LAMP-TR-130 February 2006
CS-TR-4787
UMIACS-TR-2006-11

The Role of Information Retrieval in Answering Relationship Questions

Jimmy Lin

College of Information Studies
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
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 document retrieval is often an integral component of many question answering strategies, it is important to understand the impact of different information retrieval techniques. Within an approach based on sentence retrieval, this work examines three factors that contribute to question answering performance: the use of different retrieval engines, relevance (both at the document level and at the 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 layered.

Last updated: March 2, 2006
Keywords: question answering, relationship questions, information retrieval, evaluation

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 more 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 systems remain an integral component 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, and the architecture has since evolved into a widely-accepted paradigm for building working systems (Hirschman and Gaizauskas, 2001).

Due to the reliance of question answering systems on information retrieval, the relationship between them has been an 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 question answering? The answers to these questions cannot be derived solely from first principles, but rather must be determined empirically. Indeed, many works have specifically examined the effects of document retrieval on question answering (Monz, 2003; Tellex et al., 2003), including a dedicated workshop at SIGIR 2004 (Gaizauskas et al., 2004). The importance of document retrieval has prompted NIST to introduce a document ranking subtask inside the 2005 TREC QA track to specifically examine IR-related issues.

However, the connection between QA and IR has mostly been explored in the context of factoid questions, which represent only a small fraction of all information needs. In contrast to factoid questions, which can be answered by short phrases found within a single document, there is a large class of questions which require a system to synthesize answers from multiple sources. The so-called "definition" or "other" questions at recent TREC evaluations (Voorhees, 2005) serve as good examples: "good answers" to these questions include interesting "nuggets" about a particular person, organization, entity, or event. It is obvious that no single document can supply all the relevant nuggets that comprise a complete answer. Techniques for addressing such information needs have been previously explored (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 works quite well. Unfortunately, such techniques do not generalize well to other types of complex questions.

This work focuses on so-called "relationship" questions, which, like "definition" questions, require extracting and composing information nuggets from multiple documents. These questions, however, represent a new and unexplored area in question answering. This paper examines the role of information retrieval in systems designed to answer relationship questions, focusing primarily on three aspects: document retrieval performance, various 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 a strong baseline on which to add linguistic processing capabilities.

The rest of this paper is organized as follows: Section 2 provides an overview of so-called relationship questions, introduced at TREC 2005. 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.

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: A relationship question and reference nuggets created by an assessor.

2 Relationship Questions

Relationship questions represent an entirely 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 (Advanced QUestion Answering for INTelligence) program, which resulted in an understanding of a "relationship" as the ability of one object to influence another. Objects in these questions can denote both entities (people, organization, countries, etc.) or events. An example is "Has pressure from China affected America's willingness to sell high-tech weaponry to Taiwan?" Evidence for a relationship includes both the means to influence something 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.

Unlike answers to factoid questions, answers to relationship questions consist of unsorted sets of passages. For assessing answers, 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 his or her own research (see example in Figure 1). Nuggets are divided into two types ("vital" and "okay"), and this distinction plays an important role in the 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 ascertain 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 evaluation cycle—the necessity of human involvement meant that system responses could not be rapidly and automatically assessed. However, the recent introduction of POURPRE, an automatic evaluation metric for the nugget-based evaluation methodology (Lin and Demner-Fushman, 2005a), fills this evaluation gap and makes possible the work reported here.

This paper describes experiments with the twenty-five relationship questions used in the 2005 TREC QA track (Voorhees, 2005), which received a total of eleven submitted runs. Systems extracted answers from the AQUAINT corpus, a three gigabyte collection of approximately a 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 would make sense to optimize document retrieval performance in isolation. The

	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.

question of how document retrieval performance affects question answering performance will be taken up in Section 4.

Document retrieval performance can be evaluated based on the assumption that documents which contain relevant nuggets (either vital or okay) are themselves relevant. Such documents can be extracted from system submissions in the TREC 2005 QA track. In this manner, we created a set of relevance judgments, which averaged 8.96 relevant documents per question (median 7, min 1, max 21).

