El discurso y la semántica como recursos para la detección de similitud textual

Brenda Gabriela Castro Rolón[1], Gerardo Sierra Martínez[1], Juan-Manuel Torres-Moreno[1, 2, 3], Iria da Cunha Fanego[4]

[1] Instituto de Ingeniería – Universidad Nacional Autónoma de México (UNAM)
Distrito Federal, México

[2] Laboratoire Informatique d'Avignon
Université d'Avignon et des Pays de Vaucluse (UAPV) – Aviñón, Francia

[3] École Polytechnique de Montréal – Montreal, Canadá

[4] Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra (UPF)
Barcelona, España

{bcastror, gsierram}@iingen.unam.mx,
juan-manuel.torres@univ-avignon.fr, iria.dacunha@upf.edu

Abstract. *Nowadays, the Internet provides easy access to a variety of written texts, thus facilitating plagiarism. The increase in cases of plagiarism in texts has attracted the attention of the academic community towards its detection. We present a first approach to a method based on discourse and semantics. The first phase of the method is the discursive annotation of the texts to be compared, by means of the Rhetorical Structure Theory. The second phase is the comparison of the discursive structures annotated in both texts, to detect similarities between them. The third phase is the calculation of semantic similarity of the lexical units included in the matching discourse segments in both texts. The results obtained in the preliminary experiments presented in this work are positive and confirm that discourse and semantics could be the basis of a textual similarity detection system.*

Resumen. *Hoy en día, Internet ofrece un acceso fácil a una variedad de textos escritos, lo que facilita el plagio. El aumento de casos de plagio en textos ha atraído la atención de la comunidad académica a su detección. Presentamos una primera aproximación a un método basado en el discurso y la semántica. La primera fase del método es la anotación discursiva de los textos que se desea comparar, por medio de la Rhetorical Structure Theory. La segunda fase es la comparación de las estructuras discursivas anotadas en ambos textos, para detectar coincidencias entre ellas. La tercera fase es el cálculo de similitud semántica de las unidades léxicas incluidas en los segmentos discursivos coincidentes en ambos textos. Los resultados obtenidos en los experimentos preliminares presentados en este trabajo son positivos y confirman que el discurso y la semántica podrían ser la base de un sistema de detección de similitud textual.*

Anais do III Workshop "A RST e os Estudos do Texto", páginas 30–42, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

1. Introducción

Actualmente, el incremento del uso de internet como fuente de información y la ayuda que proporcionan los procesadores de texto han facilitado el aumento de los casos de plagio de textos. Barrón-Cedeño *et al.* (2010:2) mencionan que "En las últimas dos décadas se ha observado un crecimiento importante en los casos de plagio, sobre todo el de tipo académico". Este aumento en los casos de plagio ha despertado la preocupación de la comunidad académica. Como respuesta a este aumento, se ha desarrollado un gran interés por la creación de métodos y herramientas que ayuden a detectar el plagio. Con su detección se espera disminuir el número de plagios que se llevan a cabo.

Actualmente la mayoría de métodos de detección de plagio se llevan a cabo mediante la búsqueda de similitud entre textos. Maurer *et al.* (2006) dividen los métodos actuales de detección de plagio en tres categorías generales: comparación basada en palabras, búsqueda en línea de párrafos característicos mediante motores de búsqueda y análisis estilístico.

Como ejemplo de los métodos basados en la comparación de palabras, tenemos el que usa la herramienta WCopyfind[1]. Esta herramienta se basa en lo que se suele llamar "huella digital" (Barrón-Cedeño *et al.,* 2010) para buscar en un corpus de documentos, localizados en la máquina donde se ejecuta, el posible documento fuente de un supuesto plagio. Los métodos basados en "huella digital" se llevan a cabo mediante la comparación de los fragmentos o "huellas digitales" en los que se dividen los documentos a comparar. Estos métodos pueden variar en la unidad de medida de sus "huellas digitales", partiendo desde grupos de caracteres hasta conjuntos de palabras.

En Maurer *et al.* (2006) se menciona que es posible llevar a cabo la detección de plagio mediante una búsqueda en línea de párrafos característicos de una manera prácticamente manual. Esto se refiere a que una persona debe tomar uno o varios fragmentos de su elección del texto sospechoso de plagio y hacer una búsqueda mediante cualquier motor de búsqueda en línea que considere adecuado. Actualmente existen servicios en línea que llevan a cabo esta exploración de manera automática. Un ejemplo es Plagiarism.com, el cual, según se menciona en Clough (2000), es el servicio de detección de similitud textual en línea más grande disponible.

