Boosting the Recall of Descriptive Phrases in Web Snippets

Alejandro Figueroa[1]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz - DFKI, Saarbrücken, Germany
figueroa@dfki.de

Abstract

WebQA is a Web Question Answering System[1] which is aimed at discovering answers to natural language questions on the web. One of its major components is the module that answers definition questions. This module searches for answers by means of a query rewriting strategy, which considerably boosts the recall of descriptive utterances. This study compares three different search strategies and deals at greater length with the challenges posed by the assessment of web-based definition Question Answering Systems.

Index Terms: Question Answering, Definition Questions, Web Mining, Search

1. Introduction

WebQA is built on top of commercial search engines like MSN Search and Yahoo. WebQA is part of sustained efforts to implement a system which extracts answers to factoid [3] and definition [4], as well as list questions [5] **exclusively** from the brief descriptions returned by these search engines, called *web snippets*.

The reason to use web snippets as an answer source is fourfold: (a) they are computed at high speed by current commercials search engines, and therefore provide a quick and contextualised response, (b) to take advantage of the current power of indexing of vanguard search engines, (c) to the user, web snippets are the first view of the response, thus highlighting answers would make them more informative, and (d) to avoid, or at least lessen, the retrieval and costly processing of a wealth of documents. In particular, web snippets have proven to be promising for answering difficult queries like definitions questions (such as "*Who is Allen Iverson?*" or "*What are fractals?*"). This sort of query is particularly important, because 27% of the questions of real user logs are a request for a definition. In order to satisfactorily answer definition questions, Question Answering Systems (QAS) must take answers from several documents and afterwards, discriminate senses, merge answers, remove redundancy, and eventually generate a final output for the user. This study focus its attention on the first step: the search or retrieval of definition answers.

2. Related Work

QAS are usually assessed in the context of the Question Answering track of the Text REtrieval Conference (TREC). During the last years, the thoroughness and difficulty of this assessment has been systematically increased by making allowances for more challenging queries, such as definition questions.

In TREC, the target collection is the AQUAINT corpus. QAS make use of several external resources of definition information in order to successfully discover the right answers in this corpus. QAS then identify descriptive phrases by projecting the obtained nuggets into the corpus. In this way, they also filter out some misleading and spurious nuggets taken from these external sources. In the jargon of definition questions, a nugget is a piece of relevant or factual information about the particular topic of the question (a. k. a. the *definiendum*).

For instance, [6] introduced a method for answering definition questions that was assisted by a wrapper for the online Merriam Webster dictionary, which retrieved about 1.5 nuggets per question. These nuggets were used as query expansion terms for retrieving promising documents from the collection afterwards. Additionally, they automatically constructed offline an extremely large relational database containing nuggets about every entity mentioned in the AQUAINT corpus. These nuggets were accordingly taken from every article within it, and therefore, answering definition questions consisted of a simple lookup for the *definiendum*. Since nuggets often seem odd and out of place without their context, [6] expanded them to surround one hundred (non-white-space) characters in order to enhance readability.

Another example, is the strategy proposed by [2], which took advantage of external resources like WordNet glossaries, online specific resources (e.g., Wikipedia) and web snippets for learning frequencies and correlation of words, especially with the *definiendum*. One of their findings was that definitional web-sites greatly improve the performance, leading to few unanswered questions: Wikipedia covered 34 out of the 50 TREC–2003 definition queries and biography.com 23 out of 30 questions regarding people, all together provided answers to 42 queries. They additionally found that web snippets, though they yielded relevant information about the *definiendum*, were not likely to supply descriptive utterances, bringing about only a marginal improvement.

Another method that takes advantage of web snippets was presented in [1]. This method uses a centroid vector that considers word dependencies learnt from the 350 most frequent stemmed co-occurring terms taken from the best **500 snippets** retrieved by Google. These snippets were fetched by expanding the original query by means of a set of five highly co-occurring terms. These terms co-occur with the *definiendum* in sentences obtained by submitting the original query plus some task specific clues, e.g.,"*biography*". As a result, this query expansion technique improved the $\mathcal{F}(5)$ score of their system from 0.511 to 0.531. They concluded that the use of multiple search engines would help to fetch more sentences containing the *definiendum*.

