Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 21-28 recibido 19-11-12 revisado 08-02-13 aceptado 19-02-13

Análisis de similitud basado en grafos: Una nueva aproximación a la detección de plagio translingüe*

Graph-Based Similarity Analysis: A New Approach to Cross-Language Plagiarism Detection

Marc Franco-Salvador, Parth Gupta y **Paolo Rosso**
Natural Language Engineering Lab - ELiRF
Departamento de sistemas informáticos y computación
Universitat Politècnica de València
{mfranco,pgupta,prosso}@dsic.upv.es

Resumen: La variante translingüe de la detección de plagio automática trata de detectar plagio entre documentos en diferentes idiomas. En los últimos años se han propuesto una serie de aproximaciones que hacen uso de tesauros, modelos de alineamiento o diccionarios estadísticos para lidiar con la similitud a través de idiomas. En este trabajo proponemos una nueva aproximación a la detección de plagio translingüe que hace uso de una red semántica multilingüe para generar grafos de conocimiento, obteniendo un modelo de contexto para cada documento, de lo cual carecen otros métodos. Para evaluar nuestra propuesta, utilizamos las particiones español-inglés y alemán-inglés del corpus PAN-PC'11, comparando nuestros resultados con dos de las aproximaciones del estado del arte. Los resultados experimentales indican su potencial como alternativa para el análisis de similitud en detección de plagio translingüe.
Palabras clave: Detección de plagio translingüe, similitud textual, red semántica multilingüe, BabelNet, grafos de conocimiento.

Abstract: Cross-language variant of automatic plagiarism detection tries to detect plagiarism among documents across language pairs. In recent years a few approaches are proposed that use thesauri, alignment models or statistical dictionaries to deal with the similarity across languages. We propose a new approach to the cross-language plagiarism detection that makes use of a multilingual semantic network to generate knowledge graphs, obtaining a context model for each document which the other methods lack. To evaluate the proposed method, we use the Spanish-English and German-English partitions of the PAN-PC'11 corpus and compare our results with two state-of-the-art approaches. Experimental results indicate its potential to be a new alternative for similarity analysis in cross-language plagiarism detection.
Keywords: Cross-language plagiarism detection, textual similarity, multilingual semantic network, BabelNet, knowledge graphs.

1 Introducción

El plagio translingüe es definido como el uso no autorizado del contenido original de la obra de otros autores desde una fuente en otro idioma. Actualmente es un grave problema para los autores que además se ha complicado a causa de Internet. Éste pone a nuestra disposición, de forma gratuita y sencilla, una gran fuente de información y las herramientas necesarias para traducir y copiar contenidos originales. La investigación dentro del campo de la detección de plagio translingüe está justificada. En una encuesta realizada recientemente sobre las actitudes y prácticas de los estudiantes (Barrón-Cedeño, 2012), se pone de manifiesto que el plagio translingüe es un problema real: un 63.75 % de los estudiantes opina que copiar y traducir fragmentos de

* Agradecer a la Conselleria d'Educació, Formació i Ocupació de la Generalitat Valenciana por la financiación por parte del programa Gerónimo Forteza, sin el cual no hubiera sido posible llevar a cabo la investigación del primer autor que ha llevado a esta publicación. Este trabajo se ha hecho dentro del ámbito del *VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems* y como parte del proyecto de la Comisión Europea WIQ-EI IRSES (no. 269180).

ISSN 1135-5948

texto desde otros documentos y incluirlos en sus trabajos no es plagio.

La detección de plagio puede ser realizada de forma manual, pero dada la gran cantidad de obras publicadas, es muy complicado detectar los casos, aun más si la fuente del plagio proviene de otro idioma. Existen una serie de aproximaciones para llevar a cabo la detección de plagio translingüe de forma automática. Éstas hacen uso de tesauros, modelos de alineamiento o diccionarios estadísticos para detectar la similitud a nivel translingüe. *Cross-language character n-gram* (CL-CNG) (Mcnamee y Mayfield, 2004) es un modelo que se basa en la sintaxis de los documentos, haciendo uso de n-gramas, que ofrece un rendimiento notable para lenguajes con similitudes sintácticas. *Cross-language explicit semantic analysis* (CL-ESA) (Potthast et al., 2011a) es un modelo de análisis de semejanzas de colecciones relativas, lo que significa que un documento está representado por sus similitudes con una colección de documentos, las cuales son comparadas con un modelo de detección de similitud monolingüe. *Cross-language alignment-based similarity analysis* (CL-ASA) (Barrón-Cedeño et al., 2008; Pinto et al., 2009) se basa en la tecnología de máquinas de traducción estadística, la cual combina traducciones estadísticas, usando diccionarios estadísticos, y análisis de similitud. Los anteriores modelos han sido comparados (Potthast et al., 2011a), ofreciendo CL-ASA y CL-CNG el mejor desempeño. Por ese motivo, en nuestra evaluación comparamos nuestra aproximación con éstos.

