Detección automática de plagio: de la copia exacta a la paráfrasis *

Alberto Barrón-Cedeño[1], Marta Vila[2] y Paolo Rosso[1]

[1]Natural Language Engineering Lab. - ELiRF
Universidad Politécnica de Valencia
{lbarron, prosso}@dsic.upv.es

[2]CLiC, Departament de Lingüística
Universitat de Barcelona
marta.vila@ub.edu

Resumen

El plagio, el reuso no autorizado y sin referencia de texto, es un fenómeno que ha cobrado gran interés debido a la enorme cantidad de recursos bibliográficos e información al alcance de la mano en Internet. Debido a la magnitud del problema, la revisión manual de los textos en busca de plagio es prácticamente imposible. Los conocidos como detectores automáticos de plagio surgen como una medida precautoria y correctiva para asistir al humano en la detección de plagio en textos, una tarea de la lingüística forense.

Debe observarse que las herramientas de detección automática de plagio buscan solamente asistir al humano en la detección, proveyéndole de las mayores pruebas posibles de un potencial caso de plagio. La decisión final, así como las acciones pertinentes, debe ser tomada por el experto.

En este capítulo se introduce brevemente el plagio y se presenta su relación con la paráfrasis. Este fenómeno lingüístico, si bien se encuentra en la base del acto de plagiar, no ha recibido atención suficiente por parte de los expertos. En este sentido, consideramos que los trabajos existentes sobre paráfrasis en el ámbito de la lingüística y el procesamiento del lenguaje natural son valiosas para la detección automática de plagio.

Palabras clave: detección de plagio, detección de paráfrasis, lingüística forense

*Esta contribución está orientada a la descripción de los conceptos y métodos subyacentes a la detección automática de plagio y no al análisis de las herramientas comerciales disponibles. Si el lector está interesado en las herramientas, puede considerar los servicios otorgados por compañías como Turnitin (iParadigms, 2010) o DOC Cop (McCrohon, 2010). Adicionalmente, sugerimos consultar (Maurer et˜al., 2006); particularmente las secciones 4 y 5. Por otro lado, este análisis está enfocado al plagio de texto. El lector interesado en el plagio de otro tipo de recursos, como por ejemplo música, puede consultar los trabajos de Robine et˜al. (2007) y Müllensiefen and Pendzich (2009).

1. Introducción

La Real Academia Española (2008) define el plagio como el acto de "*copiar en lo sustancial obras ajenas, dándolas como propias*". Si bien dicha definición es tajante, el concepto de obra resulta un tanto ambiguo. Distintos objetos pueden ser plagiados; desde un fragmento de texto o un programa informático hasta una fotografía, pintura o una pieza musical. Esta contribución se centra en el plagio de texto. De hecho, es prudente considerar una definición más completa, tal como la acuñada por la IEEE (2008):

> *plagiar es reusar las ideas, procesos, resultados o palabras de alguien más sin mencionar explícitamente a la fuente y su autor.*

Comencemos por observar algunos de los casos de plagio que más han llamado la atención: los relacionados con la literatura y la música.

Olsson (2008), en su recientemente publicado libro sobre lingüística forense, menciona el caso de Margaret Canby y Hellen Keller. Los análisis dejan ver que *Frost King*, escrito por Keller, contiene una cantidad inesperada de fragmentos similares con respecto al *Frost Fairies* de Canby; se trataba de un caso de plagio. Dentro del ámbito de la lengua española uno de los casos más polémicos es el de Camilo José Cela y su obra *La cruz de San Andrés*, la cual, presumiblemente, está basada en el libro *Carmen, Carmela, Carmiña (Fluorescencia)*, de la escritora Carmen Formoso. Luis Izquierdo, catedrático de Literatura Española de la Universidad de Barcelona ha señalado indicios de ello (Ríos, 2009). En el caso de Keller es posible observar fragmentos de texto prácticamente copiados de Canby (Olsson, 2008, pp. 101-103). Sin embargo, en el caso de Cela las similitudes son tan sutiles que parece difícil que un procesamiento automático (al menos uno de uso general) pueda dar con los fragmentos reusados.

Otros que han llamado la atención son los posibles casos de plagio de letras musicales, siendo uno de los más recientes en España el del cantautor Enrique Bunbury. Diversas estrofas de su canción "El hombre delgado que no flaqueará jamás", sencillo del disco Hellville Deluxe (Bunbury, 2008), incluyen frases de poemas del escritor Pedro Casariego. Bunbury señala que, como hace la gran mayoría de autores, se ha inspirado en Casariego tanto como en otros escritores en la creación de sus letras.

Algunos autores señalan que el reuso de texto (y por supuesto, lengua hablada) no es algo nuevo. Autores clásicos como W. Shackespeare reutilizaban las obras de otros sin darles el crédito adecuado (Coulthard and Alison, 2007), aunque en la época el concepto de plagio no había sido acuñado aún). Volviendo a nuestros días, en las últimas dos décadas se ha observado un crecimiento importante en los casos de plagio, sobre todo el de tipo académico. La razón es muy sencilla: desde el surgimiento de los ordenadores personales e Internet (con la enorme cantidad de documentos que pone al alcance de la mano), el acto de reusar texto resulta de lo más sencillo, por lo que se ha acuñado el término de *ciberplagio* (Anderson, 1999). De hecho, tanto Kulathuramaiyer and Maurer (2007) como Weber (2007) señalan la existencia del síndrome de "copia y pega".

Así, mientras algunas encuestas de la década de 1980 dejaban ver que aproximadamente el 30 % de los estudiantes admitía haber cometido plagio alguna vez en sus trabajos durante todo el periodo escolar (Haines et˜al., 1986), el estudio realizado por la Association of Teachers and Lecturers (2008) señala que los profesores estiman que el 28 % de todos los trabajos de los estudiantes

Tabla 1: Descripción de la notación usada.

