

Evaluation of the Clinical Question Answering Presentation

Yong-Gang Cao College of Health Sciences University of Wisconsin Milwaukee Milwaukee, WI 53211,USA yonggang@uwm.edu	John Ely Carver College of Medicine University of Iowa Iowa, IA 52242,USA john- ely@uiowa.edu	Lamont Antieau College of Health Sciences University of Wisconsin Milwaukee Milwaukee, WI 53211,USA antieau@uwm.edu	Hong Yu College of Health Sciences University of Wisconsin Milwaukee Milwaukee, WI 53211,USA hongyu@uwm.edu
--	--	--	--

Abstract

Question answering is different from information retrieval in that it attempts to answer questions by providing summaries from numerous retrieved documents rather than by simply providing a list of documents that requires users to do additional work. However, the quality of answers that question answering provides has not been investigated extensively, and the practical approach to presenting question answers still needs more study. In addition to factoid answering using phrases or entities, most question answering systems use a sentence-based approach for generating answers. However, many sentences are often only meaningful or understandable in their context, and a passage-based presentation can often provide richer, more coherent context. However, passage-based presentations may introduce additional noise that places greater burden on users. In this study, we performed a quantitative evaluation on the two kinds of presentation produced by our online clinical question answering system, AskHERMES (<http://www.AskHERMES.org>). The overall finding is that, although irrelevant context can hurt the quality of an answer, the passage-based approach is generally more effective in that it provides richer context and matching across sentences.

1 Introduction

Question answering is different from information retrieval in that it attempts to answer questions by providing summaries from numerous retrieved documents rather than by simply providing a list of documents for preparing the user to do even more exploration. The presentation of answers to questions is a key factor in its efficiently meeting the information needs of information users.

While different systems have adopted a variety of approaches for presenting the results of question answering, the efficacy of the use of these different approaches in extracting, summarizing, and presenting results from the biomedical literature has not been adequately investigated. In this paper, we compare the sentence-based approach and the passage-based approach by using our own system, AskHERMES, which is designed to retrieve passages of text from the biomedical literature in response to ad hoc clinical questions.

2 Background

2.1 Clinical Question Collection

The National Library of Medicine (NLM) has published a collection of 4,653 questions that can be freely downloaded from the Clinical Questions Collection website¹ and includes the questions below:

¹ <http://clinques.nlm.nih.gov/JitSearch.html>

Question 1: “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?”

Question 2: “Child has pectus carinatum. Radiologist told Dr. X sometimes there are associated congenital heart problems. Dr. X wants to study up on this. Does the patient have these associated problems?”

Such examples show that clinicians pose complex questions of a far greater sophistication than the simple term searches that typical information retrieval systems require as input. AskHERMES, however, has been designed to handle such complexity as it encounters it.

2.2 Result Presentation

In recent years, there has been an emergence of numerous search engines – both open domain and domain-specific – as well as question answering systems, and these systems have employed a variety of methods for presenting their results, including the use of metadata, sentences, snippets, and passages. PubMed (Anon 2009a) and EAGLi (Anon 2009b), for example, use article metadata to present their results, and the combination of title, author name and publication name that they use works like the citation at the end of a paper to provide users with a general idea of what the listed article is about. On the other hand, AnswerBus (Anon 2009c) and AnswerEngine (Anon 2009d) extract sentences from relevant articles, then rank and list them one by one to answer the questions that users have. In response to a query, Google and other general search engines provide the title of a work plus a snippet of text to provide metadata as well as multiple matching hints from articles. In response to user questions, Start (Anon 2009e), Powerset (Anon 2009f) and Ask (Anon 2009g) provide a single passage as output, making them ideal for answering simple questions because they do not require users to access and read extra articles in order to answer the questions they have.

