The Role of Information Retrieval in Answering Complex Questions

Jimmy Lin
College of Information Studies
Department of Computer Science
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
jimmylin@umd.edu

Abstract

This paper explores the role of information retrieval in answering "relationship" questions, a new class complex information needs formally introduced in TREC 2005. Since information retrieval is often an integral component of many question answering strategies, it is important to understand the impact of different term-based techniques. Within a framework of sentence retrieval, we examine three factors that contribute to question answering performance: the use of different retrieval engines, relevance (both at the document and sentence level), and redundancy. Results point out the limitations of purely term-based methods to this challenging task. Nevertheless, IR-based techniques provide a strong baseline on top of which more sophisticated language processing techniques can be deployed.

1 Introduction

The field of question answering arose from the recognition that the document does not occupy a privileged position in the space of information objects as the most ideal unit of retrieval. Indeed, for certain types of information needs, sub-document segments are preferred—an example is answers to factoid questions such as "Who won the Nobel Prize for literature in 1972?" By leveraging sophisticated language processing capabilities, factoid question answering systems are able to pinpoint the exact span of text that directly satisfies an information need.

Nevertheless, IR engines remain integral components of question answering systems, primarily as a source of candidate documents that are subsequently analyzed in greater detail. Although this two-stage architecture was initially conceived as an expedient to overcome the computational processing bottleneck associated with more sophisticated but slower language processing technology, it has worked quite well in practice. The architecture has since evolved into a widely-accepted paradigm for building working systems (Hirschman and Gaizauskas, 2001).

Due to the reliance of QA systems on IR technology, the relationship between them is an important area of study. For example, how sensitive is answer extraction performance to the initial quality of the result set? Does better document retrieval necessarily translate into more accurate answer extraction? These answers cannot be solely determined from first principles, but must be addressed through empirical experiments. Indeed, a number of works have specifically examined the effects of information retrieval on question answering (Monz, 2003; Tellex et al., 2003), including a dedicated workshop at SIGIR 2004 (Gaizauskas et al., 2004). More recently, the importance of document retrieval has prompted NIST to introduce a document ranking subtask inside the TREC 2005 QA track.

However, the connection between QA and IR has mostly been explored in the context of factoid questions such as "Who shot Abraham Lincoln?", which represent only a small fraction of all information needs. In contrast to factoid questions, which can be answered by short phrases found within an individual document, there is a large class of questions whose answers require synthesis of information from multiple sources. The so-called definition/other questions at recent TREC evaluations (Voorhees, 2005) serve as good examples: "good answers" to these questions include in-

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 523–530,
Sydney, July 2006. ©2006 Association for Computational Linguistics

Qid 25: The analyst is interested in the status of Fidel Castro's brother. Specifically, the analyst would like information on his current plans and what role he may play after Fidel Castro's death.

vital	Raul Castro was formally designated his brother's successor
vital	Raul is the head of the Armed Forces
okay	Raul is five years younger than Castro
okay	Raul has enjoyed a more public role in running Cuba's Government.
okay	Raul is the number two man in the government's ruling Council of State

Figure 1: An example relationship question from TREC 2005 with its answer nuggets.

teresting "nuggets" about a particular person, organization, entity, or event. No single document can provide a complete answer, and hence systems must integrate information from multiple sources; cf. (Amigó et al., 2004; Dang, 2005).

This work focuses on so-called relationship questions, which represent a new and underexplored area in question answering. Although they require systems to extract information nuggets from multiple documents (just like definition/other questions), relationship questions demand a different approach (see Section 2). This paper explores the role of information retrieval in answering such questions, focusing primarily on three aspects: document retrieval performance, term-based measures of relevance, and term-based approaches to reducing redundancy. The overall goal is to push the limits of information retrieval technology and provide strong baselines against which linguistic processing capabilities can be compared.

The rest of this paper is organized as follows: Section 2 provides an overview of relationship questions. Section 3 describes experiments focused on document retrieval performance. An approach to answering relationship questions based on sentence retrieval is discussed in Section 4. A simple utility model that incorporates both relevance and redundancy is explored in Section 5. Before concluding, we discuss the implications of our experimental results in Section 6.

