Influencia del diccionario en la traducción para la detección de plagio translingüe

Diego A. Rodríguez-Torrejón[1,2] Alberto Barrón-Cedeño[3]
Grigori Sidorov[4] José Manuel Martín-Ramos[1] Paolo Rosso[3]

[1] Universidad de Huelva [2] IES José Caballero
[3] Natural Language Engineering Lab. - EliRF, Universitat Politècnica de València
[4] Centro de Investigación en Computación, Instituto Politécnico Nacional de México

dartsystems@gmail.com lbarron@dsic.upv.es
sidorov@cic.ipn.mx jmmartin@dti.uhu.es prosso@dsic.upv.es

Resumen. Uno de los tipos de plagio que ha reclamado atención recientemente es el plagio translingüe, es decir, cuando el fragmento reutilizado ha sido traducido antes de insertarlo en un nuevo documento y no se incluye referencia alguna sobre la fuente. Este artículo analiza un modelo de detección de plagio translingüe basado en dos etapas: (a) la normalización de los textos a una lengua común y (b) la comparación de documentos. El método de normalización, basado en un sencillo mapeo de palabras de una lengua a otra, tiene un importante impacto en la detección final . Por ello, este artículo analiza los resultados obtenidos al considerar diccionarios bilingües de distinta naturaleza para detectar los casos translingües de español a inglés en el corpus *PAN-PC-11*. Los resultados obtenidos muestran que, a pesar de no ser exhaustivos, la mejor opción es utilizar diccionarios con una alta carga de entidades con nombre y el menor grado de ambigüedad posible. Dichos diccionarios pueden ser obtenidos con relativa facilidad a partir de recursos Web, tales como *Dicts.info* y *Wiktionary*, lo que permite que la aplicación de este modelo a otros pares de lenguas sea factible.

Palabras clave: detección de plagio translingüe, diccionarios bilingües

1 Introducción

Con la gran cantidad de documentos disponibles en Internet, los casos de plagio han aumentado dramáticamente, lo que hace que el desarrollo de modelos automáticos para su detección sea imperante. En este artículo estamos particularmente interesados en la detección de plagio translingüe (o traducido). Dicho plagio ocurre cuando un escritor copia un texto en una lengua, por ejemplo, español, lo traduce (sea automática o manualmente), y lo inserta en su propio documento sin incluir referencia alguna.

Una de las técnicas de recuperación de información translingüe que más se utiliza actualmente es la de traducir los documentos a una lengua común para su posterior

comparación, a nivel monolingüe. Los métodos de traducción y, en particular, los diccionarios utilizados en esta etapa son un factor clave para la calidad de la detección. Por ello, en este artículo se analiza la influencia que el diccionario tiene en el proceso de detección. Para ello, se experimenta sobre un mismo sistema de detección de plagio translingüe con dos diccionarios bilingües: uno enriquecido automáticamente a partir de un diccionario bilingüe "tradicional", y otro resultado de unir traducciones de *Wiktionary* y de *enlaces inter-lengua de Wikipedia*.

En la sección 2, se trata el estado del arte de la detección de plagio translingüe. La sección 3 describe el modelo de detección propuesto. La sección 4 plantea el marco experimental. En la sección 5 se discuten los resultados obtenidos. Finalmente la sección 6 incluye las conclusiones y propuestas futuras de investigación.

2 Trabajos relacionados

La investigación sobre detección automática de plagio ha comenzado desde hace más de una década [1-2]. Sin embargo, la detección de plagio translingüe ha recibido atención sólo recientemente [3-5]. No obstante, aquí nos centramos en los enfoques de detección translingüe que han sido aplicados en la "Competición Internacional de Detección de Plagio"[1] (en adelante *PAN-11*).

El proceso del enfoque externo[2] de detección de plagio translingüe aplicado en la *PAN-11* es el siguiente: (1) detección de la lengua en la que están escritos los documentos, (2) traducción si dicha lengua es distinta a la lengua base (inglés) y (3) comparación a nivel monolingüe del documento sospechoso con todos los *documentos fuente* o una selección de los mismos por estimación de similitud.

