PROYECTO FIN DE CARRERA

ANÁLISIS Y EVALUACIÓN DE LA HERRAMIENTA METAMAP PARA EL PROCESAMIENTO DE TEXTOS BIOMÉDICOS.

Autor: Diana García-Miguel López

Tutor: Isabel Segura Bedmar

Leganés, Septiembre de 2010

Título: Análisis y Evaluación de la herramienta MetaMap para el procesamiento de textos biomédicos

Autor: Diana García-Miguel López

Director: Isabel Segura Bedmar

EL TRIBUNAL

Presidente: _____

Vocal: _____

Secretario: _____

Realizado el acto de defensa y lectura del Proyecto Fin de Carrera el día __ de _____ de 20__ en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO PRESIDENTE

A las tres personas con las que he compartido, comparto y compartiré toda mi vida, mis padres y mi hermana, Mary, Manolo y Beatriz.

A mis padres, porque sin duda, este y cualquier otro proyecto que haya realizado o realice a lo largo de mi vida será por a ellos. Gracias por haberme motivado siempre a seguir estudiando, desde el primer que me llevaron al colegio y me negué a entrar en clase enganchándome a un radiador hasta hoy y durante los años que todavía me quedan de estudiante.

A mi hermana Beatriz, porque simplemente es la mejor y siempre está ahí cuando la necesito, hasta para leerse tres versiones distintas de este proyecto...

A Agus, porque si hoy estoy escribiendo estas líneas, él no tiene la menor culpa, gracias por haber confiado en mí siempre y por haberme ayudado en todo. Gracias por estar junto a mí, dentro y fuera de la universidad (O.N.C).

A Isabel, mi tutora y para mí la mejor, gracias por tu comprensión y tu ayuda en todo momento y gracias por haber confiado en mí y en la posibilidad de hacer este proyecto en unos meses.

“Lo que hace que un sueño sea imposible, es el miedo a fracasar” , PAULO COELHO

Resumen

Evaluación de la nueva versión de MetaMap, **MetaMap2009v2**, herramienta que lleva a cabo un análisis de textos biomédicos y que presenta un elevado grado de configurabilidad que será presentando a lo largo de esta memoria en función de diversos ejemplos, obtenidos tras procesar una selección amplia de documentos médicos extraídos de la base de datos biomédica MedLine2010. Se va a realizar un estudio de las innovaciones y mejoras introducidas en esta nueva versión, centrándonos en su evolución desde las implementaciones iniciales.

palabras clave: MetaMap2009v2, MedLine2010, procesamiento del lenguaje natural, Metatesauro UMLS, mapeo.

Índice general

Índice de tablas

Índice de ilustraciones

Ilustración 69. Salida XML obtenida tras ejecutar la opción "–% format" en MetaMap2009v2 para la evaluación de un documento con n citas en su contenido.

Índice de ecuaciones

Capítulo 1
Introducción y objetivos

1.1. INTRODUCCIÓN

Gracias a la constante evolución de los ordenadores desde finales de los años 40, se ha logrado incrementar la capacidad de almacenar y crear información en formato digital sobre cualquier campo de conocimiento conocido, como por ejemplo la Medicina. Es este crecimiento lo que ha hecho imprescindible la necesidad de crear y desarrollar todo tipo de sistemas informáticos capaces de comprender el lenguaje natural humano.

Con motivo del estudio de este tipo de lenguaje, nace en la década de los 60 una nueva disciplina, el "Procesamiento del Lenguaje Natural" (NLP), una rama dependiente de la Inteligencia Artificial (9), cuyo objeto es el desarrollo de sistemas que permitan la comunicación entre hombre y máquinas, como si de una conversación entre dos personas se tratara (15).

El objetivo principal del Procesamiento del Lenguaje Natural es resolver los problemas derivados de la generación y comprensión del lenguaje humano, puesto que

este tipo de lenguaje posee una serie de propiedades que disminuyen la efectividad de los sistemas usados para su procesamiento. Algunas de estas propiedades son, por un lado la “variación lingüística”, referida a la aparición de distintas expresiones para exponer la misma idea. Y por otro lado la “ambigüedad lingüística”, producida cuando una misma palabra o frase posee más de un significado posible (9).

Las técnicas de Procesamiento del Lenguaje Natural son una herramienta fundamental en la Recuperación de Información, usadas tanto para describir el contenido de los documentos como para llevar a cabo una representación de la consulta formulada por el usuario.

Estas técnicas se promueven a través de distintos tipos de análisis que ocupan varios niveles. Por un lado, aparece el análisis morfológico, en el que se realiza un examen de cada una de las palabras del documento, de las que se extraen toda su información gramatical (raíces, sufijos, prefijos y desinencias), además de la clase gramatical a la que pertenecen. Y por otro, el análisis sintáctico cuyo objetivo es detectar la estructura sintáctica de una oración mediante una gramática.

Como vemos en la figura existen otros niveles que completan el análisis realizado por las técnicas de PLN. En primer lugar nos encontramos con el nivel fonológico cuyo análisis se centra en el estudio de la forma de las palabras y los fonemas, este análisis requiere información de niveles superiores, al menos del morfológico y sintáctico. Otro de los niveles que podemos encontrar, es el nivel semántico que centra su estudio en el significado literal de la frase. Finalmente, aparece el nivel prágmatico y el de integración del discurso, el primero estudia el significado real de la frase y el segundo el significado en función del contexto.

Una de las técnicas más utilizadas para aumentar los resultados de los sistemas PLN, consiste en la inclusión de recursos externos que permitan obtener una información de mayor calidad, como ocurre en el caso del dominio biomédico, en el que se desarrollan cada vez más sistema de información que hacen uso de recursos externos, como las ontologías y herramientas para el análisis de los textos.

Los avances en biomedicina junto con el incesante desarrollo de la informática, han provocado un crecimiento exponencial en la creación de documentos científicos, por lo que se hace totalmente necesario el desarrollo de sistemas que permitan el acceso a este tipo de información de una manera más fácil. Para facilitar la creación de estos sistemas de "Procesamiento del Lenguaje Natural" en el dominio biomédico, el Sistema de Lenguaje Médico Unificado (UMLS) fue desarrollado siendo una de las bases de datos del conocimiento médico más completas, cuyos recursos principales son: el Metatesauro, la red semántica y el SPECIALIST lexicon.

Una de las herramientas de análisis para textos biomédicos que destaca gracias a su rigor lingüístico y su dependencia de fuentes de conocimiento como es el léxico SPECIALIST, es MetaMap, un programa de elevada configurabilidad que realiza un mapeo de textos biomédicos a conceptos del Metatesauro UMLS (1) y (2).

Este programa fue desarrollado para facilitar la identificación de conceptos referidos en un texto, puesto que esta detección de información es imprescindible para el desarrollo de numerosas aplicaciones (2) como motores de búsqueda, sistemas de búsqueda de puestas, sistemas de generación de resúmenes o sistemas de extracción de información entre otros.

MetaMap fue desarrollado por el Dr. Alan (Lan) Aronson en la Biblioteca Nacional de Medicina y cuyo objetivo era tanto el mapeo de textos biomédicos a conceptos del Metatesauro UMLS, como la localización de este tipo de conceptos referidos en un texto cualquiera. Esta herramienta ha sido usada desde 1994 y actualmente se encuentran disponibles en la Web las últimas versiones de esta, las cuales han experimentado una importante evolución a lo largo de los años, centrada en diversoso aspectos, tanto añadir algoritmos que aportan nuevas funcionalidades, como la modificación de ciertos aspectos que puedan facilitar su uso.

1.2. OBJETIVOS

Dado la elevada funcionalidad y relevancia de MetaMap en el análisis de textos biomédicos numerosos trabajos la han empleado en su desarrollo, algunos de los títulos son: “Natural language processing to extract medical problems from electronic clinical documents” (20), “A comparison of machine learning techniques for detection of drug target articles” (21) y “Drug name recognition and classification in biomedical texts: A case study outling approaches underpinning automated systemes” (22)

El objetivo de nuestro proyecto es la evaluación y análisis de la nueva versión, MetaMap2009v2, además de realizar una comparativa de esta con versiones anteriores, evaluando para tal fin las mejoras y las nuevas funcionalidades introducidas. Para lograr este objetivo, se ha procesado una selección de documentos de la colección de textos de la base de datos MedLine2010.

Para lograr alcanzar el objetivo principal ya descrito, se proponen los siguientes objetivos, cuyo fin será proporcionar un entorno adecuado de trabajo para la ejecución de la herramienta y su posterior evaluación. En concreto se van a proponer los siguientes objetivos específicos:

- Instalación de la nueva versión de MetaMap, MetaMap2009v2.
- Descarga de la colección de documentos MedLine2010.
- Estudio teórico de la base de datos MedLine2010 y de MetaMap, realizando tanto un estudio del algoritmo interno aplicado para el procesado, como un aprendizaje de su instalación y ejecución.
- Generación de una aplicación en Java que permita el procesado automático con MetaMap de toda la selección de documentos de MedLine.
- Procesado con la herramienta MetaMap de la colección de MedLine 2010 y generación de la salida en formato XML, funcionalidad ya incluida en la nueva versión de la herramienta MetaMap2009v2.

- Procesado con MetaMap2009v2 seleccionando algunas opciones específicas de procesado interesantes de estudiar.

- Estudio de las nuevas funcionalidades y mejoras incluidas en esta versión.

- Comparativa con versiones anteriores de MetaMap.

1.3. ESTRUCTURA DEL DOCUMENTO.

La memoria de este proyecto está organizada en cinco capítulos y un anexo cuya estructura se muestra a continuación:

- Capítulo 2. En primer lugar realizaremos un análisis sobre la colección MedLine 2010. A continuación se expone un estudio completo sobre el procesado de los textos con la nueva versión de MetaMap. Para lograr un estudio completo, en una primera parte se lleva a cabo un estudio del algoritmo de análisis del programa, así como cada una de las partes de la información que nos proporciona su salida.

Tras este estudio teórico del algoritmo, se trata tanto la instalación del programa, como su ejecución y finalmente se ha realizado un análisis de cada una de las opciones proporcionadas con esta nueva versión.

- Capítulo 3. Es este capítulo se realiza un estudio de las mejoras introducidas en funcionalidades implantadas en MetaMap2009v2. Se va a llevar a cabo un especial seguimiento de ciertas impletmentaciones tales como el tratamiento de negaciones o el algoritmo de “Breaking-Sentence” y de la salida XML incorporada por la nueva versión.

- Capítulo 4. La evaluación de cualquier aplicación necesita disponer de una selección de “objetos”, en este caso una colección de citas médicas, lo más amplia posible, sobre el que ser probado y evaluado. En este

capítulo de la memoria se muestran la estadísticas sobre la colección de documentos procesada, esto es: número de documentos, número de oraciones, de palabras, número de negaciones localizadas, tipos semánticos, tipos sintácticos, etc.

- Capítulo 5. Para finalizar este proyecto se expondrán nuestras conclusiones generales y las líneas posibles de trabajo futuro.

- Anexo 1. Presupuesto total del proyecto.

Capítulo 2
MedLine2010 y MetaMap2009v2

Para el correcto análisis de la herramienta MetaMap es necesario disponer de una colección de documentos completa, que permita la evaluación de todas las funcionalidades que la herramienta nos puede ofrecer. Por este motivo, para el desarrollo de nuestro proyecto se ha utilizado una amplia selección de documentos biomédicos procedentes de una base de datos de biomédica, MedLine. Las características generales de esta base se estudian en la siguiente sección.

A continuación, se presenta un estudio teórico de MetaMap, estudiando en primer lugar el comportamiento del algoritmo que modela esta herramienta y que está basado en cinco pasos fundamentales:

- División del texto en oraciones simples.
- Generación de variantes para cada término de la oración.
- Formación del conjunto de candidatos semánticos para cada variante.
- Evaluación del nivel del mapeo de cada candidato.
- Evaluación del nivel de calidad de la combinación de candidatos final.

Tras el estudio del proceso interno del programa, analizaremos de forma práctica el nuevo software proporcionado en esta nueva versión, MetaMap2009v2, atendiendo a su proceso de instalación y ejecución, así como un análisis de la estructura en la que MetaMap2009v2 devuelve el documento procesado. Finalmente, hemos realizado un estudio práctico de cada una de las opciones de procesado y salida disponibles en esta nueva versión.

2.1. MEDLINE 2010

MedLine es una base de datos creada por la Biblioteca Nacional de Medicina (NLM), que abarca distintos campos científicos: medicina, oncología, enfermería, odontología, veterinaria, salud pública y ciencias básicas. En la versión de la base de datos correspondiente al año 2010 existen 17.969.577 referencias bibliográficas de artículos de revistas publicadas en Estados Unidos y en otros 70 países más (12). Esta base de datos se encuentra dividida en las siguientes subbases: *AIDS, Bioethics, Complementary Medicine, Core Clinical Journals, Dental Journals, Nursing Journals y PubMed Central.*(8)

Para disponer de los documentos biomédicos que residen en esta base datos, se ha accedido al directorio que los contiene a través del enlace ftp://ftp.nlm.nih.gov/nlmdata/sample/medline/, cuyo contenido está formado por ocho ficheros en formato XML ordenados alfabéticamente, los cuales constan de miles de citas cada uno y son actualizados cada año. Para descargarnos estos archivos basta con seleccionar cada uno de ellos y a continuación descomprimirlos y almacenarlos en nuestro sistema.

Los archivos de MedLine se agrupan según el año de publicación. Se encuentra disponible una lista en la que figuran todos los archivos y que proporciona distintos datos útiles para su recuperación: el nombre del archivo, los años cubiertos, el tamaño del archivo y el número de registros totales almacenados en ese fichero. En torno al 78% (14.074.824) del contenido de MedLine, se encuentra escrito en inglés. Además desde

1975, esta base de datos incluye un resumen de sus referencias, llegando a un 57% del total de ellas en la actualidad (12).

El número de citas bibliográficas contenidas en MedLine ha aumentado cada año desde 1995, excepto en el 1998 que sufrió un descenso con respecto a años anteriores. La evolución del número de citas presentes en MedLine desde su creación en 1995 hasta la actualidad se presenta en el siguiente gráfico:

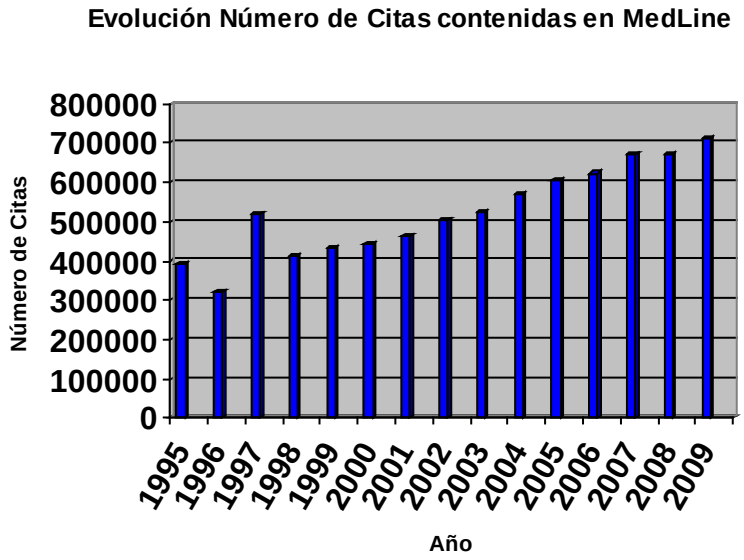


Ilustración 1. Gráfico evolución de la cantidad de citas contenidas en MedLine desde 1995 hasta 2009.

Para facilitar el acceso a esta base de datos, se crea un sistema de búsqueda, PubMed, desarrollado por "National Center for Biotechnology Information" (NCBI) en la NLM. Este proyecto permite el acceso a diversas bases de datos: MedLine, PreMedLine (contiene citas enviadas por los editores), Genbak y Complete Genoma. PubMed permite realizar búsquedas por términos, frases y autores, entre otros. Además es posible el acceso a las referencias bibliográficas y en algunos casos al texto completo.

Tanto MedLine como otras bases de datos bibliográficas de carácter biomédico, utilizan un vocabulario controlado, hablamos del Tesauro MeSH (Medical Subject Heading), para procesar la información que se introduce en ellas. Contiene encabezamientos, calificadores, definiciones, referencias cruzadas, sinónimos, etc.

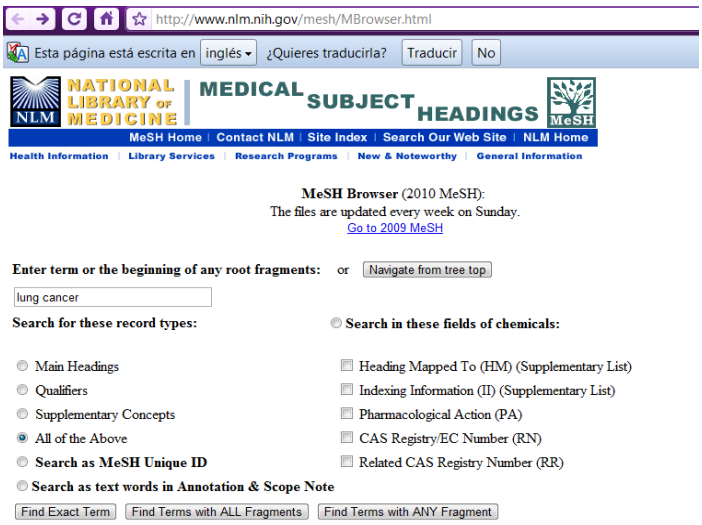


Ilustración2. Interfaz ofrecido por la Biblioteca Nacional de Medicina para la recuperación de artículos.

Cabe mencionar que, un tesauro es un lenguaje artificial que usa un vocabulario controlado. Un vocabulario de este tipo contiene términos específicos (descriptores o palabras clave) para designar cada fenómeno. De esta forma, no se podrá tratar un mismo fenómeno con más de un término diferente, como ocurre en los lenguajes naturales, eliminando problemas derivados de esta "ambigüedad lingüística".

MeSH está formado por más de 33.000 términos ordenados de forma jerárquica, que son revisados anualmente para asegurar que sean un reflejo de la terminología médica del momento. Esta jeraraquía permite que los documentos en MedLine estén clasificados, lo que facilita el acceso a estos. Cada vez que un artículo se registra en la base de datos MedLine, los autores deben asignar uno o más término MeSH a dicho artículo. Este tesauro presenta una traducción al español llamada "Descriptores en Ciencias de la Salud" (DeCS).

En 2010, el tesauro MeSH y su traducción al español, DeCS, sufren algunas adicciones y variaciones con respecto a la versión del año anterior (10 y 11) sus principales cambios serán detallados a continuación:

- Se añaden 422 descriptores MeSH.

- Son sustituidos un total de 61 descriptores por una terminología más actualizada en inglés (52 descriptores MeSH y 9 DeCS).

- 249 descriptores DeCS actualizan sus traducciones al español.

- 83 descriptores DeCS actualizan sus traducciones al portugués.

- Se eliminan 20 descriptores MeSH y 4 DeCS.

- 117 sinónimos MeSH fueron traducidos al español y portugués y 1372 sinónimos DeCS fueron añadidos.

- Finalmente, fueron actualizadas 559 definiciones y 1551 notas de indización.

A continuación, se muestra una relación del número de nuevos descriptores añadidos según su temática, cada uno de estos descriptores puede haber sido añadido a más de una categoría a la vez.

Categoría DeCS/Categoría MeSH	Número nuevos descriptores.
A / Anatomía	23
B / Organismos	45
C / Enfermedades	90
D / Compuestos Químicos y Drogas	101
E / Técnicas y Equipos Analíticos, Diagnósticos y Terapéuticos	67
F / Psiquiatría y Psicología	20
G / Ciencias Biológicas	35
H / Ciencias Naturales	12
HP / Homeopatía	0
I / Antropología, Educación, Sociología y Fenómenos Sociales	10
J / Tecnología, Industria, Agricultura	12
K / Humanidades	2
L / Ciencias de la Información	10
M / Denominaciones de grupos	3
N / Atención de Salud	29
SH / Ciencia y Salud	0
SP / Salud Pública	1
V / Características de Publicaciones	4
VS / Vigilancia Sanitaria	0
Z / Denominaciones Geográficas	4

Tabla 1. Nuevos descriptores añadidos por temática a la base de datos MedLine2010.

En la siguiente tabla, se aprecia en las cifras mostradas una evolución en el contenido de la base de datos entre el año 2009 y 2010.

	2009	2010	Variación	Variación (%)
Total descriptores MeSH	25186	25588	402	1,60%
Total clasificadores MeSH	83	83	0	0,00%
Total descriptores DeCS	4711	4698	-13	-0,30%
Media por idioma del total de sinónimos DeCS	31378	33556	2178	6,90%
Total Conceptos	29980	30369	389	1,30%

Tabla 2. Evolución del contenido de MedLine2009 a MedLine2010.

2.2. METAMAP2009v2

La Biblioteca Nacional de Medicina desarrolla un programa, MetaMap, para descubrir nuevos conceptos referidos en un texto biomédico, y posteriormente asignarlos al Metatesauro, o bien, para descubrir conceptos del Metatesauro referidos en un texto. Esta aplicación está basada en el "Procesamiento del Lenguaje Natural" y en técnicas computacionales de la lingüística. Como ya habíamos mencionado al inicio de esta memoria, MetaMap se distingue del resto de programas por su rigor lingüístico y su dependencia de las fuentes de conocimiento.

2.2.1. UMLS

En 1986, la Biblioteca Nacional de Medicina comenzó la construcción del Unified Medical Language System (UMLS), cuyo propósito era impulsar el desarrollo de sistemas para la recuperación e integración de información biomédica desde distintas fuentes, incluyendo registros electrónicos de pacientes, bases de datos bibliográficas, bases de datos actuales y otros sistemas expertos (17). Para lograr estos objetivos, UMLS integra una gran cantidad de vocabularios en una estructura superior, a través de los siguientes componentes:

– **Metatesauro**: Ontología biomédica formada por una colección de términos extraídos de diferentes vocabularios controlados y sus relaciones (18). El Metatesauro UMLS, es el mayor diccionario de sinónimos en el ámbito biomédico, esta fuente proporciona una representación del conocimiento biomédico muy útil para múltiples aplicaciones (17).

– **El lexicón especializado**: Base de datos con información sintáctica, morfológica y ortográfica, para el uso del Procesamiento del Lenguaje Natural (16).

– **La red semántica**: Posee un conjunto de categorías cuyo fin será la clasificación de las entradas en el Metatesauro (19) en función de su semántica, es decir, de su significado. En concreto, la versión actual de esta red contienen 135 tipos semánticos y todos los conceptos del Metatesauro pertenecen al menos a uno de estos tipos.

2.2.2. Algoritmo MetaMap ¿cómo analiza los textos?

MetaMap es un programa altamente configurable, cuya tarea es realizar el análisis sintácto de cualquier texto biomédico así como la detección de conceptos del Metatesauro UMLS en él. La herramienta está basada en la aplicación de un algoritmo encargado del procesamiento del texto que consta de cinco etapas distintas, hasta devolver la salida propia del programa, compuesta por el análisis morofo-sintáctico del texto y el conjunto de posibles candidatos de conceptos UMLS para cada una de las frases identificadas en el texto (2 y 3).