Elementos	
d	Documento fuente. Dicho documento contiene, presumiblemente, material original y puede ser utilizado como la fuente para otro.
d_q	Documento sospechoso. Documento analizado que puede, o no, contener fragmentos plagiados
t	Denominamos con t una palabra: la cadena de caracteres limitada por espacios y signos de puntuación
D / D_q	Colección de documentos d / d_q
s / s_q	Fragmento de texto proveniente de d / d_q
$\mathcal{A}$	El autor de un documento (sea original o plagiado)

contiene plagio. Una de las encuestas más recientes al respecto señala que, de una muestra de 900 personas, el 16 % aceptó haber plagiado alguna vez en su vida, mientras que el 25 % prefirió no responder, lo que puede implicar un alza en el porcentaje real (Potthast et˜al., 2010c). Incluso hay quien identifica a la Wikipedia[1], la enciclopedia más grande disponible en la actualidad, como una fuente recurrente en los casos de plagio (Martínez, 2009).

Dada la enorme cantidad de documentos existentes, y disponibles, la detección manual de plagio resulta imposible. Por ello, es necesario el desarrollo de herramientas computacionales que asistan al ser humano en esta tarea: los llamados detectores automáticos de plagio. Desde un punto de vista computacional, la detección de plagio es una tarea tanto del **procesamiento de lenguaje natural** (PLN) como de la **recuperación de información** (RI).

Ahora bien, el plagio no es siempre el resultado de un proceso consciente y cuyo afán es el engaño. En ocasiones es el resultado de la ignorancia (debido a la falta de instrucción sobre la adecuada referencia de fuentes) o de fenómenos tales como la criptomnesia, es decir, cuando una persona asume una idea como propia porque inconscientemente ha olvidado que tiempo atrás la ha adquirido de alguien más (Taylor, 1965). Por ello, alegamos que un sistema informático no es suficiente para aplicar una medida disciplinaria o castigo. Por el contrario, debe ser utilizado como apoyo por una persona quien, al considerar otros factores además de la evidencia proporcionada por el software, debe tomar la decisión final.

El resto de la contribución se distribuye de la siguiente manera. La sección 2 incluye los distintos tipos de plagio así como los modelos de detección automática capaces de abordarlos. La sección 3 trata la paráfrasis: tipología y modelos para su detección. Finalmente, en la sección 4 se presentan las conclusiones. Asimismo, con el afán de facilitar la comprensión del contenido, algunos conceptos se incluyen en el apéndice A (dichos conceptos se destacan en el texto en negrita). Un resumen de la notación empleada en el resto del documento se encuentra en la tabla 1.

2. Detección de plagio

Existen distintas clasificaciones en cuanto a qué partes de un texto (o lengua hablada) se pueden plagiar. Martin (2004) considera lo siguiente:

[1] `http://www.wikipedia.org`

de ideas Este tipo de plagio es independiente de las palabras. $\mathcal{A}$ adopta las ideas, pensamientos o teorías de otra persona sin darles el crédito adecuado. Maurer et˜al. (2006) acotan que este tipo de plagio ocurre cuando la idea fuente no forma parte del conocimiento común.

palabra por palabra Se trata de la copia de una (parte importante de una) frase. $\mathcal{A}$ puede realizar una copia exacta e incluso efectuar algunas modificaciones. Si no hay referencia a la fuente, se estará cometiendo plagio. Clough (2003) considera, además del plagio palabra por palabra, el plagio por paráfrasis, en el que tanto las palabras como la sintaxis son modificadas. Como veremos (*cf.* sección 3), este platea miento de la paráfrasis como un tipo de plagio no es del todo riguroso.

de fuentes $\mathcal{A}$ incluye las referencias bibliográficas que otro autor ha incluido en su propio documento d. Sin embargo, $\mathcal{A}$ no señala que dichas referencias han sido extraídas de d. En ocasiones, $\mathcal{A}$ incluye las referencias sin siquiera haberlas consultado.

de autoría $\mathcal{A}$ presume ser el autor de un documento entero que en realidad ha sido escrito por otra persona. Esto ocurre a menudo con estudiantes que entregan reportes de otras personas como suyos.

Son el plagio de ideas y el de fuentes los que resultan más complicados de descubrir. En el primer caso, la poca correlación entre las ideas y las palabras con las que se pueden expresar hacen que, a menos que se tenga un dominio suficiente del tema tratado y sus antecedentes, no se pueda descubrir la falta. En el segundo caso, a menos que se haga un análisis profundo, que quizás incluya aplicar un cuestionario a $\mathcal{A}$, no hay manera de descifrar que una persona haya leído realmente una fuente o que simplemente haya copiado las referencias de otro documento (aunque por supuesto, si las referencias en dos documentos d y d_q son prácticamente las mismas y tienen el mismo orden, pueden ser consideradas como un factor que refleje un caso de plagio (HaCohen-Kerner et˜al., 2010)).

Al observar los casos de plagio que resultan más abordables, tanto de manera manual como automática, conviene girar la vista hacia una clasificación que tome como base el tipo de operaciones realizadas al texto reusado.

Si se desea analizar no sólo los tipos de plagio, sino también los modelos existentes para su detección automática, debemos mirar a la tipología propuesta por Maurer et˜al. (2006), que está un poco más orientada a las operaciones realizadas al texto durante el proceso de plagio. A continuación se muestra dicha tipología, incluyendo ejemplos ilustrativos así como una descripción de los modelos computacionales existentes para su detección automática. En todos los casos el texto considerado como fuente es el siguiente:

$s =$ `El curioso incidente del perro a medianoche`

2.1. Copia exacta

En este caso $\mathcal{A}$ copia un fragmento de texto sin hacer una sola modificación. De esta forma, el plagio de s es simplemente:

$s' =$ `El curioso incidente del perro a medianoche`

En los casos en los que el texto reusado no sufre modificación alguna, los mejores métodos son los basados en el modelo de "huella digital" (del inglés fingerprinting). Se trata ésta de toda una familia de modelos que ha sido diseñada para realizar una comparación eficiente entre documentos (lo que es importante si consideramos la enorme cantidad de comparaciones que es necesario realizar en busca de la fuente de un posible caso de plagio).

