Creado el 15/08/2014
Última versión: 26/08/2014
Paolo Rosso, Universitat Politècnica de Valencia

Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

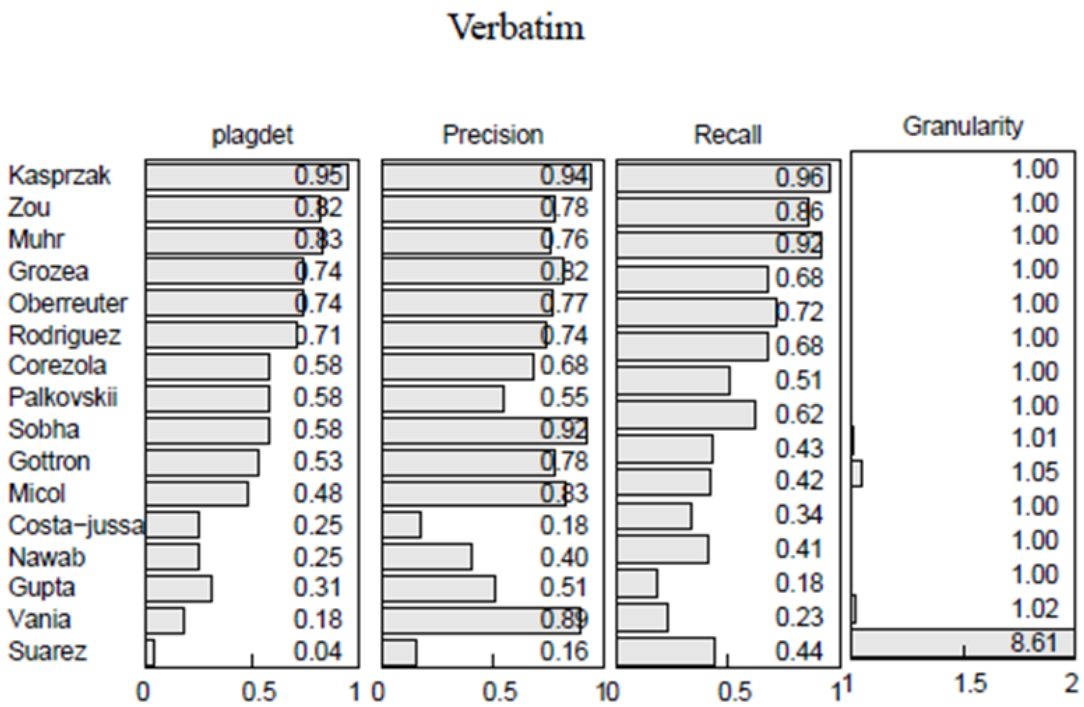
Aplicaciones de las TLH/PLAGIO Y AUTORÍA

Detección automática del plagio (y 2)

