Procesamiento del Lenguaje Natural, nº39 (2007), pp. 71-78 recibido 18-05-2007; aceptado 22-06-2007

Similitud entre documentos multilingües de carácter científico-técnico en un entorno Web

Xabier Saralegi Urizar
Elhuyar fundazioa
20170 Usurbil
xabiers@elhuyar.org

Iñaki Alegria Loinaz
IXA taldea. UPV/EHU
649 p.k., 20080 Donostia
acpalloi@si.ehu.es

Resumen: En este artículo se presenta un sistema para la agrupación multilingüe de documentos que tratan temas similares. Para la representación de los documentos se ha empleado el modelo de espacio vectorial, utilizando criterios lingüísticos para la selección de los palabras clave, la formula tf-idf para el cálculo de sus relevancias, y RSS feedback y wrappers para actualizar el repositorio. Respecto al tratamiento multilingüe se ha seguido una estrategia basada en diccionarios bilingües con desambiguación. Debido al carácter científico-técnico de los textos se han empleado diccionarios técnicos combinados con diccionarios de carácter general. Los resultados obtenidos han sido evaluados manualmente.
Palabras clave: CLIR, similitud translingüe, enlazado translingüe, RSS

Abstract: In this paper we present a system to identify documents of similar content. To represent the documents we've used the vector space model using linguistic knowledge to choose keywords and tf-idf to calculate the relevancy. The documents repository is updated by RSS and HTML wrappers. As for the multilingual treatment we have used a strategy based in bilingual dictionaries. Due to the scientific-technical nature of the texts, the translation of the vector has been carried off by technical dictionaries combined with general dictionaries. The obtained results have been evaluated in order to estimate the precision of the system.
Keywords: CLIR, cross-lingual similarity, cross-lingual linking, RSS

1 Introducción

La cantidad de información textual publicada en Internet es cada vez mayor, resultando su grado de organización todavía deficiente y caótico en muchos casos. Situándonos por ejemplo en el contexto de los medios de comunicación, observamos que los servicios que se ofrecen actualmente para una navegación integrada de información proveniente de distintas fuentes resultan escasos, y más todavía cuando se trata de información multilingüe.

Frente a este problema, proponemos una navegación organizada en base a la semejanza semántica entre contenidos, aplicada como experiencia piloto en un entorno multilingüe de sitios web de noticias científicas. Concretamente, hemos centrado nuestro experimento en el sitio web de divulgación científica en euskera Zientzia.net, combinando los siguientes idiomas: euskera, castellano e inglés. Como resultado, Zientzia.net ofrecerá para cada noticia publicada enlaces a otras noticias relacionadas, pudiendo estar publicadas en diferentes sitios web y distintos idiomas. El

ISSN: 1135-5948

objetivo final de este servicio es ofrecer al lector una navegación más completa y organizada. Una navegación similar a la ofrecida por NewsExplorer (Steinberger, Pouliquen y Ignatet, 2005) pero especializada en contenidos científico-técnicos.

Con ese objetivo, se ha diseñado y desarrollado un sistema (Fig.1) que abarca las tareas de recopilación automática de noticias procedentes de distintas fuentes, su representación mediante un modelo algebraico, y el cálculo de las similitudes entre documentos escritos en el mismo o en distintos idiomas.

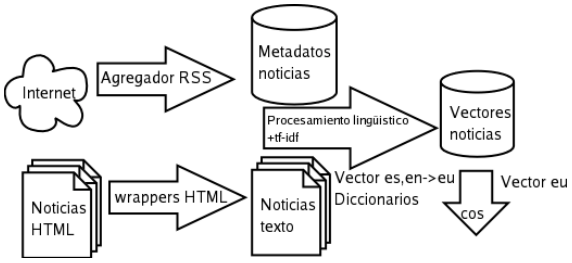

Fig 1. Esquema del flujo de información

La recopilación automática de noticias -tanto locales como remotas- la realiza un robot basado en agregadores RSS y wrappers HTML. La posterior representación de los documentos se hace según el modelo de espacio vectorial. Para la construcción de los vectores se seleccionan las palabras clave siguiendo criterios lingüísticos. Concretamente se escogen nombres comunes, entidades y términos multipalabra, y se calcula su relevancia según la ecuación tf-idf. La traducción de los vectores generados a partir de documentos escritos en distintos idiomas se hace hacia el euskera, y se utilizan tanto diccionarios técnicos como diccionarios de carácter general. Para el tratamiento de las traducciones ambiguas se ha diseñado un sencillo y efectivo método. Finalmente, el grado de similitud se estima mediante el coseno entre los vectores.