Each of these methods of presentation has strengths and weaknesses. First, a strength of using metadata is that it provides a way for discovering the general idea of an article, but it does not explain to a user why the article is relevant to the query or question, making it difficult to decide whether it is worth the time and effort to access the listed article to read more. An ap-

proach presenting a single sentence in response to a query can result in a good answer if the user is lucky but typically provides a limited idea of what the target article contains and demands that users access the source of the item to learn more. A snippet-based approach can provide a hint as to why the target article is relevant, but snippets are limited in that they are composed of segments and usually cannot be read at all; even presenting a snippet with metadata as Google does is not suitable for adequately answering many questions.

We propose a passage-based approach in which each passage is constructed by coherent sentences. The approach we propose is similar to that used by Start and Ask, but these systems have limited knowledge bases and require queries to be written using very specific question types. On the other hand, our system will be able to answer ad hoc questions (that is, questions not limited to specific types). Furthermore, the system we propose will be oriented toward answering questions in the biomedical community, a field in which automated question answering and information retrieval and extraction are in strong demand.

3 Passage-Based Approach versus Sentence-Based Approach

We define as sentence-based approaches those approaches that return a list of independently retrieved and ranked sentences. Although all the sentences are assumed to be relevant to the question, there are no assumptions of their relationship with each other. On the other hand, a passage-based approach is defined as one that returns a list of independently retrieved and ranked passages, each of which can comprise multiple tightly coupled sentences.

The passage-based approach has two benefits:

1. It provides richer context for reading and understanding.
2. It provides greater evidence for relevant ranking of the passage by matching across sentences.

For example, in Figure 1, the passage-based output of the top results of AskHERMES pertains to the question “What is the difference between the Denver ii and the regular Denver developmental screening test?” The first answer is a passage with two sentences; the first sentence in the passage informs users that there have been

criticisms of the “Denver Developmental Screening Test,” and the second sentence shows that “Denver II” addressed several concerns of the “Denver Developmental Screening Test.” The two sentences indicate that the article will mention several issues that answer the question. And the second passage directly shows the answer to the question: The criteria to select Denver II and the difference between the two tests.

If we use the sentence-based approach (see Figure 2), the sentences in the first passage will be ranked very low and might not appear in the results because both of them contain only one of

the screening tests mentioned in the question. The second passage will be reduced to only the second sentence, which is an incomplete answer to the question; consequently, the user may remain uninformed of the selection criteria between the two screening tests without further examination of the article. Figure 2 shows the sentence-based output of the same question. A comparison of the examples in the figure clearly shows how the results of the query are affected by the two approaches. The first result is incomplete, and the second and third results are irrelevant to the question although they have many matched terms.

You asked:what is the difference between the denver ii and the regular denver developmental screening test

Clustered Answers Ranked Answers Content Clustered Answers

[denver developmental screening test, denver ii, test]

- Despite widespread usage of the original **Denver Developmental Screening Test**, there have been criticisms of the tool. The updated **Denver II** tool addresses several concerns associated with the original test..[Wade;1992]{Human}

[between, denver developmental screening test, denver ii]

- The final selection of the 125 **Denver II** items was based on the following criteria: ease of administration and scoring, item appeal to child and examiner, item test-retest and inter-rater reliability, minimal "refusal" scores, minimal "no opportunity" scores, minimal subgroup differences, and a smooth step-like progression of ages at which 90% of children could perform the tasks. The major differences **between the Denver II and the Denver Developmental Screening Test** are: 1) an 86% increase in language items; 2) two articulation items; 3) a new age scale; 4) a new category of item interpretation to identify milder delays; 6) a behavior rating scale; and 7) new training materials..[Frankenburg;1992]{Human}

Figure 1. AskHERMES’ passage-based output for the question “What is the difference between the Denver ii and the regular Denver developmental screening test?”

