Detección de plagio translingüe utilizando el diccionario estadístico de BabelNet

Marc Franco-Salvador, Parth Gupta y Paolo Rosso

Natural Language Engineering Lab - ELiRF,
Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València,
España

{mfranco, pgupta, prosso}@dsic.upv.es

Resumen. En los últimos años ha habido importantes avances en el campo de la detección de plagio automática. Uno de ellos es la detección de plagio translingüe, la cual trata de detectar el plagio entre documentos en diferentes idiomas. La mayoría de aproximaciones que existen para esta tarea hacen uso de diccionarios estadísticos para lidiar con las traducciones de las palabras de los documentos. Un diccionario estadístico nos proporciona, para una palabra dada, la lista de traducciones posibles con sus respectivas probabilidades. El objetivo de este trabajo es analizar el rendimiento del diccionario estadistico de la red semántica multilingüe BabelNet para la tarea de detección de plagio translingüe. En la evaluación comparamos sus resultados con los ofrecidos por un diccionario estadístico entrenado con el conocido modelo de alineamiento IBM M1, ambos utilizando el modelo estado del arte CL-ASA como base. Los resultados de los experimentos indican que BabelNet es una buena alternativa como diccionario estadístico.

Palabras clave. Detección de plagio translingüe, similitud textual, diccionario estadístico, BabelNet.

Cross-language Plagiarism Detection Using BabelNet's Statistical Dictionary

Abstract. In recent years there have been important advances in the field of automatic plagiarism detection. One variant is cross-language plagiarism detection, which tries to detect plagiarism between documents in different languages. Most of the existing approaches to this task make use of statistical dictionaries to deal with the translations of words in the documents. A statistical dictionary provides, for a given word, the list of possible translations with their respective probabilities. The objective of this paper is to analyze the performance of the statistical dictionary of multilingual semantic network - Babelnet for cross-language plagiarism detection. In the evaluation we compare its results with those offered by a statistical dictionary trained by the well-known IBM M1 aligment model, both using state-of-the-art model CL-ASA as a base. The results of the experiments indicate that Babelnet is a good alternative as statistical dictionary.

Keywords. Cross-language plagiarism detection, textual similarity, statistical dictionary, BabelNet.

1 Introducción

El plagio es definido como el uso no autorizado del contenido original de la obra de otros autores. Es un fenómeno difícil de detectar cuyo problema se ha agravado en los últimos años a causa de Internet: una inmensa fuente de información que permite a los usuarios copiar y adueñarse, de forma muy sencilla, del contenido original de otros autores. Un estudio hecho en varias universidades españolas mostró que un 61 % de los estudiantes reconoció haber incluido fragmentos de documentos de Internet en sus trabajos al menos una vez [4]. La detección de plagio es aún más complicada si la fuente del plagio viene de documentos en otros idiomas. Recientemente se realizó una encuesta sobre las prácticas escolares y las actitudes de los estudiantes [1], también desde una perspectiva translingüe, que pone de manifiesto que el plagio translingüe es un problema real, y justifica la necesidad de investigación en este campo: un 63.75 % de los estudiantes piensa que copiar y traducir fragmentos de texto desde otros documentos y incluirlos en sus trabajos no es plagio.

En los últimos años han aparecido una serie de aproximaciones para llevar a cabo detección de plagio translingüe. Estas aproximaciones traducen los documentos sospechosos de contener plagio al lenguaje de los fuente, realizando la detección a nivel monolingüe. *Cross-language character*

n-gram (CL-CNG) [7] es un modelo de detección de plagio 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) [12] 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) [2, 11] se basa en la tecnología de máquinas de traducción estadística, la cual combina traducciones estadísticas, usando diccionarios estadísticos, y análisis de similitud. En el pasado los tres modelos anteriores fueron comparados [12], siendo CL-ASA el que produjo los mejores resultados, de ahí la motivación de comparar los resultados del diccionario estadístico de Babelnet con CL-ASA como base.

En este trabajo, haciendo uso de las herramientas de la red semántica multilingüe BabelNet [8], aplicamos diferentes técnicas sobre su diccionario estadístico para, sobre el modelo CL-ASA, traducir los documentos sospechosos al lenguaje de los fuente y detectar plagio a nivel translingüe.

