Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 111-117 recibido 29-04-2005; aceptado 01-06-2005

Un entorno para el desarrollo y la evaluación de un sistema de búsqueda de respuestas en euskera

Olatz Ansa, Xabier Arregi, Itsaso Esparza, Andoni Valverde
Lengoaia eta Sistema Informatikoak
UPV/EHU
649 p.k., 20080 Donostia
{jipanoso, xarregi, jibeslei, a.valverde}@si.ehu.es

Resumen: En este artículo se presenta una plataforma de desarrollo y evaluación de un sistema de búsqueda de respuestas (BR), para preguntas escritas en euskera y sobre una colección de documentos en euskera. La plataforma integra tres componentes: el propio sistema de BR, el componente léxico y el entorno de evaluación.
El sistema de BR se ha ideado atendiendo a los requisitos de evaluación. Se ha optado por una arquitectura distribuida, que integra servicios web autónomos.
Se describe una experiencia piloto, que ha permitido validar la adecuación del entorno y que sienta las bases para el trabajo futuro.
Palabras clave: Sistema de búsqueda de respuestas.

Abstract: This paper presents a development and evaluation platform for a question answering (QA) system. It is a monolingual system, which takes as input Basque questions and searches for answers in a Basque document collection. Three components integrate the platform: the QA system, the lexical component and the evaluation environment.
The QA system has been designed taking into account the evaluation requirements. A distributed architecture with stand-alone web services has been adopted.
A pilot experiment is also described and this experiment has allowed ratifying the environment adaptation and has laid the foundations for future work.
Keywords: Question answering system.

1 Introducción

Los sistemas de búsqueda de respuestas en colecciones de documentos abordan la tarea de la obtención de respuestas puntuales y concretas a preguntas formuladas en lenguaje natural.

Dichos sistemas integran, entre otras, técnicas de recuperación de información (IR) y de procesamiento de lenguaje (NLP), con el fin de *entender* las preguntas formuladas y generar las respuestas adecuadamente. Ello conlleva, por una parte, la necesidad de reutilizar y adaptar recursos, técnicas y herramientas de IR y NLP. Por otra parte, abre la posibilidad de evaluar el comportamiento de dichas herramientas en una aplicación real.

En este artículo se presenta una plataforma de desarrollo y evaluación de un sistema de búsqueda de respuestas.

No se conciben el desarrollo y la evaluación del sistema como tareas independientes. Al contrario, se trata de que la arquitectura de la plataforma soporte la evaluación del sistema. El objetivo es que la aplicación de técnicas alternativas o la utilización de nuevos recursos y/o herramientas, sea fácilmente evaluable, y a su vez, que el entorno de evaluación facilite la extracción de datos cuantitativos y cualitativos para posteriores desarrollos.

La versión del sistema de BR que hemos desarrollado trata con preguntas y con corpus escritos en euskera. Incorpora herramientas y recursos desarrollados por el grupo IXA[1], tales como, el analizador morfosintáctico, el lematizador, el detector/clasificador de entidades...

[1] http://ixa.si.ehu.es

ISSN: 1135-5948

Mediante el entorno de evaluación se pretende probar y evaluar, no sólo el comportamiento global del sistema de BR, sino también el comportamiento de las herramientas integradas en dicho sistema.

En el marco de las conferencias TREC[2] y CLEF[3] se han aportado entornos, criterios y métricas de evaluación para los sistemas de BR. Aunque CLEF es un foro de evaluación multilingüe, la mayoría de los sistemas monolingües están orientados al inglés. En nuestro caso, tratándose de un sistema de BR monolingüe para el euskera, se plantea la necesidad de crear un marco de evaluación propio. Pero, al mismo tiempo, la evaluación del sistema debe ser comparable con los sistemas ya existentes. Estos condicionantes han marcado la generación del banco de pruebas.

En este artículo se presenta, en primer lugar, la plataforma general, que incluye tres componentes principales: corpus y recursos léxicos, el sistema de BR y el entorno de evaluación. En el apartado tres se describe dicho entorno de evaluación y, además, se comenta la primera prueba piloto. Por último, se presentan las conclusiones más relevantes y se reseñan algunas líneas de trabajo futuro.

