Maciej Rybiński

Improving approximation of domain-focused, corpus-based, lexical semantic relatedness

– PhD Dissertation –

June 24, 2017

Under supervision of Dr José Francisco Aldana Montes
Dpto de Lenguajes y Ciencias de la Computación,
University of Málaga

AUTOR: Maciej Rybinski

http://orcid.org/0000-0002-0174-0567

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

Dr. José Francisco Aldana Montes, Profesor Catedrático del Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga.

Certifica

Que **D. Maciej Rybinski**, Ingeniero en Informática por la Politécnica de Varsovia, Polonia, ha realizado en el Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, bajo su dirección, el trabajo correspondiente a su tesis doctoral titulada

Improving approximation of domain-focused, corpus-based, lexical semantic relatedness

Revisado el presente trabajo, estimo que puede ser presentado al tribunal que ha de juzgarlo, y autorizo la presentación de esta tesis doctoral en la Universidad de Málaga.

En Málaga, enero de 2017

Firmado: **Dr. José Francisco Aldana Montes**
Profesor Catedrático del Dpto. de Lenguajes y Ciencias de la Computación de la Universidad de Málaga

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. José F. Aldana Montes for the continuous support of my Ph.D study and related research, for his patience, motivation, and help in navigating through the difficult process.

Besides my advisor, I would like to thank the external reviewers of my dissertation: Prof. Zbigniew Raś and Dr. Sara Mendes, for their insightful comments, which contributed to improving the study presented here.

My sincere thanks also goes to Prof. Palmira Marrafa and Dr. Raquel Amaro, who provided me an opportunity to join their team as intern at the University of Lisbon.

I thank my colleagues from the Khaos research group. In particular I would like to express gratitude to my fellow PhD students, for being great teammates, both inside and outside the lab.

Last but not the least, I would like to thank my family: my parents and my girlfriend for their endless support, patience and confidence in me.

Contents

Resumen 1

1 Introduction 15
1.1 Background 15
1.2 Related work 17
1.3 Contributions 20
1.3.1 Papers 20
1.3.2 Research questions 22
1.4 The methodology overview 23

2 Published works 27
2.1 SEBSem: simple and efficient biomedical semantic relatedness measure 27
2.2 Calculating semantic relatedness for biomedical use in a knowledge-poor environment 29
2.3 tESA: a distributional measure for calculating semantic relatedness . 32
2.4 DomESA: a novel approach to extended domain-oriented lexical relatedness calculations with domain-specific semantics 35
2.5 DisMatch results for OAEI 2016 37

3 Results and Discussion 39

4 Conclusions 43

Conclusiones 49

Literature 55
References 55

Resumen

Introducción

Teniendo en cuenta el volumen de las publicaciones generadas en el área de las ciencias de la vida en los últimos años, se puede observar que la comunidad científica necesita soluciones para mejorar el acceso al conocimiento contenido en los artículos científicos. La posibilidad de estimar la fuerza de la relación semántica (o proximidad semántica) entre dos conceptos, o los textos que representan a los conceptos, es un componente clave para creación de los sistemas inteligentes que podrían facilitar el acceso a la información contenida en las enormes colecciones de publicaciones científicas. Ejemplos de aplicación de las técnicas de aproximación de la relación semántica incluyen: la desambiguación, la construcción automática de resúmenes, los algoritmos de búsqueda de contenidos. Los usos de estas técnicas en escenarios orientados al dominio de las ciencias de la vida incluyen, entre otros: el análisis de datos, el descubrimiento de las enfermedades parecidas o análogas, la detección de la redundancia en los textos clínicos, etc.

Desde una perspectiva básica, considerándolo como una caja negra, un sistema de la aproximación de la relación semántica toma dos entradas y devuelve un número que refleja la fuerza de la relación semántica entre las dos entradas. Dentro de la caja negra, el sistema genera las representaciones semánticas[1] de las entradas y produce la métrica numérica tras comparar estas representaciones. En general, el concepto de una entrada es abstracto, pero en este trabajo la palabra 'entrada' significa realmente 'texto de entrada', ya que en el ámbito de este estudio, solo se contemplan aproximaciones a la relación semántica entre textos.

En este estudio, se presenta un conjunto de trabajos publicados que avalan la correspondiente tesis doctoral del autor. En las 5 contribuciones científicas se intro-

[1] En sistemas distintos se puede contemplar representaciones distintas. Por ejemplo, en un sistema basado en el conocimiento estructurado, una entrada puede estar representada o por su posición o por su importancia dentro de la estructura del conocimiento. En este estudio, donde se contempla los métodos estadísticos basados en grandes colecciones de textos, una representación es típicamente un vector de distribución generado por la entrada sobre un espacio relacionado con el corpus (documentos, vocabulario, tópicos, etc.).

ducen 3 métodos originales para la aproximación de la relación semántica entre los textos, también se describe una evaluación de un experimento consistente en aplicar uno de los métodos propuestos al problema del alineamiento de ontologías.

Los métodos para la aproximación de la relación semántica propuestos en los trabajos relacionados con la línea de la investigación presentada aquí, están enfocados principalmente en aproximar la relación semántica entre palabras/textos. En los métodos propuestos en este trabajo, solo se contempla el uso de grandes colecciones de documentos (artículos científicos, artículos enciclopédicos) como origen del conocimiento para guiar el proceso. Es de esperar que un sistema basado en este principio, además de proporcionar resultados de buena calidad, sea más flexible y robusto que los sistemas que abordan la misma clase del problema (evaluación de la proximidad entre las entradas) en base a palabras/textos.

Tradicionalmente, los sistemas de aproximación de la relación semántica se clasifican según el tipo de recurso usado como origen del conocimiento del sistema. Todos los métodos propuestos en esta tesis usan colecciones de los documentos y sus estadísticas para aproximar la relación semántica.

Los métodos presentados en este trabajo comparten un denominador común con los métodos distribucionales, que se basan en la idea de modelar la semántica de una palabra, o un texto, con sus contextos y coocurrencias con otras palabras, dentro de una colección grande de documentos. La idea en si no es nueva, ya que en su forma básica fue formulada en los años cincuenta del siglo XX.

Hoy en día, la etiqueta "semántica distribucional " une a métodos muy variados, desde los enfoques muy simples, basados en contar las coocurrencias entre las palabras, hasta los métodos bastante elaborados, basados en el aprendizaje computacional.

Los métodos propuestos en los trabajos que avalan esta tesis, están basados conceptualmente (en grado variado) en un enfoque del popular método ESA (Explicit Semantic Analysis), que ha sido usado con éxito para calcular la proximidad semántica de los textos en el dominio general.

Como ya se ha mencionado, el compendio de las publicaciones que avalan esta tesis incluye cinco posiciones. Entre las cinco, la primera y la última son contribuciones cortas a conferencias internacionales, el compendio incluye también tres publicaciones en revistas indexadas en JCR.

En la primera de las publicaciones (conferencia Nettab) se presenta el trabajo previo, que ha llevado al desarrollo del método DS, presentado luego en más detalle en el primer de los artículos de las revistas (publicado en BMC Bioinformatics). El segundo de los artículos completos (publicado en Journal of Biomedical Semantics) introduce el método tESA, una extensión de ESA que aprovecha adicionalmente la semántica de los títulos de los artículos que forman parte de un corpus de contribuciones científicas. En el tercer de los artículos (aceptado para publicación en Journal of Intelligent Information Systems) se ha propuesto el método DomESA, que, usando un enfoque computacionalmente similar al del método anterior, combina la semántica de dos colecciones de los documentos (una general y una enfocada al dominio) para obtener representaciones semánticas de las entradas. La última contribución (conferencia OM-2016) presenta los resultados de aplicación

de DomESA al problema de alineamiento de ontologías, en el cual se trata de encontrar las clases correspondientes entre dos ontologías. En este estudio, los resultados de la tarea del alineamiento de ontologías sirven como experimento con uno de los métodos propuestos. En conjunto, los trabajos presentados contribuyen con tres métodos originales para la aproximación de la relación semántica entre textos en el contexto biomédico. Cada uno de los métodos ha mejorado el estado del arte en el momento de su publicación, en cuanto la calidad de los resultados obtenidos en la evaluación directa. DomESA, siendo el método más reciente de los tres, proporciona los resultados superiores en cuanto a calidad. Sin embargo, varios aspectos del funcionamiento de los métodos hacen que DomESA no invalide a, por ejemplo, DS, ya que el DS es mucho más eficiente computacionalmente, por lo cual puede adaptarse mejor a algunos escenarios de uso.

Las contribuciones presentadas aquí están vinculadas a las siguientes cuestiones, que permiten dar perspectiva a las conclusiones sacadas al final de este estudio.

Q1 Uso de representaciones obtenidas a través de los contextos aproximados.
Q2 Mejora de la calidad de los métodos aprovechando la semántica del dominio.
Q3 Maximización de los beneficios producidos por el aprovechamiento de la semántica del dominio.
Q4 Uso novedoso de las técnicas de la aproximación de la relación semántica entre los textos.

Cada una de las cuestiones Q1-Q3 está relacionada con ciertos sub-problemas:

1.a Optimización del tamaño de las representaciones semánticas,
1.b Optimización del proceso de adquisición de las representaciones semánticas.
2.a Beneficios de usar una colección de documentos más grande.
2.b Posibilidad de usar documentos más cortos.
2.c Establecer la base de referencia con el método ESA original.
3.a Incremento de la robustez del método.
3.b Incorporación de la semántica de dominio manteniendo la generalidad del método.
3.c Mejora de la correlación lineal con las respuestas de referencia.

La calidad de la aproximación obtenida con los métodos propuestos en los trabajos relacionados se mide con respecto a cinco conjuntos de datos de referencia, que forman parte de un estándar *de facto* para la evaluación de los métodos de aproximación de la relación semántica en el contexto de las ciencias de la vida.

En la evaluación directa, se mide la similitud de las respuestas, obtenidas automáticamente a través del algoritmo evaluado, con las respuestas estandarizadas de los evaluadores humanos. Cada uno de los conjuntos de referencia contiene un cierto número de las parejas de entradas (palabras o expresiones cortas), cada una de ellas relacionada con una respuesta de modelo, siendo la respuesta una media de las respuestas asignadas por varios evaluadores humanos, cuya intuición sobre la fuerza de las relaciones semánticas tiene que reproducir el sistema. Las respuestas de referencia se han obtenido calculando la media de valores numéricos asignados por profesionales del dominio.

Para medir el éxito del sistema de aproximación de la relación semántica se suele usar los coeficientes de correlación con respecto a las respuestas de referencia. Se suele usar, sobre todo, el coeficiente de correlación de Spearman, que permite medir la calidad del ranking obtenido por el sistema. Alternativamente, se puede contemplar el uso del coeficiente de correlación de Pearson, que mide la correlación lineal con respecto a las respuestas de referencia, aunque sea la opción menos usada en la literatura. En la evaluación presentada en este estudio se ha usado las dos medidas (los coeficientes de Pearson y de Spearman) para proporcionar una base comparativa con un espectro más amplio de los trabajos de referencia.

En el trabajo presentado aquí, las comparaciones entre las correlaciones se han validado con unas pruebas estadísticas basadas en los intervalos de confianza, que bajo el nivel de confianza asumido (típicamente 0.95), permiten rechazar la hipótesis de las correlaciones siendo iguales, si los intervalos no contienen a cero.

La calidad del sistema del alineamiento de las ontologías que está basado en el uso de uno de los métodos propuestos para la aproximación de la proximidad semántica se ha medido en un proceso externo, dentro de las ramas de la iniciativa OAEI 2016. El proceso de evaluación ha combinado el uso de los otros sistemas, que participaron en el proceso, para crear un estándar de la respuesta esperada con la evaluación manual de las respuestas no alineadas con el estándar resultante.

Publicaciones

SEBSem: simple and efficient biomedical semantic relatedness measure

El resumen extendido incluido aquí ha sido publicado en EMBNet.journal como parte de la conferencia Nettab 2013. En el congreso, el trabajo se ha presentado en formato de poster.

El resumen extendido presenta un experimento (evaluado con un conjunto de datos de referencia) con una versión preliminar del método DS, y configurada para gestionar en línea todo el proceso de creación de las representaciones de los textos. En el manuscrito se describen los resultados del experimento inicial, que fueron muy prometedores.

En el método en sí, se propone la creación de representaciones de los textos de entrada en función del vocabulario clave de los documentos que están más relacionado con las entradas. Se introduce así el concepto de contexto aproximado. Los documentos que se encuentran más relacionados se extraen del corpus biomédico, para adaptar el método al conjunto de datos de referencia orientado a un dominio concreto. Para la variante del método presentado en este trabajo, la creación de la representación de una entrada ese realiza de la siguiente forma: 1. Localización de los K documentos más relevantes para la entrada del corpus; 2. Extracción del vo-

cabulario clave para cada uno de los K documentos; 3. Agregación del vocabulario como un conjunto de elementos.

El paso 1 es abstracto, o sea el método puede usar cualquier mecanismo para localizar los documentos más relevantes. En el experimento se ha usado un modelo de espacio vectorial típico para el problema de la recuperación de la información (de modo que las entradas se tratan como consultas). Para continuar de forma coherente, en el paso 2 se ha optado por un método que permite extraer el vocabulario clave del documento sin tener que usar las estadísticas globales del corpus – TGSP, un método de búsqueda de patrones frecuentes dentro de un documento. En el paso 3 se agregan las representaciones del vocabulario extraído del corpus. Las representaciones son conjuntos (no reflejaban la importancia relativa de cada uno de los elementos individuales extraídos en el paso 2). La medida final, o sea el resultado de comparar dos representaciones, se calcula como la fuerza del solapamiento entre los conjuntos dividida por el tamaño del conjunto más grande.

Como curiosidad, cabe destacar que en esta versión preliminar del sistema, se ha contemplado el uso de los Datos Vinculados (específicamente DBPedia) para estudiar la expansión semántica de las entradas del sistema. Este enfoque no ha funcionado bien en la evaluación más detallada planificada para la publicación siguiente, por lo cual no se ha vuelto a usar.