Para la evaluación de los modelos utilizamos un corpus diseñado específicamente para la tarea de detección de plagio translingüe: el corpus del PAN-PC'11 [14][1]. Dentro del ámbito de la detección de plagio, desde el año 2009 se celebra anualmente una competición internacional, *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. Dentro de la competición a nivel tenemos dos tareas: detección de plagio intrínseca y externa. La detección de plagio intrínseca se trata de, dado un documento, analizar su estructura para determinar las características del autor y detectar las secciones que no parecen propias de éste. El plagio externo en cambio trata de, dado un conjunto de documentos fuente y un conjunto de documentos sospechosos, determinar las secciones concretas de los documentos fuente que están presentes en los sospechosos. Para nuestra evaluación utilizamos la partición de detección de plagio translingüe de la tarea de detección de plagio externo.

El resto del artículo está estructurado como sigue. En la sección 2 describimos las aproximaciones del estado del arte en detección de plagio translingüe. La red semántica multilingüe BabelNet es descrita en la sección 3. En la sección 4 presentamos las diferentes aproximaciones a la detección de plagio translingüe haciendo uso del diccionario estadístico de BabelNet. En la sección 5 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. Comparamos nuestros resultados con los obtenidos por un diccionario entrenado con el modelo de alineamiento IBM M1 [3, 10], ambos sobre el modelo CL-ASA, además de comparar el modelo CL-CNG para el mismo experimento.

2 Detección de plagio translingüe

En esta sección vamos a resumir algunas de las aproximaciones del estado del arte en análisis de similitud para detección de plagio translingüe. Dichas aproximaciones nos devuelven la similitud entre secciones de un documento sospechoso d y secciones de un documento fuente d', perteneciendo d' al conjunto de documentos fuente D'.

2.1 Clases de modelos de análisis de similitud

Existen una serie de modelos de análisis de similitud que pueden ser usados en un contexto translingüe. Dichos modelos pueden ser divididos en cuatro categorías de acuerdo a la clase de propuesta: (i) los modelos basados en diccionarios, índices geográficos, reglas y thesaurus lingüísticos; (ii) los modelos basados en la sintaxis del documento; (iii) los modelos basados en corpus comparables; (iv) y los modelos basados en corpus paralelos. Los modelos de la primera clase traducen palabras aisladas y conceptos, tales como fechas, desde un lenguaje L a uno L' (e.g. CL-VSM [17] y Eurovoc [15]). La segunda clase explota las similitudes sintácticas de los lenguajes y la semejanza de las palabras para determinar la similitud (e.g. CL-CNG). Los modelos de la tercera y cuarta clase son entrenados con corpus alineados de los diferentes lenguajes.

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

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

La tercera clase necesita el corpus alineado por documentos que describan aproximadamente el mismo tema (e.g. CL-ESA), mientras que la cuarta necesita exactamente el mismo corpus, en diferentes idiomas, alineados de acuerdo a las traducciones de las palabras, ya sean manuales o con métodos estadísticos (e.g. CL-ASA, CL-LSI [5] y CL-KCCA [20]).

2.2 Modelos de análisis de similitud

Por orden de acuerdo a su clase, describimos alguno de los modelos mas destacados:

— El modelo CL-CNG [7] ofrece un rendimiento elevado para lenguajes con similitudes sintácticas y hace uso de n-gramas a nivel de caracteres para comparar los documentos en diferentes idiomas. Usualmente en esta aproximación se utilizan trigramas de caracteres (CL-C3G).

— El modelo CL-ESA [12] representa un documento sospechoso d por sus semejanzas con una colección de documentos fuente D'. El análisis de similitud entre documentos se lleva a cabo con un modelo de detección de similitud monolingüe, como los modelos de espacio vectorial [16]. En un contexto translingüe, para determinar las semejanzas entre documentos es necesario un corpus comparable multilingüe alineado por tema y idioma, como por ejemplo la enciclopedia de la Wikipedia.

