Búsqueda de respuestas basada en redundancia: un estudio para el Español y el Portugués*

Luis Villaseñor-Pineda[1], Manuel Montes-y-Gómez[1,2], Alejandro del Castillo[1]

[1] Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica, México
{villasen, mmontesg, delca}@inaoep.mx

[2] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, España

Abstract. Finding accurate information on the web has become a challenge due to the increment in the number of documents available on line. Current search engines retrieve relevant documents to general –often short– user queries, but fail extracting answers to simple and factual questions in natural language. This paper presents the basis of a statistical question answering system capable to find answers to factual questions from the web. This approach is supported on data redundancy rather than on sophisticated linguistic analysis of either questions or candidate answers. Preliminary results show that it is feasible to find concise and accurate answers from the web to factual questions made in Spanish and Portuguese languages.

1. Introducción

Gracias a la gran cantidad de documentos disponibles en Internet es posible satisfacer casi cualquier demanda de información, paradójicamente es precisamente esta gran cantidad de información el problema a enfrentar para encontrar aquella pieza de información deseada. Para abordar esta problemática se han desarrollado diversos métodos, los cuales podemos reunir en tres grandes áreas: la recuperación de información (RI), la extracción de información (EI) y, más recientemente, la búsqueda de respuestas (BR). Cada una de estas áreas aborda esta problemática de diferentes puntos de vista. Los sistemas de RI seleccionan y recuperan un conjunto de documentos a partir de las necesidades de información del usuario. Los sistemas de RI más conocidos son aquellos que localizan información a través de Internet, algunos de los "buscadores" más populares actualmente son Google, Altavista o Yahoo. Los sistemas de EI pretenden hallar información muy concreta en colecciones específicas de documentos. Esta información debe ser buscada y extraída para alimentar, por ejemplo, una base de datos y de esta forma poder ser procesada en forma automática. Por último, los sistemas de BR pretenden encontrar respuestas concisas a preguntas específicas en grandes volúmenes de información. A diferencia de los sistemas de RI, que ofrecen una lista de documentos relevantes dada una petición hecha por un usuario, un sistema de BR pretende entregar una respuesta concreta a una pregunta

* El presente trabajo fue financiado por el CONACYT México (Proyecto 43990A-1).El segundo autor agradece la beca otorgada por la Secretaría de Estado de Educación y Universidades de España.

precisa formulada por el usuario [10]. El presente trabajo se enfoca en la problemática de la búsqueda de respuestas y presenta los resultados preliminares alcanzados al aplicar una técnica basada en la redundancia de la información, es decir, hechos mencionados muchas veces y de varias maneras [1, 3]. La gran ventaja de esta técnica es la mínima dependencia en el idioma objetivo que se demuestra al aplicar dicha técnica tanto para el español como para el portugués.

La siguiente sección contrasta la técnica basada en redundancia con el enfoque tradicional de BR, además de delimitar el alcance de dicha técnica. Posteriormente se presentan los detalles para la reformulación de la pregunta, así como para el cálculo de la respuesta, tanto para el español como para el portugués. A continuación presentamos los resultados hasta ahora alcanzados y, en la última sección, de presentan las perspectivas y conclusiones de este trabajo.

2. Buscando una respuesta

Los primeros trabajos en el campo de búsquedas de respuestas (BR) adecuaron las técnicas de RI para la selección de pasajes susceptibles de contener la respuesta correcta. El rendimiento alcanzado por estos sistemas fue bastante bueno, sin embargo cuando se trato de localizar la respuesta concreta en tales pasajes estas técnicas presentaron un pobre rendimiento. El siguiente paso fue la aplicación de técnicas de Procesamiento del Lenguaje Natural (PLN) [5, 6, 8, 9]. Dado que las técnicas de PLN no son completas, una serie de trabajos, cada vez más sofisticados, han ido mejorando los sistemas de BR. Entre las herramientas de PLN utilizadas encontramos: etiquetadores léxicos, lematizadores, etiquetadores de entidades, analizadores sintácticos parciales, y hasta, complejas técnicas de análisis semántico y contextual (resolución de anáfora). A pesar de que con estas técnicas se han alcanzado resultados prometedores, se tienen dos grandes inconvenientes: (i) la construcción de tales herramientas lingüísticas es extremadamente costosa; y (ii) están fuertemente acopladas a un lenguaje humano específico.