El método, a pesar de haber usado herramientas muy simples, ha generado resultados bastante prometedores, lo cual ha indicado la validez del enfoque basado en contextos aproximados, que fue el punto de partida para la aportación presentada en la siguiente sección.

Calculating semantic relatedness for biomedical use in a knowledge-poor environment

El artículo se ha publicado en la revista BMC Bioinformatics y describe la investigación basada en las ideas introducidas en el resumen extendido presentado en la sección anterior.

En el artículo se expone la evaluación completa del método DS refinado, sobre un rango ampliado de conjuntos de datos de referencia y teniendo en cuenta un amplio rango de parámetros del método.

El método DS se define de una manera muy parecida a la del trabajo preliminar. La idea clave sigue siendo el modelado de los contextos aproximados usando un corpus biomédico: las representaciones de las entradas se crean extrayendo el vocabulario clave de una muestra relativamente pequeña (teniendo en cuenta la escala del corpus) de los documentos más relevantes para cada una de las entradas. Básicamente, se han propuesto tres cambios importantes en la estructura del procesamiento explicada en la sección anterior, que han contribuido a una mejora importante al rendimiento del método, con respecto a la investigación preliminar: (a) las representaciones se modelan como vectores, en vez de en conjuntos, de tal manera que puedan reflejar la importancia de cada elemento incluido en la representación;

(b) se incluye el uso de estadísticas globales del corpus para modelar dicha importancia; (c) se comparan las representaciones usando la similitud coseno entre los vectores, para poder incluir en los cálculos la importancia de los elementos individuales dentro de la representación..

El punto (a) simplemente implica que los pesos de los elementos individuales de las representaciones (frases, palabras) se calculan como la suma de las frecuencias en los documentos relevantes para cada una de las entradas. El punto (b) significa que en los pesos asignados eventualmente a las palabras individuales se incluyen los factores de la frecuencia inversa de los elementos en el corpus. La modificación del punto (c) permite reflejar estos cambios de modelado en el resultado numérico final.

En la evaluación presentada en el artículo se contempla una comparativa entre dos variantes del método contra el rango completo de los conjuntos de datos de referencia. Una de las variantes utiliza TGSP como método de extracción del vocabulario de los documentos individuales, la segunda variante está basada en aprovechar la representación Tf-Idf tradicional. Los resultados obtenidos en el proceso de evaluación se comparan con los resultados descritos en la literatura relevante.

En la segunda parte del artículo se presenta un experimento de análisis de datos ejecutado con una de los variantes del método propuesto. En el experimento se trata de buscar conceptos relacionados entre los conceptos provenientes de una base de datos de enfermedades raras. El experimento sirve como ilustración de las capacidades y las limitaciones del método propuesto en el artículo.

Los resultados de la evaluación directa revelaron que el método DS es capaz de proporcionar un rendimiento comparable con los métodos que por aquel entonces formaban parte del estado del arte, superando los resultados de referencia en los casos de los tres conjuntos de datos más grandes (de los cinco conjuntos de datos contemplados en el proceso de la evaluación).

Sin embargo, la evaluación también ha demostrado que el concepto de los contextos aproximados presenta ciertas desventajas en el método propuesto, principalmente por la relación muy directa entre la entrada y la representación semántica. El rendimiento del método cambia bastante en función de los parámetros usados, tamaño de los documentos, la composición del corpus utilizado, etc., por lo cual el método no es fácil de ajustar a un caso de uso particular. Estos factores dejaron un margen de mejoras que ha sido considerado en el proceso del diseño del método introducido en la siguiente contribución.

tESA: a distributional measure for calculating semantic relatedness

El artículo, publicado en la revista Journal of Biomedical Semantics, introduce el segundo de los tres métodos presentados en este estudio, junto con la evaluación completa.

tESA es una extensión del popular método ESA, que transforma las representaciones obtenidas con la metodología de ESA original con un paso computacional

adicional. La extensión está diseñada para funcionar bien con un corpus de los documentos científicos, aprovechando la representación semántica del contenido del artículo proporcionada por el vocabulario de su título.

Las representaciones de tESA se obtienen multiplicando un vector (una representación) de ESA con una matriz en la cual, las columnas corresponden a los documentos del corpus y las filas corresponden a los elementos de los títulos de los documentos que forman parte del corpus. Los valores de la dicha matriz se establecen como pesos Tf-Idf que los elementos del vocabulario tienen en los títulos de los respectivos documentos[2].

Además, en el proceso de la evaluación presentado en el artículo, se ha incluido los resultados de referencia obtenidos con el método ESA, configurado tanto con la Wikipedia (como en la implementación original), como con las colecciones de los documentos científicos del campo de biomedicina (Medline, PMC OA).

En general, tESA proporciona un enfoque parecido a ESA, pero con la representación de cada documento incluido en la representación de la entrada enriquecida con la semántica del documento. A nivel conceptual es un enfoque parecido a las otras extensiones de ESA diseñadas para 'romper' la ortogonalidad de la colección de los documentos, pero tESA conlleva un coste computacional más bajo que dichas extensiones.

tESA proporciona resultados superiores (en cuanto la calidad) a los resultados de referencia obtenidos con ESA original, superando también los resultados obtenidos con el método DS. Además, se puede observar que el método es más robusto que DS, ya que ni los pequeños cambios en los parámetros, ni pequeños cambios en el corpus, inducen cambios significativos en el rendimiento del método.

En cuanto a la eficiencia computacional, el uso de tESA conlleva un coste adicional en comparación con DS, tanto a la hora de la creación de las representaciones (el proceso es más complejo), como a la hora de comparar las representaciones (las representaciones de tESA tienen más elementos con valores distintos a cero). Sin embargo, la dimensión de los vectores de tESA es mucho más baja que la de DS, lo cual puede ser computacionalmente importante en algunos casos de uso. Si consideramos el método original, tESA conlleva un coste adicional en el proceso de la creación de las representaciones, pero las representaciones de tESA son más ligeras y de dimensión significativamente inferior.

DomESA: a novel approach to extended domain-oriented lexical relatedness calculations with domain-specific semantics

El artículo, aceptado para la publicación en la revista Journal of Intelligent Information Systems, introduce el tercer método, DomESA, presentado en este estudio.

[2] Por lo cual se trata de una matriz muy dispersa, en la que adicionalmente se considera solo los valores más importantes de cada columna.

DomESA utiliza una metodología parecida a la de tESA, siendo una extensión de ESA, en la cual el método original también está extendido con un paso adiciona, la multiplicación de los vectores de ESA por una matriz. Específicamente, para obtener los vectores de DomESA hay que multiplicar las representaciones del método ESA (configurado con un corpus biomédico) por una matriz de la similitud coseno entre los documentos del corpus biomédico y los conceptos de Wikipedia. En las columnas de la matriz solo se contempla los k mayores valores. La diferencia entre tESA y DomESA es que la idea cambia: los contenidos de los documentos individuales del corpus especializado se representa usando los k conceptos de Wikipedia más adecuados para cada uno de los documentos en vez de usar para ello el vocabulario extraído de los títulos de los documentos.

La evaluación, en la cual está incluido adicionalmente un método basado en word embeddings, demuestra que DomESA ofrece mayor calidad en los resultados que el resto de métodos incluidos en la evaluación, superando también (a nivel general, no necesariamente a nivel de los resultados individuales aislados) todos los resultados de referencia conocidos en el dominio. Adicionalmente, DomESA parece proporcionar el mejor grado de robustez, lo cual parece confirmarse en todas las pruebas con los dos tipos de correlación usados para medir el éxito de las aproximaciones.

DomESA proporciona las representaciones del tamaño comparable con las de ESA original configurado con Wikipedia[3], así que igualmente como en el caso de tESA, DomESA solo supone un coste computacional adicional a la hora de la creación de las representaciones. Adicionalmente, en el caso de DomESA no se puede ignorar el coste de preparación de la matriz, que conlleva la necesidad de calcular las similitudes entre todos los documentos de las dos colecciones.

Al ser expresados sobre el espacio de los conceptos de Wikipedia, los vectores de DomESA son sintácticamente compatibles con los del método ESA original (configurado con Wikipedia), lo cual extiende la usabilidad de DomESA. Además, esta característica abre la puerta para combinar DomESA con las otras extensiones del método original, basadas en las características de la estructura de los datos de Wikipedia.

Próximamente se va a investigar posibles extensiones de DomESA relacionadas con el aspecto composicional de las representaciones de las entradas largas, compuestas por más que una palabra.

DisMatch results for OAEI 2016

La ultima contribución, publicada electrónicamente es el acta del congreso OM-2016, que describe un experimento (y sus resultados) en la aplicación de un método de aproximación de la proximidad semántica (específicamente, DomESA) al problema del alineamiento de las ontologías. El sistema resultante, DisMatch, ha par-

[3] Pero más ligeras que las de ESA configurado con Medline.

ticipado en la campaña OAEI 2016, en la cual los sistemas de alineamiento están evaluados en un proceso externo.

El problema del alineamiento de las ontologías dentro de OAEI se reduce prácticamente al problema de buscar equivalencias entre clases de dos ontologías. En el enfoque presentado aquí, al módulo basado en DomESA, que solo utiliza las etiquetas de las clases ontológicas como entradas, se ha adjuntado un módulo de alineamiento estructural básico (Similarity Flooding).

El enfoque, aunque muy simple, ha destacado en el aspecto de descubrimiento de equivalencias únicas, o sea las equivalencias no detectadas por los otros sistemas. Aunque el diseño del DisMatch se tendría que mejorar para refinar su rendimiento, los resultados se ven prometedores en cuanto al posible impacto de métodos como DomESA como un potencial componente de un sistema de alineamiento ontológico.

Los resultados y la discusión

Este capítulo contiene el resumen global de los resultados junto con el de un experimento adicional realizado con el método DS, llevado a cabo para proporcionar una perspectiva relevante para comparar los aspectos importantes de los métodos.

Se observa, que los resultados medios calculados sobre un cierto rango de los posibles parámetros del método DS pueden dar una idea más representativa del posible rendimiento del método. Los resultados globales indican que DomESA es el método más consistente. En la evaluación de los resultados con respecto a las correlaciones de Spearman, DomESA da mejores resultados que cualquier otro método, mientras que, considerando los resultados con respecto a las correlaciones de Pearson, DomESA proporciona los mejores resultados en absoluto en cada uno de los conjuntos de datos de referencia. tESA parece ser el segundo mejor método en cuanto la calidad de los resultados, con un rendimiento alto apreciable sobre todo en los casos de los conjuntos de referencia más grandes.

Se puede observar también, que el método DS proporciona mejores resultados en cuanto a la correlación de Spearman que el método CBOW (word embeddings), utilizando las representaciones dispersas de tamaño absoluto (la cantidad de elementos mayores a cero) comparable con la dimensión de las representaciones densas de CBOW.

Este capítulo presenta los resultados de una manera breve, mientras que las conclusiones, con las partes de la discusión correspondientes, están presentadas en el capítulo siguiente.

Conclusiones

En las publicaciones relacionadas con este estudio se han introducido un conjunto de métodos nuevos para la aproximación de la proximidad semántica basada en

palabras, en el dominio biomédico. Cada uno de los tres métodos ha mejorado el estado del arte en el momento de su publicación. En este capítulo se presenta las conclusiones derivadas de las evaluaciones realizadas incluyendo el experimento adicional enfocado en el alineamiento de las ontologías.

El primer método, DS, está basado en la idea de los contextos aproximados. La clave del método es el postulado de representar la entrada con el vocabulario más importante de los documentos más adecuados, con lo cual, se consigue una aproximación del contexto de la entrada, pero sin tener que 'escanear' los documentos individuales para las ocurrencias de las entradas.

Los resultados obtenidos en la evaluación indican, que la idea de aproximar contextos con el vocabulario clave de los documentos más adecuados funciona bastante bien, con DS siendo el método más rápido y flexible a la hora de crear las representaciones. Además, en DS, las representaciones nuevas se puede crear fácilmente – usando las estadísticas del corpus y los vectores (Tf-Idf) de los documentos.

El método DS proporciona resultados inferiores que los de los métodos propuestos más adelante, pero comparables con los mejores métodos contemplados por aquel entonces. Además, DS usa las representaciones más 'ligeras' que los otros métodos propuestos aquí (las representaciones de DS tienen notablemente menos valores diferentes a cero que las de los otros métodos). La dimensión de las representaciones de DS está determinada por el tamaño del vocabulario del corpus. Generalmente, DS proporciona la aproximación de la proximidad semántica más rápida que la de los otros métodos, comparable solo con el rendimiento de los word embeddings. No obstante, el método DS proporciona la aproximación de la calidad relativamente buena (superior a la de algunos métodos establecidos, como CBOW o ESA configurada con Wikipedia). Esta metodología se puede recomendar para usar en los casos en los cuales la eficiencia en los cálculos uno a uno tiene prioridad suficientemente importante para justificar cierto nivel de compromiso en cuanto la calidad de la aproximación.

Sin embargo, el método DS sufre de cierto tipo de dependencia de los parámetros y otros factores (el corpus, el conjunto de referencia, etc.).

El método tESA es una extensión del método ESA. En tESA los vectores de representación de las entradas se obtienen multiplicando los vectores de conceptos (que serían las representaciones en el método ESA) por una matriz del vocabulario de los títulos de los documentos del corpus usado como la fuente de conocimiento. El método tESA, configurado con el corpus Medline, ha proporcionado una calidad de la aproximación un poco más alta que el método original (también configurado con el mismo corpus), pero usando las representaciones más ligeras (con menos elementos distintos a cero) y de dimensión mucho más baja.

El diseño de tESA soluciona la mayoría de los problemas del método anterior, DS. Específicamente, tESA es fácil de calibrar y es robusto en cuanto los cambios de los parámetros. Además, el método depende mucho menos de las características de los documentos individuales que forman parte del corpus. Adicionalmente, el método demuestra un buen funcionamiento con las colecciones de los documentos de tamaño variable. No obstante, el método tESA aprovecha el corpus más grande. Los resultados obtenidos con el método tESA configurado con Medline superan

tanto las bases de referencias establecidas con el método original (ESA), como la del método anterior (DS).