1. Análisis del texto en sintagmas. División del texto en frases o sintagamas, que durante el proceso serán denominadas con la palabra “Phrase”. Esta división en partes más simples, hace que el esfuerzo de procesamiento sea menor, y por lo tanto, se produce una disminución en el tiempo de procesado total de cada documento.

2. Generación de variantes para cada frase. Una variante consiste en una o más palabras de la frase nominal (llamadas generador), junto con todas sus

variantes de habla, ortográficas, acrónimos y abreviaturas, sinónimos, inflexiones, variantes derivacionales y otras combinaciones significativas.

El proceso de generación de variantes, antes de añadir las inflexiones y las variantes de ortografía, es el que se muestra en la próxima figura:

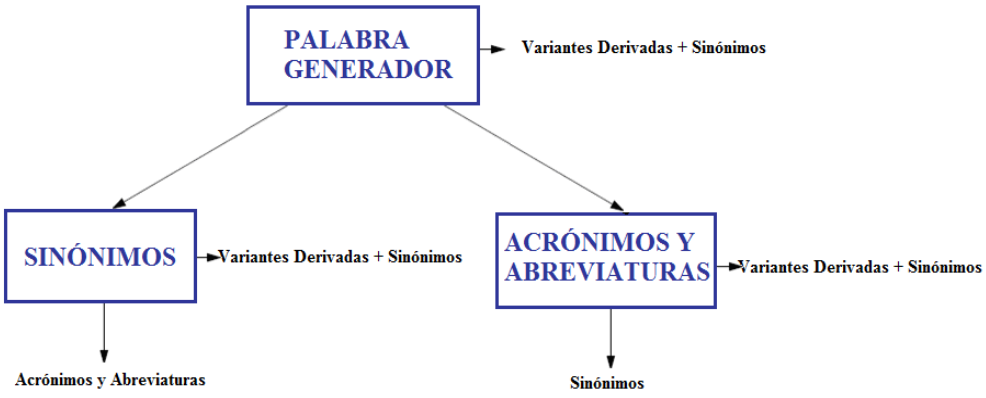


Ilustración 3. Diagrama de bloques ilustrativo del proceso seguido en la generación de variantes.

Los algoritmos de generación de variantes usan diversas fuentes de conocimiento, entre ellas:

-el léxico SPECIALIST y una tabla con las formas canónicas derivadas de este.

-una base de conocimiento SPECIALIST de acrónimos y abreviaturas.

-una base de conocimiento SPECIALIST que contiene reglas de derivación morfológica.

- dos bases de conocimiento de sinónimos, una obtenida por la extracción de sinónimos desde el Dorland s Illustrated Medical Dictionary, y una base de datos complementaria desarrollada para su uso con SPECIALIST.

En la siguiente figura, vemos un ejemplo de este proceso para la palabra “ocular”, donde se muestra el tipo de variante generada en cada momento y la

distancia total de cada una de las variantes en función de clase a la que pertenezca cada una de ellas.

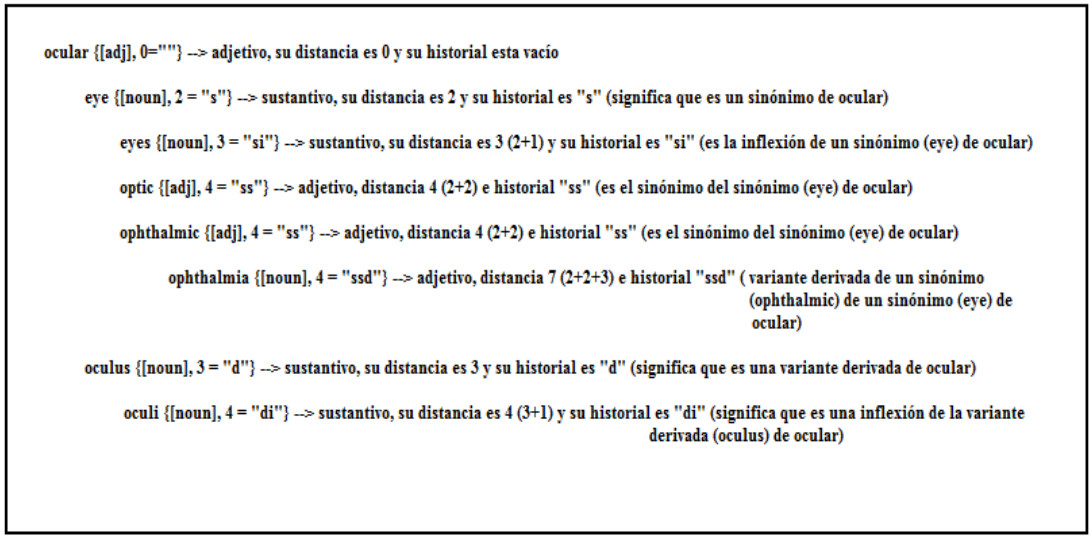


Ilustración4. Generación de variantes para la palabra "*ocular*".

3. Para cada una de las frases, se recupera el "*candidate set*", es decir, el conjunto de conceptos UMLS candidatos, que se forma tomando todas las cadenas contenidas en el Metatesauro UMLS que tengan al menos una de las variantes obtenidas de la palabra o frase. Junto a estos conceptos se devuelve el valor de la función de evaluación que expondrá en el siguiente punto. El Metatesauro contiene al menos un candidato para cada variante.

Como ejemplo, podemos observar el conjunto de candidatos para la frase "*ocular complications*":

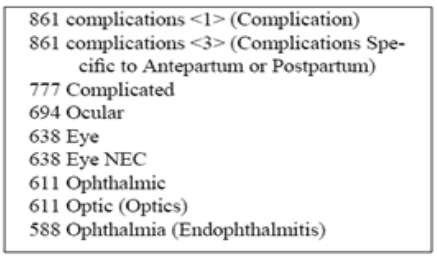


Ilustración 5. Conjunto de Candidatos devueltos para el texto "*ocular complications*" (2).

4. Evaluación de los candidatos obtenidos. En esta fase el algoritmo se encarga del cálculo de una "medida de calidad" del grado de similitud entre una palabra o frase y cada uno de los candidatos obtenidos del Metatesauro. Para el cálculo de dicha medida se realiza una media ponderada de las propiedades: centralidad, variación, cobertura y cohesión. El resultado final será un valor entre 0 y 1000, donde 0 indica que no existe ninguna similitud y por el contrario, 1000 indicará que esta es perfecta.

- Centralidad. El valor de esta propiedad será "1" si la cadena analizada es la parte principal de la frase, llamada "*head*" y por el contrario, el valor será "0" si no lo es. Veamos en la siguiente tabla el valor que le correspondería a esta propiedad en algunos de los candidatos del Metatesauro obtenidos para la cadena de texto "*ocular complications*".

Candidato	Centralidad
Complications	1
Complicated	1
Ocular	0

Tabla 3. Valores de la centralidad para algunos candidatos del texto *"ocular complications"*.

- Variación. Estima el grado en que difieren las variantes obtenidas de sus correspondientes palabras en la frase original. Para calcular esta variación, en primer lugar se halla el valor de la "*variant distance*" para cada variante. Este valor será la suma de las distancias para cada paso en la generación de las variantes, el valor para cada paso se muestra en la Tabla 4.

Tipo de Variante	Distancia
Variante del habla	0
Inflexión	1
Acrónimos, Abreviatura, Expansión y Sinónimo	2
Variantes derivadas	3

Tabla 4. Distancia asignada a cada tipo de paso en la construcción de la tabla de variantes.

A continuación, para ilustrar este paso se adjunto la siguiente figura, en la que dado el texto de entrada "*ocular complications*", se muestra la generación completa de variantes para la palabra "ocular" y el valor que toma la "*variant distance*" en que cada paso para cada una de las variantes.

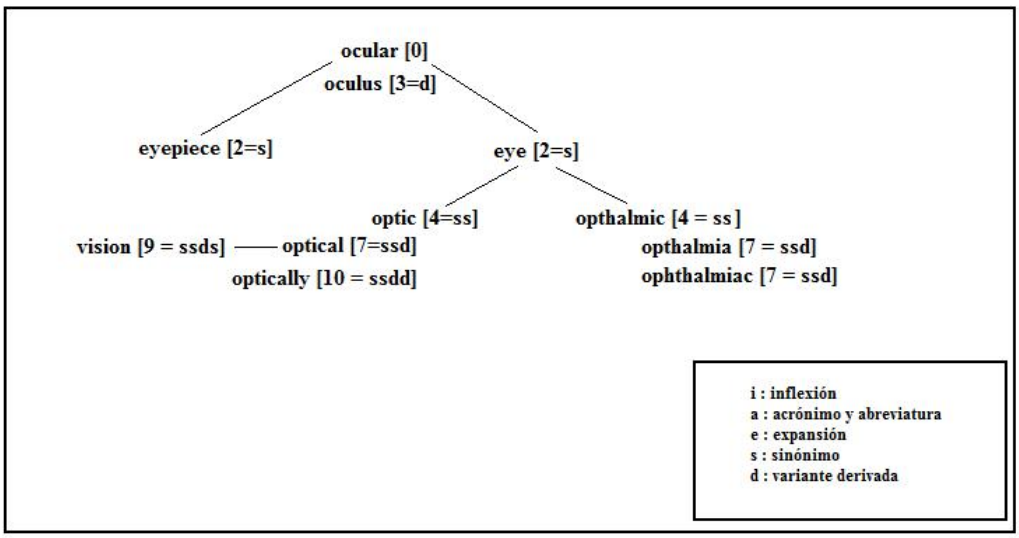


Ilustración 6. Diagrama ilustrativo de la generación de variantes para la palabra "ocular" en el que se muestra el valor de la "variant distance" en cada paso.

El grado de "variación" (V) vendrá determinado por la siguiente expresión, $V = \frac{4}{(D+4)}$, donde D es el valor de "*variant distance*" hallado anteriormente.

El valor final de la variación para un candidato, será el promedio de los valores hallados para cada una de las variantes, es decir, si un candidato esta formado por la combinación de variantes, el valor final de la variación quedará dividido por el número de estas. De esta forma, el valor final para la variación de los candidatos de "*ocular complications*" que se muestran en este ejemplo quedaría dividido por uno, puesto que sólo es usada una variante para dar lugar a cada uno de los candidatos. Los valores obtenidos serían:

Candidatos	Variación
ocular	V= 4/(0+4)=1
optical	V= 4/(7+4)=0.36
complications	V= 4/(0+4)=1

Tabla 5. Valores tomados por la Variación para candidatos del texto "*ocular complications*".

- Cobertura. Medida calculada a partir del número de términos que participan en el proceso de coincidencia ("match"), tanto de la frase que se está evaluando, como de la cadena obtenida del Metatesauro UMLS.

Para calcular esta medida, se hallan dos valores, el "*Metathesaurus span*" y el "*phrase* span". El primero es el número de palabras de la cadena devuelta por el Metatesauro que participan en el proceso, y el segundo el número de palabras de la frase original que también participan en él. Finalmente, estos dos valores son divididos, cada uno por la longitud total de la cadena a la que hacen referencia. Para obtener el valor final de la cobertura, se realiza la media ponderada de los dos valores anteriores, dándole a la cadena del Metatesauro dos veces el peso que a la frase. La expresión final para la cobertura quedaría de la siguiente forma:

$$Cobertura = \left(2 * \frac{MetathesaurusSpan}{LongitudCadenaMetatesauro} + \frac{PhraseSpan}{LongitudCadenaTextoEntrada}\right) / 3$$

Ecuación 1. Expresión del cálculo de la Cobertura para cada candidato.

De nuevo, ilustraremos el cálculo de esta otra propiedad, para ello también tomaremos el texto "*ocular complications*" y algunos de sus conceptos candidatos: "*complications*", "*ocular*" y "*complicated*".

Candidato	Cobertura
complications	(1/2 + 2*1/1)/3=0.83
complicated	(1/2 + 2*1/1)/3=0.83
ocular	(1/2 + 2*1/1)/3=0.83

Tabla 6. Valores calculados para la Cobertura de distintos candidatos del texto "ocular complications".

- Cohesión. Medida similar a la cobertura, pero da especial relevancia a la conexión entre los términos de cada una de las cadenas. Para este cálculo tendremos en cuenta que se denomina como un "*connected component*", a la mayor secuencia de palabras contiguas que participan en el mapeo de una determinada cadena.

El valor de la cohesión para cada candidato, se calcula como la suma del cuadrado del número de palabras contiguas que participan en el proceso de mapeo en los candidatos del Metatesauro, dividido por el cuadrado de la longitud total de la cadena, más el cuadrado del número de palabras contiguas del segmento de texto original que también participan en el mapeo, dividido por la longitud total de este segmento original.

Finalmente, el valor de la cohesión será la media ponderada de la cohesión para la cadena del Metatesauro y del fragmento de texto de entrada que está siendo analizado por MetaMap. La expresión final para el cálculo de la cohesión quedaría:

$$Cohesión = \left(2 * \frac{(NúmeroMáximoPalabrasContiguasCandidato)^2}{(LongitudCadenaMetatesauro)^2} + \frac{(NúmeroMáximoPalabrasContiguasPhrase)^2}{(LongitudCadenaPhrase)^2} \right) / 3$$

Ecuación 2. Expresión del cálculo de la Cohesión para cada candidato.

Para el texto "*ocular complications*", usado como ejemplo para ilustrar este apartado de la memoria, el valor de cohesión de algunos de sus candidatos será el que se muestra a continuación:

Candidatos	Cohesión
complications	$(1^2/2^2 + 2*1^2/1^2)/3=0.75$
complicated	$(1^2/2^2 + 2*1^2/1^2)/3=0.75$
ocular	$(1^2/2^2 + 2*1^2/1^2)/3=0.75$

Tabla 7. Valores calculados para la Cohesión de distintos candidatos del texto "ocular complications".

Tras obtener el valor de estas propiedades, se realiza la evaluación final de cada candidato, para ello se calcula la media ponderada de las cuatro propiedades vistas anteriormente, dando a la cohesión y a la cobertura un peso doble que a la centralidad y la variación, esta media ha de ser normalizada para que el valor final esté en el rango de 0 a 1000. Este cálculo se puede representar mediante la siguiente expresión:

$$EvaluaciónFinalCandidato = 1000 * \frac{(Centralidad + Variación + 2 * Cobertura + 2 * Cohesión)}{6}$$

Ecuación 3. Expresión del cálculo de la Evaluación Final de cada candidato.

Para los candidatos *"ocular"* y *"complications"*, el cálculo de la evaluación final se muestra a continuación, junto con la salida devuelta por MetaMap2009v2, donde aparece el resultado final del proceso de evaluación expuesto. En casos como este, el valor ha de ser redondeado al entero superior.

Candidato	Evaluación final
ocular	1000*(0+1+2*0.83+2*0.75)/6=693.33 = 694
complications	1000*(1+1+2*0.83+2*0.75)/6=861

Tabla 8. Valores obtenidos para la Evaluación Final de dos candidatos del texto *"ocular complications", "ocular" y "complications".*

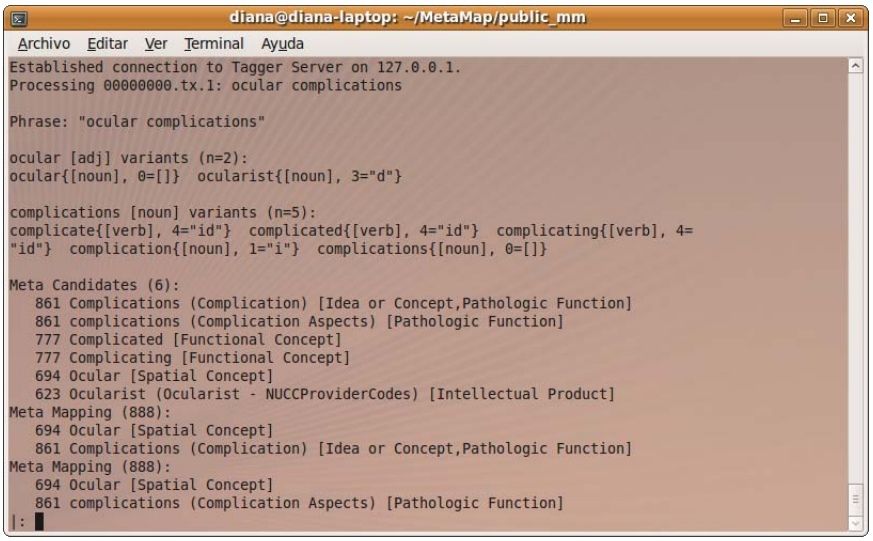


Ilustración 7. Salida de MetaMap2009v2 tras procesar el texto "ocular complications".

5. El mapeo final. Es el paso final del algoritmo de MetaMap, en el cuál se construyen los mapeos completos mediante la combinación de los distintos candidatos involucrados en cada una de las partes de la frase. A continuación, se examina cada combinación obtenida y se evalúan de la misma forma que se hace en el caso de un candidato de forma individual, como se expuso en el paso anterior.

El resultado final del mapeo será el conjunto de candidatos que mejor se ajusten al mapeo de una frase, es decir, será aquel que presente un valor más elevado tras la evaluación. Cabe mencionar, que MetaMap2009v2 presenta una opción que devuelve más de un conjunto final de candidatos y que estudiaremos más adelante.

Las expresiones correspondientes al cálculo de cada propiedad en este paso son:

$$Variación = \frac{4}{(D+4)} = \frac{4}{(D(candidatoMapeoFinal(1)) + \ldots + D(candidatoMapeoFinal(n) + 4)}$$

Ecuación 4. Expresión del cálculo de la Variación de cada combinación de candidatos.

$$Cobertura = \left(2 * \frac{NúmeroPalabrasCombinaciónCandidatos}{LongitudCadenaMetatesauro} + \frac{NúmeroPalabrasPhrase}{LongitudCadenaPhrase} \right) / 3$$

Ecuación 5. Expresión del cálculo de la Cobertura de cada combinación de candidatos.

$$Cohesión = \left(2 * \frac{(N^{o}\ MaxPalabrasContiguasMapeadasCandidato(1))^2 + \ldots + (N^{o}\ MaxPalabrasContiguasMapeadasCandidato(n))^2}{(LongitudTotalCombinaciónCandidatosMetatesauro)^2} + \frac{(N^{o}\ MaxPalabrasContiguasPhrase)^2}{(LongitudTotalCadenaPhrase)^2} \right) / 3$$

Ecuación 6. Expresión del cálculo de la Cohesión de cada combinación de candidatos.

$$EvaluaciónFinalCandidato = 1000 * \frac{(Centralidad + Variación + 2 * Cobertura + 2 * Cohesión)}{6}$$

Ecuación 7. Expresión para el cálculo de la Evaluación Final de cada combinación de candidatos.

Para ilustrar esta última fase, realizaremos la evaluación de la combinación de candidatos *"ocular" y "combinations"* para el texto de entrada *"ocular combinations":*

Característica	Valor
Centralidad	1
Variación(V)	4/(D+4)= 4/((0+0)+4)= 1
Cobertura	(2 / 2+2*(1+1) / 2) / 3 = 1
Cohesión	$(2^2 / 2^2 + 2*(1^2+1^2) / 2^2) / 3 = 2/3$
Evaluación Final	1000*(1+1+2*1+2*0.67)/6=888

Tabla 9. Valores obtenidos para la Evaluación Final de dos candidatos del texto "ocular complications", "ocular" y "complications".

Finalmente cabe mencionar, que desde las primeras versiones de la herramienta se incluye una estructura de datos que ofrece información sobre la relación que existe entre las palabras del texto de entrada que va a aser analizado con MetaMap y sus conceptos UMLS candidatos, aunque no es hasta esta nueva versión cuando esta funcionalidad adquiera mayor relevancia.

2.2.3. Instalación y Ejecución de MetaMap2009v2.

2.2.3.1. Instalación.

Esta versión sólo está disponible para Linux y Solaris, por lo que en función del sistema operativo en el que estemos trabajando, se realizará la descarga de uno de los siguientes archivos disponibles en la página web de MetaMap, tras habernos registrado en ella previamente.

- Si necesitamos la versión de MetaMap para Linux, podrá ser descargada desde: http://metamap.nlm.nih.gov/download/public_mm_linux_2009v2.tar.bz2

- Si la que necesitamos es la versión de MetaMap para Solaris, se podrá descargar desde:
http://metamap.nlm.nih.gov/download/public_mm_solaris_2009v2.tar.bz2

Tras haber descargado correctamente el fichero con la nueva versión de MetaMap, MetaMap2009v2, este ha de ser almacenado en la carpeta donde posteriormente se instalará el programa. A continuación, dado que el archivo esta comprimido, extraeremos su contenido mediante la siguiente línea de comandos en el "Terminal". Para ello, en <platform>, introducimos el nombre de nuestro sistema operativo, en este caso, Linux, y en <year> el nombre completo de la versión, 2009v2.

```
bunzip2 -c public_mm_<platform>_<year>.tar.bz2
bunzip2 -c public_mm_linux_2009v2.tar.bz2
```

Ilustración 8. Secuencia de comandos introducida en el terminal de Linux o Solaris para la extracción de los ficheros contenidos en el paquete descargado de MetaMap2009v2.

Para facilitar el proceso de instalación, introduciremos por línea de comandos el valor de la variable JAVA_HOME, este es la ruta en la que se encuentra almacenada la distribución de Java en nuestro equipo. Si no conocemos esta información, podemos obtenerla mediante la ejecución de "which java" en nuestra consola, como se muestra a continuación:

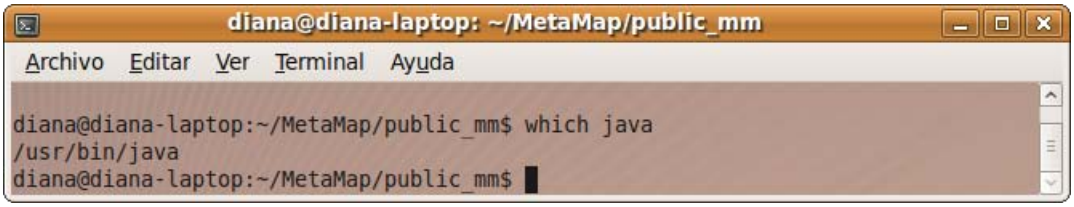


Ilustración 9. Secuencia introducida en la consola para conocer la ruta de instalación de la distribución Java.

En este caso el valor de la variable JAVA_HOME sería /usr/ (si la salida ruta devuelta hubiera sido /usr/local/jre1.4.2/bin/java el valor de esta variable sería /usr/local/jre1.4.2/).