Detección del plagio

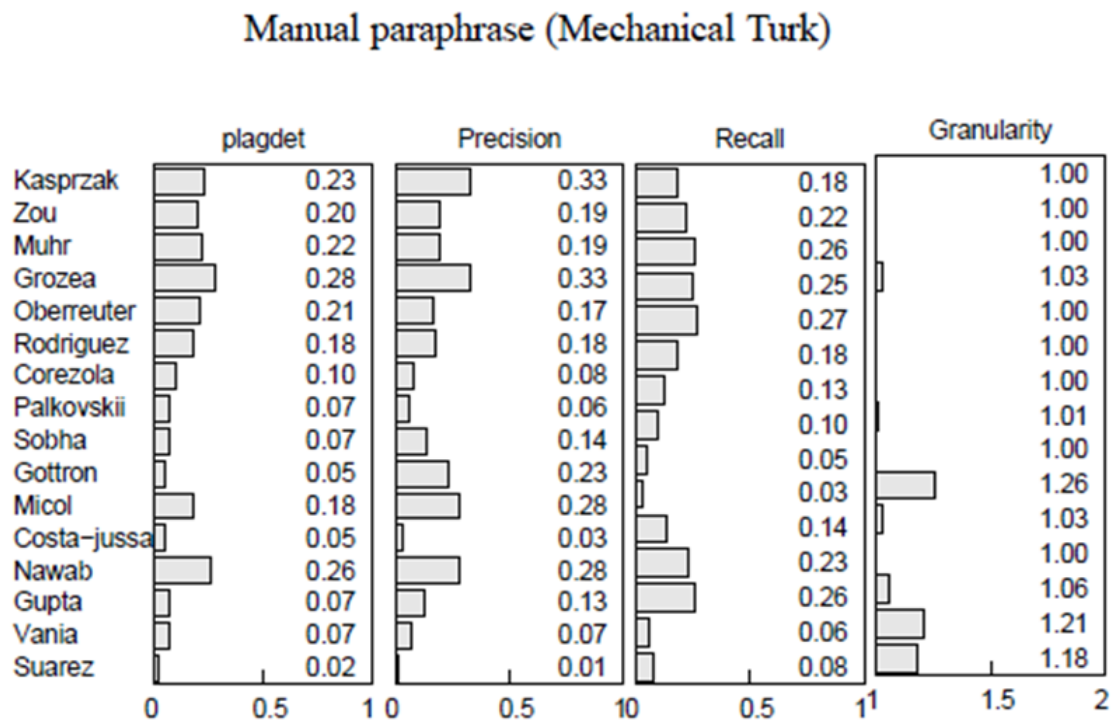
Dificultades

El plagio es bastante más difícil de detectar cuando ha ocurrido un proceso de **paráfrasis**. Si "solo" ha habido un plagio *verbatim* —"copiar-pegar"—, el estado actual de los sistemas de detección de plagio es que consiguen detectar la mayoría de los casos.



Por ejemplo, aquí veis los sistemas que han participado en una competición que se organizó en el 2010 y, sin entrar en detalle respecto a las medidas que se utilizan de "precision" y "recall" a nivel de carácter, si solo nos centramos en ésta medida que resume las dos anteriores, "plagdet" —de "detección de plagio"—, vemos que los mejores sistemas obtienen resultados muy altos, mientras que en el caso de textos que tienen paráfrasis los valores bajan a un cuarto,

es decir, los mejores sistemas no llegan ni al 30 %.



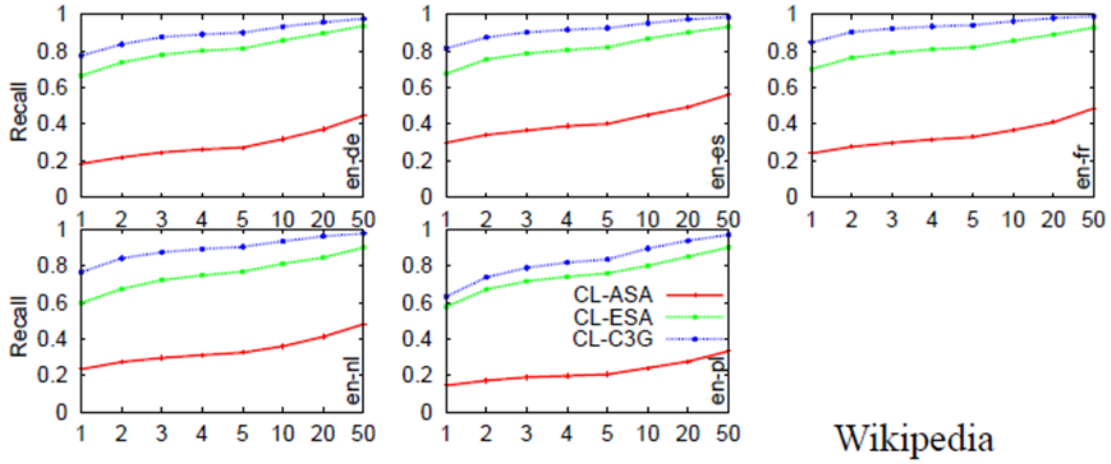
Aquí está disponible, en esta dirección, el corpus P4P de paráfrasis, que ha sido etiquetado con la ayuda de compañeras lingüistas de la Universidad de Barcelona del equipo de Toni Martí, donde cada paráfrasis ha sido etiquetada en función de una tipología desarrollada en la Universidad de Barcelona. Un vistazo nos hará entender cuáles son los tipos de paráfrasis más difíciles de detectar por parte de los sistemas actuales de detección de plagio.