Nuestra nueva aproximación, llamada *cross-language knowledge graphs analysis* (CL-KGA), proporciona un modelo de contexto de los documentos sospechosos y fuente a comparar. Para ello utiliza grafos de conocimiento generados por una red semántica multilingüe, los cuales expanden y relacionan los conceptos originales del texto. Así, la similitud entre documentos se mide mediante un método de análisis de similitud entre grafos.

Para la evaluación de los modelos utilizamos el corpus del PAN-PC'11 (Potthast et al., 2011b)[1], la competición internacional celebrada de forma anual en el marco de *Uncovering Plagiarism Authorship and Social Software Misuse* (PAN)[2], en la cual se presentan y ponen a prueba aproximaciones para la detección de plagio a nivel monolingüe y translingüe. Para nuestra evaluación utilizamos su partición de detección de plagio translingüe.

La estructura de la publicación es la siguiente: En la sección 2 explicamos en que consiste una red semántica multilingüe. En la sección 3 presentamos el modelo CL-KGA de análisis de similitud, y describimos los modelos con los que lo comparamos: CL-CNG y CL-ASA. En la sección 4 evaluamos nuestra aproximación utilizando los casos español-inglés (es-en) y alemán-inglés (de-en) de la tarea de detección de plagio externo del corpus del PAN-PC'11, comparando nuestros resultados con los obtenidos por los otros dos modelos. Finalmente, en la sección 5 presentamos nuestras conclusiones y trabajos futuros.

2 Red semántica multilingüe

Una red semántica multilingüe (RSM) consiste en un grafo dirigido y ponderado donde los nodos representan conceptos y nombres de entidades, y las aristas representan relaciones entre ellos. Además, cada uno de los nodos tiene una dimesión multilingüe con el conjunto de las lexicalizaciones del concepto en diferentes idiomas. En este trabajo, a partir de fragmentos de texto, vamos a utilizar una RSM para construir grafos de conocimiento, y compararlos entre ellos para detectar plagio translingüe.

La aproximación que describimos en la sección 3 es genérica y puede ser utilizada con cualquier RSM como ConceptNet[3] o EuroWordNet[4], pero para nuestros experimentos hemos elegido BabelNet (Navigli y Ponzetto, 2010). Ésta está formada por una base de conocimiento de gran tamaño, con el conjunto de lexicalizaciones de los conceptos disponibles en los siguientes idiomas: alemán, catalán, español, francés, inglés e italiano. Sus relaciones y conceptos provienen de WordNet, la mayor red semántica disponible, y de las entradas multilingüe de la Wikipedia[5], así BabelNet combina información lexicográfica con conocimiento enciclopédico. La lista de conceptos está formada por to-

[1] URL: http://www.uni-weimar.de/cms/medien/webis/research/corpora/corpus-pan-pc-11.html

