Revista Signos
ISSN: 0035-0451
revista.signos@ucv.cl
Pontificia Universidad Católica de Valparaíso
Chile

Ferreira, Anita; Atkinson, John
Disminución de la sobrecarga de información en la World Wide Web a partir de interacciones dialógicas hombre-computador
Revista Signos, vol. 42, núm. 69, 2009, pp. 9-27
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Revista Signos
2009, 42(69)
9-27

Disminución de la sobrecarga de información en la *World Wide Web* a partir de interacciones dialógicas hombre-computador*

Anita Ferreira
John Atkinson
Universidad de Concepción
Chile

Resumen: En este trabajo se presenta un enfoque de procesamiento automático de lenguaje natural para diálogos cooperativos basados en la web. Este enfoque se centra en las consultas del usuario, generando automáticamente interacciones lingüísticas que consideran el contexto, la retroalimentación del usuario y los resultados iniciales de la búsqueda. Se describen también los diferentes componentes para el procesamiento de lenguaje natural en el contexto de interacciones dialógicas. Finalmente, se discuten los principales resultados generados a partir de un prototipo computacional que tiene como fin reducir tanto el número de turnos conversacionales requeridos como la sobrecarga de información.

Palabras Clave: Procesamiento de Lenguaje Natural (PLN), búsquedas inteligentes en la web, Filtrado de Información (FI), Recuperación de Información (RI), Generación de Lenguaje Natural (GLN).

* Proyecto FONDECYT 1070714 *An Interactive Natural-Language Dialogue Model for Intelligent Filtering based on Patterns Discovered from Text Document*.

Recibido:
13-VIII-2007
Aceptado:
24-III-2008

Correspondencia: Anita Ferreira (aferreir@udec.cl). Departamento de Español, Facultad de Humanidades y Arte, Universidad de Concepción. Casilla 160-C, Correo 3, Concepción Chile.

Reducing information overloading on the World Wide Web by using human-computer dialogue interactions

Abstract: This work suggests a natural language processing approach to web-based cooperative dialogs. It focuses on the user's requests automatically generating language-drive interactions that take into account the context, user feedback, and the initial web search results. The different components of natural language processing in the context of dialog discourse interactions are described. The main results of a working prototype aiming to decrease both the number of conversational turns and the information overload are lastly discussed.

Key Words: Natural Language Processing (NLP), Intelligent Web Search, Information Filtering (IF), Information Retrieval (IR), Natural Language Generation (NLG).

INTRODUCCIÓN

El vertiginoso desarrollo que ha tenido la *World Wide Web* (www) ha transformado a la Red Internet en un inmenso espacio de información con contenidos diversos, a menudo, muy poco estructurados u organizados. Los usuarios deben lidiar con cantidades siempre crecientes de información, un fenómeno que se plasma en la frase 'sobrecarga de información'. Si bien una capacitación mínima parece proporcionar las habilidades necesarias para navegar sitios web individuales, la búsqueda basada en palabras clave y la navegación entre sitios requieren cada vez más de mayor experiencia (Jansen & Pooch, 2000; Jansen & Spink, 2000; Tong, Changjie & Jie, 2001; Benamara & Saint Dizier, 2003).

La utilización eficiente de motores de búsqueda como Altavista y Excite precisa de conocimientos sofisticados. Las investigaciones que se han realizado sobre el comportamiento adoptado tanto por novicios como por expertos en la búsqueda de información han sido utilizadas en diversos tipos de aplicaciones. Así, por ejemplo, el modelamiento del comportamiento que adoptan los usuarios en la búsqueda de información puede servir como base para el mejoramiento de las interfaces y funcionalidades de los sistemas de búsqueda existentes (Yahoo, Google, etc.). Estas investigaciones han demostrado que los sistemas más sofisticados del futuro podrán identificar las variadas necesidades de los usuarios novicios y expertos (Lau & Horvitz, 1999). Además, una comprensión más profunda de las dificultades que enfrentan los usuarios en el proceso de búsqueda puede ser beneficioso para los sistemas de ayuda (Holscher & Strube, 2000).

La investigación sobre el modelamiento de usuarios se ha centrado hasta cierto punto en el comportamiento de los usuarios en la web. Lau y Horvitz (1999), por ejemplo, construyeron redes bayesianas para modelar las sucesivas consultas que hacen los usuarios de motores de búsqueda. Estas redes pueden enriquecer las bitácoras de los motores de búsqueda con categorías manualmente elaboradas de metas comunicativas que permiten predecir las modificaciones que se

realizarán en las consultas. De manera similar, Zuckerman, Albrecht, Nicholson y Doktor (2000) propusieron el uso de modelos markovianos para predecir la siguiente consulta de los usuarios en la web, basándose en el tiempo y ubicación de sus consultas anteriores. Sin embargo, estos estudios no toman en cuenta el perfil del usuario ni su nivel de experiencia.

Aunque los sistemas tradicionales de búsqueda de información que utilizan palabras clave pueden constituir un primer paso en el proceso de una búsqueda general, el desafío es realizar tareas con mayor precisión e inteligencia, aplicando los conocimientos del usuario —sus intenciones, metas, etc.— para mejorar la calidad de los resultados de una búsqueda cimentada en un mínimo número de interacciones o intercambios comunicativos. De tal modo que las limitaciones de búsquedas basadas en palabras clave se pueden superar si se utilizan sistemas de diálogos en lenguaje natural entre el usuario y el motor de búsqueda (Bloedorn & Mani, 1998). En este aspecto, el estudio que aquí presentamos propone un modelo de diálogo hombre máquina (del inglés *Human Computer Interactions*) en el cual la retroalimentación (del inglés *feedback*) que se da en interacciones dialógicas en lenguaje natural puede desempeñar un rol clave en la reducción de la sobrecarga de información y la obtención de información precisa sobre lo que el usuario realmente busca. Este conocimiento lingüístico subyacente podría ayudar a hacer más específico el sistema de búsqueda al restringir los requisitos e identificar las intenciones del usuario.

