Funciones de Cercanía Semántica para la Validación de Respuestas en la Web

Antonio Juárez-González, Manuel Montes-y-Gómez & Luis Villaseñor-Pineda
Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Sta. María Tonanzintla, C.P. 72840 Puebla, México
{antjug, mmontesg, villasen}@inaoep.mx

Resumen

En este documento se presenta un método de validación de respuestas que aprovecha la información disponible en la Web. El método se basa en la idea de que la respuesta correcta a una pregunta debe ser semánticamente más cercana a los términos de ésta que una respuesta incorrecta. Se proponen cuatro funciones distintas para medir la cercanía semántica de una respuesta y una pregunta en la Web. Los resultados obtenidos confirman nuestra hipótesis, y apuntan hacia el uso de la Web como corpus textual para la validación de respuestas. Sin embargo, aún falta experimentar para encontrar la mejor función según sean los tipos de pregunta y respuesta esperada.

1. Introducción

Hoy en día existe una gran cantidad de textos digitales accesibles desde la red. Estos textos contienen información de prácticamente cualquier tema, por lo que cualquier necesidad de información puede satisfacerse realizando una búsqueda en la Web.

Los motores de búsqueda, como Google y Yahoo, son el mecanismo más usado para acceder a la información en la Web. Estos sistemas permiten encontrar documentos relevantes a una necesidad de información general expresada por un usuario, pero son incapaces de devolver una respuesta concisa para una necesidad de información concreta. Cuando este es el caso, la alternativa más apropiada son los llamados sistemas de *Búsqueda de Respuestas* (BR).

Los sistemas de BR son capaces de responder preguntas formuladas por los usuarios en un lenguaje natural [2]. Hasta el momento, estos sistemas sólo pueden dar respuesta a preguntas factuales, es decir, preguntas cuya respuesta es una entidad nombrada, una fecha o alguna cantidad. Por ejemplo, para la pregunta "¿Quién es el presidente de México?", la respuesta de un sistema de BR es la entidad nombrada "Vicente Fox Quesada".

Típicamente los sistemas de BR obtienen un conjunto de posibles respuestas a una pregunta. Gracias a un proceso de ordenamiento se determina la respuesta más probable. Esta respuesta es entregada al usuario como resultado. Sin embargo, en muchas ocasiones no se tienen los elementos suficientes para determinar la respuesta correcta a pesar de pertenecer al conjunto de respuestas candidatas. Un método ampliamente usado por los sistemas actuales de BR es la triangulación de la información. Es decir, utilizar una fuente de información adicional con la cual re-ordenar las respuestas candidatas [1, 6, 7]. En este trabajo se propone el uso de la Web como fuente de información alterna. En particular se experimenta con algunas funciones de similitud y distancia que permiten medir la cercanía semántica entre los términos de la pregunta y las respuestas, considerando únicamente su frecuencia de ocurrencia en la Web. La manera de combinar estas frecuencias se basa en el concepto general de similitud [5], y en las ideas planteadas recientemente por Cilibrasi y Vitanyi [3].

El resto del documento está organizado de la siguiente manera. La sección 2 define las funciones de cercanía semántica usadas, la sección 3 describe el método propuesto para la validación y reordenamiento de las respuestas candidatas, la sección 4 muestra los resultados obtenidos para la colección de preguntas del CLEF 2004, por último, la sección 4 presenta nuestras conclusiones.

2. Funciones de Cercanía Semántica

En las ciencias de la computación la *similitud* es un concepto fundamental y ampliamente usado. Existen muy diversas maneras de medir la similitud entre dos objetos cualesquiera, sin embargo, la gran mayoría de éstas la definen como la razón entre la información común y la información total contenida en ambos objetos [5].

Con base en esta definición se proponen las siguientes funciones para medir la relación semántica entre una pregunta dada y una posible respuesta considerando la Web como repositorio general de información. En todas ellas se aplica la siguiente notación: X denota el conjunto de palabras de la pregunta, y la respuesta candidata, $f(X)$ el número de páginas Web que contienen las palabras en X, $f(y)$ el número de páginas que citan la respuesta candidata, y $f(X \cup y)$ las páginas que contienen ambas.

$$s(X,y)=\frac{f(X\cup y)}{f(X)+f(y)-f(X\cup y)} \quad (1)$$

$$s(X,y)=\frac{f(X\cup y)}{\max\{f(X),f(y)\}} \quad (2)$$

$$s(X,y)=\frac{f(X\cup y)}{\min\{f(X),f(y)\}} \quad (3)$$

Las formulas 1 y 2 expresan una función de similitud, donde $s(X, y) = 1$ es un indicio de un fuerte vinculo entre la pregunta y la respuesta. Por su parte, la formula 3 define la cercanía semántica entre X e y como una función de contención. En este caso, dado que generalmente se tiene $f(X) << f(y)$, pues la pregunta es más especifica que la respuesta, $s(X, y) = 1$ simplemente indica que las páginas que contienen los términos de la pregunta tienden a contener la respuesta, pero no necesariamente viceversa. En otras palabras, la función 3 representa una manera más laxa de interpretar la relación semántica entre la pregunta y la respuesta.