En este caso se asume que se cuenta con un repositorio significativo de potenciales documentos fuente y, por ende, se puede contar con una base de datos de representaciones de documentos. Cada documento de la colección de referencia se divide en fragmentos s. Brin et˜al. (1995) propone que dichos fragmentos sean oraciones, mientras que otros autores sugieren considerar n-**gramas**; ya sea de palabras (Bernstein and Zobel, 2004) o de caracteres (Schleimer et˜al., 2003). Se aplica una **función hash** a cada fragmento s, lo que genera un valor numérico prácticamente único (la probabilidad de que un s' ($s' \neq s$) genere el mismo valor es prácticamente nula y cambiar únicamente un carácter de s modifica completamente el valor hash resultante. Dividir el texto en oraciones permite que la comparación se haga de manera eficiente, pero también se pueden usar n-gramas que son menos sensibles a las modificaciones del texto reusado. Así, el valor hash de d', considerando la función Karp-Rabin (Karp and Rabin, 1987), es:

$$hash(s) = 3041551560959492699$$

Para cada fragmento de texto se aplica la misma función y todos los valores resultantes se guardan en una base de datos. Algunos modelos utilizan representaciones de los documentos completos (Brin et˜al., 1995), mientras que otros realizan un sub-muestreo (Schleimer et˜al., 2003). Cuando un documento sospechoso d_q se analiza en busca de plagio, se llevan a cabo las mismas operaciones: se divide en fragmentos y a cada fragmento se le aplica la misma función hash. Los valores resultantes se buscan en la base de datos y, en caso de encontrarse, son presentados al usuario como potenciales casos de plagio.

En el ejemplo ofrecido, dado que s' ha sido copiado exactamente de s, los valores hash obtenidos son iguales, por lo que el caso de plagio es detectado (considérese, por ejemplo, que $hash$(El curioso incidente del perro a la medianoche) = 399429840814458043).

Esta familia de métodos proporciona resultados de manera muy rápida y precisa. Sin embargo, si $\mathcal{A}$ modifica un solo carácter del texto que plagia, el método no es capaz de detectar el caso. Por ello, para otros tipos de plagio es necesario considerar métodos más flexibles.

2.2. Copia modificada

En este caso, $\mathcal{A}$ realiza distintas operaciones antes de reutilizar el texto. Por ejemplo, consideremos los siguientes dos:

$s'_1 =$ `El curioso incidente del sabueso a medianoche`
$s'_2 =$ `El curioso incidente a media noche de mi perro`

Evidentemente, los modelos de huella digital son inútiles en la detección de este tipo de plagio y se necesitan representaciones más flexibles. Antes que nada, lo que se busca es estimar cuál es la similitud entre los fragmentos de texto s y

s'_n; es decir, $sim(s, s')$. Una de las medidas de similitud más utilizadas en PLN y RI es la **similitud de coseno**. Esta medida devuelve un valor real entre 0 y 1 tal que $sim(s, s') = 0$ implica que s y s' son completamente diferentes y $sim(s, s') = 1$ son exactamente iguales. Sin embargo, para simplificar la explicación de los modelos, optaremos en este caso por utilizar el **coeficiente de Jaccard**, que, a diferencia de la medida del coseno, descarta cualquier peso de las palabras y considera los textos como simples conjuntos.

Otra cuestión relevante es cómo deben representarse los documentos o fragmentos de texto. Una primera opción sería representarlos con el sencillo modelo de **bolsa de palabras**, en el que los elementos que representan a los textos son las mismas palabras. Así, la similitud estimada entre los textos es:

$$sim(s, s'_1) = \frac{6}{8} = 0{,}75 \ , \tag{1}$$

mientras que

$$sim(s, s'_2) = \frac{5}{11} = 0{,}54 \ . \tag{2}$$

Debido a que los ejemplos elegidos son sencillos, los fragmentos resultan bastante parecidos y su similitud es alta. Sin embargo, considérese que en un marco más realista, f debe compararse contra millones de f's para generar una lista ordenada con base en las similitudes estimadas. Por ello, algunos modelos proponen, antes de realizar la comparación, llevar a cabo una normalización semántica: cada palabra se expande a todas las palabras que guardan una relación semántica con él Kang et˜al. (2006); Alzahrani and Salim (2010).

Por otro lado, se ha observado que, en el caso de detección de plagio, el uso del modelo de bolsa de palabras no es siempre el mejor. Es fácil intuir que, al considerar casos reales, es muy probable que dos fragmentos sobre el mismo tema tengan una alta cantidad de palabras en común. Por ello, se ha observado que es mejor considerar n-gramas de nivel 2 o 3 (2-gramas o 3-gramas) (Clough and Gaizauskas, 2009; Barrón-Cedeño and Rosso, 2009). Al considerar 2-gramas, las similitudes resultantes son:

$$sim_2(s, s'_1) = \frac{4}{8} = 0{,}5 \ , \tag{3}$$

mientras que

$$sim_2(s, s'_2) = \frac{2}{12} = 0{,}16 \ . \tag{4}$$

Ahora considérese la siguiente reformulación:

s'_3 = `Esta noche la mascota sufrió un accidente muy extraño`

Para los modelos descritos anteriormente la detección de este tipo de plagio, generado por medio de una paráfrasis por sustitución, sinonimia, generalización y cambio de orden resulta un tanto más complicado de ser detectado. Sin juzgar si este se trata de un caso verdadero de plagio o no, eso es tarea del experto, localizar este tipo de fragmentos altamente relacionados puede ser de gran relevancia para tomar dicha decisión. Por ello, es necesario recurrir a modelos de detección de paráfrasis (*cf.* sección 3), los cuales hasta el momento se han mantenido un tanto distanciados en el desarrollo de métodos de detección de plagio, probablemente debido a su complejidad.

2.3. Plagio traducido

En este caso $\mathcal{A}$ incluye un fragmento de s' en d' que proviene (y ha sido traducido) de un documento escrito originalmente en otra lengua. Así, el fragmento generado es el siguiente:

$s' =$ `The curious incident of the dog in the Night-Time` [2]

Este tipo de plagio ha recibido atención apenas recientemente (Barrón-Cedeño et˜al., 2008; Ceska et˜al., 2008; Pinto et˜al., 2009; Potthast et˜al., 2010b), quizás debido a la complejidad que implica. Sin embargo, es importante desarrollar modelos para su detección ya que, como Barrón-Cedeño et˜al. (2010) estiman, el fenómeno del plagio translingüe es frecuente, sobre todo cuando no existen muchos recursos en la lengua de $\mathcal{A}$ que opta por plagiar.

