Published in final edited form as:
J Biomed Inform. 2010 December ; 43(6): 962–971. doi:10.1016/j.jbi.2010.07.007.

Automatically Extracting Information Needs from Complex Clinical Questions

Yong-gang Cao, PhD[1], **James J Cimino, MD**[2], **John Ely, MD**[3], and **Hong Yu, PhD**[1]
[1]University of Wisconsin-Milwaukee

[2]National Institutes of Health

[3]University of Iowa

Abstract

Objective—Clinicians pose complex clinical questions when seeing patients, and identifying the answers to those questions in a timely manner helps improve the quality of patient care. We report here on two natural language processing models, namely, automatic topic assignment and keyword identification, that together automatically and effectively extract information needs from ad hoc clinical questions. Our study is motivated in the context of developing the larger clinical question answering system AskHERMES (Help clinicians to Extract and aRrticulate Multimedia information for answering clinical quEstionS).

Design and Measurements—We developed supervised machine-learning systems to automatically assign predefined general categories (e.g., *etiology*, *procedure*, and *diagnosis*) to a question. We also explored both supervised and unsupervised systems to automatically identify keywords that capture the main content of the question.

Results—We evaluated our systems on 4,654 annotated clinical questions that were collected in practice. We achieved an F1 score of 76.0% for the task of general topic classification and 58.0% for keyword extraction. Our systems have been implemented into the larger question answering system AskHERMES. Our error analyses suggested that inconsistent annotation in our training data have hurt both question analysis tasks.

Conclusion—Our systems, available at http://www.askhermes.org, can automatically extract information needs from both short (the number of word tokens <20) and long questions (the number of word tokens >20), and from both well-structured and ill-formed questions. We speculate that the performance of general topic classification and keyword extraction can be further improved if consistently annotated data are made available.

Keywords

natural language processing; question answering; question analysis; keyword extraction

Contact Information: Hong Yu, 2240 Hartford Avenue, Enderis Hall 939, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, Phone: 414-229-3344, Fax: 414-810-0065, hongyu@uwm.edu.

1. Introduction

Clinicians have many questions when seeing patients [1], with reports of up to six questions for every patient encounter [2-7]. Identifying answers to such questions will support the practice of evidence-based medicine [8],[9] and, as a result, will improve the quality of patient care [10-12].

Although clinicians have information needs when seeing patients, studies have concluded that clinicians' information needs were often unmet (e.g., [1],[8]) and that a lack of time was the most common obstacle preventing them from meeting these needs [13],[14]. Many clinical databases (e.g., UpToDate) provide high quality summaries for answering important medical questions related to patient care. The summaries, however, are written by domain experts who manually review the literature related to specific medical topics. As such, the database may be limited by its scope and timeliness. In addition, accessing clinical databases still requires a significant amount of time. For example, one evaluation study [15] has shown that it takes over 4 minutes to search for answers to clinical questions in UpToDate.

Many clinicians use the Internet (e.g., Google) to search for answers to their questions [16-19]. However, the Internet poses challenges in information content relatedness and quality [20-29]. A survey study [30] found that 92% of physicians preferred a target site rather than a search engine (e.g., Google). A recent study published in 2009 [31] concluded that PubMed appeared to be better than Google Scholar for locating relevant primary literature articles to answer specific drug-related questions. Another recent study [17] reported that for military physicians, open-domain sites (PubMed) are more commonly used by military surgeons.

Therefore, searching answers from the published biomedical literature is important for clinicians. Although scientifically strong, clinically relevant, original research articles occur in the highest concentrations in only a few select journals (e.g., *The New England Journal of Medicine*, *Annals of Internal Medicine, JAMA*, and *Archives of Internal Medicine*), much clinical evidence appears in a wide range of other biomedical journals [32]. Without a search engine, it is unlikely for a clinician to keep up with the most recent clinical evidence as reported in an ever-increasing number of volumes in the medical literature. Literature search engines, including PubMed and Google Scholar, do not return answers in response to specific questions; rather, such search engines frequently return a large number of articles in response to a specific user query. Clinicians, however, have limited time for browsing the articles retrieved, and a study observed that clinicians spent on average two minutes or less seeking an answer to a question, and that if a search took longer, it was likely to be abandoned [1],[33].

To address the aforementioned obstacles, we are building a fully automated system AskHERMES — Help clinicians Extract and aRticulate Multimedia information from literature to answer their ad-hoc clinical quEstionS [34-45]. Although it is still at the preliminary stage, AskHERMES is currently the only online system that automatically retrieves, extracts, and integrates information from the literature and other information resources and attempts to formulate this information as answers in response to ad hoc medical questions posed by clinicians, all of which can be achieved within a time-frame that meets their demands.

Fletcher (30) identified three basic skills necessary for clinicians to manage their information needs: (1) find potentially relevant information, (2) judge the best from the much larger volume of less credible information, and (3) judge whether the best information retrieved provides sufficient evidence for clinical decisions. AskHERMES attempts to complete the first two tasks by *finding* and *filtering* clinical information. We have previously

found that AskHERMES advances several other baseline information retrieval systems (e.g., PubMed) for answering definitional questions [39],[46]. Currently, AskHERMES attempts to answer all types of clinical questions, with a preliminary capacity.

One of the a key difference between AskHERMES and other clinical question-answering related work (e.g., [47-50]) is its computational approaches for automatically extracting information needs from the questions, and that is the focus of this study.

2 Background

Question answering can be considered an advanced form of information retrieval. A variety of approaches have addressed question answering in the biomedical domain. Zweigenbaum [51],[52] surveyed the feasibility of question answering in the biomedical domain. Rinaldi and colleagues [53] adapted an open-domain question answering system to answer genomic questions (e.g., “where was spontaneous apoptosis observed?”). The EpoCare project (Evidence at Point of Care) proposed a framework that aimed to provide physicians with the best available medical information from both literature and clinical databases [47]. Infobuttons [48],[54-59] [62],[48],[56],[57],[6l],[60],[58],[59],[63]served as a medical portal to external information retrieval Systems (e.g. PubMed) and databases (e.g., UpToDate). A related project is Wilczynski et al (2001) [60] in which a biomedical article can be classified into clinically useful but distinguishing formats (e.g., Original Study and Case Report) and purposes (e.g., Diagnosis, and Treatment) and such classifications have been incorporated into the PubMed.

Other approaches related to question answering include SemRep [61],[62], which maps biomedical text to the UMLS concepts and represents concept relations with the UMLS semantic relationships (e.g., TREATS, Co-OCCURS_WITH, and OCCURS_IN), and then condenses the concepts and their semantic relations to generate a short summary. Essie is an information retrieval engine developed and used at the NLM that incorporates knowledge-based query expansion and heuristic ranking [63]. CQA-1.0 [61] attempts to capture elements related to EBM (e.g., strength of evidence). In their study, Sneiderman et al [61] integrated the three systems (SemRep, Essie, and CQA-1.0) to achieve the best information retrieval system (that outperformed each of the three systems) in response to clinical questions.

Most systems described above, however, are not available online. To our knowledge, AskHERMES (http://www.askhermes) is the only medical search engine available online that can provide answers in response to ad hoc, complex clinical questions. Figure 1 shows AskHERMES' architecture.

