2, que como se ha visto previamente permite una mayor expresividad en las propiedades y un soporte extendido para los atributos. En la tabla 6.4, aparecen las relaciones mapeadas de OBO y BioTop con sus respectivos axiomas.

Tabla 6.4 OWL 2 Axiomas de las propiedades de la ontología de relaciones de OBO y BioTop.

Relación	T	S	R	I	A	F	IF	Ontología
part_of	X		X		X			OBO
located_in	X		X					OBO
contained_in				X				OBO
adjacent_to	X	X						OBO
transformation	X							OBO
derives_from	X				X			OBO
preceded_by	X				X			OBO
has_participan	X							OBO
has_agent	X							OBO
caused_by				X	X			BioTop

Estas relaciones son binarias y se han implementado utilizando los axiomas de las propiedades que pueden definirse en OWL 2. A continuación se indican los principales axiomas de las propiedades:

Reflexiva (R). (X Relación X)

Irreflexiva (I). not(X Relación X).

Simétrica (S). (X Relación Y) ↔ (Y Relación X)

Asimétrica (A). (X Relación Y) →not(Y Relación X).

Transitiva (T). (X Relación Y) and (Y Relación Z) → (X Relación Z).

Funcional (F). (X Relación Y) and (X Relación Z) → (Y = Z)

Funcional Inversa (IF). (X Relación Y) and (Z Relación Y) → (X = Z)

El modelo formal de OWL soporta un conjunto de descripciones lógicas mediante el que se pueden realizar inferencias. Este servicio es soportado por razonadores DL (por ejemplo HermiT, Pellet2, Fact++, Racer) (Sirin et al., 2007), donde el razonador puede realizar un conjunto de tareas que pueden resumirse en las siguientes:

- Comprobar la consistencia, que asegura que la ontología no contiene ningún hecho contradictorio.
- Comprobar si una instancia satisface los requisitos necesarios para pertenecer a una clase, es decir, comprobar si es posible para una clase ser instanciada, y en caso de ser así, qué tipo de instancias son las más adecuadas. Si se inserta una instancia que no satisface los requisitos de una clase, puede causar que toda la ontología sea inconsistente.
- Clasificar, que consiste en calcular las relaciones de las subclases entre cada una de las clases nombradas para crear una jerarquía completa de clases. La jerarquía de clases puede ser utilizada para responder consultas, tales como, obtener todas las subclases directas de una clase.

- **Realización, que consiste en localizar las clases más específicas a las que pertenece un individuo, o en otras palabras, computar los tipos indirectos para cada uno de los individuos.**

Una ontología OWL se puede considerar, desde un punto de vista lógico, como una colección de axiomas que deben ser satisfechos. Esto no sólo incluye clases y propiedades, sino también restricciones como las clases disjuntas. Es por esto, por lo que la consistencia es fundamental en la ingeniería ontológica.

Se dice que una ontología es internamente inconsistente cuando alguna de las partes de la ontología son inconsistentes con otras partes de la misma. Por ejemplo, una ontología es internamente inconsistente si una de las propiedades relativa a las relaciones entre los conceptos no se satisface. Los axiomas de las propiedades definidos en la tabla 6.4 ayudan a detectar cualquier inconsistencia en la ontología poblada. Por ejemplo, la relación *part_of* posee tanto las propiedades transitiva como asimétrica, de manera que no es posible que haya un ciclo dentro de una partonomía conceptual. Eso asegura que se pueda inferir de la ontología el resultado correcto del conocimiento mediante la aplicación de los axiomas correspondientes.

Además, la existencia de tales restricciones es útil para garantizar la consistencia de la construcción de los individuos, que deben satisfacer las restricciones definidas para la clase correspondiente. Por otro lado, el razonador se puede utilizar para la clasificación automática de los individuos en la colección de condiciones definida.

6.4.2 Asignación de roles semánticos a las relaciones de la ontología.

Como apunta Friedman (2002), la interacción entre entidades en un texto se expresa fundamentalmente por medio de relaciones verbales y sus correspondientes formas nominales.

Las relaciones de un dominio se pueden formalizar en una ontología como *Object Properties,* como se ha visto anteriormente. Dichas relaciones pueden ser creadas por un experto o, en algunos casos, inferirse por el sistema a partir de un conjunto de textos. La metodología propuesta en esta tesis prevé la identificación de relaciones semánticas entre posibles instancias de forma automática. Para ello, se ha realizado un mapeo entre las relaciones ontológicas y los *frames* o marcos semánticos que pueden representar dichas relaciones a nivel textual.

La ventaja de asignar *frames* a las relaciones de la ontología, es que los *frames* semánticos ya están asociados con un conjunto de expresiones lingüísticas que los representan a nivel textual, y por otro lado, las entidades implicadas como roles de un *frame* pueden equivaler, en muchos casos, a entidades relevantes de domino y éstas, a su vez, pueden extrapolarse a instancias de la ontología.

En el caso de las ontologías de alto nivel, las relaciones poseen un alto grado de abstracción, y no siempre pueden ser extraídas directamente de un texto. Por esta razón, el mapeo entre relaciones y expresiones lingüísticas concretas presenta algunas dificultades que se describen a continuación.

Por otro lado, existen relaciones en el texto que no se pueden sistematizar en la ontología, es decir, no toda la información textual puede ser modelada en una ontología, ni todo el conocimiento modelado en una ontología está explícitamente expresado en una colección de textos. En consecuencia, se pueden distinguir tres tipos diferentes de relaciones en función de la correspondencia que existe entre textos y ontologías:

- Relaciones explícitas en el texto cuya expresión lingüística o conjunto de expresiones lingüísticas pueden ser mapeadas con una o más

relaciones modeladas en la ontología. Dos ejemplo de este tipo de relaciones son "derives from" y "caused by". Se trata de relaciones que en la ontología misma están expresadas con un verbo cuyo significado no presenta un alto grado de ambigüedad.

- Relaciones explícitas en el texto cuya expresión lingüística no se puede mapear a una relación modelada en la ontología. Normalmente, se trata de relaciones expresadas mediante verbos con un significado muy general. Por ejemplo, "do" y "have" son relaciones en las que la precisión que se obtendría con el mapeo de dichos verbos a relaciones concretas sería muy baja.
- Relaciones modeladas en la ontología que no se corresponden con una expresión o conjunto de expresiones lingüísticas particulares, al menos no de un modo que se pueda llevar a cabo una generalización. Por ejemplo la relación "hasParticipant" cumplie esta característica. A menudo, este tipo de relaciones no se expresan mediante un verbo o determinadas formas léxicas y pueden estar implícitas en el texto.

En una ontología, un concepto se refiere de manera unívoca a un modo de significar que puede estar expresado en la lengua por uno o más términos. No obstante, esta univocidad no deja de ser una convención adoptada por el ontólogo. La elección de un término y no otro para representar un concepto en la ontología supone una agrupación previa de todos los términos que pueden representar un concepto y la elección de uno de ellos como la realización lingüística de dicho concepto.

El caso de las relaciones es similar, en el sentido de que una relación puede expresarse por medio de diferentes términos o combinaciones de términos. El problema es, como se ha mencionado, que las relaciones en muchos casos no están explícitas en el texto y se deben inferir del contenido global del documento.

No obstante, existen algunos géneros discursivos que facilitan la extracción de relaciones. Por ejemplo, las definiciones de glosario son un ejemplo de estructura lingüística a partir de la que se pueden extraer relaciones, fundamentalmente de hiperonimia-hiponimia.

Por ejemplo:

***[NF kappa B**$_{DEFINENDUM}$**]** **is a** [heterodimer $_{\text{DEFINICIÓN}}$] consisting of a 50kDa DNA.

Este tipo de relaciones taxonómicas, a pesar de las dificultades (Navigli et al., 2010) son las más propicias a la sistematización, ya que, cuando se trata de una definición, el verbo que indica la relación suele aparecer explícito.

En el lado opuesto, tenemos las relaciones causales que, en muchas ocasiones, requieren algún tipo de inferencia por parte del lector. Por ejemplo en la siguiente frase:

[$_{\text{CAUSA}}$The storm] left [$_{\text{CONSECUENCIA}}$ more than a million people in three states without power and submerged highways even hundreds of miles from its center]

El verbo "leave" en este caso indica una relación causal del tipo *A causa B*, pero no se puede generalizar asociando la expresión lingüística *leave* a cualquier relación causal.

En este trabajo, nos centramos en obtener relaciones explícitas en el texto expresadas mediante un verbo (conjugado o no) o mediante una forma nominalizada o lexicalizada del mismo que se puedan representar de manera ontológica. Ya que, como se ha indicado, las interacciones entre instancias se expresan fundamentalmente mediante verbos o sus formas nominales correspondientes.

En dominios especializados, el grado de ambigüedad es menor que en dominios de carácter general. Por este motivo, la sistematización de las expresiones lingüísticas que expresan algunas relaciones se ha podido llevar a cabo.

Una vez desarrollado el modelo ontológico que se ha descrito en el apartado 6.4, se procede a la asignación de roles semánticos susceptibles de representar a cada una de las relaciones. Dado que un marco semántico tiene asociadas un conjunto de unidades léxicas, dichas unidades serán las formas lingüísticas que la relación pueda adoptar en el texto. No obstante, no todas las relaciones se pueden asociar a un *frame* semántico.

En el apartado 3.2.4, se han descrito algunos proyectos que han desarrollado sistemas de etiquetado de roles semánticos para el dominio de la bioinformática. Sin embargo, en dichas aproximaciones, el conjunto de verbos que se ha tenido en cuenta está restringido a ciertos subdominios dentro de la biomedicina. En concreto, la mayoría de los sistemas se refieren únicamente al dominio de la biología molecular. Consecuentemente, no dan cobertura a las relaciones expresadas en un rango más amplio de textos biomédicos. Por este motivo, para la metodología descrita en este trabajo se han utilizado los *frames* generales que proporciona FrameNet (Baker et al., 1998), que, en los casos que se ha creído conveniente, se han mapeado con los de VerbNet (Kipper-Schuler, 2005), y WordNet (Miller, 1995, Fellbaum, 1998), del que es posible obtener nuevas unidades léxicas a partir del conjunto de *synsets*.

Para la selección de los marcos semánticos, en primer lugar se ha realizado un análisis estadístico de una muestra significativa del corpus GENIA y del corpus Biotext (Schwartz & Hearst, 2003). En total, se han considerado unas 10.000 palabras, de las que se han extraído los verbos más frecuentes y se ha estudiado si dichos verbos se podrían ajustar de modo general a las relaciones expresadas en la ontología. En la siguiente tabla (tabla 6.5), se muestran las frecuencias de algunos de los verbos que podrían representar relaciones ontológicas.

Tabla 6.5 Frecuencias de los verbos en los corpora biomédicos analizados.

Verbo	Frecuencia
Induce	309

Contain	140
Cause	95
Change	80
Include	77
Locate	40
Localize	11

De entre los verbos más frecuentes que expresan una relación, se ha seleccionado el *frame* al que pertenecen y a su vez, partiendo del modelo ontológico, se han seleccionado aquellas relaciones que pueden ser expresadas mediante los *roles* seleccionados.

La asociación entre marcos semánticos y relaciones de la ontología se lleva a cabo manualmente. Para cada *frame* asociado a una relación en la ontología, se buscan en el corpus ejemplos que sustenten dicha relación. Es un proceso de retroalimentación, como se puede ver en la figura 6.6.

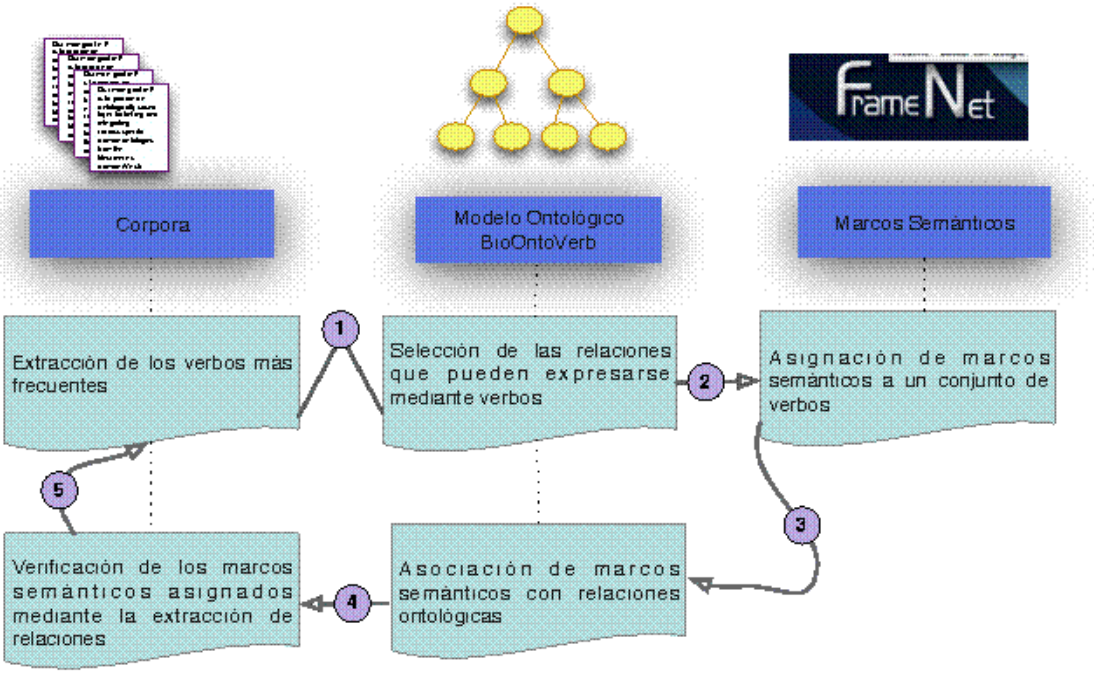


Figura 6.6 Proceso de asociación de Verbos-Marcos semánticos- Relaciones ontológicas.

Los principales problemas que existen para llevar a cabo esta tarea son los siguientes:

1. No existe una correspondencia clara y unívoca entre las unidades lingüísticas de los marcos semánticos y los verbos que indican las relaciones en la ontología de alto nivel.
2. El mismo verbo no siempre indica el mismo tipo de relación.
3. Las relaciones en la ontología de alto nivel pueden estar expresadas mediante:

   a) un participio verbal con una preposición, como por ejemplo *located_in*, *contained_in*, de modo que en el texto suelen aparecer en forma pasiva.

   b) están nominalizadas, un fenómeno frecuente en textos biomédicos. Por ejemplo, *transformation_of*, *instance_of*.

   c) mediante un verbo con una preposición *derives_from*.
4. El número de verbos de los proyectos que describen exclusivamente roles del dominio biomédico es bastante reducido, por lo que hay que recurrir a proyectos de anotación de roles de carácter general, como FrameNet, lo que conlleva problemas de ambigüedad, dado que cada verbo tiene numerosos sentidos, de los cuales, al menos uno es aplicable al dominio biomédico.
5. Los roles semánticos relacionan dos o más entidades, de modo que la correcta identificación de las entidades resulta fundamental para poder establecer una relación entre ellas.
6. Las relaciones que se pueden representar en la ontología mediante *frames* o marcos semánticos son de carácter binario, es decir, cada relación se establece entre dos roles que a su vez se corresponden con dos entidades. Si en un *frame* están implicados más de dos roles, es sólo el núcleo del *frame* el que se tiene en cuenta. Una excepción es el *fram Transformation*

que, como se explica más adelante se ha dividido en dos frames diferentes, cada uno de los cuales relaciona dos roles.

En la figura 6.7 se muestra un ejemplo del *frame Locating*, en el que se relacionan dos entidades. Por un lado, la entidad para la que se indica la localización, es decir *Sought entity* y, por otro, la entidad que indica la localización, etiquetada como *Location*.

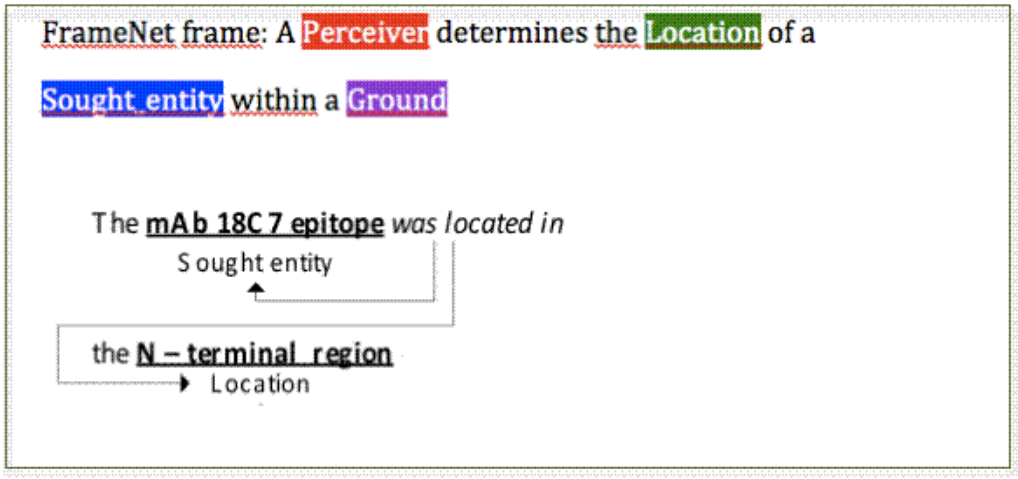


Figura 6.7 Ejemplo de un *frame* aplicado a una relación en el corpus.

El núcleo del *frame* consta de cuatro roles, pero para el dominio biomédico sólo se han considerado dos, a saber, *Location* y *Sought entity*, ya que *Perceiver* o el Perceptor no es relevante (generalmente, se sobreentiende que es un científico) y no aparece reflejado en los textos ni en la ontología. Por otro lado, los roles *Ground* o Terreno tampoco tienen cabida en este contexto.

Este rol semántico relaciona dos entidades. Por un lado, la entidad [SOUGHT ENTITY mAb 18C7 epitope] está relacionada mediante la unidad léxica *was located in* con la entidad [LOCATION N-terminal región], de manera que, en este *frame* existe una relación ontológica del tipo *Located_in* entre la Entidad$_1$ y la Entidad$_2$, cuyos roles son el *Sought Entity* y *Location*, respectivamente. La ventaja que presenta el dominio de la biomedicina con respecto a otros dominios es la existencia de ontologías de alto nivel o *top Level ontologies* que proporcionan unas relaciones de carácter general ya definidas, y presentes en las ontologías de

dominio biomédico. Definir los roles para cada una de las ontologías de dominio no es factible por el tiempo y esfuerzo necesarios.

En la siguiente tabla (tabla 6.6), se muestra el mapeo que se ha llevado a cabo entre las relaciones ontológicas, los *frames*, las unidades léxicas asociadas y algunos ejemplos extraídos del corpus GENIA y Biotext.

