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

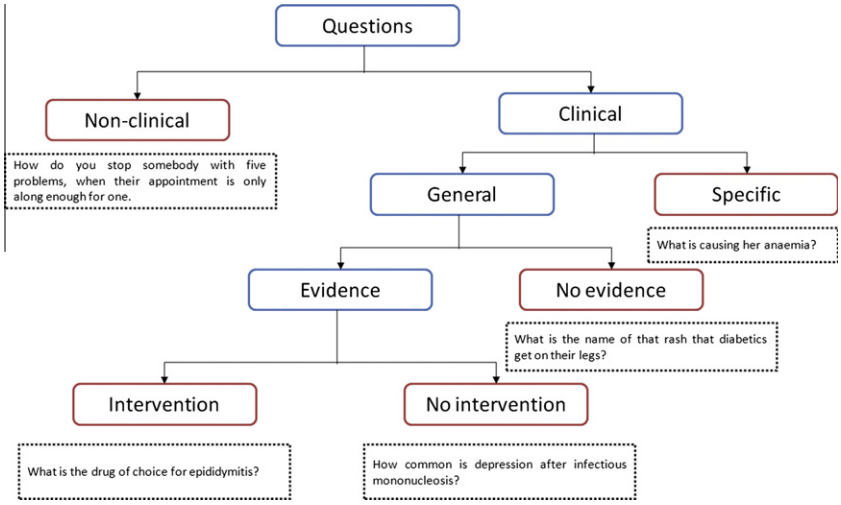


Fig. 1. Evidence taxonomy designed by Ely and colleagues.

SVM ladder approach is the best classifier in both methods, which achieved nearly 60% overall accuracy by adopting bag of words, UMLS concepts and semantic types as learning features.

The initial question collection was extended to nearly 4700 questions which were collected in the later studies [28–30]. This question collection was maintained by the National Library of Medicine and assigned with 12 general topics. Most recently, the automatic question topic assignment was investigated [31,32] by exploring a supervised machine learning approach (SVM) on this collection. Since one question can be assigned to multiple topics, a binary SVM classifier was adopted for each of the 12 topics by exploring some simple learning feature sets. Finally, an *F*-score of 76.5% was obtained by using bag of words, bigrams, UMLS concepts and semantic types as learning features.

In the open domain, the question studies are mainly based on the answer type classification since the answer types were specified by TREC, such as whether the question is asking for a person's name, date and location. The generic question template is not quite realistic for the open domain to generalize the question set, since the question set is too large to design such a taxonomy. Moreover, the answer can be found by using the answer type and grammatical relationship between the question and answer. Consequently, the generic question template is not quite helpful to the open domain. As well the work (generic question template assignment) reported here is rarely involved in the open domain. However, it is important to stress that generic question template assignment does play an essential role in medical QA systems, especially those systems that focus on some particular question templates [33,34]. Since only a medical QA framework was described by Athenikos et al. [34], no detailed computational classifier was designed. Furthermore, in Terol's work [33] a pattern based approach was adopted, but this is worrisome because firstly, the reliability of two training pattern generation approaches (manual and supervised) has not been discussed; secondly, the final evaluation result was doubtful because the size of the training pattern was unspecified, meanwhile the test set was not balanced as it contains 50 manually created questions based on the ten most frequent medical question types and 200 open domain questions. We argue that the appearance of medical terms in the open domain questions is less frequent than in medical questions, and the predicate structures in the ten most frequent medical question templates are less commonly present in the open domain questions. Hence Terol's strategy is less representative of the real world problem than it needs to be to yield meaningful results.

Compared to these sparse works on the medical knowledge based questions, the studies on the patient records based questions are even smaller. Before the 2010 publication of Neilsen et al. [35], medical QA researchers [24,36–38] focused exclusively on the literature based clinical question since physicians are urged to the use of the best evidence from scientific and medical research when faced with questions about how to care for their patients [39–42]. However, we believe the previous clinical diagnoses or findings for a patient also makes a significant contribution to the current clinical decision that an attending physician needs to make. If this information can be rapidly and reliably obtained by simply asking a question, on average a better clinical decision can be made. The importance of the patient records based QA system has been discussed in [35], but only an architecture of a multi-source (patient records and online biomedical resources) clinical QA system has been described, called MiPACQ. Unfortunately, no practical study of the patient records based questions has been published. In the following sections, we will introduce our investigation into issues such as question collection, question categorization, and question annotation, aimed at identifying the best strategies for these tasks, so as to provide a comprehensive study of questions of patient notes. As well, two elaborate classification models will be described which were designed by exploring rich learning features, such as the open domain and medical domain knowledge resources.

2. Materials and methods

As the specific questions of either patients or clinicians have not been explored in depth in the existing studies (Currie et al. and Ely et al.), we need to take steps to help fill this knowledge gap by designing an ontology for clinical questions appropriate to the contents of patient notes. Our methodology follows the previous studies, however, it is more systematic as to question collection, question analysis and question classification.

2.1. Question collection

This study was conducted at the Intensive Care Unit (ICU) of the Royal Prince Alfred Hospital (RPAH), Sydney, Australia. Three main methods were applied to collect questions, namely face to face interviews, ICU visits, and a web based question collection form (integrated into the Intelligent Clinical Notes System (ICNS) [1]).

