Desarrollo de un corpus de entrenamiento para sistemas de Búsqueda de Respuestas basados en aprendizaje automático*

Empar Bisbal y **Lidia Moreno**
Depto. de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia (España)
{ebisbal,lmoreno}@dsic.upv.es

David Tomás y **José. L. Vicedo**
Depto. de Lenguajes y Sistemas Informáticos,
Universidad de Alicante (España)
{dtomas,vicedo}@dlsi.ua.es

Resumen: En este trabajo se describe el desarrollo de un corpus de preguntas y respuestas factuales similares a las utilizadas en las conferencias TREC. Dicho corpus consta de más de 70.000 muestras, cada una de ellas con la siguiente información: una pregunta, el tipo de esa pregunta, la respuesta exacta, el párrafo del que ha sido extraída la respuesta, el documento del que ha sido extraído el párrafo y una etiqueta indicando si la respuesta es correcta (muestra positiva) o no (muestra negativa) en el contexto proporcionado. El corpus desarrollado puede ser utilizado, por ejemplo, para entrenar un clasificador binario que decida de forma automática si la respuesta proporcionada por un determinado sistema es correcta o no. Hasta donde conocemos, éste es el primer corpus que puede ser utilizado para entrenar todos y cada uno de los módulos de un sistema de Búsqueda de Respuestas: clasificación de la pregunta, recuperación de información, extracción de la respuesta y validación de la misma. El proceso para la obtención del corpus ha sido realizado de forma semi-automática.
Palabras clave: corpus, aprendizaje automático, búsqueda de respuestas

Abstract: This paper describes the development of an English corpus of factoid TREC-like question-answer pairs. The corpus obtained consists of a set of more than 70,000 samples, containing each one the following information: a question, its question type, an exact answer to that question, the different context levels (sentence, paragraph and document) where the answer occurs inside a document, and a label indicating whether the answer is correct (a positive sample) or not (a negative sample). For instance, this corpus can be used for training a binary classifier in order to decide if a given answer is correct (positive) to the question formulated or not (negative). To our knowledge, this is the first corpus that can be used to train each one of the modules of a trainable Question Answering system: question classification, information retrieval, answer extraction and answer validation. The process carried out to obtain the corpus was semi-automatic.
Keywords: corpora, machine learning, question answering

1. Introducción

Las aproximaciones empíricas al Procesamiento del Lenguaje Natural (PLN), sugieren que es posible aprender las complicadas y desordenadas estructuras del lenguaje mediante el estudio de grandes cantidades de ejemplos en lenguaje natural, utilizando para ello técnicas como la estadística, el reconocimiento de patrones o los métodos de aprendizaje automático. Estas aproximaciones se basan en la utilización de grandes corpus textuales.

Numerosas investigaciones muestran que se pueden realizar importantes progresos en el campo de la comprensión del lenguaje mediante la extracción automática de información a partir de grandes corpus de texto. Por esta razón se han desarrollado gran cantidad de recursos para ayudar en la tarea del aprendizaje automático. El área de aplicación de cada recurso depende principalmente del nivel de anotación que presente. Existen corpus como el Proyecto Gutemberg[1] que no proporcionan ningún tipo de información adicional, sólo el texto plano. Existen también

* Este trabajo ha sido desarrollado en el marco del proyecto CICYT R2D2 (TIC2003-07158-C04).

[1] http://www.gutenberg.org

corpus como el de noticias de la agencia española EFE de los años 1994 y 1995 (véase CLEF[2]), con anotaciones sobre formato que permiten identificar información acerca de la edición, los autores, las cabeceras de la noticia o los párrafos en los que se subdivide el corpus. Finalmente, existen corpus anotados como el Penn Treebank (Marcus et al., 1993), que proporcionan información más elaborada acerca de la estructura sintáctica y gramatical del texto. Todos estos corpus son generales, no centrándose su utilización a una tarea concreta.

En este trabajo se presenta un corpus desarrollado para aplicar aprendizaje automático en la tarea de Búsqueda de Respuestas (BR). Este tipo de sistemas trata de extraer respuestas exactas a preguntas formuladas en lenguaje natural. Se ha desarrollado un corpus de pares pregunta-respuesta en inglés adaptado al entrenamiento de estos sistemas en cualquiera de las etapas de la tarea de BR: clasificación de la pregunta, recuperación de documentos, extracción de la respuesta y validación de la misma.