— El modelo CL-ASA [2, 1] utiliza corpus en diferentes idiomas alineados a nivel de palabra para, combinando análisis de similitud y traducción estadística, estimar la similitud de documentos dados. La similitud entre dos documentos $S(d, d')$ se puede calcular mediante un diccionario estadístico para extraer una probabilidad de traducción $p(x, y)$ entre dos palabras x e y, $x \in d$ y $y \in d'$, haciendo uso de la siguiente fórmula:

$$S(d|d') = \sum_{x \in d} \sum_{y \in d'} p(x, y) \quad (1)$$

Como se introdujo en la sección 1, los modelos CL-C3G, CL-ASA y CL-ESA han sido comparados en [12] dejando de lado las otras aproximaciones descritas que ofrecen un buen rendimiento, como CL-LSI o CL-KCCA, por tener un coste computacional de orden cúbico, lo cual es demasiado elevado para un sistema de detección de plagio en un escenario realista como la Web. Los resultados de la comparación de los modelos muestran que CL-ASA ofrece un elevado rendimiento a un bajo coste computacional, lo cual nos ha motivado a elegir CL-ASA como base para llevar a cabo nuestros experimentos sobre el diccionario estadístico de BabelNet.

Cabe mencionar algunos diccionarios estadísticos que pueden ser utilizados con algunos de los modelos anteriores: el diccionario bilingüe CoReMo [19], la base de datos en inglés WordNet [6], la base de datos multilingüe EuroWordNet [21], y la red semántica multilingüe BabelNet, la cual es descrita en la sección 3. Por otro lado, para nuestros experimentos se ha entrenado un diccionario estadístico alemán-inglés y español-inglés haciendo uso del modelo de alineamiento de palabras IBM M1 [3, 10], sobre el corpus paralelo multilingüe JRC-Acquis [18].

3 BabelNet

BabelNet [8] es una red semántica multilingüe de gran tamaño, la cual sigue la estructura tradicional de una base de conocimiento, y en consecuencia está formada por un grafo dirigido y ponderado, en el cual los nodos representan conceptos y nombres de entidades, y las aristas representan las relaciones entre ellos. Además, en BabelNet cada uno de los nodos tiene un conjunto de lexicalizaciones de los conceptos en diferentes lenguas, dándonos una dimensión multilingüe de la base de conocimiento en los siguientes idiomas: alemán, catalán, español, francés, inglés e italiano.

Los conceptos y relaciones de BabelNet son tomados de la mayor red semántica disponible, WordNet, y de las entradas multilingüe etiquetadas de la enciclopedia colaborativa Wikipedia[3], lo cual convierte a BabelNet en un "diccinario enciclopédico" multilingüe que combina información lexicográfica con conocimiento enciclopédico. El listado de conceptos de BabelNet está formado por todos los significados de palabra de WordNet y de las entradas de la Wikipedia, mientras que las relaciones entre ellos comprenden los punteros semánticos entre

[3] http://www.wikipedia.org/

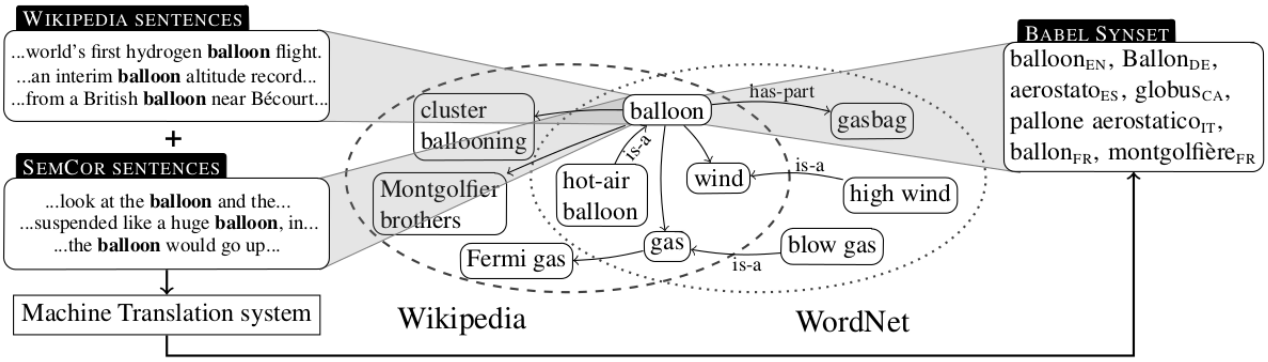

Figura 1. Ejemplo de la estructura interna de BabelNet (Figura tomada de [8]).

conceptos de los synsets[4] de Wordnet y los enlaces entre entradas de la Wikipedia. Las lexicalizaciones multilingüe de los conceptos se toman de los enlaces de entradas entre los diferentes idiomas de la Wikipedia. Podemos ver un ejemplo de como está confeccionado BabelNet en la Fig. 1.

