Universidad de Matanzas
Camilo Cienfuegos
Facultad de Informática

Título: Propuesta para la modificación de la disciplina Inteligencia Artificial con la incorporación de temas de Procesamiento del Lenguaje Natural

Autores: Katia Vila Rodríguez.

Antonio C. Fernández Orquín.

Resumen

Es muy común escuchar que nos encontramos viviendo la época de la “Sociedad de la Información”, en la que disponemos de una gran cantidad de información accesible de forma casi gratuita a través de la red conocida como Internet. De este modo, por el contrario de lo que ocurría en el pasado, el problema no está en la disponibilidad de dicha información, sino en la localización de la información que realmente le interesa al usuario. La inmensa mayoría de la información que se encuentra en Internet se encuentra de forma textual, por lo que para su tratamiento es necesario aplicar técnicas de Procesamiento del Lenguaje Natural; el cual es considerado una de las ramas más importantes de la Inteligencia Artificial y en la actualidad se están desarrollando trabajos de investigación de gran relevancia a nivel mundial en esa área.

En el trabajo se exponen elementos del Procesamiento del Lenguaje Natural y sus aplicaciones en la Web- ejemplo de ello es la Búsqueda Inteligente en la Web- por la necesidad de mejorar la calidad de los servicios que en ella utilizamos diariamente y que presentan una organización que hasta hoy aún se cataloga de caótica. Líneas tan importantes como la Recuperación de Información y los Sistemas de Búsqueda de Respuesta se destacan por sus aportes y por las perspectivas de desarrollo que presentan. Se incluye además la propuesta de su introducción en la carrera de Ingeniería Informática, tanto en pregrado como en postgrado, por la importancia que reviste.

Introducción

En las dos últimas décadas hemos asistido a un crecimiento exponencial de la cantidad de información digital disponible y a la explosión de las comunicaciones entre ordenadores como vía principal de transmisión de información entre usuarios. La investigación en sistemas de información textual que faciliten la localización, el acceso y el tratamiento de grandes cantidades de información, ha sido impulsada precisamente por la actual disponibilidad de información- principalmente de carácter textual- unido al creciente número de usuarios finales (no especialistas en tratamiento de datos ni en computadores) que disponen de acceso directo a dicha información a través de ordenadores personales.

Generalmente, cuando un usuario emplea un ordenador para buscar una información determinada, lo que realmente está intentando es encontrar respuesta a sus necesidades de información.

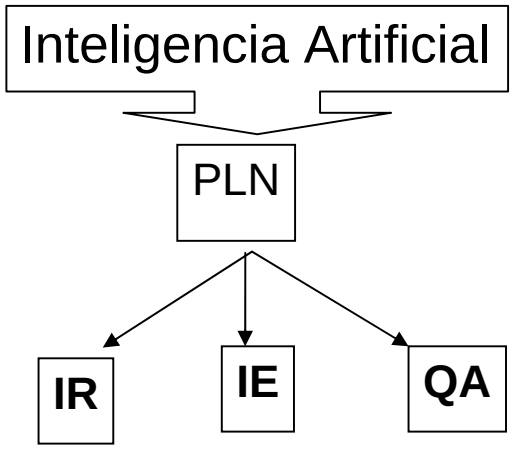
Para facilitar esta tarea, se necesitará disponer de sistemas que sean capaces de localizar la información requerida, procesarla, integrarla y generar una respuesta acorde a los requerimientos expresados por el usuario en sus preguntas. Además, estos sistemas deben ser capaces de comprender preguntas y documentos escritos en lenguaje natural en dominios no restringidos permitiendo así, una interacción cómoda y adecuada a aquellos usuarios inexpertos en el manejo de computadores. Sin embargo, y aunque las investigaciones avanzan en buena dirección, todavía no existe hoy ningún sistema operacional que cumpla todos estos requisitos.

