Detección automática de plagio en texto

Luis Alberto Barrón Cedeño

Departamento de Sistemas Informáticos y Computación

Director: Paolo Rosso

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Tesis desarrollada dentro del Máster en Inteligencia artificial, reconocimiento de formas e imagen digital

Valencia, noviembre de 2008

A quien me dejó ir y a quien me acompaña

Que la aritmética es la más baja de todas las actividades mentales se demuestra por el hecho de que es la única que puede realizarse por medio de una máquina.

Arthur Schopenhauer

Resumen

Plagiar es robar el crédito por el trabajo realizado por otra persona. En el caso de la lengua escrita, significa incluir en un documento fragmentos de texto escritos por alguna otra persona sin darle el crédito correspondiente.

La detección automática de plagio se basa en diversas técnicas de recuperación y extracción de información así como de reconocimiento de formas y teoría de la información. Esta tarea ha comenzado a generar gran interés debido a la posibilidad de crear mecanismos que puedan detectar casos de plagio de manera eficiente. Esto puede provocar, como un efecto secundario, el desaliento por caer en esta falta.

En este trabajo presentamos una descripción del estado del arte en materia de detección de plagio. Además, describimos un conjunto de técnicas (algunas de ellas ya existentes y otras diseñadas por nosotros mismos) y presentamos distintas evaluaciones sobre ellas. Los resultados que hemos obtenido hasta ahora con diversas técnicas basadas en conceptos tan variados como los modelos de lenguaje, distintas técnicas de comparación de texto y métodos estadísticos para la reducción de espacios de búsqueda, han generado resultados prometedores.

Como continuación de los experimentos realizados, planteamos las pautas que consideramos conveniente seguir dentro de esta investigación para seguir haciendo aportaciones en el área dentro futuras investigaciones.

Abstract

To plagiarise is to take the credit for another person's work. In the case of text, to plagiarise means including text fragments (and even entire documents) from other persons in a document without giving the corresponding credit.

The automatic plagiarism detection is based on different techniques of Information Retrieval and Extraction as well as Pattern Recognition and Information Theory. This task has received special attention in the last years due to the possibility of generating efficient mechanisms for the detection of plagiarism cases. The production of this kind of resources could minimise the temptation to plagiarise.

In this work, we present a description on the state-of-the-art in automatic plagiarism detection. Additionally, we describe a set of techniques (some of them already developed and other recently designed by ourselves) including some evaluations. The results that we have obtained with techniques based on different concepts such as Language Models, different text comparison techniques, and statistical methods for the search space reduction, are promising.

Finally, we establish the direction that this research work should take in order to keep providing elements to the plagiarism detection area.

Agradecimientos

Quisiera agradecer en primer lugar a Paolo Rosso, quien sencillamente se ha limitado a proporcionarme lo que podía esperar de un supervisor. Espero que podamos seguir así durante el doctorado, aunque una mejor distribución del agobio sería bastante conveniente.

En segundo lugar quisiera agradecer al Dr. Benno Stein, de la Universidad de Weimar. Sus comentarios certeros han permitido que tanto la investigación como yo vayamos madurando poco a poco. Estoy seguro de que serguirá siendo importante en futuras investigaciones.

Igualmete, quiero agradecer a los profesores que han participado directamente en las distintas etapas de esta investigación (en orden cronológico): Encarna Segarra, Ferrán Pla, Alfons Juan, Francisco Casacuberta, José Miguel Benedí y Enrique Vidal.

Por supuesto, a David Pinto quien, desde el punto de vista de un estudiante (mucho más avanzado), aportó muchísimo a todas las etapas.

En el ámbito personal, quiero agradecer a mi mamá. Porque sé que hay sentimientos y buenos deseos que son capaces de cruzar tierra y mar.

Gracias a Silvia por atreverse a estar en esta aventura y por ser tan paciente. La luz al fondo del tunel es aún tenue, pero espero que juntos podamos lograr que brille un poco más cada vez.

Sé que mis amigos y familiares en América se acuerdan de vez en cuando de mí. No lo duden, yo no los olvido.

Igualmente, gracias a los americanos y europeos (y uno que otro africano) que en esta península han compartido algo conmigo.

Índice general

Índice general v

Índice de figuras vii

Índice de tablas ix

1. Introducción 1

1.1. Descripción de la problemática 2

1.2. Motivación y objetivos 3

1.3. Planteamiento del problema 4

1.4. Detección de plagio dentro del ámbito *IARFID* 5

1.5. Organización de la tesis 6

2. Estado del arte de la detección de plagio 7

2.1. Análisis intrínseco de plagio 9

2.2. Detección de plagio con referencia 12

2.2.1. Análisis a nivel de documentos 13

2.2.2. Análisis basado en comparación de n-gramas 13

2.2.3. Determinando el tipo de plagio a nivel de sentencia 15

2.2.4. Acelerando el proceso de detección de plagio 16

2.3. Recursos disponibles 18

2.3.1. El corpus METER 18

2.3.2. El corpus de plagio *Webis* 19

3. Aproximaciones a la detección automática de plagio 23

3.1. Corpus utilizados en los distintos experimentos 23
3.1.1. Un corpus sintético basado en el proyecto Gutenberg 23
3.1.2. Un corpus sintético basado en documentos especializados 25
3.1.3. Extracto del corpus METER 25
3.2. Modelos de lenguaje aplicados a la detección de plagio 26
3.2.1. Modelos de lenguaje 27
3.2.2. Planteamiento del modelo basado en modelos de lenguaje 28
3.2.3. Evaluación del método basado en modelos de lenguaje 29
3.2.4. Discusión sobre el método basado en modelos de lenguaje 33
3.3. Búsqueda de plagio basada en n-gramas de palabras 34
3.3.1. Planteamiento del modelo basado en n-gramas 35
3.3.2. Evaluación del método basado en modelos de lenguaje 36
3.3.3. Discusión sobre el método basado en n-gramas 37
3.4. El problema del espacio de búsqueda 38
3.4.1. Planteamiento del método de reducción de espacio de búsqueda . . . 38
3.4.2. Evaluación del método de reducción del espacio de búsqueda 41
3.4.3. Discusión sobre el método de reducción de espacio 44

4. Líneas de investigación abiertas **45**
4.1. Diseño de métodos eficientes 45
4.2. Diseño de métodos translingües 46
4.3. Creación de corpus adecuados para las investigaciones en detección de plagio 47

Bibliografía **51**

A. Descripción de símbolos **57**

B. Publicaciones en el marco de la investigación **59**

Índice de figuras

2.1. Ejemplo de texto plagiado detectable por la simple lectura. 10

2.2. Distribución de n-gramas en un conjunto de textos sobre el mismo tema (n = grado el n-grama) 14

3.1. Perplejidad obtenida para los fragmentos de test en el corpus especializado. (+ = sentencia original, ■ = sentencia plagiada) 30

3.2. Perplejidad obtenida para los fragmentos de test en el corpus literario. (+ = sentencia original, ■ = sentencia plagiada) 32

3.3. Un ejemplo de sentencia plagiada. ($d_{\{a,b\}}$ es el fragmento original que sirvió como fuente de los fragmentos plagiados) 35

3.4. Evaluación del método de detección de plagio basado en n-gramas (n = grado del n-grama, , t =umbral 37

3.5. Proceso de reducción del espacio de búsqueda. 41

3.6. Evaluación del proceso de reducción del espacio de búsqueda. ($\{tf, tfidf, tp\}$ = técnicas de extracción de características, l = longitud de los términos) 42

Índice de tablas

2.1. Número de n-gramas que varios documentos tienen en común (normalizado por el número total de n-gramas en todos los documentos) 14

2.2. Producto punto entre una frase original y una potencialmente plagiada . . . 17

2.3. Nota de la PA en conjunto con la nota de *The Telegraph* correspondiente . . 20

2.4. Ejemplo de documento plagiado del corpus Webis 21

3.1. Estadísticas del corpus sintético literario. 24

3.2. Estadísticas del corpus especializado. 25

3.3. Estadísticas del extracto del corpus METER considerados en los experimentos 26

3.4. Sentencias lematizadas del corpus literario con las más altas perplejidades . . 33

3.5. Comparación de resultados: búsqueda exhaustiva contra reducción de espacio + búsqueda exhaustiva. (P =Precision, R =Recall, F = F-me asure, t = tiempo promedio de procesamiento (seg.)) 43

4.1. Características deseables en un corpus para la detección de plagio 48

Capítulo 1

Introducción

El plagio es una falta grave que se puede ver reflejada en muchos ámbitos. La definición proporcionada por la Real Academia Española acerca del acto de plagiar [17] es, además de simple, tajante:

> *plagiar. Copiar en lo sustancial obras ajenas, dándolas como propias.*

Existen muy diversos tipos de plagio: de música, de imágenes, de palabras e incluso de ideas. El tipo de plagio que nos interesa en este momento es el de palabras, en particular aquellas que se encuentran en una realización escrita, es decir, el plagio en texto. Dentro de este ámbito, plagiar implica incluir fragmentos de texto que se encuentran en documentos escritos por otro autor en un documento propio sin incluir el crédito correspondiente.

En nuestros días, los problemas concernientes al plagio, y en particular al plagio de textos, se han visto incrementados debido al fácil acceso a grandes fuentes de información a través de medios electrónicos. Desafortunadamente, su detección es prácticamente imposible de forma manual. Por ello, es importante desarrollar mecanismos automatizados que permitan realizar la tarea de detección de plagio en un tiempo razonable. De esta manera, se puede combatir la enorme tentación de plagiar textos provenientes de cualquier lado. Por fortuna, como podrá observarse a través de la lectura de este trabajo, este problema que se ha visto beneficiado por el desarrollo de la tecnología informática, puede ser atacado por ella misma.

Antes de continuar, es prudente señalar que a lo largo del trabajo se utilizará la primera persona del plural para hacer referencia a la investigación que hemos realizado debido a que no ha sido el resultado de un trabajo individual. En ella, además de mis esfuerzos propios, hay horas de trabajo del supervisor de esta investigación, Paolo Rosso, y de varios profesores del Máster en Inteligencia artificial, reconocimiento de formas e imagen digital (*IARFID*) con quienes, a través de los pequeños proyectos relacionados con esta tarea desarrollados en el marco de sus asignaturas, la investigación ha ido adquiriendo forma.

1.1. Descripción de la problemática

El plagio de textos es un gran problema en la actualidad, sobre todo dentro de círculos estudiantiles. Sólo por dar un ejemplo, en una prueba realizada a un conjunto de alrededor de 300 estudiantes [21], se observó que más de la mitad aceptaba haber hecho alguna trampa durante el año escolar. Por supuesto, existen casos de plagio dentro de otros ámbitos, como el laboral o el científico.

Los casos de plagio de documentos (ya sea completos o de los fragmentos que los componen), se han visto incrementados de manera importante en los últimos años. Una de las razones principales para el desarrollo de este fenómeno, como se señaló anteriormente, es la facilidad con la que es posible acceder a documentos en soporte electrónico. Existen dos maneras principales de obtener fragmentos de texto en formato electrónico que pueden ser utilizados de manera ilícita. La primera de ellas es a través de la Web. A menudo muchos artículos científicos así como reportes e incluso libros enteros pueden ser accedidos a través de un buscador suficientemente sofisticado como el ofrecido por sitios como Google o Yahoo. La segunda consiste en obtener los documentos de personas conocidas, como compañeros de clase.

Por otro lado, existe un fenómeno de particular interés dentro del ámbito de Internet que puede ser considerado como un plagio, se trata del *reuso* de contenidos. Dicho fenómeno puede observarse de manera muy sencilla. Cuando se realiza una petición de información por medio de cualquier buscador, es cada vez más común que se hallen copias de una misma página (específicamente de su contenido), alojadas en distintos servidores. Frecuentemente sólo existen entre las páginas implicadas ciertos cambios en los encabezados y pies de página, los cuales no afectan en nada al contenido. Existen también algunos cambios de formato (tales como el color del fondo o la fuente del texto) que evidentemente no aportan mayor información. En este caso en particular, resulta irrelevante si la copia de una página hace referencia a su fuente, y ello sin señalar que dicha referencia no existe en la gran mayoría de los casos. Estas duplicidades de información a través de Internet no hacen más que obstruir al usuario en su búsqueda de información y su detección oportuna podría ahorrar tiempo de búsqueda.

El Departamento de Química de la Universidad de Kentucky ha delimitado distintos tipos de plagio[1]. El caso más sencillo se refiere a la copia directa de material original sin realizar alguna modificación. Otro de los casos, que a primera vista se refleja más complicado de analizar de manera automática es la reescritura (acto conocido en inglés como *rewording* o *rewrite*). En este caso, en lugar de simplemente copiar algún fragmento de texto, éste se reescribe por medio del uso de sinónimos o del cambio en el orden de las palabras que lo conforman. Aunque el segundo caso de plagio es más complicado de detectar que el primero, varias de las técnicas que se describen a través de este trabajo son buenas aproximaciones a lograr el resultado deseado: la detección de los fragmentos plagiados en un documento sin

[1]La descripción completa de dicho análisis puede consultarse en la página http://www.chem.uky.edu/Courses/common/plagiarism.html

importar que se presenten modificados.

Por su parte, Iyer y Singh [24] han destacado tres tipos de plagio (existe un cuarto tipo que es el de ideas, pero queda completamente fuera del contexto de esta investigación):

1. Palabra por palabra. Se trata de la copia, incluso modificada, de fragmentos de texto sin incluir su fuente.
2. De referencias. Se da cuando una referencia es observada en un documento y se transcribe al que se está escribiendo sin haber leído realmente la fuente.
3. De autoría. Ocurre cuando un autor dice ser creador de un trabajo que fue sustancialmente realizado por otro.

Dentro de esta clasificación, los plagios que consideramos factibles de ser detectados por medio del análisis de texto son el plagio de palabra por palabra y el plagio de autoría. Esto se debe a que se ven reflejados en la lectura de un texto. Por su parte, el plagio de referencias es prácticamente imposible de detectar. ¿Si tanto la cita como la referencia son incluidas correctamente, cómo podría detectarse un caso de plagio?

Por último, debe destacarse que, si bien el lenguaje utilizado en un texto está limitado por el tema tratado o por el público a quién va dirigido (entre otros factores), el autor aún es libre de escribir como quiera hacerlo. Un escritor es libre, por ejemplo, de hacer oraciones o párrafos de la longitud que desee, de usar el vocabulario y los signos de puntuación como lo considere conveniente o de usar el tiempo verbal que más le agrade. Estos son algunos de los factores que pueden ser explotados en la tarea de detección de plagio.

1.2. Motivación y objetivos

El reuso de la información no es un fenómeno nuevo y definitivamente no siempre es algo malo. Un ejemplo clásico de reuso de información es el de los filósofos griegos Platón y Sócrates. En este caso, Platón fue uno de los más asiduos seguidores de Sócrates y el encargado de dar a conocer su filosofía. Un caso mucho más reciente (y vano, cabe mencionar) es el de Enrique Bunbury, que ha sido calificado de plagiador por su último sencillo "El hombre delgado que no flaqueará jamás"[2]. Dicha canción está inspirada en un poema escrito por el desaparecido Pedro Casariego e incluso extrae algunas frases de él.

Sin embargo, estos casos, al igual que muchos otros, contemplan el reuso de información con el afán de enriquecerla. No se trata simplemente de un proceso de extracción de información, sino que implica su obtención, raciocinio y generación de una nueva versión, ya sea totalmente diferente o enriquecida de manera significativa. Desafortunadamente la información no siempre es reusada con estas "buenas intenciones". Cuando una persona extrae información de alguna fuente y la incluye en algún otro lugar, sin ningún tipo de procesamiento racional, se comete un acto de plagio que debe ser reprobado. ¡Y dentro del conjunto

[2]http://www.elpais.com/articulo/cultura/apropiacion/indecente/elpepucul/20080909elpepicul_2/Tes

de procesamientos racionales, no debe incluirse la modificación de la información con el afán de ocultar un plagio cometido!

Por fortuna, tecnología similar a la que ha hecho del acto de plagiar un proceso sencillo, puede detectarlo y por ende, evitarlo. Muchos de los métodos desarrollados para resolver tareas relacionadas con la búsqueda de información, el almacenamiento eficiente de datos e incluso la traducción estadística, pueden ser explotados, en ocasiones tras algunas adaptaciones, para la detección automática del plagio.

Los objetivos generales de esta investigación se detallan continuación:

1. Estudiar la problemática de la detección del plagio desde el punto de vista estadístico con el afán de observar cuáles son las ventajas y carencias de los métodos existentes hasta ahora.
2. Analizar los recursos existentes que puedan cubrir dichas carencias y comenzar con su adaptación y aplicación en la mejora de los resultados obtenidos en esta tarea.
3. Sentar las bases para la creación de los recursos necesarios para la futura generación de metodologías que ataquen la problemática dentro de entornos tanto monolingües como translingües.

Algunos objetivos particulares que se desprenden son:

1. El estudio del estado del arte en materia de detección automática de texto plagiado.
2. La búsqueda de recursos lingüísticos (corpus) que puedan ser utilizados tanto en el diseño como en la evaluación de los métodos existentes y futuros.
3. La generación de experimentos que comprueben el funcionamiento de algunas de la técnicas existentes.
4. El diseño de métodos preliminares que ataquen algunas de las carencias en los métodos desarrollados hasta el momento.
5. La delimitación de la brecha que se abre para la investigación a futuro, la cual será abordada dentro del marco de la investigación doctoral.