2 Relationship Questions

Relationship questions represent a new class of information needs formally introduced as a subtask in the NIST-sponsored TREC QA evaluations in 2005 (Voorhees, 2005). Previously, they were the focus of a small pilot study within the AQUAINT program, which resulted in an understanding of a "relationship" as the ability for one object to influence another. Objects in these questions can denote either entities (people, organization, countries, etc.) or events. Consider the following examples:

- Has pressure from China affected America's willingness to sell weaponry to Taiwan?
- Do the military personnel exchanges between Israel and India show an increase in cooperation? If so, what are the driving factors behind this increase?

Evidence for a relationship includes both the means to influence some entity and the motivation for doing so. Eight types of relationships ("spheres of influence") were noted: financial, movement of goods, family ties, co-location, common interest, and temporal connection.

Relationship questions are significantly different from definition questions, which can be paraphrased as "Tell me interesting things about *x*." Definition questions have received significant amounts of attention recently, e.g., (Hildebrandt et al., 2004; Prager et al., 2004; Xu et al., 2004; Cui et al., 2005). Research has shown that certain cue phrases serve as strong indicators for nuggets, and thus an approach based on matching surface patterns (e.g., appositives, parenthetical expressions) works quite well. Unfortunately, such techniques do not generalize to relationship questions because their answers are not usually captured by patterns or marked by surface cues.

Unlike answers to factoid questions, answers to relationship questions consist of an unsorted set of passages. For assessing system output, NIST employs the nugget-based evaluation methodology originally developed for definition questions; see (Voorhees, 2005) for a detailed description. Answers consist of units of information called "nuggets", which assessors manually create from system submissions and their own research (see example in Figure 1). Nuggets are divided into two types (“vital” and “okay”), and this distinction plays an important role in scoring. The official metric is an F_3-score, where nugget recall is computed on vital nuggets, and precision is based on a length allowance derived from the number of both vital and okay nuggets retrieved.

In the original NIST setup, human assessors were required to manually determine whether a particular system’s response contained a nugget. This posed a problem for researchers who wished to conduct formative evaluations outside the annual TREC cycle—the necessity of human involvement meant that system responses could not be rapidly, consistently, and automatically assessed. However, the recent introduction of POURPRE, an automatic evaluation metric for the nugget-based evaluation methodology (Lin and Demner-Fushman, 2005), fills this evaluation gap and makes possible the work reported here; cf. Nuggeteer (Marton and Radul, 2006).

This paper describes experiments with the 25 relationship questions used in the secondary task of the TREC 2005 QA track (Voorhees, 2005), which attracted a total of eleven submissions. Systems used the AQUAINT corpus, a three gigabyte collection of approximately one million news articles from the Associated Press, the New York Times, and the Xinhua News Agency.

3 Document Retrieval

Since information retrieval systems supply the initial set of documents on which a question answering system operates, it makes sense to optimize document retrieval performance in isolation. The issue of end–to–end system performance will be taken up in Section 4.

Retrieval performance can be evaluated based on the assumption that documents which contain one or more relevant nuggets (either vital or okay) are themselves relevant. From system submissions to TREC 2005, we created a set of relevance judgments, which averaged 8.96 relevant documents per question (median 7, min 1, max 21).

Our first goal was to examine the effect of different retrieval systems on performance. Two freely-available IR engines were compared: Lucene and Indri. The former is an open-source implementation of what amounts to be a modified *tf.idf* weighting scheme, while the latter employs a language modeling approach. In addition, we experimented with blind relevance feedback, a retrieval technique commonly employed to improve performance (Salton and Buckley, 1990). Following settings in typical IR experiments, the top twenty terms (by *tf.idf* value) from the top twenty documents were added to the original query in the feedback iteration.

	MAP	**R50**
Lucene	0.206	0.469
Lucene+brf	0.190 (−7.6%)$^{\circ}$	0.442 (−5.6%)$^{\circ}$
Indri	0.195 (−5.2%)$^{\circ}$	0.442 (−5.6%)$^{\circ}$
Indri+brf	0.158 (−23.3%)$^{\triangledown}$	0.377 (−19.5%)$^{\triangledown}$

Table 1: Document retrieval performance, with and without blind relevance feedback.

