UCES Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 260-274, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310

La detección de mensajes salientes de *papers* sobre investigación en ciencias sociales y su aplicación en búsqueda de documentos*

The detection of salient messages from social science research papers and its application in document search

Ágnes Sándor y Angela Vorndran*****

Resumen

El procesamiento del lenguaje natural proporciona instrumentos eficaces para ayudar a investigadores a enfrentarse con el cuerpo creciente de literatura científica. Uno de los usos más acertados y bien establecidos es la extracción de la información, por ejemplo, la extracción de entidades y hechos. Esta aplicación, sin embargo, no es del todo aplicable a las ciencias sociales, ya que los mensajes principales de las publicaciones no son hechos sino argumentos. En este artículo proponemos una metodología de procesamiento del lenguaje natural destinado a detectar oraciones que comunican mensajes salientes en trabajos de investigación pertenecientes a las ciencias sociales. Consideramos dos tipos de oraciones que contienen mensajes salientes: oraciones que resumen el artículo en su totalidad o partes del artículo y las oraciones que comunican cuestiones de investigación. Tales oraciones son detectadas usando un analizador gramatical de dependencia y reglas especiales de "unión de conceptos". En un experimento de prueba-de-concepto hemos mostrado la eficacia de nuestra proposición: buscando artículos en la base de documentos de ciencia educativa construida por el proyecto EERQI hemos descubierto que la presencia de la(s) palabra(s) de pregunta en las oraciones salientes detectadas por nuestro instrumento es un indicador importante de la importancia del artículo. Hemos comparado la importancia de los artículos recuperados con nuestro método con aquellos recuperados por el motor de búsqueda

* Este trabajo está subsidiado por el 7º Programa Marco Europeo para la Investigación en Ciencias Socioeconómicas y Temas sobre Humanidades (*Framework Programme for Research in the Socio-economic Sciences and Humanities Theme*) (SSH).
Los autores agradecen a Alexander Botte, Caroline Hagège, Aaron Kaplan y Sybille Peters por las discusiones útiles, comentarios y ayuda técnica sin lo cual no se podría haber llevado a cabo este trabajo.

** Centro de Investigación Europeo Xerox. Dirección: 6. Chemin Maupertuis, 38240 Meylan, Francia. E-mail: Agnes.Sandor@xrce.xerox.com

*** Instituto Alemán para la Investigación Educativa Internacional (DIPF). Dirección: Schlossstrasse 29, 60486 Frankfurt/M., Alemania. E-mail: vorndran@dipf.de

Lucene como configurado para la base de contenido de EERQI, con el ranking de importancia de omisión, que está basado en medidas de frecuencia de palabras. Los resultados son complementarios, lo cual señala la utilidad de la integración de nuestro instrumento en el Lucene.

Palabras clave: extracción de información, mensajes salientes, ciencias sociales, argumentos, errores sistemáticos.

Summary
Natural language processing provides effective tools to help researchers cope with the growing body of scientific literature. One of the most successful and well-established applications is information extraction, i.e. the extraction of named entities and facts. This application, however, is not well suited to social sciences, since the main messages of the publications are not facts, but rather arguments. In this article we propose a natural language processing methodology in order to detect sentences that convey salient messages in social science research papers. We consider two sentence types that bear salient messages: sentences that sum up the entire article or parts of the article and sentences that convey research issues. Such sentences are detected using a dependency parser and special "concept-matching" rules. In a proof-of-concept experiment we have shown the effectiveness of our proposition: searching for articles in the educational science document base built by the EERQI project we have found that the presence of the query word(s) in the salient sentences detected by our tool is an important indicator of the relevance of the article. We have compared the relevance of the articles retrieved with our method with those retrieved by the Lucene search engine as configured for the EERQI content base with the default relevance ranking which is based on word frequency measures. The results are complementary, which points to the utility of the integration of our tool into Lucene.

Key words: information extraction, salient messages, social sciences, arguments, systematic errors.

1. El problema

Los instrumentos de extracción de información proporcionan fragmentos estructurados de información fáctica comunicada por textos digitales. Una de las aplicaciones de la extracción de la información es el refinamiento del resultado de búsqueda en los géneros de texto donde los mensajes del texto son sobre todo fácticos. Entre las disciplinas académicas esto sucede -o al menos se supone que generalmente sucede- en el campo de las ciencias exactas. Ha habido esfuerzos recientes para identificar el rol argumentativo que los fragmentos de información extraídas juegan en el texto para permitir al usuario evaluar la relevancia de la información extraída (ver por ejemplo, de Waard et al., 2009).