Las herramientas de BabelNet nos permiten utilizarlo como diccionario, traductor, diccionario estadístico, para desambiguación del sentido de las palabras [9] y construir grafos de conocimiento: dado un conjunto de conceptos, como las palabras de una frase, podemos utilizar la red semántica multilingüe BabelNet para encontrar los caminos que conectan y relacionan dichos conceptos y crear un grafo que contenga, no solo los conceptos y relaciones originales, sino también los que los relacionan, creando un "modelo del contexto" de la frase original.

3.1 El diccionario estadístico de BabelNet

Para el presente trabajo hemos tomado de BabelNet su diccionario estadístico. Dada una palabra x escrita en un lenguaje $A \in L = \{CA, DE, EN, ES, FR, IT\}$, el diccionario estadístico nos permite obtener el conjunto $\{(x_1, w_1), (x_2, w_2), ...\}_B$, $B \in L$, de traducciones posibles de la palabra x en el lenguaje B, siendo w_i el peso de la traducción de la palabra x_i. Los pesos de las traducciones son computados de forma que las más probables tienen un peso mayor, y su valor está en función del número de conceptos que se relacionan con dicha palabra dentro de la red semántica multilingüe de BabelNet.

4 Detección de plagio translingüe con el diccionario estadístico de BabelNet

Como hemos comentado en la Sección 3, el diccionario estadístico de BabelNet (BN-dict) nos proporciona una lista de traducciones ponderadas para una palabra dada, y aunque dichos pesos están relacionados con la traducción más probable, no son probabilidades ni están normalizados. Teniendo en cuenta lo anterior, vamos a proponer diferentes formas de estimar la similitud $S(d|d')$ entre dos documentos d y d', partiendo de la Ecuación 1 del modelo CL-ASA, que utilizan diferentes métodos de normalización. En nuestro caso, podemos atender a dos métodos de normalización distintos:

— Normalización del peso de traducción $w(x, y)$: Dada una palabra x en un lenguaje A y una palabra y en un idioma B, basta con dividir el peso de traducción $w(x, y)$ por la suma de todos los pesos de traducción posibles $w(x)$ de la palabra x al idioma B, para dar lugar a una probabilidad de traducción $p(x, y)$:

$$p(x, y) = \frac{w(x, y)}{w(x)} \quad (2)$$

— Normalización de la similitud $S(d|d')$: Dado un documento d con un número total de de palabras $|d|$, podemos normalizar su

[4] Synset: nodo que comprende todos los posibles significados de una palabra y sus lexicalizaciones multilingües (en el caso de BabelNet).

similitud respecto al número de palabras del documento fuente del siguiente modo:

$$S_{norm}(d|d') = \frac{S(d|d')}{|d|} \quad (3)$$

Los métodos de normalización descritos en las Ecuaciones 2 y 3 son compatibles entre sí y combinados con la Ecuación 1 se pueden utilizar para dar lugar a cuatro ecuaciones de similitud diferentes:

— BN-dict$_1$: Normalización en función de los pesos de traducciones y del número de palabras del documento fuente:

$$S_{full\_normalization}(d|d') = \frac{\sum_{x\in d}\sum_{y\in d'} p(x,y)}{|d|} \quad (4)$$

— BN-dict$_2$: Normalización en función de los pesos de traducciones:

$$S_{weight\_normalization}(d|d') = \sum_{x\in d}\sum_{y\in d'} p(x,y) \quad (5)$$

— BN-dict$_3$: Normalización en función del número de palabras del documento fuente:

$$S_{size\_normalization}(d|d') = \frac{\sum_{x\in d}\sum_{y\in d'} w(x,y)}{|d|} \quad (6)$$

— BN-dict$_4$: Sin normalización:

$$S_{no\_normalization}(d|d') = \sum_{x\in d}\sum_{y\in d'} w(x,y) \quad (7)$$

5 Experimentos y evaluación

En nuestra evaluación vamos a comparar las cuatro aproximaciones descritas en la sección 4 que hacen uso del diccionario estadístico de BabelNet con un diccionario entrenado con el modelo de alineamiento IBM M1, ambos sobre un modelo CL-ASA como base, para la tarea de detección de plagio translingüe ES-EN y DE-EN. Además, mostraremos los resultados del modelo CL-C3G para la misma tarea.

