Using the Weighted Keyword Model to Improve Information Retrieval for Answering Biomedical Questions

Hong Yu, PhD,1,2,3 Yong-gang Cao, PhD1,3

1Departments of Health Sciences, 2Computer Science, and 3Medical Informatics, University of Wisconsin-Milwaukee

Abstract
Physicians ask many complex questions during the patient encounter. Information retrieval systems that can provide immediate and relevant answers to these questions can be invaluable aids to the practice of evidence-based medicine. In this study, we first automatically identify topic keywords from ad hoc clinical questions with a Condition Random Field model that is trained over thousands of manually annotated clinical questions. We then report on a linear model that assigns query weights based on their automatically identified semantic roles: topic keywords, domain specific terms, and their synonyms. Our evaluation shows that this weighted keyword model improves information retrieval from the Text Retrieval Conference Genomics track data.

1. Introduction
Clinicians and biomedical researchers often need to search a vast body of literature in order to make informed decisions [1,2]. Information retrieval and question answering systems (e.g., [3]) facilitate clinicians and biomedical researchers in accessing relevant information.

Most existing information retrieval systems require users to enter query terms, which are then used to search for relevant documents. However, observational studies (e.g., [1,4-6]) have shown that clinicians typically have complex information needs and ask complex questions. Questions 1 and 2 are two examples from a collection of 4,653 questions posed by more than 100 primary care physicians [1,4-6] that is maintained and published by the National Library of Medicine (NLM).1

Question 1: "Thirty-eight-year-old woman with bloody diarrhea, worse over the past week. I treated her with Flagyl empirically. I saw her two days later and she was lots better. No more blood, no fever. Now her report comes back and the clostridium difficile is negative but she's growing salmonella. Should I finish the Flagyl or discontinue it?"

Question 2: "The maximum dose of estradiol valerate is 20 milligrams every 2 weeks. We use 25 milligrams every month which seems to control her hot flashes. But is that adequate for osteoporosis and cardiovascular disease prevention?"

Similarly, biomedical scientists also pose complex questions that require complex answers [7,8]. Question 3 is an example of such that appeared in the TREC Genomics Track evaluation data.

Question 3: "What effect does the insulin receptor gene have on tumorigenesis?"

In this paper, we first report on applying natural language processing approaches to automatically extract topic keywords from complex biomedical questions. In the above three examples, the keywords are salmonella infections for question 1, estradiol valerate and osteoporosis and cardiovascular disease prevention for question 2, and insulin receptor gene and tumorigenesis for question 3. We then report on a weighted keyword model for query-term weight assignment.

We have implemented this model into our clinical question answering system AskHERMES. Section 2, below, reviews the background of this research. Section 3 describes the model. The evaluation methods, results and discussion are in Sections 4, 5 and 6, respectively. Section 7 briefly describes the AskHERMES system in which the weighted keyword model has been implemented. Conclusions and future work are described in Section 8.

2. Background
Although the literature has reported different models for weighing query terms for question answering (see articles in the TREC evaluation) and it is common knowledge to assign weights based on the perceived importance of a query term, methods for identifying the importance of query terms are, to our knowledge, ad hoc: most models incorporate simple algorithms (e.g., ranking query terms based on the IDF value

1 Available at [http://clinques.nlm.nih.gov/About.html](http://clinques.nlm.nih.gov/About.html)
In contrast, we weigh query terms based on automatically identified keywords and domain-specific terminology. We then developed a linear model incorporating the identified keywords to improve information retrieval.

3. Model

The weighted keyword model begins by automatically identifying semantically rich topic keywords, as shown in questions 1–3. Query term weights are based on the identified keywords, and the UMLS concepts and their synonyms. In this section, we first briefly describe our approaches for automatic keyword identification and then describe our weighted keyword model.