Un aspecto clave en el proceso descrito es el paso (2), al que denominamos *normalización*. Prácticamente todos los enfoques aplicados en la *PAN-11* basan esta etapa en la explotación del servicio de traducción de *Google*[3]. Si bien dicha etapa permitió a los sistemas participantes obtener buenos resultados [7], éstos deben ser valorados cuidadosamente. Por un lado, la realidad es que, como al generar los casos de plagio, fue utilizado el mismo traductor (*Google Translator*) que en la detección, esta última tarea se reduce a la de detectar copias exactas, lo cual se considera prácticamente resuelto. Sin embargo, en el momento en que otro traductor se utiliza para plagiar (o incluso dicha traducción es realizada manualmente), el problema es el de detectar casos de plagio con un alto nivel de paráfrasis, un problema que se considera aún sin solu-

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

[2] En este enfoque, un documento sospechoso se compara a otros documentos conocidos, llamados "*documentos fuente*", con el fin de encontrar fragmentos que sean más similares de lo esperado y por tanto puedan haber sido plagiados. En contraste, el enfoque intrínseco intenta determinar si un documento sospechoso contiene algún fragmento con un estilo o complejidad inesperados con respecto al resto del documento, lo cual podría ser el resultado de la inserción de texto de una fuente externa [9-11].

[3] http://translate.google.com

ción [8]. Por otro lado, la dependencia de sistemas externos de traducción puede representar un inconveniente por su potencial costo económico y limitada disponibilidad [13-14], lo que afecta de manera importante al modelo de detección. Además, dado el alto volumen de texto que en ocasiones se requeriría traducir el proceso puede ser prohibitivo en términos de tiempo, desde horas hasta días [12-14]. El único sistema translingüe participante en la *PAN-11* que no depende de un traductor externo es *Crosslingual CoReMo* [15], cuya arquitectura se describe en la siguiente sección.

3 Modelo

El sistema estudiado está enfocado en la tarea de detección externa de plagio translingüe, y está basado en la comparación de n-gramas contextuales[4] [16-17] (*CTnG*), previa normalización al inglés de los documentos. Según [15], la utilización de un diccionario más exhaustivo en la normalización, mejoraría la detección.

3.1 Etapa de normalización

Para esta tarea, se emplean dos diccionarios específicos (*direct2stemSs2Nn* y *stem2stemSs2Nn*), creados para obtener directamente *stems*[5] [18] en la lengua de normalización *Nn*, a partir de palabras completas o sus *stems* en el idioma original *Ss*.

3.1.1 Diccionarios *direct2stem* y *stem2stem* del módulo de normalización. Para minimizar el procesamiento y optimizar el uso de memoria, como el sistema solo necesita el *stem* de las palabras, se generaron dos tipos de diccionario especiales con una sola traducción (sin ambigüedad), directamente reducida a su *stem*:

El tipo ***direct2stem*** procede de la reducción al *stem* solo en el término traducido (inglés) de un diccionario translingüe normal, tras eliminar las expresiones, y eliminar la multiplicidad de términos por entrada. El ***stem2stem***, se obtiene de la reducción al *stem* del término de entrada en el diccionario *direct2stem*, y eliminación de las consecuentes duplicaciones y ambigüedades resultantes. Aunque es más reducido e impreciso, consigue aumentar la cobertura sobre los términos del texto cuando no se encuentran en el *direct2stem*.

La efectividad de estos diccionarios dependerá tanto del diccionario de origen (los de estos experimentos se detallan en 4.1.1 y 4.1.2), como de la estrategia empleada para la desambiguación de entradas múltiples. Si ésta no existe o es insuficiente, como último recurso, se emplea la mayor frecuencia del *stem* en inglés en un corpus de documentos dado, como el propio corpus de documentos conocidos (sería una generación específica para el ámbito del problema).

[4] Abreviados *CTnG*. Agrupación de n *stems* de palabras relevantes (no vacías) ordenadas entre sí y en minúscula. Para obtenerlos: (a) las palabras "vacías" (conocidas como *stopwords*) son eliminadas, (b) se aplica un proceso de *stemming*, (c) se extraen los n-gramas, y (d) las palabras en cada uno de los n-gramas resultantes son ordenadas alfabéticamente.

[5] Raíces de las palabras, obtenidas aplicando el algoritmo Porter's Stemmer [18].