5.1 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 y un conjunto de documentos sospechosos D', 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 líneas de longitud sobre pares de documentos (d,d'), $d \in D$ y $d' \in D'$, y detectamos plagio translingüe sobre ellos con los sistemas comentados anteriormente. En la Tabla 1 podemos ver la estadísticas de los documentos utilizados para la evaluación.

Tabla 1. Estadísticas de la tarea de detección de plagio externo del corpus PAN-PC'11

Docs. partición ES-EN		Docs. partición 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

5.2 Experimentos y medidas

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 $plagdet$:

$$plagdet(S,R) = F_1/log2(1 + granularity(S,R))$$

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. Una descripcción más detallada del corpus y las medidas se puede encontrar en el resumen del PAN-PC'11 [13].

En la Tabla 2 podemos observar los resultados de detección de plagio ES-EN. Vemos como el modelo CL-C3G ofrece los resultados más bajos, siendo el $baseline$ para este tipo de experimentos. Por otro lado, las dos aproximaciones con el

Tabla 2. Resultados de la detección de plagio translingüe ES-EN

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-ASA (BN-dict$_1$)	0.254	0.198	0.458	1.132
CL-ASA (BN-dict$_2$)	0.264	0.205	0.518	1.160
CL-ASA (BN-dict$_3$)	0.554	0.491	0.663	1.030
CL-ASA (BN-dict$_4$)	**0.563**	**0.499**	**0.662**	**1.015**
CL-ASA (IBM M1)	0.517	0.448	0.689	1.071
CL-C3G	0.170	0.128	0.617	1.372

diccionario de BabelNet normalizando los pesos de las traducciones, BN-dict$_1$ y BN-dict$_2$, pese a haber superado al CL-C3G, no muestran un buen desempeño comparados con los mejores, que han superado un $plagdet$ de 0.5. Así para futuros trabajos queda descartada la normalización de los pesos de las traducciones, ya que acotarlos a un rango [0,1] suaviza demasiado el valor de la similitud $S(d|d')$ para casos positivos de plagio. Finalmente, podemos ver como las dos aproximaciones restantes, BN-dict$_3$ y BN-dict$_4$, normalizando la similitud en función del número de palabras del documento y sin normalización, han superado los resultados obtenidos con el diccionario entrenado con el modelo IBM M1. En concreto, BN-dict$_4$ (el mejor), ha obtenido un 10.31 % más de $plagdet$, lo cual asociado a los otros valores de $recall$, $precision$ y $granularity$ (cuanto más cercana a 1 mejor) que podemos observar, indica que se producen más detecciones correctas y menos falsos positivos. En vista de lo anterior, podemos afirmar que el diccionario estadístico de BabelNet, es una buena alternativa para la detección de plagio Español-Inglés.

Tabla 3. Resultados de la detección de plagio translingüe DE-EN

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-ASA (BN-dict$_1$)	0.104	0.075	0.246	1.152
CL-ASA (BN-dict$_2$)	0.103	0.074	0.246	1.151
CL-ASA (BN-dict$_3$)	0.289	0.222	0.595	1.172
CL-ASA (BN-dict$_4$)	0.219	0.164	0.460	1.152
CL-ASA (IBM M1)	**0.406**	**0.344**	**0.604**	**1.113**
CL-C3G	0.078	0.047	0.330	1.089

