Moreiro González, J. A. (16-18 de septiembre, 1992). Perspectiva documental del procesamiento del lenguaje natural. [Comunicación en congreso]. VIII Congreso de la Sociedad Española del Lenguaje Natural, Granada

PERSPECTIVA DOCUMENTAL DEL PROCESAMIENTO DEL LENGUAJE NATURAL

José A. Moreiro González

Universidad Carlos III
Madrid

Planteamiento teórico en torno al paralelismo existente entre el procesamiento de la lengua para la traducción y el que se da en la transformación documental. Partiendo de una vinculación común a las tendencias científicas de los años sesenta, se explica el origen de las aplicaciones semánticas al manejo de la información. Se destacan las coincidencias en el tratamiento automatizado entre la traducción y la representación documental. Para considerar desde este convencimiento la coincidencia en los avances y la identidad de problemas. Lo que conlleva al reconocimiento de las aplicaciones lingüísticas a la intermediación documental automatizada y un breve análisis de la situación presente en España.

1.- Razones históricas de una vinculación

Como consecuencia de las necesidades informáticas surgidas durante la II Guerra Mundial apareció en los años siguientes un movimiento documental que en poco tiempo logró cambiar la perspectiva teórica e incluso la propia denominación de nuestro campo de actividad. Pasaba a un primer plano la recuperación urgente de información mediante dispositivos más baratos y seguros. En su logro se volvió imprescindible alcanzar un uso eficiente de la documentación almacenada y una mayor capacidad de transferencia comunicativa.

El movimiento se gestó en empresas de titularidad privada como la Documentation Incorporated, una de las firmas comerciales dedicadas a la consultoría y a la recuperación de la información. De ella partieron las iniciativas para las primeras indizaciones coordinadas: los unitérminos, antecedente de los actuales sistemas de recuperación. En esta nueva tendencia los primeros protagonistas se relacionaron tanto con la recuperación por conceptos como con la automatización. Justifica desde entonces el maridaje indisoluble que surgió en su aplicación.

Taube, Perry, Casey, Berry, Garfield, Luhn y Kent se interesaron desde los tempranos sesenta por los procesadores de palabras, por la indización automatizada desde los títulos (índices KWIC y KWOC), por la elaboración de thesauri mediante ordenador. En especial Luhn se propuso transformar el conocimiento humano mediante la indización y el resumen en la persecución de reducciones homomórficas. Para ello aplicó la automatización al manejo de los textos documentales. Desde entonces la semántica documental y la intervención de los ordenadores en el proceso informativo se establecieron como las características más definitorias para entender qué sea la «Information Science». Sin su concurso el análisis de contenido resulta imposible.

— Si quisiésemos caracterizar la «Information Science» respecto a sus precedentes documentales deberíamos destacar como aportaciones más originales las siguientes:

- Planteamiento de los problemas de recuperación de la información desde indizaciones coordinadas sobre términos extraídos del lenguaje natural.

- Aplicación de los ordenadores al control y gestión de la información.

- Como consecuencia de ésta, aparición de las Bases de Datos, en cuanto alternativa a los medios de difusión impresos.

- Sometimiento de los hechos documentales a medición estadística.

- Sujección de la actividad y la producción informativas a la ley de la oferta y de la demanda.

De esta forma vemos como lo más peculiar de nuestra actividad tiene que ver desde tres décadas con el lenguaje natural y las actuaciones para su automatización.

Traducción automatizada y transformación documental. Relaciones de inclusión.

La representación documental de base semántica muestra un claro paralelismo con la traducción automatizada. Demostrable incluso desde una coincidencia temporal en sus orígenes. E implicándose hasta el punto de que las aplicaciones lingüísticas a la Documentación no pueden concebirse lejos de las exigencias teóricas y prácticas de la traducción. ¿Porqué?:

- Por la obligación de aproximarse ambas al texto íntegro si lo que se desea es un tratamiento automático aceptable.