El corpus consta de más de 70.000 muestras. Cada una de estas muestras contiene información que relaciona la pregunta y la respuesta a cuatro niveles contextuales diferentes: a nivel de documento, a nivel de párrafo, a nivel de oración y a nivel de coincidencia exacta. Cada muestra está etiquetada como positiva o negativa en función de si la respuesta se puede considerar correcta o no en el contexto proporcionado. Las instancias negativas son útiles para identificar el contexto en el que una respuesta no es correcta. De esta manera, el corpus puede ser empleado para el entrenamiento de un clasificador binario que permita decidir de forma automática si una respuesta dada responde correctamente o no a la pregunta formulada. Además, para que el corpus pueda ser utilizado en la fase de clasificación de la pregunta, se ha incluido el tipo de ésta como información.

Existen otros corpus empleados para el entrenamiento de alguna de las fases de un sistema de BR. Sin embargo, éste es el primero que puede ser utilizado para entrenar todos los componentes de este tipo de sistemas, incluyendo además tanto muestras positivas como negativas.

El resto del artículo está organizado de la siguiente manera: en la sección 2 se presenta la situación actual en el desarrollo de corpus y sistemas de BR basados en aprendizaje automático; la sección 3 describe el formato del corpus y la información que contiene cada muestra; en la sección 4 se comentan los recursos que han sido necesarios para la construcción del corpus, así como los detalles acerca del proceso de generación del mismo; la sección 5 muestra las estadísticas del corpus y, por último, en la sección 6 se comentan las posibles aplicaciones y trabajos futuros relacionados con el corpus presentado.

2. Trabajos Relacionados

Existen diversos sistemas de BR basado en corpus, que emplean técnicas de aprendizaje automático en alguna de las fases del proceso de búsqueda.

En (Ravichandran et al., 2003), se desarrolló un corpus de pares pregunta-respuesta llamado *KM database*. Este corpus contiene preguntas del trivial (el juego de cartas) junto con sus correspondientes respuestas. Las preguntas fueron filtradas de forma que quedaron aquellas con formato similar a las del TREC[3]. Finalmente, se obtuvo un corpus de 16.228 pares. A partir de este corpus, se extrajeron de forma automática patrones que fueron utilizados en la fase de extracción de la respuesta.

En (Soricut and Brill, 2004), se desarrolló un sistema de BR basado en una arquitectura *noisy-channel* que hace uso tanto de un modelo del lenguaje para las respuestas, como de un modelo de transformación para los términos de la respuesta/pregunta. Para poder aplicar técnicas de aprendizaje automático, primero se construyó un corpus de pares pregunta-respuesta con una amplia cobertura léxica. Se obtuvieron aproximadamente 1 millón de pares a partir de FAQs y el corpus resultante fue aplicado en las fases de análisis de la pregunta y de extracción de la respuesta. El sistema se centra principalmente en preguntas no factuales.

El sistema desarrollado por (Agichtein et al., 2001) utiliza una colección de aproximadamente 30.000 pares pregunta-respuesta para el entrenamiento, obtenidas a partir de más de 270 FAQs sobre diversos temas del proyecto FAQFinder (Burke et al., 1997). Se

[2]http://www.clef-campaign.org

[3]Preguntas con diez palabras o menos que no sean de tipo test

utilizó dicho corpus en aprender características textuales para clasificar las preguntas y generar reformulaciones, evaluando dichas transformaciones en sistemas de recuperación de información.

La aproximación de (Berger et al., 2000) está fuertemente basada en aprendizaje automático. A partir de una colección de preguntas y respuestas, el algoritmo aprende correlaciones léxicas entre unas y otras. Para el aprendizaje se recopilaron dos conjuntos diferentes de pares pregunta-respuesta: 1.800 muestras extraídas de FAQs de Usenet y 5.145 extraídas a partir de diálogos telefónicos.

Todos estos corpus presentan alguno de los siguientes problemas al ser utilizados por sistemas de BR basados en aprendizaje:

- No se proporciona el tipo de la pregunta, por lo que el corpus no puede ser utilizado en la fase de clasificación de la pregunta.
- No contiene muestras negativas, que resultan de utilidad a la hora de aprender el contexto en el que una respuesta es incorrecta.
- El contexto proporcionado para las respuestas no es adecuado para alguna de las etapas de un sistema de BR: demasiado breve para la fase de recuperación de información o demasiado extenso para la de extracción de la respuesta.