Con el propósito de evaluar el sistema, se ha escogido un grupo de documentos al azar de una colección previamente procesada por el mismo, y se ha calculado la precisión analizando manualmente los cuatro primeros semejantes detectados automáticamente (*cutoff* 4).

2 Obtención de documentos

Nuestro sistema se especializa en la recolección e interrelación de documentos pertenecientes al dominio científico-técnico dentro del genero periodístico o divulgativo. Se ha confeccionado una lista de sitios web referentes dentro de la divulgación científica que sirvan de fuentes de información.

Para la creación y continua actualización de la colección de noticias provenientes de las distintas fuentes, se ha implementado un lector basado en sindicación RSS. Mediante la sindicación RSS obtenemos de manera periódica resúmenes de las noticias que se publican en un determinado sitio-web. Los resúmenes suelen contener adicionalmente el título y la URL de cada noticia. Esto implica que, si deseamos acceder al contenido de la noticia, debemos acudir al documento HTML y extraer su contenido.

Sin embargo esta última tarea no es trivial, ya que el texto del contenido suele estar mezclado con otros elementos textuales añadidos -tales como menús de navegación, publicidad, información corporativa...-. [1] Para realizar esta limpieza se proponen generalmente técnicas de carácter automático basadas en aprendizaje supervisado (Lee, Kan y Lai, 2004), pero los resultados no llegan a ser óptimos. Por esa razón, y teniendo además en cuenta que la lista de sitos web a tratar no es muy amplia, hemos decidido implementar los wrappers de manera manual. Concretamente se ha analizado manualmente la estructura HTML de las noticias publicadas en cada sitio web, y se han

[1]Con el objetivo de impulsar trabajos enfocados a la limpieza de documentos web SIGWAC ha programado para Junio del 2007 una tarea (CLEANEVAL) en formato de competición.

implementado parsers empleando el modelo XPath en base a los patrones observados en cada sitio web.

La obtención de noticias publicadas se lleva a cabo, por tanto, en dos pasos: Primero, mediante el agregador RSS obtenemos los metadatos de las noticias publicadas en unos sitios web determinados y, a continuación, extraemos el contenido textual del documento HTML señalado en los metadatos mediante el wrapper HTML correspondiente al sitio web.

Como paso añadido, debido a que algunos sitios web publican noticias en varios idiomas, detectamos el idioma del documento utilizando LangId[2]. Esta identificación es necesaria para poder determinar posteriormente el sentido en el que será traducido el vector generado.

3 Representación de los documentos multilingües

En este trabajo se ha experimentado únicamente con el modelo de espacio vectorial. Pese ha existir modelos más avanzados (Ponte y Croft, 1998), hemos considerado que trabajar con este modelo nos proporcionará un robusto prototipo que podrá ser mejorado en el futuro.

Para la construcción de los vectores, hemos partido de los documentos en formato texto que en el sistema son suministrados según el método explicado en el punto 2.1. Como primer paso se ha realizado una selección del léxico representativo según criterios lingüísticos. Para ello, previamente se ha etiquetado automáticamente cada texto. El etiquetado POS y lematizado se ha llevado a cabo con las herramientas Eustagger para el caso del euskera, y Freeling para el caso del castellano e inglés. A partir del texto lematizado se han podido identificar determinadas unidades léxicas que hemos estimado como más representativas del contenido, descartando el léxico que no aportaría más que ruido para el caso que nos ocupa: modelar el contenido semántico. Así, se han seleccionado nombres comunes, entidades y términos multipalabra. El caso de los adjetivos y verbos no es claro (Chen y Hsi, 2002), y en nuestro caso su ausencia se debe fundamentalmente a que, al estar poco representados en los diccionarios técnicos bilingües, su traducción resultaba limitada. De todas formas, realizamos una serie de experimentos (no concluyentes) que apuntaban a que la no inclusión de verbos y adjetivos implicaba una casi nula mejora en la detección de documentos similares.