Ahora bien, en función del lenguaje de programación usado por el intérprete de comandos en la consola, introducimos dos de las secuencias que se muestran a continuación:

1. Primera secuencia, en nuestro caso seleccionamos la segunda:

```
# in C Shell (csh or tcsh)
setenv JAVA_HOME /usr/

# in Bourne Again Shell (bash)
export JAVA_HOME=/usr/

# Bourne Shell (sh)
JAVA_HOME=/usr/
export JAVA_HOME
```

2. Segunda secuencia:

```
# in C Shell (csh or tcsh)
setenv PATH MetaMap/public_mm/bin:$PATH

# in Bourne Again Shell (bash)
export PATH=MetaMap/public_mm/bin:$PATH

# Bourne Shell (sh)
PATH=MetaMap/public_mm/bin:$PATH
export PATH
```

Finalmente, podemos iniciar el proceso de instalación, para ello simplemente debemos ejecutar el archivo "i**nstall.sh**", situándonos en su carpeta contenedora.

Tras ejecutar este fichero, durante el proceso de instalación se nos pide introducir la ruta de acceso a la carpeta en la que se encuentra almacenado la distribución completa de la nueva versión de MetaMap y la ruta de acceso a la distribución de Java, que será la obtenida tras ejecutar el comando "which java".

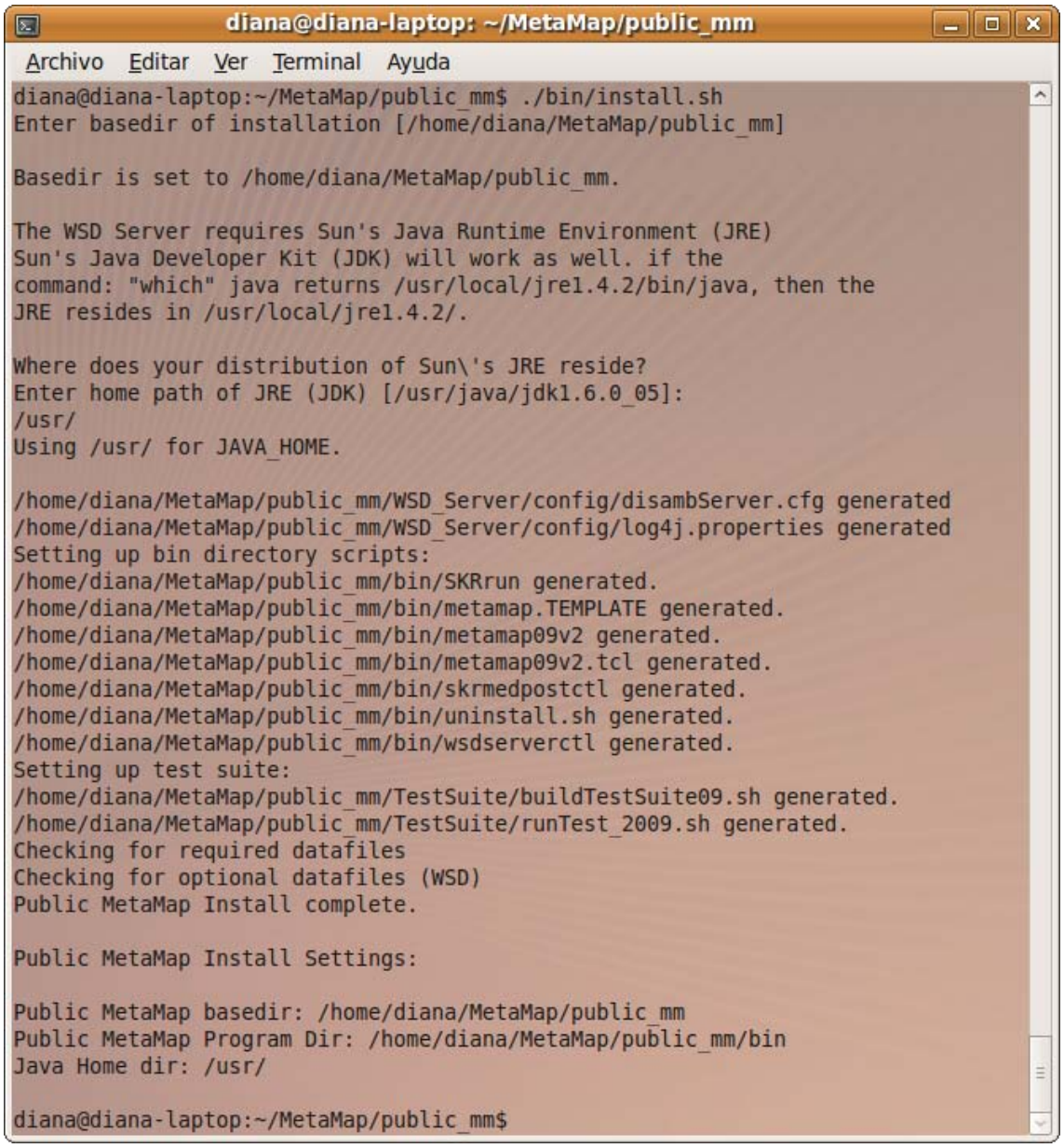


Ilustración 10. Salida de MetaMap2009v2 si la herramienta ha sido instalada correctamente.

2.2.3.2. Ejecución.

Existen dos tipos de servidores diferentes que han de ser activados, en función del modo de uso escogido para MetaMap. En primer lugar está el servidor "SKR/MedPost Part-of-Speech Tagger Server", el cual ha de ser activado siempre, y otro servidor de carácter opcional, que es activado si se quiere usar la opción de desambiguación (-y). Para activar estos servidores la secuencia introducida necesaria es la que sigue:

- Para activar SKR/Medpost Part-of-Speech Tagger Server:

```
./bin/skrmedpostctl start
```

- Para activar Word Sense Disambiguation (WSD) Server:

```
./bin/wsdserverctl start
```

Tras haber activado los servidores necesarios, estamos listos para ejecutar el programa, la estructura general en la que han de ser introducido los comandos para el correcto funcionamiento de la herramienta se muestra a continuación:

```
./bin/metamap09v2 [Options] InputFile OutputFile
```

Para ejecutar MetaMap2009v2, siempre se ha de introducir el siguiente comando, **"./bin/metamap09v2"**, el resto de los términos que aparecen en la figura anterior son de carácter opcional. En el caso de **[Options]**, si ninguna opción es introducida, el programa seguirá funcionando correctamente y el texto será analizado por MetaMap2009v2 usando un conjunto de opciones por defecto.

En cuanto a los ficheros de entrada y salida (**InputFile y OutputFile**), si se desea almacenar en un archivo el resultado obtenido tras el procesado de MetaMap, se ha de indicar al menos la ruta de acceso a un fichero de entrada. Si no se indica una ruta de salida, MetaMap creará por defecto uno con el mismo nombre y ruta que el de entrada, pero con extensión "**.out**". En el caso de indicarse una ruta para el fichero de salida, el resultado será almacenado en esta.

Si por el contrario no se indica ningún fichero de entrada "**InputFile**", la salida no se almacenará en ningún fichero, si no que se mostrará en la consola. En este caso, MetaMap permite introducir más adelante el texto que se desea procesar.

A continuación, veremos en detalle, las dos formas posibles en las que los textos analizados por MetaMap2009v2 pueden ser introducidos, la primera cuando se

introduce en la línea de comandos la ruta de acceso al fichero deseado y una segunda forma en la que una vez ejecutado el programa se introduce el fragmento que se desee analizar.

1. Esta primera forma, está indicada para analizar pequeños fragmentos de texto, puesto que han de ser introducidos manualmente tras la cadena “ **: |** ”. Si observamos el siguiente ejemplo, en el que tras arrancar el programa mediante la introducción en la línea de comandos de “**/bin/metamp09v2**”, sin opciones, se realiza un análisis del texto introducido tras la cadena “ **: |** ”.

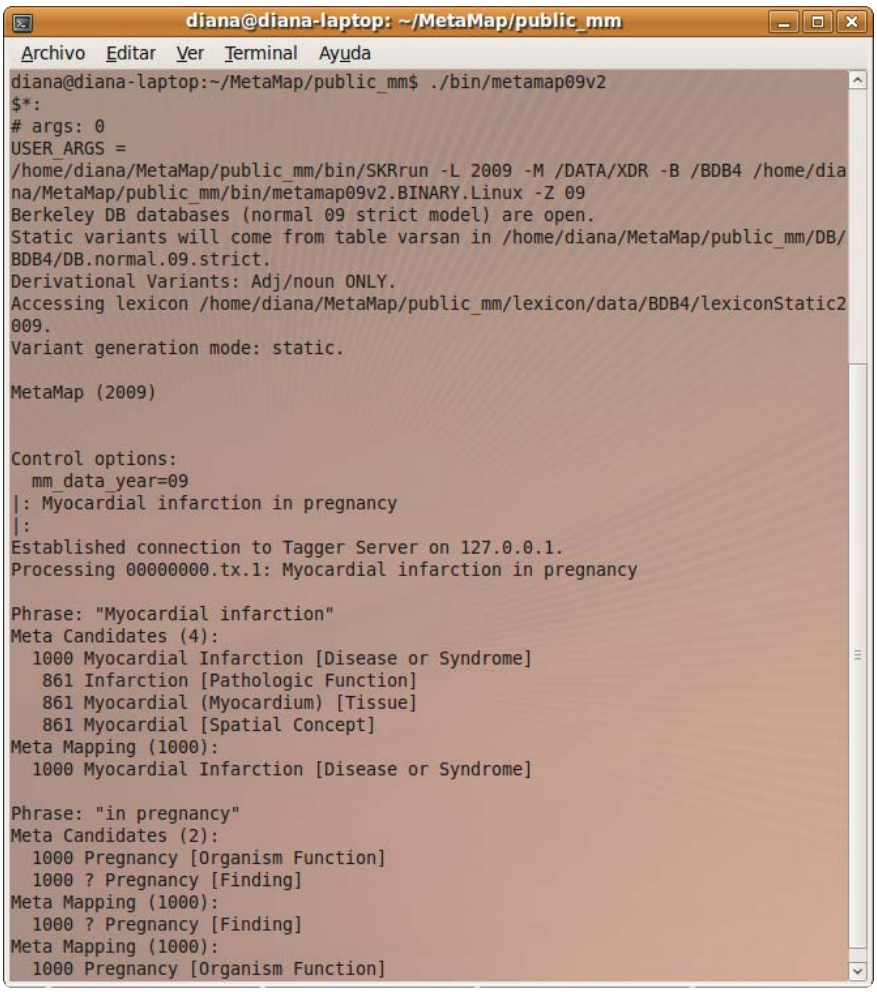


Ilustración 11. Ejecución de MetaMap200v2 procesando el texto “Myocardial infarction in pregnancy” y salida devuelta con el análisis completo de la oración.

2. Si se le indica al programa la dirección del fichero que contiene el texto biomédico que ha de ser procesado, la salida no será mostrada por pantalla como en el caso anterior, si no que es almacenada en un fichero, cuya ruta bien es indicada explícitamente en la introducción de la línea de comandos, o bien MetaMap lo crea por defecto, usando la misma ruta y nombre que el de entrada, añadiéndole la extensión ".out".

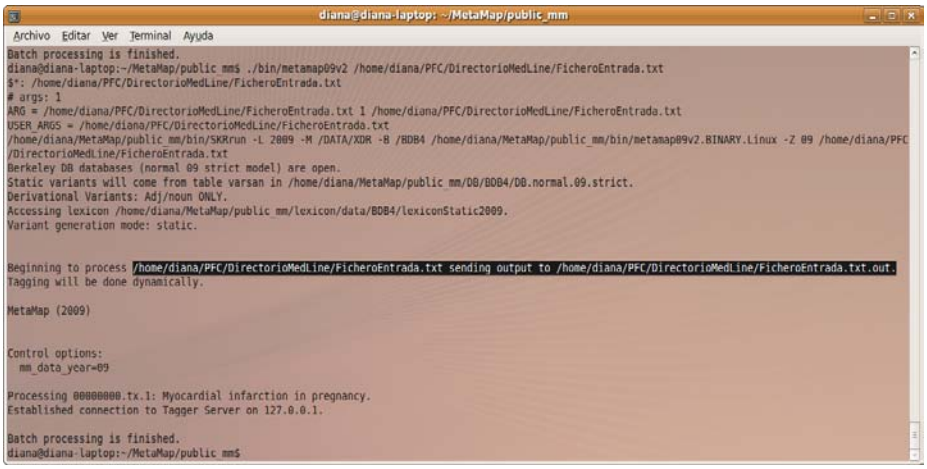


Ilustración 12. Ejecución de MetaMap200v2 que procesa el texto contenido en el fichero de entrada especificado y crea uno de salida por defecto.

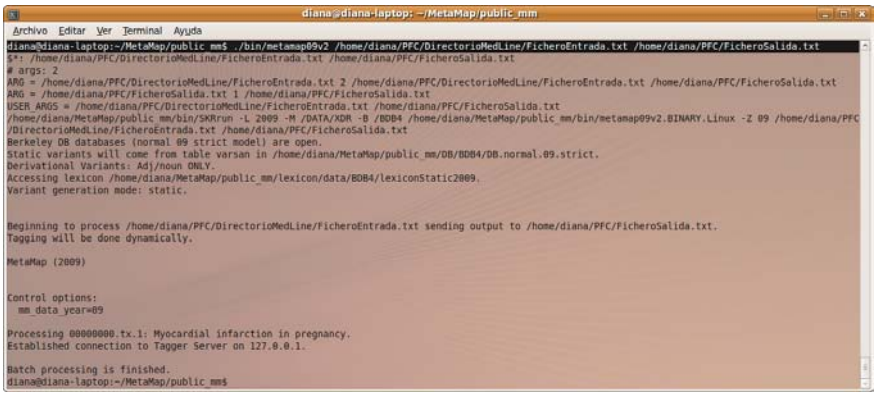


Ilustración 13. Ejecución de MetaMap200v2 que procesa el texto contenido en el fichero de entrada y lo almacena en uno salida, ambos especificados.

En la primera figura, sólo se ha introducido la ruta del fichero de entrada, por lo que MetaMap creará por defecto el archivo de salida. En la parte subrayada se muestra la asignación del nombre

por defecto al fichero de salida. Por el contrario, en la segunda figura se han indicando ambas rutas de entrada y salida.

Tanto en la primera ilustración, como en la segunda, podemos observar que el resultado del procesado no es impreso por pantalla, como ocurre cuando el texto es introducido manualmente en el "terminal".

2.2.4. Información proporcionada por MetaMap2009v2.

El resultado ofrecido por MetaMap variará en función de las opciones seleccionadas, existiendo una amplia gama de estas. Vamos a distinguir entre los siguientes tipos de opciones:

1. **Opciones de procesamiento**, controlan el comportamiento interno de MetaMap.

2. **Opciones de salida**, permiten controlar como MetaMap muestra sus resultados.

Evaluaremos cada una de las opciones incluidas en esta nueva versión, pero en primer lugar definiremos las partes fundamentales y comunes en cada una de las salidas de este programa.

2.2.4.1. Rasgos generales.

Para realizar una exposición más clara de cada uno de los términos principales de una salida MetaMap, observaremos la siguiente figura en la que se representa la salida de MetaMap2009v2 tras procesar el texto "*Myocardial infarction in pregnancy*", estudiando de forma jerárquica cada una de las partes de esta salida.

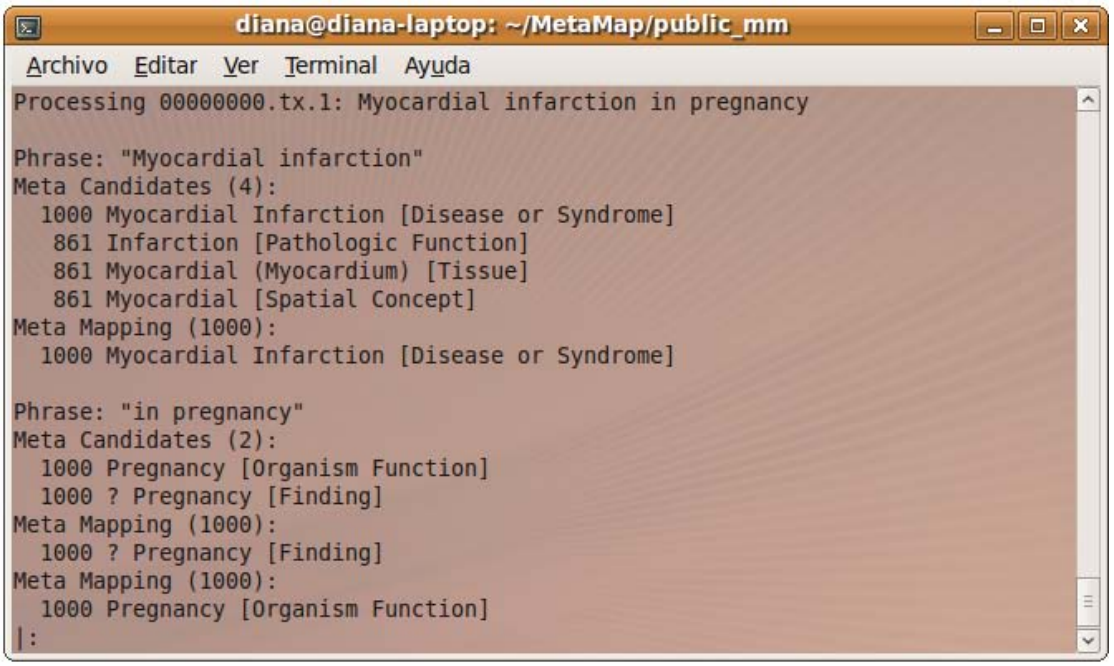


Ilustración 14. Ejemplo salida ofrecida por MetaMap200v2 que va a ilustrarla explicación de sus componentes.

- **Processing 00000000.tx.numero**: Esta cadena precede a cada una de las oraciones en que es dividido el texto de entrada que va a ser procesado. Es decir, podemos considerar a esta cadena como un marcador de oraciones, quedando determinado en todo momento la frase analizada por MetaMap.

El valor que se encuentra al final de esta cadena indica la posición ocupada por la frase en el texto completo de entrada. En la figura adjunta, dado que la entrada sólo tiene una frase, aparece un único "*Processing 00000000.tx.1*", que muestra la oración que se va a analizar "*Myocardial infarction in pregnancy*" y que es la primera del texto completo.

- **Phrase**: Cada uno de los conjuntos formados por una o más palabras, en los que se divide una oración y sobre los que a continuación, se obtienen sus variantes y candidatos. En este ejemplo, tenemos dos "*Phrases*": "*Myocardial infarction*" y "*in pregnancy*".

- **Meta Candidates**: Tras la generación de variantes que lleva a cabo el algoritmo de MetaMap se forma un conjunto de candidatos, este es mostrado a la salida del programa, indicándolo mediante la etiqueta "*Meta Candidates*" y ordenado de manera descendente según el valor obtenido en la función de evaluación descrita en el algoritmo de análisis de MetaMap. Cabe mencionar que este término aparece seguido de un número entre paréntesis, este indica el número total de candidatos devueltos.

En las líneas que siguen a esta etiqueta aparecen los conceptos candidatos obtenidos y cada uno de ellos muestra por defecto un conjunto de datos, estos son:

Evaluación Candidato (Nombre preferido para el concepto, si es diferente al candidato) [Tipo/s Semántico]

En el ejemplo, para "*Myocardial infarction*" se han devuelto cuatro candidatos, en primer lugar aparece "*Myocardial*" puesto que es el que alcanza un valor mayor tras ser evaluado frente al concepto original. Cabe mencionar que cuando se devuelve un valor igual "1000", significa que existe una coincidencia idéntica entre el concepto y el candidato del Metatesauro UMLS.

- **Meta Mapping**: Muestra el resultado final del mapeo, este será la combinación de candidatos que halla obtenido un valor más alto tras ser evaluado frente a la frase original. A diferencia de lo que ocurre en el caso de los Meta Candidate, el valor que aparece entre paréntesis corresponde el mayor nivel de coincidencia obtenido tras evaluar todas las combinaciones posibles de los candidatos.

En nuestro ejemplo, "*Meta Mapping*" es seguido de un valor entre paréntesis igual a "1000", cuyo significado es la coincidencia absoluta entre la frase construida tras la combinación de los candidatos devueltos por el Metatesauro y la cadena original. En la cadena "Myocardial infarction" sólo es devuelto un "*Meta Mapping*", mientras que con "in pregnancy", aparecen dos. Esto se debe a que en el segundo caso, dos

posibles combinaciones han alcanzado el valor más alto, mientras que en la primera cadena sólo lo ha hecho una de las combinaciones candidatas.

2.2.4.2. Opciones de Procesado.

Estas opciones modelan el comportamiento interno de MetaMap:

-+ (--bracketed output)

Marca entre "<<<<<<" y ">>>>>>" cada una de las secciones que componen la salida devuelta por MetaMap (Phrase, Candidates y Mapping).

```
Phrase: "Association"
>>>>> Phrase
association
<<<<< Phrase
>>>>> Candidates
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
  928 Associated [Qualitative Concept]
  928 Associate [Idea or Concept,Professional or Occupational Group]
  928 Associationism [Idea or Concept]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
<<<<< Mappings
```

Ilustración 15. Salida MetaMap2009v2 aplicando la opción -+ para la evaluación de "Association".

-@ (--WSD) <option>

Especifica el nombre del host que ejecuta el Servidor WSD (Word Sense Disambiguation), usado en los procesos de desambiguación.

-8 (--dynamic_variant_generation)

La generación de variantes se hace de forma dinámica, en lugar de buscar las posibles variantes de una palabra en una tabla, como se hace habitualmente. Esta opción es usada sólo en acciones de depuración.

-a (--all_acros_abbrs)

Permite añadir al conjunto de candidatos variantes de tipo Acrónimo y Abreviatura. Estas son las variantes menos fiables, puesto que sólo una de las expansiones del acrónimo o abreviatura es la correcta.

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   944 ASSOC (Associated) [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 16. Salida MetaMap2009v2 aplicando la opción -a para la evaluación de "Association".

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or  Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 17. Salida MetaMap2009v2 sin aplicar la opción -a para la evaluación de "Association".

-d (--no_derivational_variants)

MetaMap no usará "variantes derivadas" para la creación de la tabla de variantes, y por lo tanto, ningún componente del conjunto de candidatos será una variante de este tipo.

Para ilustrar este caso mostraremos tanto la salida obtenida aplicando esta opción, como la devuelta en el caso de no usarla. Al no usarse estas variantes, el

número de candidatos disminuye significativamente y los que permanecen presentarán un valor del resultado de evaluación muy elevado, dado que ningún será una “variante derivada”.

```
Phrase: "Association"
Meta Candidates (4):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 18. Salida MetaMap2009v2 tras aplicar la opción -d para la evaluación de “Association”.

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Sociar del Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 19. Salida MetaMap2009v2 sin aplicar la opción -d para la evaluación de “Association”.

-D (--all_derivational_variants)

MetaMap usa “variantes derivadas” para la generación de la tabla de variantes, aunque sólo se permitirá el uso de este tipo si son procedentes de nombres y adjetivos, mientras que si no usamos esta opción, MetaMap tomará todas las “variantes derivadas” de la palabra. Al igual que en el caso anterior, ilustraremos este caso con las dos figuras que se muestran a continuación.

```
Phrase: "of maternal"
Meta Candidates (4):
  1000 Maternal [Clinical Attribute]
   944 Mother (Mothers) [Family Group]
   928 Materna [Organic Chemical,Pharmacologic Substance,Vitamin]
   916 MATER (NLRP5 gene) [Gene or Genome]
Meta Mapping (1000):
  1000 Maternal [Clinical Attribute]
```

Ilustración 20. Salida MetaMap2009v2 tras aplicar la opción -D para la evaluación de “of maternal”.