Tabla 6.6 Asociación de las relaciones de BioOntoVerb OM con los *frame* de FrameNet

Relaciones Ontológicas de BioOntoVerb	**Frame de FrameNet**	**Unidades Léxicas**	**Ejemplos de los corpora**
CausedBy	Cause: A Cause causes an Effect	Cause, Induce, Make, Create, Incite, Provoke, Engender, Instigate, Generate, Produce, Result, Develop	nur77 *may cause* apoptosis by monomer oncoproteins *might cause* AML TG and PMA *induce* IL - 2R alpha The use of 9 - mer peptides (BCR1 / 9) *would generate* a CD8 / HLA class I - restricted response Anti - CD28 - stimulated T cells *produced* significant amounts of IL – 8 K30p (p13K30) and K34p (p13K34) *result in* a Trp - Arg substitution patients with acquired immunodeficiency syndrome (AIDS) *develop* glucocorticoid resistance
Located_in (OBO)	Locating: A Perceiver determines the Location of a Sought_entity within a Ground	Locate, Situate, Place, Localize, Localized in, Be present In, Localized to, Find	One type , EA / S , *is located in* the upstream promoter region the mAb 18C7 epitope *was located in* the N - terminal region I kappa B gamma - 1 *is found in* both the cytoplasm and nucleus kappaB regulatory elements *located in* the U3 region oxidase component cytochrome b558 *is localized in* gelatinase granules
Contained_in	Contain: Container holds within its physical boundaries the Contents.	Contain, Include, Have as a component, Incorporate, Hold inside/ in.	Murine B lymphocytes, adipocytes, and olfactory neurons *contain* a DNA-binding protein IL - 1 beta and IL - 6 genes *contain* AP - 1 NF - kappa B *is held in* the cytoplasm

Derives_from	Create: An (Agent) or Cause leads to the formation of a Created entity.	Create, Originate from, Derive from, Generate (inversa), Come from, Develop from, Alter, Transcribe from, Stems from	HL60 and normal cells *are able to generate* estrone (E1) cytotoxicity of indolocarbazoles *derives from* Top1p The SM protein *derived from* the spliced RNA B cell mutant *derived from* the Burkitt lymphoma Raji cell B cell hybrids carrying an AIR - 1 locus *derived from* CIITA TCR genes *developed from* Id3 EBNA - 1 gene in infected thymocytes *was transcribed from* the Fp promoter p45 mRNA *is transcribed only from* aNF - E2 promoter
Transformation_of	Transform: An Agent or Cause causes an Entity to change,	Transform, Modify, Change	the ability of Bcr - Ab1 to *transform* fibroblasts Human T - cell leukemia virus type 1 (HTLV - 1) Tax *transforms* normal T – cells I kappa B alpha *is* selectively *modified in* LMP - 1 - expressing B cells GATA sites *have been mutated into* human T lymphocytes the putative transmembrane domain of Ost4p *were changed into* an ionizable amino acid
	either in its category membership or in terms of the value of an Attribute. In the former case, an Initial_category and a Final category may be expressed (...).	Transform in/into, Change in/into Mutate in/into, Change into, Convert into, Turn into.	

En la tabla anterior (tabla 6.6), aparecen, por un lado, las relaciones ontológicas que se han podido asociar con roles semánticos, y, por otro, el marco semántico de FrameNet, adecuado a esa relación ontológica, en el que sólo se representan los roles que constituyen el núcleo del *frame* .

En cuanto a las expresiones lingüísticas, se han incluido los verbos que se han considerado pertinentes, es decir, no sólo se han incluido aquellos asociados al *frame* de FrameNet. Por otra parte, si un verbo asociado a un *frame* no se ha considerado relevante, se ha eliminado.

Para enriquecer el número de expresiones lingüísticas asociadas a un *frame,* se ha utilizado WordNet, en el que se han buscado los sinónimos, así como VerbNet, que agrupa los verbos en función de su significado. El único requisito que se ha seguido en el modelo propuesto para la agrupación de expresiones lingüísticas es que se ajusten al esquema del *frame* de FrameNet en el contexto de la biomedicina.

Finalmente, en la última columna de la tabla se muestran algunos ejemplos extraídos del corpus en donde aparecen subrayadas las entidades relacionadas y, en cursiva, la expresión lingüística que las relaciona.

Por ejemplo, la relación *Contained_in* está asociada con el marco semántico *Contain*, y los roles semánticos que constituyen el núcleo de la misma son *Container* y *Contents*. A su vez, las unidades léxicas asociadas al *frame* son *Contain*, *Include*, *Have as a component*, *Incorporate, Hold inside* y *Hold in*.

Además, hay que tener en cuenta que si se localiza una relación entre una entidad y ésta a su vez está conectada a una o más entidades mediante la conjunción copulativa *and* o disyuntiva *or*, todas las entidades conectadas son incluidas en la ontología como candidatos a relaciones, ya que el rol semántico que desempeñan en el texto es el mismo.

Las relaciones que se localizan en el texto son de carácter binario, es decir, entre dos entidades que desempeñan roles semánticos diferentes. Como se ha explicado, la mayoría de los marcos semánticos se adaptan a esta relación de tipo binario, manteniendo los dos roles principales del marco. Sin embargo, en el caso del *frame Transform* existen cuatro roles básicos: (1) *Agent or Cause*, (2) *Entity*,

(3) *Initial category* y (4) *Final category*. En este caso, el *frame* se ha dividido en dos partes. La primera incluye los roles *Agent or Cause* y *Entity*, y los verbos *Transform*, *Modify* y *Change*. La segunda parte incluye los roles *Intial Category* y *Final category,* e incluye *Transform in/into, Change in/into Mutate in/into, Change into, Convert into, Turn into*, es decir, aquellas unidades léxicas que indican transformación y que van seguidas por las preposiciones *in* o *into*.

6.4.3 Recursos para el reconocimiento de Entidades Nombradas

La extracción de entidades nombradas es un prerrequisito para muchos aspectos de aprendizaje automático de ontologías a partir de texto, como se indica en (Buitelaar et al., 2005). En biomedicina, las entidades nombradas pueden coincidir con los términos del dominio, entendiendo término como la realización lingüística de un concepto específico del domino.

El desarrollo de un sistema de reconocimiento y extracción de entidades nombradas está fuera de los límites del trabajo propuesto aquí. En la literatura, existen varios sistemas de extracción de entidades adaptados al domino biomédico, a algunos de los cuales se ha hecho mención en el apartado 4.4.

Dado que existen estas alternativas para la identificación de entidades nombradas en biomedicina, no se ha desarrollado ningún recurso específico, siendo el identificador de entidades de GENIA el que se ha utilizado, incorporándolo para ello en el framework GATE.

El identificador de entidades nombradas de GENIA incluido en GENIA *tagger* (Ver apartado 6.5.1) es de libre distribución y está especialmente adaptado al dominio de la biología molecular. En particular, las entidades que reconoce son:

- Protein
- DNA
- RNA

- Cell Line
- Cell Type

Por otra parte, la elaboración de terminologías biomédicas normalizadas es un campo ampliamente desarrollado, existiendo recursos terminológicos de gran cobertura en los que se recoge la terminología del dominio desde diversas perspectivas, incluyendo las prácticas profesionales.

Uno de los principales recursos terminológicos es UMLS (NLM, 2008), que combina la información de más de 100 vocabularios del dominio biomédico. Por este motivo, se han incluido algunas de las listas terminológicas de UMLS en GATE. A pesar de las limitaciones que implica el reconocimiento de entidades mediante listas, como ya se ha señalado, se ha creído conveniente la inclusión de listas extraídas de UMLS para ampliar la cobertura del reconocedor de entidad nombrada de GENIA. Por último, se han mapeado las entidades etiquetadas en el corpus GENIA con las listas de UMLS pertinentes. Las listas se han añadido a GATE en forma de Gazetteers.

En la siguiente tabla (tabla 6.7), se puede ver el mapeo que se ha realizado entre las entidades anotadas en GENIA corpus y las listas de UMLS. Como se puede observar, existe una correspondencia directa entre varias de las clases y las listas de UMLS. Por ejemplo, la lista *lipid* se corresponde directamente con la clase *lipid* de la taxonomía GENIA. Otras listas, como *ingorganic_compound* o *unicellular organism* en GENIA, se denominan *inorganic* y *mono_cell,* respectivamente, pero hacen referencia al mismo tipo de entidades.

Tabla 6.7 Mapeo entre listas de UMLS y GENIA tagger.

Lista de UMLS	Genia Corpus
amino_acid_monomer	amino_acid_monomer
carbohydrate	carbohydrate
cell_component	cell_component
cell_cultured	cell_line
DNA	DNA_N/A

inorganic_compound	Inorganic
lipid	lipid
multicellular_organism	multi_cell
organic_compound_other	other_organic_compound
peptide	peptide
polynucleotide	polynucleotide
RNA	RNA_N/A
tissue	tissue
unicellular_organism	mono_cell
virus	virus

Además, se han añadido algunas listas de carácter general de UMLS, como *Syndrome_or_Disease,* con las que se han obtenido buenos resultados en la extracción de entidades menos específicas. La incorporación de nuevas listas al sistema de reconocimiento de entidades nombradas se puede realizar de manera sencilla y se adapta a las necesidades del corpus y de la ontología que se pretenda instanciar.

6.5 Validación de la metodología en el domino de la biomedicina

Una vez que se han descrito los recursos creados y el marco de trabajo en el que se integran, se procede en este apartado a la descripción de la metodología para la instanciación automática de ontologías basado en el modelo ontológico definido.

La metodología que aquí se propone tiene como objetivo poblar una ontología de dominio biomédico a partir de texto en lenguaje natural y haciendo uso de las herramientas y recursos de PLN descritos anteriormente.

La arquitectura del sistema, que se muestra en la figura 6.8, está compuesta por tres fases secuenciales denominadas (1) Fase de PLN, (2) Fase de Reconocimiento

y Extracción de entidades nombradas, extracción de relaciones, y finalmente (3) Fase de Instanciación de la Ontología.

Brevemente descrito, el procedimiento que se lleva a cabo es el siguiente, durante la fase de NLP se analiza morfo-sintácticamente del corpus. Es decir, se obtienen las categorías gramaticales de las palabra y de los constituyentes de cada oración.

En una segunda fase, se anotan las posibles entidades y las relaciones semánticas que existen entre ellas mediante GENIA tagger que identifica tanto los verbos como las entidades.

En la fase final, se obtienen las instancias de la ontología de dominio a partir de las entidades nombradas previamente anotadas y de las relaciones que existen entre ellas, procediendo a su clasificación en la ontología. Además, durante esta fase se comprueba la consistencia de las instancias obtenidas. Al final de esta etapa, el resultado es una ontología de dominio que ha sido enriquecida con nuevas instancias.

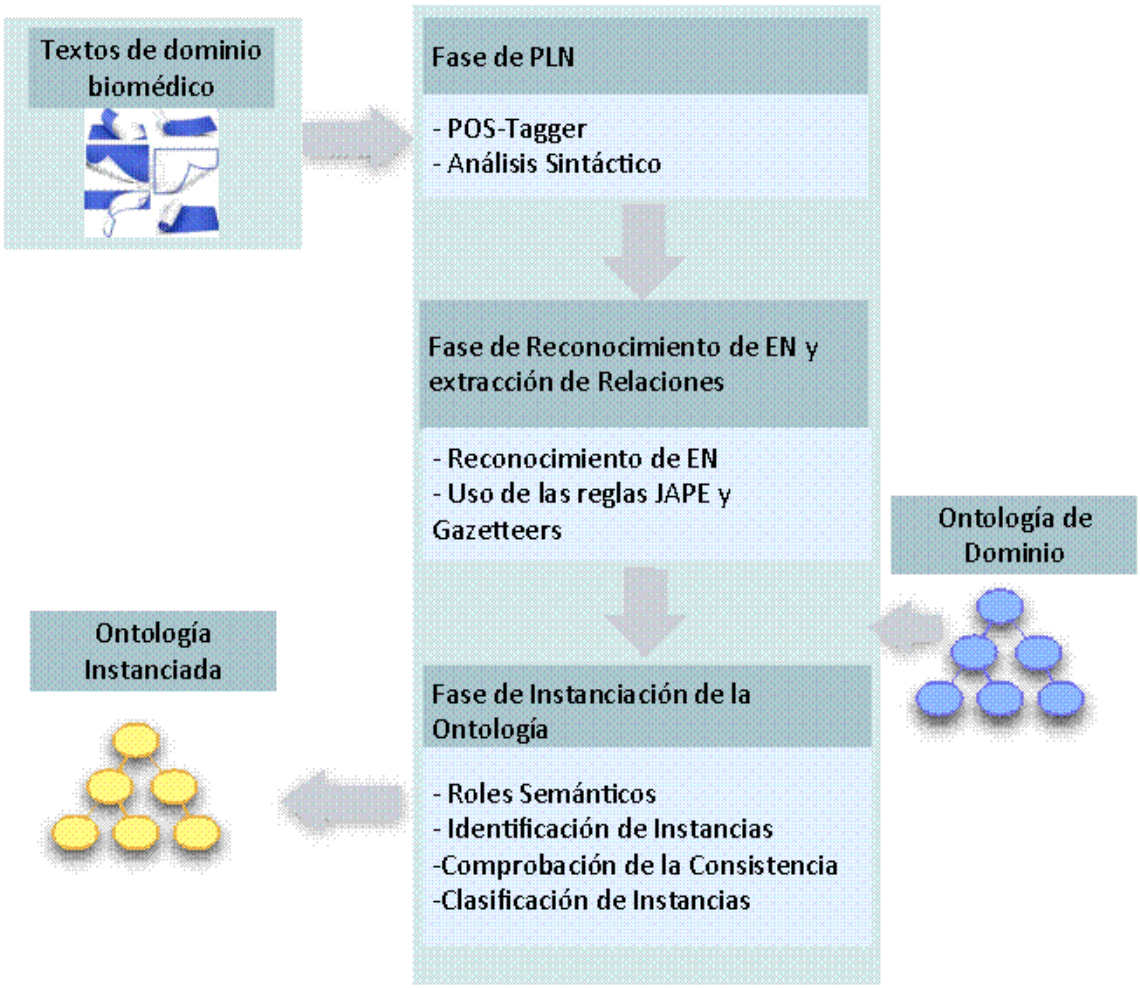


Figura 6.8 Arquitectura del sistema.

6.5.1 Fase de PLN

El principal objetivo de esta fase es obtener la estructura morfosintáctica de cada oración, para lo que se ha hecho uso de GENIA tagger (Tsuruoka et al., 2005) a través de GATE.

Como se ha visto previamente, el lenguaje biomédico cuenta con algunas particularidades que pueden afectar al PLN, como por ejemplo las ambigüedades

causadas por nombres y abreviaturas que comienzan por letras mayúsculas, la inclusión de caracteres no-alfanuméricos en expresiones químicas y numéricas tales como comas, paréntesis y guiones, los participios de verbos poco frecuentes que describen eventos específicos del domino, etc.

El etiquetador de GENIA es capaz de gestionar estos problemas de forma más eficiente que los analizadores morfológicos de dominio general. No obstante, no todas las ambigüedades se resuelven y, en algunos casos, aquellas ambigüedades no resueltas pueden afectar a etapas posteriores del proceso.

A continuación, en la tabla 6.8 y en la tabla 6.9 se muestra un ejemplo de la siguiente oración analizada mediante GENIA tagger y mediante Freeling (ver apartado 3.3), un etiquetador morfo-sintáctico de carácter general:

After treatment with 5-azacytidine, the adult mensenchymal stem cells were transformed into cardiomyocytes.

Tabla 6.8 Etiquetación morfológica con GENIA tagger.

Palabra	**Lema**	**Etiqueta Morf. (POS-Tag.)**
After	After	IN
treatment	treatment	NN
with	with	IN
5-azacytidine	*5-azacytidine*	NN
,	,	,
the	the	DT
adult	adult	JJ
mensenchymal	*mensenchymal*	JJ
stem	stem	NN
cells	cell	NNS
were	be	VBP
transformed	transform	VBN
into	into	IN
cardiomyocytes	*cardiomyocyte*	**NNS**
.	.	.

Tabla 6.9 Etiquetación morfológica con FreeLing.

Palabra	Lema	Etiqueta Morf. (POS-Tag.)
After	After	IN
treatment	treatment	NN
with	with	IN
5-azacytidine	*5-azacytidine*	Z
,	,	Fc
the	the	DT
adult	adult	NN
mensenchymal	*mensenchymal*	JJ
stem	stem	NN
cells	cell	NNS
were	be	VBD
transformed	transform	VBN
into	into	IN
cardiomyocytes	*cardiomyocytes*	NNS
.	.	Fp

Como se puede observar, los resultados obtenidos presentan algunas divergencias. La diferencia más significativa es la etiquetación de *5-azacytine* como *Z*, es decir, como una cifra, mientras que con GENIA tagger *5-azacytine* ha sido etiquetada como NN, es decir, como Sustantivo. Como se ha visto, una de las particularidades del lenguaje biomédico es la combinación de caracteres alfanuméricos en un mismo término. Por este motivo, la incorrecta identificación de dichos términos puede suponer una pérdida significativa de información, en este caso la pérdida de una entidad nombrada.

Por otro lado, *adult* en Freeling ha sido etiquetado como NN, es decir, como sustantivo. Sin embargo, en este caso la correcta etiquetación sería JJ (Adjetivo), que es la que le asigna GENIA. En biomedicina es frecuente encontrar un sintagma nominal formado por un sustantivo y dos o más adjetivos, tratándose en ocasiones de entidades nombradas.

En cuanto al análisis sintáctico, es necesario para la correcta identificación de los roles. Durante el análisis morfológico, se determina si la diátesis de un verbo es pasiva o activa, y durante el análisis sintáctico se identifican los roles o función sintáctica de los constituyentes. Esto es importante porque una entidad que en una oración activa es el sujeto, en una oración pasiva será el agente, sin que cambie el significado del evento. Un ejemplo simple es la oración *Juan regó las plantas* en donde *Juan* es el Agente y *las plantas* el Paciente. Un cambio en la diátesis no afecta a estos roles, es decir en *Las plantas fueron regadas por Juan*, Agente y Paciente siguen siendo los mismos.

En la siguiente figura (figura 6.9), se muestra el análisis morfológico y el análisis sintáctico superficial (división en *chunks*) realizado con GENIA tagger. En la oración (A) *I kappa B alpha is selectively modified by LMP-1 expressing B cells,* los sintagmas nominales etiquetados como NP (Noun Phrase) se clasifican en fases posteriores como entidades. Si se analiza la misma oración en ditátesis pasiva (B), el resultado del análisis es el mismo. Dado que el POS tagger de GENIA incorpora el reconocedor de entidades nombradas, es suficiente con realizar un análisis sintáctico superficial de la oración que permita la identificación de dichas entidades, puesto que la metodología propuesta prevé el uso de verbos en diátesis activa o pasiva y cuenta con el uso del razonador ontológico para determinar en qué sentido se produce una relación.

