EL PROCESAMIENTO DEL LENGUAJE NATURAL APLICADO AL ANÁLISIS DEL CONTENIDO DE LOS DOCUMENTOS

Isidoro GIL LEIVA y José Vicente RODRÍGUEZ MUÑOZ
Departamento de Información y Documentación
de la Universidad de Murcia

Resumen Se presentan las técnicas del procesamiento del lenguaje natural a través de su definición, desarrollo histórico, niveles de análisis, así como algunos de los problemas que surgen en el análisis automático de textos utilizando estas técnicas. Posteriormente, se muestran las distintas aplicaciones del procesamiento del lenguaje, exponiendo más extensamente, las dirigidas al análisis del contenido documental, es decir, la elaboración automática de resúmenes y la indización automática de documentos. Finalmente, se emiten las conclusiones.

Palabras clave: Procesamiento del lenguaje natural / análisis automático del contenido documental / software de análisis del contenido documental / resumen automático / indización automática

1. INTRODUCCIÓN

Las tecnologías de la información están alcanzando cotas cada vez más altas en la vertiente de análisis automático de los documentos. El análisis del contenido documental (resumen e indización) ya se puede perpetrar de modo automático gracias al procesamiento del lenguaje natural (PLN), si bien es cierto, que no se han alcanzado soluciones finales. El PLN sigue siendo una disciplina desconocida para profesionales, e incluso investigadores, del área de la Biblioteconomía y la Documentación a pesar de que interviene directamente en campos propios de este dominio como la Recuperación o Análisis de la Información. En España son pocos los investigadores del área que, directa o indirectamente, han indagado en este tema. Si bien, en (García, 1996) encontramos un reciente e interesante estudio experimental, tanto por el tema abordado como por la metodología utilizada, donde se ha originado un sistema

para el análisis automático de documentación periodística con la finalidad última, de efectuar una recuperación más intuitiva y cercana a los usuarios de este tipo de documentación. Otras aportaciones, en este caso teóricas son las de (Moreiro, 1992 y 1993) donde reflexiona sobre las relaciones entre el PLN y la transformación documental y entre la Lingüística y la Documentación respectivamete.

Partiendo de conocimientos informáticos y lingüísticos (lingüistas computacionales), se están desarrollando sistemas para la confección de resúmenes y la indización automática. Este tipo de investigaciones se lleva practicando desde hace décadas, y se comienza a recoger los frutos de años de inspección, por lo que se debe permanecer atentos a su evolución de cara a la posible implantación en un futuro próximo.

La complejidad del procesamiento del lenguaje natural es algo que queda patente, por lo que cabe reseñar que es un campo no exento de dificultad para los lingüistas que deben adquirir la instrumentación de los informáticos, y para los informáticos, ya que deben hacer suyos conocimientos lingüísticos. De la misma forma, estas técnicas se tornan doblemente intrincadas para los documentalistas porque hay que tomar los conceptos, las herramientas y maneras de proceder tanto de lingüistas como de informáticos. A pesar de lo espinoso que resulta para los documentalistas, consideramos necesario comenzar a constituir, o incoporarse, a grupos de trabajo con objeto de potenciar el incremento de este tipo de herramientas de análisis automático del contenido documental para la lengua española.

Por otro lado, este tipo de actividades no debe sobresaltar a los que consideran que las tareas de resumir e indizar son operaciones pura y exclusivamente ejecutables por el intelecto humano, puesto que se está demostrando que, a pesar de no existir todavía logros definitivos, y esto hay que resaltarlo, se pueden alcanzar a corto o mediano plazo. Y aunque nos mostramos optimistas en cuanto a las posibilidades de estas nuevas herramientas, hay que señalar que somos conscientes que hasta que los resultados no sean los deseados no deben incorporase en las labores documentales. Igualmente, estas tecnologías no se han de tomar como un nuevo «competidor» para los documentalistas, puesto que el tiempo ahorrado en el análisis del contenido se podrá dedicar a otros cometidos, como por ejemplo, la difusión de la información, que es sin duda la razón de ser de todo el quehacer documental.

