Detección de plagio translingüe utilizando una red semántica multilingüe

Marc Franco Salvador

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS
Y COMPUTACIÓN

Dirigido por:
Paolo Rosso

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Trabajo Final de Máster desarrollado dentro del Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Valencia, Febrero 2013

Cuanto más cercana a la verdad, mejor será la mentira, y la misma verdad, cuando puede utilizarse, es la mejor mentira.

Isaac Asimov

Resumen

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 apropiarse, de forma muy sencilla, del contenido original de otros autores. Aunque el plagio se puede detectar de forma manual, dada la gran cantidad de contenidos que se publican, es una tarea prácticamente imposible de llevar a cabo, aún más si las fuentes de plagio vienen de documentos en otros idiomas.

Actualmente existe un gran interés, dentro de la literatura y la ciencia, por investigar y desarrollar sistemas de detección de similitud a nivel monolingüe y translingüe que sean capaces de detectar de forma automática las secciones de plagio entre documentos. La comunidad académica también se ve beneficiada por dichos sistemas, ya que permite la detección y disuasión por parte de los profesores hacia su alumnado, de las prácticas habituales de copiar y pegar, sin referencia alguna a la fuente de procedencia, de contenidos originales obtenidos de la Web.

En la presente tesis describimos el estado del arte en materia de detección de plagio textual a nivel monolingüe y translingüe. Además, se estudia la utilización de una red semántica multilingüe para crear dos modelos de detección de plagio translingüe: utilizando un diccionario estadístico, y mediante grafos de conocimiento a modo de modelos de contexto para modelar fragmentos de documento. Los resultados experimentales resultan muy prometedores. Como trabajos futuros, se definen diferentes líneas de investigación haciendo uso de grafos de conocimiento.

Abstract

Plagiarism is defined as the unauthorized use of the original content of other authors. It is a difficult phenomenon to detect whose problem has worsened in recent years because of the Internet: a vast source of information that allows users to copy and take possession, very simply, of the original content of other authors work. Although plagiarism can be detected manually, given the large amount of content published, it is virtually impossible to carry out, even more if the source of plagiarism comes from documents in other languages.

Currently, literature and science have strong interest in research and development of automatic monolingual and cross-language similarity detection systems, capable of detecting plagiarism among sections between documents. The Academic Community also benefits by such systems. It allows teachers to detect and discourage their students of the usual practice of copy and paste, without reference to its source, from original content obtained from Internet.

In this thesis we describe the state-of-the-art in text plagiarism detection at monolingual and cross-language level. In addition, we study the use of a multilingual semantic network to create two cross-language plagiarism detection models: using a statistical dictionary, and using knowledge graphs as context models from document fragments. Experimental results are very promising. As future work, we define different research lines using knowledge graphs.

Agradecimientos

Me gustaría dar mi más sincero agradecimiento a algunas personas sin las que no hubiera sido posible llevar a cabo este trabajo de final de máster.

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 mi investigación. 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), quiero agradecer su financiación para mi estancia de un mes en México, en el Centro de Investigación en Computación del Instituto Politécnico Nacional (México D.F.) y en el Instituto Nacional de Óptica Electrónica y Astrofísica (Puebla).

A mi director de tesina, el doctor Paolo Rosso, el cual depositó su confianza en mí y me ofreció toda la ayuda y apoyo necesarios para seguir adelante en mi trabajo. A los doctores Grigori Sidorov y Manuel Montes por su amabilidad y ayuda durante mi estancia en México. A mi compañero de laboratorio Parth Gupta, por prestarme su ayuda y consejo en tantas ocasiones, y por ser tan paciente conmigo cuando apenas comenzaba mi labor de investigación. A Enrique Flores por su ayuda durante las fases finales de esta redacción. También a Roberto Navigili por haber desarrollado BabelNet y ofrecer su ayuda para familiarizarnos con el API sistema.

A todos mis amigos de dentro y fuera del máster que han tenido que soportar todas mis quejas e inquietudes en los peores momentos, además de interesarse por mi trabajo cuando no estaban relacionados con la investigación.

Quiero agradecer especialmente a mis padres su apoyo y por ser tan pacientes y comprensivos conmigo durante todo el proceso que duró esta investigación. Tampoco olvidar a la familia que se ha preocupado de saber como me iba durante todo este tiempo.

Por último, agradecer a todo aquel que haya olvidado nombrar en este apartado y que haya influido de alguna manera en el éxito de esta dura etapa de mi vida.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Descripción del problema, motivación y objetivos

En los últimos años, Internet ha tenido un impacto profundo en el mundo laboral, el ocio y el conocimiento a nivel mundial. Gracias a la Web, millones de personas tienen acceso fácil e inmediato a una cantidad extensa y diversa de información *online*. Desde su creación ha supuesto toda una revolución en la manera de pensar y actuar de las personas. En contraposición, si bien ha contribuído a mejorar enormemente a nuestra sociedad, su concepción también ha significado la aparición de toda una nueva serie de delitos o infracciones que hacen uso de este como conducto: suplantación de identidad, robo de cuentas bancarias, redes de pederastia, infracciones del *copyright*, etc. Gracias a la Web, también ha surgido una práctica nueva para los autores: la reutilización. Esta hace uso de fragmentos parciales o totales de otras fuentes para elaborar "nuevo" contenido. Dicha práctica no siempre es ilegal ya que puede venir acompañada de referencias a su fuente original, o ser parte de una publicación de libre disposición, siendo libre su reutilización. El problema ocurre cuando los contenidos provienen de fuentes no citadas, y por tanto, se está asumiendo la autoría de los mismos. Este fenómeno es conocido como plagio.

En la RAE se define el plagio (*plagiarism*) como "la acción de copiar en lo sustancial obras ajenas, dándolas como propias" [13].

Este hecho ha ocasionado graves problemas de autoría en la literatura y la ciencia. Un claro ejemplo sería cuando en abril de 2011 saltó a la prensa una noticia que supuso un escándalo a nivel internacional: el ex ministro alemán de Defensa Karl Theodor zu Guttenberg fue condenado por la vía penal y tubo que dimitir de todos sus cargos por plagiar gran parte del texto de su tesis doctoral[1]. Dentro de la literatura también podemos encontrar listas de presuntos casos de "plagio célebres"[2] con nombres de autores hispanos de reconocido renombre: Garcilaso de la Vega, Camilo José Cela, Pablo Neruda, Gabriel García Márquez, etc.

Dentro de la ciencia, concretamente en las publicaciones de artículos científicos, aparece el término de auto-plagio (*self-plagiarism*). Este consiste en copiar partes de tamaño significativo de trabajos propios, publicados previamente, sin citar la fuente de la publicación original. Ciertamente no se incurre en delito de autoría si tomamos parte de un contenido propio, pero se está cometiendo una falta con respecto a la comunidad científica. Actualmente no existe una política consensuada respecto al auto-plagio [22]. Por ejemplo, IEEE[3] no admite reutilización de porciones largas del texto de trabajos previos, mientras que ACM[4] lo permite si viene acompañado de la referencia a la fuente original.

Si bien es físicamente posible la detección de plagio de forma manual por parte de personas expertas, actualmente, debido a la gran cantidad de publicaciones que se producen diariamente, su detección manual en la práctica se hace imposible; aún más si el origen del plagio proviene de una fuente en otro idioma. Lo cual es conocido como el plagio translingüe (*cross-language plagiarism*).

La investigación dentro del campo de la detección de plagio translingüe está justificada. En una encuesta realizada recientemente so-

[1] El 23 de febrero de 2011 Guttenberg ha sido desposeído de su título de doctorado por la Universidad de Bayreuth tras las pruebas de plagio detectadas en su tesis doctoral: `http://www.elpais.com/articulo/internacional/Dimite/ministro/Defensa/aleman/plagiar/tesis/doctoral/elpepuint/20110301elpepuint_6/Tes`

[2] Lista de conocidos autores que han cometido plagio literario abiertamente: `http://elplagio.com/?page_id=27`

[3] `http://www.ieee.org/publications_standards/publications/rights/ID_Plagiarism.html`

[4] `http://www.acm.org/publications/policies/plagiarism_policy`

bre las actitudes y prácticas de los estudiantes en las universidades [2], 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 texto desde otros documentos y incluirlos en sus trabajos no es plagio. De todo lo anterior podemos deducir que no solo el plagio se produce en una gran medida dentro de la comunidad académica, sino que además la población no está concienciada de que lo que están cometiendo es un grave delito. Para la detección y disuasión de esta práctica, es necesaria la investigación y el desarrollo de herramientas que permitan la detección de plagio, a nivel monolingüe y translingüe, de forma automática.

Existen modelos de detección de plagio translingüe que traducen el texto completo antes de comenzar un análisis posterior para determinar si existe plagio [39, 19], lo cual conlleva una pérdida de recuperación (*recall*) de casos de plagio y no es realista en un escenario como la Web. Por ello, existen una serie de modelos de análisis de similitud, que pueden utilizarse a nivel translingüe para llevar a cabo la detección de plagio de forma automática sin tener que pasar por esta traducción previa. 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) [30] es un modelo que se basa en la sintaxis de los documentos, haciendo uso de n-gramas a nivel de caracter, que ofrece un rendimiento notable para lenguajes con similitudes sintácticas. *Cross-language conceptual thesaurus based similarity* (CL-CTS) [21], como su nombre indica, utiliza un tesauro lingüístico para analizar la similitud entre documentos. *Cross-language explicit semantic analysis* (CL-ESA) [43] 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) [4, 41] 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 [43, 21], ofreciendo CL-ASA y CL-CNG el mejor desempeño. Por ese motivo, en nuestra

evaluación comparamos nuestra aproximación con éstos.

Dentro del ámbito de la detección de plagio automática, desde el año 2009 se celebra anualmente una competición internacional, *Uncovering Plagiarism Authorship and Social Software Misuse* (PAN)[5], en la cual se presentan y ponen a prueba aproximaciones para la detección de plagio a nivel monolingüe y translingüe. Dentro de la competición tenemos dos tareas: detección de plagio intrínseco y externo. La detección de plagio intrínseco se trata de, dado un documento, analizar su estructura para determinar las características del autor y detectar las secciones que no parecen propias de este. 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.

El objetivo principal del presente trabajo es estudiar el estado del arte en materia de análisis de similitud textual, a nivel monolingüe y translingüe, para su posterior aplicación en detección de plagio. Además, estudiamos la utilización de un recurso lingüístico como es la red semántica multilingüe de BabelNet [33], para su aplicación en detección de plagio translingüe. En concreto, haciendo uso de ésta, se proponen dos nuevas aproximaciones de análisis de similitud translingüe. En primer lugar se propone una aproximación utilizando el diccionario estadístico de BabelNet sobre el modelo CL-ASA como base. A continuación se propone un nuevo modelo, llamado *cross-language knowledge graphs analysis* (CL-KGA), que mediante grafos de conocimiento generados por una red semántica multilingüe, los cuales expanden y relacionan los conceptos originales del texto, proporciona un modelo de contexto de los documentos sospechosos y fuente a comparar. 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 [44][6]. En concreto, en nuestra evaluación utilizamos la partición de detección de plagio translingüe de la tarea de detección

[5] `http://pan.webis.de/`

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

de plagio externo, para comparar los modelos del estado de arte, CL-CNG y CL-ASA, con nuestras aproximaciones.

1.2. Estructura de la tesis

Además del actual capítulo introductorio, el presente trabajo consta de otros cinco capítulos y un apéndice, los cuales se describen a continuación:

- **Capítulo 2** Estado del arte.

  Este capítulo describe el estado del arte en detección de plagio textual. En primer lugar se definen los tipos de plagio existentes y los enfoques que se utilizan en texto para su detección: intrínseco y externo. A continuación se describen los principales modelos de detección de plagio textual monolingüe, y finalmente describimos los modelos de detección de plagio textual translingüe.

- **Capítulo 3** Red semántica multilingüe.

  En este capítulo se describe en qué consisten las redes semánticas, y dentro de estas, su extensión como redes semánticas multilingües. A continuación describimos la red semántica multilingüe BabelNet, la cual utilizamos en nuestro trabajo, y los diferentes usos que se le pueden dar. Finalmente, dentro de estos, explicamos en qué consisten los grafos de conocimiento que se pueden generar a partir de una red semántica, y sus aplicaciones.

- **Capítulo 4** Modelos propuestos.

  En esta parte describimos las dos aproximaciones propuestas en este trabajo: el modelo CL-ASA utilizando el diccionario estadístico de Babelnet, y el modelo CL-KGA, que utiliza los grafos de conocimiento de BabelNet para llevar a cabo el análisis de similitud translingüe. Para el diccionario estadístico proponemos diferentes métodos de normalización de pesos, en función de la longitud del documento y del número de traducciones. Para el modelo CL-KGA proponemos un método de análisis de similitud de grafos y un método de normalización de pesos en grafos tras una intersección.

- **Capítulo 5** Evaluación de los modelos.

  En este capítulo describimos la competición de detección de plagio PAN, el corpus utilizado para nuestra evaluación, el PAN-PC'11, y las unidades de medida que se emplean. A continuación evaluamos los modelos propuestos contra los modelos CL-ASA y CL-CNG, utilizando su partición de detección de plagio externo, la cual comprende particiones de detección de plagio translingüe español-inglés y alemán-inglés.

- **Capítulo 6** Conclusiones y trabajos futuros.