Los mensajes de textos salientes de ciencia sociales no son típicamente hechos sino argumentos, interpretaciones, análisis, etc.

En consecuencia, las tecnologías de extracción de información tradicionales no son adecuadas para captarlos. Proponemos un instrumento que detecta las oraciones de las publicaciones de ciencia sociales que comunican tales mensajes salientes. El objetivo de nuestro instrumento es refinar la búsqueda en un motor de búsqueda operativo de ciencias sociales.

El 7º Programa Marco Europeo (Framework European Project) EERQI (http: // www.eerqi.eu/) está desarrollando un motor de búsqueda disponible al público (http: // www.eerqi.eu/page/index_raw.jsp) dedicado a la recuperación de trabajos de investigación en ciencias educativas, tanto de fuentes heterogéneas[1] de recolección de datos como de internet. El motor de búsqueda recupera literatura de investigación en ciencias educativas a modo de ejemplo, pero se propone ser útil en la recuperación de literatura de investigación en ciencias sociales en general. El motor de búsqueda EERQI usa la biblioteca disponible al público de Lucene. En el presente, Lucene clasifica los resultados según un algoritmo que usa medidas TF-IDF, por ejemplo, toma en cuenta la frecuencia de las palabras de consulta en el documento y la frecuencia inversa del documento que se relaciona con la incidencia de términos de consulta en todos los documentos de la recopilación.

En el marco del proyecto EERQI proponemos mejorar tanto el ordenamiento de jerarquías como la presentación de los fragmentos de información de los documentos recuperados, mediante la integración en Lucene de nuestro instrumento, el cual detecta mensajes salientes en artículos en inglés[2]. Para justificar nuestra propuesta hemos realizado un experimento de prueba-de-concepto, cuyos resultados son prometedores.

En esta comunicación brevemente describiremos una propuesta de determinar mensajes salientes en artículos de investigación en ciencias sociales, el instrumento de lenguaje natural que los detecta, la evaluación de nuestros resultados experimentales, y finalmente sacaremos algunas conclusiones para la integración del instrumento en el motor de búsqueda.

2. Mensajes salientes en artículos de investigación en ciencias sociales

Nuestra proposición para determinar mensajes salientes está inspirada, en parte, por el empleo de metainformación en la biblioteconomía, en parte por consideraciones conceptuales respecto de la esencia de discurso científico.

[1] La investigación en estas recopilaciones no es de acceso público.

[2] En el futuro serán incluidos el francés y el alemán también.

El experimento descrito en la sección 4 ha sido diseñado como una prueba-de-concepto que justifica la opción de estas dos clases de oraciones detectadas como mensajes salientes[3].

2.1. Mensajes salientes en bibliotecas digitales

digitales científicas aplican metainformación orientada por contenido como la definida en el campo de biblioteconomía (Prasad y Madalli, 2008). La metainformación orientada por contenido -títulos de artículos y resúmenes- consiste en fragmentos estructuralmente separables de artículos científicos que resumen los mensajes principales del artículo. Así la mera presencia de la palabra(s) de consulta en estos campos sugiere la relevancia con alta probabilidad (Morato et al., 2003; Wang et al., 2009). Contrariamente a esto la presencia de la palabra(s) de consulta en el texto completo es un indicador más débil de relevancia debido al hecho de que el texto completo naturalmente contiene más palabras, y así un término no recibe tanto peso como cuando este es parte de los metadatos (Schatz, 1997; Spärck Jones, 2006).

El hecho que las respuestas a las consultas que aparecen en el título o el resumen son sumamente relevantes no significa, sin embargo, que no haya mensajes salientes en el cuerpo del artículo: también ha sido demostrado que el contenido relevante de los artículos científicos no puede ser reducido al título y al resumen y así, limitando la consulta al título y el resumen, el usuario se arriesga a omitir documentos relevantes (Voorbij, 1998; Kim et al., 2008).

Por otra parte, en nuestra recopilación de documentos -como en muchas otras recopilaciones de documentos, por ejemplo, en Internet- no siempre tenemos metadatos, y entonces es difícil restringir la consulta al título o el resumen.