3.1 Automatic Topic Keyword Identification

We developed a probabilistic model to automatically identify topic keywords from ad hoc clinical questions. Our model is trained and tested on the NLM’s 4,653 clinical questions, which have been annotated by physicians who assigned one to three keywords for each clinical question. Using the annotated questions, we trained a supervised machine-learning system that is based on conditional random fields. Our ten fold cross validation results showed that the system achieved 67.6% precision, 50.8% recall, and 58% F-score for automatic keyword identification. Details of the approaches are described in Yu and Cao (2009) [10].

3.2 The Weighted Keyword Model

To judge whether a query term is biomedical, domain-specific, we applied the tool MMTx, the implementation of the MetaMap [11], to map the question to concepts in the UMLS. The UMLS incorporates concept synonyms, which are used for query expansion. We used the methods described in Section 3.1 to identify the topic keywords. Details of the approaches are described in Yu and Cao (2009) [10].

Each query term is assigned the baseline weight of the IDF value. We calculated the IDF values from more than 17 million citations in the MEDLINE collection. Our weighted keyword increases the baseline IDF value if the query term is identified as a keyword of the question. In addition, we experimented with increasing the weights of query terms based on which group they belong to. Our experiments with different weighting models concluded that most have similar impacts on information retrieval. One of the models is shown below:

- **Original Words**: the baseline IDF value
- **UMLS Synonym**: 2*IDF
- **UMLS Concept**: 3*IDF
- **Keyword Synonym**: 4*IDF
- **Keywords**: 5*IDF

4. Evaluation Methods

Currently, there is no evaluation data available for clinical information retrieval and question answering. The only available biomedical information retrieval evaluation data is the Genomics Track of the Text RETrieval Conference (TREC)². TREC Genomics incorporates more than 160,000 full-text biomedical articles [7]. The 2006 and 2007 tasks focused on information retrieval for question answering [7,12]; a sample question from the tasks is “What is the role of IDE in Alzheimer’s disease?” We therefore evaluated the weighted keyword model using the TREC Genomics evaluation.

Systems The purpose of this study is to compare different weighted keyword models for information retrieval. LUCENE is a high performance, full-featured text search engine [13] that has shown to be robust in biomedical texts [3]. We therefore implemented all our systems with LUCENE. The top 1,000 sentences of output from each system were used for evaluation. The following weighted keyword models were evaluated:

- **A. Original Words**: In this system, only the non-stop words embedded in the original question were used as query terms. There were no weighted keywords.
- **B. Reweight**: In this system, we increased the weight of keywords.
- **C. Query Expansion**: In this system, we expanded the queries with the UMLS synonyms.
- **D. Query Expansion & Reweight**: In this system, we included query terms from all five groups and weighed each group differently as described in Section 3.2.

² [http://trec.nist.gov/](http://trec.nist.gov/)
Figure 1: The mean average precision (MAP) scores of 19 TREC Genomics questions for four systems. The original words system takes in all non-stop words of an ad hoc question as bag-of-word queries to return relevant documents. Reweight is built on top of the original words system; it increases the weights of terms that are identified as keywords of the question. Query expansion incorporates synonyms from the UMLS. Expansion & reweight assigns different weights to different groups of query terms as described in Section 3.2.

Data There were 28 and 36 questions posed in TREC Genomics 2006 and 2007, respectively. However, two questions were excluded by the TREC Genomics organizers [7,8]; 19 questions returned no result for related questions. The purpose of our study is to evaluate the effectiveness of the weighted keyword model for information retrieval. We used the remaining 43 questions for our evaluation.

Evaluation Metrics We used the evaluation package published by the TREC Genomics Track (a Python script, available at http://ir.ohsu.edu/genomics/) to report the document-level retrieval performance. As stated in [8], the TREC Genomics judges returned a document as relevant if any text in that document was relevant to a question. A character-based mean average precision (MAP) measure is used by TREC Genomics to compare the accuracy of the extracted answers.