Los términos multipalabra en todos los idiomas a tratar (euskera, inglés y castellano) se han identificado a partir de una lista de términos (Euskalterm[3], ZT hiztegia[4]) sobre el texto lematizado. Hemos descartado utilizar técnicas de detección automática de terminología para evitar la generación de ruido y también simplificar la posterior traducción mediante diccionarios. Para el caso de la identificación de entidades hemos utilizado un heurístico sencillo pero a la vez eficiente en cuanto a la precisión u omisión de ruido. Concretamente se han marcado como entidades las series de palabras escritas en mayúscula y que, o son palabras desconocidas, o aparecen en un repertorio de entidades monopalabra previamente elaborado.

Para calcular la relevancia de cada palabra clave se ha experimentado con distintas variantes de tf-idf. Según nuestros experimentos aplicando el logaritmo a tf (1)

$$\text{tf-idf} = \log(\text{tf}) \cdot \text{idf} \qquad (1)$$

hemos obtenido mejores resultados, ya que se ha observado que la similitud entre

[2]Un identificador de idioma basado en palabras y frecuencias de trigramas desarrollado por el grupo IXA de la UPV/EHU.

[3]Diccionario terminológico que contiene al rededor de 100.000 fichas terminológicas en euskera con equivalencias en español, francés, inglés y latín.

[4]Diccionario enciclopédico de ciencia y tecnología que consta aproximadamente de 15.000 entradas en euskera con equivalencias en español, francés, inglés.

documentos con muy pocas claves (con valores tf-idf altos) en común obtenía puntuaciones demasiado altas, generando en muchos casos similitudes imprecisas (falsos positivos).

4 Similitud multilingüe

4.1 Medidas de similitud

Para el cálculo de la similitud entre documentos representados según el modelo espacio vectorial existen distintas métricas. La más extendida es el coseno. Otras métricas también utilizadas son Jackar, Dice... En el modelo OKAPI se toma en consideración el tamaño del documento y la colección proporcionando mejores resultados. (Robertson et al., 1994)

Las métricas mencionadas son aplicables directamente a vectores que representan textos de un mismo idioma pero, para el caso de vectores que corresponden a distintos idiomas, es necesario realizar previamente un proceso de traducción. Para llevar a cabo esa tarea dos son las principales estrategias que se proponen en la literatura: traducción del vector mediante un modelo estadístico entrenado a partir de un corpus bilingüe (Hiemstra, 2001) (basada en corpus), o traducción del vector mediante diccionarios bilingües (Pirkola, 1998) (basada en diccionarios).

En la traducción mediante diccionarios la traducción obtenida puede resultar muy ruidosa ya que la traducción de una palabra resulta ambigua en muchos casos. En tal caso, si aceptamos todas las traducciones posibles y calculamos su tf-idf según la frecuencia de la palabra original, podemos introducir traducciones erróneas que desdibujan la representación del documento original. Esto resulta realmente peligroso ya que las traducciones extrañas, al tener un alto idf, pueden fácilmente distorsionar la representación del vector, y en consecuencia el cálculo de similitudes. Como posible solución se plantean las "consultas estructuradas" (Pirkola, 1998). Originalmente pensadas para tratar "query expansión" en un entorno monolingüe, ponderan según una estrategia prudente las posibles traducciones de cada palabra penalizando el peso tf-idf de todas si el valor df de alguna de ellas es alto.

Un tipo de traducción basada en corpus es la guiada por modelos estadísticos (Hiemstra, 2001). La traducción de los vectores se lleva a cabo mediante el uso de un modelo de traducción -entrenado a partir de un corpus bilingüe en los idiomas a tratar-. De esta forma, se obtiene la traducción del vector más probable según el modelo de traducción y el modelo de lenguaje del idioma objetivo.

De todas formas, tanto la cobertura como la precisión de las técnicas mencionadas no son óptimas. Esto hace que en el proceso de traducción se pierda información -o se introduzca ruido-, de forma que la representación siempre vaya a ser inferior al original. Con el objetivo de reforzar la representación se pueden utilizar técnicas de "query expansion", de manera que se añadan nuevas palabras clave relacionadas semánticamente con el conjunto de términos del vector.