You asked:what is the difference between the denver ii and the regular denver developmental screening test

Clustered Answers Ranked Answers Content Clustered Answers

[between, denver developmental screening test, denver ii]

- The major differences **between the Denver II and the Denver Developmental Screening Test** are: 1) an 86% increase in language items; 2) two articulation items; 3) a new age scale; 4) a new category of item interpretation to identify milder delays; 6) a behavior rating scale; and 7) new training materials.. [Frankenburg;1992]{Human}

[denver ii developmental screening test, developmental]

- The **Denver II Developmental Screening Test** was used to assess **developmental** skills.. [Kinzler;2001]{Human}

[denver developmental screening test, ii, test]

- This was achieved by a prospective evaluation, from November/1995 to April/2000, of 43 HIV infected children (group I) and 40 HIV seroreverters children (group II) through neurological exam and neurodevelopmental tests: **Denver Developmental Screening Test (DDST)** and **Clinical Adaptive Test/Clinical Linguistic and Auditory Milestone Scale (CAT/CLAMS)**..[Bruck;2001]{Human}

Figure 2. AskHERMES’ sentence-based output for the question “What is the difference between the Denver ii and the regular Denver developmental screening test?”

While the results shown in Figures 1 and 2 suggest that a passage-based approach might be better than a sentence-based approach for question answering, this is not to say that passage-based approaches are infallible. Most importantly, a passage-based approach can introduce noisy sentences that place an additional burden on users as they search for the most informative answers to their questions. In Figure 3, the first sentence in the output of sen-

tence-based approach answers the question. However, the passage-based approach does not answer the question until the fourth passage, and when it does, it outputs the same core answer sentence that was provided in the sentence-based approach. Additionally, the core sentence is nested within a group of sentences that on their own are only marginally relevant to the query and in effect bury the answer.