El presente trabajo se basa en el trabajo desarrollado por Brill [1]. Este enfoque contrasta fuertemente con los sistemas anteriores pues no depende de costosas herramientas lingüísticas. La idea fundamental de este trabajo descansa en la suposición de que los términos usados para expresar la pregunta son los mismos términos que se usaron para escribir la respuesta. Es decir, dada una pregunta, el sistema genera una serie de reformulaciones con los términos usados en la pregunta – estas reformulaciones son simples manipulaciones de palabras. Muchas de estas reformulaciones coincidirán con ciertos extractos en la colección de documentos dada y a partir de éstos se obtiene la respuesta observando los términos más frecuentes. Por supuesto, mientras más grande sea la colección se tiene una mayor posibilidad de encontrar la respuesta correcta. Es así como tomamos ventaja de la redundancia de la información en la WEB (hechos mencionados muchas veces y de varias formas, es decir, múltiples ocurrencias de las respuestas). Cabe mencionar que esta idea también ha sido explorada por otros sistemas de BR [2, 7] con pequeñas variantes y siempre para el idioma inglés.

El presente trabajo presenta un estudio al aplicar este enfoque al español y al portugués, extendiendo el enfoque inicial de Brill. La diferencia principal radica en la reformulación de la pregunta. Básicamente el trabajo de Brill usa un lexicón para

determinar las partes de la oración y las variantes morfológicas de palabras claves. En nuestro caso, las reformulaciones no dependen de un lexicón y se basan solamente en la manipulación de las palabras de la pregunta, sin tener casi ningún conocimiento previo acerca de dichas palabras. A diferencia del trabajo de Brill, no se hace uso de ningún conjunto de patrones léxicos por tipo de pregunta, para extender las reformulaciones con palabras no contenidas en la pregunta original. En nuestro enfoque no hace uso de conocimiento externo, específico del idioma, y sólo se manipulan directamente las palabras de la pregunta, aplicando un método puramente estadístico para la selección de las respuestas. Así se intenta llevar a sus últimas consecuencias la aplicación del concepto de redundancia con lo que se obtiene una gran independencia del lenguaje usado.

Por ahora este enfoque sólo es posible usarlo en preguntas factuales y, también se encuentra limitado en el tipo de información a responder. Por el momento, sólo se pueden responder cuatro tipos de preguntas definidas por el pronombre o adverbio interrogativo utilizado para formular la pregunta. Esta categorización nos es indispensable para el cálculo de la respuesta la cual se sirve de elementos léxicos o tipográficos de los textos (véase el párrafo 3.3).

3. Módulos del Sistema

Los párrafos subsecuentes describen cada uno de los módulos del sistema de BR propuesto (véase la Fig. 1).

3.1 Reformulación de la Pregunta

Durante una primera etapa de experimentación se probó con todas las posibles reformulaciones de las preguntas, es decir, todas las permutaciones de sus palabras. Estos experimentos demostraron dos cosas: (i) que el esquema no es funcional para analizar preguntas con más de 5 palabras; (ii) que la gran mayoría de las reformulaciones construidas son inadecuadas, i.e. son sintácticamente incorrectas. A partir de estos resultados iniciales se seleccionó un conjunto de reformulaciones, aquellas con mejores resultados. Como es de imaginarse las mejores reformulaciones correspondieron a aquellas que presentan una estructura sintáctica correcta. Cabe mencionar, que esta primera etapa se realizó exclusivamente para el español, sin embargo, es presumible que las reformulaciones son aplicables a lenguas cuya gramática es similar, como el caso del portugués. Finalmente, se puede afirmar que dichas reformulaciones son las más comunes para escribir la respuesta a una pregunta dada.

Los siguientes párrafos presentan cinco reformulaciones posibles tanto para el español como para el portugués. En todos los casos el primer paso consistió en eliminar la primera palabra de la pregunta la cual, por lo general, es un adverbio o pronombre interrogativo. Para efectos de la exposición mostraremos las diferentes reformulaciones mediante estas dos preguntas ejemplo: “¿Quién obtuvo el premio Nóbel de la paz en 1992?” y “Qual o monte mais alto do mundo?”.

