Monografía sobre Acceso a Información Multilingüe: presentación

Anselmo Peñas, Julio Gonzalo

Dpto. Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Juan del Rosal, 16
28040 Madrid
{anselmo,julio}@lsi.uned.es

Con el gran crecimiento y progresivo carácter multilingüe de la información digitalizada (especialmente en Internet), el área de investigación en *Acceso a la Información Multilingüe* ha recibido una atención especial por parte de la comunidad científica. En la actualidad se desarrollan periódicamente varias competiciones de ámbito internacional en las que se evalúan comparativamente sistemas multilingües de búsqueda de información, extracción de resúmenes, búsqueda de respuestas o traducción automática, como TREC (Text REtrieval Conference), CLEF (Cross-Language Evaluation Forum), NTCIR (NII-NACSIS Test Collection for Information Retrieval systems), DUC (Document Understanding Conference), etc.

La idea de este monográfico surge como consecuencia del Taller sobre Acceso Multilingüe a la Información y Procesamiento de Lenguaje Natural que tuvo lugar en Sevilla en el marco de IBERAMIA 2002[1]. El objetivo es dar una perspectiva de la investigación en torno a esta área multidisciplinar en la que convergen el Procesamiento de Lenguaje Natural, la Recuperación de Información o la investigación en Bibliotecas Digitales, con temas como:

- Recuperación y filtrado de información.
- Extracción de resúmenes mono y multi-documento.
- Representación, almacenamiento y recuperación en Bibliotecas Digitales.
- Traducción automática.
- Sistemas de búsqueda de respuestas.
- Extracción de información y minería de textos.
- Sistemas de diálogo.
- Evaluación de sistemas de acceso a la información.

En este número especial se recogen dos tipos de artículos: por un lado, revisiones en alguno de los temas del monográfico y, por otro, trabajos originales de investigación en el área. En la primera modalidad, se incluyen cuatro revisiones del estado del arte:

- El artículo de López-Ostenero et al. resume las aportaciones más significativas al campo de la *Búsqueda de Información Multilingüe,* entendido como la búsqueda de documentos en idiomas distintos al de la consulta. Los autores describen los aspectos monolingües de la tarea (cuáles son las características de cada idioma que pueden dar lugar a estrategias de indexación y recuperación específicas) y los aspectos de traducción (ya sea mediante la proyección de la consulta en el idioma de los documentos, mediante la traducción de los documentos, o mezclando ambas en estrategias alternativas). Finalmente, se hace un énfasis especial en esta revisión en los aspectos interactivos de la búsqueda multilingüe.

- En contraposición a los sistemas de búsqueda de documentos (como los del artículo anterior), los sistemas de *Búsqueda de Respuestas* o de pregunta-respuesta reciben una pregunta en lenguaje natural (por ejemplo, "*¿Cuándo murió Bob Marley?")* y dan como resultado una respuesta exacta (por ejemplo, *"Robert Nesta Marley murió el 11 de mayo de 1981*") tras

[1] *http://nlp.uned.es/ia-mlia/iberamia2002*

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. Vol. VIII, No. 22 (2004), pp. 7-9
ISSN: 1137-3601 © AEPIA (http://www.aepia.org/revista).

realizar una búsqueda en una colección textual determinada. El artículo de Vicedo describe el estado actual y los retos futuros para este tipo de sistemas, entre los cuales se encuentra la capacidad de responder preguntas con independencia del idioma de las fuentes documentales.

- Dentro del campo del Procesamiento del Lenguaje Natural, se llama *Extracción de Información* (EI) al proceso automático en el que se obtienen elementos de información de estructura predeterminada a partir de texto libre, normalmente en un dominio específico. En las evaluaciones MUC (Message Understanding Conference) se utilizó, por ejemplo, el dominio de las finanzas: los sistemas debían identificar organizaciones, compañías, localizaciones, etc., y descubrir en los textos sucesos como, por ejemplo, qué compañías se fusionaron y cuándo. Estos sistemas pueden proporcionar, por tanto, elementos de acceso y síntesis de información que van más allá de la simple localización de documentos relevantes. Pueden permitir, por ejemplo, preguntas estructuradas como las que se hace a una base de datos, y se han utilizado también como fuente de información para la extracción automática de resúmenes. El artículo de Turmo resume los temas de interés dentro de este campo, prestando atención a los retos que supone la construcción de sistemas de EI multilingües.

