Contribución de la información semántica en un sistema de aprendizaje automático para resolver la implicación textual*

Sonia Vázquez, Zornitsa Kozareva y **Andrés Montoyo**
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{svazquez, zkozareva, montoyo}@dlsi.ua.es

Resumen: La forma en que podemos expresar un mismo pensamiento puede variar dependiendo de las circunstancias, del objetivo que pretendemos alcanzar, etc. Esta variabilidad del lenguaje se convierte en un problema cuando una aplicación de procesamiento del lenguaje natural intenta extraer el mismo contenido semántico a partir de dos construcciones distintas. Para resolver este problema se ha creado una tarea denominada "Reconocimiento de Implicación Textual" o *"Textual Entailment Recognition"*, cuyo objetivo es establecer si la semántica de un texto se puede inferir de la semántica de otro texto. En este artículo, presentamos una nueva aproximación que utiliza la semántica latente y la medida de similitud del coseno para resolver el problema de la Implicación Textual. Para realizar la evaluación se han utilizado diferentes corpus y recursos, realizando posteriormente un estudio sobre el impacto de combinar la información semántica obtenida con un sistema de aprendizaje automático.
Palabras clave: Implicación Textual, Semántica Latente, Aprendizaje Automático

Abstract: The variability of semantic expression is a special characteristic of natural language. This variability is challenging for many natural language processing applications that try to infer the same meaning from different text variants. In order to treat this problem a generic task has been proposed: Textual Entailment Recognition. In this paper, we present a new Textual Entailment approach based on Latent Semantic Indexing (LSI) and the cosine measure. This proposed approach extracts semantic knowledge from different corpora and resources. Our main purpose is to study how the acquired information can be combined with an already developed and tested Machine Learning system. The carried out experiments show that the combination of MLEnt, LSI and cosine measure improves the results of the initial approach.
Keywords: Textual Entailment, Latent Semantic Indexing, Machine Learning

1. Introducción

La variabilidad semántica es una característica propia del lenguaje, es decir, podemos expresar de diferentes formas un mismo pensamiento. Por ello, es necesario identificar correctamente aquellas frases o fragmentos que aunque construidos con estructuras diferentes, tienen el mismo contenido semántico.

En Procesamiento del lenguaje natural (PLN) existen diferentes tareas, tales como, Extracción de Información (EI), Búsqueda de Respuestas (BR), Resumen Automático (RA) o Traducción Automática (TA) entre otras, que necesitan resolver correctamente la variabilidad semántica. Por lo tanto, como respuesta a este tipo de problema se ha creado una tarea denominada Resolución de la Implicación Textual"(RTE) (Dagan, Glickman, y Magnini, 2005) (Dagan y Glickman, 2004), la cual, se ocupa de detectar si dos fragmentos de texto con diferente estructura tienen el mismo contenido semántico. En la Resolución de la Implicación Textual cada par de fragmentos se compone de una parte denominada Texto (T) y otra parte denominada Hipótesis (H), y el objetivo final es demostrar si el contenido del texto proporciona la misma información que el contenido de la hipótesis. Por ejemplo, T(*"He died of blood loss"*) y H(*"He died bleeding"*) tienen el mismo significado pero diferente estructura. Por lo tanto, podemos decir que la semántica de una sentencia se puede inferir de la semántica de la otra.

* Este trabajo ha sido parcialmente financiado por los proyectos CICyT número TIC2003-07158-C04-01 y PROFIT número FIT-340100-2004-14 y por el Gobierno Valenciano GV04B-276.

Para resolver el problema de la variabilidad semántica se han desarrollado diferentes aproximaciones basadas en: solapamiento léxico, WordNet (Miller, 1995), medidas estadísticas, mapeo de n-gramas, mapeo sintáctico, etc (Bar-Haim et al., 2005). Utilizando estas técnicas y aplicándolas sobre sistemas de PLN ya desarrollados, se pueden mejoran los resultados finales (Kozareva y Montoyo, 2006b).