En la Tabla 3 tenemos los resultados de detección de plagio DE-EN. Una vez más, tenemos CL-C3G como $baseline$ ofreciendo los resultados más bajos, y las dos aproximaciones con normalización de pesos de traducciones, BN-dict$_1$ y BN-dict$_2$, a continuación, en esta ocasión muy cerca de CL-C3G. Las dos aproximaciones restantes no han conseguido para DE-EN alcanzar al diccionario del modelo IBM-M1. La mejor de las dos, BN-dict$_3$, ofrece un valor de $plagdet$ un 29 % inferior a éste, como consecuencia de unos valores de $recall$, $precision$ y $granularity$ peores. Cabe señalar que observando los valores de $recall$ y $precision$ de BN-dict$_3$, se puede deducir que está habiendo muchos falsos positivos en la detección, lo cual nos lleva a pensar en la posibilidad de que no se estén procesando fragmentos de documentos lo sufientemente representativos de su contenido como para ser comparables, o dicho de otra manera, que se estén perdiendo muchas palabras porque el diccionario de BabelNet no las encuentra. Para confirmar dicha hipótesis hemos realizado un nuevo experimento que mida el porcentaje de uso del diccionario utilizado en cada prueba. Así, en la Tabla 4, vemos como se confirma nuestra teoría. Mientras que para detección ES-EN el porcentaje de uso de ambos diccionarios es similar, en torno al 70 %, siendo BabelNet el que más encuentra, para DE-EN la utilización de BabelNet no llega ni a un 50 %, así que estamos perdiendo la mitad de palabras, a diferencia del diccionario entrenado con el modelo IBM M1, que encuentra un 70 %. Podemos deducir que el problema es con el lenguaje alemán, el cual requerirá un procesamiento especial del texto, extrayendo los lemas de las palabras en su forma infinitiva, para poder aumentar el número de coincidencias dentro de BabelNet, porque así fue tratado el lenguaje durante su desarrollo.

Tabla 4. Estadísticas del uso de los diccionarios

Diccionario ES-EN	% palabras encontradas
BabelNet	71.10 %
IBM M1	68.35 %
Diccionario DE-EN	% palabras encontradas
BabelNet	49.34 %
IBM M1	69.45 %

6 Conclusiones y trabajos futuros

En este trabajo hemos mostrado la efectividad y potencial del diccionario estadístico de la red semántica multilingüe BabelNet para la detección de plagio translingüe ES-EN, superando incluso los resultados de un diccionario entrenado con el modelo de alineamiento IBM M1. Por otro lado, aunque no hemos alcanzado tan buenos resultados en la detección DE-EN, hemos

analizado los resultados argumentando que la disminución del porcentaje de detección es debida al procesamiendo de las palabras previo al que fueron sometidas las palabras en BabelNet para el lenguaje Alemán. Un futuro estudio tratará de la efectividad de los diccionarios estadísticos previo procesamiento del texto por parte de un analizador morfosintáctico.

En futuros trabajos seguiremos investigando dentro del campo de la detección de plagio translingüe haciendo uso de BabelNet. En concreto, además de ampliar nuestro sistema de detección de plagio con los idiomas disponibles en BabelNet, futuras investigaciones girarán en torno a la detección de similitud, a nivel monolingüe y translingüe, haciendo uso de los grafos de conocimiento descritos en la sección 3, los cuales permiten generar un modelo del contexto de las frases.

Agradecimientos

En primer lugar agradecer a la Consellería 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). Por otro lado agradecer a Roberto Navigili por haber desarrollado BabelNet y ofrecer su ayuda para familiarizarnos con el API sistema. Finalmente a Alberto Barrón Cedeño por desarrollar la versión inicial del modelo CL-ASA.

Referencias

1. **Barrón-Cedeño, A.** (**2012**). *On the mono- and cross-language detection of text re-use and plagiarism*. Ph.D. thesis, Universitat Politènica de València.
2. **Barrón-Cedeño, A., Rosso, P., Pinto, D., & Juan, A.** (**2008**). On cross-lingual plagiarism analysis using a statistical model. In *proceedings of the ECAI'08 workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, PAN'08.
3. **Brown, P., Della Pietra, S., Della Pietra, V., & Mercer, R.** (**1993**). The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, volume 19(2). 263–311.
4. **Comas, R. & Sureda, J.** (**2008**). Academic cyberplagiarism: tracing the causes to reach solutions. *Digithum*, 10, 1–6.
5. **Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K.** (**1997**). Automatic cross-language retrieval using latent semantic indexing. In *AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval*. Hull & D. Oard (eds.), 18–24.
6. **Fellbaum, C.** (**1998**). Wordnet: An electronic lexical database. MIT Press.
7. **Mcnamee, P. & Mayfield, J.** (**2004**). Character n-gram tokenization for european language text retrieval. *Inf. Retr.*, 7(1-2), 73–97. ISSN 1386-4564.
8. **Navigli, R. & Ponzetto, S. P.** (**2010**). Babelnet: building a very large multilingual semantic network. In *proceedings of the 48th annual meeting of the Association for Computational Linguistics*, ACL '10. Stroudsburg, PA, USA, 216–225.
9. **Navigli, R. & Ponzetto, S. P.** (**2012**). Multilingual wsd with just a few lines of code: The babelnet api. In *50th annual meeting of the Association for Computational Linguistics*.
10. **Och, F. J. & Ney, H.** (**2003**). A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29(1). 19–51.
11. **Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., & Rosso, P.** (**2009**). A statistical approach to crosslingual natural language tasks. *journal of algorithms*, 64(1), 51–60. doi:10.1016/j.jalgor.2009.02.005.
12. **Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P.** (**2011**). Cross-language plagiarism detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 45(1).
13. **Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P.** (**2010**). An evaluation framework for plagiarism detection. In *proc. of the 23rd int. conf. on Computational Linguistics*, COLING-2010. Beijing, China, 997–1005.
14. **Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P.** (**2011**). Overview of the 3rd international competition on plagiarism detection. In *CLEF (Notebook Papers/Labs/Workshop)*.
15. **Pouliquen, B., Steinberger, R., & Ignat, C.** (**2003**). Automatic annotation of multilingual text collections with a conceptual thesaurus. In *workshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology—its Potential and Practicalities'*, EUROLAN'2003. 9–28.