De acuerdo con lo anterior, en este estudio se explora la generación de diálogos interactivos en lenguaje natural para búsquedas bibliográficas en la web con el fin de mejorar el proceso y filtrado de información mediante un número reducido de intercambios lingüísticos con el usuario. Nuestro enfoque se centra en mejorar el paradigma de búsqueda de información, a través de un modelo basado en la lingüística computacional y un agente de búsqueda inteligente.

1. Sistemas de búsqueda de información

Muchos motores de búsqueda tales como Google, Yahoo y Altavista usan un *software* automatizado que obtiene el contenido de cada servidor que encuentra en la web, creando bases de datos indexadas en la medida en que encuentra documentos. Sin embargo, los usuarios pueden enfrentar serios problemas cuando usan esas enormes bases de datos, entre otros, gastan un tiempo significativo en verificar si los resultados recuperados contienen lo que ellos buscaban. Desde la perspectiva de la recuperación de información, los usuarios pueden encontrar demasiada información relacionada con un ámbito temático por lo que ellos, a menudo, no consultan todos los enlaces y abandonan la búsqueda con un conjunto limitado de recuperación de documentos.

Estudios sobre las conductas de los usuarios en la web y temas sobre la usabilidad demuestran que se deben superar numerosos obstáculos para buscar información de manera eficaz (Holscher

& Strube, 2000). Como se muestra en la Figura 1, estos obstáculos van desde problemas de experiencia (¿cómo recuperar una página existente?) hasta temas de diseño (el navegador está diseñado inadecuadamente o es difícil de utilizar eficazmente). Para un porcentaje significativo de usuarios web cuesta mucho tiempo buscar información como documentos específicos o páginas web determinadas. Otros temas se relacionan directamente con la dificultad de recuperar información útil y comprensible.

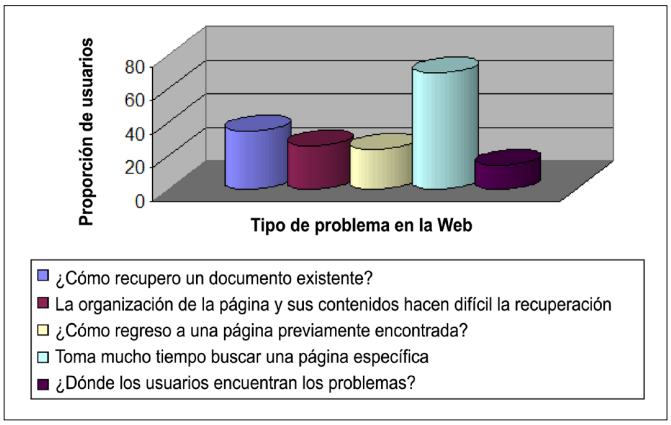

Figura 1. Factores de dificultad en la Búsqueda de Información en la web. Adaptado de Holscher y Strube (2000).

Un análisis, en detalle, de estos factores revela dos problemas específicos:

a) Los sistemas de búsqueda actual no pueden examinar el comportamiento, intenciones o perfil del usuario web con el fin de recabar información que sería de ayuda en la automatización de tareas rutinarias.

b) La representación basada en palabras clave que utilizan los motores de búsqueda y los sistemas de recuperación de información es restrictiva.

De hecho, muchos sistemas de búsqueda no pueden capturar información subyacente que es relevante para el usuario. Esto ocurre en parte porque no es factible que los sistemas de búsqueda tradicionales capturen el conocimiento implícito del diálogo con el cual los usuarios a menudo se comunican. Por ejemplo, si una consulta en lenguaje natural utiliza información de trasfondo implícita o referencias pronominales ("Busca documentos que contengan…" o "¿Cuáles son los planes de viaje de la presidenta de Chile para el próximo año?"), el motor de búsqueda no ha-

llará dicha información. En cuanto a la precisión y a la recuperación, el sistema también pasará por alto miles de documentos relevantes si las consultas no contienen suficientes palabras clave para permitir que se emitan juicios de similitud. Ahora bien, en los ejemplos dados, el primer problema se relaciona con la 'inteligencia' y capacidad de adaptación del motor de búsqueda, mientras que el segundo involucra la representación de las consultas, la interacción del usuario y la capacidad del sistema de captar la información subyacente (e implícita) expresada en lenguaje natural.

1.1. Agentes de búsqueda inteligentes

El área de la inteligencia artificial denominada como 'agentes de búsqueda inteligentes' (Ardissono, Boella & Lesmo, 2000; Levy & Weld, 2000; Salter & Antonopoulos, 2006) permite abordar los problemas anteriormente descritos. Estos agentes utilizan la tradicional tecnología de los 'buscadores' de una manera distinta. Normalmente, estas herramientas son 'robots' que el usuario puede entrenar para buscar recursos informativos específicos en la web. Este agente se puede personalizar para desarrollar perfiles individuales o satisfacer necesidades informativas precisas. Un agente inteligente también puede ser autónomo, esto es, puede emitir juicios sobre la probable relevancia del material.

Otro rasgo importante de estos sistemas es que su uso frecuente como herramientas de búsqueda aumenta su efectividad. Los agentes aprenden de la experiencia y el usuario tiene la oportunidad de revisar los resultados de la búsqueda y rechazar cualquier fuente de información irrelevante o inútil. El agente almacena esta información en el perfil del usuario y la utiliza cuando realiza búsquedas, aprende de sus incursiones anteriores en la web y produce una agenda de búsqueda más nítidamente definida si así se desea. No obstante, el empleo de cualquiera de los enfoques actuales conlleva problemas prácticos para el usuario web, incluyendo la sobrecarga de información, la cantidad de tiempo destinado a la realización de la búsqueda y la obtención de resultados relevantes.