Ante la creciente necesidad de aplicaciones que faciliten, al menos en parte el acceso y tratamiento de toda esta información, la comunidad científica concentra sus esfuerzos en la resolución de problemas más especializados y por tanto, más fácilmente abordables. Esta circunstancia propició el desarrollo de campos de investigación que afrontaron el problema desde diferentes puntos de vista: la recuperación de información (RI) y, posteriormente, la búsqueda de respuestas (BR). En el trabajo se destacarán aquellos aspectos más relevantes de cada una de estas líneas de investigación, además se abordarán otros campos de investigación que tienen estrecha relación con esas líneas. Por último se expondrán algunas ideas para incorporar el estudio de esas líneas de investigación en la Carrera de Informática por la importancia actual de las mismas.

Desarrollo

1. Estado del Arte.

En este apartado se revisarán brevemente los aspectos fundamentales de las líneas de investigación de interés a incorporar en los estudios de la Carrera Informática; así como algunos datos de su estado actual de desarrollo:



- Recuperación de Información *(Information Retrieval)*.
- Extracción de Información *(Information Extraction)*
- Búsqueda de respuestas *(Question Answering)*.

1.1 Campos de investigación

Entre los campos de investigación que se relacionan para desarrollar y dar solución a las cuestiones del lenguaje se encuentran:

- *Lingüística Computacional: s*e ocupa de la aplicación de métodos computacionales en el estudio científico del lenguaje.
- *Procesamiento del Lenguaje Natural (Natural Language Processing)*: se le llama habitualmente a las aproximaciones de los lingüistas computacionales a los lenguajes.
- *Language Engineering (LE)* o *Human Language Technologies (HLT)*: se le denomina habitualmente a las aproximaciones de la ingeniería al lenguaje.

Enlaces con otras disciplinas fuera de la Lingüística:

- *Computer Science*
- *Artificial Intelligence*
- *Cognitive Psychology.*

Las líneas de investigación que analizamos aúnan esfuerzos de los investigadores lingüistas teóricos y de los investigadores empíricos con el objetivo de desarrollar las

teorías necesarias en el ordenador para dar solución a las problemáticas que presentan.

1.2 La recuperación de información.

Los sistemas de recuperación de información (RI) tienen como tareas la selección y recuperación de aquellos documentos que son relevantes a necesidades de información arbitrarias formuladas por los usuarios. Estos sistemas devuelven como resultado una lista de documentos que suele presentarse ordenada en función de valores que intentan reflejar en qué medida cada documento contiene información que responde a las necesidades expresadas por el usuario.

En la actualidad los sistemas de RI más conocidos son aquellos que permiten –con algún grado de éxito- localizar información a través de Internet. Sirvan como ejemplos algunos de los motores de búsqueda más utilizados actualmente como *Google*, Alta Vista o *Yahoo*.

Una característica importante de estos sistemas reside en la necesidad de procesar grandes cantidades de texto en un tiempo muy corto, del orden de milisegundos para búsquedas en Internet. La misma constituye a su vez una limitación, ya que impone una severa restricción en cuanto a la complejidad de los modelos y técnicas de análisis y tratamiento de documentos que pueden emplearse.

Se pueden destacar la aparición de dos líneas de investigación- dentro del ámbito de la RI- orientadas a mejorar el rendimiento de estos sistemas: La recuperación de pasajes (RP) y la aplicación de técnicas de procesamiento del lenguaje natural (PLN) al proceso de RI.

La RP nace como alternativa a los modelos clásicos de RI [5]. Estos sistemas miden la relevancia de un documento con respecto a una pregunta en función de la relevancia de los fragmentos contiguos de texto (pasajes) que lo conforman, un ejemplo de este tipo de sistemas es IR-n, [15]. Esta aproximación facilita la detección, dentro de documentos grandes, de aquellos extractos que pueden ser muy relevantes para el usuario y que, debido a estar inmersos en un documento mayor, pueden pasar desapercibidos cuando el sistema considera el documento completo como una unidad de información. Como demuestran diversos estudios [11], aunque estos sistemas resultan computacionalmente más costosos que los de RI, las mejoras de rendimiento alcanzadas justifican, en la mayoría de los casos, la adopción de este tipo de aproximaciones.