For each question, fifty documents from the AQUAINT collection were retrieved, representing the number of documents that a typical QA system might consider. The question itself was used verbatim as the IR query (see Section 6 for discussion). Performance is shown in Table 1. We measured Mean Average Precision (MAP), the most informative single-point metric for ranked retrieval, and recall, since it places an upper bound on the number of relevant documents available for subsequent downstream processing.

For all experiments reported in this paper, we applied the Wilcoxon signed-rank test to determine the statistical significance of the results. This test is commonly used in information retrieval research because it makes minimal assumptions about the underlying distribution of differences. Significance at the 0.90 level is denoted with a $^{\wedge}$ or $^{\vee}$, depending on the direction of change; at the 0.95 level, $^{\vartriangle}$ or $^{\triangledown}$; at the 0.99 level, $^{\blacktriangle}$ or $^{\blacktriangledown}$. Differences not statistically significant are marked with $^{\circ}$. Although the differences between Lucene and Indri are not significant, blind relevance feedback was found to hurt performance, significantly so in the case of Indri. These results are consistent with the findings of Monz (2003), who made the same observation in the factoid QA task.

There are a few caveats to consider when interpreting these results. First, the test set of 25 questions is rather small. Second, the number of relevant documents per question is also relatively small, and hence likely to be incomplete. Buckley and Voorhees (2004) have shown that evaluation metrics are not stable with respect to incomplete relevance judgments. Third, the distribution of relevant documents may be biased due to the small number of submissions, many of which used

Lucene. Due to these factors, one should interpret the results reported here as suggestive, not definitive. Follow-up experiments with larger data sets are required to produce more conclusive results.

4 Selecting Relevant Sentences

We adopted an extractive approach to answering relationship questions that views the task as sentence retrieval, a conception in line with the thinking of many researchers today (but see discussion in Section 6). Although oversimplified, there are several reasons why this formulation is productive: since answers consist of unordered text segments, the task is similar to passage retrieval, a well-studied problem (Callan, 1994; Tellex et al., 2003) where sentences form a natural unit of retrieval. In addition, the TREC novelty tracks have specifically tackled the questions of relevance and redundancy at the sentence level (Harman, 2002).

Empirically, a sentence retrieval approach performs quite well: when definition questions were first introduced in TREC 2003, a simple sentence-ranking algorithm outperformed all but the highest-scoring system (Voorhees, 2003). In addition, viewing the task of answering relationship questions as sentence retrieval allows one to leverage work in multi-document summarization, where extractive approaches have been extensively studied. This section examines the task of independently selecting the best sentences for inclusion in an answer; attempts to reduce redundancy will be discussed in the next section.

There are a number of term-based features associated with a candidate sentence that may contribute to its relevance. In general, such features can be divided into two types: properties of the document containing the sentence and properties of the sentence itself. Regarding the former type, two features come into play: the relevance score of the document (from the IR engine) and its rank in the result set. For sentence-based features, we experimented with the following:

- Passage match score, which sums the *idf* values of unique terms that appear in both the candidate sentence (S) and the question (Q):

$$\sum_{t \in S \cap Q} idf(t)$$

- Term *idf* precision and recall scores; cf. (Katz et al., 2005):

$$\mathcal{P} = \frac{\sum_{t \in S \cap Q} idf(t)}{\sum_{t \in A} idf(t)}, \mathcal{R} = \frac{\sum_{t \in S \cap Q} idf(t)}{\sum_{t \in Q} idf(t)}$$

- Length of the sentence (in non-whitespace characters).

Note that precision and recall values are bounded between zero and one, while the passage match score and the length of the sentence are both unbounded features.

Our baseline sentence retriever employed the passage match score to rank all sentences in the top *n* retrieved documents. By default, we used documents retrieved by Lucene, using the question verbatim as the query. To generate answers, the system selected sentences based on their scores until a hard length quota has been filled (trimming the final sentence if necessary). After experimenting with different values, we discovered that a document cutoff of ten yielded the highest performance in terms of POURPRE scores, i.e., all but the ten top-ranking documents were discarded.