Adicionalmente se aplicó una métrica de distancia que permite establecer la relación semántica entre dos objetos a partir de la información disponible en la Web [3]. Esta métrica, al igual que las funciones 1 y 2, es estricta, dado que la condición $s(X, y) = 0$ se presenta cuando las páginas vinculadas con la pregunta y la respuesta son las mismas, mientras que $s(X, y) \rightarrow \infty$ cuando $f(X\cup y) \rightarrow 0$.

$$s(X,y)=\frac{G(X\cup y)-\min\{G(X),G(y)\}}{\max\{G(X),G(y)\}} \quad (4)$$

$$G(i)=\log\frac{M}{f(i)}$$

En esta métrica, el factor de normalización M representa el tamaño de la Web, es decir, el número de páginas indexadas por el buscador utilizado. En los experimentos presentados en la sección 4 el valor de M se definió como el número de páginas que incluyen la preposición "*de*"; palabra más común del español.

3. Validación de Respuestas

La validación de respuestas se realiza mediante un sistema que consta de los siguientes tres módulos:

- Generación de peticiones
- Medición de la cercanía semántica
- Reordenamiento de las respuestas

La figura 1 muestra un diagrama de bloques del sistema desarrollado.

El módulo de generación de peticiones recibe como entrada la pregunta realizada por el usuario al sistema de BR, así como la lista de respuestas candidatas obtenidas por éste. Para cada una de las parejas pregunta-respuesta se generan tres peticiones: una construida mediante las palabras no vacías de la pregunta (X), otra conformada exclusivamente por las palabras de la respuesta (y), y finalmente una tercera obtenida por la concatenación de las dos anteriores ($X\cup y$). Por ejemplo, si se tiene la pregunta "*¿Cuál es el nombre de pila del juez Borsellino?*", y la respuesta "*Paolo"*, entonces se generarían las siguientes tres peticiones:

1. +"pila" +"juez" +"Borsellino"
2. +"Paolo"
3. +"pila" +"juez" +"Borsellino" +"Paolo"

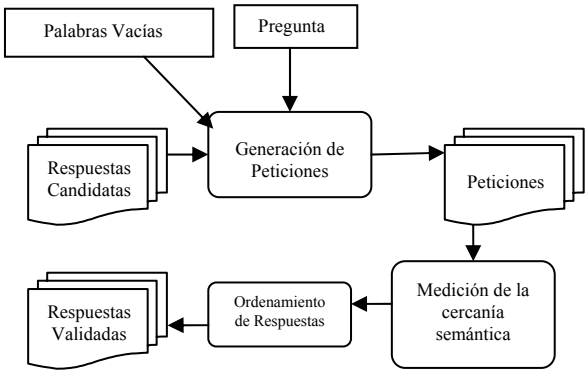

Figura 1. Sistema de Validación de Respuestas

El segundo módulo se encarga de la medición de la cercanía semántica entre la pregunta y cada una de las respuestas candidatas. Para ello envía las peticiones generadas a la Web, usando algún buscador de referencia (en nuestro caso Google), y después combina los resultados de la búsqueda mediante las funciones de similitud y distancia descritas en la sección 2.

Una vez calculada la cercanía semántica de cada respuesta candidata con la pregunta, éstas simplemente se reordenan de mayor a menor (o menor a mayor, cuando se usa la función 4), y la mejor respuesta se entrega al usuario.

Cabe señalar que el sistema de BR es un módulo completamente independiente al sistema de validación de respuestas, pues este último hace caso omiso de cualquier tipo de orden o ponderado de las respuestas candidatas generadas por el primero.

3. Resultados Experimentales

En los experimentos se usaron las 200 preguntas en español correspondientes al corpus del CLEF[1] 2004. Las respuestas candidatas se obtuvieron aplicando un sistema de BR basado en pasajes [4]. La tabla 1, en su segunda columna, muestra los resultados obtenidos por dicho sistema.

[1] Cross Language Evaluation Forum.

