Automatically Extracting Information Needs from Ad Hoc Clinical Questions

Hong Yu[1,2,3], PhD, Yong-gang Cao[1,3], PhD

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

Abstract

Automatically extracting information needs from ad hoc clinical questions is an important step towards medical question answering. In this work, we first explored supervised machine-learning approaches to automatically classify an ad hoc clinical question into general topics. We then explored both unsupervised and supervised methods for automatically extracting keywords from an ad hoc clinical question. Our methods were evaluated on the 4,654 clinical questions maintained by the National Library of Medicine. Our best systems or methods showed F-score of 76% for the task of question-general topic classification and of 58% for extracting keywords from ad hoc clinical questions.

1. Introduction

Physicians have many ad hoc clinical questions at the moment of patient care [1]. And yet, physicians have limited time and resources to search for answers to these questions. Medical question answering systems [2] apply natural language processing approaches to automatically generate answers in response to ad hoc questions; such systems may be promising for clinical information access at point of patient care.

An important step for developing a medical question answering system is to automatically extract information needs from ad hoc questions. Ely and associates [1] collected thousands of clinical questions from more than 100 family doctors, which the National Library of Medicine (NLM) has published with annotations[1] [1,3-5]. Two examples of questions are shown below:

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?"

Here, we represent the information needs by two means. First, a question can be classified by *general topic* to facilitate information retrieval. For example, question 1 represents a *pharmacological* question, and we can therefore identify the Micromedex pharmacological database as the resource for answer extraction.

Secondly, a question incorporates *keywords* (specific topic terms) that capture the most important content of the question. In our example, the keywords for question 1 are "estradiol valerate" and "osteoporosis and cardiovascular disease prevention," while the keywords for question 2 are "pectus carinatum" and "associated congenital heart problems." The keywords can be used as query terms for retrieving relevant documents. They can also be used as the anchor terms for answer extraction.

In this paper, we report our efforts to automatically identify both general topics and keywords from ad hoc clinical questions.

2. Background

Physicians typically asked complex questions. There is a good body of work that manually analyzes complex medical questions with domain knowledge. For example, Ely and associates [1,6] created a taxonomy to map 1,396 clinical questions to a set of 69 question types (e.g., "What is the cause of symptom X?", "What is the dose of drug X?") and 63 topics (e.g., *drug* or *cardiology*). The authors further classified medical questions on the basis of a high-level hierarchical taxonomy, the evidence taxonomy [7]. Specifically, a medical question can be classified as either *clinical* or *non-clinical*. A *clinical* question can be subclassified as either *general* or *patient-specific*; a *general* question can then be classified as either *evidence* or *no evidence*, and an *evidence* question can be classified as either *intervention* or *no intervention*. Seol and associates [8] identified four question types: *treatment*, *diagnosis*, *etiology*, and *prognosis*. Huang and associates [9] manually

[1] http://clinques.nlm.nih.gov/JitSearch.html

evaluated whether clinical questions can be formulated by *problem/population, intervention, comparison*, and *outcome (PICO)*.

Cimino and associates [10] proposed a framework for automatically analyzing ad hoc medical questions. In this framework, an ad hoc question (e.g., "How do I treat Hansen's disease?") can be mapped to a generic question (e.g., "What is treatment for disease?") for which a specific retrieval strategy can be developed. The approach of mapping and decomposing ad hoc questions into generic questions has been widely used in open-domain question answering systems (e.g., [11,12]). Previously, we developed supervised machine-learning techniques to automatically classify medical questions into Ely and associates' evidence taxonomy [7], and we reported ~50% F-score for classifying a clinical question into five categories defined by the evidence taxonomy [13,14].

This new work represents the largest analysis of clinical questions (over 4,600) to date. We classified questions into 12 general topics. We also present the first model to automatically identify keywords from clinical questions.

3. The Clinical Questions Collection

The 4,654 clinical questions maintained by the NLM (available at http://clinques.nlm.nih.gov/About.html) are the evaluation text collection for our study. Those over four thousand questions were collected from healthcare providers across US. Each question was assigned one or more general topics by physicians; there are a total of 13 general topics. For example, question 1 was assigned three general topics *management, treatment & prevention,* and *pharmacological.* Question 2 was assigned two: *management* and *diagnosis.*

Table 1 shows the 13 general topics and the number of questions assigned to each. 3,559 questions were assigned one general topic, 386 questions were assigned two topics, 700 questions were assigned three topics, 4 questions were assigned four topics, and 5 questions were assigned five topics.