Otras técnicas que no necesitan de traducción por ser independientes del lenguaje, y que resultan apropiadas cuando los pares de idiomas a tratar son muy numerosos, son todas aquellas en las que la selección de palabras clave del documento se realice mediante lexicones o tesauros multilingües tales como WordNet o Eurovoc. En (Steinberger, Pouliquen y Hagman, 2002) por ejemplo, se asignan descriptores independientes del idiomas del tesauro Eurovoc a cada vector mediante un modelo estadístico entrenado mediante aprendizaje supervisado. WordNet, por ejemplo, es utilizado en (Stokes y Carthy, 2001) para representar los documentos mediante cadenas léxicas.

4.2 Diccionarios

Para el caso de vectores en distintos idiomas hemos seguido una traducción mediante diccionarios bilingües.

Debido al carácter científico de los documentos -es decir, un dominio amplio pero acotado- hemos estimado apropiado el uso de recursos lingüísticos específicos (Rogati y Yang, 2004). Hemos combinado diccionarios técnicos (Euskalterm, ZT hiztegia) con diccionarios generales (Elhuyar[5], Morris[6]). No hemos hecho una traducción estadística basada en corpus paralelos por falta de recursos. No disponemos ni de corpus bilingües de carácter científico para todos los pares de lenguas, ni de un alineador a nivel de palabra de precisión notable.

		Dic. técnicos	Dic. generales
tf-idf medio	en	4.483	4.229
	es	5.036	4.871

Tabla 1: tf-idf medio arit. para palabras clave

Mediante el uso de diccionarios técnicos hemos logrado obtener un alto grado de cobertura del léxico especializado. Justamente el léxico que puede ser más representativo del tema del documento. El la tabla 1 se muestra los valores tf-idf de las palabras clave en inglés con traducción en los diccionarios técnicos frente a los tf-idf de las palabras clave con traducciones contenidas en los diccionarios generales. Las palabras clave se han agrupado por lemas y provienen de una colección de documentos reales (tabla 4). Se observa que, según el valor medio aritmético tf-idf, el grado de representatividad es ligeramente mayor en el léxico especializado. Parece, por tanto, que el uso de diccionarios técnicos es una estrategia apropiada. Más aún si también tenemos en cuenta su menor grado de ambigüedad medio en las traducciones de las palabras clave (tabla 2).

		Dic. técnicos	Dic. generales
# traduc. palabra	en->eu	1.72	2.827
	es->eu	1.805	4.243

Tabla 2: Ambigüedad media en traducciones

De todas formas, hemos observado que la cobertura respecto al léxico total podía tener una incidencia negativa en la representación de los textos, ya que algunas palabras generales pueden jugar un papel representativo en los documentos. Adicionalmente, la inclusión exclusiva de palabras técnicas también desfiguraba la dimensión del vector, debido a que las demás palabras del documento no estaban en modo alguno representadas.

Decidimos combinar de manera secuencial los diccionarios técnicos con diccionarios de carácter general. En la tabla 3 se puede observar las coberturas para las palabras clave (agrupadas en lemas) de una colección (tabla 4) obtenidas con las distintas combinaciones de diccionarios.

	diccion. técnicos	diccion. general	diccion. técnico + general
en	55,52%	61,65%	74,48%
es	77,12%	89,02%	91,57%

Tabla 3: Cobertura para las palabras clave

[5]Diccionario castellano/vasco que consta de 88.000 entradas, 144.000 acepciones y 19.000 subentradas.

[6]Diccionario inglés/vasco que consta de 67.000 entradas y 120.000 acepciones.

4.3 Traducciones ambiguas

Como hemos comentado antes, la traducción por medio de diccionarios conlleva una posible ambigüedad que redunda en traducciones incorrectas que desfiguran el vector traducido.

El uso de diccionarios técnicos reduce en cierta medida este problema, ya que el nivel de polisemia y ambigüedad en la traducción es menor (tabla 2). Aun así, el ruido generado sigue siendo un problema como hemos comentado antes. Frente a ello, y teniendo como prioridad la precisión de los resultados del sistema final, planteamos una sencilla estrategia de selección de traducción.