Con el corpus desarrollado en este trabajo pretendemos solventar las carencias enunciadas anteriormente. Nuestro corpus contiene pares de pregunta-respuesta obtenidos de forma semi-automática a partir de recursos del TREC[4] (más concretamente, a partir de las preguntas y los corpus de la tarea de BR). De esta manera, se ha obtenido un corpus de preguntas factuales similares a las de TREC, con sus correspondientes respuestas, plenamente orientado a la tarea de BR. A diferencia de otras aproximaciones, cada muestra ha sido etiquetada con el tipo de la pregunta para que el corpus se pueda emplear también en la fase de clasificación de la pregunta. El corpus presenta cuatro niveles de contexto diferentes para cada pregunta, con el objetivo de que se pueda emplear en todas las etapas de un sistema de BR: el documento y el párrafo para recuperación de información, la oración para validación de la respuesta, y la respuesta exacta para extracción de la misma. Además, el corpus contiene tanto respuestas correctas como incorrectas, es decir, disponemos de muestras positivas y negativas que pueden ser muy útiles a la hora de entrenar clasificadores binarios. Por último, el número de muestras obtenidas es suficientemente grande (aproximadamente 70.000) para que el corpus resulte apropiado para el entrenamiento de sistemas de aprendizaje automático.

[4]Text REtrieval Conference, http://trec.nist.gov

3. Descripción del Corpus

El corpus desarrollado está formado por un conjunto de pares pregunta-respuesta en inglés, donde cada muestra incluye los siguientes campos:

- El número de muestra, empleado como identificador.
- El número de pregunta en la colección del TREC.
- La pregunta en sí.
- El tipo de la pregunta, indicando la clase de respuesta esperada según una taxonomía de 15 clases (Bisbal et al., 2005), como LOCATION, PROPER_NAME, EVENT, ORGANIZATION, ACRONYM, ... Esta información resulta útil en la fase de clasificación de la pregunta, donde se asigna una clase o categoría a cada pregunta introducida al sistema.
- La respuesta exacta. Esta información es necesaria en la fase de extracción de la respuesta, donde se deben extraer únicamente los términos exactos que forman parte de la respuesta a la pregunta formulada.
- El contexto, a nivel de oración, donde se encuentra la respuesta. Esta información, junto con la pregunta, puede ser utilizada para entrenar sistemas de implicación textual (*textual entailment*) (Dagan et al., 2005) con los que abordar la fase de validación de respuestas.
- El contexto, a nivel de párrafo, donde se ha extraído la respuesta. Este dato es útil para entrenar sistemas de recuperación de pasajes.

- El identificador del documento del que se extrajo la respuesta. Esta información se puede emplear para el entrenamiento de sistemas de recuperación o reordenación de documentos, ayudando a descartar aquellos que no tengan relación con la respuesta.
- Una etiqueta indicando si la respuesta es correcta (muestra positiva) o incorrecta (muestra negativa). De esta forma, se pueden entrenar clasificadores binarios para determinar si una respuesta exacta, una oración o un párrafo, son válidos para una pregunta dada.

La Figura 1 presenta tres muestras del corpus para la pregunta "*Who is Tom Cruise married to?*". Esta pregunta es de tipo PROPER_NAME, indicando que espera recibir como respuesta el nombre de una persona. En el primer ejemplo, la respuesta "*Nicole Kidman*" es correcta (usemos nuestra imaginación y situémonos a principios de los 90) y el contexto (a nivel de oración y párrafo) la justifica. Por lo tanto, la muestra es etiquetada como POSITIVA. En el segundo ejemplo, la respuesta también es "*Nicole Kidman*" pero en este caso el contexto no da soporte a la respuesta. Esta muestra es etiquetada como NEGATIVA. En el último ejemplo, la respuesta es "*Bill Harford*" y el contexto no la justifica en ningún caso, por lo que esta muestra también será etiquetada como NEGATIVA.

4. Construcción del Corpus

En este apartado, la primera subsección describe los recursos necesarios para la construcción del corpus. El siguiente punto explica el proceso llevado a cabo para obtener el conjunto de muestras del corpus.