Por otra parte, en Rosas (2011) se describe la detección de plagio mediante el análisis estilométrico con el uso del programa *Signature*[2]. Este programa permite comparar distintas características entre los textos de un corpus localizado en la máquina donde se ejecuta. Entre las características que posee este programa se encuentran la posibilidad de hacer búsquedas por medio de palabras clave, la comparación de longitudes oracionales y la prueba estadística llamada ji cuadrada (x^2), entre otras.

Finalmente, existen otros métodos basados en diversas características de los textos. Como ejemplo de estos métodos se encuentran los basados en semántica. Tanto en Jun-Peng (2004a, 2004b) como en Chi-Hong (2007) se describen métodos de detección de plagio que se basan en las características semánticas de los textos para compararlos y llevar a cabo la detección de plagio.

[1] http://plagiarism.phys.virginia.edu/
[2] http://www.philocomp.net/humanities/signature

Sin embargo, consideramos que estos métodos no bastan, ya que los plagios no sólo se llevan a cabo copiando palabras o frases exactas de un texto.

Uno de los métodos de plagio que más dificultan la detección es la paráfrasis, ya que mediante ella se puede llevar a cabo un plagio sin utilizar las mismas palabras del texto original. Mediante una paráfrasis incluso es posible modificar la estructura sintáctica de un texto plagiado. Es por esto que para llevar a cabo la detección de plagio mediante paráfrasis es necesario un análisis más fino.

Dado que todo texto posee estructuras localizadas a varios niveles, en este trabajo realizamos una primera aproximación a un método de detección de similitud textual que toma en cuenta niveles más profundos de la lengua: el discurso y la semántica. Por una parte, para atender al nivel discursivo de la lengua, empleamos como marco teórico la Rhetorical Structure Theory (RST) de Mann y Thompson (1988). Con base en la RST comparamos las estructuras discursivas de pares de textos para buscar similitudes entre ellas. Por otra parte, nos basamos en la fórmula para calcular la similitud semántica propuesta por Maynard (1999)[3] para, mediante la ontología en línea EuroWordNet[4], llevar a cabo una comparación semántica entre las unidades léxicas que contienen las estructuras discursivas de los textos.

Este trabajo tiene principalmente dos objetivos. El primero es comparar las estructuras discursivas de un corpus de textos originales y textos parafraseados a dos niveles, bajo (variación léxica) y alto (variación léxica y sintáctica, fusión o separación de oraciones), para constatar si estas estructuras son parecidas. El segundo objetivo es calcular la similitud semántica entre las unidades discursivas cuyas relaciones son iguales en los textos originales y los textos parafraseados.

En la sección 2 exponemos la metodología del trabajo. En la sección 3 presentamos los resultados obtenidos. Y, finalmente, en la sección 4 mostramos las conclusiones a las que se llegaron.

2. Metodología

La metodología empleada en este trabajo incluyó varias fases. La primera fase fue la recopilación de un corpus de paráfrasis. La segunda fase fue la realización del análisis discursivo de los textos del corpus a partir de la RST. La tercera fase fue un análisis manual en el cual se compararon las estructuras discursivas de los textos analizados mediante la RST para buscar similitudes. La cuarta fase fue el análisis automático en el que se llevó a cabo un cálculo de similitud semántica para verificar si el contenido incluido en los segmentos discursivos similares era el mismo.

2.1. Corpus

La primera fase de la metodología fue la construcción de un corpus constituido por textos originales y paráfrasis de ellos. El corpus se compone de 12 textos en español. Para la construcción de nuestro corpus se reunieron textos de diversas fuentes (Wikipedia, revistas científicas y periódicos) y temáticas (sushi, sexualidad y astronomía). Éstos se denominaron textos originales (OR). Posteriormente se solicitó a

[3] Utilizada posteriormente por Vivaldi *et al*. (2010) en su trabajo sobre resumen automático de textos especializados.
[4] http://www.illc.uva.nl/EuroWordNet

varios voluntarios (estudiantes de licenciatura, licenciados o doctores) que intencionalmente reformularan o parafrasearan dichos textos. La paráfrasis de estos textos se llevó a cabo en dos niveles:

Nivel bajo: Variación únicamente léxica.

Nivel alto: Variación léxica, sintáctica, de organización textual o discursiva y fusión o separación de oraciones.

Como paso adicional en la construcción del corpus para este trabajo se reunieron textos de las mismas temáticas y fuentes que los textos originales. Estos textos adicionales se compararon con los originales para comprobar que las coincidencias no fueran resultado de la casualidad.