Algunos sistemas de búsqueda no pueden realizar un análisis lingüístico más profundo de los requisitos del contexto como para ayudar a proporcionar la información de calidad que el usuario realmente quiere. Para abordar estas limitaciones, los sistemas incorporan tanto variables estadísticas como variables lingüísticas en el proceso de búsqueda. Sin embargo, este enfoque está todavía en su fase preliminar y se basa, principalmente, en la representación de documentos bajo el paradigma *bag of words* (bolsa de palabras), un supuesto subyacente de muchos sistemas de Recuperación de Información (RI) (del inglés *Information Retrieval*) de última generación. Al contrario de la RI, el Filtrado de Información (FI) (del inglés *Information Filtering*) selecciona documentos sobre la base de su contenido. Ejemplos de este tipo son el sistema de FI cognitivo

de Landauer, McNamara, Dennis y Kintsch (2007), que utiliza el 'análisis semántico latente' (del inglés *Latent Semantic Analysis*) para filtrar artículos de noticias y el sistema *Infoscope* (Fischer & Stevens, 1991) que utiliza agentes basados en reglas para monitorear el comportamiento del usuario y hacerle sugerencias.

Para enfrentar la dificultad de crear perfiles adecuados durante el proceso dialógico, algunos sistemas actuales de filtrado de información permiten que el usuario indique uno o más documentos que reflejan sus intereses (Tong, Changjie & Jie, 2001), en vez de exigirle una definición directa y explícita de sus intereses. Otros sistemas intentan elaborar un perfil a partir del comportamiento del usuario. Sin embargo, este enfoque no es totalmente práctico porque el usuario a menudo no está centrado en una meta clara. Es decir, los usuarios suelen navegar la web sin tener un objetivo claro, lo cual puede llevar a los agentes de búsqueda a hacer suposiciones incorrectas.

Desde el punto de vista del Procesamiento del Lenguaje Natural (PLN), estos problemas podrían superarse hasta cierto punto si se precisara más específicamente el tópico que busca el usuario, o si se generara más interacciones explicativas para así hacer que el usuario se centrara más en lo que le interesa (Jurafsky & Martin, 2000). Aunque algunos investigadores han utilizado tecnologías de procesamiento del lenguaje natural para elaborar un perfil del usuario, solo las han aplicado a dominios restringidos que utilizan *WordNet* o recursos más antiguos, y principalmente se han centrado en los problemas de recuperación y de la generalización de conceptos para responder proactivamente a las exigencias del usuario (Bloedorn & Mani, 1998).

El PLN puede ser de utilidad con las estrategias de búsqueda que contemplan las interacciones con el usuario. En particular, los investigadores pueden utilizar técnicas de Generación de Lenguaje Natural (GLN) para permitir que el sistema dialogue de una manera efectiva con el usuario (Moore, Foster, Lemon & White, 2004; Reitter & Moore, 2007). Un problema importante en esta área es cómo reducir el número de turnos de conversación o interacción que se generan para que el usuario pueda obtener la información que busca. Algunas líneas de investigación consideran que este discurso dialógico es una totalidad que está estructurada en distintos niveles en los cuales la comunicación constituye una acción indirecta (Reitter & Dale, 2000).

Las tareas de la GLN incluyen:

- Determinación de contenidos: decidir qué decir, lo cual tiene un impacto tanto en el nivel macro (cómo determinar el contenido de un enunciado o de un turno dialógico) como en el nivel micro (cómo determinar el contenido de expresiones de referencia apropiadas).
- Estructuración de textos: identificar las estructuras más apropiadas para determinadas circunstancias.

- Realización superficial: mapear el contenido oracional a palabras y a su representación morfológica y gramatical.

El diseño de sistemas de GLN se ha focalizado fuertemente en la generación de textos (Benamara & Saint Dizier, 2003) y de sus contenidos a nivel del discurso (Reiter & Dale, 2000), en el cual tareas complejas como la planificación discursiva desempeñan un papel clave en la generación de textos eficaces (Reiter & Moore, 2007; Reiter, Moore & Keller, 2006). Otros enfoques incorporan la teoría de actos de habla en la descripción de sistemas computacionales que producen planes (Ardissono, Boella & Lesmo, 2000) que contienen secuencias de habla (Cohen & Levesque, 1990). Cuando el procesamiento del discurso involucra la gestión de interacciones dialógicas entre el usuario y el sistema, los sistemas de GLN pueden captar conocimientos subyacentes, como, por ejemplo, los turnos conversacionales, para así proporcionar respuestas acordes con los conocimientos y objetivos del usuario, reaccionar ante errores o enfrentar una reacción inesperada del usuario (Jurafsky & Martin, 2000).

1.2. Búsquedas inteligentes en la www que utilizan retroalimentación en lenguaje natural

La mayoría de los sistemas de búsqueda web que utilizan tecnología de 'Procesamiento del Lenguaje Natural' se han diseñado como 'Sistemas de Preguntas y Respuestas' (del inglés *Question-Answering Systems*). Estos sistemas utilizan el procesamiento lingüístico y otros recursos, tales como el etiquetado o la desambiguación de significados, para recuperar documentos que contengan párrafos específicos que respondan a las consultas en Lenguaje Natural. Sin embargo, debido a que en estos sistemas no existe diálogo estos esfuerzos se centran en la recuperación de párrafos precisos dentro de documentos, y no en las necesidades del usuario. En algunos casos, las técnicas lingüísticas deben ser óptimas para que puedan ayudar al proceso de selección de la información. Otras veces, las consultas de búsqueda son difíciles de formular. No obstante, esta dificultad se puede mitigar si se enfocan los diálogos y consultas a dominios y tareas restringidas.

Nuestro enfoque de búsqueda y filtrado inteligente en la web incorpora tecnologías de agentes inteligentes y técnicas de PLN para explotar información sobre el contexto, los intereses, objetivos y comportamientos del usuario en un diálogo adaptativo. El punto central de nuestro modelo metodológico considera la planificación discursiva como dependiente de la tarea y de las capacidades de análisis de diálogos como parte de un 'sistema de búsqueda interactivo'.