En este artículo se presenta una nueva aproximación para resolver la Implicación Textual utilizando Semántica Latente (*"Latent Semantic Indexing"*, LSI) (Deerwester et al., 1990) y la medida de similitud del coseno. Tradicionalmente, la técnica de la semántica latente ha utilizado una matriz conceptual obtenida a partir de extensos corpus. Sin embargo, en nuestra aproximación utilizaremos esta técnica con diferentes recursos, como por ejemplo, a partir de los textos de la colección de datos de RTE2 y con el recurso WordNet Domains (Magnini y Cavaglia, 2000). La obtención de las matrices conceptuales a partir de estos recursos se explica con más detalle en la sección 2. Además, adicionalmente a la información aportada por la técnica de la Semántica Latente hemos añadido la información obtenida a partir de la medida de similitud del coseno, con dos variantes: a partir de la frecuencia de aparición de palabras en un corpus y a partir de la información suministrada por el recurso léxico Dominios Relevantes (Montoyo, Vázquez, y Rigau, 2003), este recurso se explica con más detalle en la sección 3.2.

La evaluación del método propuesto se ha realizado sobre la colección de datos de RTE2. Posteriormente, la información semántica obtenida se ha combinado con un sistema de aprendizaje automático. En los distintos experimentos realizados se han comparado los resultados obtenidos por la nueva aproximación, con los resultados obtenidos con el sistema inicial (Kozareva y Montoyo, 2006a), realizando un estudio sobre los efectos producidos tras la incorporación de la nueva información semántica.

El objetivo de este trabajo es estudiar los efectos de incorporar la información semántica sobre un sistema, ya evaluado, que resuelve la Implicación Textual, así como establecer un sistema sencillo e independiente del idioma que pueda ser utilizado como fuente de información para otros sistemas.

2. *Representación del conocimiento mediante semántica latente*

La Semántica Latente o *"Latent Semantic Indexing"* (LSI) es un modelo computacional que explota una propiedad del lenguaje natural: las palabras del mismo campo semántico suelen aparecer en el mismo contexto. Este modelo establece relaciones entre palabras a partir de un extenso corpus, utilizando un espacio semántico vectorial donde todos los términos son representados con una matriz [términos-documentos]. Para obtener información útil, los términos deben estar distribuidos en documentos, párrafos o frases. Esta distribución determinará cuál es la co-ocurrencia entre diferentes términos y la probabilidad de utilizar otros términos en el mismo contexto. Una vez obtenida la matriz, LSI utiliza una variante del análisis factorial denominada: Singular Value Decomposition (SVD). Esta técnica utiliza un algoritmo recursivo para descomponer la matriz inicial, en tres nuevas matrices que contienen vectores y valores singulares. Estas matrices disminuyen el número de datos originales creando factores linealmente independientes. Una gran parte de estos factores son muy pequeños y pueden ser ignorados de forma que se obtiene un modelo aproximado reduciendo el número de factores. El resultado final es un modelo reducido de la matriz inicial de [términos-documentos] que puede ser utilizado para establecer relaciones de similitud entre palabras.

En esta sección presentamos cómo se ha extraído información a partir de diferentes recursos lingüísticos y cómo se ha aplicado el método de LSI para establecer la similitud entre palabras a partir de un espacio semántico. Para evaluar el funcionamiento de esta técnica se han desarrollado diferentes experimentos utilizando tres tipos de corpus: el British National Corpus (BNC) (Aston, 1996), un corpus obtenido a partir de las frases de Text e Hipótesis de RTE2 y un corpus obtenido a partir del recurso léxico WordNet Domains. La forma en que se ha utilizado la información procedente de estos corpus para construir la matriz conceptual se explica con más detalle en las siguientes secciones.

2.1. Espacio semántico con BNC

En esta sección se explica con detalle cómo se ha obtenido el espacio semántico a partir

de la información proporcionada por el British National Corpus. Este corpus contiene una colección de alrededor de 4000 documentos obtenidos a partir de periódicos nacionales, artículos especializados de diferentes áreas e intereses, etc. De esta forma, este corpus proporciona información útil para establecer relaciones entre palabras a partir de su frecuencia de aparición en los distintos documentos.

Para el estudio que nos ocupa hemos construido una matriz [términos-documentos], donde las filas representan todos los términos posibles hallados en el corpus y las columnas representan todos los documentos del corpus. En nuestra primera aproximación, hemos extraído todas las palabras previamente lematizadas y hemos calculado la frecuencia de aparición en cada documento. Esta información es la entrada del módulo de LSI para obtener el nuevo espacio conceptual.