2. EL PROCESAMIENTO DEL LENGUAJE NATURAL

El procesamiento del lenguaje natural consiste en el estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos. Un sencillo ejemplo de PLN es un corrector ortográfico de un procesador de textos que todos hemos empleado alguna vez, aunque hay otras herramientas más

complicadas como veremos más adelante. Menciona Verdejo (1994, p.5) que el lenguaje natural se distingue de los lenguajes artificiales por su riqueza (en vocabulario y construcciones), flexibilidad (reglas con múltiples excepciones), ambigüedad (pudiendo darse diversos significados de una palabra o una frase según el contexto), indeterminación (permitiendo referencias y elipsis) y posibles interpretaciones del sentido literal según la situación en que se produce. Lo que son ventajas para la comunicación humana se convierten en problemas a la hora de un tratamiento computacional, ya que implican conocimiento y procesos de razonamiento que aún no sabemos ni cómo caracterizarlos ni cómo formalizarlos.

Continuando con esta misma autora (p. 16) hay que señalar que el PLN surge en la década de los cincuenta, y su historia se entrelaza con las investigaciones que sobre el lenguaje se llevan a cabo en otras disciplinas, principalmente lingüística formal, psicología cognitiva, lógica y la informática, pero sobre todo, la inteligencia artificial, dando lugar a una disciplina denominada *lingüística computacional*. La lingüística computacional es la intersección de la lingüística y la informática con el fin de procesar o generar las lenguas. Se distinguen distintas etapas en el despliegue teórico y práctico del PLN:

1. De los años cuarenta a mitad de los sesenta. Coinciendo con la aparición de los ordenadores se extendió la idea de que el procesamiento del lenguaje se lograría en muy poco tiempo, pero paulatinamente fueron surgiendo las incógnitas que conllevan los intentos en este sentido. Por estos motivos se dejaron de financiar proyectos encaminados principalmente, a la traducción automática ruso-inglés. Asimismo, se iniciaron experimentos para comprender el lenguaje en ámbitos muy específicos.
2. Desde principios de los setenta hasta comienzos de los ochenta. El tratamiento de la sintaxis, en términos de formalismos y algoritmos, experimentó un incremento considerable, aunque realmente la teoría lingüística y la práctica computacional pocas veces convergieron.
3. De la década de los ochenta hasta la actualidad. Se ha llevado a cabo la unión entre las teorías lingüísticas y los mecanismos de parsing, a la vez que se ampliaron los estudios del PLN a nivel de la pragmática y del discurso. Por otro lado, se comenzó a hablar también de las llamadas «Industrias de la lengua», que propició que diversas compañías colocaran en el mercado productos en donde integran la informática y la lingüística. Estamos hablando de correctores ortográficos automáticos incorporados a los programas de procesamiento de textos, la traducción automática, sistemas de análisis y recuperación de información, o programas de reconocimiento y síntesis del habla, sistemas en los cuales se escudriña en el presente para perfeccionarlos.

2.1. Niveles en el PLN

Las técnicas del procesamiento del lenguaje se promueven a través de diferentes análisis ocupando cada uno de ellos distintos niveles, directamente relacionados con los inconvenientes de éstos. El primero que apuntaremos es el morfológico, en el cual, ya se han alcanzado cotas de efectividad importantes, rozando en algunos analizadores el 95% de efectividad. El analizador sintáctico toma como fuente de análisis el producto de la etapa anterior para, principalmente, desambigüar aquellas palabras que no se han desambigüado en el morfológico. Por último, veremos los niveles semánticos y pragmáticos que aún comportan más dificultad que los anteriores.

Análisis morfológico. Efectúa un examen de cada vocablo para obtener toda la información gramatical de la misma como prefijos, raíces, sufijos y desinencias, así como la clase gramatical o clases a la(s) que pertenece. A continuación se ofrece la solución del analizador morfológico que Rank Xerox [1] tiene accesible a través de la red de redes Internet para esta frase: «Esta nota para mañana».