Ya que la razón de la relevancia de las respuestas a consultas encontradas en el título y el resumen se debe al hecho de que ellos sintetizan el contenido de artículo, y el texto completo también contiene algunos pasajes de resumen (como la introducción o la conclusión de las secciones y del artículo completo), podemos captar mensajes salientes en estas partes del texto completo. Llamamos a estas oraciones "oraciones de resumen", y son la primera clase de mensaje saliente que proponemos detectar. Los ejemplos siguientes ilustran las oraciones de resumen:

El propósito de este artículo es desarrollar la idea de…

La perspectiva que usaré en este ensayo reside fuertemente en la visión…

[3] Estos dos tipos de oración son detectados para otra aplicación en el proyecto EERQI también (ver Sándor y Vorndran, 2009).

Este trabajo explora…

En su conjunto, el estudio indica…

Las oraciones de resumen expresan funciones argumentativas y anuncian temas a lo largo de todo el artículo, mediante expresiones metadiscursivas, como las oraciones ilustradas más arriba. Ellas expresan de manera explícita el desarrollo discursivo del artículo y de este modo, se espera que reiteren el desarrollo anunciado en el resumen: declaran objetivos, reclamos, conclusiones, presentan la materia, problemas, métodos, etc. Lo que llamamos oraciones de resumen corresponde a varias categorías similares reconocidas en otros sistemas como comunicación de mensajes relevantes (Sección 3).

Todos los roles argumentativos de las oraciones de resumen implican la presencia de mensajes salientes en ellos, pero también proponemos otras clases de oraciones como las portadoras de mensajes salientes. La razón de esto es doble: por un lado, los autores no usan sistemáticamente las oraciones de resumen cuando desarrollan su artículo, y por otro lado parte, la detección automática nunca es exhaustiva. Entonces para cubrir mensajes más relevantes, aplicamos también otra estrategia, que describimos en la siguiente sección.

2.2. Mensajes salientes específicos en el discurso científico
Además de las oraciones de resumen, nuestro instrumento está diseñado para detectar otra clase de oraciones como portadoras de mensajes salientes.

La definición de esta clase de oración es motivada por la consideración de que la *raison d'être* de cada artículo de investigación es contribuir al desarrollo o la solución de un tema de investigación. Sin embargo, la expresión explícita del tema de investigación por metadiscurso, similar a la de sintetizar en las oraciones de resumen, es relativamente poco frecuente (Ruiying y Allison, 2004). Tomamos las expresiones salientes concernientes a la contribución del autor al desarrollo o solución de cuestiones de investigación en oraciones que hablan de refutar, cuestionar o señalar como significantes o nuevas las ideas relacionadas con investigación, los hechos o teorías, indican un vacío en el conocimiento, o señalan cualquier error o contraste relacionado al tema de investigación. Nos referiremos a estas oraciones como oraciones de temas de investigación. Las siguientes oraciones expresan un tema de investigación en nuestro sentido:

El interés por esta averiguación surgió en 1997de una nueva idea sobre pedagogía escolar y pedagogía deportiva.

Esta oración señala la nueva idea del autor en pedagogía escolar y pedagogía deportiva, que será detallada en las oraciones subsecuentes del artículo:

Con la ausencia de trabajos detallados sobre masculinidad y deportes en escuelas primarias de Sudáfrica (para una excepción, ver Bhana, 2002) este trabajo apunta, de algún modo, a abordar los temas en torno a niños desarrollando su relación con el deporte.

Esta oración describe un error respecto de la investigación anterior y propone llevar a cabo algunas de las tareas faltantes.

Sin embargo, el efecto resto, el efecto de primer orden y el efecto de segundo orden, algunos son efectos positivos y otros son negativos, lo cual contrasta con los resultados de investigaciones previas debido a dos razones centrales que...

En esta oración el autor ofrece razones por las cuales algunos conceptos contradicen las investigaciones previas.

Mientras que la categoría y el rol de las oraciones de resumen corresponden a funciones tradicionalmente reconocidas como retóricas o discursivas, el concepto de oraciones de temas de investigación se define de modo menos obvio.

Contrariamente a las oraciones de resumen, como ilustran los ejemplos expuestos más arriba, las oraciones de temas de investigación no cumplen con funciones retóricas, argumentativas o discursivas, consideradas en análisis textuales.