Entonces, tomando en cuenta las temáticas de los textos, se puede dividir el corpus en tres sub-corpus con los temas sushi, sexualidad y astronomía. A su vez cada uno de esos sub-corpus está formado por tres tipos de texto:

1. Textos originales (OR)

2. Textos parafraseados:

a) Nivel bajo (Pb)

b) Nivel alto (Pa)

3. Textos de la misma temática que los textos originales (Pno)

Se guardaron los textos en formato de archivo .txt y se eliminaron los títulos, ya que no se analizaría la similitud en ellos y no se le pidió a los participantes que hicieran una paráfrasis de los mismos. Las estadísticas del corpus en cuanto a textos y palabras se muestran en la Tabla 1.

Tabla 1 Constitución del corpus de paráfrasis

Corpus total		
Temática	Textos	Palabras
Sushi	4	2513
Sexualidad	4	2436
Astronomía	4	3618
Total	12	8567

2.2. Anotación discursiva

El siguiente paso de nuestra metodología es la anotación discursiva de los textos del corpus. Para la primera aproximación que se presenta en este trabajo, el análisis discursivo de los textos se lleva a cabo por un solo anotador con base en los criterios mencionados en da Cunha e Iruskieta (2010), que describimos de manera general a continuación.

Con respecto a la segmentación discursiva, la primera característica que deben cumplir las Unidades Discursivas Mínimas (EDUs) es contener un verbo, ya sea en forma conjugada, en infinitivo o en gerundio. La única excepción a esto se hace en el caso de los títulos, ya que rara vez tendrán verbo, aunque, como mencionan Calsamiglia y Tusón (2007), los títulos son enunciados con fuerza retórica. No se segmentan oraciones de relativo, sustantivas, de complemento directo o indirecto, ni completivas.

Con respecto a la anotación, la lista de relaciones discursivas utilizada es la misma que se muestra en da Cunha e Iruskieta (2010), la cual contiene veintinueve relaciones discursivas. Se utilizó la relación Same-Unit propuesta por Carlson y Marcu (2001). Esta relación tiene la función utilitaria de unir los segmentos separados de una EDU en la que se inserta otra, por lo que no se tomó en cuenta para las comparaciones entre textos.

Para llevar a cabo la anotación de los textos se empleó la herramienta RSTTool (O'Donnell, 2000) puesto que es la más utilizada para el análisis de textos basado en la RST. Con los textos del corpus anotados con ayuda de la RSTTool, se contó con doce árboles discursivos con un conteo de EDUs y de relaciones discursivas como se muestra en la Tabla 2.

Tabla 2 Número de EDUs y relaciones que forman el corpus

Corpus total		
Temática	EDUs	Relaciones
Sushi	183	165
Sexualidad	139	130
Astronomía	161	133
Total	483	428

2.3. Análisis manual

Con el corpus anotado con base en la RST, se compararon las estructuras discursivas de los textos analizados mediante la RST para buscar similitudes. Se realizaron tres tipos de comparaciones entre los textos de la misma temática:

A) OR con Pa

B) OR con Pb

C) OR con Pno

Entonces, para el sub-corpus de sushi, el sub-corpus de sexualidad y el sub-corpus de astronomía se hicieron tres comparaciones respectivamente, es decir, se hizo un total de nueve comparaciones.

Para esta primera aproximación a la detección de similitud textual decidimos no analizar todas las relaciones que aparecieran en los textos. Por una parte, no se analizaron las relaciones multinucleares. Esto se debe a que las comparaciones se hacen entre parejas de textos y sería necesario tener un número estable de EDUs por relación, pero este tipo de relaciones no lo tiene. Tampoco se tomaron en cuenta las relaciones de Elaboración, ya que es la relación más frecuente en la lengua, por lo que no es característica de ningún discurso en particular[5]. Si se atiende a la Figura 1, es posible ver que del total de relaciones discursivas que no se analizaron para este trabajo casi el 60% fueron relaciones de Elaboración, lo cual nos da indicio de la gran frecuencia de uso de esta relación. En esta primera aproximación a la detección de similitud textual creímos pertinente comparar las relaciones que pudieran caracterizar un discurso para así evitar que la similitud encontrada fuera resultado de la casualidad. Por último,

[5] Se puede ver la alta frecuencia de aparición de las relaciones de Elaboración en los corpus existentes anotados con la RST. Un ejemplo es el RST Spanish Treebank (da Cunha et al., 2011), en donde se observa un 24.53% de relaciones de Elaboración.

puesto que los análisis se hacen a manera de comparaciones entre pares de textos y no es posible comparar relaciones que no se encuentren en ambos textos, no fueron analizadas las relaciones que no aparecieran en el par de textos a comparar.