Primera reformulación: “bolsa de palabras”. Básicamente con esta reformulación obtenemos los mismos resultados que con un sistema de RI, así la búsqueda de extractos usa todas las palabras de la pregunta excluyendo las palabras vacías: (“obtuvo”, “premio”, “Nóbel”, “paz”, “1992”) (“monte” “mais” “alto” “mundo”).

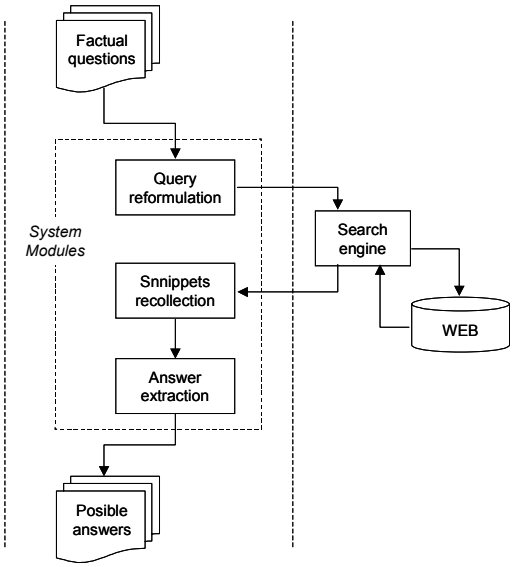

Fig. 1. Módulos del Sistema de Búsqueda de Respuestas

Segunda reformulación: "eliminación primera palabra". Entre las primeras observaciones al examinar una lista de preguntas factuales, notamos que, con frecuencia, inmediatamente después del pronombre o adverbio interrogativo se encuentra el núcleo verbal. Al colocar el verbo en posición final (o eliminarlo) es posible transformar la frase interrogativa a su forma declarativa. Es de suponer que dicha forma declarativa será abundante en los documentos analizados. Dado que no se desea utilizar ningún recurso lingüístico para determinar el verbo, se generan una serie de reformulaciones manipulando la primera palabra de la pregunta (después de eliminar la partícula interrogativa). Dado que en ciertas ocasiones es posible encontrar verbos auxiliares también se generarán reformulaciones manipulando la segunda palabra. Ejemplos de estas reformulaciones son: "obtuvo el premio Nóbel de la paz en 1992", "el premio Nóbel de la paz en 1992", "premio Nóbel de la paz en 1992", "el premio Nóbel de la paz en 1992 obtuvo", "premio Nóbel de la paz en 1992 obtuvo el"; "o monte mais alto do mundo", "monte mais alto do mundo", "monte mais alto do mundo o", "mais alto do mundo o monte".

Es interesante notar que gracias a este tipo de reformulaciones es posible recopilar extractos que usan sinónimos verbales, por ejemplo, extractos como "recibió el premio Nóbel de la paz en 1992" o "consiguió el premio Nóbel de la paz en 1992".

Tercera reformulación: "por componentes". En este caso, la pregunta es segmentada en componentes. Un componente es una expresión delimitada por preposiciones. A partir de combinaciones de estos componentes se construirán nuevas reformulaciones. Ejemplos de estas reformulaciones son: "obtuvo el premio Nóbel de la paz en 1992", "obtuvo el premio Nóbel en 1992 de la paz", "de la paz obtuvo el premio Nóbel en 1992", "de la paz en 1992 obtuvo el premio Nóbel", "en 1992 obtuvo el premio Nóbel de la paz", "en 1992 de la paz obtuvo el premio Nóbel",

"obtuvo el premio Nóbel" "de la paz" "en 1992"; "o monte mais alto do mundo", "do mundo o monte mais alto", "o monte mais alto" "do mundo".

Es evidente que en algunos casos la reformulación no tiene sentido ("en 1992 de la paz obtuvo el premio Nóbel") y no habrá extractos resultantes, sin embargo en otros casos ("en 1992 obtuvo el premio Nóbel de la paz"), la reformulación será apropiada para la recolección de extractos relevantes.