Este concepto es una extensión del concepto de "potential breakthrough" tal como es definido en Lisacek et al. (2005): "Un "potential breakthrough" indica una debilidad en el modelo actual, por ejemplo, observaciones que aparecen como contradictorias, competición entre múltiples hipótesis, resultados inesperados, o nuevas ideas que deben ser integradas en el modelo". El concepto de oraciones sobre temas de investigación modera la significancia atribuida a oraciones que tienen el potencial de comunicar nuevos descubrimientos en el sentido de que un tema de investigación no es necesariamente un descubrimiento potencial.

El reconocimiento de esta categoría de oraciones está motivado por la teoría del progreso científico desarrollada por Kuhn (1962), que se basa en la concepción de la ciencia esencialmente como una actividad de resolución de problemas. Con la categoría de oración de temas de investigación nos proponemos capturar las expresiones de algunos tipos de resúmenes de la actividad de resolución de problemas a nivel oraciones, como expresado en el discurso científico. Estas oraciones también se pueden considerar oraciones sintéticas, pero no en el nivel del desarrollo argumentativo del artículo sino en un nivel de argumentación independiente, que apunta a capturar las cuestiones teóricas discutidas en el artículo.

3. Trabajos relacionados

Las oraciones que expresan mensajes salientes son extraídas para resúmenes automáticos, extracción de información y navegación, que son aplicaciones similares a las nuestras.

En los sistemas de resumen los mensajes salientes son requeridos para representar el desarrollo de los artículos exhaustivamente y de manera coherente. Esto significa que deben seguir el desarrollo del artículo a lo largo de una línea consistente, que es la estructura del discurso (por ejemplo, Teufel & Moens, 2002), la estructura argumentativa (por ejemplo, Teufel & Moens, 1999) o el desarrollo del tema (por ejemplo, Saggion, 2002). Los mensajes salientes extraídos para navegación y extracción de información también están muchas veces basados en la argumentación (por ejemplo, Ruch et al., 2007) o en las funciones del discurso (por ejemplo, Teufel & Moens, 1998).

No nos preocupa la coherencia, pero sí apuntamos a detectar las ideas importantes, y así combinamos dos niveles complementarios: uno discursivo y uno conceptual. Definimos los mensajes salientes tanto en términos de fórmulas discursivas como desde una perspectiva conceptual que es independiente de la retórica, la argumentación o los temas de los artículos. La ventaja del enfoque conceptual puede residir en el hecho de que es independiente de la manera en que el autor construye el artículo, y se concentra en el acto de resolver problemas. No obstante, no hemos diseñado ningún estudio comparativo entre estos dos tipos de enfoque.

En los siguientes párrafos mencionamos brevemente los tipos de oraciones reconocidas como oraciones clave, que son similares o se solapan con nuestros dos tipos de oraciones.

Como mencionamos en el punto 2.1., las oraciones de resumen cumplen con roles discursivos particulares. Estas oraciones corresponden entonces a numerosos tipos de oraciones extraídos en varios sistemas: “propósito”, “solución/método”, “conclusión/reclamo” (Teufel & Moens, 1998) y “objetivo” (Teufel & Moens, 2002).

Las oraciones de temas de investigación, por otro lado, no corresponden a categorías en sistemas que extraen mensajes salientes, algunas de estas categorías, sin embargo captan descripciones de esfuerzos de investigación.

En el esquema de anotación de Teufel and Moens (2002) nuestra categoría de oraciones de tema de investigación sería parte de la categoría “contraste”.

Mizuta et al. (2006) extraen una zona argumentativa llamada “planteamiento del problema” y “propio”, que están también relacionadas a nuestra categoría.

De Waard (2007) identifica numerosos movimientos consecutivos en artículos de biología celular incluyendo "metas de investigación", que son muy parecidas a nuestras oraciones de temas de investigación. En una publicación posterior Waard y Kircz (2008) establecen un marco donde "Contribuciones" se refiere al trabajo del autor.

Kando (1997) también menciona "tema de investigación" como una parte de la descripción del problema en su marco de estructuras de artículos de investigación.

En el contexto de estudios sobre análisis de género son consideradas también las intenciones del autor de presentar el tema de investigación. El establecimiento de los antecedentes de la pregunta de investigación cumple un rol mayor en el modelo para introducciones de artículos CARS de Swale (1990) donde los movimientos para establecer el tema en el artículo son rigurosamente descriptos. Ruiying y Allison (2004) analizan la estructura de artículos de investigación lingüística y descubren "Preguntas de investigación/foco" como una parte de la introducción del artículo, pero muestran que está expresado explícitamente en solo algunos de los artículos analizados.