Sobre todo, parece interesante considerar la comparación entre tESA y el método original ESA (configurado también con Medline). La ventaja de tESA en cuanto la calidad parece indicar que la incorporación de la semántica del dominio (biomédico) puede ser más compleja que coger un método del dominio general y configurarlo con el corpus más adecuado. En tESA se ha utilizado la observación que, un título de un artículo científico proporciona una descripción optimizada de los contenidos del dicho artículo. En ESA, la idea básica es que las entradas parecidas son las que co-ocurren mucho en los documentos mutuamente importantes para las entradas. En tESA esta idea básica se convierte en una formulación más relajada: las entradas parecidas son las que, a través de los contenidos de los documentos, ' activan' el mismo vocabulario del espacio de los títulos. Esta característica contribuye, a que tESA disminuya el impacto de la ortogonalidad del corpus. Cabe destacar que es una característica propia de las representaciones de tESA (que son generalmente más ligeras que las representaciones de ESA) y que el cálculo de la proximidad semántica prácticamente se resuelve de la misma manera, tanto en tESA, como en ESA, o sea calculando la similitud coseno entre las parejas de las representaciones de las entradas. Por lo tanto, tESA es eficiente en cuanto el rendimiento en línea, sobre todo en comparación con los otros métodos que reducen el impacto de la ortogonalidad del corpus. Además, el hecho, de que los vectores de tESA estén expresados sobre un espacio relativamente pequeño (siendo el vocabulario de los títulos de los documentos del corpus) hace que las representaciones de tESA sean más aplicables (que las del método original ESA) a los escenarios específicos, para los cuales este tipo de optimización es necesario, y sin tener que sacrificar mucho en cuanto a la calidad de la aproximación de la proximidad semántica.

La idea de extender el uso de la semántica del corpus enfocada en un dominio especializado, presente en tESA, ha sido también la clave en el proceso del diseño del tercer método – DomESA. DomESA es un método computacionalmente parecido a tESA, específicamente DomESA también extiende el método ESA con un paso adicional de multiplicar los vectores por una matriz. En DomESA los vectores de representaciones de las entradas se obtienen multiplicando los vectores del método original por una matriz de similitud entre los documentos del corpus especializado (Medline) y los del corpus del dominio general (Wikipedia). El diseño de DomESA extiende el aspecto disperso de tESA. En el método ESA original, dos entradas se consideran próximas, si aparecen juntas en los mismos documentos (siendo importantes dentro de los contenidos de esos documentos). En tESA las dos entradas se consideran próximas, si los documentos en los cuales aparecen las entradas comparten el mismo vocabulario en los títulos. En DomESA las dos entradas se consideran similares, si aparecen en conjuntos de documentos similares, específicamente, si los documentos en los cuales aparecen son similares a los mismos conceptos/artículos de Wikipedia.

DomESA proporciona resultados de correlaciones de Spearman por encima de los resultados obtenidos con los otros métodos contemplados en este estudio, solo siendo la segunda mejor opción detrás de tESA en un conjunto de referencia. En

cuanto los resultados de correlaciones de Pearson, DomESA sobrepasa a los otros métodos incluidos en la evaluación. En general, DomESA parece ser el método más consistente de todos los métodos contemplados aquí en cuanto la calidad de los resultados.

Aparte de los buenos resultados de la aproximación de la proximidad semántica, se puede observar que el uso de la semántica del corpus especializado en DomESA está restringido al problema de asignar los pesos a los conceptos/artículos del recurso del dominio general. Específicamente, los eventuales vectores de representación de las entradas están expresados sobre el espacio de los artículos de Wikipedia. Concretamente, en DomESA las representaciones son compatibles con las representaciones del método original (configurado con Wikipedia), con el procesamiento de DomESA afectando nada más que la distribución de los valores dentro de las representaciones. Este factor de compatibilidad potencialmente extiende el rango del uso de los métodos basados en las colecciones de los documentos, porque añade un mecanismo que proporciona la estimación de la proximidad semántica entre las entradas del dominio especializado con las entradas del dominio general. Adicionalmente, el hecho de que las entradas de DomESA estén representadas sobre el espacio de los artículos de Wikipedia, significa que estas representaciones pueden ser interpretadas usando una base del conocimiento estructurada, ya que se trata de unos vectores que a cada concepto de Wikipedia les tienen asignado un grado numérico de relevancia (o sea, un peso). Esta característica de los vectores de DomESA significa que potencialmente se puede extender DomESA de la misma manera que se ha extendido el método original.

Además, los resultados presentados en este estudio indican, que las representaciones usadas por DomESA no son mucho más 'grandes' que las del método ESA original configurado con Wikipedia y más ligeros que las de ESA configurado con Medline (en cuanto los elementos de los vectores con los valores diferentes a cero). Esto significa, que, exactamente como en el caso de tESA, el coste adicional implicado en el procesamiento de DomESA se paga solo a la hora de crear las representaciones, ya que es necesario el paso adicional de multiplicación por la matriz.

Adicionalmente, el trabajo presentado aquí parece indicar que superar (o reducir) la ortogonalidad del corpus mejora los resultados de los métodos derivados de ESA.

El último trabajo presenta a un experimento (junto con la evaluación de los resultados obtenidos) basado en aplicar DomESA al problema del alineamiento de las ontologías. Aunque los resultados proporcionados por el sistema propuesto – DisMatch - son inferiores a los de los mejores sistemas evaluados en la edición 2016 de la iniciativa, nuestro método ha destacado en la capacidad de descubrir las correctas relaciones únicas, o sea las relaciones correctas que no se habían descubierto con los otros sistemas. Los resultados de la evaluación indican que un módulo de la aproximación de la proximidad semántica podría ser un componente importante de un sistema complejo diseñado para resolver el problema del alineamiento de las ontologías. Los sistemas del alineamiento de las ontologías tradicionalmente usan las métricas basadas en las cadenas de textos para aproximar la proximidad entre los conceptos. Los resultados demostrados en el artículo, parecen indicar que la in-

clusión de un componente basado en el concepto de la proximidad semántica puede mejorar el rendimiento de un sistema complejo.

Además, se va a valorar la posibilidad de mejorar DomESA, a través de modificar la estrategia composicional de las representaciones del método, el algoritmo Word's Mover Distance siendo una de las opciones interesantes para considerar.

Chapter 1
Introduction

1.1 Background

Given the massive amount of new research that has been published in the life sciences in recent years, the scientific community needs solutions that lead to the creation of self-readable document repositories, which could automatically classify and provide structural representation of the knowledge expressed 'implicitly' (from a machine-based perspective) in the scientific articles. Achieving this goal would be an important step, that eventually could lead to improving current information access and retrieval methods, which are mostly based on keyword queries. The inadequacy of currently available tools and approaches has been mentioned in the domain literature, e. g. Bellazzi et al (2012). Establishing semantic relatedness between concepts or their textual representation is one of the key enabling components in automated knowledge extraction from texts, as many text processing applications need a numerical equivalent of how the concepts fit together. The ultimate goal of a relatedness measure is to assign a numerical approximation of the relatedness strength to a given pair of *inputs* (usually represented by *input texts*, especially in the context of lexical relatedness measures contemplated in this study), a simple black-box style overview is presented in Figure 1.1. According to Budanitsky (1999), successful applications of approximations of semantic relatedness include general domain tasks such as: word sense disambiguation (Agirre and Rigau, 1996), text summarization (Barzilay and Elhadad, 1997) and information retrieval (Rada et al, 1989). According to McInnes et al (2013), applications of semantic similarity (which is a more specific concept) and relatedness measures in life sciences include direct data analysis (discovery of protein – pathway interactions (Guo et al, 2006), discovering similar diseases (Mathur and Dinakarpandian, 2012)), semantic search (Sahay and Ram, 2011), redundancy detection in clinical records (Zhang et al, 2011), sense disambiguation (McInnes et al, 2011). Applications of semantic similarity to compare gene products have been reviewed by Pesquita et al (2009).

Most state-of-the-art measures use some kind of pre-existing knowledge base in order to produce a numerical approximation of semantic relatedness between two

concepts or their lexicalizations, as reflected in a relatively recent survey presented by Zhang et al (2012). This trend has also been reflected in a recent review presented by Couto and Pinto (2013). It can be argued that the notion of semantic relatedness is connected with such resources due to its concept-based nature (Budanitsky and Hirst, 2006). However, as pointed out by Zhang et al (2012); Turney and Pantel (2010); Agirre et al (2009), corpus-based methods have been used with some success as approximations of semantic relatedness. To the best of our knowledge, most state-of-the-art measures that are applied in life sciences make use of highly specialized knowledge-rich resources, which makes them barely suitable for scenarios in which the knowledge base does not exist, or is extremely large or subject to frequent changes. Furthermore, their methodologies are not easy to repeat for even a slight change of settings, as reflecting each change in settings in the knowledge resource requires substantial effort.

This study focuses on the problem of improving domain-focused, corpus-based measures of semantic relatedness, with the methods evaluated in the biomedical domain. As a result of the work presented in this study, three new methods for approximating semantic relatedness in domain-oriented settings have been proposed. The presentation is completed with an application of one of the measures to the problem of ontology alignment, which also provides another perspective on the added value of the contributions presented hereby. The scope of the work presented in this study can be motivated along the following dimensions: domain-focus, resources and domain choice. Domain-focused relatedness approximation is especially interesting, because, as opposed to the general-domain relatedness approximation, the problem seems relatively far from being resolved. Specifically, the results obtained in the domain-oriented scenarios are much worse than those of general domain. The focus on corpus-based methods is motivated by the fact that, given the current state of sharing scientific knowledge, free-text collections, although noisy, are the most complete and up-to-date source of knowledge, that can be used ‘as it is’ in many different possible usage scenarios. Finally, the focus on Life Sciences is principally motivated by the domain’s need for better methods for intelligent support in knowl-

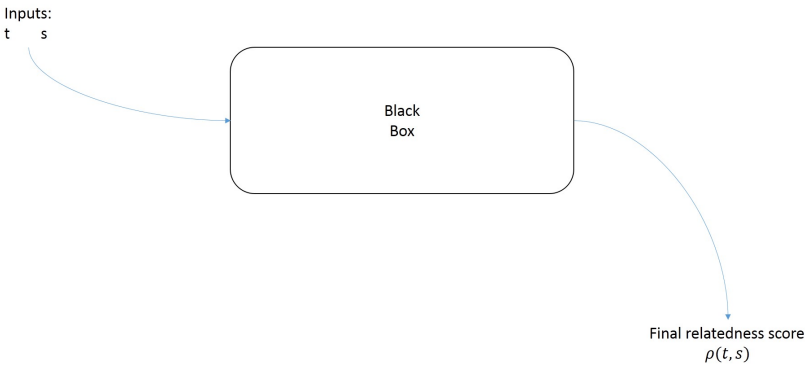

Fig. 1.1 A black box view of a relatedness approximation system.

edge processing. This need, in turn, results in an exceptional availability of materials that are key enabling factors for the research focused on the domain (corpora, reference datasets, etc.).

The rest of this chapter is organized as follows. In the next section the related work is presented and the scope of this study is defined w.r.t. the state-of-the-art. In the subsequent section the contributions of this study are presented, with the main focus on the presentation of important research questions and their links to the publications relative to this dissertation. In Section 1.4 an overview of the methodology is presented, with special focus on the setup of experiments included in the publications.

Chapter 2 includes the published manuscripts, as this dissertation is presented as a collection of published contributions. The articles presented in Sections 2.1-2.4 present original research and methods focused on improving the quality of corpus-based semantic relatedness approximation within the biomedical domain. The paper included in Section 2.5 demonstrates a direct application of one of the methods to the problem of biomedical ontology alignment, i.e. to finding concept-to-concept equivalence mappings between pairs of ontologies.

Then the next chapter contains a global overview of the results obtained in published work, together with a brief discussion of these results. The final chapter presents the main conclusions drawn from the research presented in this study, with the main focus on those aspects related to the research questions presented in Section 1.3.

The additional, unnumbered chapter presents the conclusions (contents of Chapter 4) in Spanish.

1.2 Related work

The methods for calculating semantic relatedness can be roughly divided into two main groups: those that rely entirely on a specialized and structured knowledge-rich resource (Batet et al, 2011; Budanitsky and Hirst, 2006; Cross, 2004), and distributional measures that rely on implicit statistical features of a large document collection (Sahami and Heilman, 2006; Landauer et al, 1998).

The basic idea, reflected in the methods presented in this dissertation, of leveraging the distributional hypothesis, that words with similar contexts will have similar meanings ('you shall know a word by the company it keeps' (Firth, 1957)), has been explored thoroughly in the field of distributional semantics research. Techniques vary from basic word co-occurrence matrices, where a vector associated with a given word is created by counting the words that appear in its immediate context (determined by a window of a certain size), to latent semantic indexing (Dumais, 2004). More recently, dense word representations derived through machine learning techniques, e.g. word2vec methods (Mikolov et al, 2013), have gained popularity. The label of 'distributional semantics' covers an extremely wide scope of methods and models that share the feature conveyed in the label itself — semantics being

modeled as a distribution over a large set of features derived from a large body of linguistic evidence (i.e. a large corpus of documents).

With the increased popularity of using Wikipedia as a Knowledge Base (KB) for semantic relatedness estimation the aforementioned division has become less clear, as Wikipedia combines the features of both worlds. It does implicate a structure, as it comprises a set of topic-oriented and categorized entries, which are also interconnected with hyperlinks. It can also be treated as a large collection of documents, as it contains over 2M articles of at least 150 words each.

ESA (Explicit Semantic Analysis), by Gabrilovich and Markovitch (2007), relies on associating words with vectors of relevance, expressed over the space of Wikipedia articles. Numerous extensions of ESA have already been proposed, many of which combine the original approach with the Wikipedia-specific features, through the concept-to-concept feature/similarity matrices, e.g. the works by Scholl et al (2010); Polajnar et al (2013); Haralambous and Klyuev (2013). Some of those extensions, e.g. NESA (Non - Orthogonal ESA) by Asooja et al (2015), also provide variants that are generic enough to be used with any document collection.