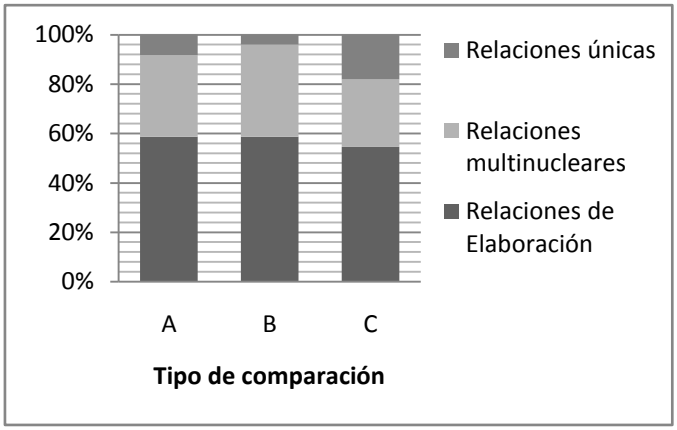

Figura 1 Composición de las relaciones discursivas que no se compararon

Una vez establecidas las relaciones objeto de análisis, el análisis manual se divide en dos fases: análisis específico y análisis general. En el análisis específico se lleva a cabo la comparación del contenido de las EDUs cuyas relaciones se determinó, dadas las características previamente mencionadas, eran objeto de análisis para este trabajo. Esta comparación de contenido se llevó a cabo para denotar las unidades en las que se intuía coincidencia en el contenido. Es decir, se revisaron las EDUs cuya relación discursiva coincidía para detectar la existencia de similitud en el contenido de éstas. Gracias a esto fue posible identificar, para cada comparación, las EDUs coincidentes en tanto a relación y contenido.

Para realizar este paso de la metodología se hicieron varias tablas comparativas tales como la Tabla 3. En estas tablas se colocó el contenido de las EDUs de los textos a comparar en columnas paralelas. Además, se colocó a la izquierda de esas EDUs el nombre de la relación que mantienen. Esto se hizo con el fin de verificar la coincidencia no sólo de contenido sino de relaciones discursivas.

Si se encontraban EDUs que coincidieran en tanto a relación y contenido, se colocaban en paralelo y, en una columna en el extremo izquierdo, se marcaban con un 1. En caso de que se encontraran EDUs sin coincidencia, se colocaban al lado de un espacio vacío y en la columna del extremo izquierdo se marcaban con un 0. Estas marcas se colocaron no sólo para distinguir las EDUs que tenían coincidente, sino también para facilitar el análisis general que se haría a continuación.

Tabla 3 Comparación tipo B del contenido de las EDUs en el sub-corpus de astronomía

	Comparación B astronomía					
	OR			Pb		
	Relación discursiva		EDUs	Relación discursiva		EDUs
1	Propósito	Núcleo	El festival astronómico lleva como título "¡Haz Química con el Universo!"	Propósito	Núcleo	El evento lleva como título "¡Haz Química con el Universo!"
		Satélite	para unirse a las celebraciones del Año Internacional de la Química, declarado por la Asamblea General de la Organización de las Naciones Unidas (ONU) el 30 de diciembre de 2008, a propuesta de la Unión Internacional de Química Pura y Aplicada.		Satélite	para unirse a las celebraciones del Año Internacional de la Química, declarado por la Asamblea General de la ONU el 30 de diciembre de 2008, a propuesta de la Unión Internacional de Química Pura y Aplicada.
1	Resultado	Núcleo	La primera noche de las estrellas se realizó el 31 de enero de 2009 en 26 sitios arqueológicos y plazas públicas de 22 estados de la República.	Resultado	Núcleo	La primera vez se realizó el 31 de enero de 2009 en 26 sitios arqueológicos y plazas públicas de 22 estados de la República Mexicana.
		Satélite	Esta celebración logró convocar a más de 210,000 personas (el doble de la capacidad del Estadio Azteca);		Satélite	Este evento logró convocar a más de 210,000 personas (que corresponde al doble de la capacidad del Estadio Azteca);
0	Resultado	Núcleo	La segunda noche de las estrellas se realizó el 17 de abril de 2010 con el tema "Nuestro Universo en Movimiento"	Resultado	Núcleo	
		Satélite	reuniendo en esa ocasión más de mil 320 telescopios en 31 sedes		Satélite	

En el análisis general se contabilizaron los resultados obtenidos del análisis específico para así poder observar un panorama de éstos. Por una parte, se contabilizaron las coincidencias y diferencias dentro de cada comparación realizada. Con esta contabilización fue posible observar el tipo de comparación en el que se tuvo un número más elevado de coincidencias de EDUs en tanto a relación y contenido. Además se contabilizó dentro de cada comparación, como se observa en la Tabla 4, el número de coincidencias por relación discursiva para así observar en cuáles hay mayor cantidad de coincidencias.