5. Evaluation Results
Table 1 shows the average MAPs of four systems for document retrieval for question answering using the TREC Genomics data. The baseline system is the original words which achieved a 0.042 MAP score. Query expansion improved the average MAP score by 28.6%. The reweight system improved the average MAP score by 9.5%. The absolute MAP improvements and their statistical significances are shown in Table 2. The improvement of reweight was statistically significant. Query Expansion and Expansion & Reweight both had larger standard deviations, which made the performance differences statistically non-significant. Figure 1 shows the MAP scores of a subset of TREC Genomics questions for the four systems. The MAP score differences by four systems; we only report in Figure 2 those systems with the MAP scores >0.03. As shown in Figure 2, the MAP scores ranged from close to zero to close to 0.7 in response to different questions. The variations in the MAP scores lead to the large value of standard deviation as shown in Table 1.

Table 1: Average MAP scores (standard deviations in parentheses) of four systems for document retrieval for question answering using the TREC Genomics data.

| System                  | Original Words | Query Expansion | Reweight | Expansion & Reweight |
|-------------------------|----------------|-----------------|----------|----------------------|
| Average MAP (St. Dev)   | .042 (.085)    | .054 (.117)     | .046 (.092) | .053 (.116)         |

Table 2: Improvement in MAP scores of three systems (query expansion, reweight, and expansion & reweight) over the original words system.

| Improvement            | Query Expansion | Reweight | Expansion & Reweight |
|------------------------|-----------------|----------|----------------------|
| Average MAP (St. Dev)  | .012 (.051)     | .004 (.099) | .011 (.054)         |
| p-value                | .119            | .005     | .183                |

6. Discussion
Our work shows that, for most of the questions, a reweight system significantly outperforms a non-reweight system (p<0.005). We have tried different reweight combinations and found that in all cases, increasing the weights of keywords has significant improvements (data not shown). Our results clearly demonstrate the effectiveness of weighted keywords for improving information retrieval. We do not compare our absolute MAP scores with those who participated in the TREC Genomics competition, as the absolute MAP scores depend upon many other factors, including data preprocess and passage ranking.
Our results show that although query expansion has improved the MAP scores for most of the questions, these improvements were not statistically significant. Our results are consistent with the reports in TREC Genomics. Query expansion was widely used in both the 2006 and 2007 TREC Genomics competitions [7,8]. Few teams have reported that query expansion statistically improves information retrieval. Teams report that the performance of query expansion varies for different topics (e.g., [14]). Reasons for this include failure in identifying synonyms [15], which depends upon the correct mapping to external knowledge resources. The variations in performance in query expansion explain our results, in which the improvement in weighted keywords diminished after query expansion.

Figure 2: AskHERMES system components

Our topic keyword model was trained over thousands of clinical questions, and it is interesting that the model can be used directly to capture the keywords in genomics questions and to improve the information retrieval in the genomic domain. The results demonstrate the generalizability of both our keyword identification model and the weighted keyword model. On the other hand, the question of whether the weighted keyword model can actually improve information retrieval and question answering in the clinical domain still needs to be tested.

7. Implementing the Weighted Keyword Model in the AskHERMES System

Our long-term goal is to develop an advanced medical question answering system to assist physicians in their clinical decision making. We have created such a prototype system called AskHERMES (Help physicians to Extract and aRticulate Multimedia information for answering clinical quEstionS), which can be accessed at http://www.askhermes.org. Figure 2 shows the AskHERMES system components. We have previously shown AskHERMES to outperform several other baseline information retrieval systems for answering definitional questions [3,16]. Currently, AskHERMES attempts to answer all types of clinical questions.

In this study, we have integrated the weighted keyword model into the AskHERMES system, and our preliminary observation shows that the model slightly increases AskHERMES’ performance for question answering. Figure 3 shows the answers of two models (with and without weighted keywords) to a sample clinical question. A physician (Dr. Andrew Bennett) examined the outputs of both models. He concluded that none of the text outputs directly answered the questions, although the answers can be identified from the source articles. He also concluded that the weighted output is more on target than the unweighted one in both text outputs and source answers. The evaluation seems to support that the weighted model outperforms the unweighted one. On the other hand, a formal evaluation is required to draw any general conclusions.