In addition, we built a linear regression model that employed the above features to predict the nugget score of a sentence (the dependent variable). For the training samples, the nugget matching component within POURPRE was employed to compute the nugget score—this value quantified the "goodness" of a particular sentence in terms of nugget content.[1] Due to known issues with the vital/okay distinction (Hildebrandt et al., 2004), it was ignored for this computation; however, see (Lin and Demner-Fushman, 2006b) for recent attempts to address this issue.

When presented with a test question, the system ranked all sentences from the top ten retrieved documents using the regression model. Answers were generated by filling a quota of characters, just as in the baseline. Once again, no attempt was made to reduce redundancy.

We conducted a five-fold cross validation experiment using all sentences from the top 100 Lucene documents as training samples. After experimenting with different features, we discovered that a regression model with the following performed best: passage match score, document score, and sentence length. Surprisingly, adding

[1] Since the count variant of POURPRE achieved the highest correlation with official rankings, the nugget score is simply the highest fraction in terms of word overlap between the sentence and any of the reference nuggets.

Length	1000	2000	3000	4000	5000
F-Score					
baseline	0.275	0.268	0.255	0.234	0.225
regression	0.294 (+7.0%)°	0.268 (+0.0%)°	0.257 (+1.0%)°	0.240 (+2.5%)°	0.228 (+1.6%)°
Recall					
baseline	0.282	0.308	0.333	0.336	0.352
regression	0.302 (+7.2%)°	0.308 (+0.0%)°	0.336 (+0.8%)°	0.343 (+2.3%)°	0.358 (+1.7%)°
F-Score (all-vital)					
baseline	0.699	0.672	0.632	0.592	0.558
regression	0.722 (+3.3%)°	0.672 (+0.0%)°	0.632 (+0.0%)°	0.593 (+0.2%)°	0.554 (−0.7%)°
Recall (all-vital)					
baseline	0.723	0.774	0.816	0.834	0.856
regression	0.747 (+3.3%)°	0.774 (+0.0%)°	0.814 (−0.2%)°	0.834 (+0.0%)°	0.848 (−0.8%)°

Table 2: Question answering performance at different answer length cutoffs, as measured by POURPRE.

Length	1000	2000	3000	4000	5000
F-Score					
Lucene	0.275	0.268	0.255	0.234	0.225
Lucene+brf	0.278 (+1.3%)°	0.268 (+0.0%)°	0.251 (−1.6%)°	0.231 (−1.2%)°	0.215 (−4.3%)°
Indri	0.264 (−4.1%)°	0.260 (−2.7%)°	0.241 (−5.4%)°	0.222 (−5.0%)°	0.212 (−5.8%)°
Indri+brf	0.270 (−1.8%)°	0.257 (−3.8%)°	0.235 (−7.8%)°	0.221 (−5.7%)°	0.206 (−8.2%)°
Recall					
Lucene	0.282	0.308	0.333	0.336	0.352
Lucene+brf	0.285 (+1.3%)°	0.308 (+0.0%)°	0.319 (−4.2%)°	0.322 (−4.2%)°	0.324 (−7.9%)°
Indri	0.270 (−4.1%)°	0.300 (−2.5%)°	0.306 (−8.2%)°	0.308 (−8.1%)°	0.320 (−9.2%)°
Indri+brf	0.276 (−2.0%)°	0.296 (−3.6%)°	0.299 (−10.4%)°	0.307 (−8.5%)°	0.312 (−11.3%)°

Table 3: The effect of using different document retrieval systems on answer quality.

the term match precision and recall features to the regression model decreased overall performance slightly. We believe that precision and recall encodes information already captured by the other features.

Results of our experiments are shown in Table 2 for different answer lengths. Following the TREC QA track convention, all lengths are measured in non-whitespace characters. Both the baseline and regression conditions employed the top ten documents supplied by Lucene. In addition to the F$_3$-score, we report the recall component only (on vital nuggets). For this and all subsequent experiments, we used the (count, macro) variant of POURPRE, which was validated as producing the highest correlation with official rankings. The regression model yields higher scores at shorter lengths, although none of these differences were significant. In general, performance decreases with longer answers because both variants tend to rank relevant sentences before non-relevant ones.

Our results compare favorably to runs submitted to the TREC 2005 relationship task. In that evaluation, the best performing automatic run obtained a POURPRE score of 0.243, with an average answer length of 4051 character per question.