Una vez se ha obtenido el nuevo espacio conceptual, podemos establecer la similitud entre cada par de frases de RTE2 (T-H) (Texto-Hipótesis). Para este propósito y utilizando el método de LSI hemos realizado dos experimentos diferentes:

- **Documentos Relevantes.** Se extraen los 20 primeros documentos más relevantes para cada frase y se determina el grado de similitud a partir del número de documentos compartidos.
- **Palabras relevantes.** Se extraen las 800 primeras palabras más relevantes para cada frase y se determina el grado de similitud a partir del número de palabras compartidas.

El resultado final para cada tipo de experimento es un valor normalizado entre 0-1 que muestra el grado de similitud entre cada par de frases (cuanto más cercano a 1 sea el resultado, más similares serán las frases).

2.2. Espacio semántico con Text-Hipótesis

En esta sección presentamos otro tipo de experimentos utilizando como fuente de información las palabras obtenidas a partir de las frases de Texto (T) e Hipótesis (H). En este caso, estudiamos la similitud entre cada par T-H utilizando como espacio conceptual las frases proporcionadas como Texto o las frases proporcionadas como Hipótesis, respectivamente. En otras palabras, vamos a construir dos tipos diferentes de matrices: una utilizando las frases de Texto como corpus y otra utilizando las frases de Hipótesis como corpus. Con este experimento tratamos de comprobar, si la información contenida en las frases de RTE2 es suficiente para establecer la implicación textual.

La matriz [Hipótesis-Texto] contiene una columna para cada frase de las proporcionadas como Texto y una fila para cada frase de las proporcionadas como Hipótesis. Por tanto, para establecer la similitud entre H-T calculamos para cada frase de Hipótesis cuáles son las frases de Texto más relevantes de acuerdo a nuestro espacio conceptual. El resultado es una lista de las 20 frases más relevantes de Texto junto con su valor de similitud asociado. Si el par Texto-Hipótesis que estamos buscando se encuentra entre las 20 frases extraídas, obtenemos el valor de similitud obtenido por el método LSI (entre 0-1), en otro caso, H-T tienen un valor de similitud de 0.

El mismo procedimiento se sigue para obtener la matriz [Texto-Hipótesis] y calcular la similitud entre los pares T-H.

2.3. Espacio semántico con WordNet Domains

En esta sección vamos a detallar cómo crear una matriz [términos-dominios], a partir de la información proporcionada por WordNet Domains (WND), una base de datos léxica. En WND cada palabra tiene asociada una glosa, equivalente a la definición de un diccionario. Debido a la polisemia, una palabra podrá tener diferentes glosas o definiciones, y asociadas a cada definición encontraremos una o varias etiquetas de dominio (la colección de dominios está organizada jerárquicamente y agrupa los sentidos de las palabras según su contenido semántico). Es decir, para un sentido de una palabra tendremos uno o varios dominios y una glosa asociada. Por ejemplo, para los diferentes sentidos de la palabra *"music"* tendremos los siguientes dominios asociados: *music#1:Music, music#2:Acoustics, music#3:Free_time*, etc. (Para *"music#1:Music"*: su glosa sería: *"An artistic form of auditory communication incorporating instrumental or vocal..."*).

El primer paso para obtener la matriz conceptual es extraer el conjunto de etiquetas de dominio, ya que, cada dominio va a represen-

tar una columna de nuestra matriz. En este caso, tenemos una jerarquía de dominios con alrededor de 200 etiquetas. Una vez sabemos cuántos dominios tenemos, necesitamos extraer qué palabras están relacionadas con cada dominio. Para obtener esta información extraemos las palabras de las glosas de WordNet Domains y asignamos a cada una de estas palabras el dominio asociado al sentido que están describiendo. Es decir, hemos asignado a todas las palabras de WordNet Domains su dominio o dominios correspondientes, creando así pares de palabras-dominios. Una vez obtenidos todos los pares de palabras-dominios, construimos una matriz [palabras-dominios] para obtener el espacio conceptual.