<Esta>	Pronombre+Demostrativo+Femenino+Singular
	Determinante+Demostrativo+Femenino+Singular
<nota>	Verbo(notar)+Presente indicativo+3ª persona+Singular
	Verbo(notar)+Imperativo+2ª persona+Singular
	Sustantivo+Femenino+Singular
<para>	Preposición
	Verbo(parir)+Presente subjuntivo+1ª persona+Singular
	Verbo(parir)+Presente subjuntivo+3ª persona+Singular
	Verbo(parir)+Imperativo+3ª persona+Singular
	Verbo(parar)+Presente indicativo+3ª persona+Singular
	Verbo(parar)+Imperativo+2ª persona+ Singular.
<mañana>	Sustantivo+Femenino+Singular
	Adverbio.

Como se puede comprobar, la conclusión del análisis automático es una lista con las posibles categorías gramaticales junto a los valores gramaticales de cada una de las palabras que aparecen en la frase. La incertidumbre del lenguaje queda visible, por lo que se tendrán que emplear otros mecanismos para lograr una desambigüación lo más completa posible.

Análisis sintáctico. Tiene por objeto comprobar si los vocablos del texto están bien coordinados y unidos, es decir, averiguar si las oraciones son gra-

[1] Esta compañía tiene disponible a través de la red Internet un analizador morfológico que se puede probar en la siguiente dirección de World Wide Web URL:http://www.xerox.fr/grenoble/mltt/Tools.html (mayo, 1996).

maticalmente correctas. Además, en este estadio se pretende también la resolución de problemas no solucionados en el análisis anterior como la reiterada ambigüedad gramatical de las palabras, destinando para ello varios mecanismos como las posiciones de las voces en las oraciones o los contrastes gramaticales. Para esto se dispone de una base de datos donde se recogen todas las combinaciones gramaticalmente aceptables entre los vocablos, aunque otra opción es al contrario, recoger las posibilidades prohibidas. Una construcción que no se permite es ésta: «Las clasificación es ...», ya que una de las reglas es que el artículo y el sustantivo deben coincidir en género y número. El criterio sería así:

ARTI (SING)+ SUST (PLUR) = PROHIBIDO

Por otro lado, habrá reglas que señalan que un verbo no puede ir precedido de un artículo.

ARTI+VERB=PROHIBIDO
ARTI+HOMOGRAFÍA [SUST, VERB]=SUST

Análisis semántico-pragmático. La semántica estudia la significación de las palabras, por tanto, un análisis semántico tratará de averiguar el significado de las oraciones de un texto, y por extensión el del mismo texto. En el PLN el conocimiento semántico se representa (Smeaton, 1995) especificando conceptos simples y sencillos que se combinan en estructuras más intrincadas como redes semánticas, dependencias conceptuales o estructuras. Como el resto de los análisis del lenguaje, esta fase toma el resultado de otros análisis precedentes que puede ser ambigüo (ambigüedades sintácticas en este caso) para tratar de eliminarlas. Uno de los rompecabezas en el procesamiento de la fase semántica es la gran cantidad de conocimiento que se necesita acerca de los vocablos y conceptos en el universo del discurso, en orden a formalizar tales interpretaciones, de ahí su complejidad.

2.2. Problemas en el análisis de los textos

Sin profundizar en el PLN se puede deducir que los programas que acometan un análisis automático de textos se van a encontrar con una serie de dificultades más o menos engorrosas, que sin embargo, no ofrecen ningún obstáculo para su comprensión en la comunicación humana, aunque sí para un analizador automático. Veamos algunos ejemplos:

1. La aparición en el texto de un guión puede tener varios significados, uno que está poniendo en relación dos palabras o también lo podemos hallar

dividiendo un vocablo al final de una línea, indicando que el resto de la voz está en el siguiente renglón.