4.1. Recursos

Los recursos necesarios para la elaboración del corpus se han obtenido de las colecciones empleadas en la tarea de BR de las conferencias TREC.

A la hora de recopilar los pares de pregunta-respuesta para la construcción del corpus, nos centramos en aquellas preguntas que tuvieran respuesta exacta. Por esta razón, únicamente las preguntas formuladas del TREC 2002 al TREC 2005 fueron tenidas en cuenta. En años anteriores de la tarea de BR (TREC 1999 al 2001), a los sistemas se les exigía que devolvieran pasajes, no respuestas exactas, por lo que las descartamos para nuestro corpus. Por esta misma razón, sólo las preguntas de la subtarea principal (*"main" subtask*) fueron tenidas en cuenta, desechando las preguntas de las subtareas *"list"* y *"passage"*. Finalmente se obtuvo un conjunto de 1.505 preguntas típicas del TREC.

Se utilizó también la colección de documentos AQUAINT[5], empleada igualmente en la tarea de BR del TREC, que nos permitió obtener los diferentes niveles de contexto en los que aparecían las posibles respuestas al conjunto de preguntas seleccionado. El corpus AQUAINT consta de 1.033.461 documentos en inglés, con aproximadamente 375 millones de palabras, obtenidos a partir de tres fuentes: el Xinhua News Service, el New York Times News Service y el Associated Press Worldstream News Service. Este conjunto de documentos fue el empleado en las últimas ediciones del TREC, del año 2002 al 2005.

Finalmente, también se utilizaron los ficheros de juicios proporcionados por la organización del TREC 2002 al TREC 2005. Estos ficheros contienen información sobre todas las respuestas dadas por los sistemas presentados a competición. Un juicio está formado por cuatro campos:

- El identificador de la pregunta.
- El identificador del documento de la colección AQUAINT donde se halla la respuesta.
- El juicio de los evaluadores.
- La respuesta.

El juicio de los asesores indica si la respuesta es correcta, incorrecta, inexacta o no soportada. "No soportada" indica que la respuesta es correcta, pero que a partir del documento no se puede deducir que ésa es la respuesta a la pregunta formulada. "Inexacta" indica que la respuesta es correcta y el documento la soporta pero que, o bien la respuesta contiene más palabras (o letras) de las necesarias, o bien carece de ellas. Véase (Voorhees99 et al., 1999) para una descripción detallada acerca de cómo se juzgaron las

[5]Linguistic Data Consortium (LDC) catalog number LDC2002T31 and ISBN1-58563-240-6.

Muestra 1

Id Muestra:	26821
Id Pregunta:	1395
Pregunta:	Who is Tom Cruise married to?
Tipo:	PROPER_NAME
Respuesta:	Nicole Kidman
Oración:	The drama is said to be about a pair of married psychiatrists (played by the married Tom Cruise and Nicole Kidman) and their sexual lives, but only a few Warner executives, Cruise and Kidman, and Pat Kingsley, a top public relations executive, have seen the film.
Párrafo:	Along the way, Kubrick's secretive methods generated a continual buzz. Actors had to sign agreements not to talk to the press, and shooting scripts were kept under strict security. The drama is said to be about a pair of married psychiatrists (played by the married Tom Cruise and Nicole Kidman) and their sexual lives, but only a few Warner executives, Cruise and Kidman, and Pat Kingsley, a top public relations executive, have seen the film.
Documento:	NYT19990326.0303
Clase:	POSITIVE

Muestra 2

Id Muestra:	26824
Id Pregunta:	1395
Pregunta:	Who is Tom Cruise married to?
Tipo:	PROPER_NAME
Respuesta:	Nicole Kidman
Oración:	The film itself, starring Tom Cruise and Nicole Kidman as a married couple in New York on a sexual odyssey, received wildly mixed reviews.
Párrafo:	The film itself, starring Tom Cruise and Nicole Kidman as a married couple in New York on a sexual odyssey, received wildly mixed reviews. After strong box office sales in its first weekend, attendance has dropped sharply.
Documento:	NYT19990727.0184
Clase:	NEGATIVE

