Resource Analysis for Question Answering

Lucian Vlad Lita
Carnegie Mellon University
llita@cs.cmu.edu

Warren A. Hunt
Carnegie Mellon University
whunt@andrew.cmu.edu

Eric Nyberg
Carnegie Mellon University
ehn@cs.cmu.edu

Abstract

This paper attempts to analyze and bound the utility of various structured and unstructured resources in Question Answering, independent of a specific system or component. We quantify the degree to which gazetteers, web resources, encyclopedia, web documents and web-based query expansion can help Question Answering in general and specific question types in particular. Depending on which resources are used, the QA task may shift from complex answer-finding mechanisms to simpler data extraction methods followed by answer re-mapping in local documents.

1 Introduction

During recent years the Question Answering (QA) field has undergone considerable changes: question types have diversified, question complexity has increased, and evaluations have become more standardized - as reflected by the TREC QA track (Voorhees, 2003). Some recent approaches have tapped into external data sources such as the Web, encyclopedias, databases in order to find answer candidates, which may then be located in the specific corpus being searched (Dumais et al., 2002; Xu et al., 2003). As systems improve, the availability of rich resources will be increasingly critical to QA performance. While on-line resources such as the Web, WordNet, gazetteers, and encyclopedias are becoming more prevalent, no system-independent study has quantified their impact on the QA task. This paper focuses on several resources and their inherent potential to provide answers, without concentrating on a particular QA system or component. The goal is to quantify and bound the potential impact of these resources on the QA process.

2 Related Work

More and more QA systems are using the Web as a resource. Since the Web is orders of magnitude larger than local corpora, redundancy of answers and supporting passages allows systems to produce more correct, confident answers (Clarke et al., 2001; Dumais et al., 2002). (Lin, 2002) presents two different approaches to using the Web: accessing the structured data and mining the unstructured data. Due to their complementary nature of these approaches, hybrid systems are likely to perform better (Lin and Katz, 2003).

Definitional questions (*"What is X?"*, *"Who is X?"*) are especially compatible with structured resources such as gazetteers and encyclopedias. The top performing definitional systems (Xu et al., 2003) at TREC extract kernel facts similar to a question profile built using structured and semi-structured resources: WordNet (Miller et al., 1990), Merriam-Webster dictionary *www.m-w.com*), the Columbia Encyclopedia (*www.bartleby.com*), Wikipedia (*www.wikipedia.com*), a biography dictionary (*www.s9.com*) and Google (*www.google.com*).

3 Approach

For the purpose of this paper, resources consist of structured and semi-structured knowledge, such as the Web, web search engines, gazetteers, and encyclopedias. Although many QA systems incorporate or access such resources, few systems quantify individual resource impact on their performance and little work has been done to estimate bounds on resource impact to Question Answering. Independent of a specific QA system, we quantify the degree to which these resources are able to directly provide answers to questions.

Experiments are performed on the 2,393 questions and the corresponding answer keys provided through NIST (Voorhees, 2003) as part of the TREC 8 through TREC 12 evaluations.

4 Gazetteers

Although the Web consists of mostly unstructured and loosely structured information, the available structured data is a valuable resource for question answering. Gazetteers in particular cover several frequently-asked factoid question types, such as

"*What is the population of X?*" or "*What is the capital of Y?*". The CIA World Factbook is a database containing geographical, political, and economical profiles of all the countries in the world. We also analyzed two additional data sources containing astronomy information (*www.astronomy.com*) and detailed information about the fifty US states (*www.50states.com*).

Since gazetteers provide up-to-date information, some answers will differ from answers in local corpora or the Web. Moreover, questions requiring interval-type answers (e.g. "*How close is the sun?*") may not match answers from different sources which are also correct. Gazetteers offer high precision answers, but have limited recall since they only cover a limited number of questions (See Table 1).