- El artículo de Alonso et al. está dedicado a los sistemas de *Resumen Automático*, y trata aspectos como la naturaleza de los resúmenes, los enfoques computacionales más utilizados y los retos inmediatos que se plantean en este área, haciendo especial énfasis en el resumen multilingüe.

En la segunda modalidad (trabajos originales de investigación) se recogen siete artículos. Los dos primeros tratan sobre distintos aspectos de la recuperación de información multilingüe:

- El artículo de Abdelali et al. describe un sistema interactivo de recuperación de información multilingüe en idiomas como el árabe, el ruso y el inglés, que permite refinar la consulta e identificar información relevante sin necesidad de conocer el idioma de los documentos.

- El artículo de Martínez et al. describe un sistema de recuperación de información multilingüe que ha participado en el foro internacional de evaluación CLEF (Cross-Language Evaluation Forum), y que hace aportaciones originales y eficientes a problemas como el de la fusión de listas de relevancia provenientes de colecciones en distintos idiomas.

Los dos artículos siguientes (Alonso et al. y Figuerola et al.) tratan aspectos específicos de la Recuperación de Información en castellano, y experimentan con técnicas como la normalización de términos atendiendo a criterios morfológicos que, a priori, podrían ser más útiles para el castellano que para otros idiomas menos flexivos como el inglés.

Los tres artículos que cierran esta colección tratan, cada uno, distintos aspectos del acceso a información multilingüe:

- El artículo de Gómez et al. describe un sistema de categorización de textos para el filtrado de información multilingüe en Internet. Por *categorización de textos* se entiende la asignación automática de descriptores dentro de una clasificación preestablecida. El filtrado de textos es una variante de la recuperación de información en la que no existe un flujo de consultas, sino un perfil de usuario más o menos estable y un flujo de documentos que son clasificados y filtrados de acuerdo con ese perfil. En el artículo se citan aplicaciones como el filtrado de contenidos pornográficos.

- En el artículo de López-Cozar se describe un sistema con capacidades multilingües de diálogo, en el que se accede de forma natural a la información de una base de datos de medios de transporte con horarios, precios, etc.

- Finalmente, Casañ et al. tratan de algunos aspectos específicos de codificación de palabras en *Traducción Automática*, una tecnología que, aunque tiene tantos años como la Inteligencia Artificial, y su uso se ha popularizado en Internet, tiene aún muchos retos por resolver, y constituye una pieza clave dentro del Acceso a Información Multilingüe.

Esta colección de artículos no es – no podría serlo – un recorrido exhaustivo de esta área de investigación, pero sí creemos que ofrece una panorámica de temas de investigación que están alcanzando ya un cierto grado de madurez, y cuyos resultados influirán en la próxima generación de asistentes para la búsqueda, organización y síntesis de información.

Agradecimientos

Queremos agradecer la colaboración de la Asociación Española para la Inteligencia Artificial (AEPIA), de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), del Instituto Cervantes, de la Red Iberoamericana de Tecnologías del Software para la década del 2000 (RITOS2) y de la European Network of Excellence in Human Language Technologies (ELSNET), por la difusión de la convocatoria y por el apoyo prestado en todo momento.

También queremos agradecer a los autores su participación, a los revisores su trabajo y a Ana García Serrano su disposición en la elaboración de la monografía.

Finalmente, queremos agradecer a Felisa Verdejo, Nieves Brisaboa, Isabel Bermejo y Alexander Gelbukh su iniciativa y apoyo en la organización del taller sobre Acceso Multilingüe a la Información y Procesamiento de Lenguaje Natural.

Editores de la monografía

Anselmo Peñas Padilla y Julio Gonzalo Arroyo
Dpto. Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
anselmo@lsi.uned.es, julio@lsi.uned.es

Revisores

Miguel A. Alonso	U. de A Coruña
Manuel de Buenaga	U. Europea de Madrid
Antonio Ferrández	U. de Alicante
Carlos Figuerola	U. de Salamanca
Pablo de la Fuente	U. de Valladolid
Ana M. García	U. Politécnica de Madrid
Alexander Gelbukh	Inst. Polit. Nac. de México
José M. Gómez	U. Europea de Madrid
Fernando López	U.N.E.D
Manuel J. Maña	U. de Vigo
Fernando Martínez	U. de Jaén
Horacio Rodríguez	U. Politécnica de Cataluña
L. Alfonso Ureña	U. de Jaén
José Luis Vicedo	U. de Alicante