1.3. Planteamiento del problema

El problema de la detección automática de plagio puede verse como un problema de búsqueda y/o clasificación (en el análisis intrínseco de plagio, enfoque abordado en la sección 2.1, la tarea es abordada como un problema de clasificación pura).

El primero de los elementos a considerar en esta tarea es el corpus de referencia D, el cual está conformado por un conjunto de documentos de referencia[3]. Por extraño que parezca en principio, no existe ninguna condición de que dichos documentos de referencia sean realmente

[3]En el apéndice A contiene un resumen de la simbología utilizada a través de este trabajo

originales. La razón es que basta con que un fragmento de un texto sospechoso sea hallado en otro para determinar que se trata de un caso de plagio. La extensión y cobertura que pueda alcanzar el corpus de referencia es uno de los factores clave en el éxito de los sistemas de detección de plagio basados en corpus. Cada documento $d \in D$ es una potencial fuente del texto incluido en un documento sospechoso, es decir, un texto plagiado. El segundo elemento a considerar es precisamente el documento sospechoso, el cual será conocido en adelante somo s. Este documento puede ser original, contener fragmentos plagiados o, de hecho, estar enteramente plagiado. Estos dos elementos son suficientes para describir el planteamiento de la tarea de detección de plagio:

> Sean s un documento sospechoso y D un conjunto de documentos de referencia, el objetivo de la detección automática de plagio es encontrar aquel documento $d \in D$ que haya sido utilizado como fuente para obtener el documento s, el cual presumiblemente es un caso de plagio. Dicha búsqueda puede llevarse a un nivel más específico: Sea $s_i \in s$ un fragmento plagiado, el objetivo es encontrar aquel fragmento $d_j \in d$ tal que d_j es la fuente del fragmento plagiado s_i.

Encontrar la fuentes de un fragmento sospechoso es una prueba adecuada para determinar que se trata de un fragmento plagiado.

1.4. Detección de plagio dentro del ámbito *IARFID*

Esta tesis se ha desarrollado dentro del marco del Máster IARFID de la Universidad Politécnica de Valencia. A continuación, describimos cómo se integra la presente investigación en el marco de IARFID.

Cuando hablamos de la tarea de detección de plagio, definitivamente no nos referimos a una tarea sencilla. Esta tarea implica la aplicación de métodos de recuperación y extracción de información, además de reconocimiento de formas y procesamiento de lenguaje natural.

A través de este documento se verá que los conceptos probabilísticos de Bayes, impartidos en asignaturas como *Aprendizaje y percepción* e *Introducción al reconocimiento de formas*, son vitales para esta tarea. Igualmente se observará cómo los modelos de lenguaje, primero introducidos en la asignatura de *Lingüística computacional* y luego aplicados en las de *Reconocimiento de escritura* y *Reconocimiento automático del habla*, podrían ser una buena herramienta para caracterizar los textos analizados. También se observará cómo diversos métodos estadísticos, como los vistos en las asignaturas de *Métodos estadísticos en tecnologías del lenguaje* y *Aplicaciones de la lingüística computacional*, pueden ser de gran ayuda en la implementación de métodos eficientes.

Se verá también que el algoritmo *EM* aplicado en la alineación de textos (en particular el modelo *IBM-1*), visto a profundidad en las asignaturas de *Análisis estadístico de formas* y *Traducción automática* y luego aplicado en la de *Sistemas y herramientas de traducción*,

puede ser útil para abordar un enfoque de la detección de plagio totalmente novedoso: la detección de plagio translingüe.

Algunos de los conocimientos adquiridos en otras asignaturas, como las relacionadas con las redes neuronales y otros métodos de clasificación, no han sido aplicados de manera explícita todavía en esta investigación. Sin embargo, no nos queda duda de que más adelante surgirán para resolver los problemas que poco a poco nos vayamos planteando.

Las investigaciones realizadas en el marco de algunas de estas asignaturas han dado lugar a algunas publicaciones científicas, las cuales se incluyen en el apéndice B.

1.5. Organización de la tesis

Adicionalmente a este capítulo introductorio, el presente trabajo consta de tres capítulos más, los cuales son descritos a continuación:

Capítulo 2 Estado del arte de la detección de plagio.

En este capítulo se describen los dos principales enfoques de la detección de plagio: el análisis intrínseco de plagio y la detección de plagio con referencia. Debido a que nuestras investigaciones actuales están más orientadas al segundo enfoque, es el que se describe con mayor detalle. Además del estado del arte en la detección de plagio, incluye la descripción de dos corpus estándares para el diseño, puesta a punto y evaluación de los métodos diseñados para abordar la tarea de detección de plagio.

Capítulo 3 Aproximaciones a la detección automática de plagio.

Este capítulo contiene descripciones de nuestros primeros esfuerzos en esta tarea. Debido a que en el desarrollo de varios de ellos se utilizaron corpus conformados *ad hoc* por nosotros mismos (algunos de ellos adaptados de otros existentes), se incluye su descripción. Los métodos abordados en este capítulo incluyen uno basado en modelos de lenguaje, otro basado en la comparación exhaustiva de n-gramas y un último basado en el cálculo de distancias entre distribuciones de probabilidad.

Capítulo 4 Líneas de investigación abiertas.

En este capítulo se describen las brechas que, a nuestra consideración, siguen abiertas en el tema de la detección automática de plagio. Incluye el planteamiento de la generación de métodos de reducción en el espacio de búsqueda así como de métodos capaces de detectar plagios translingües, en los que los documentos originales y sospechosos están escritos en diferentes idiomas. Dada la necesidad de la conformación de corpus con características especiales para el desarrollo de esta área en general, se plantea también la creación de nuevos corpus. Dada la complejidad de estas tres tareas, serán abordadas con mayor profundidad dentro de la investigación doctoral.

Capítulo 2

Estado del arte de la detección de plagio

Cuando nos referimos a la tarea de detección automática de plagio, no podemos hacerlo de manera totalmente independiente a su "tarea hermana": la atribución de autoría [48]. La tarea de atribución de autoría consiste en determinar cuál es el autor de un texto dado por medio de información recabada previamente. Esta información proviene principalmente de otros textos escritos tanto por el mismo autor del texto en cuestión como por otros autores. En ambos casos, un texto es analizado con el afán de determinar quién es el autor de un fragmento determinado. Algunos casos célebres de la detección de autoría incluyen la búsqueda por esclarecer la autoría de obras relacionadas con W. Shakespeare o la atribución de la autoría de los conocidos como *Federalist papers*[1], que fueron publicados en el siglo XVIII para persuadir en la ratificación de la constitución estadounidense.

En el caso que nos atañe en este momento, la detección de plagio, la tarea no implica únicamente el análisis de textos completos para determinar si han sido escritos por un autor determinado o no, sino que busca analizar un documento (o uno de sus fragmentos) para intentar determinar si realmente fue escrito por el autor que reclama haberlo hecho. Existen dos vertientes principales que buscan dar solución a este problema que, debido a su naturaleza, no son capaces de ofrecer el mismo tipo de información tras el análisis realizado. El primero de ellos es el conocido como *análisis intrínseco de plagio*, en el que el único recurso utilizado es el texto sospechoso por sí mismo. El segundo de ellos es la *detección de plagio con referencia* en donde se requiere contar con un conjunto de documentos originales con el afán de buscar el origen de los fragmentos potencialmente plagiados dentro de un texto sospechoso. En el caso del análisis intrínseco de plagio (sección 2.1), sólo es posible hallar fragmentos que son sospechosos de ser plagiados. Por medio de la detección de plagio con referencia (sección 2.2), es posible obtener además el origen potencial de un fragmento de texto considerado como candidato a ser un caso de plagio.

[1] http://www.foundingfathers.info/federalistpapers/

Para la certera detección automática de plagio, es de vital importancia seleccionar un conjunto de características del texto que sean capaces de discriminar textos plagiados de originales. Clough [11] ha delimitado un conjunto de características que pueden ser explotadas para localizar potenciales casos de plagio. Si bien Clough las orienta al ámbito académico, su aplicación se extiende de manera directa a cualquier otro. Las características son:

1. Vocabulario utilizado. Analizar el vocabulario utilizado en alguna tarea con respecto a documentos escritos previamente por el mismo estudiante. La existencia de una alta cantidad de vocabulario nuevo podría ayudar a determinar si un estudiante realmente escribió un texto o no.
2. Cambios de vocabulario. Si el vocabulario utilizado en un texto cambia significativamente a través de un documento.
3. Texto incoherente. Si un texto fluye de manera inconsistente o confusa.
4. Puntuación. Es muy poco probable que dos autores utilicen los signos de puntuación exactamente de la misma manera.
5. Cantidad de texto común entre documentos. Es poco frecuente que dos documentos escritos de manera independiente compartan grandes cantidades de texto.
6. Errores en común. Resulta muy improbable que dos textos independientes tengan los mismos errores de escritura (errores de dedo, por ejemplo).
7. Distribución de las palabras. Es poco frecuente que la distribución en el uso de las palabras a través de textos escritos independientemente sea la misma.
8. Estructura sintáctica del texto. Un indicador de plagio es que dos textos compartan una estructura sintáctica común.
9. Largas secuencias de texto en común. Es poco probable que dos textos independientes (incluso cuando traten el mismo tema), compartan largas secuencias de caracteres o palabras consecutivas.
10. Orden de similitud entre textos. Si existe un conjunto significativo de palabras o frases comunes en dos textos, puede haber un caso de plagio.
11. Dependencia entre ciertas palabras y frases. Un autor tiene preferencias sobre el uso de ciertas palabras y frases. Encontrarlas en un trabajo realizado por otro, debe ser considerado sospechoso.
12. Frecuencia de palabras. Es poco común que las palabras halladas en dos textos independientes sean usadas con la misma frecuencia.
13. Preferencia por el uso de sentencias cortas o largas. Los autores pueden tener una marcada preferencia sobre la longitud de las sentencias. Dicha longitud podría ser poco usual en compañía de otras características.
14. Legibilidad del texto. Resulta improbable que dos autores compartan las mismas medidas de legibilidad, tales como los índices de *Gunning* [51], *Flesch* [16] o *SMOG* [53].
15. Referencias incongruentes. La aparición de referencias en el texto que no se encuentran en la bibliografía o viceversa son disparadores de un posible caso de plagio.

La utilidad de la característica 1 puede ser puesta en entredicho. Se supone que un estudiante aprende a través del tiempo y, debido a esta razón, su vocabulario debe verse incrementado. Por otra parte, las características 1, 5 y 9 representan buenos ejemplos de las dificultades enfrentadas en la tarea de detección automática de plagio: la necesidad de realizar comparaciones de manera exhaustiva entre documentos sospechosos y originales.

Las características señaladas en esta lista, en conjunto con algunas otras, han sido explotadas por diferentes enfoques para la detección de plagio. De hecho, Maurer et al. [32] han realizado una clasificación de métodos para la detección de plagio que está conformada por tres categorías principales:

- La comparación exhaustiva entre documentos sospechosos y documentos de referencia.
- La definición de un fragmento de texto característico en un documento sospechoso para buscarlo en la Web.
- La realización de un análisis de estilo, el cual es conocido como *estilometría*.

Esta clasificación lleva a dividir la tarea de detección automática de plagio en los dos conjuntos que hemos definido previamente: análisis intrínseco de plagio y detección de plagio con referencia.

2.1. Análisis intrínseco de plagio

Como se observará más adelante (específicamente en la sección 3.4), una de las mayores dificultades en la detección de textos plagiados es la enorme cantidad de documentos originales que deben ser considerados con el objetivo de determinar si pueden ser su fuente. Por ello, algunas investigaciones se han basado en el análisis de plagio sin tomar en cuenta ningún documento original con el cual hacer alguna comparación. Este es el principio básico del *análisis intrínseco de plagio*.

En análisis intrínseco de plagio se basa en un hecho muy común en el ámbito académico (e incluso fuera de él): una persona es capaz de detectar que un documento es irregular (sospechoso de plagio), por el simple hecho de leerlo. Para explicar la manera en la que esto ocurre, la figura 2.1 muestra un ejemplo[2].

No es necesario realizar un análisis profundo de este texto para considerar que contiene fragmentos plagiados, una simple lectura hace sospecharlo. Por ejemplo, al principio del primer párrafo se habla del trabajo en primera persona del plural -*hemos hecho*- mientras que en la última sentencia se habla en primera persona singular -*mi teoría*-. Por si fuera poco, en el último párrafo se vuelve a utilizar el plural -*nos parece*-. Además, la complejidad y estilo entre el primero y último fragmentos con respecto al segundo y tercero, definitivamente no

[2]Este texto es completamente sintético y ha sido escrito con afanes enteramente ilustrativos. Algunos fragmentos fueron obtenidos de las páginas Web http://www.aquimama.com/bebes-0a12-meses/agua01.shtml y http://es.wikipedia.org/wiki/Mineral_(nutriente).

...
En este trabajo, hemos hecho una investigación acerca de la influencia que tiene la cantidad de sales minerales en el humor de las mujeres. Para la investigación he trabajado con 5 mujeres que han tomado agua con distinta cantidad de sales minerales. Mi teoría es que entre más sales minerales haya en el agua, las mujeres son más volubles.
...
Las sales minerales son moléculas inorgánicas de fácil ionización en presencia de agua y que en los seres vivos aparecen tanto precipitadas como disueltas. Las sales minerales disueltas en agua siempre están ionizadas. Estas sales tienen función estructural y funciones de regulación del pH, de la presión osmótica y de reacciones bioquímicas, en las que intervienen iones específicos. Participan en reacciones químicas a niveles electrolíticos.
...
El agua mineral de mineralización muy débil (inferior a 50 mg/l de residuo seco). Es muy diurética y está indicada en el tratamiento de trastornos como la hipertensión y los cálculos renales. El agua oligometálica (menos de 500 mg/l de residuo seco). Es la más adecuada para un consumo diario. El agua mineral de mineralización fuerte (más de 1.500 mg/l de residuo seco). Su consumo puede complementar al de las aguas oligometálicas en determinados períodos, por ejemplo, en verano, cuando, a través del sudor, el organismo pierde una mayor cantidad de minerales.
...
Nos parece que los resultados son buenos.
...

Figura 2.1: Ejemplo de texto plagiado detectable por la simple lectura.

son los mismos. Mientras los fragmentos de los extremos se muestran sencillos y hasta cierto punto coloquiales, los centrales son mucho más formales y técnicos. El estilo de escritura y la complejidad de un texto son características claves para detectar un posible caso de plagio.

La idea principal en el análisis intrínseco de plagio es precisamente capturar el estilo y la complejidad a través de un documento sospechoso con el afán de encontrar fragmentos inusuales que sean candidatos a ser casos de plagio. Uno de los pocos trabajos basados en este enfoque es el realizado por Meyer zu Eissen et al. [34]. En esta investigación el estilo y complejidad de un texto son medidas con base en un conjunto de parámetros que intentan medir los aspectos anteriormente señalados. Dado un documento sospechoso s, los parámetros a considerar son:

1. Promedio de clases de palabras basado en la frecuencia. Cada palabra $w \in s$ es asignada a una clase denominada $c(w)$. La clase asignada a cada palabra depende de su frecuencia en el documento. La palabra (o conjunto de palabras) que presente la máxima frecuencia de aparición en s es denominada w^* y se asocia a la clase c_0. El resto de palabras en s es asignado a la clase cuyo subíndice se determina por $\lfloor log_2(f(w^*)/f(w)) \rfloor$, en donde $\lfloor \cdot \rfloor$ es la función piso. Esta medida refleja la complejidad y el tamaño del vocabulario

de un documento. Se utiliza debido a que suele ser bastante estable cuando se analiza un documento original, escrito por un único autor, sin importar su longitud.

2. Longitud de las sentencias. El promedio de la longitud de sentencias suele ser relativamente uniforme a través de un documento escrito por un autor.
3. Partes de la oración. Las categorías gramaticales utilizadas hablan del estilo de escritura de un autor.
4. Número promedio de palabras de paro. El uso de determinados artículos, preposiciones y otras palabras que no aportan demasiado significado a un texto puede ser completamente diferente de un autor a otro.
5. Índice de confusión de Gunning. Esta medida ha sido diseñada para determinar qué tan comprensible es un texto escrito (particularmente en inglés). El valor obtenido por dicha medida es una aproximación al número de años de educación formal que una persona requiere para comprender un texto leyéndolo una sola vez.

   Dicho índice se calcula tomando una muestra de texto de alrededor de 100 palabras por medio de la siguiente ecuación:

   $$I_G = 0.4 \left(\frac{|palabras|}{|sentencias|} + 100 * \frac{|palabras\, complejas|}{|palabras|} \right) \tag{2.1}$$

   donde $|\cdot|$ es el número de elementos $\cdot$ en la muestra de texto. Se considera que una palabra compleja contiene al menos tres sílabas (a excepción de los nombres propios, palabras compuestas o con sufijos como *es*, *ed* o *ing*).