[2] URL: http://pan.webis.de/

[3] URL: http://csc.media.mit.edu/conceptnet/

[4] URL: http://www.illc.uva.nl/EuroWordNet/

[5] URL: http://www.wikipedia.org/

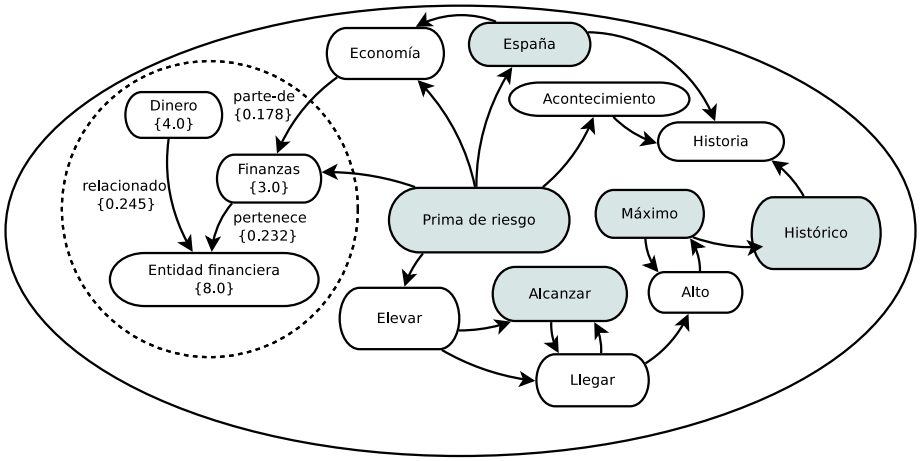

Figura 1: Ejemplo simplificado, sin dimensión multilingüe, del grafo de conocimiento de la frase "la prima de riesgo española alcanza máximos históricos" (las relaciones y los pesos de nodos y aristas se incluyen sólo dentro de la línea discontinua).

dos los significados de palabra en WordNet y de las entradas etiquetadas de la Wikipedia, por otro lado las relaciones entre conceptos las forman los punteros semánticos entre conceptos en WordNet y los enlaces entre entradas en la Wikipedia. Por último, las lexicalizaciones multilingüe se obtienen a partir de las entradas en diferentes idiomas de la Wikipedia.

El API de BabelNet permite utilizar esta RSM, entre otros usos, como diccionario estadístico, traductor, para desambiguación del sentido de las palabras (Navigli y Ponzetto, 2012), y para construir grafos de conocimiento.

2.1 Grafos de conocimiento

Un grafo de conocimiento consiste en un grafo dirigido y ponderado, generado a partir de un conjunto de palabras como las de una frase, que contiene los conceptos originales expandidos y relacionados entre ellos, dando lugar a un "modelo de contexto" de la frase o conjunto de palabras original. El peso de un concepto es el número de relaciones salientes, mientras que el peso de una relación está en función del número de relaciones que conectan con sus conceptos origen y destino[6]. Para comprender mejor en que consiste un grafo de conocimiento, vamos a poner un ejemplo. Supongamos que tenemos la frase "La prima de riesgo española alcanza máximos históricos". Los conceptos de la frase son $C = \{$prima de riesgo, españa, alcanzar, máximo, histórico$\}$. En BabelNet podemos construir un grafo de conocimiento g a partir de C, el cual contendrá un nuevo listado de conceptos $C_g = (C \cup C')$, siendo la lista de conceptos expandidos $C' = \{$economía, finanzas, historia...$\}$. Además, entre los conceptos de C_g, aparecerán una serie de conexiones R que los relacionan, $R \in \{$relacionado, parte de, pertenece, equivalente, opuesto...$\}$. Además, como se ha comentado anteriormente, cada uno de los conceptos y relaciones del grafo tiene una dimensión multilingüe, por lo que dos fragmentos de texto similares en diferentes lenguajes, deberían tener unos grafos también similares. En la fig. 1 podemos ver el contenido del grafo g en español.

3 Modelos de detección de plagio translingüe

De acuerdo a su paradigma de resolución, existen diferentes categorías de modelos de análisis de similitud que pueden ser utilizados para detección de plagio translingüe: (i) modelos que hacen uso de diccionarios, *gazetteers*, reglas o tesauro lingüísticos para realizar las traducciones de los conceptos desde un lenguaje origen L a uno destino L'. En esta categoría tenemos por ejemplo CL-VSM (Steinberger, Pouliquen, y Ignat, 2004), que utiliza modelos de espacio vectorial (Stein y Anderka, 2009) o CL-CTS (Gupta, Barrón-Cedeño, y Rosso, 2012), que usa el tesauro conceptual Eurovoc para las traducciones. (ii) Modelos que se basan en la sintaxis del documento y su estructura para comparar los documentos. El CL-CNG (Mcna-

[6]En (Navigli y Ponzetto, 2010) tenemos la fórmula de estimación de pesos de relaciones.