2 Descripción de la plataforma de BR

La plataforma de BR está constituida por tres componentes principales y una interfaz común que permite interactuar con el usuario. El entorno de evaluación se describe en el apartado tres. A continuación mencionamos los aspectos más relevantes de los otros dos componentes.

2.1 Léxico y corpus.

Este componente engloba:

- Un corpus del periódico "Euskaldunon Egunkaria" de los años 2000, 2001 y 2002. Este corpus se etiquetó en el marco del proyecto Hermes[4] e Hizking21[5]. Contiene 23 millones de palabras, ya lematizadas. Se han marcado, además, las unidades léxicas complejas y las entidades.
- BasqueWN: Versión del EuroWN para el euskera. Actualmente, consta de 24247, 3503 y 103 synsets correspondientes a nombres, verbos y adjetivos respectivamente. En los cuales hay 20515 nombres, 3285 verbos y 43 adjetivos diferentes. Este recurso léxico está integrado en el MCR o "Multilingual Central Repository" (Atserias, 2004) (desarrollado en el marco del proyecto *Meaning*[6]), lo que posibilita su uso multilingüe, es decir, el acceso a toda la información contenida para otros idiomas.
- EEBL: Base de conocimiento léxico extraído del diccionario "Euskal Hiztegia" (Agirre et al., 2003; Lersundi, 2005).

2.2 Sistema de BR.

Se ha desarrollado un prototipo en el que se prima la versatilidad y la adaptabilidad de su arquitectura. El sistema se basa en servicios web, integrados mediante el protocolo de comunicación SOAP. La adopción de este modelo permite incorporar al sistema herramientas ya creadas en el grupo IXA, tales como el analizador morfológico, el lematizador o el reconocedor y clasificador de entidades.

Todas estas herramientas son servicios web que funcionan autónomamente. El sistema de BR es el cliente que accede a esos servicios cuando los necesita. Este modelo distribuido permite parametrizar las herramientas lingüísticas y adecuar de un modo organizado el comportamiento del sistema en la fase de desarrollo y test.

La comunicación entre los distintos servicios se lleva a cabo mediante documentos XML. Cada servicio recibe, además de los datos, un documento XML dónde se especifica la *configuración* o conjunto de parámetros para su ejecución.

2.2.1 Arquitectura modular

La versión actual del sistema es muy básica. Incluye los tres módulos principales de los sistemas de BR:

1. Análisis de la pregunta: en este servicio se procesa la pregunta y se extrae la información necesaria para las siguientes tareas. En el análisis de la pregunta se recuperan los términos de búsqueda y el tipo de respuesta esperada. En esta tarea se hace uso del analizador morfológico para el euskera, *Morfeus* (Alegria et al., 2002), y el reconocedor/clasificador de entidades, *Eihera* (Alegria et al., 2004). Un estudio de

[2] http://trec.nist.gov
[3] http://www.clef-campaign.org
[4] http://nlp.uned.es/hermes
[5] http://www.hizking21.org
[6] http://www.lsi.upc.es/~nlp/meaning/meaning.html

la tipología de las preguntas en euskera ha permitido deducir con bastante precisión el tipo de respuesta esperada. Opcionalmente, se pueden expandir los términos de búsqueda mediante la sinonimia, hiponimia y/o hiperonimia. Para este proceso de expansión léxica el sistema recurre al servicio de consulta léxico-semántica, que da acceso al BasqueWN.

2. Recuperación de pasajes: la unidad de recuperación es el párrafo y no el documento entero. En este prototipo se ha utilizado el motor de búsqueda *swish-e*[7]. Se trabaja sobre el corpus periodístico del "Euskaldunon Egunkaria" que ha sido previamente procesado.
3. Extracción de respuestas: de los párrafos recuperados se extraen todas las posibles respuestas candidatas. Estas respuestas son siempre entidades del mismo tipo que la respuesta esperada en el análisis de la pregunta. Se crean ventanas de dimensión variable teniendo como eje las respuestas candidatas, y se establece un peso para cada respuesta en función de la proximidad de los términos de la pregunta (Moldovan et al., 1999). A la hora de computar la cercanía de los términos se diferencia entre los propios términos de la pregunta y entre sus expansiones (Viñedo, 2003). El cómputo de estos valores devuelve una lista ordenada y ponderada de respuestas. El sistema, por último, selecciona las cinco primeras, como máximo.