Tabla 4 Conteo de coincidencias y diferencias en el sub-corpus de sexualidad

Sub-corpus Sexualidad			
	Relación	Coincidencias	Diferencias
A	Concesión	0	2
	Interpretación	1	1
	Resultado	0	2
Total		1	5
B	Causa	1	0
	Concesión	2	2
	Fondo	0	1
	Interpretación	0	2
	Resultado	2	0
Total		5	5

2.4. Análisis automático

El paso siguiente de la metodología efectúa un análisis automático. Éste realiza el cálculo de la similitud semántica entre las EDUs que el análisis manual identificó como coincidentes. Este cálculo permite comprobar los resultados del análisis manual y obtiene una medida de similitud semántica entre las EDUs coincidentes.

Para este cálculo se desarrolló un programa perl que utiliza la fórmula de Maynard (1999). Además se utilizaron varios recursos del sistema de resumen

automático Cortex (Torres-Moreno *et al*., 2001; Torres-Moreno *et al.*, 2009). Específicamente se utilizaron las listas de palabras funcionales y el reagrupamiento de familias léxicas.

Se optó por incluir el reagrupamiento de familias léxicas porque es una alternativa a la lematización. Mediante este proceso, en lugar de lugar de normalizar cada palabra por su lema, éstas se sustituyen por un representante de su familia léxica. A pesar de que la lematización disminuye la cantidad de palabras a comparar, esta disminución no resulta tan significativa como con el reagrupamiento de familias léxicas. Por ejemplo, la lista {cantaríamos, cantante, cantantes, cantos, canción, canciones} es lematizada a {cantar}, {cantante}, {canto}, {canción}; en cambio la misma lista puede ser reagrupada en una familia representada por la única palabra {**cantar**}. Este proceso de reagrupar en familias léxicas aumenta las probabilidades del algoritmo de encontrar palabras en EuroWordNet, recurso en constante desarrollo.

El cálculo de similitud semántica se realizó por pares de EDUs. De estas EDUs, una proviene del texto original y la otra del texto a comparar. Se compararon núcleos con núcleos y satélites con satélites para mantener la correspondencia del análisis discursivo.

El primer paso del algoritmo es un pre-procesamiento de la EDU del texto original O y la EDU del texto con el que se compara el original P. El pre-proceso filtra O y P para remover signos de puntuación. Después se normalizan O y P a minúsculas, se eliminan las palabras funcionales de ambas EDUs usando las listas de Cortex y finalmente se reagrupan las palabras en familias léxicas, también usando los recursos de Cortex. Este pre-proceso genera dos listas de palabras O' y P'.

En el segundo paso, cada palabra de la lista O' se compara con cada palabra de la lista P'. La lista O' contiene i=1,2,...,m palabras y la lista P' contiene j=1,2,...,n palabras. n no necesariamente es igual a m por lo que se comparan en total n x m parejas.

Para el cálculo de similitud, el algoritmo busca en EuroWordNet los términos O_i y P_j a comparar. Sólo se comparan verbos con verbos y sustantivos con sustantivos, dada la arquitectura actual de EuroWordNet. Entonces, usando la información proveniente de las rutas hiperonímicas de esta ontología se calcula la similitud semántica entre pares de palabras usando la ecuación propuesta por Maynard (1999):

$$similitud\ (synset_1, sinset_2) = 2 \times número\ de\ nodos\ comunes\ (synset_1, sinset_2) / profundidad(synset_1) + profundidad(synset_2)$$

Este cálculo se traduce como: la similitud entre dos *synsets* es igual al número de nodos comunes de los *synsets* multiplicado por 2 y dividido entre la suma de las profundidades de los dos *synsets*. Vivaldi *et al*. (2010) proporcionan un ejemplo de este cálculo usando las palabras en inglés “vas” y “gland” (Figura 2)[6].

[6] La figura se tomó de Vivaldi *et al*. (2010)