En cuanto a la aplicación de técnicas de PLN, la comunidad científica consideró a priori que su utilización reportaría considerables beneficios a la tarea de RI. Muchos y diversos intentos llevaron a cabo utilizando diversas técnicas y herramientas [18] sin embargo, el esfuerzo empleado no fue recompensado con mejoras de rendimiento sustanciales.

El principal foro de investigación en sistemas de RI lo constituye la serie anual de conferencias *Text REtrieval Conference* [19], en ellas se diseñan una serie de tareas con la finalidad de evaluar y comparar el rendimiento de los diferentes sistemas de RI. A través de las actas de estas conferencias se puede observar con detalle la evolución de las investigaciones desarrolladas en este campo.

1.3 La búsqueda de respuestas.

La investigación en sistemas de RI facilitó el tratamiento de grandes cantidades de información, sin embargo, las características de esta línea de investigación presentaban serios inconvenientes a la hora de facilitar la obtención de respuestas concretas a preguntas muy precisas formuladas de forma arbitraria por los usuarios. Los sistemas de RI se vieron incapaces por sí solos de afrontar tareas de este tipo. De hecho, una vez que el usuario recibía la lista de documentos relevantes a su pregunta, todavía le quedaba pendiente una ardua tarea. Necesitaba revisar cada uno de estos documentos para comprobar en primer lugar, si esos documentos estaban realmente relacionados con la información solicitada y en segundo lugar, localizar en su interior la información puntual deseada.

Por esos inconvenientes y principalmente, por un creciente interés en sistemas que afrontaran con éxito la tarea de localizar respuestas concretas en grandes volúmenes de información, aparece un nuevo campo de investigación conocido como búsqueda de respuestas (BR) o *Question Answering* (QA).

A continuación se definen los sistemas de búsqueda de respuestas y sus características, se presentan las diferentes líneas de investigación que se están desarrollando en este campo.

Se puede definir la BR como la tarea automática realizada por las computadoras que tiene como finalidad la de encontrar respuestas concretas a necesidades precisas de información formuladas por los usuarios. Los sistemas de BR son especialmente útiles en situaciones en las que el usuario final necesita conocer un dato muy específico y no dispone de tiempo -o no necesita- leer toda la documentación referente al tema de la búsqueda para solucionar su problema. A modo de ejemplo, algunas aplicaciones prácticas podrían ser las siguientes:

- Sistemas de ayuda en línea de software.
- Sistemas de consulta de procedimientos y datos en grandes organizaciones.
- Interfaces de consulta de manuales técnicos.
- Sistemas búsqueda de respuestas generales de acceso público sobre Internet.
- etc.

El primer acercamiento a las características de un sistema de BR y la primera aproximación a un sistema funcional (QUALM) fueron introducidos por Wendy Lehnert a finales de los 70, [12] y [13]. En esos trabajos se definieron las características ideales de un sistema de BR. Estos sistemas deberían entender la pregunta del usuario, buscar la respuesta en una base de datos de conocimiento y posteriormente componer la

respuesta para presentarla al usuario. En consecuencia, estos sistemas deberían integrar técnicas relacionadas con el Entendimiento del Lenguaje Natural, la Búsqueda de Conocimiento- incluyendo posiblemente técnicas de inferencia- y la Generación de Lenguaje Natural.

Los inicios de la investigación en sistemas de BR tuvieron lugar en la comunidad científica relacionada con la Inteligencia Artificial (IA). Desde esta perspectiva, la investigación desarrollada consideró requisito indispensable que los sistemas de BR tenían que satisfacer todas y cada una de las características ideales anteriormente citadas. Sin embargo, hasta la fecha únicamente se han podido obtener algunos resultados a costa de restringir mucho los dominios sobre los que se realizan las consultas.

La investigación en sistemas de BR recientemente también se ha afrontado desde el punto de vista de la comunidad especializada en sistemas de RI. Sin embargo, desde esta perspectiva, el poder desarrollar la tarea sobre dominios no restringidos constituye el requisito básico e innegociable a cumplir. Partiendo de este requerimiento inicial, las investigaciones se han orientado hacia el desarrollo de sistemas que van incorporando progresivamente herramientas más complejas que permiten la evolución de estos sistemas hacia la consecución de las características ideales propuestas por Lehner.