2. La anáfora es un tipo de señalamiento que desempeñan ciertas palabras como por ejemplo «ésta», «lo», «le»,»aquí», «allí» etc., para asumir el significado de una parte del discurso ya emitida:

> «Hay diferencias entre la indización manual y la indización automática, la primera la acomete una persona y la segunda la ejecuta un programa»

«primera» = indización manual «segunda» = indización automática

Para (Grishman, 1991, p.148) una de las características de las lenguas naturales, que constituye a la vez la razón para que su análisis sea tan difícil, es que gran parte de lo que se comunica está implícito en el discurso. Este fenómeno ha sido muy atendido tanto por lingüistas teóricos como por los que se dedican a la lingüística computacional. La resolución de la anáfora es un proceso de sustitución: reemplazamos el sintagma anafórico por una descripción más completa que permita la interpretación del sintagma nominal por medio de etapas sucesivas de procesamiento semántico. Se distinguen (Rico, 1994a) dos tipos de expresiones anafóricas: las endofóricas, es decir, las que aparecen dentro de los textos y por tanto, su antecedente se localiza en el mismo texto, y las exofóricas, cuyo antecedente no es posible detectarlo dentro del texto. Rico (1994b) examina un procedimiento de resolución de este fenómeno.

3. La elipsis es otro de los inconvenientes que surgen en el análisis automático y se define como una figura de construcción que consiste en omitir en la oración una o más palabras, necesarias para la recta construcción gramatical, pero no para que resulte claro el sentido. Las expresiones elípticas (Palomar, 1994) son fenómenos muy habituales en la comunicación humana y en contextos hombre-máquina como elementos de cohesión, coherencia y economía discursivas. Continua manifestando que una diferencia fundamental entre ciertas construcciones elípticas y los fenómenos anafóricos (por ej., en casos de anáfora pronominal), es que en las primeras no aparece ninguna entidad lingüística que deba ser vinculada con un antecedente mediante una relación de correferencia, sino que, simplemente, se da un «vacío» en la estructura sintáctica de la frase en cuestión. Desde la perspectiva del tratamiento computacional se han distinguido tres tipos de elipsis: intrasentenciales (producidas en el interior de una oración u oraciones coordinadas), fragmentos (ocurren fundamentalmente en situaciones de diálogo), y elipsis semánticas (aquellas que no se manifiestan como estructuras incompletas sino como proposiciones semánticamente incompletas). Entre las clases de vocablos que se pueden elidir encontramos verbos, sustantivos, adjetivos y adverbios.

«El mismo documentalista realiza (i) y (ii) corrige la indización de los artículos»
(i) «la indización de los artículos»
(ii) «El mismo documentalista»

En definitiva, si se pretende comprender íntegramente un texto con la finalidad bien de resumirlo o indizarlo por medios automáticos, sin duda alguna se tendrán que destinar mecanismos para la resolución de los dilemas que representa la anáfora y la elipsis, porque de lo contrario, cuando un sistema automático que busca los conceptos esenciales de un texto localice términos como: «ésta», «aquélla», o «primera» no sabrá que se está haciendo referencia a un vocablo mencionado anteriormente. Lo mismo que en el caso de una palabra elidida, la cual habrá que restituirla al pasaje.

4. Nombres propios y siglas. En un texto aparecen continuamente nombres propios (a veces, en distintos idiomas) por lo que se está ante un trance añadido que hay que resolver. Además, se detectan siglas, que para complicar esta empresa se pueden escribir de modos variados. En los textos nos podemos encontrar oraciones como:

«El sistema CLARIT, diseñado por D.A Evans, es una aproximación a la indización automática vía procesamiento del lenguaje natural»
«Desde la S.E.P.L.N. se está potenciando la lingüística computacional», o también, «La SEPLN es la Sociedad Española para el Procesamiento del Lenguaje Natural»

El reconocimiento de los nombres propios y su posterior análisis y clasificación (Miranda, 1994) es una tarea bastante compleja, debido fundamentalmente, a su elevado número, a la gran variedad de formas que adoptan y a la ambigüedad que algunos de ellos presentan. El simple reconocimiento ya resulta difícil, porque el único distintivo con el que se cuenta es el empleo de la mayúscula, pista que no es válida para las palabras que van detrás de punto.