- Por las necesidades de comunicación que fomentan su presencia y hasta justifican su existencia.

- Por el impulso que para ambos hechos supuso la llegada de los primeros ordenadores.

La analogía es tan clara que podemos llamar traducción a la versión de un texto desde una lengua natural a otra, igual que al proceso de transformación textual desde el documento completo a su resumen e indización. Para que ambas acciones puedan ser diligenciadas, se hace necesario profundizar en la comprensión de los discursos. Lo que viene a ser corroborado por una nueva coincidencia. Desde los años sesenta asistimos a un gran avance en el conocimiento y la conceptualización del texto imprescindible para un procesamiento ulterior.

Además, porque las dos principales aplicaciones del procesamiento automático de las lenguas naturales tienen que ver con la Documentación (puesto que la mayor parte de las informaciones parten de los textos):

- Las de índole terminológica, punto de partida para los bancos de datos desde los que se nutren los elementos de los thesauri.

- Las relativas al estudio de las estructuras y estrategias textuales. Ya que no en vano la descripción sustancial se expresa íntegramente en desarrollos lingüísticos naturales.

Centrándonos en las tareas de intermediación informativa, y desde la perspectiva del conocimiento, la transformación textual plantea los mismos problemas si lo que pretendemos es lograr el paralelismo semántico entre texto y descriptores como entre texto y resumen. En los dos casos obtenemos la representación del contenido mediante términos o mediante un nuevo texto. Ambas describen el texto, sólo les diferencian los fines documentales de recuperación o de explicación. Desde esta perspectiva representan al texto, lo que obliga a redefinir el concepto de recuperación y el de resumen fuera de los caminos divergentes que hasta ahora han solido explicarles.

La representación del significado en lo que se ha venido denominando «sistema de recuperación documental» cumplía las dos misiones de indizar los documentos, tras captar el significado de los documentos controlados, y la de ayudar a la comprensión de las preguntas que plantea el usuario mediante la recuperación propiamente dicha. Lo que implica una representación de la carga sustantiva de información de los textos. Este diálogo documental también se puede cumplir mediante el resumen, al que sin duda se trasladan las macroestructuras textuales con mayor coherencia y exactitud.

La situación parece abordable desde una panorámica lingüística entroncada con los «lenguajes científico-técnicos» de las áreas cognitivas humanas. Si el marco en el que la C.E.E. dió nombre a las Industrias de la Lengua implicaba someter a procesos industriales la «Generación de Documentos multilingües» (Proyectos EUROTRA y SPRIT), no cabe duda que la Documentación tiene que ver con el tratamiento informativo de las lenguas naturales. Pues su filosofía de actuación consiste fundamentalmente en oponerse a la dispersión informativa mediante la reducción de los datos. Con la condición de que en ésta queden reflejadas los trazos semánticos de los documentos de origen. Se trata por tanto de una traducción con fidelidad a las fuentes. O sea una buena traducción.

Coincidencia en los logros y en los problemas

El manejo automático de los textos en actuaciones documentales empezó con posturas en que primaba la practicidad. Ir a lo más barato y sencillo: fueron dos décadas donde los criterios de identificación terminológica y de frecuencia estadística marcaron el camino a seguir. La reiteración de términos claves era suficiente para elegir un concepto como representativo, y cuando varios de ellos coincidían en una oración se suponía que era relevante para el contenido global, o de una parte del texto. Esta situación parece un calco de lo que sucedía con los primeros ordenadores aplicados a traducir: procedían palabra por palabra y haciendo exclusivamente consideraciones sintácticas. Recordemos el programa SYSTRAN. Junto al criterio estadístico-terminológico se utilizaban (utilizaron) después pistas discriminadoras sobre la información del texto, por atención a las fuentes más expresivas: título, subtítulos y encabezamientos. E incluso captando las locuciones indicadoras, cuando el propio texto explicita la importancia de una oración, advirtiendo el interés de lo que a una expresión se sigue.

