The Effect of Document Retrieval Quality on Factoid Question Answering Performance

Kevyn Collins-Thompson
Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
{kct,callan}@cs.cmu.edu

Egidio Terra
Charles L.A. Clarke
School of Computer Science
University of Waterloo
Canada
{elterra,claclark}@plg2.uwaterloo.ca

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Performance, Experimentation.

Keywords: Information Retrieval, Question Answering.

1. INTRODUCTION

A widely-used architecture for factoid question answering (QA) involves the use of a multi-step pipeline consisting of: 1) initial question analysis, 2) document and/or passage retrieval, and 3) answer extraction. In this study, we examine the relationship between the quality of document retrieval and the overall accuracy of QA systems. We evaluate two QA systems using TREC 2002 test set questions [9]: Carnegie Mellon's JAVELIN system [7] and Waterloo's MultiText QA system [2]. We adapt the two QA systems in order to use different sets of documents as input, and seven different document retrieval methods to create the list of documents including a combination of different systems. The set of known relevant documents was used as a baseline to compare the different retrieval methods. Documents with exact or inexact judgments are considered relevant.

Our main hypothesis for this study is that there is a positive relationship between improved document retrieval and QA accuracy across systems.

2. EMPIRICAL EVALUATION

To test our hypothesis we checked document retrieval effectiveness and the impact of the different document retrieval methods on complete QA systems. In the first experiment, 445 questions of the TREC 2002 QA track (those with answers in the AQUAINT collection) were given as input to each of five retrieval systems, as shown in Table 1. For systems 1 to 3, the retrieval component is already used in the context of a QA system. In addition to the JAVELIN and MultiText QA systems, we used the passage retrieval component of UMass Amherst's QA system for initial retrieval. The underlying retrieval models for these systems are radically different: JAVELIN uses a combined structured-query and tf.idf document retrieval model, Waterloo a cover density passage model, and UMass a language model approach. These systems also perform question analysis before retrieving documents and/or passages. The adhoc retrieval systems 4 and 5 did no relevance feedback and used standard stopping and stemming methods.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

Table 1: Description of retrieval methods/sytems

	System	Summary Description
1	CMU IR	JAVELIN question processing and document retrieval.
2	Waterloo IR	MultiText question processing and passage retrieval.
3	UMass IR	U. Mass. (Amherst) question processing and passage retrieval.
4	SABIR	The SMART [1] document retrieval system (adhoc, vector-space model).
5	CLARIT	The Claritech [3] document retrieval system (adhoc, vector space model).
6	Fusion	The top 30 documents based on the COMB-MNZ [4] algorithm.
7	Oracle Set	Perfect retrieval: all and only documents containing an answer.

All retrieval systems produced 30 ranked documents or passages per question from the AQUAINT collection. Two other sets of documents were created: a Fusion run from the top 30 merged results of systems 1–4 using a variant of COMB-MNZ [4]; and set of relevant documents (oracle) using judgments from TREC. In this experiment, we are only concerned with the relationship between document retrieval and overall QA accuracy. While two of the five systems (Waterloo, UMass) retrieve passages instead of documents, to make for consistent comparison, we do not examine passage retrieval accuracy in this study and only make use of each passage's document identifier. For an evaluation of passage retrieval in context of QA see Tellex *et al.* [8].

In the second experiment, we ran CMU's JAVELIN and Waterloo's MultiText QA systems end-to-end, using each of the seven above IR runs as input to the answer extraction phase. These two systems adopt very different strategies for answer extraction. JAVELIN uses support vector machines as the core of its answer extraction component while MultiText uses a named-entity recognizer and statistical features to select exact answers.

2.1 Document Retrieval Effectiveness

We used two statistics for each retrieval system: 1) Question coverage, the fraction of questions for which the system

Table 2: Retrieval metrics based on TREC 2002 questions using the top 30 retrieved documents

Retrieval Method	Question Coverage	MAP	p-value
Oracle Set	1.0000	1.0000	$< \mathbf{10^{-16}}$
Fusion (Comb-MNZ)	0.8315	0.2804	$< \mathbf{10^{-16}}$
UMass IR	0.7568	0.2383	0.5862
Waterloo IR	0.7506	0.2373	0.2852
CMU IR	0.7455	0.2175	$< \mathbf{10^{-10}}$
SABIR	0.6402	0.1327	$< \mathbf{10^{-5}}$
CLARIT	0.4449	0.1016	

found at least one document containing the answer; and 2) Mean Average Precision (MAP). Table 2 shows these statistics for all seven IR runs ranked by MAP. The 4th column in table 2 contains the p-value of the Wilcoxon signed-rank test calculated over MAP of adjacent retrieval sets.