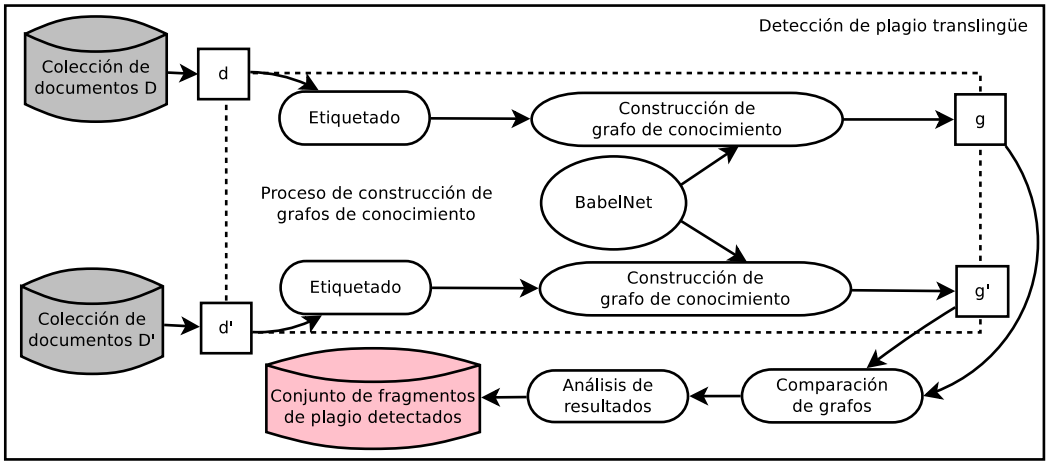

Figura 2: Proceso de detección de plagio translingüe utilizando grafos de conocimiento.

mee y Mayfield, 2004) está incluido en esta categoría. (iii) Modelos que utilizan corpus comparables, como CL-ESA (Potthast et al., 2011a). Éste utiliza corpus alineados por tema y idioma, como la enciclopedia de la Wikipedia, y analiza la similitud con un modelo monolingüe como los modelos de espacio vectorial. (iv) Los modelos basados en un corpus paralelo alinean los corpus en diferentes idiomas a nivel de documento y palabra. Los modelos CL-ASA (Barrón-Cedeño et al., 2008; Barrón-Cedeño, 2012), CL-LSI (Dumais et al., 1997) y CL-KCCA (Vinokourov, Shawe-Taylor, y Cristianini, 2003), quedan dentro de esta categoría. Además, existen modelos que pueden utilizar combinaciones de las categorías anteriores, para lo cual nuestra aproximación CL-KGA es el ejemplo perfecto, siendo la RSM BabelNet la unión de (i) un tesauro (WordNet) y de (iii) un corpus comparable (Wikipedia).

Dejando de lado aproximaciones como CL-LSI y CL-KCCA que ofrecen un elevado rendimiento a un alto coste computacional, existen trabajos (Potthast et al., 2011a; Gupta, Barrón-Cedeño, y Rosso, 2012) que han comparado algunos de los anteriores modelos: CL-ASA, CL-ESA, CL-CNG y CL-CTS. En sus resultados se refleja como CL-CNG es un buen *baseline* para tomar como partida en la detección de plagio translingüe, y CL-ASA ofrece en promedio los mejores resultados. Por esa razón hemos elegido CL-CNG y CL-ASA como las aproximaciones a comparar, en la evaluación, con nuestro modelo.

A continuación vamos a describir nuestra nueva propuesta, CL-KGA, y los dos modelos con los que la comparamos.

3.1 Análisis de similitud basado en grafos de conocimiento

La aproximación que proponemos en esta publicación, CL-KGA, utiliza grafos de conocimiento generados a partir de una RSM para obtener una similitud entre dos textos, como por ejemplo documentos o fragmentos de texto. Dado un conjunto de documentos D en un lenguaje L_1 y un conjunto de documentos D' en un lenguaje L_2, para comparar dos documentos $d \in D$ y $d' \in D'$, en primer lugar debemos realizar un procesado previo del texto para extraer y etiquetar morfológicamente sus conceptos. Además, es conveniente lematizar el texto. Para todas estas tareas, en nuestra investigación hemos hecho uso de la herramienta TreeTagger [7]. Una vez procesado el texto, podemos construir, utilizando la RSM BabelNet, los grafos de conocimiento g y g' a partir de los documentos d y d'. En la fig. 2 podemos ver un esquema del proceso de detección de plagio translingüe utilizando grafos de conocimiento. Para obtener una similitud $S(g, g')$ entre g y g', tomando como base la aproximación de comparación flexible de grafos conceptuales[8] (Montes y Gómez et al., 2001), hemos propuesto la ecuación 1 para trabajar con grafos de conocimiento.