Since the vital/okay nugget distinction was ignored when training our regression model, we also evaluated system output under the assumption that all nuggets were vital. These scores are also shown in Table 2. Once again, results show higher POURPRE scores for shorter answers, but these differences are not statistically significant. Why might this be so? It appears that features based on term statistics alone are insufficient to capture nugget relevance. We verified this hypothesis by building a regression model for all 25 questions: the model exhibited an R^2 value of only 0.207.

How does IR performance affect the final system output? To find out, we applied the baseline sentence retrieval algorithm (which uses the passage match score only) on the output of different document retrieval variants. These results are shown in Table 3 for the four conditions discussed in the previous section: Lucene and Indri, with and without blind relevance feedback.

Just as with the document retrieval results, Lucene alone (without blind relevance feedback) yielded the highest POURPRE scores. However, none of the differences observed were statistically significant. These numbers point to an interesting interaction between document retrieval and question answering. The decreases in performance at-

Length	**1000**	**2000**	**3000**	**4000**	**5000**
F-Score					
baseline	0.275	0.268	0.255	0.234	0.225
baseline+max	0.311 (+13.2%)$^{\wedge}$	0.302 (+12.8%)$^{\blacktriangle}$	0.281 (+10.5%)$^{\blacktriangle}$	0.256 (+9.5%)$^{\vartriangle}$	0.235 (+4.6%)$^{\circ}$
baseline+avg	0.301 (+9.6%)$^{\circ}$	0.294 (+9.8%)$^{\wedge}$	0.271 (+6.5%)$^{\wedge}$	0.256 (+9.5%)$^{\vartriangle}$	0.237 (+5.6%)$^{\circ}$
regression+max	0.275 (+0.3%)$^{\circ}$	0.303 (+13.3%)$^{\wedge}$	0.275 (+8.1%)$^{\circ}$	0.258 (+10.4%)$^{\circ}$	0.244 (+8.4%)$^{\circ}$
Recall					
baseline	0.282	0.308	0.333	0.336	0.352
baseline+max	0.324 (+15.1%)$^{\wedge}$	0.355 (+15.4%)$^{\vartriangle}$	0.369 (+10.6%)$^{\vartriangle}$	0.369 (+9.8%)$^{\vartriangle}$	0.369 (+4.7%)$^{\circ}$
baseline+avg	0.314 (+11.4%)$^{\circ}$	0.346 (+12.3%)$^{\wedge}$	0.354 (+6.2%)$^{\wedge}$	0.369 (+9.8%)$^{\vartriangle}$	0.371 (+5.5%)$^{\circ}$
regression+max	0.287 (+2.0%)$^{\circ}$	0.357 (+16.1%)$^{\wedge}$	0.360 (+8.0%)$^{\circ}$	0.371 (+10.4%)$^{\wedge}$	0.379 (+7.6%)$^{\circ}$

Table 4: Evaluation of different utility settings.

tributed to blind relevance feedback in end–to–end QA were in general *less* than the drops observed in the document retrieval runs. It appears possible that the sentence retrieval algorithm was able to recover from a lower-quality result set, i.e., one with relevant documents ranked lower. Nevertheless, just as with factoid QA, the coupling between IR and answer extraction merits further study.

5 Reducing Redundancy

The methods described in the previous section for choosing relevant sentences do not take into account information that may be conveyed more than once. Drawing inspiration from research in sentence-level redundancy within the context of the TREC novelty track (Allan et al., 2003) and work in multi-document summarization, we experimented with term-based approaches to reducing redundancy.

Instead of selecting sentences for inclusion in the answer based on relevance alone, we implemented a simple utility model, which takes into account sentences that have already been added to the answer A. For each candidate c, utility is defined as follows:

$$\text{Utility}(c) = \text{Relevance}(c) - \lambda \max_{s \in A} sim(s, c)$$

This model is the baseline variant of the Maximal Marginal Relevance method for summarization (Goldstein et al., 2000). Each candidate is compared to all sentences that have already been selected for inclusion in the answer. The maximum of these pairwise similarity comparisons is deducted from the relevance score of the sentence, subjected to λ, a parameter that we tune. For our experiments, we used cosine distance as the similarity function. All relevance scores were normalized to a range between zero and one.