3.1.2 Algoritmo de normalización a la lengua *Nn* del documento en lengua *Ss*

Simbología:
d_{Ss} *= documento en lengua original*
t_{Xx} *= token (palabra) en lengua Xx*
$stem_{Xx}(t_{Ss})$ *= stem en lengua Xx de* t_{Ss}

```
FOR EACH tSs ∈ dSs:
   IF tSs ∉ stopWordsSs:
      IF tSs ∈ direct2stem_Ss2Nn:
         tNn = traduce_direct2stem(tSs)
      ELSE:
         IF stemSs(tSs) ∈ stem2stem :
            tNn = traduce_stem2stem(stemSs(tSs))
         ELSE:
            tNn = stemNn(tSs)  #(1)
         END IF
      END IF
      tNn.posición = tSs.posición  #(2)
      tNn.longitud = tSs.longitud  #(3)
      registrar tNn en modelo normalizado de dSs
   END IF
END FOR
```

(1) Cuando el token no se encuentra en el diccionario.
(2) y (3) facilitan la localización exacta de la detección sobre el original, sin necesidad de extrapolar la detección sobre el documento normalizado.

3.2 Etapa de Detección

Combinando los unigramas procedentes de la normalización, se obtiene un modelo en *CTnG* del documento original, que se usa tanto para la indexación conjunta de los documentos conocidos, como para la comparación directa entre documentos.

El proceso de detección es el siguiente: (a) Se selecciona el documento conocido más similar para cada *chunk*[6] del documento sospechoso, pero (b) solo se comparan al documento conocido, aquellos *chunks* consecutivos que apunten al mismo (*Poda por Monotonía Referencial*). (c) Partiendo de esta monótona fracción del documento sospechoso, se delimita la correspondencia de las zonas probablemente plagiadas entre el documento sospechoso y su respectiva fuente, según la similitud de tamaño y cantidad de *CTnG* coincidentes. (d) Opcionalmente, se unen detecciones solapadas o cercanas al mismo tiempo en el documento sospechoso y la posible fuente, para dar una salida más limpia al usuario. A este paso se le llama "*filtro de granularidad*".

Entre las virtudes más destacables de *CoReMo*, están su rapidez y la escasez e independencia de recursos necesarios, siendo probablemente el más rápido en analizar

[6] Subdivisiones del documento basadas exclusivamente en un número fijo de n-gramas consecutivos. Este número, en la práctica se corresponde (aproximadamente) al doble de palabras.

PAN-PC-11, en especial para casos translingües (normalización en menos de 3 minutos y buena calificación en la detección) y traduce e indexa los 203 documentos en español (33 MB) en menos de 2 minutos, analizando el posible plagio translingüe de los 11.093 documentos sospechosos (1.6 GB) en menos de 5 minutos en modo *CT3G* (trigramas contextuales), y en menos de 9 en modo *SC3G* (variante de los trigramas contextuales, menos sensible al ruido y paráfrasis) sobre un PC a 3GHz.

4 Marco experimental

En esta sección, se detallan los objetos sujetos a prueba y del entorno de evaluación.

4.1 Diccionarios fuente

Son los recursos utilizados para generar los diccionarios de la etapa de normalización.

4.1.1 Diccionario Extendido de Flexiones de español a inglés [19]**.** Contiene información para cada palabra de entrada, de su lema[7], de las posibles variantes de palabras traducidas y de sus correspondientes lemas en inglés (excepto adverbios), acompañadas de una medida de correspondencia gramatical entre cada par {clave, posible traducción}, normalizada para el conjunto de posibles traducciones con el mismo lema. Dicha medida, es útil para determinar, dentro de la variedad de flexiones del mismo lema, la(s) que probablemente sea(n) más adecuada(s).

Sus más de cuatro millones de entradas, lo hacen bastante completo, es decir, presumiblemente contiene una entrada para toda palabra hallada en un texto.

4.1.2 Diccionarios del proyecto *Dicts.Info.* Obtenido de la unión de dos diccionarios descargados en abril de 2010 del proyecto *Dicts.info:* (a) *Wiktionary*[8] (5.860 entradas español-inglés con escasas traducciones alternativas a desambiguar) y (b) base de datos de enlaces inter-lengua entre artículos de *Wikipedia*, cuyas entradas corresponden a entidades con nombre (7.450 entradas español-inglés), normalmente no incluidas en el diccionario anterior. De esta forma se consigue un considerable número de términos, en su mayoría, en la forma canónica (lemas), sin flexiones de género, número, tiempo verbal, derivados adverbiales, etc.