Otra dificultad ocurre en la detención cuando el plagio es **translingüe**, es decir, se ha cometido plagio desde un texto escrito en otro idioma.

Se podría pensar en traducir luego de analizar el plagio a nivel monolingüe pero esto no siempre es posible. Además, no es muy realista pensar en traducir todos los documentos de Internet a un idioma.

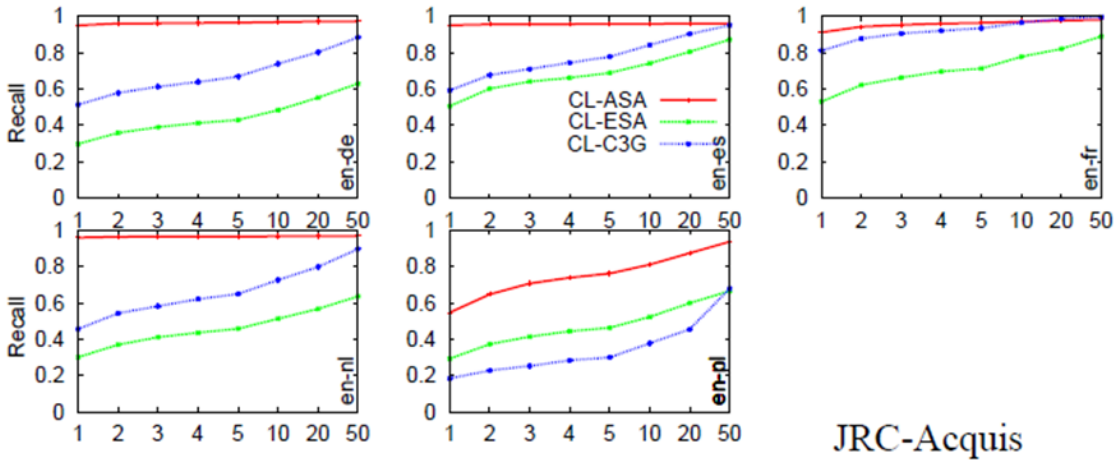
Hay diferentes aproximaciones a este problema concreto, desde la sencilla que se basa en n-gramas de caracteres, el CL-ASA que se basa en técnicas estadísticas sobre corpus paralelos, hasta el CL-ESA que necesita un corpus comparable como, por ejemplo, Wikipedia. O técnicas que se han desarrollado recientemente que se basan de en grafos de conocimiento o tesauros a nivel conceptual.

Aquí vemos una comparación que se hizo entre CL-ASA y CL-ESA, es decir, las dos técnicas que se basan en corpus comparables o paralelos, y con la de trigramas de caracteres.



La comparación se ha hecho utilizando parte de Wikipedia para pares de lenguas como inglés-alemán, inglés-holandés, inglés-polaco, inglés-español e inglés-francés.

Como podéis ver, de manera bastante sorprendente técnicas sencillas como los trigramas suelen tener unas prestaciones hasta mejor que las técnicas que se basan en corpus comparables como Wikipedia.



Pero cuando esta comparación ocurre con corpus paralelos como JRC-Acquis, métodos que han sido entrenados con corpus paralelos como CL-ASA suelen portarse mejor, por lo menos con los pares de idiomas más similares; un poco menos cuando los lenguajes en cuestión no son tan similares en sintaxis y alfabeto, como el inglés y el polaco.