   Por ejemplo, la revista Newsweek tiene un índice $I_G(Newsweek) = 10$ mientras que un comic tiene aproximadamente $I_G(comic) = 6$ [51].
6. Índice de Flesch-Kincaid. Este índice es muy parecido al anterior y de nuevo intenta calcular los años de educación necesarios para comprender un documento. Su cálculo se hace por medio de la siguiente ecuación:

   $$I_{FK} = 1.599\lambda - 1.015\beta - 31.517 \tag{2.2}$$

   donde λ es el número promedio de palabras de una sílaba por cada cien palabras y β es la longitud promedio de las sentencias en número de palabras [16].
7. Índice de Dale-Chall. Este índice fue diseñado en los años cuarenta para determinar de nuevo los años de estudio necesarios para leer un texto [15]. La fórmula para su cálculo es:

   $$I_{DC} = 0.0496\beta + 0.1579\phi + 3.6365 \tag{2.3}$$

   donde ϕ el porcentaje de "palabras difíciles" en el texto (es necesario definir previamente un vocabulario con estas palabras).
8. Función R. Esta medida propuesta por Honore [23] intenta capturar la variedad en el vocabulario de un autor. Se calcula por medio de la siguiente ecuación:

   $$R = \frac{100\, log(M)}{M^2} \tag{2.4}$$

   donde M es el número de palabras en el texto analizado.

9. Función K. Esta función definida por Yule [55] es en realidad una alternativa para el cálculo de la riqueza de vocabulario obtenida por medio de la función R. Se calcula de la siguiente manera:

$$K = \frac{10^4 \left(\sum_{i=1}^{\infty} i^2 V_i - M\right)}{M^2} \tag{2.5}$$

donde V_i es el número de palabras que aparece i veces en el texto. M tiene el mismo significado que en el punto anterior[3] .

Estas medidas son calculadas primero considerando el texto completo del documento sospechoso, luego se calculan para cada párrafo. Con los resultados obtenidos, se realiza una comparación para buscar las variaciones que puedan ser reflejo de fragmentos que no fueron escritos por el mismo autor, es decir, candidatos a estar plagiados.

En un segundo trabajo publicado por los mismos autores [49], este método se complementa con un proceso de identificación de autoría. Dicho proceso se lleva a cabo con base en el análisis de los fragmentos considerados sospechosos con respecto a los que no lo son. De esta manera, se busca confirmar el resultado obtenido previamente. Si tomamos en cuenta las características que Clough considera útiles en la tarea de detección de plagio [11], el análisis intrínseco de plagio considera las características 2, 7, 12 y 14.

Hay un aspecto importante en este enfoque al que se debe poner especial atención: de ninguna manera el análisis intrínseco de plagio es capaz de *demostrar* que un fragmento de texto está plagiado. La razón es simple, dado que por su esencia este enfoque no considera algún tipo de comparación de los documentos sospechosos con respecto a documentos originales, no es posible encontrar el potencial documento original que sirvió como fuente de un fragmento plagiado.

2.2. Detección de plagio con referencia

Una de las primeras ideas que pueden surgir cuando se busca resolver la tarea de detección de plagio es realizar una comparación de un texto sospechoso con un conjunto de textos originales. Éste es precisamente el principio básico de esta aproximación a la detección de plagio: la *detección de plagio con referencia*.

Dado un corpus D conformado por un conjunto de documentos originales[4] y un documento sospechoso s, la tarea de detección de plagio puede reducirse a realizar una comparación exhaustiva del texto en s sobre el corpus D para responder a la pregunta: ¿Existe algún fragmento $s_i \in s$ que esté incluido en algún documento de D?

[3]Las fórmulas para el cálculo de las funciones R y K se han obtenido de [46, sección 2].

[4]En realidad el hecho de que dichos documentos deban ser originales está en entredicho. Basta con que un fragmento del texto analizado sea hallado en otro para determinar que se trata de un caso de plagio.

2.2.1. Análisis a nivel de documentos

Existen varias investigaciones que abordan la detección de casos de plagio desde este punto de vista. Quizás uno de los primeros desarrollos a este respecto sea *SCAM* [44], el cual fue desarrollado en 1995 con el objetivo de detectar documentos duplicados. Los autores de SCAM describieron dos entornos en los que su herramienta sería útil. El primero de ellos es el de una librería con venta de libros en formato electrónico a través de Internet. Una persona que compre un libro podría ponerlo disponible en algún otro sitio de manera gratuita y la librería tendría gran interés en saberlo. Otro entorno que resulta aún más interesante es el de la búsqueda de información. A menudo es común encontrar, por medio de la búsqueda de contenidos en cualquier buscador de Internet actual, un conjunto de páginas alojadas en distintos servidores que contienen prácticamente la misma información. En este caso, el detector de duplicados ahorraría tiempo de búsqueda al usuario[5].

SCAM es capaz de detectar relaciones de *plagio*, *subconjunto*, *copia* y *relación*, los cuales significan que un documento contiene algunas partes de otro, está contenido en otro, es una copia de él o está fuertemente relacionado. Esto se hace con base en lo que sus autores han llamado *modelo de frecuencia relativa*. Este enfoque se basa principalmente en una adaptación de la medida de similitud de coseno. El método se basa en definir un conjunto compuesto por aquellas palabras que muestren un número de ocurrencias similar en cada uno de los dos documentos analizados. Si dos documentos muestran frecuencias similares de ocurrencia de un conjunto de palabras, es muy probable que se trate de distintas versiones de un mismo texto.

El análisis realizado por SCAM se da principalmente a nivel de documento. Sin embargo, esto no es siempre suficiente. Para que exista un caso de plagio, no es necesario que se halle un duplicado de un documento entero, basta con que un fragmento de texto sea extraído de otro. Por ello, se han desarrollado otros enfoques que realizan un análisis a un nivel inferior, los cuales se abordan en las siguientes secciones.

Como ya se ha señalado, los fragmentos plagiados pueden aparecer mezclados entre texto nuevo e incluso modificados a distintos niveles. Por ejemplo, el corpus METER (que será descrito con mayor detalle en la sección 2.3), considera dos niveles de reuso de texto a nivel de fragmento: la copia exacta (conocida como *verbatim*) y la reescritura (*rewording*). Para encontrar fragmentos plagiados con estas características, un análisis a un nivel más bajo debe ser realizado.

2.2.2. Análisis basado en comparación de n-gramas

Para realizar una estrategia de búsqueda flexible, Lyon et. al [29] basan la comparación de documentos en los n-gramas contenidos en ellos. La justificación para realizar esto es que

[5]Curiosamente, SCAM fue desarrollado, entre otros, por Héctor García-Molina, quien fue uno de los profesores de los fundadores de Google. Sin embargo, parece que por alguna razón este buscador no cuenta con esta característica.

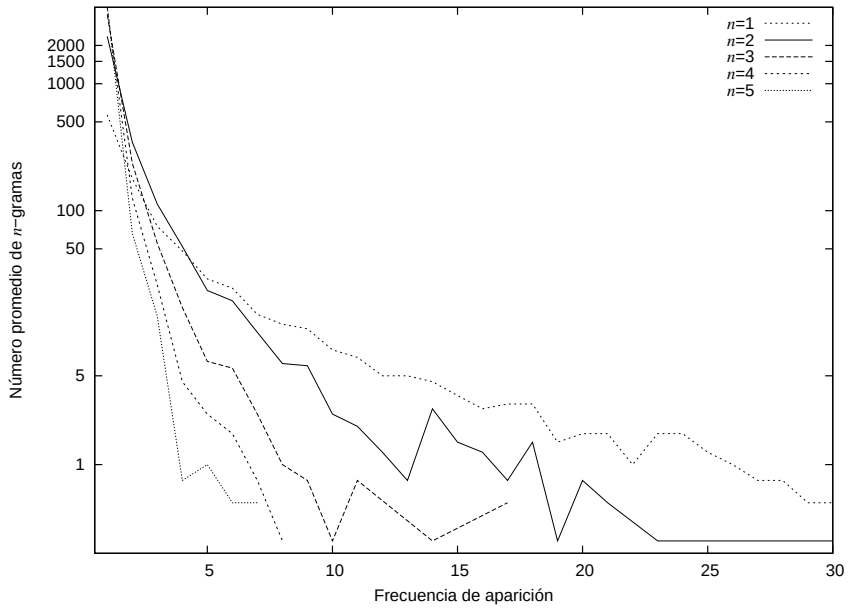


Figura 2.2: Distribución de n-gramas en un conjunto de textos sobre el mismo tema (n = grado el n-grama)

Tabla 2.1: Número de n-gramas que varios documentos tienen en común (normalizado por el número total de n-gramas en todos los documentos)

Documentos	1-gramas	2-gramas	3-gramas	4-gramas
2	0.1692	0.1125	0.0574	0.0312
3	0.0720	0.0302	0.0093	0.0027
4	0.0739	0.0166	0.0031	0.0004

dos textos independientes tienen un nivel muy bajo de n-gramas en común, siempre y cuando se considere un valor $n > 1$. De hecho, la frecuencia de aparición de n-gramas en un mismo documento suele ser muy baja. Este fenómeno se puede observar con claridad en la figura 2.2. En ella se muestra el número promedio de n-gramas ocurrido con una determinada frecuencia en cuatro documentos escritos por el mismo autor y sobre el mismo tema.

A medida que el grado de los n-gramas considerados se eleva, la mayoría de ellos tiende a ser único. Por ende, la probabilidad de aparición de un n-grama en documentos distintos (incluso escritos por el mismo autor) es menor mientras el valor de n sea mayor. Prueba de ello es la tabla 2.1, que muestra el número de n-gramas que aparecen en varios de los cuatro documentos. Resulta claro que la probabilidad de encontrar un bigrama en varios documentos es mucho más alta que la de encontrar un tetragrama.

Los documentos utilizados en este análisis, los cuales fueron seleccionados de entre los que aparecen en el apéndice B, contienen en promedio 3,728 palabras.

Bajo esta justificación, Lyon et al. [30, 29] basan su detección de potenciales plagios en la comparación exhaustiva de n-gramas, lo cual se ve reflejado en su prototipo *Ferret*. En particular, estos investigadores señalan que los mejores resultados se obtienen considerando trigramas[6]. Así, tanto el documento sospechoso s como cada uno de los documentos de referencia $d \in D$ son codificados en forma de n-gramas para luego compararlos. Con el objetivo de determinar si el documento s puede estar plagiado del documento d se han propuesto dos medidas principales: semejanza (R) y contención (C) [30] (en la literatura original estas medidas se denominan *resemblance* y *containment* respectivamente).

La medida de semejanza es útil cuando los conjuntos de n-gramas a comparar provienen de textos de longitud equiparable (comparaciones documento a documento, sentencia a sentencia, etc.). Considerando un documento de referencia d y uno sospechoso s, la semejanza se define por medio de la ecuación 2.6. La R hace referencia a Resemblance.

$$R(s \mid d) = \frac{|N(d) \cap N(s)|}{|N(d) \cup N(s)|} \tag{2.6}$$

donde $N(\cdot)$ es el conjunto de n-gramas en el documento $\cdot$.

Claramente, la semejanza es simplemente el cálculo del conocido como coeficiente de Jaccard [25] entre los conjuntos de n-gramas de ambos documentos.

En caso de que los textos que se deseen comparar no tengan una longitud equiparable (comparaciones sentencia a documento, por ejemplo), la opción es la medida de contención, la cual se define en la ecuación 2.7.

$$C(s_i \mid d) = \frac{|N(s_i) \cap N(d)|}{|N(s_i)|} \tag{2.7}$$

donde s_i es alguna de las sentencias en el documento sospechoso s.

Tanto la semejanza como la contención son valores dentro del intervalo $[0, 1]$. Es necesario definir un umbral dentro de este intervalo tal que al ser superado, se considere que el texto sospechoso es un candidato a haber sido plagiado a partir del texto de referencia. En este caso se consideran las características 5, 9 y 10 de Clough.

Esta técnica de detección de plagio ha dado buenos resultados. Por ello, hemos realizado varios experimentos con base en la medida de contención. Dichos experimentos se describen en la sección 3.3.

2.2.3. Determinando el tipo de plagio a nivel de sentencia

Para Kang et al. [27] el proceso de detección de plagio debe realizarse a nivel de sentencia (análisis de una sentencia sospechosa s_i con respecto a cada una de las sentencias de referencia

[6]En la sección 3.3 se discute con mayor profundidad la conveniencia de considerar otros grados de n-gramas.

d_j). Uno de los aspectos que diferencian esta investigación es que *PPChecker*, la herramienta que estos investigadores han desarrollado, no sólo busca encontrar fragmentos plagiados, sino que intenta determinar de qué tipo de plagio se trata. En particular, se consideran los siguientes tipos de sentencia plagiada:

1. Copia exacta (*verbatim*). En este caso, s_i y d_j son idénticas.
2. Copia con inserción de palabras. A la sentencia d_j se le han agregado palabras para generar s_i.
3. Copia con eliminación de palabras. Lo inverso al caso anterior, en lugar de agregar palabras, se eliminan.
4. Reescritura. La sentencia s_i expresa exactamente lo mismo que d_j. Esto se debe al cambio de palabras por sus sinónimos o la sustitución de ciertas palabras (como preposiciones o artículos) por otras.

Este enfoque también considera la longitud de las sentencias comparadas. La razón es simple: dadas las sentencias de referencia d_1 y d_2 y las sentencias sospechosas s_1 y s_2 con longitudes $| \cdot_1 | = 2$ y $| \cdot_2 | = 10$ ($\cdot$ es un comodín que puede ser interpretado como d o s), si s_1 es exactamente igual a d_1 y 8 de las 10 palabras de s_2 coinciden con las de d_2, es más factible que s_2 sea un verdadero caso de plagio que s_1. Sin embargo, medidas como la del coseno concluirían lo contrario. De manera similar a *Ferret*, *PPChecker* considera tanto la intersección como la unión de los vocabularios. Con esto sería suficiente para detectar copias exactas además de aquellas que se hayan hecho con inserción o eliminación de palabras.

Uno de los aspectos más interesantes de la investigación de Kang et al. es el método para detectar copias reescritas. Una de las etapas iniciales del método diseñado consiste en la expansión del vocabulario de las sentencias. Dicha expansión se hace por medio de la base léxica de Wordnet[7]. De esta manera, las comparaciones de vocabulario entre las sentencias no se realizan únicamente sobre las palabras que aparecen en ellas, sino además con todas las que se les relacionen en Wordnet. En esta investigación se consideran las características 5, 9 y 10 de Clough.

2.2.4. Acelerando el proceso de detección de plagio

Un factor a considerar en todos los enfoques anteriores es que ninguno de ellos se preocupa por el serio problema del tiempo de procesamiento. No obstante que los resultados obtenidos sean buenos en términos de calidad, no se menciona nada con respecto a la velocidad con la que son obtenidos. Como ya se ha señalado, uno de los factores más importantes para la detección de plagio cuando se considera un corpus de referencia es precisamente la longitud de dicho corpus. Entre más documentos de referencia se tengan, más probable será que un texto plagiado sea detectado. Sin embargo, esta es un arma de dos filos. Si el corpus de

[7] http://wordnet.princeton.edu/

Tabla 2.2: Producto punto entre una frase original y una potencialmente plagiada

	Plagiar	es	robar	el	trabajo	ajeno
Plagiar	■	·	·	·	·	·
es	·	■	·	·	·	·
robar	·	·	■	·	·	·
el	·	·	·	■	·	·
trabajo	·	·	·	·	■	·
de	·	·	·	·	·	·
otro	·	·	·	·	·	·

referencia es muy grande, el tiempo necesario para realizar las búsquedas (sea con base en documentos enteros o cualquier tipo de fragmento), puede no ser adecuado en la práctica.

El enfoque de Si et al. [45] a la detección de plagio trata implícitamente de evitar este problema. Antes de describir el método, cabe señalar que *CHECK*, el prototipo desarrollado por Si et al., surgió por la "necesidad de detener las copias de texto tan fáciles de hacer gracias a la existencia de la Web y los navegadores como Netscape" [45, sección 1]. Este hecho refleja un poco la época en la que CHECK fue desarrollado: 1997.

El proceso de búsqueda de CHECK se basa en la estructura de los documentos. En este caso el corpus de referencia D está compuesto por un conjunto de archivos escritos en LaTeX. De esta manera, dado un documento sospechoso s (que también debe ser un documento LaTeX), el primer análisis se basa en la comparación de la estructura de s con respecto a la de los documentos en D. Cuando esta comparación, la cual se realiza por medio de estructuras arbóreas, coincide hasta llegar a una hoja, los fragmentos hallados se comparan de manera exhaustiva. Esta comparación se realiza por medio de la técnica producto punto, la cual tiene una gran similitud (más que gráfica) con los métodos de alineación desarrollados para la traducción automática de carácter estadístico [9]. Un ejemplo de comparación de vocabulario de dos sentencias con base en la técnica de producto punto puede observarse en la tabla 2.2.

Si el resultado del producto punto excede cierto umbral de similitud, s_i se considera un caso de plagio. El método utilizado por CHECK se basa en las características 4, 5 y 8 de Clough.

Como ya se habrá inferido, CHECK tiene una enorme debilidad: únicamente es capaz de procesar documentos LaTeX y es inútil con cualquier otro formato de texto, como el simple texto plano.

Una opción totalmente diferente para intentar acelerar el proceso de búsqueda de fragmentos plagiados es el uso de "huellas digitales" (*fingerprint*) [47]. En vez de realizar las comparaciones sobre trozos cadenas de texto (*chunks*), éstas se realizan sobre valores numéricos que son asociados a ellas. La comparación entre números es mucho más rápida que la de

cadenas de texto.

2.3. Recursos disponibles

El principal recurso necesario en el diseño, puesta a punto y evaluación de los métodos de detección de plagio es un conjunto de documentos originales que puedan ser utilizados como corpus de referencia así como un conjunto de documentos que tengan fragmentos plagiados (por supuesto, provenientes de los documentos originales señalados previamente), los cuales puedan ser utilizados como documentos sospechosos.