  En el último capítulo se plantean las conclusiones que se pueden extraer de los resultados experimentales de nuestra evaluación. Además se proponen diferentes líneas de investigación para el futuro relacionadas con nuestro trabajo: utilización de diferentes redes semánticas multilingües para la extensión de nuestro modelo a más idiomas, y aplicación de las aproximaciones propuestas para detección de plagio o análisis de similitud a nivel monolingüe. Por último, en vista de los resultados obtenidos, se proponen líneas de investigación que utilicen grafos de conocimiento para tareas diferentes al análisis de similitud como minería de opiniones, clasificación de textos translingüe, adaptación de dominios y generación de resúmenes textuales.

- **Apéndice A** Publicaciones y charlas invitadas.

  En este apéndice mostramos las publicaciones a las que ha dado lugar este trabajo de investigación, además de las charlas invitadas en las que se ha divulgado su contenido.

Capítulo 2

Estado del Arte

2.1. Detección de reutilización y plagio en texto

El plagio se define como "la acción de copiar en lo sustancial obras ajenas, dándolas como propias" [13]. A la hora de hablar de plagio, es importante diferenciar antes la reutilización. La reutilización de texto se define como copiar en lo sustancial obras ajenas, dándolas como propias o no, por tanto, la reutilización puede verse como un hiperónimo del plagio. Si bien ambos tienen una marcada diferencia, esto no afecta a la hora de su detección, ya que el resultado sobre el papel es similar, y también lo será el proceso de detección a seguir [9]. Por este motivio, a partir de ahora nos referiremos directamente al término "plagio" y trataremos únicamente la detección de este.

Dada la gran cantidad de contenidos y obras disponibles actualmente, la detección de plagio por parte del ser humano se convierte en una tarea prácticamente imposible. Por este motivo es necesario el desarrollo de modelos de detección de plagio automático. Dentro de esta clase podemos encontrar el uso de técnicas de detección de plagio en tres vertientes claramente diferenciadas: lingüística forense, procesamiento del lenguaje natural (PLN) y recuperación de información (RI).

La lingüística forense aplica la detección de plagio, sobre todo, para determinar la autoría de notas de suicidio, amenazas, etc. Dada una nota de suicidio escrita "supuestamente" por un fallecido, el objetivo es analizar si la nota realmente ha sido escrita por él. Den-

tro de esta tarea se pueden realizar análisis tipográficos, en el caso de escritura manual, o se puede atribuir la autoría también analizando el estilo y características propias de la escritura del autor en el contenido del texto.

El procesamiento del lenguaje natural utiliza la detección de plagio para, mediante un análisis sintáctico y estructural del texto, determinar su similitud total o parcial con otro texto. Esta práctica trata de detectar el plagio a pesar de intentar "engañar" al detector mediante técnicas de paráfrasis, reformulando el fragmento de plagio, o bien si se ha producido un resumen o redistribución del contenido. Por otro lado, las técnicas de procesamiento del lenguaje natural conllevan un alto coste computacional para el análisis textual.

La recuperación de información se define como la "aplicación de las técnicas computacionales para la adquisición, organización, almacenamiento, recuperación y distribución de la información" [25]. El plagio dentro de esta vertiente se utiliza para tareas de clasificación y categorización de documentos [40], o bien para la detección de duplicados casi idénticos entre documentos [45]. Las técnicas que aquí se emplean suelen implicar un coste computacional reducido a cambio de una pérdida de precisión, ya que no son capaces de detectar paráfrasis. Es imporante señalar, que su aplicación para la detección de plagio es muy válida, ya que en una gran parte de los casos de plagio que se producen en texto, se han limitado a realizar una copia exacta de la fuente [47].

De acuerdo al tipo de análisis que se realiza en el texto, la detección de plagio automática se puede dividir en dos clases: intrínseco y externo.

La detección de plagio intrínseco trata de, dado un documento, analizar su estructura para determinar las características del autor y detectar las secciones que no parecen propias de este. Para ello extrae las características y estilo de fragmentos del texto, y las compara con otros fragmentos del mismo texto para determinar si pertenecen al mismo autor [8]. Características que se suelen extraer son: el rango

del vocabulario, la longitud de las oraciones y de las palabras. Para su análisis será necesario el uso de técnicas de PLN o RI, siendo PLN más común si lo que se desea es una obtención en detalle del estilo y características propios del autor. Cabe señalar que esta clase de detección de plagio es muy utilizada dentro de la lingüística forense, comparando características y estilo con otros documentos escritos por el autor.

La detección de plagio externa 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 llevar a cabo esta clase de detección, se suele realizar un proceso de tres fases [56]: En primer lugar una clasificación temática de los documentos para descartar los que no tienen relación; a continuación una comparación entre documentos de una misma clase para determinar un valor de similitud entre ellos o fragmentos de ellos; finalmente se realiza un análisis de todos los valores de similitud obtenidos para descartar los que no son realmente casos de plagio. Dentro de la clase de detección de plagio externo se utilizan técnicas de PLN y RI, tanto en conjunto (detección de fragmentos sospechosos de plagio con RI y análisis de similitud con PLN), como por separado según la clase de plagio que queramos detectar (*copy-paste*, paráfrasis, resumen...) y de lo importante que es la velocidad en el proceso de detección.

De acuerdo al lenguaje del texto de los documentos a analizar, se pueden distinguir dos clases de detección de plagio: monolingüe y translingüe. La detección de plagio monolingüe, como su nombre indica, detecta plagio entre documentos pertenecientes a un mismo lenguaje. En cambio, la detección de plagio translingüe es capaz de detectar el plagio entre documentos pertenecientes a lenguajes distintos. En la primera clase, al trabajar en un mismo lenguaje, el plagio es más fácil de detectar y no requerirá de modelos tan complejos. En la segunda clase se van a requerir técnicas de análisis para poder trabajar a nivel translingüe, ya sea traduciendo previamente todo el documento mediante una máquina de traducción estadística, para trabajar a nivel monolingüe [39, 19], o bien utilizando modelos que trabajen directamente a nivel translingüe utilizando análisis de sin-

taxis [30], tesauros [21], corpus comparables [43] o alineados [4, 41].

En general, independientemente del tipo de detección de plagio o de los lenguajes de los documentos, para todo proceso de comparación de documentos se suele seguir un proceso de tres etapas: preproceso, representación de la información y aplicación de una medida de similitud. En el preproceso se suele hacer una tokenización eliminando signos de puntuación, además de eliminar la capitalización del texto. En algunos casos también se aplica una lematización de las palabras, siendo más común si se trabaja con técnicas de PLN. También puede resultar conveniente eliminar palabras muy comunes con poco valor como los artículos o las preposiciones. En la representación de la información se pueden emplear diferentes técnicas, como el uso de n-gramas, bolsas de palabras (palabras de un fragmento de texto agrupadas sin orden), o fragmentos de texto de tamaño de una o más frases. La medida de similitud viene en función del tipo de representación que se haya utlizado, siendo lo más común utilizar modelos probabilísticos, o modelos de espacio vectorial [3].

En las siguientes dos secciones vamos a estudiar los modelos de detección de plagio monolingüe y translingüe para las clases de detección intríseco y externo. Para más información sobre el estado del arte en detección de plagio, la tesis doctoral de Alberto Barrón-Cedeño [2], también enfocada en la detección de plagio, ofrece una recopilación muy ampliada del estado del arte y de todo el trabajo anterior relacionado con la detección de plagio.

2.2. Detección automática de plagio textual monolingüe

En el trabajo publicado en [42], centrado en las medidas de evaluación en detección de plagio, se ha hecho una recopilación de publicaciones relacionadas con detección de plagio y reutilización en general. En este trabajo nos vamos a centrar a los modelos más representativos del estado del arte.

En esta sección trataremos la detección de plagio textual, intrínseco y externo, a nivel monolingüe.

2.2.1. Detección de plagio intrínseco monolingüe

A continuación se muestran las aproximaciones a la detección de plagio intrínseco monolingüe que mejores resultados han ofrecido.

Media de frecuencias de clases de palabras

Una de los métodos que mejores resultados ofrece en análisis intrínseco a nivel monolingüe es el uso de medias de frecuencias de clases de palabras (*Averaged Word Frequency Class*). En el trabajo [31] se analizaron a través de 450 documentos, creados a partir de un corpus artificial, estadísticas y características como:

- *La media de frecuencias de clases de palabras* proporciona una estimación de la complejidad y diversidad del vocabulario.
- *La longitud media de las frases* da una medida de la complejidad de las frases en los documentos.
- *Las características de la estructura gramatical (Part-of-Speech)* ofrecen una medida de la variedad en el lenguaje.
- *El promedio del número de palabras de paro (Stopwords)* da una medida del número de artículos, preposiciones, etc, en el texto.

A las anteriores medidas se les aplicó un análisis discriminativo, buscando características que generasen una separación entre diferentes clases, siendo la media de frecuencias de clases de palabras la que mejores resultados ofreció.

Perfiles de *n*-gramas de caracteres

Esta aproximación [54], haciendo uso de n-gramas, dada su simplicidad, ofrece un buen rendimiento en detección de plagio intrínseco.

Haciendo uso de *3*-gramas, crea perfiles p_d de diferentes documentos d, y perfiles p_s de fragmentos de texto $s \in d$, utilizando una ventana deslizante de tamaño m y desplazamiento n ($m = 1000$ y $n = 200$ en el trabajo original). La disimilitud entre p_d y p_s, se estima mediante la d_1 *normalizada*:

$$nd_1(p_s, p_d) = \frac{\sum_{t \in p_s} \left(\frac{2(tf_{t,p_s} - tf_{t,p_d})}{tf_{t,p_s} + tf_{t,p_d}} \right)^2}{4|p_s|} , \quad (2.1)$$

donde $tf_{t,x}$ es la frecuencia normalizada del término t (un n-grama de caracteres) en x. El resultado $nd_1(p_s, p_d)$ viene acotado en $0 \leqslant nd_1 \leqslant 1$, siendo 0 el mejor valor posible de similitud.

Medidas de complejidad de Kolmogorov

Las medidas de complejidad de Kolmogorov [29] también han sido usadas para detección de plagio intrínseco en trabajos como [50], donde se representa el texto como estadísticas de características sintácticas y gramáticas, sobre una cadena binaria en la que se comparan nombres contra no-nombres. También realizan el experimento comparando palabras largas contra cortas. Las cadenas binarias se comprimen mediante un algoritmo con un alto porcentaje de compresión [29], siendo la codificación y compresión un clasificador de plagio, ya que permite diferenciar entre diferentes secciones del texto según los ratios de compresión.

2.2.2. Detección de plagio externo monolingüe

En general, el proceso de detección de plagio externo tiene las siguientes etapas: preproceso, análisis de similitud y postproceso.

El preproceso puede formar parte, o no, de un sistema de detección de plagio externo. Tareas típicas que se pueden realizar en el preproceso son una tokenización, lematización, eliminación de la capitalización y eliminación de palabras pertenecientes a determinadas categorías gramaticales, como las preposiciones o los artículos.

El postproceso es la etapa en la que analizamos todos los valores de similitud obtenidos por el detector, para determinar cuales se salen de la media lo suficiente para determinar que es plagio. Por ejemplo se puede determinar que es plagio si se sale de la media tres desviaciones típicas. Otra tarea que se realiza en el postproceso es

tomar diferentes fragmentos de plagio d_i reportadas por un detector sobre un documento d, y combinarlas como una sola en el caso en que se solapen o estén lo suficientemente cerca, por debajo de un umbral establecido. La técnica de análisis detallado que se utiliza en la competición de detección de plagio PAN a nivel monolingüe es descrita en la sección 5.1.2.

A continuación se muestran las aproximaciones de análisis de similitud que mejores resultados ofrecen y pueden utilizarse para la tarea de detección de plagio externo.

Modelos basados en huella digital

Una huella digital (*fingerprint*) de un documento de texto es una representación resumen de su contenido, a modo de firma. Se suele modelar utilizando funciones *hash*, que proporcionan números únicos a partir de secuencias de caracteres [57]. Las secuencias de caracteres pueden ser conjuntos de caracteres, palabras o frases (conocidos como *shingles*, una secuencia continua de *tokens* en un documento).

Algunos modelos utilizando huella digital son los siguientes:

- **COPS** [6] fue uno de los primeros modelos en utilizar huella digital en texto. El sistema se usaba para controlar los lugares en los que un documento podía ser republicado. El sistema COPS sigue el siguiente esquema: (i) cuando un nuevo documento d se crea, es registrado en un servidor; (ii) d se segmenta en *shingles* (generalmente frases). Cada *shingle* se almacena en una gran base de datos. Cuando un nuevo documento d' entra en el sistema, utilizamos el algoritmo descrito en la fig. 2.1. Si un *shingle* de d' tiene una coincidencia con un *shingle* de un documento d de la base de datos, el sistema alertará de una coincidencia.
- **Winnowing** [48] hace uso de n-gramas como *shingles* para generar un conjunto de valores *hash* del texto. Los n-gramas pueden ser a nivel de caracter, palabra o frase. El modelo trabaja con una ventana deslizante que guarda siempre el menor

1: **Given** d' **and** $DB_{\mathcal{H}}$**:**
2: break d' into shingles d'_i
3: for each chunk d'_i in d':
4: Compute $\mathcal{H}(d'_i)$
5: if $\exists\, \mathcal{H}(d'_i)$ in $\mathcal{H}$*
6: return $d \mid \mathcal{H}(d'_q) \in d$

Figura 2.1: **Algoritmo COPS**, donde d' es un documento sospechoso de ser plagio, $\mathcal{H}$ es una función de *hash*, y $\mathcal{H}$* es la base de datos de *hash* de *shingles* generados previamente de una colección de documentos D.

valor de *hash* de la ventana en su base de datos, para posteriormente detectar similitudes con un nuevo documento entrante.