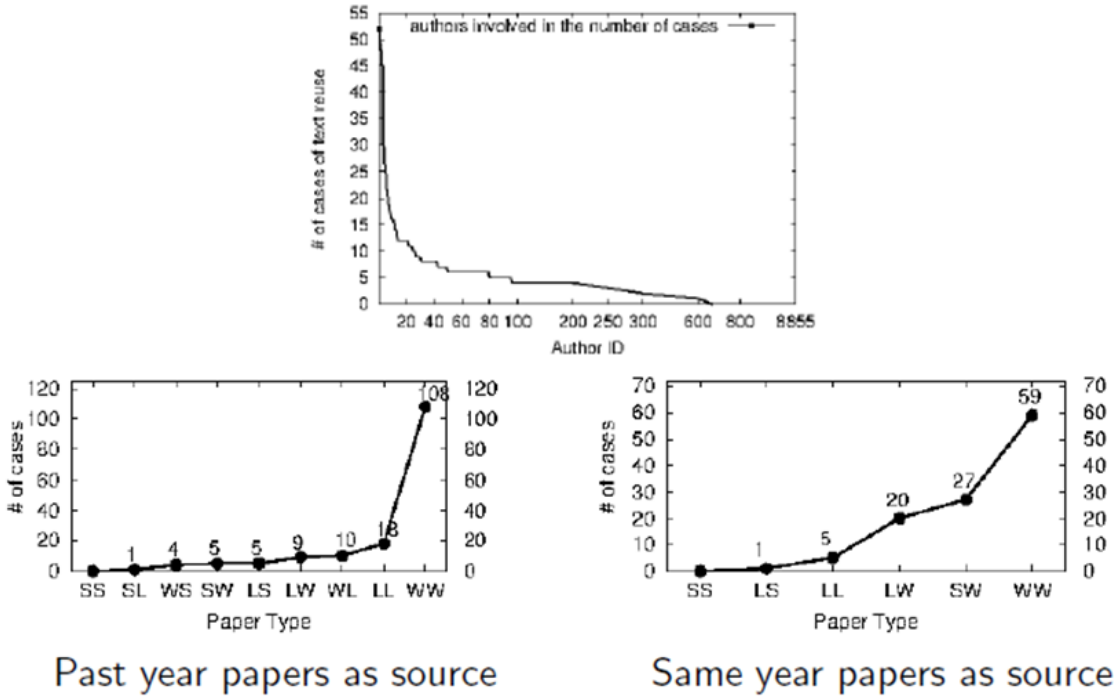
Competiciones

En estos años se han desarrollado diferentes competiciones sobre detección de plagio, la primera en el 2009 en el marco de la SEPLN, y actualmente se está siguiendo con esta competición en el marco del CLEF.

Otras competiciones más a nivel transligüe donde se han considerado idiomas como el indi o el bujarati han sido organizadas en otro foro, FIRE, donde este año (2014) organizaremos una competición sobre la detección de plagio en código fuente, SOCO.

Reuso de texto en la investigación

Y por último, hablar algo de reuso de texto en la investigación. Esto es un estudio que hicimos en el marco de la ACL para ver cuántos autores han reutilizado texto de sus propios artículos publicados anteriormente, o de artículos publicados por otros autores.



También se ha hecho un estudio con respecto a qué tipo de artículos y como veis y sobre todo esto ocurre entre artículos publicados en workshops de la ACL, o entre artículos "long"[1].

Esto ocurre también en el mismo año porque los artículos se pueden presentar como "long", como "short" o en el workshop[2]. La mayor parte de casos se refiere a plagio entre artículos que se han enviado a diferentes workshops.

Cuando hablamos de plagio, mejor hablar de reutilización de texto porque en este caso la mayor parte de los casos se refieren a autoplagio y no a "*cross*-plagio", plagio por parte de autores que reutilizan texto de otros.

Podéis investigar más en estos temas en algunas de las referencias que os propongo. Y también podéis contactarme para saber más sobre la competición de detección de plagio en código fuente; aquí tenéis una "demo" de una herramienta que ha sido desarrollada en la Universidad Politécnica de Valencia. También para hacer ejercicios sobre detección de plagio

---

[1] En la mayoría de congresos se suelen aceptar varios tipos de artículo: *long* y *short* hacen referencia a la longitud en páginas, el comité científico suele proponer que ciertos trabajos que consideren de menor relevancia se publiquen en formato corto y su presentación en el congreso y en las actas del mismo es diferente a la de los artículos largos.

[2] También es habitual que, además de la conferencia principal, se organicen talleres o conferencias satélite que se realizan en paralelo.

en castellano.

Por último, quiero agradecer a quien hizo o quien está haciendo un doctorado sobre detección de plagio aquí en la Universidad Politécnica de Valencia, y a los compañeros del PAN, en el marco del cual estamos organizando estas competiciones que he comentado anteriormente.

Muchas gracias por vuestra atención