Se puede realizar una primera clasificación de los sistemas de BR, teniendo en cuenta las orientaciones planteadas con anterioridad, en dos tipos: sistemas de BR en dominios restringidos y sistemas de BR en dominios no restringidos.

La serie de conferencias TREC se puede considerar como uno de los referentes para conocer el estado del arte de este tipo de sistemas de BR. Otro de los eventos de gran importancia en el área es el FORUM: *Cross-Language Evaluation Forum* (CLEF), [7]. Asimismo, durante el desarrollo de la conferencia TREC-9 (2000) se publicaron dos documentos que pretenden marcar el estado actual y líneas futuras para este tipo de sistemas:

- Vision Statement to Guide Research in Q&A and Text Summarization [6].
- Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A) [4].

Es a partir del TREC-8 (1999) cuando se realiza la primera evaluación exhaustiva de sistemas de BR, extendiendo así las evaluaciones de los sistemas de RI a los de BR, concretamente sobre los denominados *Open-Domain Question Answering*, es decir, sistemas de BR que no trabajan sobre dominios restringidos, sino sobre grandes colecciones de textos de diferentes dominios. El tamaño de las colecciones de textos tratadas es del orden del millón de documentos y tres millones de *bytes*. Respecto al número de preguntas evaluadas hablamos de diferentes cantidades según el año de la evaluación: 200, 693, 500 y 500, para los TREC-8, 9, 10 y 11 respectivamente.

También cabe destacar el interés creciente de Universidades y empresas en este tipo de conferencias, expresado por el alto número de participantes año tras año: 20, 28, 36 y 34 organizaciones respectivamente, destacándose empresas del potencial de IBM,

Microsoft o Xerox. Lo que resalta la importancia de incorporar estas temáticas en los estudios de pregrado y postgrado de la carrera de Informática, para incorporarnos al movimiento mundial de investigación en la rama de Procesamiento del Lenguaje Natural.

1.3 Campos de investigación relacionados

Además de los ya citados, se han desarrollado investigaciones en otros campos también cercanos a la búsqueda de respuestas: el proceso de *tests* de lectura y comprensión de textos (*Reading Comprehension Tests*) y la búsqueda de preguntas frecuentes (*Frequently Asked Questions Finding*).

Proceso de tests de lectura y comprensión de textos. Los *tests* de lectura y comprensión de textos conforman una herramienta tradicionalmente utilizada para evaluar el nivel de comprensión que un lector adquiere al leer un documento.

Un test de lectura y comprensión está formado por dos elementos: un texto en el que se narra una historia o noticia y un conjunto de preguntas de *test* relativas a dicha narración. La complejidad de estas preguntas suele ser creciente. Esto permite evaluar el nivel y la capacidad de comprensión del texto alcanzado por el lector mediante la comprobación de sus respuestas.

El proceso automático de este tipo de *tests* presenta varias vertientes de interés. La primera de ellas reside en el uso de estos *tests* como material de evaluación de sistemas automáticos de comprensión del lenguaje natural [17]. Su uso se está considerando como alternativa a los sistemas actuales utilizados para evaluar técnicas avanzadas de PLN. En particular, pueden utilizarse como banco de pruebas para medir el rendimiento de los sistemas de BR si bien, cabría tener en cuenta el escaso volumen de información que contienen.

Quizás el ámbito de aplicación más interesante se basa en la construcción de sistemas que permitan evaluar de forma automática el nivel de comprensión que un lector o bien, un sistema automático, alcanzan al leer un documento. Este proceso se realizaría mediante la comparación de las respuestas correctas incluidas en el *test* con las que suministra el lector o el sistema automático que procesa dicho *test*. La eficiencia de este método radica básicamente en la obtención de medidas de similitud que permitan determinar de forma fiable cuando ambas respuestas (la correcta y la suministrada) son equivalentes.