Como se señaló en el capítulo 1, uno de los ámbitos en los que con mayor frecuencia se encuentran casos de plagio es el escolar. Diversas investigaciones tales como [30] y [54] se han valido de los reportes de estudiantes para conformar su corpus, ya sea para ajustar o para evaluar sus métodos. Sin embargo, no es sencillo obtener uno de estos corpus. Además, por razones éticas, un investigador no está dispuesto a distribuir los trabajos de sus alumnos en los que se demuestre que han hecho trampa.

Por ello, suele ser necesario que los corpus explotados sean de distinta naturaleza al verdadero plagio o que los casos se construyan de manera artificial (secciones 2.3.1 y 2.3.2 respectivamente).

2.3.1. El corpus METER

El corpus *METER* [12], cuyo nombre proviene del inglés *MEasuring TExt Reuse* (midiendo el reuso de texto) fue creado dentro del proyeco METER[8] en la Universidad de Sheffield. El principal objetivo de este proyecto era trabajar en la detección y medida del reuso de texto. En realidad no se trata de un corpus de casos de plagio, sino de notas periodísticas. La primera sección del corpus está conformada por un conjunto de noticias escritas por el organismo inglés *Press Association* (PA), una agencia de noticias del Reino Unido. Distintos periódicos, como *The Times*, *The Guardian* o *The Independent*, sólo por mencionar algunos, tienen acuerdos con la PA para utilizar sus notas como la fuente para las publicadas en sus propios periódicos. No existe ninguna limitación para que un periódico publique tal cual la nota proveniente de la PA o para que lo haga previa modificación de ella. Son precisamente las notas publicadas en los periódicos las que conforman la segunda sección del corpus.

Es interesante señalar cómo se clasifican las notas publicadas en los periódicos. Clough et al. [12] las han clasificado dentro de tres niveles principales de reuso de texto que reflejan la relación de cada nota con respecto a la correspondiente de la PA. Este trabajo ha sido realizado por un periodista experto, por lo que puede considerarse bastante confiable. Los tres niveles de reuso en estas notas periodísticas se resumen a continuación:

[8] http://www.dcs.shef.ac.uk/nlp/meter/

1. Completamente derivada. Significa que la nota de periódico fue creada considerando a la versión de la PA como la única fuente.
2. Parcialmente derivada. Implica que la versión de la PA fue utilizada como una de las fuentes de la nota hallada en el periódico.
3. No derivada. Señala que la versión de la PA no ha sido considerada para escribir la nota del periódico.

Si bien estas anotaciones a nivel de documento (nota periodística) son de bastante utilidad, algunas de las notas tienen cada fragmento de texto identificado con respecto a su relación con la nota de la PA. En este caso, los elementos de cada sentencia pueden pertenecer a tres clases diferentes:

1. Copia exacta (*verbatim*). Significa que la nota de periódico fue creada considerando a la versión de la PA como la única fuente.
2. Reescritura (*rewrite*). Implica que la versión de la PA fue utilizada como una de las fuentes de la nota hallada en el periódico.
3. Nueva (*new*). Señala que la versión de la PA no ha sido considerada para escribir la nota del periódico.

El corpus está conformado por alrededor de 1,700 textos acerca de dos temas: leyes y cortes y espectáculos. Los artículos fueron publicados entre julio de 1999 y junio de 2000. Existen varias versiones del corpus, una en texto plano, una en formato SGML y otra en formato XML. Con el objetivo de ilustrar la manera en la que está conformado el corpus, se incluye en la tabla 2.3 un fragmento de nota de la PA seguido de la misma noticia, pero en la versión de *The Telegraph*. Los fragmentos en la nota del Telegraph están identificados como *Rewrite* y *Verbatim*. Su fragmento de origen en la nota de la PA está identificado con **negritas** o *itálicas*, respectivamente.

Como Clough et al. lo señalan, el corpus Meter puede ser utilizado en varias tareas, como es el resumen automático de (múltiples) documentos, la generación automática de encabezados, la estimación de reuso de texto y, por supuesto, la detección automática de plagio. Para el caso particular de la tarea de detección de plagio, las notas "originales" de la PA conforman el corpus de referencia. Mientras tanto, las versiones publicadas por los distintos periódicos conforman el corpus de documentos sospechosos. Cabe señalar que estos no son casos reales de plagio, pero para los fines de las investigaciones al respecto, encajan perfectamente.

2.3.2. El corpus de plagio *Webis*

El corpus Webis [33] fue creado por el Grupo de Sistemas de Información y tecnología Web de la Universidad de Weimar[9]. Dada las complejidad de crear un corpus conformado por plagios reales, este corpus fue generado a partir de plagios sintéticos, es decir, los plagios

[9]http://www.uni-weimar.de/cms/medien/webis/home.html

Tabla 2.3: Nota de la PA en conjunto con la nota de *The Telegraph* correspondiente

Versión PA: Titanic restaurant case discontinued

Celebrity **chef Marco Pierre White** *today* **won** *the battle of* **the Titanic and Atlantic restaurants**. **Oliver Peyton, owner of the Atlantic** Bar and Grill, **had tried to** *sink Marco's* **new Titanic restaurant housed in the same West End hotel in London by seeking damages against landlords Forte Hotels and** *an* **injunction in the High Court**. But today the Atlantic announced in court it had reached a confidential agreement with the landlords and was discontinuing the whole action.
Mr Peyton, whose action began on Monday, had **claimed that the Titanic was a replica of the Atlantic**, with the same art deco style and attracting the same clientele **and should not be allowed to trade in competition** because he has exclusive rights under his lease **at the Regent Palace Hotel** off Piccadilly Circus.

Versión *The Telegraph*

```
<Rewrite PAsource="">  THE </Rewrite>
<Verbatim PAsource=""> chef Marco Pierre White </Verbatim>
<Rewrite PAsource="">  yesterday </Rewrite>
<Verbatim PAsource=""> won </Verbatim>
<Rewrite PAsource="">  a dispute over </Rewrite>
<Verbatim PAsource=""> the Titanic and Atlantic restaurants. </Verbatim>
<Verbatim PAsource=""> Oliver Peyton, owner of the Atlantic, had tried
                       to </Verbatim>
<Rewrite PAsource="">  close White's </Rewrite>
<Verbatim PAsource=""> new Titanic restaurant, housed in the same West End hotel
                       in London, by seeking damages against </Verbatim>
<Rewrite PAsource="">  the </Rewrite>
<Verbatim PAsource=""> landlords, Forte Hotels, and </Verbatim>
<Rewrite PAsource="">  a </Rewrite>
<Verbatim PAsource=""> High Court injunction.</Verbatim>

<Rewrite PAsource="">  He </Rewrite>
<Verbatim PAsource=""> claimed that the Titanic was a replica of the Atlantic
                       and should not be allowed to trade in competition at the
                       Regent Palace Hotel. </Verbatim>
```

Tabla 2.4: Ejemplo de documento plagiado del corpus Webis

··· Specifically what we want to do is investigate if and how parsing can be used as an aid to the retrieval of natural language text.
<inserted source="http://portal.acm.org/..."type="modified">
To assist understanding our results, there is a providing of two baselines for each query/data set. These baseline tests are run without query expansion. For each query set, our system run in both single database mode (search all documents as a single database) and distributed mode (search the highest ranked 10 of 100 collections).
...
tests are necessary to isolate the performance of query expansion without the influence other factors such as merge algorithms.
</inserted>
An overview of what parsing is, can be found in ···

han sido creados de manera artificial. Este corpus fue generado originalmente durante el desarrollo de la investigación descrita en [34]. Los documentos utilizados para su creación son un conjunto de artículos sobre ciencias de la computación provenientes de la biblioteca digital de la ACM[10].

Un conjunto de documentos es utilizado como fuente de fragmentos plagiados que son insertados en aquellos documentos que están destinados a conformar el corpus de documentos sospechosos. En la versión actual existen alrededor de 100 documentos sospechosos que pueden contener casos de plagio conformados por copias exactas o modificadas. La ventaja de este corpus es que, debido precisamente a que es sintético, los casos de plagio están relativamente controlados. Además de versiones en texto plano, apropiadas para realizar las pruebas de los métodos diseñados, contiene versiones XML con los fragmentos plagiados identificados. Por ello, la evaluación de los resultados obtenidos puede resultar una tarea sencilla. Un extracto de uno de los documentos que conforman este corpus se reproduce en la tabla 2.4

[10]http://portal.acm.org/dl.cfm

Capítulo 3

Aproximaciones a la detección automática de plagio

En este capítulo se describen algunos de los trabajos que hemos realizado hasta el momento en la tarea de detección de plagio. Si bien no todos los experimentos han obtenido resultados competitivos, han servido para comprender la problemática a enfrentar. En la sección 3.1 se describen los corpus que se han utilizado en los distintos experimentos. La sección 3.2 contiene los detalles del primer trabajo realizado al respecto, el cual se basa en modelos de lenguaje. La sección 3.3 contiene los resultados de un conjunto de experimentos inspirados en los trabajos de Lyon et al. [30, 29]. Finalmente, la sección 3.4 aborda uno de los subproblemas que menos se han tratado en la literatura: el manejo eficiente del espacio de búsqueda, representado por el corpus de referencia, en la detección de plagio.

3.1. Corpus utilizados en los distintos experimentos

En las investigaciones realizadas se han utilizado principalmente tres corpus: dos sintéticos (secciones 3.1.1 y 3.1.2) y uno real, extracto del corpus METER (sección 3.1.3).

3.1.1. Un corpus sintético basado en el proyecto Gutenberg

Al principio de las investigaciones se creó un pequeño corpus de carácter literario. Este corpus fue conformado con base en obras escritas por W. Shakespeare y por L. Carroll. Todas las obras fueron obtenidas de la página Web del proyecto Gutenberg[1].

El corpus D de documentos de referencia está conformado por las versiones en inglés de los libros *Macbeth* y *Romeo y Julieta*, ambos escritos por W. Shakespeare. El único documento

[1] http://www.gutenberg.org

Tabla 3.1: Estadísticas del corpus sintético literario.

Característica	Valor
Tamaño del corpus de referencia (kb)	150
Tokens	33,173
Tipos	3,335
Tamaño del corpus sospechoso (kb)	263
Tokens	43,993
Tipos	10,652
Tokens en el corpus entero	77,166
Tipos en el corpus entero	12,867

sospechoso es *Alicia en el país de las maravillas*, de L. Carroll. Una breve descripción de estos textos se incluye en la tabla 3.1.

Dentro del documento originalmente escrito por Carroll, hemos insertado de manera aleatoria un conjunto de sentencias "plagiadas". Dichas sentencias provienen de textos escritos por W. Shakespeare y son las que se reproducen a continuación:

1. *I had thought to haue let in some of all Professions, that goe the Primrose way to th' euerlasting Bonfire.*
2. *Mac. We will proceed no further in this Businesse: He hath Honour'd me of late, and I haue bought Golden Opinions from all sorts of people, Which would be worne now in their newest glosse, Not cast aside so soone*
3. *The hearing of my Wife, with your approach: So humbly take my leaue*
4. *If thou could'st Doctor, cast The Water of my Land, finde her Disease, And purge it to a sound and pristine Health, I would applaud thee to the very Eccho*
5. *I had thought to haue let in some of all Professions, that goe to the Primrose way to the euerlasting Bonfire.*
6. *Mac. We will proceed no further in your Businesse: He hath Honour'd me of late, and I haue bought a lot of Golden Opinions from all sorts of people, Which would be worne now in their newest glosse, Not cast aside so soone*
7. *The hearing of my dear Wife, with your approach: So humbly take my leaue*
8. *If thou could'st Doctor, cast The Water from my Land, finde her Disease, And purge it to a sound and pristine Health, I would applaud thee at the very Eccho*
9. *I would give you some violets, but they wither'd all when my father died.*
10. *What is't but to be nothing else but mad?*
11. *To keep my name ungor'd. But till that time I do receive your offer'd love like love, And will not wrong it.*
12. *What is he that builds stronger than either the mason, the shipwright, or the carpenter?*

Los fragmentos 1 al 4 provienen de Macbeth, el cual es uno de los documentos del corpus de referencia, y son copias exactas. Los fragmentos 5 al 8 son los mismos, pero con modificaciones. Finalmente, los fragmentos 9 al 12 son copias exactas de fragmentos de Hamlet, documento que no está incluido en el corpus de referencia.

Tabla 3.2: Estadísticas del corpus especializado.

Característica	Valor
Tamaño del corpus de referencia (kb)	52
Tokens	7,845
Tipos	2,207
Tamaño del corpus sospechoso (kb)	22
Tokens	3,410
Tipos	1,257
Tokens en el corpus entero	11,255
Tipos en el corpus entero	2,905

Si bien este podría ser considerado un corpus de juguete y que los textos escritos por Shakespeare y Carroll son demasiado lejanos y ésta no es una situación de plagio muy realista, este corpus ha servido para realizar algunos experimentos preliminares.

3.1.2. Un corpus sintético basado en documentos especializados

Este corpus fue creado a partir de un conjunto de documentos (específicamente artículos científicos) adscritos dentro del área de la lingüística[2]. Al igual que en el caso anterior, el corpus de referencia está conformado por un conjunto de documentos escritos por un único autor $\mathcal{A}_1$. Por otro lado, el corpus de test está formado por texto del mismo autor $\mathcal{A}_1$ al que manualmente le fueron insertados un conjunto de fragmentos provenientes de documentos escritos por un distinto autor $\mathcal{A}_2$.

Algunas estadísticas de este corpus se incluyen en la tabla 3.2. Como puede observarse, se trata de otro minicorpus que fue diseñado para realizar algunas pruebas iniciales.

3.1.3. Extracto del corpus METER

El otro corpus utilizado fue conformado por una fracción del corpus METER [12] el cual es descrito en la sección 2.3.1. Sólo se ha utilizado la sección de documentos acerca de leyes y corte, la cual es la que se encuentra en formato XML.

El corpus de referencia está compuesto por las 771 notas de la PA halladas en esta versión del corpus. Por su parte, el corpus de documentos sospechosos se compone por 444 notas de periódico. Estas notas se han seleccionado debido a que están anotadas a nivel de fragmento de texto como *verbatim*, *rewrite* o *new* (véase la sección 2.3.1 para más detalles al respecto).

[2]Debido a cuestiones de derechos de autor, en este caso no se dan a conocer abiertamente los autores de los documentos.

Tabla 3.3: Estadísticas del extracto del corpus METER considerados en los experimentos

Característica	Valor
Tamaño del corpus de referencia (kb)	1,311
Número de notas de la PA	771
Tokens	226k
Tipos	25k
Tamaño del corpus sospechoso (kb)	828
Número de notas de periódico	444
Tokens	139k
Tipos	19k
Tokens en el corpus entero	366k
Tipos en el corpus entero	33k

Los fragmentos etiquetados como *verbatim* y *rewrite* son considerados generadores de sentencias plagiadas. Una sentencia s_i en un documento sospechoso s es considerada plagiada si cumple con la desigualdad $|w_v \cap w_r| > 0.4|s_i|$, donde $|\cdot|$ significa longitud, en este caso medida en palabras. $|w_v|$ y $|w_r|$ son el número de palabras en los fragmentos *verbatim* y *rewrite* respectivamente. El contemplar esta condición evita considerar de manera errónea como sentencias plagiadas aquellas que sólo contengan fragmentos comunes incidentales, como entidades nombradas (nombres propios, fechas o lugares, por ejemplo).

Algunas estadísticas de los corpus de referencia y sospechoso extraidos del corpus METER se incluyen en la tabla 3.3. El preprocesamiento en ambos subcorpus consiste en la división de palabras y signos de puntuación ($w, \rightarrow [w][,]$) y un proceso de *stemming* [40][3].

3.2. Modelos de lenguaje aplicados a la detección de plagio

Si bien el enfoque "tradicional" de la detección de plagio con referencia es contar con un corpus de referencia compuesto por documentos originales de diversos autores para luego compararlos con un documento sospechoso escrito por cualquier autor, en este caso hemos diseñado un planteamiento diferente. Dado un documento sospechoso s escrito por el autor $\mathcal{A}$, el corpus de referencia se conforma por un conjunto de documentos escritos previamente por el mismo autor $\mathcal{A}$. De esta manera, se espera que los fragmentos plagiados hallados en s presenten diferencias importantes con respecto a los del resto de documentos escritos por el autor sospechoso.

[3]Para ello, se ha utilizado la implementación del stemmer de Porter de Vivake Gupta, la cual está disponible en http://tartarus.org/∼martin/PorterStemmer/

Los modelos de lenguaje son comúnmente utilizados en tareas de reconocimiento del habla [26] y recuperación de información [39, 22]. Igualmente han sido utilizados en reconocimiento óptico de caracteres [8, 42] y traducción automática [14, 56]. Por si fuera poco, también han sido utilizados en la tarea hermana a la detección de plagio: la atribución de autoría de lengua escrita [35, 13] e incluso de código fuente [18]. En el primer caso, se utilizan modelos de lenguaje de n-gramas de caracteres y perplejidad para determinar la autoría de un documento analizado. En el segundo, la frecuencia de n-gramas a nivel de *byte*.

En nuestro caso, hemos tratado de explotar modelos de lenguaje (tanto n-gramas como perplejidad) a niveles léxico y gramatical para detectar fragmentos plagiados en un texto.