En vez de proporcionar muestras o buscar información en la web, nuestro enfoque utiliza un sistema de búsqueda dialógico que considera los intereses específicos del usuario con el objetivo de mejorar los requisitos de la búsqueda. Esto permite que el sistema pueda obtener a través de las interacciones dialógicas conocimientos implícitos para su uso en el refinamiento y filtrado de

los resultados de la búsqueda. Un sistema que utiliza este enfoque debería poder determinar las necesidades del usuario rápidamente, mediante una combinación de información proporcionada por el usuario y la retroalimentación basada en la interacción en Lenguaje Natural.

La Figura 2 grafica el Modelo del Agente de Búsqueda y Filtrado mediante Retroalimentación en Lenguaje Natural. La operación se inicia con una consulta en Lenguaje Natural proporcionada por el usuario. La consulta pasa, luego, a la fase de procesamiento del discurso, en la cual se generan los intercambios de interacción (turnos) relevantes en forma de enunciados en LN, convirtiéndose en una consulta de búsqueda más elaborada y específica. En la medida que avanza el diálogo, el sistema genera una consulta de búsqueda cada vez más refinada, la cual finalmente se envía a un agente de búsqueda.

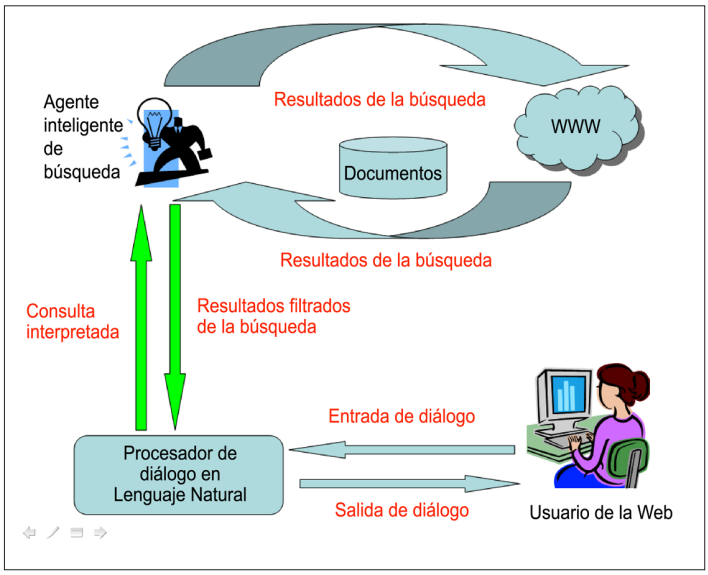

Figura 2. Modelo de un Agente de Búsqueda y Filtrado mediante Retroalimentación en Lenguaje Natural.

2. Estudio experimental

Con el objeto de desarrollar y probar estructuras de diálogo en el contexto situacional de búsquedas bibliográficas en la web, utilizamos la técnica experimental 'Mago de Oz' (WoZ) (Jurafsky & Martin, 2000), la cual permite simular en ambientes computacionales interacciones en lenguaje natural con usuarios, a quienes se les hace creer que pueden interactuar dialógicamente con un sistema computacional cuando en verdad lo están haciendo con una persona. De esta forma, se puede obtener muestras de diálogo muy similares a la interacción hombre-máquina. A través de dicha técnica, se obtuvo un corpus significativo de diálogos que se grabó, etiquetó y analizó lingüísticamente para determinar su estructura discursiva con el fin de establecer un modelo de generación de lenguaje natural basado en estruturas de diálogo que orienten la planificación textual y generación de un discurso interactivo computacional. El sistema simula la interacción entre un ser humano y un sistema que procesa el lenguaje natural de manera real. El proceso interactivo continúa hasta que el modelo cumple con las expectativas o con determinados requerimientos de diseño.

2.1. Muestra

Un grupo de 22 sujetos humanos no expertos utilizó el simulador WoZ para buscar información en la web. Las interacciones entre los sujetos y el sistema fueron registradas y posteriormente analizadas lingüísticamente con la finalidad de determinar su estructura enunciativa y discursiva. El proceso de refinamiento, basado en la información disponible, continuó hasta que se satisficieran las necesidades y metas de cada sujeto. En la práctica, se estableció un umbral de 20 minutos para revisar hasta qué punto el sujeto cumplió con el objetivo comunicativo.

Después de realizar la búsqueda, se les pidió a los participantes de la interacción que proporcionaran 'explicaciones' y 'descripciones' de los resultados que obtuvieron. Esto se hizo para poder desarrollar un modelo computacional que utilice documentos extraídos del la web para producir descripciones y explicaciones en LN.

2.2. Modelo del generador de diálogos interactivos en lenguaje natural (LN)

El modelo de generación de diálogos interactivos en LN considera diversos componentes, incluyendo el contexto, los conocimientos de los participantes (los usuarios y el sistema) y la situación en la cual el diálogo se concibe (por ejemplo, interacción para buscar información en la web). También contempla un conjunto de módulos para los cuales la entrada y salida se delimitan de acuerdo con las distintas etapas de información lingüística y no lingüística que se define en el diálogo. Este componente de procesamiento se basa en modelos lingüísticos de procesamiento del discurso computacional (Moore, 1994).

En la Figura 3, el modelo de PLN propuesto genera salidas de discurso a partir de los resultados de una búsqueda bibliográfica en la web. El proceso se inicia con la entrada del usuario (consulta en Lenguaje Natural) y produce una salida compuesta de un enunciado en LN para desencadenar el diálogo y orientar al usuario o bien para enviar una solicitud de búsqueda al agente de búsqueda (si el usuario se encuentra en la última etapa del diálogo y toda la información se ha filtrado).