Muestra 3

Id Muestra:	26831
Id Pregunta:	1395
Pregunta:	Who is Tom Cruise married to?
Tipo:	PROPER_NAME
Respuesta:	Bill Harford
Oración:	The story follows the descent of Bill Harford (Cruise, toothy as ever), a successful young doctor on the Upper West Side of Manhattan, into a perilous, secretive netherworld.
Párrafo:	At the same time ''Eyes Wide Shut'' is a sternly anti-erotic movie that regards its sexual license with a cold puritanical hauteur. The movie is not a turn-on (it is really a horror film without gore), and the sexual chemistry between its married stars, Tom Cruise and Ms. Kidman, is tepid at best. The story follows the descent of Bill Harford (Cruise, toothy as ever), a successful young doctor on the Upper West Side of Manhattan, into a perilous, secretive netherworld. The catalyst is a confession by his wife, Alice (Ms. Kidman), about the fierce, unconsummated desire she once felt for a young naval officer. In black-and-white sequences that punctuate the movie, Bill torments himself with visions of Alice and her would-be lover in bed together, and these images drive him to examine his own wayward impulses.
Documento:	NYT19990719.0343
Clase:	NEGATIVE

Figura 1: Tres ejemplos representativos del corpus desarrollado.

respuestas. La Figura 2 muestra un fragmento de uno de estos ficheros de juicios.

4.2. Proceso

El corpus fue obtenido de forma semi-automática a partir de los recursos descritos en el punto anterior. Primero, se seleccionaron todas las preguntas factuales de la

1395	NYT19991220.0294	-1	Julia Roberts
1395	NYT19991101.0416	1	Nicole Kidman
1395	APW19990712.0006	3	actress Nicole Kidman
1395	NYT19991101.0416	3	actress Nicole Kidman
1395	APW19990712.0006	1	Nicole Kidman
1395	APW19990423.0019	2	Tom Cruise and Nicole Kidman
1395	NYT19990628.0254	2	Nicole Kidman

Figura 2: Fragmento de un fichero de juicios. La tercera columna indica si la respuesta es incorrecta (-1), correcta (1), no soportada (2) o inexacta (3).

tarea de BR de los TREC de 2002 a 2005. Estas preguntas fueron etiquetadas manualmente con el tipo de respuesta esperado siguiendo la clasificación presentada en (Bisbal et al., 2005).

Para cada pregunta se realizó un proceso automático de búsqueda de todas las posibles respuestas dentro de los ficheros de juicios. Estos ficheros muestran las respuestas enviadas por los sistemas que participaron en dichas competiciones. Cada juicio fue procesado de acuerdo a los siguientes pasos:

1. Lectura de la respuesta.
2. Lectura del juicio dado por los evaluadores.
3. Lectura del documento indicado y búsqueda de apariciones de la respuesta dentro del mismo.
4. Obtención y almacenamiento de todos los párrafos del documento que contienen la respuesta. Ya que en el corpus AQUAINT los párrafos de los documentos aparecen etiquetados, se hizo uso de dichas etiquetas para extraerlos fácilmente. Por cada párrafo obtenido se generó una muestra.
5. Extracción de la oración donde aparece la respuesta a partir del párrafo.

En este punto del proceso automático, se dispone de una conjunto de muestras que relacionan cada pregunta con una posible respuesta y con los diferentes contextos (oración, párrafo y documento) en los que ha sido localizada. Si los evaluadores juzgaron la respuesta como "incorrecta", la muestras es etiquetada como NEGATIVA. Si el juicio fue "correcta", la muestra es etiquetada como POSITIVA. Si la muestra fue juzgada como "no soportada", la muestra se etiqueta como NEGATIVA, puesto que aunque la respuesta es correcta, ésta no se puede deducir a partir del contexto.

Las respuestas que fueron juzgadas como "inexactas" requirieron un tratamiento especial. En este caso, el contexto sí justifica la respuesta, pero la respuesta no se puede considerar correcta porque, o bien contiene información superflua, o bien carece de alguna parte importante de la respuesta. Para solucionar este problema, se recogieron de forma automática todas las respuestas correctas (aquellas cuyo juicio es "correcta") y se intentaron localizar en el documento en el que se había encontrado la repuesta "inexacta". Por ejemplo, supongamos que la Figura 2 muestra todos los juicios para la pregunta número "1395". Al procesar la tercera línea se detecta que la respuesta dada fue "*actress Nicole Kidman*" y que el juicio dado por los evaluadores fue "inexacta". En este caso el proceso buscará todas las respuestas etiquetadas como "correcta" que aparecen en los juicios para la pregunta "1395", tratando de localizarlas en el documento "APW19990712.0006". Las respuestas encontradas son entonces etiquetadas como POSITIVAS.