Uno de los modelos más sencillos para detectar este tipo de plagio es traducir el texto sospechoso. Usando una herramienta de traducción online[3], s' se convierte en:

$s'_t =$ `El curioso incidente del perro en la noche-tiempo`

Una vez traducido, cualquiera de las técnicas señaladas anteriormente puede ser aplicada. Sin embargo, como se puede observar en el ejemplo, los traductores automáticos suelen cometer errores. Si bien en este caso no afecta demasiado a la estimación de similitud, en general su influencia es mucho más negativa. Por ese motivo, Potthast et˜al. (2010b) han propuesto el uso de tres modelos que, de cierta manera, pueden ser considerados complementarios[4]. El primero de ellos se conoce como CL-ESA (del inglés Cross-Language Explicit Semantic Analysis)(Potthast and Stein, 2008). Dicho modelo explota el multilingualismo de Wikipedia. s y s' son comparados primero con un conjunto de artículos de Wikipedia en su correspondiente lengua (en este caso inglés y español). La única condición es que dichos artículos aborden exactamente el mismo tema. Una vez se ha hecho la comparación de s (s') con los artículos correspondientes, se forma un vector con el que se puede estimar $sim(s, s')$. Los indicios del modelos son mostrados gráficamente en la figura 1.

El segundo modelo se conoce como CL-ASA (del inglés Cross-Language Alignment-based Similarity Analysis) (Barrón-Cedeño et˜al., 2008). Dicho modelo está basado en los principios estadísticos de la traducción automática (Brown et˜al., 1990) pero sin llevar a cabo realmente una traducción. Para cada palabra en el documento sospechoso se consideran todas las posibles traducciones disponibles en un **diccionario probabilístico** previamente estimado. De esa manera, se intenta reducir el error causado por elegir una traducción que no era la adecuada para el contexto. Además, el modelo considera la longitud de los textos a comparar. Para darse una idea de este parámetro, hay que considerar el hecho de que, por ejemplo, dado un texto en inglés y su traducción al francés, en general, el texto en francés será más largo. Es este precisamente el factor

[2] En realidad este fragmento de texto es el título del libro escrito por Mark Haddon (2004). Nos hemos permitido utilizarlo en este caso para ejemplificar los distintos tipos y modelos

[3] El servicio de traducción de Google

[4] Debido a su complejidad, el cálculo de similitud con base en los dos primeros modelos no se incluye.

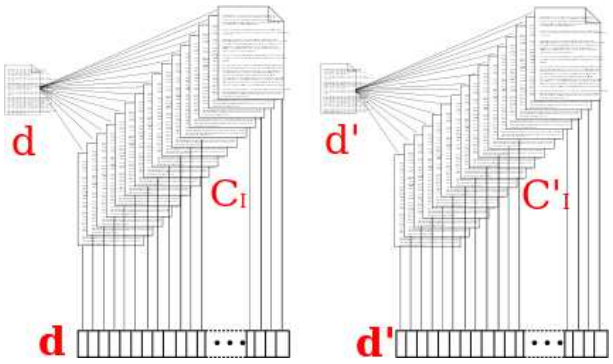

Figura 1: Representación gráfica del proceso de estimación de similitud entre textos en distintos idiomas con el modelo CL-ESA. d y d' representan los fragmentos de texto a comparar. C_I y C'_I son las colecciones de artículos de Wikipedia en la lengua correspondiente. Para cada c_i existe un artículo sobre el mismo tema c'_I. Los vectores resultantes pueden compararse por medio de la similitud del coseno.

que se considera en este caso. La mayor complejidad de CL-ASA es que requiere una colección importante de documentos traducidos para aprender tanto el diccionario como el modelo de longitud; una tarea que no es trivial.

Por último, el modelo más sencillo es CL-CNG (del inglés Cross-Language Character n-grams)(Mcnamee and Mayfield, 2004). Este modelo ha mostrado muy buenos resultados en aquellos casos en los que las lenguas implicadas guardan alguna relación por ejemplo, comparten las mismas raíces. Los pasos son muy sencillos: (*i*) los espacios y signos de puntuación son descartados; (*ii*) las cadenas resultantes se dividen en 3-gramas a nivel de caracteres; y (*iii*) los conjuntos resultantes se comparan con base en la similitud del coseno o el coeficiente de Jaccard. Tras el primer paso, las cadenas resultantes son las siguientes:

s' = el*curio*so*incident*edelperroamedianoche
s' = the*curio*us*incident*ofthedogin*the*night*time*

Los fragmentos con 3-gramas en común están en itálicas. Si bien la similitud estimada por este método devuelve valores mucho menores a aquellos que abordan el problema monolingüe, al final estas bajas similitudes logran diferenciar pares de textos que son potencialmente traducciones de los que no lo son.

El estado del arte en cuestión de detección de plagio se basa en un proceso dividido en tres pasos (Stein et˜al., 2007); dados d_q y D (donde D suele ser una colección enorme de documentos, por ejemplo, Internet mismo): (*i*) recuperación heurística: aquellos documentos en D que son más similares a d_q, ya sea en cuanto a temática abordada o estilo, son recuperados; (*ii*) comparación exhaustiva: d_q es comparada con a cada uno de los documentos d de la subcolección previamente recuperada, dando como resultado pares de fragmentos s_q, s, es decir, el potencial plagio acompañado de su fuente; y finalmente (*iii*) postprocesamiento: aquellos casos que no sean verdaderos casos de plagio, por ejemplo los que incluyen la referencia adecuada o no son suficientemente similares, son descartados. A este esquema se pueden reducir los modelos diseñados por la mayoría de los dieciocho participantes en la segunda competencia sobre detección automática de plagio, PAN 2010 (Potthast et˜al., 2010a), patrocinada por

Yahoo! Research. Algunos casos de plagio, sobre todo los de mayor grado de reformulación, aún son difíciles de hallar automáticamente.