We compared two freely-available document retrieval engines: Lucene[1] and Indri[2]. 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 (Metzler and Croft, 2004). In addition, we experimented with blind relevance feedback, a commonly-employed technique in information retrieval to improve performance (Salton and Buckley, 1990). Following typical settings used in 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.

For each question, fifty documents from the AQUAINT collection were retrieved (using the question verbatim as the query), representing the number of documents that a typical QA system might consider. Performance is shown in Table 1. We measured Mean Average Precision, 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, $^{\triangle}$ 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 statistically 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 discovered that blind relevance feedback hurt retrieval performance in the factoid task.

There are a few caveats that one should consider when interpreting these results. First, the test set of twenty-five questions is rather small; in *ad hoc* retrieval, approximately fifty TREC topics are required to obtain confident results (Voorhees and Buckley, 2002). Second, the number of relevant documents per question is also 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 and the popularity of Lucene in question answering systems. Due to these three factors, one should interpret the results reported here as suggestive, not definitive. Larger data sets and more detailed analyses are required to produce conclusive results.

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

[2]http://www.lemurproject.org/

4 Selecting Relevant Sentences

We adopt an extractive approach to answering relationship questions that views the task as sentence retrieval. This conception is much in line with the thinking of many researchers today. There are several reasons why such a formulation is productive: since answers consist of unordered sets of text segments, the task is similar to passage retrieval, a well-studied problem (Callan, 1994; Mochizuki et al., 2000; Tellex et al., 2003) where sentences form a natural unit of retrieval. In addition, the novelty track at TREC has specifically tackled the questions of relevance and redundancy at the sentence level (Harman, 2002).

Empirically, an IR-based 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 run (Voorhees, 2003).[3] 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, without regard to any of the other already-selected sentences (which may naturally result in sentences conveying redundant information). 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 major features come into play: the relevance score of the document (from the IR engine) and its rank in the retrieved 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):

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

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

- 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 simply employs 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 selects sentences based on their score until a hard length quota has been filled (trimming sentences if necessary). After experimenting with different values, we discovered that a document cutoff of ten yielded the highest performance in terms of POURPRE scores.

[3] Albeit recall was more heavily favored at that time.

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.

In addition, we constructed 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.[4] The distinction between vital and okay nuggets was not taken into account in computing the nugget score, since problems associated with this division have been pointed out (Hildebrandt et al., 2004). Generally, interannotator agreement between vital and okay nuggets is low (Lin and Demner-Fushman, 2005b), so it would be unproductive to attempt to learn this inherently unstable distinction. When presented with a question, the system ranks sentences from the top ten retrieved documents using the regression model. Answers are generated by filling a quota of characters, just as in the baseline.

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

Results of our experiments are shown in Table 2 for answers of different lengths. Following the TREC QA track convention, all lengths are measured in terms of non-whitespace characters. As previously mentioned, both the baseline and regression conditions employed the top ten documents supplied by Lucene. In addition to the F-score (β=3), 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 length cutoffs, although none of the differences are statistically significant. In general, performance decreases as the length of the answer increases because both variants tend to place relevant sentences before non-relevant ones; i.e., the density of nuggets decreases as the answer length increases. This is exactly what we expect, since our term-based features are capturing at least some of the variance of sentence-level relevance.

Since training samples presented to our regression model did not preserve the vital/okay distinction, 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 the differences are not significant.

Why might this be so? It appears that features based on term statistics alone are insufficient to capture the variance exhibited by nugget relevance. We verified this hypothesis by building a regression

[4]Since the count variant of POURPRE was reported to yield 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					
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.

model for all twenty five questions; its R^2 value was merely 0.207. Although it may be possible to devise more term-based features to capture additional variance (e.g., taking into account term density), we strongly suspect that significantly better performance can only be achieved by attempts to actually understand language.

Our results compare favorably to runs submitted in to the 2005 TREC QA track. In that evaluation, the best performing automatic run obtained a POURPRE F-score of 0.243, with an average answer length of 4051 non-whitespace character per question.