(A) [**NP** *I kappa B alpha*] [**VP** *is selectively modified*] [**PP** *by*] [**NP** *LMP - 1*]
(B) [**NP** *LMP - 1*] [**ADVP** *selectively*] [**VP** *modifies*] [**NP** *I kappa B alpha*][11]

Figura 6.9 Ejemplo de análisis sintáctico ligero con GENIA tagger.

[11] GENIA tagger utiliza una notación estándar en donde, NP (Noun Phrase), VP (Verbal Phrase), PP (Prepositional Phrase) y ADVP (Adverbial Phrase)

6.5.2 Fase de reconocimiento y extracción de entidades nombradas

Durante esta fase, en primer lugar se identifican las menciones de entidades mediante GENIA *tagger* integrado en GATE. La salida producida por cada componente de GATE es un conjunto de anotaciones, es decir, metadatos asociados a una sección específica del contenido del documento. Cada anotación identificada en el texto va unida al correspondiente tipo de entidades nombradas (ver figura 6.10).

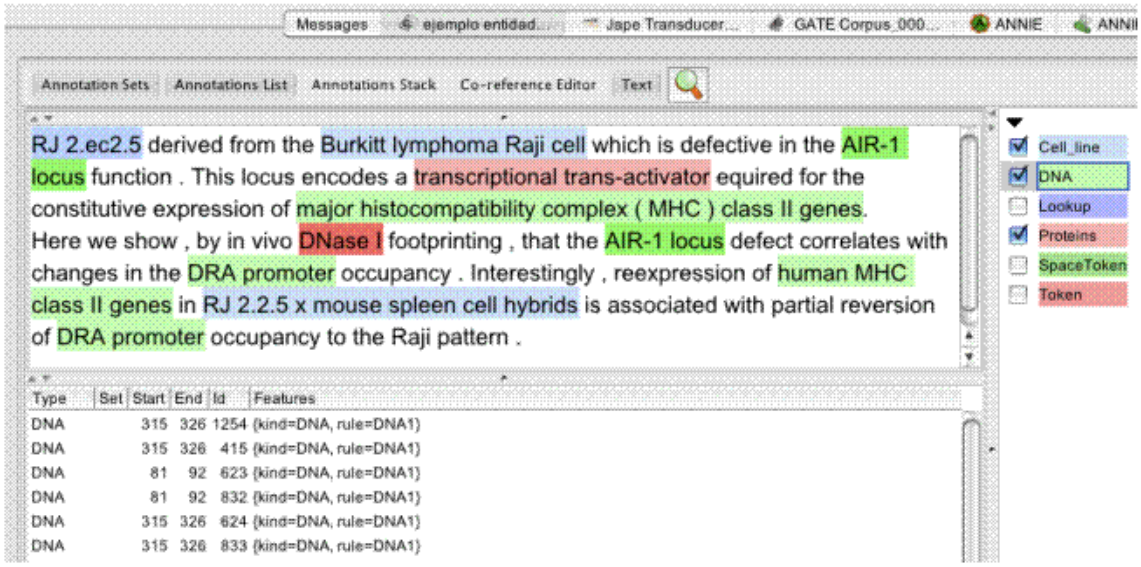


Figura 6.10 Entidades nombradas obtenidas mediante GATE.

En la figura 6.10, se muestra un ejemplo de anotación de entidad nombrada, en este caso las entidades identificadas son "Cell_line", "DNA" y "Proteins", de las que se han localizado varias menciones en el texto. Por ejemplo, en la primera línea de la figura se han identificado dos entidades nombradas del tipo "Cell_line", *RJ 2.ec2.5* y *Burkitt lymphoma Raji Cell*.

6.5.3 Fase de instanciación de la ontología

Una vez que las entidades han sido extraídas, se procede a detectar en el texto las posibles relaciones entre las entidades. Es decir, se localizan aquellas unidades léxicas asociadas a los *frame* de FameNet que se encuentran entre dos entidades.

Un *frame* conecta las dos entidades más cercanas a dicho marco semántico, siempre y cuando cumplan con ciertos requisitos en cuanto a la distancia que existe entre ellas. La distancia es un parámetro configurable por el usuario. No obstante, las unidades léxicas que conectan entidades no suelen aparecer a una distancia mayor de 3 *tokens* de dichas entidades. Además, si a la izquierda de la primera entidad relacionada o a la derecha de la segunda entidad relacionada hay una conjunción copulativa o coordinada, entonces la distancia se actualiza y se contabilizan de nuevo tres términos desde la conjunción.

Por otro lado, si a la izquierda o a la derecha de la relación aparece un punto, la relación no se considera válida aunque haya una entidad a menos de tres *tokens* de distancia.

En la figura 6.11, se puede ver una representación gráfica del proceso.

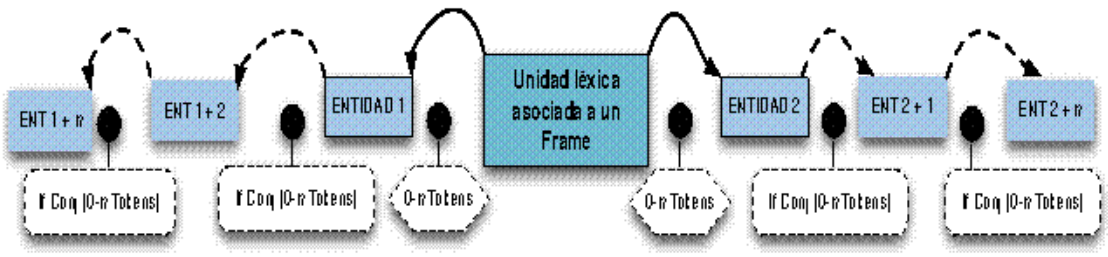


Figura 6.11 Extracción de relaciones en el texto.

Por otro lado, la figura 6.12 representa una visión general del proceso en el que las entidades candidatas a instancias de la ontología se incluyen provisionalmente en la clase correspondiente de la ontología. Una vez que el razonador comprueba si cumplen con axiomas ontológicos definidos, las entidades pasan a formar parte de la ontología como instancias.

En el texto que aparece en la figura 6.12, se encuentra la unidad léxica *be transformed into*, que está asociada al *frame* de FrameNet *transform*. Este *frame*, a su vez, está mapeado con la relación ontológica *transformation_of*, de manera que se pueden obtener las entidades más cercanas a la derecha y a la izquierda del marco semántico, esto es, *cardiomyocytes* y *adult mesenchymal stem cells*.

Por lo tanto, la relación que se establece entre ellas es:

cardiomyocytes TRANSFORMATION_OF *adult mesenchymal stem cells*.

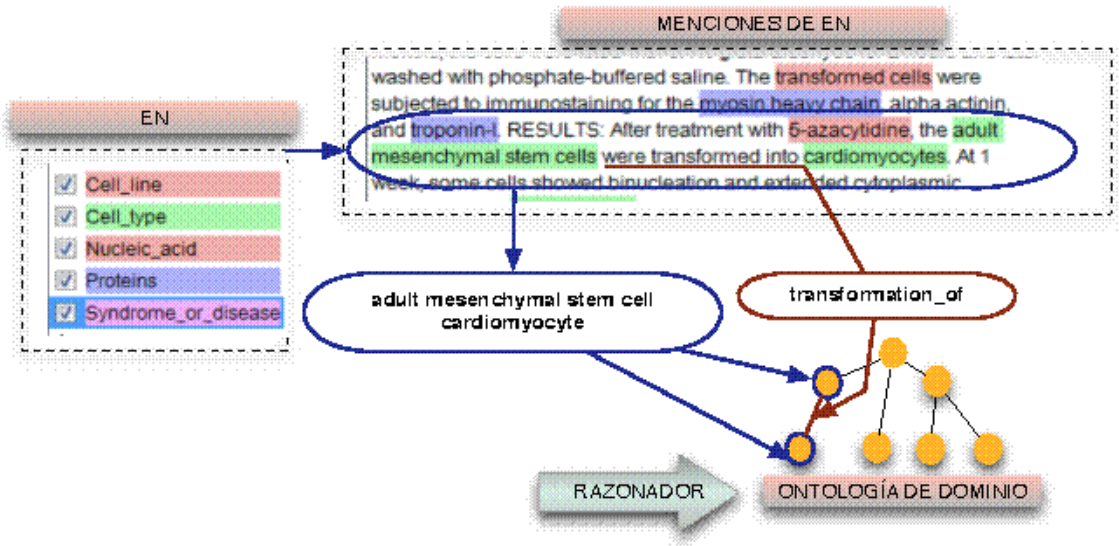


Figura 6.12 Ejemplo de instanciación de la ontología.

A partir de un conjunto de entidades nombradas y las relaciones ontológicas detectadas entre ellas, el sistema determina si son individuos o instancias de la ontología. En primer lugar, la ontología de dominio que se pretende instanciar debe definirse siguiendo el modelo ontológico BioOntoVerb OM. Las propiedades definidas entre las clases deben estar asociadas con el máximo de relaciones posibles del modelo ontológico.

Una vez que las entidades han sido anotadas en el texto y que se han localizado las posibles expresiones lingüísticas asociadas con un *frame*, se obtienen los

individuos que participan en cada una de las relaciones, es decir, las entidades que han sido previamente anotadas en el texto. Estas entidades se insertan de forma provisional en la ontología como individuos de aquellas clases relacionadas por medio de la *Object Property* identificada.

Dado que una relación se puede establecer entre una o más clases, la elección de la clase correcta se lleva a cabo comprobando si un tipo de relación entre dos entidades existía previamente. En caso de ser así, esta relación no se inserta de nuevo.

Además, un razonador comprueba que la ontología es consistente después de la inserción de las instancias en las clases. Si los individuos cumplen los axiomas previamente definidos, entonces pasan a ser instancias de una clase. Es probable que, debido a las relaciones de una instancia en la ontología, esta instancia pueda clasificarse en otra clase mediante las opciones de cálculo de tipos inferidos que permiten los razonadores basados en lógica descriptiva.

Por otro lado, algunos tipos de entidades nombradas están asociados a clases de la ontología. Este tipo de asociación sólo es posible cuando el tipo de entidad coincide con el nombre de una clase de la ontología. Entonces, aquellas anotaciones de entidades que en la fase previa han sido clasificadas dentro de un tipo de entidad, se convierten en candidatos a instancias de la ontología. Estos candidatos se insertan provisionalmente en la clase o clases correspondientes. Una vez que el razonador comprueba la consistencia de la ontología, se eliminan aquellos individuos que son inconsistentes.

Por ejemplo, *Myosin heavy chain* y *troponin I* son menciones de la entidad nombrada del tipo *Proteína*. Si la entidad nombrada *Proteína* está asociada en la ontología con la clase *Protein*, entonces esas menciones de entidades se convierten en individuos de la clase *Protein*.

Al finalizar todo el proceso de instanciación, el razonador comprueba de nuevo si son consistentes y si se puede inferir nuevo conocimiento. Si la ontología es inconsistente, se elimina la última relación insertada. En el caso de que la

ontología sea consistente, y que el razonador haya inferido que alguno de los individuos implicado en una relación se puede clasificar en una clase nueva, entonces se lleva a cabo la nueva clasificación. En el ejemplo mencionado más arriba, tanto *Myosin heavy chain* como *troponin I* pasan a ser instancias de la clase *Protein*, y heredan todas las relaciones de esta clase.

6.6 Evaluación de la metodología en el dominio de la biomedicina

La evaluación de la metodología descrita se ha llevado a cabo mediante el desarrollo de un prototipo que implementa dicha metodología. La plataforma se ha desarrollado en Java y OWLAPI y se ha utilizado para procesar el contenido de las ontologías. La librería OWLAPI no sólo provee una API enriquecida para tratar ontologías OWL, sino que también facilita el uso de razonadores OWL, de manera que se pueden emplear lenguajes de consulta tales como SPARQL[12] o DLquery[13]. La descripción detallada de la API utilizada para la evaluación está fuera de los límites de la investigación en cuestión, centrándonos aquí en el desarrollo de la metodología y los recursos necesarios para el correcto funcionamiento del prototipo.

Para el experimento, se ha utilizado una ontología extraída de xGENIA (Rak et al., 2007). Como se ha visto, xGENIA es una ontología OWL-DL basada en la taxonomía de GENIA y que fue desarrollada como resultado de la anotación

[12] SPARQL http://www.w3.org/TR/rdf-sparql-query/ es un lenguaje de consulta utilizado para obtener información de los grafos RDF. SPARQL se puede utilizar para expresar consultas en diferentes fuentes de datos, siempre que los datos se hayan almacenado originalmente como RDF o que se puedan visualizar como RDF vía middleware. El resultado de las consultas SPARQL puede ser un conjunto de resultados o gráficos RDF.

[13] http://protegewiki.stanford.edu/wiki/DL_Query

manual del corpus de GENIA. Tanto la ontología como el corpus se han utilizado como punto de referencia para testear y desarrollar las herramientas extracción de instancias biomédicas.

La metodología que se describe en esta memoria, contempla como un primer paso para llevar a cabo la instanciación, el mapeo de la ontología de dominio, en este caso xGENIA, con BioOntoVerb OM. No todas las *Object Properties* entre las clases de la ontología xGENIA se agrupan bajo las la relaciones genéricas representadas en el modelo ontológico BioOntoVerb OM (son relaciones mucho más específicas), de manera que se han realizado algunos cambios en las propiedades de la ontología para representar las relaciones como un subconjunto de las relaciones propuestas en nuestro modelo ontológico. En la tabla 6.10, se muestran las correspondencias entre la ontología xGENIA y BioOntoVerb.

Tabla 6.10 Mapeo BioOntoVerb y xGENIA.

Relaciones en BioOntoVerb OM	**Relaciones en xGENIA**
part_of	-
located_in	-
contained_in	Contain in
adjacent_to	-
transformation_of	producer of
derives_from	stems from
preceded_by	-
has_participant	-
has_agent	-
caused_by	Generated by induced by produced by promoted by stimulated by activated by affected by

La relación de BioOntoVerb que ha obtenido un mayor número de asociaciones es *caused_by*, a la que se han mapeado relaciones específicas cuyo significado indica causalidad.

Por otro lado, la relación *derives_from* se ha asociado con *stems_from*. Como se ha indicado previamente, la relación *stems_from* es una relación artificial creada en xGENIA con el objetivo de indicar la superclase de una clase semántica más específica aprovechando las características del corpus GENIA, que contiene etiquetas anidadas. No obstante, en el proceso de adaptación de la ontología xGENIA al modelo ontológico que proponemos, esta relación ha sido transformada en una relación regular, definiéndose para ella los axiomas correspondientes y asociándose a la relación de BioOntoVerb *derives_from*. En la figura 6.13, se muestra la ontología que se ha utilizado para la validación.

La ontología representa dos jerarquías, Substance y Source.

- *Substance* (Sustancia): en biología hace referencia al material por el cual está constituido un organismo. En la jerarquía existen dos tipos principales de sustancias, Atom (Atómicas) y Compound (Compuestos). Estas últimas se dividen a su vez en Inorganic (inorgánico) y Organic (orgánico).
- *Source* (Origen/Fuente): se refiere a una localización biológica donde se hallan las sustancias y tienen lugar las reacciones. En esta jerarquía, se han considerado tanto las fuentes naturales (Natural source) como las artificiales (Artificial source). Algunos de los hijos de Natural_source son Body_part, Tissue o Cell_type

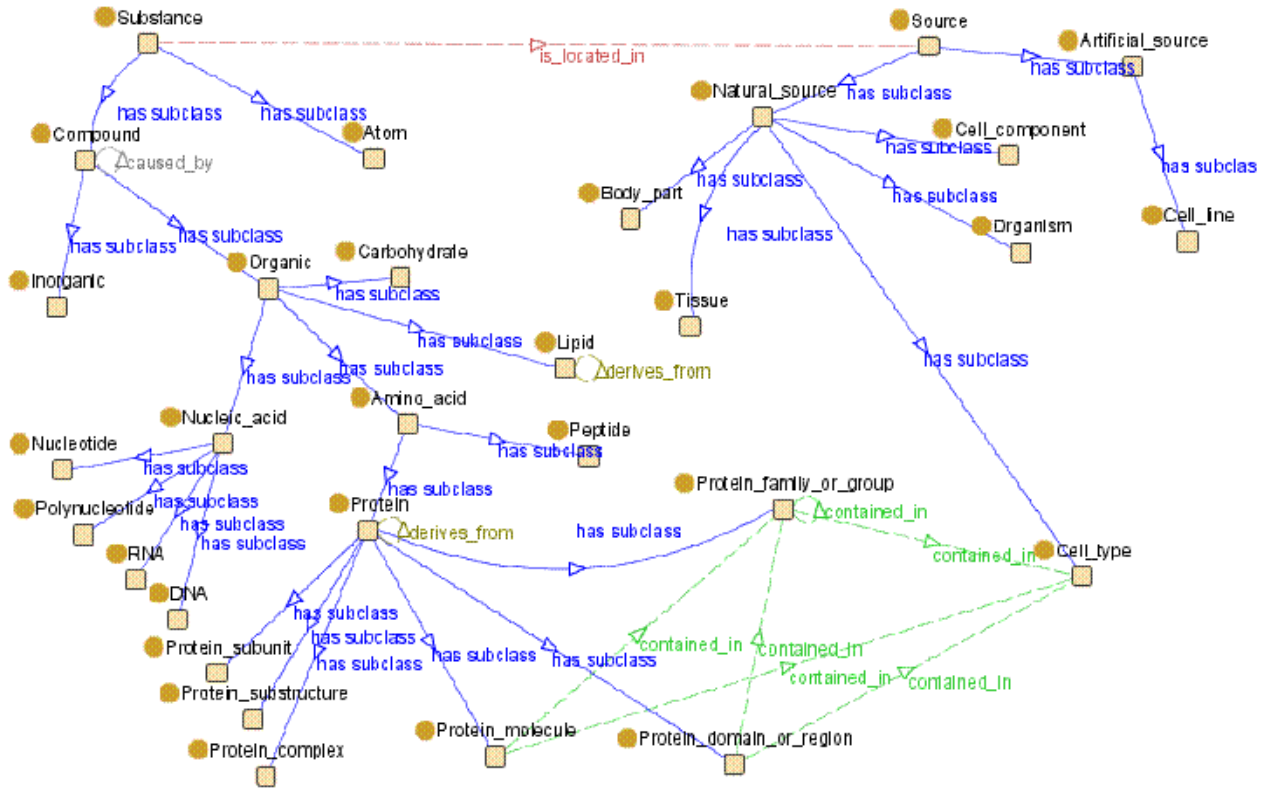


Figura 6.13 Extracto de la ontología instanciada.