```
Phrase: "of maternal"
Meta Candidates (12):
  1000 Maternal [Clinical Attribute]
  944 Mother (Mothers) [Family Group]
  928 Materna [Organic Chemical,Pharmacologic Substance,Vitamin]
  916 MATER (NLRP5 gene) [Gene or Genome]
  893 Mats [Medical Device]
  893 mate (Partner in relationship) [Family Group]
  893 mate (Yerba mate) [Plant]
  893 MAT ([acyl-carrier-protein] S-malonyltransferase activity) [Molecular Function]
  893 mating [Organism Function]
  893 MAT (ACAT1 gene) [Gene or Genome]
  893 MAT (MAT1A gene) [Gene or Genome]
  893 Mate (Yerba mate preparation) [Pharmacologic Substance]
Meta Mapping (1000):
  1000 Maternal [Clinical Attribute]
```

Ilustración 21. Salida MetaMap2009v2 sin aplicar la opción -D para la evaluación de "of maternal".

-g (--allow_concept_gaps)

Devuelve como candidatos conceptos con "gaps", es decir, "con diferencias", como ocurre con el candidato "*Mental impairment disorders*" para "*mental disorders*".

```
Phrase: "on  eventual mental disorders."
Meta Candidates (18):
  901 Mental disorders [Mental or Behavioral Dysfunction]
  837 Psychic disorder NOS (Psychic disease) [Mental or Behavioral Dysfunction]
  827 Disorders (Disease) [Disease or Syndrome]
  827 MENTAL (Psyche structure) [Mental Process]
  812 Diaphragmatic disorders (Disease of diaphragm) [Disease or Syndrome]
  755 Mentum (Chin) [Body Location or Region]
  743 Psychic [Functional Concept]
  728 Mental impairment disorders [Mental or Behavioral Dysfunction]
  727 Diaphragm (Respiratory Diaphragm) [Body Part, Organ, or Organ Component]
  727 DIAPHRAGM [Medical Device]
  727 Diaphragm (Entire diaphragm) [Body Part, Organ, or Organ Component]
  727 Diaphragm (Device Diaphragm) [Medical Device]
  727 DIAPHRAGM (Diaphragm Dosage Form) [Biomedical or Dental Material]
  727 Diaphragms [Medical Device]
  721 PSYCH (Psychiatric problem) [Mental or Behavioral Dysfunction]
  711 Mental+behavioral disorder [Mental or Behavioral Dysfunction]
  699 Psyche [Invertebrate]
  670 Phrenic nerve disorder [Disease or Syndrome]
Meta Mapping (901):
  901 Mental disorders [Mental or Behavioral Dysfunction]
Processing 00000000.tx.2: The mental defectives had a greater amount of recorded
complications of pregnancy and delivery, prematurity, and abnormal neonatal
experiences.
```

Ilustración 22. Salida MetaMap2009v2 tras aplicar la opción -g para la evaluación de "on eventual disorders".

```
Phrase: "on  eventual mental disorders."
Meta Candidates (15):
   901 Mental disorders [Mental or Behavioral Dysfunction]
   837 Psychic disorder NOS (Psychic disease) [Mental or Behavioral Dysfunction]
   827 Disorders (Disease) [Disease or Syndrome]
   827 MENTAL (Psyche structure) [Mental Process]
   812 Diaphragmatic disorders (Disease of diaphragm) [Disease or Syndrome]
   755 Mentum (Chin) [Body Location or Region]
   743 Psychic [Functional Concept]
  727 Diaphragm (Respiratory Diaphragm) [Body Part, Organ, or Organ Component]
   727 DIAPHRAGM [Medical Device]
   727 Diaphragm (Entire diaphragm) [Body Part, Organ, or Organ Component]
   727 Diaphragm (Device Diaphragm) [Medical Device]
   727 DIAPHRAGM (Diaphragm Dosage Form) [Biomedical or Dental Material]
   727 Diaphragms [Medical Device]
   721 PSYCH (Psychiatric problem) [Mental or Behavioral Dysfunction]
   699 Psyche [Invertebrate]
Meta Mapping (901):
   901 Mental disorders [Mental or Behavioral Dysfunction]
Processing 00000000.tx.2: The mental defectives had a greater amount of recorded
complications of pregnancy and delivery, prematurity, and abnormal neonatal experiences.
```

Ilustración 23. Salida MetaMap2009v2 sin ser aplicada la opción -g para la evaluación de "on eventual disorders".

-i (--ignore_word_order)

Con esta opción MetaMap ignora el orden de las palabras en las frases procesadas. Ignorar el orden de las palabras hace que el valor de la evaluación de los candidatos varíe con respecto a la obtenida sin aplicar esta opción:

```
Phrase: "The mental defectives"
Meta Candidates (17):
   907 DIAPHRAGM DEFECT (Defect of diaphragm) [Anatomical Abnormality]
   896 Diaphragmatic defect [Finding]
   799 Defective [Functional Concept]
   750 defects (defects aspect) [Functional Concept]
   750 Defect [Functional Concept]
  750 DEFECT (THYROID HORMONE PLASMA MEMBRANE TRANSPORT
DEFECT) [Disease or Syndrome]
   666 MENTAL (Psyche structure) [Mental Process]
   595 Mentum (Chin) [Body Location or Region]
  583 Psychic [Functional Concept]
  566 Diaphragm (Respiratory Diaphragm) [Body Part, Organ, or Organ Component]
  566 DIAPHRAGM [Medical Device]
  566 Diaphragm (Entire diaphragm) [Body Part, Organ, or Organ Component]
  566 Diaphragm (Device Diaphragm) [Medical Device]
  566 DIAPHRAGM (Diaphragm Dosage Form) [Biomedical or Dental Material]
  566 Diaphragms [Medical Device]
  560 PSYCH (Psychiatric problem) [Mental or Behavioral Dysfunction]
  539 Psyche [Invertebrate]
Meta Mapping (907):
   907 DIAPHRAGM DEFECT (Defect of diaphragm) [Anatomical Abnormality]
```

Ilustración 24. Salida MetaMap2009v2 tras aplicar la opción -i para la evaluación de "The mental defectives.".

```
Phrase: "The mental defectives"
Meta Candidates (17):
  907 DIAPHRAGM DEFECT (Defect of diaphragm) [Anatomical Abnormality]
  896 Diaphragmatic defect [Finding]
  827 Defective [Functional Concept]
  777 defects (defects aspect) [Functional Concept]
  777 Defect [Functional Concept]
  777 DEFECT (THYROID HORMONE PLASMA MEMBRANE TRANSPORT DEFECT)
[Disease or Syndrome]
  694 MENTAL (Psyche structure) [Mental Process]
  623 Mentum (Chin) [Body Location or Region]
  611 Psychic [Functional Concept]
  594 Diaphragm (Respiratory Diaphragm) [Body Part, Organ, or Organ Component]
  594 DIAPHRAGM [Medical Device]
  594 Diaphragm (Entire diaphragm) [Body Part, Organ, or Organ Component]
  594 Diaphragm (Device Diaphragm) [Medical Device]
  594 DIAPHRAGM (Diaphragm Dosage Form) [Biomedical or Dental Material]
  594 Diaphragms [Medical Device]
  588 PSYCH (Psychiatric problem) [Mental or Behavioral Dysfunction]
  566 Psyche [Invertebrate]
Meta Mapping (907):
  907 DIAPHRAGM DEFECT (Defect of diaphragm) [Anatomical Abnormality]
```

Ilustración 25. Salida MetaMap2009v2 sin aplicar la opción -i para la evaluación de "The mental defectives.".

-l (--allow_large_n)

Devuelve candidatos no sólo de la forma habitual, sino también para las palabras compuestas por uno o dos caracteres, existiendo en el Metatesauro, 1000 cadenas de un carácter y 2000 de dos. Normalmente estos candidatos van a ser conjunciones, determinantes y preposiciones. Como se aprecia en el ejemplo al seleccionar esta opción, se va a producir también una variante correspondiente a "to" mientras que si no es seleccionada, no aparece ninguna variante de "to".

```
Phrase: "of a control group to"
Meta Candidates (15):
  901 Control Group (Control Groups) [Group]
  827 control (control aspects) [Qualitative Concept]
  827 Group (Groups) [Idea or Concept]
  827 Group (Population Group) [Population Group]
  827 Control (control substance) [Substance]
  827 Group (Specialty Group) [Health Care Related Organization]
  827 Group (Group Object) [Conceptual Entity]
  827 Group (User Group) [Population Group]
  827 Control (Scientific Control) [Conceptual Entity]
  827 CONTROL (Control function) [Functional Concept]
  793 Grouped [Spatial Concept]
  793 Grouping [Functional Concept]
  793 groups [Idea or Concept]
  793 Controlling [Phenomenon or Process]
  660 To [Qualitative Concept]
Meta Mapping (901):
  901 Control Group (Control Groups) [Group]
  660 To [Qualitative Concept]
```

Ilustración 26. Salida MetaMap2009v2 aplicando la opción -l para la evaluación de "of a control group to".

-K (--ignore_stop_phrases)

Opción usada para generar una nueva tabla de frases de parada tras un cambio en los datos de UMLS.

-o (--allow_overmatches)

Devuelve como candidatos todas las variantes que se encuentren en el Metatesauro UMLS, que tengan en uno o en ambos extremos palabras que coincidan con la palabra o palabras generador, a estos casos se les llama "**overmatches**". El uso de esta opción incrementa el número de candidatos devueltos y por tanto hace más lenta la ejecución de MetaMap. Para evaluar esta opción hemos obtenidos los resultados que se muestran a continuación, aunque sólo se muestra un fragmento, puesto que es muy extensa.

```
Phrase: "Association"
Meta Candidates (1550):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
   722 Association Learning [Mental Process]
   722 Free Association [Therapeutic or Preventive Procedure]
   722 Association, Library (Library Associations) [Professional Society]
   722 VATER association [Congenital Abnormality,Disease or Syndrome]
   722 Malformation association [Disease or Syndrome]
   722 MVRCS association [Disease or Syndrome]
 722 CHARGE association (CHARGE association disorder) [Congenital
Abnormality,Disease or Syndrome]
   722 Thought association [Mental Process]


.....(continúa)
```

Ilustración 27. Salida MetaMap2009v2 aplicando la opción -o para la evaluación de "Association".

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 28. Salida MetaMap2009v2 si no es aplicada la opción -o para la evaluación de "Association".

-P (--composite_phrases)

MetaMap tomará para su mapeo frases compuestas en lugar de otras más simples. Una frase compuesta es una frase simple seguida de un sintagma preposicional que a su vez puede ir seguido de una o más frases preposicionales. Esta opción está en fase experimental, puesto que actualmente el análisis devuelto no es del todo correcto. Un ejemplo de esta nueva funcionalidad se muestra a continuación, es importante anotar que el procesamiento es más lento al usar esta opción.

```
Phrase: "Association of maternal"
Meta Candidates (100):
   790 Association (Mental association) [Mental Process]
   790 Association (Relationships) [Qualitative Concept]
   790 Association (Chemical Association) [Phenomenon or Process]
   790 association (Relationship by association) [Social Behavior]
   718 Associated [Qualitative Concept]
   718 Associate [Idea or Concept,Professional or Occupational Group]
   718 Associationism [Idea or Concept]
   662 Loosening of associations [Mental or Behavioral Dysfunction]
   662 Speed of associations [Mental or Behavioral Dysfunction]
   662 Structure of associations [Mental or Behavioral Dysfunction]
   658 ASSOCIATE OF MYC (MYCBP gene) [Gene or Genome]
    ..... .....(continúa)
```

Ilustración 29. Salida MetaMap2009v2 tras aplicar la opción -P en la evaluación de "Association of maternal".

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]

Phrase: "of maternal"
Meta Candidates (12):
  1000 Maternal [Clinical Attribute]
   944 Mother (Mothers) [Family Group]
   928 Materna [Organic Chemical,Pharmacologic Substance,Vitamin]
   916 MATER (NLRP5 gene) [Gene or Genome]
   893 Mats [Medical Device]
   893 mate (Partner in relationship) [Family Group]
   893 mate (Yerba mate) [Plant]
 893 MAT ([acyl-carrier-protein] S-malonyltransferase activity) [Molecular Function]
   893 mating [Organism Function]
   893 MAT (ACAT1 gene) [Gene or Genome]
   893 MAT (MAT1A gene) [Gene or Genome]
   893 Mate (Yerba mate preparation) [Pharmacologic Substance]
Meta Mapping (1000):
  1000 Maternal [Clinical Attribute]
```

Ilustración 30. Salida MetaMap2009v2 sin aplicar la opción -P en la evaluación de "Association of maternal".

-Q (--quick_composite_phrases)

Versión de la opción estudiada "**--composite_phrases**", diseñada para mejorar la ineficiencia de esta. Al igual que la anterior analiza frases compuestas, pero incluye un número menor de candidatos, puesto que a diferencia que la otra, esta incluye en la tabla de candidatos conceptos del Metatesauro formados por una sola palabra. Esta opción presenta una mayor eficiencia de la opción anterior y disminuye velocidad de procesamiento.

```
Phrase: "Association of maternal"
Meta Candidates (19):
   790 Association (Mental association) [Mental Process]
   790 Association (Relationships) [Qualitative Concept]
   790 Association (Chemical Association) [Phenomenon or Process]
   790 association (Relationship by association) [Social Behavior]
   718 Associated [Qualitative Concept]
   718 Associate [Idea or Concept,Professional or Occupational Group]
   718 Associationism [Idea or Concept]
   623 Maternal [Clinical Attribute]
   567 Mother (Mothers) [Family Group]
   552 Materna [Organic Chemical,Pharmacologic Substance,Vitamin]
   540 MATER (NLRP5 gene) [Gene or Genome]
   517 Mats [Medical Device]
   517 mate (Partner in relationship) [Family Group]
   517 mate (Yerba mate) [Plant]
   517 MAT ([acyl-carrier-protein] S-malonyltransferase activity) [Molecular Function]
   517 mating [Organism Function]
 Meta Mapping (746):
   790 Association (Chemical Association) [Phenomenon or Process]
   623 Maternal [Clinical Attribute]
Meta Mapping (746):
   790 Association (Mental association) [Mental Process]
   623 Maternal [Clinical Attribute]
Meta Mapping (746):
   790 association (Relationship by association) [Social Behavior]
   623 Maternal [Clinical Attribute]
Meta Mapping (746):
   790 Association (Relationships) [Qualitative Concept]
   623 Maternal [Clinical Attribute]
```

Ilustración 31. Salida MetaMap2009v2 tras aplicar la opción -Q en la evaluación de "Association of maternal".

-S (--tagger OPTION)

Especifica el nombre del host que va a ejecutar el "Tagger Server" usado durante el proceso de etiquetado.

-t (--no_tagging)

El analizador SPECIALIST usará los resultados de su "etiquetador" por defecto durante el proceso de análisis, división y asignación de etiquetas a las parte del texto que se va a evaluar con MetaMap. Actualmente se usa el etiquetador Med-Post/SKR. Este servidor fue desarrollado por la NCBI, específicamente para el análisis de textos biomédicos.

-u (--unique_acros_abbrs_only)

Restringe la generación de variantes de tipo acrónimo y abreviatura, sólo a aquellas que tengan una única expansión. Esta será la opción que produzca mejores resultados en cuanto a la devolución de variantes de este tipo, pero sigue siendo más eficiente que no se generen este tipo de variantes.

```
Phrase: "of mental deficiency."
Meta Candidates (23):
  1000 Mental deficiency [Mental or Behavioral Dysfunction]
  944 MDS (Dysmyelopoietic Syndromes) [Neoplastic Process]
  944 MDS (Miller Dieker syndrome) [Disease or Syndrome]
  928 mentally deficient (Mentally Disabled Persons) [Patient or Disabled Group]
  916 DPH DEFICIENCY [Disease or Syndrome]
  861 Deficiency [Functional Concept]
  861 Deficiency (Malnutrition) [Disease or Syndrome]
  861 MENTAL (Psyche structure) [Mental Process]
  861 deficiency (deficiency aspects) [Qualitative Concept]
  789 Mentum (Chin) [Body Location or Region]
  789 % deficient [Quantitative Concept]
  777 Psychic [Functional Concept]
  761 Diaphragm (Respiratory Diaphragm) [Body Part, Organ, or Organ Component]
  761 DIAPHRAGM [Medical Device]
  761 Diaphragm (Entire diaphragm) [Body Part, Organ, or Organ Component]
  761 Diaphragm (Device Diaphragm) [Medical Device]
  761 DIAPHRAGM (Diaphragm Dosage Form) [Biomedical or Dental Material]
  761 Diaphragms [Medical Device]
  755 PSYCH (Psychiatric problem) [Mental or Behavioral Dysfunction]
  738 DiA (DIASP) [Organic Chemical,Pharmacologic Substance]
  738 DIA (Desmoplastic infantile astrocytoma) [Neoplastic Process]
  738 DIA (DIAPH2 gene) [Gene or Genome]
  733 Psyche [Invertebrate]
Meta Mapping (1000):
  1000 Mental deficiency [Mental or Behavioral Dysfunction]
```

Ilustración 32. Salida MetaMap2009v2 tras aplicar la opción -u en la evaluación de "of mental deficiency".

```
Phrase: "of mental deficiency."
Meta Candidates (17):
  1000 Mental deficiency [Mental or Behavioral Dysfunction]
  928 mentally deficient (Mentally Disabled Persons) [Patient or Disabled Group]
  861 Deficiency [Functional Concept]
  861 Deficiency (Malnutrition) [Disease or Syndrome]
  861 MENTAL (Psyche structure) [Mental Process]
  861 deficiency (deficiency aspects) [Qualitative Concept]
  789 Mentum (Chin) [Body Location or Region]
  789 % deficient [Quantitative Concept]
  777 Psychic [Functional Concept]
  761 Diaphragm (Respiratory Diaphragm) [Body Part, Organ, or Organ Component]
  761 DIAPHRAGM [Medical Device]
  761 Diaphragm (Entire diaphragm) [Body Part, Organ, or Organ Component]
  761 Diaphragm (Device Diaphragm) [Medical Device]
  761 DIAPHRAGM (Diaphragm Dosage Form) [Biomedical or Dental Material]
  761 Diaphragms [Medical Device]
  755 PSYCH (Psychiatric problem) [Mental or Behavioral Dysfunction]
  733 Psyche [Invertebrate]
Meta Mapping (1000):
  1000 Mental deficiency [Mental or Behavioral Dysfunction]
```

Ilustración 33. Salida MetaMap2009v2 si no se aplica la opción -u en la evaluación de "of mental deficiency".

-U (--allow_duplicate_concept_names)

Permite la aparición en el conjunto de candidatos de conceptos duplicados. Es decir, de conceptos que poseen el mismo CUI (Identificador de Concepto Único). Al usar esta opción la eficiencia de la herramienta disminuirá, puesto que se realiza el mismo mapeo en múltiples ocasiones, al ser devuelto un mimo concepto como varios candidatos.

```
Phrase: "neonatal records"
Meta Candidates (6):
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
  694 Neonatal (Infant, Newborn) [Age Group]
  694 Neonatal (Neonatal Clinical Nurse Specialist) [Professional or Occupational Group]
  694 Neonatal (Neonatal Nurse Practitioner) [Professional or Occupational Group]
Meta Mapping (888):
  694 Neonatal (Infant, Newborn) [Age Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Infant, Newborn) [Age Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Infant, Newborn) [Age Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Clinical Nurse Specialist) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Clinical Nurse Specialist) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Clinical Nurse Specialist) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Nurse Practitioner) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Nurse Practitioner) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Nurse Practitioner) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
```

Ilustración 34. Salida MetaMap2009si se aplica la opción -U en la evaluación de "neonatal records".

```
Phrase: "neonatal records"
Meta Candidates (4):
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
  694 Neonatal (Infant, Newborn) [Age Group]
  694 Neonatal (Neonatal Clinical Nurse Specialist) [Professional or Occupational Group]
  694 Neonatal (Neonatal Nurse Practitioner) [Professional or Occupational Group]
Meta Mapping (888):
  694 Neonatal (Infant, Newborn) [Age Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Clinical Nurse Specialist) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
Meta Mapping (888):
  694 Neonatal (Neonatal Nurse Practitioner) [Professional or Occupational Group]
  861 Records [Idea or Concept,Intellectual Product,Qualitative Concept]
```

Ilustración 35. Salida MetaMap2009v2 sin aplicar la opción -U en la evaluación de "neonatal records".

-y (--word_sense_disambiguation)

Opción usada para solucionar problemas de ambigüedad, en la que varios candidatos posean un mismo valor de evaluación. Para seleccionar esta opción se ha de activar el servidor WSD (Word Sense Disambiguation), encargado de solventar dichos problemas.

-Y (--prefer_multiple_concepts)

Esta opción es usada para descubrir el grado de cohesión que existe entre los conceptos encontrados en un texto.

-z (--term_processing)

Cuando se invoca esta opción, MetaMap va a tratar cada entrada como una sola frase. En estos casos MetaMap va necesitar mucho más tiempo intentando combinar todos los "meta candidates" obtenidos para dar un mapeo final. Es decir, esta opción lo que hace es tomar una frase completa, en lugar de fragmentos formados por una o varias o palabras.

2.2.4.3. Opciones de Salida.

Las opciones de salida nos permiten controlar la forma en la que MetaMap va a mostrar sus resultados:

-b (--compute_all_mappings)

Esta opción muestra todos los mapeos finales obtenidos tras la combinación de todos los candidatos devueltos y no sólo aquellos que presentan un resultado más elevado tras su evaluación. Esta opción, es raras veces usadas, puesto que al mostrarse todos los mapeos la salida resulta demasiado larga.

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 asociamiento (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
Meta Mapping (928):
  928 Associate [Idea or Concept,Professional or Occupational Group]
Meta Mapping (928):
  928 Associated [Qualitative Concept]
Meta Mapping (928):
  928 Associationism [Idea or Concept]
```

Ilustración 36. Salida MetaMap2009v2 tras aplicar la opción -b en la evaluación de "Association".

-c (--hide_candidates)

Esta opción deshabilita la lista de los Meta Candidates, es decir, no se va a mostrar el conjunto de candidatos extraídos de las variantes halladas para un concepto.

```
Phrase: "Association"
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 37. Salida MetaMap2009v2 tras aplicar la opción -c en la evaluación de "Association".

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 38. Salida MetaMap2009v2 sin aplicar la opción -c en la evaluación de "Association".

-G (--sources)

Muestra entre "{ }" para cada uno de los candidatos, la fuente o fuentes del Metatesauro UMLS de las que proviene cada uno.