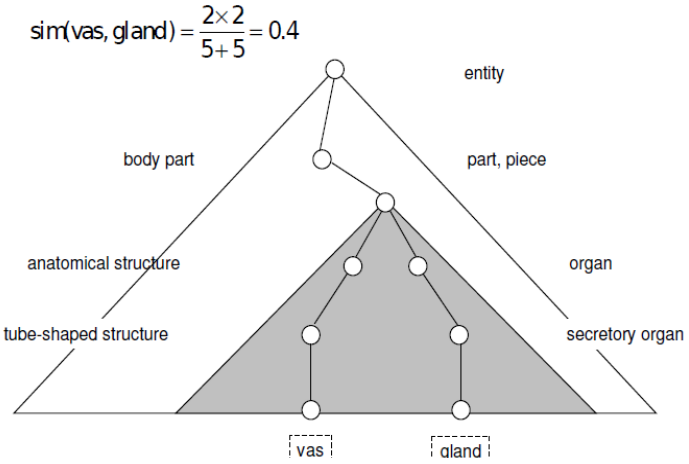

Figura 2 Ejemplo de cálculo de similitud semántica

Alguna de las palabras O_i o P_j pueden no encontrarse en EuroWordNet. El resultado de la similitud de esta pareja se tomó como 0. Si ambas palabras a comparar resultaban idénticas, aún si no se encontraran en EuroWordNet, su similitud resultante sería un valor de 4. Este último valor se definió dado que, en primer lugar, la utilización repetida de palabras idénticas es una marca de alto grado de paráfrasis y, en segundo lugar, empíricamente encontramos que este valor compensa la gran cantidad resultante de valores 0 de similitud. Si no se llegara a dar ninguno de estos dos casos, se calcula la similitud siguiendo la fórmula propuesta por Maynard (1999) descrita anteriormente.

El cálculo de similitud semántica se realiza para cada par posible de palabras y se acumula el resultado para obtener un valor único de similitud entre todas las EDUs. El resultado final se obtiene mediante la división de este valor único entre las *n*x*m* parejas de palabras comparadas. El resultado se encuentra normalizado entre 0 y 1. 0 significa ausencia de similitud y 1 significa similitud absoluta entre los fragmentos textuales O y P.

3. Resultados

En este apartado expondremos los resultados obtenidos mediante el análisis manual y el automático descritos anteriormente.

3.1. Análisis manual

Gracias al análisis general que se llevó a cabo, dentro del análisis manual pudimos observar de manera global los resultados obtenidos.

El cálculo de coincidencias por relación nos muestra las relaciones en las que se encontró mayor número de coincidencias de contenido entre las EDUs del texto original y las del texto con el que se comparó. Gracias a este cálculo pudimos ver que para todas las comparaciones que se hicieron en el corpus, en promedio, 39.5% de las EDUs en las que se encontró coincidencia poseían la relación de Resultado. También se observó que para el sub-corpus de temática sushi la relación de Causa representa el 28.5% de las coincidencias.

Por otra parte, la Tabla 5 contiene el promedio de las coincidencias y diferencias encontradas para cada tipo de comparación. En esta tabla podemos ver que, en promedio, la mayor cantidad de coincidencias se encontró en la comparación de tipo B.

Por otra parte, es posible también observar que en la comparación de tipo C no se encontró ninguna coincidencia. Estos resultados se pueden ver como un primer indicio de la similitud textual.

Tabla 5 Promedio de resultados del análisis general

Análisis general			
		Coincidencias	Diferencias
A	OR – Pa	2.67	10.33
B	OR – Pb	4.67	5.33
C	OR – Pno	0.00	6.33
Total		7.33	22.00

3.2. Análisis automático

Mediante el cálculo de similitud semántica que se explica en el apartado 3.2 se obtuvieron los resultados que se observan en la Figura 3. Estos resultados se muestran por nivel de paráfrasis. Cada nivel de paráfrasis corresponde a uno de los tipos de comparación: baja corresponde al tipo B, alta al tipo A y nula al tipo C.

La Figura 3 muestra el promedio de similitud semántica que se obtuvo en cada comparación. En esta figura se puede observar la diferencia entre los resultados obtenidos del cálculo de similitud semántica para cada nivel de paráfrasis.

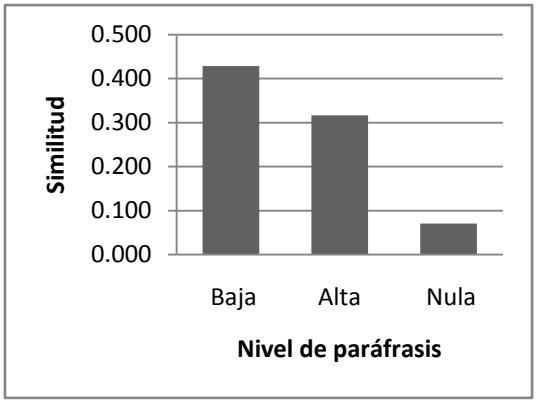

Figura 3 Gráfico de resultados de similitud semántica

Para mostrar un resultado de similitud semántica para la comparación de tipo C se calculó la similitud semántica entre EDUs tomadas al azar que poseyeran la misma relación discursiva. Esto dado que en el análisis manual no se encontraron coincidencias para este tipo de comparación y se creyó necesario tener los datos de la similitud semántica entre los textos que no eran paráfrasis. Esta comparación se usó para comprobar si, a pesar de que estos textos poseían el mismo origen textual y hacían referencia al mismo tema, existía diferencia alguna con los resultados obtenidos de las paráfrasis. Como se puede observar en la Figura 3, es evidente la diferencia entre la similitud semántica obtenida para las paráfrasis y sus originales, y los pares de textos que no son paráfrasis. Esto sucede a pesar de que estos pares de textos refieren a la misma temática y poseen un origen similar.