The module of WebQA that answers definition questions was described firstly in [4]. Contrary to QAS in TREC, WebQA searches for definition sentences only on the web, in particular in web snippets. The advantage of descriptive phrases extracted from web snippets is that they provide an adequate unit of contextual information [4], being comparable in size with the enhanced nuggets obtained by [6]. For the purpose of markedly

[1]http://experimental-quetal.dfki.de/

increasing the recall of definition sentences within web snippets, WebQA biases the search engine in favour of some lexico-syntactic structures that often convey definitions by means of a purpose-built query rewriting strategy. As a result, WebQA finished with $\mathcal{F}(5)$ score of 0.53 for the TREC 2003 data-set, which is "competitive" with the best systems, which achieve a value between 0.5 and 0.56 [1, 6, 7, 8].

However, a key point for correctly interpreting these results is the completeness of the assessor's list. It is known that systems in TREC were able to find relevant nuggets, which were not included in this list (cf. [6] for details). In the case of web-based systems like WebQA, this vital fact is more likely to happen, because they discover many additional nuggets seen as relevant by the user, but excluded from the assessor's list. This exclusion actually brings about a decrease in the $\mathcal{F}(5)$ score, because these extra nuggets enlarge the response without increasing precision. Moreover, WebQA must determine exclusively from the context whether or not a certain nugget conveys definition information; this means it lacks a target corpus that could act like a filter for some spurious and misleading answers. This kind of evaluation is, nonetheless, the unique current way to have an objective reference to the performance of several systems.

This study shows two search strategies that boost the recall of sentences that convey definitions, and consequently, they better the performance of the definition module of WebQA. These strategies: (a) take into consideration the prior knowledge provided by Google n-grams while rewriting the query, and (b) take up the suggestion of [1] by adding an extra search engine[2]. Another thing minutely examined in this work, is the impact of the assessor's list on the evaluation of web-based definition QAS.

3. Mining the Web for Definitions

Currently, the definition component of WebQA makes use of ten purpose-built search queries, which are based on some local lexico-syntactic constructions that often convey definitions (cf. [4] for details). These ten search queries help WebQA to substantially increase the recall of descriptive utterances within web snippets (δ stands for the *definiendum*):

q_1="δ"
q_2="δ is a " $\vee$ "δ was a " $\vee$ "δ were a " $\vee$ "δ are a "
q_3="δ is an " $\vee$ "δ was an " $\vee$ "δ were an " $\vee$ "δ are an "
q_4="δ is the " $\vee$ "δ was the " $\vee$ "δ were the " $\vee$ "δ are the "
q_5="δ has been a " $\vee$ "δ has been an " $\vee$ "δ has been the " $\vee$ "δ have been a " $\vee$ "δ have been an " $\vee$ "δ have been the "
q_6="δ, a " $\vee$ "δ, an " $\vee$ "δ, the " $\vee$ "δ, or "
q_7=("δ" $\vee$ "δ also " $\vee$ "δ is " $\vee$ "δ are ") $\wedge$ (called $\vee$ nicknamed $\vee$ "known as")
q_8="δ became " $\vee$ "δ become " $\vee$ "δ becomes "
q_9="δ which " $\vee$ "δ that " $\vee$ "δ who "
q_{10}="δ was born " $\vee$ "(δ)"

The drawback to this query rewriting strategy is that these search queries are statically built, causing that two promising lexico-syntactic clauses could be submitted in the same query, lessening the retrieval of descriptive phrases. A good illustrative example is δ="*Allen Iverson*" and q_2. In this case, "*Allen Iverson is a*" and "*Allen Iverson was a*" are two clauses likely to yield definitions. Consequently, they should be separately submitted in order to avoid weakening the recall. Further, clauses such as "*Allen Iverson were a*" and "*Allen Iverson are a*" only bring about misleading sentences:

[2]http://www.yahoo.com/

- Cheers for visiting ***Allen Iverson were a*** *slap in the face to the Clippers.*
- Carmelo Anthony and ***Allen Iverson were a*** *combined 1 for 10 in the third, when Denver committed 9 turnovers.*

Analogously, 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. Nevertheless, these patterns observe a local lexico-syntactic dependency with the *definiendum*, specifically, they are unlikely to contain additional words in between. This is an important fact, because off-line n-grams counts supplied by Google can be used to transform this static query construction into a more dynamic one. In our working example, an excerpt of Google 4-grams counts is as follows:

Allen Iverson is a 209
Allen Iverson is an 68
Allen Iverson is the 425
Allen Iverson was a 57
Allen Iverson was the 101

The first beneficial aspect of Google n-grams is that, in some cases, the grammatical number can be inferred. In particular, in the case of "*Allen Iverson*", singular lexico-syntactic clues are most promising. However, it is not always possible to draw a clear distinction. A good example is "*fractals*":

fractals are a 176 (e.g. "Fractals are a powerful tool...")
fractals are an 86 (e.g. "Fractals are an exquisite...")
fractals are the 215 (e.g. "Fractals are the place...")
fractals is a 124 (e.g. "Fractals is a new branch of...)
fractals is the 148 (e.g. "Fractals is an innovative...)

Then, a strategy was designed (S-I), which selects a grammatical number whenever more than three keywords corresponding to one grammatical number exist, and zero to the another. The second favourable aspect is that the frequencies give hints about the hierarchy within the lexico-syntactic patterns. S-I takes advantage of this hierarchy for configuring the ten queries. First, the search queries q_7 and q_{10} are merged into one query $q_7^{'}$. This query is composed of the following clauses:

"δ also called ", "δ also nicknamed", "δ also known", "δ is called", "δ stands for", "δ is known", "δ are called", "δ are nicknamed", "δ are known", "δ was born", "δ was founded", "δ was founded", "δ is nicknamed"

Accordingly, $q_7^{'}$ consists merely of the clauses that can be found in Google n-grams. If any clause cannot be found, $q_7^{'}$ is set to $\emptyset$. In any case, $q_{10}^{'}$ remains as $\emptyset$. It is worth pointing out that, the term "*stands for*" replaces the parentheses in q_{10}. 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 $\emptyset$. Third, the clauses included in the queries q_2 and q_3, as well as q_4, are dynamically sorted across the available queries, as highlighted in table 1.

Table 1: Dynamic queries (grammatical number known).

$q_7^{'} = \emptyset$	$q_7^{'} \neq \emptyset$
$q_1^{'}$:"δ R_1" $q_2^{'}$:"δ R_2" $q_3^{'}$:"δ R_3" $q_4^{'}$:"δ R_4" $q_5^{'}$:"δ R_5" $q_7^{'}$:"δR_6"	$q_1^{'}$:"δ R_1" $q_2^{'}$:"δ R_2" $q_3^{'}$:"δ R_3" $q_4^{'}$:"δ R_4" $q_5^{'}$:"δ R_5" $\vee$ "δ R_6"

where R_1 and R_6 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" $\vee$ "δ were an" $\vee$ "δ was the"
$q_2^{'}$:"δ was a" $\vee$ "δ are an"
$q_3^{'}$:"δ are a" $\vee$ "δ was an" $\vee$ "δ were the"
$q_4^{'}$:"δ were a" $\vee$ "δ is an"
$q_{10}^{'}$:"δ is the" $\vee$ "δ are the"

In the case $q_{10}^{'} = \emptyset$, the following queries are reformulated:

$q_1^{'}$:"δ is a" $\vee$ "δ were an"
$q_3^{'}$:"δ are a" $\vee$ "δ was an"
$q_7^{'}$:"δ was the" $\vee$ "δ were the"

Every query is eventually surrounded with the feature "*inbody:*" in order to avoid matching a clause with the title of a web page.

4. Experiments

S-I and the static query rewriting strategy (S-O) were assessed by means of the defintion 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). Figure 1 compares the $\mathcal{F}(5)$ score per question for the three strategies.

WebQA with the static query rewriting finished with an average $\mathcal{F}(5)$ score of 0.5472, while the dynamic query rewriting improved the average value to 0.5792, and this rewriting along with an additional search engine, improved to 0.5842. Here, the first aspect to point out is the increase to 0.5472 with respect to the $\mathcal{F}(5)$ value (0.53) reported in [4]. We interpret this increase as a change in the fetched content from the web. As well as that, it is worth remarking that 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 to the additional search engine. It is also worth noting that each submission is done as described in [4], and hence, S-O and S-I fetch a maximun of 300 snippets, while S-II 600. These sets of snippets are comparable in size with the 500 snippets retrieved by [1]. Overall, the $\mathcal{F}(5)$ values, achieved by WebQA with our rewriting strategies incorporated, are "competitive" with the best definition QAS. These systems obtain a value between 0.5 and 0.56 [1, 6, 7, 8].