- **SPEX** [5] compara subcadenas de documentos. Las subcadenas son secuencias de n-gramas de palabras presentes en más de un documento. El algoritmo genera, para cada documento, n-gramas en un rango [1,l] y los almacena en una base de datos utilizando una función *hash*. Cada documento sospechoso de contener plagio será descompuesto en n-gramas del mismo modo y comparado con las firmas almacenadas en la base de datos. En la fig. 2.2 se muestra el algoritmo SPEX.

1: **Given a collection of documents** D**:**
2: for each $d \in D$:
3: for each 1-gram $g \in d$
4: $\mathcal{H}_1^* \leftarrow \mathcal{H}(g)$
5: for $n = \{2, \ldots, l\}$:
6: for each $d \in D$:
7: for each n-gram in $g \in d$
8: if $cnt(\mathcal{H}^*_{n-1}, g_{[0,n-1]}) == cnt(\mathcal{H}^*_{n-1}, g_{[1,n]}) == 2$:
9: $\mathcal{H}_n^* \leftarrow \mathcal{H}(g)$

Figura 2.2: **Algoritmo SPEX**, donde $\mathcal{H}_n^*$ es la tabla *hash* de n-gramas de palabra. Para cada n-grama en $\mathcal{H}_n^*$ se asocia un contador. La función *cnt* devuelve un contador para un *hash* dado.

Modelos de espacio vectorial

Un modelo de espacio vectorial [55] utiliza vectores de características para modelar los fragmentos de documento. En el caso de plagio en texto las características pueden ser palabras, n-gramas o estadísticas del documento. Para analizar la similitud entre dos

vectores A y B, se puede utilizar el coeficiente de Jacaard 2.2.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{2.2}$$

donde $J(A, B)$ está acotado en [0,1], siendo 1 cuando A y B sean idénticos.

Expansión de vocabulario con tesauros

Existe una clase de modelos que hacen uso de conceptos de tesauros[1] para modelar los documentos. La ventaja de utilizar un tesauro para buscar las palabras de los documentos es que es un buen método para detectar plagio cuando se ha utilizado paráfrasis, ya que en el tesauro podremos encontrar los sinónimos de un concepto y darlo como válido para una palabra dada.

En el trabajo de [26] utilizan la red semántica WordNet [14] para realizar una expansión del vocabulario de acuerdo a las relaciones semánticas de los conceptos originales presentes en un documento. Los autores utilizan seis métodos diferentes para medir la similitud entre los conjuntos de vocabulario expandido de dos documentos dados. Entre los métodos encontramos medir la diferencia de vocabularios y medir la convergencia con un mismo concepto de WordNet.

En el presente trabajo, en la sección 4.2 se hace uso de la red semántica multilingüe BabelNet [33], la cual posee los conceptos de WordNet además de las entradas etiquetadas de la Wikipedia tomadas como conceptos, para realizar una expansión del vocabulario representando los documentos mediante grafos de conocimiento, lo cual es una técnica relacionada con esta categoría pero desde un enfoque translingüe.

Búsqueda de similitudes en grandes colecciones de datos

El método de búsqueda de similitudes en grandes colecciones de datos (BDCSS) [20], fue el ganador de la competición internacional

[1] Tesauro: "Nombre dado por sus autores a ciertos diccionarios, catálogos o antologías." [13]

de detección de plagio PAN'11[2] descrita en la sección 5.1, en su tarea de detección de plagio externo.

El método utiliza un proceso de tres etapas para la detección:

- *Preproceso:* En primer lugar se convierten todos los documentos a texto plano. A continuación se transforma el texto en una estructura de datos más eficiente de menor tamaño y más fácil acceso. Para ello se utiliza un algoritmo de mapeo que traduce todo el texto al inglés (si fuera necesario), lo lematiza y convierte todos los sinónimos en una misma palabra para simplificar la comparación. Además se transforman todas las palabras del texto a códigos binarios.
- *Detección de pasajes sospechosos:* Para detectar los casos de plagio, en primer lugar se utiliza un método para contabilizar el número de palabras N_{MW} en común (sin repeticiones) entre dos fragmentos I_S e I_R, sin importar el órden ni número de palabras de los fragmentos, mediante la ecuación 2.3.

$$N_{MW}(I_S, I_R) = |I_S \cap I_R| \tag{2.3}$$

  Dados dos fragmentos a comparar I_S e I_R, utilizamos los subfragmentos de estos I_{Si} e I_{Rj}, siendo i y j el tamaño del subfragmento comenzando desde su primera palabra. Para que un caso sea plagio tiene que cumplir la ecuación 2.4, siendo $q_{min} = 0{,}5$. Además, cada par de subfragmentos a comparar tienen que cumplir que $N_{MW}(I_{Si}, I_{Rj}) \leqslant 15$.

$$q_{Si} = \frac{N_{MW}(I_S i, I_R)}{N_{MW}(I_S i, I_S i)} \leqslant q_{min} \wedge \frac{N_{MW}(I_S, I_R j)}{N_{MW}(I_R j, I_R j)} \leqslant q_{min} \tag{2.4}$$

  Por último, si dos casos de plagio son adyacentes son unidos en uno solo.

- *Postproceso:* En la última etapa se eliminan los solapamientos de casos de plagio en un mismo fragmento, además de mejorar

[2] `http://www.webis.de/research/events/pan-11`

el *recall* y *precision* estableciendo tres umbrales debajo de los cuales tienen que encontrarse los fragmentos de plagio detectados para ser válidos: t_1 = número de palabras del fragmento $< T_1$ (cuyo valor óptimo establecen en 70), t_2 = número de caracteres de cada palabra $< T_2$ (valor óptimo = 60), y t_3 = número de caracteres de los fragmentos de plagio $< T_3$ (valor óptimo = 200).

Detección de plagio por comparación detallada

El método de búsqueda de detección de plagio por comparación detallada [27], fue el ganador de la competición internacional de detección de plagio PAN'12[3], en su tarea de detección de plagio externo.

Dado un conjunto de documentos sospechosos y fuentes de contener plago, el método utiliza un proceso de tres etapas para la detección:

- *Preproceso:* En esta etapa se eliminan los signos de puntuación, las palabras irrelevantes (como los artículos), los espacios en blanco consecutivos, y se lematiza el texto. Además se ponderan las palabras de los documentos utilizando un método de ponderamiento del software Lucene[4].

- *Comparación detallada:* Este método toma como casos a comparar fragmentos de una frase de longitud, así el primer paso es separar todos los documentos fuente y sospechosos en frases. A continuación se seleccionan las frases cuya distancia del coseno sea superior a un umbral t_1 = 0,42, como se puede ver en la ecuación 2.5.

$$sim(S,R) = \cos\Theta = \frac{\sum_{k=1}^{n} w_{Sk} * w_{Rk}}{\sqrt{(\sum_{k=1}^{n} w_{Sk}^2)(\sum_{k=1}^{n} w_{Rk}^2)}} > t_1, \qquad (2.5)$$

[3] http://www.webis.de/research/events/pan-12
[4] http://lucene.apache.org/

siendo S una frase sospechosa, R una frase fuente, w_{Sk} el peso de la palabra k de la frase S y w_{Rk} el peso de la palabra k de la frase R.

El siguiente paso es filtrar los casos posibles haciendo que su comparación estructural T sea superior a un umbral $t_2 = 0{,}32$, como se puede ver en la ecuación 2.6.

$$T(S,R) = \frac{2 * \sum_{t \in (S \cup R)} min(N_S(t), N_R(t))}{|S| + |R|} > t_2, \qquad (2.6)$$

donde $N_S(t)$ y $N_R(t)$ es el número de veces que se repite el término común t en las frases S y R.

Finalmente se mezclan los casos de plagio que están solapados o adyacentes mediante un algoritmo de ordenación.

- *Postproceso:* En la última fase se vuelve a pasar la ecuación 2.6 sobre los casos de plagio detectados que ya han sido unidos si era necesario, estableciendo $t_2 = 0{,}3$ en esta ocasión.

2.3. Detección automática de plagio textual translingüe

En esta sección trataremos la detección de plagio textual, intrínseco y externo, a nivel translingüe.

2.3.1. Detección de plagio intrínseco translingüe

Si bien a priori prodría parecer que la detección de plagio intrínseco translingüe no tiene sentido, ya que en detección intrínseca se trabaja sobre un solo documento para detectar secciones que no son propias del estilo y características del autor, y por tanto se trabajaría a nivel monolingüe, existen trabajos que detectan secciones de documento que no son propias del estilo del autor al ser traducciones de otros documentos.

En [1] detectan documentos traducidos desde otro lenguaje, con la asunción de que la traducción conservará parte de la esencia de la estructura sintáctica del lenguaje original, y por tanto un conteo de categorías morfosintácticas, nombres y adverbios sería suficiente para reverlar secciones impropias del lenguaje.

En el trabajo descrito en [51] se realiza una clase de detección de plagio intrínseco translingüe. Los autores estudian la detección del uso de traductores online, para traducir documentos que debían ser traducidos a mano, por parte de estudiantes de lenguas.

En general, para detectar plagio intrínseco translingüe, todos los modelos descritos a nivel monolingüe son válidos, ya que son capaces de detectar secciones impropias del estilo del autor, como serán las que provengan de documentos traducidos desde otro lenguaje.

2.3.2. Detección de plagio externo translingüe

De acuerdo a su paradigma de resolución, existen cuatro categorías diferentes 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 tesauros 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 [58] , que utiliza modelos de espacio vectorial [55] o CL-CTS [21], que usa el tesauro conceptual Eurovoc[5] para las traducciones. (ii) Modelos que se basan en la sintaxis del documento y su estructura para comparar los documentos. El modelo CL-CNG [30] está incluido en esta categoría. (iii) Modelos que utilizan corpus comparables, como CL-ESA [43]. Éste utiliza corpus alineados por tema y lenguaje, 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 [4, 2], *Cross-language latent semantic indexing* (CL-LSI) [12] y *Cross-language kernel canonical correlation analysis*

[5] `http://eurovoc.europa.eu/`

(CL-KCCA) [60], quedan dentro de esta categoría. Además, existen modelos que pueden utilizar combinaciones de las categorías anteriores, para lo cual nuestra aproximación CL-KGA, descrita en la sección 4.2, es un buen ejemplo de ello, siendo la red semántica multilingüe BabelNet la unión de (i) un tesauro (WordNet) y de (iii) un corpus comparable (Wikipedia[6]).

A continuación se muestran algunas de las aproximaciones de análisis de similitud translingüe, ordenadas por categoría, que mejores resultados han ofrecido y pueden utilizarse para la tarea de detección de plagio externo.

Análisis de similitud translingüe basado en un tesauro conceptual

El modelo CL-CTS [21], *Cross-language conceptual thesaurus based similarity*, representa los documentos por sus conceptos presentes dentro de un tesauro multilingüe, modelados mediante vectores conceptuales $\vec{c}$. Para asignar un concepto e a un documento d, se tiene que cumplir la regla $v(e,d) > 0$, siendo $v(e,d)$ el peso asignado a e en d y se estima en la ecuación 2.7.

$$v(e,d) = \sum_{t \in e, T_e} f(t,d), \tag{2.7}$$

donde T_e se refiere al vocabulario de conceptos en Eurovoc y $f(t,d)$ representa la frecuencia del término t en el documento d.

Para estimar la similitud $S(d,d')$ entre dos documentos se utiliza la ecuación 2.8.

$$S(d,d') = \frac{\alpha}{2} * \left(\frac{\vec{c_d}.\vec{c_{d'}}}{|d||d'|} + l(d,d') \right) + (1-\alpha) * \zeta(d,d'), \tag{2.8}$$

donde $\vec{c_x}$ representa el vector conceptual de un documento x, $l(d,d')$ es el factor de longitud definido en 2.9, y α es un factor de normalización para que $S(d,d')$ esté acotada en $[0,1]$. $\zeta(d,d')$ es la similitud

[6] www.wikipedia.org

del coseno entre los *3*-gramas a nivel de caracter de los nombres de entidades en d y d', para los cuales se da un tratamiento especial debido al hecho de que no están presentes en el tesauro, pero suponiendo similitudes sintácticas entre lenguajes, se puede cubrir la carencia utilizando n-gramas a nivel de caracter.

$$l(d, d') = exp\left(-0{,}5\left(\frac{|d'|/|d| - \mu}{\sigma}\right)^2\right), \tag{2.9}$$

donde μ y σ son la media y desviación típica de la longitud de caracteres entre traducciones de documentos desde el lenguaje L_1 al lenguaje L_2. Mediante esta fórmula se tiene en cuenta el hecho de que un mismo texto puede ocupar distinto espacio según el lenguaje en el que esté escrito. Más información sobre la ecuación se puede encontrar en [46].

En el trabajo original, para la adquisión de los conceptos, utilizan el terauro Eurovoc, el cual tiene 6.797 conceptos multilingües representados en 22 idiomas.

Análisis de silimitud translingüe basado en n-gramas de caracteres

El modelo CL-CNG [30], *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) [43].