A parte de esto, se han definido algunas relaciones no-taxonómicas, que son:

- *Substance is_located_in Source*
- *Compound caused_by Compound*
- *Protein derives_from Protein*
- *Lipid derives_from Protein*
- *Protein_molecule part_of Protein_complex*
- *Protein_subunit part_of Protein_complex*
- *Protein_molecule contained_in Cell_type*
- *Protein_domain_or_region contained in Cell_type*

- *Protein_family_or_group contained in Cell_type*
- *Protein_molecule contained_in Protein_family_or_group*
- *Protein_domain_or_region contained in Protein_family_or_group*
- *Protein_family_or_group contained in Protein_family_or_group*

Para extraer las instancias y las relaciones entre ellas, se ha utilizado una parte del corpus GENIA y BioText, en concreto 340.000 palabras.

Dado que el prototipo que se ha diseñado utiliza el identificador de entidades de GENIA, las medidas de exhaustividad y precisión obtenidas en la extracción de dichas entidades son altas, casi del 97%.

La meta del experimento es la evaluación de la exhaustividad y precisión en la detección de relaciones entre instancias durante el proceso de instanciación.

La ontología instanciada resultante se ha comparado con parte de la ontología xGENIA, que como se ha visto, estaba previamente instanciada. La ventaja es que, en este caso, no es necesaria la intervención de un experto para validar si las instancias obtenidas son correctas.

La evaluación se ha realizado sobre aquellas relaciones del modelo ontológico descrito en este trabajo, que se han mapeado con la ontologías de dominio, es decir, las relaciones que se muestran en la tabla 6.10.

Las medidas que se han utilizado para la evaluación son la precisión, exhaustividad y la medida-F.

La precisión se ha definido como el número de relaciones ontológicas extraídas que ya existían en la ontología xGENIA dividido por el número total de relaciones extraídas:

$$precisión = \frac{\text{Re}lacionesCorrectamenteExtraídas}{N°TotalDe\,\text{Re}lacionesExtraídas}$$

Otro parámetro que se ha utilizado para la evaluación es la exhaustividad, que se define como el número de relaciones ontológicas extraídas divididas por el número total de relaciones ontológicas que existen en el corpus.

$$exhaustividad = \frac{\text{Re}lacionesCorrectamenteExtraídas}{N°TotalDe\,\text{Re}lacionesEnElCorpus}$$

Finalmente, se ha evaluado la medida F (*F-measure*) que es el promedio ponderado de la exhaustividad y la precisión

$$medida - F = \frac{2(Exhaustividad.Precisión)}{Exhaustividad + Precisión}$$

El prototipo ha alcanzado una precisión total de 79.59%, una exhaustividad del 69.02% y una medida de F de 73,93%. Estos valores son altos, debido a que el experimento se ha llevado a cabo en un dominio específico, el de la biomedicina, y los roles semánticos seleccionados están estrechamente relacionados con los verbos y otras unidades léxicas que indican las relaciones en dicho dominio. En la tabla 6.11, se pueden ver los resultados obtenidos para cada relación.

Tabla 6.11 Resultados de la evaluación de la metodología.

Relaciones	Nº de relaciones en el corpus	Precisión	Exhaustividad	Medida de F
part_of	35	57,78%	74,29%	65,00%
located_in	51	73,33%	86,27%	79,28%
contained_in	26	57,50%	88,46%	69,70%
adjacent_to	0	-	-	-
transformation_of	34	68,42%	76,47%	72,22%
derives_from	254	85,99%	70,08%	77,22%
preceded_by	0	-	-	-

has_participant	0	-	-	-
has_agent	0	-	-	-
caused_by	436	81,18%	63,30%	71,13%
TOTAL	836	79,59%	69,02%	73,93%

La frecuencia de un verbo en el corpus, el número de roles asociados con una relación y los posibles mapeos entre BioOntoVerb y la ontología de dominio, son los factores que influyen en la obtención de mejores o peores resultados. Por ejemplo, para la relación *caused_by* se han obtenido 456 instancias. Esto se debe, por una parte, a que existe un elevado número de relaciones en el corpus GENIA y, por otra, a que el número de roles asociados con esta relación es también más elevado que la media. Esto implica, que la probabilidad de que aparezca una unidad léxica que represente este tipo de relación en el corpus sea también más elevado.

En el extremo opuesto se encuentran relaciones como *has_participant* o *adjacent_to,* que, aunque están contempladas en el modelo ontológico que aquí se propone, no están representadas en la ontología xGENIA. En consecuencia, el número de relaciones en el corpus que se han obtenido es de cero.

Es remarcable también la diferencia en la medida de precisión obtenida por algunas relaciones. Por ejemplo, *contained_in* y *part_of* han obtenido una precisión baja (57,50 y 57,78%, respectivamente). Esto se debe al hecho de que las unidades léxicas que representan a estas relaciones en el corpus no se corresponden con las unidades léxicas asociadas a los roles del modelo ontológico.

En cuanto a la exhaustividad, la relación *caused_by* es la que ha obtenido el porcentaje más bajo, es decir, 63,30%. Como se ha dicho, esta relación tiene asociados un elevado número de *frames* y, en consecuencia, de unidades léxicas,

las cuales no siempre representan una relación de causalidad. Los mejores resultados en cuando a exhaustividad se refiere, son los de las relaciones *contained_in* y *located_in*, con unos porcentajes de 88,46 y 86,27%, respectivamente. Estas relaciones están representadas por unidades léxicas con un grado de ambigüedad bajo. Además, esta relación suele aparecer explícita en el texto.

Finalmente, la metodología propuesta ha obtenido los mejores resultados en cuanto a la medida de F se refiere, en la relación *located_in*, con un porcentaje del 79,28%.

El total de las relaciones de instancias extraídas es de 836, una cifra alentadora, ya que supone el 69,02% de las relaciones en el corpus. Además, se ha obtenido una precisión total del 79,59% y una medida de F del 73,93%.

BLOQUE III

CONCLUSIONES Y TRABAJO FUTURO

CAPÍTULO 7

CONCLUSIONES, TRABAJO FUTURO Y CONTRIBUCIONES

Resumen. En este capítulo se ponen de manifiesto las principales aportaciones de las metodologías propuestas, así como sus limitaciones y las posibles soluciones.
Finalmente se enumeran las publicaciones que se han realizado, fruto de las investigaciones descritas en esta memoria.

7.1 Conclusiones y Contribuciones

Es indiscutible la revolución que ha supuesto internet y la era digital en cuanto al acceso a la información se refiere. La Web, a la que se conectan y aportan contenidos diariamente millones de personas, contiene información relevante tanto a nivel personal como profesional, información a la que no siempre es posible acceder de un modo eficiente. Esto se debe, a que gran parte de la información disponible en la Web no posee una estructura semánticamente inteligible para el ordenador, ya que se ha diseñado para el consumo humano. La Web Semántica tiene como meta proveer un mecanismo formal que permita organizar los datos de la web de manera que las máquinas sean capaces de explotarlos fácilmente. El dilema al que se ha enfrentado la Web Semántica desde su nacimiento es, que si dicha web no posee un contenido sustancial, se desarrollarán pocas herramientas para consumirlo, sin dichas herramientas, la posibilidad de publicar contenidos adaptados a la Web Semántica no es atractiva (Huyh et al., 2005). Es en este punto en donde entran en juego las ontologías. Las ontologías, que se presentan como potentes estructuras para organizar, compartir, reutilizar e inferir nuevo conocimiento, son costosas de crear y mantener. De este

modo, de forma paralela al nacimiento de la ingeniería ontológica, surge un campo de investigación dedicado a la creación automática de ontologías, de la que la instanciación automática forma parte.

Con frecuencia, el punto de partida de un sistema de instanciación de ontologías es una ontología parcialmente instanciada o una lista de individuos (generalmente en forma de entidades nombradas) y sus relaciones. Entonces, haciendo uso de analogías léxico estructurales a partir de un conjunto de patrones semilla, se procede a extraer, ya sea de la Web o de un conjunto de documentos previamente seleccionado, nuevas instancias de las clases y/o de las relaciones, que serán incluidas en la ontología.

El reconocimiento y clasificación de entidades nombradas es una subtarea de la extracción de información y tiene como objetivo la localización de ciertos términos relevantes para un dominio y su clasificación en unas categorías predeterminadas, como por ejemplo nombres de personas, localizaciones, direcciones, etc. Las entidades nombradas, se consideran un importante componente tecnológico para muchas de las aplicaciones de PLN, entre las que se incluyen, la extracción de información, los sistemas de pregunta-respuesta, o la población de ontologías.

En consecuencia, la instanciación de ontologías requiere de una infraestructura en la que se aúna el esfuerzo de varías tipologías de profesionales, de manear que, el punto de vista lingüístico en las tareas de PLN o el punto de vista de un documentalista en las tareas de gestión de la información, poseen la misma relevancia que el punto de vista tecnológico- computacional.

En esta tesis confluyen fundamentalmente estas tres disciplinas, en distintos grados, por un lado la lingüística, por otro la informática y finalmente la documentación. Y es desde la perspectiva de todas ellas desde las que se ha abordado el desarrollo de las metodologías propuestas.

Desde una perspectiva lingüística, esta tesis se interesa en los mecanismos que la lingüística ha utilizado tradicionalmente para el estudio científico del lenguaje humano en general, y de las lenguas en particular. Dando cuenta de los mecanismos de construcción del discurso, y de cómo los distintos lenguajes de

especialidad influyen en la configuración de herramientas de PLN, aplicadas a las producciones textuales de un dominio concreto. El análisis lingüístico tradicional puede conducir a conclusiones que permitan la optimización de las herramientas para el procesamiento y comprensión del lenguaje natural.

Desde una perspectiva documentalista esta tesis se interesa en los sistemas de organización del conocimiento en la Web, fundamentalmente en las ontologías, proponiendo nuevos mecanismos para la obtención de conocimiento a partir de texto y su organización ontológica. Como se ha visto, las ontologías presentan ciertas ventajas con respecto a otros sistemas de organización del conocimiento, independientemente de que su uso se centre sobre todo en el ámbito de la Web Semántica.

Desde un punto de vista informático, esta tesis se interesa en los mecanismos computacionales existentes para la extracción de conocimiento. Se realiza un análisis de las herramientas y recursos existentes.

En cuanto a la instanciación de ontologías, la realidad es que ninguna de las metodologías descritas en la literatura es capaz de extraer instancias de cualquier tipo de documento y cualquier dominio. Si bien es cierto, que partimos de la premisa de que no todo el conocimiento se pude modelar mediante ontologías. Las ontologías de carácter general son proclives a la ambigüedad y a una visión subjetiva de la estructura organizativa. Es difícil llegar a un consenso sobre cómo representar el conocimiento disponible en dominios generales, por este motivo, se ha creído conveniente restringir la aplicación de la metodología a dos de los dominios donde el uso de ontologías está más extendido, que son el turismo por un lado y la biomedicina por otro.

Los servicios turísticos son los más demandados por los usuarios de internet, como señala el Servicio Europeo de Estadística en una encuesta llevada a cabo en 2006. El 47% de los usuarios de internet habían utilizado servicios turísticos en los tres meses anteriores a la encuesta, y, con una media del 96%, el alojamiento, es el sector con el nivel de acceso más alto de internet. De hecho, el uso de servicios relacionados con los viajes y el alojamiento es la actividad más frecuente

que los usuarios llevan a cabo en internet después del envío de mails y de buscar información sobre bienes y servicios en general.

Por otra parte, para la comunidad biomédica, los dominios de la biología molecular y genética son de gran relevancia, prueba de ello es el creciente número de recursos, entre ellos ontológicos, que encontramos disponibles en la web.

En esta tesis se han propuesto dos metodologías para la instanciación automática de ontologías. La primera de ellas, descrita en el capítulo 5, permite poblar ontologías a partir de documentos en lenguaje natural utilizando la distancia cotextual y la ganancia de conocimiento. Se basa tanto en las tecnologías de la Web Semántica y como en las técnicas de PLN. Es un método basado en patrones pero con la ventaja de que:

- No es necesario disponer de un corpus anotado previamente
- Una vez que se han creado los recursos lingüísticos, el sistema es completamente automático.

Además, el sistema permite solucionar los casos de ambigüedad que se produce cuando existe más de una posible anotación para una entidad, eligiendo la mejor opción en función del contexto.

La segunda metodología, descrita en el capítulo 6, permite poblar una ontología ya existente que se haya mapeado con el modelo ontológico desarrollado. Este modelo, denominado BioOntoVerb, está basado en la asignación de *frames* semánticos a las relaciones de una ontología de alto nivel. Además, cuenta con un módulo para el reconocimiento e identificación de entidades nombradas basado en GENIA *tagger*, al que se han incorporado algunas listas de UMLS, lo que otorgan al sistemas una mayor cobertura. La ventaja es que el sistema no se limita a la extracción de EN del tipo proteínas y genes, sino que el conjunto de entidades nombradas que potencialmente se puede extraer es mucho mayor y se puede modificar y extender.

Finalmente, la consistencia de la ontología resultante está garantizada en ambas metodologías, gracias al razonador que se ejecuta al final del proceso, este paso es fundamental en un método que pretende ser completamente automático.

Entre las **principales ventajas** de las metodologías propuestas, se encuentra su modularidad, el hecho de que estén divididas en fases, y que reutilicen recursos ya desarrollados y disponibles gratuitamente, de amplia difusión, facilita su adaptación a diversos escenarios de aplicación.

En el caso de la metodología que se ha presentado para la biomedicina, la arquitectura modular otorga a dicha metodología una gran versatilidad, ya que el proceso de instanciación de la ontología no depende directamente de reglas lingüísticas desarrolladas a partir de un corpus específico. El modelo ontológico desarrollado es adaptable a aquellas ontologías biomédicas que contengan relaciones más específicas que se puedan englobar en otras más generales. El sistema es completamente modular y sus componentes se pueden adaptar incluso a otros dominios, siempre y cuando exista una ontología de alto nivel disponible para ser mapeada con los *frames* semánticos.

En cuanto a la metodología propuesta basada en la distancia cotextual y la ganancia de conocimiento, los recursos lingüísticos para el PLN se integran en GATE, lo que aporta flexibilidad al sistema, puesto que los plug-ins de GATE pueden ser modificados, bien reemplazando los existentes o desarrollando otros nuevos, dependiendo de las necesidades del dominio y de la naturaleza de los textos. Por ejemplo, el módulo de entidades nombradas se puede ampliar fácilmente mediante la inclusión de nuevas listas. En consecuencia, la metodología propuesta es flexible y portable a otros dominios, siempre y cuando se disponga de los recursos lingüísticos necesarios. De igual manera, el proceso de población de la ontología se ha diseñado para ser independiente del domino y del idioma

Por otro lado, en esta tesis se plantean una serie de procedimientos y técnicas lingüísticas que, mediante el análisis de un corpus en lenguaje natural representativo del dominio, pretenden, optimizar el desarrollo de herramientas y recursos para el Procesamiento de Lenguaje Natural y apoyar el desarrollo de ontologías.

Se parte de la premisa de que con la información que nos aporta el texto a todos los niveles (estructural, morfológico, sintáctico, pragmático y semántico), es

posible el establecimiento de reglas para la instanciación de ontologías de dominio. La búsqueda de regularidades de tipo léxico y morfosintáctico es una constante en los estudios enfocados a la enseñanza-aprendizaje de segundas lenguas, acentuándose este estudio en el caso del discurso de especialidad (Fernández Toledo, 1999), por lo que consideramos el uso de estas técnicas aporta información valiosa para el desarrollo de herramientas del procesamiento de lenguaje natural en general y para el desarrollo de recursos lingüísticos para el poblamiento de ontologías en particular.

Otro de los elementos innovadores que aporta esta tesis es la conjunción de diversos recursos para llevar a cabo la instanciación de la ontología. En la metodología validada en el dominio de la biomedicina se ha procedido a la asignación de roles semánticos a las relaciones de una ontología de alto nivel. Creemos que la reutilización e integración de recursos ya desarrollados es una solución válida para agilizar los procesos en la creación e instanciación de ontologías. En la actualidad, herramientas como las descritas en los capítulos 3 y 4, han alcanzado una madurez suficiente, permitiendo no tener que partir de cero a la hora de desarrollar un sistema para la extracción de conocimiento a partir de texto.

7.2 Trabajo Futuro

En este apartado, se mencionan cuáles podrían ser la líneas de trabajo futuras para cada una de las metodologías, teniendo en cuenta sus ventajas y limitaciones.

7.2.1 Metodología basada en la distancia cotextual y ganancia de conocimiento

La metodología presentada se ha validado en el dominio del turismo en documentos escritos en idioma español, obteniéndose resultados prometedores. Entre el trabajo futuro se encuentra la ampliación del rango de documentos utilizados a otros ámbitos del sector turístico, como por ejemplo la oferta cultural relacionada con determinados lugares.

Además, el análisis del discurso realizado muestra que las entidades normalmente aparecen en los textos en un orden particular y agrupadas por categorías. Por esta razón y porque la pérdida de información debida a entidades que no siguen este esquema es mínima, se pretende desarrollar un mecanismo que permita calcular el peso de las entidades en diferentes partes del documento, como se propone en (Zhu *et al.*, 2010). Sería un parámetro adicional para facilitar el proceso de desambiguación de entidades nombradas.

Por otra parte, dentro de la fase de reconocimiento de Entidades Nombradas podrían incorporarse nuevas técnicas como la extracción de Menciones nombradas. Dentro de la tipología textual con la que se ha validado la metodología, al tratarse de textos descriptivos breves, los elementos cohexivos del texto, entre de los que se encuentran las menciones, no son necesarios, pero para otros dominios podría ser interesante su inclusión.

Entre las líneas de trabajo futuro se incluye también la adaptación del sistema a la evolución de ontologías. La evolución de ontologías u *ontology evolution* se puede definir como la actividad de adaptar una ontología existente al nuevo conocimiento producido como resultado de los cambios de un domino, preservando al mismo tiempo su consistencia. (Zablith et al., 2010)

La metodología aquí presentada está concebida para su aplicación a documentos de naturaleza cambiante, como por ejemplo los corpora del turismo, esto es, documentos cuya información es susceptible de ser modificada durante su ciclo de vida. Por ejemplo, se pueden añadir o quitar servicios, pueden aparecer nuevos hoteles y desaparecer otros, etc. Con la evolución de ontologías se podrá acceder periódicamente a documentos más actualizados, añadiendo, eliminando o modificando las entidades de conocimiento de la ontología necesarias. De esta manera las instancias erróneas u obsoletas podrán corregirse, dando como resultado una ontología siempre actualizada.

7.2.2 Metodología basada en *frames* semánticos

El proceso de reconocimiento y extracción de EN se lleva a cabo utilizando el módulo para la extracción de EN que posee GENIA, que está basado en técnicas de aprendizaje automático. Actualmente existen muchas bases de conocimiento en el dominio de la biomedicina, tales como UMLS y GO, y el uso de vocabularios controlados puede ser de gran utilidad para la identificación de entidades nombradas y otros términos en el dominio de la biomedicina. Por este motivo y dado que para la metodología propuesta es fundamental contar con un módulo de reconocimiento de entidades lo más preciso y exhaustivo posible, ya que de ello depende la posterior instanciación de la ontología, se está planeando la integración de nuevos recursos terminológicos en el módulo de reconocimiento de entidades nombradas. De hecho, entre las principales limitaciones del sistema se encuentra la pérdida de conocimiento o la adquisición de conocimiento erróneo durante la fase de reconocimiento de entidades.