There are Wikipedia-based methods for approximating semantic relatedness other than ESA, such as WikiRelate (Strube and Ponzetto, 2006). It is worth noting however, that in the case of algorithms, which by default rely on Wikipedia-specific features (such as link structure), incorporating semantics of an unstructured domain-focused corpus is a much more difficult task.

There is also a significant number of papers that weigh in on the application of established methods within biomedical settings. Muneeb et al (2015) explore the performance of the aforementioned word2vec and GloVe (Pennington et al, 2014) methods with models trained on a large biomedical corpus. Similar, but more detailed experiments are presented by Pakhomov et al (2016) and Chiu et al (2016). In a recent paper by Sajadi et al (2015), the authors present an original hybrid node ranking based method, based on the Wikipedia structure, applied in biomedical settings. Furthermore they provide an extensive evaluation of their method in comparison with other state-of-the-art methods, which constitutes an important reference perspective.

Finally, there is a large group of methods for the approximation of biomedical semantic relatedness, which rely, to different extents, on the use of structured domain specific resources. In the papers by Pedersen et al (2007); Liu et al (2012), the authors propose extracting concept representations from a large biomedical corpus document, in a process, which is guided by a structured knowledge resource. The methods rely on 'scanning' the corpus for words co-occurring with words associated with specific concepts of the knowledge source. Sánchez and Batet (2011) showcase the performance of a wide spectrum of ontology-based Information Content (IC) methods, which use SNOMED CT as the knowledge resource. SNOMED CT is the largest medical vocabulary collection, with over 400K systematically organized concepts with their lexical representations and additional information. The IC measures presented by Sánchez and Batet (2011) use the ontological structure of SNOMED (positions of concepts in the ontology, distance between them, number of sub-concepts, etc.) to compute a semantic score between a pair of concepts.

Martinez-Gil (2016) proposes a fuzzy logic-based method for combining scores achieved by the ontology-based methods. Our methods, although computationally dependent on a specific corpus[1], do not rely on high-level KB representations of the domain, which makes them more flexible and easier to adapt to non-standard use cases.

It can be argued that the notion of semantic relatedness is connected inherently with structured resources due to its concept-based nature (Budanitsky and Hirst, 2006), i.e. that the concept of semantic relatedness has to do with structurally represented knowledge, rather than with the statistics of a collection of documents. This argument is also raised by Gabrilovich and Markovitch (2007). This means, that, in strict terms, the distributional, corpus-based measures can only be used as proxies for semantic relatedness approximation, *de facto* being measures of relatedness between texts. Furthermore, these measures of relatedness between texts (or words) depend on the linguistic resource (typically a corpus) they are set up with. However, as pointed out by Zhang et al (2012); Turney and Pantel (2010); Agirre et al (2009), the corpus-based methods have been used with some success as approximations of semantic relatedness of concepts. This can be attributed to the fact that the concepts can also be represented by, for example, associated textual labels (in the simplest scenario). In this study, the focus is on the lexical proxies of semantic relatedness[2]. The evaluation presented here seems to confirm both the validity and usefulness of the chosen approach.

In this study, a direct evaluation of the proposed methods for relatedness calculation is presented, i.e. the methods are primarily evaluated against reference datasets that capture the scores assigned manually to pairs of textual inputs by human annotators. We rely on a collection of reference datasets that have become a standard in the evaluation of biomedical semantic relatedness (Pedersen et al, 2007; Pakhomov et al, 2010, 2011).

It is worth noting, that the sub-problem of producing a semantic representation of a given input text is not unique to the problem of textual semantic relatedness, and has been widely addressed in the field of semantic annotation (or semantic indexing). In the semantic annotation problem, an input text (typically longer than an input for the semantic relatedness problem, e.g. a scientific article) gets assigned to a representation expressed over a space of concepts of a knowledge resource. Especially methods, which treat the problem of assigning the representations as an information retrieval (i.e. concept retrieval) task, bear a certain resemblance to methods like ESA. Berlanga et al (2010) postulate using an IC-based method to assign UMLS concepts to biomedical texts. ESA is used directly by Ślezak et al (2011) for a similar task of annotating biomedical documents with Wikipedia concepts. Although the problem of obtaining the representations is similar, the semantic indexing systems are subsequently evaluated differently, either by the applicability

[1] This dependence practically means that they rely on a presence of a sufficiently large (in terms of semantic coverage, information) corpus that can be used to represent the domain.

[2] It could be argued, that some of the methods proposed in this study provide semantics identical to those of ESA (where input texts are expressed over a vector of relevance weights over the space of Wikipedia concepts/articles).

of the resulting annotations to semantic search/retrieval problem, their usability in organizing the collection (e.g. semantic clustering), or the capacity of the systems to reproduce the annotations of the human annotators.

The experiment on ontology alignment presented in Section 2.5 is included here principally to illustrate a potentially interesting application of lexical corpus-based semantic relatedness measures (and is discussed as such), so the general problem of ontology matching is beyond the scope of this study. Nevertheless, it can be observed, that the potential of using textual resources as a source of background knowledge for the task of mapping classes between pairs of ontologies has not been researched thoroughly, with the exception of several systems that use Wikipedia-based relatedness measures. Notably, CIDER-CL system (Gracia and Asooja, 2013) uses CL-ESA (Potthast et al, 2008), a cross lingual extension of ESA, in its lexical matching process for concepts described in different languages. Using an ESA variant for lexical matching is the feature, that CIDER-CL has in common with the approach presented in Section 2.5, while the general composition of the matching strategies differs substantially between the two systems. In our approach we use the semantic relatedness calculations combined with the classic Similarity Flooding algorithm (Melnik et al, 2002). The evaluation has been performed within the *Phenotype* track of the Ontology Alignment Evaluation Initiative 2016 campaign (Achichi et al, 2016; Harrow et al, 2016).

1.3 Contributions

1.3.1 Papers

The article included in Section 2.2 describes an approach towards approximating lexical semantic relatedness based on representing the input texts with vectors of the key vocabulary extracted from a relatively small sample of the most relevant documents of a given corpus. The paper presents a side-by-side analysis of two methods for extracting the vocabulary – truncated Tf-Idf vectors and most frequently appearing word sequences that follow a certain grammatical pattern. Both methods are evaluated against the best scores reported (at that point) in the literature. The paper also presents a use case of a direct application of the relatedness measures to data analysis, i.e. to the problem of finding pairs of semantically related entities within a given set of concepts (in this particular case, the concepts were extracted from *Orphanet*[3]). The article included in Section 2.1 describes the pilot experiment that eventually led to the extended evaluation presented in Section 2.2. The pilot experiment was also presented at the poster session of Nettab 2013 workshop.

The article in Section 2.3 describes tESA – an approach for approximating semantic relatedness in biomedical domain, which extends the ESA method by rep-

[3] Orphanet is a database, which represents knowledge on rare/orphaned diseases. See: www.orpha.net for more information.

resenting inputs with vectors obtained by multiplying the ESA-style vectors with a column-based matrix of document titles. The method improves the original method through by-passing the corpus orthogonality through the use of the semantics of the titles of the scientific articles. Furthermore, the paper also presents a (most) detailed evaluation of the original ESA method in biomedical settings, i.e. the original method is evaluated in setups involving major biomedical corpora and against the full array of relevant reference datasets.

The paper included in Section 2.4 presents another extension of ESA, which is computationally close to the approach presented in Section 2.3. Specifically, the study concentrates on improving the original ESA method in biomedical settings by combining the semantics of two corpora – multiplying ESA-style vectors obtained with a domain focused corpus with a column matrix of document-to-document similarities obtained through pairwise comparisons of domain-focused (here: Medline) and general-domain corpus elements (here: Wikipedia articles).

Finally, the paper in Section 2.5 describes an approach towards incorporating the lexical semantic relatedness approximation into the problem of ontology alignment. The proposed system, DisMatch, combines the lexical relatedness-based mapping with a structural mapping strategy (Similarity Flooding). The goal of the experiment was to check whether a relatedness-driven approach provides improvements over the most typically used methods. The evaluation presented in the paper has been carried out externally within the OAEI 2016 campaign. DisMatch was evaluated in a track dedicated to alignment of phenotype ontologies, which involved two pairwise alignment tasks of large biomedical ontologies.

The papers included in Sections 2.1 – 2.4 introduce three domain-focused methods for calculating lexical semantic relatedness based on the use of textual resources. The paper included in Section 2.5 describes an experiment, which involves using the method described in Section 2.4. Each of the proposed methods address some of the research questions outlined below. Table 1.1 presents a short overview of the published articles. Table 1.2 presents a summary of the implemented methods and systems.

Table 1.1 Summary of the work presented with the dissertation.

Section	Reference	Venue	Type
2.1	Rybiński and Aldana-Montes (2013)	EMBnet.journal	Extended abstract
2.2	Rybiński and Aldana-Montes (2014)	BMC Bioinformatics	Research paper
2.3	Rybiński and Aldana-Montes (2016)	JBMS	Research paper
2.4	Rybiński and Aldana-Montes (2017)	JIIS	Research paper
2.5	Rybiński et al (2016)	OM-2016	Results report (workshop)

Navas-Delgado et al (2015) present an integrative effort to create a metabolic pathway database. Although the semantic integration of the resources was an important aspect of the paper, the integration itself did not involve the approximation

Table 1.2 Summary of the original relatedness approximation methods presented within the published papers.

Reference	Method	Input represented with
Rybiński and Aldana-Montes (2013, 2014)	DS	Words extracted from the best-fit documents
Rybiński and Aldana-Montes (2016)	tESA	Relevance-weighted words of the document titles
Rybiński and Aldana-Montes (2017)	DomESA	Relevance-weighted Wikipedia concepts

of the semantic relatedness, so the contribution is beyond the scope of the work presented here.

1.3.2 Research questions

As already mentioned, the aim of this study was to research ways of improving corpus-based methods for approximating semantic relatedness for domain-oriented use. Therefore, the scope was narrowed down to algorithms that only take advantage of collections of textual resources. The research presented here was undertaken with the goal of answering the following, main questions:

Q1 Is it enough to use roughly approximated contexts to model input texts for the task of relatedness approximation?
Q2 Can the quality of the methods be improved in the domain-focused setup by incorporating more information?
Q3 How to maximize the benefit of incorporating domain-focused semantics?
Q4 What are the possible new benefits of using corpus-based semantic relatedness?

The main questions can be roughly associated with the contributions[4] included in this study. Q1 is covered in the articles included in Sections 2.1 and 2.2. An attempt to answering Q2 is presented in the article included Section 2.3. The method introduced in Section 2.4 is directly related to Q3.

The methods for approximation of lexical semantic relatedness were introduced in the publications included in Sections 2.2 (with the publication included in Section 2.1 describing the related preliminary work), 2.3 and 2.4. The paper included in Section 2.5 presents an experimental system for ontology matching, based on the method presented in Section 2.4. The ontology matching experiment was evaluated externally and it provides interesting insights into both the promising aspects and limitations of applications of the corpus-based semantic relatedness approximation methods (Q4). Table 1.3 shows the relationships between papers and the main research questions.

[4] Contributions here mean the published works included in Chapter 2

Table 1.3 Summary of the relationships between the published papers and the main research questions.

Reference	Research question(s)
(Rybiński and Aldana-Montes, 2014)	Q1, Q4
(Rybiński and Aldana-Montes, 2016)	Q2
(Rybiński and Aldana-Montes, 2017)	Q3, Q2
(Rybiński et al, 2016)	Q4

Moreover, the main research questions Q1-Q3 can be associated with several subproblems, which were also addressed in the corresponding publications, namely the following:

1.a Optimizing the size of the representation of the input texts.
1.b Optimizing the process of acquisition of the representation of the input texts.
2.a Benefits of using a larger corpus.
2.b Making the best of using shorter documents.
2.c Establishing a baseline score obtained with explicit semantic analysis (ESA) set up with a domain-focused corpus.
3.a Improving the robustness (w.r.t. the changes of parameters, reference datasets).
3.b Incorporating domain-focused semantics, while maintaining general domain syntactic compatibility.
3.c Improving the'linear' correlation with model scores.

1.4 The methodology overview

As mentioned in the Related Work section, the methods for approximating semantic relatedness were evaluated with a set of reference datasets, that have become a *de facto* standard for evaluating the relatedness measures in biomedical settings. The summary of the characteristics of each of the reference datasets is presented in Table 1.4.

Each of the datasets contains pairs of inputs with an associated human-assigned relatedness score. The relatedness scores included in the datasets were calculated as the average between answers supplied by a given number of human annotators.

It can be observed that all but one of the reference datasets are focused on semantic relatedness. The umnsrsSim dataset, however, is focused on the notion of semantic similarity, which is a narrower concept. Specifically, similarity will not include some types of relations, e.g. antonymy or meronymy, while relatedness includes all the possible relations. For example, *scalpel* will be semantically related to *tonsillectomy*, while not being semantically similar. Nonetheless, the similarity-oriented dataset can still be used in the evaluation presented here, as none of the evaluated methods is capable of distinguishing between relationship types, so they are all equally 'handicapped' w.r.t. the notion of similarity. Furthermore, in practical

Table 1.4 Summary of the features of the reference datasets. The features are: the number of evaluated input pairs, the number of distinct individual inputs (referred to as 'items'), the area of focus, the profile of the annotators and the corresponding literature reference.

Dataset	No. of pairs	No. of items	Focus	Annotators	Reference
umnsrsSim	566	375	Similarity	Med. residents	Pakhomov et al (2010)
umnsrsRel	587	397	Relatedness	Med. residents	Pakhomov et al (2010)
mayo101	101	191	Relatedness	Med. coders	Pakhomov et al (2011)
mayo29c	29	56	Relatedness	Med. coders	Pedersen et al (2007)
mayo29ph	29	56	Relatedness	Physicians	Pedersen et al (2007)

applications it is often difficult to separate the two concepts. As a result, relatedness approximation methods may be used to approximate similarity and vice versa.