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association {MTH, MSH, RCD, SNM, SNOMEDCT, CSP}) [Mental
Process]
  1000 Association (Relationships {MTH, RCD, SNOMEDCT, HL7V2.5, ICNP, NCI})
[Qualitative Concept]
  1000 Association (Chemical Association {MTH, NCI, CSP}) [Phenomenon or Process]
  1000 association (Relationship by association {MTH, AOD}) [Social Behavior]
   928 Associated {MTH, RCD, SNMI, SNOMEDCT, CCPSS, LNC, NCI} [Qualitative Concept]
   928 Associate {MTH, HL7V2.5} [Idea or Concept,Professional or Occupational Group]
   928 Associationism {PSY} [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association {MTH, NCI, CSP}) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association {MTH, MSH, RCD, SNM, SNOMEDCT, CSP}) [Mental
Process]
Meta Mapping (1000):
  1000 association (Relationship by association {MTH, AOD}) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships {MTH, RCD, SNOMEDCT, HL7V2.5, ICNP, NCI})
[Qualitative Concept]
```

Ilustración 39. Salida MetaMap2009v2 aplicando la opción -G en la evaluación de "Association".

-e (--exclude_sources) <list>

Usada para excluir de la lista de candidatos, todos aquellos conceptos que pertenezcan a las fuentes indicadas tras esta opción (las fuentes han de ir separadas por comas y sin espacio tras estas).

En este ejemplo se ha eliminado de la lista de candidatos, aquellos conceptos que pertenezcan a las fuentes "PSY" y "AOD". Desaparece el candidato "Associationism", pero no "association", puesto que este último pertenece a más de una fuente a la vez. Es decir, un candidato deja de aparecer si todas las fuentes a las que pertenece son excluidas.

```
Phrase: "Association"
Meta Candidates (6):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated (Associated with) [Qualitative Concept]
   928 Associate (Associate - relationship) [Idea or Concept,Professional or Occupational
Group]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 40. Salida MetaMap2009v2 aplicando la opción -e PSY,AOD en la evaluación de "Association".

-F (--formaltaggeroutput)

Muestra al comienzo del análisis de cada oración (utterance) la información de etiquetado que le es asignada a cada término del texto cuando es evaluada por el analizador SPECIALIST. En un fichero existirán tantos análisis de este tipo como oraciones contenga el texto completo.

```
Processing 00000000.tx.1: Association of maternal and fetal factors with development of mental deficiency.
[
 ['Association',noun],
 [of,prep],
 [maternal,adj],
 [and,conj],
 [fetal,adj],
 [factors,noun],
 [with,prep],
 [development,noun],
 [of,prep],
 [mental,adj],
 [deficiency,noun],
 ['.',pd]
]
…...(a continuación aparece la evaluación de cada concepto y sus candidatos)
```

Ilustración 41. Salida MetaMap2009v2 aplicando la opción -F en la evaluación de "Association".

-I (--show_cuis)

Muestra el número CUI (Identificador de Concepto Único) de cada uno de los conceptos, tanto en cada uno de los candidatos evaluados, como en los mapeos finales realizados.

```
Phrase: "Association"
Meta Candidates (7):
  1000 C0004083:Association (Mental association) [Mental Process]
  1000 C0439849:Association (Relationships) [Qualitative Concept]
  1000 C0596306:Association (Chemical Association) [Phenomenon or Process]
  1000 C0699792:association (Relationship by association) [Social Behavior]
   928 C0332281:Associated [Qualitative Concept]
   928 C0750490:Associate [Idea or Concept,Professional or Occupational Group]
   928 C0871654:Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 C0596306:Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 C0004083:Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 C0699792:association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 C0439849:Association (Relationships) [Qualitative Concept]
```

Ilustración 42. Salida MetaMap2009v2 aplicando la opción -I en la evaluación de "Association".

-j (--dump_aas)

Muestra los Acrónimos o Abreviaturas descubiertos por MetaMap en el texto, siguiendo esta estructura:

AA| PMID| Acrónimo/Abreviatura(AA)| Expansión| Nºpalabras aa| Nºcaracteres en aa| Nº palabras expansión| Nº caracteres expansión

Ilustración 43. Estructura análisis de AAs devuelta por MetaMap2009v2 tras usar la opción -j.

-J (--restrict_to_sts) <list>

Sólo son considerados como candidatos aquellos conceptos cuyo tipo o tipos semánticos estén especificados en la lista que se ha de introducir tras esta opción (los componentes de esta lista han de estar separados por comas).

El tipo semántico de un determinado concepto aparece entre corchetes a continuación del candidato al que hace referencia. Pero el tipo que se ha de poner en esta lista no es el nombre completo, tal y como aparece tras los candidatos, sino que cada tipo semántico esta representado por un conjunto de cuatro letras que se muestra en la Tabla 16.

En el ejemplo que mostramos para la cadena “Association”, vamos a restringir la salida de programa sólo a aquellos conceptos que pertenezcan al tipo semántico “Qualitative concepts”. Para lograr esta salida, la secuencia ejecutada ha de ser: “**-J qlco**”

```
Phrase: "Association"
Meta Candidates (2):
  1000 Association (Relationships) [Qualitative Concept]
   928 Associated [Qualitative Concept]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 44. Salida MetaMap2009v2 aplicando la opción -J qlco en la evaluación de “Association”.

-k (--exclude_sts) <list>

Esta opción realiza la acción inversa a la última vista, es decir, excluye de la lista de candidatos devuelta aquellos conceptos que pertenezcan al tipo semántico indicado. Si usamos el mismo ejemplo que en el caso anterior, al utilizar esta opción ninguno de los candidatos devueltos debería pertenecer al tipo “Qualitative Concept”.

```
Phrase: "Association"
Meta Candidates (5):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
```

Ilustración 45. Salida MetaMap2009v2 aplicando la opción -k qlco en la evaluación de "Association".

-m (--hide_mappings)

Sólo se muestra la lista de candidatos obtenidos para un determinado concepto, pero no así sus mapeos finales.

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
```

Ilustración 46. Salida MetaMap2009v2 aplicando la opción -m en la evaluación de "Association".

-n (--number_the_candidates)

Numera el conjunto de candidatos en función del número de estos que existan de forma ascendente. A continuación se muestra un ejemplo.

```
Phrase: "Association"
Meta Candidates (7):
   1.  1000 Association (Mental association) [Mental Process]
   2.  1000 Association (Relationships) [Qualitative Concept]
   3.  1000 Association (Chemical Association) [Phenomenon or Process]
   4.  1000 association (Relationship by association) [Social Behavior]
   5.   928 Associated [Qualitative Concept]
   6.   928 Associate [Idea or Concept,Professional or Occupational Group]
   7.   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 47. Salida MetaMap2009v2 aplicando la opción -n en la evaluación de "Association".

-N –fielded_mmi_output

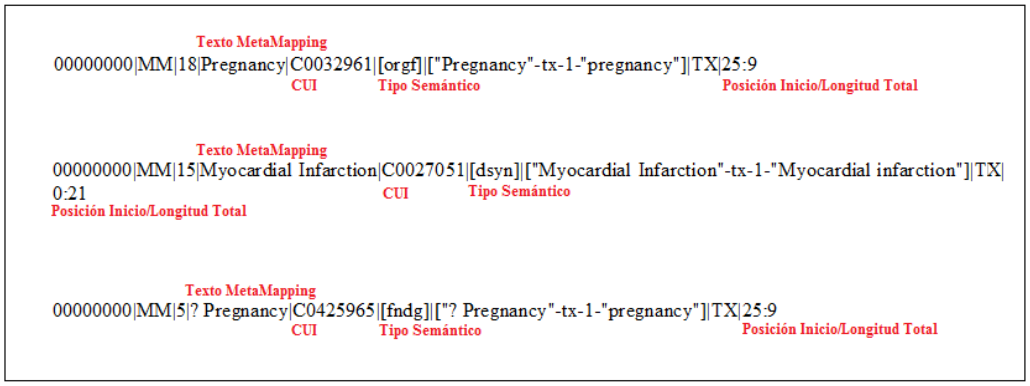
Devuelve una lista de todos los mapeos realizados, es decir, con los “Meta Mapping” y para cada uno de estos mapeos devuelve diversa información. A continuación se muestra en primer lugar la salida normal de MetaMap y tras esta la salida obtenida por la aplicación de esta opción, ante el texto de entrada “Myocardial infarction in pregnancy”.

```
Processing 00000000.tx.1: Myocardial infarction in pregnancy.

Phrase: "Myocardial infarction"
Meta Candidates (4):
  1000 Myocardial Infarction [Disease or Syndrome]
   861 Infarction [Pathologic Function]
   861 Myocardial (Myocardium) [Tissue]
   861 Myocardial [Spatial Concept]
Meta Mapping (1000):
  1000 Myocardial Infarction [Disease or Syndrome]

Phrase: "in pregnancy."
Meta Candidates (2):
  1000 Pregnancy [Organism Function]
  1000 ? Pregnancy [Finding]
Meta Mapping (1000):
  1000 ? Pregnancy [Finding]
Meta Mapping (1000):
  1000 Pregnancy [Organism Function]
```

Ilustración 48. Salida MetaMap2009v2 sin aplicar la opción -N en la evaluación del texto “Myocardial infarction in pregnancy”.



lustración 49. Salida MetaMap2009v2 tras aplicar la opción -N en la evaluación del texto “MyocardialInfarction in pregnancy”.

-O (--show_preferred_names_only)

Esta opción muestra en la salida sólo los nombres preferidos para un determinado concepto en el caso de tenerlo, si no lo tiene se devuelve el propio candidato.

Como ocurre en el ejemplo que ilustra este caso, el primer candidato "Association" tiene como nombre preferido "Mental association", por lo que al activar esta opción será devuelto dicho preferido. En cambio, el candidato "Associated" que no tiene un nombre preferido, será devuelto el mismo nombre del candidato.

```
Phrase: "Association"
Meta Candidates (7):
  1000 Mental association [Mental Process]
  1000 Relationships [Qualitative Concept]
  1000 Chemical Association [Phenomenon or Process]
  1000 Relationship by association [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Chemical Association [Phenomenon or Process]
Meta Mapping (1000):
  1000 Mental association [Mental Process]
Meta Mapping (1000):
  1000 Relationship by association [Social Behavior]
Meta Mapping (1000):
  1000 Relationships [Qualitative Concept]
```

Ilustración 50. Salida MetaMap2009v2 tras aplicar la opción -O en la evaluación del texto "Association".

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process]
  1000 Association (Relationships) [Qualitative Concept]
  1000 Association (Chemical Association) [Phenomenon or Process]
  1000 association (Relationship by association) [Social Behavior]
   928 Associated [Qualitative Concept]
   928 Associate [Idea or Concept,Professional or Occupational Group]
   928 Associationism [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept]
```

Ilustración 51. Salida MetaMap2009v2 sin aplicar la opción -O en la evaluación del texto "Association".

-p (--hide_plain_syntax)

La salida es mostrada sin indicar que palabra o palabras (que "phrase") corresponde con los candidatos y el mapeo que se muestra a continuación. Es decir, desaparece el término "**Phrase:** " de la salida de MetaMap y los "Meta Candidates" aparecen inmediatamente después del último "Meta Mapping" de la "Phrase" anterior.

```
Processing 00000000.tx.1: Myocardial infarction in pregnancy.
Meta Candidates (4):
  1000 Myocardial Infarction [Disease or Syndrome]
   861 Infarction [Pathologic Function]
   861 Myocardial (Myocardium) [Tissue]
   861 Myocardial [Spatial Concept]
Meta Mapping (1000):
  1000 Myocardial Infarction [Disease or Syndrome]
Meta Candidates (2):
  1000 Pregnancy [Organism Function]
  1000 ? Pregnancy [Finding]
Meta Mapping (1000):
  1000 ? Pregnancy [Finding]
Meta Mapping (1000):
  1000 Pregnancy [Organism Function]
```

Ilustración 52. Salida MetaMap2009v2 tras aplicar opción -p para la evaluación de "Myocardial Infarction in pregnancy".

```
Processing 00000000.tx.1: Myocardial infarction in pregnancy.

Phrase: "Myocardial infarction"
Meta Candidates (4):
  1000 Myocardial Infarction [Disease or Syndrome]
   861 Infarction [Pathologic Function]
   861 Myocardial (Myocardium) [Tissue]
   861 Myocardial [Spatial Concept]
Meta Mapping (1000):
  1000 Myocardial Infarction [Disease or Syndrome]

Phrase: "in pregnancy."
Meta Candidates (2):
  1000 Pregnancy [Organism Function]
  1000 ? Pregnancy [Finding]
Meta Mapping (1000):
  1000 ? Pregnancy [Finding]
Meta Mapping (1000):
  1000 Pregnancy [Organism Function]
```

Ilustración 53. Salida MetaMap2009v2 sin aplicar la opción -p en la evaluación del texto "Myocardial Infarction in pregnancy".

-q (--machine_output)

La salida devuelta por esta opción corresponde con la MMO (MetaMap Machine Output). Esta salida es el resultado ofrecido por MetaMap/SKR tras haber procesado una “cita” o cualquier otro elemento de texto libre, la cual proporciona una gran cantidad de información sobre los elementos de entrada. Cada frase nominal es desglosada y etiquetada usando el “etiquetador” MedPost/SKR que se encarga de identificar a los nombres, verbos, preposiciones, adjetivos, puntuación, localiza la idea principal o “header” de la frase. A continuación, se muestra una salida de este tipo ofrecida por MetaMap2009v2.

```
args('MetaMap -Z 09 -q /home/diana/PFC/DirectorioPruebasMed/13218007.txt
/home/diana/PFC/SalidaMetaMap/13218007.txt',[mm_data_year-'09',machine_output-[],infile-
'/home/diana/PFC/DirectorioPruebasMed/13218007.txt',outfile-'/home/diana/PFC/SalidaMetaMap/13218007.txt']).

aas([]).

neg_list([]).

utterance('00000000.tx.1',"Myocardial infarction in pregnancy.",0/35,[]).
phrase('Myocardial infarction',[head([lexmatch(['myocardial infarction']), inputmatch(['Myocardial',infarction]),
tag(noun), tokens([myocardial,infarction])])],0/21,[]).

candidates([ev(-1000,'C0027051','Myocardial Infarction','Myocardial
Infarction',[myocardial,infarction],[dsyn],[[[1,2],[1,2],0]],yes,no,['MEDLINEPLUS', 'MSH', 'MTH', 'NCI', 'NDFRT',
'ICPC2P', 'ICPC2ICD10ENG', 'CCPSS', 'COSTAR','CST', 'WHO', 'AIR', 'MDR', 'OMIM', 'RCD', 'SNM',
'SNOMEDCT', 'AOD', 'BI', 'CSP', 'PSY', 'SNMI', 'MTHICD9', 'DXP'],[0/21]),ev(-
861,'C0021308','Infarction','Infarction',[infarction],[patf],[[[2,2],[1,1],0]],yes,no,['LCH','MDR','MSH','MTH','NCI','ND
FRT','SNM','SNOMEDCT','SNMI','CCPSS','CST','CSP'],[11/10]),ev(-
861,'C0027061','Myocardial','Myocardium',[myocardial],[tisu],[[[1,1],[1,1],0]],yes,no, ['FMA', 'HL7V2.5',
'LNC','MSH','MTH','NCI','PSY','RCD','SNM','SNOMEDCT','UWDA','AOD','CSP','SNMI'],[0/10]),ev(-
861,'C1522564','Myocardial','Myocardial',[myocardial],[spco],[[[1,1],[1,1],0]],yes,no,['MTH','NCI'],[0/10])]).
mappings([map(-1000,[ev(-1000,'C0027051','Myocardial Infarction','Myocardial
Infarction',[myocardial,infarction],[dsyn],[[[1,2],[1,2],0]],yes,no,['MEDLINEPLUS','MSH','MTH','NCI','NDFRT','ICP
C2P','ICPC2ICD10ENG','CCPSS','COSTAR','CST','WHO','AIR','MDR','OMIM','RCD','SNM','SNOMEDCT','AOD','B
I','CSP','PSY','SNMI','MTHICD9','DXP'],[0/21])])]).
phrase('in pregnancy.',[prep([lexmatch([in]), inputmatch([in]), tag(prep), tokens([in])]), head([lexmatch([pregnancy]),
inputmatch([pregnancy]),tag(noun),tokens([pregnancy])]),punc([inputmatch(['.']),tokens([])])],22/13,[]).
candidates([ev(-
1000,'C0032961','Pregnancy','Pregnancy',[pregnancy],[orgf],[[[1,1],[1,1],0]],yes,no,['ICNP','ICPC2EENG','LCH','LNC'
,'MDR','MEDLINEPLUS','MSH','MTH','NCI','NDFRT','OMS','PSY','RCD','SNM','SNOMEDCT','CCPSS','COSTAR','
DXP','SNMI','WHO','AOD','BI','CSP','ICPC2ICD10ENG','MEDCIN','GO'],[25/9]),ev(-1000,'C0425965','?
Pregnancy','?
Pregnancy',[pregnancy],[fndg],[[[1,1],[1,1],0]],yes,no,['MTH','SNOMEDCT','ICPC2P','RCD','ICPC','ICPC2EENG','SN
MI','CCPSS','MEDCIN'],[25/9])]).
mappings([map(-1000,[ev(-1000,'C0425965','? Pregnancy','?
Pregnancy',[pregnancy],[fndg],[[[1,1],[1,1],0]],yes,no,['MTH','SNOMEDCT','ICPC2P','RCD','ICPC','ICPC2EENG','SN
MI','CCPSS','MEDCIN'],[25/9])]),map(-1000,[ev(-
1000,'C0032961','Pregnancy','Pregnancy',[pregnancy],[orgf],[[[1,1],[1,1],0]],yes,no,['ICNP','ICPC2EENG','LCH','LNC'
,'MDR','MEDLINEPLUS','MSH','MTH','NCI','NDFRT','OMS','PSY','RCD','SNM','SNOMEDCT','CCPSS','COSTAR','
DXP','SNMI','WHO','AOD','BI','CSP','ICPC2ICD10ENG','MEDCIN','GO'],[25/9])])]).
'EOU'.
```

Ilustración 54. MetaMap Machine Output, ofrecida por MetaMap2009v2 tras aplicar la opción -q en la evaluación del texto “Myocardial Infarction in pregnancy”.

-r (--threshold) <integer>

Limita la salida sólo a aquellos candidatos cuyo valor obtenido tras su evaluación frente al concepto original sea superior al umbral tomado en cada caso. Para ilustrar esta opción, de nuevo tomaremos un ejemplo, para el cual se restringirá la salida a aquellos candidatos que superen el valor "950", usando la secuencia "-r 950"

```
Phrase: "Abnormalities"
Meta Candidates (3):
  1000 Abnormalities (Congenital Abnormality) [Congenital Abnormality]
  1000 abnormalities (teratologic) [Functional Concept]
   966 Abnormality [Finding]
Meta Mapping (1000):
  1000 Abnormalities (Congenital Abnormality) [Congenital Abnormality]
Meta Mapping (1000):
  1000 abnormalities (teratologic) [Functional Concept]
```

Ilustración 55. Salida ofrecida por MetaMap2009v2 tras aplicar la opción -r 950 en la evaluación del texto "Association".

```
Phrase: "Abnormalities"
Meta Candidates (5):
  1000 Abnormalities (Congenital Abnormality) [Congenital Abnormality]
  1000 abnormalities (teratologic) [Functional Concept]
   966 Abnormality [Finding]
   916 Abnormal [Functional Concept,Idea or Concept]
   916 % Abnormal (Percent of Abnormal Cells) [Quantitative Concept]
Meta Mapping (1000):
  1000 Abnormalities (Congenital Abnormality) [Congenital Abnormality]
Meta Mapping (1000):
  1000 abnormalities (teratologic) [Functional Concept]
```

Ilustración 56. Salida ofrecida por MetaMap2009v2 sin aplicar la opción -r 950 en la evaluación del texto "Association".

-R (--restrict_to_sources) <list>

Restringe la salida de los candidatos, sólo a aquellos que pertenezcan a una de las fuentes seleccionadas en la lista que se ha de poner a continuación de la opción "-R". Para ilustrar esta opción, mostramos en la siguiente figura la salida obtenida si restringimos la salida a los candidatos que pertenezcan a la fuente "PSY", la segunda figura mostrada muestra las fuentes a las que pertenece cada candidato para verificar el correcto funcionamiento de esta opción.

```
Phrase: "Association"
Meta Candidates (1):
  928 Associationism [Idea or Concept]
Meta Mapping (928):
  928 Associationism [Idea or Concept]
```

Ilustración 57. Salida ofrecida por MetaMap2009v2 tras aplicar la opción -R PSY en la evaluación del texto "Association".

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association {MTH, MSH, RCD, SNM, SNOMEDCT, CSP}) [Mental Process]
  1000 Association (Relationships {MTH, RCD, SNOMEDCT, HL7V2.5, ICNP, NCI}) [Qualitative Concept]
  1000 Association (Chemical Association {MTH, NCI, CSP}) [Phenomenon or Process]
  1000 association (Relationship by association {MTH, AOD}) [Social Behavior]
  928 Associated {MTH, RCD, SNMI, SNOMEDCT, CCPSS, LNC, NCI} [Qualitative Concept]
  928 Associate {MTH, HL7V2.5} [Idea or Concept,Professional or Occupational Group]
  928 Associationism {PSY} [Idea or Concept]
Meta Mapping (1000):
  1000 Association (Chemical Association {MTH, NCI, CSP}) [Phenomenon or Process]
Meta Mapping (1000):
  1000 Association (Mental association {MTH, MSH, RCD, SNM, SNOMEDCT, CSP}) [Mental Process]
Meta Mapping (1000):
  1000 association (Relationship by association {MTH, AOD}) [Social Behavior]
Meta Mapping (1000):
  1000 Association (Relationships {MTH, RCD, SNOMEDCT, HL7V2.5, ICNP, NCI}) [Qualitative Concept]
```

Ilustración 58. Salida ofrecida por MetaMap2009v2 sin aplicar la opción -R PSY en la evaluación del texto "Association".

-s (--hide_semantic_types)

No muestra en la salida del programa el tipo semántico al que pertenece cada concepto. Por defecto, en la salida de MetaMap siempre va a aparecer tras cada candidato y cada mapeo el tipo semántico del concepto entre corchetes "[ ]", a menos que esta opción sea utilizada.

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association)
  1000 Association (Relationships)
  1000 Association (Chemical Association)
  1000 association (Relationship by association)
  928 Associated
  928 Associate
  928 Associationism
Meta Mapping (1000):
  1000 Association (Chemical Association)
Meta Mapping (1000):
  1000 Association (Mental association)
Meta Mapping (1000):
  1000 association (Relationship by association)
Meta Mapping (1000):
  1000 Association (Relationships)
```

Ilustración 59. Salida ofrecida por MetaMap2009v2 tras aplicar la opción -s en la evaluación del texto "Association".

-T (--tagger_output)

Muestra la salida de MetaMap de forma habitual, pero incorporando al inicio del análisis de cada sentencia una línea en la que aparece la oración y debajo de esta las etiquetas asignadas por el analizador MedPost/SKR.

```
Association of     maternal and  fetal factors with development of     mental deficiency.
noun           prep adj         conj  adj   noun   prep noun             prep adj      noun        pd
```

Ilustración 60. Salida ofrecida por MetaMap2009v2 tras aplicar la opción -T en la evaluación del texto "Association of maternal and fetal factors with development of mental deficiency".