As shown in Figure 1, automatically analyzing clinical questions is the first step towards answering clinical questions. Clinicians typically ask complex questions, and there is a wealth of research proposing ways for structuring those ad hoc questions. Ely and colleagues [1] studied the 1,396 medical questions they collected in one study (1) to manually map to a set of 69 question types (e.g. “What is the cause of symptom X?” and “What is the dose of drug X?”) and 63 medical topics (e.g. *drug* or *cardiology*). Cimino and associates [64] predefined a set of generic questions (e.g., “What is treatment for disease?”) and then mapped ad hoc clinical questions to those generic questions. Seol and associates [65] identified four question types: *treatment*, *diagnosis*, *etiology*, and *prognosis*.

Niu and colleagues [47],[66] applied the PICO framework (Problem/Population, Intervention, Comparison, Outcome) to analyze clinical questions. Demner-Fushman and colleagues [48],[56-61] extracted the PICO components from texts for question answering. Previously, we developed supervised machine-learning techniques to automatically classify

medical questions into the evidence taxonomy constructed by Ely and associates [67], and we reported a ~50% F1 score for classifying a clinical question into five categories defined by the evidence taxonomy [34],[35]. In this paper, we report on models for computationally identifying both the general topics and keywords incorporated in these questions.

3 Data

Ely and associates (1) collected thousands of clinical questions from more than 100 family doctors. Until 2009, the National Library of Medicine (NLM) maintained the 4,654 questions collected in four studies [1],[14],[68],[69]. Furthermore, each question was annotated by the investigator who recorded the question[1],[14],[72],[73]; a subset of those questions are shown in Table 1.

When the annotated data was released by NLM, each question was assigned one or more general topics, and Table 2 shows the 13 general topics and the number of questions assigned to the 4,654 clinical questions. As shown in Table 2, three highest assigned topics are pharmacology (1,594), management (1,403), and diagnosis (994). Of the total 4,654 questions, 3,484 were assigned one general topic, 386 were assigned two topics, 700 were assigned three topics, 4 were assigned four topics and 5 were assigned five topics. 75 questions were not assigned any topics. An example of a question that was not assigned any topics is “What are some general facts on inflammatory bowel disease?” One question assigned five topics (*treatment & prevention, test, diagnosis, pharmacology*, and *management*) was “What is the right interval for checking the thyroid stimulating hormone level on patients on thyroid replacement, and if you make a change in dose, when should you check the thyroid stimulating hormone?”

In addition, each clinical question was assigned from one to three keywords: 4,169 questions were assigned one keyword, 471 were assigned two keywords and 14 were assigned three keywords. For example, the question as shown above was assigned two keywords: *thyroid function* and *tests*.

4. Methods

The information needs of ad hoc clinical questions can be represented by two means. First, clinical questions can be classified by general topics including *etiology, procedure, diagnosis, prognosis*, and *treatment and prevention*. Automatic topic assignment may improve information retrieval. For example, we may return a pre-classified *treatment* article in response to a *treatment* question.

Secondly, each clinical question incorporates specific topics (or keywords) that indicate the main content of the question. For example, the question *“In this patient with back pain, how do you make a diagnosis of arachnoiditis and how do you treat it?”* concerns *treatment* and *diagnosis* as general topics, and its keywords are *back pain* and *arachnoiditis*. The keywords can be used as query terms for retrieving relevant documents. They can also be used as the anchor terms for answer extraction.

We distinguish between *extractive keywords* and *indicative keywords. Extractive keywords* are those that appear explicitly in a question. The two extractive keywords for the question “In this patient with back pain, how do you make a diagnosis of arachnoiditis and how do you treat it?” are *back pain* and *arachnoiditis*. Indicative keywords are those that do not appear explicitly in the original question but are implied by other words in the question. For example, *insomnia* is the assigned keyword for the question “Is Melatonin good for anything? I don't know anything about Melatonin. I need to know the dose.” In this case, the

assignment was based on the knowledge that *Melatonin* is a medication for insomnia. Figure 2 shows three types of indicative keywords.

In this work, we report on two computational models for identifying general topics and explicit keywords, two types of information needs from ad hoc clinical questions. Automatically identifying indicative keywords is a challenging but crucial task and will comprise much of our future research. We are currently focused on the simpler task of identifying extractive keywords, and we have developed both unsupervised and supervised approaches to automatically identify extractive keywords from complex clinical questions.

4.1. General Topics Identification

We explored supervised machine-learning (ML) approaches to automatically assign a question one or more general topics shown in Table 2. All ML models were trained on the 4,654 annotated clinical questions. Since a question can be assigned to multiple topics, a multi-category classifier would prohibit such a multi-category question in inclusion. We therefore developed a binary ML classifier (Yes or No) for each of the 12 topics; we excluded the category *unspecified* because it was empty.

4.1.1 Supervised Machine Learning Systems—We experimented with several commonly used machine-learning algorithms for question classification, including naïve Bayes, decision tree, and support vector machines (SVMs), and the results of 10-fold cross validation showed that SVMs performed the best. The results were consistent with our previous studies [34],[35]. We therefore report the results for SVMs, the best classifier. For comparison, we also report the results for naïve Bayes.

4.1.2 Machine Learning Features—We explored different features for machine learning, including words and n-grams, part-of-speech (POS) and stemming. We used the Stanford Parser (http://nlp.stanford.edu/downloads/lex-parser.shtml) for the POS tagging, as recent research demonstrates a good performance of the Stanford Parser in the biomedical domain [70]. In our previous work, we found that adding the UMLS concepts and semantic types as additional features led to enhanced performance in question classification [34],[35]; we therefore mapped terms in questions to the UMLS concepts and semantic types and explored them as additional features. We applied the tool MMTx [71] identify appropriate UMLS concepts and semantic types in a question string. We also applied mutual information to select top features for question classification.

4.1.3 Training and Testing—As shown in Table 2, the distribution of clinical questions to different topics is skewed, with a large majority of questions assigned to the top three topics (pharmacology, management, and diagnosis). To compare the performance of different binary classifiers, we arranged a baseline of 50% for each classifier, so that each classifier was trained on the same number of “positive” and “negative” questions. For example, when we trained a binary classifier for *diagnosis*, we had 994 questions that were assigned to this topic (Table 2). This set of 994 questions represents the “positive” training data. To generate “negative” training data, we randomly selected 994 questions from among the remaining topics.

We report the classification performance by 10-fold cross-validation, in which we divided data into 10 folds and used 9 randomly selected folds for training and the remaining fold for testing. We repeated the classification 10 times and here report the average and standard deviation.

4.1.4 Evaluation Metrics—We report the performance by recall, precision, and F1 score, all of which are commonly used as evaluation metrics for text categorization, and we report the average F1 scores. Each F1 score (F) is calculated by F = (2 * Precision * Recall) / (Precision + Recall), where recall is the number of correctly predicted medical questions divided by the total number of annotated questions in the same category, and precision is the number of correctly predicted medical questions divided by the total number of predicted questions in the same category.

4.2 Keyword Identification

In the biomedical domain, unsupervised approaches (e.g., pattern matching, or syntactic and semantic grammar-based) have been the foundation of many successful clinical natural language processing systems (e.g., [72],[73]). Supervised machine-learning approaches have also shown success (e.g., [74]). However, the success depends upon the quality and the quantity of annotated data (e.g., [75]). Supervised machine-learning approaches may under-perform unsupervised one. For example, term variations (e.g., abbreviations [76]) allow data used for training to be sparse. In this study, we therefore explored both unsupervised and supervised approaches for keyword identification. The unsupervised approaches explored shallow parsing, noun phrase identification, domain-knowledge, and corpus statistics, while the supervised approaches explored the machine learning models logistic regression and conditional random fields.