		CIA		**All**	
Q-Set	**#qtions**	**R**	**P**	**R**	**P**
TREC8	200	4	100%	6	100%
TREC9	693	8	100%	22	79%
TREC10	500	14	100%	23	96%
TREC11	500	8	100%	20	100%
TREC12	500	2	100%	11	92%
Overall	2393	36	100%	82	91%

Table 1: *Recall (R)*: TREC questions can be directly answered directly by gazetteers - shown are results for **CIA** Factbook and **All** gazetteers combined. Our extractor precision is *Precision (P)*.

5 WordNet

Wordnets and ontologies are very common resources and are employed in a wide variety of direct and indirect QA tasks, such as reasoning based on axioms extracted from WordNet (Moldovan et al., 2003), probabilistic inference using lexical relations for passage scoring (Paranjpe et al., 2003), and answer filtering via WordNet constraints (Leidner et al., 2003).

Q-Set	**#qtions**	**All**	**Gloss**	**Syns**	**Hyper**
TREC 8	200	32	22	7	13
TREC 9	693	197	140	73	75
TREC 10	500	206	148	82	88
TREC 11	500	112	80	29	46
TREC 12	500	93	56	10	52
Overall	2393	641	446	201	268

Table 2: Number of questions answerable using WordNet glosses (*Gloss*), synonyms (*Syns*), hypernyms and hyponyms (*Hyper*), and all of them combined *All*.

Table 2 shows an upper bound on how many TREC questions could be answered *directly* using WordNet as an answer source. Question terms and phrases were extracted and looked up in WordNet glosses, synonyms, hypernyms, and hyponyms. If the answer key matched the relevant WordNet data, then an answer was considered to be found. Since some answers might occur coincidentally, we these results to represent upper bounds on possible utility.

6 Structured Data Sources

Encyclopedias, dictionaries, and other web databases are structured data sources that are often employed in answering definitional questions (e.g., *"What is X?", "Who is X?"*). The top-performing definitional systems at TREC (Xu et al., 2003) extract kernel facts similar question profiles built using structured and semi-structured resources: WordNet (Miller et al., 1990), the Merriam-Webster dictionary *www.m-w.com*), the Columbia Encyclopedia (*www.bartleby.com*), Wikipedia (*www.wikipedia.com*), a biography dictionary (*www.s9.com*) and Google (*www.google.com*).

Table 3 shows a number of data sources and their impact on answering TREC questions. N-grams were extracted from each question and run through Wikipedia and Google's *define* operator (which searches specialized dictionaries, definition lists, glossaries, abbreviation lists etc). Table 3 show that TREC 10 and 11 questions benefit the most from the use of an encyclopedia, since they include many definitional questions. On the other hand, since TREC 12 has fewer definitional questions and more procedural questions, it does not benefit as much from Wikipedia or Google's *define* operator.

Q-Set	**#qtions**	**WikiAll**	**Wiki1st**	**DefOp**
TREC 8	200	56	5	30
TREC 9	693	297	49	71
TREC 10	500	225	45	34
TREC 11	500	155	19	23
TREC 12	500	124	12	27
Overall	2393	857	130	185

Table 3: The answer is found in a definition extracted from Wikipedia *WikiAll*, in the first definition extracted from Wikipedia *Wiki1st*, through Google's **define** operator *DefOp*.

7 Answer Type Coverage

To test coverage of different answer types, we employed the top level of the answer type hierarchy used by the JAVELIN system (Nyberg et al., 2003). The most frequent types are: definition (e.g. "*What is viscosity?*"), person-bio (e.g. "*Who was Lacan?*"), object(e.g. "*Name the highest mountain.*"),

process (e.g. *"How did Cleopatra die?"*), lexicon (*"What does CBS stand for?"*)temporal(e.g. *"When is the first day of summer?"*), numeric (e.g. *"How tall is Mount Everest?"*), location (e.g. *"Where is Tokyo?"*), and proper-name (e.g. *"Who owns the Raiders?"*).