Una vez finalizado el proceso automático, se disponía de un enorme conjunto de muestras con la información mostrada en la Figura 1. Pero el proceso de construcción del corpus todavía no estaba acabado, al haber de revisar algunas de las muestras manualmente. En algunos casos la respuesta aparecía más de una vez dentro del documento. El proceso automático extrae todas las apariciones de la respuesta en el documento y genera una muestra para cada una, etiquetándolas como POSITIVA o NEGATIVA según el criterio descrito anteriormente. No existe problema alguno si la etiqueta es NEGATIVA, puesto que se puede asegurar entonces que la respuesta es incorrecta o que el documento no la soporta. El problema aparece cuando la etiqueta es POSITIVA, ya que no se puede garantizar que todos los párrafos en los que aparezca la respuesta le den soporte. Estas muestras tuvieron que ser, por tanto, revisadas manualmente. La tarea de revisión fue llevada a cabo por dos revisores. El Cuadro 1 muestra en detalle el número de muestras que tuvieron que se revisadas. Sobre el total de 6.309 muestras que hubo que revisar manual-

mente, los anotadores alcanzaron un índice de concordancia Kappa (*Kappa agreement*) del 0.94. La concordancia esperada se computó según (Fleiss, 1971), siendo igual para todos los evaluadores la probabilidad de que una muestra pertenezca a una u otra categoría. En caso de desacuerdo, el desempate lo realizó un tercer anotador.

5. Estadísticas del Corpus

El corpus desarrollado consta de 72.679 muestras extraídas a partir de 1.505 preguntas diferentes (una media de 48.29 respuestas por pregunta). Tenemos un total de 8.699 muestras POSITIVAS frente a 63.980 NEGATIVAS. El porcentaje de muestras NEGATIVAS (88 %) es mucho mayor que el de las POSITIVAS (12 %). Se decidió mantener la proporción al ser éste el resultado real de los sistemas actuales.

El Cuadro 2 muestra un resumen de las estadísticas del corpus. Se incluyen también los resultados parciales para cada competición del TREC. La columna "Conjunto" indica la edición de la que provienen los recursos. "Preguntas" indica el número de preguntas utilizadas para extraer las muestras. "Juicios" muestra el número de juicios, esto es, el número total de respuestas enviadas por los participantes en la competición. "Positivas" indica el número de muestras etiquetadas como POSITIVAS, mientras que "Negativas" muestras el de NEGATIVAS. Por último, la columna "Total Muestras" resume el número total de muestras, tanto POSITIVAS como NEGATIVAS.

Los resultados obtenidos reflejan una diferencia notable entre el número de muestras extraídas en cada edición del TREC. Mientras que en 2003 y 2005 los resultados son similares (17.705 y 18.681 respectivamente), las muestras obtenidas del TREC de 2002 superan ampliamente las de 2004 (26.679 y 9.614 respectivamente). El principal motivo de esta diferencia es el número de juicios disponibles. Este número depende principalmente de tres factores:

- El número de preguntas formuladas a los sistemas. Por ejemplo, en 2002 fueron 500 mientras que en 2004 fueron sólo 230.
- El número de sistemas en competición y el número de resultados enviados. En 2002 se enviaron 67 resultados, mientras que hubo 54 en 2003.
- La convergencia de los sistemas: sólo se tienen en cuenta los juicios distintos. Si dos sistemas encuentran la misma respuesta en el mismo documento, sólo se obtiene una muestra.

6. Conclusiones y Trabajo Futuro

Muchas aplicaciones de procesamiento del lenguaje natural extraen información a partir de grandes corpus de texto con el objetivo de aprender fenómenos lingüísticos. Las aproximaciones basadas en corpus han demostrado su fácil adaptabilidad a nuevos lenguaje y dominios. En este trabajo se ha descrito el desarrollo de un corpus diseñado para su utilización en todas y cada una de las etapas de un sistema de BR basado en aprendizaje. El corpus ha sido obtenido de forma semi-automática, siendo mínimo el esfuerzo humano necesario para su desarrollo. El corpus ha sido desarrollado en inglés, ya que para este idioma los recursos son mucho más numerosos que para otras lenguas. Las muestras, al haber sido obtenidas a partir de recursos de la tarea de BR del TREC, hacen que el corpus resultante se ajuste perfectamente a las necesidades de los sistemas de BR.