4.2.1 Unsupervised Approaches—A simple baseline system entails extracting every word in a question as a keyword. For unsupervised keyword identification, we first parsed each question to extract noun phrases as candidate keywords. For this task, we applied MMTx, which incorporates a biomedical, domain-specific shallow parser. We then ranked the noun phrases based on Inverse Document Frequency (IDF), which is commonly used in information retrieval [77]. The principle idea of the IDF model is that a noun phrase (e.g., *hypertension*) has a more important semantic role if the noun phrase is less frequently used across documents than those (e.g., the stop words, including *a* and *the*) that commonly appear in documents. The IDF of a single word noun phrase W_i is:

$$IDF(W_i)=\log\left(\frac{N}{N(W_i)}\right) \quad (1)$$

Where N is the total number of documents–a total of 17 million MEDLINE records (1966–2008), and N(W_i) is the number of documents in which W_i appears.

If a noun phrase incorporates a sequence of words, $W_1W_2\ldots W_n$, we adapted the formula in [78] to calculate the IDF value for that noun phrase as the sum of the IDF values of each word:

$$IDF(W_1W_2\ldots W_n)=\sum_{i=1}^{n} IDF(W_i) \quad (2)$$

We consider that the higher the IDF value, the higher priority the keyword. A baseline model is used to randomly select noun phrases as keywords.

We speculated that medical, domain-specific terms are likely to be keywords. For example, *hypertension* is clearly a more important term than the personal last name *Wilbur* despite

Wilbur having a higher IDF value based on its number of Google hits. We therefore built a domain-filtering model, which means that a term is only included as a keyword if it can be mapped to a UMLS concept. We integrated domain filtering with the IDF model. For the integrated UMLS+IDF model, we first applied the MMTx to map a question string to the corresponding UMLS concepts and then ranked the mapped concepts based on the IDF model.

The IDF and domain-filtering methods rank the candidate keywords, but they do not make decisions on how many top ranked keywords will be included. To determine the number of top-ranked keywords that should be included, we used a heuristic formula (3) that assigns the number of keywords as a function of the total number of word tokens in the question. The formula was based on our observation that the number of keywords increases when the question length increases. We found that this formula—although simple and heuristic—performed well for automatic keyword extraction.

$$N=\begin{cases} \frac{\textit{Number of words in the question}}{6}+1 \\ 3, \textit{if the number of words in the question is more than } 12 \end{cases} \quad (3)$$

4.2.2 Supervised Machine-Learning—We applied two representative, but different supervised machine-learning methods–logistic regression [79] and conditional random fields (CRFs) [80] – to identify automatically keywords in a medical question. Logistic regression (LR) is a multivariable method for modeling dichotomous outcomes — in our application, keyword or non-keyword. LR has been widely used in medicine [81]. Conditional random fields is relatively new [80] but has shown to be the best ML algorithm (surpassing SVM) for named entity recognition in the biomedical domain [82]. In our study, we treated *keyword* as a named entity.

4.2.2.1 Machine Learning Algorithms: The logistic regression model [79] predicts the probability of occurrence of an event by fitting data to a logistic curve. The posterior probability Prob (X) is the logistic of a linear function of the feature vector $Y_1, Y_2, \ldots, Y_n$:

$$\text{Prob}(X)=\frac{1}{1+e^{-y}}, \text{y}=\alpha_0+\alpha_1 Y_1+\alpha_2 Y_2+\alpha_3 Y_3+\ldots+\alpha_n Y_n \quad (4)$$

Conditional Random Fields (CRTs) [80] are generative probabilistic models used to segment and label sequence data and offer advantages over hidden Markov models because they are able to relax the strong independence assumption. To apply the CRFs model, we consider a clinical question as a sequence of word tokens and then identify its keywords based on the sequence of tokens that appear in the question.

4.2.2.2 Learning Features: For this study, we explored the learning features that have been described in Section 4.1.2. In addition, we added word length (i.e., the number of characters in every word) as a feature because domain-specific words (e.g., “gastrectomy”) tend to be lengthy when compared to common English words, and there is a correlation between the length of a word and its IDF value. We also added the position of a word in a question string as an additional feature, because we have observed that an important term sometimes appears toward the end of a clinical question. For example, “corneal foreign body” and “immune deficiency” appear towards the end of the questions “What is a good protocol for diagnosis and treatment of corneal foreign body?” and “What tests should be done to screen

for immune deficiency?" Note that each feature on its own will not be sufficient enough for discrimination; however, combining multiple features will improve performance in keyword identification.

4.2.2.3 Evaluation: We used the manually assigned keywords as the gold standard for evaluating our automatic keyword identification approach. In addition to the types of indicative keywords shown in Figure 2, we found that extractive keywords also have variations. These variations include simple morphosyntactic variation (*Marfan syndrome* and *Marfan's syndrome*), orthographic variation (*obsessive-compulsive disorder* and *obsessive compulsive disorder*), synonyms (*bladder infections* and *urinary tract infections*), and knowledge inferencing (*lead* and *lead poisoning*). To map a question string to its varied keywords, we used a simple approximation-matching approach that was built upon edit distance [83] with an empirical cut-off threshold. Note that keywords derived from synonyms and knowledge inferencing can also be identified as "indicative keywords." However, in this application, we separate extractive keywords from indicative keywords if there is a single word in the assigned keyword that is in common with a word in the clinical question.

A total of 3,155 questions that have a total of 3,353 associated keywords assigned to them were used to evaluate our keyword identification task. The question collection incorporates a total of 55,129 words and 13,060 noun phrases identified by MMTx.

The evaluation reflects Recall, Precision, and F score. Recall is the number of correctly predicted keywords divided by the total number of assigned keywords, and Precision is the number of correctly predicted keywords divided by the total number of predicted keywords. We calculated the F1 score which is 2*Recall*Precision/(Recall+Precision).

4.3 Error Analysis

We performed multiple error analysis steps to understand the source of errors in question classification and keyword extraction. For question classification, we first plotted the classification performance on the basis of the training size. Our hypothesis is that question classification performance increases with an increase in training size. We also plotted the classification performance as a function of the number of assigned categories. As described in Section 3, our annotated collection has assigned 1—5 categories to each question. If Question A is assigned more categories than Question B, common wisdom suggests that the performance of the question classification of A will be worse than the performance of B. We will test this hypothesis. We plotted the performance of keyword extraction as a function of question length. We also manually examined the cases for both question classification and keyword extraction.

5. Results

Table 3 shows the SVM results for automatically classifying an original clinical question into general topics. The results show that the best system was trained on bag-of-words, bigrams, POS, and the UMLS concepts and semantic types, which led to an average of 76.1% recall, 77.0% precision, and 76.5% F1 score.

Using bag-of-words as features, the average F1 score was 70.7%. We found that other features have an impact on this performance. Stemming enhanced the performance (an absolute increase of 1.6%). Bigrams also slightly enhanced the performance (an absolute increase of 0.5%), with a slight further enhancement from POS (an absolute increase of 0.01%). UMLS concepts and semantic types significantly improved the performance to its highest value — a 76.47% F1 score – which is statistically significant compared to bag-of-

words ($p<0.0001$, t-test). When POS was also added, the overall performance decreased to 75.6%, although the decrease was not statistically significant. We found that feature selection also slightly decreased the performance (an absolute decrease of 0.4%); however, the decrease was not statistically significant.