El concepto de De Liddo y Buckingham Shum (2010) de Cuestionamiento de Inteligencia Colectiva capta la esencia de la instauración de sentido en la unión de ideas a lo largo de una dimensión argumentativa. Han construido la herramienta de anotación Cohere, que permite al usuario establecer y semánticamente etiquetar vínculos entre ideas. Definen instauración de sentido como resolución de problemas, y así las ideas y los vínculos establecidos por los usuarios del instrumento Cohere están muy cerca de lo que se expresa en las oraciones de problemas de investigación.

4. La detección de mensajes salientes

La mayoría de los métodos citados arriba para identificar mensajes clave en textos dependen de métodos de aprendizaje automático y de métodos estadísticos. Tales métodos pueden ser utilizados efectivamente para detectar oraciones de resumen, ya que usualmente estas se expresan en metadiscursos más o menos formulables, cuyos marcadores pueden ser revelados por estos métodos. Es sin embargo más difícil encontrar marcadores de oraciones de temas de investigación en un nivel lingüístico, ya que no despliegan elementos discursivos o estructurales recurrentes, tal como ilustran las oraciones de ejemplo más arriba.

Descubrimos los marcadores de oraciones de temas de investigación en un nivel conceptual, y las expresamos en una estructura de conceptos coincidentes. Debido a que la estructura es aplicable también a las oraciones de resumen, detectamos ambos tipos de oraciones con el mismo método. Hemos usado este método de manera exitosa anteriormente en otras aplicaciones también (Lisacek et al., 2005; Sándor et al., 2006; Sándor, 2009).

Nuestro enfoque codifica reglas manuscritas utilizando un analizador sintáctico dependiente del lenguaje natural, el Xerox Incremental Parser (Aït et al., 2002). Los mensajes clave se detectan mediante un sistema de reglas de co-ocurrencia utilizado además del *parsing* de propósito general que aplica el método de coincidencias de conceptos (Sándor, 2007).

Sintéticamente, la coincidencia de conceptos consiste en fragmentar conceptos determinados tales como "resumen" o "tema de investigación" en conceptos constituyentes, compilando[4] listas de palabras que las expresan y estableciendo reglas que definen qué co-ocurrencias de los conceptos constituyentes expresan el significado del concepto definido. Los conceptos constituyentes son, por ejemplo, OPERACIÓN MENTAL, DOMINIO DE INVESTIGACIÓN, CONTRASTE, AUTOREFERENCIA, ÉNFASIS, ARTÍCULO. La mayoría de las co-ocurrencias deben estar en relaciones de dependencia sintáctica.

De este modo, el concepto de resumen se fragmenta como OPERACIÓN MENTAL en ARTÍCULO AUTOREFERENCIADO y el concepto de tema de investigación como CONTRASTE en OPERACIÓN MENTAL, CONTRASTE en DOMINIO DE INVESTIGACIÓN o ÉNFASIS en un DOMINIO DE INVESTIGACIÓN, etc.

En las siguientes oraciones las palabras en negrita son instancias de los conceptos constituyentes, están en relaciones sintácticas entre sí, y de ese modo expresan los conceptos definidos. Podemos observar en estas oraciones que estas palabras representan los conceptos definidos como secuencias textuales semánticamente coherentes.

RESUMEN
El propósito [FUNCIÓN_DISCURSO] *de este* [AUTOREFERENCIA] *artículo* [ARTÍCULO] *es desarrollar la idea de…*

La perspectiva que yo [AUTOREFERENCIA] *he de usar en este* [AUTOREFERENCIA] *ensayo* [ARTÍCULO] *depende* [OPERACIÓN_MENTAL] *fuertemente de la observación…*

Este [AUTOREFERENCIA] *trabajo* [ARTÍCULO] *explora* [OPERACIÓN_MENTAL]...

En conjunto [FUNCIÓN_DISCURSO]*, el* [AUTOREFERENCIA] *estudio* [ARTÍCULO] *indica* [OPERACIÓN MENTAL]...

[4] Las listas son compiladas manualmente en el presente.