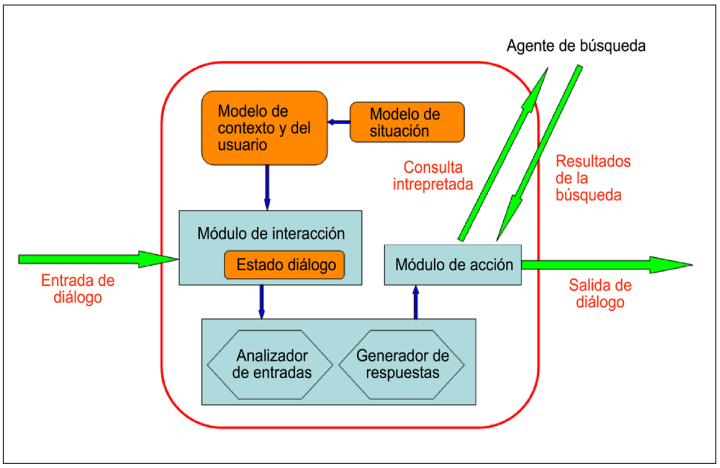

Figura 3. Modelo del Procesador de Diálogo en Lenguaje Natural.

El **modelo de contexto y del usuario** se encarga de la información relacionada con los participantes del diálogo: el 'usuario' que necesita información de la web y el 'sistema' que realiza la búsqueda. El modelo expresa el tipo de situación social, como 'búsquedas bibliográficas en la web', junto con los objetivos de los participantes: 'hallar información sobre un determinado tópico' y 'ayudar al usuario a realizar el objetivo declarado a través de la búsqueda y de la interacción dialógica'. Aquí, el modelo de usuario considera el perfil de la persona que interactúa con el sistema.

El **modelo de situación** establece las características de la situación comunicativa en la cual se lleva a cabo el diálogo. Debido a que este modelo permite interacción, las conversaciones se restringen a los requerimientos y limitaciones propios de la situación de búsquedas bibliográficas en la web. Esto implica el uso tanto de registros de estado de diálogos como de estructuras de enunciados (léxicas, sintácticas, semánticas, pragmáticas) para representar dicho modelo.

El **módulo de interacción en LN** se basa en el principio de cooperación y las máximas de conversación (de cantidad, modo, pertinencia y calidad) de Grice. Contempla estructuras de intercambio de dos posiciones, tales como pregunta-respuesta, saludo-saludo, etc. Estas estructuras de intercambio están sujetas a limitaciones de la conversación del sistema con respecto a la capacidad de la transmisión bidireccional de mensajes apropiados y comprensibles que sirvan como actos de confirmación. El **estado del diálogo** mantiene la coherencia dialógica entre el sistema y la entrada del usuario, y almacena toda la información relacionada con las interacciones entre el sistema y el usuario durante el diálogo.

El **analizador de entradas en LN** recibe la consulta del usuario y analiza la información que contiene con el fin de definir las condiciones relevantes para la generación de respuestas por parte del sistema. El resultado de este módulo es el análisis sintáctico y morfológico llevado a cabo por el sistema con el objeto de definir lo que el usuario requiere. El analizador *(parsing)* determina el tipo de acto de habla apropiado tanto para el componente de la estructura del diálogo que se está generando como para las limitaciones y condiciones declaradas. Para ello, este analizador utiliza la información de los modelos situacional y contextual.

El **generador de respuesta en LN** utiliza la información que se obtiene del agente de búsqueda y del estado del diálogo para generar un enunciado coherente acorde con la secuencia dialógica que se está llevando a cabo. El diálogo comienza con la generación del enunciado "consulta sobre cierta información solicitada por el usuario", el cual es inicialmente muy general. Luego, el sistema considera dos posibles generaciones: una consulta específica para la situación comunicativa (¿Qué tópico quiere buscar?) y una consulta general sobre el contexto de los distintos tipos de información que están disponibles en la web (¿Qué tipo de información necesita?). Las solicitudes del usuario se pueden dividir en cuatro categorías generales: petición de información, confirmación positiva o negativa, especificación de rasgos, y especificación del tópico. El analizador de discurso procesa la entrada del usuario para obtener la información que el agente de búsqueda necesita para llevar a cabo dicha acción performativa. El generador de LN puede utilizar la información obtenida —por ejemplo, referencias a documentos— para guiar el diálogo hacia la generación de LN explicativo de acuerdo con los criterios básicos identificados en el análisis de las estructuras dialógicas obtenidas en la etapa experimental de simulación dialógica hombre-computador.

Sobre la base de varias muestras obtenidas en los estudios experimentales con usuarios que realizaban búsquedas en la web, se delimitó un conjunto de criterios básicos iniciales para restringir el proceso de generación de LN. Por ejemplo, identificamos los siguientes casos, donde R representa el número de referencias que el agente de búsqueda obtiene como resultado:

- R > 100 (la consulta es demasiado general);

- R < 100 (se consideran otros temas, como la lengua en que están los documentos encontrados —las páginas están escritas en otro idioma— y el tipo de documento);
- R < 30 (adecuado para mostrar los resultados de la búsqueda).

El **generador de respuesta en LN** puede producir dos tipos de respuestas para explicar los resultados de la búsqueda: una que apunta a obtener una especificación más detallada de la consulta del usuario ('su consulta es demasiado general, ¿Podría definirla más específicamente?'), o una que requiere que el usuario proporcione algún rasgo del tópico que se busca ('Encontré demasiada información' o 'Encontré *N* referencias a documentos sobre este tópico, ¿Cuál le interesa más?'). El analizador de discurso utiliza la respuesta específica del usuario para realizar una búsqueda más refinada. El agente de búsqueda realiza la tarea de nuevo, buscando la información específica y el tópico buscado.

El **generador de respuesta en LN** produce tres tipos de enunciados descriptivos proveyendo diferentes sugerencias al usuario:

a. Precisa que los resultados de la búsqueda están en diferentes lenguas: La información está escrita en diferentes lenguas, ¿Prefiere Ud. los documentos que están en inglés?
b. Ofrece desplegar todos los documentos encontrados en la búsqueda: Encontré veinte documentos sobre el tópico, ¿Desea Ud. revisarlos todos?
c. Muestra los resultados de las búsquedas acorde a un parámetro de frecuencia: Encontré información acerca de grupos de investigación, cursos, y páginas web, ¿Está Ud. interesado en alguna información específica?