The task of the system is to replicate the average scores assigned by the human annotators and the success of a system is measured by how close its scores are to the human-assigned average scores. This 'closeness' of the automatically assigned scores to the model scores can be represented by a correlation coefficient between the respective sets of scores. Most studies advocate using Spearman's rank correlation for this purpose.

Using Spearman's rank correlation can be justified by the fact that rating an entire set of input pairs is actually ranking them from the most to the least related. In most of the papers presented in this study, the methods are only evaluated with the Spearman's rank correlation (i.e. the methods presented in Sections 2.2 and 2.3 are only evaluated with the rank correlations).

Some sources, however (e.g. Muneeb et al (2015); Martinez-Gil (2016)), present the evaluation of relatedness approximation methods with the success rate measured as the Pearson's correlation coefficient with the model score, rather than the Spearman's rank correlation coefficient. In simple terms, an evaluation that uses Pearson's correlation coefficient takes into account whether the correlation is linear, rather than monotonic. To provide a fuller perspective on the performance, the method presented in Section 2.4 is evaluated with both correlation coefficients.

The evaluations include comparative experiments with relevant state-of-the-art measures, as well as with the related previous work. For this purpose, the paper included in Section 2.4 also presents an evaluation of the tESA method (Sec. 2.3) with Pearson's correlation coefficient. Table 1.5 presents an example reference dataset with scores generated by the DomESA method presented next to the model scores.

Regardless of the actual method of calculating correlations, comparing the results obtained with the different methods effectively involves comparing the correlations those methods generate w.r.t. the model answers. For this reason, the methods, for which the quality improvements have been claimed over a given baseline [5], are presented together with the tests of the statistical significance of the presented findings.

We evaluate the statistical significance of correlation comparisons using a methodology presented by Zou (2007). Specifically we construct a 0,95 confidence level

[5] I.e. tESA, described in section 2.3, and DomESA, described in section 2.4.

Table 1.5 Pair-by-pair list of human and DomESA generated scores for the mayo29c reference dataset; the output of DomESA accounts for correlation scores of 0.84 (Spearman's correlation coefficient) and 0.863 (Pearson's correlation coefficient).

Input 1	Input 2	Model score	DomESA score
Renal failure	Kidney failure	4	0.89
Abortion	Miscarriage	3.3	0.35
Heart	Myocardium	3	0.52
Stroke	Infarct	2.8	0.26
Delusion	Schizophrenia	2.2	0.36
Calcification	Stenosis	2	0.43
Tumor metastasis	Adenocarcinoma	1.8	0.43
Congestive heart failure	Pulmonary edema	1.4	0.4
Pulmonary fibrosis	Malignant tumor of lung	1.4	0.08
Diarrhea	Stomach cramps	1.3	0.12
Mitral stenosis	Atrial fibrillation	1.3	0.18
Brain tumor	Intracranial hemorrhage	1.3	0.1
Antibiotic	Allergy	1.2	0.02
Pulmonary embolus	Myocardial infarction	1.2	0.07
Carpal tunnel syndrome	Osteoarthritis	1.1	0.05
Rheumatoid arthritis	Lupus	1.1	0.27
Acne	Syringe	1	0.01
Diabetes mellitus	Hypertension	1	0.08
Cortisone	Total knee replacement	1	0.03
Cholangiocarcinoma	Colonoscopy	1	0.06
Lymphoid hyperplasia	Laryngeal cancer	1	0.12
Appendicitis	Osteoporosis	1	0.01
Depression	Cellulitis	1	0.01
Hyperlipidemia	Tumor metastasis	1	0.01
Multiple sclerosis	Psychosis	1	0.05
Peptic ulcer disease	Myopia	1	0
Rectal polyp	Aorta	1	0.01
Varicose vein	Entire knee meniscus	1	0.01
Xerostomia	Alcoholic cirrhosis	1	0.02

confidence intervals (CI) for dependent overlapping correlations (as for a pair of methods, both of them produce their correlation against the same reference dataset). This test allows us to refute, under the assumed confidence level, the *null hypothesis* of the two correlations being equal[6].

Before describing the external OAEI evaluation process that DisMatch system was subject to, it is best to first, briefly, define the simplified version of the alignment problem addressed by the system. Given two ontologies O_1 and O_2, an alignment system produces a set of mappings between the classes of the respective ontologies. Each mapping, therefore, consists of a pair of classes (with one class of O_1 and the other of O_2) and the type of relationship between the classes. The relationship type in the case of the system presented here is restricted to equivalence, which simpli-

[6] See https://seriousstats.wordpress.com/2012/02/05/comparing-correlations/ for discussion and code.

fies the problem even further. In layman's terms, the system takes two ontologies and decides, which of their classes/concepts are equivalent, to produce a list of the equivalence pairs.

The OAEI evaluation process is set in motion for a given task (here the *Phenotype* track) and it involves the mappings provided by all the systems participating in the task. Precision[7] and recall[8] of each system are approximated with respect to a, so called, silver standard, according to which a mapping is considered correct, if it appears in at least 2 (silver 2) or 3 (silver 3) sets of mappings supplied by different systems. A sample of the 'unique' mappings (considered incorrect by the silver standard) produced by each of the systems has been evaluated manually to approximate the number of correct unique mappings supplied by each of the systems (regardless of the consensus ones from the silver standard). The performance of the systems is evaluated additionally with recall calculated w.r.t. a small subset of manually created mappings[9].

The specific resources and methods of the systems presented here are described in detail, in the respective scientific contributions.

[7] A percentage of correct mappings among those supplied by the system.

[8] A percentage of all the correct mappings included in the output of a system.

[9] An overview of the evaluation of the *Phenotype* track is presented and explained in more detail here: http://oaei.ontologymatching.org/2016/results/phenotype/.

Chapter 2
Published works

2.1 SEBSem: simple and efficient biomedical semantic relatedness measure

SEBSem: Simple And Efficient Biomedical Semantic Relatedness Measure

Maciej Rybiński and José F.Aldana-Montes

Dept. de Lenguajes y Ciencias de la Computación, University of Malaga,
ETSI Informática, Campus de Teatinos, Malaga - 29071, Spain

Abstract. Calculating semantic relatedness between terms is crucial in numerous knowledge and information processing tasks highly relevant to the biomedical domain. Examples include semantic search and automated processing of scientific texts. Most available methods rely heavily on highly specialised resources, which substantially limits their reusability in various applications within the domain. In this work we present a simple semantic relatedness measure that relies only on very general resources and its design features allow minimising the costs of online computations. The relatedness is computed through comparing automatically extracted key-phrases relevant to respective input terms. This simple strategy provides a method that gives promising early test results, comparable to those of human annotators and state-of-the-art methods, on a well established benchmark.

Keywords: bioinformatics; semantic relatedness; semantic similarity; knowledge extraction

1 Full reference

Rybinski, Maciej, and José Francisco Aldana-Montes. "SEBSem: simple and efficient biomedical semantic relatedness measure." *EMBnet. journal* 19.B (2013): pp-82.

2 DOI

http://dx.doi.org/10.14806/ej.19.B.738

2.2 Calculating semantic relatedness for biomedical use in a knowledge-poor environment

Calculating semantic relatedness for biomedical use in a knowledge-poor environment

Maciej Rybiński and José F.Aldana-Montes

Dept. de Lenguajes y Ciencias de la Computación, University of Malaga,
ETSI Informática, Campus de Teatinos, Malaga - 29071, Spain

Abstract. Background
Computing semantic relatedness between textual labels representing biological and medical concepts is a crucial task in many automated knowledge extraction and processing applications relevant to the biomedical domain, specifically due to the huge amount of new findings being published each year. Most methods benefit from making use of highly specific resources, thus reducing their usability in many real world scenarios that differ from the original assumptions. In this paper we present a simple resource-efficient method for calculating semantic relatedness in a knowledge-poor environment. The method obtains results comparable to state-of-the-art methods, while being more generic and flexible. The solution being presented here was designed to use only a relatively generic and small document corpus and its statistics, without referring to a previously defined knowledge base, thus it does not assume a 'closed' problem.
Results
We propose a method in which computation for two input texts is based on the idea of comparing the vocabulary associated with the best-fit documents related to those texts. As keyterm extraction is a costly process, it is done in a preprocessing step on a 'per-document' basis in order to limit the on-line processing. The actual computations are executed in a compact vector space, limited by the most informative extraction results. The method has been evaluated on five direct benchmarks by calculating correlation coefficients w.r.t. average human answers. It also has been used on Gene - Disease and Disease- Disease data pairs to highlight its potential use as a data analysis tool. Apart from comparisons with reported results, some interesting features of the method have been studied, i.e. the relationship between result quality, efficiency and applicable trimming threshold for size reduction. Experimental evaluation shows that the presented method obtains results that are comparable with current state of the art methods, even surpassing them on a majority of the benchmarks. Additionally, a possible usage scenario for the method is showcased with a real-world data experiment.
Conclusions
Our method improves flexibility of the existing methods without a notable loss of quality. It is a legitimate alternative to the costly construction of specialized knowledge-rich resources.

Keywords: Bioinformatics; Semantic relatedness; Semantic similarity; Distributional linguistics; Knowledge extraction

1 Full reference

Rybinski, Maciej, and José Francisco Aldana-Montes. "Calculating semantic relatedness for biomedical use in a knowledge-poor environment." *BMC bioinformatics* 15.14 (2014): S2.

2 DOI

10.1186/1471-2105-15-S14-S2

2.3 tESA: a distributional measure for calculating semantic relatedness

tESA: a distributional measure for calculating semantic relatedness

Maciej Rybiński and José F.Aldana-Montes

Dept. de Lenguajes y Ciencias de la Computación, University of Malaga,
ETSI Informática, Campus de Teatinos, Malaga - 29071, Spain

Abstract. Background

Semantic relatedness is a measure that quantifies the strength of a semantic link between two concepts. Often, it can be efficiently approximated with methods that operate on words, which represent these concepts. Approximating semantic relatedness between texts and concepts represented by these texts is an important part of many text and knowledge processing tasks of crucial importance in the ever growing domain of biomedical informatics. The problem of most state-of-the-art methods for calculating semantic relatedness is their dependence on highly specialized, structured knowledge resources, which makes these methods poorly adaptable for many usage scenarios. On the other hand, the domain knowledge in the Life Sciences has become more and more accessible, but mostly in its unstructured form - as texts in large document collections, which makes its use more challenging for automated processing. In this paper we present tESA, an extension to a well known Explicit Semantic Relatedness (ESA) method.

Results

In our extension we use two separate sets of vectors, corresponding to different sections of the articles from the underlying corpus of documents, as opposed to the original method, which only uses a single vector space. We present an evaluation of Life Sciences domain-focused applicability of both tESA and domain-adapted Explicit Semantic Analysis. The methods are tested against a set of standard benchmarks established for the evaluation of biomedical semantic relatedness quality. Our experiments show that the propsed method achieves results comparable with or superior to the current state-of-the-art methods. Additionally, a comparative discussion of the results obtained with tESA and ESA is presented, together with a study of the adaptability of the methods to different corpora and their performance with different input parameters.

Conclusions

Our findings suggest that combined use of the semantics from different sections (i.e. extending the original ESA methodology with the use of title vectors) of the documents of scientific corpora may be used to enhance the performance of a distributional semantic relatedness measures, which can be observed in the largest reference datasets. We also present the impact of the proposed extension on the size of distributional representations.

Keywords: Bioinformatics; Semantic relatedness; Semantic similarity; Distributional linguistics; Knowledge extraction; Explicit semantic analysis; Biomedical semantics

1 Full reference

Rybinski, Maciej, and José Francisco Aldana-Montes. "tESA: a distributional measure for calculating semantic relatedness." *Journal of biomedical semantics* 7.1 (2016): 67.

2 DOI

10.1186/s13326-016-0109-6

2.4 DomESA: a novel approach to extended domain-oriented lexical relatedness calculations with domain-specific semantics

DomESA: a novel approach for extending domain-oriented lexical relatedness calculations with domain-specific semantics

Maciej Rybiński and José F.Aldana-Montes

Dept. de Lenguajes y Ciencias de la Computación, University of Malaga,
ETSI Informática, Campus de Teatinos, Malaga - 29071, Spain

Abstract. Being able to correctly model semantic relatedness between texts, and consequently the concepts represented by these texts, has become an important part of many intelligent information retrieval and knowledge processing systems. The need for such systems is especially evident within the biomedical domain, where the sheer amount of scientific publishing contributes to an information overflow. In this paper we present a novel method to approximate semantic relatedness in domain-focused settings. The approach is an extension to a well-known ESA (Explicit Semantic Analysis) method. Our extension successfully leverages the semantics of a domain-specific document corpus. We present the evaluation of the proposed method on a set of reference datasets, that are a de facto reference standard for the task of approximating biomedical semantic relatedness. The proposed method is evaluated in comparison with other state-of-the-art methods, as well as the baselines established with the original ESA method. The results of the experiments suggest that the proposed method combines the semantics of a general and domain-specific corpora to provide significant improvements over the original method.

Keywords: Semantic relatedness; Biomedicine; Distributional linguistics; Semantic similarity; ESA; Text analytics

1 Full reference

Rybiński, Maciej, and José Francisco Aldana Montes. "DomESA: a novel approach for extending domain-oriented lexical relatedness calculations with domain-specific semantics." *Journal of Intelligent Information Systems*: 1-17.

2 DOI

0.1007/s10844-017-0442-y

2.5 DisMatch results for OAEI 2016

DisMatch results for OAEI 2016

Maciej Rybiński, María del Mar Roldán-García, José García-Nieto, and José F.Aldana-Montes

Dept. de Lenguajes y Ciencias de la Computación, University of Malaga, ETSI Informática, Campus de Teatinos, Malaga - 29071, Spain

Abstract. DisMatch is an experimental ontology matching system based on the use of corpus based distributional measure for approximating semantic relatedness. Through the use of a domain-related corpus, the measure can be applied to a problem focused on the domain of the corpus, here being the Disease and Phenotype track. In this paper, we aim to briefly present the proposed approach and the results obtained in the evaluation, as well as some early conclusions regarding the performance of DisMatch.