Para la comunidad biomédica, los dominios de la biología molecular y genética son de gran relevancia, no obstante, no cubren las necesidades de sistemas de extracción de información enfocados a un rango de usuarios más amplio (profesionales de la salud, estudiantes, etc.). Por este motivo se pretende ampliar la metodología a otro tipo de textos de dominio médico de carácter más general.

Como se ha visto, prácticamente la totalidad de las herramientas desarrolladas para el dominio biomédico están en inglés, lo que no sorprende, ya que la lengua de la mayoría de las publicaciones científicas en este dominio es también el inglés. No obstante, existe una ingente cantidad de información de tipo biomédico publicada por organismos oficiales, tales como la Organización Mundial de la Salud o la biblioteca de Medicina de EEUU, dirigida a un público más amplio y que no tiene cabida dentro de estos sistemas tan especializados. En consecuencia, se pretende expandir la metodología a otros idiomas, como el español, ya que, por un lado la traducción de las ontologías de alto nivel existentes es factible, y por otro la capa de recursos de PLN existente para español, aunque no tan desarrollada como la del inglés, es suficiente para llevar a cabo de forma fiable las

tareas de procesamiento necesarias. Por ejemplo, existen proyectos que mapean verbos en español con los *frames* de FrameNet, así como VerbNet y PropBank (Taulé et al., 2011), de manera que es posible la ampliación del modelo ontológico añadiendo la unidades léxicas de los *frames* en español.

7.3 Publicaciones y contribuciones a congresos

En este apartado se enumeran las distintas publicaciones realizadas en relación con lo expuesto en esta memoria de tesis.

7.3.1 Publicaciones JCR

Ruiz-Martínez, J.M., Miñarro-Giménez, J.A., Castellanos-Nieves, D., García-Sánchez, F. and Valencia-García, R. Ontology Population: An Application for the E-Tourism Domain. *International Journal of Innovative Computing, Information and Control*. 7(11) pp. 6115-6134, (2011) (impact factor 2010: 1.664)

Ruiz-Martínez J.M., Valencia-García R., Fernández-Breis J.T., García-Sánchez, F., Martínez Béjar, R. Ontology Learning from Biomedical Natural Language Documents Using UMLS. *Expert Systems with Applications, VOL. 38, 12365--12378, (2011)* (impact factor 2010: 1.926)

Valencia-García R., Fernández-Breis J.T., Ruiz-Martínez J.M., García-Sánchez, F., Martínez-Béjar, R. A Knowledge Acquisition Methodology to Ontology Construction for Information Retrieval from Medical Documents. *Expert Systems with Applications, VOL. 25, 314--334, (2008)* (impact factor 2009: 1.231)

7.3.2 Contribuciones a Congresos

Ruiz-Martínez J.M., Valencia-García R., Martínez-Béjar, R. BIOONTOVERB Framework: Integrating Top Level Ontologies and Semantic Roles to Populate Biomedical Ontologies. NLDB 2011. *Lecture Notes In Computer Science, VOL. 6716, 282--285, (2011)*

Ruiz Martínez J.M., Valencia García R., Martínez Béjar, R., Hoffmann A. Populating Biomedical Ontologies from Natural Language Texts. *IC3K 2010, 2nd International Joint Conference On Knowledge Discovery, Knowledge Engineering And Knowledge Management*, Valencia, Spain, 2010.

Lupiani-Ruiz E., Ruiz-Martínez J.M., Valencia-García R., Vivancos-Vicente P.J., Castejón-Garrido J.S. Invoca: Consultando la Linked Open Data en Lenguage Natural. *Procesamiento De Lenguaje Natural, VOL. 45, 323--324, (2010)*

Navigli R., Velardi P., Ruiz-Martínez J.M., An Annotated Dataset For Extracting Definitions and Hypernyms from the Web. LREC 2010: 7th Language Resources And Evaluation Conference, Valletta, Malta, 2010.

Ochoa-Hernandez J.L., Almela-Sánchez-Lafuente A., Ruiz-Martínez J.M., Valencia-García R. Efficient Multiword Term Extraction in Spanish. Application to the Financial Domain. International Conference On Intelligence And Information Technology ICIIT 2010, Lahore, Pakistan, 2010.

Ruiz Martínez J.M., Castellanos Nieves D., Valencia García R., Fernández Breis J.T., García-Sánchez, F., Vivancos Vicente P.J., Castejon Garrido J.S., Martínez Béjar, R. Accessing Touristic Knowledge Bases Through a Natural Language Interface. *Lecture Notes In Computer Science, VOL. 5465, 147--160, (2009)*

Ruiz-Martínez J.M., Guillen-Cárceles L., Valencia-García R., Martínez-Béjar, R. Método Para Poblar Ontologías en El Dominio del eTurismo. En: *Actas De La XIII Conferencia De La Asociación Española Para La Inteligencia Artificial (CAEPIA-TTIA 2009).* 179--189 (2009)

Ruiz Martínez J.M., Castellanos Nieves D., Valencia García R., Fernández Breis J.T., García-Sánchez, F., Vivancos Vicente P.J., Castejon Garrido J.S. Querying Ontology-Based Systems in Natural Language in *The E-Tourism Domain. Pacific Knowledge Acquisition Workshop*, Hanoi, Vietnam, 2008.

Ruiz Martínez J.M., Miñarro Gimenez J.A., Guillen Cárceles L., Valencia García R., García-Sánchez, F., Fernández Breis J.T., Martínez Béjar, R. Populating ontologies in the eTourism domain. *Workshop On Natural Language Processing and Ontology Engineering, IEEE/WIC/ACM International Conference On Web Intelligence and Intelligent Agent Technology,* Sydney, Australia, 2008.

Ruiz Martínez J.M., Valencia García R., Fernández Breis J.T., Martínez Béjar, R. La Anotación Semántica para la Consulta De Ontologías En Lenguaje Natural 249--258. En: *Acceso Y Visibilidad De Información Multilingüe En La Red: El Rol De La Semántica.* 978-84-362-5609-3, 2008

BLOQUE IV

SUMMARY

RESUMEN EN INGLÉS

CHAPTER 8 (CAPÍTULO 8)

ENGLISH SUMMARY (RESUMEN EN INGLÉS)

Abstract. The Semantic Web aims to extend the current Web standards and technologies so that the semantics of Web contents is machine processable. For the Semantic Web vision to become real, methods and mechanisms that assist in the creation of an initial pool of semantically described Web resources must be developed. On that scenario, the provision of valuable ontology-based knowledge services is an important step towards. Ontology population is a knowledge acquisition activity that relies on (semi-) automatic methods to transform un-structured, semi-structured and structured data sources into instance data.

The Ontology population task implies linguists, computer specialists and researches, among others, and several resources related with Natural Language Processing and Ontology Engineering.

In this research, two methodologies for ontology population are presented. The description of the methodologies emphasize on the linguistic aspects of the process, such as the discourse analysis and the use of linguistic resources such as WordNet or FrameNet.

The first methodology is based on the co-textual distance and the knowledge gain. Once a discourse analysis is performed and by using the GATE framework, the system obtains, a set of semantic annotations, which are considered as ontology instance candidates. In a second stage, the semantic ambiguities are solved, and the annotations are related with their corresponding ontological entities. The methodology has been tested in the tourism domain.

The second methodology is based on mapping semantic frames onto top-level ontologies. The result is an ontological model utilized for extracting relations between named entities. The entities implied into these relationships become ontology instance candidates. The methodology has been tested in the biomedical domain

8.1 Introduction

The information contained on Web pages was originally designed to be human-readable, and so most of the knowledge currently available on the Web is kept in large collections of textual documents. As the Web grows in both size and

complexity, there is an increasing need for automating some of the time consuming tasks related to Web content processing and management. In 2001, Tim Berners-Lee and his colleagues defined the Semantic Web as an extension of the current Web, in which information is given well-defined meaning, enabling computers and people to work better in cooperation (Berners-Lee, et al., 2001). The Semantic Web vision is based on the idea of explicitly providing the knowledge behind each Web resource in a manner that is machine processable. Ontologies (Studer et al., 1998) constitute the standard knowledge representation mechanism for the Semantic Web. The formal semantics underlying ontology languages enables the automatic processing of the information in ontologies and allows the use of semantic reasoners to infer new knowledge. In this work, an ontology is seen as "a formal and explicit specification of a shared conceptualization" (Studer et al., 1998). Ontologies provide a formal, structured knowledge representation, and have the advantage of being reusable and shareable. They also provide a common vocabulary for a domain and define, with different levels of formality, the meaning of the terms and the relations between them. Knowledge in ontologies is mainly formalized using five kinds of components: classes, relations, functions, axioms and instances (Gruber, 1993). Ontology Web Language (OWL) is the W3C standard for representing ontologies in the Semantic Web and, in this thesis, it has been used to represent the knowledge extracted from texts.

Ontologies are thus the key for the success of the Semantic Web vision. The use of ontologies can overcome the limitations of traditional natural language processing methods such as text classification (Yang et al., 2009). They are also relevant in the scope of the mechanisms related, for instance, with Information Retrieval (Park et al., 2009; Rung-Ching et al., 2010), Service Discovery (Zhang et al., 2009) or Question Answering (Yang et al., 2009). However, creating and manually populating ontologies is a very time-consuming and labor-intensive task. Several methodologies for ontology learning and ontology population have been created in order to assist in building ontologies. Yet, none of the current

proposals is scalable enough to deal with the ontologization of the Web content bulk.

Ontology Learning (also named ontology generation or ontology extraction) is a knowledge acquisition activity that relies on (semi-) automatic methods to transform unstructured (e.g. corpora), semi-structured (e.g. folksonomies, html pages, etc.) and structured data sources (e.g. databases) into conceptual structures. Ontology Population, on the other hand, is a knowledge acquisition activity that relies on (semi-) automatic methods to transform un-structured, semi-structured and structured data sources into instance data. Thus, while Ontology Learning deals with the acquisition of new concepts and relations with the consequence of changing the definition of the ontology itself (Tanev & Magnini, 2006), the goal of Ontology Population is the extraction and classification of instances of the concepts and relationships defined in the ontology. The instantiation of the ontology with new knowledge is a relevant step towards the provision of valuable ontology-based knowledge services.

8.2 Aims of the thesis

The two methodologies proposed in this thesis target the automatic ontology population in a way that combines the traditional linguistic analysis and technology for textual knowledge extraction. Moreover, we make use of a combination of already developed linguistic and ontological resources in order to carry out the instantiation process. With all, the objectives of this thesis are:

- To analyse, from a linguistic point of view, the characteristics of a specialized language. Linguistics, as the scientific study of the human language, has developed around the centuries. Even if statistic and computational methods are required for developing tools for natural language processing, these methods will be considerably more effective taking into account the linguistic features of texts.
- To design and implement a methodology for automatic instantiation of ontologies based on the co-textual distance and the knowledge gain. The

co-textual distance is defined as the physical distance between two linguistic units in the text, while the knowledge gain is the measure of the quantity of knowledge acquired by the system. In other words, the more knowledge acquired in the form of instances and relationships, the more the knowledge gained.

- To design and implement a methodology for ontology instantiation based on semantic roles. The methodology's framework requires also the design of an ontological model within which top-level ontologies are combined with both linguistic and ontological resources.
- To validate the methodology based on the co-textual distance and knowledge gain for tourism domain. The validation involves extracting instances from annotated texts concerning the restoration and hospitality domains with the objective of instantiating an ontology.
- To validate the methodology based on semantic roles. This operation requires the mapping between our ontological model and a biomedical domain ontology. Relations are extracted from a biomedical domain corpus.

8.3 Ontology population based on co-textual distance and knowledge gain

Chapter 5 presents an ontology population methodology based on co-textual distance and knowledge gain. The co-text, which is always an explicit element, refers to the linguistic set that surrounds a passage (Aznar, 1991), i.e. the words or sentences present before and after it (Bustos Gisbert, 1996).

The knowledge gain refers to the amount of information that the inclusion of a particular instance gives to the ontology. The gain metric helps to determine what class or property is assigned to an entity. All the possible classifications are calculated and the system chooses the one able to produce the highest gain.

A preliminary discourse analysis is executed on the training *corpora* in order to customize the NLP and ontological resources for our particular application domain.

Ontology population requires the existence of certain linguistic resources where to obtain the instances from, i.e. a corpus. *A corpus is a collection of pieces of language that are selected and ordered, according to explicit linguistic criteria, in order to be used as a representative language sample* (Sinclair, 1996). In this work, two corpora, *Hotels Corpus* and *Restaurants Corpus*, obtained from an official tourism Web page[1] have been compiled.

The *Restaurants corpus* consists of a Spanish language description of 848 restaurants (67,500 words approximately). The *Hotels corpus* comprises the description, in Spanish language, of 112 hotels (14,000 words approximately). For the development of the linguistic resources, 200 restaurant descriptions and 40 hotel descriptions have been considered. The remaining descriptions have been used for system testing.

A particular domain or subject matter is characterized by a specialized vocabulary, semantic relations and syntax (Sabou et al., 2005).

The development of language resources, customized for the needs of a sublanguage, requires an exhaustive knowledge of the discursive practices of the domain.

A deep discourse analysis, based on the parameters proposed by Bhatia (1993), was performed. The results of the analysis showed that these typologies of tourism texts contain repetitive discourse patterns. Some basic features of the tourism sublanguage are:

- Specialized vocabulary: *Check-in, overbooking, charter.*
- The vocabulary is related to others disciplines such as art, architecture or activities.
- Evaluative vocabulary: *Excellent situation, Magnificent views, Friendly helpful service.*

[1] http://www.murciaturistica.com

- A frequent use of nominal style "*Only 2 minutes from Paddington Station with the Heathrow Express service, and close to Hyde Park*"
- The usage of simple sentences and lists of characteristics and services.

The lexical, syntactic and semantic regularities are useful for the creation of the rules and lists identifying named entities. Some of the most relevant named entities identified in the context of the tourism domain are: Hotel, Services, Address, Postal Zip, Telephone, Fax, Country, City, Municipality, Beach, e-mail, Web Page, Monuments, Architectonical style, Airport, Restaurant, Menu, Meals, etc.

With the inferred information, we created the linguistic resources and a domain tourism ontology, which will allow to validate the methodology.

8.3.1 Description of the methodology

The methodology is comprised of four sequential stages: (i) the NLP and Corpus processing, (ii) the Named Entity recognition (NER), (iii) the Ontology population, and (iv) the Consistency checking. In a nutshell, the system works as follows. The NLP and corpus-processing module parses the corpus, in order to extract the linguistic information. The aim of the second stage is to obtain named entities from the text. NER is a subtask of information extraction that aims to locate and classify atomic elements within the text, into predefined categories such as names of persons, organizations, locations, expressions of times, quantities, monetary values or percentages. The more NEs are obtained the more information is gathered in the ontology population phase. During the third stage, the identified named entities occurrences are disambiguated and the ontology is populated. Each named entity occurrence becomes a candidate for one or various individuals of one or more classes in the ontology. If the entity had not yet been added, the system populates the ontology with the information extracted, creating new individuals. Otherwise, the pre-existing individuals are enriched by adding to

it the new attributes/relationships. A check of the populated ontology is finally performed by an OWL-DL reasoner.

NLP and Corpus processing phase. The main objective of this phase is to obtain the morphologic and syntactic structure of each sentence of the corpus. A set of NLP tools, including a sentence detection component, tokenizer, POS taggers, lemmatizers and syntactic parsers were developed using the GATE[2] framework (Cunningham, 2002). GATE provides for developing and deploying software components that process human language. GATE helps scientists and developers in three ways: (i) by specifying an organizational structure for language processing software; (ii) by providing a class library, that implements the architecture and can be used to embed language processing capabilities in diverse applications; (iii) by providing a development environment, built on top of the framework, made up of convenient graphical tools for developing components.

In particular, a Freeling POS-tagger (Asterias et al., 2006) plug-in has been developed and integrated into GATE. Freeling is an open source language analysis tool suite that provides language analysis services such as morphological analysis, PoS tagging and syntactic analysis. In this phase, the grammar category of each word of the sentence is identified, tokens are lemmatized and a syntactic analysis is performed.

Named Entity Recognition (NER) phase. During this second phase, the named entity candidates are identified by GATE. The output produced by each component of GATE is a set of annotations, that is, metadata associated with a particular section of the document content. Because the *corpora* used for testing are written in Spanish, it has required to building specific resources to deal with them.

Two main components are needed in the NER phase: a Gazetteer, and a JAPE Transducer. The role of the gazetteer is to identify named entities in the text based

[2] http://gate.ac.uk/

on candidate lists. The system obtains annotations for every word that appears in the gazetteer lists of entities that are relevant in the domain being analysed. Several lists with a variable degree of generality were created. Examples of general lists are: *locations, zip code, career, first names, surnames, and address identifiers* (e.g. '*street*', '*avenue*', '*C/*', '*square*', etc.). *Restaurant facilities, meals* or *occupational* categories are examples of more specific lists.

The JAPE transducer is a module for executing JAPE grammars. JAPE is a rich and flexible, regular expression-based, rule mechanism offered by the GATE framework. Hence, a set of JAPE rules to obtain occurrences of zip codes, telephone, fax or mobile numbers, urls, emails, addresses, restaurants, person names or money references has been implemented.

Each annotation obtained by the JAPE transducer is considered as a NE. All the occurrences of the identified NEs in the text are candidates to be instances or values of the attributes of an instance. A NE representing a Hotel for example, *Hotel la Manga Golf*, can be considered as a candidate of an instance of the class Hotel, and the NEs that represent emails or phone numbers are then considered as candidates of possible values of instance attributes of the ontology (e.g. the phone number of a hotel).

The system classifies the annotations in the form of groups (i.e. Activity, Hotel, Restaurant, etc.). Each group can have one or more NE's annotations and some may be ambiguous. This happens when the same fragment of text is annotated into two different groups. This kind of ambiguity is resolved in the next phase, where the system understands whether they match with some instance, relationship or property of the ontology.

Ontology Population Phase. During this phase the system determines whether they are instances, attributes or relationships for the ontology model. The system must associate each annotation to a particular ontology entity. In OWL, the main type of resources are Classes, 'Subclass of' relationships, Datatype Properties, Object Properties and Individuals. The ontology model is defined by classes,

relationships that connect those classes, and datatype properties, which are the attributes belonging to each class.

The methodology for performing the ontology population consists of four main phases: (i) gathering the NEs identified in previous phases, (ii) creating a tree of combinations of ambiguous NEs, (iii) calculating the score of all combinations, and (iv) inserting individuals in the ontology model.

The first step takes as input the list of annotated NEs that the Corpus Processing and NER phases have identified in the text. At this point it is necessary to resolve the conflict introduced by the previous phase, not related to recognition mistakes but to language ambiguity. Different kinds of ambiguities might appear:

a) An annotation is related with more than one NE. For example: "Guggenheim" can be a surname and a museum name.
b) Several NEs overlaps in the text. For example, "Chelsea football club" is a NE which has an overlapped NE "Chelsea"
c) An annotated NE is related to several resources in the ontology. For example, the number "22358897" may be a phone number or a fax number.