TEMA DE INVESTIGACIÓN
Mi interés [ACTITUD] *de indagar* [DOMINIO DE INVESTIGACIÓN] *emergió en 1997a partir de una nueva* [ÉNFASIS] *idea* [DOMINIO DE INVESTIGACIÓN] *en pedagogía escolar y pedagogía del deporte*.

Con la ausencia [CONTRASTE] *de trabajos detallados* [DOMINIO DE INVESTIGACIÓN] *sobre masculinidad y deportes en escuelas primarias de Sudáfrica (para una excepción, ver Bhana, 2002) este trabajo apunta, de algún modo, a focalizar* [OPERACIÓN MENTAL] *los temas* [DOMINIO DE INVESTIGACIÓN, CONTRASTE] *en torno a la relación que desarrollan los niños con el deporte*. *Sin embargo* [CONTRASTE]*, este efecto resto, el efecto de primer orden y el efecto de segundo orden, algunos son efectos negativos y otros positivos, lo cual contrasta con* [CONTRASTE] *los resultados* [OPERACIÓN_MENTAL] *de investigaciones previas* [DOMINIO DE INVESTIGACIÓN] *debido a dos razones centrales que...*

Cuando las mismas palabras no están en relación sintáctica, no necesariamente expresan los conceptos definidos. Este es el caso en las siguientes oraciones que surgen de artículos recuperados para nuestro experimento descripto en la siguiente sección:

*La exploración [*OPERACIÓN_MENTAL) *de género y poder es otro tema que se observa en los artículos* [ARTÍCULO] *en este* [AUTOREFERENCIA] *tema de Género y Educación*.

El Hombre Acción mantiene [OPERACIÓN MENTAL] *que en contraste* [CONTRASTE] *con los alumnos de escuelas medias inglesas, solo un par de alumnos fumaba regularmente en su año (líneas 1 y 4)*.

Experimento de prueba de concepto
En la sección previa hemos delineado nuestro método para la definición y detección de mensajes salientes en artículos de investigación sobre ciencias sociales. Para mostrar en un experimento empírico que las oraciones detectadas realmente conllevan mensajes salientes, hemos usado estas oraciones como metadatos adicionales en el motor de búsqueda Lucene, de manera similar a como la metadata orientada por el contenido es utilizada en bibliotecas digitales: testeamos si los mensajes salientes pueden ser usados de manera exitosa como material de soporte para la recuperación.

El algoritmo de búsqueda aplicado por el motor de búsqueda Lucene, que fue usado en la selección de los documentos del proyecto de base de contenido EERQI, incluye frecuencias de términos (FT) y frecuencias de documento inverso (FDI) para categorizar los documentos recuperados. Estas medidas están basadas en la frecuencia de ocurrencia de términos de búsqueda en los documentos. La así llamada fórmula FT-FDI evalúa la cantidad de veces que un término de búsqueda ocurre en un documento en relación

al número de veces que un término ocurre en toda la colección del documento. Si, de ese modo, un término de búsqueda aparece frecuentemente en un documento pero solo rara vez o nunca en la mayoría de los otros documentos en la colección, el documento es altamente rankeado (cf. Manning et al., 2009).

El método desarrollado en este estudio está destinado a sostener el ranking de documentos recuperados al asignar un valor más alto a los términos de consulta recuperados en oraciones detectadas por nuestro instrumento. Suponemos que como consecuencia, la precisión respecto de la relevancia de los documentos recuperados se incrementará, ya que aumenta la posibilidad de que el término de consulta represente el contenido de todo el documento. Mientras que un término recuperado con el método FT-FDI puede ser localizado en cualquier parte del documento y puede así ser irrelevante al significado y contenido esencial del artículo, un término recuperado en una oración de resumen o una oración de tema de investigación conlleva gran semejanza con el tema general el artículo.

En lo que sigue proporcionamos indicaciones para comparar los resultados provistos por Lucene y aquellos de nuestro instrumento.

Hemos recuperado 1.200 documentos de investigación con el motor de búsqueda EERQI con la consulta "deportes Y escuela". Hemos evaluado[5] un artículo como relevante si su tema central estaba relacionado tanto con escuela como con deportes.

Hemos evaluado la relevancia de los 15 primeros artículos regresados por Lucene. Nuestra evaluación humana encontró 3 artículos relevantes respecto de la consulta. Ninguno de estos artículos fue seleccionado como relevante por nuestra herramienta.