2.3. Desarrollo de herramientas para el PLN

Coincidimos plenamente con (Verdejo, 1994, p. 19), por razones obvias, cuando reconoce que lo ideal sería que se dispusiera de una biblioteca básica de herramientas como corpus, lexicones o analizadores para poder iniciar investigaciones que precisan estos utensilios. No obstante, hemos observado que la red de redes Internet es un medio útil para conocer la existencia de instrumentos creados en este campo, el contacto con investigadores de esta disciplina, o incluso, el aprovechamiento de analizadores en la misma red.

3. APLICACIÓN DEL PROCESAMIENTO DEL LENGUAJE NATURAL

3.1. Aplicaciones Generales

Las aplicaciones del PLN en este caso son variadas. A continuación desplegamos algunas de ellas:

1. Traducción automática. Los primeros ensayos surgieron a mitad de los cincuenta pero una década más tarde bastantes proyectos financiados hasta ese momento se abandonaron, principalmente en Norteamérica porque no se conseguía la calidad deseada. A partir de los ochenta se potenció de nuevo este tipo de experimentación a través de grandes programas en Europa, Japón y Estados Unidos, no obstante, no se ha alcanzado aún la calidad de la traducción humana.
2. Comunicación hombre-máquina. Las experiencias han ido encaminadas principalmente, a la consecución de la comunicación humano-robots, lo que ha posibilitado que, en pocos años y de forma generalizada, por medio de un micrófono se diga a un ordenador: «Abre el procesador de textos X».
3. Enseñanza asistida por ordenador.
4. Procesadores de textos. Llevan incorporados correctores ortográficos, diccionarios de sinónimos más o menos completos, así como los verificadores automáticos de las palabras que se van tecleando en el procesador de textos.

3.2. Aplicaciones Específicas en Documentación

En cuanto a la instrumentación donde interviene el PLN en el análisis y recuperación de información destaca:

I. La interrogación de Bases de datos en lenguaje natural. De este modo se simplifican las consultas, puesto que proporciona la interrogación en una base de datos de una manera totalmente natural como por ejemplo: «Qué documentos hay sobre la difusión de la información en las bibliotecas universitarias», en lugar de hacerlo con un lenguaje intercalado con operadores booleanos.

II. La generación automática de tesauros, que posibilita la identificación de relaciones sintácticas y semánticas entre palabras y frases.

III. La categorización y difusión de la información. Un programa manejando técnicas del PLN y teniendo previamente establecidos una serie de perfiles con las necesidades de información de un conjunto de usuarios, analiza y filtra la información que ingresa en el sistema de información, generalmente a través del correo electrónico, haciéndola llegar a cada usuario de forma totalmente automática. Este instrumento se está integrando actualmente en algunos

sistemas de información de organizaciones norteamericas. Un programa con estas características es Intell X[x] de la compañía norteamericana DataTimes.

IV. Elaboración automática de resúmenes.
V. Indización automática de documentos.

Estas dos últimas aplicaciones del procesamiento del lenguaje natural al análisis del contenido documental, se describen más ampliamente en el siguiente apartado.

Elaboración automática de resúmenes

De los años sesenta hasta la actualidad se ha venido indagando, con más o menos ímpetu, en la mejora de mecanismos para la confección automática de resúmenes, tomando como base el procesamiento del lenguaje natural con objeto de reducir el coste del procedimiento documental. La sucesión que siguen estos sistemas para alcanzar sus propósitos, es de forma generalizada la siguiente: se lee el texto con formato electrónico y se aplican distintos análisis lingüísticos para su reducción. Para practicar estos pasos se dispone de una serie de analizadores con sus reglas gramaticales, así como de lexicones más o menos amplios dependiendo del sistema.