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 2.10:

$$S(d, d') = \frac{\vec{d}.\vec{d'}}{|d|.|d'|}, \tag{2.10}$$

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

Análisis de similitud translingüe basado en corpus comparable

El modelo CL-ESA [43], *Cross-language explicit semantic analysis*, se basa en el modelo ESA [17], *explicit semantic analysis*, el cual representa dos documentos a comparar d y d' por sus similitudes con un conjunto de documentos conocido como colección de índices C_I. Las similitudes entre los documentos son representadas mediante los vectores $\vec{d}$ y $\vec{d'}$ estimados como en la ecuación 2.11.

$$\vec{d} = \{sim(d, c) \forall c \in C_I\}, \tag{2.11}$$

donde $sim(d, c)$ computa las similitudes entre el documento d y el documento c de la colección C_I, siendo el vector resultante $\vec{d}$ de la siguiente forma:

$$\vec{d} = [sim(d, c_0), sim(d, c_1), ..., sim(d, c_n)]$$

El análisis de similitud entre vectores se lleva a cabo con un modelo de detección de similitud monolingüe, como los modelos de espacio vectorial [55] definidos en la sección 2.2.2.

En un contexto translingüe, CL-ESA, para determinar las semejanzas entre documentos en diferentes lenguajes, necesita un corpus comparable multilingüe alineado por tema y lenguaje en el que exista una correspondencia $\{C_I, C'_I\}(C_I \in L_1, C'_I \in L_2)$. Generalmente con este modelo se utiliza el corpus comparable de la enciclopedia de la Wikipedia.

Análisis de similitud translingüe basado en alineamiento

El modelo CL-ASA, *Cross-language alignment-based similarity analysis*, mide la similitud entre dos documentos d y d', en dos idiomas diferentes L_1 y L_2, alineándolos 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 2.12:

$$S(d, d') = l(d, d') * t(d|d'), \tag{2.12}$$

donde $l(d, d')$ es el factor de longitud definido en 2.9 y $t(d|d')$ es el modelo de traducción definido en la ecuación 2.13.

$$t(d|d') = \sum_{x \in d} \sum_{y \in d'} p(x, y), \tag{2.13}$$

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 los experimentos del capítulo 5.2.2 se ha entrenado un diccionario estadístico alemán-inglés y español-inglés haciendo uso del modelo de alineamiento de palabras IBM M1 [7, 37], sobre el corpus paralelo multilingüe JRC-Acquis [59], además de probar también el diccionario estadístico de la red semántica multilingüe BabelNet, como se explica en la sección 4.1.

Dejando de lado aproximaciones como CL-LSI y CL-KCCA que ofrecen un buen rendimiento a un alto coste computacional, existen trabajos [43, 21] 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 el capítulo de evaluación 5, con nuestros modelos.

Capítulo 3

Redes semánticas

3.1. Red semántica

Una red semántica es una forma de representación de conocimiento lingüístico en la que los conceptos y sus interrelaciones se representan mediante un grafo, también conocido como base de conocimiento de la red. Las redes semánticas han sido muy utilizadas en Inteligencia Artificial para representar el conocimiento, utilizándolas, entre otras cosas, para representar mapas conceptuales y mentales, e incluso para modelar tesauros lingüísticos completos mediante un grafo de grandes dimensiones.

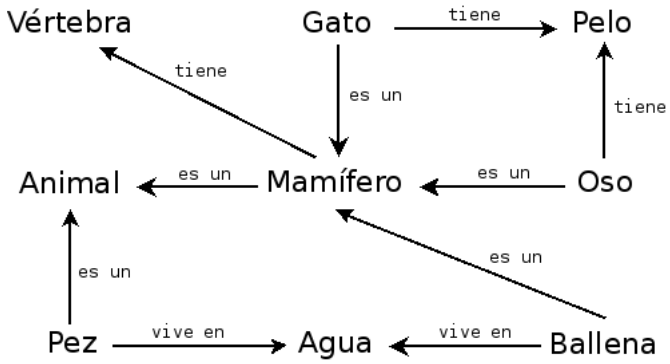

Figura 3.1: Ejemplo de red semántica sobre el mundo animal.

En una base de conocimiento o red semántica, los elementos semánticos (conceptos) se representan mediante nodos. Dos elementos semánticos entre los que exista algún tipo de relación, estarán

unidos en la red mediante una arista (relación). Dependiendo del conocimiento que se esté representando en la red, los grafos podrán ser dirigidos, de modo que existan relaciones no simétricas entre conceptos. Además, atendiendo a la clase de contenido de la red, habrá diferentes tipos de relaciones, siendo ejemplos de posibles relaciones la hiponimia, hiperonimia, la meronimia, etc. En la figura 3.1 tenemos un ejemplo de red semántica. Existen diversas redes semánticas que modelan tesauros lingüísticos, como por ejemplo WordNet.

En el año 1985, el profesor de psicología George A. Miller, del *Cognitive Science Laboratory* de la Universidad de Princeton, comenzó la dirección del proyecto WordNet, el cual actualmente es una enorme base de datos léxica del idioma inglés. WordNet agrupa las palabras en conjuntos de sinónimos llamados 'synsets', proporcionando definiciones cortas y generales, y almacenando las relaciones semánticas entre estos conjuntos de sinónimos. El propósito del proyecto es doble: por un lado producir una combinación de diccionario y tesauro cuyo uso es más intuitivo, y ayudar al análisis automático de textos y a las aplicaciones de Procesamiento del Lenguaje Natural. Además, la base de datos y las herramientas pueden ser descargadas y usadas libremente[1]. Por otro lado, la base de datos puede consultarse online en su sitio Web[2].

El proyecto original de WordNet, para la lengua inglesa, se ha extendido a través de la Asociación Global de WordNet[3], la cual promueve la creación y unión de diferentes WordNets de lenguas y países del mundo. Además de WordNet, existen otras redes semánticas que modelan tesauros, como por ejemplo la que utiliza los modelos de Gellish[4] para crear la taxonomía-diccionario inglés Gellish, la cual es una gran red semántica, de código abierto, que representa y relaciona conceptos de la lengua inglesa.

Existe otra categoría de redes semánticas, conocidas como redes semánticas multilingües, las cuales fundamentan el trabajo de investigación de este informe, y de las que hablaremos a continuación.

[1] `http://wordnet.princeton.edu/wordnet/download/`
[2] `http://wordnetweb.princeton.edu/perl/webwn`
[3] `http://www.globalwordnet.org/`
[4] `http://gellish.sourceforge.net/`

3.2. Red semántica multilingüe

Una red semántica multilingüe sigue el esquema de una base de conocimiento tradicional, y por tanto el descrito en la sección anterior. La diferencia principal respecto a una red semántica es que cada uno de los nodos del grafo, tiene una dimesión multilingüe, de modo que existe un conjunto de lexicalizaciones del concepto en diferentes idiomas.

Actualmente están disponibles para su uso diferentes redes semánticas multilingües, como por ejemplo ConceptNet [23], EuroWordNet [61], el cual realiza una unión de todos los 'synset' equivalentes entre los WordNet de la Unión Europea, o BabelNet [33], el cual hemos elegido como red semántica multilingüe para la realización de nuestra experimentación, y del que hablaremos a continuación.

3.2.1. BabelNet

BabelNet [33] está formado 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.

Las relaciones y conceptos provienen de WordNet, la mayor red semántica disponible, y de las entradas multilingüe de la Wikipedia, así BabelNet combina información lexicográfica con conocimiento enciclopédico. La lista de conceptos está formada por todos los significados de palabra en WordNet ('synsets') y las etiquetas de las entradas 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.

Las lexicalizaciones multilingües se obtienen a partir de las entradas en diferentes idiomas de la Wikipedia, utilizándolas como corpus comparable y combinándolo con traducción estadística. Por otro lado, se utiliza el corpus semCor[5], el cual está compuesto por frases en inglés con los conceptos de WordNet etiquetados, de modo

[5] http://www.gabormelli.com/RKB/SemCor_Corpus

que mediante un traductor estadístico podemos obtener las traducciones de dichos conceptos en los diferentes lenguajes de BabelNet. Por último, cabe señalar que cada uno de los conceptos dentro de la base de conocimiento tiene una de las siguientes categorías gramaticales asociadas: adjetivo, adverbio, nombre y verbo. En la figura 3.2 tenemos un esquema de la estructura interna de BabelNet.

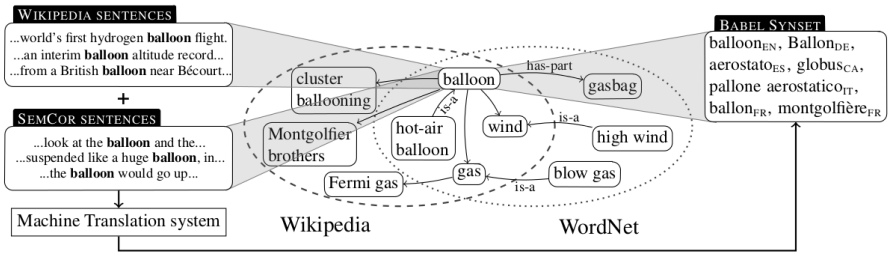

Figura 3.2: Ejemplo de la estructura interna de BabelNet (figura tomada de [33]).

BabelNet puede ser explorado de forma gráfica [35] utilizando la herramienta online BabelNetXplorer[6]. El API[7] de BabelNet permite utilizar esta red semántica multilingüe, entre otros usos, como traductor, como desambiguador del sentido de las palabras [36], como diccionario estadístico y para construir grafos de conocimiento. Estos dos últimos usos son los que utilizamos en nuestra investigación y de los que hablaremos a continuación.

Diccionario estadístico de BabelNet

Dada una palabra x escrita en un lenguaje $A \in L = \{CA, DE, EN, ES, FR, IT\}$, el diccionario estadístico de BabelNet 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.

Las traducciones y sus pesos se toman de la base de conocimiento de BabelNet del siguiente modo: (i) para una palabra x_A dada en

[6] `http://lcl.uniroma1.it/bnxplorer/`
[7] `http://lcl.uniroma1.it/babelnet/`

un lenguaje A, buscamos dentro de la base de conocimiento la lista de conceptos $C_A = \{x_A \cup \text{equivalentes}(x_A)\}$ en el lenguaje A, donde equivalentes(x_A) son los conceptos que guardan una relación de equivalencia de algún tipo (sinonimia, pertenencia, parte-de...) con x_A dentro de la base de conocimiento; (ii) para cada concepto de C_A devolvemos la lista de traducciones C_L disponibles en su dimensión multilingüe para cada uno de los lenguajes L; (iii) el peso w_L de cada traducción x_L será el número de relaciones salientes de cada uno de sus conceptos asociados que se dirija a otro concepto de la lista C_L. Utilizando este método las traducciones más probables serán las de mayor peso (las más relacionadas), si bien no están normalizadas.

Para comprender mejor como funciona el diccionario estadístico vamos a poner un ejemplo: dada la palabra x_{en} = "*car*" en el lenguaje inglés, el sistema busca la lista de conceptos iguales o equivalentes C_{en} = {*car, auto, motor, motorcar, cab...*}$_{en}$. A continuación obtenemos la lista de traducciones C_L = {{*coche, automovil, auto, carro...*}$_{es}$,{*auto, wagen...*}$_{de}$,...} a todos los lenguajes L de BabelNet utilizando la dimensión multilingüe de cada uno de los conceptos de C_{en}. Finalmente obtenemos los pesos de cada una de las traducciones dentro de C_L contando el número de relaciones que existen dentro de la base de conocimiento de BabelNet entre cada uno de los conceptos de C_L para cada lenguaje L, devolviendo la lista de traducciones ponderadas C_L = {{*(coche, 8), (automovil,6), (auto, 5), (carro, 3)...*}$_{es}$,{*(auto, 9), (wagen, 6)...*}$_{de}$,...}. Notese que no tienen porqué ser los mismos pesos para cada lenguaje, ya que se puede dar el caso de que algún concepto tenga más de una traducción equivalente en un mismo synset, lo cual será contabilizado como dos conceptos diferentes que están igualmente relacionados.

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 su número

de relaciones salientes, mientras que el peso de una relación es el peso original de la relación en la base de conocimiento de BabelNet, la cual parte de la base de conocimiento de WordNet a la que se le han añadido los conceptos de la Wikipedia utilizando un algoritmo de *mapping* [34].

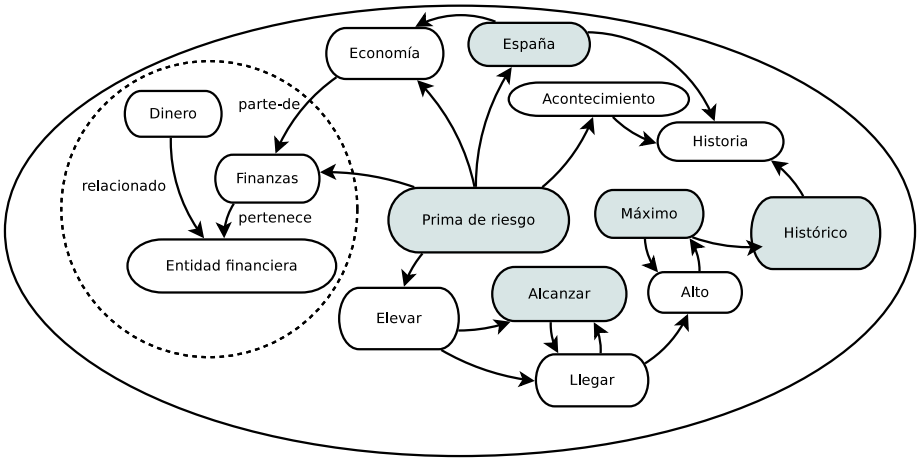

Figura 3.3: Ejemplo simplificado, sin pesos en nodos y aristas, ni dimensión multilingüe, del grafo de conocimiento de la frase "la prima de riesgo española alcanza máximos históricos" (los nombre de las relaciones se incluyen sólo dentro de la línea discontinua).