How do different document sets affect question answering performance? To find out, we applied the baseline sentence retrieval algorithm (which uses the passage match score only) on the output of different document retrieval engines. 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. Nevertheless, these figures suggest that document retrieval performance does indeed affect end–to–end performance on relationship questions. In contrast to factoid questions, which require only one correct answer instance, answers to relationship questions require a system to extract nuggets from multiple documents, thereby placing more importance on the overall quality of the result set.

5 Reducing Redundancy

The methods described in the previous section for choosing relevant sentences do not take into account redundant 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 the Maximal Marginal Relevance method for multi-document summarization (Goldstein et al., 2000), we experimented with attempts to reduce redundancy using term-based similarity measures.

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

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

The candidate sentence is compared to all sentences that have thus far been selected in the answer. The maximum of these pairwise similarity comparisons is deducted from the relevance score of the sentence, subjected to a redundancy penalty λ, a parameter that we tune. For our experiments, we

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%)$^{\triangle}$	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%)$^{\triangle}$	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%)$^{\triangle}$	0.369 (+10.6%)$^{\triangle}$	0.369 (+9.8%)$^{\triangle}$	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%)$^{\triangle}$	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.

used cosine distance as the similarity function. All relevance scores are 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 (sentence are trimmed if necessary).

We experimented with two different sources for the relevance scores: the baseline sentence retriever (using the 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 are compared against the baseline relevance score without a redundancy penalty. To compute the optimal value for 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 the POURPRE F-score across all length intervals. For all three conditions, we discovered 0.4 to be the optimal value.

Statistically significant gains in performance can be attributed to a simple term-based approach to reducing redundancy. This result is not surprisingly since similar techniques have proven effective in multi-document summarization and related tasks. Empirically, the max operator was found to outperform the avg operator in quantifying the degree of redundancy. The observation that performance improvements are more noticeable for shorter answer lengths confirms our intuitions. Redundancy is better tolerated in longer answers because a redundant answer has less of a chance to "squeeze out" a relevant nugget that is also novel.

The conclusions from these experiments are fairly clear: while it is productive to model answering relationship question as independent decisions about sentence-level relevance, this simplification fails to capture the overlap in information content that results in redundant answers. A simple term-based approach to tacking this issue was found to be highly effective.

6 Discussion

Overall, this work presents two take-away messages. First, *ad hoc* retrieval, document retrieval for factoid question answering, and document retrieval for answering relationship questions all represent distinctive tasks, despite superficial similarities. Techniques that work well for one do not necessarily work for the others—blind relevance feedback being an illustrative example. This finding supports the need for component-level analyses as part of an overall research agenda.

Second, while information retrieval techniques form a strong baseline for answering relationship

questions, there are clear limitations of term-based approaches. Although we have certainly not tried every possible method, this work represents an exploration of the "obvious" techniques. As our regression results suggest, a variety of exclusively term-based features is unable to capture the variance in sentence-level relevance. On the other hand, however, simple IR-based techniques appear to work well at reducing redundancy, suggesting that determining information content overlap is simpler than determining relevance with respect to an information need.

We believe that to answer relationship questions well, language processing techniques must take over where information retrieval techniques leave off. Yet, there are a number of challenges, the biggest of which is that question classification and named-entity recognition techniques, which have worked well for factoid questions, are not applicable to relationship questions, since answer types are difficult to anticipate. Recent work on applying semantic models to QA (Narayanan and Harabagiu, 2004) provide a promising direction, since they can provide computational models for different types of "relationships".

The biggest contribution of this work is that it provides a solid foundation for a system devoted to complex information needs. As far as we are aware, this is the first in-depth study of relationship questions in the literature. Since information retrieval techniques are generally applicable to other domains and information needs, this work can be leveraged to tackle other types of complex questions, e.g., opinion questions such as "How does the Chilean government view attempts at having Pinochet tried in Spanish Court?", which were the focus of a pilot study within the AQUAINT program in 2005.

