Evaluación de heurísticas para un Sistema de Recuperación de Pasajes

Ernesto Miñón, Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
ernesto_minon@yahoo.com.mx, hgimenezs@gmail.com

Resumen En este trabajo presentamos resultados preliminares sobre la selección de los mejores pasajes a partir de un conjunto de pasajes recuperados. Nuestro enfoque es el uso de algunas heurísticas basadas en el análisis de preguntas factuales. Estas heurísticas consideran las distancias entre el *pivote* de la pregunta y términos que ocurren en la pregunta y en los pasajes recuperados. Se presentan resultados de la evaluación sobre un subconjunto de CLEF 2004.

1 Introducción

El creciente volumen de información en la web, así como las nuevas exigencias de los usuarios han superado a los ofrecimientos de los sistemas existentes al tratar de localizar información relevante y puntual en grandes volúmenes de documentos. Esto ha provocado la aparición de nuevas herramientas, así como de un nuevo campo de investigación llamado Búsqueda de Respuestas (BR). Tales sistemas normalmente consideran las siguientes etapas en su desarrollo: (i) Análisis de la pregunta, (ii) Recuperación de documentos relacionados, (iii) Selección de pasajes relevantes, y (iv) Extracción de la respuesta [1] [2] [3].

Como se observa la selección de pasajes involucra la previa extracción de los mismos, además es una parte medular de algunos sistemas de búsqueda de respuestas debido a que permiten la localización de la respuesta en el documento de forma más eficiente. Por ello nuestro interés en trabajar en un sistema de Recuperación de Pasajes (RP) orientado a BR para la lengua Española, que nos permita extraer fragmentos contiguos de texto y que contengan la respuesta esperada a una consulta formulada en Lenguaje Natural (LN), teniendo en cuenta que se pretende desarrollar un sistema práctico y sencillo.

Actualmente existen diferentes paradigmas de RP, el primero de ellos tiene un buen desempeño al integrar técnicas de Procesamiento de Lenguaje Natural (PLN) [4][5][6], sólo que existe el inconveniente de la dificultad para adaptarlo a tareas multilingüe; es decir en este tipo de sistemas se observa una alta dependencia con el lenguaje. Otra aproximación es una adecuación del Modelo de Espacio Vectorial (MEV) [7][8][9], la cual no aborda una problemática específica de BR. Para atacar está problemática se procedío a realizar un análisis en el

corpus CLEF[1], con la finalidad de conocer el comportamiento en las preguntas, y de esta forma elaborar una propuesta para la obtención de pasajes relevantes. El análisis de la pregunta, y la propuesta para la selección de pasajes son las dos secciones siguientes. Posteriormente, se presentan los resultados del análisis y las conclusiones.

2 Análisis de la pregunta

La presente etapa del proyecto es la parte medular del mismo; es decir, si el análisis se realiza de manera incorrecta los pasajes obtenidos serán irrelevantes. Cabe resaltar que en el presente trabajo se presenta un avance preliminar a la que sería la propuesta final del proyecto, se pretende concluir con la propuesta final contemplando todas las preguntas factuales del *corpus* CLEF. El análisis consistió en proponer diversas medidas entre pregunta y pasaje tomando como base un *pivote*. El pivote es el punto de referencia que permite calcular la distancia de cada término t en la pregunta. Intuitivamente, la respuesta a una pregunta contendrá buena parte de los términos que constituyen la pregunta. Aparte del uso de sinónimos, es esencial que el pivote ocurra en la respuesta. Así, la base de las heurísticas es medir la similitud entre la pregunta y el pasaje considerando, a través de la distancia, cierta libertad a los términos del pasaje.

El criterio para la obtención del pivote surgió del análisis de las preguntas del *corpus* CLEF. Para obtenerlo es necesario determinar el verbo principal de la pregunta, posteriormente tomamos el sustantivo más cercano al verbo en la frase nominal. Se pretende que la heurística permita obtener pasajes utilizando el contexto de la respuesta. Las etapas que conforman el análisis de la pregunta son las siguientes:

- Lematización. Permite determinar las partes del discurso de la pregunta, para lograrlo hacemos uso del sistema AGME [10].
- Obtención del pivote ρ en una determinada pregunta q.
- Obtención de la distancia del *keyword* en la pregunta $d(kw)_q$, para ello utilizamos la siguiente heurística:

$$d(kw)_q = \frac{1}{\delta(kw, \rho)}, kw \in q, \tag{1}$$

  donde $\delta(kw, \rho)$ es la distancia de kw respecto al pivote ρ; número de palabras más uno entre estas palabras.
- Obtención de la distancia de *stopword* respecto a la consulta q:

$$d(sw)_q = \frac{1}{\delta(sw, \rho)}, sw \in q. \tag{2}$$

[1] *Cross Language Evaluation Forum.*

3 Algoritmo de selección de pasajes

Una vez que se realiza el análisis de la pregunta, se procede a recuperar los pasajes que contienen el pivote obtenido en el análisis de la pregunta p. En cada pasaje p se reusa ρ convirtiéndose en nuestro punto de referencia que permitirá calcular la distancia de cada término t en el pasaje. Así, se calcula la distancia de kw en el pasaje p:

$$d(kw)_p = \frac{1}{\delta(kw, \rho)}, kw \in p, \tag{3}$$

de igual manera obtenemos la distancia de los *stopwords* en el pasaje:

$$d(sw)_p = \frac{1}{\delta(sw, \rho)}, sw \in p. \tag{4}$$

Se observa que entre mayor sea la distancia entre el término y el pivote la densidad será menor. Una vez que obtenemos la distancia correspondiente a cada *keyword* y *stopword* en la consulta y el pasaje, se procede a obtener la distancia conjunta entre el *keyword* de la pregunta y el del pasaje $d_{kw}(q, p)$, así como la distancia conjunta entre el *stopword* de la pregunta y el del pasaje $d_{sw}(q, p)$, como se indica en las siguientes expresiones:

$$d_{kw}(q, p) = d(kw)_q \cdot d(kw)_p, \ (kw \in q)\&(kw \in p), \tag{5}$$

$$d_{sw}(q, p) = d(sw)_q \cdot d(sw)_p, \ (sw \in q)\&(sw \in p). \tag{6}$$

Por último se obtiene la suma de las distancias aportadas por cada *keyword* en la pregunta y el pasaje $d_{kw}(q, p)$, así como las distancias aportadas por cada *stopword* en la pregunta y el pasaje $d_{sw}(q, p)$. De esta forma se obtiene la similitud entre la pregunta y el pasaje:

$$sim(q, p) = \sum_r d_r(q, p) + \sum_t d_t(q, p). \tag{7}$$

Una mayor similitud indica que la distancia entre los términos de la consulta y el pasaje es menor, es decir que los términos se encuentran más cercanos entre sí. Como observamos el cálculo de similitud nos generará una serie de pasajes con un *ranking*, donde el de mayor puntuación será seleccionado como el pasaje respuesta.

4 Resultados

La colección de documentos utilizada fue el *copus* CLEF[2], el cual está conformado por un conjunto de 454,045 noticias obtenidas de la *EFE News Agency*, las cuales fueron escritas en 1994 y 1995. Se cuenta con 200 preguntas supervisadas, clasificadas de la siguiente forma: preguntas factuales, de definición y nulas (aquellas que no tienen respuesta en la colección). Para cada una de ellas

[2] *Corpus* amablemente proporcionado por el INAOE.

existe una referencia hacia un identificador del documento y la noticia adonde se encuentra la respuesta, esto nos permite entrenar al sistema a desarrollar. Para nuestro caso de estudio tomamos una muestra de 17 preguntas factuales, repartidas entre diferentes tipos (*where*, *what*, etc.).

Debido a que el sistema de RP está en construcción, se procedió a probar las heurísticas en forma manual. Para ello se extrajo una muestra de preguntas al azar tratando de cubrir los diferentes tipos de ellas. Los resultados que se obtuvieron durante el experimento, fueron los siguientes. Dado un conjunto de 17 preguntas factuales, se obtuvo el pasaje correcto para 11 de ellas (64.7%), los pasajes obtenidos para las 6 preguntas restantes no contenian la respuesta correcta (35.3%). A continuación se muestran algunos resultados obtenidos.

Pregunta 0002. ¿Cuánto aumenta la población mundial cada año? Respuesta (EFE19940601-00641; 1,6 por ciento)

Pregunta 0026. ¿Cuánto costó el Túnel del Canal? Respuesta (EFE19940226-15738; 10.000 millones de libras; 15.000 millones de dólares)

Pregunta 0066. ¿Qué transbordador estadounidense llevó por primera vez un astronauta ruso a bordo? Respuesta (EFE19950131-18141; El Discovery)

Como una muestra de los resultados obtenidos se presentan las tablas 1, 2 y 3. En la tabla 1 se observa que el pasaje obtenido como más relevante responde la pregunta planteada, el segundo pasaje también lo hace pero de manera indirecta. La tabla 2 muestra los pasajes obtenidos, los cuales no contienen la respuesta esperada por CLEF; esta es una pregunta difícil debido a que supone conocimiento sobre el contexto. Para el caso de la tabla 3, el pasaje obtenido contiene la respuesta correcta, aunque no es el pasaje indicado por CLEF.

5 Conclusiones

Aún cuando no es posible dar resultados concluyentes de las heurísticas practicadas, tenemos confianza en que ellas puedan arrojar resultados prometedores, ya que está previsto utilizar las heurísticas para enfrentar las preguntas nulas. Además las heurísticas pueden ser adaptadas a los diferentes tipos de preguntas dentro de la clase de factuales. También, está en curso la mejora de la heurística; específicamente estamos analizando el efecto del pivote en cada uno de los tipos de pregunta factual. Con lo anterior hemos proyectado alcanzar un porcentaje alto de pasajes correctos. Por otro lado, el sistema tiene una base muy simple y natural, además es bastante "ligero" en el sentido que no utiliza grandes recursos; solamente se basa en un lematizador y el cálculo de la similitud (ec. 7).