El **módulo de generación de acción** lleva a cabo la correspondiente acción de búsqueda. Para mantener la coherencia dialógica, este módulo recibe la información entregada por el analizador de la entrada del usuario y registra la respuesta que el usuario provee después de realizar la acción. El analizador de entrada procesa la respuesta del usuario y el generador de respuesta produce una salida que confirma o refuta la acción realizada (por ejemplo: '¿Encontraste lo que estabas buscando?').

Para verificar que se ha logrado el objetivo comunicativo, el analizador procesa la entrada. Si se obtiene una respuesta positiva, el sistema generará un enunciado que le dará al usuario la oportunidad de seleccionar otro tópico de conversación para una nueva búsqueda o bien le sugerirá hacer una búsqueda en profundidad sobre algún aspecto de la actual búsqueda. El proceso completo se inicia estableciendo un objetivo a partir del cual se construye una estructura completa en el nivel enunciativo. En general, los objetivos subsecuentes se dividen en funciones lingüísticas que permitan iniciar el diálogo, responder a una pregunta, preguntar por el logro de un objetivo de discurso, solicitar un nuevo tópico, etc.

2.3. Agente de búsqueda adaptativo

A diferencia de motores de búsqueda tradicionales en la www (Google, Yahoo, etc.), se diseñó un agente de búsqueda que no entrega al usuario toda la información que encuentra en la web. El agente espera hasta que se logre suficiente conocimiento acerca de la retroalimentación del usuario, los objetivos y otros datos que tienen impacto positivo para reducir la sobrecarga de información. A medida que procede la interacción, el agente refina las consultas y filtra la información inicial obtenida desde la búsqueda en la web y la retroalimentación del usuario hasta que este pueda mostrar una cantidad reducida y apropiada de información.

El agente de búsqueda se compone de tres componentes:

a. Un motor de búsqueda.

b. Un analizador de criterios, que procesa la información obtenida de acuerdo con la retroalimentación del usuario y el conocimiento del contexto actual.

c. El registro del estado de la información, que gestiona la información obtenida y el conocimiento obtenido de diálogos anteriores con el usuario.

Posteriormente, el generador de LN utiliza los resultados del agente para producir enunciados adaptativos acordes con el estado de la conversación y las limitaciones del momento. En este contexto, el término 'criterios' se refiere a la representación subyacente declarada por los documentos y en el perfil del usuario. Son similares a los criterios que se utilizan en 'recuperación de información', pero son incrementados con rasgos basados en vectores con finalidades específicas, para proporcionar la expresión requerida.

Los documentos y las consultas del usuario están representados en un espacio multidimensional, para que se puedan traducir en un patrón que representa un vector de criterios una vez que el procesador de LN los haya procesado. El agente de búsqueda utiliza métricas de distancia y algunos motores existentes para recuperar los documentos apropiados. Un documento *D* estructurado de esta manera se representa como un vector *i*: $V_d = X_0 X_1 \ldots X_n$ donde X_i expresa el valor extraído de los usuarios o los resultados de búsqueda para el *i*-ésimo criterio de los documentos que se están recuperando. Estos criterios representan información contextual relevante relacionada con páginas web que puedan ser de utilidad para entrenar los patrones y filtrar los resultados de búsqueda. Inicialmente, el criterio X_0 se relacionará con el tema principal (el tópico principal del input en LN) y el resto del vector permanecerá vacío. En la medida que avanza el diálogo y se obtienen nuevos resultados de búsqueda, se van llenando estas celdas.

Utilizando la información obtenida de los componentes del agente de búsqueda, las muestras de diálogo, y la información sobre el contexto extrajimos y sintetizamos los patrones de búsquedas más frecuentes. Estos patrones incluyen la URL de la página web seleccionada, el autor del

documento, el idioma en el cual está escrito el documento, el país de origen del documento, el tipo de documento o página (comercial, educacional, etc.), documentos relacionados con eventos, documentos técnicos, grupos de investigación o productos y servicios. Por ejemplo, la celda X_4 puede almacenar el criterio 'idioma'; si en una interacción posterior se determina que el usuario quiere documentos en español, la celda X_4 tendría el valor 'español' y se incluiría en las búsquedas posteriores. Cada criterio también contempla una ponderación que representa su contribución al documento recuperado.

Independientemente de si los criterios consisten en información del diálogo o del actual estado de la búsqueda, el agente previamente entrenado utiliza los vectores correspondientes para realizar la consulta de búsqueda en la web. Si no se puede filtrar más, el agente muestra los resultados; de lo contrario, el generador de diálogos emite nuevas solicitudes al agente. Cuando la información en los vectores y la retroalimentación del usuario son insuficientes, el agente puede utilizar un sencillo y entrenable mecanismo de inferencias bayesiano para tomar decisiones simples mediante la predicción de las acciones que se realizarán con mayor probabilidad. Es decir, dada alguna información sobre el contexto ¿cuál es la acción más probable? El resultado tiene dos consecuencias básicas: una que afecta la información que se filtra y otra que ayuda con la generación de oraciones para buscar criterios omitidos o incompletos. En la práctica, el generador de LN traduce estas acciones en limitaciones pragmáticas de alto nivel que hacen que el agente genere un determinado tipo de diálogo en LN, como una pregunta, una solicitud, retroalimentación, etc.