This work also represents the first known use of POURPRE for system development 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 this work demonstrates its usefulness in doing just that.

Nevertheless, there are a number of limitations of our approach that should be mentioned. Most stem from the nature of the nugget-based evaluation paradigm, indirectly reflected in POURPRE. The conception of answers to complex questions as unordered sets of strings means that coherence is not a factor that comes into play when assessing the quality of system output. Other issues such as anaphora are not adequately handled by the current automatic evaluation methodology, which is exclusively based on term overlap. Although coherence and ordering are issues that have been studied within the context of multi-document summarization, they are not explicitly addressed in most current question answering systems. However, we do see more dialog between the two communities in the future, given the shift from generic to query-focused summaries in the 2005 DUC evaluation (Dang, 2005). The convergence between QA and multi-document summarization (Amigó et al., 2004) will surely spur the development of more capable systems that take into account a wider range of user needs.

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 revealed techniques that can be employed to improve performance, but more importantly, this work highlights limitations of purely IR-based methods. With a strong baseline in hand, the door is wide open for the integration of natural language understanding techniques.

8 Acknowledgments

This work has benefited from discussions with Bonnie Dorr, David Zajic, and Rich Schwartz. I would like to thank Esther and Kiri for their kind support.

References

James Allan, Courtney Wade, and Alvaro Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *Proceedings of SIGIR 2003*.

Enrique Amigó, Julio Gonzalo, Victor Peinado, Anselmo Peñas, and Felisa Verdejo. 2004. An empirical study of information synthesis task. In *Proceedings of ACL 2004*.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR 2004*.

James P. Callan. 1994. Passage-level evidence in document retrieval. In *Proceedings of SIGIR 1994*.

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. In *Proceedings of SIGIR 2005*.

Hoa Dang. 2005. Overview of DUC 2005. In *Proceedings of DUC 2005 Workshop at HLT/EMNLP 2005*.

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

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Jamie Callan. 2000. Creating and evaluating multi-document sentence extract summaries. In *Proceedings of CIKM 2000*.

Donna Harman. 2002. Overview of the TREC 2002 novelty track. In *Proceedings of TREC 2002*.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of HLT/NAACL 2004*.

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

Boris Katz, Gregory Marton, Gary Borchardt, Alexis Brownell, Sue Felshin, Daniel Loreto, Jesse Louis-Rosenberg, Ben Lu, Federico Mora, Stephan Stiller, Ozlem Uzuner, and Angela Wilcox. 2005. External knowledge sources for question answering. In *Proceedings of TREC 2005*.

Jimmy Lin and Dina Demner-Fushman. 2005a. Automatically evaluating answers to definition questions. In *Proceedings of HLT/EMNLP 2005*.

Jimmy Lin and Dina Demner-Fushman. 2005b. Will pyramids built of nuggets topple over? Technical Report LAMP-TR-127/CS-TR-4771/UMIACS-TR-2005-71, University of Maryland, College Park, December.

Donald Metzler and W. Bruce Croft. 2004. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735–750.

Hajime Mochizuki, Makoto Iwayama, and Manabu Okumura. 2000. Passage-level document retrieval using lexical chains. In *Proceedings of RIAO 2000*.

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

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of COLING 2004*.

John Prager, Jennifer Chu-Carroll, and Krzysztof Czuba. 2004. Question answering using constraint satisfaction: QA–by–Dossier–with–Constraints. In *Proceedings of ACL 2004*.

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

Stefanie Tellex, Boris Katz, Jimmy Lin, Gregory Marton, and Aaron Fernandes. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of SIGIR 2003*.

Ellen M. Voorhees and Chris Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proceedings of SIGIR 2002*.

Ellen M. Voorhees. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of TREC 2003*.

Ellen M. Voorhees. 2005. Overview of the TREC 2005 question answering track. In *Proceedings of TREC 2005*.

Jinxi Xu, Ralph Weischedel, and Ana Licuanan. 2004. Evaluation of an extraction-based approach to answering definition questions. In *Proceedings of SIGIR 2004*.