En este caso, el espacio conceptual se ha obtenido a partir de una base de datos léxica y no de un corpus determinado. Por tanto, podemos decir que las palabras se van a relacionar de acuerdo a sus relaciones semánticas, independientemente del contexto en el que aparezcan.

3. Aplicación de la medida del coseno

Para poder determinar la relación semántica entre palabras necesitamos establecer una forma de medir el grado de similitud. En este trabajo, se ha utilizado una aproximación vectorial, la medida de similitud del coseno. Esta aproximación mide la distancia entre dos palabras utilizando vectores de co-ocurrencia. Cada palabra se representa mediante un vector y el grado de similitud entre dos palabras se obtiene midiendo la distancia entre sus vectores asociados. Para obtener los vectores de co-ocurrencia hay diferentes tipos de relaciones léxicas que pueden ser utilizadas. La aproximación tradicional basada en corpus, construye un tipo de vectores llamados vectores de co-ocurrencia de palabras. Este tipo de vectores representan una palabra a partir de los patrones formados con otras palabras del corpus. Es decir, podemos medir la similitud entre dos palabras utilizando relaciones gramaticales (co-ocurrencia de palabras con una relación sintáctica específica) o utilizando relaciones no gramaticales (co-ocurrencia de palabras en una ventana de n-palabras). Sin embargo, podemos considerar otro tipo de vectores de co-ocurrencia: vectores de co-ocurrencia de documentos. En este caso, las relaciones entre palabras se obtienen a partir de un conjunto de documentos y la similitud entre cada par de palabras se calcula midiendo su solapamiento en el conjunto de documentos.

Las siguientes secciones presentan una descripción de cómo se han obtenido los vectores de co-ocurrencia para medir la similitud entre cada par de frases de T-H. En nuestro estudio, hemos utilizado dos tipos de vectores de co-ocurrencia, uno basado en la información proporcionada por un corpus, y el otro basado en el recurso léxico Dominios Relevantes, obtenido a partir de la base de datos léxica WordNet Domains.

3.1. Frecuencia de aparición en documentos

En nuestra primera aproximación hemos estudiado el efecto de utilizar la medida del coseno con la información proporcionada por el British National Corpus (BNC). Este corpus proporciona un conjunto de alrededor de 4000 documentos, a partir de los cuales, podemos establecer la similitud entre dos palabras utilizando vectores de co-ocurrencia de documentos. Nuestra intención es establecer la similitud entre las frases de Texto y las de Hipótesis midiendo su distancia semántica.

El primer paso es representar las frases de T y H mediante vectores. Cada vector tendrá alrededor de 4000 atributos, uno por cada documento del corpus BNC. En nuestro caso, cada vector representará una frase (T o H). Los valores asignados a cada atributo de los vectores se han obtenido a partir de la frecuencia de aparición de todas las palabras de la frase en cada documento del corpus. Debido a que el número de palabras en T y H es diferente, debemos normalizar los resultados obtenidos de acuerdo al número de palabras de cada frase. Además, una vez obtenida la información proporcionada por la frecuencia de las palabras en los documentos hemos calculado la *"Inverse Document Frequency"* $\left(idf_w = \log\left(\frac{N}{n_w}\right)\right)$. Donde N es el número total de documentos y n_w es el número de documentos que contienen la palabra w. Con esta medida evitamos que palabras muy comunes tengan demasiada importancia y damos mayor peso a aquellas palabras menos comunes y que por tanto, nos aportan mayor información semántica.

Una vez obtenidos los vectores de co-ocurrencia de documentos, podemos

medir la similitud entre dos sentencias (T y H) utilizando el valor del coseno $\left(\cos(T,H) = \frac{T \cdot H}{|T||H|} = \frac{\sum_{i=1}^{n} T_i \cdot H_i}{\sqrt{\sum_{i=1}^{n} {T_i}^2} \cdot \sqrt{\sum_{i=1}^{n} {H_i}^2}}\right)$.

En nuestro estudio, hemos utilizado diferentes umbrales para establecer a partir de cuál se obtienen los mejores resultados, es decir, a partir de qué valor del coseno detectamos correctamente si T y H infieren el mismo significado o no. La Tabla 2 muestra los resultados obtenidos para el mejor umbral.

