Journal of Biomedical Informatics 45 (2012) 292–306

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

An ontology for clinical questions about the contents of patient notes

Jon Patrick *, Min Li

Health Information Technology Research Laboratory, School of IT, Faculty of Engineering and IT, The University of Sydney, Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 16 July 2011
Accepted 17 November 2011
Available online 28 November 2011

Keywords:
Classification
Electronic health records
Natural language processing

ABSTRACT

Objective: Many studies have been completed on question classification in the open domain, however only limited work focuses on the medical domain. As well, to the best of our knowledge, most of these medical question classifications were designed for literature based question and answering systems. This paper focuses on a new direction, which is to design a novel question processing and classification model for answering clinical questions applied to electronic patient notes.
Methods: There are four main steps in the work. Firstly, a relatively large set of clinical questions was collected from staff in an Intensive Care Unit. Then, a clinical question taxonomy was designed for question and answering purposes. Subsequently an annotation guideline was created and used to annotate the question set. Finally, a multilayer classification model was built to classify the clinical questions.
Results: Through the initial classification experiments, we realized that the general features cannot contribute to high performance of a minimum classifier (a small data set with multiple classes). Thus, an automatic knowledge discovery and knowledge reuse process was designed to boost the performance by extracting and expanding the specific features of the questions. In the evaluation, the results show around 90% accuracy can be achieved in the answerable subclass classification and generic question templates classification. On the other hand, the machine learning method does not perform well at identifying the category of unanswerable questions, due to the asymmetric distribution.
Conclusions: In this paper, a comprehensive study on clinical questions has been completed. A major outcome of this work is the multilayer classification model. It serves as a major component of a patient records based clinical question and answering system as our studies continue. As well, the question collections can be reused by the research community to improve the efficiency of their own question and answering systems.

1. Introduction

The large amount of information available in electronic patient records make it an attractive resource for answering a variety of questions that users may have. Current information retrieval (IR) techniques have proven quite successful at locating patient records that might be relevant to a user's query [1]. However, in the present scenario, the set of retrieved documents represents an answer size that is still too large to identify the best matches readily. These solutions leave the user with a relatively large amount of text to review. This phenomenon requires a more efficient retrieval technique to retrieve only the part of the document which is relevant. The usefulness of a solution to this problem can be seen from a previous study that reported an average of six medical questions was asked by family doctors in a half day practice [2].

Question Answering (QA) technology for clinical needs relies on pinpointing, relevant matches so small as to be just answer-sized according to the semantic interpretation of the question. Consequently, this technology can help doctors to use limited time to browse the retrieved information and improve their productivity and efficiency thus contributing to patient quality and safety. Research in the task of QA has recently become one of the fastest growing topics in computational linguistics, especially since the launch of the QA track at the Text REtrieval Conference (TREC) in 1999 [3]. One of the essential components in QA is question classification, which can not only indicate the possible answers but also suggests different processing strategies. Since the contribution of TREC, the open domain question classification work has been intensively explored. For example: an hierarchical classifier was designed by Li and Roth [4], which is based on the SNoW (Sparse Network of Winnows) [5] learning approach. In their work, the corpus consisted of 5500 training and 500 test questions compiled from four main sources: the 4500 English questions published by USC [6], 849 TREC 8 [7] and TREC 9 [8] questions, and 500 TREC 10 [9] questions, as well, a sequence of two classifiers was adopted to classify questions into six coarse classes and fifty fine classes (see Table 1). The accuracy of 91.0% for the coarse grained classes and 84.2% for the fine grained classes was achieved by using lexical items, part of speech tags, chunks (non-overlapping parses), named

* Corresponding author. Fax: +61 2 9351 3838.
E-mail address: jonpat@it.usyd.edu.au (J. Patrick).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.11.008

Table 1
Taxonomy defined by Li and Roth.

Abbreviation	Entity		Description	Human	Location	Numeric
Abbreviation	Animal	Other	Definition	Group	City	Code
Expression	Body	Plant	Description	Individual	Country	Date
	Color	Product	Manner	Title	Mountain	Distance
	Creative	Religion	Reason	Description	Other	Money
	Currency	Sport			State	Order
	Disease/medicine	Substance				Other
	Event	Symbol				Period
	Food	Technique				Percent
	Instrument	Tern				Speed
	Language	Vehicle				Temp
	Letter	Word				Size
						Weight

entities, head chunks (the first noun chunk in a sentence) and semantically related words (words that often occur with a specific question class) as learning features.