The IR components of the three QA systems achieved almost identical coverage of about 75%, despite their very different retrieval models. This was significantly higher than the two adhoc retrieval methods used. Of all questions, 83 were hard: that is, not covered by any retrieval method except the oracle, giving a best possible combined coverage of 83.4%. The COMB-MNZ fusion run of 83.15% is very close to this limit, missing only one question covered by individual systems. We also evaluated the Condorcet-fuse [5] method on all possible subsets of the four systems. Merging all systems gave slightly lower coverage of 82.02%, but we saw slightly better performance for some combinations of two and three systems. A similar trend is observed in MAP results, where IR components of QA systems had a substantially better performance than adhoc IR. The COMB-MNZ fusion MAP performance is significantly better than all individual systems.

2.2 QA System Effectiveness

We used the above 7 retrieval runs as input to the answer-extraction component of MultiText and JAVELIN. We state overall QA system performance by Mean Reciprocal Rank (MRR) within the top 5 answers. The UMass system does not currently perform exact answer extraction and so was not used in this experiment. Note that for IR systems that provided passages, these were passed to the answer-extraction components. The MRR for all 14 combinations is given in Table 3. These figures show a significant gap between actual and optimal system performance.

When the improvement on MAP is significant there is a corresponding increasing in JAVELIN's MRR, with the exception of SABIR and CLARIT. The MultiText QA system has a clear preference for passages, especially for MultiText own passages, and its behavior when applied to documents is not conclusive. Nevertheless, MultiText QA improves with better document retrieval, as demonstrated by its performance on the oracle documents.

3. DISCUSSION AND CONCLUSIONS

We have presented results that illustrate the effect of seven document retrieval methods on the overall accuracy of two QA systems. We found that the retrieval methods specialized for QA, which typically included a question analysis phase, obtained better question coverage and MAP than adhoc IR methods. In an evaluation of IR in the TEQUESTA QA system, Monz [6] reported a similar preference for QA-tailored IR methods. A simple COMB-MNZ fusion run of the top four IR systems gave a 7.47% absolute increase in coverage and a significant improvement in MAP over the best individual system. The observed gap between actual and optimal QA performance for both systems, as measured by the oracle run, illustrates the potential gains from future improvements to document retrieval.

Drawing general conclusions about the relative effects of different retrieval methods when used with different QA systems is difficult, due to the complex interaction between retrieval and answer extraction. We did find that, while the response varied for different systems, there was a consistent relationship between the quality of initial document retrieval, and the performance of the overall QA system.

Table 3: MRR of QA systems

Retrieval Method	MultiText QA	JAVELIN QA
Oracle Set	0.5525	0.6214
Fusion (Comb-MNZ)	0.3112	0.3212
Waterloo IR	0.4946	0.3066
CMU IR	0.3115	0.2481
UMass IR	0.3368	0.2460
CLARIT	0.3200	0.2436
SABIR	0.2739	0.2357

Acknowledgments

This work was part of the NRRC RIA summer workshop.

4. REFERENCES

[1] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New Retrieval Approaches Using SMART: TREC-4. In *Proc. of TREC-4*, pp. 25–48, NIST, 1996.

[2] C. Clarke, G. Cormack, G. Kemkes, M. Laszlo, T. Lynam, E. Terra, and P. Tilker. Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002). In *Proc. of TREC*, pp. 823–831, NIST, 2002.

[3] D. A. Evans and R. G. Lefferts. CLARIT-TREC Experiments. *Information Processing and Management*, 31(3):385–395, 1995.

[4] E. Fox and J. Shaw. Combination of Multiple Searches. In *Proc. of TREC-2*, pp. 243–252, NIST, 1994.

[5] M. Montague and J. A. Aslam. Condorcet Fusion for Improved Retrieval. In *Proc. of CIKM*, pp. 538–548, 2002.

[6] C. Monz. *From Document Retrieval to Question Answering*. Ph. D. dissertation, Universiteit Van Amsterdam, December 2003.

[7] E. Nyberg, T. Mitamura, J. Carbonell, J. Callan, K. Collins-Thompson, K. Czuba, M. Duggan, L. Hiyakumoto, N. Hu, Y. Huang, J. Ko, L. Lita, S. Murtagh, V. Pedro, and D. Svoboda. The JAVELIN Question-Answering System. In *Proc. of TREC*, pp. 128–137, NIST, 2002.

[8] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR*, pp. 41–47. ACM Press, 2003.

[9] E. Voorhees. The TREC 2002 question answering track. In *Proc. of TREC*, pp. 115–123, NIST, 2002.