Estamos ante una fase de proliferación de este tipo de herramientas, por lo que se actúa, en estos momentos, desde grupos de investigación institucionales y compañías comerciales como por ejemplo Oracle. El programa ConTex de esta compañía es uno de estos sistemas. No está disponible como un producto autónomo en la fecha, sino que se debe adquirir con una licencia de base de datos Oracle7. ConText dispone del procesamiento del lenguaje natural para la identificación de temas y contenido de los textos, y no de consideraciones estadísticas. Se ha concebido inicialmente para resumir información relacionada con la Empresa para aligerar la gran cantidad de información que consumen los responsables de la toma de decisiones en las organizaciones. Para una profundización mayor se puede consultar (Summarizing, 1995), (Pinto, 1992).

Un sistema de producción automática de resúmenes puede ser de gran utilidad para los centros de documentación en las organizaciones (Gil, 1996) puesto que los directivos de las empresas si hay algo de lo que carecen es de tiempo, por lo que hay que hallar la manera para que la información que les llegue sea en el instante y con la extensión adecuadas, para que puedan digerirla y convertirla en acciones. El tipo de información a resumir sería datos sobre competidores y sus productos, legislación, informes sobre sectores o mercados, noticias aparecidas en la prensa diaria bien en papel o electrónicamente sobre economía o sobre su sector, así como informes a los que se tenga acce-

so, capítulos de monografías de reciente publicación o de Tesis doctorales relacionadas con el área de actividad.

Indización automática

Su origen hay que buscarlo a finales de los años cincuenta con la aparición de los ordenadores y por la denominada explosión de la información. La indización, ya sea en su vertiente manual o automática, ha despertado desde siempre más interés -desde el punto de vista de la investigación- que la automatización del proceso de resumir. Los primeros pasos en indización automática (Gil, 1995) se emprendieron al amparo de medios estadísticos, aunque al poco tiempo se comprendió que se podían intercalar métodos lingüísticos. Finalmente, fue la lingüística la que se convirtió, hasta la fecha, en fundamento insustituible para el logro de una indización automática de documentos.

Las principales tentativas en este área proceden de Estados Unidos donde hay una larga tradición de más de cuatro décadas, por lo que subrayaremos sólo algunos intentos como (Cohen, 1995), (Silvester, 1994) o (Evans, 1991). Pero no obstante, también se han proyectado prototipos experimentales para lenguas distintas del inglés como el español (Simón, 1990) , ruso (Pozhariskii, 1991), francés (Orban, 1993), árabe (Hmeide, 1995), chino (Wan, 1995), coreano (Seo, 1993) o portugués (Robledo, 1991).

La mayoría de los sistemas de indización mencionados se ha diseñado al amparo de grupos de investigación universitarios o desde grandes instituciones, que por la gran cantidad de documentación que manejan se ven en la necesidad de diseñar una herramienta automática para agilizar la indización de documentos. Por otro lado, también hay sistemas de indización automática en el mercado como SPIRIT (de la sociedad SIEMENS) que puede indizar textos en francés, inglés y alemán; Golem, que es un sistema de almacenamiento y recuperación documental, pero dispone de un módulo llamado PASSAT que prepara la indización automática para el alemán, holandés e inglés; ALETH (de ERLI) para textos en francés; DARWIN (de CORA) realiza sus análisis en francés, inglés y alemán; INDEXICON (de la corporación norteamericana ICONOVEX) que está disponible para el inglés; y programas de indización asistida por ordenador como SINTEX o ALEXDOC.

Todos los programas de indización automática no tienen las mismas características, no consuman idénticos procedimientos ni toman las mismas herramientas para llevar a cabo sus análisis. Así, los hay que disponen casi exclusivamente, de una lista de autoridades y cuando el sistema localiza en el texto una o varias palabras pertenecientes a esta lista, las considera como términos de indización. De igual modo, hay otros más complicados que se sirven de los diferentes planos del análisis lingüístico, unidos o no, a principios estadísti-

cos. En general, se puede advertir que todos los sistemas de indización automática aplican alguno o varios de estos instrumentos:

Bases de datos con :
- — raíces de las palabras
- — desinencias
- — expresiones idiomáticas
- — vocablos vacíos
- — lista de autoridades

Analizadores:
- — morfológicos
- — sintácticos
- — semánticos

Técnicas estadísticas
Normalización de términos
Autoreenvíos

El uso de este tipo de programas no tiene una trascendencia tan directa de cara a los usuarios como la creación automática de resúmenes, sino que su conveniencia reside, siempre y cuando el análisis realizado tenga una buena calidad, en el ahorro de tiempo y esfuerzo para el documentalista, y así, al «descargarse» de las faenas del análisis, puede dedicar más atención a otras.