The input to the second step is the list of annotated NEs described above. This step deals with the occurrences of ambiguities of type 'a'. Those ambiguities are avoided by defining all possible groups of non-ambiguous annotations.

For example, let us suppose that we have a Hotel description where several entities occurrences were annotated as Hotel, Address, Phone, Activity and Location, respectively.

Let us suppose that in the Hotel description we find the expression "Ritz-Carlton New York Central Park", which overlaps three different labels, namely

1- "Ritz-Carlton New York Central Park" as a Hotel;

2- "New York" as a Location; and

3- "Central Park" as a Location.

Ambiguity is present because the Hotel annotation is overlapped with the Location annotations. In order to remove this ambiguity, it is necessary to create

two separate groups of annotations, one with the Hotel annotation and the other with the Location annotations.

The non-ambiguous groups are represented in the form of a tree, where each level of the tree represents a NE. Sibling nodes represent ambiguous annotations related to a NE. The annotations in sibling nodes are incompatible between them. In the example, "Ritz-Carlton New York Central Park" is a node, and its sibling node is "New York, Central Park". The other levels are the nodes: "NE mention Address", "NE mention Phone", "NE mention Activity".

In the third step of the methodology, a score is assigned to each branch of the tree representing a group of non-ambiguous annotations. We have developed an algorithm to evaluate all possible groups of NEs. The algorithm is based on the co-textual distance and the knowledge gain. The assessment of each group is based on the number of annotations that can be mapped on to the ontology and the number of relationships that may be created among them.

The NE occurrences in text are usually surrounded by other annotations within the same scope that can be linked. The distance between two annotations in the text is a measurable parameter.

On other hand, the more the number of NEs and relationships are created in the ontology, the better the score that this group can achieve. Following with the example above, "Ritz-Carlton New York Central Park", according to the ontology, could be related with Address (Hotel has_Adress), Phone (Hotel has_Phone), Activity (Hotel has_Activity) and Location (Hotel has_Location). However, "New York" and "Central Park" cannot be related with any ontological property.

Briefly, the algorithm works as follows:

(1) The input parameter of the algorithm is the "*NE_list*". This list contains all the annotations of NEs identified.

(2) The tree represents all the allowed NEs groups. For each incompatible annotation, a sibling node is added to the tree. The depth of the tree is the number of NEs.

(3) The algorithm visits all nodes from the root to the leaves, in depth-first order. When the algorithm reaches a leaf node it calculates the score of

the group of annotations. Each group of non-ambiguous annotations is mapped provisionally into the corresponding ontology classes and properties. New instances are linked to the closest instances surrounding them in the text. When two instances can be combined, a new relationship is created. The score of each annotation depends also on the distance of the entities that are linked in the text.

At the end, when all groups are generated and scored, the one with the highest score is returned. Once ambiguity problems have been removed and the group of annotations with the highest score has been identified, ie. the group with the highest knowledge gain, it is possible to initiate the population of the ontology.

(4) Finally, a reasoner checks the consistency of the ontology.

Evaluation. A part of both copora was used to measure the precision, recall and f-measure of the methodology. The measurable elements are individuals, object properties and data type properties. The best value obtained was a 98.66%, refering to the precision of the individual of Restaurants corpus. On the contrary, the worst value was the recall obtained for the data type properties of the Hotel corpus (81.19%). The main reasons for these significant values are that (1) both domains are quite specific, and (2) the linguistic analysis performed has allowed the creation of fitted linguistic resources.

8.4 Ontology population based on semantic roles

Chapter 6 presents a framework, called BioOntoVerb, for ontology population from biomedical natural language text. The framework attempts to support domain experts in building ontologies from natural language texts. It allows for several semantic relations to be used and reduces the degree of expert participation during the ontology construction process in the biomedical domain.

It is based on the association of semantic frames from FrameNet with an ontological model created from top-level ontologies.

In (Smith et al., 2005), the most common relations used in biomedical domain ontologies are presented and formalized. As a result of this effort, the OBO ontology of biomedical relations was generated. The OBO Relation Ontology comprises ten different types of relations including taxonomic and partonomic ones and other approach to formally represent relations between ontologies is BioTop, which is an upper-domain ontology for the life science domain intended to link and integrate various specific domain ontologies (Beisswanger et al., 2008). BioTop follows the formal design principles described by the OBO Foundry and it is implemented in OWL-DL.

On the other hand, *a semantic role is the relationship between a syntactic constituent and a predicate. It defines the role of a verb argument in the event expressed by the verb* (Moreda et al, 2010). In FrameNet (Baker et al., 1998) roles are defined for each semantic frame, that is, a schematic representation of situations involving various participants, properties, and other conceptual roles.

The BioOntoVerb framework is based on the integration of the above referenced ontological and linguistic resources. Its architecture is logically divided into three different layers, as the figure 8.1 shows:

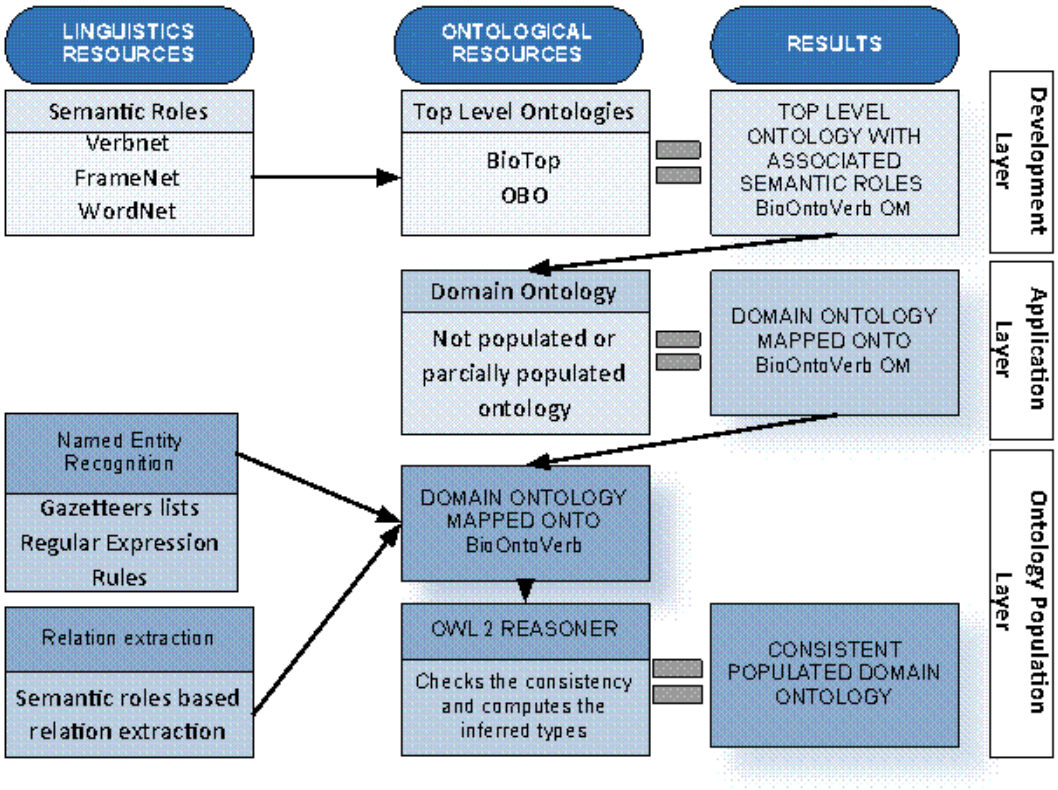


Figure 8.1 Architecture of the BioOntoVerb framework

Development Layer or Ontological model definition. In this layer, a top level ontological model, based on the different types of relations described in OBO Relations Ontology (Smith et al., 2005) and the relationships defined in the BioTop ontology (Beisswanger et al, 2008) is defined. This ontological model, called BioOntoVerb OM, allows to define domain ontologies based on the semantic relationships that are of common usage in biomedical domains. Here, relationships are expressed by means of Object Properties in OWL 2 and some property axioms have been assigned to each relation including transitivity, asymmetry or symmetry. For each of these Object Properties, a set of frames extracted from FrameNet was assigned in order to detect ontological semantic relations between instances from natural language texts. Each frame is associated with a set of verbal expressions or lexical units representing the relationships in a

textual level. These verbal expressions were obtained from Wordnet[3] and VerbNet[4].

The association, between roles and relationships of the ontology are manually assigned. For each role associated with a relationship in the ontology, some examples supporting this relationship are searched in the corpus. It is a feedback-based process until the list of ontological relations corresponding to linguistic expressions and roles has been configured.

The result is an ontological model integrated into the populating framework.

Application Layer. Here, the domain ontology, i.e the ontology to be populated, is defined using our ontological model. For doing that, as many relations from the domain ontology as possible are mapped onto the ontological model. The relationships, the defined properties, as well as the associated frames from the ontological model become part of the domain ontology.

Ontology Population Layer. This layer implements the ontology population process through three stages: NLP Phase, NER Phase and the Ontology Population Phase.

(1) NLP Phase. The main objective of this phase is to obtain the morphologic and syntactic structure of each sentence. Some aspects of biomedical texts may affect the morphologic and syntactic analysis, such as the ambiguity caused by names and abbreviations that begin with capital letters; chemical and numeric expressions including non-alphanumeric characters such as commas, parentheses, and hyphens; participles of unfamiliar verbs that describe domain-specific events; and fragments of words (Tateisi & Tsujii, 2004). The GENIA tagger (Tsuruoka et al., 2005) is able to manage these problems more efficiently than a general POS Tagger. GENIA tagger was integrated in GATE framework.

(2) NER Phase. Named Entities are identified by the GATE Framework where the GENIA Named Entities module was integrated. A

[3] http://wordnet.princeton.edu/

[4] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

combination of JAPE rules and lists of Gazetteers are also used to perform the processes. Lists, containing biological terms extracted from GeneOntology[5] and UMLS[6], were added as Gazetteers.

All the occurrences of identified NEs are considered as candidate instances.

(3) Semantic frames are detected in the text in order to extract possible relationships between candidate instances. For example, in the text we find the phrase "NF-kappaB gamma-1 is found in cytoplasm". Here, "NF-kappaB gamma-1" and "cytoplasm" are NEs related by the lexical unit "is found in". This lexical unit, according to our ontological model, belong to the frame *Locating*, which is related with the object property *Located_in*.

If the relationship does not already exist in the ontology, it is then created and added.

(4) A reasoner, such as Hermit, is executed in order to (i) check for the consistency of the ontology and (ii) compute inferred types. If the ontology is inconsistent, the last relationship inserted into the ontology is removed. In case the ontology is consistent and the reasoner has inferred that one individual belonging to the relationship can be classified into a new class, this new classification is done.

Evaluation. The framework has been validated by using an ontology extracted from the xGENIA ontology (Rak et al. 2007), which is an OWL-DL ontology based on the GENIA taxonomy. Some changes in its properties were required in order to represent them as a subset of the relations proposed in the ontological model. It achieved a precision value of 79.02% and a recall value of 65.4%. These values are significant because (1) the domain is a quite specific one, and (2) the semantic roles used were adapted to biomedical domain.

[5] http://www.geneontology.org/
[6] http://www.nlm.nih.gov/research/umls/

8.5 Related work and Discussion

In the last few years, a number of approaches have been applied to Ontology Population from unstructured text. Many of them combine natural language processing techniques (such as linguistic pattern recognition and extraction, POS-tagger and syntactic analysis) while others use machine learning techniques. In a deliverable of the BOEMIE project (Petasis et al., 2007), an analysis of the most prominent ontology population systems is provided. Here, different dimensions to compare these ontology population approaches are defined. These dimensions are subsequently explained and compared to our approaches.

Concerning the initial requirements, namely, resources or background knowledge, most of the approaches make use of Named Entity Recognition and Classification (NERC) modules. NERC is a subtask of information extraction that aims at locating and classifying the text atomic elements into predefined categories such as names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages (Murata et al., 2010). For example, in (Tanev & Magnini, 2006), the starting point is a training set of named entities (NEs) instances for each class under consideration. SPRAT (Maynard et al., 2009) considers the NEs identified by GATE as candidates for instances of the Ontology. In (Magnini et al., 2006) the system identifies the NEs mentions and connects them to concepts and relations already defined in the ontology. The methodology described in (Giuliano & Gliozzo, 2008) is also based on NEs substitution.

In our proposals, the NEs identified in the text are considered instance candidates of a predetermined ontology. Consequently, NERC has a major importance in the proposed frameworks, since the quality of this process leads to a more accurate and complete final ontology.

Other systems also need an annotated training corpus. For example, the main input of Ontoshopie (Celjuska & Vargas-Vera, 2004) is an XML annotated corpus where each entity is associated to the corresponding class in the ontology. In

(Tanev & Magnini, 2006), a syntactically parsed corpus containing training entities is needed. (Navigli & Velardi, 2006) makes use of parsed glossary definitions and a set of manually defined linguistic patterns, and OntoPop (Amardeilh et al., 2006) utilizes semantic annotations of texts. Finally, (McDowell & Cafarella, 2008) requires an ontology with a root class and a few simple linguistic patterns in order to extract concept instances and taxonomic relationships from the web.

Some approaches use Machine learning to populate the ontology. For example, Ontosophie (Celjuska & Vargas-Vera, 2004) relies on a conceptual dictionary that generates extraction rules. These rules are then used for the system training. In (Tanev & Magnini, 2006) an unsupervised algorithm, based on vector-feature similarity, is employed. The algorithm is applied to syntactically parsed corpus containing each training entity at least twice. For each occurrence in the corpus, syntactic features are obtained and used to construct the feature vector. The new feature vector is then compared to the existing ones, and the new instance is inserted in the class with the most similar feature vector. (Giuliano & Gliozzo, 2008) use a supervised machine learning approach to instantiate a semi-populated ontology from the Web.

Other approaches make use of manually constructed patterns as input, as in (Maynard et al., 2009), (Amardeilh et al., 2006), (McDowell & Cafarella, 2008). This is also the case of (Navigli & Velardi, 2006), where a set of rules defines regular expressions in order to annotate certain gloss fragment with the ontology properties (conceptual relations). The result is an annotated fragment where a pair of terms are associated by means of an ontology relation. These terms are considered the domain and range of the relationship. After a disambiguation process, the eligible terms are inserted in the ontology as individuals of the concept defining the annotated gloss. Finally, (Magnini et al., 2006) describes a manually created benchmark for ontology population where NEs mentions are assigned to concepts and relations already defined in the ontology.

The methodology presented here, based on co-textual distance and knowledge gain, requires predefined patterns for detecting NEs and a heuristic algorithm for populating ontologies, so no machine learning approach is applied.

On the other hand, the methodology based on semantic roles, requires a mapping between our ontological model, BioOntoVerb and a domain ontology.

Another parameter to classify ontology population approaches is the degree of automation. Some systems such as (Maynard et al., 2009; Giuliano & Gliozzo, 2008; Navigli & Velardi, 2006; McDowell & Cafarella, 2008) are unsupervised or weakly supervised (Tanev & Magnini, 2006), while others such as (Celjuska & Vargas-Vera, 2004) or (Amardeilh et al., 2006) need to be guided by an expert. The ontology population processes proposed here are fully automatic.

Some systems have been tested in a specific domain collection. For example, the framework described in (Navigli & Velardi, 2006) has been tested in the domain of cultural heritage and its portability requires new linguistic patterns to be developed in accordance with the domain and language. This is also the case of (Maynard et al., 2009; Celjuska & Vargas-Vera, 2004). Other systems extract generic NEs types like persons or locations (Tanev & Magnini, 2006; Magnini et al., 2006; Giuliano & Gliozzo, 2008; Amardeilh et al., 2006). Finally, Ontoshypon (McDowell & Cafarella, 2008) is a domain independent methodology.

The two methodologies above described have been tested in the tourism domain and in the biomedical domain. The first methodology is quite portable to other domains, while, the portability of the second one depends on the existence of a top-level ontology in the target domain.

Few frameworks declare whether or not a consistency check of the ontology is performed during or at the end of the process (Tanev & Magnini, 2006; Magnini et al., 2006; Giuliano & Gliozzo, 2008; Celjuska & Vargas-Vera, 2004; Navigli & Velardi, 2006; McDowell & Cafarella, 2008). In (Amardeilh et al., 2006), a manual maintenance of the knowledge acquisition rules is required, and they do not use any reasoner to check for the consistency of the ontology. Finally, in (Maynard et al., 2009), the GATE plugin used to insert the instances in ontology checks for the consistency of them before the insertion.

The methodologies proposed here include a step to verify the consistency of the populated ontology by using OWL-DL reasoners such as Pellet or Hermit.

Some systems perform some disambiguation tasks during the ontology population process. For example, in (Celjuska & Vargas-Vera, 2004), confidence values are assigned to the extracted entities and in the case of ambiguity, they select the value with the highest confidence. Other systems, such as (Giuliano & Gliozzo, 2008; Amardeilh et al., 2006), use context features to disambiguate. In (Maynard et al., 2009), even though the system can detect and warn about possible ambiguities, the disambiguation process largely depends on the end user. In (Tanev & Magnini, 2006), ambiguous NEs are allowed within the training corpus. However, if ambiguous NEs are found during the system execution, they are not included in the ontology. Some disambiguation strategies based on the inclusion of more information during the search of instances on the Web are proposed in (McDowell & Cafarella, 2008). Finally, other systems apply disambiguation methods during the process, but not necessarily concerning NEs. For example, (Navigli & Velardi, 2006) apply a semantic disambiguation algorithm based on structural patterns to the annotated glosses, and (Magnini et al., 2006) use different co-reference measures to address the problem of mentions disambiguation.

The methodology based on co-textual distance and knowledge gain runs an entity disambiguation process. However, the methodology based on semantic roles does not perform any disambiguation task during the ontology population process. Instead, it makes use of the ontology reasoner to infer if the ontology has been correctly instantiated.

Almost all the methods examined in this research provide support for English resources (Maynard et al., 2009; Giuliano & Gliozzo, 2008; Celjuska & Vargas-Vera, 2004; Navigli & Velardi, 2006; McDowell & Cafarella, 2008). Nonetheless, the language dependency degree is variable according to the portability of their linguistic components. It is possible to distinguish between strongly language-dependent systems like (Maynard et al., 2009), (Celjuska & Vargas-Vera, 2004; Navigli & Velardi, 2006), and weakly language dependent ones like (Giuliano &

Gliozzo, 2008), (McDowell & Cafarella, 2008). Only (Magnini et al., 2006) and (Amardeilh et al., 2006) take into consideration other languages such as Italian and French, respectively.

The system based on co-textual distance and knowledge gain has been tested with Spanish documents, but it could be easily ported to other languages by merely changing the initial language resources requirements.