$$S(g, g') = S_c(g, g') * (a + b * S_r(g, g')) \qquad (1)$$

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

[8] Un grafo conceptual es un grafo finito dirigido bipartido con dos clases de nodos: conceptos y relaciones (Sowa, 1984; Sowa, 1999).

$$S_c(g,g') = \frac{\left(2 * \sum_{c \in g_u} w(c)\right)}{\left(\sum_{c \in g} w(c) + \sum_{c \in g'} w(c)\right)} \quad (2)$$

$$S_r(g,g') = \frac{\left(2 * \sum_{r \in N(c,g_u)} w(r)\right)}{\left(\sum_{r \in N(c,g)} w(r) + \sum_{r \in N(c,g')} w(r)\right)} \quad (3)$$

donde $S_c(g,g')$ es la similitud entre los conceptos de los grafos, $S_r(g,g')$ es la similitud entre las relaciones, g_u es el grafo resultante de la intersección de g y g', c es un concepto, r es una relación, $w(c)$ y $w(r)$ son sus pesos, y $N(c,g_i)$ es el conjunto de relaciones conectadas al concepto c en el grafo g_i. Las variables a y b se utilizan en la ecuación 1 para dar la apropiada relevancia a los conceptos y relaciones, ya que sus pesos no se calculan del mismo modo y, por tanto, valores de similitud iguales no tienen porqué tener el mismo significado. Además, para la resolución de determinados problemas, no son igual de relevantes conceptos que relaciones, por este motivo se suele utilizar la regla $a + b = 1$, y se toman a y b como porcentajes de relevancia. En la sección 4 analizaremos cuales son los porcentajes de relevancia adecuados para conceptos y relaciones en detección de plagio translingüe utilizando BabelNet.

Es importante señalar que después de la intersección $g_u = (g \cap g')$, los pesos del grafo g_u tendrán que ser recalculados. El cálculo del peso de un concepto es trivial, pues es el número de relaciones salientes. Recalcular el peso de las relaciones requiere de coste cúbico siguiendo su proceso de creación en BabelNet, ya que para cada relación sería necesario recorrer todos los conceptos dos veces[9]. Por ese motivo, en la ecuación 4 proponemos un algoritmo genérico de reestimación del peso $w(r,c,g_u)$, siendo r una relación saliente de un concepto c en el grafo de intersección g_u. El nuevo peso se calcula en función del antiguo y del nuevo valor del peso de c en los grafos g, g' y g_u,

$$w(r,c,g_u) = \frac{w(c,g) * d(c,g,g_u) + w(c,g') * d(c,g',g_u)}{2} \quad (4)$$

[9]Donde $t(n,m) \in O(n^2 * m)$, siendo n el número de conceptos y m el número de relaciones entre ellos.

$$d(x,g_1,g_2) = \frac{|R(g_1,x)|}{|R(g_2,x)|} \quad (5)$$

donde $w(c,g_i)$ es el peso del concepto c en el grafo g_i, y $R(g_i,x)$ es el conjunto de relaciones salientes del concepto x en el grafo g_i.

3.2 Análisis de silimitud basado en n-gramas de caracteres

El modelo CL-CNG, *cross-language character n-gram*, ha demostrado ofrecer un rendimiento elevado para lenguajes europeos con similitudes sintácticas y hace uso de n-gramas a nivel de caracteres para comparar los documentos en diferentes idiomas. En este modelo se utilizan normalmente trigramas de caracteres (CL-C3G) (Potthast et al., 2011a).

Dado un documento fuente d en un lenguaje L_1 y un documento sospechoso d' en un lenguaje L_2, la similitud $S(d,d')$ entre los dos documentos se mide como se muestra en la ecuación 6:

$$S(d,d') = \frac{\vec{d}.\vec{d'}}{|d|.|d'|} \quad (6)$$

donde $\vec{d}$ y $\vec{d'}$ son las proyecciones vectoriales de d y d' en un espacio de n-gramas de carácter.