3.2. Dominios Relevantes

En esta sección presentamos una segunda aproximación para establecer si T y H infieren el mismo significado. En esta aproximación, utilizamos el recurso léxico Dominios Relevantes (DR). En este caso, la información vendrá representada por vectores de co-ocurrencia de dominios.

En primer lugar, vamos a explicar cómo se ha obtenido el recurso léxico Dominios Relevantes y cómo se ha utilizado esa información para obtener los vectores de co-ocurrencia de dominios. El recurso léxico DR se ha obtenido a partir de las palabras de las glosas de WordNet Domains. Para cada palabra de la glosa tenemos un dominio asignado, éste dominio es el asociado al sentido semántico que describe la glosa. Esta información será utilizada para calcular la relevancia de una palabra con respecto a un dominio. Es decir, una vez sabemos la frecuencia de aparición de una palabra con un dominio, podemos establecer con la fórmula del Ratio de Asociación (RA) $\left(RA(w,D) = Pr(w|D)\, log_2 \frac{Pr(w|D)}{Pr(w)}\right)$ (Church y Hanks, 1990), cuáles son los dominios más relevantes de esa palabra.

Por lo tanto, el recurso léxico DR contiene todas las palabras de WordNet Domains con sus dominios asociados y ordenados según el RA. Por ejemplo, para la palabra *"organ"* tenemos un listado de cuáles serían sus dominios relevantes ordenados por el RA: *Organ{Surgery:0.189502, Radiology:0.109413, Sexuality:0.048288, Optics:0.048277, Anatomy:0.047832, Physiology:0.029388, ...:...}*.

Los vectores de dominio obtenidos tienen tantos atributos como dominios existen en WordNet Domains. Para establecer la distancia semántica entre T-H extraemos todas las palabras de cada sentencia (previamente lematizadas), obtenemos cuáles son sus DR asociados y construimos los vectores de co-ocurrencia de dominios. Por tanto, una vez hemos obtenido el par de vectores de co-ocurrencia, podemos calcular su distancia semántica utilizando la medida del coseno.

La utilización de este tipo de recurso se debe a que queremos determinar si la elección de un corpus u otro influye a la hora de establecer la distancia semántica. Es decir, podemos extraer frecuencias de aparición de palabras a partir de diferentes corpus y obtener diferentes resultados al realizar el cálculo de la distancia semántica entre el mismo par de sentencias. Por tanto, con el recurso DR tratamos de evitar la dependencia de un corpus porque los pares palabra-dominio se han extraído a partir de una base de datos léxica y las relaciones han sido obtenidas a partir de los significados de las palabras no a partir de un campo específico o de una clasificación de documentos.

4. Experimentación

En esta sección presentamos un conjunto de experimentos realizados con las aproximaciones presentadas en las secciones previas. Por una parte, se han realizado experimentos con LSI utilizando como corpus BNC, WordNet Domains, frases de Texto y frases de Hipótesis, estudiando el efecto producido al utilizar lematización o no. Y por otra parte, se han realizado experimentos utilizando la medida de similitud del coseno con frecuencia de aparición en documentos y con el recurso Dominios Relevantes. Además, se han combinado estas diferentes aproximaciones con un sistema de aprendizaje automático previo, desarrollado y evaluado para resolver la Implicación Textual, para estudiar los resultados tras añadir esta nueva información.

4.1. RTE2

Para la realización de los diferentes experimentos se ha utilizado el conjunto de datos del *"development"* y el *"test"* del *Second Recognizing Textual Entailment Challenge* (RTE2) [1]. Los ejemplos de este conjunto de datos han sido obtenidos a partir de aplicaciones reales de Extracción de Información, Recuperación de Información, Búsqueda de Respuestas y Resumen automático. El corpus proporciona 1600 ejemplos de Texto-Hipótesis, de los cuales, 800 han sido utiliza-

[1] http://www.pascal-network.org/Challenges/RTE2/Evaluation/

dos como datos para el *"development"* y los restantes como datos para el *"test"*.

Los resultados de nuestros experimentos han sido determinados de acuerdo al *script* de evaluación de RTE2. De acuerdo con este script, los sistemas han sido comparados por sus resultados de accuracy $((n^o\, ejemplos\, correctos)/(n^o\, total\, ejemplos))$.

