Indexing Low Frequency Information for Question Answering

Abolfazl Keighobadi Lamjiri, Julien Dubuc, Leila Kosseim, and Sabine Bergler

CLaC Laboratory, Department of Computer Science, Concordia University
Montreal, Québec, Canada, H3G-1M8

Abstract

This paper presents our experiments with a low-frequency approach to information retrieval for question answering over a small, closed domain corpus and a variety of question types. With a corpus of 255 questions categorized into simple, average and challenging, we compared the performance of our question answering system (QASCU) when used with two different information retrieval systems, Lucene and BioKI. Lucene uses a standard tf.idf weighting scheme on documents, while BioKI uses a weighted keyword occurrence optimization scheme on paragraphs, that does not bias against low-frequency terms. While IR with Lucene yields better IR results at the document level than BioKI, running QASCU on BioKI output achieves higher precision. This indicates that for closed domain QA with an IR component, the basic F-measure performance of the IR component at the document level is not necessarily indicative of the overall performance. We contend that the findings are relevant also to retrieval from video, text, and sound collections that usually feature low redundancy in the text snippets used for retrieval.

1. Introduction

Much recent research in question answering (QA) has been guided by the evaluation standards of shared tasks, such as TREC QA [VT99]. While the TREC gold standards have been a major contribution to the development of many systems, they also introduce a normative bias towards their task and particularities of their dataset. Thus, due to the availability of standard data sets, there has been a strong bias towards the use of newspaper articles in information retrieval and for QA the bias has been towards simple factoid questions that can be answered with a single word or phrase.

Traditionally, sample IR output has been made available to TREC QA participants who do not have access to an IR system, suggesting that improved QA can be largely achieved independent of IR and that IR engines do not need tailoring for the QA task. In this paper, we report on an experiment that probes this assumption in the context of a closed domain question answering system over questions formulated by proxy users. To test the influence of the IR component in the context of QA, we substituted the IR component of BioKI [BSDL06] for Lucene[1] in the QASCU QA system. BioKI's IR component takes a baseline approach of weighted keyword scoring that is not biased against low frequency information. While it shows consistently lower document level F-measure compared to Lucene, the results of QA with BioKI are slightly better. We stipulate that large collections of video, images, and sound with associated texts used for retrieval will feature low redundancy and thus behave more like a small, closed domain corpus in this respect and have to investigate low frequency IR techniques rather than standard, redundancy dependent models.

2. Previous Work

It is well understood in the IR community that one IR strategy cannot serve all purposes equally well [GHG04]. In particular, there is a marked difference in behavior of different IR strategies depending on the availability of redundancy. While many QA systems are able to capitalize on the redundancy of the Web and bypass the document collection for answering a question (e.g. [BLBDN01, LK03]), research on small document collections has been hampered by its absence (e.g. [HK04, [RHDMS04]). We observe the same lack of redundancy in the small, usually decontextualized text snippets used to index video, image, and sound data for retrieval purposes.

[1] Available at http://lucene.apache.org/

Most QA systems use a general-purpose IR system to return documents or relevant passages (e.g. [KEW01]); while some use a specialized IR method tailored specifically for the task of QA (e.g. [IFR01]). Recently, researchers have investigated the effect of alternative IR techniques for the specific task of QA [CM03], and specifically, the sensitivity of answer extraction performance to the initial document retrieval precision [JLin06]. While most work has been done on short factoid-type questions using the TREC data (e.g. [CXB04]), Zhuo, et al. in [ZDDLN04] use domain-specific knowledge to improve QA in a closed domain scenario, using a specialized ontology to identify domain terms in documents, and then associate them with more general semantic labels that are included in the indexation of the document collection. Their results show an improvement in mean reciprocal rank (MRR) performance. The difficulty with dealing with closed domains, however, is that textual genre and complexity differ significantly across studies, making comparisons difficult, and making results and techniques difficult to port from one domain to another. The best system working on the reading comprehension task, for example, reports an MRR of 40% [Mol03] for short and grammatically simple sentences. Our own QA system [KKR07] using Lucene also achieves an MRR of 39% for simple questions; but since the document collection and question set are significantly different, results are not comparable.

3. The IR and the QA Systems

Our QA system, called QASCU, was designed as a closed domain QA system but uses no domain-specific information in order to be more portable. Given the output of an IR system on a given document collection, QASCU extracts candidate answer sentences and then ranks them based on their semantic and syntactic similarity with the question.