3.2.1. Modelos de lenguaje

Antes de entrar de lleno a la descripción del método diseñado, vale la pena dar una breve introducción a los modelos de lenguaje (*LM*, por sus siglas en inglés). Un modelo de lenguaje estadístico "intenta predecir una palabra dadas las palabras previas" [31]. Para predecir cuál será la siguiente palabra, la mejor opción sería considerar todas las palabras antes de ella en un texto (o los símbolos en una imagen o resultado de un procesamiento de lengua hablada). La probabilidad de una sentencia $w_1 w_2 \dots w_n$, si se conoce $w_{\{1,2,\dots,n-1\}}$ pero no w_n, está dada por la probabilidad condicional de Bayes con base en la regla de la cadena: $P(W) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2) \cdots P(w_n|w_1 \cdots w_{n-1})$. Desafortunadamente, el conjunto de entrenamiento necesario para definir apropiadamente estas probabilidades debe ser extremadamente grande y, sin importar su extensión, nunca se tendría una representación para todas las sentencias posibles en un texto.

La mejor opción es simplemente considerar modelos de lenguaje de n-gramas. Dentro de este marco, el modelo se basa en cadenas conformadas sólo por n palabras, incluyendo la que se busca predecir (valores comunes para n son 2 y 3). La definición de la probabilidad de un n-grama para $n = 3$ es la de la ecuación 3.1.

$$P_3(W) = P(w_{n-2}) \cdot P(w_{n-1}|w_{n-2}) \cdot P(w_n|w_{n-2} w_{n-1}) \tag{3.1}$$

Nuestra idea principal en esta investigación preliminar fue calcular las probabilidades de los n-gramas en un corpus conformado por documentos escritos por un único autor. De esta manera, se contaría con una representación de su vocabulario, frecuencia gramatical y, por ende, su estilo de escritura. Estas representaciones, en forma de modelos de lenguaje, pueden ser luego utilizadas para analizar otros textos con el objetivo de buscar candidatos de fragmentos plagiados.

La cuestión ahora es cómo determinar si un texto es similar a otro. En acorde con Peng et al. [35], hemos optado por utilizar la perplejidad, una manera de expresar la entropía, que es frecuentemente utilizada para evaluar qué tan bien un modelo de lenguaje describe un lenguaje. En nuestro caso, éste es el lenguaje del autor $\mathcal{A}$.

La perplejidad (PP) se calcula por medio de la siguiente ecuación:

$$PP = \sqrt[M]{\prod_{i=1}^{M} \frac{1}{P(w_i|w_i-1)}} \tag{3.2}$$

donde M es el número de palabras en el texto analizado y $P(w_i|w_i - 1)$ es la probabilidad de una palabra w_i dada w_{i-1}. Este es el caso de la perplejidad para modelos de lenguaje de bigramas.

Entre más baja es la perplejidad de un modelo de lenguaje con respecto a un texto, más predecibles son sus palabras. En otras palabras, entre más alta es la perplejidad, mayor es la incertidumbre sobre la siguiente palabra en un texto (véase [31, pp. 60-78] para profundizar en este tema).

Los modelos de lenguaje de los experimentos de esta sección fueron obtenidos por medio de la herramienta SRILM [50][4].

3.2.2. Planteamiento del modelo basado en modelos de lenguaje

Como se ha visto en la sección 3.2.1, una baja perplejidad implica que, dada una secuencia de palabras, un modelo de lenguaje está suficientemente preparado para predecir, con una baja tasa de error, cuál será la siguiente. Bajo esta consideración, hemos definido la siguiente hipótesis:

Hipótesis Sea LM un modelo de lenguaje de un corpus compuesto por un conjunto de documentos D_T escritos por un único autor $\mathcal{A}$, las perplejidades de los fragmentos $s_1, s_2 \in s$, dado que s_1 ha sido escrito por $\mathcal{A}$ y s_2 ha sido plagiado serán claramente diferentes. Específicamente, $PP(s_1) \ll PP(s_2)$.

Hemos considerado tres versiones de los textos en estos experimentos:

i Texto original
ii Etiquetado de partes de la oración del texto
iii Versión lematizada del texto

Dichas versiones buscan representar el estilo de escritura en el texto analizado. Específicamente, se trata de caracterizar el vocabulario del autor y su riqueza sintáctica, (*i*) y (*iii*), así como su estilo morfosintáctico (*ii*). Las etiquetas de partes de la oración y los lemas han sido obtenidos con Treetagger [43].

[4] http://www.speech.sri.com/projects/srilm/

Los modelos de lenguaje fueron calculados de manera independiente para las tres versiones del corpus de entrenamiento considerando $\{2-4\}$-gramas. Los documentos de test fueron divididos en sentencias, incluyendo aquellas que fueron plagiadas.

Dado el modelo de lenguaje obtenido durante la etapa de entrenamiento, se calcula la perplejidad de cada uno de los fragmentos del corpus de test. Como ya se señaló, se espera que los fragmentos plagiados tengan una mayor perplejidad que los escritos por el mismo autor.

3.2.3. Evaluación del método basado en modelos de lenguaje

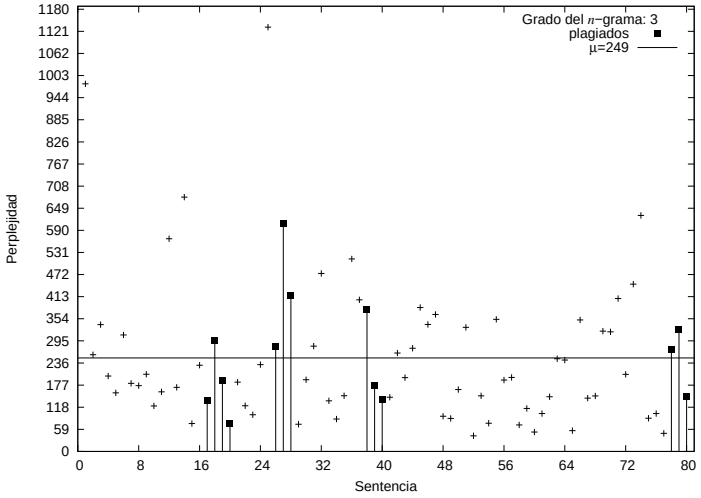
Con el objetivo de probar (o rechazar) nuestra hipótesis, hemos realizado dos experimentos: uno sobre un corpus conformado por artículos científicos (sección 3.1.2) y otro sobre un corpus literario (sección 3.1.1. El primero de ellos se realizó con un corpus pequeño (alrededor de 11,000 tokens), el segundo se realizó sobre un corpus significativamente más extenso (alrededor de 77,000 tokens).

Si bien existen ya trabajos en los que se consideran muchas características en la tarea de detección automática de plagio (tal como [34]), estos experimentos preliminares sólo consideran una característica: la perplejidad. Por esta razón no puede realizarse una comparación directa de los resultados obtenidos por esta técnica con los obtenidos por medio de técnicas más robustas. Nuestro objetivo con estos experimentos no ha sido mejorar el estado del arte en esta tarea, sino determinar si este tipo de caracterización resulta de utilidad para luego combinarla con otras características.

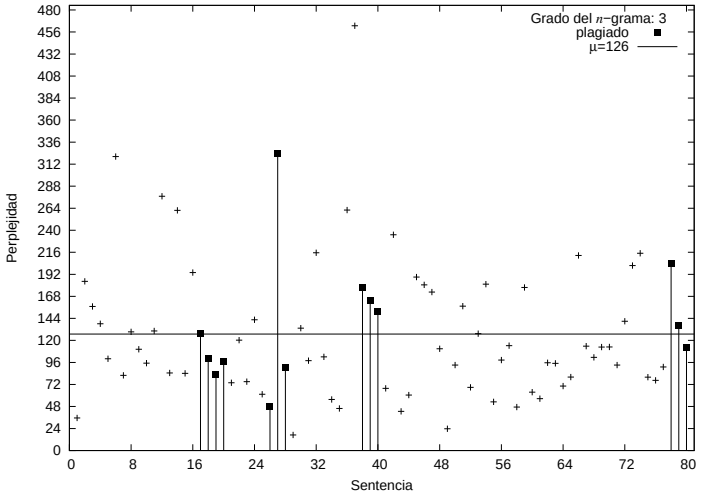
Consideremos el primer experimento, realizado con el corpus de textos especializados (sección 3.1.2). La figura 3.1 muestra la perplejidad de cada sentencia con base en el modelo de lenguaje de trigramas. Debido a que al procesar el documento sospechoso en su versión de texto original se consideran palabras singulares y plurales, femeninos y masculinos y tiempos verbales, las perplejidades obtenidas en este caso (figura 3.1(a)) son las más altas. En particular, los dos valores mayores son $PP_{25} = 1132.15$ y $PP_1 = 980$, donde 25 y 1 representan el número de sentencia en el documento. La sentencia s_{25} está conformada por sólo siete palabras, y contiene una cita del tipo "*autor (2001)*". Dado que dicha cadena no apareció en el corpus de entrenamiento, la probabilidad $P(autor_a \in n-grama)$ tiende a 0. En cuanto al caso de la sentencia s_1, ésta es el título del documento, incluyendo los datos del autor (sólo hemos considerado al punto como separador de sentencias). El autor del documento es francófono, por lo que esta sentencia está llena de palabras escritas en una lengua distinta a la que fue utilizada para calcular el modelo de lenguaje correspondiente.

Si las sentencias se ordenan de manera descendente con base en su perplejidad calculada, la primera sentencia plagiada aparece en la sexta posición. Se trata de la sentencia s_{27} cuya perplejidad es $PP_{27} = 608.21$. Esta sentencia contiene seis palabras que no aparecieron en el corpus de entrenamiento.

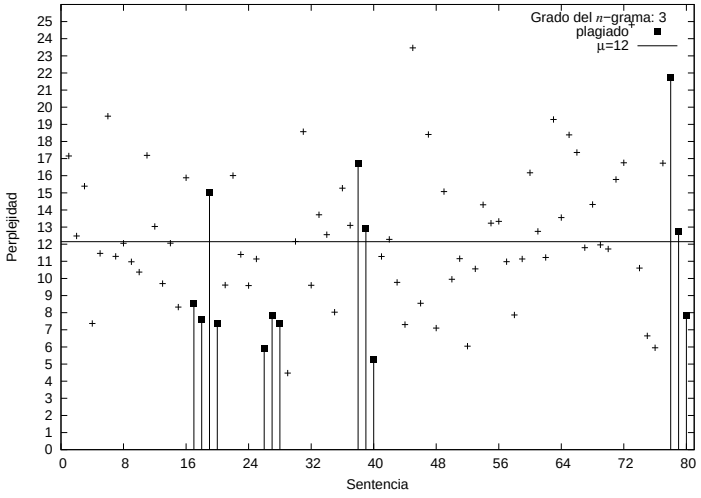
Al experimentar con la versión lematizada del texto (figura 3.1(b)), sólo se considera



(a) original



(b) lematizado



(c) partes de la oración

Figura 3.1: Perplejidad obtenida para los fragmentos de test en el corpus especializado. (+ = sentencia original, ■ = sentencia plagiada)

la riqueza del vocabulario y el estilo de escritura del autor, sin considerar las características adicionales contempladas en el caso anterior (las cuales sólo generan ruido). Las perplejidades más altas en este caso son $PP_{37} = 462.78$ y $PP_{27} = 323.46$. Si bien la sentencia S_{37} no es un caso de plagio, se trata de una cita realizada en el texto hacia otro autor, lo que se refleja de inmediato en la perplejidad obtenida. Como lo señalamos anteriormente, la sentencia s_{27} está plagiada.

Sólo queda analizar el último caso de este experimento, es decir, el realizado con el etiquetado de partes de la oración (figura 3.1(c)). Como es evidente, el vocabulario en este caso es mucho más pequeño que en los anteriores (alrededor de 40 palabras definidas por las categorías gramaticales del etiquetador[5]). Como resultado, los valores de las perplejidades se encuentran dentro de un intervalo mucho menor.

En este caso, las tres perplejidades más altas son $PP_{45} = 23.36$, $PP_{78} = 21.90$ y $PP_6 = 19.46$. En la sentencia s_{45}, compuesta por 20 tokens, hay tres cadenas conformadas por paréntesis y cardinales que son poco comunes (por ejemplo, la cadena *(2)*). Dichas cadenas fueron etiquetadas como *(LS)* (*LS =list item*). La sentencia s_{78} es un caso de plagio que contiene el trigrama *DT NN IN*. Este trigrama es el tercero con menor probabilidad en el corpus de referencia. Por si fuera poco, otros ni siquiera se incluyen en el modelo de lenguaje correspondiente. Es el caso de los trigramas *RB VVZ DT* y *DT RBR JJ* [6], cuya probabilidad está muy cercana a 0.

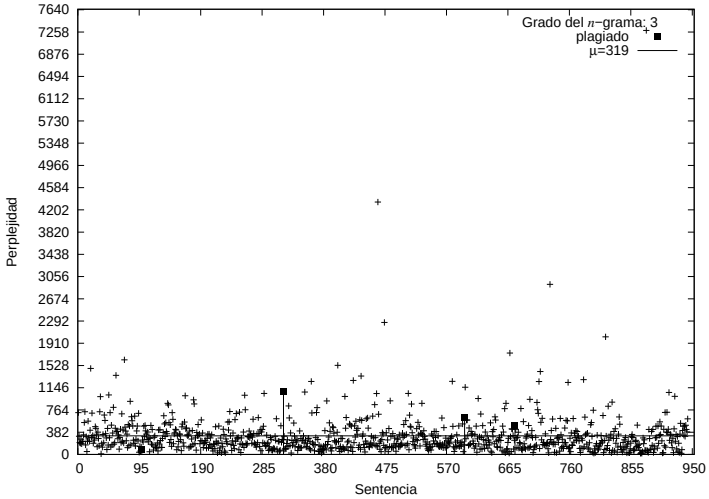
El segundo experimento se realizó con el corpus sintético de textos literarios (sección 3.1.1). En este caso se contemplaron exactamente las mismas condiciones que en el anterior, es decir, utilizar las versiones originales, lematizadas y de partes de la oración del texto, calcular los respectivos modelos de lenguaje con el conjunto de entrenamiento y luego calcular las perplejidades del conjunto de test, el cual contiene fragmentos plagiados. Los resultados pueden observarse en la figura 3.2.

Los resultados obtenidos al analizar el texto original (figura 3.2(a)) confirman lo visto anteriormente: no es una buena idea considerar el texto original en el método basado en modelos de lenguaje puesto que no existen diferencias significativas entre los fragmentos plagiados y los originales. Por otro lado, se puede observar que en el caso en el que se considera la versión lematizada del texto así como sus etiquetas morfosintácticas (figuras 3.1(b) y 3.2(c), respectivamente), los fragmentos plagiados generalmente obtienen altos valores de perplejidad con respecto a los obtenidos para los fragmentos originales. Sin embargo, en los tres casos siguen existiendo fragmentos originales cuya perplejidad es mayor. Hay que considerar, sin embargo, que algunas de las altas perplejidades en estos casos se deben a errores en el proceso de etiquetado y lematización.

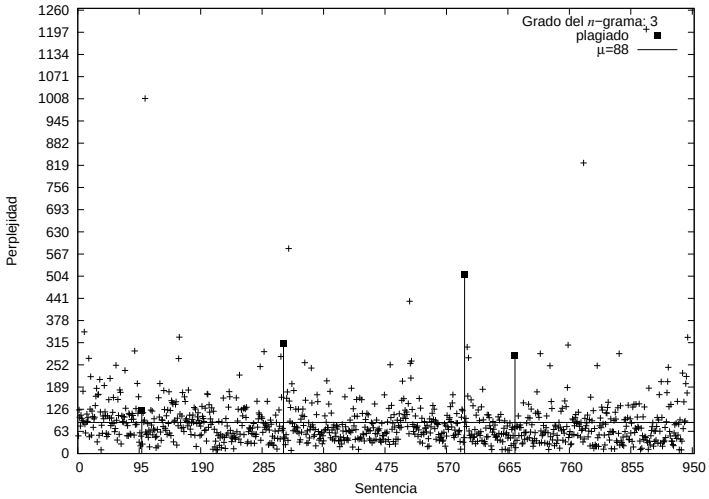
Con el afán de observar los resultados con mayor detalle, la tabla 3.4 incluye las sentencias con mayor perplejidad en la versión de texto lematizado. La primera sentencia contiene palabras comenzadas con mayúscula que fueron erróneamente consideradas como nombres

[5]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps

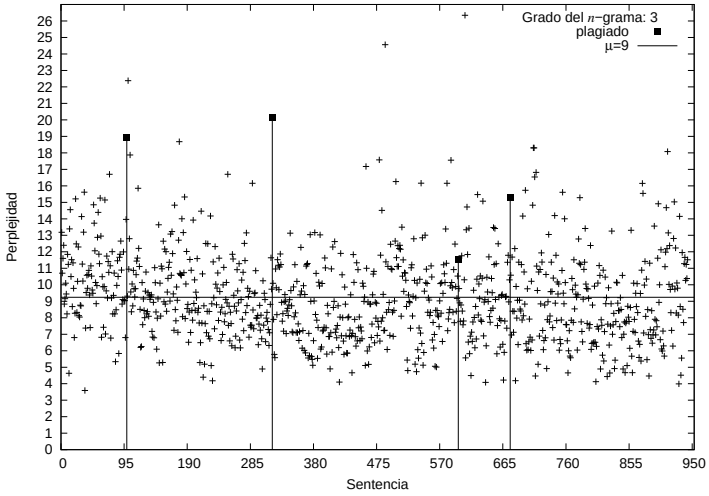
[6]DT=determinante; NN=nombre; IN=preposición; RB=adverbio; VVZ=verbo; RBR=adverbio comparativo; JJ=adjetivo.