3. Detección de paráfrasis

Son paráfrasis aquellas expresiones lingüísticas diferentes en la forma pero con (aproximadamente) el mismo significado. Los siguientes son ejemplos prototípicos de paráfrasis, ya que, a pesar de las modificaciones en la forma, el significado se mantiene:

1. Me dijo que no pensaba participar en el concurso
   Me comentó que no pensaba participar en el concurso

2. Joan Ponç pintó *Suite última tauromaquia* en 1982
   Suite última tauromaquia fue pintada en 1982 por Joan Ponç

Frecuentemente, plagiar consiste en aplicar la estrategia discursiva de la paráfrasis. Así, cuando plagiamos, podemos parafrasear; y todo plagio puede ser considerado, en esencia, una paráfrasis.

Comparando los ejemplos 1 y 2 con los de plagio con modificación (*cf.* sección 2.2), se observa claramente como, en realidad, nos encontramos ante un mismo fenómeno. Los únicos casos de plagio que cabría situar fuera del ámbito de la paráfrasis son la copia exacta y la traducción. En el primero, la forma de los dos miembros del par es la misma; en el segundo, interviene más de una lengua.

Con todo, tanto la copia exacta como la traducción también pueden considerarse, de algún modo, casos de paráfrasis. El primero puede situarse en uno de los extremos de un continuo de paráfrasis que iría desde la identidad semántica absoluta hasta la ausencia de identidad. La copia exacta se situaría en el primero de esto extremos y la no-paráfrasis, en el segundo. A lo largo de este continuo tendrían cabida todas las maneras de expresar formas parafrásticas. Por otro lado, Milićević (2007, p. 56), propone considerar la traducción como un caso particular de paráfrasis: la paráfrasis intralingüística.

Desde la perspectiva del PLN, y teniendo en cuenta el planteamiento que acabamos de exponer, el plagio puede verse como la aplicación de mecanismos parafrásticos orientados a un determinado fin: copiar lo que han escrito otros autores sin que se note. Así, la paráfrasis está en la base del plagio que va más allá de la copia exacta. En concreto, consideramos relevantes las tipologías de paráfrasis, que no dejan de ser tipologías de plagio, y las distintas aproximaciones al tratamiento computacional de la paráfrasis, que podrían ser aplicadas (algunas de ellas) a la detección automática de plagio.

3.1. Tipologías de paráfrasis

La aparente simplicidad de la paráfrasis (diferente forma, mismo significado) se desvanece cuando nos damos cuenta de que, en realidad, nos encontramos ante un fenómeno complejo, de límites difusos y con una amplia variedad de manifestaciones que pueden implicar conocimiento de tipo morfológico, léxico, sintáctico, semántico y pragmático. Para dar cuenta de esta complejidad, en el marco de la lingüística y también del PLN, se han construido varias tipologías de paráfrasis: Dras (1999), Fujita (2005), Bhagat (2009), Žolkovskij and Mel'čuk

(1965) y Milićević (2007), entre otras. No obstante, estas tipologías no cubren el fenómeno de la paráfrasis en su totalidad y/o lo analizan desde una perspectiva muy minuciosa de casos concretos. Asimismo, algunas de ellas se centran en un tipo de paráfrasis determinado —la paráfrasis sintáctica en el caso de Dras (1999)— o se enmarcan en una teoría lingüística específica difícilmente implementable —la teoría significado-texto en los casos de Žolkovskij and Mel'čuk (1965), y Milićević (2007).

La tipología que presentamos a continuación pretende ofrecer una visión amplia e inclusiva del fenómeno de la paráfrasis, sin centrarse en los mecanismos morfológicos, léxicos o sintácticos concretos[5]. Se organiza en cinco grandes tipos en función de la operación que se ha realizado para la obtención de la forma parafrástica: (*i*) sustitución de una pieza léxica por otra, (*ii*) eliminación de piezas léxicas, (*iii*) transformación estructural, (*iv*) modificación de la segmentación en oraciones y (*v*) cambio de orden de las piezas léxicas. Cada uno de estos cinco tipos alberga diversos fenómenos, de los que aquí solo se citan algunos[6].

Hay que señalar que, si bien estos tipos de paráfrasis se presentan de forma independiente, normalmente aparecen combinados. Las transformaciones, por ejemplo, suelen ir acompañadas de algún tipo de eliminación[7].

3.1.1. Sustitución

Sustitución de una pieza léxica por otra.

Sinonimia Sustitución de una pieza léxica por uno de sus sinónimos.

> Me *dijo* que no pensaba participar en el concurso
> Me *comentó* que no pensaba participar en el concurso

Antonimia Sustitución de una pieza léxica por su antónimo. Dicha sustitución se acompaña de otro tipo de modificación(es) (cambio de orden, en el ejemplo).

> Las ciudades del norte son más *ricas* que la zona costera
> La zona costera es más *pobre* que las ciudades del norte

Generalización Sustitución de una pieza léxica por otra de contenido más genérico, el hiperónimo en muchos casos.

> El curioso incidente del *sabueso* a media noche
> El curioso incidente del *perro* a media noche

Sustitución acción-actante Sustitución de una pieza léxica que representa la acción por otra que representa uno de los actantes de dicha acción.

> No soporto la *conducción* imprudente
> No soporto a los *conductores* imprudentes

[5] Algunos tipos de paráfrasis son bidireccionales (e.g., generalización-especificación). No obstante, los nombramos señalando sólo una de estas direcciones (e.g., generalización).

[6] Dado que la descripción de la tipología no es el objetivo final de este artículo, no nos extendemos en la exposición de dichos fenómenos.

[7] Los ejemplos de esta sección han sido creados ad hoc o extraídos y adaptados de otras fuentes: Bhagat (2009), Dras (1999), Fujita (2005), Milićević (2007), Žolkovskij and Mel'čuk (1965) y Pustejovsky (1995).

Sustitución palabra-definición Sustitución de una pieza léxica por su definición.

Necesito *cuerda*
Necesito *algo para atar*

3.1.2. Eliminación

Eliminación de una o más piezas léxicas.

Eliminación de contenido no proposicional Eliminación de una o más piezas léxicas de contenido no proposicional.

Juan *hizo* un intento para dejar de fumar
Juan intentó dejar de fumar

Eliminación de argumentos Eliminación de una o más piezas léxicas que representan uno de los argumentos del predicado.

Joan Ponç pintó Suite última tauromaquia en 1982
Suite última tauromaquia fue pintada en 1982

Eliminación de adjuntos Eliminación de una o más piezas léxicas que constituyen elementos adjuntos del predicado.