4.2 Diccionarios para la normalización

Para evaluar la influencia de los diccionarios de origen, se generaron los conjuntos de diccionarios del módulo de normalización *FLEX* y *WIKI.*

---

[7] Forma base (forma normalizada) que representa el paradigma completo de una palabra (su conjugación o declinación), es decir, un conjunto de formas morfológicas con sus características gramaticales respectivas (e.g. el lema “mesa” representa las formas “mesa” y “mesas”).

[8] http://www.wiktionary.org/

4.2.1 Diccionarios *WIKI.* Son los antiguos diccionarios de *Crosslingual CoReMo*, procedentes de *Wikipedia Interlanguage Links* y *Wiktionary* de Abril de 2010, teniendo *direct2stem* y *stem2stem* respectivamente 7.404 y 6.821 entradas, desambiguadas por la frecuencia del *stem* en los documentos fuente del *PAN-PC-09*.

4.2.2 Diccionarios *FLEX*. Tomando como origen el citado diccionario extendido de flexiones, se generaron varios conjuntos de diccionarios *direct2stem* y *stem2stem* para español-inglés, marcados como A, B y C, con diferentes estrategias de preprocesamiento para evaluar los resultados obtenidos de cada una de ellas al usarlos por el modelo detector propuesto en el entorno de pruebas:

A. A partir del diccionario completo sin desambiguar (4.834.447 entradas).
B. A partir de las 442.261 entradas con máxima afinidad gramatical.
C. Similar al anterior, pero usando nuevas *stopwords* españolas e inglesas que incluyen flexiones de los verbos auxiliares, y alguna palabra más para mejorar la correspondencia entre *stopwords* de ambos idiomas.

La generación *FLEX-C*, estuvo motivada por el estudio de las palabras no encontradas por *FLEX*-A y B : se encontraron algunas *stopwords*, no incluidas en la lista de generación, muchas entidades con nombre, y algunos adverbios derivados de participios verbales o adjetivos. La primera medida evidente fue ampliar la lista de palabras vacías para generar el diccionario tipo C. Con ello se volvió a mejorar la prestación, aunque apenas una centésima, e incrementó también la ratio de palabras encontradas frente a las perdidas.

La desambiguación se llevó a cabo por frecuencia del *stem* en los documentos fuente del *PAN-PC-11*, y el número final de entradas en los diccionarios generados resultó de 317.066 (A), 309.441(B) y 304.485 (C) en el *direct2stemEs2En.dic*, y 34.366 (A), 33.811 (B) y 33.483 (C) en el *stem2stemEs2En.dic*. Comparados con *WIKI* son 41 y 5 veces más extensos respectivamente.

4.3 Corpus

En 2009, surge la Competición Internacional de Detección de Plagio, en el marco del taller *PAN*. Su objetivo es fomentar el desarrollo de herramientas automáticas para la detección de plagio. Para ello, esta competición proporciona un marco de evaluación compuesto por un corpus con casos (simulados) de plagio y medidas de evaluación especialmente diseñadas para esta tarea. Se elige el corpus *PAN-PC-2011* [20], procedente de la edición *PAN-11*, por tratarse del estándar de referencia actual para la evaluación y desarrollo de sistemas de detección de plagio. Para las pruebas de traducción y detección se emplean solo los 203 documentos fuente en español (33,3 MB), sin embargo, se han de analizar los 11.093 documentos sospechosos (1,6 GB).

4.4 Medidas de evaluación

Para la evaluación del sistema, se utilizan las medidas propuestas en el marco de la competición *PAN* [18]. Se componen de versiones adaptadas (macropromediadas[9]) de la *Precisión* (*P*) y de la *Cobertura* (*R*) --medidas estándar usadas en *RI*-- así como una nueva medida, llamada "*Granularidad*"[10] (*G*) de las detecciones válidas[11].