Here, we test the effectiveness of an IR technique that targets low-frequency information on closed-domain QA. Developed for a literature navigation system in the micro-biology domain, BioKI explicitly addresses low-frequency information. It segments texts into paragraphs, then scores each paragraph using a scoring function based on weighed keywords that rewards closeness of the keywords in the paragraph and coverage of the keywords in the query found in the paragraph. This scoring strategy is lenient, it allows some keywords to be omitted, provided the score of a paragraph is above a threshold which is specified as part of the query. BioKI returns an ordered list of paragraphs for which keyword cooccurrence exceeds the threshold. In this experiment, all keywords were given equal weight. The paragraphs are output to the QA system for further processing. BioKI achieved average performance on the TREC Genomics 2006 document and passage retrieval tasks [BSDL06].

The QA system scores each sentence of the IR results, using the semantic relatedness of the question's main verb to the target verb, which is the verb in the candidate sentence that includes the question head (the most important noun phrase in the question [KKR07]). For this, we use Leacock and Chodorow's similarity measure from WordNet::Similarity [PPPM04]. Once the target verb is found to be semantically similar to the question, we try to match the target parse tree to the parse tree of the question by trying to unify the Minipar parses [Lin93] with a fuzzy unification method. The unification uses two features: the number of overlapping words based on a bag-of-words approach and the number of overlapping syntactic links [KKR07] for its ranking. QASCU participated in the TREC 2006 QA task [KKBR06] and ranked 10th out of 59 on factoid questions with an accuracy of 0.194. The version of the system used at TREC uses a combination of Web searching through a modified version of Aranea [LK03] and the parse-tree unifier described above using Lucene for IR on the AQUAINT corpus. On our closed domain corpus of 340 documents (see Section 4), using Lucene on the document collection, the parse-tree matcher alone achieves an accuracy of 0.24.

4. The Corpus and the Question Set

To create our closed domain collection, we asked 15 students to formulate questions and answers given a corporate document collection of Bell Canada. The document collection consists of 340 Web pages, internal documents and a few technical manuals (750KB of text). The questions produced vary in style, length and complexity; most are long and complex ‘what’ and ‘how’ type questions. Some questions require domain knowledge for acronym or synonym expansion, while others are lexically and syntactically similar to the answer sentence. We randomly chose 35% for development and kept the rest for testing. To assess the complexity of the QA collection, we compared the documents to several other textual genres. The most popular readability measures [Gre04] show that the Bell documents have on average longer sentences and are of lower reading ease compared to short stories[2], news articles from the AQUAINT corpus used in open domain QA [VT99], and grade 5 reading comprehension texts often used in QA [Mol03].

We categorized the questions according to the level of processing required to answer them with respect to the document collection. This helps in a step by step problem solving approach: we focus on the simple questions first, and build upon their solution for more complex questions.

Three categories were defined: simple, average, and challenging.

Simple Questions have most of their keywords appear in the answer sentence. In addition, the question and answer sentences share a simple and conventional grammatical structure. For example:

> Q: What does IP Messaging Enterprise Standard service include?
> A: IP Messaging Enterprise Standard service includes POP3, IMAP and Webmail mailbox access.

Average Questions are typically longer and exhibit a more complex grammatical structure that is not shared by the answer sentence. Question keywords might not appear in the answer sentence in the same lexical form and thus a lexical knowledge base such as WordNet [Fel98], or a dictionary is required to find semantic equivalence. For example, in the following, synonymy between *allow* and *provide* has to be established.

> Q: What does Virtual FaxTM provide?
> A: Virtual FaxTM allows you to send and receive faxes anywhere as quickly and easily as regular e-mail, using your PC or wireless device, such as a PDA or RIM Blackberry, through a web based or client based e-mail interface.

Challenging Questions require advanced NLP techniques such as coreference resolution and domain knowledge. In the following example, the phrases *no answer transfer* and *you don’t answer* are semantically equivalent in the domain, but the lexical and syntactic structures of the question and answers differ.

> Q: What is no answer transfer?
> A: When someone calls your mobile, it will ring there first. If you don’t answer, it will ring through to a second number predetermined by you.

The Bell Corpus question set contains 102 simple questions, 98 average questions, and 55 challenging questions, for a total of 255 questions.