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:

$$\text{C} = \{\text{prima de riesgo, España, alcanzar, máximo, histórico}\}$$

Utilizando las herramientas de 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:

$$C' = \{\text{economía, finanzas, historia...}\}$$

Además, entre los conceptos de C_g, aparecerán una serie de co-

nexiones R que los relacionan:

$R \in$ {relacionado, parte de, pertenece, equivalente, opuesto...}

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, lo cual explotaremos en la sección 4.2 para generar nuestro modelo de análisis de similitud translingüe. En la fig. 3.3 podemos ver el contenido del grafo g en español.

Un grafo de conocimiento se genera en BabelNet del siguiente modo: (i) dada una lista de conceptos etiquetada morfológicamente, buscamos caminos para cada uno de ellos, dentro de la base de conocimiento de BabelNet, hasta el resto de conceptos de la lista[8]; (ii) una vez tenemos todos los caminos entre conceptos, estos son suministrados a un constructor de grafos que los fusiona generando un grafo de conocimiento, el cual posee en su interior conceptos y relaciones expandidos respecto a la lista de conceptos original.

El constructor de grafos que utiliza BabelNet es una clase en Java que, dado un conjunto P de caminos entre conceptos, los fusiona generando un grafo g del siguiente modo: (i) en primer lugar tomamos todos los conceptos presentes en la lista de caminos P y realizamos su unión C para eliminar repeticiones; (ii) tomamos todas las relaciones entre conceptos presentes en P y realizamos su unión R para eliminar repeticiones; (iii) finalmente conectamos los conceptos de C mediante las relaciones de R obteniendo el grafo $g = \{C\ x\ R\}$.

[8]La búsqueda viene delimitada por una profundidad máxima (tres conceptos de profundidad por defecto) y puede filtrar, entre otras cosas, conceptos de longitud inferior a determinado tamaño.

Capítulo 4

Modelos propuestos

En este capítulo vamos a proponer dos modelos de análisis de similitud para uso en detección de plagio externo translingüe. En primer lugar, en la sección 4.1 sobre el modelo CL-ASA como base, utilizaremos el diccionario estadístico de BabelNet, aplicando diferentes métodos de normalización sobre los pesos de las traducciones de palabras. A continuación, en la sección 4.2 presentaremos el nuevo modelo de análisis de similitud basado en grafos de conocimiento, CL-KGA, que expande y relaciona los conceptos originales de fragmentos de texto, para proporcionar un modelo de contexto de su contenido y realizando la comparación de fragmentos de texto a nivel de grafos.

4.1. CL-ASA con el diccionario estadístico de BabelNet

Como hemos comentado en la sección 3.2.1, 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 2.13 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 L_1 y una palabra y en un lenguaje L_2, 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 lenguaje L_2, para dar lugar a una probabilidad de traducción $p(x, y)$:

$$p(x, y) = \frac{w(x, y)}{w(x)} \tag{4.1}$$

- **Normalización de la similitud *S(d,d')***: Dado un documento d con un número total de palabras $|d|$, podemos normalizar su similitud respecto al número de palabras del documento fuente del siguiente modo:

$$S_{norm}(d|d') = \frac{S(d, d')}{|d|} \tag{4.2}$$

Los métodos de normalización descritos en las ecuaciones 4.1 y 4.2 son compatibles entre sí y combinados con la ecuación 2.13 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:

$$S_{full\_normalization}(d, d') = \frac{\sum_{x \in d} \sum_{y \in d'} p(x, y)}{|d|} \tag{4.3}$$

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

$$S_{weight\_normalization}(d, d') = \sum_{x \in d} \sum_{y \in d'} p(x, y) \tag{4.4}$$

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

$$S_{size\_normalization}(d, d') = \frac{\sum_{x \in d} \sum_{y \in d'} w(x, y)}{|d|} \tag{4.5}$$

- **BN-dict**$_4$: Sin normalización:

$$S_{no\_normalization}(d, d') = \sum_{x \in d} \sum_{y \in d'} w(x, y) \tag{4.6}$$

El método descrito ha sido publicado en [15]. En nuestra evaluación, en la sección 5.2.2, compararemos los cuatro métodos propuestos con los modelos estado del arte CL-CNG y CL-ASA, además de con nuestra otra propuesta, CL-KGA, de la que hablaremos a continuación.

4.2. Análisis de similitud basado en grafos de conocimiento

En esta sección vamos a presentar el modelo *Cross-language knowledge graphs analysis* (CL-KGA), el cual utiliza grafos de conocimiento generados a partir de una red semántica multilingüe 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 $d' \in D'$ utilizando grafos de conocimiento, debemos de seguir los siguientes pasos:

- En primer lugar debemos realizar un procesado previo del texto para extraer y etiquetar morfológicamente sus conceptos, eliminando también los tipos de palabra que no aportan mucha información útil para la detección de plagio, como son los

artículos. Además, es conveniente lematizar las palabras. Para estas tareas en nuestra investigación hemos hecho uso de la herramienta TreeTagger[1] [49], la cual es compatible con diferentes idiomas entre los cuales se encuentran el alemán, español e inglés, que son los que utilizamos en nuestra evaluación. Cabe señalar que como en BabelNet solo existen cuatro categorías gramaticales (adjetivo, adverbio, nombre y verbo), todas las categorías etiquetadas por TreeTagger que se salgan de estas, serán re-etiquetadas como nombres, la categoría más común en BabelNet.

- Una vez realizado el preprocesamiento del texto, podemos construir, utilizando la red semántica multilingüe BabelNet, los grafos de conocimiento g y g' a partir de los documentos d y d'.
- Finalmente, para obtener una similitud $S(g, g')$ entre g y g', tomando como base la aproximación de comparación flexible de grafos conceptuales[2] [32], hemos propuesto la ecuación 4.7 para trabajar con grafos de conocimiento.

$$S(g, g') = S_c(g, g') * (a + b * S_r(g, g')) \tag{4.7}$$

$$S_c(g, g') = \frac{\left(2 * \sum_{c \in g_{int}} w(c)\right)}{\left(\sum_{c \in g} w(c) + \sum_{c \in g'} w(c)\right)} \tag{4.8}$$

$$S_r(g, g') = \frac{\left(2 * \sum_{r \in N(c, g_{int})} w(r)\right)}{\left(\sum_{r \in N(c,g)} w(r) + \sum_{r \in N(c,g')} w(r)\right)} \tag{4.9}$$

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

[2] Un grafo conceptual es un grafo finito dirigido bipartido con dos clases de nodos: conceptos y relaciones [52, 53].

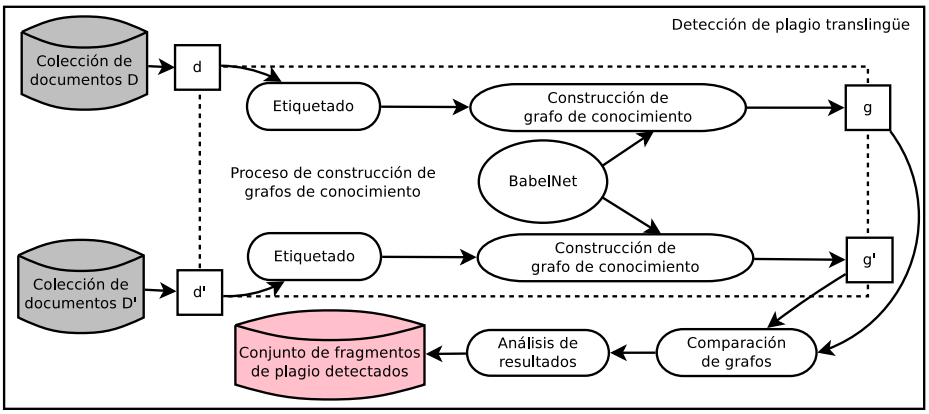

Figura 4.1: Proceso de detección de plagio translingüe utilizando grafos de conocimiento.

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_{int} 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 4.7 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, aunque del modo que están dispuestas las variables en la ecuación los conceptos siempre serán más importantes que las relaciones, debido a que estas últimas no podrían existir sin los conceptos. En la sección 5.2.1 analizaremos cuales son los porcentajes de relevancia adecuados para conceptos y relaciones en detección de plagio translingüe utilizando BabelNet.

En la fig. 4.1 podemos ver un esquema completo del proceso de detección de plagio translingüe utilizando grafos de conocimiento. Es importante señalar que después de la intersección de grafos $g_{int} = (g \cap g')$, los conceptos y relaciones del grafo g_{int} probablemente se hayan visto reducidos, y por tanto, sus pesos tendrán que

ser recalculados. El cálculo del peso de un concepto es trivial, pues es el número de relaciones salientes del mismo. En cambio, 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 siguiendo su algoritmo de *mapping* entre conceptos de la Wikipedia a WordNet[3]. Por ese motivo, en la ecuación 4.10 proponemos un algoritmo genérico de reestimación del peso $w(r, c, g_{int})$, siendo r una relación saliente de un concepto c en el grafo de intersección g_{int}. El nuevo peso se calcula en función del antiguo y del nuevo valor del peso de c en los grafos g, g' y g_{int}.

$$w(r, c, g_{int}) = \frac{w(c, g) * d(c, g, g_{int}) + w(c, g') * d(c, g', g_{int})}{2} \quad (4.10)$$

$$d(x, g_1, g_2) = \frac{|R(g_1, x)|}{|R(g_2, x)|} \quad (4.11)$$

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.

El método descrito ha sido publicado en [16]. En la evaluación de la sección 5.2.2 compararemos el uso de grafos de conocimiento, en detección de plagio translingüe, con los métodos propuestos utilizando el diccionario estadístico de BabelNet, además de con los modelos CL-ASA (utilizando un diccionario entrenado con el modelo IBM M1) y CL-CNG.

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

Capítulo 5

Evaluación

Este capítulo está dedicado a la evaluación de los modelos propuestos en el capítulo 4. En primer lugar, en la sección 5.1 describiremos el corpus PAN-PC'11, el cual utilizaremos en nuestros experimentos. A continuación, en la sección 5.1.1 explicaremos las unidades de medida que vamos a emplear para medir la calidad de nuestros resultados. El método de análisis detallado de similitud es explicado en la sección 5.1.2. Por último, una vez descritos todos los detalles necesarios para comprender correctamente la evaluación que se va a llevar a cabo, en la sección 5.2 se realizarán diferentes experimentos para evaluar nuestros modelos frente al estado del arte en detección de plagio externo translingüe.

5.1. Corpus PAN-PC'11

Desde el año 2009, en el marco del PAN *Uncovering Plagiarism Authorship and Social Software Misuse* (PAN)[1], se celebra una competición internacional sobre detección de plagio. En la edición del año 2013 se presentan y ponen a prueba aproximaciones en los siguientes ámbitos: detección de plagio, identificación del autor (dado un documento, ¿quien lo ha escrito?) y *profiling* del autor (dado un documento, ¿qué rasgos caracterizan a su autor?). Para los experimentos de nuestra evaluación, vamos a utilizar el corpus de su edición del año 2011: el PAN-PC'11, debido a que es el corpus que se

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

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

Tabla 5.1: Estadísticas de la tarea de detección de plagio externo translingüe del corpus PAN-PC'11

ha utilizado para la comparación de los modelos CL-ASA, CL-CNG y CL-CTS en [21].

Del corpus PAN-PC'11, tomamos las particiones español-inglés (es-en) y alemán-inglés (de-en) por las cuales está formada su tarea de detección de plagio translingüe externo, la cual se describe en la sección 2.3.2. En la tabla 5.1 podemos ver las estadísticas de los documentos que forman la partición.

5.1.1. Unidades de medida

Para medir la calidad de los resultados experimentales vamos a tomar las medidas utilizadas en la competición del PAN:

- *recall* a nivel de caracter. La medida nos da un porcentaje de la cantidad de fragmentos de plagio correctos detectados respecto al total existente en el corpus. La fórmula atiende al hecho de que se trabaja con fragmentos de texto de los cuales podemos detectar solo una parte, por ello la clásica fórmula $tp/(tp + fn)$ es reescrita, para trabajar a nivel de caracter, del siguiente modo:

$$recall(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R}(s \sqcap r)|}{|s|},$$

siendo

$$s \sqcap r = \begin{cases} s \cap r & si\ r\ detecta\ s. \\ \emptyset & en\ otro\ caso. \end{cases}$$

donde S es el conjunto de casos de plagio existentes en el corpus, R es el conjunto de casos de plagio detectados en el corpus, s es un caso de plagio de S y r es un caso de plagio de R.

- *precision* a nivel de caracter. La medida da un porcentaje que mide, de los casos de plagio que se han detectado, cuantos eran verdaderamente plagio. La fórmula clásica $tp/(tp + fp)$ se modifica para trabajar al nivel de caracter del siguiente modo:

$$precision(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S}(s \sqcap r)|}{|r|},$$

- *granularity*. La medida sirve para detectar el hecho de que en ocasiones los detectores solapen o deporten multiples detecciones para un mismo caso de plagio, lo cual no es posible determinar mediante *precision* y *recall*. La fórmula es la siguiente:

$$granularity(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s|,$$

donde S_R es el conjunto de casos correctamente detectados en el corpus y R_s es una ocurrencia de s dentro del conjunto de casos de plagio detectados R. Dado el modo en que está concebida la fórmula, el valor óptimo de *granularity* será 1, y este irá aumentando en caso de que se produzcan solapamientos o fragmentaciones de casos de plagio en la detección.

- *plagdet*. Con el fin de obtener una medida de detección de plagio global de un detector sobre un curpus, se combinan las tres medidas anteriores para dar lugar al *plagdet*, el cual se calcula mediante la siguiente fórmula:

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

donde F_1 es la media armónica de *precision* y *recall* ponderadas equitativamente. [2]

[2] Una descripcción más detallada de las medidas se puede encontrar en [42].

