Componente Basado en FAQ de Sistema Pregunta-Respuesta para la Interaccion de Ciudadanos con el Estado Colombiano

Adán Beltran Gomez
Universidad Manuela Beltran, Bogota, Colombia.
adan.beltran@docentes.umb.edu.co

Santiago Camargo Marín
Universidad Manuela Beltran,Bogota, Colombia.
santiago.camargo@umb.edu.co

Wilson Ferney Lemus
Universidad Manuela Beltran, Bogota, Colombia
wilson.lemus@umb.edu.co

ABSTRACT

Question- Answer Systems are systems that improve access to large amount of textual information on the Web because they use the natural language interface and give a concrete answer and no a list of relevant documents. These systems make use of different techniques ranging from traditional Information Retrieval through Semantic Web elements , and general techniques in natural language processing and text mining . Research results related to the design and construction of the recovery module FAQ (Frequently Asked Questions) associated with SQA in the context of e- government Colombian Ubaté Mayor , Cundinamarca are presented. Techniques , data sources used , proposed architecture , followed methodology and test plan used in constructing the FAQ recovery module on a specific domain SQA are detailed.

Keywords: Electronic Government, Question Answering Systems, Natural Language Processing

RESUMEN

Los Sistemas Pregunta-Respuesta (SQA- por sus siglas en inglés, System *Question-Answering*) son sistemas que permiten mejorar el acceso a la gran cantidad de información textual en Internet ya que utilizan el lenguaje natural como interfaz y dan una respuesta concreta y no un listado de documentos relevantes. Estos sistemas hacen uso de diferentes técnicas que van desde las tradicionales en Recuperación de la Información hasta elementos de Web Semántica, y en general técnicas del Procesamiento del Lenguaje natural y minería de texto. Se presentan los resultados de investigación relacionados con el diseño y construcción del módulo de recuperación de FAQ (Preguntas Frecuentes) relacionado con un SQA en el contexto de Gobierno electrónico colombiano en la Alcaldía de Ubaté, Cundinamarca. Se detallan las técnicas, recursos de datos utilizados, arquitectura propuesta, metodologías seguidas y el plan de pruebas utilizado en la construcción del módulo de recuperación de FAQ en un SQA de dominio específico.

Palabras claves: Gobierno Electrónico, Sistemas Pregunta Respuesta, Procesamiento Lenguaje Natural

1. INTRODUCTION

La tendencia actual de los sistemas de información empresariales es la interoperabilidad entre sus aplicaciones, clientes y proveedores. Y apoyando este enfoque tecnologías como SOAP, SOA , Web Services y otras se han

convertido en estándares apoyados por la industria, pero a pesar de estos esfuerzos aun la interoperabilidad está limitada al conocimiento previo de estos servicios por parte del personal técnico y otros aspectos. El consumo de estos servicios actualmente no se hace en forma transparente, es decir se necesita tener conocimiento previo para que la interacción sea posible.
Adicionalmente, en la mayoría de los recursos Web se codifica la información pero no el conocimiento. En el caso de las búsquedas en la Web es el ser humano quien decide si los resultados están relacionados a la temática buscada, consumiendo recursos y tiempo considerables, ya que las aplicaciones entre si no tienen la capacidad de "hablar" en forma automática y semántica para su entendimiento.

Por otro lado, en los sitios web existe información especializada que obliga a que su recuperación no sea fácil, usualmente en sitios web con alto contenido de información en forma textual, exigiendo que el usuario indague profundamente para encontrar la respuesta deseada, generando que las personas desistan de la búsqueda y escojan otro método como por ejemplo llamar telefónicamente a un Call Center o servicio al cliente o preguntar a otras personas, llevando a que los sitios web asignen recursos para una atención personalizada.
Como medida para resolver esta problemática muchos sitios web han dispuesto de una sección de respuestas a preguntas frecuentes, (FAQ: Frequently Asked Questions, por sus siglas en ingles), pero los usuarios deben leer en forma secuencial las preguntas hasta encontrar la pregunta que ellos tratan de formular, en algunos sitios web el listado de las FAQs son tan extensos que los usuarios ni siquiera intentan leerlas