Figure 3 shows the classification performance of topic assignment as a function of training size. The category *history* has the lowest classification performance (67.7% F1 score), and *pharmacology* has the highest classification performance (89.3% F1 score). The Pearson's correlation shows a P value of 0.07 (one tail), which is not statistically significant.

Figure 4a shows the number of questions as a function of the number of topic assignments; the relation resembles a power law distribution, with the highest percentage (76.5%) being questions that were assigned only one topic and much fewer being questions that were assigned two labels or more. Specifically, 8.3% of the questions were assigned two topics, 15.0% questions were assigned three, four questions were assigned four, and five questions were assigned five topics.

The questions that were assigned five topics include “What are the indications for getting a digoxin level?”, which was assigned *Treatment and Prevention*, *Test*, *Diagnosis*, *Pharmacology*, and *Management*, and “Can this rash be a drug reaction to 1% hydrocortisone or some inert ingredient in the preparation?”, which was assigned *Treatment and Prevention*, *Diagnosis*, *Physical Finding*, *Pharmacology*, and *Management*. Figure 4b shows the classification performance of topic assignment as a function of number of topics assigned to a question.

Table 4 shows the performance of keyword extraction with both supervised and unsupervised approaches. The results show that the baseline of a random word achieves an 11.4% F1 score. Limiting to noun phrase increases the performance to 17.6%. IDF has further boosted the performance to a still very poor 29.6%. When the UMLS concepts were used for filtering keywords, the performance increased to 53.8%, the highest in the unsupervised machine learning approaches. For comparison, if all the UMLS concepts that appeared in a question were considered keywords, the F1 score was only 29.5%, comparable to the method of noun phrase+IDF. Statistical analysis shows that all performance increases were significant ($p<0.01$, t-test).

Table 4 also shows that both supervised approaches outperformed unsupervised ones. Logistic regression increased the F1 score 2.4% when compared to the best unsupervised approach (UMLS concept+IDF), yielding a 55.1% F1 score in keyword identification, although Statistical analysis shows that the increase is not statistically significant (t-test). Conditional random fields performed the best with a 58.0% F1 score, which showed an absolute increase of 7.8% over the UMLS concept+IDF approach. The results show that although the difference in keyword identification between the CRF model and the logistic regression model is not statistically significant (t-test), the difference between the CRF model and the UMLS+IDF is statistically significant ($p<0.01$, t-test).

Figure 5 (a) shows recall, precision and F1 score of keyword extraction as a function of question length. The results show that the recall is mostly high (>80%) and mostly independent of question length. In fact, when the number of word tokens of a question is more than 60, the recall increases with question length. On the other hand, precision decreases with question length, and as a result, the F1 score decreases as well.

Figure 5(b) shows the distribution of questions as a function of number of word tokens in a question. As shown, the number of questions decreases when the number of word tokens increases and a majority of questions (92%) have a number of word tokens below 40. The

longest question is 114 words, as shown here: “In a patient with infectious mononucleosis, if the spleen is going to get enlarged, how long does it take to do that? That is, how long should they stay out of sports? And then the other question is, how long does it take to go back to normal. Also, how big is a 10-year-old's spleen supposed to be on ultrasound? Do you have to wait until it goes back to normal size to go back to sports or just wait 4 to 6 weeks and then go back regardless? Also, is feeling the spleen good enough or do you need to do an ultrasound to see if someone should stay out of sports?”

We manually analyzed inconsistent assignments and found many inconsistent annotations. For example, *Management* can refer to patient management, the scope of which can include *Diagnosis*, *Treatment & Prevention*, *Pharmacology*, or *Test*. If a question was assigned to those categories, it seems that *Management* might also be co-assigned. In fact, we found that many annotated questions followed this rule. For example, “What is the dose of neurontin?” was assigned three categories: *Management, Treatment and Prevention*, and *Pharmacology*. However, there was inconsistency, as some questions did not follow this rule. For example, “What is the dose of aspirin needed to prevent TIA's (transient ischemic attacks)?” was assigned to only one category: *Pharmacology*.

Similar to the inconsistent annotation of *Management*, we found that inconsistency appeared in many other category assignments. For example, the question “35-year-old female with ‘throat feels funny’ possible allergies. She couldn't describe the feeling other than her throat feels funny. The question is what is going on? Could it be related to her beta blocker?” was assigned the category “History” by the NLM, but we believe it should be assigned two categories: *Diagnosis* and *Management*.

We also found inconsistency in keyword assignment. For example, “Is cow's milk a risk for mad cow disease?” was assigned only the keyword “milk,” and we believed that “mad cow disease” should be added. Our system has correctly identified both “milk” and “mad cow disease” as keywords. Some inconsistency is due to keyword composition. For example, the question “What are the clinical signs of neonatal myasthenia gravis?” was assigned two keywords – *myasthenia gravis* and *neonatal*, we consider that one keyword *neonatal myasthenia gravis* is a better model for this question, and our system correctly identified this as the keyword.

6. Discussion

Overall, as shown in Table 3, the average performance for automatically assigning a category to a question was a 76.5% F1 score, as opposed to the baseline of 50% attained by random guessing. Our results show that feature selection impacted question classification performance. While stemming and bi-grams improved the F1 score slightly (the absolute increase of 0.5%~1.6%), the UMLS concepts and semantic types had the highest F1 score increase (the absolute increase of 5.3%). The results were consistent with our previous work in question classification (45, 46). Our results show that neither POS nor feature selection improved the performance of question classification, and we speculate that data sparseness is the cause.

For the task of keyword identification with unsupervised machine-learning approaches, we assigned the number of top-ranked keywords as a function of the total number of word tokens in the question. Despite the simplicity, we found the formula performs quite well. Our results show the baseline of randomly selecting words achieved an F1 score of only 11.4%; the results indicate that keyword identification is a challenging research task. Our results show that limiting keyword candidates to noun phrases only helped increase the performance to an F1 score of 17.6%, which is still very poor performance. We found a

significant increase (an absolute increase of 12%) in performance when IDF prioritization was introduced. Our results support the method of query prioritization with the IDF value, a commonly used technique in open-domain question answering (e.g.,[78]). After domain-filtering, we obtained the highest performance (53.8% F1 score) in unsupervised approaches, which was an 81.8% increase over noun phrase+IDF. Our results show the importance of domain-specific knowledge in keyword identification in medical questions.

As stated earlier, in the clinical domain, unsupervised approaches (e.g., pattern matching, or syntactic and semantic grammar-based) have been the foundation of many successful clinical natural language processing systems (e.g., [74],[75]). Our unsupervised approaches are competitive approaches that were built upon previous work [78],[84]. Nevertheless, both supervised machine-learning approaches (i.e., logistic regression and CRFs) outperformed unsupervised ones. Specifically, CRFs outperformed the best unsupervised approach – UMLS+IDF ($P<0.01$) – increasing its F1 score by an absolute value of 4.2%, to 58.0%. Several factors may contribute to the results. Our clinical question collection has shown that many questions were ill-formed with grammatical errors, which has made linguistically-driven and rule-based approaches a challenge. Supervised learning has the ability to learn the relations from multiple features and was therefore robust in this specific task.

Another advantage of supervised machine-learning approaches is that the number of keywords is automatically predicted. This is in contrast to our unsupervised machine-learning approaches in which the number of keywords is based on the question length. Figure 5(a) shows that the recall of keyword prediction is high (in most questions, it is >80%). Furthermore, the recall remains high or increases when question length increases. The results further demonstrated the robustness of supervised machine learning approaches for keyword extraction.