Para la validación de respuestas se usaron las cinco primeras respuestas entregadas por el sistema de BR. La tabla 1, en las columnas 3-6, muestra los resultados obtenidos empleando las cuatro funciones de similitud y distancia descritas en la sección 2. Cabe resaltar que, dado que el propósito del experimento fue evaluar la validación de respuestas en la Web y no el sistema de BR empleado, solamente se consideraron las 104 preguntas con alguna respuesta candidata, además de que se aseguró la existencia de la respuesta correcta en la lista de respuestas candidatas. En los casos necesarios se agregó manualmente la respuesta correcta a dicha lista.

Tabla 1. Resultados de la Validación de Respuestas

Preguntas	BR	(1)	(2)	(3)	(4)
Sin respuesta	96				
Con respuestas candidatas	104				
Con respuesta correcta	31	53	57	39	60
Precisión	.29	.50	**.**54	.37	**.57**

La medida de evaluación usada fue la precisión, definida como el porcentaje de preguntas contestadas correctamente del conjunto de preguntas con alguna respuesta candidata.

Los resultados obtenidos son contundentes respecto a la conveniencia del uso de la Web para validar las respuestas candidatas de un sistema de BR. Con todas las funciones empleadas se obtuvieron mejores precisiones que la del sistema de BR. En particular, la medida de distancia (fórmula 4) duplicó la precisión de referencia.

Un caso interesante es la función 3, la cual mide la similitud de una manera más laxa. Su pobre mejoría sobre el sistema de referencia hace suponer que para establecer una relación semántica entre un par pregunta-respuesta es necesario distinguir una fuerte correlación entre ellas, y no basta con apreciar solamente una relación unidireccional, por más fuerte que ésta sea.

Finalmente es importante mencionar que las funciones utilizadas parecen ser complementarias. Los resultados obtenidos indican que combinando las cuatro funciones se pueden contestar correctamente 83 preguntas, es decir, un 79% de las preguntas con respuestas candidatas.

5. Conclusiones

El esquema propuesto para la validación de respuestas es simple, intuitivo y eficiente. Los resultados obtenidos demuestran que la precisión de un sistema de BR puede incluso duplicarse validando –ordenando– sus respuestas candidatas mediante la información contenida en la Web. En particular, los experimentos realizados nos permitieron concluir lo siguiente: (i) la aplicación del método en la Web es factible, (ii) la Web, por su tamaño y diversidad de contenidos, resulta ser una colección adecuada para la validación de respuestas a preguntas de todo tipo, y (iii) la Web, también por su gran tamaño, requiere de la aplicación de medidas estrictas de similitud para determinar adecuadamente la relación semántica entre dos conceptos, dígase, un par pregunta-respuesta.

Además el esquema propuesto es muy flexible, pues puede usarse con cualquier lenguaje, e incluso considerando cualquier conjunto de documentos como repositorio de referencia.

En el trabajo se usaron cuatro funciones para medir la cercanía semántica entre una pregunta y una respuesta candidata. Los resultados, aunque prometedores, aun no son suficientes para concluir sobre la validez de las funciones propuestas. Sin embargo, algo interesante es que las funciones utilizadas parecen ser complementarias (i.e., el conjunto de respuestas correctas no fue el mismo). Posiblemente a través de aprendizaje automático podamos aprovechar el potencial de cada función, y seleccionar la mejor de ellas para cada una de las preguntas.

Agradecimientos

Agradecemos al CONACYT por su apoyo económico (beca xxx y proyecto 43990) y al comité organizador del CLEF por los recursos textuales facilitados.

6. Referencias

[1] E. Brill, J. Lin, M. Banko, S. Dumais and A. Ng. “Data-intensive question answering”. In TREC 2001, 2001.

[2] Burger, J. et al. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST 2001.

[3] Cilibrasi R., P. Vitanyi, Automatic meaning discovery using Google, Manuscript CWI, 2004. http://arxiv.org/abs/cs.CL/0412098

[4] Montes-y-Gómez M., Villseñor-Pineda L., Pérez-Coutiño M. A., Gómez-Soriano J. M., Sanchis-Arnal E., Rosso P.. The INAOE and UPV Joint Participation at CLEF 2005. To appear in the proceedings of the CLEF 2005 Workshop.

[5] Lin, An Information-Theoretic Definition of Similarity, Proc. of the International Conference on Machine Learning, Madison, Wisconsin, 1998.

[6] Sisay Fissaha Adafre, Willem Robert van Hage, Jaap Kamps, Gustavo Lacerda de Melo and Maarten de Rijke. The University of Amsterdam at CLEF 2004. Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK, 2004.

[7] Vicedo, J.L., Izquierdo R., Llopis F. and Muñoz R., Question Answering in Spanish. CLEF 2003 Workshop, Springer-Verlag, 2003.