16. **Stein, B. & Anderka, M.** (**2009**). Collection-relative representations: A unifying view to retrieval models. In *20th international conference on Database and Expert Systems Applications*, DEXA'09. A.M. Tjoa & R.R. Wagner (eds.), 383–387.
17. **Steinberger, R., Pouliquen, B., & Ignat, C.** (**2004**). Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In *4th Slovenian Language Technology Conference*, IS'2004. Information Society.
18. **Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D.** (**2006**). The jrc-acquis: A multilingual aligned parallel corpus with +20 languages. In *5th international conference on Language Resources and Evaluation*. LREC'2006.
19. **Torrejón, D. & Ramos, J.** (**2011**). Crosslingual coremo system (contextual reference monotony). In *CLEF (Notebook Papers/Labs/Workshop)*.
20. **Vinokourov, A., Shawe-Taylor, J., & Cristianini, N.** (**2003**). Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS-02: Advances in Neural Information Processing Systems*. S. Becker, S. Thrun, & K. Obermayer (eds.), 1473–1480.
21. **Vossen, P.** (**2004**). Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. In *international journal of Lexicography*, volume 17.

Marc Franco-Salvador is a M.Sc. student of Artificial Intelligence, Pattern Recognition and Digital Image at Universitat Politècnica de València and a member of Natural Language Engineering Lab of the ELiRF research group. His master's thesis is focused on similarity retrieval on cross-language plagiarism. His scientific interests include Pattern Recognition in general and Natural Language Processing applied to plagiarism detection from cross-language perspective.

Parth Gupta is a Ph.D. student at Universitat Politècnica de València and a member of Natural Language Engineering Lab of the ELiRF research group. He received M.Tech degree in Informaion and Communication Technology from DA-IICT India. His master's thesis was focused on Learning to Rank technologies of Information Retrieval. His research interests are Information Retrieval and Statistical Natural Language Processing from a cross-language perspective. He also serves the organising committee of Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) Lab at the Conference and Labs of the Evaluation Forum (CLEF): http://pan.webis.de/ and Cross-Language !ndian News Story Search track at the Forum for Information Retrieval Evaluation (FIRE): http://www.dsic.upv.es/grupos/nle/clinss.html

Paolo Rosso received his Ph.D. degree on Computer Science (1999) from the Trinity College Dublin, University of Ireland. He is currently an Associate Professor at Universitat Politècnica de València, Spain, where he leads the Natural Language Engineering Lab. of the ELiRF research group. He has published approx. 200 papers in conferences, workshops and journals; and he has been involved in several national and international research projects. His research interests are mainly focused on plagiarism detection, irony detection in social media, and short texts analysis. He is one of the organisers of PAN activities on Uncovering Plagiarism, Authorship, and Social Software Misuse at the Conference and Labs of the Evaluation Forum (CLEF): `http://pan.webis.de/` and at the Forum for Information Retrieval Evaluation (FIRE): `http://www.dsic.upv.es/grupos/nle/clinss.html`

Article received on 25/10/2012; accepted on 26/11/2012.