En este marco se encuentra la propuesta presentada en [3]. Este trabajo desarrolla un sistema automático de evaluación de sistemas de BR basado en un proceso de comparación de las respuestas correctas suministradas por humanos con las devueltas automáticamente por el sistema de BR.

***Búsqueda de preguntas frecuentes*.** Los sistemas de búsqueda de preguntas frecuentes tienen como objetivo localizar y devolver pasajes de texto como respuesta a preguntas de los usuarios.

Las principales diferencias de estos sistemas con los de BR radican en las características de la base de datos documental sobre la que realizan el proceso, y la forma de búsqueda de la respuesta [2]. Estos sistemas utilizan bases documentales formadas por conjuntos de preguntas que tienen asociadas sus correspondientes respuestas. Ejemplos de estas bases de datos pueden ser los conjuntos de preguntas más frecuentes (*Frequently Asked Questions* - FAQs) disponibles en Internet y que versan sobre temas muy diversos.

Estos sistemas realmente no realizan una búsqueda de respuestas tal y como se ha definido previamente. Simplemente localizan aquellas preguntas incluidas en la base documental que son similares a la realizada por el usuario y como resultado, presentan sus correspondientes respuestas asociadas.

Sistemas como FAQ Zinder [9] o Askjeeves [1] constituyen ejemplos de algunas implementaciones que están actualmente disponibles en Internet.

Se pueden citar además otros campos que entran dentro de la línea general de investigación: Búsqueda Inteligente de Información en la Web.

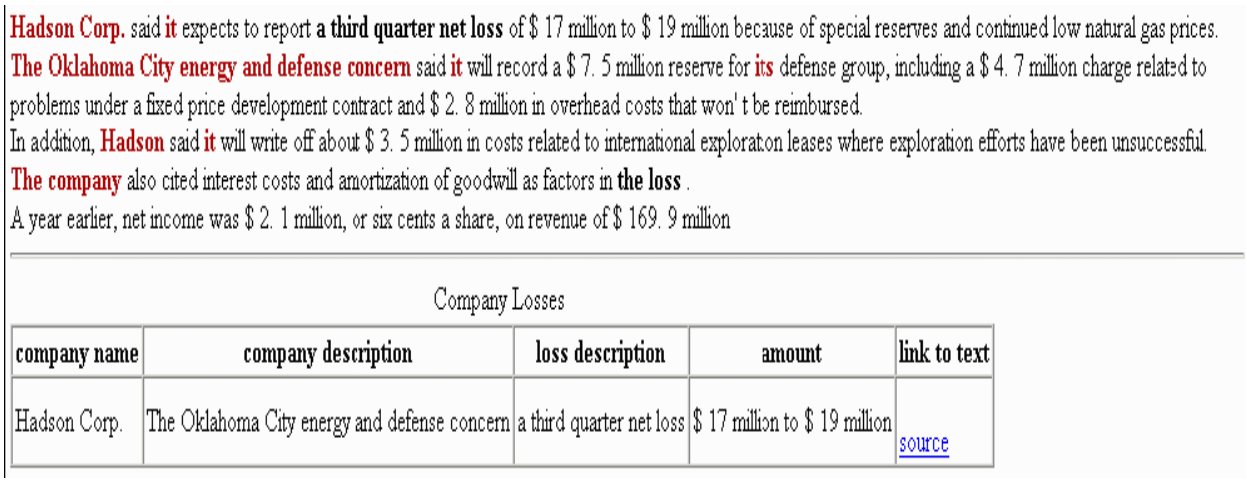
Extracción de información. Los sistemas de extracción de información (EI) realizan la tarea de buscar información muy concreta en colecciones de documentos. Su finalidad consiste en detectar, extraer y presentar dicha información en un formato que sea susceptible de ser tratado posteriormente de forma automática.

Estos sistemas se diseñan y construyen de forma específica para la realización de una tarea determinada, en consecuencia, dispondremos de un sistema diferente en función del tipo de información a extraer en cada caso. Un ejemplo podría ser un sistema de EI orientado a la extracción de la Información relevante de los Programas de Disciplina [10]. Este sistema opera de forma que cada vez que aparece uno de los datos definidos como relevantes, lo extrae y lo incorpora en el campo correspondiente de una base de datos creada a tal efecto. Como puede deducirse, estos sistemas necesitan aplicar técnicas complejas de PLN debido la gran precisión que se requiere en los procesos de detección y extracción del tipo de información que les es relevante.