Our results show that the use of CRFs outperformed logistic regression, although not at a statistically significant level (t-test). The results suggest that sequential information (or word order) may contribute to the performance difference.

As described in the results, we found inconsistent topic and keyword assignments. We speculate that the inconsistency is caused by the fact that there has not been an annotation guideline for the annotated clinical question collection and there isn't any report for annotation agreement. We found ambiguity in the scope and scope overlap between different categories. For example, *Management* might or might not refer to patient management, the scope of which may include *Diagnosis, Treatment & Prevention, Pharmacology*, or *Test*. A lack of annotation guideline also leads to inconsistency in keyword composition as shown in the example of “What are the clinical signs of neonatal myasthenia gravis?”

Inconsistency in topic assignment may be responsible for the relation between the training size and the topic classification performance, as shown in Figure 3. Typically, there is a positive relation between a training size and a classification performance: the larger the training size, the better a classifier performs. Our Pearson's correlation analysis, however, concluded that there is only a weak correlation ($p=0.07$) between the training size and the topics classification performance. Although the results show that best performing category, *pharmacological*, had the largest number of question instances (1,594), and the worst performing category *history* (67.7% F1 score) had the least number of question instances (43), *Management* did not perform well (71.4% F1 score, comparing with the highest 89.3% F1 score for *pharmacological*) even though the number of instances available for training was high (1,403). *Procedure*, on the other hand, has only 122 instances, but our classifier achieved an 80.5% F1 score, the 3rd highest classification performance. We speculate that

procedure is an unambiguous category for assignment, and therefore it has a highly consistent annotation.

We examined the topic assignment performance by the number of topics assigned to a question. Intuitively, we speculate that there is an inverse relation between the number of topics that are assigned to a question and the classification performance for topics assignment: when the number of topics assigned to a question is higher, the classification performance of that question can be lower, and vice versa. However, our results, as shown in Figure 4, show that the classification performance of a question did not correlate with the number of categories assigned to the question. In addition to inconsistent topics assignment, difference in training data may additionally explain this result. As described earlier, the number of questions that corresponds to the number of categories assigned to questions shows a power law distribution: a large number of questions were assigned to one topic only, and a much fewer number of questions were assigned two or more topics. As a result, our observation may not be generalizable.

Nevertheless, a lack of correlation between the number of topics and the classification performance in our question data collection supports our binary classification strategy — to select a negative set of data randomly from all other categories other than the classification category – because the binary classification performance does not depend upon the number of assigned categories. Typically, the performance of a multi-classifier decreases when the number of categories increases.

Despite the noisy data, our results show a good and reliable performance for question classification and keyword identification. In fact, we observe that in some cases our system outperformed the original annotation data to automatically assign the correct topics for question classification and to automatically identify the correct keywords for keyword identification.

Question classification and keyword extraction can improve information retrieval and question answering, because resource searches are based on content (topic areas and keywords), not the simpler bag-of-words that treat each word mostly equally. Although validation of this assertion remains a future work, we have performed a pilot evaluation as reported in [85]. In our study, we showed whether a simple model in which we increased the weight of keywords in a question may lead to improved information retrieval. Since there is no evaluation data available for clinical information retrieval and question answering, we evaluated our approach using the text collection of the Genomics Track of the Text REtrieval Conference (TREC), which incorporates more than 160,000 full-text biomedical articles [86]. The 2006 and 2007 tasks focused on information retrieval for question answering [87],[86]; a sample question from the tasks is “What is the role of IDE in Alzheimer's disease?” We employed a simple model in which we increased the weight of keywords and the results showed an improvement in information retrieval that was statistically significant [85].

7. Conclusions and Future Work

We report here on two natural language processing models, namely, automatic topic assignment and keyword identification, that together automatically and effectively extract information needs from ad hoc clinical questions. Both models can be accessed from the AskHERMES system (http://www.AskHERMES.org). The first model automatically assigns general topics (e.g., *Diagnosis* and *Treatment and Prevention*) and the second model automatically extracts keywords or semantic content. Our evaluation of 4,654 annotated clinical questions has shown an average performance of 76.5% F1 score for the first model

and 58.0% F1 score for the second model. We found that a significant amount of inconsistent annotation lowered the performance for both models, and we anticipate improved models if consistent annotations can be achieved. Future work will focus on the annotation and will investigate and evaluate how the two models improve clinical question answering. For example, we may integrate the work of [60] to categorize the MEDLINE articles as question-type specific– *etiology*, *diagnosis*, and *treatment*, to improve information retrieval. We may also explore probabilistic models [88] to incorporate automatic keyword identification for improving clinical question answering.

Acknowledgments

The authors acknowledge support from the National Library of Medicine, grant number 1R01LM009836, and from the University of Wisconsin-Milwaukee, a MiTAG grant. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the NIH and UWM.

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;319:358–61. [PubMed: 10435959]
2. Timpka T, Arborelius E. The GP's dilemmas: a study of knowledge need and use during health care consultations. Methods Inf Med 1990;29:23–9. [PubMed: 2407930]
3. Bergus GR, Randall CS, Sinift SD, Rosenthal DM. Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? Arch Fam Med 2000;9:541–7. [PubMed: 10862217]
4. Ely JW, Burch RJ, Vinson DC. The information needs of family physicians: case-specific clinical questions. J Fam Pract 1992;35:265–9. [PubMed: 1517722]
5. Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med 1991;114:576–81. [PubMed: 2001091]
6. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med 1985;103:596–9. [PubMed: 4037559]
7. Smith R. What clinical information do doctors need? Bmj 1996;313:1062–8. [PubMed: 8898602]
8. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998;280:1347–1352. [PubMed: 9794316]
9. Hersh W. Evidence-based medicine and the Internet. ACP J Club 1996;125:A14–6. [PubMed: 8963525]
10. Westbrook JI, Gosling AS, Coiera E. Do clinicians use online evidence to support patient care? A study of 55,000 clinicians. J Am Med Inform Assoc 2004;11:113–20. [PubMed: 14662801]
11. Westbrook JI, Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? J Am Med Inform Assoc 2005;12:315–21. [PubMed: 15684126]
12. Gosling AS, Westbrook JI. Allied health professionals' use of online evidence: a survey of 790 staff working in the Australian public hospital system. Int J Med Inform 2004;73:391–401. [PubMed: 15135758]
13. Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, Pifer EA. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ 2002;324:710–713. [PubMed: 11909789]
14. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc 2005;12:217–24. [PubMed: 15561792]