La selección se aplica cada vez que se calcula la similitud (coseno) entre dos vectores de distintos idiomas ($\vec{v}$ y $\vec{w}$). Basándonos en la hipótesis de que la probabilidad de que muchas traducciones ($(i,j)\in D$) incorrectas ocurran en el otro vector es baja, resolvemos la desambiguación eligiendo para cada traducción ambigua aquella que esté presente en el otro vector:

$$\cos(\vec{v}, \mathrm{tr}(\vec{w})) = \frac{\sum_{(i,j)\in D}(v_i w_j)}{|\vec{v}||\vec{w}|} \quad (2)$$

Así, evitamos el ruido que generaría la inclusión de las traducciones incorrectas. Frente al caso de utilizar técnicas de ponderación equitativa de las traducciones, nuestra técnica también se debe mostrar más efectiva en cuanto a la precisión final, ya que el posible ruido afectará solamente a parejas de documentos con baja semejanza mutua. Como hemos dicho anteriormente, suponemos que la probabilidad de que muchas traducciones incorrectas concurran en el otro vector es baja.

En el sistema, el cálculo de similitudes entre documentos se realiza cada vez que el robot recoge una nueva colección de noticias. Se calculan las distancias entre los documentos recientemente recogidos y los documentos de Zientzia.net tanto nuevos como previamente almacenados.

5 Evaluación

En la evaluación hemos querido analizar únicamente los resultados obtenidos en el sistema final. Debido a la dificultad de calcular la cobertura y, siendo la precisión el principal requisito del sistema, hemos evaluado únicamente esta última. Concretamente, hemos calculado la precisión analizando por cada documento de la colección sus cuatro primeros semejantes según el sistema (*cutoff)*.

La colección base de noticias se ha obtenido y procesado mediante los procesos explicados en los anteriores apartados. Consta de todos los artículos publicados hasta la fecha en Zientzia.net, y de artículos publicados en los otros sitios web durante un periodo de un mes (tabla 4). Aunque la idea del sistema es mostrar los semejantes a partir de la navegación de los documentos en euskera, la evaluación se ha hecho en sentido inverso debido a la superioridad numérica del los artículos de Zientzia.net. De la otra forma, la probabilidad de encontrar semejantes se reduciría notablemente.

	# docs	# palabras	# palab/doc
es	108	71.366	661
eu	3146	1.249.255	397
en	550	284.317	517

Tabla 4: Colección de noticias procesada

Para la evaluación formamos 3 grupos (uno para cada idioma) de 10 documentos escogidos aleatoriamente de la colección base. Tras procesar toda la colección mediante el sistema analizamos por cada documento los 4 primeros más semejantes (de entre los de Zientzia.net) según el sistema. El método de análisis propuesto consistió en valorar el grado de semejanza del contenido en base a una escala de relevancia dividida en cuatro categorías y

basada en el esquema utilizado en (Braschler y Schäuble, 1998).

(a) Comparten el tema principal: Los documentos hablan sobre el mismo tema.
(b) Tema principal relacionado o comparten temas: Los documentos tratan de temas muy relacionados o mantienen en común temas no principales.
(c) Comparten área: Los documentos pertenecen ha una determinada área sin llegar a ser general.
(d) Parecido remoto: Las relaciones entre los documentos son remotas o inexistentes.

De esta forma, se pretende valorar como más positivas las relaciones de gran parecido. Sabemos que esta escala es discutible, ya que de cara al usuario puede ser más útil una referencia que complemente el artículo en curso que un artículo sobre el mismo tema. Además, asignar a cada documento una categoría de esta escala resulta en muchos casos una tarea de difícil precisión.

El análisis fue llevado a cabo por un profesional en el campo de la divulgación científica, y se hizo para dos prototipos distintos:

1) distribuyendo equitativamente el peso entre las traducciones .
2) aplicando la desambiguación propuesta anteriormente.

Quisimos comprobar si el método diseñado para resolver casos de traducción ambiguos mejoraba la precisión del sistema.

En las tablas 5, 6 y 7 se muestran las distintas precisiones (*cutoff* 4) acumulando las categorías según la escala de relevancia comentada. Se observa que los resultados varían según el idioma, siendo evidente la perdida de información tras la traducción. Este hecho influye en mayor medida a las relaciones inglés-euskera debido a la menor cobertura de los diccionarios bilingües inglés-euskera.