Diferentes disciplinas han propuesto mecanismos para mejorar o ayudar a encontrar información relevante al usuario, tales como Recuperación de la Información (IR-*Information Retrieval* por sus siglas en inglés) que han propuesto las bases teóricas de los motores de búsqueda actuales, Extracción de información (IE *Extraction Information*, por sus siglas en inglés) que proponen sistemas que extraen el conocimiento que están en los textos y Web Semántica que propone toda una arquitectura basada en estándares que facilita la interacción entre aplicaciones en forma automática a través de etiquetas semánticas entre los datos. Dentro de los sistemas de Recuperación de Información están los sistemas de pregunta-respuesta (QA Question-Answer) que buscan devolverle al usuario una respuesta concreta a su pregunta y no un listado de documentos relevantes. Estos sistemas al estar desarrollados con una interfaz en lenguaje natural generalmente tienen una estructura de complejidad algorítmica que involucra la utilización de técnicas de inteligencia artificial (IA), Aprendizaje Automático (*ML-Machine Learning*, en inglés) y procesamiento del lenguaje natural (PLN, *Processing Language Natural*, en inglés), entre otros.

Aunque los SQA han evolucionado considerablemente desde sus inicios en la década de los 60 con sistemas como Eliza [1], Baseball[2] y Lunar[3] entre otros, su precisión en la generación de respuesta es aun baja y no es igual para todos los idiomas, por ejemplo para el inglés más del 77% de las preguntas se contestan correctamente según el análisis realizado por Voorhees [4] basado en los datos de la Conferencia TREC-2004 , pero para el español está en 55% según el análisis de los datos de la conferencia CLEF 2006 [5], lo anterior a pesar de que el español es el tercer idioma con más presencia en la web y el segundo idioma utilizado para consultas, lo anterior debido en parte a los pocos corpus disponibles para este idioma.

En el presente artículo se muestran los resultados de investigación en el desarrollo de un SQA para el contexto del Gobierno en línea colombiano, en la primera parte se muestran los conceptos generales de los SQA, en la segunda parte los resultados de investigación en el diseño de un SQA de dominio específico para la interacción del ciudadano con el Estado Colombiano; se detallan las metodologías utilizadas para las diferentes tareas, la arquitectura utilizada, las técnicas utilizadas, recursos, plan de pruebas y finalmente se presentan las conclusiones

2. SISTEMAS PREGUNTA-RESPUESTA

Los SQA, son un tipo Sistema de Recuperación de Información, que procesan una pregunta o consulta en lenguaje natural y retornan o extraen una respuesta de fuentes estructuradas (bases de datos) o desestructuradas (textos). A diferencia de los sistemas de Recuperación de Información la consulta se hace en lenguaje natural y no en

palabras claves, lo que implica que el sistema debe reconocer el tipo de respuesta que el usuario espera, para dar la respuesta concreta o un párrafo donde se encuentre. En general, la complejidad de estos sistemas consiste en establecer la relación implícita entre la pregunta y la respuesta, existen aproximaciones que los consideran como sistemas en el camino hacia la Web semántica [6].

A igual que otros sistemas que intentan la extracción y análisis del conocimiento semántico de los textos, los sistemas SQA tienen diferentes componentes que requieren de las habilidades de lingüistas, estadísticas, ciencias de la computación, expertos del dominio, entre otros. Lo anterior al estado actual en que se encuentra la extracción del conocimiento semántico en forma automática, enfoques como los de minería de texto, aprendizaje máquina, procesamiento del lenguaje natural, Recuperación de la Información e inteligencia artificial han dado las bases para el desarrollo de estos sistemas.

Existen básicamente dos enfoques en el desarrollo de SBR, el semántico y el estadístico. Desde el punto de vista semántico Sowa [7] propone tres niveles del conocimiento semántico que se puede obtener: Fuerte, donde la información esta etiquetada utilizando lógica formal permitiendo gran capacidad de razonamiento, Medio la información tiene notación semi-formal permitiendo cierto nivel de razonamiento y Ligero donde existen etiquetas de clasificación y restricciones pero no es posible hacer razonamiento. Los sistemas pueden estar en uno o combinar varios de estos niveles en sus componentes.