The system based on semantic roles has been tested in English document, and its portability depends on existing top-level ontologies and semantic frames in the target language.

Regarding exclusive applicability in the biomedical domain, some systems just consider taxonomic and partonomy relationships. For example, the system proposed by (He et al., 2006) uses machine learning techniques and discourse analysis methods in order to extract protein-to-protein interactions. However, in (Sánchez & Moreno, 2008), a domain ontology is enriched by discovering non-taxonomic relationships from the web using patterns based on verb phrases. The methodology here proposes, populates as many relationship kinds as relations of the ontological model have been mapped onto the domain ontology.

In Ray and Craven (2001), the entities Proteins and Locations which hold the relationships subcellular-location and the entities Gene and Disorder in the relationships disorder-association are obtained by means of Hidden Markov Models. This methodology is also used in Rosario and Hearst (Rosario & Hearst 2004), where the authors identify semantic relations between "treatment" and "disease" in bioscience texts using graphical models and neural networks.

In (Bundschus et al., 2008), semantic relations between diseases and treatments are classified using Conditional Random Fields. In (Chun et al., 2006), relationships between genes and diseases from MedLine abstracts are obtained by studying the co-occurrence of terms. The relationships extracted in our approach are not limited to a few entities. The diversity of entities extracted by the system can be easily extended by means of gazetteer lists or other named entity recognition systems.

The algorithm presented in (Sharma, 2010) extracts five types of entities and the relationships between them. The verbs they use are from UMLS and the relationships are extracted based on the syntax of each sentence. In our methodology, we adopt a semantic approach, instead of a syntactic one. The advantage is that the entity extraction is not dependant on the structure of the sentence, so that for example the methodology is able to manage the diathesis alternation.

Verbs and their syntactic-semantic characteristics are used in other approaches in order to extract relationships. For example, in (Tsai et al., 2007) thirty verbs are annotated with predicate-argument structures. These structures are used for generating argument-type templates in order to extract new argument types.

In (Sahay et al., 2008), the authors manually construct some lexico-semantic patterns, as for example "d is caused by e", which are able to extract basic information from biomedical web snippets. With the so obtained information, they enrich some existing resources such as UMLS. The use of patterns is not a novel approach. The morpho-syntactic structures can give valuable information about the structures where the instances are inserted. However, with a semantic approach like the one here presented, it is possible to extract a wider range of information.

8.6 Conclusions and Future work

Most of the information currently available on the Web is not directly machine- accessible, because it has been designed for human consumption. The Semantic Web aims to provide a formal mechanism for organizing data on the Web in such a way that machines can easily access them. Disappointingly, the chicken-or-egg dilemma has accompanied the Semantic Web from its very conception: *without substantial Semantic Web content, few tools will be written to consume it; without many such tools, there is little appeal to publish Semantic Web content* (Huynh et al., 2005). In a fundamental step towards alleviating this

problem, in this work two methodologies for populating ontologies from unstructured web documents are proposed.

These methodologies allow for the automatic population of ontologies on the basis of Semantic Web Technologies and Natural Language Processing techniques.

The first methodology, based on co-textual distance and knowledge gain, is a simple and scalable methodology for ontology population from textual resources based on lightweight NLP techniques and ontological engineering. The methodology has been implemented in the form of a software prototype and tested in the tourism domain. It is worth pointing out that, although the prototype has been customized to deal with tourism-related texts, the methodology remains domain-independent.

The methodology under question is a pattern-based approach, but with the advantages that (1) there is no need for a previously annotated corpus, and (2) once the linguistic resources have been created, the system is completely automatic. The linguistic framework is integrated in GATE, which is widely used by the computational linguistics community. GATE allows the easy integration of linguistics resources depending on the needs of the domain and the nature of the texts under consideration. Consequently, the proposed approach is flexible and rather portable to other domains. On a related note, the ontology population process has been designed in order to be domain, language-independent. Moreover, the consistency of the ontology is checked at the end of the process, which is essential in a fully automatic method.

The validation of the proposed methodology in the scope of other application domains such as the financial domain is left for future work. Furthermore, the conducted analysis of discourse shows that NEs usually appear in texts in a particular order and grouped by categories. For this reason, a mechanism to calculate the weight of NEs in different parts of texts like the one proposed in (Zhu et al., 2010) could be used as an additional parameter to facilitate the NEs disambiguation process.

The second methodology, presented in chapter 6, is a semantic frame-based process for ontology population which provides a suitable framework for textual knowledge acquisition in the biological domain. In particular, with this approach a given ontology is enriched by adding instances gathered from biological natural language texts.

In this work, the NER process is performed using the GENIA Named Entities module, which is based on machine learning techniques. Currently, there exist many knowledge bases in the biomedical domain such as UMLS and GeneOntology and the use of such controlled vocabularies would be very helpful for identifying NE and terms in biomedical texts. It is planned to develop a new Named Entity module that can also use these ontologies. Indeed this system is not limited to protein and gene extraction but the set of Named Entities can be modified and extended.

Failure to perform a knowledge extraction process and/or erroneous knowledge acquisition one during the NER are the major limitations of the described system due to the fact that NEs not identified or misidentified affect the system's recall and precision.

A set of predefined semantic relations based on the OBO relation and BioTop ontology have therefore been defined. Unfortunately, some of the relationships between the instances of the GENIA corpus cannot be modelled with this set. As future work, the integration of some of the most commonly used relationships proposed in (Gómez-Pérez, et al 2000) such as, the associated_with relationship will be investigated.

REFERENCIAS BIBLIOGRÁFICAS

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. *Proceedings of the Fifth ACM Conference on Digital Libraries*, (pp. 85-94). Disponible en línea en http://dl.acm.org/citation.cfm?id=336644

Agirre, E., Ansa, O., Hovy, E., & Martinez, D. (2000). Enriching very large ontologies using the WWW. *Proceedings of the ECAI Ontology Learning Workshop in Conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000)*, Berlín, Alemania. Disponible en línea en http://arxiv.org/abs/cs/0010026v1

Albertrazzi, L. (1996). Material and formal ontology. en R. Poli, P. Simons (eds.), *Formal Ontology*. Kluwer, Dordrecht, (pp. 199-232).

Alcántara Plá, M. (2007). *Introducción al análisis del estructuras lingüísticas en el corpus. Aproximación semántica*. Madrid: Universidad Autónoma de Madrid.

Alcaraz Varó, E., Mateo Martínez, J., & Yus Ramos, F. (2006). *Las lenguas profesionales y académicas*. Barcelona: Ariel.

Alfonseca, E. (2008). Reconocimiento de entidades, resolución de correferencia y extracción de relaciones. En Felisa Verdejo (Ed.) *Curso de Tenologías Ligüísticas*.Publicaciones UNED (Universidad Nacional de Educación a Distancia) Colección Actas y Congresos.

Almela Pérez, R. (2002). *Morfología del español*. Murcia: DM.

Amardeilh, F. (2006). OntoPop or how to annotate documents and populate ontologies from texts. *Proceedings of the Workshop on Mastering the Gap: From Information Extraction to Semantic Representation* (ESWC'06), Budva, Montenegro.

Ananiadou, S., & McNaught, J. (2006). *Text mining for biology and biomedicine*: Artech House bioinformatics series.

Antoniou, G., & Van Harmelen, F. (2004). *A semantic web primer*: The MIT Press.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., .Eppig, J. T. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics, 25*(1), 25.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, Ll., Padró, M. (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 (LREC2006)*; ELRA; 2006.
Recurso on line http://www.lsi.upc.edu/~nlp/freeling

Aznar, E., Cros, A., & Quintana, L. (1991). *Coherencia textual y lectura*: Universidad de Barcelona, Instituto de Ciencias de la Educación.

Baclawski, K., & Niu, T. (2005). *Ontologies for bioinformatics*: MIT press. Cambridege.

Baeza-Yates, R., & Ribeiro, B. d. A. N. (1999). *Modern information retrieval*. Harlow, England: Addison-Wesley.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley FrameNet project. Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL-98), (pp. 86-90).

Barta, R., Feilmayr, C., Pröll, B., Grün, C., & Werthner, H. (2009). Covering the semantic space of tourism: An approach based on modularized ontologies. *Proceedings of the 1st Workshop on Context, Information and Ontologies*.

Baud, R. H., Ceusters, W., Ruch, P., Rassinoux, A. M., Lovis, C., & Geissbuhler, A. (2007). Reconciliation of ontology and terminology to cope with linguistics. *Medinfo. 12*(1), (pp. 796-801).

Beisswanger, E., Schulz, S., Stenzhorn, H., & Hahn, U. (2008). BioTop: An upper domain ontology for the life sciencesA description of its current structure, contents and interfaces to OBO ontologies. *Applied Ontology, 3*(4), 205-212.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American Magazine, 284*(5), (pp. 34-43).

Bhatia, V. K. (1993). *Analysing genre. language use in professional settings.* London: Longman.

Biomed. (2011). *Using BioMed central's open access full-text corpus for text mining research.* Recuperado en Septiembre, 2011, de http://www.biomedcentral.com/info/about/datamining/

Black, W. J., & Vasilakopoulos, A. (2002). Language independent named entity classification by modified transformation-based learning and by decision tree induction. *Proceedings of the 6th Conference on Natural Language Learning-Volume 20,* (pp. 1-4).

Blake, J. (2004). Bio-ontologies—fast and furious. *Nature Biotechnology, 22*(6), (pp. 773-774).

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2011). Relation adaptation: Learning to extract novel relations with minimum supervision. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence,* (pp. 2205-2210).

Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., & Cunningham, H. (2002). Shallow methods for named entity coreference resolution. *Chaines De Références Et Résolveurs d'anaphores, Workshop TALN,* Nancy, France.

Borrega, O., Taule, M., & Martı, M. A. (2007). What do we mean when we speak about named entities. *Proceedings of Corpus Linguistics*. Disponible en línea en http://dl.acm.org/citation.cfm?id=336644

Borst, W. N. (1997). Construction of engineering ontologies for knowledge sharing and reuse: *CTIT Ph.D-Thesis Series*.

Bosch Abarca, E., Giménez Moreno, R., & Montañés Brunet, E. (2005). Protocolos comunicativos en la promoción de servicios turísticos. *Quaderns De Filologia. Estudis Lingüístics,* (10), (pp. 205-224).

Brin, S. (1999). Extracting patterns and relations from the world wide web. *The World Wide Web and Databases,* (pp. 172-183).

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: An overview. *Ontology Learning from Text: Methods, Evaluation and Applications,* (pp. 3-12).

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., & Kriegel, H. (2008). Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics, 9*(1), 207.

Bustos Gisbert, J. M. (1996). *La construcción de textos en español*. Universidad de Salamanca.

Calsamiglia Blancafort, H. (1999). *Las cosas del decir. manual de análisis del discurso*. Barcelona: Ariel.

Calvi, M. T. (2006). Lengua y comunicación en el español de turismo. Madrid: Arco/Libro.

Cardoso, J. (2007). The semantic web vision: Where are we? *Intelligent Systems, 22*(5), (pp. 84-88).

Cardoso, J. (2005). E-tourism: Creating dynamic packages using semantic web processes. *W3C Workshop on Frameworks for Semantics in Web Services*. Disponible en línea en http://www.w3.org/2005/04/FSWS/Submissions/16/paper.html

Carreras, X., Marquez, L., & Padro, L. (2002). Named entity extraction using AdaBoost. *Proceedings of the 6th Conference on Natural Language Learning-Volume 20,* (pp. 1-4).

Castano, S., Ferrara, A., Montanelli, S., & Lorusso, D. (2008). Instance matching for ontology population. *In Proc. of the Sixteenth Italian Symposium on Advanced Database Systems (SEBD 2008)* (pp. 121-132).

Castells, P. (2003). La web semántica. Sistemas Interactivos y Colaborativos en la Web, *Ediciones de La Universidad De Castilla-La Mancha*, (pp. 195-212).

Celjuska, D., & Vargas-Vera, M. (2004). Ontosophie: A semi-automatic system for ontology population from text. *Proceedings International Conference on Natural Language Processing ICON, 4.*

Ceusters, W., Smith, B., & Flanagan, J. (2003). Ontology and medical terminology: Why description logics are not enough. *Proceedings of TEPR, (pp. 10* 14).

Chen, J., Ji, D., Tan, C. L., & Niu, Z. (2006). Relation extraction using label propagation based semi-supervised learning. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics,* (pp. 129-136).

Chou, W. C., Tsai, R. T. H., Su, Y. S., Ku, W., Sung, T. Y., & Hsu, W. L. (2006). A semi-automatic method for annotating a biomedical proposition bank. *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006,* (pp. 5-12).

Chun, H., Tsuruoka, Y., Kim, J., Shiba, R., Nagata, N., Hishiki, T., & Tsujii, J. (2006). Extraction of gene-disease relations from MedLine using domain dictionaries and machine learning. *Pac Symp Biocomput, 11*, (pp. 4–15).

Cimiano, P., Pivk, A., Schmidt-Thieme, L., & Staab, S. (2005). Learning taxonomic relations from heterogeneous evidence. En Buitelaar, P., Cimiano, P., & Magnini, B. (Eds.), *Ontology Learning from Text: Methods, Applications and Evaluation*: IOS Press. (pp. 59–73).

Cimiano, P., Handschuh, S., & Staab, S. (2004). Towards the self-annotating web. *Proceedings of the 13th International Conference on World Wide Web,* New York, NY, USA. (pp. 462-471).

Codina, L., & Rovira, C. (2006). La web semántica. En J. Tramullas (Ed.), *Tendencias en documentación digital*: Trea.

Cohen, K. B., & Hunter, L. (2006). A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics, 7*, S5.

Coseriu, E. (1987). Gramática, semántica, universales. *Estudios de lingüística funcional*. Madrid: Gredos.

Couturat, L. (1903). *Opuscules et fragments inédits de Leibniz*. París, Francia.

Culotta, A., & Sorensen, J. (2004). Dependency tree kernels for relation extraction. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics,* (ACL-04), Barcelona, Spain, 2004, (pp. 423–429).

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities, 36*(2), 223-254.

Curse, D. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Danger, R., & Berlanga, R. (2009). Generating complex ontology instances from documents. *Journal of Algorithms, 64*(1), (pp. 16-30).

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? *AI Magazine, 14*(1), 17.

de Boer, V., van Someren, M., & Wielinga, B. J. (2007). Relation instantiation for ontology population using the web. *Ki 2006: Advances in Artificial Intelligence, Proceedings, 4314*, (pp. 202-213).

Delahousse, J. (2003). *Semantic web use case: An application for sustainable tourism development*. Paris: Mondeca. Recuperado en Junio de 2011 de http://www.mondeca.com/sw-tourism-ontoweb-sig4-V2.pdf

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The automatic content extraction (ACE) program–tasks, data, and evaluation. *Proceedings of LREC, 4* (pp. 837–840).

Dolbey, A., Ellsworth, M., & Scheffczyk, J. (2006). BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. *Biomedical Ontology in Action KR-MED 2006 Proceedings,* (pp. 87-94).

Doran, C., Egedi, D., Hockey, B. A., Srinivas, B., & Zaidel, M. (1994). XTAG system: A wide coverage grammar for english. *Proceedings of the 15th Conference on Computational Linguistics*. (2), (pp. 922-928).

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence, 165*(1), (pp. 91-134).

Faure, D., & Poibeau, T. (2000). First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. *Ontology Learning, ECAI-2000 Workshop,* , (pp. 7–12).

Fellbaum, C. (1998). WordNet: An electronic lexical database: The MIT press.

Fernández Toledo, P. (1999). *Conocimiento previo, esquemas de género y comprensión lectora del inglés como lengua extranjera*. Tesis Doctoral. Universidad De Murcia.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences, 280*(1), (pp. 20-32).

Firth, J. R. (1961). *Papers in linguistics, 1934-1951* Oxford University Press.

Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). Named entity recognition through classifier combination. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (4),* (pp. 168-171).

Friedman, C., Kra, P., & Rzhetsky, A. (2002). Two biomedical sublanguages: A description based on the theories of Zellig Harris. *Journal of Biomedical Informatics, 35*(4), (pp. 222-235).

Fukuda, K., Tamura, A., Tsunoda, T., & Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. *Pacific Symposium on Biocomputing,* (pp. 707-718).

Gacitua, R., Sawyer, P., Piao, S., & Rayson, P. (2007). Ontology acquisition process: A framework for experimenting with different techniques. En *Proceedings of the UK e-Science all Hands,* Nottingham, UK, (pp. 561-567).

Giuliano, C., & Gliozzo, A. (2008). Instance based lexical entailment for ontology population. En *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), (pp.265-272).

Gómez Díaz, R. (2005). *La lematización en español: Una aplicación para la recuperación de información*: Trea.

Gómez-Pérez, A., & Manzano-Macho, D. (2003). OntoWeb D.1.5. A survey of ontology learning methods and techniques. *OntoWeb Deliverable, 1* Disponible en Línea en www.sti-innsbruck.at/fileadmin/documents/deliverables/Ontoweb/D1.5.pdf

Gómez-Pérez, A., Moreno, A., Pazos Sierra, J., & Sierra-Alonso, A. (2000). Knowledge maps: An essential technique for conceptualization. *Data & Knowledge Engineering. Special Issue on Conceptual Modelling, 33*, 1 (pp. 69-190).

Grishman, R. (2006). Named entity extraction. En Keith Brown (Ed.), *Encyclopedia of language & linguistics,* (pp. 434-436). Oxford: Elsevier.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), (pp. 199-220).

Gruninger, M., & Fox, M. S. (1995). Methodology for the design and evaluation of ontologies. *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI. 95*.Disponible en Línea en http://ibict.phlnet.com.br/anexos/grninger95methodology.pdf

Guarino, N. (1998). Formal ontology and information systems. En *Proceedings of FOIS'98*, Trento, Italy. Amsterdam: IOS Press, (pp. 3-15)

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human and Computer Studies, 43(5/6)*, (pp. 625-640).

Gutiérrez Losada, I. (2010). Ontologías turísticas geográficas: Creación de una ontología sobre rutas turísticas (a pie o en bicicleta) por espacios naturales.

Hahn, U., Schulz, S., & Romacker, M. (1999). Part-whole reasoning: A case study in medical ontology engineering. *Intelligent Systems and their Applications, IEEE, 14*(5), (pp. 59-67).

Hasegawa, T., Sekine, S., & Grishman, R. (2004). Discovering relations among named entities from large corpora. En *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics,* (pp. 415-424).

He, T., Xu, C., Li, J., & Zhao, J. (2006). Named entity relation extraction method based on seed self-expansion. *Computer Engineering, 32*(21), (pp. 183-193).

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th Conference on Computational Linguistics, 2*, (pp. 539-545).

Hernández Terrés, J,M. & Escavi Zamora, R. (1999). *Lingüística general*. Universidad de Murcia: DM.

Huynh, D., Mazzocchi, S., & Karger, D. (2005). Piggy bank: Experience the semantic web inside your web browser. *The Semantic Web–ISWC 2005,* (pp. 413-430).