5. Experiment

The first experiment was performed using the Lucene IR engine; while for the second, we used BioKI [BSDL06]. To overcome the problem of not finding all keywords expressed as stated, we designed a feedback loop for Lucene to relax the query iteratively until a minimum number of documents are retrieved if very few documents are initially returned. BioKI has an equivalent built-in lenient scoring mode. Once the top *n* documents (Lucene) or paragraphs (BioKI) are returned by the IR engine, they are sent to QASCU which will identify those sentences containing an answer. Only sentences that contain *x*% of the question keywords are kept. Experiments show

[2] We took 5 classic short stories from http://www.bnl.com/shorts/

that varying the threshold (from *x*=35% to *x*=75%) only affects the QA system's run time, but not its accuracy[3], showing relative robustness of our parse-tree ranking method towards noise. The result of the IR for our QA task is ultimately measured by the mean reciprocal rank (MRR) of the ranked list of candidate sentences. The baseline accuracy for the ranking task (i.e. ranking sentences at random) depends on the number of candidates extracted in the IR phase: the more candidates in the set, the harder the ranking task. With the output of BioKI, this baseline would yield a very low MRR of 5.9%; while Lucene would yield an even lower MRR of 3.1%.

6. Results

Figure 1 shows the results of Lucene and BioKI at the document level. Recall that Lucene extracts documents while BioKI extracts paragraphs from the corpus. Regardless of the IR system used, the F-measure decreases consistently for harder questions. This is in line with our observations that simple questions share more keywords with the sentences containing the answer, yielding better IR results. The figure also shows that Lucene does much better at the document level than BioKI; on average, twice the F-measure. BioKI here simply uses the sum of the internal scores of the paragraphs to approximate document rank, disregarding, for instance, the length of the document (thus a longer document with occurrence of the keywords throughout will score higher than a one paragraph document that contains the answer.)

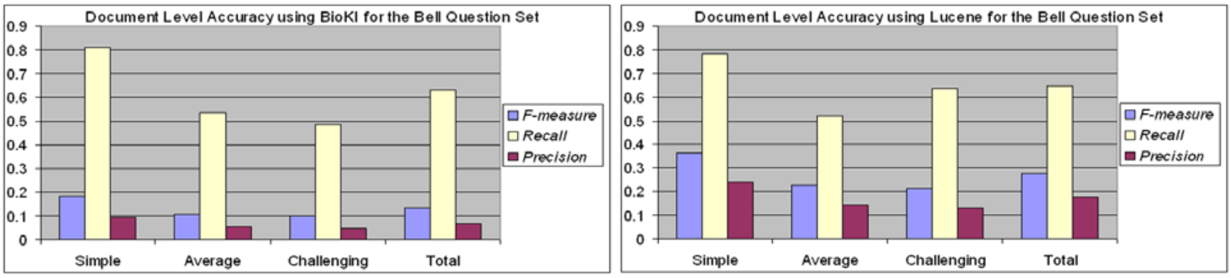

Figure 1: IR results with BioKI (left) and Lucene (right) at the document level

Figure 2 (right) shows the F-measure of both systems at the sentence-level. Here again, the relation between the F-measure and the question type holds. Both IR systems retrieve relevant sentences more often for simple questions compared to challenging questions. However, surprisingly, BioKI does systematically better than Lucene. BioKI returns a set of relevant paragraphs, selected and ordered from multiple documents; Lucene however has no filtering over the contents of a document. Better F-measure is actually a result of higher precision at the sentence level. This highlights that BioKI's focus on paragraphs is appropriate for QA. [JLin06] similarly report that recovery from low ranks of relevant documents is possible.