AType	#qtions	WikiAll	DefOp	Gaz	WN
object	1003	426	92	58	309
lexicon	50	25	3	0	26
defn	178	105	9	11	112
pers-bio	39	15	11	0	17
process	138	23	6	9	16
temporal	194	63	14	0	50
numeric	121	27	13	10	18
location	151	69	21	2	47
proper	231	76	10	0	32

Table 4: Coverage of TREC questions divided by most common answer types.

Table 4 shows TREC question coverage broken down by answer type. Due to temporal consistency, numeric questions are not covered very well. Although the process and object types are broad answer types, the coverage is still reasonably good. As expected, the definition and person-bio answer types are covered well by these resources.

8 The Web as a Resource

An increasing number of QA systems are using the web as a resource. Since the Web is orders of magnitude larger than local corpora, answers occur frequently in simple contexts, which is more conducive to retrieval and extraction of correct, confident answers (Clarke et al., 2001; Dumais et al., 2002; Lin and Katz, 2003). The web has been employed for pattern acquisition (Ravichandran et al., 2003), document retrieval (Dumais et al., 2002), query expansion (Yang et al., 2003), structured information extraction, and answer validation (Magnini et al., 2002) . Some of these approaches enhance existing QA systems, while others simplify the question answering task, allowing a less complex approach to find correct answers.

8.1 Web Documents

Instead of searching a local corpus, some QA systems retrieve relevant documents from the web (Xu et al., 2003). Since the density of relevant web documents can be higher than the density of relevant local documents, answer extraction may be more successful from the web. For a TREC evaluation, answers found on the web must also be mapped to relevant documents in the local corpus.

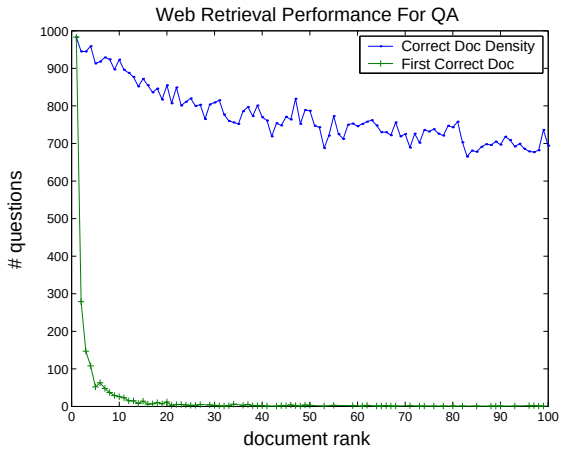

Figure 1: Web retrieval: relevant document density and rank of first relevant document.

In order to evaluate the impact of web documents on TREC questions, we performed an experiment where simple queries were submitted to a web search engine. The questions were tokenized and filtered using a standard stop word list. The resulting keyword queries were used to retrieve 100 documents through the Google API (*www.google.com/api*). Documents containing the full question, question number, references to *TREC*, *NIST*, *AQUAINT*, *Question Answering* and similar content were filtered out.

Figure 1 shows the density of documents containing a correct answer, as well as the rank of the first document containing a correct answer. The simple word query retrieves a relevant document for almost half of the questions. Note that for most systems, the retrieval performance should be superior since queries are usually more refined and additional query expansion is performed. However, this experiment provides an intuition and a very good lower bound on the precision and density of current web documents for the TREC QA task.

8.2 Web-Based Query Expansion

Several QA systems participating at TREC have used search engines for query expansion (Yang et al., 2003). The basic query expansion method utilizes pseudo-relevance feedback (PRF) (Xu and Croft, 1996). Content words are selected from questions and submitted as queries to a search engine. The top n retrieved documents are selected, and k terms or phrases are extracted according to an optimization criterion (e.g. term frequency, n-gram frequency, average mutual information using corpus statistics, etc). These k items are used in the expanded query.