Question: which infants are candidates for synagis (respiratory-syncytial-virus immune globulin) and when should it start?	
Sentence-based Approach	
1.	Palivizumab (Synagis) and respiratory syncytial virus intravenous immune globulin are licensed by the US Food and Drug Administration for the prevention of severe lower respiratory tract infections caused by respiratory syncytial virus in infants with bronchopulmonary dysplasia, infants with a history of premature birth (< or =35 weeks gestational age) and children with hemodynamically significant congenital heart disease . [Silvia,2005](Human)
2.	To determine outcome and length of stay (LOS) for infants younger than 2 years of age admitted to hospital for respiratory syncytial virus (RSV) infection and to perform a cost-benefit analysis of prophylaxis with RSV gamma-globulin (Respigam; CSL Laboratories , Melbourne, Victoria, Australia) and monoclonal antibody (Synagis ; Abbott Australasia, Kurnell, NSW, Australia) in the Australian context. [Numa,2001](Human)
Passage-based Approach	
1.	Human metapneumovirus (hMPV) is a newly recognized pathogen that like its better-known relative, human respiratory syncytial virus (hRSV), appears to be ubiquitous and an important cause of respiratory disease in diverse subpopulations. No antivirals or vaccines are currently approved for the treatment or prevention of hMPV infections. However, ribavirin is licensed to treat serious hRSV-induced infections in children and immune globulin designed for intravenous administration (i.v.IG) and palivizumab (Synagis), a humanized monoclonal antibody preparation, have been utilized as alternatives to vaccines for preventing or reducing the severity of infections caused by this virus . Because both ribavirin and i.v.IG have broad viral specificities, studies were performed to compare the ability of these two agents to inhibit the replication of hRSV and hMPV in tissue culture-based assays. [Philip,2003](Human)
2.	The Food and Drug Administration recently approved the use of palivizumab (pal-vizh-mab), an intramuscularly administered monoclonal antibody preparation. Recommendations for its use are based on a large , randomized study demonstrating a 55% reduction in the risk of hospitalization attributable to respiratory syncytial virus (RSV) infections in high-risk pediatric patients . Infants and children with chronic lung disease (CLD), formerly designated bronchopulmonary dysplasia, as well as prematurely born infants without CLD experienced a reduced number of hospitalizations while receiving palivizumab compared with a placebo. Both palivizumab and respiratory syncytial virus immune globulin intravenous (RSV-IGIV) are available for protecting high-risk children against serious complications from RSV infections. Palivizumab is preferred for most high-risk children because of ease of administration (intramuscular), lack of interference with measles-mumps-rubella vaccine and varicella vaccine , and lack of complications associated with intravenous administration of human immune globulin products. RSV-IGIV, however, provides additional protection against other respiratory viral illnesses and may be preferred for selected high-risk children including those receiving replacement intravenous immune globulin because of underlying immune deficiency or human immunodeficiency virus infection. For premature infants about to be discharged from hospitals during the RSV season, physicians could consider administering RSV-IGIV for the first month of prophylaxis. Most of the guidelines from the American Academy of Pediatrics for the selection of infants and children to receive RSV-prophylaxis remain unchanged. Palivizumab has been shown to provide benefit for infants who were 32 to 35 weeks of gestation at birth. RSV-IGIV is contraindicated and palivizumab is not recommended for children with cyanotic congenital heart disease . The number of patients with adverse events judged to be related to palivizumab was similar to that of the placebo group (11% vs 10%, respectively); discontinuation of injections for adverse events related to palivizumab was rare. [null,1998](Human)
3.	Respiratory syncytial virus (RSV) continues as an emerging infectious disease not only among infants and children , but also for the immune-suppressed, hospitalised and the elderly . To date, ribavirin (Virazole, ICN Pharmaceuticals, Inc.) remains the only therapeutic agent approved for the treatment of RSV. However, its clinical benefits are small and occur only in a fraction of RSV-infected patients . The prophylactic administration of palivizumab (Synagis , MedImmune, Inc.) is problematic and costly and, therefore, only recommended for use in high-risk infants . Clearly, the need for an effective and safe drug remains high . This review discusses several different antisense approaches and compares them with traditional strategies, such as RSV-targeting antibodies and antivirals, as well as developments in vaccine research. [Hagen,2005](Human)
4.	Therefore, protection of high-risk infants is possible only by passive prophylaxis with specific antibodies. Palivizumab (Synagis) and respiratory syncytial virus intravenous immune globulin are licensed by the US Food and Drug Administration for the prevention of severe lower respiratory tract infections caused by respiratory syncytial virus in infants with bronchopulmonary dysplasia, infants with a history of premature birth (< or =35 weeks gestational age) and children with hemodynamically significant congenital heart disease . Palivizumab is a humanized monoclonal antibody produced by recombinant DNA technology, directed to an epitope in the A antigenic side of the F-protein of the respiratory syncytial virus . [Silvia,2005](Human)

Figure 3. An example comparing the sentence-based approach and passage-based approach

4 Evaluation Design

To evaluate whether the passage-based presentation improves question answering, we plugged two different approaches into our real system by making use of either the passage-based or the sentence-based ranking and presentation unit constructor. Both of them share the same document retrieval component, and they share the same ranking and clustering strategies. In our system, we used a density-based passage retrieval strategy (Tellex et al.

2003) and a sequence sensitive ranking strategy similar to ROUGE (F. Liu and Y. Liu 2008). An in-house query-oriented clustering algorithm was used to construct the order and structure of the final hierarchical presentation. The difference between the two approaches is the unit for ranking and presentation. A passage-based approach takes the passage as its primary unit, with each passage consisting of one or more sentences. Those sentences in the passage are extracted from the adjacent matching sentences in the original article.

To evaluate the difference between the passage-based presentation and sentence-based presentation, we randomly selected 20 questions from 4,653 clinical questions. A physician (Dr. John Ely) was shown the corresponding passage-based and sentence-based outputs of every question and was then asked to judge the relevance of the output and which output had the higher quality answer. Because physicians have little time in clinical settings to be sifting through data, we presented only the top five units (sentences or passages) of output for every question.

