Journal of Biomedical Informatics 44 (2011) 1032–1038

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain

Feifan Liu [a,*], Lamont D. Antieau [a], Hong Yu [a,b]

[a] *Department of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, United States*
[b] *Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI, United States*

ARTICLE INFO

Article history:
Received 7 March 2010
Accepted 5 August 2011
Available online 12 August 2011

Keywords:
Question classification
Medical question answering
Supervised machine learning
Support vector machines
Natural language processing

ABSTRACT

Objective: Both healthcare professionals and healthcare consumers have information needs that can be met through the use of computers, specifically via medical question answering systems. However, the information needs of both groups are different in terms of literacy levels and technical expertise, and an effective question answering system must be able to account for these differences if it is to formulate the most relevant responses for users from each group. In this paper, we propose that a first step toward answering the queries of different users is automatically classifying questions according to whether they were asked by healthcare professionals or consumers.

Design: We obtained two sets of consumer questions (∼10,000 questions in total) from Yahoo answers. The professional questions consist of two question collections: 4654 point-of-care questions (denoted as PointCare) obtained from interviews of a group of family doctors following patient visits and 5378 questions from physician practices through professional online services (denoted as OnlinePractice). With more than 20,000 questions combined, we developed supervised machine-learning models for automatic classification between consumer questions and professional questions. To evaluate the robustness of our models, we tested the model that was trained on the Consumer–PointCare dataset on the Consumer–OnlinePractice dataset. We evaluated both linguistic features and statistical features and examined how the characteristics in two different types of professional questions (PointCare vs. OnlinePractice) may affect the classification performance. We explored information gain for feature reduction and the back-off linguistic category features.

Results: The 10-fold cross-validation results showed the best F1-measure of 0.936 and 0.946 on Consumer–PointCare and Consumer–OnlinePractice respectively, and the best F1-measure of 0.891 when testing the Consumer–PointCare model on the Consumer–OnlinePractice dataset.

Conclusion: Healthcare consumer questions posted at Yahoo online communities can be reliably classified from professional questions posted by point-of-care clinicians and online physicians. The supervised machine-learning models are robust for this task. Our study will significantly benefit further development in automated consumer question answering.

1. Introduction

The use of computer technology is an integral part of meeting healthcare information needs. While the earliest use of computers to search for medical information was generally performed by healthcare professionals, the Internet has paved the way for laypeople with healthcare concerns such as diseases, symptoms, and treatments to search for information [1–3], and the number of people using the Internet for such information is growing. Tu and Cohen [4] report, for instance, that the number of those searching for health-related information online doubled between 2001 and 2007 to include nearly one-third of all adult Internet users. This growth has not only expanded the healthcare topics that users search for, but also expanded in the myriad sources available to them; recently, for instance, there has been a boom in healthcare-related social networking [5], particularly among question–answering sites [6]. These trends suggest that healthcare consumers will become increasingly dependent on Internet resources for answers to their medical questions in the years ahead.

In light of these developments, this paper reports on efforts to improve the medical question–answering system, *AskHERMES* (http://www.askhermes.org/). AskHERMES (Help clinicians Extract and aRticulate Multimedia information from literature to answer their ad hoc clinical quEStions) is a fully automated system that uses natural language processing tools to retrieve, extract, analyze, and integrate information from medical literature and other

[*] Corresponding author. Address: 2400 E Hartford Ave., Room 953, Milwaukee, WI 53211, United States. Fax: +1 414 229 2619.
E-mail address: liuf@uwm.edu (F. Liu).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.08.008

information resources to provide answers in response to questions posed by healthcare professionals, e.g. physicians. To date, much of the research that has been done to benefit the system has focused on extracting information needs from the complex questions that physicians often ask [7] and to present responses to such questions in the most effective way [8]. More recently, the use of automatic speech recognition tools that enable physicians to pose questions to the system as spoken rather than typewritten input has been explored [9]. The research direction presented in the current paper, however, reflects the belief that AskHERMES could also benefit healthcare consumers, but only after the development of several subcomponents will it accurately interpret the information needs of lay users (i.e. healthcare consumers), quickly find information appropriate to their literacy level and technical expertise (or lack thereof), and then summarize the information that is retrieved and present it in the most useful way.