La medida establecida para determinar la calidad en la detección de un plagio se denomina *Plagdet*, y se determina por las fórmulas (1) y (2).

$$F = \frac{2\,R\,P}{R+P} \qquad (1) \qquad\qquad Plagdet = \frac{F}{\log_2(1+G)} \qquad (2)$$

4.5 Experimentos propuestos

Se comparan los cuatro modos de trabajo del detector del modelo: el *CT3G* (normal), y la variante *SC3G* (menos sensible al ruido y paráfrasis), con y sin uso del Filtro de Reducción de Granularidad. Esto permitirá determinar el progreso de los parámetros obtenidos del análisis (Precisión, Cobertura, Granularidad y Plagdet) para cada diccionario, según el modo de trabajo y tamaño de *chunk* seleccionados, localizando el ajuste óptimo y prestación máxima para cada combinación.

5 Resultados

La tabla 1, muestra la efectividad obtenida con los distintos diccionarios (sin valorar la corrección) en el proceso de normalización común a todos los modos de trabajo. De los resultados de WIKI, se observa que queda mucho por mejorar, pues la traducción directa (la más fiable), cubre apenas el 27% de las palabras, y que casi la mitad del total no son encontradas. Esto provoca que el sistema necesite largas longitudes de *chunk* para referenciar con garantías suficientes a algún documento fuente en caso de plagio. Las expectativas de resultados de *FLEX* en cuanto a cobertura de las palabras buscadas se cumplieron ampliamente: triplica la traducción directa y reduce a un tercio respecto a *WIKI las no halladas*.

---

[9] Promedio de las coberturas o precisiones, en su caso, obtenidas individualmente con base en caracteres de detecciones válidas, para cada uno de los plagios simulados (cobertura) o de las detecciones existentes (precisión).

[10] Medida de usabilidad del sistema que indica el grado de fragmentación de las detecciones válidas. Se obtiene de la relación entre el número de detecciones válidas y el número de casos de plagios detectados (aunque sea parcialmente). Su valor mínimo y óptimo es 1.

[11] Se considera que una detección es válida para sistemas externos solo si tiene al menos un carácter en común tanto en el documento sospechoso como en el documento fuente. En caso contrario toda la detección es considerada como falso positivo.

***Tabla 1**. Efectividad del diccionario en subcorpus de fuentes en español (no stopwords)*

	direct2stem	*stem2stem*	*no halladas*	*total*
WIKI	702.084 (26,91%)	663.497 (25,43%)	1.243.601 (47,66%)	2.609.182 (100%)
FLEX A y B	1.915.338 (73,41%)	247.477 (9,48%)	446.367 (17,11%)	2.609.182 (100%)
FLEX C	1.915.338 (74,54%)	236.168 (9,19%)	417.923 (16,27%)	2.569.429 (100%)

Sin embargo, contrario a lo que se podría esperar, las pruebas realizadas por los diccionarios extraídos por las estrategias A, B y C, fueron aceptables y comparables en cuanto a niveles de Plagdet, y presentando una distribución similar de su efectividad en cuanto al tamaño de *chunk* a los obtenidos por *WIKI*, pero siempre inferiores. En resumen: El peor resultado se obtuvo con el diccionario *FLEX* A, siempre entre 0.04 y 0.08 puntos por debajo de la prestación *Plagdet* de *WIKI*. *FLEX B, me*joró esta diferencia, aún estando entre 0.04 y 0.06 puntos por debajo d*e WIKI, s*iguiendo la misma distribución de respuesta. El diccionario *FLEX* C, reduce la diferencia hasta 0.03 ~ 0.05. Este es el diccionario utilizado en los experimentos reportados en las figuras 1 y 2.

En la figura 1, se observan cuatro pares de curvas (coincidentes cada par con el modo de trabajo) que presentan una ligera una homotecia vertical, indicando que la respuesta del algoritmo con ambos diccionarios es muy similar. Los máximos obtenidos por *FLEX* A, B y C para cada algoritmo son aceptables, y comparables a los de *WIKI.* Los tamaños de *chunk* no han experimentado el acortamiento esperado: los mejores resultados se obtienen con tamaños de *chunk* de 325 y de 100 para los modos *CT3G* y *SC3G* respectivamente (sin usar el *filtro de granularidad*), muy superiores a los valores de *chunk* óptimos obtenidos en experimentos monolingües sin necesidad de normalización, de 55 en ambos casos. Esto indica que plagios de longitud inferior se escapan, objetivo que se persigue mejorar con un nuevo diccionario.