El siguiente paso en este proyecto es conocer la efectividad de las medidas aquí presentadas en todo el conjunto de preguntas de CLEF y confrontarlo con otros sistemas de RP.

Table 1. Pasajes ordenados por *ranking* correspondientes a la pregunta 0002.

Ranking	**Pasaje Obtenido**
0.611	La población mundial aumenta a tasas anuales del 1,6 por ciento y, segnún los cálculos de la ONU, pasar de los actuales 5.600 millones de habitantes a 6.260 en el año 2000 y a 8.500 millones en el 2025.
0.5833	Según Buettner, el 12 de mayo vivían en el planeta 5.650.379.690 personas, si es que es posible calcular con exactitud algo así como la población mundial, que aumenta a un ritmo diario de 260.000 personas.
0.566	Por último, Wirth subrayó que lo que está en juego es el futuro del mundo, pues, si el crecimiento actual se mantiene, cada diez años la población mundial aumentará el equivalente a China, y a mediados del siglo que viene se podrían haber alcanzado los 12.500 millones de seres humanos, más del doble de los 5.700 actuales.EFE
0.0625	Hace una semana, la prensa italiana se sorprendió ante las afirmaciones de ese documento, que sugería, basándose en datos de la ONU, la necesidad de que aumentase la población mundial a razón de una media de 2,3 hijos por familia.

Table 2. Pasajes ordenados por *ranking* correspondientes a la pregunta 0026.

Ranking	**Pasaje Obtenido**
0.375	Un tren de la compañía "British Rail" llevó a los invitados desde la estación Victoria, en Londres, hacia el comienzo del túnel, donde otro ferrocarril les trasladó al lugar de la comida.
0.125	Según aparecía en la invitación, se pretendia emular una cena celebrada en noviembre de 1827 para celebrar la finalización del túnel del Támesis.

Table 3. Pasajes ordenados por *ranking* correspondientes a la pregunta 0066.

Ranking	**Pasaje Obtenido**
0.712	1994.- Lanzamiento del transbordador "Discovery" con un astronauta ruso a bordo, primero en un vehículo espacial estadounidense.
0.160	<TEXT> Cabo Cañaveral (EEUU), 20 mar (EFE).- Norman Thagard, primer astronauta norteamericano que participa en un vuelo espacial ruso, declaró hoy que no encontró "ninguna sorpresa importante" al comienzo de los tres meses que pasará a bordo de la estación MIR.
0.111	"Esta es una misión muy complicada, pues vamos a unir los programas espaciales de dos de las naciones mas grandes del mundo", declaró James Wetherbee, quien capitaneará la misión junto a Eileen Collins y una tripulación compuesta por los especialistas Bernard Harris, Michael Foale, la astronauta Janice Voss y el ruso Titov.

Bibliografía

1. Moldovan, Dan, Sanda Harabagiu, Roxana Girju, Paul Morarescu, Finley Lacatusu, A. Novischi, A. Badulescu, y O. Bolohan., LCC Tools for Question Answering, En Eleventh Text REtrieval Conference, volumen 500-251 de NIST Special publication, Gaithersburg, USA, nov. National Institute of Standards and Technology, 2002.
2. Soubbotin, M. y S. Soubbotin, Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach, En Eleventh Text REtrieval Conference, volumen 500-251 de NIST Special Publication, Gaithersburg, USA, nov. National Institute of Standards and Technology, 2002.
3. Yang, Hui y Tat-Seng Chua, The Integration of Lexical Knowledge and External Resources for Question Answering, En Eleventh Text REtrieval Conference, volumen 500-251 de NIST Special Publication, Gaithersburg, USA, nov. National Institute of Standards and Technology, 2002.
4. Mark A. GreenWood, Using Pertainyms to Improve Passage Retrieval for Question Answering Information About a Location, University of Sheffield, Portobello Road S1 4DP UK, 2004.
5. Xiaoyong. Liu, W. Bruce Croft, Passage Retrieval Based On Language Models, University of Massachussetts, Amherst, MA 01003, ACM 1-58113-492-4, 2002.
6. Dell Zhang and Wee Sun Lee, A Language Modeling Approach to Passage Question Answering, , National University of Singapore, TREC2003, 2003.
7. Callan, J.P, Passage-level evidence in document retrieval, In proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 302-309. New York: ACM, 1994.
8. Kaszkiel, Justin Zobel, Efficient Passage Ranking for Document Databases, ACM Transactions on Information Systems, RMIT University, Vol. 17, No. 4, pages 406-439, October 1999.
9. Vicedo, SEMQA: Un modelo semántico aplicado a los sistemas de Búsqueda de Respuestas, Ph.D. tesis, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante, Ctra. de San Vicente s/n. Mayo 2002.
10. A. Gelbukh, G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: Computational Linguistics and Intelligent Text Processing (CICLing-2003), Lecture Notes in Computer Science, N 2588, Springer-Verlag, 2003, pp. 215220.