Cuarta reformulación: "por componentes excluyendo la primera palabra". Como vimos en la segunda reformulación, generalmente la primera palabra es un verbo. En esta ocasión repetimos la tercera reformulación eliminando la primera palabra. Ejemplos de este tipo de reformulaciones son: "el premio Nóbel de la paz en 1992", "el premio Nóbel en 1992 de la paz", "de la paz el premio Nóbel en 1992", "de la paz en 1992 el premio Nóbel", "en 1992 el premio Nóbel de la paz", "en 1992 de la paz el premio Nóbel", "el premio Nóbel" "de la paz" "en 1992"; "monte mais alto do mundo", "do mundo monte mais alto", "monte mais alto" "do mundo".

Quinta reformulación: "por componentes excluyendo las dos primeras palabras". En este caso, suponemos la presencia de un verbo auxiliar. Ejemplos de reformulaciones de este tipo son: "premio Nóbel de la paz en 1992", "premio Nóbel en 1992 de la paz", "de la paz premio Nóbel en 1992", "de la paz en 1992 premio Nóbel", "en 1992 premio Nóbel de la paz", "en 1992 de la paz premio Nóbel", "premio Nóbel" "de la paz" "en 1992"; "mais alto do mundo", "do mundo mais alto", "mais alto" "do mundo".

Como puede observarse, las reformulaciones son sencillas manipulaciones de los términos de la pregunta, que finalmente tratan de aprovechar cierta estructura sintáctica presente en las preguntas factuales –estructura muy similar tanto para el español como para el portugués. Por supuesto, estás reformulaciones son ciegas y se aplican de manera indiscriminada. Esto provoca que muchas reformulaciones no tengan sentido, en cuyo caso es poco probable la recopilación de extractos de interés. Sin embargo, en otros casos la reformulación coincidirá con alguno o varios documentos con la consecuente recopilación de extractos apropiados.

3.2 Recolección de extractos

Este modulo toma las reformulaciones anteriores y lanza las búsquedas sobre la Web apoyándose en algún motor de búsqueda ya existente. En nuestro caso, está recopilación de extractos se realiza usando Google. Por ejemplo, para el español usando el primer tipo de reformulación se obtuvieron extractos como:

> *Edición Especial Aniversario - 30 Años*
> … Ganador del premio Nobel de la Paz (1993). 7.- Rigoberta Menchu (1959) Lider indígena guatemalteca, recibió el premio Nobel de la Paz en 1992 …

De igual forma usando la primera reformulación para el portugués se obtuvieron extractos como éste:

> *Canal Kids - Viagem - Você Sabia? - A Corrida ao Teto do Mundo*
> … O monte Everest, na fronteira do Nepal e do Tibete, e o teto do mundo Não existe montanha mais alta no planeta e chegar la no alto sempre foi um sonho …

3.3 Cálculo de la respuesta

Después de obtener para cada reformulación un conjunto de extractos se calculan las frecuencias de los términos. Para ello se calculan los n-gramas para n={1..5} considerando los signos de puntuación como límites de frase.

Posteriormente se obtiene una lista con cinco respuestas candidatas ordenadas en función de su frecuencia, es decir, el término, o términos, con mayor presencia será el primero en considerarse como la respuesta correcta. Por supuesto, es necesario aplicar una serie de criterios para determinar con mayor precisión la respuesta correcta. Hasta ahora el método más eficaz ha sido el método de *frecuencia compensada con expresiones regulares*. Para una exposición detallada de los diferentes métodos evaluados véase [4].

El método de frecuencia compensada con expresiones regulares filtra los 20 1-gramas más frecuentes bajo criterios tipográficos (mes del año, palabras con mayúscula inicial, números, etc.) usando expresiones regulares. A partir de estos 1-gramas se obtienen todos los n-gramas, con n={2..5}, compuestos de estos 1-gramas. Posteriormente las frecuencias de los n-gramas se suman. De esta manera, una expresión de cinco términos que claramente por su longitud tendrá una frecuencia relativa pobre se verá mejorada al compensarla con las frecuencias relativas de los 2, 3 y 4-gramas que la conforman. La Tabla 1 muestra los resultados de las respuestas candidatas de nuestras dos preguntas ejemplo. Un peso alto significará que se tiene una mayor presencia de dicha secuencia palabras, así como las subsecuencias de palabras contenidas.