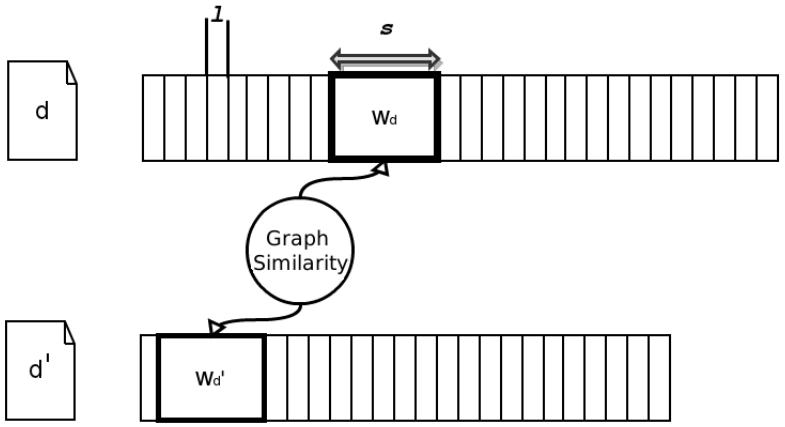

Figura 5.1: Esquema del proceso de comparación de fragmentos de texto entre dos documentos d y d' utilizando ventana deslizante, donde s es el tamaño de la ventana (5 en nuestro caso), y w_d es un fragmento de texto del documento d.

5.1.2. Análisis detallado de similitud

Hay que tener en cuenta, que todos los modelos de detección de plagio descritos, tanto en el estado del arte de la sección 2.3.2 como en nuestras propuestas del capítulo 4, devuelven un valor de similitud entre pares de fragmentos de texto, y ese valor de similitud todavía no será un indicativo de que exista plagio hasta que se puedan analizar todos los valores de similitud obtenidos entre cada par de fragmentos de texto comparados.

En esta sección vamos a explicar el método de análisis detallado de similitud y post-procesamiento heurístico utilizado en la competición del PAN, el cual también utilizaremos en nuestros experimentos, mediante el que se deciden los fragmentos de texto que son plagio una vez se tienen todos valores de similitud.

Dado un conjunto de documentos fuente D en el lenguaje L_1 y un conjunto de documentos sospechosos D' en el lenguaje L_2, para poder comparar diferentes fragmentos de documento obteniendo una similitud entre ellos, utilizamos una ventana deslizante de cinco frases de longitud (lo cual consideramos aproximadamente un párrafo) y salto de una una frase, sobre pares de documentos (d,d'), $d \in D$ y $d' \in D'$, y detectamos plagio sobre ellos con los modelos de detección de plagio translingüe deseados. En la figura 5.1 podemos ver

un esquema de como se utiliza la ventana deslizante para comparar fragmentos de documento.

Una vez se hayan comparado todos los pares de documentos en el corpus, para cada documento d tendremos un fichero que contendrá un valor de similitud entre cada uno de sus fragmentos de texto, y cada uno de los del resto de documentos comparados. Para determinar qué fragmentos son plagio, atendiendo a sus valores de similitud, en primer lugar nos quedamos con los primeros 50 documentos d' más similares para cada documento d. A continuación, para cada fragmento s, $s \in d$, nos quedamos con los cinco fragmentos de texto más similares s', $s' \in d'$. Si la distancia entre dos fragmentos de texto de un mismo documento es inferior a un umbral $thres_1$, los unimos (en nuestro caso $thres_1 = 1500$). Finalmente, para decidir que fragmentos contienen plagio, se considera que solo aquellos fragmentos p que estén compuestos por un mínimo de $thres_2$ (en nuestro caso $thres_2 = 3$) fragmentos son plagio. En la figura 5.2 podemos ver el algoritmo completo de análisis detallado y post-procesamiento heurístico.

1: Given d **and** D'**:**
// **Detailed analysis**
2: $S \leftarrow \{split(d, w, l)\}$ $S' \leftarrow \{split(d', w, l)\}$
3: for every $s \in S$:
4: $P_{s,s'} \leftarrow argmax^5_{s' \in S'} sim(s, s')$
// **Post-processing**
5: until no change:
6: for every combination of pairs $p \in P_{s,s'}$:
7: if $\delta(p_i, p_j) < thres_1$:
8: $merge\_fragments(p_i, p_j)$
// **Output**
9: return $\{p \in P_{s,s'} \mid |p| > thres_2\}$

Figura 5.2: **Análisis detallado de similitud y post-procesamiento**, donde $split(d, w, l)$ corta d en fragmentos de longitud w con salto l. $argmax^5_{s' \in S'} sim(s, s')$ devuelve los 5 fragmentos más similares $s \in S$ con respecto a s'. $\delta(p_i, p_j)$ mide la distancia, en caracteres, entre los fragmentos fuentes y suspechosos p_i y p_j. $merge\_fragments(p_i, p_j)$ une los fragmentos de plagio p_i, y p_j. $thres_1$ representa la distancia máxima permitida entre p_i y p_j para que sean unidos. $thres_2$ es el mínimo número de fragmentos de los que tiene que estar compuesto p para ser considerado plagio.

5.2. Experimentos

En esta sección vamos a realizar los experimentos para evaluar la calidad de los modelos propuestos en el capítulo 4. En primer lugar, en la sección 5.2.1 vamos a realizar unos experimentos para determinar cual es la relación de porcentajes adecuados para los valores de relevancia de conceptos y relaciones con el modelo CL-KGA. A continuación, en la sección 5.2.2 compararemos los modelos propuestos con dos de los modelos estado del arte en detección de plagio translingüe, CL-C3G (CL-CNG con trigramas) y CL-ASA, además de compararnos contra el propio ganador de la competición de detección de plagio PAN del año 2011, de la cual utilizamos el corpus.

5.2.1. Valores de relevancia de conceptos y relaciones

En esta sección vamos a comparar el rendimiento del modelo CL-KGA utilizando la red semántica multilingüe 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 la partición 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

Tabla 5.2: Relevancia de conceptos y relaciones en el modelo CL-KGA utilizando BabelNet.

En vista de los resultados de la tabla 5.2, podemos deducir que las relaciones utilizando la red semántica multilingüe BabelNet son prácticamente igual de importantes que los conceptos para es-en,

mientras que para de-en tienen poca o ninguna importancia[3]. La diferencia puede estar producida por unos conceptos muy conectados en grafos es-en, mientras que en de-en podemos estar teniendo un problema de falta de contexto en los grafos, lo cual se produce si los conceptos que provienen de un lenguaje aglutinativo como el alemán, no están preprocesados adecuadamente, y por tanto, al buscarlos dentro de BabelNet no los encuentra todos. Si este efecto se produce de forma muy acusada, se acaba perdiendo el contexto de las frases, del mismo modo que ocurriría si se le suministrasen a una persona los conceptos incompletos de una oración. En los experimentos que realizaremos a continuación investigaremos más profundamente este fenómeno.

Para nuestros siguientes experimentos tomaremos las configuraciones de conceptos y relaciones que mejor han funcionado para ambas particiones.

5.2.2. Detección de plagio externo translingüe

En este experimento vamos a comparar los modelos decritos en el capítulo 4, CL-KGA [16] y CL-ASA utilizando el diccionario estadístico de BabelNet [15] con diferentes normalizaciones (BN-dict$_1$,BN-dict$_2$,BN-dict$_3$ y BN-dict$_4$), con los modelos estado del arte CL-C3G y CL-ASA (con un diccionario entrenado con el modelo IBM M1) descritos en la sección 2.3.2, para las particiones completas es-en y de-en del corpus del PAN-PC'11. Además, despues de comparar nuestros modelos con el estado del arte, también compararemos nuestros resultados con los del ganador de la edición del PAN 2011.

En la tabla 5.3 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 diccionario de BabelNet normalizando los pesos de las traducciones, BN-dict$_1$ y BN-dict$_2$,

[3]Tengase en cuenta que como se describió en 4.2, del modo que está concebida la fórmula de similitud de grafos, los conceptos siempre tendrán más relevancia que las relaciones, aunque los porcentajes aquí sean los mismos

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-KGA	**0.594**	**0.518**	**0.706**	**1.008**
CL-ASA (BN-dict$_4$)	0.563	0.499	0.662	1.015
CL-ASA (BN-dict$_3$)	0.554	0.491	0.663	1.030
CL-ASA (IBM M1)	0.517	0.448	0.689	1.071
CL-ASA (BN-dict$_2$)	0.264	0.205	0.518	1.160
CL-ASA (BN-dict$_1$)	0.254	0.198	0.458	1.132
CL-C3G	0.170	0.128	0.617	1.372

Tabla 5.3: Resultados de la detección de plagio translingüe es-en

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 utilizando diccionario), ha obtenido un 10.31 % más de *plagdet*, lo cual asociado a los otros valores de *recall*, *precision* y *granularity* 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. Finalmente, podemos observar como CL-KGA, utilizando grafos de conocimiento, 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 (BN-dict$_4$)- 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.

En la tabla 5.4 tenemos los resultados de detección de plagio

Modelo	Plagdet	Rec.	Prec.	Gran.
CL-KGA	**0.514**	**0.443**	**0.631**	**1.018**
CL-ASA (IBM M1)	0.406	0.344	0.604	1.113
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 (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-C3G	0.078	0.047	0.330	1.089

Tabla 5.4: Resultados de la detección de plagio translingüe de-en

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 utilizando diccionario 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, al igual que en los experimentos de la sección 5.2.1. 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 5.5, 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 %,

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

Tabla 5.5: Estadísticas del uso de los diccionarios

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 al ser aglutinativo, extrayendo los lemas de las palabras con un tagger específico[4], para poder aumentar el número de coincidencias dentro de BabelNet. Finalmente, en la tabla 5.4 vemos también unos buenos resultados para de-en en nuestro modelo CL-KGA, el cual ha superado a todos los otros, superando a 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.

Llegados a este punto, observando que el corpus PAN-PC'11 existen casos de plagio creados mediante traducción automática (auto.), y casos creados mediante traducción automática además de corrección manual (man.), los cuales serán más complicados de detectar a causa de posibles paráfrasis empleadas, vamos a realizar unos experimentos midiendo el *recall* y la *precision* para observar cómo se portan los modelos según el caracter de las traducciones.

Modelo	de-en				es-en			
	Recall		Precision		Recall		Precision	
	auto.	man.	auto.	man.	auto.	man.	auto.	man.
CL-KGA	**0.538**	**0.247**	**0.698**	**0.098**	**0.601**	**0.221**	**0.774**	**0.098**
CL-ASA (M1)	0.538	0.126	0.642	0.041	0.596	0.180	0.741	0.068
CL-ASA (BN-d*)	0.472	0.092	0.631	0.033	0.599	0.198	0.720	0.076

Tabla 5.6: Resultados de detección de plagio separados por lenguajes y tipo de traducción del caso de plagio. *Nota: CL-ASA (BN-d*) es el mejor método con diccionario de BabelNet para cada partición de lenguajes: BN-dict$_3$ para de-en y BN-dict$_4$ para es-en.*

En la tabla 5.6 podemos ver separados por lenguajes y tipo de traducción en la creación del caso de plagio, los resultados de *precision* y *recall* para detección de plagio translingüe. Vemos como los casos

[4]En nuestros experimentos utilizamos la herramienta TreeTagger para extraer y etiquetar conceptos.

de traducción automática obtienen los mejores valores con todos los modelos, lo cual es lógico si tenemos en cuenta que se han generado de forma automática por una máquina de traducción estadística, sin realizar ningún tipo de paráfrasis o resumen. Además cabe señalar que el número de casos con traducción automática es diez veces superior a los de corrección manual, lo cual puede haber influido, ya que estaríamos ante muy pocos casos de plagio manual a encontrar en un corpus muy grande en comparación, y por tanto el detector devolverá pocos valores de similitud "relevantes" cuyos valores se verán paliados por el resto de valores de similitud "no relevantes", ya que según el método de análisis detallado de la sección 5.1.2, tomamos los primeros 50 fragmentos más similares por documento. Cabe señalar como CL-KGA ha superado en un 96 % el *recall* de CL-ASA (IBM M1) en casos de traducción manual para de-en, y un 20 % para es-en. Por otro lado, la *precision* ha sido un 139 % superior para de-en y un 44 % para es-en comparando el CL-KGA con el CL-ASA (IBM M1). Los métodos utilizando el diccionario estadísico de BabelNet han sido ligeramente mejores para es-en, pero como se había visto en los experimentos anteriores, no se ha acercado a los mejores en de-en.

Por último, una vez hemos compados nuestros modelos con los modelos estado del arte, solo nos queda compararnos con el ganador de la competición internacional PAN 2011 en su tarea de detección de plagio externo translingüe: BDCSS[5] descrito en la sección 2.2.2. La comparación con BDCSS se realiza a continuación del resto ya que en el *overview* no se pueden observar los resultados separados por particiones según el lenguaje, y en su lugar se les ha aplicado un promedio.

En la tabla 5.7 tenemos los resultados promediados de las particiones de-en y es-en de nuestros modelos[6] frente al estado del arte y al ganador del PAN-PC'11 en su tarea de detección de plagio externo translingüe. Cabe señalar que BDCSS ha superado al CL-KGA en *recall* un 18 %, *precision* un 23 % y *granularity* un 1 %, lo cual ha

[5]Más información sobre los resultados de la competición se puede encontrar en el *overview* [44]

[6]Solamente hemos promediado el método utilizando el diccinario estadístico de BabelNet que mejores resultados ha ofrecido en promedio: BN-dict$_3$.