Nuestro instrumento selecciona un artículo como relevante respecto a la consulta si contiene al menos un mensaje saliente (por ejemplo, una oración de resumen o de "tema de investigación") que contiene ambas palabras de consulta. Hemos evaluado nuestro instrumento en los 330 artículos (de los 1.200 recuperados por Lucene) que contienen al menos una oración con ambas palabras de consulta.

Entre los 330 artículos 85 fueron seleccionados por nuestro programa. La siguiente lista muestra la evaluación humana de estos 85 artículos:

- Número de artículos relevantes de acuerdo a la evaluación humana: 23 (la mayoría de ellos son clasificados con bajo puntaje por Lucene). En cuatro de estos artículos la oración saliente es detectada en una oración erróneamente seleccionada.

[5] La evaluación fue llevada a cabo de manera independiente por los dos autores. El acuerdo interanotador fue casi del 100%.

- Número de artículos no relevantes de acuerdo a la evaluación humana: 62.

Análisis de los errores:
- Error debido a la transformación de formato[6]: 29
- La detección automática del tipo de oraciones es correcta pero la oración no es relevante con respecto a la consulta: 15
- La detección automática del tipo de oraciones es correcta y la oración es relevante con respecto a la consulta, pero el artículo completo no es relevante: 7
- Detección errónea del tipo de oración: 11

De los restantes 245 artículos 35 han sido evaluados como relevantes a la consulta. Chequeamos si en esos artículos las oraciones que contienen ambas palabras de consulta expresan mensajes salientes que no fueron detectados por el instrumento, y encontramos un ejemplo de este tipo.

En total encontramos 58 artículos relevantes mientras evaluábamos nuestro instrumento. Todos fueron clasificados con bajo puntaje (menos de 100) por Lucene.

6. Conclusiones

En este experimento hemos comparado los resultados de la clasificación relevante de Lucene con nuestro método de selección basado en el contenido. Los resultados muestran que los artículos relevantes recuperados por Lucene entre los mejor clasificados y los seleccionados por nuestro instrumento están desconectados, es decir, los dos enfoques son complementarios. Ya que nuestro instrumento, a pesar de su muy estricta regla de selección (la presencia de ambas palabras de consulta en una oración etiquetada como portadora de un mensaje saliente), recupera un número considerable de artículos relevantes que aparecerían al final en la lista de clasificación de Lucene[7], consideramos que nuestro enfoque es prometedor y que la integración de los dos instrumentos puede ser beneficiosa para el usuario. Este experimento nos ayudó a identificar un número de errores sistemáticos tipo que esperamos será posible enmendar, y así mejorar la precisión, antes de la integración. Trabajos posteriores también tendrán que definir el algoritmo de integración.

Referencias

Aït-Mokhtar, S.; Chanod, J.P. & Roux, C. (2002). "Robustness beyond shallowness: incremental dependency parsing". *Natural Language Engineering*, 8(2/3), pp. 121-144.

[6] Los artículos recuperados (pdf, html, doc) son transformados a texto sencillo para el análisis PNL.

[7] Los casos más obvios donde el ránking estadístico de recuperación no-estructurada de colecciones de documentos está destinado a fracasar son artículos relevantes en una recopilación de artículos sobre temas varios en lengua extranjera, que contienen un *abstract* en inglés.

De Liddo A. & Buckingham Shum, S. (2010). "Cohere: A prototype for contested collective intelligence". In *ACM Computer Supported Cooperative Work (CSCW 2010). Workshop: Collective Intelligence In Organizations. Toward a Research Agenda*. February 6-10, 2010, Savannah, Georgia, USA.

de Waard, A. (2007). "A Pragmatic Structure for Research Articles". In *2nd International Conference on the Pragmatic Web*. Oct 22-23, 2007. Tilburg, pp. 83-89.

de Waard, A. & Kircz, J. (2008). "Modeling Scientific Research Articles-Shifting Perspectives and Persistent Issues". In *Proceedings ELPUB2008. Conference on Electronic Publishing*. Toronto, Canada.

de Waard, A.; Buckingham Shum, S.; Park, J.; Samwald, M. & Sándor, Á. (2009). "Hypotheses, Evidence and Relationships: The HypER Model of Representing Scientific Knowledge". *ISWC 2009, the 8th International Semantic Web Conference*. Westfields Conference Center near Washington, DC., USA. 25-29 October 2009.