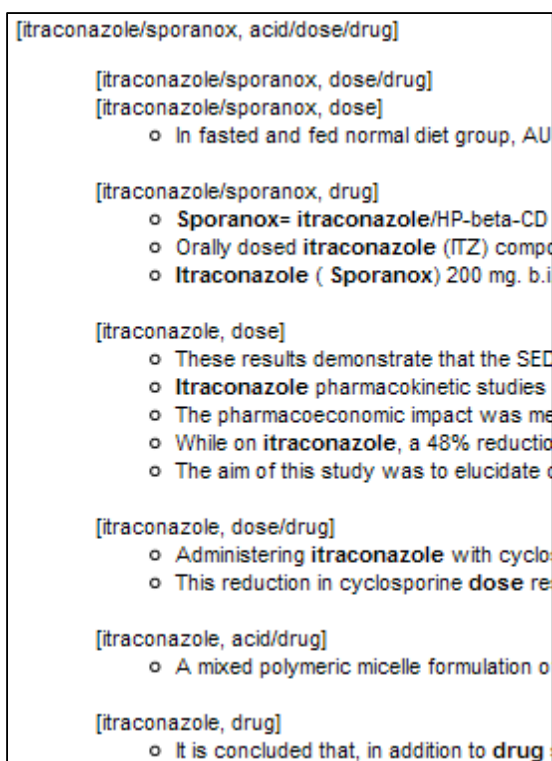


Figure 4. A partial screenshot of AskHERMES illustrating hierarchical clustering based on the question “What is the dose of sporanox?”

For answer extraction, we built a hierarchical weighted-keyword grouping model (Yu and Cao 2008; Yu and Cao 2009). More specifically, in using this model we group units based on the presence of expanded query-term categories: keywords, keyword synonyms, UMLS concepts, UMLS synonyms, and original words, and we then prioritize the groups based on their ranking. For example, units that incorporate keywords are

grouped into the first cluster, followed by the cluster of units that incorporate keyword synonyms, UMLS concepts, etc. The units that appear synonymous are in the clusters with the same parent cluster. Figure 4 shows an example of the top branch of the clusters for the question “What is the dose of sporanox?” in which the answers are organized by *sporanox* and *dose* as well as their synonyms.

5 Evaluation Result and Discussion

We classify physician evaluations as being of the following four types and plot their distribution in Figure 5:

- **Hard Question:** The question is considered difficult because it is patient-specific or unclear (that is, it is a poorly formed question), e.g., “Multiple small ulcers on ankles and buttocks. No history of bites. I sent him for a complete blood count (cbc) and blood sugar but I don’t know what these are.”
- **Failed Question:** Neither approach can find any relevant information for the question.
- **Passage Better:** Passage-based approach presents more useful information for answering the question.
- **Sentence Better:** Sentence-based approach provides the same amount of useful information while reducing the effort required by the passage-based approach.

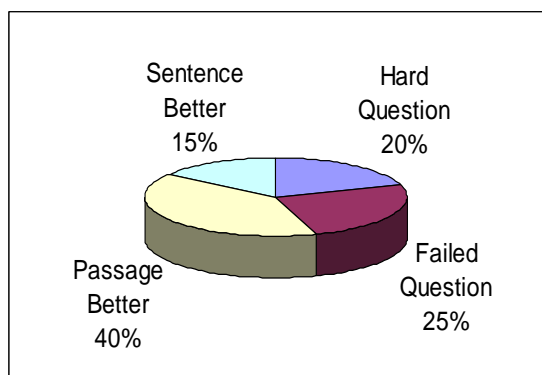


Figure 5. Distribution of the defined Evaluation categories

The evaluation data is shown in Table 1. In our study, the score range is set from 0 to 5 with the value 0 referring to answers that are totally irrelevant to the question and the value 5 meaning there is enough information to fully answer the question. Our results show that the passage-based approach is better than the sentence-based approach (p-value < 0.05).

