Discovering answers to definition questions in web snippets

Alejandro G. Figueroa A.
DFKI GmbH – LT Lab, Stuhlsatzenhausweg 3, Saarbruecken, Germany

Introduction

Definition questions are queries like:

- *Who is George Bush?*
- *What is epilepsy?*
- *What is AI?*

This sort of query is particularly important because 27% of the questions raised by real user logs are a request for a definition. Answering definition questions involves discovering as much descriptive information about the target concept (*definiendum*) as possible:

- *Who is George Bush?*
  - *born July 6, 1946.*
  - *forty-third President of USA.*
  - *Forty-six Governor of Texas.*

Our Questions Answering System (WebQA) extracts these pieces of descriptive information (nuggets) from the brief descriptions returned by commercial search engines (such as MSN Search). An illustrative *web snippet* is as follows:

Biography of President George W. Bush
George W. Bush is the 43rd President of the United States. He was sworn into office on January 20, 2001, re-elected on November 2, 2004, and sworn in for a ...
http://www.whitehouse.gov/president/biography.html

Booting recall of descriptive phrases

The first step in the answering process is retrieving as much snippets as possible that contain descriptive information. For this purpose, WebQA builds search queries by taking advantage of local-syntactic structures that often convey definitions:

- *George Bush [is|are|has been|have been|was|were] [a|the|an] …*
- *George Bush, [a|an|the] , …*
- *George Bush [become|became|becomes]..*
- *George Bush, who ….*
- *George Bush was born …*
- *George Bush, or ….*
- *George Bush, nicknamed …*
- *George Bush (1946) …*

Using these structures, WebQA generates ten search queries (q_1 to q_{10}):

1. *"George Bush"*
2. *"George Bush is a" ∨ "George Bush was a" ∨ "George Bush were a" ∨ "George Bush are a"*
3. *"George Bush is an" ∨ "George Bush was an" ∨ "George Bush were an" ∨ "George Bush are an"*
4. *"George Bush is the" ∨ "George Bush was the" ∨ "George Bush were the" ∨ "George Bush are the"*
5. *"George Bush has been the" ∨ "George Bush has been an" ∨ "George Bush has been the" ∨ "George Bush have been a" ∨ "George Bush have been an" ∨ "George Bush have been the"*
6. *"George Bush, a" ∨ "George Bush, an" ∨ "George Bush, or"*
7. *("George Bush" ∨ "George Bush also" ∨ "George Bush is" ∨ "George Bush are") ∧ (called ∨ nicknamed ∨ "known as")*
8. *"George Bush became" "George Bush became" "George Bush become" "George Bush becomes"*
9. *"George Bush which" "George Bush that" "George Bush who"*
10. *"George Bush was born" "(George Bush)"*

Improving retrieval

A disadvantage of the previous query rewriting strategy is that it is static:

- two promising lexico-syntactic clauses could be submitted in the same query, lessening the recall of definitions.
- a set of unpromising lexico-syntactic patterns can be set in the same query and hence, bring about an unproductive retrieval, diminishing the number of descriptive utterances.

Off-line n-grams counts supplied by Google are used to transform this static query construction into a more dynamic one. An excerpt from Google 4-grams counts is as follows:

George Bush is a 20515
George Bush is an 3019
George Bush is the 10029
George Bush was a 2163
George Bush was an 240
George Bush was the 1810

In some cases, the grammatical number can be inferred. In particular, in the case of *"George Bush"*, singular lexico- syntactic clues are most promising. However, it is not always possible to draw a clear distinction:

fractals are a 176
fractals are an 86
fractals are the 215
fractals is a 124
fractals is the 148

S-I is a strategy, which selects a grammatical number whenever more than three keywords corresponding to one grammatical number exist, and zero to the another.