15. Hoogendam A, Stalenhoef AF, Robbe PF, Overbeke AJ. Answers to questions posed during daily patient care are more likely to be answered by UpToDate than PubMed. J Med Internet Res 2008;10:e29. [PubMed: 18926978]
16. Cullen RJ. In search of evidence: family practitioners' use of the Internet for clinical information. J Med Libr Assoc 2002;90:370–9. [PubMed: 12398243]
17. Stephens MB, Von Thun AM. Military medical informatics: accessing information in the deployed environment. Mil Med 2009;174:259–64. [PubMed: 19354089]
18. Tang H, Ng JH. Googling for a diagnosis–use of Google as a diagnostic aid: internet based study. Bmj 2006;333:1143–5. [PubMed: 17098763]
19. Kitchin DR, Applegate KE. Learning radiology a survey investigating radiology resident use of textbooks, journals, and the internet. Acad Radiol 2007;14:1113–20. [PubMed: 17707320]
20. Purcell GP, Wilson P, Delamothe T. The quality of health information on the internet. Bmj 2002;324:557–8. [PubMed: 11884303]
21. Jadad AR, Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? Jama 1998;279:611–4. [PubMed: 9486757]
22. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor–Let the reader and viewer beware. Jama 1997;277:1244–5. [PubMed: 9103351]
23. Glennie E, Kirby A. The career of radiography: information on the web. Journal of Diagnostic Radiography and Imaging 2006;6:25–33.
24. Childs S. Judging the quality of internet-based health information. Performance Measurement and Metrics 2005;6:80–96.
25. Griffiths KM, Christensen H. Quality of web based information on treatment of depression: cross sectional survey. Bmj 2000;321:1511–5. [PubMed: 11118181]
26. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001;16:671–92. [PubMed: 11780707]
27. Benigeri M, Pluye P. Shortcomings of health information on the Internet. Health Promot Int 2003;18:381–6. [PubMed: 14695369]
28. Wyatt JC. Commentary: measuring quality and impact of the World Wide Web. Bmj 1997;314:1879–81. [PubMed: 9224133]
29. McClung HJ, Murray RD, Heitlinger LA. The Internet as a source for current patient information. Pediatrics 1998;101:E2. [PubMed: 9606244]
30. De Leo, G.; LeRouge, C.; Ceriani, C.; Niederman, F. Websites most frequently used by physician for gathering medical information. AMIA Annu Symp Proc; 2006; p. 902
31. Freeman MK, Lauderdale SA, Kendrach MG, Woolley TW. Google Scholar versus PubMed in locating primary literature to answer drug-related questions. Ann Pharmacother 2009;43:478–84. [PubMed: 19261965]
32. Fletcher RH, Fletcher SW. Evidence-based approach to the medical literature. J Gen Intern Med 1997;12 2:S5–14. [PubMed: 9127238]
33. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P, Mosbaek C, Kraemer D. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. J Am Med Inform Assoc 2002;9:283–93. [PubMed: 11971889]
34. Yu, H.; Sable, C. Being Erlang Shen: Identifying answerable questions. Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasoning for Answering Questions; 2005;
35. Yu, H.; Sable, C.; Zhu, HR. Classifying Medical Questions based on an Evidence Taxonomy. Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains; 2005;
36. Yu, H.; Wei, Y. The semantics of a definiendum constrains both the lexical semantics and the lexicosyntactic patterns in the definiens. Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL; New York, USA. 2006; p. 1-8.
37. Yu, H. Towards Answering Biological Questions with Experimental Evidence: Automatically Identifying Text that Summarize Image Content in Full-Text Articles. AMIA Annu Symp Proc; 2006; p. 834-8.

38. Lee, M.; Cimino, J.; Zhu, HR.; Sable, C.; Shanker, V.; Ely, J.; Yu, H. Beyond information retrieval-medical question answering. AMIA Annu Symp Proc; 2006; p. 469-73.
39. Yu H, Lee M, Kaufman D, Ely J, Osheroff J, Hripcsak G, Cimino J. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. J Biomed Inform 2007;40:236–51. [PubMed: 17462961]
40. Yu, H.; Cao, YG. Automatically extracting information needs from ad hoc clinical questions. AMIA Annu Symp Proc; 2008; p. 96-100.
41. Yu, H.; Cao, YG. Using the weighted keyword models to improve clinical question answering. IEEE International Conference on Bioinformatics & Biomedicine Workshop NLP Approaches for Unmet Information needs in Health Care; 2009;
42. Kim, D.; Yu, H. Hierarchical Image Classification in the Bioscience Literature. Proc AMIA Symp; 2009;
43. Cao YG, Ely J, Antieau L, Yu H. Evaluation of the clinical question answering presentation. Submitted to BioNLP. 2009
44. Cao, YG.; Ely, J.; Yu, H. Using weighted keywords to improve clinical question answering. Proceeding of IEEE BIBM Workshop in NLP Approaches in Unmet Information Needs in Health Care; 2009;
45. Agarwal, S.; Yu, H. Automatically generating structured text summaries for figures in biomedical literature. AMIA Annu Symp Proc; 2009;
46. Yu, H.; Kaufman, K. A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. Pacific Symposium on Biocomputing; 2007; p. 328-339.
47. Niu, Y.; Hirst, G.; McArthur, G.; Rodriguez-Gianolli, P. Answering clinical questions with role identification. ACL workshop on natural language processing in biomedicine; 2003;
48. Cimino, JJ.; Borotsov, DV. Leading a horse to water: using automated reminders to increase use of online decision support. AMIA Annu Symp Proc; 2008; p. 116-20.
49. Huang, X.; Lin, J.; Demner-Fushman, D. Evaluation of PICO as a Knowledge Representation for Clinical Questions. AMIA Annu Symp Proc; 2006; p. 359-63.
50. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 2007;33:63–103.
51. Zweigenbaum, P. Question-answering for biomedicine: Methods and state of the art. MIE 2005 Workshop; 2005;
52. Zweigenbaum, P. Question answering in biomedicine. EACL Workshop on Natural Language Processing for Question Answering; Budapest. 2003; p. 1-4.
53. Rinaldi, F.; Dowdall, J.; Schneider, G.; Persidis, A. Answering questions in the genomics domain. ACL 2004 Workshop on Question Answering in Restricted Domain; 2004;
54. Cimino, J. Infobuttons: anticipatory passive decision support. AMIA Annu Symp Proc; 2008; p. 1203-4.
55. Cimino JJ, Li J, Allen M, Currie LM, Graham M, Janetzki V, Lee NJ, Bakken S, Patel VL. Practical considerations for exploiting the world wide web to create infobuttons. Medinfo 2004;11:277–81.
56. Cimino, JJ.; Li, J. Sharing infobuttons to resolve clinicians' information needs. AMIA Annu Symp Proc; 2003; p. 815
57. Cimino, JJ.; Li, J.; Graham, M.; Currie, LM.; Allen, M.; Bakken, S.; Patel, VL. Use of online resources while using a clinical information system. AMIA Annu Symp Proc; 2003; p. 175-9.
58. Cimino, JJ.; Li, J.; Bakken, S.; Patel, VL. Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users. Proc AMIA Symp; 2002; p. 170-4.
59. Cimino, JJ.; Elhanan, G.; Zeng, Q. Supporting infobuttons with terminological knowledge. Proc AMIA Annu Fall Symp; 1997; p. 528-32.
60. Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. Medinfo 2001:393–3.