La investigación en este campo ha sido muy intensa. En particular, la serie de conferencias *Message Understanding Conference* (MUC) han constituido uno de sus principales foros de promoción. Estas conferencias han permitido la evaluación y comparación de diversos sistemas, realizando para la EI la misma función que las conferencias TREC para la recuperación de información.

En la Figura 1 se muestra un ejemplo de un sistema de extracción de información disponible en la Web, en http://nlp.shef.ac.uk/research/areas/ie.html . En el sistema se observa que dado un texto no estructurado en lenguaje natural, es capaz de extraer las entidades (aparecen resaltadas en color rojo), y de rellenar la plantilla denominada *Company Losses*, la cual se ha de conocer previamente y que consta de los campos: *company name, company description, loss description, amount* y *link to text*. Se podría

efectuar el relleno de otras plantillas, pero siempre se deberían conocer previamente al diseño del sistema de extracción de información, lo que le resta flexibilidad frente a lo que ofrece un sistema de BR ante las diferentes peticiones de información de un usuario.



Hadson Corp. said it expects to report a third quarter net loss of $ 17 million to $ 19 million because of special reserves and continued low natural gas prices.
The Oklahoma City energy and defense concern said it will record a $ 7. 5 million reserve for its defense group, including a $ 4. 7 million charge related to problems under a fixed price development contract and $ 2. 8 million in overhead costs that won' t be reimbursed.
In addition, Hadson said it will write off about $ 3. 5 million in costs related to international exploraton leases where exploration efforts have been unsuccessful.
The company also cited interest costs and amortization of goodwill as factors in the loss .
A year earlier, net income was $ 2. 1 million, or six cents a share, on revenue of $ 169. 9 million

Company Losses

company name	company description	loss description	amount	link to text
Hadson Corp.	The Oklahoma City energy and defense concern	a third quarter net loss	$ 17 million to $ 19 million	source

Figura 1. Ejemplo de un sistema de extracción de información.

Interfaz de Acceso a Bases de Datos en Lenguaje Natural. Con referencia a las interfaces de acceso a bases de datos en lenguaje natural (ILNBD), mientras que éstas trabajan sobre información estructurada, es decir, bases de datos en las que se conoce la estructura de sus campos (número, tipo, contenido, longitud, etc.), la BR trabaja sobre información no estructurada, o sea, documentos en lenguaje natural que no suelen tener una estructura predefinida. Además, la BR tiene como objetivo el trabajar sobre dominios no restringidos, por lo que pueden realizar búsquedas de información sobre documentos de diferentes estilos (narrativo, manuales técnicos, periodísticos, etc.), en tanto que estos interfaces sólo pueden buscar información sobre el dominio concreto de la base de datos. Como ejemplo de una interfaz de este tipo podemos observar la Figura 2, en la que se accede a una serie de bases de datos con la información de la Universalización de la Enseñanza Superior de la Provincia de Matanzas [21], y brinda la posibilidad de realizar las búsquedas en lenguaje natural. La base de datos almacena información del proceso de Universalización, concretamente sobre las sedes, su localización, carreras que se ofertan, asignaturas por carrera, matrícula de las sedes, etc.

Figura 2. Ejemplo de una ILNBD

Clasificación de textos. La clasificación de textos o *Automated Text Categorization* no más que la asignación a documentos de categorías previamente establecidas y su aplicación fundamental se encuentra en el filtrado de textos o *Content/Text Filtering (TF)*. Es un sistema que dado un documento de entrada decide qué documentos son relevantes para el usuario en función de su perfil previamente establecido.

Podemos citar algunos ejemplos de este tipo de sistema:

- Filtrado de correo (anti-spam)
- Filtrado de páginas Web con contenidos violentos.