4.2. LSI

Los experimentos basados en la información proporcionada por el método de LSI muestran que la elección de diferentes corpus influye a la hora de establecer si dos sentencias infieren el mismo significado.

Tal y como se describió en la Sección 2, hemos construido diferentes matrices a partir de diferentes tipos de corpus. Por lo tanto, para cada tipo de corpus utilizado hemos obtenido diferentes resultados. La información utilizada en cada experimento se explica a continuación[2]:

- **BNC corpus** *(LSI_BNC_NT)*. Extraemos las palabras lematizadas del corpus BNC y construimos una matriz [palabras-documentos].
- **H_sentences** *(LSILemaH, LSINoLemaH)*. Utilizamos como corpus las frases de H y construimos dos tipos de matrices: una con las palabras lematizadas y otra con las palabras no lematizadas. El resultado final son dos matrices [T_frases-H_frases].
- **T_sentences** *(LSILemaT, LSINoLemaT)*. Utilizamos como corpus las frases de T y construimos dos tipos de matrices: una con las palabras lematizadas y otra con las palabras no lematizadas. El resultado final son dos matrices [H_frases-T_frases].
- **Relevant Domains** *(LSI_RD)*. Extraemos los dominios relevantes de cada T_frase y H_frase y construimos una matriz [DR_H_frases-DR_T_frases].

La Tabla 1 muestra los resultados obtenidos en los diferentes experimentos con LSI.

Los mejores resultados se han obtenido utilizando las frases del Texto como corpus y los Dominios Relevantes. Con LSI tomando como corpus las frases del Texto, los resultados son 56.87 % para el *development* y 54.25 % para el *test*. Esta aproximación obtiene mejores resultados que la aproximación que utiliza las frases de la Hipótesis como corpus porque el Texto proporciona más información. Es decir, para establecer si dos frases infieren el mismo significado es necesario establecer un mapeo completo entre cada par de frases. Esto se consigue utilizando como corpus el Texto, ya que, contiene mayor información que la Hipótesis, de otra forma, se perdería información.

Sets	Acc.	IE	IR	QA	SUM
devLSI_BNC_NT	49.90	49.87	49.15.	50.15	50.43
devLSI_LemaH	53.25	52.00	48.00	54.00	59.00
devLSI_NoLemaH	50.17	50.15	50.03	50.22	50.28
devLSI_LemaT	**56.87**	**51.50**	**58.00**	**56.50**	**61.50**
devLSI_NoLemaT	52.88	50.50	53.00	48.00	60.00
devLSI_RD	**56.98**	**52.25**	**58.60**	**56.83**	**60.25**
testLSI_BNC_NT	49.67	49.43	49.00	50.02	50.24
testLSI_LemaH	49.38	52.50	48.50	49.00	47.50
testLSI_NoLemaH	53.37	50.50	54.00	49.00	60.00
testLSI_LemaT	**54.25**	**50.50**	**48.00**	**57.00**	**61.50**
testLSI_NoLemaT	53.63	52.50	50.00	50.00	62.00
testLSI_RD	**54.51**	**50.55**	**48.53**	**56.73**	**62.25**

Tabla 1: *Resultados para LSI*

La segunda aproximación utiliza como corpus el recurso DR. En este caso, la matriz inicial de LSI se ha obtenido a partir de la información de WordNet Domains. Una vez obtenido el espacio conceptual se ha establecido la similitud entre cada par de frases T-H. Como resultado final, hemos obtenido un porcentaje de 56.98 % para el *development* y 54.51 % para el *test*. En este caso, los resultados mejoran debido a que el corpus sobre el que se ha construido la matriz viene determinado por las relaciones semánticas de las palabras en una base de datos léxica, evitando así el problema de la dependencia del contexto cuando se utilizan corpus.

Los resultados obtenidos con el corpus BNC demuestran que la información proporcionada no es suficiente para una correcta determinación de la Implicación Textual. En la Tabla 1 se puede observar que los resultados oscilan entorno al 50 %.