61. Sneiderman CA, Demner-Fushman D, Fiszman M, Ide NC, Rindflesch TC. Knowledge-based methods to help clinicians find answers in MEDLINE. J Am Med Inform Assoc 2007;14:772–80. [PubMed: 17712086]
62. Srinivasan, P.; Rindflesch, T. Exploring text mining from MEDLINE. Proc AMIA Symp; 2002; p. 722-6.
63. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14:253–63. [PubMed: 17329729]
64. Cimino JJ, Aguirre A, Johnson SB, Peng P. Generic queries for meeting clinical information needs. Bull Med Libr Assoc 1993;81:195–206. [PubMed: 8472005]
65. Seol YH, Kaufman DR, Mendonca EA, Cimino JJ, Johnson SB. Scenario-based assessment of physicians' information needs. Medinfo 2004;11:306–10.
66. Niu, Y.; Hirst, G. Analysis of semantic classes in medical text for question answering. ACL 2004 Workshop on Question Answering in Restricted Domains; 2004;
67. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, Stavri PZ. A taxonomy of generic clinical questions: classification study. Bmj 2000;321:429–32. [PubMed: 10938054]
68. 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;45:382–8. [PubMed: 9374962]
69. D'Alessandro DM, Kreiter CD, Peterson MW. An evaluation of information-seeking behaviors of general pediatricians. Pediatrics 2004;113:64–9. [PubMed: 14702450]
70. Lin J, Wilbur WJ. Syntactic sentence compression in the biomedical domain: facilitating access to related articles. Information Retrieval. 2007 in press.
71. MMTx. 2005. Available at http://mmtx.nlm.nih.gov/docs.shtml
72. Friedman C, Hripcsak G, Shagina L, Liu H. Representing information in patient reports using natural language processing and the extensible markup language. Journal of the American Medical Informatics Association 1999;6:76. [PubMed: 9925230]
73. Schwartz, AS.; Hearst, MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput; 2003;
74. Patrick J, Li M. A cascade approach to extract medication event (i2b2 challenge 2009). 2009
75. Li Z, Cao Y, Antieau L, Agarwal S, Zhang Q, Yu H. A Hybrid Approach to Extract Medication Information from Medical Discharge Summaries.
76. Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc 2002;9:262–72. [PubMed: 11971887]
77. Jones SK. A satistical interpretation of term specificity and its application in retrieval. Journal of Documentation 1972;28:11–21.
78. Pradhan, SS.; Illouz, G.; Blair-Goldensohn, SJ.; Schlaikjer, AH.; Krugler, V.; Filatova, E.; Duboue, PA.; Yu, H.; Passonneau, RJ.; Ward, W.; Hatzivassiloglou, V.; Jurafsky, D.; McKeown, K.; Martin, JH. Building a foundation system for producing short answers to factual questions. Eleventh Text Retrieval Conference (TREC-11); Washington, DC. 2002;
79. Kleinbaum, DG.; Klein, M. Logistic Regression. 2nd. Springer; 2005.
80. Lafferty, J.; McCallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. ICML-01; 2001; p. 282-9.
81. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature::: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology 2001;54:979–985. [PubMed: 11576808]
82. Settles, B. Biomedical named entity recongition using conditional random fields and rich feature sets. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA); Geneva, Switzerland. 2004; p. 104-7.
83. Ristad ES, Yianilos PN, Inc MT, Princeton NJ. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998;20:522–532.
84. Johnson, SB.; Aguirre, A.; Peng, P.; Cimino, J. Interpreting natural language queries using the UMLS. Proceedings of the Annual Symposium on Computer Application in Medical Care; 1993; p. 294

85. Yu, H.; Cao, YG. Using the weighted keyword models to improve biomedical information retrieval. AMIA Summit on Translational Bioinformatics; San Francisco, USA. 2009;

86. Hersh, W.; Cohen, AM.; Roberts, P.; Rekapalli, HK. TREC 2006 Genomics Track overview. TREC Genomics Track conference; 2006;

87. Hersh, W.; Cohen, A.; Ruslen, L.; Roberts, P. TREC 2007 Genomics Track overview. The TREC Genomics Track Conference; 2007;

88. Bendersky, M.; Croft, WB. Discovering key concepts in verbose queries. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval; 2008; p. 491-498.

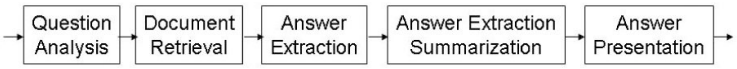

Figure 1.
AskHERMES' system architecture. AskHERMES takes as input a question posed by a clinician. *Question Analysis* automatically extracts information needs. *Document Retrieval* retrieves relevant documents (MEDLINE and WWW). *Answer Extraction* automatically identifies the sentences that provide answers to questions. *Summarization* condenses the text by removing the redundant sentences and by generating a coherent summary. *Answer Presentation* presents the summary to the user who posed the question.

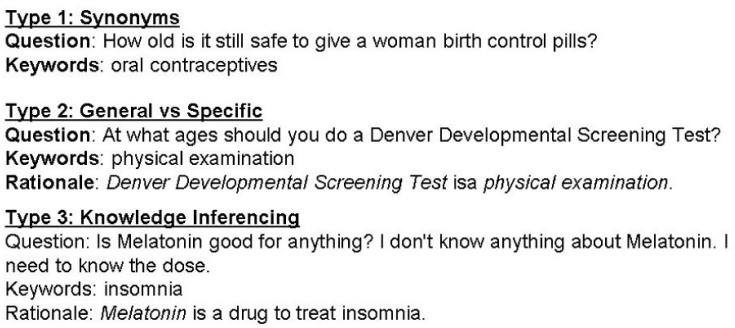

Figure 2.
Three types of indicative keywords and examples

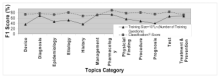

Figure 3.
The classification performance of topic assignment and the corresponding training size (training size =10*Ln (Number of Training Questions)).

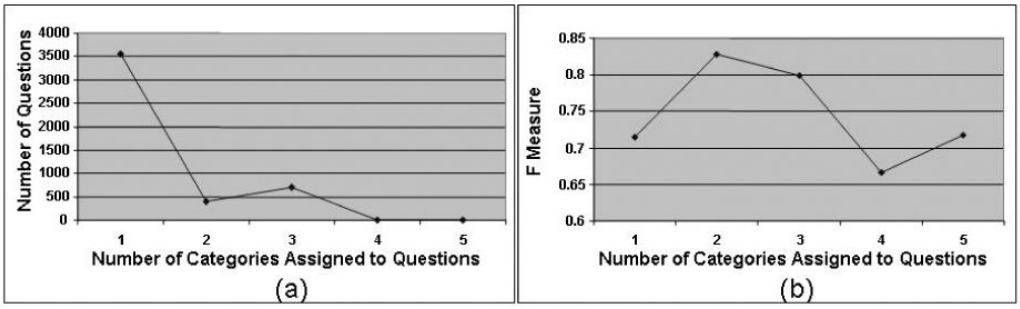

Figure 4.

(a) Number of questions as a function of number of categories assigned to questions of the ClinicalQuestions Collection; (b) Classification performance of topic assignment as a function of number of categories assigned to a question

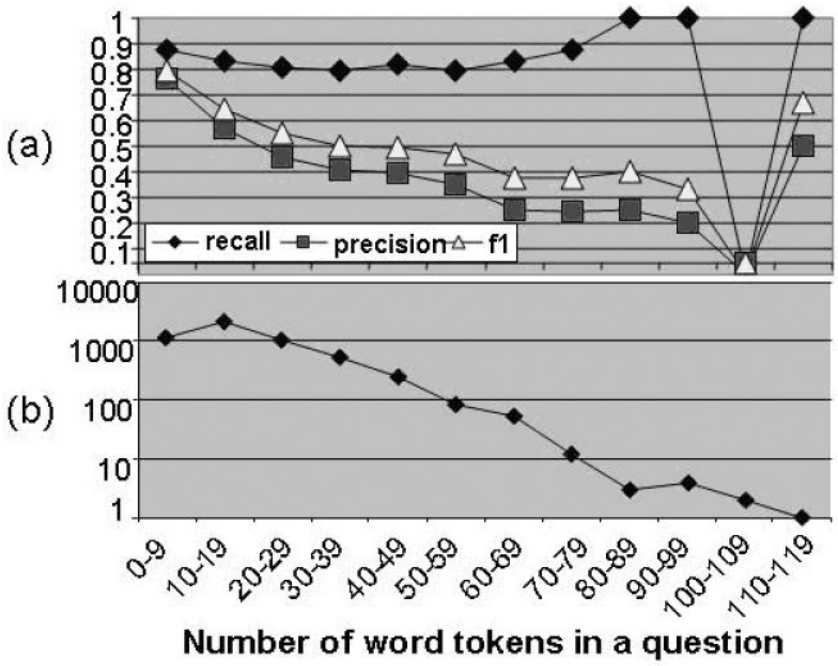

Figure 5.
(a) Recall, precision, and F1 score of keyword extraction (using CRFs) as a function of question length. (b) Number of questions as a function of question length.