(a) original



(b) lematizado



(c) partes de la oración

Figura 3.2: Perplejidad obtenida para los fragmentos de test en el corpus literario. (+ = sentencia original, ■ = sentencia plagiada)

Tabla 3.4: Sentencias lematizadas del corpus literario con las más altas perplejidades

Perplejidad	Sentencia
1205.6	all Persons more Than A Mile High TO leave the court.
1009.1	William 's conduct at first be moderate.
825.6	the twelve juror be all write very busily on slate.
582.5	‘ oh , there go his precious nose ’ ; as an unusually large saucepan fly close by it , and very nearly carry it off.
508.1	the hearing of my wife , with your approach : so humbly take my

propios. La probabilidad de que estos “falsos” nombres propios ocurran en otros documentos es prácticamente 0. Esta es una de las debilidades de considerar las versiones originales y lematizadas de los documentos sospechosos: dado que están conformados por un lenguaje abierto, resulta difícil que un modelo de lenguaje contemple todo el vocabulario y combinaciones posibles.

En el caso de la segunda sentencia de la tabla, todas sus palabras estaban también en el corpus de entrenamiento. Sin embargo en éste *William* nunca apareció al principio de una sentencia y el trigrama *William 's conduct* tampoco. La tercera sentencia contiene la palabra *juror*, que el modelo de lenguaje ignora. Además contiene la palabra *busily*, que tiene una muy baja probabilidad: $P('busily') = 0.0000191$ (para compararla, considérese la probabilidad de *the*: $P('the') = 0.03869$).

La primera sentencia plagiada aparece en la quinta posición. Lo interesante de este caso es que el modelo de lenguaje conoce todas las palabras de la sentencia, sin embargo, los trigramas que la componen tienen muy bajas probabilidades.

Con respecto al análisis del etiquetado morfosintáctico, la mayor perplejidad obtenida es $PP_{608} = 26.34$. La sentencia s_{608} contiene, por ejemplo, el trigrama *DT NN RBR* (determinante, nombre, adverbio comparativo). Dicho trigrama corresponde al fragmento $(that)_1\,(is--''The)_2\,\,(more)_3$, el cual, debido a un error en la separación de las palabras, no fue etiquetado correctamente, resultando en un trigrama con una probabilidad muy baja.

La cuarta sentencia en la lista ordenada por la perplejidad en este caso tiene $PP_{318} = 20.132$. En esta sentencia tanto el estilo como el vocabulario son completamente diferentes a los del resto del documento. Esto se debe a que fue escrita por W. Shakespeare y es uno de los casos de plagio que han sido correctamente detectados.

3.2.4. Discusión sobre el método basado en modelos de lenguaje

Luego de haber realizado los distintos experimentos, se puede observar que los resultados obtenidos por este método no son suficientemente acertados como para determinar correcta-

mente si un fragmento de texto debe ser considerado como original o plagiado.

Es cierto que las perplejidades obtenidas con las tres versiones de los textos analizados (texto original, texto lematizado y etiquetas morfosintácticas) han conducido a la detección de conjuntos de sentencias poco comunes que no están realmente plagiadas. Sin embargo, en la mayoría de los casos estos conjuntos incluyen las plagiadas. Además, las tres variantes del método no siempre localizan las mismas sentencias como candidatas a estar plagiadas. Por ello, consideramos que es necesario que el proceso se componga por las tres variantes para obtener los mejores resultados posibles.

El cálculo de modelos de lenguaje para caracterizar el estilo y vocabulario de un autor $\mathcal{A}$ para luego determinar si un nuevo texto, supuestamente escrito por el mismo $\mathcal{A}$, contiene fragmentos ajenos, no ha obtenido precisamente los mejores resultados. A la salida se obtiene una combinación excesiva de sentencias originales y plagiadas en el conjunto de sentencias consideradas sospechosas. Sin embargo, las sentencias plagiadas tienden a tener algunas de las mayores perplejidades.

Por los resultados es evidente que el espacio de características de la perplejidad no es completamente separable (en cuanto a fragmentos plagiados y originales). Sin embargo, consideramos que los resultados obtenidos al considerar otras características pueden verse mejorados al incluirla.

Esta investigación ha permitido escribir un artículo que fue presentado y publicado en las memorias del taller *PAN: Uncovering Plagiarism, Authorship and Social Software Misuse*, llevado a cabo en Grecia este año [1].

3.3. Búsqueda de plagio basada en n-gramas de palabras

Antes de comezar con la descripción formal de este enfoque, considérese el siguiente ejemplo[7]. Consideremos un escenario clásico de proceso de plagio: un autor $\mathcal{A}$ se encuentra trabajando en un reporte s sobre el oceanógrafo francés *Jacques Cousteau*. $\mathcal{A}$ busca el artículo en Wikipedia sobre Cousteau [52] e incluye en su reporte la sentencia de la figura 3.3, la cual es la sentencia sospechosa $s_i \in s$. Asumamos que el corpus de referencia D incluye un documento d, que es precisamente la página de Wikipedia que fue utilizada como una de las fuentes del trabajo escrito por $\mathcal{A}$. El documento d contiene, entre otros, los fragmentos d_a y d_b:

d_a *"In 1950: he founded [...] and he leased a ship called Calypso from Thomas Loel Guinness for a symbolic one franc a year [...]"*

d_b *"[...] 1996 Calypso was rammed and sunk in Singapore harbor[...]".*

[7]El ejemplo se presenta en un caso en inglés debido a que hasta ahora todos los experimentos que hemos hecho han sido con documentos en este idioma.

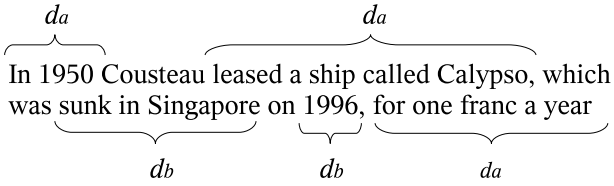


Figura 3.3: Un ejemplo de sentencia plagiada. ($d_{\{a,b\}}$ es el fragmento original que sirvió como fuente de los fragmentos plagiados)

Una técnica de búsqueda podría implicar como etapa inicial, dado que s ya ha sido dividido en sentencias, hacer lo mismo con los documentos de referencia $d \in D$ para luego buscar un par de sentencias s_i, d_j que sean iguales. Como resulta evidente en el ejemplo, esta aproximación no daría buenos resultados debido a que, aunque s_i es un verdadero caso de plagio, se trata de una reescritura a partir de la fuente y no de una copia exacta. Por ello, una comparación de n-gramas en lugar de sentencias completas es la mejor opción. Si se considerara $n = 2$, por ejemplo, la comparación de $N_2(s_i)$ con respecto a $N_2(d_a)$, siendo $N_2(\cdot)$ el conjunto de n-gramas en $\cdot$, los siguientes bigramas serían encontrados: *"In 1950"*, *"leased a"*, *"a ship"*, *"ship called"*, *"called Calypso"*, *"for one"*, *"one franc"*, *"franc a"*, y *"a year"*. Esta intersección entre $N_2(s_i)$ y $N_2(d_a)$ podría llevarnos a considerar que s_i es un plagio de s_a. Sin embargo, el resto de bigramas en $N_2(s_i)$ serían considerados originales, sin importar que hayan sido plagiados del mismo documento (aunque de un fragmento distinto). Cuando se compare $N_2(s_i)$ con $N_2(d_b)$, ocurrirá un caso similar.

3.3.1. Planteamiento del modelo basado en n-gramas

El ejemplo descrito en la sección anterior nos lleva a considerar tres aspectos principales:

1. Conviene que las comparaciones entre un fragmento sospechoso y uno de referencia (sin importar si se trata de sentencias, párrafos o incluso documentos completos), se hagan basadas en n-gramas.
2. Es una buena idea dividir un documento sospechoso s en sentencias para que sean éstas las que se comparen con el corpus de referencia D.
3. Los documentos de referencia $d \in D$ no deben dividirse en sentencias, sino que simplemente deben ser codificados en forma de n-gramas

Estas ideas son básicamente una combinación de los conceptos descritos en [29] y [27]. El modelo basado en n-gramas se define formalmente a continuación:

Sea D un conjunto de documentos originales (el corpus de referencia). Cada documento $d \in D$ es codificado en forma de n-gramas de un orden conveniente para conformar conjuntos de n-gramas $N(d)$ (la manera de elegir un orden apropiado se muestra experimentalmente en

la siguiente sección). Sea s un documento sospechoso de plagio, s se divide en sentencias. El conjunto de n-gramas $N(s_i)$ para cada sentencia $s_i \in s$ es comparado con los conjuntos $N(d)$ con el objetivo de relacionar las sentencias plagiadas con aquel documento que potencialmente haya sido utilizado como su fuente.

Debido a la diferencia en las dimensiones de los conjuntos $N(s_i)$ con respecto a $N(d)$, la comparación es realizada con base en la medida de contención [30], ecuación 2.7 (sección 2.2.2). Si la máxima contención $C(s_i \mid d)$, luego de considerar todos los documentos de referencia $d \in D$, es mayor a un cierto umbral, s_i se considera un candidato a ser un plagio de d.

3.3.2. Evaluación del método basado en modelos de lenguaje

El objetivo de los experimentos realizados con este método ha sido determinar cuál es el mejor valor de n, es decir, el grado de los n-gramas a considerar para realizar la comparación de sentencias sospechosas con respecto a documentos de referencia. Los valores de n que hemos probado están dentro del intervalo $[1, \cdots, 5]$. Para realizar este experimento, hemos utilizado la parte del corpus METER descrita en la sección 3.1.3. La evaluación fue realizada en las medidas estándar de Precision, Recall y F-measure[8]. La figura 3.4 muestra los valores obtenidos con los distintos grados de n-gramas contemplados.

Los conjuntos de bolsas de palabras ($n = 1$) no consideran ninguna información acerca del contexto de las palabras ni sobre el estilo sintáctico de los textos. Estos factores provocan la obtención de un buen nivel de Recall en estos experimentos, que es prácticamente constante hasta que el umbral considerado es de 0.7 (recordemos que la medida de contención está dentro del umbral $[0, 1]$). Sin embargo, la probabilidad de que un documento $d \in D$ tenga el vocabulario completo de una sentencia s_i es demasiado alta. Por esta razón, la Precision obtenida en este caso es la más baja de todos los experimentos. En el otro extremo, considerar n-gramas de grado 4 (e incluso mayores) produce una estrategia de comparación demasiado rígida. En estos casos, cambios menores en s_i evitan que sea considerada como una sentencia plagiada, lo que resulta en los valores de Recall más bajos.

Los mejores resultados se obtienen considerando bigramas y trigramas (los mejores valores de F-measure obtenidos fueron 0.68 y 0.66, respectivamente). En ambos casos los n-gramas son suficientemente cortos para manejar modificaciones en las sentencias plagiadas y a la vez suficientemente largos para componer cadenas cuya probabilidad de aparecer en un documento, a excepción de la fuente que haya sido utilizada para realizar el plagio, sea muy baja. Una búsqueda basada en trigramas es más rígida, lo que permite una mejor Precision. Una búsqueda basada en bigramas resulta más flexible, generando mejores valores de Recall. La diferencia se ve reflejada en el umbral con el que se obtienen los mejores valores de F-measure: 0.34 para bigramas y 0.17 para trigramas.

[8] Si bien existen traducciones de los nombres de estas medidas, se ha optado por mantener los nombres en inglés para evitar posibles confusiones.

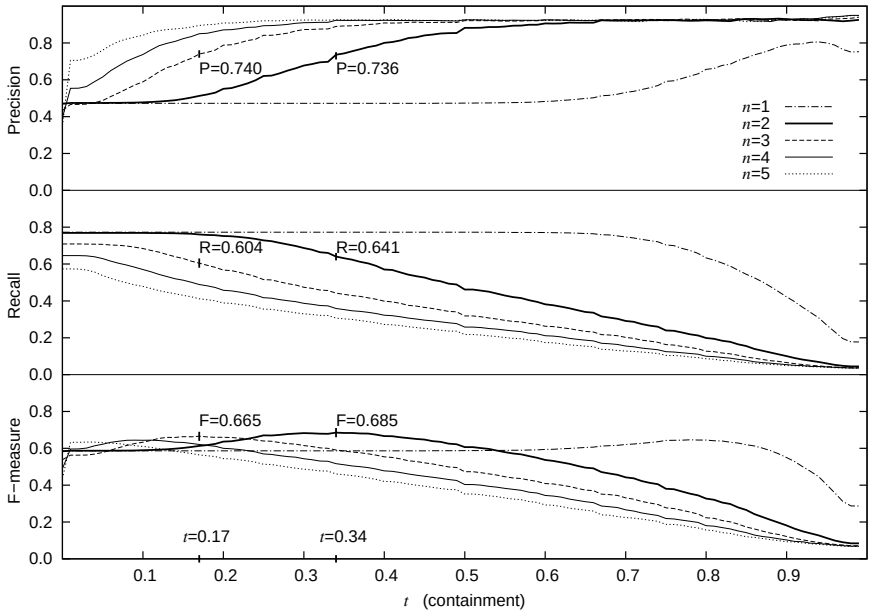


Figura 3.4: Evaluación del método de detección de plagio basado en n-gramas (n = grado del n-grama, , t =umbral

3.3.3. Discusión sobre el método basado en n-gramas

La estrategia de comparación basada en n-gramas ha mostrado ser suficiente flexible y certera. Los resultados sobre un corpus estándar en este tipo de tareas, el corpus METER, lo han demostrado. En cuanto al tipo de n-gramas que se deben considerar sólo se encuentran los de grado 2 y 3. Los bigramas favorecen la medida de Recall mientras que los trigramas favorecen la Precision. La razón por la que este modelo funciona se puede encontrar en la sección 2.2.2.

Aún queda pendiente realizar el proceso de extensión de vocabulario para poder detectar mejor aquellos casos de plagio generados por medio de copias reescritas. En este momento nos encontramos aún analizando si la mejor opción es utilizar los mismos recursos que en [27] o existen mejores posibilidades.

Cabe señalar que hemos escrito un artículo describiendo estos experimentos realizados sobre el corpus METER. Este artículo [2] ha sido enviado para su valoración a la edición 2009 de la Conferencia Europea sobre Recuperación de Información (*ECIR*)[9]. Desafortunadamente aún no se han anunciado los resultados de dicha valoración.

[9]http://ecir09.irit.fr/access.php

3.4. El problema del espacio de búsqueda

En las publicaciones sobre la tarea de detección de plagio a menudo se asume, como se puede ver en los métodos descritos en el capítulo 2, que el espacio de búsqueda (que consiste en el conjunto de documentos de referencia D) es suficientemente pequeño como para que cualquier estrategia de búsqueda genere resultados aceptables en un corto tiempo. Sin embargo, en general esto no es verdad. Los corpus de referencia suelen estar compuestos por grandes cantidades de documentos originales, lo que afecta directamente al tiempo de procesamiento necesario para analizar un documento sospechoso.

Por ello, consideramos que antes de realizar cualquier tipo de búsqueda (como la descrita en la sección anterior, por ejemplo), es necesario reducir tanto como sea posible el espacio de búsqueda representado por los documentos hallados en el corpus de referencia. El método esbozado a continuación, el cual hemos probado con el mismo extracto del corpus METER descrito en la sección 3.1.3, ha dado resultados prometedores.

3.4.1. Planteamiento del método de reducción de espacio de búsqueda

Dado el corpus de referencia D y un documento sospechoso s, nuestros esfuerzos están ahora orientados a localizar, de manera eficiente, un subconjunto D' de documentos de referencia tal que $|D'| \ll |D|$. El subconjunto D' debe contener aquellos documentos d con la más alta probabilidad de contener la fuente de los posibles fragmentos plagiados en s. Luego de esta reducción del corpus de referencia, puede realizarse un proceso exhaustivo de búsqueda de las sentencias de s sobre D'.

El método de reducción del espacio de búsqueda se basa en la distancia simétrica de Kullback-Leibler. Hemos determinado usar la distancia de Kullback-Leibler debido a que su aplicación en tareas relacionadas como la de agrupación de documentos (*clustering*) ha dado buenos resultados [7, 36].

La distancia de Kullback-Leibler

En 1951 Kullback y Leibler propusieron la que después sería conocida como la *divergencia de Kullback-Leibler* (KL_d) [28], también conocida como entropía cruzada. Dado un espacio de eventos, la KL_d se define por la ecuación 3.3. Su objetivo es medir qué tan diferentes son dos distribuciones de probabilidades P y Q sobre un vector de características $\mathcal{X}$.

$$KL_d(P \mid\mid Q) = \sum_{x \in \mathcal{X}} P(x) log \frac{P(x)}{Q(x)} \tag{3.3}$$

Sin embargo, KL_d no es simétrica, es decir, $KL_d(P \mid\mid Q) \neq KL_d(Q \mid\mid P)$. Por ello,

varios autores (incluyendo los mismos Kullback y Leibler) han propuesto diferentes versiones simétricas de KL_d, conocidas como distancia simétrica de Kullback-Leibler (KL_δ). Entre dichas versiones hemos optado por considerar la de Bigi [7] (ecuación 3.4). Hemos seleccionado esta versión porque, a diferencia de otras [28, 19, 6] que contemplan un doble cálculo de KL_d, ésta sólo implica una adaptación de la ecuación 3.3, agregando una resta.

$$KL_\delta(P \mid\mid Q) = \sum_{x \in \mathcal{X}} (P(x) - Q(x)) log \frac{P(x)}{Q(x)} \tag{3.4}$$

Dado $d \in D$ y s, calculamos la distancia KL_δ de la distribución de probabilidad P_d con respecto a Q_s. Estas distribuciones están compuestas por un conjunto de características de los documentos implicados con el objetivo de definir un conjunto reducido de documentos de referencia D'.