La diferencia principal entre la RI y el TF, es que la RI resuelve necesidades de información específica y temporal clasificando los documentos en relevantes o no y las necesidades de información son dinámicas (definidas a corto plazo): varían en cada petición del usuario, lo cual no puede ser resuelto con los sistemas de clasificación de textos.

Recuperación de información multilingüe. Los sistemas de recuperación multilingüe o *Cross Language Information Retrieval* (CLIR) funcionan igual que los sistemas de RI para un único lenguaje la diferencia entre ellos estriba en que en los CLIR la pregunta y/o los documentos pueden estar en diferentes idiomas. Entre los métodos que se usan se encuentran: los traductores automáticos y el uso de mecanismos interlingua.

2. Propuesta

Tras presentar la evolución de las investigaciones producidas en los diferentes campos relacionados con el desarrollo de sistemas orientados a facilitar la búsqueda, localización y extracción de información textual, esta sección se centra en el objetivo fundamental del trabajo, que es la exposición de los objetivos que consideramos deben ser seguidos en la elaboración de la propuesta de los contenidos a incorporar en la carrera de Informática.

Objetivos

- Estudiar el funcionamiento de los buscadores de información.
- Mejorar su funcionamiento: precisión y calidad de los resultados devueltos.
- Aplicar técnicas de procesamiento del lenguaje natural a estos buscadores.
- Permitir búsquedas sobre preguntas concretas escritas en lenguaje natural, y no únicamente limitarnos a búsquedas por palabras clave.
- Obtener sistemas que logren ir un paso más allá de lo que realizan los buscadores tradicionales, para ello, contestar a las preguntas del usuario. Es decir, en lugar de devolver el documento completo, devolver la zona del texto donde se encuentra la información requerida.
- Introducir las técnicas que permiten añadir una capacidad de trabajo multilingüe a los buscadores.
- Estudiar los campos relacionados con la Búsqueda de Respuesta y sus diferencias.
- Profundizar en todos los campos que se engloban dentro de la línea general de investigación: Búsqueda Inteligente de Información en la Web.

Proposición de contenidos

A continuación se propone incluir en los estudios de la carrera de Informática, específicamente en la disciplina de Inteligencia Artificial, los siguientes contenidos:

- La Búsqueda Inteligente de Información en la Web, con todas las líneas de investigación que abarca.
- Técnicas de procesamiento del lenguaje natural (PLN).
- Sistemas de búsqueda de información (*Information Retrieval*). Estado del arte.
- Incorporación de técnicas de lenguaje natural a los buscadores de información.
- Sistemas de búsqueda de respuesta (*Question* Answering).
- Sistemas de Recuperación de Pasajes.
- Recuperación de información multilingüe (Cross *Language Information Retrieval*).

Conclusiones

- El estudio y desarrollo de herramientas de Procesamiento de Lenguaje Natural es de gran importancia para resolver los problemas existentes en el acceso y tratamiento de la gran cantidad de información disponible en la Web.
- Los sistemas de búsqueda de respuesta facilitan la obtención de respuestas concretas a preguntas muy precisas formuladas de forma arbitraria por los usuarios, sin embargo los sistemas de recuperación de información sólo son capaces de devolver la lista de documentos relevantes a una pregunta formulada por el usuario. Por tanto, los dos sistemas se complementan y juntos logran de manera eficiente los resultados que espera el usuario.
- Existen otros campos dentro del PLN y de la Búsqueda Inteligente de Información en la Web que son de gran importancia: Extracción de Información, Clasificación de Textos, Búsqueda de Preguntas Frecuente, Proceso de Comprensión de Textos, Recuperación de Información Multilingüe, entre otros.
- Se presentó una propuesta para incluir el estudio del PLN en la disciplina de Inteligencia Artificial en la carrera de Ingeniería Informática, por la importancia que reviste esta rama de la informática en la actualidad.

Recomendaciones

En el marco del perfeccionamiento de la propuesta de los temas que se deben incorporar en los estudios de la carrera de Informática se debe continuar trabajando por elevar la calidad en varias direcciones, tales como:

- Reforzar aún más el trabajo investigativo en un área tan importante como el Procesamiento del Lenguaje Natural con el fin de elevar la calidad del profesional e investigador graduado como Ingeniero Informático.
- Proponer a la comisión de carrera esta propuesta y en dependencia de los resultados definir finalmente su implementación.