Modelo	Plagdet	Rec.	Prec.	Gran.
BDCSS	**0.64**	**0.57**	**0.83**	**1.00**
CL-KGA	0.55	0.48	0.67	1.01
CL-ASA (IBM M1)	0.46	0.40	0.65	1.09
CL-ASA (BN-dict$_3$)	0.42	0.36	0.63	1.10
CL-C3G	0.12	0.09	0.47	1.23

Tabla 5.7: Resultados de la detección de plagio translingüe en promedio comparados con el ganador del PAN-PC'11: BDCSS. *Nota: redondeamos a dos decimales como en el overview del PAN'11 [44].*

llevado a un *plagdet* un 16 % superior. A pesar de que CL-KGA no haya superado al ganador de la competición, su rendimiento en comparación es notable, ya que para la competición del PAN se dispone con anterioridad del corpus de la competición, además de conocer las herramientas de traducción estadística que se utilizan para crear los casos de plagio. Por tanto, es posible realizar un estudio concreto sobre ese corpus para entrenar un modelo específico que sea capaz de alcanzar altas puntuaciones en ese dominio concreto (de hecho los valores de configuración que se muestran en la sección 2.2.2 son los óptimos para el corpus PAN-PC'11). En cambio, nuestros modelos están desarrollados para ofrecer la misma calidad de resultados en cualquier ámbito, y no requieren de ninguna configuración ni entrenamiento específico para ningún dominio, lo cual los convierte en modelos adecuados para trabajar sobre un escenario de contexto realista como la Web.

En vista de todos los resultados anteriores, podemos afirmar cómo hacer uso de grafos de conocimiento es una buena alternativa para la detección de plagio translingüe. Cabe señalar que el coste computacional de crear un grafo de conocimiento es notablemente más elevado que el de otros métodos que realizan búsquedas en diccionarios. Con las máquinas existentes actualmente es perfectamente posible el uso de grafos de conocimiento para detectar similitud, pero es un hecho que hay que tener en cuenta según la tarea y la importancia del coste computacional. En cambio, los modelos utilizando el diccionario estadístico de BabelNet ofrecen un coste computacional reducido, a costa de un rendimiento más reducido (sobre todo para de-en, sobre cuyos lenguajes habrá que seguir trabajando para

mejorar su rendimiento), pero se convierten en una buena opción para análisis de similitud es-en.

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusiones

El campo de la detección de plagio translingüe automática está en plena expansión. A conocidos modelos de análisis de similitud como CL-CNG, CL-ESA y CL-ASA [43] que pueden ser utilizados para la tarea de detección de plagio externo, cada año se le suman nuevas aproximaciones gracias también a la competición internacional de detección de plagio: el PAN [42].

En este trabajo se ha hecho un resumen del estado del arte actual en detección de plagio, tanto a nivel monolingüe como translingüe. A continuación se ha definido en que consisten las redes semánticas, haciendo hincapié en su variante multilingüe. Una vez se han introducido los conceptos necesarios, se han propuesto nuevos modelos para la detección de plagio translingüe externo. Dichos modelos hacen uso de la red semántica multilingüe BabelNet para llevar a cabo la detección. Por un lado se han propuesto cuatro modelos que combinan el modelo de análisis de similitud CL-ASA con el diccionario estadístico de BabelNet, ofreciendo cada uno un tipo de normalización en los pesos de las traducciones del diccionario: atendiendo a la logitud del texto y al número de traducciones posibles de una palabra. Además, se ha propuesto un nuevo modelo de análisis de similitud basado en grafos, *cross-language knowledge graphs analysis* (CL-KGA), el cual utiliza grafos de conocimiento creados a partir

de BabelNet, que expanden y relacionan los conceptos originales de un fragmento de texto, proporcionando un modelo de contexto de su contenido. La ventaja de dicha aproximación, a parte de la expansión y relación del vocabulario, es el hecho de que un grafo de conocimiento tiene una dimensión multilingüe en cada uno de sus nodos y aristas, de modo que dos grafos creados a partir de dos textos en diferentes lenguajes que hablen del mismo contenido serán similares.

En el aspecto del coste computacional, CL-ASA combinado con el diccionario estadístico de BabelNet resulta igual de eficiente que el CL-ASA clásico, mientras que CL-KGA tiene un coste superior, dada la complegidad de elaborar grafos de conocimiento, pero siendo una dificultad salvable con las máquinas existentes actualmente, y que no impide que CL-KGA sea un eficiente modelo de análisis de similitud.

Para la evaluación de los modelos propuestos se ha hecho uso de un corpus diseñado específicamente para la tarea: el corpus de la competición PAN-PC'11, del cual se ha tomado su tarea de detección de plagio translingüe externo, la cual incluye detección de plagio de-en y es-en. Toda la metodología de la evaluación ha seguido estrictamente las pautas que en esta competición se dictan. En los experimentos se han comparado los modelos propuestos con los del estado del arte CL-CNG (utilizando trigramas) y CL-ASA (utilizando un diccionario estadístico entrenado con el modelo IBM M1), además de compararnos con el ganador de dicha competición.

Los resultados experimentales han resultado clarificadores. Por un lado los modelos que combinan CL-ASA con el diccionario estadístico de BabelNet han ofrecido resultados dispares. Las puntuaciones utilizando normalización en función del número de traducciones de una palabra han resultado en un punto medio entre el CL-ASA (IBM M1) y CL-CNG. Para los modelos que combinan CL-ASA con el diccionario de BabelNet y normalizan en función de la longitud del texto o que no normalizan, los resultados han sido superiores al CL-ASA (IBM M1) y al CL-CNG para detección de plagio es-en, mientras que para de-en los resultados han sido inferiores. El problema aquí ha sido con en lenguaje aglutinativo alemán,

que requiere de herramientas específicas para la extracción, lematización y etiquetado de sus conceptos. Sin embargo, en vista de los resultados podemos afirmar que el diccionario estadístico de BabelNet permite su utilización para una detección de plagio translingüe eficiente en un contexto es-en.

Por otro lado, el modelo CL-KGA ha ofrecido unos resultados excelentes en todas sus pruebas, superando a todos los otros modelos. Para es-en ha obtenido una puntuación global de detección de plagio (*plagdet*) un 14.9 % superior al modelo CL-ASA (IBM M1) y un 250 % superior al CL-CNG, para de-en ha superado al CL-ASA (IBM M1) en un 26.6 % y al CL-CNG en un 559 %.

Tras los experimentos anteriores en los que se ha visto la calidad de los modelos, se ha pasado a un análisis más exhaustivo en el que se han detectado por separado los casos de plagio en el corpus generados por una máquina de traducción estadística y los casos que han recibido corrección manual. En dichas pruebas hemos podido observar como, si bien se detectan más del doble de casos traducidos de forma automática, CL-KGA ha superado en todas las pruebas al resto de modelos, doblando en algunos casos los valores. Por último se han comparado los modelos evaluados con el ganador de la edición del PAN del año 2011, BDCSS, para el cual solamente están disponibles resultados promediados de las dos particiones es-en y de-en. El modelo CL-KGA ha resultado un 16 % inferior a este, CL-ASA (IBM M1) un 39 % inferior y el mejor CL-ASA combinado con el diccionario estadístico de BabelNet un 52 % inferior. Dichos resultados resultan muy positivos, ya que CL-KGA no necesita una fase previa de entrenamiento sobre un corpus para estimar parámetros, y es capaz de ofrecer todo su rendimiento desde el primer momento en cualquier excenario, incluso en un contexto realista como la Web, cosa que no ocurre con BDCSS.

En conclusión, en este trabajo hemos presentado diferentes modelos de análisis de similitud en un contexto translingüe, resaltando el modelo CL-KGA, que utilizando grafos de conocimiento a modo de modelos de contexto, ha demostrado ser una alternativa que ofrece mejores resultados que los modelos estado del arte que se uti-

lizan actualmente para dicha tarea, y modelo sobre el cual todavía quedan muchas vías por explorar.

6.2. Líneas de investigación abiertas

A raíz de las conclusiones que hemos podido extraer de los resultados experimentales, existen diversas ideas para mejorar la investigación actual, además de otras líneas de investigación que nacen tras los satisfactorios resultados que los grafos de conocimiento han proporcionado en análisis de similitud a nivel translingüe. Estas ideas podrían llevarse a cabo en el futuro en el marco de una tesis doctoral.

Las líneas de investigación abiertas para mejorar los modelos propuestos en el presente trabajo son las siguientes:

- **Utilización de un *tagger* para la lematización y etiquetado de las palabras a buscar en el diccionario estadístico de BabelNet:** En el trabajo actual las palabras de un texto a buscar en el diccionario se le proporcionaban directamente a BabelNet, sin recibir ningún tipo de tratamiento, lo cual puede influir al fallo en la localización de palabras derivadas de su forma infinitiva, pero que realmente sí existen dentro del diccionario. Una herramienta que realice un análisis sintáctico del texto para etiquetar correctamente las palabras, y que posteriormente las lematice, nos permitiría un automento considerable del porcentaje de aciertos dentro del diccionario de BabelNet, además de hacer uso de la búsqueda con categoría gramatical, que actualmente se omite.
- **Utilizar el método de construcción de grafos de conocimiento solamente para realizar una expansión del vocabulario, y buscar luego los conceptos en el diccionario con el modelo CL-ASA:** Esta aproximación sería una combinación de los clásicos métodos utilizando diccionario con la expansión de vocabulario que proporcinan los grafos de conocimiento, descartando por otro lado las relaciones entre conceptos. Una ventaja de dicho método sería que una vez se tenga la expansión de vocabulario, el coste sería siempre el mismo que

utilizando CL-ASA, en lugar del coste superior de los grafos de conocimiento.

- **Utilizar una herramienta específica para la extracción y etiquetado de conceptos en alemán:** Actualmente, en la elaboración de grafos de conocimiento se hace uso de la herramienta TreeTagger por su capacidad para etiquetar texto de multiples lenguajes, pero por los resultados vistos en la evaluación, la extracción de conceptos y etiquedado se podría mejorar notablemente para el alemán, ya que parece que estamos perdiendo un elevado porcentaje de conceptos por un mal etiquetado. Dentro de las posibles opciones nos encontramos con las herramientas del kit OpenNLP[1] que incluye, entre otras herramientas, etiquetador, tokenizador, *parser* y *chunker*. Por otro lado, el *tagger* de la universidad de Stanford[2] también podría ser una buena alternativa para etiquetar el lenguaje alemán.
- **Mejorar algoritmo de similitud de grafos:** Existen diversas mejoras que se podrían aplicar al algoritmo de comparación de grafos propuesto en la sección 4.2. Por un lado se podría hacer una versión que comparase grafos por el método de los X vecinos más cercanos a un concepto original del texto, descartando el resto. Por otro lado, se podría modificar el algoritmo actual para que diera más valor a los conceptos originales del texto, y los que se hayan añadido a la hora de generar el grafo de conocimiento fueran perdiendo valor conforme se alejasen de los conceptos originales.
- **Utilizar otras redes semánticas multilingües para dotar a CL-KGA de más lenguajes:** Existen otras redes semánticas multilingües como ConceptNet [23] o EuroWordNet [61] que podrían ser utilizadas para la creación de grafos de conocimiento, dotando al modelo CL-KGA de más lenguajes dentro de su dimensión multilingüe.
- **Método híbrido entre CL-ASA y CL-KGA:** Para reducir el coste computacional del CL-KGA, podemos sacrificar precisión utilizando CL-ASA para localizar las secciones del tex-

[1] `http://opennlp.apache.org/`
[2] `http://nlp.stanford.edu/software/tagger.shtml`

to sospechosas de contener plagio, y a continuación pasar CL-KGA sobre ellas y determinar si efectivamente existe plagio.

Los grafos de conocimiento han demostrado ser efectivos para su uso en detección de plagio translingüe, pero su utilidad va mucho más allá. En mayo de 2012 salió la noticia de que Google comenzaba a utilizar su propio grafo de conocimiento[3] para mejorar las búsquedas en inglés dotándolas de contexto. Aunque sus herramientas y el grafo todavía no están disponibles públicamente para el desarrollo, podemos hacer uso de la red semántica multilingüe BabelNet para la creación de grafos de conocimiento y su aplicación en nuevas tareas. Las nuevas líneas de investigación que nacen a partir de este trabajo son las siguientes:

- **Utilizar CL-KGA para otras tareas diferentes de detección de plagio translingüe:** El modelo CL-KGA puede ser utilizado para todo tipo de tareas de análisis de similitud, tanto a nivel monolingüe como translingüe. Podría ser interesante evaluar su calidad en detección de similitud y reutilización en textos a nivel monolingüe en el corpus METER [4] por ejemplo.
- **Emplear grafos de conocimiento para *opinion mining:*** Dentro del ámbito del análisis de opiniones [38] en texto, podríamos utilizar la expansión de vocabulario que ofrecen los grafos de conocimiento, para luego buscar sobre dicho vocabulario las palabras sobre un diccionario de sentimientos [10], lo cual proporcionaría un aumento del vocabulario y contexto más relacionado con los sentimientos que manifieste el texto, y podría aumentar la precisión del análisis.
- **Aplicar grafos de conocimiento para la elaboración de resúmenes:** La tarea de elaboración de resúmenes de texto [18] se podría ver también beneficiada de la información que los grafos de conocimiento aportan. Elaborando grafos de fragmentos de texto, combinandolo con nuestro método de análisis de similitud en grafos, podríamos detectar qué fragmentos de texto

[3] `http://www.fayerwayer.com/2012/05/google-presenta-el-grafo-del-conocimiento-para-darle-sentido-a-las-busquedas/`

[4] `http://nlp.shef.ac.uk/meter/`

son los más representativos de un texto o tema. Por otro lado, analizando los complementos de los grafos intersección en el algoritmo de similitud[5], podríamos detectar qué partes son más o menos frecuentes o incluso contrarias entre diferenes resúmenes. Finalmente, al tener los grafos una dimensión multilingüe, el sistema podría trabajar a partir de textos en diferentes lenguajes de forma nativa.