Arturo se fue corriendo a casa *a eso de las 12*
Arturo se fue corriendo a casa

Cambio en la estructura argumental Cambio en el tipo de argumento regido por el verbo. En el ejemplo, el verbo empezar exige una oración subordinada, en la forma parafrástica, se omite el predicado.

María empezó a *leer* el libro
María empezó el libro

3.1.3. Transformación

Transformación de la estructura oracional o sintagmática (paso de activa a pasiva, en el ejemplo).

Mamá escribió la nota
La nota fue escrita por mamá

3.1.4. Segmentación

Segmentación de la estructura oracional o sintagmática en dos o más estructuras independientes.

Michael Phelps batió el record mundial de los 100 m
mariposa en un tiempo de 49 segundos y 82 centésimas
Michael Phelps batió el record mundial de los 100 m
mariposa. Hizo un tiempo de 49 segundos y 82 centésimas

3.1.5. Cambio de orden

Cambio de orden de las piezas léxicas.

Antes irse a su casa, Blanca pasó por la biblioteca
Blanca pasó por la biblioteca antes de irse a su casa

3.2. Aproximaciones al tratamiento computacional de la paráfrasis

El tratamiento computacional de la paráfrasis aplica métodos y técnicas de naturaleza muy diversa. A continuación, se presentan cinco tipos de aproximación al tratamiento de la paráfrasis[8].

3.2.1. Hipótesis distribucional

Aquellas expresiones lingüísticas que aparecen en contextos similares tienden a compartir el significado (Harris, 1954). En el ejemplo, estos dos fragmentos pueden considerarse como paráfrasis por el hecho de compartir el contexto: a y b (Bhagat and Ravichandran, 2008; Lin and Pantel, 2001; Vila et˜al., 2010).

{la lluvia}$_a$ *volvió a intervenir, interrumpiendo el duelo durante* {casi dos horas}$_b$
{casi dos horas}$_b$ de *suspensión que tuvo el partido a causa de* {la lluvia}$_a$

3.2.2. Matching

Aquellas expresiones lingüísticas que comparten un gran número de unidades lingüísticas tienden a compartir también el significado. En el ejemplo, dado que estos dos fragmentos comparten un gran número de unidades, pueden considerarse como paráfrasis.

El matching puede realizarse mediante bolsa de palabras, donde las **entidades nombradas** tienen un papel relevante dada su estabilidad, o n-gramas (Barzilay and Lee, 2003).

Con la victoria de Rafa Nadal en la final del Abierto de los Estados Unidos, el español consigue los {cuatro}$_a$ Grandes Torneos del tenis mundial (los denominados {Grand Slam}$_b$), una hazaña que hasta ahora sólo habían logrado {seis}$_c$ jugadores: {Fred Perry}$_d$, Donald {Budge}$_e$, {Roy Emerson}$_f$, {Rod Laver}$_g$, {Andre Agassi}$_h$ y {Roger Federer}$_i$.

Anteriormente sólo {seis}$_c$ jugadores habían conseguido completar los {cuatro}$_a$ torneos del {Grand Slam}$_b$: los estadounidenses {Andre Agassi}$_h$ (1999) y Don {Budge}$_e$ (1938); los australianos {Rod Laver}$_g$ (1962) y {Roy Emerson}$_f$ (1964); el inglés {Fred Perry}$_d$ (1935); y el suizo {Roger Federer}$_i$ (2009).

[8]Los ejemplos de esta sección han sido extraídos de las versiones online de RTVE, SPORT, ABC y Clarín (14/09/2010), o de las referencias citadas.

3.2.3. Distancia de edición

Aquellas expresiones lingüísticas separadas por una distancia de edición pequeña tienden a compartir el significado. Existen varios algoritmos para calcular la distancia de edición. Uno de los utilizados en paráfrasis es la **distancia de Levenshtein**. En el ejemplo, podemos observar dos expresiones lingüísticas con una distancia de edición baja (Dolan et˜al., 2004).

```
The leading indicators measure the economy...
The leading index measures the economy...
```

3.2.4. Traducción múltiple

Aquellas expresiones lingüísticas resultantes de la traducción de un mismo fragmento de texto en otra lengua (fragmento original, en el ejemplo) pueden verse como paráfrasis (traducciones 1 y 2 en el ejemplo) (Zhao et˜al., 2009; Barzilay and McKeown, 2001).

orig `Emma pleurait, et il s'efforçait de la consoler, enjolivant de calembours ses protestations (Flaubert, *Madame Bovary*)`

$trad_1$ `Emma burst into tears and he tried to comfort her, saying things to make her smile.`

$trad_2$ `Emma cried, and he tried to console her, adorning his words with puns.`

3.2.5. Aplicación de reglas

Comprobar si los candidatos a paráfrasis cumplen una serie de reglas creadas manual o automáticamente (Barzilay et˜al., 1999).

```
Head omission: group of students/students
Ordering of sentence components: Tuesday they met.../They
met ... Tuesday
```

4. Conclusiones

En este capítulo hemos presentado el panorama actual de los modelos automáticos para la detección de plagio, una tarea en la que se combinan métodos de recuperación de información y procesamiento del lenguaje natural. Desde la perspectiva de este último, y teniendo en cuenta el planteamiento que acabamos de exponer, podemos considerar que el plagio es la aplicación de mecanismos parafrásticos orientados a un determinado fin: copiar lo que han escrito otros autores pero procurando que el lector no lo note. Así, la paráfrasis está en la base de distintos tipo de plagio.

Hemos observado que existen modelos automáticos, que de hecho se aplican ya en sistemas comerciales, para detectar casos de copia exacta y copia con ligeras modificaciones. No obstante, éste no es el caso del plagio en que las modificaciones van más allá de simples cambios por sinónimos o cambios de

orden. En este sentido, consideramos que los trabajos existentes sobre paráfrasis, así como recuperación de información translingüe y traducción estadística, constituyen una fuente de conocimiento y herramientas para la mejora de los sistemas actuales de detección automática de plagio, por lo que es necesario investigar la mejor manera de aplicarlos.