In addition, each clinical question was assigned 1–3 keywords: 4,167 questions were assigned one keyword, 471 were assigned two keywords and 14 were assigned three keywords. For question 1, the keywords assigned were "estrogen replacement therapy," "osteoporosis," and "coronary arteriosclerosis;" for question 2, the keywords assigned were "thorax," "abnormalities," and "heart defects, congenital."

Table 1: Of the 4,654 clinical questions maintained by the NLM, the general topics and the number of questions assigned.

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%)
Pharmacological	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%)

4. Method

With 4,654 annotated clinical questions as the training and testing sets, we explored supervised machine-learning approaches to automatically classify a question by general topics. We developed a binary classifier for each of the 12 topics (We excluded the category *unspecified* because it was empty); a binary classifier would allow multiple topics to be assigned to a question. We also developed approaches to automatically extract keywords from ad hoc clinical questions.

4.1 Question Classification

We used the freely available Weka 3 (http://www.cs.waikato.ac.nz/ml/weka/) for supervised machine-learning systems. We experimented with various machine-learning algorithms, including naïve Bayes, decision tree, neural network, and support vector machines (SVMs), and found that SVMs performed the best, which is consistent with our previous studies [13,14]. We therefore only report the results for SVMs.

We explored different features for machine learning, including words and bigrams. We experimented with part-of-speech (POS) and with/without stemming as additional features. We used the Stanford Parser (http://nlp.stanford.edu/downloads/lex-parser.shtml) for the POS tagging. Additionally, we mapped terms in questions to Unified Medical Language System (UMLS) [15] concepts and semantic types and explored them as additional features. The UMLS incorporates the Metathesaurus, a database that contains more than one million biomedical concepts. It also incorporates the Semantic Network, a top-down semantic knowledge representation containing 135 nodes or semantic types (e.g., *disease or syndrome*); each semantic type represents a category to which certain UMLS concepts can be mapped.

Each UMLS concept in the Metathesaurus is assigned one or more semantic types. We have applied the tool MMTx [16] to identify appropriate UMLS concepts and semantic types in a question string. Our previous work showed that adding the UMLS concepts and semantic types as additional features lead to enhanced performance in question classification [13,14]. We also experimented with feature selection using mutual information [17].

As shown in Table 1, the distribution of clinical questions to different topics is skewed, with a large majority of questions assigned to the top three topics. In order to compare the performance of different binary classifiers, we have arranged that each classifier has a baseline of 50%; this indicates that each classifier is trained on the same number of positive and negative data. For example, when we trained a binary classifier for *diagnosis,* we had 994 questions that were assigned to the topic (Table 1). 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. We also repeated the classifications 10 times when we randomly selected negative data. We then report the average F scores. Each F-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

We developed both unsupervised and supervised approaches to automatically identify keywords from each original question and rank them. The two unsupervised approaches we explored are the IDF model and domain filtering. We explored two supervised machine-learning approaches: logistic regression and conditional random fields (CRFs).

4.2.1 Unsupervised Approaches

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 the Inverse Document Frequency (IDF) model that was originally proposed in [18]. Specifically, if a noun phrase incorporates a sequence of words, $W_1\ W_2\ ...W_n$, the IDF value for that noun phrase is the sum of the IDF values of each word:

$$IDF(W_1W_2...W_n) = \sum_{i=1}^{n} IDF(W_i) \qquad (1)$$

The higher the IDF value, the higher priority the keyword. A baseline model is to randomly select noun phrases as keywords.

We speculated that medical, domain-specific terms are more likely to be keywords, and we therefore further enhance keyword identification with the biomedical knowledge resource the UMLS. For this task, we first applied the MMTx to map a question string to the corresponding UMLS concepts and then ranked the concepts based on the IDF model.