At each step in the answer generation process, utility values are computed for all candidate sentences. The one with the highest score is selected for inclusion in the final answer. Utility values are then recomputed, and the process iterates until the length quota has been filled.

We experimented with two different sources for the relevance scores: the baseline sentence retriever (passage match score only) and the regression model. In addition to taking the max of all pairwise similarity values, as in the above formula, we also experimented with the average.

Results of our runs are shown in Table 4. We report values for the baseline relevance score with the max and avg aggregation functions, as well as the regression relevance scores with max. These experimental conditions were compared against the baseline run that used the relevance score only (no redundancy penalty). To compute the optimal λ, we swept across the parameter space from zero to one in increments of a tenth. We determined the optimal value of λ by averaging POURPRE scores across all length intervals. For all three conditions, we discovered 0.4 to be the optimal value.

These experiments suggest that a simple term-based approach to reducing redundancy yields statistically significant gains in performance. This result is not surprising since similar techniques have proven effective in multi-document summarization. Empirically, we found that the max operator outperforms the avg operator in quantifying the degree of redundancy. The observation that performance improvements are more noticeable at shorter answer lengths confirms our intuitions. Redundancy is better tolerated in longer answers because a redundant nugget is less likely to "squeeze out" a relevant, novel nugget.

While it is productive to model the relationship task as sentence retrieval where independent decisions are made about sentence-level relevance, this simplification fails to capture overlap in information content, and leads to redundant answers. We found that a simple term-based approach was effective in tackling this issue.

6 Discussion

Although this work represents the first formal study of relationship questions that we are aware of, by no means are we claiming a solution—we see this as merely the first step in addressing a complex problem. Nevertheless, information retrieval techniques lay the groundwork for systems aimed at answering complex questions. The methods described here will hopefully serve as a starting point for future work.

Relationship questions represent an important problem because they exemplify complex information needs, generally acknowledged as the future of QA research. Other types of complex needs include analytical questions such as "How close is Iran to acquiring nuclear weapons?", which are the focus of the AQUAINT program in the U.S., and opinion questions such as "How does the Chilean government view attempts at having Pinochet tried in Spanish Court?", which were explored in a 2005 pilot study also funded by AQUAINT. In 2006, there will be a dedicated task within the TREC QA track exploring complex questions within an interactive setting. Furthermore, we note the convergence of the QA and summarization communities, as demonstrated by the shift from generic to query-focused summaries starting with DUC 2005 (Dang, 2005). This development is also compatible with the conception of "distillation" in the current DARPA GALE program. All these trends point to same problem: how do we build advanced information systems to address complex information needs?

The value of this work lies in the generality of IR-based approaches. Sophisticated linguistic processing algorithms are typically unable to cope with the enormous quantities of text available. To render analysis more computationally tractable, researchers commonly employ IR techniques to reduce the amount of text under consideration. We believe that the techniques introduced in this paper are applicable to the different types of information needs discussed above.

While information retrieval techniques form a strong baseline for answering relationship questions, there are clear limitations of term-based approaches. Although we certainly did not experiment with every possible method, this work examined several common IR techniques (e.g., relevance feedback, different term-based features, etc.). In our regression experiments, we discovered that our feature set was unable to adequately capture sentence relevance. On the other hand, simple IR-based techniques appeared to work well at reducing redundancy, suggesting that determining content overlap is a simpler problem.

To answer relationship questions well, NLP technology must take over where IR techniques leave off. Yet, there are a number of challenges, the biggest of which is that question classification and named-entity recognition, which have worked well for factoid questions, are not applicable to relationship questions, since answer types are difficult to anticipate. For factoids, there exists a significant amount of work on question analysis—the results of which include important query terms and the expected answer type (e.g., person, organization, etc.). Relationship questions are more difficult to process: for one, they are often not phrased as direct *wh*-questions, but rather as indirect requests for information, statements of doubt, etc. Furthermore, since these complex questions cannot be answered by short noun phrases, existing answer type ontologies are not very useful. For our experiments, we decided to simply use the question verbatim as the query to the IR systems, but undoubtedly performance can be gained by better query formulation strategies. These are difficult challenges, but recent work on applying semantic models to QA (Narayanan and Harabagiu, 2004; Lin and Demner-Fushman, 2006a) provide a promising direction.