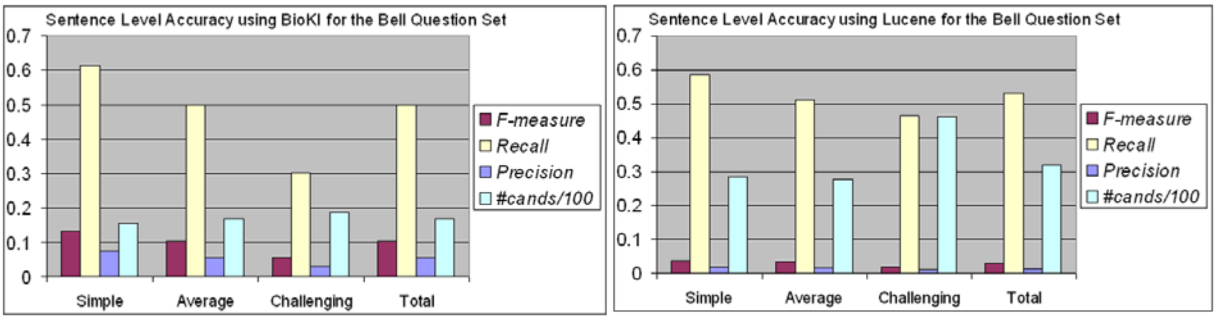

Figure 2: IR results with BioKI (left) and Lucene (right) at the sentence level

[3] How much this fact is due to the small size of the Bell Corpus is still under investigation.

Figure 3 shows the final accuracy of the QA system when using Lucene and BioKI. In Figure 3 (right), MRR refers to the traditional Mean Reciprocal Rank measure that scores the first correct answer based on its rank. The score of a correct answer is the inverse of its rank (1 for rank 1, ½ for rank 2, 1/3 for rank 3…). Figure 3 (left), however, shows the accuracy of the QA system, i.e. when only the first returned answer in the list (rank 1) is considered (1 point if the correct answer is at rank 1, 0 otherwise). A more precise set of candidate sentences results in higher MRR for BioKI (30%). The noise removed by BioKI helps to find the answer sentence more easily: on average, 17 sentences are extracted to be ranked for answer status from BioKI's results, as opposed to 32 sentences from Lucene's[4]. Finally, the fact that QASCU shows the same level of accuracy independent of the number of candidates returned illustrates the robustness of our parse-tree ranking [KKR07]. Again, MRR and MRR-1 are consistently higher for simple questions as opposed to more challenging questions.

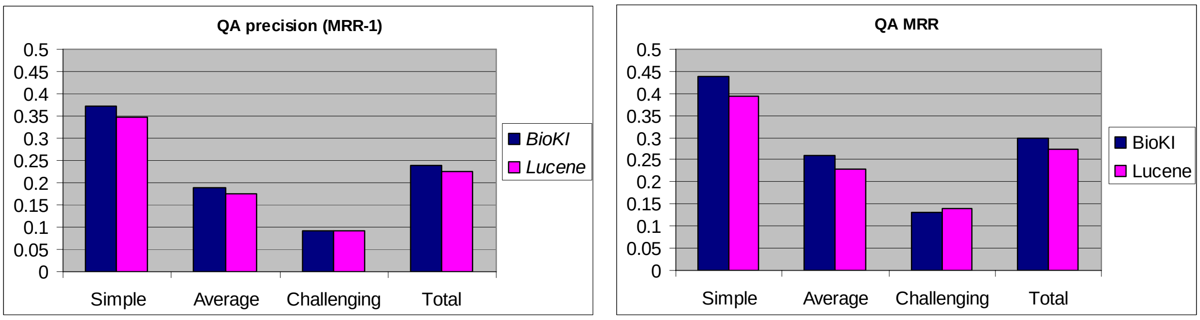

Figure 3: Accuracy (MRR-1) (left) and MRR (right) for the overall QA

7. Conclusions and Future Work

In this paper, we presented the results of a small experiment using two alternative IR methods to feed sentences to our QA system. Our experiment indicates several important axes of research. Firstly, looking at QA and IR mainly from the point of view of large shared tasks can blind research to the features of certain applications. This experiment shows that a basic proximity-based, weighted keyword approach to IR without feedback or term expansion is competitive to Lucene and has certain niche advantages. The potential of using domain resources to enable term expansion to raise the overall low score for answering challenging questions is an important open question. Secondly, this experiment has shown the robustness of our QA approach that uses linguistic features of the question and answer candidates for the final answer sentence ranking. Lucene and BioKI return results at different granularity (document vs. paragraph) and in different numbers. We would like to run Lucene on paragraphs in order to compare it with BioKI at the same level of granularity. Given that the different IR techniques have different advantages for QA, a combination model would be interesting, where different IR techniques are chosen depending on the question type. Whether our classification of simple, average, and challenging questions could be predicted in advance from the question/document set is another open question.

In summary, we believe that this small study has shown that a low frequency approach to IR and QA is important for document collections that exhibit a low degree of redundancy. While the Bell corpus is small, we predict similar behavior in large collections with low redundancy, including video, image, and sound databases and advocate similar studies of particular features to better address their requirements.