The IDF and domain-filtering methods rank the candidate keywords. To determine how many top-ranked keywords, *N*, for inclusion, we used a heuristic formula:

$$N = \begin{cases} \frac{Number of\ words in\ the\ question}{6} + 1 \\ 3,\ if\ the\ number of\ words in\ the\ question is\ more than 12 \end{cases} \qquad (2)$$

4.2.2 Supervised Machine-Learning

We applied two supervised machine-learning methods–logistic regression and conditional random fields (CRFs), to automatically identify keywords in a medical question. Logistic regression model is widely used in the biomedical domain and CRFs has shown to be one of the best models for biomedical named entity recognition. We used Weka 3 for logistic regression. For CRFs, we used Mallet (http://mallet.cs.umass.edu). The features we explored include word position, character length, POS, idf, and semantic types.

4.3 Evaluation of Keyword Identification

The NLM-assigned keywords were used as the gold standard for evaluating our automatic keyword identification approach. However, when we manually examined the NLM keywords, we found we couldn't use them directly. We found in many cases, an assigned keyword does not appear in the original question. For example, the assigned keyword for the question "how do you use redux? what are the indications?" was "obesity." In this case, the assignment was based on the knowledge that "redx" is a medication of obesity. Inferring keywords based on domain knowledge is a challenging text-mining task. We currently focus on the simpler task of keyword extraction only.

Table 2: F scores (10-fold cross validation) for applying support vector machines to automatically assign topics to ad hoc clinical questions. We explored different features and their combinations, e.g., part-of-speech (POS) and the UMLS concepts and semantic types (CSTY).

General Topics	Bag-of-Words	With Stemming	With Feature Selection (top-2000 features)	Words+ Bigrams	Words+ Bigrams +POS	Words+Bigrams +CSTYs+ Stemming	Words+Bigrams+ POS+CSTYs+ Stemming
Device	56.90%	64.96%	62.35%	61.72%	61.07%	73.07%	71.21%
Diagnosis	73.68%	73.75%	75.15%	75.88%	75.23%	76.78%	77.17%
Epidemiology	70.58%	68.47%	65.83%	67.93%	67.96%	72.23%	70.26%
Etiology	79.22%	81.59%	78.22%	82.43%	79.68%	80.38%	82.64%
History	54.29%	59.23%	51.30%	53.76%	57.91%	67.71%	61.70%
Management	68.44%	68.12%	67.96%	71.38%	71.35%	71.11%	70.98%
Pharmacological	82.56%	82.97%	82.90%	84.04%	83.81%	89.25%	88.71%
Physical Finding	71.74%	72.38%	72.72%	71.14%	69.58%	77.75%	76.67%
Procedure	70.39%	71.34%	69.18%	66.57%	65.35%	80.45%	80.32%
Prognosis	72.95%	74.44%	68.39%	69.16%	73.84%	74.25%	74.26%
Test	79.06%	80.62%	78.95%	79.17%	78.18%	83.04%	82.37%
Treatment & Prevention	68.01%	68.82%	69.78%	70.30%	69.55%	71.56%	70.46%
Average	70.65%	72.22%	70.23%	71.12%	71.13%	76.47%	75.56%

We developed a heuristic algorithm to automatically map a term to its variation. Our heuristic algorithm is a simplified version of the one we previously developed for mapping different variations of full-forms [19]. A total of 3,155 questions that have assigned a total of 3,353 associated keywords 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. We also evaluated by Recall, Precision, and F-score. Here 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.

5. Results and Discussion

Table 2 shows the SVMs results for automatically classifying an original clinical question into the general topics. Our results show that *history* was the topic most difficult to identify, while *pharmacological* had the best classification. As was expected, the more questions available for training, the better the performance. Using bag-of-words as feature, the average F score was 70.65%. We explored different feature combinations. Our results show that stemming enhanced the performance (+1.57%), while feature selection slightly decreased the performance (-0.42%). We found that bigrams slightly enhanced the performance (+0.47%), with a slight further enhancement from POS (+0.01%). The UMLS concepts and semantic types had the highest performance increase which is statistically significant (the two-tailed P value is less than 0.0001 comparing to bag-of-words); the combined features led to the highest F score—76.47%. Interestingly, when the features included bag-of-words, bigrams and the UMLS concepts and semantic types, POS decreased the overall performance (-0.91%).