Rigoberta Menchu	0.07418	*Everest*	0.00843
Rigoberta Menchu Tum	0.05753	Antes Everest	0.00733
Menchu	0.05541	Buenos Aires	0.004666

Tabla 1. Respuestas candidatas con el método de frecuencia compensada.

4 Resultados del estudio

Para evaluar el comportamiento de nuestro sistema se utilizaron dos métricas de evaluación comúnmente usadas para evaluar los sistemas de BR: (i) the Mean Reciprocal Rank (MRR) y (ii) la precisión. Nuestro sistema arroja como resultado una lista ordenada de cinco posibles respuestas.

Para calcular el MRR, a cada pregunta se le asigna una calificación que corresponde al inverso de la posición donde se encuentra la respuesta correcta entre las cinco opciones arrojadas por el sistema. Si la respuesta correcta no está entre esas cinco opciones, a la pregunta se le asigna una calificación de cero. Finalmente las calificaciones de todas preguntas son promediadas para obtener el MRR.

Para calcular la *precisión,* se obtiene el porcentaje de preguntas para las cuales la respuesta correcta está entre las cinco respuestas posibles arrojadas por el sistema.

El presente estudio se evaluó usando un corpus de cuarenta preguntas para el español y otro corpus de cuarenta preguntas para el portugués. En ambos casos se tocaron muy diversos temas y se consideraron únicamente preguntas factuales. También se limitó la variedad de preguntas en función del adverbio o pronombre interrogativo usado, por ahora sólo es posible hacer preguntas del tipo: quién, cuándo, dónde, cuál (y para el portugués: quem, quando, onde, qual). Ejemplo de estas

preguntas son: ¿Quién es el Gobernador del Banco de México?, ¿Cuándo fue lanzado el Apolo 11?, ¿Dónde nació Pitágoras?, ¿Cuál fue el nombre real de Marilyn Monroe?, Onde era o campo de concentração de Auschwitz?, Qual a antiga capital da Polónia?, Quando foi a independência de Cabo Verde?, Quem é a ministra sueca do ambiente?

Para cada una de las preguntas se generaron las reformulaciones según se explicó en el párrafo 3.1. A partir de cada una de las reformulaciones se recopilaron extractos de interés, a partir de los cuales se calculó una lista ordenada de cinco posibles respuestas. Actualmente el sistema se limita a recopilar como máximo cincuenta extractos para cada reformulación, por supuesto, en algunos casos, cuando la reformulación no tiene sentido, no se recuperó ni un solo extracto. Por razones de eficiencia y dada la volatilidad de los datos en la Web estos extractos se recopilaron una sola vez.

La Tabla 2 muestra los resultados hasta ahora obtenidos usando el método de frecuencia compensada sobre cada tipo de reformulación. Cada columna corresponde a la aplicación del mismo método de cálculo de respuesta sobre diferentes conjuntos de extractos recuperados, dependiendo del tipo de reformulación usado.

TIPO DE PREGUNTA	Bolsa de palabras		Eliminación primera palabra		Componentes		Componentes excluyendo una palabra		Componentes excluyendo dos palabras	
IDIOMA	*ESP*	*POR*	*ESP*	*POR*	*ESP*	*POR*	*ESP*	*POR*	*ESP*	*POR*
Quién / Quem	90%	**70%**	**100%**	50%	80%	10%	**100%**	30%	**100%**	**70%**
Cuándo / Quando	**70%**	10%	50%	10%	10%	10%	40 %	**20%**	50%	10%
Dónde / Onde	**100%**	40%	**100%**	**70%**	30%	0%	70%	50%	60%	40%
Cuál / Qual	80%	10%	70%	**50%**	40%	30%	80%	**50%**	**90%**	40%
Precisión	**85%**	32%	80%	**45%**	40%	12%	73%	37%	75%	40%
MRR	0.6821	0.275	**0.7175**	**0.4333**	0.4	0.083	0.6404	0.3125	0.6542	0.3395

Tabla 2. Resultados para el español y el portugués

Como puede observarse los resultados son alentadores, sin embargo, las diferencias entre el portugués y el español son significativas. Esto puede explicarse por la cantidad de documentos presentes en la Web en ambos idiomas. La Tabla 3 muestra el promedio de reformulaciones realizadas, así como el promedio de extractos recuperados para ambos idiomas.