3 Entorno de evaluación

El entorno de evaluación da soporte al proceso de test y mejora. Se cuenta, para ello, con un banco de preguntas que interesa que sean equiparables a las preguntas utilizadas en los foros internacionales. Por tanto, el banco de pruebas se ha diseñado en función de esa premisa.

3.1 Diseño y creación de la base de datos de preguntas-respuestas.

Teniendo en cuenta que la mayoría de los sistemas monolingües se han evaluado en las distintas ediciones del TREC, hemos tomado como conjunto de partida una colección de 400 preguntas seleccionadas de forma aleatoria de entre las distintas ediciones del TREC.

Tras la elección del conjunto de preguntas, éstas fueron traducidas al euskera. Sin embargo, estas preguntas TREC traducidas no son adecuadas para el corpus sobre el que vamos a trabajar, dado que, como es lógico, la mayoría de ellas no se responden en ese corpus. Ha sido necesario, por tanto, *localizar* las preguntas traducidas, es decir, reescribir su contenido para que se puedan responder en nuestra colección de documentos. Para este trabajo se ha utilizado el corpus desarrollado en el proyecto Hermes que corresponde al "Euskaldunon Egunkaria" del año 2000. En la localización se ha preservado la estructura formal de la pregunta, tanto su estructura sintáctica como el tipo de respuesta. Pero se han adecuado los términos de forma que la pregunta tuviera respuesta en nuestro corpus.

Los tipos de respuesta esperados para las preguntas son los siguientes: PERSON, ORGANIZATION, LOCATION, NUMERIC, TIME, PROPERTY, DEFINITION, MANNER y OTHER. El tipo NUMERIC se especializa como NUMERIC-MEASURE-DISTANCE y NUMERIC-MEASURE-DURATION. El tipo TIME se especializa en TIME-DATE.

Ilustremos con un ejemplo este trabajo. Una pregunta seleccionada ha sido la 642 del TREC9, que es: "*Who's the lead singer of the Led Zeppelin band?*", la cual se traduce al euskera como: "*Nor da Led Zeppelin taldeko kantari nagusia?*". No se responde a esta pregunta en nuestro corpus. Sin embargo, se puede localizar la pregunta de forma que sí tenga respuesta. Quedaría de esta forma: "*Nor da Oskorri taldeko kantari nagusia? (¿Quién es el cantante principal del grupo Oskorri?)* ".

Para cada pregunta tratada, se almacena la siguiente información en la base de datos:

- Pregunta original del TREC
- Edición del TREC de la que se ha seleccionado
- Traducción de la pregunta
- Pregunta localizada
- Tipo de respuesta esperado para la pregunta original
- Tipo de respuesta esperado para la pregunta localizada
- Respuesta(s) de la pregunta localizada
- Documentos en los que se ha(n) encontrado la(s) respuesta(s).

En total se han traducido y localizado 396 preguntas. La mayoría, 327, son de tipo *factoid*.

[7] http://swish-e.org

La tabla 1 muestra qué partículas interrogativas se han recogido en esa colección de preguntas.

Part. Interrogativa	Nº de preguntas	Porcentaje
What	200	50.50%
When	45	11.36%
How	55	13.88%
Where	23	5.80%
Who	47	11.86%
Otros	26	11.86
Total	396	100%

Tabla 1: Porcentajes de partículas interrogativas de las preguntas originales seleccionadas.

3.2 Primera evaluación piloto.

Muchas son las causas que pueden provocar que no se encuentre la respuesta correcta en el corpus: la respuesta no se encuentra en la colección de documentos utilizada, el análisis de la pregunta es erróneo, ha fallado la recuperación de documentos...

Una evaluación basada en una visión monolítica del sistema aporta poca luz sobre su comportamiento. Por ello se ha pretendido que la evaluación sea escalonada y que aporte información sobre cada una de las fases del proceso, aprovechando de esta forma la organización distribuida del sistema.