A lo largo de este artículo se ha presentado una introducción con los aspectos más relevantes que intervienen en el procesamiento del lenguaje, con la finalidad de servir de base para futuros trabajos en donde se tratarán más extensamente las distintas experiencias (descripción de sistemas, metodologías empleadas, tiempos de ejecución, resultados obtenidos, etc.) que se están llevando a cabo en el análisis automático del contenido documental tomando como fundamento el procesamiento del lenguaje natural.

NOTA

Para quienes deseen profundizar en algunos de los aspectos tratados anteriormente proporcionamos a continuación una pequeña gúia referencial. Así, para una visión general y completa de lo que es la Lingüística Computacional ver (Grishman, 1991). Además, a través de[2] se puede obtener información sobre Universidades que ofrecen programas de lingüística computacional y procesamiento del lenguaje, los mayores laboratorios no académicos de ensayo sobre PLN, las publicaciones más importantes en este campo, listas de correo

[2] En el grupo de noticias de Internet llamado comp.ai.nat-lang se discute sobre problemas relativos al lenguaje natural y los ordenadores, como análisis y generación del lenguaje natural o traducción automática.

electrónico y cómo subscribirse, grupos de noticias, organizaciones y asociaciones profesionales, Congresos y Conferencias, así como una extensa bibliografía parcelada en PLN.

Si se buscan estudios sobre la aplicación de herramientas básicas en el PLN la publicación dirigida por José Vidal Beneyto titulada *Las industrias de la lengua*[3] pero principalmente, los *Boletines de la Sociedad Española del Procesamiento del Lenguaje Natural* (SEPLN).

Con respecto al ámbito de la obtención automática de resúmenes, se puede consultar un monográfico de la revista *Information Processing & Management*, 31 (5), 1995. Y finalmente, para la indización automática, no hay ninguna monografía sobre este asunto, sino aportaciones diseminadas por gran cantidad de fuentes sobre Documentación, aunque destacan la mencionada *Information Processing & Management* y el *Journal of the American Society for Information Science* (JASIS).

4. CONCLUSIONES

La utilidad del procesamiento del lenguaje natural para el análisis y recuperación de información se está haciendo cada vez más presente como ha quedado demostrado. No obstante, también hemos podido comprobar que practicar un análisis automático de textos es una misión no exenta de problemas, que hay que ir salvando si queremos conseguir un resumen o indización de modo automático que nos merezca una total confianza. Igualmente, el avance en el aumento de sistemas de elaboración de resúmenes y de indización de modo automático ha ido en los últimos años, e irá, paralelo a los avances en el procesamiento del lenguaje, y al mismo tiempo, a la disponibilidad de potentes ordenadores en los que procesar grandes cantidades de texto en el menor tiempo posible.

Por otro lado, ante los escasos equipos de investigación en el área de Biblioteconomía y Documentación en este campo de actuación, sería recomendable que se iniciara una incorporación o incluso, se crearan grupos de trabajo interdisciplinares (lingüistas-informáticos-documentalistas) ya que el diseño y auge de herramientas automáticas para el análisis del contenido documental así lo requiere. Pero además, se le daría sentido una vez más, a la interdisciplinariedad de la Documentación.

[3] Las Industrias de la lengua. Madrid: Fundación Germán Sánchez Ruipérez, 1991.