3.3 Análisis de similitud basado en alineamiento

El modelo CL-ASA mide la similitud entre dos documentos d y d', en dos idiomas diferentes L_1 y L_2, alineandolos a nivel de palabra, determinando la probabilidad de que un documento d' sea una una traducción del documento d. La similitud $S(d,d')$ se calcula haciendo uso de la ecuación 7:

$$S(d,d') = l(d,d') * t(d|d') \quad (7)$$

donde $l(d,d')$ es el factor de longitud definido en (Pouliquen, Steinberger, y Ignat, 2003) y $t(d|d')$ es el modelo de traducción definido en la ecuación 8:

$$t(d|d') = \sum_{x \in d} \sum_{y \in d'} p(x,y) \quad (8)$$

donde $p(x,y)$ es la probabilidad de que una palabra x en el lenguaje L_1 sea una traducción de la palabra y del lenguaje L_2. Dichas probabilidades de traducción pueden obtenerse mediante un diccionario estadístico. Para nuestros experimentos se ha entrenado

un diccionario estadístico alemán-inglés y español-inglés haciendo uso del modelo de alineamiento de palabras IBM M1 (Brown et al., 1993; Och y Ney, 2003), sobre el corpus paralelo multilingüe JRC-Acquis (Steinberger et al., 2006), además de probar también el diccionario estadístico de la RSM BabelNet.

4 Experimentos y evaluación

En esta sección vamos a evaluar el rendimiento de nuestro modelo CL-KGA, para la tarea de detección de plagio translingüe es-en y de-en, utilizando la RSM BabelNet como base de conocimiento, frente a los modelos estado del arte CL-ASA y CL-C3G. Para el modelo CL-ASA realizaremos las pruebas con dos diccionarios estadísticos diferentes: un diccionario entrenado con el modelo de alineamiento IBM M1, y el diccionario estadístico de BabelNet (BN_{dict}), que ya ha demostrado anteriormente ofrecer un buen rendimiento para la tarea de detección de plagio translingüe (Franco-Salvador, Gupta, y Rosso, 2012). Además, previamente a la comparación de los modelos, vamos a realizar unos experimentos para determinar cual es la relación de porcentajes adecuados para los valores de relevancia de conceptos y relaciones en el CL-KGA.

4.1 Corpus y definición de la tarea

Del corpus PAN-PC'11, tomamos las particiones es-en y de-en para su tarea de detección de plagio externo: dado un conjunto de documentos fuente D en el lenguaje L_1 y un conjunto de documentos sospechosos D' en el lenguaje L_2, la tarea es determinar los fragmentos concretos de los documentos fuente que están presentes en los sospechosos. Para ello utilizamos una ventana deslizante de cinco oraciones de longitud sobre pares de documentos (d,d'), $d \in D$ y $d' \in D'$, y detectamos plagio translingüe sobre ellos con los modelos comentados anteriormente. En la Tabla 1 podemos ver las estadísticas de los documentos utilizados para la evaluación.

Documentos es-en		Documentos de-en	
Sospechosos	304	Sospechosos	251
Fuentes	202	Fuentes	348
Casos de plagio {es,de}-en			
Traducción automática	5.142		
Traducción automática + corrección manual	433		

Cuadro 1: Estadísticas de la tarea de detección de plagio externo del corpus PAN-PC'11

4.2 Unidades de medida

Para medir la calidad de los resultados vamos a tomar las medidas utilizadas en la competición del PAN: *recall* (rec.) y *precision* (prec.) a nivel de carácter, además de *granularity* (gran.), la cual tiene en cuenta el hecho de que en ocasiones los detectores solapan o deportan multiples detecciones para un mismo caso de plagio. Las tres medidas son combinadas con el objetivo de obtener una medida global de la detección de plagio, el *plagdet*:

$$plagdet(S,R) = \frac{F_1}{log_2(1 + granularity(S,R))}$$

Donde S es el conjunto de casos de plagio del corpus, R es el conjunto de casos de plagio reportados por el detector, y F_1 es la media armónica de *precision* y *recall* ponderadas equitativamente. [10]

4.3 Experimento 1

En primer lugar vamos a comparar el rendimiento del modelo CL-KGA utilizando la RSM BabelNet, según sus valores de relevancia para conceptos y relaciones. Para ello hemos diseñado un experimento, midiendo solamente el *plagdet*, utilizando una porción aleatoria del 20 % del corpus PAN-PC'11, tanto para es-en como de-en, en el que probaremos los siguientes porcentajes de relevancia para conceptos (c) y relaciones (r): $(c,r) \in \{(100,0),(80,20),(50,50),(20,80),(0,100)\}$.