- **Utilizar grafos de conocimiento para clasificación temática translingüe:** El problema de la clasificación temática translingüe [24] consiste en, dado un conjunto de textos en diferentes lenguajes, clasificarlos según su tema. Utilizando grafos de conocimiento podemos tomar dos líneas de investigación diferentes: (i) Dado un conjunto de textos de los que conocemos el tema, podemos crear un grafo de conocimiento para cada uno de ellos, y combinandolos (bien mediante intersección o mediante métodos más complejos como la detección de los conceptos más representativos) obtendríamos un grafo prototipo. Si creamos un grafo prototipo para cada una de los temas/clases, para cada nuevo documento a clasificar en el sistema será cuestión de crear su grafo de conocimiento y analizar su similitud con los grafos prototipo de cada clase. (ii) Podemos generar los subgrafos comunes más frecuentes [28] de los documentos de cada clase, y una vez tenemos estos subgrafos, los podemos tratar como características para combinarlos con otros métodos de clasificación como Bayes o máquinas de soporte vectorial. Para cada nuevo texto a clasificar, podríamos extraer los subgrafos comunes más frecuentes con cada una de las clases y analizar su similitud mediante un clasificador.

  La utilización de grafos de conocimiento para esta tarea podría comportar dos mejoras sustanciales respecto a otras aproximaciones: (i) Por un lado los grafos de conocimiento tienen una dimensión multilingüe, y gracias a ello no habría que tratar con máquinas de traducción estadística y perder precisión en el proceso, lo cual es uno de los principales problemas de la tarea. (ii) Por otro lado, en ocasiones si los textos están hablando de

[5]Los complementos de un grafo intersección serían aquellas partes del grafo que no se encuentran presentes en la intersección, al no ser comunes entre los grafos fuente.

un tema muy propio de un lenguaje (debido a la cultura de su país de origen), se puede perder mucha precisión, ya que la traducción del texto llevaría a la pérdida total del contexto, p.e: un texto en castellano que hable de pelota vasca. En cambio, si se utilizan grafos de conocimiento, al crear un modelo de contexto del contenido del texto, podremos aportar la suficiente información mediante nuevos conceptos y relaciones, como para comprender de qué estamos hablando, a pesar de cambiar de idioma.

- **Emplear grafos de conocimiento para *domain adaptation*:** La tarea de *domain adaptation* [11] se refiere a, dado un corpus con unas características concretas sobre un dominio, adaptar dichas características para su utilización en otro dominio diferente con el que guarde relación, como por ejemplo adaptar un corpus sobre noticias deportivas de futbol al dominio de noticias deportivas de baloncesto. Las ventajas que aportarían los grafos de conocimiento respecto a otras aproximaciones serían similares a las explicadas en el problema anterior de clasificación temática. Por un lado la utilizació de grafos de conocimiento aportaría una mayor calidad en las traducciones que otros métodos que utilizan máquinas de traducción estadística, por otro, para conceptos con los que una simple palabra no fuera suficiente para comprender su significado, la expansión de vocabulario en los grafos de conocimiento aportaría una definición más clara de lo que represente dicho concepto en el texto.

Bibliografía

[1] M. Baroni and S. Bernardini. A new approach to the study of translationese: machine learning the difference between original and translated text. 21(3):259–274, 2006.

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

[3] A. Barrón-Cedeño, A. Eiselt, and P. Rosso. Monolingual text similarity measures: A comparison of models over wikipedia articles revisions. In *Proc. of 7th Int. Conf. on Natural Language Processing*, pages 29–38. ICON-2009, 2009.

[4] A. Barrón-Cedeño, P. Rosso, D. Pinto, and A. Juan. On cross-lingual plagiarism analysis using a statistical model. In *Proc. of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, PAN'08, 2008.

[5] Y. Bernstein and J. Zobel. A scalable system for identifying coderivative documents. In *Proc. of the Symposium on String Processing and Information Retrieval*, 2004.

[6] S. Brin, J. Davis, and H. García-Molina. Copy detection mechanisms for digital documents. In *Proc. of the 1995 ACM SIGMOD Int. Conference on Management of Data*, pages 398–409. ACM Press, 1995.

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

[8] P. Clough and M. Stevenson. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis*, 45(1):5–24, 2010.

[9] R. Comas and J. Sureda. Academic cyberplagiarism: tracing the causes to reach solutions. *Digithum*, 10:1–6, 2008.

[10] T. Crawford and A. Ellis. A dictionary of rational-emotive feelings and behaviors. *Journal of Rational-Emotive and Cognitive-Behavior Therapy*, 7(1):3–28, 1989.

[11] H. Daumé, III, A. Kumar, and A. Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 53–59, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

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

[13] R. A. Española. *Real academia española. Diccionario de la lengua española. Vigésima segunda edición.* 2008.

[14] C. Fellbaum. *WordNet: An electronic lexical database*. Bradford Books, 1998.

[15] M. Franco-Salvador, P. Gupta, and P. Rosso. Cross-language plagiarism detection using BabelNet's statistical dictionary. *Computación y Sistemas, Revista Iberoamericana de Computación*, 16(4):383–390, 2012.

[16] M. Franco-Salvador, P. Gupta, and P. Rosso. Cross-language plagiarism detection using a multilingual semantic network. In *Proc. of the 35th European Conference on Information Retrieval (ECIR'13)*. Springer-Verlag, LNCS(7814), 2013.

[17] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In

Proc. of the 20th int. joint conference on Artifical intelligence, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[18] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 121–128, New York, NY, USA, 1999. ACM.

[19] T. Gottron. External plagiarism detection based on standard IR technology and fast recognition of common subsequences. In *Lab Report for PAN at CLEF 2010*, 2010.

[20] J. Grman and R. Ravas. Improved implementation for finding text similarities in large sets of data - Notebook for PAN at CLEF 2011. In V. Petras, P. Forner, and P. D. Clough, editors, *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[21] P. Gupta, A. Barrón-Cedeño, and P. Rosso. Cross-language high similarity search using a conceptual thesaurus. In *Proc. 3rd Int. Conf. of CLEF Initiative on Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics*. CLEF 2012, 2012.

[22] P. Gupta and P. Rosso. Text reuse with acl: (upward) trends. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pages 76–82, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[23] C. Havasi. Conceptnet 3: A flexible, multilingual semantic network for common sense knowledge. In *Proc. of the 22nd Conf. on Artificial Intelligence*, 2007.

[24] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Proc. of the ICDM 03, The 2003 IEEE International Conference on Data Mining*, pages 541–544, 2003.

[25] P. Jackson and I. Moulinier. *Natural language processing for online applications: Text retrieval, extraction and categorization.* John Benjamins Publishing Company, 2002.

[26] N. Kang, A. Gelbukh, and S. Han. PPChecker: Plagiarism Pattern Checker in document copy detection. In *Text, Speech and Dialogue (TSD 2006)*, volume LNAI (4188), pages 661–667. Springer-Verlag, 2006.

[27] L. Kong, H. Qi, S. Wang, C. Du, S. Wang, and Y. Han. Approaches for candidate document retrieval and detailed comparison of plagiarism detection. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[28] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 313–320, Washington, DC, USA, 2001. IEEE Computer Society.

[29] M. Li and P. M. Vitnyi. *An introduction to Kolmogorov complexity and its applications.* Springer Publishing Company, Incorporated, 3 edition, 2008.

[30] P. Mcnamee and J. Mayfield. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1):73–97, 2004.

[31] S. Meyer zu Eissen and B. Stein. Intrinsic plagiarism detection. In *Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, volume LNCS(3936), pages 565–569. Springer-Verlag, 2006.

[32] M. Montes y Gómez, A. F. Gelbukh, A. López-López, and R. A. Baeza-Yates. Flexible comparison of conceptual graphs. In *Proc. DEXA*, pages 102–111, 2001.

[33] R. Navigli and S. P. Ponzetto. Babelnet: building a very large multilingual semantic network. In *Proc. of the 48th annual meeting of the association for computational linguistics*, ACL '10, pages 216–225, Stroudsburg, PA, USA, 2010.

[34] R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, Dec. 2012.

[35] R. Navigli and S. P. Ponzetto. BabelNetXplorer: A platform for multilingual lexical knowledge dase access and exploration. In *Companion Volume to the Proceedings of the 21st World Wide Web Conference,* Lyon, France, 16–20 April 2012, pages 393–396, 2012.

[36] R. Navigli and S. P. Ponzetto. Multilingual wsd with just a few lines of code: The babelnet api. In *Proc. 50th annual meeting of the association for Computational Linguistics*, 2012.

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

[38] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.

[39] R. C. Pereira, V. P. Moreira, and R. Galante. UFRGS@PAN2010: Detecting external plagiarism - Lab report for PAN at CLEF 2010. In M. Braschler, D. Harman, and E. Pianta, editors, *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[40] D. Pinto, J. Benedí, and P. Rosso. Clustering narrow-domain short texts by using the kullback-leibler distance. In *Computational Linguistics and Intelligent Text Processing*, pages 611–622, 2007.

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

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

[43] M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso. Cross-language plagiarism detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 45(1):45–62, 2011.

[44] M. Potthast, A. Eiselt, A. Barrón-Cedeño, B. Stein, and P. Rosso. Overview of the 3rd int. competition on plagiarism detection. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[45] M. Potthast and B. Stein. New issues in near-duplicate detection. *Data Analysis, Machine Learning and Applications*, pages 601–609, 2008.

[46] B. Pouliquen, R. Steinberger, and C. Ignat. Automatic linking of similar texts across languages. In *Proc. Recent Advances in Natural Language Processing III*, pages 307–316. RANLP'03, 2003.

[47] P. Scanlon and D. Neumann. Internet plagiarism among college students. *J Coll Student Dev*, 43:375–385, 2002.

[48] S. Schleimer, D. S. Wilkerson, and A. Aiken. Winnowing: Local algorithms for document fingerprinting. In *ACM SIGMOD Conference*, pages 76–85, 2004.

[49] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proc. Int. Conf. on new methods in language processing*, 1994.

[50] L. Seaward and S. Matwin. Intrinsic plagiarism detection using complexity analysis. In *Agirre (Eds.). PAN'09*, pages 56–61, 2009.

[51] H. Somers, F. Gaspari, and A. Niño. Detecting inappropriate use of free online machine translation by language students – a special case of plagiarism detection. In *Proc. of the Eleventh Annual Conference of the European Association for Machine Translation*, pages 41–48, 2006.

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

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

[54] E. Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. In *Proc. of the 3rd Int. Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*, pages 38–46, 2009.

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

[56] B. Stein, M. Koppel, and E. Stamatatos. Plagiarism analysis, authorship identification, and near-duplicate detection. *SIGIR Forum (PAN 2007)*, 41(2):68–71, 2007.

[57] B. Stein and M. Potthast. Applying hash-based indexing in text-based information retrieval. In *Proc. of the 7th Dutch-Belgian Information Retrieval Workshop*, pages 29–35. M. Moens, T. Tuytelaars, and A. de Vries, editors, 2007.

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

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

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

[61] P. Vossen. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-

lingual index. In *Proc. Int. Journal of Lexicography*, volume 17, 2004.

Apéndice A

Publicaciones y charlas invitadas

Las investigaciones descritas en esta tesis han permitido la publicación de los siguientes artículos:

- M. Franco-Salvador, P. Gupta and P. Rosso. Cross-Language Plagiarism Detection using a Multilingual Semantic Network. *In 35th European Conference on Information Retrieval (ECIR'13). Springer-Verlag, LNCS(7814), Moscow, Russia, 2013.* **(CORE B)**

- M. Franco-Salvador, P. Gupta and P. Rosso. Cross-language Plagiarism Detection Using BabelNet's Statistical Dictionary. *Computación y Sistemas, Revista Iberoamericana de Computación, ISSN 1405-5546, vol. 16, num. 4, pp. 383-390, 2012.* **(included in the Index of Excellence of CONACyT, Scopus, Redalyc, E-Journal, Latindex, Biblat, Periódica, DBLP and SciELO)**

- M. Franco-Salvador, P. Gupta and P. Rosso. Graph-Based Similarity Analysis: A New Approach to Cross-Language Plagiarism Detection. *Journal of the Spanish Society of Natural Language Processing (Sociedad Española de Procesamiento del Languaje Natural), 2013. (submitted)* **(included in the Index of Excellence of RECYT, Scopus, DICE, RESH, Biblio-**

teca.Net, LATINDEX, CARHUS PLUS+, CINDOC-CSIC, e-Revistas, RUA, Dialnet and INIST)

El contenido de este trabajo de investigación también se ha divulgado mediante las siguientes charlas invitadas:

- **Graph-Based Similarity Analysis: A New Approach to Cross-Language Plagiarism Detection.**
  Centro de Investigación en Computación (CIC). Instituto Politécnico Nacional. México D.F., México. Nov. 2012

- **Cross-Language Plagiarism Detection using a Multilingual Semantic Network.**
  Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Puebla, México. Dec. 2012