El primer nivel de evaluación trata sobre si los términos de la pregunta para la búsqueda de documentos se han seleccionado correctamente o no y si el tipo de respuesta esperada es correcto o no. La evaluación del tipo de respuesta puede ser automática ya que disponemos en la base de datos del tipo de respuesta esperada para cada pregunta. En cambio, la evaluación de los términos seleccionados debe ser manual y está ligada a la recuperación de los pasajes del corpus.

Basándonos a modo ilustrativo en el ejemplo anterior, el tipo de respuesta que devuelve el sistema es *PERSONA* y los términos que selecciona son *Oskorri, talde (grupo), kantari (cantante), nagusi (principal)*. En este caso, el tipo de respuesta es correcto y se han seleccionado todos los términos relevantes de la pregunta. Falta comprobar si los documentos seleccionados son correctos o no.

En el segundo nivel se analiza si se han recuperado los pasajes que contienen la respuesta. Se trata, a su vez, de un indicador de la calidad de los términos seleccionados y, en su caso, de las técnicas de expansión utilizadas. Se trata realmente de la evaluación de un sistema de IR.

En el ejemplo, con los términos seleccionados se recupera un único pasaje, que no contiene la respuesta. El hecho de que no se haya recuperado ningún pasaje "válido" se puede deber a que los términos seleccionados (*Oskorri, talde, kantari, nagusi*) no son los utilizados en los documentos que contienen la respuesta, o bien, a que la búsqueda ha sido demasiado estricta. El empleo de técnicas de relajación y/o expansión léxica puede ser válido para recuperar pasajes candidatos (Bilotti, 2004). Así podríamos obtener el término *abeslari* (sinónimo de *kantari*) desde la base de datos léxica MCR, lo que permitiría recuperar tres documentos, entre los que sí estaría la respuesta adecuada

Otra técnica interesante en este contexto es la relajación de los términos de la pregunta en base al valor IDF (*Inverse Document Frequency*) obtenido en el corpus de consulta. La base de esta técnica es la eliminación sucesiva del término con menor peso según los diferentes algoritmos que se describen en Bilotti (2004).

En este nivel puede darse el caso de que se recuperen pasajes válidos que no se habían almacenado en la base de datos de preguntas-respuestas. Esto nos permite enriquecer la propia base de datos.

En un tercer nivel, se evalúa si entre las posibles respuestas obtenidas desde los pasajes seleccionados se encuentra la respuesta adecuada a la pregunta. Es decir, si teniendo en cuenta el tipo de respuesta esperado, se seleccionan bien las respuestas candidatas. En la versión actual del sistema, la selección de las respuestas es muy dependiente de las entidades marcadas en el corpus, por lo que, indirectamente, también se evalúa el rendimiento del reconocedor/clasificador de entidades. En el ejemplo que nos ocupa, la expansión léxico-semántica ha permitido la inclusión de la respuesta correcta entre las candidatas.

En el cuarto nivel se evalúa la precisión con la que se selecciona la respuesta correcta, es decir, se valora la adecuación de los criterios para establecer un ranking entre las respuestas candidatas.

En el ejemplo, el sistema ha asignado el mayor peso a la respuesta "Natxo de Felipe", que es la respuesta adecuada.

Aunque en esta primera experiencia piloto, la evaluación ha sido manual, la complejidad del proceso y su alto costo invita a aplicar métodos automáticos. La información contenida en la base de datos posibilita este tipo de tratamientos, aunque siempre haya que contar con un cierto margen de error. Pretendemos que sólo cuando haya una necesidad clara, por cuestiones de fiabilidad y precisión, se aplique la evaluación manual. Pero, mientras tanto, la evaluación automática puede servir para analizar el comportamiento del sistema en sus distintas fases de proceso. Y este tipo de evaluación, mucho menos costosa, sería la base para una mejora progresiva del sistema.

3.3 Resultados de la prueba piloto.

Mediante esta primera experiencia piloto se ha pretendido observar la adecuación del entorno de evaluación.

Se han efectuado varias pruebas con las preguntas de la base de datos que obedecen a estos dos patrones: "¿Quién es...? y ¿Dónde está...? En total han sido 38 las preguntas que siguen esos patrones.