Table 1. Quantitative measurement of the answers generated by both approaches to the 20 questions

No.	Passage-based approach score	Sentence-based approach score
1	3	1
2	2	0
3	2	0
4	0	0
5	0	0
6	1	0
7	3	1
8	3	0
9	0	0
10	0	0
11	1	2
12	1	2
13	3	4
14	0	0
15	1	0
16	2	1
17	0	0
18	1	0
19	0	0
20	0	0
mean	1.15	0.55
s.deviation	1.18	1.05
p-value	0.01	

Through further analysis of the results, we found that 70% of the sentences yielded by the sentence-based approach did not answer the question at all (the score is zero), while this was true for only 40% of the output of the passage-based approach. This indicates that the passage-based approach provides more evidence for answering questions by providing richer context and matching across sentences.

On the other hand, if the question was too general and included a plethora of detail and little focus, both approaches failed. For example, in the ques-

tion "One year and 10-month-old boy removed from his home because of parental neglect. Care-taker says he often cries like he's in pain, possibly abdominal pain. Not eating, just drinking liquids, not sleeping. The big question with him: "is it something physical or all adjustment disorder?" there is a great deal of description of the boy, and a variety of common symptoms are also provided. AskHERMES found a passage containing all of the following extracted words: "availability, because, before, between, changes, children, decrease, disorder/disorders, drug, eating, going, increase, indications/reasons, intake, laboratory, level, may, often, one, patient/patients, physical, recommended, routinely, specific, still, symptom/symptoms, two, urine, used, women, treat/treated/treating/therapy/treatment/treatments, and work." But since these words are so commonly used in a variety of scenarios, the output passage is off-topic.

For very simple questions, the sentence-based approach works well for providing answers in a very concise form. For example, the question "what is the dose of zyrtec for a 3-year-old?" can be answered by the dosage amount for the target age group, and the query resulted in this answer: "...children of both sexes aged between 2 to 6 years with allergy rhinitis (AR) were included in this study, who were randomly selected to be treated with Zyrtec (Cetirizine 2 HCL) drops 5 mg daily for 3 weeks." From a literal view, this looks like an answer to the question because it discusses the dosage of Zyrtec for the specific age group; however, it actually describes an experiment and does not necessarily provide the suggested dosage that the user is seeking. This leads to an interesting problem for clinical question answering: how should experimental data be distinguished from suggestion data for recommended daily usage? People tend to ask for the best answer instead of the possible answers. This is one of the main reasons why in Table 1, there is no perfect score (5).

Our result looks similar to the conclusion of Lin et al (Jimmy Lin et al. 2003), whose study on open-domain factoid question answering indicates a preference among users for the answer-in-paragraph approach rather than the three other types of presentation: exact-answer (that is, answer entity), answer-in-sentence, and answer-in-

document. The results of both Lin's research and our own indicate the usefulness of context, but Lin's work focuses on how surrounding context helps users to understand and become confident in answers retrieved by simple open-domain queries, while our research reveals that adjacent sentences can improve the quality of answers retrieved using complex clinical questions. Our results also indicate that context is important for relevance ranking, which has not been thoroughly investigated in previous research. Furthermore, our work places emphasis on proper passage extraction from the document or paragraph because irrelevant context can also be a burden to users, especially for physicians who have limited time for reading through irrelevant text. Our continuous sentence-based passage extraction method works well for our study, but other approaches should be investigated to improve the passage-based approach.

With respect to the quality of the answer, the content of the output is not the only important issue. Rather, the question itself and the organization of content are also important issues to consider. Luo and Tang (Luo and Tang 2008) proposed an iterative user interface to capture the information needs of users to form structured queries with the assistance of a knowledge base, and this kind of approach guides users toward a clearer and more formal representation of their questions. DynaCat (Pratt and Fagan 2000) also uses a knowledge-based approach to organize search results. Thus, applying domain-specific knowledge is promising for improving the quality of an answer, but the difficulty of the knowledge-based approach is that building and updating such knowledge bases is human labor intensive, and furthermore, a knowledge-based approach restricts the usage of the system.