Keywords: Ontology Matching, Bench-marking, Lexical Semantic Relatedness

1 Full reference

Rybiński, Maciej, et al. "DisMatch results for OAEI 2016." *OM@ ISWC*. 2016.

Chapter 3
Results and Discussion

In this chapter a global overview of the results obtained in the published contributions related with this study is presented. An overall perspective on all of the semantic relatedness approximation algorithms proposed is shown.

Tables 3.1 and 3.2 present an overview of best-case scenario (optimal parameter combination) Spearman's and Pearson's correlation coefficients obtained by the respective methods in the evaluation performed over the range of reference data sets presented in Chapter 1.

It can be noted that each of the methods provides a range of scores over the datasets used in the evaluation, i.e. each methods performs better on some reference datasets and worse on the others. This is due to the fact that some for some datasets it is 'harder' to order the input pairs 'correctly' (accordingly to the mean human scores) by relatedness than in others. Specifically, the mayo29 datasets are of high inter-annotator-agreement (all the annotators were close on the consensual ordering), so the problem is relatively easier than in the case of the other three datasets – the 'correct' ordering was more intuitive to human annotators, so it should also be more attainable to the automated approaches (which is reflected in the scores of the evaluated systems).

Table 3.1 Spearman's correlation results obtained for different reference datasets.

	umnsrsSim	umnsrsRel	mayo101	mayo29c	mayo29ph
ESA (Wiki)	0.501	0.501	0.549	0.722	0.822
ESA (Medline)	0.621	0.608	0.547	0.734	0.835
tESA (Medline titles)	0.639	**0.649** (G)	0.549	0.687	0.783
Word2vec CBOW	0.529	0.454	0.416	0.757	0.757
DS (best)	0.58	0.54	0.6	**0.85**	0.76
DS (average)	0.55	0.52	0.56	0.79	0.7
DomESA	**0.691** (G)	0.63	**0.708** (G)	0.84	**0.881** (G)
Reference score	0.66	0.601	0.63	0.9 (G)	0.84
Reference	McInnes and Pedersen (2016)	Chiu et al (2016)	Sajadi et al (2015)	Sánchez and Batet (2011)	Pedersen et al (2007)
Method type	Vector	Word2vec	Word2vec	ICC	Vector

The results of the DS method are presented for the best case scenario and an average of results calculated over a range of parameter values. It is worth noting that

Table 3.2 Pearson's correlation results obtained for different reference datasets.

	umnsrsSim	umnsrsRel	mayo101	mayo29c	mayo29ph
ESA (Wiki)	0.342	0.282	0.429	0.709	0.757
ESA (Medline)	0.274	0.228	0.361	0.749	0.711
tESA (Medline titles)	0.391	0.381	0.36	0.796	0.79
Word2vec CBOW	0.57	0.473	0.454	0.744	0.805
DomESA	**0.581** (G)	**0.508** (G)	**0.682** (G)	**0.863**	**0.889**

the best case results in DS are obtained with different parameter combinations for distinct reference datasets, so the average results provide a more realistic estimate of the method's performance. Additionally, it can be observed that the DS method has not been evaluated in terms of Pearson's correlation coefficient, neither has the average representation vector size been measured in the course of the experiments presented in the published works. For this reason, we provide an additional experiment with the DS method set up to work reasonably well with the optimal PMC corpus (which contains full texts of scientific articles). Its results and parameters are presented in Table 3.3.

Table 3.3 Additional results obtained for the DS method. These results were obtained for $N = 200$, $c = 0.005$

	umnsrsSim	umnsrsRel	mayo101	mayo29c	mayo29ph
Spearman's correlation	0.6	0.539	0.594	0.756	0.736
Pearson's correlation	0.416	0.331	0.23	0.806	0.694
Vector avg. size	334.6	327.4	326.4	320.8	320.8

The results presented in Table 3.3 are different from those reported in the original paper. This is due to the change in the document corpus, i.e. the PMC Open Access has grown from around 400k to more than 1 million articles.

Apart from the original methods presented in this study, the evaluation has been completed with relevant state-of-the-art methods for approximating lexical semantic relatedness: ESA and word2vec-CBOW. Additionally, for each of the reference datasets, a top reported rank relatedness score is included. The reported score is the highest score reported in the literature (to the best of the knowledge of the author). The top reference scores are only presented for the rank correlation results, as all the methods of comparable quality were only evaluated w.r.t. the rank correlation. The method type column value 'vector' denotes methods, which use second order co-occurrence vectors, while ICC denotes a method based on the information content.

It can be observed, that DomESA provides the best overall results, quality-wise, across the range of reference data sets, only being outperformed slightly by tESA on the *unmsrsRelatedness* data set and very slightly on the small *mayo29c* dataset by two other methods (the latter being statistically insignificant, at least in the case of

the DS method). Furthermore, its results seem more 'similar' to human judgment, as it achieves the best scores in terms of Pearson's correlation coefficient across the board, which means that the scores produced by DomESA are the most linearly similar to those produced by human annotators on all of the reference data sets (among the methods included in the experimental evaluation carried out within this study).

tESA provides second-best overall results in terms of ranking correlation (behind DomESA), scoring especially well in the experiments involving the larger datasets. tESA also provides third-best results in terms of Pearson's correlations (being outperformed by DomESA and CBOW).

In comparison to the state-of-the-art methods, the algorithms presented in this study surpass the best reported results by a considerable margin on the larger reference datasets. In terms of Spearman's rank correlation coefficient tESA provides the overall best result on the *unmsrsRelatedness* dataset, while DomESA provides globally best correlations on the *unmsrsSimilarity* and *mayo101* datasets. For the smaller reference datasets, the methods provide results comparable to the best reported scores, with the difference being almost insignificant given the size of these reference datasets. Furthermore, the Pearson's correlation coefficients achieved by the DomESA method are significantly higher than those obtained by other state-of-the-art methods, for all of the reference datasets.

The good performance of the methods presented here on the larger datasets can be (to a certain extent) explained by the fact that their semantic representations are derived from a corpus of a fairly general scope, thus being independent from specific views of the domain. This 'average' domain knowledge works well for approximating consensus mean annotator scores, independently of the lower inter-annotator-agreement inherent in the larger datasets.

The DS method is capable of achieving results comparable to the ESA (Medline) baseline, however it does need a certain degree of fine tuning. When used with consensus parameters the method does perform reasonably well, but it does not seem to surpass the ESA (Medline) baseline consistently. It does however use smaller representations to model input texts than ESA, tESA and DomESA (smaller in terms of the number of non-zero elements per representation). Curiously, when we take into account the number of non-zero elements of DS vectors, DS will yield better results (in terms of Spearman's rank correlation) than those of Word2vec CBOW, with data structures of similar size (with size being the total number of non-zero elements in the DS method's representations compared to the dimensionality of dense CBOW vectors).

tESA and DomESA, being extensions of the original ESA method, provide a solution towards adapting ESA to a domain-specific usage scenario, other than simply setting up ESA with a domain-focused background corpus. As stated in Chapter 2, this is accomplished by the additional operations on the vector that would have been the output of the original ESA method (i.e. vector-matrix multiplication). This comes at a certain computational cost, nonetheless the experimental data shows that the actual representations are of comparable size to the corresponding ESA vectors, while their number of dimensions is being reduced. In other words, tESA and

DomESA use structures of smaller size (number of non-zero elements) to those of the original ESA set up with a biomedical corpus (Medline), with DomESA's structure size being between that of the structures of ESA set up with Wikipedia and that of the structures of ESA set up with Medline. Additionally, tESA's and DomESA's vectors are of lower dimension than those of ESA (Medline). tESA's vectors have the dimension of the title vocabulary size, DomESA's vectors have the dimension of the general corpus' size (here:Wikipedia), while the original ESA set up with Medline will have their dimension equal to the size of specialized corpus.

DS in turn provides weaker rank correlation with human judgment, but it does so with much lighter representations of size comparable to the dimension of a CBOW representation.

The results presented in Section 2.5, describe the results of the DisMatch system in the OAEI 2016 campaign. The system uses DomESA representation vectors to produce relatedness scores between concepts of two ontologies, with textual labels, describing the concepts, used as inputs for the relatedness approximation. The relatedness score is then used together with a structural matcher to find the correspondences between the ontologies. The experiment consisted of two mapping tasks. In one of the tasks the structural matcher could not cope with the structure of one of the inputs, so its results were discarded. It resulted in relatively poor alignment, but also provided an interesting perspective on the application of lexical semantic relatedness to the problem, presented in the following section.

DisMatch did not provide especially good results w.r.t. to the silver standard, possibly due to the overall structure of the alignment system not being especially elaborate (the state-of-the-art systems tend to use more sophisticated matching strategies, which often involve combining a higher number of matchers or knowledge-rich resources). Nonetheless, the system stands out in the aspect of discovering the correct, unique mappings (i.e. mappings not detected by other systems), which points to a potential advantage of using a relatedness-based component in the problem of ontology matching. Specifically, recurring to corpus-based semantic relatedness may provide an added value in cases in which entire related segments of the ontologies to be matched are lexicalized differently (e.g. due to a slightly different perspective/purpose of a given ontology), thus limiting the applicability of matchers that propagate traditional string-based similarity.

Chapter 4
Conclusions

Within the work associated with this study a group of new methods for approximating lexical semantic relatedness has been introduced. Each of the three proposed methods improved the state-of-the-art in terms of quality of the approximation at the time of being published. In this chapter we present the conclusions derived from the evaluations and from the additional ontology matching experiment. The relationship between the conclusions and the research questions formulated in Chapter 1 are quite self explanatory, but they are highlighted with the corresponding references in parenthesis, e.g. (Q1), (1.b), etc.

The first of the methods, DS (described in the works included in Sections 2.1 and 2.2), is based on the idea of representing the input texts with vectors derived from the vocabulary present in a small sample of the documents of the corpus, that are the most relevant to the given input. The key to DS is in representing inputs with the most important vocabulary of the most important documents, which is a way of roughly approximating the context of the input, but without having to scan through the entire corpus for all the occurrences of the input words.

The results show that the concept of roughly approximating context through the key vocabulary of the most relevant documents works well (as DS performed better or similar to state-of-the art methods), while being much faster than scanning the corpus for occurrences and the actual contexts. Also, new representations can be created easily, by leveraging the corpus statistics and vector representations of the documents (Q1).

The DS method provides results inferior to those of the methods introduced later on, but it did provide results comparable or better than those of the then-known methods (it is also worth noting that the ESA baselines (Medline setup) were not contemplated until the paper presented in Section 2.3). DS also uses 'lighter' representations than the other proposed methods, with vectors with fewer non-zero elements. DS vectors' dimension is determined by the size of the vocabulary of the corpus, so tESA's vectors are of lower dimension (and more non-zero values), and ESA (Wiki) and DomESA representations are of comparable dimension (while being less sparse). Overall, DS offers the fastest (comparable only to those of word embeddings) one-to-one comparisons, through using very sparse high dimensional

vectors, with reasonable relatedness approximation quality (rank correlation higher than the one of word embeddings/CBOW, ESA, set up with Wikipedia). Its methodology is recommendable for use in those cases in which the speed of one-to-one comparisons is crucial, and some approximation quality can be sacrificed (1.a). Also, the acquisition of the representations is much faster than in the other methods (1.b).

An eventual user would also have to be prepared to perform some fine tuning of the method, as its performance seems to depend heavily on the combination of the reference dataset, background corpus and parameters, e.g. consider the change in performance with the updated PMC corpus, shown in the previous chapter. On a more positive note, the method's performance did not change much on the three larger reference datasets. The drop in performance on the smaller reference datasets can be connected to the fact that the method relates the inputs with the vocabulary quite directly, and for a different corpus it would lead to different representations, which in turn would lead to different results. With a larger reference dataset it is reasonable to expect this difference not to be very significant (because the method is the same after all), but for a smaller dataset it can result in a significant 'sway'.

The tESA method, introduced in the article included in Section 2.3, is an extension of the well-known ESA method. In tESA the representation vectors are obtained by multiplying ESA vectors by a matrix of titles of the documents of the corpus, with which the ESA vectors were obtained. tESA set up with the Medline corpus of scientific abstracts performed slightly better than the original ESA method (also set up with the same corpus), while operating on lighter representations of much lower dimension.

The tESA method addresses most of the shortcomings of the DS method. Specifically, it is easy to tune the method's parameter and its even important changes to its value do not greatly affect the performance. Also, the method is much less dependent on the length of individual documents within the corpus (2.b), as it works equally well with both abstracts and full scientific papers. Furthermore, it also shows a degree of robustness w.r.t. the corpus size, as tESA has shown good results both with Medline and PMC OA corpus, but takes advantage of a larger corpus (2.a). Medline-based tESA performed better than the established ESA baselines (set up with general domain and biomedical corpora; 2.c) and significantly better than the DS method.

The comparison with the original ESA seems especially interesting, as the slight advantage of tESA suggests, that incorporating domain specific semantics goes beyond setting up a general domain method with a domain specific resource (which provides an answer to question Q2). In tESA, we have used the fact that the titles of scientific documents provide a highly optimized description of the contents of these documents. In ESA the main principle is that words that 'trigger' the same documents are related. In tESA this translates to a 'fuzzier' notion, that words that are related should trigger the same title vocabulary. This contributes to tESA's bypassing of the corpus orthogonality. It is worth noting, that this notion is inherent to tESA's representation vectors (which have fewer dimensions and non-zero values than corresponding ESA vectors) and that the relatedness computation is resolved,

as in ESA, with cosine similarity between the representation vectors. This makes tESA a resource efficient approach, as opposed to some other methods designed to work around the corpus orthogonality (e.g. NESA). Also, the fact that tESA vectors are expressed over a relatively low-dimension space (of the key title vocabulary), means that tESA vector representations are a good fit for scenarios that need this kind of optimization (e.g. centroid-based classification or clustering), without having to sacrifice the relatedness approximation quality.