Table 1

A typology of question types with representative samples of clinical questions collected by Ely and his associates (1). The left column represents generic question proportions. For example, of the 4,654 clinical questions, “What”, “How”, “Do”, and “Can” account for 2,231 (or 48%), 697 (or 15%), 320 (or 7%), and 187 (or 4%) of the questions, respectively. Question examples (1-6) are in the right column.

General Question Type and (Percentage)	Sample Questions
“What …” (48%)	1. What is the cause and treatment of this old man's stomatitis? 2. What should you do with someone who is not getting better from epicondylitis after physical therapy and nonsteroidal anti-inflammatory drugs have not worked?
“How …” (15%)	3. How long should you leave a patient on Coumadin and heparin?
“Do …” (7%)	4. Do angiotensin II inhibitors work like regular angiotensin converting enzyme inhibitors to preserve kidney function in mild diabetes?
“Can …” (4%)	5. Can Lorabid cause headaches?
Others (25%)	6. His last (and first) Hepatitis B vaccine was last August (6 months ago). I'm going to bunch the series in reverse; ie, give second injection today and 3rd in a month or two. Is that okay to bunch them in reverse?

Table 2

The general topics and the number of questions assigned to the 4,654 clinical questions

General Topics	Number of Questions (percentage of the total questions)
Device	37 (0.8%)
Diagnosis	994 (21.4%)
Epidemiology	104 (2.2%)
Etiology	173 (3.7%)
History	42 (0.9%)
Management	1403 (30.1%)
Pharmacology*	1594 (34.3%)
Physical Finding	271 (5.8%)
Procedure	122 (2.6%)
Prognosis	53 (1.1%)
Test	746 (16.0%)
Treatment & Prevention	868 (18.7%)
Unspecified	0 (0%)

* The original category is “Pharmacological” in the ClinicalQuestions Collection

Table 3

Binary classification results–precision (top) and F1 score (bottom), on 10-fold cross validation for applying support vector machines to automatically assign general topics to ad hoc clinical questions. We explored different combinations of features: bag-of-words (W), stemming (S), top 2000 features (T2000), bag of words+bigrams (Bi), bag-of-words+bigrams+part of speech (W+Bi+POS), bag-of-words+bigrams+UMLS concepts and semantic types (W+Bi+CSTY), and bag-of-words+bigrams+part-of-speech+UMLS concepts and semantic types (W+Bi+CSTY+POS). The baseline system that randomly assigns a topic is 50%.

General Topics	W	S	T2000	Words+Bigrams	W+Bi+POS	W+Bi+CSTY	W+Bi+CSTY+POS
Device	57.8%	67.4%	61.9%	64.8%	62.4%	**71.4%**	71.2%
	56.9%	65.0%	62.4%	61.7%	61.1%	**73.0%**	71.2%
Diagnosis	73.1%	73.0%	73.8%	75.6%	75.2%	76.0%	**76.4%**
	73.7%	73.8%	75.2%	75.9%	75.2%	76.8%	**77.2%**
Epidemiology	71.8%	68.8%	65.8%	69.2%	68.2%	**72.8%**	69.7%
	70.6%	68.5%	65.8%	67.9%	68.0%	**72.2%**	70.3%
Etiology	80.6%	84.5%	78.6%	**87.0%**	82.4%	84.8%	86.4%
	79.2%	81.6%	78.2%	82.4%	79.7%	80.4%	**82.6%**
History	57.3%	60.1%	54.6%	55.3%	59.9%	**68.2%**	63.5%
	54.3%	59.2%	51.3%	53.8%	57.9%	**67.7%**	61.7%
Management	69.4%	68.8%	69.9%	**73.8%**	73.2%	73.1%	72.5%
	68.4%	68.1%	68.0%	**71.4%**	71.4%	71.1%	71.0%
Pharmacology	83.4%	83.7%	83.6%	84.5%	84.5%	**89.7%**	89.0%
	82.6%	83.0%	82.9%	84.0%	83.8%	**89.3%**	88.7%
Physical Finding	71.9%	71.5%	69.8%	72.1%	69.7%	**77.6%**	76.2%
	71.7%	72.4%	72.7%	71.1%	69.6%	**77.8%**	76.7%
Procedure	69.2%	70.2%	68.2%	67.1%	66.2%	**80.4%**	80.1%
	70.4%	71.3%	69.2%	66.6%	65.4%	**80.5%**	80.3%
Prognosis	72.8%	73.9%	66.8%	68.9%	72.6%	73.5%	**74.3%**
	73.0%	74.4%	68.4%	69.2%	73.8%	74.3%	**74.3%**

J Biomed Inform. Author manuscript; available in PMC 2011 December 1.

General Topics	W	S	T2000	Words+Bigrams	W+Bi+POS	W+Bi+CSTY	W+Bi+CSTY+POS
Test	79.5%	81.3%	79.5%	81.3%	78.7%	**84.4%**	83.4%
	79.1%	80.6%	79.0%	79.2%	78.2%	**83.0%**	82.4%
Treatment & Prevention	67.8%	68.6%	68.8%	71.1%	70.1%	**71.8%**	70.5%
	68.0%	68.8%	69.8%	70.3%	69.6%	**71.6%**	70.5%
Average	71.2%	72.7%	70.1%	72.6%	71.9%	**77.0%**	76.1%
	70.7%	72.2%	70.2%	71.1%	71.1%	**76.5%**	75.6%

Bold indicates the highest performance

Table 4

Performance of keyword extraction. “IDF” indicates keyword prioritization with IDF value; otherwise a random selection of keywords is applied. “UMLS concept” indicates that we only use the text as a keyword if the text can be mapped by MMTx to a UMLS concept. We experimented with selecting all UMLS concepts and then prioritizing with its IDF value. We also report the results of logistic regression and conditional random fields.

Method	Precision	Recall	F1-score
Random words	11.2%	11.6%	11.4%
Noun phrase	16.5%	18.9%	*17.6%
Noun phrase+IDF	28.0%	31.4%	*29.6%
All UMLS concepts	17.5%	95.0%	29.5%
UMLS concept+IDF	44.3%	68.6%	*53.8%
Logistic regression	68.7%	46.0%	55.1%
Conditional random fields	67.6%	50.8%	58.0%

*indicates that the F1 score is statistically significant ($p<0.01$, t-test) compared with the score in the previous method. The F1 score of conditional random fields is statistically significant compared with the UMLS concept+IDF. The difference between Logistic regression and Conditional random field and between UMLS concept+IDF and Logistic regression is not statistically significant.