Figure 3: The outputs of two models, with and without weighted keywords in response to a sample clinical question. The keyword “head trauma” was automatically identified by AskHERMES. Each answer can be linked to its source page. “Human” indicates that the source page is a human study.
8. Conclusions and Future Work

Our contributions include a robust keyword identification system that is trained on thousands of ad hoc clinical questions and a linear model for incorporating the identified keywords as a way to improve information retrieval. Our evaluation results with the TREC Genomics data show an improvement in information retrieval with the weighted keyword model. We also demonstrate that the weighted keyword model can be easily integrated into a clinical question answering system.

The evaluation of the effectiveness of the weighted keyword model for improving clinical question answering remains as our future work. The key is to create evaluation data, which is an important but long-term challenging task. In addition, we hope to explore our weighted keyword models in open-domain information retrieval and question answering.

Acknowledgement: The authors acknowledge the support 1R01LM009836-01A1 to Hong Yu. The authors also acknowledge the National Library of Medline for making the 4,654 clinical questions and their annotations freely available. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the NIH’s views.

We thank Dr. Andrew Bennett for evaluating the AskHERMES system. Yu supervised the project and wrote the paper; Cao performed the experiments and built AskHERMES; Bennett provided the judgment of the AskHERMES’ outputs.

References

1. Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, Evans ER: Analysis of questions asked by family doctors regarding patient care. BMJ 1999;319:358-361.
2. Yu H, Lee M: Accessing bioscience images from abstract sentences. Bioinformatics 2006;22:e547-556.
3. Yu H, Lee M, Kaufman D, Ely J, Osheroff JA, Hripcsak G, Cimino J: Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. J Biomed Inform 2007;40:236-251.
4. Ely JW, Osheroff JA, Ferguson KJ, Chambliss ML, Vinson DC, Moore JL: Lifelong self-directed learning using a computer database of clinical questions. J Fam Pract 1997;45:382-388.
5. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME: Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc 2005;12:217-224.
6. D'Alessandro DM, Kreiter CD, Peterson MW: An evaluation of information-seeking behaviors of general pediatricians. Pediatrics 2004;113:64-69.
7. Hersh W, Cohen A, Roberts P, Rekapalli H: TREC 2006 Genomics Track overview. In TREC Genomics Track conference. 2006.
8. Hersh W, Cohen A, Ruslen L, Roberts P: TREC 2007 Genomics Track overview. In The TREC Genomics Track Conference. 2007.
9. Pradhan S, Illouz G, Blair-Goldensohn S, Schlaikjer A, Krugler V, Filatova E, Duboue P, Yu H, Passonneau R, Ward W, Hatzivassiloglou V, Jurafsky D, McKeown K, Martin J: Building a foundation system for producing short answers to factual questions. In Eleventh Text Retrieval Conference (TREC-11). Washington, DC, 2002.
10. Yu H, Cao YG: Automatically extracting information needs from ad hoc clinical questions. AMIA Annu Symp Proc 2008:96-100.
11. Aronson AR: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21.
12. Hersh W, Cohen A, Yang J, Bhupatiraju R, Roberts P, Hearst M: TREC 2005 Genomics Track overview. In TREC Genomics Track conference. 2005.
13. Lucene A: A high-performance, full-featured text search engine library. Available at http://lucene.apache.org/java/docs/ In, 2005.
14. Jimeno A, Pezik P: Information retrieval and information extraction in TREC Genomics 2007. In The TREC Genomics Conference. 2007.
15. Cohen A, Yang J, Fisher S, Roark B, Hersh W: The OHSU biomedical question answering system framework. In The TREC Genomics Conference. 2007.
16. Yu H, Kaufman K: A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. In Pacific Symposium on Biocomputing. 2007:328-339.