The notion of extending the use of the semantics of a domain-focused corpus, which is manifested in tESA, was also the key to designing the third method, DomESA, introduced in the paper included in Section 2.4. DomESA is mathematically close to tESA, i.e. DomESA also extends the traditional ESA representations in a matrix multiplication step. Conceptually, the intuition behind the tESA's processing can be presented as extending ESA with individual documents represented by their titles (while in the 'original' ESA a document is only represented by its identifier). Titles provide a highly optimized way of describing the contents of an article. The idea behind DomESA was to use a more powerful and less noisy representation scheme for individual documents to maximize the benefits of incorporating domain-focused semantics. Our original assumption was that a vector of a number of the most similar (vocabulary-wise) Wikipedia articles could provide a better representation of the semantics of a scientific article than its title vocabulary. In DomESA, the original method is extended by obtaining the vector-based representations of inputs by multiplying the ESA representation vectors by a truncated matrix of similarities between a domain-focused (Medline) and general-domain (Wikipedia) corpora. This extends the 'fuzzy' aspect of tESA. In the original ESA two input texts are related if they appear (and are important, i.e. yield a high cosine similarity score) in the same articles. In tESA two inputs are related if they appear (and are important) in the articles with the same words in the titles. In DomESA two inputs are related if they appear in *similar* documents, i.e. if the documents they appear in are similar to the same Wikipedia articles (again, in terms of cosine similarity).

DomESA provides very good results w.r.t. the methods evaluated in the papers in terms of rank-based relatedness with the model scores, only being slightly outperformed by tESA on one dataset. It also provides a lower score to the one reported in Sánchez and Batet (2011) on one of the smaller datasets. It also surpasses all the other methods included in the evaluation in terms of Pearson's correlation scores (3.c). Overall, DomESA seems to be the most consistent method across the board (3.a), regardless of the correlation coefficient used to measure the approximation success.

Apart from the good relatedness scores (which are more relative to Q2), it is worth noting that DomESA's use of domain-related semantics is restricted to the problem of assigning weights in a space of general-domain concepts, i.e. its final representation vectors are expressed over the Wikipedia articles (Q3). More specifically, DomESA's representation vectors are compatible with traditional ESA vectors (with the method set up with Wikipedia), with DomESA's processing only affecting the value distribution within the vectors (3.b). This compatibility feature potentially expands the range of use of the corpus based measures in general, as it allows for

comparisons between domain-focused inputs (e.g. for inputs which do not appear in Wikipedia) and general-domain inputs (for which we assume that the general-domain semantic representation is correct). Additionally, the fact, that DomESA's representations are expressed over Wikipedia concepts, means that the representations are interpretable w.r.t. a rich knowledge base. Specifically, the representations are vectors, with each element corresponding to its relevance to a specific Wikipedia concept/article. This feature of the DomESA representations makes it possible to extend DomESA in the same way that ESA has been extended, for example, with thematic reinforcements. An especially interesting extension would involve adding multi-language support to DomESA as has been done in CL-ESA.

Moreover, the experimental results presented here suggest that DomESA's representations are not that much heavier (in terms of having more non-zero valued elements) than those of the original ESA set up with Wikipedia, while being lighter than the representation vectors of ESA set up with Medline. This means, that just like with tESA, DomESA's additional computational cost is only significant at the point of creating the representation vectors, as the matrix multiplication step is necessary.

Additionally, the work presented in this study seems to suggest that overcoming (or reducing) the corpus orthogonality improves the results of ESA-style methods. The performance gap between the variants of the original ESA, set up with either Wikipedia or Medline corpus, can be easily explained by the difference in the domain suitability of the resources, rather than by the difference in orthogonality between the resources (Medline is more domain-specific and less orthogonal). Nonetheless, the argument about domain suitability is not valid in the case of comparing the performance of ESA (set up with Medline) with either tESA or DomESA, which suggests that reducing the corpus orthogonality may indeed improve the results.

The paper included in 2.5 presents an experiment (together with the evaluation of the results) of applying the DomESA method to the problem of ontology alignment. Although the results of the proposed system, DisMatch, were not entirely on par with those of the best systems participating in the competition, our system stood out in discovering correct, unique mappings, i.e. correct mappings that were not discovered by any other system. This seems to suggest that a lexical relatedness-based matcher could be a valuable component within a more elaborate structure of an ontology matching system. Traditional ontology matching systems normally use string-based metrics of similarity and vocabulary overlap measures, combined with structural matchers. Our results seem to suggest that including a relatedness-based component could boost the recall of a full-scale matching system (Q4).

Additionally, as mentioned in the paper included in Section 2.4, our experience with the ontology matching task seems to point to a specific deficiency of DomESA inherited from the original ESA method, which consequently results in an increased rate of false positive results. Specifically, the relatedness approximations for longer input texts often result in relatively high scores due to the inputs sharing a large part of common vocabulary. DomESA operates without accounting for the dissimilarity between the non-common parts of the inputs, so two inputs may yield a high relat-

edness score because of sharing some words of a relatively low semantic impact, even though the non-common parts of the inputs are highly dissimilar. This tends to occur often for the pairs of inputs with the respective non-common elements being poorly represented within the background corpus. The possibility of improving this aspect of the method through revisiting the compositional aspect of the representation vectors will be studied in the near future. Accounting for the dissimilarity of the non-similar elements will first be addressed along the lines of the methodology designed for word embeddings, i.e. *Word's Mover Distance* (Kusner et al, 2015), which seems applicable to DomESA.

Conclusiones

En las publicaciones relacionadas con este estudio, se han introducido un grupo de nuevos métodos para la aproximación de la proximidad semántica basada en palabras en el dominio biomédico. Cada uno de los tres métodos ha mejorado el estado del arte en el momento de su publicación. En este capítulo, se presenta las conclusiones derivadas de las evaluaciones y del experimento adicional enfocado en el alineamiento de ontologías. Entre los paréntesis se indica las relaciones con las cuestiones planteadas en el Capítulo 1, por ejemplo (Q1), (Q2), etc.

El primer método, DS (introducido en las publicaciones incluidas en las Secciones 2.1 y 2.2), está basado en la idea de representar los textos de entrada con los vectores derivados del vocabulario presente en los subconjuntos de los documentos del corpus, que son los más adecuados para cada una de las entradas. La clave del método es la idea de representar la entrada con el vocabulario más importante de los documentos más adecuados, con lo cual se consigue una aproximación del contexto de la entrada, pero sin tener que 'escanear' los documentos individuales para las ocurrencias de las entradas.

Los resultados obtenidos en la evaluación indican, que la idea de aproximar contextos con el vocabulario clave de los documentos más adecuados funciona bastante bien (el método DS ha funcionado mejor o comparable con respecto a los métodos de referencia), siendo el método más rápido y flexible a la hora de crear las representaciones. Además, en DS las representaciones nuevas se pueden crear fácilmente – usando las estadísticas del corpus y los vectores (Tf-Idf) de los documentos (Q1; 1.b).

El método DS proporciona resultados inferiores que los de los métodos propuestos más adelante (siendo importante reconocer, que la evaluación detallada del método ESA en el contexto biomédico no se había contemplado antes de lo expuesto en el trabajo incluido en la Sección 2.3), pero comparables con los mejores métodos contemplados por aquel entonces. Además, DS usa las representaciones más 'ligeras' que los otros métodos propuestos en esta tesis (las representaciones de DS tienen notablemente menos valores diferentes a cero que las de los otros métodos). La dimensión de las representaciones de DS está determinada por el tamaño del vocabulario del corpus: DS usa vectores de dimensiones más alta que la de los vectores

de tESA (con las representaciones de DS teniendo menos valores diferentes a cero), mientras que las representaciones de ESA y DomESA tienen dimensiones comparables con las de las representaciones de DS, pero las representaciones de DS son mucho menos densas. Generalmente, DS proporciona la aproximación de la proximidad semántica más rápida que la de los otros métodos, comparable solo con el rendimiento de los word embeddings, a través del uso de los vectores muy dispersos de alta dimensión. No obstante, el método DS proporciona la aproximación de la calidad relativamente buena (superior a la de algunos métodos establecidos, como CBOW o ESA configurada con Wikipedia). Esta metodología se puede recomendar para usar en los casos, en los cuales la eficiencia de los cálculos uno a uno tiene prioridad, hasta el punto que justifica cierto nivel del compromiso en cuanto la calidad de la aproximación (1.a).

Un hipotético usuario del método tendría que estar preparado para ajustar los parámetros del método DS a su propio caso de uso, porque el rendimiento del método puede variar según la combinación del corpus usado, la fuente del conocimiento, datos a los cuales se aplica el método y los parámetros. Como ejemplo, se puede considerar el cambio del rendimiento del método por la expansión del corpus PMC, presentado en el capítulo anterior. El rendimiento del método no ha cambiado mucho en los casos de los 3 conjuntos de datos más grandes, pero ha bajado en los conjuntos mayo29ph y mayo29c. Ese cambio en los resultados obtenidos para conjuntos pequeños se puede explicar con el hecho de que el método relaciona las entradas con los elementos del vocabulario de una manera relativamente directa, así que un cambio importante en el corpus puede inducir un cambio importante en las representaciones, lo cual puede reflejarse en los resultados obtenidos para un conjunto tan pequeño. Para un conjunto más grande, se puede esperar que el efecto de estos cambios esté 'suavizado' estadísticamente por el tamaño de la prueba, mientras que los resultados obtenidos para un conjunto más pequeño pueden estar más distorsionados por los cambios inducidos.

El método tESA, introducido en el artículo incluido en la Sección 2.3, es una extensión de un método establecido: ESA. En tESA los vectores de representación de las entradas se obtiene multiplicando los vectores de conceptos (que son las representaciones en el método ESA) por una matriz del vocabulario de los títulos de los documentos del corpus usado como la fuente de conocimiento. El método tESA, configurado con el corpus Medline, ha proporcionado una calidad de la aproximación un poco más alta que el método original (también configurado con el mismo corpus), pero usando las representaciones más ligeras (con menos elementos distintos a cero) y de dimensión mucho más baja.

El diseño de tESA soluciona la mayoría de los problemas del método anterior, DS. Específicamente, tESA es fácil de calibrar y es robusta en cuanto los cambios de los parámetros. Además, el método depende mucho menos de las características de los documentos individuales que forman parte del corpus (2.b): funciona bien tanto con los textos completos, como con los resúmenes de los artículos científicos. Adicionalmente, el método demuestra un buen funcionamiento con las colecciones de los documentos de variable tamaño, ya que proporciona resultados parecidos en la configuración con PMC y con Medline. No obstante, el método tESA aprovecha

el corpus más grande, la configuración con Medline siendo la óptima (2.a). Los resultados obtenidos con el método tESA configurado con Medline superan tanto las bases de referencia establecidas con el método original (ESA), como la del método anterior (DS).

Sobre todo, parece interesante considerar la comparación entre tESA y el método original ESA (configurado también con Medline). La ventaja de tESA en cuanto a la calidad parece indicar que la incorporación de la semántica del dominio (biomédico) puede ser más compleja que coger un método del dominio general y configurarlo con el corpus más adecuado (Q2). En tESA se ha utilizado la observación, que un título de un artículo científico proporciona una descripción optimizada de los contenidos del dicho artículo. En ESA, la idea básica es que las entradas parecidas son las que co-ocurren mucho en los documentos mutuamente importantes para las entradas. En tESA esta idea básica se convierte en una formulación más relajada: las entradas parecidas son las que, a través de los contenidos de los documentos, ' activan' el mismo vocabulario del espacio de los títulos. Esta característica contribuye a que tESA disminuya el impacto de la ortogonalidad del corpus. Notablemente, es una característica propia de las representaciones de tESA (que son generalmente más ligeras que las representaciones de ESA) y que el cálculo de la proximidad semántica prácticamente se resuelve de la misma manera, tanto en tESA, como en ESA, o sea calculando la similitud coseno entre las parejas de las representaciones de las entradas. Por lo tanto, tESA es eficiente en cuanto el rendimiento en línea, sobre todo en comparación con los otros métodos que reducen el impacto de la ortogonalidad del corpus (por ejemplo, NESA). Además, el hecho de que los vectores de tESA están expresados sobre un espacio relativamente pequeño (siendo el vocabulario de los títulos de los documentos del corpus) hace que las representaciones de tESA son más aplicables (que las del método original ESA) a los escenarios específicos, para los cuales este tipo de optimización es necesario (por ejemplo construcción de un clasificador basado en los centroides), sin tener que sacrificar mucho en cuanto a la calidad de la aproximación de la proximidad semántica.

La idea de extender el uso de la semántica del corpus enfocado en un dominio especializado, presente en tESA, ha sido también la clave en el proceso del diseño del tercer método – DomESA (introducida en el artículo adjuntado en la Sección 2.4). DomESA es un método computacionalmente parecido a tESA, específicamente DomESA también extiende el método ESA con el paso de multiplicar los vectores por una matriz. Conceptualmente, la explicación intuitiva del procesamiento de tESA es de extender el método original ampliando la representación de los documentos individuales incluidos en el proceso de superposición de los vectores para obtener el vector de la representación semántica de la entrada – representando cada uno de los documentos con un vector extraído de su título (mientras que en el método ESA un documento en este contexto se representa con un vector binario activado en la posición correspondiente al identificador del documento). Los títulos en este contexto sirven para reflejar la semántica del documento de una manera optimizada. La idea base de DomESA es representar los documentos de una manera menos ruidosa y más expresiva, que permita aprovechar al máximo la información semántica incluida en el corpus especializado. La suposición original era que un

vector de un cierto número de conceptos de Wikipedia más similares (en cuanto el vocabulario) al contenido del artículo, serviría mejor para representar la semántica del dicho artículo, que el vector del vocabulario extraído de su título[1]. En DomESA los vectores de representaciones de las entradas se obtienen multiplicando los vectores del método original por una matriz de similitud entre los documentos del corpus especializado (Medline) y los del corpus del dominio general (Wikipedia). El diseño de DomESA extiende el aspecto disperso de tESA. En el método ESA original, dos entradas se consideran próximas si aparecen juntas en los mismos documentos (siendo importantes dentro de los contenidos de esos documentos). En tESA las dos entradas se consideran próximas si los documentos en los cuales aparecen las entradas comparten el mismo vocabulario en los títulos. En DomESA las dos entradas se consideran similares si aparecen en los conjuntos de documentos similares, específicamente, si los documentos en los cuales aparecen, son similares a los mismos conceptos/artículos de Wikipedia.