% (c,r)	Plagdet(es-en)	Plagdet(de-en)
(100,0)	0.617	**0.636**
(80,20)	0.616	0.6247
(50,50)	**0.655**	0.620
(20,80)	0.642	0.581
(0,100)	0.612	0.522

Cuadro 2: Relevancia de conceptos y relaciones en el modelo CL-KGA.

En vista de los resultados de la tabla 2, podemos deducir que las relaciones utilizando la RSM BabelNet son prácticamente igual de importantes que los conceptos para es-en, mientras que para de-en tienen poca o ninguna importancia. La diferencia puede estar producida por unos conceptos muy conectados en grafos es-en, mientras que en de-en podemos estar ante un elevado número de conceptos parte de un grafo menos conexo. Para nuestros siguientes experimentos tomaremos las mejores configuraciones de ambas particiones.

[10] Una descripcción más detallada de las medidas se puede encontrar en (Potthast et al., 2010).

4.4 Experimento 2

En este experimento vamos a comparar CL-KGA, con los modelos descritos anteriormente, para las particiones completas es-en y de-en del corpus del PAN-PC'11.

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-KGA	**0.594**	**0.518**	**0.706**	**1.008**
CL-ASA$_{BNdict}$	0.567	0.499	0.662	1.015
CL-ASA$_{IBMM1}$	0.517	0.448	0.689	1.071
CL-C3G	0.170	0.128	0.617	1.372

Cuadro 3: Resultados de la detección de plagio translingüe es-en

En la tabla 3 podemos observar como CL-KGA ha superado en todos los valores al resto de modelos para la detección de plagio translingüe es-en. El modelo CL-ASA que más se le ha aproximado -utilizando el diccionario del propio BabelNet- tiene un *plagdet* un 4.7 % inferior. Además, aparte de observar el aumento de los valores de *precision* y *recall*, es importante señalar que se ha alcanzado un valor de *granularity* muy próximo a 1, lo cual es el mejor valor posible, e indica que no existen solapamientos en la detección interpretando una sección de plagio como varias, o viceversa.

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-KGA	**0.514**	**0.443**	**0.631**	**1.018**
CL-ASA$_{IBMM1}$	0.406	0.344	0.604	1.113
CL-ASA$_{BNdict}$	0.289	0.222	0.595	1.171
CL-C3G	0.078	0.047	0.330	1.089

Cuadro 4: Resultados de la detección de plagio translingüe de-en

En la tabla 4 vemos también unos buenos resultados para de-en en nuestro modelo. CL-KGA ha superado al CL-ASA$_{IBMM1}$, el más cercano, en un valor de *plagdet* del 26.6 %, lo cual supone una excelente mejora respecto al estado del arte actual. Los otros valores también han mejorado, destacando un incremento del *recall* de un 28 %, lo cual indica un considerable aumento en el número de detecciones positivas. En esta ocasión el diccionario de BabelNet no se ha comportado tan bien como para es-en[11].

En vista de los resultados anteriores, podemos afirmar cómo hacer uso de grafos de conocimiento es una buena alternativa para la detección de plagio translingüe.

[11]Lo cual viene justificado en (Franco-Salvador, Gupta, y Rosso, 2012) como consecuencia del procesamiento previo de las palabras en alemán al construir BabelNet

5 Conclusiones y trabajos futuros

En este trabajo hemos presentado un nuevo modelo para el análisis de similitud a nivel translingüe, el CL-KGA, que hace uso de una RSM para construir grafos de conocimiento a modo de modelos de contexto de documentos. El modelo propuesto ha demostrado ofrecer un rendimiento superior a otros modelos estado del arte como CL-ASA y CL-CNG, evaluados sobre la partición translingüe del corpus PAN-PC'11.