Los resultados de la evaluación para ambos diccionarios son muy similares, sin llegarse a una conclusión concreta de para qué propósito es uno mejor que otro. En la figura 2, pueden compararse en el modo de funcionamiento *CT3G* sin *filtro de granularidad* (como muestra del efecto similar observable en los ocho modos posibles), y en ella observamos, que prácticamente son iguales en cualquier medida, normalmente algo mayor para *WIKI*, excepto en granularidad. Esto se debe a que *WIKI*, con algo mayor cobertura, detecta fragmentos más pequeños cuando hay menor coincidencia de n-gramas, y que quedan dispersos sin conectar con otras detecciones.

A pesar de la elevada ratio de palabras encontradas por *FLEX*, en comparación con *WIKI*, los resultados son inferiores. Todo esto hace replantearse las causas de porqué

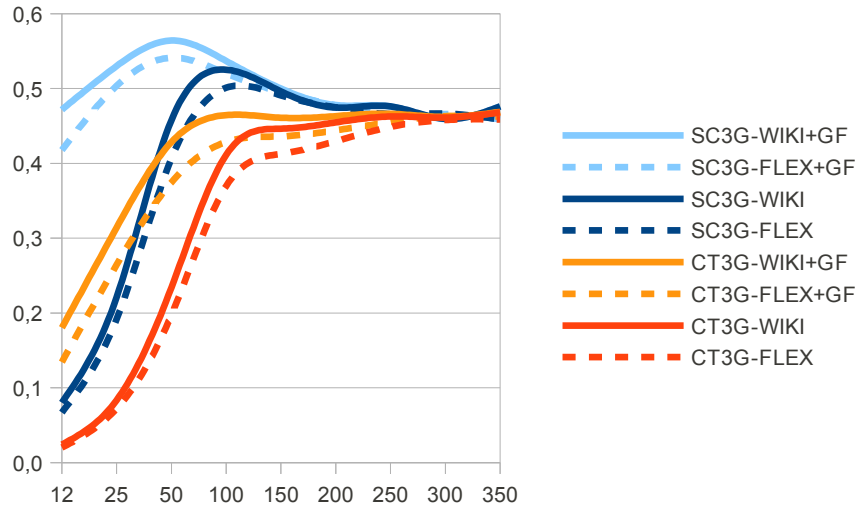

Fig. 1. Evolución del Plagdet en función de tamaño de chunk al ejecutar el modelo con los diccionarios WIKI (línea continua) y FLEX-C (línea discontinua) con los modos de funcionamiento disponibles: CT3G (rojizos), SC3G (azulados). El empleo del filtro de granularidad se indica mediante tonos más claros.

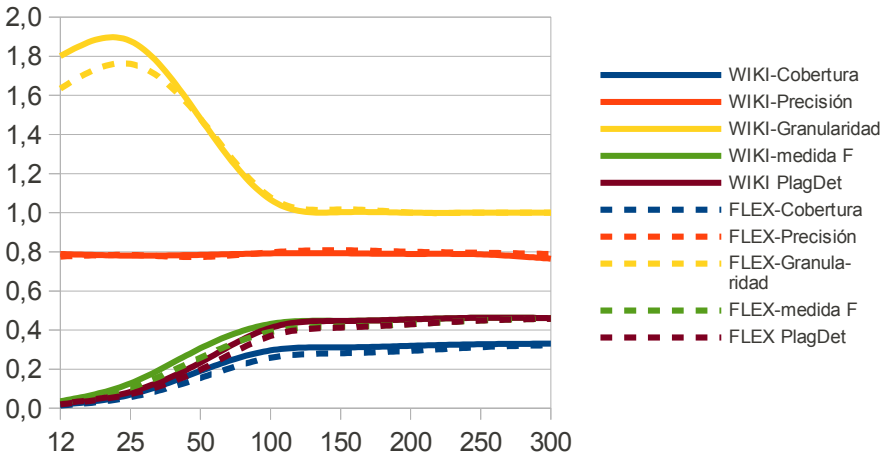

Fig. 2. Resultados de la evaluación utilizando los diccionarios WIKI y FLEX-C con el modelo en modo CT3G sin filtro de granularidad (el modo con inferiores prestaciones, pero mayores diferencias entre sus respuestas de Cobertura, Precisión y Granularidad).