Los SQA, utilizan diferentes recursos como taxonomías, Archivos de descripción de recursos (RDF), Ontologías, Tesauros, archivos de Preguntas Frecuentes, acceso a Bases de Datos y demás. Otros enfoques utilizan solamente el aprendizaje automático a partir de ejemplos etiquetados para las diferentes categorías de las preguntas que se pueden clasificar y contestar.

Dadas las potencialidades de aplicación de SQA, en las últimas décadas se han desarrollado diferentes propuestas tanto de sistemas de dominio abierto [8] , como de dominio específico: educación [9, 10] , tutores virtuales [11, 12], legal [13], turismo[14] , tratamiento de enfermedades [15], acceso a Bodegas de Datos [16], medico[17, 18], entre otros. A pesar de los avances y sofisticación de estos sistemas propuestos tales como el sistema Watson de IBM, aun presentan inconsistencias en el proceso de validación de respuestas para cierto tipo de preguntas [19].

2.1 SISTEMAS SQA BASADOS EN FAQ

Uno de los enfoques que se han utilizado últimamente es la utilización de los archivos de preguntas frecuentes (FAQ por sus siglas en inglés- Frecuency Answering Questions) donde se han utilizado diferentes elementos de los sistemas basados en IR donde sus documentos de extracción de respuestas son los archivos FAQ que muchas páginas tienen.

Dentro de los inconvenientes de estos sistemas es la construcción de la base de datos de preguntas frecuentas en forma manual que generalmente requieren muchos recursos, por lo que se han propuesto sistemas para generar dichos archivos en forma automática como en [20] con la utilización de técnicas de agrupamiento o clustering.

En [21] se muestra una metodología para desarrollar estos sistemas basados en FAQ en dominios cerrados donde se alcanzan índices de precisión hasta del 90%.

2.1.1 CRITERIOS DE SIMILITUD ENTRE PREGUNTAS

El metodo mas utilizado para la busqueda de la pregunta dentro del FAQ es el Modelo Vectorial propuesto por G. Salton[22], donde cada documento es representado por un vector de términos, las consultas formuladas en lenguaje natural son representadas como un vector de términos. Este método es la base de muchos sistemas de

recuperación. Existen diferentes propuestas de criterios de similitud dentro del modelo vectorial, en general es fácil aplicar alguna función de similitud que estime la semejanza entre el vector de la consulta y el de cada uno de los documentos.

Similitud basada en la relación Coseno. Desde el punto de vista geométrico, si ambos vectores, consulta y documento están próximos, es factible asumir que el documento es similar a la consulta. El modelo vectorial plantea medir la similitud de un documento d_j y una consulta q en base a la proximidad de sus vectores en base al coseno del ángulo que tales vectores forman.

$$sim\ (d_i, q) = \frac{\cos\theta = \vec{d_i} \cdot \vec{q}}{\|\vec{d_i}\| \times \|\vec{q}\|} = \frac{\sum_{k=1}^{n} w_{ki} \times w_{kq}}{\sqrt{\sum_{k=1}^{t} w^2{}_{ki}} \times \sqrt{\sum_{k=1}^{t} w^2{}_{kq}}} \quad (1)$$

En este modelo [23], a diferencia del modelo booleano, no se limita a comprobar si los términos especificados están o no , sino que permite la existencia de correspondencias parciales y el grado de similitud o relevancia conforme a la cual los documentos pueden ser devueltos de mayor a menor relevancia.

Frecuencia en Documentos: antes de calcular el grado de similitud entre los vectores, es necesario calcular los pesos de los términos. Dichos pesos pueden ser calculados de múltiples maneras, si bien el esquema de pesos *TF–IDF* y sus derivados se han convertido en los más populares.