S-I makes use of hierarchy within the lexico-syntactic patterns, given by the frequency hints, for configuring the ten queries. First, the search queries q_7 and q_{10} are merged into one query q'_7, composed of the following clauses: *" also called ", " also nicknamed", "also known", "is called", "stands for", "is known", "are called", "are nicknamed", "are known", "was founded", "was founded", "is nicknamed", "was born".*

q'_7 consists merely of the clauses that can be found in Google n-grams. If any clause cannot be found, q'_7 is set to ∅. In any case, q'_{10} remains as ∅. Second, $q'_5 = q_5$, $q'6 = q_6$ and $q'_8 = q_8$ as well as $q'_9 = q_9$. Additionally, the q'_1 is set to ∅. Third, the clauses included in the queries q_2 and q_3, as well as q_4, are dynamically sorted across the available queries:

q'7 = ∅	q'7 ≠ ∅
q'_1="δ R1" q'_2="δ R2" q'_3="δ R3" q'_4="δ R4" q'_5="δ R5" q'_6="δ R6"	q'_1="δ R1" q'_2="δ R2" q'_3="δ R3" q'_4="δ R4" q'_5="δ R5"∨"δ R6"

Table 1. Dynamic queries (Grammatical number known).

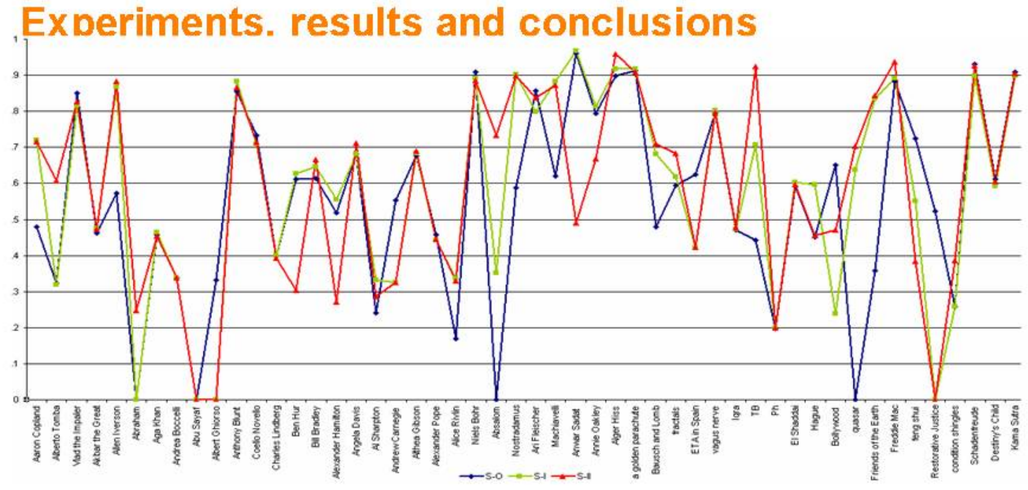
where R1 and R6 correspond to the highest and lowest frequent lexico-syntactic patterns according to Google frequency counts. In the case that the grammatical number cannot be distinguished, the queries are as follows:

q_1:"δ is a" ∨ "δ were an" ∨ "δ was the"
q_2:"δ was a" ∨ "δ are an"
q_3:"δ are a" ∨ "δ was an" ∨ "δ were the"
q_4:"δ were a" ∨ "δ is an"
q_{10}:"δ is the" ∨ "δ are the"

In the case q_{10} = ∅, the following queries are reformulated:

q_1:"δ is a" ∨ "δ were an"
q_3:"δ are a" ∨ "δ was an"
q_7:"δ was the" ∨ "δ were the"

Experiments. results and conclusions



S-I and the static query rewriting strategy (S-O) were assessed by means of the definition question set supplied by TREC 2003. Following the suggestion of [1], S-I was additionally tested together with the use of an extra search engine (S-II).

- *WebQA with S-I finished with an average F(5) score of 0.5472, while S-II improved the average value to 0.5792, and S-II, improved to 0.5842.*
- *S-I obtained an improvement without increasing the number of submitted queries, whereas the marginal increase achieved by S-II with respect to S-I, is at the expense of sending ten extra queries.*

An interesting finding: Google n- grams can be used for optimising the retrieval of definitions in web snippets, and thus, they can also assist QAS in fetching more promising full-documents.