un diccionario muy inferior en entradas consigue una mejor puntuación global en el sistema de detección que el extendido, que proporciona una cobertura directa de casi el triple de palabras. Se entiende que la efectividad de las traducciones de los diccionarios *WIKI*, ha sido muy superior, decantándose el resultado en favor de la calidad en lugar de la cantidad. Volviendo a analizar los orígenes de ambos diccionarios, se observa que los antecesores de *WIKI*, tenían una sola, o muy pocas alternativas de traducción (escasa ambigüedad), e incluyen gran variedad de entidades con nombre. El Diccionario de Flexiones tiene una gran variedad de salidas que desambiguar, y carece prácticamente de entidades con nombre.

6 Conclusiones y trabajo a futuro

Se han comparado los resultados de un modelo de detección de plagio externo translingüe, empleando distintos diccionarios de normalización, generados a partir de diccionarios bilingües procedentes de recursos Internet (*WIKI*) y de un diccionario extendido mediante flexiones gramaticales (*FLEX*) con millones de entradas. Contrario a lo esperado, los resultados fueron mejores con los diccionarios más sencillos. Del análisis de esos resultados se exponen las siguientes concusiones:

- La inclusión de entidades con nombre en los diccionarios, es a priori una buena idea, pues suelen tener pocas variaciones o sinónimos, convirtiéndose en una poderosa fuente de contextualización/detección que puede mejorar la efectividad global del sistema. Se sugiere un nuevo cóctel con otro diccionario para ampliarlo.

- El método utilizado para desambiguar (frecuencia del *stem* en un corpus del idioma destino), no es muy idóneo, en especial cuando se disponen de muchas alternativas. Suele decantarse a palabras frecuentes (similares a las *stopwords*, con poca capacidad de definir el contexto). En especial, ante la gran variedad de alternativas del diccionario, la probabilidad de acierto, disminuye bastante.

- El *Diccionario de Flexione*s, a pesar de estos resultados, sigue siendo una fuente potencialmente interesante para la elaboración de los diccionarios orientados a la detección de plagio, por el nivel de cobertura de palabras mostrado. No obstante, se precisan de nuevas estrategias que faciliten una desambiguación más idónea.

- La afinidad gramatical incluida en el diccionario de Flexiones, no sirve para desambiguar fuera del mismo lema, por lo que no ayuda a elegir el lema correcto para aplicar luego la mejor flexión. Tampoco lo es su producto por la frecuencia del término o de su *stem*, ya que sesga la elección al lema con menos flexiones.

- Un indicador extra, a incluir en el diccionario de flexiones, de *correlación de traducción estadística entre lemas de clave y de traducción*, normalizado para el conjunto de lemas alternativos, ayudaría notablemente a la elección final del mejor

candidato en base primero a dicha correlación por lema, y luego afinada por la similitud gramatical de clave/traducción.

- Una estrategia prometedora con la estructura de dicho diccionario, es usar la similitud gramatical para la elección de la mejor flexión en el subconjunto del lema en inglés coincidente con la traducción del lema de la palabra española en un diccionario menos ambiguo, como por ejemplo, los mencionados de *Wiktionary* y *Wikipedia*. Se podría así obtener una interesante combinación de diccionarios, y no solo para el propósito de detección de plagio.
- El diccionario de flexiones podría enriquecerse con la inclusión de adverbios de modo, generados automáticamente, del mismo modo que el resto de flexiones.
- Actualmente, los diccionarios generados, descartan entradas que contengan expresiones de salida, o combinaciones de dos o más palabras más simples. En idiomas como el alemán, son muy frecuentes, y el proceso automatizado de generación de diccionarios, actualmente las descarta. Es posible modificar el proceso de normalización para incluir las expresiones como varios *stems* consecutivos con delimitaciones repetidas de la posición de la palabra original en el documento sospechoso.