En el esquema *TF–IDF* básico el peso w_{ij} de un término *i* en un documento *j* viene dado por la formula: $w_{ij} = tf_{ij} \times idf_{i}$, donde *tf* es la frecuencia del término *t* en el documento y $idf = log(N/n_i)$, donde *N* es el total de documentos y n_i es el número de documentos donde aparece el termino *t*

A la serie de fórmulas para el computo de pesos formada por este esquema inicial y las variantes a las que da lugar se las denomina esquemas *TF–IDF*.
Los buenos resultados obtenidos con el modelo vectorial, unidos a la simplicidad a nivel de concepto e implementación, su bondad a la hora de aceptar consultas en lenguaje natural, y su capacidad para permitir correspondencias parciales y ordenamiento por relevancia, han hecho de este modelo una de las principales bases sobre la que se han desarrollado gran parte de los experimentos y sistemas en todo el ámbito de la Recuperación de Información.
Sus buenas características, unidas al hecho de que sea uno de los modelos de representación más utilizados, le han convertido frecuentemente en el sistema de referencia respecto al cual comparar resultados a la hora de desarrollar nuevos modelos de recuperación.
Adicionalmente existente otras medidas de similitud en el modelo vectorial, que se utilizan en menor medida:[24]

$$\textit{Producto escalar:} \quad \sum_{j=1}^{t} w_{dij} \cdot w_{qj}$$

$$\textit{Coeficiente de Dice:} \quad \frac{2 \cdot \sum_{j=1}^{t} w_{dij} \cdot w_{qj}}{\sum_{j=1}^{t} w^2{}_{dij} + \sum_{j=1}^{t} w^2{}_{qj}}$$

$$\textit{Coeficiente de Jaccard:} \quad \frac{\sum_{j=1}^{t} w_{dij} \cdot w_{qj}}{\sum_{j=1}^{t} w^2{}_{dij} + \sum_{j=1}^{t} w^2{}_{qj} - \sum_{j=1}^{t} w_{dij} \cdot w_{qj}} \quad (2)$$

3. RESULTADOS

A continuación se presentan los resultados más relevantes en cuanto al diseño del SQA de dominio específico para las entidades del Estado Colombiano donde se toma como caso de estudio la Alcaldía de Ubaté, se muestran los componentes de la arquitectura general y posteriormente se detallan los pormenores del componente de recuperación de FAQ, que es el objetivo del presente artículo.

3.1 ARQUITECTURA PROPUESTA

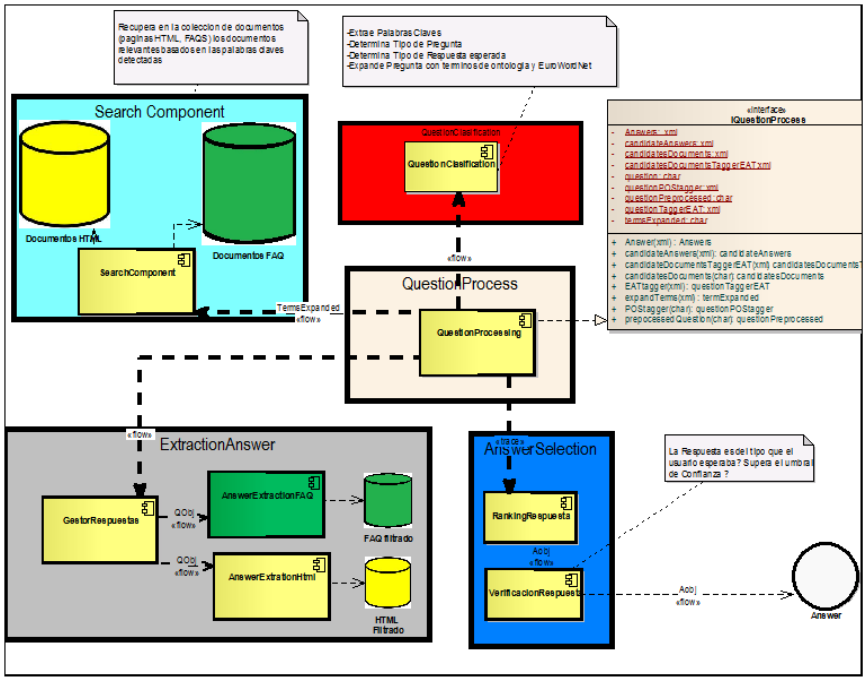

Figura. 1. Arquitectura del Framework

En la figura 1 se describen los componentes que interactúan en la arquitectura del Framework que permiten procesar la pregunta y resolverla con la mejor opción.