4. Conclusiones

Estos resultados confirman la posibilidad de encontrar similitudes entre las estructuras discursivas de un texto original y de un texto que lo parafrasea. Además se puede

observar en la Tabla 5 que, cuanto mayor sea la complejidad de la paráfrasis, es mayor la diferencia entre las estructuras discursivas.

También, como se puede observar en la Figura 3, hay una diferencia significativa entre la similitud semántica obtenida de la comparación entre textos sin paráfrasis y la comparación de los textos originales con paráfrasis. Esto indica que con esta metodología es posible comparar textos que tratan el mismo tema y distinguir entre ellos cuáles son paráfrasis. Atendiendo a esto, se puede ver que cuanto menor es la complejidad de una paráfrasis, mayores son los resultados del cálculo de similitud semántica. Lo que indica que además es posible hacer una distinción entre niveles de complejidad de paráfrasis según los resultados del cálculo de similitud textual.

Dado que existen técnicas de detección de similitud textual que utilizan como único método la comparación de textos a nivel de palabra, es posible pensar que sería adecuado seguir un método que sólo haga un cálculo de similitud semántica o que sólo compare los textos a nivel de estructura discursiva. Pero atendiendo a las aportaciones y las carencias que presentan las dos partes del método presentado en este trabajo, tanto el análisis manual como el automático, creemos que no se podrían utilizar como métodos de detección de similitud textual individualmente.

Por una parte, el análisis discursivo previo de los textos aporta tres cosas: confirma la sospecha de similitud entre textos, indica los fragmentos de texto que son similares en estructura discursiva y, con esto, permite discriminar ciertos fragmentos textuales para su posterior comparación semántica. Pero es cierto que dos textos pueden compartir una misma estructura discursiva a pesar de tener contenidos semánticos distintos. Por otra parte, el cálculo de similitud semántica especifica la medida en que son similares los textos y, como podemos ver por los resultados obtenidos (ver Figura 3), permite discriminar las paráfrasis según el nivel de complejidad que presentan. Pero si solamente se tomara en cuenta la similitud semántica entre textos, es posible que textos que refieran a un mismo tema, a pesar de no ser plagio o paráfrasis uno de otro, tuvieran un grado muy alto de similitud semántica.

Teniendo esto en cuenta podemos concluir que los resultados de este trabajo no sólo podrían ser un punto de partida para la implementación de una nueva técnica de detección de similitud textual, sino que cabe la posibilidad de utilizar los resultados del cálculo similitud semántica como base para un clasificador de textos parafraseados según la complejidad de la paráfrasis.

4.1. Aportaciones

La principal aportación de este trabajo deriva del hecho que, hasta donde los autores tienen conocimiento, no hay metodologías que se basen tanto en análisis discursivo como en cálculos semánticos para la detección de similitud textual. Aquí se presenta un método que sí lo hace, por lo que es posible tomar como innovador el método propuesto.

4.2. Trabajo futuro

En esta primera aproximación a una nueva metodología de detección de similitud textual, no se consideró más importante la coincidencia de una relación discursiva en particular, sino que se tomaron estas relaciones como distintivos de las estructuras discursivas de los textos a comparar. En trabajos futuros podría ser interesante dar

mayor o menor importancia a las EDUs coincidentes según la relación discursiva que poseen.

Además, dado que esta metodología ha arrojado los resultados esperados, como trabajo futuro se podría plantear la verificación de los resultados obtenidos en un corpus de mayor tamaño. Finalmente, dado que en este trabajo no se hizo, se podría considerar el análisis tomando en cuenta las relaciones multinucleares y de Elaboración.

4.3. Otras posibles aplicaciones

Teniendo en cuenta los resultados obtenidos en este trabajo, es posible observar que, además de que la detección de similitud textual se puede utilizar en la detección de plagio, esta metodología puede utilizarse en otras aplicaciones.

Primero, sería posible utilizar este método para la atribución de autoría. Esto es, comparar las estructuras de varios textos de un autor con un texto supuestamente de su autoría; en este caso, si hubiera similitud textual, sería probable que fuera del mismo autor. Segundo, si se compararan las respuestas de los exámenes de las personas examinadas con un texto previamente resuelto por el examinador sería posible utilizar este método para la evaluación de exámenes. Si hubiera similitud alta, las respuestas serían presumiblemente correctas.