Acknowledgements

This project was funded by Bell University Laboratories and NSERC.

[4] This difference could be attributed to Lucene returning documents and BioKI paragraphs.

References

[BSDL06] Bergler S., Schuman J., Dubuc J. and Lebedev A. *BioKI, A general literature navigation system at TREC Genomics 2006.* In Notebook Proceedings of the 15th Text Retrieval Conference (TREC-15), pp 379–382, Gaithersburg, November 2006.

[BLBDN01] Brill E., Lin J., Banko M, Dumais S. and Ng A. *Data-intensive question answering*. In Proceedings of the 10th Text Retrieval Conference (TREC 2001). pp 393–400, Gaithersburg, November 2001.

[CM03] Monz C., *From Document Retrieval to Question Answering*. Ph.D. Thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2003.

[CXB04] Chang Y., Xu H., Bai S. *A Re-examination of IR Techniques in QA System*. First International Joint Conference on Natural Language Processing (IJCNLP), pp 71-80, China, March 2004.

[Fel98] Fellbaum C. WordNet: An Electronic Lexical Database, MIT Press, 1998.

[GHG04] Gaizauskas R., Hepple M. and Greenwood M. *Information Retrieval for Question Answering*, ACM SIGIR Forum 38(2), 2004.

[Gre04] Greenfield J. *Readability Formulas for EFL.* In JALT Journal, 2004.

[HK04] Doan-Nguyen H. and Kosseim L. *Improving the Precision of a Closed-Domain Question-Answering System with Semantic Information*. In Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO-2004), pp. 850-859, Avignon, April 2004.

[IFR01] Ittycheriah A., Frans M. and Roukos S. *IBM's statistical question answering system*, In Proceedings of the Tenth Text Retrieval Conference (TREC 2001). pp. 258–264, Gaithersburg, November 2001.

[KKR07] Keighobadi Lamjiri A., Kosseim L., Radhakrishnan T. *A Hybrid Unification Method for Question Answering in Closed Domains*, (to appear) In the 3rd International Workshop on Knowledge and Reasoning for Answering Questions (KRAQ'07), Hyderabad, India, January 2007.

[KBKR06] Kosseim L., Beaudoin A., Keighobadi Lamjiri A. and Razmara M. *Concordia University at the TREC-QA Track*. In Notebook Proceedings of the 15th Text Retrieval Conference (TREC-15), pp. 383-393, Gaithersburg, November 2006.

[KEW01] Kwok C., Etzioni O. and Weld D.S. *Scaling question answering to the web.* In Proceedings of the 10th International Conference on World Wide Web, pp 150–161, Hong Kong, 2001.

[Lin93] Dekang Lin. *Principle-based Parsing without Overgeneration.* In Proceedings of ACL-93, pp 112–120, USA, 1993.

[JLin06] Jimmy Lin. *The Role of Information Retrieval in Answering Complex Questions*. In Proceedings of COLING/ACL 2006, pp 523-530, Australia, 2006.

[LK03] Lin J. and Katz B. *Question Answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques.* In Proceedings of CIKM'03, USA, 2003.

[Mol03] Molla D. *Towards semantic-based overlap measures for Question Answering.* In Proceedings of Australasian Language Technology Workshop (ALTW), Melbourne, 2003.

[PPM04] Pedersen T., Patwardhan S. and Michelizzi J.. WordNet::Similarity -measuring the relatedness of concepts. In Proceedings of NAACL-04, Boston, USA, 2004.

[RHDMS04] F. Rinaldi, M. Hess, J. Dowdall, D. Mollá and R. Schwitter. Question Answering in Terminology-rich Technical Domains (2004)**.** In Mark Maybury (Ed.) *New Directions in Question Answering*. AAAI Press, pp. 71-82.

[VT99] Voorhees E. and Tice D. *The TREC-8 Question Answering Track Evaluation.* In Proceedings of the 8th Text REtrieval Conference (TREC-8), Gaithersburg, November 1999.

[ZDDLN04]Zhuo Z., Da Sylva L., Davidson C., Lizarralde G., Nie J-Y. Domain-Specific QA for the Construction Sector. In *Proceedings of the Workshop on Information Retrieval for Question Answering (IR4QA)*, SIGIR'04, pp. 65-71, Sheffield, September 2004.