IEEE. (2011). *Suggested upper merged ontology (SUMO)*. Recuperado en Septiembre de 2011 de http://www.ontologyportal.org/index.html

Isozaki, H., & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. *Proceedings of the 19th International Conference on Computational Linguistics* (1,) (pp.1-7).

Izquierdo, R., Ferrández, O., Ferrández, S., Tomás, D., Vicedo, J. L., Martínez, P., & Suárez, A. (2007). QALL-ME: Question answering learning technologies in a multiLingual and multiModal envinroment. *Procesamiento Del Lenguaje Natural, 38*, (pp. 43-47).

Jackson, P., & Schilder, F. (2006). Natural language processing: Overview. En Keith Brown (Ed.), *Encyclopedia of language & linguistics* (pp. 503-518). Oxford: Elsevier.

Jakkilinki, R., Georgievski, M., & Sharda, N. (2007). Connecting destinations with ontology-based e-tourism planner. *14th Annual Conference of the International Federation for IT&Travel and Tourism*.

Jannach, D., Shchekotykhin, K., & Friedrich, G. (2009). Automated ontology instantiation from tabular web sources.The AllRight system. *Web Semantics: Science, Services and Agents on the World Wide Web, 7*(3), (pp. 136-153).

Jurafsky, D. & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech Recognition* (2nd Edition): Pearson Prentice Hall

Kanellopoulos, D. N., & Panagopoulos, A. A. (2008). Exploiting tourism destinations' knowledge in an RDF-based P2P network. *Journal of Network and Computer Applications, 31*, (pp. 179-200).

Kant, I. (2001). *Lectures on metaphysics - part III metaphysik L2*: *Cambridge Univerity Press,*

Kim, J. D., Ohta, T., Teteisi, Y., & Tsujii, J. (2006a). *GENIA Ontology*. Technical report, Tsujiilab, University of Tokyo. Disponible en Línea en

http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Ontology

Kim, J. D., Ohta, T., Teteisi, Y., & Tsujii, J. (2006b). GENIA corpus manual. Journal of Biomedical Informatics. 39(3) (pp. 333-349)

Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications,* 70-75.

Kim, J. -., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus. A semantically annotated corpus for bio-textmining. *Bioinformatics, 19*(1), (pp. 180-182).

Kim, J., Ohta, T., & Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics, 9*(1), 10.

Kingsbury, P., & Palmer, M. (2002). From treebank to propbank. *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC-2002), (pp. 1989–1993).

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of english verbs. *Language Resources and Evaluation, 42*(1), (pp. 21-40).

Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon.* PhD. Thesis. University of Pennsylvania. Philadelphia.

Klein, D., Smarr, J., Nguyen, H., & Manning, C. D. (2003). Named entity recognition with character-level models. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (4), (pp. 180-183).

Knublauch, H. (2004). Travel.owl. Recurso on-line. Recuperado en Marzo de 2011 de : *Http://protege.Stanford.Edu*.

Korhonen, A., krymolowski, Y., & Collier, N. (2006). Automatic classification of verbs in biomedical texts. *Proceedings of ACL-COLING 2006.* Sydeney. Australia.

Kozareva, Z., Ferrández, O., Montoyo, A., Munoz, R., & Suárez, A. (2005). Combining data-driven systems for improving named entity recognition. *Natural Language Processing and Information Systems,* (pp. 169-194).

Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., White, P. (2004). Integrated annotation for biomedical information extraction. *Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL),* (pp. 61-68).

Lassila, O., & Swick, R. R.Resource description framework (RDF) model and syntax. World Wide Web Consortium. Recuperado en Junio de 2011 de *Http://www.w3.org/TR/WD-Rdf-Syntax*.

Lavelli, A., Califf, M., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., Ireson, N. (2008). Evaluation of machine learning-based information extraction algorithms: Criticisms and recommendations. *Language Resources and Evaluation, 42,* (pp. 361-393).

Lee, K., Hwang, Y., Kim, S., & Rim, H. (2004). Biomedical named entity recognition using two-phase model based on SVMs. *Journal of Biomedical Informatics, 37*(6), 436-447.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* Chicago.

Lin, D., & Pantel, P. (2001). Induction of semantic classes from natural language text. En *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 317-322).

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The unified medical language system. *Methods of Information in Medicine, 32*(4), (pp. 281-291).

Llisterri, J. (2003). Lingüística y tecnologías del lenguaje. *Panorámica de Estudios Lingüísticos, 2*, (pp. 9-71).

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems, 16 (2)*, (pp. 72-79).

Magnini, B., Pianta, E., Popescu, O., & Speranza, M. (2006). Ontology population from textual mentions: Task definition and benchmark. *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, (pp. 26-32).

Mandel, M. (2006). Integrated annotation of biomedical text: Creating the PennBioIE corpus. *Text Mining Ontologies and Natural Language Processing in Biomedicine, Manchester, UK.*

Márquez, L. (2008). Etiquetado de roles semánticos. En Felisa Verdejo (Ed.), *Acceso y visibiliad de la informción multilingüe en la red: El rol de la semántica:* Publicaciones UNED (Universidad Nacional de Educación a Distancia), Colección Actas y Congresos.

Mayfield, J., McNamee, P., & Piatko, C. (2003). Named entity recognition using hundreds of thousands of features. *Proceedings of the Seventh Conference on Natural Language Learning en HLT-NAACL 2003* (4), (pp. 184-187).

Maynard, D., Bontcheva, K., & Cunningham, H. (2003). Towards a semantic extraction of named entities. *Recent Advances in Natural Language Processing, Bulgaria.*

Maynard, D., Funk, A., & Peters, W. (2009). SPRAT: A tool for automatic semantic pattern-based ontology population. *International Conference for Digital Libraries and the Semantic Web,* Trento, Italy.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy markov models for information extraction and segmentation. *Proceedings of the Seventeenth International Conference on Machine Learning,* (pp. 591-598).

McDowell, L. K., & Cafarella, M. (2008). Ontology-driven, unsupervised instance population. *Web Semantics: Science, Services and Agents on the World Wide Web, 6 (3)*, (pp. 218-236).

McGuinness, D. L., & van Harmelen, F. *OWL web ontology language overview*. Recuperado en Junio de 2011 de http://www.w3.org/TR/owl-features/

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM, 38*(11), (pp. 39-41).

Missikof, M., Werthner, H., Höpken, W., Dell'Erba, M., Fodor, O., Formica, A., & Taglino, F. (2003). Harmonise towards interoperability in the tourism domain. *10th International Conference on Information Technologies in Tourism (ENTER 2003),* (pp. 29-31).

Mizoguchi, R., Vanwelkenhuysen, J., Ikeda, M. (1995). *Task Ontology for Reuse of Problem Solving Knowledge. Towards Very Large Knowledge Bases.* KnowledgeBuilding and Knowledge Sharing, (pp. 46-59).

Moreda Pozo, P. (2008). *Los roles semánticos en la tecnología del lenguaje humano: Anotación y aplicación.* Tesis doctoral. Universidad de Alicante.

Moreiro Gonzáles, J. A. (2007). La representación de los contenidos digitales: De los tesauros automáticos a las folksonomías. *Actas Del VI Workshop CALSI.*

Moreno Ortiz, A. (1997). Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. *Estudios De Lingüística Española,* (9).

Murata, M., Shirado, T., Torisawa, K., Iwatate, M., Ichii, K., Ma, Q. & Kanamaru, T. (2010) Extraction and Visualization of Numerical and Named Entity Information from a very large number of Documents Using Natural Language Processing, *International Journal of Innovative Computing; Information and Control (IJICIC)*; *(6)*; 3(B); (pp. 1549-1568)

Nadeau, D., Turney, P., & Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *19th Canadian Conference on Artificial Intelligence.*

Nakamura-Delloye, Y. (2011). Named entity extraction for ontology enrichment. Disponible en Línea en http://hal.inria.fr/hal-00606077_v1/

Navigli, R., & Velardi, P. (2006). Enriching a formal ontology with a thesaurus: An application in the cultural heritage domain. *COLING• ACL 2006,* Sydney.

Navigli, R., Velardi, P., & Ruiz-Martınez, J. M. (2010). An annotated dataset for extracting definitions and hypernyms from the web. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010),* Valletta, Malta.

NCBI, National Center for Biotechnology Information, US. (2011). *PubMed.* Recuperado en Septiembre de 2011 de http://www.ncbi.nlm.nih.gov/pubmed/?ncbi_mmode=std

NLM, National Library of Medicine. (2008). *UMLS knowledge sources.* United States. Recurso on line https://uts.nlm.nih.gov/

NLM, National Library of Medicine US. (2011). *MeSH, medical subject headings.* Recuperado en Junio de 2011 de http://www.nlm.nih.gov/mesh/MBrowser.html

Ono, T., Hishigaki, H., Tanigami, A., & Takagi, T. (2001). Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics, 17*(2).

Park, H. M., Lee, Y. L., Noh, B. N & Lee, H. H. (2009) Ontology-Based Generic Event Model for Ubiquitous Environment; *International Journal of Innovative Computing; Information and Control (IJICIC);* 5 (11)(B); (pp. 4317-4326).

Pasquier, C. (2008). Biological data integration using semantic web technologies. *Biochimie, 90*(4), (pp. 584-594).

Pastor Sánchez, J. (2009). *Diseño De Un Sistema Colaborativo Para La Creación y Gestión De Tesauros En Internet Basado En SKOS*. Tesis doctoral. Universidad De Murcia.

Pedraza-Jiménez, R., Codina, L., & Rovira, C. (2007). Web semántica y ontologías en el procesamiento de la información documental. *El Profesional De La Información, 16*(6), (pp. 569-578).

Persidis, A. (1999). Bioinformatics. *Nature Biotechnology, 17*, (pp. 828-830).

Petasis, G., Karkaletsis, V., & Paliouras, G. (2007). *Ontology population and enrichment: State of the art*. (deliverable público No. D4.3:). BOEMIE Bootstrapping Ontology Evolution with Multimedia Information Extraction.

Poli, R. (2000). Levels of reality. *Cognitive Analysis Dependence and Dynamic Categories*. BISCA 2000: Bolzano international schools.

Poli, R. (2001). *ALWIS: Ontology for knowledge engineers*. Tesis doctoral. Leiden-Utrecht Research Institute of Philosophy.

Poli, R. (2003). Descriptive, formal and fomalized ontologies. *Contributions to Phenomenology, 48*, (pp. 183-210).

Prantner, K. (2004). OnTour: The ontology. *Deri Insbruck,* Recurso on line. Recuperado en Marzo de 2011 de http://e-tourism.deri.at/ont/docu2004/OnTour%20-%20The%20Ontology.pdf

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., & Salakoski, T. (2007). BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics, 8*(1).

Rak, R., Kurgan, L., & Reformat, M. (2007). xGENIA: A comprehensive OWL ontology based on the GENIA corpus. *Bioinformation, 1*(9).

Ramshaw, L. A., & Weischedel, R. M. (2005). Information extraction. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference,* (5).

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning,* (pp. 147-155).

Ravichandran, D., & Hovy, E. (2002). Learning surface text patterns for a question answering system. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* (pp. 41-47).

Ray, S., & Craven, M. (2001). Representing sentence structure in hidden markov models for information extraction. *International Joint Conference on Artificial Intelligence, 17*(1) (pp.1273-1279).

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. *Proceedings of the National Conference on Artificial Intelligence,* (pp. 474-479).

Rosario, B., & Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

Ruiz-Martínez, J. M., Valencia-García, R., Martínez-Béjar, R., & Hoffmann, A. (2010). Populating biomedical ontologies from natural language texts. *Proceedings of the International Conference Knowledge Engineering and Ontology Development (KEOD),* Valencia, España.

Rung-Ching, C., Cho-Tsan, B., & Ming-Yung, T. (2010) Web Pages Cluster Based On The Relations Of Mapping Keywords To Ontology Concept Hierarchy; *International Journal of Innovative Computing; Information and Control (IJICIC)*; 6 (6)(B); (pp. 2749–2760)

Sabou, M., Wroe, C., Goble, C., & Stuckenschmidt, H. (2005). Learning domain ontologies for semantic web service descriptions. *Web Semantics: Science, Services and Agents on the World Wide Web, 3*(4), 340-365.

Sager, J. C., & Nkwenti-Azeh, B. (1990). *A practical course in terminology processing*.

Sahay, S., Mukherjea, S., Agichtein, E., Garcia, E. V., Navathe, S. B., & Ram, A. (2008). Discovering semantic biomedical relations utilizing the web. *ACM Transactions on Knowledge Discovery from Data (TKDD), 2*(1), (pp. 1-15).

Sánchez-Jiménez, R., & Gil-Urdiciain, B. (2007). Lenguajes documentales y ontologías. *El Profesional De La Información, 16*, (pp. 551-560).

Sánchez, D., & Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data & Knowledge Engineering, 64*(3), (pp. 600-623).

Santamaría, C., Gonzalo, J., & Verdejo, F. (2003). Automatic association of web directories with word senses. *Computational Linguistics, 29*(3), (pp. 485-502).

Saquete, E., Ferrández, O., Ferrández, S., Martínez-Barco, P., & Muñoz, R. (2008). Combining automatic acquisition of knowledge with machine learning approaches for multilingual temporal recognition and normalization. *Information Sciences, 178*(17), (pp. 3319-3332).

Schwartz, A. S., & Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing, , 8* (pp. 451-462).

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. *Proceedings of the Language Resources and Evaluation Conference (LREC),* (pp. 1977–1980).

Sekine, S., Sudo, K., & Nobata, C. (2002). Extended named entity hierarchy. *Proceedings of the LREC-2002 Conference,* (pp. 1818–1824).

Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (NLPBA), 1, (pp. 104-107).

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

Sharma, A., Swaminathan, R., & Yang, H. (2010). A verb-centric approach for relationship extraction in biomedical text. *The Fourth IEEE International Conference on Semantic Computing (ICSC2010)*.

Shen, D., Zhang, J., Zhou, G., Su, J., & Tan, C. L. (2003). Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, 13*, (pp. 49-56).

Shinyama, Y., & Sekine, S. (2006). Preemptive information extraction using unrestricted relation discovery. Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, (pp. 304-311).

Sinclair, J. (1996). EAGLES:Preliminary recommendations on text typology. *Document EAG-TCWG-CTYP/P*.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics, 5*(2), (pp. 51-53).

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology, 6*(5), R46.

Sowa, J. F. (2000). Knowledge representation: Logical, philosophical, and computational foundations Pacific Grove : Brooks/Cole, c2000.

Spasic, I. (2004). A Machine Learning Approach to Term Classification. Tesis doctoral. University of Salford.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering, 25*(1-2), 161-197.

Stumme, G., Hotho, A., & Berendt, B. (2006). Semantic web mining: State of the art and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web, 4*(2), (pp. 124-143).

Tanev, H., & Magnini, B. (2006). Weakly supervised approaches for ontology population. *Proceedings of EACL-2006, Trento,* (pp. 3-7).

Tateisi, Y., & Tsujii, J. (2004). Part-of-speech annotation of biology research abstracts. *Proceedings of LREC04*.

Taulé, M., Borrega, O., & Martí, M. A. (2011). AnCora-net: Integración multilingüe de recursos lingüísticos semánticos. *Procesamiento De Lenguaje Natural, 47*(0), (pp. 153-160).

Thomsen, E.H., & Madsen, B. N. (2009). International Standard ISO 704: 2009: Terminology Work-Principles and Methods.

Toral, A., & Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. *NEW TEXT Wikis and Blogs and Other Dynamic Text Sources,* (pp. 56- 65).

Tsai, R., Dai, H. J., Huang, C. H., & Hsu, W. L. (2008). Semi-automatic conversion of BioProp semantic annotation to PASBio annotation. *BMC Bioinformatics, 9*(Suppl 12), S18.

Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. *Advances in Informatics,* , 382-392.

Tsai, R., Chou, W., Su, Y., Lin, Y., Sung, C., Dai, H., Hsu, W. (2007). BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model

with automatically generated template features. *BMC Bioinformatics, 8*(1), (pp. 325-335)

Uschold, M., & King, M. (1995). Towards a methodology for building ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing.*

Valencia-García, R., Nieves, D. C., Vicente, P. J. V., Breis, J. T. F., Martínez-Béjar, R., & García Sánchez, F. (2004). An approach for ontology building from text supported by NLP techniques. *Current Topics in Artificial Intelligence: 10th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2003, and 5th Conference on Technology Transfer, TTIA 2003, San Sebastian, Spain.*

van Heijst, G., Schreiber, A. T., & Wielinga, B. J. (1997). Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies, 46 (2/3)*(2/3), (pp. 183-292).

Verdejo, F. (2008). Introducción y metodología de trabajo. En Felisa Verdejo (Ed.), *Acceso y visibiliad de la informción multilingüe en la red: El rol de la semántica:* Publicaciones UNED (Universidad Nacional de Educación a Distancia), Colección Actas y Congresos.

Wächter, T., Alexopoulou, D., Dietze, H., Hakenberg, J., & Schroeder, M. (2008). Searching biomedical literature with anatomy ontologies. *Anatomy Ontologies for Bioinformatics,* (pp. 177-194).

Wattarujeekrit, T., Shah, P., & Collier, N. (2004). PASBio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics, 5*(155).

Welty, C., & Guarino, N. (2001). Supporting ontological analysis of taxonomic relationships. *Data & Knowledge Engineering, 39*(1), (pp. 51-74).

WTO. (2001). *Thesaurus on tourism and leisure activities of the world tourism organization.* Organización Mundial Turismo.

Yang, N. Sun, T. L. Sun, X. Y. Cao and X. J. Zheng; (2009) The Application of Latent Semantic Indexing and Ontology in Text Classification; *International Journal of*

Innovative Computing; Information and Control (IJICIC); 5 (12A), (pp. 4491-4499)

Zablith, F., d'Aquin, M., Sabou, M., & Motta, E. (2010). Using ontological contexts to assess the relevance of statements in ontology evolution. *Knowledge Engineering and Management by the Masses,* (pp. 226-240).

Zelenko, D., Aone, C., & Richardella, A. (2003). Kernel methods for relation extraction. *The Journal of Machine Learning Research, 3*, (pp. 1083-1106).

Zhang, Y., Huang, H., Yang, D. & Zhang, H. (2010). A Hierarchical and Chord-Based Semantic Service Discovery System in the Universal Network; *International Journal of Innovative Computing; Information and Control (IJICIC)*; 5 11(A), (pp. 3745-3753).

Zhou, G., Su, J., Zhang, J., & Zhang, M. (2005). Exploring various knowledge in relation extraction. En ACL-5 (pp. 427–434).

Zhou, L. (2007). Ontology learning: State of the art and open issues. *Information Technology and Management, 8*(3), (pp. 241-252).

Zhu, Z., Liu, P., Zhao, L., & Lv, T. (2010). Research of feature weights adjustment based on semantic paragraphs matching. *ICIC Express Letters, 4*(2), (pp. 559-564).