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 $\mathcal{F}(5)$ value equal to zero. For instance, S-II scored zero for three questions; in particular, for the following output concerning "*Albert Ghiorso*":

- said ***Albert Ghiorso, a*** *veteran Berkeley researcher, who holds the Guinness world record.*
- ***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 who packed.*
- ***Albert Ghiorso is an*** *American nuclear scientist who helped discover several elements on the periodic table.*

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

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

Like [6] also noticed, "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, at the time of writing, one finds:

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

Further, some relevant nuggets, including *veteran Berkeley researcher*, are unconsidered, enlarging the response, and thus decreasing the $\mathcal{F}(5)$ score. We hypothesise that a nugget can be seen as "vital" or "okay" according to how often its **type** (birthplace, birthdate, occupation, outstanding achievement) occurs across abstracts and/or bodys of online encyclopedias, such as Encarta or Wikpedia. We deem that this sort of **type-oriented** evaluation would be more appropriate to web-based definition QAS. Only in one *definiendum* were the three strategies unable to discover any nugget in the assessor' list: "*Abu Sayaf*". The reason is uncovered when the following frequencies on Google n-grams are checked:

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

In this case, the spelling of the *definiendum* in the query is unlikely to occur in the web, causing an $\mathcal{F}(5)$ equals to zero. Conversely, when WebQA processes "*Abu Sayyaf*", the scores obtained by each method are: 0.844 (S-O), 0.8794 (S-I) and 0.8959 (S-II). Accordingly, the new average $\mathcal{F}(5)$ values are: 0.564 (S-O), 0.59679 (S-I) and 0.602 (S-II).

Another complicated problem is that the list of the assessor is aimed predominantly at one possible sense of the *definiendum*. Therefore, discovered descriptive utterances concerning additional senses, similar to the uncosidered nuggets, bring about a decrease in the $\mathcal{F}(5)$ value. To illustrate this, a descriptive sentence found by S-II regarding "*Nostradamus*":

- ***Nostradamus is a*** *neural network-based, short-term demand and price forecasting system, utilized by electric and gas utilities, system operators and power pools, electric...*

Indeed, it is highly frequent to find ambiguous terms. For example, Wikipedia contains more than 19000 different disambiguation pages. In this case, the list of the assessor only accounts for the reference to the French astrologer/prophet. When sentences concerning other senses are manually removed, the $\mathcal{F}(5)$ values for this concept increase as follows: from 0.5871 to 0.5936 (S-O), from 0.9028 to 0.9182 (S-I) and from 0.8977 to 0.9167 (S-II). Obviously, a more noticeable difference is due to *definiendums* with more senses such as "*Absalom*".

Another difficulty that QAS encounter when they extract definition phrases from the web, is that opinions are also given like definitions. A good example is given by the *definiendum* "*Charles Lindberg*":