Bibliografía

1. Askjeeves. http://www.askjeeves.com
2. Berger, Adam, Rich Caruana, David Cohn, Dayne Freitag y Vibhu Mittal (2000).Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding, en Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Question Answering, págs. 192-199, Athens, Greece.
3. Breck, Eric, John Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light y Inderjeet Mani (2000b). How to Evaluate Your Question Answering System Every Day ... and Still Get Real Work Done, en Proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000, Athens, Greece.
4. Burger, John; Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, Ralph Weishedel. *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A).* http://wwwnlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc. 2000.
5. Callan, James P. (1994). Passage-Level Evidence in Document Retrieval, en Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval, págs. 302-310, Springer Verlag, London, UK.
6. Carbonell, Jaime; Donna Harman, Eduard Hovy, Steve Maiorano, John Prange, Karen Sparck-Jones. *Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization.* http://wwwnlpir.nist.gov/projects/duc/papers/Final-Vision-Paper-v1a.pdf. 2000.
7. Cross-Language Evaluation Forum (CLEF): http://clef.iei.pi.cnr.it:2002/.
8. Curso *IR & NLP Course* (UNED): http://rayuela.ieec.uned.es/~ircourse/ .
9. FAQ Finder. http://faqfinder.cs.uchicago.edu:8001/
10. *García, Yaniseth. "Herramienta para la Extracción de Información de los Programas de Disciplina"*. Trabajo de Diploma para optar por el título de Ingeniero Informático, Universidad de Matanzas Camilo Cienfuegos, Cuba. 2005.
11. Kaszkiel, Marcin y Justin Zobel (2001). Effective Ranking with Arbitrary Passages, Journal of the American Society for Information Science (JASIS), 52(4), 344-364.
12. Lehnert, Wendy G. (1977). Human and computational question answering, Cognitive Science, (1), 47-63.
13. Lehnert, Wendy G. (1980). Question answering in natural language procesing, en Carl Hansen Verlag, editor, Natural Language Question Answering Systems, págs. 9-71.
14. Libro INFORMATION RETRIEVAL BOOK: C. J. van RIJSBERGEN: http://www.dcs.gla.ac.uk/Keith/Preface.html.
15. Llopis, F. Pascual. IR-n: Un sistema de Recuperación de Información basado en Pasajes. Tesis Doctoral de Fernando Llopis Pascual. 2003.

16. Martínez, F. Santiago. El problema de la fusión de colecciones en la recuperación de información multilingüe y distribuida: cálculo de la relevancia documental en dos pasos. Tesis Doctoral Fernando Martínez Santiago. 2004.
17. Riloff, E. y M. Thelen (2000). A Rule-based Question Answering System for Reading Comprehension Tests, en ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, págs. 13-19, Seattle, Washington.
18. Strzalkowski, T., G. Stein, G. Bowden Wise, J. Perez-Carballo, P. Tapananinen,T. Jarvinen, A. Voutilainen y J. Karlgren (1998). Natural language information retrieval: TREC-7 report, en Seventh Text REtrieval Conference, vol. 500-242 de NIST Special Publication, págs. 217-226, National Institute of Standards and Technology, Gaithersburg, USA.
19. Text REtrieval Conference (TREC) Home Page: http://trec.nist.gov/ .
20. Vicedo, J. Luis. El modelo SEMQA: un modelo semántico aplicado a los sistemas de Búsqueda de Respuestas. Tesis Doctoral de José Luis Vicedo. 2002.
21. Vila, Katia Rodríguez. *Interfaz de Acceso en Lenguaje Natural a la Información de la Municipalización en Matanzas*. Trabajo de Diploma para optar por el título de Ingeniero Informático, Universidad de Matanzas Camilo Cienfuegos, Cuba. 2005.
22. Wesley, Addison. Modern Information Retrieval. Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Addison Wesley.1999.