Selección de características

Para definir las distribuciones de probabilidad P_d hemos probado las siguientes tres técnicas no supervisadas de selección de términos:

1. *Frecuencia de términos* (*tf* por sus siglas en inglés). La relevancia del i-ésimo término t_i en el j-ésimo documento d_j es proporcional a la frecuencia de t_i en d_j. Se define como:

$$tf_{i,j} = \frac{f_{i,j}}{\sum_k f_{k,j}} \tag{3.5}$$

   donde $f_{i,j}$ es la frecuencia de t_i en d_j y se normaliza por la frecuencia total de los términos t_k en d_j.

2. *Frecuencia de términos - frecuencia invertida de documentos* (*tfidf* por sus siglas en inglés). El peso *tf* de un término t_i se normaliza por el número de documentos en el corpus en los que aparece. Se calcula como :

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i = tf_{i,j} \cdot log \frac{|D|}{|\{d_j | t_i \in d_j\}|} \tag{3.6}$$

   donde $|D|$ es la cantidad de documentos en D y $|\{d_j | t_i \in d_j\}|$ es el número de documentos en D que contienen t_i.

3. *Punto de transición* (*tp* por sus siglas en inglés). El punto de transición tp^* se obtiene por medio de la siguiente ecuación:

$$tp^* = \frac{\sqrt{8 \cdot I_1 + 1} - 1}{2} \tag{3.7}$$

donde I_1 es la cantidad de términos t_k con frecuencia de aparición 1 en d_j [38]. Con el objetivo de dar mayor relevancia a los términos cercanos al punto de transición, los pesos finales de los términos se calculan como:

$$tp_{i,j} = (\langle tp^* - f(t_i, d_j)\rangle + 1)^{-1} \tag{3.8}$$

Para garantizar valores positivos, $\langle\cdot\rangle$ es la función valor absoluto.

El objetivo del proceso de selección es crear una lista ordenada de términos. Cada distribución de probabilidad P_d se compone de los términos de la parte más alta de esta lista. Suponemos que dichos términos son los que mejor caracterizan a un documento d. Hemos experimentado caracterizando a los documentos con un porcentaje dentro del intervalo de [10,90] % de sus términos con los valores más altos de $\{tf, tfidf, tp\}$. Los resultados se verán más adelante.

Esta técnica sólo se utiliza para seleccionar los términos que caractericen a cada documento $d \in D$. La manera de calcular sus probabilidades asociadas se describe a continuación.

Cálculo de probabilidades para los vectores de características

La probabilidad (peso) de cada término incluido en P_d es simplemente calculada por la ecuación 3.5, es decir, $P(t_i, d) = tf_{i,d}$. Estas distribuciones de probabilidad son independientes de cualquier otro documento de referencia o sospechoso y sólo se calculan una vez.

Dado un documento sospechoso s, una distribución de probabilidad preliminar Q'_s se calcula de la misma manera, es decir, $Q'(t_i, s) = tf_{i,s}$. Sin embargo, con el afán de evitar valores infinitos al calcular la distancia entre las distribuciones de un documento s con respecto a d, lo cual se podría dar cuando exista un término t_i tal que $t_i \in d$ y $t_i \notin s$ (y viceversa), realizamos un proceso de suavizado para obtener Q_s, que es la distribución de probabilidad final que caracterizará al documento s, con respecto a cada distribución P_d. Q_s debe estar compuesta por los mismos términos que P_d. Por ello, si existe un término t_i tal que $t_i \in P_d \cap Q'_s$, la probabilidad $Q(t_i, s)$ es el resultado de un proceso de suavizado de $Q'(t_i, s)$, de lo contrario, $Q(t_i, s) = \epsilon$. Se trata de un simple proceso de suavizado de tipo *back-off*. Como en el caso de [7], la probabilidad $Q(t_i, s)$ se calculan de la siguiente manera:

$$Q(t_i, s) = \begin{cases} \gamma \cdot Q'(t_i \mid s) & \text{si } t_i \text{ ocurre en } d \text{ y } s \\ \epsilon & \text{si } t_i \text{ sólo ocurre en } d \end{cases} \tag{3.9}$$

Es importante notar que los términos que ocurran en s, pero no d, no son considerados relevantes. γ es un coeficiente de normalización estimado de la siguiente forma:

$$\gamma = 1 - \sum_{t_i \in d, t_i \notin s} \epsilon$$

respetando la condición:

$$\sum_{t_i \in s} \gamma \cdot Q'(t_i, s) + \sum_{t_i \in d, t_i \notin s} \epsilon = 1$$

El valor de ϵ es menor que la mínima probabilidad de un término del documento d.

Luego de calcular $KL_\delta(P_d \mid\mid Q_s)$ para todo $d \in D$, es posible definir un subconjunto de documentos de referencia D' que tengan una alta probabilidad de ser las fuentes de los posibles casos de plagio en s. El conjunto D' incluye los diez documentos de referencia con la menor KL_δ con respecto a s.

Un proceso de búsqueda exhaustiva como el descrito en la sección 3.3 puede entonces realizarse únicamente sobre el subconjunto de documentos de referencia D'. El proceso luego de haber obtenido las distribuciones P_d para todo $d \in D$ se resume en el algoritmo de la figura 3.5.

Algoritmo 1: Dado un corpus de referencia D y un documento sospechoso s:

// C
Calcular $Q'_s(t_k) = tf_{k,s}$ para todo $t_k \in s$
Para cada documento $d \in D$
    Definir la dist. de probabilidad Q_s dada P_d
    Calcular $KL_\delta(P_d \mid\mid Q_s)$
// Definiendo el subconjunto D' del corpus de referencia
$D' = \{d\}$ tal que $KL_\delta(P_d \mid\mid Q_s)$ es una de las 10 menores distancias obtenidas
$N_{s_i} = [n\text{-gramas en } s_i]$ para todo $s_i \in s$
// B
Para cada documento d en D'
    $N_d = [n\text{-gramas en } d]$
    Para cada sentencia s_i en s
        Calcular $C(N_{s_i} \mid N_d)$
    Si $\text{máx}_{d \in D'}(C(N_{s_i} \mid N_d)) \geq Threshold$
        s_i es un candidato de plagio proveniente de $\arg\text{máx}_{d \in D'}(C(N_{s_i} \mid N_d))$

Figura 3.5: Proceso de reducción del espacio de búsqueda.

3.4.2. Evaluación del método de reducción del espacio de búsqueda

Para evaluar éste método hemos utilizado el extracto del corpus METER descrito en la sección 3.1.3. Nuestros dos experimentos buscan comparar tanto la velocidad del proceso (en segundos) como la calidad de los resultados (de nuevo en términos de Precision, Recall y F-measure).

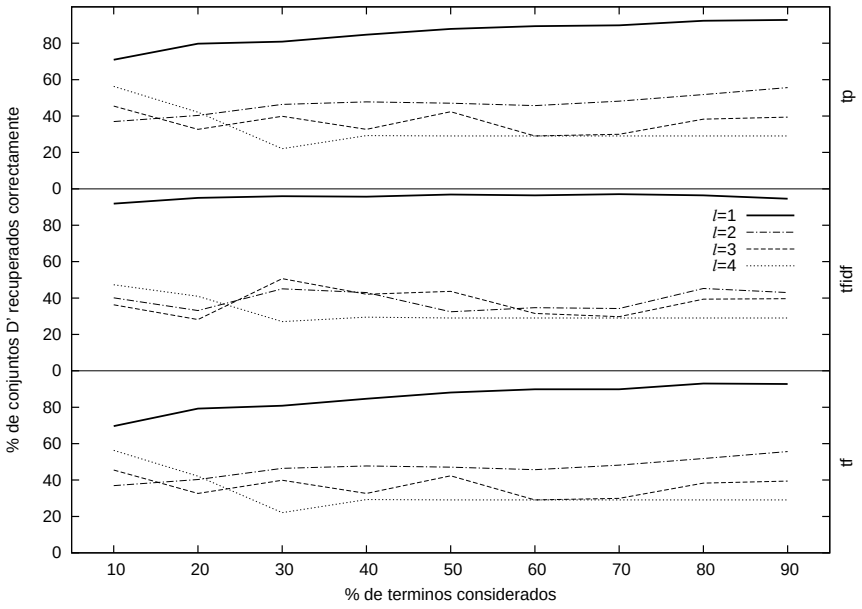


Figura 3.6: Evaluación del proceso de reducción del espacio de búsqueda. ($\{tf, tfidf, tp\}$ = técnicas de extracción de características, l = longitud de los términos)

Nuestro primer experimento busca analizar el impacto del proceso de reducción, comparamos el proceso de detección con y sin reducción. Los experimentos exploran los siguientes tres parámetros:

1. Longitud de los términos de las distribuciones de probabilidad: $l = \{1, 2, 3\}$
2. Técnica de selección de características: tf, $tfidf$ y tp.
3. Porcentaje de términos en d_j considerados para definir P_d: $[10, 90]\,\%$

Los resultados de la variación de estro parámetros se muestran gráficamente en la figura 3.6. Recordemos que los documentos sospechosos son notas publicadas en distintos periódicos y que el corpus de referencia está compuesto por documentos de la PA (sección 3.1.3). Dado un documento sospechoso s consideramos un acierto si la nota de la PA que fue considerada para generarlo está incluida en el conjunto reducido de documentos de referencia D'.

Para las tres técnicas de selección de características, los mejores resultados se obtienen al considerar n-gramas de grado 1. Mayores grados producen distribuciones de probabilidad demasiado uniformes y cercanas a 1. Estas distribuciones no permiten que KL_δ determine adecuadamente qué tan cerca se encuentra un documento de otro. Con respecto a la mejor técnica de selección de características, considerar simplemente tf no da buenos resultados. En este caso un alto número de palabras funcionales (preposiciones y artículos, por ejemplo), que no pueden caracterizar a un documento, son consideradas. Los resultados obtenidos con tp son parecidos y se deben a las mismas razones. Sin embargo, consideramos que estos podrían

Experimento	umbral	P	R	F	t
Con reducción	0.34	0.73	0.63	0.68	2.32
Sin reducción	**0.25**	**0.77**	**0.74**	**0.75**	**0.19**

Tabla 3.5: Comparación de resultados: búsqueda exhaustiva contra reducción de espacio + búsqueda exhaustiva. (P =Precision, R =Recall, F = F-me asure, t = tiempo promedio de procesamiento (seg.))

mejorar ante documentos de mayor longitud. Los mejores resultados se obtienen con $tfidf$. Las palabras funcionales (y otras) que no caracterizan al documento no son consideradas en las distribuciones de probabilidad y éstas caracterizan correctamente a los documentos de referencia y luego a los sospechosos. Con respecto a la longitud de las distribuciones de probabilidad, la calidad de la recuperación es prácticamente constante cuando se usa $tfidf$ sobre n-gramas de grado 1. La única mejora se observa al pasar de un 10 % a un 20 % del vocabulario en el documento (el porcentaje de documentos recuperados correctamente se incrementa de 91.89 % a 95.04 %). Por ello, consideramos que la mejor opción es tomar en cuenta el 20 % del vocabulario en d con el mayor $tfidf$ para componer P_d. De esta manera, se obtiene un buen porcentaje de documentos de referencia correctamente recuperados con una dimensión suficientemente baja en las distribuciones probabilidad.

El segundo experimento muestra la mejora obtenida al realizar una etapa de reducción de espacio antes de la de búsqueda exhaustiva de fragmentos plagiados (3.3). La tabla 3.5 muestra los resultados obtenidos cuando la búsqueda exhaustiva se realizó basada en bigramas sobre D y D' (los corpus de referencia original y reducido). Aunque la técnica de contención descrita en la sección 3.3 por sí misma da buenos resultados, el considerar demasiados documentos de referencia, los cuales en ocasiones no tienen ninguna relación con el sospechoso, genera mucho ruido al proceso de búsqueda, lo que afecta a los valores de Precision y Recall obtenidos. Una mejora importante se obtiene cuando $s_i \in s$ se busca solamente sobre D', luego del proceso de reducción del espacio de búsqueda.

Con respecto al tiempo de procesamiento, el tiempo promedio que se requiere para analizar un documento s sobre el corpus de referencia entero D es de 2.32 segundos en promedio, mientras que el proceso entero de reducción del espacio de búsqueda y el análisis de s con respecto al conjunto reducido D' necesita sólo 0.19 segundos[10]. Esta diferencia de tiempo se debe a tres factores: (1) las distribuciones P_d son precalculadas una sola vez para cada documento d, (2) La distribución $Q'(s)$ dado s sólo se calcula una vez y se va adaptando a cada distribución P_d y (3) en vez de buscar las sentencias $s_i \in s$ en D (compuesto por más de 700 documentos), sólo se buscan en D', que sólo contiene 10 documentos.

[10]El experimento se realizó con una implementación en Python sobre una PC Linux con 3.8GB de RAM y procesador de 1600 MHz.

3.4.3. Discusión sobre el método de reducción de espacio

La inclusión de una etapa de reducción del espacio de búsqueda en el proceso de detección de plagio ha mejorado los resultados obtenidos tanto en términos de calidad de la salida como de velocidad. Un buen corpus de referencia debe tener una cantidad importante de documentos, pero para hacer una búsqueda adecuada de las posibles fuentes de los casos plagiados en un documento sospechoso no es necesario realizar comparaciones exhaustivas sobre el corpus de referencia entero.

Un proceso de preselección de documentos de referencia es lo más lógico en esta tarea y los resultados obtenidos al basar dicho proceso en la distancia de Kullback-Leibler ha dado resultados prometedores.

Hemos escrito el par de artículos [3] y [4] sobre lo descrito en esta sección . Dichos artículos han sido enviados para su evaluación a la edición 2009 de la Conferencia del Capítulo Europeo de la Asociación para la Lingüística computacional que se celebrará en Grecia y la Conferencia sobre Procesamiento de texto Inteligente y Lingüística computacional, que se llevará a cabo en México[11]. Desafortunadamente aún no se han anunciado los resultados de la correspondientes valoraciones.

[11] http://www.eacl2009.gr/conference/ y http://www.cicling.org/2009/ respectivamente

Capítulo 4

Líneas de investigación abiertas

Si bien se han desarrollado técnicas que han mostrado buenos resultados en la tarea de la detección de plagio, existe aún una brecha muy grande como para poder considerar que se trata de un problema resuelto. Entre las principales necesidades que hemos detectado, las cuales bien vale la pena abordar en futuras investigaciones, se encuentran las siguientes:

1. El diseño de métodos eficientes para la detección de plagio que sean capaces de dar buenos resultados en tiempos prácticamente adecuados.
2. La generación e implementación de metodologías para la detección de plagio dentro de un contexto translingüe, es decir, en los que el corpus de referencia y los documentos sospechosos estén escritos en distintos idiomas.
3. La creación de corpus de plagios estándares cuyo objetivo sea específicamente el diseño y evaluación de métodos para la detección automática de plagio.

Actualmente hemos comenzado ya a trabajar sobre el primer punto y, parcialmennte, sobre el segundo y tercero. Consideramos que aún hay mucho trabajo pendiente al respecto. En cuanto al tercer punto, en este momento estamos comenzando a trabajar en la generación de corpus con las características necesarias para el desarrollo de los métodos de detección de plagio. Nuestros primeros esfuerzos se describen a través de las siguientes secciones.

La complejidad de estas tres tareas, incluso de manera individual, nos ha llevado a programar su investigación dentro de la continuación de esta investigación: en la etapa de doctorado.

4.1. Diseño de métodos eficientes

Como lo señalamos en la sección 3.4, en las publicaciones sobre la tarea de detección de plagio a menudo se asume que el espacio de búsqueda, que consiste en el conjunto de documentos de referencia D, es suficientemente pequeño como para que cualquier estrategia de búsqueda genere resultados aceptables en un corto tiempo. Esto no suele ser verdad.

Hasta ahora, hemos probado atacar este problema por medio de una reducción del espacio de búsqueda basada en la distancia de Kullback-Leibler [3, 4]. Aunque dicha investigación ha dado muy buenos resultados, consideramos que es necesario considerar otros métodos para estos mejoren más.

Incluso, nos planteamos la aplicación de la distancia de Kullback-Leibler para la tarea de análisis semántico explícito (*ESA* por sus siglas en inglés) [20]. Creemos que este tipo de análisis, el cual está basado en la comparación distribuciones de probabilidad de documentos prototipo, puede ser útil tanto para la reducción del espacio de búsqueda como para el mismo análisis de plagio. Esta técnica ha sido utilizada también en problemas translingües y en ella se basa uno de los escasos trabajos sobre la detección de plagio translingüe [41].

4.2. Diseño de métodos translingües

Hay un tipo de plagio muy importante cuya detección automática no ha sido poco tratado: el plagio translingüe. Previamente se ha definido que "un texto es considerado un plagio de otro si sus contenidos son considerados semánticamente similares, sin importar que estén escritos en idiomas diferentes, y la correspodiente cita no está incluida" [5].