- Question Processing. Gestiona los procesos de las preguntas relacionándolas con los respectivos componentes. Ingresa la respectiva pregunta y devuelve un objeto con el texto etiquetado.
- Question Clasification. Entra el objeto etiquetado, determina las palabras claves, el tipo de pregunta, determina tipo de respuesta esperada y expande la pregunta con términos de ontología y EuroWordNet, devolviendo documento XML con las preguntas expandidas.
- Component Search. Entra documento XML con las preguntas pre procesadas y expandidas, busca en las bases de datos de las preguntas frecuentes y documentos HTML, las respectivas respuestas a las preguntas. Retorna documento XML con las preguntas expandidas y sus respectivas respuestas.
- Extraction Answer. Entra objeto de la pregunta, el componente AnswerExtractionFAQ extrae la respuesta de la base de datos de FAQ filtrados, a sus ves el componente AnswerExtractionHTML extrae la respuesta de la base de datos de documentos HTML filtrados. Se retorna el documento XML con las respuestas a la pregunta.
- Answer Selection. Recibe los documentos XML procesados en los componentes anteriores, mediante algoritmos de ranking extrae la mejor respuesta del documento, se verifica si la respuesta seleccionada corresponde a la pregunta ingresada inicial mente, retorna objeto con la respuesta.

3.2 COMPONENTE FAQ

El sistema que realiza la búsqueda de la similitud de preguntas se representa en la Figura 2. Aquí encontramos el proceso de validación de las preguntas que realiza el usuario.

La pregunta ingresa el sistema y es dirigida al Question Process el cual la envía al componente Search Componenet, donde realiza el proceso de busque de preguntas relevantes en el repositorio de FAQ's y a estas FAQ's relevantes encontradas se les realiza el cálculo de la similitud coseno, la(s) FAQ(s) que superó el lumbral de precisión es enviada al Question Process. Luego la FAQ es enviada al Selection Answers para un verificación de respuesta esperada y así retornar la validación de la FAQ