DomESA proporciona los resultados de las correlaciones de Spearman por encima de los resultados obtenidos con los otros métodos contemplados en este estudio, solo siendo la segunda mejor opción detrás de tESA en un conjunto de referencia. DomESA proporciona también un segundo mejor resultado (tras un método citado – (Sánchez and Batet, 2011)) en un conjunto de referencia más pequeño. En cuanto los resultados de las correlaciones de Pearson, DomESA sobrepasa todos los otros métodos incluidos en la evaluación (3.c). En general, DomESA parece ser el método más consistente de todos los métodos contemplados aquí, en cuanto a la calidad de los resultados (3.a).

Aparte de los buenos resultados de la aproximación de la proximidad semántica (Q2), se puede observar que el uso de la semántica del corpus especializado en DomESA está restringido al problema de asignar los pesos a los conceptos/artículos del recurso del dominio general. Específicamente, los eventuales vectores de representación de las entradas están expresados sobre el espacio de los artículos de Wikipedia (Q3). Las representaciones de DomESA son compatibles con las representaciones del método original (configurado con Wikipedia), con el procesamiento de DomESA afectando nada más que la distribución de los valores dentro de las representaciones (3.b). Este factor de compatibilidad potencialmente extiende el rango del uso de los métodos basados en las colecciones de los documentos, porque proporciona un mecanismo que permite estimación de la proximidad semántica entre las entradas del dominio especializado con las entradas del dominio general. Adicionalmente, el hecho, de que las entradas de DomESA estén representadas sobre el espacio de los artículos de Wikipedia, significa, que estas representaciones pueden ser interpretadas usando una base de conocimiento estructurada, ya que se trata de los vectores que al cada concepto de Wikipedia tienen asignado un grado numérico de relevancia (o sea, un peso). Esta característica de los vectores de DomESA significa, que potencialmente se puede extender DomESA de la misma manera que se ha extendido el método original, por ejemplo ampliando las representaciones con la información temática/categórica. Una extensión especialmente interesante seria

[1] Se puede observar que la idea es parecida a representar los documentos del corpus especializado con ESA tradicional configurada con el corpus del dominio general (Wikipedia)

añadir el soporte multi-lingüístico a DomESA, de la misma manera que se ha propuesto en CL-ESA.

Además, los resultados presentados en este estudio indican, que las representaciones usadas por DomESA no son mucho más 'grandes' que las del método ESA original configurado con Wikipedia y más ligeros que las de ESA configurado con Medline (en cuanto los elementos de los vectores con valores diferentes a cero). Esto significa, que, exactamente como en el caso de tESA, el coste adicional implicado en el procesamiento de DomESA se paga solo a la hora de crear las representaciones, ya que es necesario el paso adicional de multiplicación por la matriz.

Adicionalmente, el trabajo presentado aquí parece indicar que superar (o reducir) la ortogonalidad del corpus mejora los resultados de los métodos derivados de ESA. La diferencia en el rendimiento entre los dos variantes de ESA (Wikipedia/Medline) puede ser explicada fácilmente con la idoneidad del corpus en cuanto el dominio de uso/evaluación, aunque Medline sea menos ortogonal. Sin embargo, el argumento sobre la idoneidad de la colección de los textos no parece valido en el caso de comparar el rendimiento inducido por DomESA con el método ESA (configurado con Medline), lo cual parece indicar que la reducción de la ortogonalidad puede traer un mejora en cuanto a la calidad de los resultados.

El trabajo incluido en la Sección 2.5 presenta a un experimento (junto con la evaluación de los resultados obtenidos) basado en aplicar DomESA al problema del alineamiento de las ontologías. Aunque los resultados proporcionados por el sistema propuesto, DisMatch, son inferiores a los de los mejores sistemas evaluados en la edición 2016 de la iniciativa, nuestro método ha destacado en la capacidad de descubrir correctas relaciones únicas, o sea las relaciones correctas que no se habían descubierto con los otros sistemas. Los resultados de la evaluación indican que un módulo de la aproximación de la proximidad semántica podría ser un componente importante de un sistema complejo diseñado para resolver el problema del alineamiento de las ontologías. Los sistemas de alineamiento de las ontologías tradicionalmente usan las métricas basadas en las cadenas de textos para aproximar la proximidad entre los conceptos. Los resultados demostrados en el artículo, parecen indicar que la inclusión de un componente basado en el concepto de la proximidad semántica puede mejorar el rendimiento de un sistema complejo (Q4).

Además, como ya se ha mencionado en el artículo de la Sección 2.4, nuestra experiencia con el problema del alineamiento de las ontologías indica un problema importante del mecanismo de DomESA, heredado del método original, que resulta en un elevado (comparando con los resultados obtenidos para los típicos conjuntos de referencia) porcentaje de los falsos positivos. En concreto, DomESA a menudo devuelve los valores altos para las parejas de entradas que comparten una parte del vocabulario. El método tiende a ignorar la relación entre los elementos únicos de las entradas, sobre todo ignorando las disimilitudes entre los elementos. Como resultado, dos entradas pueden generar una aproximación de la proximidad semántica bastante alta por el hecho de compartir una parte del vocabulario, aunque sean elementos de poca importancia semántica, mientras que los elementos propios de las entradas sean incompatibles. Esto suele ocurrir para parejas de las entradas para las cuales los elementos propios no están muy bien reflejados en el corpus. La posibil-

idad de mejorar este aspecto de DomESA, a través de modificar la estrategia composicional de las representaciones del método, se va a investigar pronto. La inclusión de las disimilitudes entre los elementos de las entradas se va a implementar según la metodología propuesta para los 'word embeddings' – Word's Mover Distance. El algoritmo propuesto por Kusner et al (2015) parece aplicable a las representaciones de DomESA.

Literature

References

Achichi M, Cheatham M, Dragisic Z, Euzenat J, Faria D, Ferrara A, Flouris G, Fundulaki I, Harrow I, Ivanova V, et al (2016) Results of the ontology alignment evaluation initiative 2016. In: 11th ISWC workshop on ontology matching (OM), No commercial editor., pp 73–129

Agirre E, Rigau G (1996) Word sense disambiguation using conceptual density. In: Proceedings of the 16th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pp 16–22

Agirre E, Alfonseca E, Hall K, Kravalova J, Paşca M, Soroa A (2009) A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp 19–27

Asooja NAK, Bordea G, Buitelaar P (2015) Non-orthogonal explicit semantic analysis. Lexical and Computational Semantics (* SEM 2015) p 92

Barzilay R, Elhadad M (1997) Using lexical chains for text summarization. In: Proceedings of the ACL workshop on intelligent scalable text summarization: July 1997; Madrid, Spain, Association for Computational Linguistics, pp 10–17

Batet M, Sánchez D, Valls A (2011) An ontology-based measure to compute semantic similarity in biomedicine. Journal of biomedical informatics 44(1):118–125

Bellazzi R, Masseroli M, Murphy S, Shabo A, Romano P (2012) Clinical bioinformatics: challenges and opportunities. BMC bioinformatics Suppl 14:S1

Berlanga R, Nebot V, Jimenez E (2010) Semantic annotation of biomedical texts through concept retrieval. Procesamiento del Lenguaje Natural 45:247–250

Budanitsky A (1999) Lexical semantic relatedness and its application in natural language processing. University of Toronto

Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics 32(1):13–47

Chiu B, Crichton G, Korhonen A, Pyysalo S (2016) How to train good word embeddings for biomedical nlp. ACL 2016 p 166

Couto FM, Pinto HS (2013) The next generation of similarity measures that fully explore the semantics in biomedical ontologies. Journal of bioinformatics and computational biology 11(05)

Cross V (2004) Fuzzy semantic distance measures between ontological concepts. In: NAFIPS'04, IEEE, vol 2, pp 635–640

Dumais ST (2004) Latent semantic analysis. Annual review of information science and technology 38(1):188–230

Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI, vol 7, pp 1606–1611

Gracia J, Asooja K (2013) Monolingual and cross-lingual ontology matching with cider-cl: evaluation report for oaei 2013. In: Proceedings of the 8th International Conference on Ontology Matching-Volume 1111, CEUR-WS. org, pp 109–116

Guo X, Liu R, Shriver CD, Hu H, Liebman MN (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 22(8):967–973

Haralambous Y, Klyuev V (2013) Thematically reinforced explicit semantic analysis. International Journal of Computational Linguistics and Applications 4(1):79

Harrow I, Jimenez-Ruiz E, Splendiani A, Romacker M, Negru S, Woollard P, Markel S, Alam-Faruque Y, Koch M, Younesi E, et al (2016) Introducing the disease and phenotype oaei track. In: 11th ISWC workshop on ontology matching (OM)

Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ (2015) From word embeddings to document distances. In: Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), pp 957–966

Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. Discourse processes 25(2-3):259–284

Liu Y, McInnes BT, Pedersen T, Melton-Meaux G, Pakhomov S (2012) Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, umls and wordnet. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, ACM, pp 363–372

Martinez-Gil J (2016) Accurate semantic similarity measurement of biomedical nomenclature by means of fuzzy logic. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 24(02):291–305

Mathur S, Dinakarpandian D (2012) Finding disease similarity based on implicit semantic similarity. Journal of biomedical informatics 45(2):363–371

McInnes B, Liu Y, Pedersen T, Melton G, Pakhomov S (2013) Umls::similarity: Measuring the relatedness and similarity of biomedical concepts. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 9-14 June 2013; Atlanta, Association for Computational Linguistics, pp 28–31

McInnes BT, Pedersen T (2016) Improving correlation with human judgments by embedding second order vectors with semantic similarity. arXiv preprint arXiv:160900559

McInnes BT, Pedersen T, Liu Y, Melton GB, Pakhomov SV (2011) Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2011, p 895

Melnik S, Garcia-Molina H, Rahm E (2002) Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: Data Engineering, 2002. Proceedings. 18th International Conference on, IEEE, pp 117–128

Mikolov T, Chen K, Corrado GS, Dean J (2013) Efficient estimation of word representations in vector space

Muneeb T, Sahu SK, Anand A (2015) Evaluating distributed word representations for capturing semantics of biomedical concepts. ACL-IJCNLP 2015 p 158

Navas-Delgado I, García-Godoy MJ, López-Camacho E, Rybinski M, Reyes-Palomares A, Medina MÁ, Aldana-Montes JF (2015) kpath: integration of metabolic pathway linked data. Database 2015:bav053

Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB (2010) Semantic similarity and relatedness between clinical terms: an experimental study. In: AMIA annual symposium proceedings, American Medical Informatics Association, vol 2010, p 572

Pakhomov SV, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG (2011) Towards a framework for developing semantic relatedness reference standards. Journal of biomedical informatics 44(2):251–265

Pakhomov SV, Finley G, McEwan R, Wang Y, Melton GB (2016) Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics 32(23):3635–3644

Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG (2007) Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics 40(3):288–299

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: EMNLP, vol 14, pp 1532–43

Pesquita C, Faria D, Falcao AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS computational biology 5(7):e1000,443

Polajnar T, Aggarwal N, Asooja K, Buitelaar P (2013) Improving esa with document similarity. In: Advances in Information Retrieval, Springer, New York, pp 582–593

Potthast M, Stein B, Anderka M (2008) A wikipedia-based multilingual retrieval model. In: European Conference on Information Retrieval, Springer, pp 522–530

Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19(1):17–30

Rybiński M, Aldana-Montes JF (2013) Sebsem: simple and efficient biomedical semantic relatedness measure. EMBnet journal 19(B):pp–82

Rybiński M, Aldana-Montes JF (2014) Calculating semantic relatedness for biomedical use in a knowledge-poor environment. BMC bioinformatics 15(14):1

Rybiński M, Aldana-Montes JF (2016) tESA: a distributional measure for calculating semantic relatedness. BMC Journal of Biomedical Semantics – accepted for publication

Rybiński M, Aldana-Montes JF (2017) DomESA: a novel approach for extending domain-oriented lexical relatedness calculations with domain-specific semantics. JIIS

Rybiński M, del Mar Roldán-Garcıa M, Garcıa-Nieto J, Aldana-Montes JF (2016) Dismatch results for OAEI 2016

Sahami M, Heilman TD (2006) A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international conference on World Wide Web, AcM, pp 377–386

Sahay S, Ram A (2011) Socio-semantic health information access. In: AAAI Spring Symposium: AI and Health Communication, AAAI

Sajadi A, Milios EE, Kešelj V, Janssen JC (2015) Domain-specific semantic relatedness from wikipedia structure: A case study in biomedical text. In: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, pp 347–360

Sánchez D, Batet M (2011) Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of biomedical informatics 44(5):749–759

Scholl P, Böhnstedt D, García RD, Rensing C, Steinmetz R (2010) Extended explicit semantic analysis for calculating semantic relatedness of web resources. In: Sustaining TEL: From Innovation to Learning and Practice, Springer, New York, pp 324–339

Ślezak D, Janusz A, Świeboda W, Nguyen HS, Bazan JG, Skowron A (2011) Semantic analytics of pubmed content. In: Symposium of the Austrian HCI and Usability Engineering Group, Springer, pp 63–74

Strube M, Ponzetto SP (2006) Wikirelate! computing semantic relatedness using wikipedia. In: AAAI, vol 6, pp 1419–1424

Turney PD, Pantel P (2010) From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37(1):141–188

Zhang R, Pakhomov S, McInnes BT, Melton GB (2011) Evaluating measures of redundancy in clinical texts. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2011, p 1612

Zhang Z, Gentile AL, Ciravegna F (2012) Recent advances in methods of lexical semantic relatedness–a survey. Natural Language Engineering 1(1):1–69

Zou GY (2007) Toward using confidence intervals to compare correlations. Psychological methods 12(4):399