En concreto, consideramos relevantes las tipologías de paráfrasis, que no dejan de ser tipologías de plagio, y las distintas aproximaciones al tratamiento computacional de la paráfrasis, que pueden ser aplicadas (algunas de ellas) a la detección automática de plagio.

Agradecimientos

Agradecemos a M. Antònia Martí y Horacio Rodríguez por sus valiosos comentarios sobre las versiones preliminares de este documento. Este trabajo ha sido parcialmente financiado por las becas CONACYT-Mexico 192021 y FPU AP2008-02185 (Ministerio de Educación), así como los proyectos MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03, TEXT-MESS 2.0 TIN2009-13391-C04-04, ANCORA-NET FFI2009-06497-E/FILO y ClInt FFI2009-06252-E/FILO (Plan I+D+i).

Referencias

Salha Alzahrani and Naomie Salim. Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection. In Braschler and Harman (2010).

Gregory L. Anderson. Cyberplagiarism. a look at the web term paper sites. *College & Research Libraries News*, 60(5):371–373, 1999.

Association of Teachers and Lecturers. School Work Plagued by Plagiarism - ATL Survey. Technical report, Association of Teachers and Lecturers, London, UK, 2008. Press release.

Alberto Barrón-Cedeño and Paolo Rosso. On Automatic Plagiarism Detection based on n-grams Comparison. *Advances in Information Retrieval. Proceedings of the 31st European Conference on IR Research*, LNCS (5478):696–700, 2009.

Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. On Cross-lingual Plagiarism Analysis Using a Statistical Model. In Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, *ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008)*, pages 9–13. CEUR-WS.org, 2008.

Alberto Barrón-Cedeño, Paolo Rosso, Eneko Agirre, and Gorka Labaka. Plagiarism Detection across Distant Language Pairs. In Huang and Jurafsky (2010).

Regina Barzilay and Lillian Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*, pages 16–23, 2003.

Regina Barzilay and Kathleen McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the ACL 2001*, pages 50–57, 2001.

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Information fusion in the context of multi-document summarization. In *Proceedings of the ACL 1999*, pages 550–557, 1999.

Yaniv Bernstein and Justin Zobel. A Scalable System for Identifying Co-Derivative Documents. In *Proceedings of the Symposium on String Processing and Information Retrieval*, pages 55–67. Springer, 2004.

Rahul Bhagat. *Learning Paraphrases from Text*. PhD thesis, University of Southern California, 2009.

Rahul Bhagat and Deepak Ravichandran. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of the ACL 2008*, pages 674–682, 2008.

Martin Braschler and Donna Harman, editors. *Notebook Papers of CLEF 2010 LABs and Workshops*, September 2010.

Sergey Brin, James Davis, and Hector Garcia-Molina. Copy Detection Mechanisms for Digital Documents. In Michael J. Carey and Donovan A. Schneier, editors, *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 398–409. ACM Press, 1995.

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.

Enrique Bunbury. Hellville Deluxe. CD, 2008.

Zdenek Ceska, Michal Toman, and Karel Jezek. Multilingual Plagiarism Detection. In *Proceedings of the 13th International Conference on Artificial Intelligence*, pages 83–92. Springer Verlag Berlin Heidelberg, 2008.

Paul Clough. Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service, 2003. URL `http://ir.shef.ac.uk/cloughie/papers/pasplagiarism.pdf`.

Paul Clough and Robert Gaizauskas. Corpora and Text Re-Use. In Anke Lüdeling, Merja Kytö, and Tony McEnery, editors, *Handbook of Corpus Linguistics*, Handbooks of Linguistics and Communication Science, pages 1249—1271. Mouton de Gruyter, 2009.

Malcolm Coulthard and Johnson Alison. *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge, Oxon, UK, 2007.

Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING 2004*, pages 350–356, 2004.

Mark Dras. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. PhD thesis, Macquarie University, 1999.

Atsushi Fujita. *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. PhD thesis, Nara Institute of Science and Technology, 2005.

Yaakov HaCohen-Kerner, Aharon Tayeb, and Natan Ben-Dror. Detection of Simple Plagiarism in Computer Science Papers. In Huang and Jurafsky (2010), pages 421–429. URL `http://www.aclweb.org/anthology/C10-1048`.

Mark Haddon. *The Curious Incident of the Dog in the Night-Time*. Vintage, 2004.

Valerie J. Haines, George M. Diekhoff, George M. LaBeff, and Robert E. Clarck. College Cheating: Inmaturity, Lack of Commitment, and the Neutralizing Attitude. *Research in Higher Education*, 25(4):342–354, 1986.

Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

Chu-Ren Huang and Dan Jurafsky, editors. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, August 2010. Coling 2010 Organizing Committee.

IEEE. A plagiarism FAQ. http://www.ieee.org/web/publications/rights/plagiarism_FAQ.htm, 2008. [Online; accessed 3-March-2010].

iParadigms. Turnitin, 2010. URL `http://www.turnitin.com`. [Online; accessed 3-March-2010].

Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

NamOh Kang, Alexander Gelbukh, and SangYong Han. PPChecker: Plagiarism pattern checker in document copy detection. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proceedings of the Text, Speech and Dialogue, 10th International Conference (TSD 2006)*, volume LNCS (LNAI) (4188), pages 661–667. Springer-Verlag, 2006.

Richard M. Karp and Michael O. Rabin. Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, 31(2):249–260, mar. 1987. doi: 10.1147/rd.312.0249.

Narayanan Kulathuramaiyer and Hermann Maurer. Coping With the Copy-Paste-Syndrome. In *E-Learn 2007*, pages 1072—1079, Quebec, CA, 2007.

Dekang Lin and Patrick Pantel. DIRT-discovery of inference rules from text. In *Proceedings of ACM SIGKDD 2001*, pages 323–328, 2001.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

Brian Martin. Plagiarism: policy against cheating or policy against learning? `http://www.uow.edu.au/arts/sts/bmartin/`, 2004.

Iván A Martínez. Wikipedia usage by Mexican students. The constant usage of copy and paste. In *Wikimania 2009*, Buenos Aires, Argentina, 2009.

Hermann Maurer, Frank Kappe, and Bilal Zaka. Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8):1050–1084, 2006.

Mark McCrohon. DOC Cop, 2010. URL `http://doccop.com`. [Online; accessed 10-March-2010].