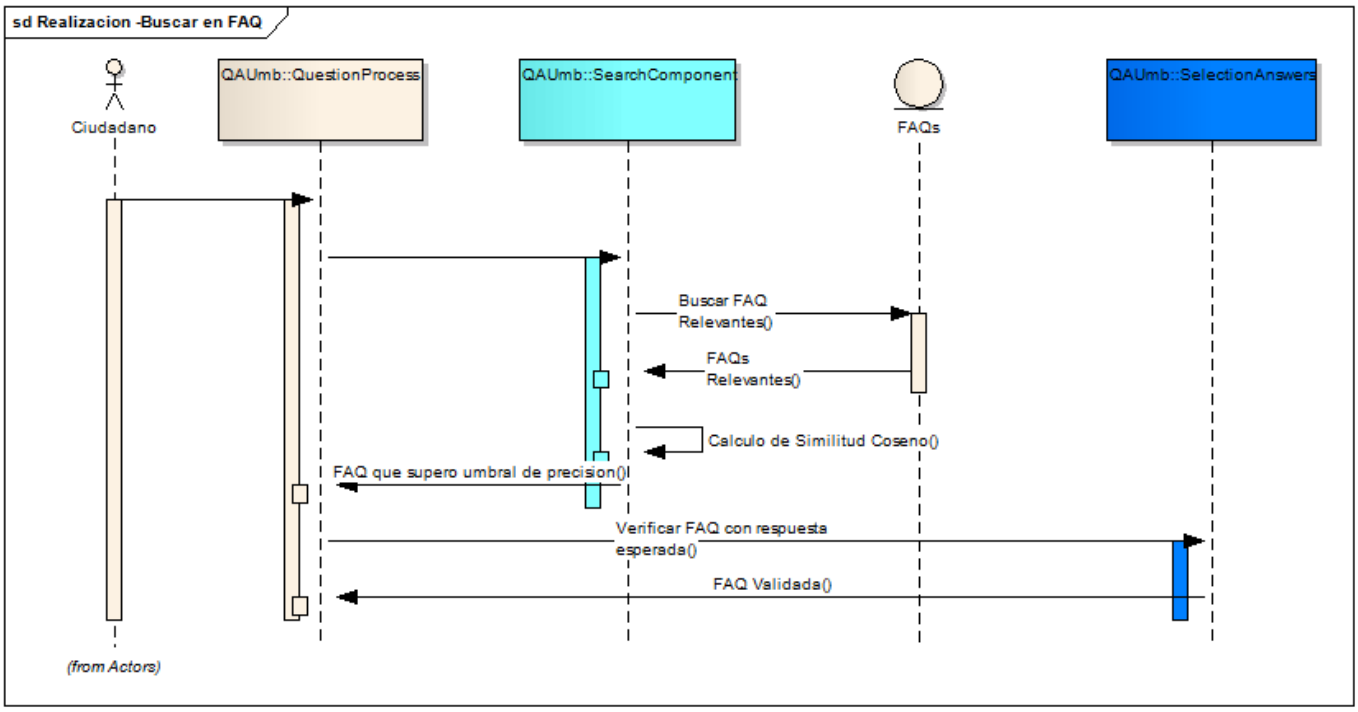

Figura. 2. Diseño del componente de recuperación de FAQ

RECOPILACION DE CORPUS DE FAQ

El corpus es recolectado de forma manual, mediante el ingreso a las distintas paginas web gubernamentales, allí se recopila una a unas las preguntas frecuentes que puede realizar el ciudadano. Ademas de realizar un trabajo de campo con los ciudadanos del municipio de Ubate – Cundinamarca, en donde se realiza un recolección mas

exacta puesto que son preguntas hechas por los mismos ciudadanos, esto nos permite tener un corpus mas robusto y mas exacto para su respectiva comparacion al momento ejecutar el sistema FAQ.

Despues de esta recopilación se le realiza un analisis en donde se le asigana un Id unico, el texto que resuelve la pregunta, taxonomia de la pregunta, tematica de la pregunta, taxonomia de la respuesta y subtematica de la pregunta.

3.3 PRUEBAS

Se creó el archivo CorpusPruebas.XML, como colección para probar el sistema. Este archivo fue desarrollado a partir de las preguntas recolectadas en Ubaté, donde se etiquetaron los valores esperados, el tema, la intención, el IndiceFAQ que podría contestar dicha pregunta, urls donde podría contestar la pregunta, entre otros, en la figura 3 se muestra parcialmente el archivo generado para probar el sistema.

Figura 3. Archivo CorpusPruebas.XML utilizado para probar el componente FAQ

El procedimiento para probar el sistema en forma automatica se hizo a travez de un programa en Python que recorría las 1400 preguntas del archivo de CorpusPruebas.XML enviándolas al componente principal del del sistema *QuestionProcess*, según la arquitectura detallada en la primera parte del articulo.

En la Tabla 1 se muestran parcial los resultados de los casos de pruebas, en general se obtuvo una precisión superior al 70%.

Tabla 1: Muestra parcial de un caso de prueba realizado al componente FAQ

PREGUNTA	ID FAQ CORRECTO	ID FAQ RECUPERADO	TEST
Como hago los aportes de un empleado con novedad de licencia no remunerada	2	2	Passed
Los juegos de bingo necesitan un mínimo de operación	6	6	Passed
Cuanto tiempo tiene de vigencia un juego localizado	7	7	Passed
Como hago tramites en línea	47	2326	Fail
Que Alternativa se tomó frecuente al paro agrario en Ubate	0	0	Passed
Que es un ministerio	70	760	Fail
Que es la Contraloría	80	1578	Fail

4. CONCLUSIONES

Se ha mostrado el proceso de investigación y desarrollo del componente FAQ de un SQA en un dominio o contexto especifico como es el gobierno electrónico . Debido a la gran cantidad de sectores que un gobierno atiende, la especificidad es poca debido a que no se puede seleccionar solo un sector, por lo que el término "dominio especifico", es discutible en este contexto.

El porcentaje de precision alcanzado por el sistema es comparable con sistemas similares que muestran la literatura. Uno de los detalles por mejorar de dicho componente es el desempeño ya que en la medida que el Corpus de FAQ aumentaba considerablemente el tiempo para recuperar la pregunta correctamente aumentaba considerablemente.

5. AGRADECIMIENTOS

Proyecto de Investigación co-financiada por Colciencias, Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia (MinTic) y el Consorcio Universidad Manuela Beltrán, Multimedia Service ltda e Innova&IP Ltda - Proyecto No. 1263-595-37045.