Kando, N. (1997). "Text-level structure of research papers: Implications for text-based information processing systems". In *Proceedings of the 19th British Computer Society Annual Colloquium of Information Retrieval Research*. Sheffield University, Sheffield, UK, pp. 68-81.

Kim, S.S.; Myaeng, S.H. & Yoo, J. (2005). "A Hybrid Information Retrieval Model Using Metadata and Text". In *Digital Libraries: Implementing Strategies and Sharing Experiences. 8th International Conference on Asian Digital Libraries. LNCS 3815*. Bangkok: ICADL, pp. 232-241.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. Chicago: Univ. of Chicago Pr.

Lisacek, F.; Chichester, C.; Kaplan, A. & Sándor, Á. (2005). "Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases". *First International Symposium on Semantic Mining in Biomedicine*, Cambridge, UK. April 11-13, 2005.

Manning, C.D.; Raghavan, P. & Schütze, H. (2009). *Introduction to Information Retrieval*. Online edition. Cambridge: Cambridge University Press.

Mizuta, Y.; Korhonen, A.; Mullen, T. & Collier, N. (2006). "Zone analysis in biology articles as a basis for information extraction". *International Journal of Medical Informatics*, *75*(6), pp. 468-87.

Morato, J.; Llorens, J.; Genova, G. & Moreiro, J.A. (2003). "Experiments in discourse analysis impact on information classification and retrieval algorithms". *Information Processing and Management*, *39*(6), pp. 825-851.

Prasad, A.R.D. & Madalli, D.P. (2008). "Faceted infrastructure for semantic digital libraries". *Library Review, 57*(3), pp. 225-234.

Ruiying, Y. & Allison, D. (2004). "Research articles in applied linguistics: structures from a functional perspective". *English for Specific Purposes*, *23*(3), pp. 264-279.

Ruch, P.; Boyer, C.; Chichester, C.; Tbahriti, I.; Geissbühler, A.; Fabry, P.; Gobeill, J.; Pillet, V.; Rebholz-Schuhmann, D.; Lovis, C. & Veuthey, A.-L. (2007). "Using argumentation to extract key sentences from biomedical abstracts". *International Journal of Medical Informatics*, *76*(2-3), pp. 195-200.

Saggion, H. (2002). "Shallow-based Robust Summarization". *Automatic Summarization: Solutions and Perspectives*. ATALA.

Sándor, Á.; Kaplan, A. & Rondeau, G. (2006). "Discourse and citation analysis with concept-matching". *International Symposium: Discourse and document (ISDD)*. Caen, France. June 15-16, 2006.

Sándor, Á. (2007). "Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts". *Revue Française de Linguistique Appliquée, 200*(2), pp. 97-109.

Sándor, Á. (2009). "Automatic detection of discourse indicating emerging risk". To appear in: *Critical Approaches to Discourse Analysis across Disciplines*. *Risk as Discourse*. *Discourse as Risk: Interdisciplinary perspectives*.

Sándor, Á. & Vorndran, A. (2009). "Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences". In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, Suntec, Singapore*. 7 August 2009. Singapore (2009), pp. 36-44.

Schatz, B.R. (1997). Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science,* 275, pp. 327-334.

Spärck Jones, K. (2006). Information retrieval and digital libraries: lessons of research. In *Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries*. Kolkata, India.

Swales, J.M. (1990). *Genre analysis: English in academic and research settings*, Cambridge: University Press.

Teufel, S., Moens, M. (1998). "Sentence Extraction and Rhetorical Classification for Flexible Abstracts". In *AAAI Technical Report SS-98-06*, pp. 16-25.

Teufel, S., Moens, M. (1999). "Argumentative classification of extracted sentences as a first step towards flexible abstracting". In Mani, I. & Maybury, M. (eds.), *Advances in automatic text summarization*. Cambridge, Mass: MIT Press.

Teufel, S. & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, *28*(4), pp. 409-445.

Voorbij, H.J. (1998). "Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences". *Journal of Documentation*, *54*(4), pp. 466-476.

Wang, H.-C.; Huang, T.-H.; Guo, J.-L. & Li, S-C. (2009). "Journal article topic detection based on semantic features". In *Proceedings of the 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems: Next-Generation Applied Intelligence*. *Lecture Notes In Artificial Intelligence,* Vol. 5579, pp. 644-652.

Fecha de recepción: 15/12/09
Fecha de aceptación: 10/05/10