4.3. Coseno

En la Tabla 2 se muestran los resultados de la medida tradicional de similitud del coseno utilizando la frecuencia de aparición en documentos y los resultados obtenidos utilizando el recurso DR. Usando la frecuencia de

[2]Cada experimento viene precedido de *dev* (*development data set*) o de *test* (*test data set*)

aparición en documentos obtenemos un 52 % de *accuracy*. Sin embargo, utilizando DR tanto el *development* como el *test* alcanzan un 54 %. Los resultados obtenidos demuestran que la información contextual proporcionada por las frases de T-H no es muy representativa y no proporciona suficiente conocimiento. Por tanto, el coseno por sí solo no puede establecer una correcta inferencia de significado, pero puede ser útil combinado con otros recursos.

Sets	Acc.	IE	IR	QA	SUM
devCosine_DF	52.60	48.63	47.32	55.13	59.32
devCosine_RD	54.25	50.50	48.00	57.00	61.50
testCosine_DF	52.18	46.13	49.43	55.34	57.83
testCosine_RD	54.00	46.50	56.50	56.00	57.00

Tabla 2: *Resultados del coseno*

4.4. Combinación de MLEnt con LSI y el coseno

Los experimentos realizados revelan que el método de LSI y la medida del coseno no son lo suficientemente potentes para establecer una correcta inferencia entre un par de frases. Sin embargo, ambas medidas pueden proporcionar información útil. Por ello, se han combinado los valores obtenidos por estas técnicas con un sistema existente de aprendizaje automático (MLEnt) basado en SVM y técnicas de stacking y voting. Las diferentes aproximaciones se han modelado como atributos.

En la Tabla 3 se muestran varios experimentos con los resultados de la combinación del sistema MLEnt con LSI y coseno. Se pueden distinguir dos tipos de experimentos: uno con el sistema previo MLEnt y otro con la combinación de LSI y coseno. Cada experimento realizado se detalla a continuación:

- **MLEnt con las características iniciales** *(ML_Lex, ML_Sem)*. MLEnt con características Léxicas o Semánticas.
- **MLEnt con LSI** *(ML_LSI_Lex, ML_LSI_Sem)*. Combinación de MLEnt con el método de LSI con las sentencias de Texto como corpus.
- **MLEnt con coseno** *(MLcosLex, MLcosSem)*. Combinación de MLEnt con la medida del coseno a partir de los Dominios Relevantes.
- **MLEnt con LSI y coseno** *(ML_LSI_cosL, ML_LSI_cosS)*. Combinación de MLEnt con LSI y coseno.

Sets	Acc.	IE	IR	QA	SUM
devML_Lex	56.87	49.50	55.50	51.00	71.50
devML_Sem	60.12	54.00	61.00	59.00	66.50
devML_LSI_Lex	**62.03**	**56.13**	**62.53**	**60.32**	**69.15**
devML_cos_Lex	56.91	49.45	55.62	52.13	70.43
devML_LSI_cosL	57.13	49.50	55.50	52.50	71.00
devML_LSI_Sem	**62.56**	**57.13**	**62.83**	**60.54**	**69.75**
devML_cos_Sem	60.21	54.13	61.06	59.14	66.54
devML_LSI_cosS	**61.75**	**56.00**	**59.50**	**62.50**	**69.00**
testML_Lex	51.75	52.00	53.50	55.50	46.00
testML_Sem	54.25	50.00	55.50	47.50	64.00
testML_LSI_Lex	**55.01**	**51.23**	**55.83**	**47.96**	**65.03**
testML_cos_Lex	52.57	49.50	44.95	53.73	62.13
testML_LSI_cosL	54.87	46.50	53.00	56.00	64.00
testML_LSI_Sem	**56.18**	**52.03**	**56.53**	**50.14**	**66.03**
testML_cos_Sem	54.42	50.22	55.62	47.61	64.25
testML_LSI_cosS	**56.50**	**53.00**	**58.00**	**57.50**	**57.50**

Tabla 3: *Resultados obtenidos en la combinación de MLEnt con LSI y coseno*

Como muestra la Tabla 3, los experimentos realizados combinando LSI y la información del coseno mejoran los resultados previos del sistema MLEnt. La adición de esta información como una nueva característica a nuestro sistema mejora los resultados. Por lo tanto, la información semántica es una buena forma de mejorar los resultados en un sistema de aprendizaje automático. De hecho, el mejor resultado obtiene un porcentaje de 62 % para el *development* y un 57 % para el *test*. Estos resultados son los obtenidos en el experimento que combina el sistema MLEnt con ambas medidas LSI y coseno.