Keyword identification is a challenging task, as we see the baseline (randomly selecting words) had an F-score of only 11.4%. Noun phrases helped to increase the performance to 17.6%—still very poor. We found a significant increase (+12%) in performance when the IDF prioritization was introduced. After domain-filtering, meaning that we only include a term as a keyword if the term can be mapped to the UMLS concept, we obtained the highest performance (53.8%) in unsupervised approaches. Both supervised machine-learning approaches outperformed unsupervised ones. The CRFs outperformed logistic regression, although not statistically significant (t-test). On the other hand, CRFs outperformed the best unsupervised approach UMLS+IDF (t-test, $p<0.01$), increasing 4.2% in F-score.

Table 3: Performance of keyword extraction. "IDF" indicates keyword prioritization with IDF value; otherwise a random selection of keywords is applied. "UMLS concept" indicates we only use the text as keyword if the text can be mapped by MMTx to a UMLS concept. We experimented with selecting all UMLS concepts and then prioritization with IDF value. We also report the results of logistic regression and conditional random fields.

	Precision	Recall	F
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%

Error analysis showed that there are mainly two factors that contributed to the disagreement with the keywords. First, we found inconsistency in keyword assignment. For example, the question "is cow's milk a risk for mad cow disease?" was assigned only the keyword "milk," while our system correctly identified "mad cow disease" as another keyword.

We also found disagreement in keyword composition. For example, the question "what are the clinical signs of neonatal myasthenia gravis?" were assigned two keywords "myasthenia gravis" and "neonatal" while our system extracted only "neonatal myasthenia gravis."

Contributions and Acknowledgement: Hong Yu supervised the research and wrote the paper. Yong-Gang Cao carried out the experiments and implemented the system. We thank James Cimino and John Ely for fruitful discussions. The authors acknowledge the support from the National Library of Medicine to Hong Yu, grant number 1R01LM009836-01A1.

References

1. Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, Evans ER: Analysis of questions asked by family doctors regarding patient care. BMJ 1999;319:358-361.
2. Yu H, Lee M, Kaufman D, Ely J, Osheroff JA, Hripcsak G, Cimino J: Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. J Biomed Inform 2007;40:236-251.
3. Ely JW, Osheroff JA, Ferguson KJ, Chambliss ML, Vinson DC, Moore JL: Lifelong self-directed learning using a computer database of clinical questions. J Fam Pract 1997;45:382-388.
4. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME: Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc 2005;12:217-224.
5. D'Alessandro DM, Kreiter CD, Peterson MW: An evaluation of information-seeking behaviors of general pediatricians. Pediatrics 2004;113:64-69.
6. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, Stavri PZ: A taxonomy of generic clinical questions: classification study. Bmj 2000;321:429-432.
7. 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.
8. Seol YH, Kaufman DR, Mendonca EA, Cimino JJ, Johnson SB: Scenario-based assessment of physicians' information needs. Medinfo 2004;11:306-310.
9. Huang X, Lin J, Demner-Fushman D: Evaluation of PICO as a Knowledge Representation for Clinical Questions. AMIA Annu Symp Proc 2006:359-363.
10. Cimino JJ, Aguirre A, Johnson SB, Peng P: Generic queries for meeting clinical information needs. Bull Med Libr Assoc 1993;81:195-206.
11. Hermjakob U: Parsing and question classification for question answering. In ACL Workshop on Open-Domain Question Answering., 2001.
12. Harabagiu S, Lacatusu F, Hickl A: Answering complex questions with random walk models. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006.
13. Yu H, Sable C: Being Erlang Shen: Identifying answerable questions. In Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasoning for Answering Questions. 2005.
14. Yu H, Sable C, Zhu H: Classifying Medical Questions based on an Evidence Taxonomy. In Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains. 2005.
15. Humphrey B, Lindberg D, Schoolman H, Barnett G: The Unified Medical Language System: An Informatics Research Collaboration. Journal of the American Medical Association 1998;5:1-11.
16. Aronson AR: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21.
17. Yang Y, Pedersen J: A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference (ICML'97). 1997.
18. Pradhan S, Illouz G, Blair-Goldensohn S, Schlaikjer A, Krugler V, Filatova E, Duboue P, Yu H, Passonneau R, Ward W, Hatzivassiloglou V, Jurafsky D, McKeown K, Martin J: Building a foundation system for producing short answers to factual questions. In Eleventh Text Retrieval Conference (TREC-11). Washington, DC, 2002.
19. Yu H, Kim W, Hatzivassiloglou V, Wilbur W: Disambiguating biomedical abbreviations. ACM Transactions on Information Systems (TOIS) 2006;24:380-404.