- ***Charles Lindberg was a*** *true American hero.*

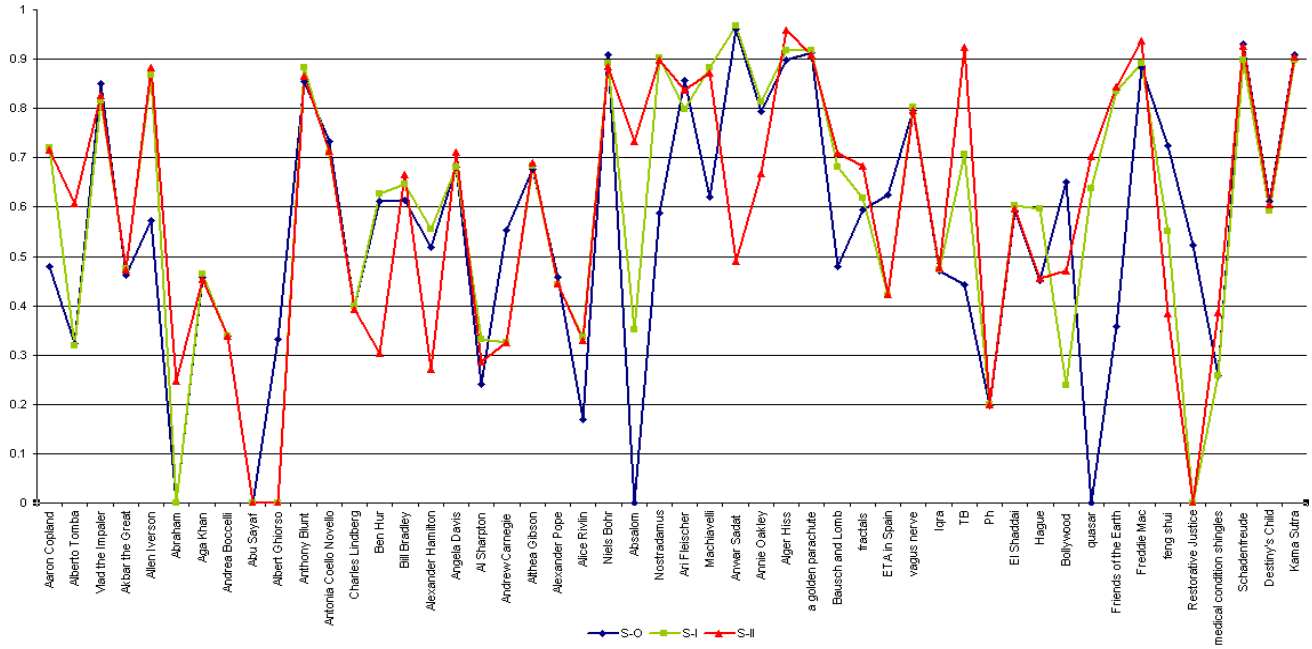

Figure 1: Comparison between F(5) scores obtained by each strategy for each *definiendum* in the TREC 2003 question-set.

This sentence does not syntactically differ from the definition "*Charles Lindberg was a famous American pilot.*" We envisage that a large-scale redundancy and the use of opinion mining techniques would help to discriminate opinions from facts.

Our ongoing research is aimed at incorporating more linguistic information into the query rewriting strategy. Specifically, promising verb phrases can be interpreted as definition lexico-syntactic patterns, and therefore, appended to the "*definiendum*". These verb phrases can be determined by means of retrieved descriptive sentences, a chunker, and the corresponding recalls can be estimated by inspecting the frequency of these new clauses on Google n-grams. This sort of strategy would help to fetch more and diverse descriptive information about the *definiendum*.

5. Conclusions

This study compares three query rewriting strategies that are aimed at boosting the recall of descriptive sentences in web snippets and consequently, at improving the performance of definition QAS. One interesting finding is that Google n-grams can be used particularly for optimising the retrieval of defintions in web snippets, and accordingly, they can also assist QAS in fetching more promising full-documents.

This paper additionally discusses the major challenges posed by web-based definition QAS, and it sketches accordingly some directions that could help to face these challenges.

6. Acknowledgements

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.

7. References

[1] Chen, Y., Zhon M. and Wang, S. "Reranking Answers for 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. "Language Independent Answer Prediction from the Web", Proceedings of the FinTAL 5th International Conference on Natural Language Processing, August 23-25 in Turku, Finland, 2006.

[4] Figueroa, A. and Neumann, G. "A Multilingual Framework for Searching Definitions on Web Snippets", KI 2007: Advances in Artificial Intelligence', LNCS, Volume 4667/2007, p. 144-159.

[5] Figueroa, A. and Neumann, G. "Mining Web Snippets to Answer List Questions", In AI07: the 2nd International Workshop on Integrating AI and Data Mining, 2nd-6th December 2007, Gold Coast, Queensland, Australia.

[6] Hildebrandt W., Katz B. and Lin J. "Answering Definition Questions Using Multiple Knowledge Sources", Proceedings of HLT-NAACL 2004, pp. 49–56.

[7] Voorhees, E., M. "Evaluating Answers to Definition Questions", Proceedings of HLT-NAACL 2003, pp. 109–111, 2003.

[8] Xu, J., Licuanan, A. and Weischedel, R. "TREC2003 QA at BBN: Answering definitional questions", Proceedings of the Twelfth Text REtrieval Conference, 2003.