We experimented by using the top 5, 10, 15, 20,

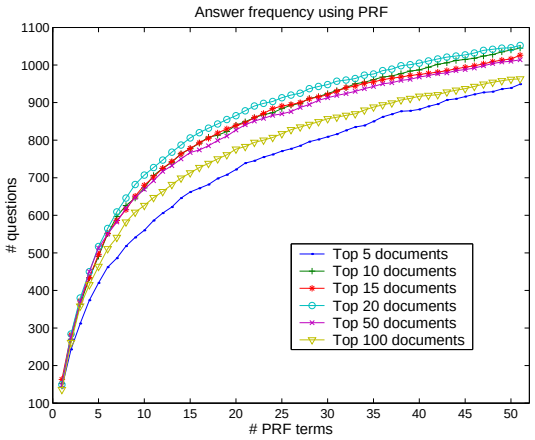

Figure 2: Finding a correct answer in PRF expansion terms - applied to 2183 questions for witch answer keys exist.

50, and 100 documents retrieved via the Google API for each question, and extracted the most frequent fifty n-grams (up to trigrams). The goal was to determine the quality of query expansion as measured by the density of correct answers already present in the expansion terms. Even without filtering n-grams matching the expected answer type, simple PRF produces the correct answer in the top n-grams for more than half the questions. The best correct answer density is achieved using PRF with only 20 web documents.

8.3 Conclusions

This paper quantifies the utility of well-known and widely-used resources such as WordNet, encyclopedias, gazetteers and the Web on question answering. The experiments presented in this paper represent loose bounds on the direct use of these resources in answering TREC questions. We reported the performance of these resources on different TREC collections and on different question types. We also quantified web retrieval performance, and confirmed that the web contains a consistently high density of relevant documents containing correct answers even when simple queries are used. The paper also shows that pseudo-relevance feedback alone using web documents for query expansions can produce a correct answer for fifty percent of the questions examined.

9 Acknowledgements

This work was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

References

C.L.A. Clarke, G.V. Cormack, and T.R. Lynam. 2001. Exploiting redundancy in question answering. *SIGIR*.

S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. 2002. Web question answering: Is more always better? *SIGIR*.

J. Leidner, J. Bos, T. Dalmas, J. Curran, S. Clark, C. Bannard, B. Webber, and M. Steedman. 2003. Qed: The edinburgh trec-2003 question answering system. *TREC*.

J. Lin and B. Katz. 2003. Question answering from the web using knowledge annotation and knowledge mining techniques. *CIKM*.

J. Lin. 2002. The web as a resource for question answering: Perspectives and challenges. *LREC*.

B. Magnini, M. Negri, R. Pervete, and H. Tanev. 2002. Is it the right answer? exploiting web redundancy for answer validation. *ACL*.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Five papers on wordnet. *International Journal of Lexicography*.

D. Moldovan, D. Clark, S. Harabagiu, and S. Maiorano. 2003. Cogex: A logic prover for question answering. *ACL*.

E. Nyberg, T. Mitamura, J. Callan, J. Carbonell, R. Frederking, K. Collins-Thompson, L. Hiyakumoto, Y. Huang, C. Huttenhower, S. Judy, J. Ko, A. Kupsc, L.V. Lita, V. Pedro, D. Svoboda, and B. Vand Durme. 2003. A multi strategy approach with dynamic planning. *TREC*.

D. Paranjpe, G. Ramakrishnan, and S. Srinivasan. 2003. Passage scoring for question answering via bayesian inference on lexical relations. *TREC*.

D. Ravichandran, A. Ittycheriah, and S. Roukos. 2003. Automatic derivation of surface text patterns for a maximum entropy based question answering system. *HLT-NAACL*.

E.M. Voorhees. 2003. Overview of the trec 2003 question answering track. *TREC*.

J. Xu and W.B. Croft. 1996. Query expansion using local and global analysis. *SIGIR*.

J. Xu, A. Licuanan, and R. Weischedel. 2003. Trec 2003 qa at bbn: Answering definitional questions. *TREC*.

H. Yang, T.S. Chua, S. Wang, and C.K. Koh. 2003. Structured use of external knowledge for event-based open domain question answering. *SIGIR*.