-v (--variants)

Muestras las variantes generadas para cada uno de los conceptos, a partir del análisis y evaluación de estas, se obtiene la tabla de candidatos para cada concepto.

En este ejemplo se muestra una salida de este tipo, en la que en primer lugar se indica el número total de variantes obtenidas y a continuación, un listado de todas ellas, indicando en cada caso su función, su "distance variant" y su historial.

```
Phrase: "of maternal"

maternal [adj] variants (n=15):
mat{[verb], 7="dss"}  mate{[verb], 7="dss"}  mated{[verb], 7="dss"}  mater{[noun], 4="ss"}
materna{[noun], 3="d"}  maternal{[adj], 0=[]}  maternally{[adv], 3="d"}  maters{[noun], 4="ss"}
mates{[verb], 7="dss"}  mating{[verb], 7="dss"}  mats{[verb], 7="dss"}  matted{[verb], 7="dss"}
matting{[verb], 7="dss"}  mother{[noun], 2="s"}  mothers{[noun], 2="s"}

Meta Candidates (12):
  1000 Maternal [Clinical Attribute]
   944 Mother (Mothers) [Family Group]
   928 Materna [Organic Chemical,Pharmacologic Substance,Vitamin]
   916 MATER (NLRP5 gene) [Gene or Genome]
   893 Mats [Medical Device]
   893 mate (Partner in relationship) [Family Group]
   893 mate (Yerba mate) [Plant]
   893 MAT ([acyl-carrier-protein] S-malonyltransferase activity) [Molecular Function]
   893 mating [Organism Function]
   893 MAT (ACAT1 gene) [Gene or Genome]
   893 MAT (MAT1A gene) [Gene or Genome]
   893 Mate (Yerba mate preparation) [Pharmacologic Substance]
Meta Mapping (1000):
  1000 Maternal [Clinical Attribute]
```

Ilustración 61. Salida ofrecida por MetaMap2009v2 tras aplicar la opción -v en la evaluación del texto "Association".

-W (--preferred_name_sources)

Al igual que la opción ya vista "-G", esta va a mostrar las fuentes del Metatesauro UMLS, pero al contrario que la anteior opción que mostraba las fuentes a las que pertenecían los candidatos, esta mostrará a las que pertenecen los "nombres preferidos" para cada concepto.

En el caso de no existir un nombre de este tipo para un determinado candidato, será mostrada la fuente a la que pertenece este, como ocurría usando la función "-G".

```
Phrase: "Association"
Meta Candidates (7):
  1000 Association (Mental association) [Mental Process] [MTH]
  1000 Association (Relationships) [Qualitative Concept] [MTH,RCD,SNOMEDCT]
  1000 Association (Chemical Association) [Phenomenon or Process] [MTH,NCI]
  1000 association (Relationship by association) [Social Behavior] [MTH]
   928 Associated [Qualitative Concept] [MTH,RCD,SNMI,SNOMEDCT]
   928 Associate [Idea or Concept,Professional or Occupational Group] [MTH]
   928 Associationism [Idea or Concept] [PSY]
Meta Mapping (1000):
  1000 Association (Chemical Association) [Phenomenon or Process] [MTH,NCI]
Meta Mapping (1000):
  1000 Association (Mental association) [Mental Process] [MTH]
Meta Mapping (1000):
  1000 association (Relationship by association) [Social Behavior] [MTH]
Meta Mapping (1000):
  1000 Association (Relationships) [Qualitative Concept] [MTH,RCD,SNOMEDCT]
```

Ilustración 62. Salida ofrecida por MetaMap2009v2 tras aplicar la opción -W en la evaluación del texto "Association".

-x (--syntax)

Muestra la salida de MetaMap de la forma habitual, añadiendo antes del conjunto de candidatos información sobre el análisis sintáctico realizado por el analizador SPECIALIST.

```
Phrase: "in the prenatal"

msu
  prep([lexmatch([in]),inputmatch([in]),tag(prep),tokens([in])])

  det([lexmatch([the]),inputmatch([the]),tag(det),tokens([the])])

  head([lexmatch([prenatal]),inputmatch([prenatal]),tag(adj),tokens([prenatal])])

Meta Candidates (2):
  1000 Prenatal (Prenatal care) [Health Care Activity]
  1000 prenatal [Age Group]
Meta Mapping (1000):
  1000 Prenatal (Prenatal care) [Health Care Activity]
Meta Mapping (1000):
  1000 prenatal [Age Group]
```

Ilustración 63. Salida ofrecida por MetaMap2009v2 tras aplicar la opción -x en la evaluación del texto "in the prenatal".

-X (--truncate_candidates_mappings)

Muestra en la salida los cien primeros candidatos que forman el conjunto de estos. Ahora bien, normalmente el número de candidatos devueltos para un determinado concepto no es tan elevado, por eso este tipo de opciones son útiles para aquellos casos en los que la salida tras aplicar alguna otra opción, devuelve un número de candidatos muy elevado.

Por ejemplo, si se ha ejecutado MetaMap2009v2 para que nos devuelva como candidatos todas aquellas cadenas encontradas en el Metatesauro que contengan en alguno de sus extremos el concepto deseado ("-o"), esta opción hace que la cantidad de candidatos aumente significativamente, por lo que si se utiliza "-X" el número de candidatos se reducirá a los cien primeros.

-% (--XML) <option>

Genera una salida en formato XML que estudiaremos más en detalle en el siguiente capítulo. Para esta salida podemos seleccionar uno de los siguientes formatos: format, format1, noformat, noformat1.

--negex

Muestra al comienzo del fichero de salida de MetaMap una lista con los conceptos UMLS negados encontrados y las cadenas asociadas que han provocado tal negación, así como un análisis de las mismas.

2.2.4.4. Incompatibilidad entre opciones.

MetaMap 2009 v2 nos permite ejecutar varias opciones al mismo tiempo, pero existen diversos conjuntos de opciones que no pueden ser usados en una misma salida de programa. Los casos principales son:

- Tagger Output (-T) con Formal Tagger Output (-F). Estas opciones no pueden ser usadas conjuntamente ya que ambas devuelven el mismo tipo de salida, un análisis sintáctico de cada palabra de la oración, pero con distinto formato.

Mientras que la primera presenta cada palabra y su función sintáctica en una línea, la segunda devuelve toda la oración en la misma línea y debajo de esta en otra línea el análisis sintáctico.

- Hide Candidates (-c) con Number Candidates (-n). En este caso dado que la primera opción deshabilita que sean mostrados en la salida los conceptos UMLS candidatos, no será posible usar la segunda puesto que la función de esta es numerar la lista de candidatos obtenida.

- No Derivational Variants (-d) con All Derivational Variants (-D). Estas opciones son excluyentes, ya que seleccionando la primera se evita el uso de variantes derivadas, mientras que la segunda permite el uso de algunos tipos de estas variantes.

- Allow Acronym/Abbreviation Variants (-a) con Unique Acronym/Abbreviation Variants Only (-u). La primera opción permite incluir todo tipo de abreviaturas y acrónimos en el conjunto de candidatos, mientras que la segunda restringe el uso a abreviaturas y acrónimos que sólo presenten una expansión, por lo que ambas opciones no podrían aparecer juntas, ya que supondría una contradicción.

- Restrict to Sources (-R) con Exclude Sources (-e). Estas opciones son excluyentes una de la otra, puesto que realizan tareas completamente opuestas e incompatibles, mientras que la primera restringe el uso a las fuentes de vocabularios señalados, la segunda excluye los vocabularios que se le son indicados.

- Restrict to Semantic Type(s) (-J) con Exclude Semantic Type(s) (-k). De la misma forma que en el caso anterior, estas opciones son excluyentes, puesto que la primera sólo devuelve candidatos cuyo tipo semántico se especifique y la segunda por el contrario excluye los tipos semánticos indicados.

Capítulo 3
Nuevas funciones y mejoras en MetaMap2009v2

En el diseño de la nueva versión de la herramienta, MetaMap2009v2, se pueden apreciar diversos cambios que implican tanto la introducción de mejoras como la corrección de errores en versiones pasadas.

Una de las mejoras más significativas es la nueva implementación de NegEx, así como la incorporación de otros importantes cambios: la aparición de un nuevo método de “Sentence-Breaking”, la consolidación de MatchMap y la evolución en la generación de una salida en formato XML.

Todas estas variaciones se analizarán a continuación y se realizará un estudio de su evolución desde la implementación inicial y a través de las distintas versiones, hasta llegar a la que nos ocupa en este proyecto.

3.1. Algoritmo Sentence-Breaking

La implementación inicial de este algoritmo aparece en la versión MetaMap2008v2, hasta entonces, en versiones previas para indicar el final de una frase era necesario un espacio en blanco al final de cada oración, pero durante el desarrollo de esta versión, se observó que en múltiples citas de MedLine, los finales de frase no estaban marcados con un espacio en blanco, por lo que el programa no controlaría casos como los que se muestran a continuación:

serum markers for each of those subgroups.The mathematical simulation

neurogenesis in the adult nervous system.These findings may have

informed choices.Patients who are ready to make changes must be provided

affiliation with a trade union.Although still shut-out by the general

Dado que este espacio en blanco, indicador de fin de oración, deja de aparecer en los textos de la biblioteca MedLine, se hace necesaria la aparición de un nuevo método para identificar el final de oración, este consistirá en observar si una frase va seguida de una nueva cadena que comienza por una letra mayúscula, sí es así, esta se considerará como otra oración diferente.

Es durante el desarrollo de la versión de MetaMap2008v2, cuando tendrá lugar la creación de un nuevo algoritmo que divida las oraciones del texto de entrada en diversas frases ante la aparición de ciertos caracteres.

En esta versión inicial la lista de caracteres estaba compuesta por:“<”, “>” y “=”. Cabe mencionar, que este cambio fue motivado por la aparición del siguiente texto, el cual claramente no debía ser analizado como una sola frase:

yeast extract > peptone > magnesium
sulfate > vitamin C = potassium phosphate > calcium chloride = ammonium sulfate.

En la implementación de este algoritmo se añade la habilidad de mostrar información sobre las frases en que es dividida la entrada, esta salida se mostraba

mediante la opción **--debug phrase**. Concretamente la salida consistía en la palabra **Phrase** seguida del **PMID** (número de identificación del texto en MedLine), **Section Type** (ti (título) o ab (resumen)), **utterance number** (posición de la frase en el texto), el n**úmero de "nombres"** que tiene dicha frase y finalmente las **palabras que la componen,** estos campos aparecían separados por la una barra vertical "|":

```
Phrase | 16755936 | ab |4 | 3 | [of,the,individual,nutrient,component]
```

En la siguiente versión, MetaMap2009, además de los caracteres vistos hasta ahora, se añaden: "**;**", "**.**", "**!**" y "**?**". Pero es importar destacar, que estos últimos, serán considerados límites de frase, sólo si no están precedidos por un espacio en blanco.

Finalmente, en la nueva versión de MetaMap2009v2 y la que nos ocupa en la realización de este proyecto, se añade un nuevo carácter a la lista ya existente, "*****".

Este nuevo tipo aparece porque en el caso de listas como las que se muestran a continuación, si no es incluido, todas las frases serían consideradas como una sola, es decir, al no incorporar "*" a la lista de caracteres de partición de frases, este texto sería considerado como una cadena larga que necesitaría unos 45 minutos para su procesamiento total debido a la masiva necesidad de combinación de posibilidades, mientras que si el asterisco es tratado como un carácter de troceado, MetaMap podría procesar todo el texto en apenas unos pocos segundos.

```
* Fermentation technology
* Enzyme technology
* Recombinant DNA technology
* Cell culture technology
* Ex vivo and therapeutic stem cell culture technology
* Metabolic engineering
* Tissue engineering
* Nanobiotechnology
* Quantitative systems biotechnology
* Biosensor development
* Food processing with special focus on the safety of the nation's food supply
```

Como conclusión a este apartado, podemos decir que la versión de MetaMap2009v2, presenta como mejora ante versiones anteriores, el incremento del número de caracteres de ruptura de frase a ocho, estos son: "<", ">", "=", ";", ".", "!" , "?" y "*", aumentando así la eficiencia de la herramienta, debido al aumento la velocidad de procesamiento de algunos tipos de textos.

3.2. MatchMap

Otra de las mejoras introducidas por esta nueva versión es la referida a la consolidación de MatchMap. Esta estructura de datos se comienza a incluir en MetaMap 2008, pero no será hasta la versión actual cuando esta funcionalidad adquiera mayor relevancia.

MatchMap es una lista que ofrece un conjunto de datos informativos que muestran la relación entre las palabras del texto de entrada original y sus respectivos candidatos. Pero para comprender mejor su funcionamiento, estudiaremos dos casos significativos, en primer lugar veremos un ejemplo en el cual un texto de entrada es mapeado usando sólo uno de los conceptos candidatos, mientras que en el segundo para mapear completamente la entrada será necesario el uso de varios conceptos candidatos, uno para cada segmento que conforma el texto completo de entrada.

Para ilustrar el primer caso tomaremos como ejemplo el texto "*obstructive sleep apnea*" y su candidato "*sleep apnea*". Como ya habíamos dicho, en este tipo de situaciones sólo es necesario un candidato para el mapeo completo de todo el texto de entrada, por lo que su salida MatchMap sería [[[2,3],[1,2],0]], esto es, las posiciones de las palabras mapeadas en el texto de entrada, seguidas de las posiciones que ocupan las mismas en el candidato, y finalmente la variación léxica entre ambos textos.

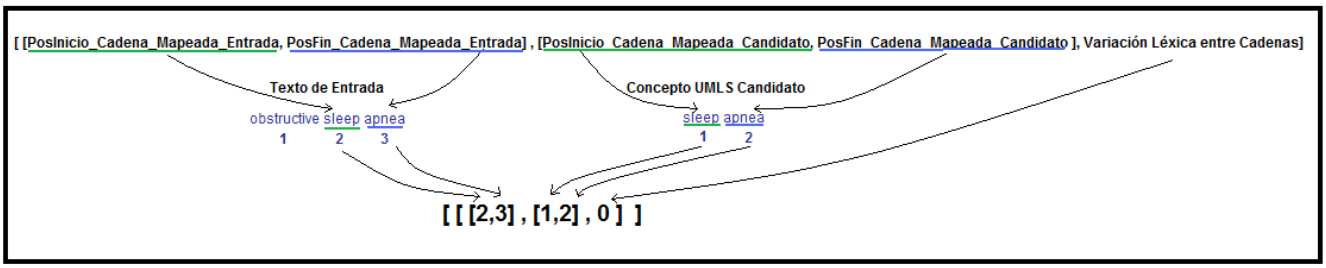


Ilustración 64. Lista MatchMap ofrecida por MetaMap2009v2 tras evaluar el texto "obstructive sleep apnea" y su concepto candidato UMLS "sleep apnea".

El segundo caso que podemos encontrar, es en el que para llevar a cabo el mapeo completo de un texto es necesaria la combinación de varios candidatos, esto ocurre con el texto “protein synthesis”, cuyo mapeo completo usa el candidato de la palabra protein, “protein”, y el candidato de la palabra synthesis, “synthesizers”.

En estos casos, MatchMap nos ofrecerá una relación para cada una de las palabras mapeadas, [[[1,1], [1,1], 0], [[2,2], [2,2], 6]], es decir, nos devuelve las posiciones de inicio y fin del segmento de texto mapeado en cada ocasión, seguido de las mismas posiciones esta vez en la cadena correspondiente a su concepto candidato, tras cada una de las correspondencias podemos observar la variación léxica en ambas ocasiones entre el concepto original y el candidato ofrecido por el Metathesaurus UMLS.

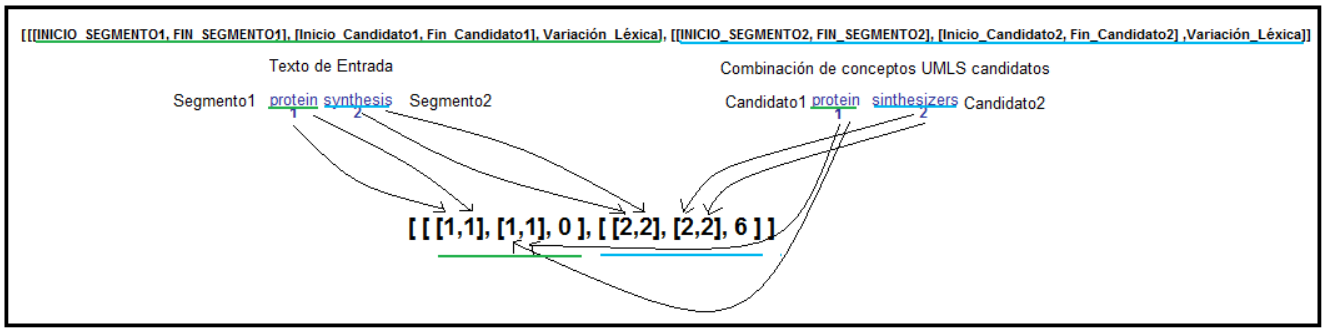


Ilustración 65. Lista MatchMap ofrecida por MetaMap2009v2 tras evaluar el texto “protein síntesis” y sus conceptos candidatos UMLS “protein” y “syntehsizers”.

La consolidación de MatchMap en la nueva versión de MetaMap, consiste en comprobar si los segmentos del texto original de entrada, así como los candidatos de cada uno de estos segmentos en su agrupación final son adyacentes, sí es así, MatchMap no devuelve una lista para cada segmento como se presenta en la ilustración anterior, si no que es devuelta una sola lista en la que se indica la posición de comienzo del primer segmento y la de fin del segundo, seguidas de la posición de inicio del primer candidato y la que indica el final del segundo. La variación léxica total en esta nueva lista MatchMap vendrá determinada en función del valor de las variaciones obtenidas independientemente, es decir, se ha de hacer la media ponderada de todas variaciones.

Para el ejemplo visto en la imagen anterior esta nueva funcionalidad es aplicada puesto que tanto los segmentos mapeados del texto de entrada, como los conceptos UMLS candidatos que corresponden a cada segmento son adyacentes. La lista MatchMap final que resultaría es, [[[1,2], [1,2], 3]].

3.3. NegEx

Una nueva funcionalidad que cabe destacar en el estudio de MetaMap2009v2, es NegEx, algoritmo creado por Wendy Champman (23).

Anteriormente, NegEx aparecía sólo en la salida MMO (MetaMap Machine Output), pero a partir de la versión de MetaMap2009v2 se crea una opción específica para invocar esta funcionalidad (**--negex**), la cual si es solicitada por el usuario devuelve en la parte superior del fichero de salida, un análisis de las negaciones encontradas en **una cita completa**.

Este algoritmo puede recibir como entrada una oración cualquiera, pero sólo devuelve una salida si alguno de los términos que la componen aparece explícitamente negado o se menciona como una posibilidad hipotética.

Para localizar las negaciones en el texto, NegEx va a identificar entre una de las dos posibles expresiones regulares que pueden aparecer en el documento de entrada. Estas expresiones pueden venir dadas por tres tipos distintos de frases negativas.

3.3.1. Expresiones Regulares y "Frases Negativas"

Las expresiones regulares, consideradas como negaciones, siguen alguna de estas dos estructuras, donde el asterisco puede ser tanto una palabra como una frase del Metathesaurus UMLS.

Regular Expression 1: <negation phrase> * <indexed term>

Regular Expression 2: <indexed term> * <negation phrase>

Existen tres tipos de “frases negativas”, que pueden ser identificadas por NegEx, para saber de que tipo se trata, se ha creado un conjunto de listas en las cuales estas frases aparecen agrupadas por tipos: oraciones Pseudo-Negativas, oraciones negativas Pre-UMLs y oraciones negativas Post-UMLS:

1. Oraciones Pseudo-Negativas, frases que parecen negaciones, pero que realmente no lo son. Si una oración de este tipo es encontrada, NegEx automáticamente la omite y salta a la siguiente frase negativa que encuentre. Algunas de estas son: “no increase”, “no suspicious change”, “no significant change”, “no change”,....

2. Oraciones negativas Pre-UMLS, en estas el término que hace negativa la oración aparece antes de la palabra o frase UMLS, son las usadas en la Regular Expression 1. Este tipo de oraciones, a su vez, se encuentra subdividida en dos tipos: frases que niegan una oración (como “absence of”, “cannot”, “declined”, “denied”,... entre otras) y frases que la modelan como una posibilidad condicionada (“rule him out”, “what must be ruled out is”, “it to be ruled out for”,...).

3. Oraciones Post-UMLS, al contrario que en caso anterior, en estas el término que niega la oración aparece a continuación del contenido UMLS, corresponden a la Regular Expression 2. Este tipo también esta subdividido en diversas categorías, al igual que el punto 2 encontramos frases que niegan la oración y frases que la formulan como una posibilidad, pero además en este caso, también encontramos frases o conjunciones que disminuyen el alcance de la frase (“but”, “however”, “yet”, “still”, “except”, “secondary to”... entre otras).

3.3.2. Algoritmo de Funcionamiento de NegEx

El funcionamiento de este algoritmo sigue los pasos que describimos a continuación:

1. Para cada una de las oraciones de entrada, NegEx localiza todas las “frases negativas” que hay en ellas, a partir de las listas mencionadas más arriba.

2. Tras este proceso de identificación, el algoritmo toma cada una de las frases negativas localizadas y en función del tipo al que pertenezcan realizan una acción u otra:

2.1. Si se trata de una Pseudo-Negativa, esta se omite y se pasa a analizar la siguiente.

2.2. En el caso de las frases tipo Pre-UMLS, se crea una "ventana o hueco" con capacidad para seis palabras tras la frase negativa analizada en ese momento. Cabe destacar que una palabra o frase UMLS, es considerada una sola palabra de las seis mencionadas como capacidad.

2.3. Finalmente, si se trata de un tipo Post-UMLS, se realiza la misma acción que con las del tipo anterior, salvo que la "ventana" aparece delante de la frase de negación.

Ahora bien, la ventana que se crea puede contener en su interior más "frases negativas", si esto ocurre, la capacidad de esta ventana disminuirá su tamaño inmediatamente. Esta variación de la ventana tendrá lugar tanto si lo que se encuentra es una frase negativa como si es una conjunción. Veamos más en detalle a continuación este proceso.

a) Si otra "frase negativa" aparece en el interior de la ventana marcada, su tamaño decrece, poniendo como fin la palabra que precede a la negación encontrada.

Observemos este ejemplo, en la oración "**The patient denies [chest pain and has no shortness of breath]**" se tienen dos frases negativas "**denies**" y "**no**", donde la ventana inicial es la contenida entre los corchetes.

Dado que la segunda negación aparece en el interior de la ventana inicial, se disminuye el tamaño de esta variando su posición de fin hasta la palabra que precede a la segunda negación encontrada, la nueva ventana quedaría marcada por los corchetes de la siguiente forma "**The patient denies [chest pain and has] no shortness of breath**".

Finalmente, este algoritmo marcará "**chest pain**", como concepto negado por "**denies**" y "**shortness of breath**" lo será por "**no**".

b) Por otro lado, si dentro de la ventana inicial aparece una conjunción de las descritas en la lista correspondiente, el tamaño de la ventana disminuirá y su final lo marcará la palabra que se encuentre justo antes de la conjunción.