Es una realidad que muchos casos de plagio no están escritos en el mismo idioma que los documentos que fueron utilizados como fuente. Si consideramos el mismo ejemplo de Wikipedia planteado en la sección 3.3, la frase "*En 1950 Cousteau arrendó un barco llamado Calypso, el cual fue hundido en Singapur en 1996, por un franco al año*" es definitivamente un caso de plagio. En particular, plagio translingüe.

Para resolver este problema, no es suficiente con aplicar alguna de las técnicas existentes para detectar plagios monolingües. Es necesario diseñar nuevos métodos que impliquen etapas de recuperación de información multilingüe [36] e incluso métodos de traducción automática [10].

Entre los escasos trabajos al respecto se encuentra uno basado en el cálculo de similitud entre documentos por medio de artículos de Wikipedia [41]. El otro es el que conforma nuestro primer intento en esta tarea en particular [5], el cual está basado en el modelo de alineación IBM-1 [9], comúnmente usado en tareas de traducción estdística.

Desafortunadamente, nuestro modelo es dependiente de un conjunto de textos originales y plagiados alineados para realizar el entrenamiento estadístico necesario. Hemos realizado pruebas con dos pequeños corpus generados por nosotros mismos (uno inglés-español y otro inglés-italiano) y los resultados no son del todo malos. Sin embargo, para validar este método es necesario contar con un corpus más significativo. La conformación de un corpus con las características adecuadas es por sí mismo un reto complicado, tal como se verá en la sección 4.3.

Cabe señalar que con esta investigación ha sido posible publicar 2 artículos: [5] y [37]. Dichos artículos se presentaron en el taller *PAN: Uncovering Plagiarism, Authorship and*

Social Software Misuse, llevado a cabo en Grecia y en el Fourth Latin American Workshop on Non-Monotonic Reasoning 2008.

4.3. Creación de corpus adecuados para las investigaciones en detección de plagio

La necesidad de corpus útiles para el diseño y puesta a punto de los métodos de detección de plagio es evidente. Si bien existen corpus que pueden ser útiles, como el corpus METER [12] (véase la sección 2.3.1), en realidad este corpus no está compuesto por plagios y, dada su naturaleza periodística, sus documentos son cortos. Dicha longitud no es muy realista cuando se busca atacar verdaderos problemas de plagio.

Por ello, en conjunto con el Grupo de Sistemas de Información y tecnología Web de la Universidad de Weimar, hemos comenzado el planteamiento de corpus que cumplan con las características necesarias[1].

Un corpus adecuado para el desarrollo de tecnología en esta área debe tener ciertas características mínimas. Una descripción de dichas características, así como las opciones que hasta ahora hemos planteado para cubrirlas se encuentran en la tabla 4.1.

Un aspecto interesante a considerar es si el corpus debe estar conformado por documentos reales o sintéticos. Lo ideal sería que se tratara del primer caso. Sin embargo el conformar el corpus con casos de plagio reales implica una buena cantidad de dificultades. Entre ellas, podemos destacar el hecho de que habría que aplicar técnicas de detección de plagio para hallar los fragmentos plagiados. Luego de esto, sería necesario realizar una revisión manual de los fragmentos identificados como plagiados (al igual que para todos los demás) para asegurarse de que no hay errores. Dadas las dimensiones del corpus planteado, esto resulta imposible. Una opción sería utilizar otros corpus, como el corpus METER [12], sin embargo, en este caso existen problemas de licencia y derechos de autor que convierten esta tarea en algo impráctico. Por ello, hemos decidido que el corpus sea sintético. De esta manera, existe pleno control de las características que se desea garantizar.

Para conformar un corpus de plagios sintéticos, será necesario diseñar métodos para la generación de plagios de manera automática. Si se buscara incluir solamente copias exactas, no habría mucha dificultad al respecto. Se requieren distintos niveles de reescritura de los plagios, por lo que es necesario diseñar los métodos que los generen, lo cual por sí mismo es una tarea complicada.

[1]Cabe señalar que en conjunto con este grupo alemán, nos proponemos organizar la edición 2009 (tercera) del taller internacional *PAN: Uncovering Plagiarism, Authorship and Social Software Misuse* (http://www.aisearch.de/pan-08/). Hasta el momento se plantea su celebración en torno al XXV congreso de la Sociedad Española para el Procesamiento de Lenguaje Natural, que se llevará a cabo en el mes de septiembre de 2009 en San Sebastián. Dicho taller deberá incluir una competencia de detección de plagio tanto monolingüe como translingüe, por lo que la necesidad de la creación de un corpus adecuado para realizarla tiene aún mayor interés

Tabla 4.1: Características deseables en un corpus para la detección de plagio

Característica	Planteamiento inicial
Suficientemente extenso para implicar un problema tanto de búsqueda como de rendimiento	Consideramos que un volumen adecuado sería contar con alrededor de 10,000 documentos (entre sospechosos y de referencia).
Con distintos tipos de plagio cuya dificultad de localización sea distinta	Planteamos la existencia de copias exactas y copias con distintos niveles de reescritura. Para el caso del corpus multilingüe, dichos niveles pueden ser traducciones profesionales (realizadas por un humano) y traducciones automáticas (realizadas por medio de algún recurso electrónico)
Que todos los casos de plagio estén perfectamente delimitados	Si un documento sospechoso de plagio dentro del corpus tuviera más fragmentos plagiados que los identificados, la evaluación de los diversos métodos sobre él sería erronea.
Un corpus que permita identificar de manera eficiente cada uno de los fragmentos plagiados	Para ello, hemos determinado que los corpus estén etiquetados con el formato XML. De esta manera no sólo será posible identificar los fragmentos plagiados, sino que que se podrá agregar mayor información, como el tipo de plagio del que se trate o su documento de origen.

Para el caso de los corpus monolingües, se plantea crear uno de documentos escritos en inglés y otro en español (de momento no se descarta crear también uno de documentos en alemán). En cuanto al corpus multilingüe, el idioma principal, es decir, el del corpus de referencia, será el inglés. Los idiomas de los documentos sospechosos serán el español y el alemán.

Como se puede inferir, para la generación de estos corpus será necesario explotar recursos propios de la recuperación de información, así como el diseño de nuevas técnicas que permitan la generación automática de casos de plagio. Además, para diversas tareas tales como la medición del nivel de reescritura de un fragmento plagiado, será necesario incluso aplicar las mismas técnicas diseñadas para la detección de plagio.

Bibliografía

[1] Alberto Barrón-Cedeño and Paolo Rosso. Towards the exploitation of statistical language models for plagiarism detection with reference. In *Proceedings of the ECAI'08 PAN Workshop Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 15–19, Patras, Greece, 2008.

[2] Alberto Barrón-Cedeño and Paolo Rosso. On automatic plagiarism detection based on n-grams comparison. In *European Conference on Information Retrieval 2009*, 2009, enviado para su evaluación.

[3] Alberto Barrón-Cedeño and Paolo Rosso. On the importance of search space reduction in automatic plagiarism detection. In *Conference of the European Chapter of the Association for Computational Linguistics 2009*, 2009, enviado para su evaluación.

[4] Alberto Barrón-Cedeño and Paolo Rosso. Reducing the plagiarism detection search space on the basis of the Kullback-Leibler distance. In *Conference on Intelligent Text Processing and Computational Linguistics*, 2009, enviado para su evaluación.

[5] Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. On cross-lingual plagiarism analysis using a statistical model. In *Proceedings of the ECAI'08 PAN Workshop Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 9–13, Patras, Greece, 2008.

[6] Charles H. Bennet, Péter Gács, Ming Li, Paul M. B. Vitányi, and Wojciech H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[7] Brigitte Bigi. Using Kullback-Leibler distance for text categorization. In *Proceedings of the 25th ECIR'03*, volume LNCS (2633) Advances in Information Retrieval, pages 305–319, Pisa, Italy, 2003.

[8] Thomas M. Breuel. The ocropus open source ocr system. *Proceedings of the IS&T/SPIE 20th Annual Symposium 2008*, 2008.

[9] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vicent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[10] Jorge Civera and Alfons Juan. Mixtures of ibm model 2. *Proceedings of the EAMT Conference*, pages 159–167, 2006.

[11] Paul Clough. Plagiarism in natural and programming languages: an overview of current tools and technologies. *Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK*, 2000.

[12] Paul Clough, Robert Gaizauskas, and Scott Piao. Building and annotating a corpus for the study of journalistic text reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-02)*, volume V, pages 1678–1691, Las Palmas de Gran Canaria, Spain, 2002.

[13] Rosa Maria Coyotl-Morales, Luis Villaseñor Pineda, Manuel Montes-y Gómez, and Paolo Rosso. Authorship attribution using word sequences. *Proc. of the 11th Iberoamerican Congress on Pattern Recognition, (CIARP 2006)*, LNCS (4225):844–853, 2006.

[14] Josep M. Crego, José B. Mariño, and Adriá de Gispert. An n-gram-based statistical machine translation decoder. In *Interspeech'2005 - Eurospeech*, pages 2534–2544, 2005.

[15] E. Dale and J. S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27:37–53, 1948.

[16] William H. DuBay. The principles of readability, 2004. Impact Information, "www.impact-information.com/impactinfo/readability02.pdf", Última consulta: Noviembre de 2008.

[17] Real Academia Española. Diccionario de la lengua española. vigésima segunda edición.

[18] Georgia Frantzeskou, Efstathios Stamatatos, and Stefanos Gritzalis. Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *International Journal of Digital Evidence*, 6(1), 2007.

[19] Bent Fuglede and Flemming Topse. Jensen-Shannon divergence and Hilbert space embedding. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT'04)*, page 31, Chicago, IL, 2004.

[20] E. Gabrilovich. Feature generation for textual information retrieval using world knowledge. *Phd thesis, Israel Institute of Technology*, 2006.

[21] Valerie J. Haines, George M. Diekhoff, Emily E. LaBeff, and Robert E. Clark. College cheating: Immaturity, lack of commitment, and the neutralizing attitude. *Research in Higher Education*, 25(4):342–354, 1986.

[22] Djoerd Hiemstra. A linguistically motivated probabilistic model of information retrieval. *Proc. of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, (ECDL 1998)*, LNCS (1513):569–584, 1998.

[23] A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing bulletin*, 7(2):172–179, 1979.

[24] Parvati Iyer and Abhipsita Singh. Document similarity analysis for a plagiarism detection system. In *Proceedings of the 2nd Indian Int. Conf. on Artificial Intelligence (IICAI-2005)*, pages 2534–2544, 2005.

[25] P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[26] Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, 1997.

[27] NamOh Kang, Alexander Gelbukh, and SangYong Han. PPChecker: Plagiarism pattern checker in document copy detection. In *Proceedings of the TSD-2006: Text, Speech and Dialogue*, volume LNAI (4188), pages 661–667, Brno, Czech Republic, 2006.

[28] Solomon Kullback and Richard Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[29] Caroline Lyon, Ruth Barrett, and James Malcolm. A theoretical basis to the automated detection of copying between texts, and its practical implementation in the ferret plagiarism and collusion detector. In *Proceedings of Plagiarism: Prevention, Practice and Policies Conference*, Newcastle, UK, 2004.

[30] Caroline Lyon, James Malcolm, and Bob Dickerson. Detecting short passages of similar text in large document collections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 118–125, Pennsylvania, 2001.

[31] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press Publisher, Cambridge Massachusetts and London, England, 2000.

[32] Hermann Maurer, Frank Kappe, and Bilal Zaka. Plagiarism - a survey. *Journal of Universal Computer Science*, 12(8):1050–1084, 2006.

[33] Sven Meyer zu Eissen, Benno Stein, M. Kulig, and (editors). Plagiarism corpus webis-pc-08, 2008. Web Technology & and Information Systems Group, Bauhaus University Weimar, "http://www.uni-weimar.de/medien/webis/research/corpora".

[34] Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. Plagiarism detection without reference collections. *Reinhold Decker and Hans J. Lenz, editors, Advances in Data Analysis*, pages 359–366, 2007.

[35] Fuchim Peng, Dale Schuurmans, Vlado Keselj, and Shaojun Wang. Automated authorship attribution with character level language models. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary*, 2003.

[36] David Pinto, José-Miguel Benedí, and Paolo Rosso. Clustering narrow-domain short texts by using the Kullback-Leibler distance. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, CICLING 2007*, volume LNCS (4394), pages 611–622, Mexico City, Mexico, 2007.

[37] David Pinto, Jorge Civera, Alfons Juan, Paolo Rosso, and Alberto Barrón-Cedeño. A statistical approach to crosslingual natural language tasks. In *Fourth Latin American Workshop on Non-Monotonic Reasoning*, Puebla, México, 2008.

[38] David Pinto, Héctor Jiménez-Salazar, and Paolo Rosso. Clustering abstracts of scientific texts using the transition point technique. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, CICLING 2006*, volume LNCS (3878), pages 536–546, Mexico City, Mexico, 2006.

[39] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. *Croft, Moffat, van Rijsbergen, Wilkinson, and Zobel, Eds., 21st Annual International ACM SIGIR Conference*, pages 275–281, 1998.

[40] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[41] Martin Potthast, Benno Stein, and Maik Anderka. A wikipedia-based multilingual retrieval model. In *30th European Conference on IR Research, ECIR 2008*, volume 4956 LNCS, Glasgow, UK.

[42] Verónica Romero, Vicente Alabau, and José-Miguel Benedí. Combination of n-grams and stochastic context-free grammars in an offline handwritten recognition system. In *3rd Iberian Conference on Pattern Recognition and Image Analysis, Springer-Verlag*, volume LNCS (4477), pages 467–474, Girona, Spain, 2007.

[43] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[44] Narayanan Shivakumar and Héctor García-Molina. SCAM: A copy detection mechanism for digital documents. In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*, 1995.

[45] Antonio Si, Hong Va Leong, and Rynson W. H. Lau. Check: a document plagiarism detection system. In *Proceedings of the 1997 ACM Symposium on Applied Computing*, pages 70–77, San Jose, CA, 1997.

[46] Efstathios Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35:193–214, 2001.

[47] Benno Stein. Principles of hash-based text retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference*, pages 527–534, Amsterdam, Netherlands, 2007.

[48] Benno Stein, Moshe Koppel, and Efstathios Stamatatos. Plagiarism analysis, authorship identification, and near-duplicate detection (pan' 07). *SIGIR Forum*, 41(2):68–71, 2007.

[49] Benno Stein and Sven Meyer zu Eissen. Intrinsic plagiarism analysis with meta learning. In *Proceedings of the SIGIR'07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*, pages 45–50, Amsterdam, Netherlands, 2007.

[50] Andreas Stolcke. Srilm - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 527–534, Denver, Co., 2002.

[51] Wikipedia The free encyclopedia. Gunning fog index.

[52] Wikipedia The free encyclopedia. Jacques-Yves Cousteau.

[53] Wikipedia The free encyclopedia. Smog.

[54] Daniel R. White and Mike S. Joy. Sentence-based natural language plagiarism detection. *Journal on Educational Resources in Computing (JERIC)*, 4(4), 2004.

[55] G. Yule. The statistical study of literary vocabulary. 1944.

[56] Richard Zens and Hermann Ney. N-gram posterior probabilities for statistical machine translation. 2006.

Apéndice A

Descripción de símbolos

Símbolo	Descripción
w	Una palabra gráfica. Aunque extrictamente se trata de un conjunto de caracteres hallados entre espacios, consideramos que los signos de puntuación son independientes de las palabras y que de hecho, para fines prácticos, funcionan como una palabra más
s	Documento sospechoso. Es un documento que es analizado con el objeto de determinar si tiene casos de plagio, es decir, fragmentos de texto que han sido copiados (aún adaptándolos) de otros textos sin incluir la cita adecuada.
D	Corpus de referencia. Compuesto por un conjunto de documentos presumiblemente originales que son fuente potencial de los casos de plagio.
d	Documento de referencia. Uno de los documentos que componen el corpus de referencia
$N(\cdot)$	Conjunto de n-gramas en $\cdot$
$\mathcal{A}$	El autor de algún texto

Apéndice B

Publicaciones en el marco de la investigación

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

1. Alberto Barrón-Cedeño and Paolo Rosso. Towards the exploitation of statistical language models for plagiarism detection with reference. In Proceedings of the ECAI'08 PAN Workshop Uncovering Plagiarism, Authorship and Social Software Misuse, pages 15-19, Patras, Greece, 2008.

2. David Pinto, Jorge Civera, Alfons Juan, Paolo Rosso, and Alberto Barrón-Cedeño. A statistical approach to crosslingual natural language tasks. In Fourth Latin American Workshop on Non-Monotonic Reasoning, Puebla, México, 2008

3. Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. On cross-lingual plagiarism analysis using a statistical model. In Proceedings of the ECAI'08 PAN Works- hop Uncovering Plagiarism, Authorship and Social Software Misuse, pages 9-13, Patras, Greece, 2008.

Igualmente, los siguientes trabajos han sido enviados para su valoración:

1. Alberto Barrón-Cedeño and Paolo Rosso. On the importance of search space reduc- tion in automatic plagiarism detection. Enviado a: Conference of the European Chapter of the Association for Computational Linguistics 2009.

2. Alberto Barrón-Cedeño and Paolo Rosso. Reducing the plagiarism detection search space on the basis of the Kullback-Leibler distance. Enviado a: Conference on Intelligent Text Processing and Computational Linguistics, 2009.

3. Alberto Barrón-Cedeño and Paolo Rosso. On automatic plagiarism detection based on n-grams comparison. Enviado a: European Conference on Information Retrieval 2009.