6 Conclusion and Future Work

In this study, we performed a quantitative evaluation on the two kinds of presentation produced by our online clinical question answering system, AskHERMES. Although there is some indication that sentence-based passages are more effective for some question types, the overall finding is that by providing richer context and matching across sentences, the passage-based approach is generally a

more effective approach for answering questions. Compared to Lin's study on open-domain factoid questions (Jimmy Lin et al. 2003), our study addresses the usefulness of context for answering complex clinical questions and its ability to improve answer quality instead of just adding surrounding context to the specific answer.

While conducting this investigation, we noticed that simple continuous sentence-based passage constructions have limitations in that they have no semantic boundary and will form too long a passage if the question contains many common words. Therefore, we will take advantage of recent advances we have made in HTML page analysis components to split documents into paragraphs and use the paragraph as the maximum passage, that is, a passage will only group sentences that appear in the same paragraph. Furthermore, by setting the boundary at a single paragraph, we can loosen the adjacency criterion of our current approach, which requires that the sentences in a passage be next to each other in the original source, and instead adopt a requirement that they only be in the same paragraph. This will enable us to build a model consisting of one or more core sentences as well as several satellite sentences that could be used to make the answer more complete or understandable.

Acknowledgments

The authors acknowledge support from the National Library of Medicine to Hong Yu, grant number 1R01LM009836-01A1. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the NIH.

References

- Anon. 2009a. PubMed Home. <http://www.ncbi.nlm.nih.gov/pubmed/> (Accessed: 10. March 2009).
- Anon. 2009b. EAGLi: the EAGL project's biomedical question answering and information retrieval interface. <http://eagl.unige.ch/EAGLi/> (Accessed: 6. March 2009).
- Anon. 2009c. AnswerBus Question Answering System. <http://www.answerbus.com/index.shtml> (Accessed: 6. March 2009).
- Anon. 2009d. Question Answering Engine. <http://www.answers.com/bb/> (Accessed: 6. March 2009).

- Anon. 2009e. The START Natural Language Question Answering System. <http://start.csail.mit.edu/> (Accessed: 6. March 2009).
- Anon. 2009f. Powerset. <http://www.powerset.com/> (Accessed: 19. April 2009).
- Anon. 2009g. Ask.com Search Engine - Better Web Search. <http://www.ask.com/> (Accessed: 6. March 2009).
- Lin, Jimmy, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh Boris, Boris Katz and David R Karger. 2003. What Makes a Good Answer? The Role of Context in Question Answering Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi. *PROCEEDINGS OF INTERACT 2003*: 25--32. doi:10.1.1.4.7644, .
- Liu, F. and Y. Liu. 2008. Correlation between rouge and human evaluation of extractive meeting summaries. In: *The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008)*.
- Luo, Gang and Chunqiang Tang. 2008. On iterative intelligent medical search. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 3-10. Singapore, Singapore: ACM. doi:10.1145/1390334.1390338, <http://portal.acm.org/citation.cfm?id=1390338> (Accessed: 13. March 2009).
- Pratt, Wanda and Lawrence Fagan. 2000. The Usefulness of Dynamically Categorizing Search Results. *Journal of the American Medical Informatics Association* 7, Nr. 6 (December): 605-617.
- Tellex, S., B. Katz, J. Lin, A. Fernandes and G. Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 41-47. ACM New York, NY, USA.
- Yu, Hong and Yong-Gang Cao. 2008. Automatically extracting information needs from ad hoc clinical questions. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*: 96-100.
- Yu, Hong and Yong-Gang Cao. 2009. Using the weighted keyword models to improve information retrieval for answering biomedical questions. In: *To appear in AMIA Summit on Translational Bioinformatics*.