3. Análisis y resultados

Para investigar el impacto de la sobrecarga de información y las capacidades de la búsqueda inteligente en búsquedas web, diseñamos e implementamos un modelo de búsqueda que utiliza la tecnología de agentes inteligentes y técnicas del PLN. Los resultados que produce nuestro sistema se definen en términos de dos temas: el tipo de enunciado en LN que el sistema genera automáticamente, que sugiere que la generación de discurso basada en búsquedas puede ser plausible; y los beneficios que brinda este tipo de interacción para reducir la sobrecarga de información y el tiempo que el usuario destina a búsquedas en la web. El análisis de los resultados se basa en la generación de 1.000 muestras de estructuras dialógicas que se obtuvieron del procesamiento discursivo que llevó a cabo el sistema. El procesador de discurso produce estructuras dialógicas e interactúa con el usuario partiendo con objetivos comunicativos específicos, como se detalla a continuación (S = Salida del sistema; U = Entrada del usuario):

- *Acción*: El sistema genera una oración o consulta sobre un tópico que el usuario solicita.
  - S: ¿Qué le interesa?

- U: Necesito información básica sobre la lingüística.

- *Acción*: Acorde con los resultados de la búsqueda sobre un tópico en particular se requiere una mayor precisión del tópico.
  - S: Su consulta es demasiada amplia, ¿Podría hacerla más específica?
  - U: Bien.

- *Acción*: Resultados de búsqueda y generación de diálogos dependientes del contexto (el enunciado generado es completamente adaptativo, por lo cual variará entre interacciones).
  - S: La información obtenida está escrita en distintos idiomas, ¿Prefiere información en español?
  - U: Me parece bien
  - S: Hay 20 referencias distintas sobre este tópico, ¿Quiere verlas todas?
  - U: Sí
  - S: Hallé información sobre grupos de investigación, cursos, etc., ¿Cuál de estos le interesa?
  - U: Grupos de investigación

Desde el punto de vista del filtrado, se analizó el comportamiento del sistema para evaluar el número de turnos conversacionales que se necesita para obtener requisitos más precisos e información filtrada, frente al número de referencias o documentos que satisfagan dicho requerimiento. Inicialmente, el conjunto de posibles documentos candidatos incluía más de 30.000 referencias, pero para el presente estudio se redujo el número a menos de 1.000 referencias. Tal como se indica en la Figura 4, realizamos tres experimentos: uno centrado en información sobre *Java*, otro en *Cartoons* y el último en *Computer Networks*. En los tres casos, cada interacción consistió en uno o más diálogos (intercambios) entre el usuario y el sistema.

Como se observa en la Figura 4(a), las interacciones dialógicas sobre el tópico *Java* muestran un incremento en el número de documentos relevantes hallados después de tres intercambios. Esto ocurre debido a que tanto el contexto como el tipo de preguntas cambia según la situación y el contenido de los documentos. El mismo número de interacciones produce resultados diferentes porque el tipo de documento que se buscaba cambió en la medida que se restringieron otros rasgos. De modo similar, cuando en el diálogo se establece una limitación sobre la lengua en que están los documentos, muchas de las referencias a documentos dejan de corresponder.

En la Figura 4(b) se puede revisar que los resultados fueron similares en el segundo experimento. Se observa aumentos repentinos incluso en diálogos con tres intercambios, incrementando de

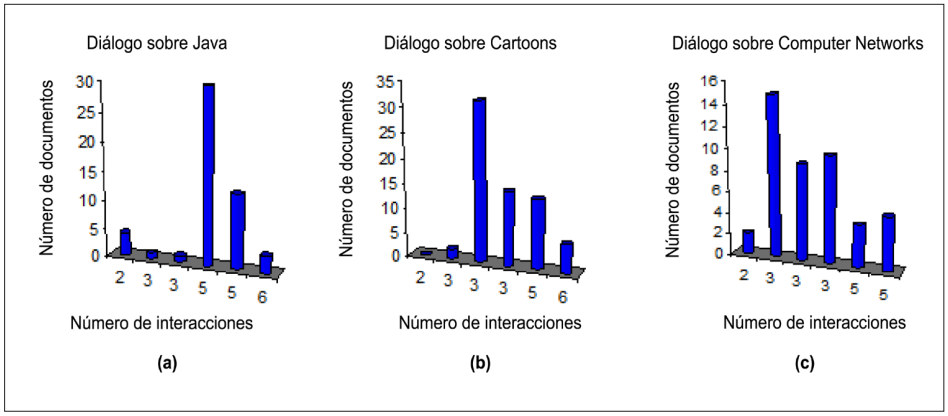

Figura 4. Resultados de búsqueda interactiva basada en Diálogos sobre *Java* (a), *Cartoons* (b) y *Computer Networks* (c).

uno a casi 35 referencias. Este cambio se produjo porque el agente hizo una inferencia y el usuario aplicó una restricción relacionada con la naturaleza del documento. Para las interacciones sobre *Computer Networks*, la Figura 4(c) muestra un número de documentos relevantes recuperados mayor que para los experimentos sobre *Java*. Aún así, el filtrado al término del diálogo es mejor que para las evaluaciones previas: el máximo número de documentos obtenidos es menor que para las interacciones del primer experimento y la reducción en los documentos filtrados es mejor para diálogos de 3 interacciones.

De todos los experimentos, puede observarse que existe una reducción significativa en los resultados obtenidos con un número mínimo de turnos conversacionales debido a las restricciones en la naturaleza de la información que se entrega finalmente. En general, el promedio de interacciones y el resultado del proceso de filtrado requeridos para satisfacer cada objetivo comunicativo en los experimentos se muestra en la Tabla 1.

Tabla 1. Resultados generales del filtrado basado en el Modelo de Diálogo.

Tópico	Promedio		
	Interacciones	Documentos	Reducción (%)
Java	3,6	9,8	72,0
Cartoons	3,7	11,5	67,2
Computer Networks	3,5	7,5	78,5
Promedio	3,6	9,6	72,6

En la Tabla 1 se muestra un filtrado de casi 73% con un número promedio de 3.6 interacciones dialógicas. Es decir, en promedio, menos del 27% de los documentos inicialmente recuperados se despliegan realmente al usuario. Es importante también destacar que los experimentos no requirieron más de cuatro turnos conversacionales para filtrar al menos 67.2% de los documentos. Estos tres experimentos revelan un decremento significativo en los resultados de búsquedas a partir de un número reducido de turnos conversacionales debido a los deslindes de la naturaleza de la información que finalmente se proporcionó.