REFERENCIAS

[1] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, pp. 36-45, 1966.
[2] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery, "Baseball: an automatic question-answerer," in *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, 1961, pp. 219-224.
[3] W. A. Woods, "Progress in natural language understanding: An application to lunar geology," in *Proceedings of the June 4-8, 1973, national computer conference and exposition*, 1973, pp. 441-450.
[4] E. M. Voorhees, "Overview of the TREC 2004 question answering track," ed, 2004, pp. 52–62.
[5] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, *et al.*, "Overview of the CLEF 2006 multilingual question answering track," in *Evaluation of Multilingual and Multi-modal Information Retrieval*, ed: Springer, 2007, pp. 223-256.
[6] M. Konopík and O. Rohlík, "Question answering for not yet semantic web," vol. 6231 LNAI, ed. Brno, 2010, pp. 125-132.
[7] J. F. Sowa, "Future Directions for Semantic Systems," ed, 2011.
[8] B. Magnini, "Open Domain Question Answering: Techniques, Systems and Evaluation," in *Tutorial of the Conference on Recent Advances in Natural Language Processing-RANLP. Borovetz, Bulgaria*, 2005.
[9] W. W. Z. L. S. Wang, "Preference Survey and System Design on Mobile Q&A for Junior High School Students," ed, 2012.
[10] D. Wen, J. Cuzzola, L. Brown, and Kinshuk, "Exploiting semantic roles for asynchronous question answering in an educational setting," vol. 7310 LNAI, ed. Toronto, ON, 2012, pp. 374-379.
[11] A. M. Olney, A. C. Graesser, and N. K. Person, "Tutorial Dialog in Natural Language," ed, 2010.
[12] A. Tewari, I. Liu, C. Cai, and J. Canny, "An analysis of the dialogic complexities in designing a question/answering based conversational agent for preschoolers," vol. 7502 LNAI, ed. Santa Cruz, CA, 2012, pp. 36-45.
[13] T. Hughes, C. Hughes, and A. Lazar, "Epistemic structured representation for legal transcript analysis," in *Advances in Computer and Information Sciences and Engineering*, ed: Springer, 2008, pp. 101-107.
[14] Y. Todorova and M. Gelfond, "Toward question answering in travel domains," vol. 7265, E. Esra, L. Joohyung, L. Yuliya, and P. David, Eds., ed, 2012, pp. 311-326.
[15] C. Pechsiri, S. Painuall, and U. Janviriyasopak, "Medicinal Property Knowledge Extraction from Herbal Documents for Supporting Question Answering System," in *New Frontiers in Applied Data Mining*. vol. 7104, L. Cao, J. Huang, J. Bailey, Y. Koh, and J. Luo, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 431-443.
[16] N. Kuchmann-Beauger and M. A. Aufaure, "A natural language interface for data warehouse question answering," vol. 6716 LNCS, ed. Alicante, 2011, pp. 201-208.
[17] W.-H. Lu, C.-M. Tung, and C.-W. Lin, "Question Intention Analysis and Entropy-Based Paragraph Extraction for Medical Question Answering," in *6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore*. vol. 31, C. T. Lim and J. C. H. Goh, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 1582-1586.
[18] C. Yong-Gang, J. Ely, and Y. Hong, "Evaluating the weighted-keyword model to improve clinical question answering," in *Bioinformatics and Biomedicine Workshop, 2009. BIBMW 2009. IEEE International Conference on*, 2009, pp. 331-335.
[19] D. Zhang, "Inconsistency in Multi-Agent Systems," ed, 2012.
[20] M. Jintao and Z. Jian, "FAQ Auto Constructing Based on Clustering," in *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*, 2012, pp. 468-472.
[21] E. Sneiders, "Automated FAQ answering with question-specific knowledge representation for web self-service," in *Human System Interactions, 2009. HSI '09. 2nd Conference on*, 2009, pp. 298-305.
[22] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 613-620, 1975.
[23] C. W. Cleverdon, J. Mills, and M. Keen, "Factors determining the performance of indexing systems," 1966.

[24] R. Peña, R. Baeza-Yates, and J. V. R. Muñoz, *Gestión digital de la información: de bits a bibliotecas digitales y la web*: Ra-Ma, 2002.