BIBLIOGRAFÍA

BARRERAS MUNDET, J. Resolución de elipsis y técnicas de parsing en una interficie de lenguaje natural. Boletín de la SEPLN, febrero 1993, nº 13, p. 247-258

COHEN, J. D. Highlights: Language-and domain- independent automatic indexing terms for abstracting. Journal of the American Society for Information Science. 1995, vol. 46, nº 3, p. 162-174

EVANS, D. A. (et al.) Automatic indexing of abstracts via natural processing using a simple thesaurus. Med Decis Making. 1991, vol. 11, nº 4, p. 108-115

GARCÍA GUTIERREZ, A. Procedimientos de análisis documental automático. Estudio de caso. Sevilla: Junta de Andalucía, Consejería de Cultura, 1996

GIL LEIVA, I; RODRÍGUEZ MUÑOZ, J. V. Tecnologías aplicadas en la gestión de la información en las organizaciones. III Jornadas Nacionales de Información y Documentación Empresarial, Mayo 1996

—, De la indización humana a la indización automática. Actas II Encuentro de ISKO-España, Noviembre 1995 (en prensa)

GRISHMAN, R. Introducción a la lingüística computacional. Madrid: Visor distribuciones, 1991

HMEIDE, I.I. Design and implementation of automatic word and phrase indexing for information retrieval with arabic documents. PH.D., Illinois Institute of Technology, 1995

MIRANDA GARCÍA, A.; RODRÍGUEZ-SÁNCHEZ TORRES, L. Analizador morfosintáctico de nombres propios y siglas. Actas del X Congreso SEPLN, julio 1994

MOREIRO GONZÁLEZ, J. A. Perspectiva documental del procesamiento del lenguaje natural. Boletín de la SEPLN, febrero 1993, nº 13, p. 41-45

—, Implicaciones documentales en el procesamiento del lenguaje natural. Ciencias de la Información, marzo 1993, vol. 24, nº 1, p. 48-54

ORBAN DE XIVRY, D. Le traitement de l'information textuelle: utilisation du système SPIRIT (Système Probabiliste d'Indexation et de Recherche d'Informations Textuelles). Cahiers de la Documentation. 1993, vol. 47, p. 15-23

PALOMAR, M. (et al.). Formalización de la coordinación mediante la Gramática de Huecos. Actas del X Congreso SEPLN, julio 1994

PINTO MOLINA, M. El resumen documental. Principios y métodos. Madrid: Fundación Germán Sánchez Ruiperez, 1992, págs. 323-332

POZHARISKII, I. F. Automatic document indexing (case study of an automated geological information system). Automatic Documentation and Mathematical Linguistics. 1991, vol. 25, nº 10, p. 30-37

RICO PÉREZ, C.(a) Estudio de la incidencia de la diferentes fuentes de información en

el establecimiento de relaciones anafóricas. Boletín de la SEPLN, marzo 1994, nº 14, p. 63-75

—, (b) Resolución de la anáfora discursiva mediante una estrategia de inspiración vectorial. Actas del X Congreso SEPLN, julio 1994

ROBLEDO, J. Indexaçao automática de textos: uma abordagem otimizada e simple. Ciência da Informaçao. 1991, vol. 20, nº 2, p. 130-136

SEO, E-G. An experiment in automatic indexing with korean text: a comparison of syntactic-statistical and manual methods. PH.D., University of Illinois at Urbana-Champaign, 1993

SILVESTER, J. P. (et al.) Machine-aide indexing at NASA. Information Processing & Management, 1994, vol. 30, nº 5, p. 631-645

SIMÓN GRANDA, J.; LEMA GARZÓN, de J. Primeras experiencias sobre el análisis de textos en castellano aplicado a la indexación automática de información. Jornadas Españolas de Documentación Automatizada, mayo 1990, p. 1255-1269

SMEATON, A. F. Natural language processing used in information retrieval taks: an overview of achievements to date. Encyclopedia of Library and Information Science. 1995, vol. 55, supplement 18, p. 220-238

SUMMARIZING text. Information Processing & Management, september 1995, vol. 31, nº 5, p. 625-784

VERDEJO MAILLO, M. F. Procesamiento del lenguaje natural: fundamentos y aplicaciones. UNED, Curso de Verano, julio 1994

WAN, T-L. Experiments with automatic indexing and a relational thesaurus in a chinese information retrieval system. PH.D., Illinois Institute of Technology, 1995