Agradecimientos

La investigación de los miembros de la Universitat Politècnica de València y del Instituto Politècnico Nacional se ha desarrollado en el marco del proyecto EC WIQ-EI (IRSES n. 269180). La investigación del segundo autor es llevada a cabo gracias a la beca CONACyT-México 192021/302009. Este trabajo se ha desarrollado en el marco del VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems y es financiado parcialmente por los proyectos MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (plan I+D+i), CONACyT-México 83270 SNI, SIP-IPN 20111146 y 20120418, ICYT-DF PICCO10-120.

7 Referencias

1. Clough, P.D. (2003), Measuring Text Reuse, PhD thesis, University of Sheffield.
2. Maurer, H., Kappe, F., and Zaka, B. (2006). Plagiarism - A Survey. Journal of Universal Computer Science, 12(8), 1050-1084.
3. 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 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse, pp. 9-13. Patras, Greece.
4. Pinto D., Civera J., Barrón-Cedeño A., Juan A., Rosso P (2009). A statistical approach to crosslingual natural language tasks. In: Journal of Algorithms, vol. 64, num. 1, pp. 51-60.
5. Ceska, Z., Toman, M., and Jezek, K. (2008). Multilingual Plagiarism Detection. In Proceedings of the 13th International Conference on Artificial Intelligence (ICAI 2008), page83-92, Varna, Bulgaria. Springer-Verlag.

6. Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Cross-Language Plagiarism Detection. Language Resources and Evaluation (LRE), 45 (1): 45-62, 2011.
7. Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Overview of the 3rd International Competition on Plagiarism Detection. In [22]
8. Stein B., Potthast M., Barrón-Cedeño A., Rosso P., Stamatatos E., Koppel M. Fourth International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. In: SIGIR Forum, 45 (1), pp. 45-48, June 2011. ACM
9. Meyer zu Eissen, Sven and Benno Stein. (2006). Intrinsic plagiarism detection. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006),London, ISBN 3-540-33347-9, volume 3936 of Lecture Notes in Computer Science, pages 565–569. Springer.
10. Stamatatos, E. (2009). Intrinsic Plagiarism Detection Using Character n-gram Profiles. In Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors (2009). SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), volume 502, San Sebastian, Spain. CEUR-WS.org. http://ceur-ws.org/Vol-502.
11. Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE), 45 (1): 63-82, 2011.
12. Neil Cooke, Lee Gillam, Henry Cooke Peter Wrobel, and Fahad Al-Obaidli. A High-performance Plagiarism Detection Sy*stem*: Notebook for PAN at CLEF 2011. In [22]
13. Cristian Grozea and Marius Popescu. The Encoplot Similarity Measure for Automatic Detection of Plagiarism: Notebook for PAN at CLEF 2011. In [22].
14. Du Zou, Wei-Jiang Long, and Ling Zhang. A Cluster-Based Plagiarism Detection Method: Lab Report for PAN at CLEF 2010. In Braschler et al. [21].
15. Diego Antonio Rodríguez Torrejón y José Manuel Martín Ramos. Crosslingual CoReMo Sy*stem*: Notebook for PAN at CLEF 2011. In [22].
16. Diego Antonio Rodríguez Torrejón y José Manuel Martín Ramos (2010). Detección de plagio en documentos. Sistema externo monolingüe de altas prestaciones basado en n-gramas contextuales . Procesamiento del Lenguaje Natural, 45:49–57 .
17. Diego Antonio Rodríguez Torrejón y José Manuel Martín Ramos. CoReMo System (Contextual Reference Monotony): a Fast, Low Cost and High Performance Plagiarism Detection System. Lab Report for PAN at CLEF 2010. In [21].
18. Martin F. Porter. An algorithm for suffix stripping (Porter stemmer). Program, 14(3):130-137. (1980) http://tartarus.org/~martin/Porter*Stem*mer/index.html
19. Grigori Sidorov, Alberto Barrón-Cedeño and Paolo Rosso. English-Spanish Large Statistical Dictionary of Inflectional Forms. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA), 2010, pp. 277-281
20. Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In 23rd International Conference on Computational Linguistics (COLING 10), August 2010. Association for Computational Linguistics
21. Braschler, Harman, and Pianta, editors. Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy, 2010. ISBN 978-88-904810-0-0.
22. Petras, Forner, and Clough, editors, Notebook Papers of CLEF 2011 LABs and Workshops, 19-22 September, Amsterdam, The Netherlands, September 2011.