CONCLUSIONES

A manera de conclusión, en este estudio, hemos presentado un enfoque de búsqueda en la web para disminuir los problemas de la sobrecarga y filtrado de información tomando en cuenta las interacciones y la retroalimentación entre el usuario web y un sistema de búsqueda. Nuestro sistema de filtrado de documentos web combina métodos del PLN con la tecnología de agentes inteligentes para abordar el problema de la sobrecarga de información.

Los experimentos diseñados para observar la retroalimentación del usuario y las capacidades de inferencia del agente de búsqueda revelan que proporcionar criterios de exploración para la información recuperada, según su importancia o grado de uso, puede ahorrar tiempo y mejorar la precisión en las búsquedas bibliográficas. Sin embargo, se deben tomar en cuenta las contribuciones del usuario en las decisiones que toma el sistema. Pese a la complejidad moderada de los experimentos y a las limitaciones de su diseño, los temas identificados en nuestro estudio no deberían cambiar drásticamente en el caso de requerimientos e implementaciones más avanzados (por ejemplo, distintos idiomas, distintas capacidades de búsqueda, etc.).

Desde el punto de vista del lenguaje, el modelo de interacción dialógica se presenta prometedor como estrategia para enfrentar requisitos de búsqueda de información más específicos, en los cuales tanto el diseño como la implementación de un sistema de generación de LN pueden adaptarse fácilmente a diversas situaciones comunicativas. Aunque los sistemas de generación de LN son relativamente conocidos, no lo es el hecho de integrar en un mismo enfoque dos tecnologías con el fin de abordar los problemas de búsqueda y filtrado en la web.

Al contrario de otros enfoques que emplean el PLN para enfrentar problemas similares, nuestro modelo identifica los intereses y objetivos del usuario web en la medida que avanza el diálogo. Otros enfoques utilizan distintos parámetros para establecer de antemano información como el grado de interés o la relevancia. En algunos modelos de búsqueda que emplean el PLN, la suposición de trabajo subyacente es que un agente inteligente debe analizar un perfil del usuario para extraer parte de su 'historia' y así llegar a una respuesta final. Debido a que nuestro modelo utiliza documentos reales de la web, no puede filtrar el conjunto total de documentos

relevantes; en vez de esto, utiliza diálogos interactivos para obtener información relevante del usuario para luego filtrar las referencias precisas.

REFERENCIAS BIBLIOGRÁFICAS

Ardissono, L., Boella, G. & Lesmo, L. (2000). Plan based agent architecture for interpreting natural language dialogue. *International Journal of Human-Computer Studies*, 52, pp.583-636.

Benamara, F. & Saint Dizier, P. (2003). Webcoop: A cooperative question-answering system on the web. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*. European Chapter Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 63-66 [en línea]. Disponible en: DOI= http://dx.doi.org/10.3115/1067737.1067749

Bloedorn, E. & Mani, I. (1998). Using NLP for machine learning of user profiles, *Intelligent Data Analysis*, *3*(2), 3-18.

Cohen, P. & Levesque, H. (1990). Performatives in a rationally based speech act theory. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*. Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 79-88 [en línea]. Disponible en: DOI= http://dx.doi.org/10.3115/981823.981834

Holscher, C. & Strube, G. (2000). Web search behavior of Internet experts and newbies. En *Proceedings of the 9th International World Wide Web Conference*. Amsterdam: North-Holland Publishing Co. (pp. 337-346).

Jansen, B. & Spink, A. (2000). Real life, real users and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, *36*(2), 207-227.

Jansen, B. & Pooch, U. (2000). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*, *53*(2), 235-246.

Jurafsky, D. & Martin, J. (2000). *An introduction to natural language processing, computational linguistics and speech processing*, Prentice Hall.

Landauer, T., McNamara D., Dennis, S. & Kintsch, W. (2007). *Handbook of latent semantic analysis*, Hillsdale, NJ: Erlbaum.

Lau, T. & Horvitz, E. (1999). Patterns of search: Analyzing and modeling web query refinement. En *Proceedings of the Seventh international Conference on User Modeling*. J. Kay, Ed. Springer-Verlag New York, Secaucus, NJ, 119-128.

Levy, A. & Weld, D. (2000). Intelligent internet systems. *Artificial Intelligence*, *11*(8), 1-14.

Fischer, G. & Stevens, C. (1991). Information access in complex, poorly structured information spa-

ces. En *Proceedings Eighth ACM Conference on Human Factors in Computing Systems*. New Orleans, Louisiana, pp. 63-70.

Moore, J. (1994). *Participating in explanatory dialogues: Interpreting and responding to questions in context*, MIT Press, Cambridge, MA.

Moore, J., Foster, M., Lemon, O. & White, M. (2004). Generating tailored comparative descriptions in spoken dialogue. FLAIRS Conference, Florida, USA.

Reitter, E. & Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

Reitter, E. & Moore, J. (2007). Predicting success in dialogue. En *Proceedings of 44th Annual Meeting of the Association for Computational Linguistics*, *1*(45), 808-815.

Reitter, E., Moore, J. & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. En *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 685-690.

Salter, J. & Antonopoulos, N. (2006). Cinema screen recommender agent: Combining collaborative and content-based filtering. IEEE *Intelligent Systems and Their Applications*, *21*(1), 35-41.

Tong, L., Changjie, T. & Jie, Z. (2001). Web document filtering technique based on natural language understanding, *Int'l J. Computer Processing of Oriental Languages*, *14*(3), 279-291.

Zukerman, I., Albrecht, D., Nicholson, A. & Doktor, K. (2000). Trading off granularity against complexity in predictive models for complex domains. *Proc. 6th Int'l Pacific Rim Conf. Artificial Intelligence*, Springer-Verlag, pp. 1274-1279.