La siguiente oración nos sirve de ejemplo para ilustrar el caso, "**The patient denies [chest pain but has experienced some shortness of breath]**", dónde encontramos como frase negativa "**denies**" y como conjunción "**but**", por lo que de acuerdo con el algoritmo NegEx, "**chest pain**" queda negado por "**denies**", pero en este caso "**shortness of breath**" no es negado por "**but**", aunque si queda modificado su alcance.

Tras marcar todos los términos UMLS contenidos en una misma ventana, como negados o posibles, si pertenecen a alguno de estos dos tipos, se repite de nuevo el paso número dos del algoritmo, hasta haber analizado todas las "frases negativas" de una oración. Y de nuevo se comienza a aplicar el algoritmo desde el primer paso hasta acabar con todas las oraciones del texto de entrada que nos ocupa.

3.3.3. Análisis Salida NegEx

En este último apartado, estudiaremos la estructura general de la salida ofrecida por el algoritmo cuando se localiza una negación y los datos que se extraen

de su análisis. Finalmente, se mostrará un ejemplo de la salida devuelta por MetaMap2009v2 al seleccionar la opción –**negex.**

Cuando llevamos a cabo un procesado de textos aplicando esta opción, en la parte superior del fichero de salida de MetaMap, aparece la palabra "NEGATIONS:", y en las líneas siguientes, pero sólo en las citas que contengan algún tipo de negación, se realizará una evaluación de cada negación. Los elementos que forman este bloque de salida se detallan a continuación, junto con una figura con la que se ilustra esta estructura, cabe mencionar que en una salida existirán tantos bloque de análisis de negación, como frases negativas sean localizadas.

-Negation Type: Indica el tipo de negación extraído del texto, existen cinco tipos posibles en función de la lista de negaciones a la que pertenezcan.

-nega: Frases negativas, cuyo uso convierten un término en negativo. Aparecen en oraciones Pre-UMLS.

-pnega: Frases de posibilidad condicionada, es decir, su uso señala un término como posible. Aparecen en oraciones de tipo Pre-UMLS.

-negb: Frases negativas, hacen negativo el significado de un término. Usadas en oraciones Post-UMLS, es decir, aparecen tras el término que es negado.

-pnegb: Frases de posibilidad condicionada, señalan un término como posible. Aparecen en oraciones Post-UMLS.

-**pseudoneg**: Frases pseudo negativas.

-Negation Trigger: Palabra o grupo de palabras, que pertenece a uno de los cinco tipos que hemos señalado en el apartado anterior, que constituyen la frase negativa.

-Negation PosInfo: Posición que ocupa el primer carácter de la frase negativa en la oración que está siendo evaluada, se considera como posición

"0" al primer carácter de la oración. Tras esta posición, se indica la longitud total de la frase negativa.

-Negated Concept: Expone el concepto que queda negado por la frase negativa, así como su Identificador de Concepto Único (CUI).

-Concept PosInfo: Posición en la oración procesada en la que aparece el primer carácter el término que queda negado junto con la longitud total de este.

Tras haber analizado la función de cada una de las partes que constituyen esta salida, en la siguiente figura se muestra la estructura en la que se muestra el análisis realizado a cada una de las negaciones localizadas en un mismo texto.

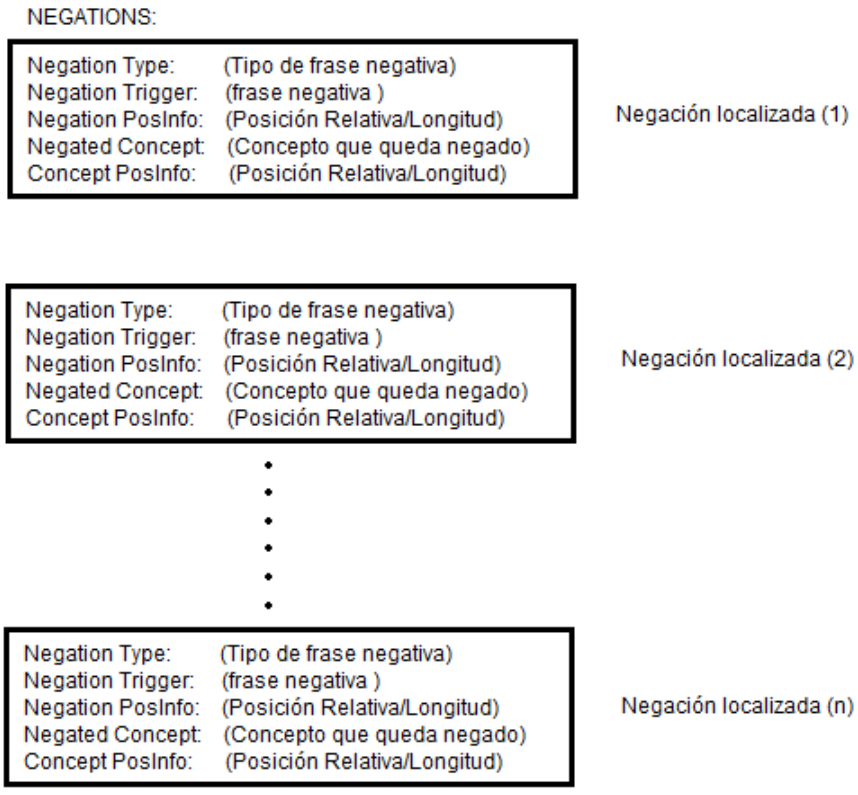


Ilustración 66. Estructura de salida para el análisis de n negaciones localizadas tras ser procesado un determinado texto con el algoritmo NegEx.

Tras este análisis de cada una de las partes de la salida ofrecida por NegEx, se muestra un ejemplo de la salida devuelta por MetaMap2009v2 al seleccionar la opción **--negex**:

```
NEGATIONS:
Negation Type:     nega
Negation Trigger:  no
Negation PosInfo:  0/2
Negated  Concept:  C0032285:Pneumonia
Concept  PosInfo:  3/10

Negation Type:     nega
Negation Trigger:  checked for
Negation PosInfo:  14/11
Negated  Concept:  C0332448:Infiltrates
Concept  PosInfo:  26/12


Processing 00000000.tx.1: No pneumonia.

Phrase: "No pneumonia."
Meta Candidates (3):
  1000 Pneumonia [Disease or Syndrome]
   907 Lung [Body Part, Organ, or Organ Component]
   907 Lung (Entire lung) [Body Part, Organ, or Organ Component]
Meta Mapping (1000):
  1000 Pneumonia [Disease or Syndrome]
Processing 00000000.tx.2: Checked for infiltrates.

Phrase: "Checked"
Meta Candidates (1):
   966 Check (Checking) [Qualitative Concept]
Meta Mapping (966):
   966 Check (Checking) [Qualitative Concept]

Phrase: "for infiltrates."
Meta Candidates (3):
  1000 Infiltrates (Infiltration) [Pathologic Function,Therapeutic or Preventive Procedure]
   966 Infiltrate [Intellectual Product]
   966 Infiltrate (Administration Method - Infiltrate) [Functional Concept]
Meta Mapping (1000):
  1000 Infiltrates (Infiltration) [Pathologic Function,Therapeutic or Preventive Procedure]
```

Ilustración 67. Salida obtenida tras ejecutar la opción –negex en MetaMap2009v2 para la evaluación del texto “no pneumonia”.

Una de las utilidades más destacadas de la salida NegEx, es la que da la Machine Output, en concreto la salida XML, que se estudiará en el apartado siguiente.

3.4. Salida XML

MetaMap2009v2 incluye una salida XML, ya implantada desde la versión de MetaMap2008. Pero en el estudio de la actual versión, se ha observado un cambio considerable en el nombre, cantidad y tipo de etiquetas usadas para la salida, con el fin de lograr que esta sea más legible y fácil de comprender, pero no así un cambio en la estructura de la misma.

En la primera versión de esta funcionalidad, en MetaMap2008, se generaba un fichero con una sola salida XML para cada archivo de entrada, sin tener en cuenta si en cada uno de estos ficheros había una sola cita (un documento biomédico) o varias, por lo que uno de los cambios más importantes en esta salida es la generación de una salida XML para cada cita encontrada en un mismo fichero. MetaMap introduce una línea en blanco para diferenciar cada salida XML generada en un fichero, si existe más de una cita en un documento. Es importante notar que MetaMap interpreta como citas independientes a aquellos textos que aparecen separados por una línea en blanco.

Esta salida es invocada mediante la etiqueta **-%** o **--XML**, seguida de una de las opciones de formato proporcionada: **format**, **noformat, format1 o noformat1**. Las dos primeras aparecen desde la primera versión de esta función, donde **format** ofrece una salida fácil de "comprender", puesto que cada uno de los componentes de la salida XML aparece en una línea distinta, mientras que la opción **noformat** muestra una salida XML sin espacios en blanco ni retornos de carro. Si usamos este último formato se ahorra en torno a un 50% de espacio en memoria, con respecto a la opción format.

En esta última versión, se incorporan dos nuevas opciones de formato, **format1** y **noformat1**, que ofrecen la misma estructura de salida que **format** y **noformat** respectivamente, pero que a diferencia de estos, generan un elemento <MMO> dentro de la <MMOs>, en vez de un <MMOs> para cada cita diferente encontrada en un mismo documento.

Para mostrar gráficamente la diferencia principal existente entre estos dos formatos, se muestran las siguientes figuras. La primera figura ilustra la salida

obtenida si se usa el formato tipo **format1,** mientras que en la segunda aparece la salida XML usando el otro formato proporcionado **format,** ambas salidas son resultado de procesar un documento que contiene varias citas.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE MMO PUBLIC "-//NLM//DTD MetaMap Machine Output//EN"
                "http://ii-public.nlm.nih.gov/DTD/MMOtoXML_v2.dtd">
<MMOlist>
 <MMO>
   . . . salida XML para cita(1). . . .
 </MMO>
 <MMO>
   . . . salida XML para cita(2). . . .
 </MMO>
  .....
  .....
 <MMO>
   . . . salida XML para cita(n). . . .
 </MMO>
</MMOlist>
```

Ilustración 68. Salida XML obtenida tras ejecutar la opción "–% format1" en MetaMap2009v2 para la evaluación de un documento con n citas en su contenido.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE MMO PUBLIC "-//NLM//DTD MetaMap Machine Output//EN"
                "http://ii-public.nlm.nih.gov/DTD/MMOtoXML_v2.dtd">
<MMOlist>
 <MMO>

   . . . salida XML para cita(1). . . .

 </MMO>
</MMOlist>

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE MMO PUBLIC "-//NLM//DTD MetaMap Machine Output//EN"
                "http://ii-public.nlm.nih.gov/DTD/MMOtoXML_v2.dtd">
<MMOlist>
 <MMO>

   . . . salida XML para cita(2). . . .

 </MMO>
</MMOlist>
.....
.....

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE MMO PUBLIC "-//NLM//DTD MetaMap Machine Output//EN"
                "http://ii-public.nlm.nih.gov/DTD/MMOtoXML_v2.dtd">
<MMOlist>
 <MMO>

   . . . salida XML para cita(n). . . .

 </MMO>
</MMOlist>
```

Ilustración 69. Salida XML obtenida tras ejecutar la opción "–% format" en MetaMap2009v2 para la evaluación de un documento con n citas en su contenido.

A continuación realizaremos, siguiendo una estructura jerárquica, una comparativa entre las etiquetas XML en la versión anterior y la actual, exponiendo

claramente el cambio de nombres y su uso en la presente versión.

Etiqueta MetaMap2009v2	Etiqueta anterior	Descripción
<MMOs> **<MMO>**	**<MMOlist>** **<MMO>**	MMO (MetaMap Machine Output), esta etiqueta marca el principio y el fin de cada salida XML generada por una entrada de texto. Toda salida Xml está estructura en cuatro partes fundamentales: **- <CmdLine>** **- <AA>** **- <Negation>** **- <Utterance>**
<CmdLine>	**<Args>**	En este apartado se describen todos los datos usados en la línea de comandos para obtener un tipo de evaluación u otra, esta descripción consta de los siguientes atributos: **- <Command>** **- <Option>** Se aprecia un cambio en el nombre de la etiqueta que hace más comprensible su función en la salida XML.
<Command>	**<Command>**	Indica las opciones de datos, opciones de procesado de documentos y opciones de salida, introducidas por el usuario, y por defecto, aplicadas en una determinada ejecución de MetaMap. ***MetaMap -Z 09 -% format RUTA_FICHERO_ENTRADA.txt RUTA_FICHERO_SALIDA_XML.txt*** En esta secuencia se muestra la versión de MetaMap usada, las opciones -Z y -%. y la ruta del fichero de origen del texto y de destino de la salida.
<Options> **<Option>**	**<Options>** **<Option>**	Se trata de un campo no obligatorio para el correcto funcionamiento del programa, es decir, si no se introduce ninguna opción, MetaMap procesará y emitirá una salida de acuerdo con los parámetros por defecto. Existirán tantas etiquetas <option> en la salida en formato XML, como opciones haya en la línea de comandos, dentro de cada una de estas etiquetas se realiza un análisis completo de la opción **- <OptName>** **- <OptValue>**
<OptName>	**<OptName>**	Se muestra el nombre completo de la opción usada, por ejemplo, en “-% format”,

		el valor de esta etiqueta sería “XML”, o en el caso de la ruta del fichero de entrada o de salida, el valor de <OptName> sería “Infile” o “Outfile” respectivamente.
<OptValue>	**<OptValue>**	Indica el valor del nombre de las opciones expuestos anteriormente, es decir, en el caso de las rutas de los ficheros de salida y entrada, el valor de esta etiqueta contendrá la ruta completa de acceso a los archivos.
<AAs> **<AA>**	**<AAs>** **<AA>**	Aparecen detalladas todas las Abreviaturas y Acrónimos(AAs) encontrados tras el análisis completo del texto. En el inicio de esta sección hay un contador en el que se indica el número de AAs totales. Esta sección consta de: **- <AAText>** **- <AAExp>** **- <AATokenNum>** **- <AALen>** **- <AAExpTokenNum>** **- <AAExpLen>** **- <AACUI>**
<AAText>	**<AAName>**	El acrónimo o abreviatura (AA) encontrada en el texto de entrada.
<AAExp>	**<AAExpansion>**	La expansión de las siglas o la abreviatura detectada en el análisis del texto.
<AATokenNum>	**<NumAATokens>**	Número de palabras que contiene cada uno de los AAs encontrados.
<AALen>	**<AALen>**	Longitud de la cadena de texto que representa al acrónimo o abreviatura en cuestión.
<AAExpTokenNum>	**<NumDefnTokens>**	Número de palabras que forman la expansión de nuestra AA. (Los espacios en blanco equivalen a una palabra).
<AAExpLen>	**<DefnLen>**	Cantidad de caracteres existentes en la expansión de AA. (Los espacios en blanco se contabilizan como un carácter).
<AACUIs> **<AACUI>**	**<CUIs>** **<CUI>**	El Identificador de Concepto Único (CUI) asociado con la expansión de AA.
<Negations> **<Negation>**	**<Negations>** **<Negation>**	En esta sección, se muestran todos los datos relacionados con la negación

		encontrada en el estudio de un documento. Las partes de las que consta son: **- <NegType>** **- <NegTrigger>** **- <NegTriggerPI>** **- <NegConcept>** **- <NegConcPI>**
<NegType>	**<NegType>**	Tipo de negación, esta puede ser: pseudoneg, nega, pnega, negb, pneg.
<NegTrigger>	**<NegTrigger>**	Palabra o palabras identificadas como negación.
<NegTriggerPIs> **<NegTriggerPI>**	**<NTSpans>** **<NTSpan>**	Información sobre la palabras o palabras que niegan la oración. Muestra la posición de comienzo y la longitud de la secuencia de palabras de negación de la siguiente forma: **Posición_Inicio/Longitud_Total**
<NegConcepts> **<NegConcept>**	**<CUIConcepts>** **<CUIConcept>**	Sección dedicada al concepto que es negado por la secuencia de palabras identificada en los apartados anteriores. Se muestran: **- <NegConcCUI>** - <**NegConcMatched**>
<NegConcCUI>	**<NegExCUI>**	Identificador de Concepto Único asociado a la palabra negada.
<NegConcMatched>	**<NegExConcept>**	Concepto negado.
<NegConcPIs> **<NegConcPI>**	**<NCSpans>** **<NCSpan>**	Posición de inicio de la palabra negada y longitud total de esta palabra.
<Utterances> **<Utterance>**	**<Utterances>** **<Utterance>**	Cada una de las oraciones en las que es dividido el texto de entrada. Existen en una salida XML tantas secciones de este tipo como oraciones tenga el documento. Las partes en las que se divide son: **-<PMID>** **-<UttSection>** **-<UttNum>** **-<UttText>** **-<UttStartPos>** **-<UttLength>** **-<Phrase>**
<PMID>	**<PMID>**	El número identificativo es asignado por PubMed a la cita de la que ha sido extraída la oración esta siendo analizada.
UttSection>	**<Location>**	Lugar donde estaba la oración en el documento evaluado (Título o Resumen).

<UttNum>	**<SeqNo>**	Posición ocupada con respecto al número total de frases del resumen o del título, por la oración que va a ser evaluada.
<UttText>	**<UText>**	Oración que va a ser evaluada.
<UttStartPos>	**<UStartPos>**	Posición de inicio de una determinada frase, considerando esta como el lugar que ocupa el primer carácter de la oración.
<UttLength>	**<USpanLen>**	Longitud total de la frase, es decir, número de caracteres que la componen, teniendo en cuenta que los espacios en blanco son considerados como otro carácter.
<Phrases> **<Phrase>**	**<Phrases>** **<Phrase>**	Se considera una Phrase como una de las subcomponentes sintácticas que forma una oración completa. Por lo que por cada "utterances" se tendrán diversas "phrases". Las partes en las que se divide esta sección son: **-<PhraseText>** **-<SyntaxUnit>** **-<PhraseStartPos>** **-<PhraseLength>** **-<Candidate>** **-<Mapping>**
<PhraseText>	**<PText>**	Contiene el texto que forma una determinada "phrase".
<SyntaxUnits> **<SyntaxUnit>**	**<Tags>** **<Tag>**	Esta parte contiene diversos atributos, que dan información sintáctica y léxica, sobre las palabras que forman una "phrase". **-<SyntaxType>** **-<LexMatch>** **-<InputMatch>** **-<LexCat>** **-<Token>**
<SyntaxType>	**<Type>**	Tipo sintáctico de la unidad analizada: "**adv**", "**aux**", "**compl**", "**conj**", "**det**", "**head**", "**mod**", "**modal**", "**pastpart**", "**prep**", "**pron**", "**punc**", "**shapes**", y "**verb**".
<LexMatch>	**<LexMatch>**	Palabra/s analizada/s en minúsculas, (forma que MetaMap procesa).
<InputMatch>	**<InputMatch>**	Palabra/s analizada/s, tal y como aparece en el texto.

<LexCat>	**<POS>**	Categoría léxica a la que pertenece la palabra/s: "**adj**", "**adv**", "**aux**", "**compl**", "**conj**", "**det**", "**modal**", "**noun**", "**prep**", "**pron**", y "**verb**".
<Tokens> **<Token>**	**<Tokens>** **<Token>**	Palabras que constituyen la unidad léxica, cada una de ellas etiquetada como un "token", habrá tantos como palabras compongan la unidad.
<PhraseStartPos>	**<PStartPos>**	Posición que ocupa el primer carácter que forma la "phrase" con respecto al texto total de entrada, y no a una sola oración (utterance).
<PhraseLength>	**<PSpanLen>**	Número de caracteres totales que constituyen una determinada "phrase".
<Candidates> **<Candidate>**	**<Candidates>** **<Candidate>**	Delimitado por esta etiqueta se muestran todo el conjunto de candidatos generados para cada una de las "phrase" que constituyen una "utterance". Los atributos que caracterizan esta sección son: **-<CandidateScore>** **-<CandidateCUI>** **-<CandidateMatched>** **-<CandidatePreferred>** **-<MatchedWord>** **-<MatchMap>** **-<SemType>** **-<IsHead>** **-<IsOverMatch>** **-<Source>** **-<ConceptPI>**
<CandidateScore>	**<NegScore>**	Valor obtenido para cada candidato, calculado mediante el proceso de evaluación de las características: centralidad, variación, cohesión y cobertura.
<CandidateCUI>	**<UMLSCUI>**	Identificador de Concepto Único asignado al concepto del Metatesauro obtenido como candidato para una "phrase" determinada.
<CandidateMatched>	**<UMLSConcept>**	Nombre del candidato extraído del Metatesauro devuelto.
<CandidatePreferred>	**<UMLSPreferred>**	Nombre "preferido" para el candidato devuelto.

<MatchedWords> **<MatchedWord>**	**<MatchedWords>** **<MatchedWord>**	Palabra/s del texto original frente a la que se evalúa la calidad de los candidatos.
<SemTypes> **<SemType>**	**<STs>** **<ST>**	Tipo semántico del candidato, corresponde al indicado junto a cada candidato entre "[ ]".
<MatchMaps> **<MatchMap>**	**<MatchMaps>** **<MatchMap>**	Estructura de datos que representa la correspondencia entre las palabras en la frase original y las del candidato, junto con un indicador de la variación léxica entre ambas. **-<TextMatchStart>** **-<TextMatchEnd>** **-<ConcMatchStart>** **-<ConcMathEnd>** **-<LexVariation>**
<TextMatchStart>	**<TWMatchPosS>**	Posición de la primera palabra de la cadena de texto original que es usada para el mapeo.
<TextMatchEnd>	**<TWMatchPosE>**	Posición de la última palabra de la cadena de texto original que es usada para el mapeo.
<ConcMatchStart>	**<CWMatchPosS>**	Posición de la primera palabra en el concepto candidato que es usada para el mapeo.
<ConcMatchEnd>	**<CWMatchPosE>**	Posición de la última palabra en el candidato que es usada para el mapeo.
<LexVariation>	**<Variation>**	Grado de variación léxica entre las palabras del texto original frente a las correspondientes en los conceptos candidatos. Si las palabras son exactamente iguales, este valor es 0.
<IsHead>	**<IsHead>**	El valor de este apartado es YES/NO, en función de si el candidato que se está evaluando constituye la parte principal de la frase o no.
<IsOverMatch>	**<IsOverMatch>**	Devuelve YES o NO, dependiendo de si el concepto candidato se trata de uno del tipo "overmatch" o no.
<Sources> **<Source>**	**<Sources>** **<Source>**	Fuente o fuentes de vocabulario de UMLS, en las que el candidato reside.

<ConceptPIs> **<ConceptPI>**	**<Spans>** **<Span>**	Información sobre la posición del concepto en el texto de entrada, presenta dos atributos: **-<StartPos>** **-<Length>**
<StartPos>	**<StartPos>**	Posición que ocupa el carácter primero del concepto evaluado en el texto de entrada, considerando como posición cero el primer carácter de la entrada. Los espacios en blancos son considerados como un carácter.
<Length>	**<SpanLen>**	Número de caracteres totales del concepto candidato evaluado.
<Mappings> **<Mapping>**	**<Mappings>** **<Mapping>**	En esta sección se estudia el conjunto de candidatos que constituyen el mapeo final del texto de entrada. Los atributos son: **-<MappingScore>** **-<Candidate>**
<MappingScore>	**<MapNegScore>**	Valor dado por la función de evaluación explicada en el algoritmo interno de MetaMap. Aparece con un signo negativo delante.