Paul Mcnamee and James Mayfield. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97, 2004.

Jasmina Milićević. *La paraphrase*. Peter Lang, Berne, 2007.

Daniel Müllensiefen and Marc Pendzich. Court Decisions on Music Plagiarism and the Predictive Value of Similarity Algorithms. *Musicae Scientiae. Discussion Forum 4B*, pages 257–295, 2009.

John Olsson. *Forensic Linguistics*. Continuum International Publishing Group, New York, NY, 2008.

David Pinto, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso. A Statistical Approach to Crosslingual Natural Language Tasks. *Journal of Algorithms*, 64(1):51–60, 2009.

Martin Potthast and Benno Stein. New Issues in Near-Duplicate Detection. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 601—609, Berlin Heidelberg New York, 2008. Springer.

Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. Overview of the 2nd International Competition on Plagiarism Detection. In Braschler and Harman (2010).

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. Cross-Language Plagiarism Detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 2010b. doi: 10.1007/s10579-009-9114-z.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, 2010c.

James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

Real Academia Española. Diccionario de la lengua española. Vigésima segunda edición, 2008. URL `http://www.rae.es/rae.html`. Real Academia Española.

Pere Ríos. La juez ve plagio en 'La Cruz de San Andrés' de Cela, 04 2009. URL `http://www.elpais.com/articulo/cultura/juez/ve/plagio/Cruz/San/Andres/Cela/elppgl/20090421elpepicul_3/Tes`.

Matthias Robine, Pierre Hanna, Pascal Ferraro, and Julien Allali. Adaptation of String Matching Algorithms for Identification of Near-Duplicate Music Documents. In Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, *SIGIR 2007 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007)*, 2007.

Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. Winnowing: Local Algorithms for Document Fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, New York, NY, 2003. ACM.

Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. Strategies for Retrieving Plagiarized Documents. In Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen de Vries, editors, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 825–826, Amsterdam, The Netherlands, 2007. ACM.

Kraüpl Taylor. Cryptomnesia and Plagiarism. *The British Journal of Psychiatry*, 111:1111–1118, 1965.

Marta Vila, Horacio Rodríguez, and M. Antònia Martí. Wrpa: A system for relational paraphrase acquisition from wikipedia. *Procesamiento del Lenguaje Natural*, 45:11–19, 2010.

Alexander Žolkovskij and Igor Mel'čuk. O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-texničeskaja informacija*, 5:23–28, 1965.

Stefan Weber. *Das Google-Copy-Paste-Syndrom. Wie Netzplagiate Ausbildung und Wissen gefahrden.* Telepolis, 2007.

Wikipedia. Hash, 2010. URL `\url{http://es.wikipedia.org/wiki/Hash}`. [Online; accessed 17-Septiembre-2010].

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. Extracting paraphrase patterns from bilingual parallel corpora. *Natural Language Engineering*, 15 (4):503–526, 2009.

A. Conceptos básicos

bolsa de palabras Bajo este modelo un texto es representado por las palabras que contiene sin respetar orden alguno. Si bien el sentido del texto se pierde, desde un punto de vista matemático ello no afecta en los cálculos, en este caso, de similitud.

coeficiente de Jaccard Esta medida calcula la similiud entre dos conjuntos (Jaccard, 1901):

$$sim(d, d_q) = J(d, d_q) = \frac{|v_d \cap v_{d_q}|}{|v_d \cup v_{d_q}|} \ . \tag{5}$$

es decir, la intersección dividida por la unión de los conjuntos. v_k es el vocabulario contenido en k.

diccionario probabilístico Un diccionario que incluye todas las posibles traducciones de una palabra en otro idioma. Incluye la probabilidad de cada palabra de ser traducida por otra.

distancia de Levenshtein La también conocida como distancia de edición representa el número mínimo de operaciones necesarias para convertir una cadena (por ejemplo, palabra u oración), en otra. Las operaciones suelen ser sustitución, inserción y eliminación.

entidad nombrada Son entidades nombradas los nombres de personas, lugares, organizaciones y fechas.

función hash Una función para generar claves que representan de manera casi unívoca a un documento, texto o archivo Wikipedia (2010). Una de las funciones hash má conocidas es la de Karp and Rabin (1987)

n-grama Una representación redundante de texto que consiste en fragmentos de texto solapados (ya sea a nivel de caracter o de palabra) cuya longitud es n. Por ejemplo, los 3-gramas, a nivel de caracteres de "*ejemplo*" son $[eje, jem, emp, mpl, plo]$; y los 2-gramas a nivel palabra de "*éste es sólo un ejemplo*" son *[éste es, es sólo, sólo un, un ejemplo]*.

Procesamiento de lenguaje natural Un campo interdisciplinario que combina principios de lingüística y ciencias de la computación para la generación, comprensión y procesamiento de lengua hablada y, sobre todo, escrita.

Recuperación de información Un área interdisciplinaria que combina principios de ciencias de la computación, ciencias de la información y estadística, entre muchos otros, para la recuperación, adquisición y procesamiento de información de distintos tipos (texto, imágenes, sonidos, etc.)

Similitud del coseno Una medida de similitud entre dos vectores A y B. En el caso de texto, cada dimensión suele venir dada por una palabra (o cualquier otra representación definida) y un peso que representa su relevancia en el documento o fragmento (usualmente frecuencia, ya sea normalizada o no). La similitud del coseno se expresa matemáticamente como:

$$cos(d, d_q) = \frac{\sum_{t \in d \cap d_q} \left(\omega_{t,d} \cdot \omega_{t,d_q}\right)}{\sqrt{\sum_{t \in d} \left(\omega_{t,d}\right)^2 \cdot \sum_{t_q \in d_q} \left(\omega_{t,d_q}\right)^2}} \ , \tag{6}$$

donde $\omega_{t,d}$ is el peso de la palabra t en el documento d. Es decir, el numerador etá compuesto por la suma de los productos de todas las palabras que los dos documentos tienen en común (conocido como producto punto). El denominador, está basado en la magnitud de ambos vectores considerados y su función es la de normalizar la estimación final. Esta medida puede ser calculada a nivel de documento d o fragmento s. El lector interesado puede leer más al respecto en (Manning and Schütze, 1999).