A pesar de que esta metodología se diseñó pensando en el español, puesto que la RST es, hasta cierto punto, independiente de la lengua y dado que EuroWordNet existe en varios idiomas, también sería posible implementar esta metodología para distintas lenguas.

Finalmente, es de mencionar que la metodología propuesta en este trabajo no sólo se mostró útil en la detección de similitud textual, sino que también, gracias a los resultados que se presentan, se observa la posibilidad de distinguir entre niveles de paráfrasis. Es decir, gracias a este método no sólo se puede definir la existencia o no de similitud entre pares de textos, sino que se puede observar el nivel de complejidad que presenta una paráfrasis.

Bibliografía

Bao, J-P., Shen, J-Y., Liu, X-D., Liu, H-Y. y Zhang, X-D. (2004a). Semantic Sequence Kin: A Method of Document Copy Detection. En *Proceedings of Advances In Knowledge Discovery and Data Mining,* vol. 3056, 529-538. Sydney, Australia: Lecture Notes in Artificial Intelligence (LNAI).

Bao, J-P., Shen, J-Y., Liu, X-D., Liu, H-Y. y Zhang, X-D. (2004b). Finding Plagiarism Based on Common Semantic Sequence Model. En *Proceedings of the 5th International Conference on Advances in Web-Age Information Management* (WAIM)*,* vol. 3129, 640-645. Dalian, China: Lecture Notes in Computer Science.

Barrón-Cedeño, A., Vila, M. y Rosso, P. (2010). Detección automática de plagio: de la copia exacta a la paráfrasis. En *Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica. Jornadas (in)formativas de lingüística forense,* 76-96. Madrid, España: Euphonia Ediciones SL.

Calsamiglia, H. y Tusón, A. (2007). *Las cosas del decir: Manual de análisis del discurso*, Ariel, 2ª ed.

Carlson, L. y Marcu, D. (2001). *Discourse Tagging Reference Manual* [Reporte técnico]. University of Southern California, Information Sciences Institute, EUA. ISI-TR-545. http://www.isi.edu/~marcu/discourse, Junio.

Chi-Hong, L. y Yuen-Yan, C. (2007). A Natural Language Processing Approach to Automatic Plagiarism Detection. En *Proceedings of the 8th ACM Conference on Information Technology Education (SIGITE'07*), 213–218. Florida, EUA: ACM Press. ISBN: 978-1-59593-920-3.

Clough, P. (2000). *Plagiarism in natural and programming languages: an overview of current tools and technologies* [Reporte técnico]. University of Sheffield, Dept. of computer Science, Reino Unido. CS-00-05. ftp://ftp.dlsi.ua.es/people/armando/maria/Plagiarism.rtf, Junio.

da Cunha, I. e Iruskieta, M. (2010). Comparing rhetorical structures of different languages: The influence of translation strategies. *Discourse Studies, 12* (5), 563-598.

da Cunha, I., Torres-Moreno, J.-M. y Sierra, G. (2011). On the Development of the RST Spanish Treebank. En *Proceedings of the Fifth Law Workshop (LAW V)*, 1-10. Oregon, EUA: Association for Computational Linguistics. http://aclweb.org/anthology-new/W/W11/W11-0401.pdf, Junio.

Mann, W. C. y Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, *8* (3), 244-281.

Maurer, H., Kappe, F. y Zaka, B. (2006). Plagiarism - A Survey. *Journal of Universal Computer Science*, *12* (8), 1050-1084.

Maynard, D. (1999). *Term recognition using combined knowdlege sources*. Tesis doctoral, Manchester Metropolitan Univerity, Manchester, Inglaterra.

O'Donnell, M. (2000). RSTTool 2.4 -- A Markup Tool for Rhetorical Structure Theory. En *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, 253-256. Mitzpe Ramon, Israel: Association for Computational Linguistics.

Rosas, A. (2011). *Análisis estilométrico para la detección de plagio.* Tesis de licenciatura, Universidad Nacional Autónoma de México, Ciudad de México, México.

Torres-Moreno, J.-M., St-Onge, P.-L., Gagnon, M., El-Bèze, M. y Bellot, P. (2009) Automatic Summarization System coupled with a Question-Answering System (QAAS), *CoRR*. cs.IR: 0905.2990v1

Torres-Moreno, J.-M., Velázquez, P. y Meunier, J.-G. (2001). Cortex: un algorithme pour la condensation automatique des textes. En *La cognition entre individu et societé* ARCo 2001, vol. 2, 365. Lyon, Francia: Hermès Science.

Vivaldi, J., da Cunha, I., Torres-Moreno, J.-M. y Velázquez, P. (2010). Automatic Summarization Using Terminological and Semantic Resources. En *7th International Conference on Language Resources and Evaluation (LREC 2010).* Valleta, Malta.