As a first step in this process, we investigated ways of determining whether users are healthcare consumers or professionals based on quantitative linguistic differences between the questions asked by members of both groups. Although the literature on differences in communication between physicians and patients acknowledges that questions are a significant component of medical encounters [10–12], studies in this area have generally focused on questions that patients ask healthcare professionals [10–14], questions posed by healthcare professionals [13], or on the more general aspects of linguistic interaction between the two groups [15]. With respect to the literature on using computers to search for medical-related information, studies have also tended to investigate either the queries of healthcare consumers [16] or those of professionals [17] without taking into account the questions of both groups. A more rigorous, comparative analysis of these questions might reveal stylistic differences that could enable us to better meet the information needs of members of both groups.

With these aims in mind, we developed a supervised machine-learning framework to automatically distinguish the questions of healthcare consumers and professionals. Although an exploratory study of the differences between the information-seeking behaviors of the two groups revealed significant differences at every level of the grammar, we primarily focused on the shallow-level linguistic features (e.g. bag-of-words features) without deep language processing (e.g. syntactic parsing), as previous work determined that words are adequate representational units for the purposes of classification [18]. We found machine-learning approaches suitable for classifying questions by whether the question was posed by consumers or professionals. In addition to the success of a bag-of-words approach for classification, we experimented with statistical features and linguistic category features to improve the robustness of the classifiers.

2. Related work

Many studies in question classification have focused on the semantics of questions and their potential answers, and to that end, they have investigated the use of taxonomies in question classification both in the open domain [19,20] and in the medical domain [21]. Some systems have explored the use of syntactic features for classification but have generally done so as a supplement to semantics rather than as a replacement [22–24]. Other studies have identified additional dimensions that could be useful for question classification, for instance, the distinction between factual and analytical questions [25,26], factual and opinion questions [27], objective and subjective questions [28,29], and answerable and unanswerable questions [30]. We propose that the ability to distinguish between the questions asked by consumers and professionals could be a dimension worth exploring, in our case, because of its potential to tailor information retrieval and question answering systems for different users.

Different linguistic features and feature selection methods have been studied in previous work. In the area of corpus linguistics, studies focusing on readability [31–33] have explored word length, word frequency, and sentence length to determine linguistic complexity and genre. The information gain based feature selection has shown to be helpful for text and evidence classification [34,35]. Motivated by those prior works, we evaluated both linguistic features and statistical features on our task, and the proposed linguistic category features which are expected to capture language usage differences on a higher level between healthcare professionals and consumers, thereby eluding the data sparseness problem resulting from "bag of words" features.

3. Material and methods

We first discuss the collection of our data and provide a brief characterization before describing the machine-learning methods that were used for question classification.

3.1. Data

We used four representative datasets in our study: two sets of consumer questions and two sets of professional questions, as described below.

1. Consumer questions I (Consumer-I):

We downloaded 5013 consumer questions posted on Yahoo Answers between May and June 2009 (http://answers.yahoo.com, category "Health/Diseases and Conditions").

2. Consumer questions II (Consumer-II):

We reused 5499 consumer questions, which is a subset extracted from a previous study http://ir.mathcs.emory.edu/shared/. Questions in this subset were posted in the "Health/Diseases and Conditions" category on Yahoo Answers between Nov. 2007 and Jan. 2008.

3. Point-of-care clinical questions (PointCare):

A set of 4654 professional questions collected by physicians from interviews of family doctors following patient visits [13,36].

4. Online questions among physician practices (OnlinePractice):

ParkHurstExchange (http://www.parkhurstexchange.com) is an online publishing service based in Montreal, Canada, which provides credible and highly respected publications of physician practice questions and answers from healthcare professionals. All questions posted by physicians are selected and answered by professional members, which are further reviewed and approved by the Medical Editor-in-Chief. Through this service, physicians can ask their own questions, browse questions in different specialties and search them by keywords. We downloaded 5378 professional questions from the ParkHurstExchange website as of December 6, 2010.

Although all four collections of questions described above were not intended for an automated question–answering system, there are several benefits of using these question collections: (1) they are relatively large collections of questions in which each question can be attributed to a consumer or a professional with a high degree of certainty and thus are amenable to supervised