Future challenges

S-O and S-I scored zero for four different *definiendums*, despite the "okay" nuggets found by both systems. In fact, if a system does not discover any nugget assessed as "vital", it finishes with a F(5) value equal to zero. For instance, S-II scored zero for three questions; in particular, for the following output:

- *said Albert Ghiorso, a veteran Berkeley researcher, who holds the Guinness world...*
- *Albert Ghiorso is a nuclear scientist at Lawrence Berkeley National Laboratory in Berkeley, Calif.*
- *That's what Berkeley Lab's Albert Ghiorso, a man who has participated in the discovery of more atomic elements than any living person, told the students and teachers ...*
- *Albert Ghiorso is an American nuclear scientist who helped discover several elements on the periodic table.*

The "okay" nugget is underlined that matches the TREC 2003 assessors' list:

vital designed and built cyclotron accelerator
okay nuclear physicists/experimentalist
vital co-creator of 12 artificial elements
vital co-discovered element 106

"okay" nuggets, like nuclear physicists/ experimentalist can be easily interpreted as "vital". For example, if one considers abstracts supplied by Wikipedia as a third-party judgement, one finds:

Albert Ghiorso (b. 15 July 1915) is an American nuclear scientist who helped discover numerous chemical elements on the periodic table.

Some relevant nuggets are unconsidered, enlarging the response, and thus decreasing the F(5) score. We hypothesise that a nugget can be seen as "vital" or "okay" according to how often its type occurs across abstracts and/or bodys of online encyclopedias.

In the case of "Abu Sayaf", the three strategies were unable to discover any nugget in the assessor' list. The reason is uncovered when the following frequencies on Google n-grams are checked:

Future challenges

Abu Sayyaf 96204
Abu Sayyafs 89
Abu Sayaf 1156
Abu Saya 3205

The spelling of the *definiendum* is unlikely to occur in the web, causing an F(5) equals to zero. Conversely, when WebQA processes "Abu Sayyaf ", the scores obtained are: 0.844 (S-O), 0.8794 (S-I) and 0.8959 (S-II). Then, the new average F((5) values are: 0.564 (S-O), 0.59679 (S-I) and 0.602 (S-II).

Bibliography

[1] Chen, Y., Zhon M. and Wang, S. "Reranking Answers Definitional QA Using Language Modeling", Proceedings of the Coling/ACL-2006, pp. 1081–1088.
[2] Cui, T.S.C.H., Kan M.Y. and Xiao J. "A comparative study on sentence retrieval for definitional question answering", SIGIR Workshop on Information Retrieval for Question Answering (IR4QA), 2004.
[3] Figueroa, A. and Neumann, G. "A Multilingual Framework for Searching Definitions on Web Snippets", 2007 : Advances in Artificial Intelligence', LNCS, Volume 4667/2007, p. 144-159.
[4] HildebrandtW., Katz B. and Lin J. "Answering Definition Questions Using Multiple Knowledge Sources", Proceedings of HLT-NAACL 2004, pp. 49–56.
[5] Voorhees, E., M. "Evaluating Answers to Definition Questions", Proceedings of HLT-NAACL 2003, pp. 109–111, 2003.
[6] Xu, J., Licuanan, A. and Weischedel, R. "TREC2003 at BBN: Answering definitional questions", Proceedings of the Twelfth Text REtrieval Conference, 2003.

Acknowledgments

The work presented here was partially supported by a research grant from the German Federal Ministry of Education, Science, Research and Technology (BMBF) to the DFKI project HyLaP (FKZ: 01 IW F02) and the EC-funded project QALL-ME.

Further information

Please contact *figueroa@dfki.de*. More information on this and related projects can be obtained at *www.dfki.de*. An online PDF-version of this poster can be found at *http://www.dfki.de/~figueroa/*.

Our WebQA system can be found at:
http://experimental-quetal.dfki.de/