Tabla 10. Elementos de una salida XML generada por MetaMap200v2.

Capítulo 4
Estadísticas y documentos procesados

Para el estudio y evaluación de MetaMap2009v2, se ha tomado de forma aleatoria una selección de citas y referencias bibliográficas de las ofrecidas por la base de datos biomédica, MedLine 2010.

MedLine mantiene en un servidor una relación de ficheros actualizada en formato XML y en cada uno de ellos residen miles de citas biomédicas con múltiple información: título, resumen, datos de autores, fechas de publicación... entre otros.

Para la realización de este proyecto se han descargado dos de los ocho ficheros disponibles en este servidor, medline10n0024 y medline10n0348, con el primero se ha evaluado cada una de las funcionalidades proporcionadas por MetaMap2009v2 y el segundo ha sido usado para comprobar si el resultado ofrecido con la primera colección era correcto en ciertas aplicaciones.

Ambos ficheros han sidos descargados desde el enlace, "ftp.nlm.nih.gov/nlmdata/sample/medline/" y presentan un formato XML, en el

que se presentan numerosos datos referentes a cada una de las citas: el título, el resumen, autor, fechas, edición, etc.

Una correcta evaluación de MetaMap, implica el procesado independiente de cada una de las citas, por lo que se ha construido una aplicación en lenguaje JAVA, que lleva a cabo una lectura de los ficheros en formato XML extraídos de MedLine, y a continuación, crea un archivo por cada una de las citas contenidas, en el que se almacene el título y el resumen (si lo tiene) de cada una de ellas.

Dado que MetaMap, sólo lee archivos con extensión "*.txt*", es decir, sólo permite el procesado de ficheros de entrada cuya extensión sea esta, nuestra aplicación asignará a cada archivo creado la extensión "*.txt*" y como nombre el número PMID, nombre asignado a cada cita y documento biomédico existente en MedLine2010.

El número total de documentos procesados es de 7391, seleccionados aleatoriamente del fichero "medline10n0024" extraído de MedLine2010. Durante las primeras pruebas, este número de documentos era menor, en torno a los mil, pero dado que para el estudio de una de las nuevas funcionalidades que presenta especial relevancia en esta versión, NegEx, el número de oraciones con frases negativas era prácticamente inapreciable, se incrementó la cantidad de documentos.

A lo largo de este capítulo, se documentará este proceso de análisis de la colección de documentos y su evaluación presentando los datos y estadísticas resultantes del procesado de estos documentos.

4.1. Documentos analizados: Cifras y Medias.

Para obtener estas cifras, se ha construido una aplicación en Java, que lleva a cabo el cálculo de estos datos. Es importante mencionar, que la salida XML devuelta por MetaMap2009v2, añade al inicio de cada uno de estos ficheros una línea en blanco, por lo que Java no reconoce en estos ficheros la estructura XML. Para solucionar este problema, en la aplicación programada se realiza un duplicado de todos los ficheros de salida y se elimina la primera línea vacía, logrando así un correcto funcionamiento.

En la siguiente tabla, se refleja el número total de textos procesados por MetaMap2009v2, así como el número final de “utterances” (oraciones), “phrases” y “tokens”, obtenidos:

Tipo	Cantidad Total
Documentos	7391
Utterances	8173
Phrases	31391
Tokens	80525

Tabla 11. Cantidad final de documentos procesados y elementos obtenidos.

Para ilustrar el comportamiento general de los textos ofrecidos por MedLine2010, se ha calculado las siguientes medias estadísticas, en las que reflejamos el número medio de elementos por documento, oración y phrase.

Comportamiento Medio Documentos MedLine2010	
Número medio de “utterances” (oraciones) en un documento MedLine2010.	1'105 utterances/documento
Número medio de “phrases” en un documento MedLine2010.	4'427 phrases/documento
Número medio de “tokens” en un documento MedLine2010.	10'89 tokens/documento

Tabla 12. Media estadística de “utterances”, “phrases” y “tokens” en un documento.

Nota: El número de “utterances” (oraciones) es muy cercano a la unidad, esto es debido a que el número de documentos ofrecidos por la base de datos MedLine que contienen el título de la cita médica, es muy superior a aquellos en los que se muestra tanto el título, como el resumen de la cita. Es decir, es

mayor el número de documentos con una sola oración que los presentan más de una.

Comportamiento Medio “Utterances”	
Número medio de “phrases” en una “utterance”.	3'84 phrases/utterance
Número medio de “tokens” en una “utterance”.	3'85 tokens/utterance

Tabla 13. Media estadística de “phrases” y “tokens” en una oración “utterance”.

Comportamiento Medio “Phrases”	
Número medio de “tokens” en una “phrase”.	3'84 tokens/phrase

Tabla 14. Media estadística de “tokens” en una “phrase”.

Nota 2: Se puede apreciar en estas tablas que los datos calculados tanto por oración, como por phrase, son relativamente pequeños, esto es debido, al igual que señalábamos en la nota anterior, a que la mayoría de los documentos sólo presentan el título, lo que supone un valor menor para estas medias calculadas.

4.2. Documentos analizados: Negaciones.

Por otro lado, dado que una de las mejoras más relevantes en esta nueva versión es la relacionada con el algoritmo NegEx, se ha realizado de nuevo, una evaluación de todos los textos usando la opción –negex.

Para llevar a cabo el recuento del número de negaciones y de que tipo es cada una de estas, se ha construido una nueva aplicación en Java. Tanto el número de negaciones obtenidas así como el tipo de estas y la cantidad de textos que las incluyen, se muestran a continuación:

Resultados Algoritmo NegEx	
Número Documentos con "negaciones" localizadas	18
Número Total "negaciones"	32
Negaciones tipo: nega	29
Negaciones tipo: negb	3
Negaciones tipo: pnega	0
Negaciones tipo: pnegb	0
Negaciones tipo: pseudoneg	0

Tabla 15. Clasificación del número y tipo de negaciones localizadas por el algoritmo NegEx incluido en la versión MetaMap2009v2.

La mayoría de las negaciones localizadas pertenecen al tipo "nega", estas frases negativas son aquellas que niegan completamente el concepto y se localizan antes de este. Cabe mencionar, que no ha sido localizada tanto ninguna frase que indique como posibilidad condicionada, así como tampoco ninguna de tipo pseudo-negativo.

Dado que el número de negaciones localizadas resulta significativamente pequeño frente a la cantidad de documentos analizados, se ha evaluado manualmente si el algoritmo NegEx funciona correctamente. Para ello se ha tomado diez documentos que presentan un resumen (puesto que estos tienen una mayor probabilidad de contener negaciones que los artículos con sólo el título) y finalmente, se han buscado manualmente las negaciones que en ellos aperecen.

Tras llevar a cabo este análisis se han localizado diversas negaciones que no habían sido marcadas por el algoritmo NegEx. Podemos extraer como conclusión a este hecho que el funcionamiento de esta implementación no es eficiente, puesto que sólo localiza ciertas negaciones y señalar así como posible línea futura de trabajo la resolución de errores de este algoritmo.

4.3. Documentos analizados: Tipos Semánticos.

Dado que MetaMap se trata de una herramienta que proporciona información semántica sobre cada uno de los conceptos analizados, se ha llevado a cabo una toma de datos del número de cada uno de los tipos semánticos que aparecen en nuestra selección de 7391 documentos. Para este fin, se ha desarrollado una nueva aplicación Java que lleva a cabo este recuento de datos. En la siguiente tabla se presentan cada uno de los tipos semánticos que existen, así como la cantidad de cada uno de estos localizada:

Tipo Semántico	Abreviatura	Cantidad
Amino Acid, Peptide, or Protein	aapp	2125
Acquired Abnormality	acab	281
Activity	acty	1839
Age Group	aggp	791
Alga	alga	5
Amino Acid Sequence	amas	2
Amphibian	amph	127
Anatomical Abnormality	anab	463
Animal	anim	81
Anatomical Structure	anst	393
Antibiotic	antb	632
Archaeon	arch	0
Biologically Active Substance	bacs	1443
Bacterium	bact	530
Body Substance	bdsu	2142
Body System	bdsy	1121
Behavior	bhvr	121
Biologic Function	biof	355
Bird	bird	302
Body Location or Region	blor	1466
Biomedical Occupation or Discipline	bmod	2365
Biomedical or Dental Material	bodm	1267
Body Part, Organ, or Organ Component	bpoc	10641

Body Space or Junction	bsoj	609
Carbohydrate	carb	537
Cell Component	celc	369
Cell Function	celf	508
Cell	cell	807
Congenital Abnormality	cgab	500
Chemical	chem	247
Chemical Viewed Functionally	chvf	0
Chemical Viewed Structurally	chvs	338
Classification	clas	673
Clinical Attribute	clna	669
Clinical Drug	clnd	0
Conceptual Entity	cnce	4010
Cell or Molecular Dysfunction	comd	105
Carbohydrate Sequence	crbs	0
Diagnostic Procedure	diap	1916
Daily or Recreational Activity	dora	119
Disease or Syndrome	dsyn	9477
Educational Activity	edac	378
Environmental Effect of Humans	eehu	48
Eicosanoid	eico	2
Element, Ion, or Isotope	elii	1033
Experimental Model of Disease	emod	16
Embryonic Structure	emst	226
Entity	enty	82
Enzyme	enzy	385
Event	evnt	82
Family Group	famg	678
Fully Formed Anatomical Structure	ffas	15
Fish	fish	206
Finding	fndg	9567
Fungus	fngs	149
Food	food	814
Functional Concept	ftcn	22829
Genetic Function	genf	122

Geographic Area	geoa	1057
Gene or Genome	gngm	2248
Governmental or Regulatory Activity	gora	240
Group Attribute	grpa	34
Group	grup	84
Human-caused Phenomenon or Process	hcpp	96
Health Care Related Organization	hcro	689
Health Care Activity	hlca	2339
Hazardous or Poisonous Substance	hops	600
Hormone	horm	1325
Human	humn	805
Idea or Concept	idcn	8233
Immunologic Factor	imft	1340
Individual Behavior	inbe	574
Inorganic Chemical	inch	1381
Injury or Poisoning	inpo	1222
Intellectual Product	inpr	8702
Invertebrate	invt	1276
Indicator, Reagent, or Diagnostic Aid	irda	888
Language	lang	46
Laboratory Procedure	lbpr	2749
Laboratory or Test Result	lbtr	206
Lipid	lipd	168
Mammal	mamm	1783
Molecular Biology Research Technique	mbrt	5
Machine Activity	mcha	57
Medical Device	medd	4874
Mental Process	menp	1793
Manufactured Object	mnob	4144
Mental or Behavioral Dysfunction	mobd	822
Molecular Function	moft	647
Molecular Sequence	mosq	0
Neoplastic Process	neop	2277
Nucleic Acid, Nucleoside, or Nucleotide	nnon	0
Natural Phenomenon or Process	npop	1672

Neuroreactive Substance or Biogenic Amine	nsba	140
Nucleotide Sequence	nusq	2
Occupational Activity	ocac	772
Occupation or Discipline	ocdi	451
Organophosphorus Compound	opco	12
Organic Chemical	orch	3980
Organism Attribute	orga	668
Organism Function	orgf	1750
Organism	orgm	146
Organization	orgt	691
Organ or Tissue Function	ortf	902
Pathologic Function	patf	2177
Physical Object	phob	140
Phenomenon or Process	phpr	1253
Physiologic Function	phsf	1350
Pharmacologic Substance	phsu	7340
Plant	plnt	799
Patient or Disabled Group	podg	440
Population Group	popg	1667
Professional or Occupational Group	prog	3525
Professional Society	pros	20
Qualitative Concept	qlco	15169
Quantitative Concept	qnco	8729
Receptor	rcpt	4
Reptile	rept	14
Research Activity	resa	3656
Research Device	resd	18
Rickettsia or Chlamydia	rich	5
Regulation or Law	rnlw	175
Substance	sbst	993
Self-help or Relief Organization	shro	14
Social Behavior	socb	697
Sign or Symptom	sosy	1131
Spatial Concept	spco	5127
Steroid	strd	672

Tissue	tisu	1517
Temporal Concept	tmco	4648
Therapeutic or Preventive Procedure	topp	8661
Virus	virs	206
Vitamin	vita	425
Vertebrate	vtbt	8

Tabla 16. Tipos Semánticos a los que pertenecen cada uno de los conceptos UMLS candidatos localizados en nuestra selección de documentos.

Como se puede apreciar en esta tabla, la mayoría de los tipos semánticos presentan una probabilidad media de aparición. Aunque cabe mencionar, que algunos tipos semánticos presentan una elevada tasa de aparición en los documentos médicos analizados, frente a otros que apenas poseen ningún concepto perteneciente a su tipo.

Algunos de los tipos más frecuentes son: "Body Part, Organ, or Organ Componen", "Qualitative Concept", "Quantitative Concept", "Therapeutic or Preventive Procedure" y "Idea or Concept", entre otros. Mientras que algunas de las clases con una tasa menor son: "Vertebrate", "Rickettsia or Chlamydia", "Receptor" y "Reptile", entre otros.

4.4. Documentos analizados: Tipos Sintácticos y Signos de Puntuación.

Finalmente, para completar este capítulo se ha llevado a cabo un cálculo del número de conceptos analizados que pertenecen a cada tipo sintáctico, así como un cálculo del número de signos de puntuación.

Para realizar este recuento se ha programado una nueva aplicación Java que extrae de una colección de ficheros los tipos sintácticos y a continuación, realiza un recuento de cada uno de estos. Cabe mencionar, que la aplicación ha de recibir un formato concreto como fichero de entrada en el que se especifiquen los tipos sintácticos, para ello se ha de procesar la colección de citas bibliográficas usando la opción "-F" de MetaMap2009v2, la cual devuelve un análisis sintáctico del texto.

Tipo Sintáctico	Abreviatura	Cantidad
Determinante	det	5362
Nombre	noun	31722
Preposición	prep	27820
Adjetivo	adj	11077
Conjunción	conj	2413
Adverbio	adv	608
Verbo	verb	1137
Pronombre	pron	354
Complemento	compl	8
Modal	modal	17
Numeral	num	609
Auxiliar	aux	114
TOTAL		81241

Tabla 17. Tipos Sintácticos y número de conceptos que corresponden a cada uno de estos tipos, localizados en nuestra selección de documentos.

Signo de Puntuación	Abreviatura	Cantidad
“,”	cm	703
“.”	pd	7834
“/”	sl	15
“-”	hy	1152
“(“	lp	164
“)”	rp	164
“%”	pc	16
“;”	sc	755
“:”	cl	274
“[”	lk	4157
“]”	rk	4157

Tabla 18. Signos de Puntuación localizados en la selección de documentos.

Los tipos sintácticos más frecuentes son: los nombres (39'05%), las preposiciones (34'34%) y los adjetivos (13'63%). Frente a otros tipos que presentan un porcentaje de aparición bastante inferior: complementos (0'01%), modales

(0'02%), auxiliares (0'14%), pronombres(0'44%), adverbios (0'74% y numerales (0'75%).

Finalmente, los signos de puntuación más frecuentes localizados en nuestra colección de documentos son los puntos (“.”), puesto que en cada uno de estos documentos al menos existirá uno que marque el final del título.

Capítulo 5
Conclusiones y Líneas de Trabajo futuro

A lo largo de esta memoria se ha analizado de manera teórica y posteriormente de forma práctica el funcionamiento de la nueva herramienta, MetaMap2009v2. El estudio teórico incluye un análisis del algoritmo interno de este programa, con el que se logra cumplir uno de los objetivos principales de este proyecto, obtener una visión completa del proceso seguido por MetaMap hasta obtener su salida actual.

En cuanto a la parte práctica, se han alcanzado los objetivos expuestos al comienzo de nuestro proyecto. Analizamos a continuación cada uno de estos, así como los principales problemas surgidos y las soluciones aplicadas:

- Adquisición e instalación de MetaMap2009v2. Se ha tratado de ilustrar de una forma clara y comprensible estos procesos, en concreto, el de instalación, puesto que al realizarse a través de la línea de comandos del terminal de Linux, puede presentar algunas dificultades en su introducción.

- Otro de los puntos que se ha tratado con especial interés, es la ejecución de este programa y las diversas formas con las que este puede ser invocado, así como la estructura general que ha de seguir la invocación de MetaMap2009v2 (opciones, ruta fichero de entrada y ruta de salida).

- Adquisición de una colección amplia de documentos y citas biomédicas para su posterior procesado con MetaMap2009v2. Esta colección ha sido adquirida de entre los documentos disponibles en la base de datos médica MedLine2010.

- Uno de los problemas surgidos al comienzo de este proyecto gira en torno a que esta herramienta sólo procesa ficheros en formato "*.txt*", pero en cambio MedLine almacena los documentos y referencias bibliográficas en diversos ficheros con formato *".xml"*, en concreto, en cada de uno de ellos residen miles de estas referencias y documentos. Para solucionar esta incompatibilidad de formatos se ha generado una aplicación en Java que lleva a cabo una lectura de cada fichero *".xml"* y extrae cada una de las citas almacenadas (su título y resumen), para finalmente almacenar estas en un fichero *".txt"*. Esto nos ha permitido procesar con éxito la selección de documentos extraída de MedLine2010.

- Procesado de la selección de documentos extraída de MedLine, este procesado se ha realizado de dos formas diferentes. Por un lado para la evaluación de las opciones de procesado y salida se han tomado de forma aleatoria diversos textos breves que nos ilustrasen con un ejemplo la función de cada opción. Y por otro lado, se ha procesado la selección completa para dos de las principales mejoras introducidas en esta versión, la nueva salida XML y el algoritmo NegEx.

- Dado que uno de los principales motivos de este proyecto era analizar que nuevas funcionalidades nos aportaba esta nueva versión de MetaMap con respecto a las anteriores, hemos analizado la evolución de esta herramienta desde la publicada en 2006. Con ello se ha logrado un análisis pormenorizado de diversas funcionalidades, en concreto del algoritmo NegEx y la salida XML, observándose que la innovación y

resolución de errores en torno a estas se encuentra presente en todo momento para los creadores y desarrolladores de MetaMap, puesto que en cada nueva versión se producen nuevas mejoras y variaciones.

En cuanto a las líneas de trabajo futuro que pueden ser seguidas, en torno al análisis de esta herramienta y sus nuevas versiones, podemos nombrar:

- Evaluación continúa de cada nueva versión de esta herramienta que se incorpore al mercado.

- Implementación de nuevas aplicaciones Java que permitan el procesado de las salidas devueltas por MetaMap en cada nueva versión o que realicen funciones distintas a las que se han realizado en este proyecto.

- Análisis de posibles errores producidos por MetaMap, con una especial atención a los producidos por el algoritmo NegEx.

- Evaluación de la herramienta en nuevas bases de datos bibliográficas de tipo biomédico, así como la construcción de aplicaciones Java que procesen los documentos contenidos en las nuevas bases de datos en función del formato en el que estas almacenen los textos.

- Análisis de otros fenómenos de interés como la resolución de la anafóra, acrónimos y abreviaturas, ambigüedad, etc.

- Estudio de otras herramientas dedicadas al procesamiento sintáctico y semántico de textos en el dominio biomédico.

Bibliografía:

1. Alan R. Aronson: “MetaMap: Mapping Text to the UMLS Metathesaurus”, July 14, 2006 [Online]

http://skr.nlm.nih.gov/papers/references/metamap06.pdf

2. Alan R. Aronson, Phd: “Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program”, National Library of Medicina, Bethesda, MD 20894 [Online]

http://skr.nlm.nih.gov/papers/references/metamap_01AMIA.pdf

3. Alan R. Aronson: “MetaMap Evaluation”, May 8, 2001. [Online]

http://skr.nlm.nih.gov/papers/references/mm.evaluation.pdf

4. Marina Leonardo, José L.: “Extracción de conceptos en textos biomédicos”, Tesis Docotoral, Universidad Complutense de Madrid, 2009.

5. Perea Ortega, José M., Martín Valdivia, María T., Montejo Ráez, A. y Díaz Galiano, Manuel C.: “Categorización de textos biomédicos usando UMLS”, Revista nº 40, marzo de 2008, pp.121-127.

6. Segura Bedmar, I., Martínez, P. y Samy, Doaa.: “Detección de fármacos genéricos en textos biomédicos”, Revista nº 40, marzo de 2008, pp.27-34.

7. Segura Bedmar, Isabel**.** Application of Information Extraction techniques to pharmacological domain: Extracting drug-drug interactions. s.l. : Universidad Carlos III de Madrid, 2010.

8. Sobrido Prieto, María y González Guitian, Carlos: “Buscar en MedLine con PubMed”, Biblioteca del Complexo Hospitalario Universitario de A Coruña.

9. Vallez, Mari y Pedraza, Rafael: “El procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines”. [Online]

http://www.hipertext.net/web/pag277.htm

10. DeCS 2010. Biblioteca Virtual de la Salud. [Online]

http://decs.bvs.br/E/visiongeneral2010.htm

11. Medical Subject Headings (MeSH) [Online]

http://www.nlm.nih.gov/pubs/factsheets/mesh.html

12. 2010 MEDLINE/PubMed Baseline Distribution. [Online]

http://www.nlm.nih.gov/bsd/licensee/2010_stats/baseline_doc.html

13. MetaMap Output Methods and their Interaction with Other Options [Online]

http://skr.nlm.nih.gov/papers/MetaMap_Options_Chart.pdf

14. MetaMap Examples [Online].

http://skr.nlm.nih.gov/papers/references/MetaMap_Examples_06.pdf

15. Procesamiento de lenguajes Naturales, Wikipedia (The Free Encyclopedia). [Online].

http://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales

16. SPECIALIST Lexicon [Online]

http://www.nlm.nih.gov/pubs/factsheets/umlslex.html

17. Unified Medical Language System [Online]

http://www.nlm.nih.gov/pubs/factsheets/umls.html

18. UMLS Metathesaurus [Online]

http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

19. UMLS Semantic Network [Online]

http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

20. Stephane Meystre, Peter J. Haug, Natural language processing to extract medical

problems from electronic clinical documents: Performance evaluation, Journal of Biomedical Informatics, Volume 39, Issue 6, December 2006, Pages 589-599, ISSN 1532-0464

21. Roxana Danger, Isabel Segura-Bedmar, Paloma Martinez, Paolo Rosso, A comparison of machine learning techniques for detection of drug target articles, Journal of Biomedical Informatics, In Press,

22. Isabel Segura-Bedmar, Paloma Martinez, Maria Segura-Bedmar, Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems, Drug Discovery Today, Volume 13, Issues 17-18, September 2008, Pages 816-823, ISSN 1359-6446,

23. Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, Bruce G. Buchanan, A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries, Journal of Biomedical Informatics, Volume 34, Issue 5, October 2001, Pages 301-310, ISSN 1532-0464, DOI: 10.1006/jbin.2001.1029.

http://www.dbmi.pitt.edu/chapman/NegEx.html