Los textos son entidades complejas y por tanto de procesamiento dificultoso. El camino estadístico, mejorado en los años ochenta, parece estar alcanzando el límite de su actividad. Los posibles avances debían plantearse desde la perspectiva sintáctica, y desde lo contextual y pragmático.

El criterio establecido desde los términos reiterativos se fundamenta en la estabilidad denominativa de los fenómenos científicos. Por lo que resulta aplicable sólo en campos donde el uso de los conceptos es prolongado en el tiempo y unívoco y desambiguado en las referencias. Así, los documentos de contenido técnico o científico aplicado tienen clara ventaja sobre los propios del sector humanístico, literario y social. Sobre estos componentes nominales de los textos técnicos se han ajustado los algoritmos interpretativos. Sin embargo, según se avanza, nuevas dificultades salen al paso. En este caso, para llegar al convencimiento de que para lograr el conocimiento matriz no son suficientes los componentes nominales. Lo conseguido es notorio y las perspectivas amplísimas. Si bien los sistemas aplicables a la gestión automatizada del análisis documental se encuentra aún en un estado precompetitivo.

Se pueden establecer con claridad los vínculos que desde la automatización de los procesos de la lengua alcanzan los propios de la Documentación. La proximidad de ambos planteamientos alcanza extraordinaria similitud, en cuanto procesos interdisciplinares de idéntico origen y resultado similar. Que debemos comprender dentro de un desarrollo paralelo que ha supuesto implicaciones recíprocas:

- Ayudas a la traducción: apoyo tecnológico a la extracción conceptual y a los intentos de síntesis textuales. Reconocimiento del texto mediante criterios terminológicos (léxicos legibles por ordenador), sintácticos y semánticos, desde la profundización en sus estructuras. Se deben considerar los apoyos ofimáticos a la recuperación y manejo de los textos.

- Control bibliográfico: Favorecido por los interfaces entre el lenguaje natural y el lenguaje informático, nos permite tratar los textos tanto por su contenido como por los rasgos de identificación. La automatización de los atributos físicos y de la indización (KWIC, KWOC, en texto libre) y resumen (extractos, tendencia a su articulación) favorece el establecimiento de bases referenciales y la obtención de los documentos en cuanto difusión de los originales en un servicio documental final.

- Relaciones terminológicas: Criterios de fiabilidad, suministro y mantenimiento de términos desde los cuales establecer la relevancia y la capacidad de aclaración respecto a un texto dado. Es un factor de coincidencia para determinar los componentes nominales. Mediante algoritmos de interpretación y desde el conocimiento de mundo se desambiguan los posibles sentidos de un término. Una vez racionalizadas y unificadas las denominaciones se forman bancos de datos y diccionarios terminológicos. Desde ellos se establecen los elementos de los thesauri y los factores normativos de los términos.

- Almacenamiento y recuperación de información sobre un tema: La alta capacidad de almacenamiento de los ingenios ópticos ha permitido que los textos íntegros adopten formas legibles por las máquinas. Desde ahí se procede a su análisis, mientras que en los sistemas precedentes se partía de los títulos o de los resúmenes de los textos. La recuperación se ve facilitada en cuanto se puede ofertar el texto completo en que los conceptos se desarrollan. Lo permiten los sistemas de reconocimiento textual.

- La intervención de la Inteligencia Artificial: Ha permitido desarrollar procesos algorítmicos mediante códigos aritméticos y métodos para adaptarlos. Ello ha supuesto el comienzo de las aplicaciones pragmáticas que buscan comprender el texto antes de traducirlo. Se resalta así su relación con los modelos cognitivos: con los marcos de conocimiento en que se encuadran los documentos, con los conocimientos generales y al caso del autor y de los receptores (activación de marcos cognitivos) y con los procesos inferenciales.

4.-Breve apunte sobre la realidad española.