	Bolsa de palabras		Eliminación primera palabra		Componentes		Componentes excluyendo una palabra		Componentes excluyendo dos palabras	
IDIOMA	*Promedio reform.*	*Promedio Extractos*	*Promedio reform*	*Promedio Extractos*	*Promedio reform*	*Promedio Extractos*	*Promedio reform*	*Promedio Extractos*	*Promedio reform*	*Promedio Extractos*
Español	1	50	5	26.8	2.5	12.4	2.5	27	2.4	28.8
Portugués	1	40	5	10.13	3.4	2.4	3.4	9.5	3.4	16.8

Tabla 3. Información recuperada de la Web para el español y el portugués

De esta manera, se confirma la diferencia entre ambos idiomas en la Web, entre 2 o 3 veces un mayor número de extractos para el español que para el portugués. Por supuesto, a mayor cantidad de extractos una mejor precisión.

5 Conclusiones

Este estudio ha demostrado las enormes posibilidades de este tipo de técnicas estadísticas cuya gran ventaja es el poco o nulo uso de costosos recursos lingüísticos. Con la enorme perspectiva de ser usado independientemente del idioma. Por otro lado, este tipo de técnicas son útiles donde se tienen grandes cantidades de documentos con un cierto grado de redundancia. Es gracias a esta redundancia que las respuestas pueden ser ubicadas a través de simples reformulaciones de la pregunta. Dado que es un hecho que la Web seguirá siendo un enorme corpus con alto grado de redundancia estas técnicas podrán aportar mucho en el campo de la BR.

Por otro lado, en los experimentos realizados hasta ahora podemos concluir que no existe un par reformulación/cálculo de la respuesta capaz de resolver todos los tipos de pregunta. La idea a futuro es utilizar el mejor par posible dependiendo del tipo de pregunta. Por supuesto, otro trabajo pendiente es la ampliación de este estudio preliminar a un conjunto mayor de preguntas, en particular en colecciones controladas como las utilizadas en el CLEF.

Otro de los aspectos inmediatos a explorar es la relación entre el número de extractos examinados y la precisión en la búsqueda de las respuestas. Lo cual estaría orientado a establecer un criterio con respecto al nivel de redundancia necesario para determinar un adecuado comportamiento del sistema.

References

[1] E. Brill , J. Lin, M. Banko, S. Dumais, A. Ng (2001). *Data-intensive question answering*. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).

[2] S. Buchholz (2001). *Using grammatical relations, answer frequencies and the World Wide Web for TREC question answering*. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).

[3] C. Clarke, G. Cormarck, T. Lynam (2001). *Exploting redundancy in question answering*. In Proceedings of SIGIR' 2001.

[4] A. Del-Castillo, M. Montes-y-Gómez, L. Villaseñor-Pineda (2004). *QA on the Web: A Preliminary Study for Spanish Language.* International Conference on Computer Science. ENC 2004, (en prensa).

[5] S. Harabagiu , D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, P. Morarescu (2000). *FALCON : Boosting knowledge for question answering*. In Proceedings of the Ninth Text REtrieval Conference

[6] E. Hovy, U. Hermajakob, C. Lin (2001). *The use of external knowledge in factoid QA*. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).

[7] C. K. Kwok , O. Etzioni and D. Weld(2001). *Scaling question answering to the Web*. In Proceedings of WWW-01.

[8] J. Prager, E. Brown, A. Coden, D. Radev (2000). *Question Answering by Predictive Annotation*. In Proceedings of SIGIR'2000.

[9] M. Vargas-Vera, E. Motta (2004). *AQUA- Ontology-Based Question Answering System*. In MICAI-2004, LNCS 2972, Springer.

[10] J. L. Vicedo (2004). *La Búsqueda de Respuestas: Estado Actual y Perspectivas de Futuro*. Revista Iberoamericana de Inteligencia Artificial. Vol. VIII, No.22.