En futuras investigaciones se seguirá investigando en el campo de la detección de plagio translingüe para extender nuestro modelo con otras RSM que nos proporcionen una mayor variedad de lenguajes compatibles, además de investigar el potencial de nuestro nuevo modelo para análisis de similitud a nivel monolingüe.

Bibliografía

Barrón-Cedeño, Alberto. 2012. *On the mono- and cross-language detection of text re-use and plagiarism.* Ph.D. thesis, Universitat Politènica de València.

Barrón-Cedeño, Alberto, Paolo Rosso, David Pinto, y Alfons Juan. 2008. On cross-lingual plagiarism analysis using a statistical model. En *Proc. of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, PAN'08.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, y R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Dumais, S. T., T. A. Letsche, M. L. Littman, y T. K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. En *Proc. AAAI-97 spring symposium series: Cross-language text and speech retrieval*, páginas 18–24. Hull & D. Oard (Eds.).

Franco-Salvador, Marc, Parth Gupta, y Paolo Rosso. 2012. Cross-language plagiarism detection using BabelNet's statistical dictionary. *Computación y Sistemas, Revista Iberoamericana de Computación*, 16(4):383–390.

Gupta, Parth, Alberto Barrón-Cedeño, y Paolo Rosso. 2012. Cross-language high

similarity search using a conceptual thesaurus. En *Proc. 3rd Int. Conf. of CLEF Initiative on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics*. CLEF 2012.

Mcnamee, Paul y James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1):73–97.

Montes y Gómez, Manuel, Alexander F. Gelbukh, Aurelio López-López, y Ricardo A. Baeza-Yates. 2001. Flexible comparison of conceptual graphs. En *Proc. DEXA*, páginas 102–111.

Navigli, Roberto y Simone Paolo Ponzetto. 2010. Babelnet: building a very large multilingual semantic network. En *Proc. of the 48th annual meeting of the association for computational linguistics*, ACL '10, páginas 216–225, Stroudsburg, PA, USA.

Navigli, Roberto y Simone Paolo Ponzetto. 2012. Multilingual wsd with just a few lines of code: The babelnet api. En *Proc. 50th annual meeting of the association for Computational Linguistics*.

Och, F. J. y H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pinto, D., J. Civera, A. Barrón-Cedeño, A. Juan, y P. Rosso. 2009. A statistical approach to crosslingual natural language tasks. *Journal of algorithms*, 64(1):51–60.

Potthast, M., A. Barrón-Cedeño, B. Stein, y P. Rosso. 2010. An evaluation framework for plagiarism detection. En *Proc. of the 23rd Int. Conf. on Computational Linguistics*, COLING-2010, páginas 997–1005, Beijing, China.

Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein, y Paolo Rosso. 2011a. Cross-language plagiarism detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 45(1):45–62.

Potthast, Martin, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, y Paolo Rosso. 2011b. Overview of the 3rd int. competition on plagiarism detection. En *CLEF (Notebook Papers/Labs/Workshop)*.

Pouliquen, B., R. Steinberger, y C. Ignat. 2003. Automatic linking of similar texts across languages. En *Proc. Recent Advances in Natural Language Processing III*, páginas 307–316. RANLP'03.

Sowa, J. F. 1984. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman.

Sowa, J. F. 1999. *Knowledge representation: logical, philosophical and computational foundations*. Brooks/Cole Publishing Co.

Stein, B. y M. Anderka. 2009. Collection-relative representations: A unifying view to retrieval models. En *Proc. 20th Int. Conf. on database and expert systems applications*, DEXA'09, páginas 383–387. A. M. Tjoa & R. R. Wagner (Eds.).

Steinberger, R., B. Pouliquen, y C. Ignat. 2004. Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. En *Proc. 4th Slovenian language technology conference*, IS'2004. Information Society.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, y D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with +20 languages. En *Proc. 5th Int. Conf. on language resources and evaluation*. LREC'2006.

Vinokourov, A., J. Shawe-Taylor, y N. Cristianini. 2003. Inferring a semantic representation of text via cross-language correlation analysis. En *Proc. NIPS-02: Advances in neural information processing systems*, páginas 1473–1480. S. Becker, S. Thrun, & K. Obermayer (Eds.).