La valoración del momento presente en investigaciones documentales acerca del tratamiento semántico automatizado no alcanza un nivel de aplicación ni siquiera discreto en la profesión o en la Universidad españolas. La realidad nos muestra escasas aproximaciones teóricas más bien inconexas, pero que desde luego marcan el principio de un deseable desarrollo. Las causas podemos encontrarlas en la juventud disciplinar que ha obligado a estudios tendentes en preferencia al muestreo académico. Tampoco ayuda la diversidad de áreas cognitivas de proveniencia tanto de los profesionales como del profesorado. Si añadimos las dificultades generales para formar equipos interdisciplinares, comprenderemos que estamos más cerca de la meditación y la síntesis, que de auténticas actuaciones empíricas.

Los modelos de aproximación al análisis semántico parten de la descripción de los actos de recepción y comprensión humanos, para desde ahí plantear siquiera lejanamente los mismos procesos automatizadamente. Las escasas contribuciones del profesorado a estos estudios provienen casi siempre de aproximaciones modélicas de explicación procesual, más que de aplicaciones puntuales . Que atienden con preferencia a los procesos mentales antes que a su reproducción en las máquinas. No faltan los casos de aplicaciones más concretas, por más que de preferencia descriptiva. Curiosamente son de clara orientación documental investigaciones sobre elaboración de índices no muy alejadas de los conocidos modelos KWIC y KWOC, pero surgidas en la Lingüística aplicada. Lo que refuerza nuestra tesis del paralelismo de los dos campos, afirmando a la vez la mayor solidez del originado en la Lingüística por las razones antes aducidas.

5.-Requisitos para mejorar el tratamiento automático de los documentos escritos.

1.- Profundizar en los planteamientos teóricos que aborden el estudio del texto satisfactoriamente: la necesidad de compatibilizar las estructuras y estrategias que se dan en el texto se hace evidente si queremos alcanzar un modelo de análisis automatizable.

2.- Lograr que los originales se presenten dentro de formatos tipológicos convenidos, mediante los cuales se diferencien expresivamente las partes del texto. De esta manera se facilita su lectura y comprensión. A la vez que se vencen las frecuentes desviaciones de los textos respecto de las superestructuras típicas.

3.- Investigar en la posibilidad de establecer con claridad las líneas de coherencia entre los subtextos. Desde ellos se marcará la sucesividad de la información y se elegirá el nivel de profundización tanto en la indización como en el resumen. Este planteamiento se sigue de la imposibilidad de los modelos propuestos actualmente para mostrar la sucesividad de las oraciones. Para superarlo se precisa la capacidad de seguir los conectivos lógicos y retóricos que presentan los textos.

4.- Continuar ahondando en los criterios mediante los cuales el ordenador pueda identificar y seleccionar la carga sustantiva de información, tanto global como parcial. Supone el perfeccionamiento de los programas de Inteligencia Artificial. No sólo se trata de captar la expresión literal, sino de interpretarla desde el conocimiento de mundo y desde la situación comunicativa (comprender las implicaciones profundas de los textos).

5.- Lograr el establecimiento de unos resúmenes marco. Punto de partida para una congregación textual coherente.

6.- Fomentar la colaboración interdisciplinar de todos los científicos interesados en solucionar los múltiples interrogantes aún planteados: lingüistas, documentalistas, traductores, informáticos, lógicos y especialistas en el área de conocimiento manejada.

7.- Conjuntar la contribución y avanzar en la aplicación de los aspectos:

- Morfológicos,
- Lexicográficos,
- Sintácticos de modelos algorítmicos,
- De representación de los conocimientos en redes semánticas,
- De investigación en los procesos inferenciales desde:
  - el contexto,
  - lo no dicho,
  - las anáforas,
  - la referencia personal (competencia y situación comunicativa).

8.- Posibilitar el manejo de los textos hacia las tendencias que marca la actividad documental presente: el logro de los documentos globales partiendo de orígenes diversos, la integración de los textos con imágenes y voz, la oferta de los textos íntegros tras recuperaciones más exactas, en fin cuanto frene la fragmentación cognitiva fruto de la especialización y el crecimiento de la ciencia.