	(a)	(a+b)	(a+b+c)
Desam.	10%	37.5%	82.5%
No desam.	10%	30%	70%

Tabla 5: Cutoff 4 en-es

	(a)	(a+b)	(a+b+c)
Desam.	30%	37.5%	60%
No desam.	25%	32.5%	60%

Tabla 6: Cutoff 4 es-eu

(a)	(a+b)	(a+b+c)
17.5%	57.5%	85%

Tabla 7: Cutoff 4 eu-eu

Se ha observado que, quizás debido al pequeño tamaño de la colección, documentos con pocas palabras clave compartidos han sido aceptados como similares.

En cualquier caso, el método diseñado para resolver traducciones ambiguas mejora la precisión en todas las pruebas.

Relacionado con el tamaño y la variedad del contenido se ha observado que la precisión del sistema es menor frente a documentos de algún tema muy especial, resultando la comparación léxica insuficiente. Esto puede ser debido al reducido número de documentos, pero no ha podido ser evaluado al no tener constancia de la cobertura.

6 Conclusiones y trabajo futuro

Se ha desarrollado un sistema para la agrupación de documentos multilingües de contenido similar con el objetivo de integrarlo en un un sistema CLIR. Esto ha dado lugar a un sistema de navegación de noticias científico-técnicas multilingües, implantado en el sitio Zientzia.net.

Los resultados obtenidos nos deben llevar a realizar una evaluación más exhaustiva. Independientemente de esto, se ha comprobado que la traducción mediante diccionarios resulta positiva, más concretamente con el uso los diccionarios técnicos. El uso del método de desambiguación propuesto también ha sido exitoso, pero una nueva evaluación es necesaria para cuantificar mejor la mejora conseguida.

Sería muy interesante evaluar la perdida de precisión usando solamente resúmenes RSS, ya que consiguiendo un buen resultado estas técnicas podrían ser usadas para gran cantidad de fuentes sin necesidad de utilizar wrappers.

También se pretende realizar nuevos experimentos con modelos de lenguaje, preguntas estructuradas y distintas medidas de similitud. Adicionalmente queremos mejorar la traducción de entidades mediante detección de cognados, y la traducción general mediante generación de tesauros multilingües a partir de corpus comparables. De cara a algunas de estas tareas pensamos basar el motor de búsqueda en la herramienta Lemur toolkit (Ogilvie y Calla, 2001).

Agradecimientos

Este trabajo está subvencionado por el Departamento de Industria del Gobierno Vasco (proyectos Dokusare SA-2005/00272, Dokusare SA-2006/00167).

Bibliografía

Braschler, M., y P. Schäuble. 1998. Multilingual Information Retrieval Based on Document Alignment Techniques , ECDL 1998, pp. 183-197.

Chen, Y., y H. Hsi. 2002. NLP and IR approaches to monolingual and multilingual link detection. *The 19th Int'l Conf. Computational Linguistics*. Taipei, Taiwan.

Hiemstra, D. Using language models for information retrieval. *Ph.D. Thesis University of Twente*. Enschede.

Lee, C. H., M. Kan, y S. Lai. 2004. Stylistic and lexical co-training for web block classification. *WIDM 2004.* 136-143

Ogilvie, P., y J. Callan. 2001. Experiments using the Lemur toolkit. Proceedings of the Tenth Text Retrieval *Conference (TREC-10).*

Pirkola, A. 1998. The Effects of Query Structure and Dictionary setups in DictionaryBased Cross-language Information Retrieval. *Proce. of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55-63.

Ponte, J., y W. Croft. 1998. A Language Modeling Approach to Information Retrieval. In: Croft et al. (ed.): *Proceedings of the 21st Annual Interna- tional ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275{281. ACM, New York.

Robertson, S. E., S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford. 1994. Okapi at TREC-3. *NIST Text Retrieval Conference*.

Rogati, M., y Y. Yang. 2004. Resource Selection for Domain Specific Cross-Lingual IR. *SIGIR 2004*.

Steinberger, R., B. Pouliquen, y J. Hagman. 2002. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Third International Conference on Intelligent Text*.

Steinberger, R., B. Pouliquen, y C. Ignat. 2005. NewsExplorer: multilingual news analysis with cross-lingual linking. *Information Technology Interfaces.*

Stokes, N., y J. Carthy. 2001. Combining Semantic and Syntactic Document Classifiers to Improve First Story Detection. *SIGIR 2001*: 424-425.