Se ha obtenido una colección de 72.679 muestras, que parece suficientemente grande para entrenar cualquier sistema de BR basado en aprendizaje automático. Cada muestra relaciona una pregunta con su tipo y su respuesta exacta, proporcionando también el contexto de la respuesta a nivel de oración, párrafo y documento. De esta manera, el corpus se puede emplear para entrenar cualquier etapa de un sistema de BR, donde son necesarios diferentes niveles de contexto de la pregunta: el tipo de la pregunta para la etapa de clasificación de la pregunta, la respuesta exacta para la fase de extracción de la respuesta, la oración para la etapa de validación y el documento y pasaje para la fase de recuperación de información.

Otro beneficio de la aproximación presentada es que, a diferencia de otros corpus similares, el corpus obtenido no dispone sólo de muestras positivas, sino que también contiene muestras negativas que proporcionan el contexto en el cual una determinada respuesta es incorrecta. De este modo, un clasificador binario podría ser entrenado con este corpus

Conjunto	Revisadas manualmente
TREC 2002	2,314
TREC 2003	1,811
TREC 2004	745
TREC 2005	1,472
TOTAL	**6,309**

Cuadro 1: Muestras revisadas manualmente

Conjunto	Preguntas	Juicios	Positivas	Negativas	Total Muestras
TREC 2002	500	15,948	2,042	24,637	26,679
TREC 2003	413	9,841	2,769	14,936	17,705
TREC 2004	230	6,235	1,406	8,208	9,614
TREC 2005	362	11,967	2,482	16,199	18,681
TOTAL	**1,505**	**43,991**	**8,699**	**63,980**	**72,679**

Cuadro 2: Estadísticas del Corpus

para decidir de forma automática si las posibles respuestas son apropiadas para la pregunta formulada o no.

Como trabajo futuro, planteamos aplicar este corpus con técnicas de aprendizaje automático para la construcción de sistemas versátiles de BR.

Bibliografía

Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Shasberger B. *Building a large annotated corpus of English: The Penn Tree Bank*. In: Computational Linguistics. (1993) 313-330

Ido Dagan, Oren Glickman, and Bernardo Magnini. *Recognizing Textual Entailment*. In PASCAL Proceedings of the First Challenge Workshop, pages 1-8, Southampton, UK, April 2005.

Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Shasberger B. *Building a large annotated corpus of English: The Penn Tree Bank*. In: Computational Linguistics. (1993) 313-330

Eugene Agichtein, Steve Lawrence, and Luis Gravano. *Learning search engine specific query transformations for question answering*. In Proceedings of the 10th World Wide Web Conference (WWW10), 2001.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. *Bridging the lexical chasm: statistical approaches to answer-finding*. Research and Development in Information Retrieval, pages 192-199, 2000.

Empar Bisbal, David Tomás, José L. Vicedo, and Lidia Moreno. *A multilingual SVM-Based Question Classification System*. In Proceedings of MICAI, 2005.

Rovin D. Burke, Kristian J. Hammond, Vladimir A. Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg. *Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System*. AI Magazine, 18(2), pages 57-66, 1997.

J. L. Fleiss. *Measuring nominal scale agreement among many raters* Psyhological Bulletin, Psychological Bulletin, 76(5):378-382,1971.

Deepak Ravichandran, Abraham Ittycheriah, and Salim Roukos. *Automatic Derivation of Surface Text Patterns for a Maximum Entropy Based Question Answering System*. In Proceedings of the HLT-NAACL Conference, 2003.

Radu Soricut, and Eric Brill. *Automatic Question Answering: Beyond the Factoid*. In Proceedings of the HLT-NAACL Conference, 2004.

Ellem M. Voorhees, and Dawn M. Tice. *The TREC-8 Question Answering Track Evaluation*. In Proceedings of the 8th Text REtrieval Conference (TREC-8), 1999.