La evaluación ha sido manual, según las directrices del TREC (Voorhees y Tice, 1999).

A continuación se resumen los resultados obtenidos en las distintas ejecuciones (tabla 2).

Se han obtenido datos del rendimiento del sistema sobre el corpus utilizado en la localización de las preguntas y del que se han extraído todas las respuestas (9 millones de palabras, año 2000) y sobre el corpus completo de 23 millones de palabras, correspondientes a los años 2000, 2001 y 2002.

Se trata de un análisis progresivo del comportamiento del sistema, dónde se subraya la influencia de la aplicación de la técnica de relajación IDF:

1. Recuperación de pasajes (ver fila 1).

   En el caso de la no utilización de técnicas de relajación sobre el corpus reducido (columna 1), no se recupera ningún pasaje para el 37% de las preguntas.

   Posibles causas:

   - El conjunto de documentos en el que se busca la respuesta es pequeño.
   - El conjunto de términos utilizados para la búsqueda, en algunos casos, es muy restrictivo, y en esta versión se pide que todos los términos aparezcan en el pasaje seleccionado.

   Se observa en la columna 2 que el rendimiento del sistema no mejora trabajando sobre el corpus total. Se trata, por tanto, de un punto crítico en el rendimiento global del sistema.

   Sin embargo, la aplicación de técnicas de relajación IDF permite recuperar algún pasaje candidato para todas las preguntas en cualquiera de las versiones del corpus (columnas 3 y 4).

2. Extracción de las respuestas candidatas desde los pasajes. Se ha calculado el número de preguntas para las que se extrae alguna respuesta correcta desde los pasajes recuperados (ver fila 2).

   En las ejecuciones sin relajación de términos, sólo para el 50% de las preguntas se obtiene algún pasaje que contiene la respuesta correcta, mientras que este porcentaje asciende al 66-68% cuando se aplica la relajación de términos.

   Se ha observado que el reconocedor de entidades es clave en esta tarea, y que el comportamiento de esta herramienta condiciona en gran medida el rendimiento del sistema. También es necesaria la aplicación de tratamiento sintáctico para responder a cierto tipo de preguntas.

3. Selección de las respuestas. El sistema devuelve cinco respuestas como máximo. En este paso se han contado las preguntas para las que se ha incluido la respuesta correcta entre las seleccionadas (ver fila 3).

   En la ejecución sin relajación de términos, a partir del conjunto de preguntas con alguna respuesta candidata correcta, en un 83-84% se ha clasificado la respuesta correcta entre las cinco primeras.

   Cabe destacar que en la ejecución con relajación de términos los resultados son sustancialmente mejores sobre el corpus reducido, ya que en el 88% de los casos con alguna respuesta candidata correcta se selecciona dicha respuesta. Sin embargo, con la utilización del corpus completo el rendimiento desciende hasta el 57%. Esta pérdida de efectividad se debe a que el número de documentos recuperados es considerablemente superior a cualquiera de los casos anteriores para cada una de las preguntas, lo que dificulta la selección de la respuesta correcta.

Finalmente, para evaluar el sistema en su totalidad y poder hacer una comparativa con otros sistemas, se ha seguido la métrica "*Mean*

Reciprocal Rank" (MRR), utilizada en las ediciones del TREC hasta el 2001. En el MRR se asignan los valores (1, 0.5, 0.33, 0.25, 0.2, 0) en función de la posición de la respuesta correcta. Según esta métrica, el sistema obtiene los mejores resultados aplicando la relajación de términos con IDF (ver fila 4).

	Sin relajación de los términos		Relajación IDF	
	Corpus reducido	Corpus completo	Corpus reducido	Corpus completo
PRP	24	24	38	38
PRC	18	19	25	26
PRS	15	16	22	15
MRR	0.313	0.302	0.517	0.335

PRP: Nº de preguntas para las que se han recuperado pasajes
PRC: Nº de preguntas para las que se extrae alguna respuesta correcta desde los pasajes recuperados
PRS: Nº de preguntas para las que se ha incluido la respuesta correcta entre las seleccionadas

Tabla 2: Evolución de la prueba piloto sobre un total de 38 preguntas.