Many follow-up research studies are based on this dataset. Linear support vector machines (SVMs) [10] have been proven as an optimized learning algorithm in Zhang and Lee's studies [11] by considering the surface text features of questions. An accuracy of 87.4% was obtained in the coarse grained classification by applying bag-of-ngrams (all continuous word sequences in the question). As well, the bag-of-words model achieved 80.2% accuracy in the fine grained classification. Furthermore, the SVM tree kernel was designed for the coarse grained classification which enabled the SVM to gain the benefit of syntactic structures. By applying a tree kernel, the performance of the coarse grained classification was increased by 2.6%. Later the tree kernel was further studied by Moschitti and his colleges [12,13]. This time, the accuracy of the coarse grained classification reached 91.8% by applying bag-of-words and parser tree, which is slightly higher than Zhang and Lee's tree kernel model.

Apart from the syntactic information, semantic knowledge has also been investigated. In Li and Roth's later work [14], the combination of the semantic features, such as named entities, class-specific related words and distributional similarity based categories, as well as the syntactic features (word, part of speech tags, chunks and head chunk) gave 89.3% accuracy for the fine grained classification by using 21,500 training and 100 test questions. Most recently, the WordNet [15] knowledge resource was integrated into question classifiers [16,17]. In this study, the WordNet hypernyms of the head word (one single word specifies the object the question seeks), as well as the head word, unigram, word shape, and wh-word were considered in the learning feature. By using SVM, the best accuracy for coarse grained classes is 93.4%. As well, an accuracy of 89.2% was obtained in the fine grained classification.

Unlike the various investigations in the open domain, only a few works have been completed in the medical domain which have attempted to describe the information needs of clinicians. For example, the information needs while using a Clinical Information System (CIS) were classified according to event, resource, outcome, and context type by Currie et al. [18]. One of the major outcomes of this study is that, the 'Subject' event type (seeking data about the patient) was the most commonly occurring type of information need. This category has not been analysed in any further studies, while our aim is to address this problem. Another comprehensive clinical question study was carried out by Ely et al. [19–22]. Their observation concentrated on the questions about medical knowledge bases, which could be potentially answered by external resources, such as medical science articles and textbooks. During the observation, thousands of medical questions were collected from around one hundred family doctors, such as 'What is the dose of atorvastatin?', 'Does Zoloft cause stomach upset?' and 'How should I treat his epididymitis?'. Ultimately, three types of taxonomies were created in this work based on: topic, generalization and obstacle.

In Ely et al.'s topic taxonomy, approximately sixty topics were designed based on specialties, such as 'drug prescribing', 'obstetrics' and 'gynaecology', which was adapted from a family practice article filing system [23]. For the generic question template, an iterative annotation process was used to develop this taxonomy, which involved 69 generic templates. Questions with essentially identical structures were classified as one generic type. The three most common generic templates were: 'what is the drug of choice for condition x', 'what is the cause of symptom x' and 'what test is indicated in situation x?'. These two taxonomies can be used to guide which knowledge systematically fails to address specific types of questions, as well as which keywords can be used to link questions to answers.

The evidence taxonomy which is based on the obstacles that were created for asking and answering a question, such as knowledge gap reorganization, question formulation, information retrieval and answer generation were used to annotate two hundred questions (see Fig. 1). The taxonomy is a simple hierarchy with just five leaves. On the leaf level, the sub-classes are 'Non-clinical' (the questions do not belong to the medical domain), 'Specific' (the questions require the information from patient records) and 'No evidence' (the answer to the question is unknown). These questions were classified as not answerable by using medical textbooks or literature. In contrast, the evidence questions are potentially answerable with evidence. Two classification studies [24,25] were performed by Yu and colleagues based on this evidence taxonomy to produce a system called AskHermes. The first study identified answerable questions by using two hundred annotated questions as their corpus. In this work, a few simple features were evaluated by several different machine learning algorithms, such as bag of word, UMLS [26] concept and UMSL semantic types. As well, the best performance (80.5% accuracy) was obtained by adopting bag of word and UMLS concepts as learning features in probabilistic indexing [27].

Later, a similar methodology was used to assign labels to questions based on the taxonomy. Moreover, two different approaches were investigated, which were the ladder approach and the flat approach. In the ladder approach, a set of binary classifiers was considered. For example, a question is first predicated as 'Clinical' or 'Non-Clinical'. If it is a clinical question, a second classifier was applied to classify it into 'General' or 'Specific'. If it is a general question, it will be identified as 'Evidence' or 'No evidence'. Finally, the evidence question will be classified as 'Intervention' or 'No intervention'. On the other hand, a multi-label classifier was trained to assign one of the leaf labels to questions. The results show the