While our formulation of answering relationship questions as sentence retrieval is productive, it clearly has limitations. The assumption that information nuggets do not span sentence boundaries is false and neglects important work in anaphora resolution and discourse modeling. The current setup of the task, where answers consist of unordered strings, does not place any value on coherence and readability of the responses, which will be important if the answers are intended for human consumption. Clearly, there are ample opportunities here for NLP techniques to shine.

The other value of this work lies in its use of an automatic evaluation metric (POURPRE) for system development—the first instance in complex

QA that we are aware of. Prior to the introduction of this automatic scoring technique, studies such as this were difficult to conduct due to the necessity of involving humans in the evaluation process. POURPRE was developed to enable rapid exploration of the solution space, and experiments reported here demonstrate its usefulness in doing just that. Although automatic evaluation metrics are no stranger to other fields such as machine translation (e.g., BLEU) and document summarization (e.g., ROUGE, BE, etc.), this represents a new development in question answering research.

7 Conclusion

Although many findings in this paper are negative, the conclusions are positive for NLP researchers. An exploration of a variety of term-based approaches for answering relationship questions has demonstrated the impact of different techniques, but more importantly, this work highlights limitations of purely IR-based methods. With a strong baseline as a foundation, the door is wide open for the integration of natural language understanding techniques.

8 Acknowledgments

This work has been supported in part by DARPA contract HR0011-06-2-0001 (GALE). I would like to thank Esther and Kiri for their loving support.

References

J. Allan, C. Wade, and A. Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *SIGIR 2003*.

E. Amigó, J. Gonzalo, V. Peinado, A. Peñas, and F. Verdejo. 2004. An empirical study of information synthesis task. In *ACL 2004*.

C. Buckley and E. Voorhees. 2004. Retrieval evaluation with incomplete information. In *SIGIR 2004*.

J. Callan. 1994. Passage-level evidence in document retrieval. In *SIGIR 1994*.

H. Cui, M.-Y. Kan, and T.-S. Chua. 2005. Generic soft pattern models for definitional question answering. In *SIGIR 2005*.

H. Dang. 2005. Overview of DUC 2005. In *DUC 2005*.

R. Gaizauskas, M. Hepple, and M. Greenwood. 2004. *Proceedings of the SIGIR 2004 Workshop on Information Retrieval for Question Answering (IR4QA)*.

J. Goldstein, V. Mittal, J. Carbonell, and J. Callan. 2000. Creating and evaluating multi-document sentence extract summaries. In *CIKM 2000*.

D. Harman. 2002. Overview of the TREC 2002 novelty track. In *TREC 2002*.

W. Hildebrandt, B. Katz, and J. Lin. 2004. Answering definition questions with multiple knowledge sources. In *HLT/NAACL 2004*.

L. Hirschman and R. Gaizauskas. 2001. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300.

B. Katz, G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller, O. Uzuner, and A. Wilcox. 2005. External knowledge sources for question answering. In *TREC 2005*.

J. Lin and D. Demner-Fushman. 2005. Automatically evaluating answers to definition questions. In *HLT/EMNLP 2005*.

J. Lin and D. Demner-Fushman. 2006a. The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. In *SIGIR 2006*.

J. Lin and D. Demner-Fushman. 2006b. Will pyramids built of nuggets topple over? In *HLT/NAACL 2006*.

G. Marton and A. Radul. 2006. Nuggeteer: Automatic nugget-based evaluation using descriptions and judgements. In *HLT/NAACL 2006*.

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

S. Narayanan and S. Harabagiu. 2004. Question answering based on semantic structures. In *COLING 2004*.

J. Prager, J. Chu-Carroll, and K. Czuba. 2004. Question answering using constraint satisfaction: QA–by–Dossier–with–Constraints. In *ACL 2004*.

G. Salton and C. Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297.

S. Tellex, B. Katz, J. Lin, G. Marton, and A. Fernandes. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR 2003*.

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

E. Voorhees. 2005. Overview of the TREC 2005 question answering track. In *TREC 2005*.

J. Xu, R. Weischedel, and A. Licuanan. 2004. Evaluation of an extraction-based approach to answering definition questions. In *SIGIR 2004*.