A la luz de estos resultados los puntos más susceptibles de mejora son:

1. La extracción de documentos a partir de términos clave. La técnica de la relajación de términos con IDF supone una mejora sustancial en este punto. El problema es que, aún así, para un 32-34% de las preguntas no se obtiene ningún pasaje con la respuesta correcta. Se ve la necesidad de incorporar conocimiento léxico-semántico, tal y como se sugiere en el ejemplo del apartado 3.2.
2. La extracción de respuestas a partir de los pasajes candidatos. Una de las posibles actuaciones consistiría en analizar la valía de las palabras de la pregunta que no han sido seleccionadas como términos de búsqueda, y tratar de compensar su ausencia en el momento de ponderar las respuestas candidatas.

4 Conclusiones y trabajo futuro

Se ha presentado un entorno para gestionar el desarrollo y la evaluación de un sistema de BR en euskera. Para ello se han integrado tres componentes: el propio sistema de BR, el componente léxico y el de evaluación.

El diseño del sistema está muy condicionado por las exigencias de evaluación. Se ha optado por una arquitectura distribuida, que integra servicios web autónomos mediante el protocolo SOAP. Por tanto, se puede ejecutar y evaluar cualquier servicio por separado. El sistema es muy configurable.

En el proceso de evaluación se ha pretendido explotar esa adaptabilidad del sistema, para lo que se ha creado una base de datos de preguntas-respuestas, dónde se recoge toda la información relacionada con una colección de 396 preguntas. Estas preguntas tipo TREC se han localizado de forma que, por una parte, mantienen su estructura original y, por otra, tienen respuesta en el corpus periodístico que se está usando. El entorno de evaluación permite, además, actualizar y enriquecer dicha base de datos.

La primera experiencia piloto ha permitido deducir que el entorno de evaluación es adecuado no sólo para la tarea de la búsqueda de respuestas, sino para evaluar el rendimiento de las herramientas de PLN en este escenario.

De cara al futuro, una vez consolidada la plataforma general, se pretende mejorar el sistema de BR, aplicando técnicas más avanzadas, haciendo uso de otros recursos y ampliando el abanico de las preguntas tratadas. Entramos en una fase de pruebas y mejoras progresivas, en la que esperamos que el entorno descrito sea un soporte adecuado para detectar los puntos críticos y mejorar el rendimiento del sistema.

Agradecimientos

Este trabajo está subvencionado por el Departamento de Industria del Gobierno Vasco (proyectos Ihardetsi SAIOTEK S-PE03UN14 e Hizking21 ETORTEK2002HIZKING21).

Bibliografía

Agirre E., Ansa O., Arregi X., Artola X., Díaz de Ilarraza A., Lersundi M., "A Conceptual Schema for a Basque Lexical-Semantic Framework", Complex 2003, págs. 1-10.

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R., "Robustness and customisation in an analyser/lemmatiser for Basque ", LREC-2002, págs.

Alegria I., Arregi O., Balza I., Ezeiza N., Fernandez I., Urizar R., "Design and Development of a Named Entity Recognizer

for an Agglutinative Language ", IJCNLP-04.

Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B., Vossen P., "The MEANING Multilingual Central Repository", Proc. of the 2nd Global WordNet Conference. Brno, Czech Republic, 2004.

Bilotti M., "Query Expansion Techniques for Question Answering", tesis doctoral Massachusetts institute of technology, 2004.

Harabagiu S., Miller A., Moldovan D., "FALCON: Boosting Knowledge for Answer Engines", en TREC-9 2000.

Lersundi M., "Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpena eta postposizioak", tesis doctoral UPV/EHU 2005.

Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Gîrju R., Rus V. "LASSO: A Tool for Surfing the Answer Net", en TREC-8 1999, págs. 175

Snell J., Tidwell D., Kulchenco P., "Programming Web Services with SOAP", O'Reilly 2002.

Vicedo J.L., "Recuperando información de alta precisión: Los sistemas de Búsqueda de Respuestas", monografía SEPLN 2003.

Voorhees E., Tice D., "The TREC-8 Question Answering Track Evaluation", en TREC-8 1999, págs. 83-106.