En conclusión, podemos decir que LSI y coseno proporcionan información útil para un sistema de aprendizaje automático que trata de resolver el problema de la Implicación Textual.

5. Conclusiones y trabajo futuro

En este artículo presentamos una nueva aproximación para resolver la Implicación Textual basada en información de similitud semántica obtenida a partir de Semántica Latente y la medida de similitud del coseno. Inicialmente, se han comparado los resultados obtenidos a partir de diferentes corpus, tales como, BNC y el corpus obtenido a partir de las frases de Texto-Hipótesis. La información proporcionada por un corpus depende del dominio o del tema que se esté tratando y puede influir en la relevancia de una palabra o de toda una frase. Para evitar este tipo de dependencia, hemos propuesto dos aproximaciones: LSI y la medida del coseno, basados en el recurso WordNet Domains. En estas aproximaciones, se considera la infor-

mación proporcionada por un recurso estático, la base de datos léxica WordNet Domains. En este caso, la información contenida es más precisa que la proporcionada por un corpus donde la frecuencia de aparición de una palabra varía en función del tema que se esté tratando. Tras los experimentos realizados los resultados obtenidos para el *development* y el *test*, son bastante similares, alcanzando el 54 % para la medida del coseno y el 54.4 % para el método de LSI.

Una vez estudiada la contribución de LSI y de la medida del coseno, consideramos que la información proporcionada por estas dos técnicas no es suficiente para la correcta detección de la Implicación Textual. Por este motivo, hemos realizado una serie de experimentos, donde se han estudiado los resultados de combinar LSI, el coseno y un sistema existente de aprendizaje automático. Los experimentos realizados muestran cómo esta combinación mejora los resultados obteniendo un 61.75 % para el *development* y un 56.50 % para el *test*.

En conclusión, en este trabajo hemos demostrado que el método de la Semántica Latente es una herramienta muy potente para extraer información semántica y que puede mejorar los resultados de un sistema de aprendizaje automático existente. Para la realización de los diferentes experimentos hemos empleado diferentes funcionalidades del método LSI, sin embargo, en el futuro queremos utilizar sus propiedades para extraer sinónimos, antónimos y otro tipo de relaciones. Además, en este trabajo, la similitud semántica ha sido evaluada considerando la sentencia completa, lo cual introduce bastante ruido. Por tanto, nuestra intención es estudiar la influencia de utilizar información sintagmática en lugar de utilizar todas las palabras de una frase sin tener en cuenta el tipo de relación en los diferentes sintagmas.

Bibliografía

Aston, G. 1996. The british national corpus as a language learner resource. En *TALC 96*.

Bar-Haim, Roy, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, y Idan Szpektor. 2005. The second pascal recognising textual entailment challenge. En *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Church, K. y P. Hanks. 1990. Word association norms, mutual information and lexicograhy. *Computational Lingüistics*, 16(1):22–29.

Dagan, I. y O. Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. En *PASCAL Workshop on Learning Methods for Text Understanding and Mining*.

Dagan, I., O. Glickman, y B. Magnini. 2005. The pascal recognising textual entailment challenge. En *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, y Richard Harshman. 1990. Indexing by latent semantic indexing. En *Journal of the American Society for Information Science.*, volumen 41, páginas 321–407.

Kozareva, Z. y A. Montoyo. 2006a. Mlent: The machine learning entailment system of the university of alicante. En *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Kozareva, Z. y A. Montoyo. 2006b. The role and the resolution of textual entailment for natural language processing applications. En *11th International Conference on Applications of Natural Language to Information Systems (NLDB)*.

Magnini, Bernardo y Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. En M. Gavrilidou G. Crayannis S. Markantonatu S. Piperidis, y G. Stainhaouer, editores, *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, páginas 1413–1418, Athens, Greece.

Miller, G. 1995. Wordnet: a lexical database for english. *Communications of the ACM.*, 38(11).

Montoyo, Andrés, Sonia Vázquez, y German Rigau. 2003. Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes. *Procesamiento del Lenguaje Natural*, 30, september.