ICNS has served the ICU for 3 years, and was built to intensivists requirements so that they could selectively retrieve patient notes. Various search methods were implemented, e.g. keywords, concepts, the name of clinician, the medical registration number of the patient, the timestamp of the record, the admission time, etc. while extraction of useful information included SNOMED CT concepts, scores and measures, the corrections for misspelt words, the expansion of shorthand words and acronyms, etc. In this work, the ICNS provides a web based collection form for recording questions to assist the intensivists to identify their information needs in a given context. The information needs in this task are defined as, but not limited to, sentences with interrogative words, e.g. 'yes/no' questions, 'wh-' questions, and 'how' questions. At interviews if a statement, as distinct to a question, attempted to obtain information as part of the conversation, it was converted into a question. In this way a large extension to the initial range of questions was achieved.

In order to provide a good understanding of our work to all specialists in the ICU, at first, a meeting was held to introduce the idea of a clinical QA system in a review the ICNS. While the current retrieval system (ICNS) does not support QA, the main reason put forward for involving the ICNS in the observational studies was that it could help the clinicians to be more explicit about what information they could obtain from the patient records by simply asking questions of the system. As a result of this representation, all clinical specialists present at the meeting were willing to participate in this project, e.g. face to face interview, ICU visits, and online questionnaire.

The interviewees were invited to participate in the question gathering by the ICU administration 1 week prior the event on each of the 2 weeks the study was conducted. The interview method and objectives were introduced during the invitation. Also, three clinical specialists volunteers were used, a nutritionist, physiotherapy, and research coordinator. The interviews were conducted by two computational linguists while interviewees were operating the ICNS, for approximately 30 min for each specialist. The interview was semi-structured: the first 5 min was spent on introducing the task. Then, one computational linguist (the first author) communicated with the interviewee and asked questions according to the response of the interviewee, like 'What are you normally searching for?', 'What do you want to know about your patient?', 'How do you find the answer by performing a search?', 'Based on the search result, is there any other information you want to know?', 'What kind of information cannot be found from the patient notes by using ICNS', etc. At the same time, the second researcher was doing the note taking. Approximately 100 questions were collected from the specialists.

The majority of our questions, nearly 550 questions, were collected from the ICU visits. Similar to the specialist interviews, the clinicians who were to be involved in the ICU visit were informed 1 week prior to the visit. Each visit was conducted by two computational linguists without interrupting clinicians and occurred during the daily doctor/nursing handover, as well as the daily meeting, from which the clinical questions among clinicians were collected by note taking. During the handover, the clinicians went through every bed, and each patient status was introduced to the off duty clinicians. Meanwhile, a large number of questions were proposed by the new on-duty clinicians to gain a good understanding of each patient. Eight senior doctors and several junior doctors, as well some nurses were involved in the ICU visits which took place twice daily over 3 weeks.

Due to the busy practice environment, the questionnaire was not systematically organized. The questionnaire was available in the ICNS so that the clinicians can fill it in at anytime when they are free. Once it was submitted, it was stored into a database. Finally only a few questions were collected through the web based question collection form. The form was submitted by two specialists (one cardiologist while the other specialist's background not provided), and it contributed around twenty questions.

2.2. Question analysis

Based on the observational data, a comprehensive question study was conducted by developing the design of taxonomic and generic templates.

The main objective of this taxonomy is to contribute to the question processing engine in the clinical QA system by categorizing questions according to the answering requirements. In other words, each question class represents a unique answering strategy, which should be predicted in the question processing engine, and then this method can be applied in the subsequent processing. This taxonomy consists of three coarse grained classes and eleven fine grained classes in Fig. 2. The detailed descriptions for each class, as well as more examples are presented in Appendix A.

Four main steps were involved in this analysis, namely taxonomy design, taxonomy revision, dual annotation, and the generation of a gold standard annotation. The initial version of the taxonomy (taxonomy annotation guideline) was designed by a computational linguist (the second author) after the question collection was reviewed. The principle of categorization is based on the answer strategy, e.g. whether the answer can be retrieved directly from the patient notes, whether the answer can be found by database value retrieval, whether the answer needs a sophisticated statistical inference model or deductive reasoning model by analyzing multiple data, etc. Subsequently, a meeting was organized to revise this draft which involved six computational linguists and one doctor. The final decision was made by the most senior of the computational linguists and the doctor. Once the taxonomy was finalized, dual annotation was done on the whole collection by another computational linguist and the doctor. Finally, the gold standard was generated by unifying these two annotation versions together, for example, if a question was assigned the same class by two annotators, then this annotation was accepted, otherwise, a discussion was held between the first annotator, the second annotator and the doctor to make the final decision. Subsequently, 100% annotation agreement was achieved (the annotation statistics are presented in Table 2 in Section 4).

The generic question templates were defined to convert the questions into a set of canonical forms by which nearly four hundred questions were generalized into 17 templates. Thus, the burden of deep language processing was reduced significantly by building answering methods for these limited templates rather than all instances of each template. Furthermore, the element and answer type of each template can be used to contribute to the document retrieval engine and the answer extraction engine.

Similar to the iterative process in the design of the question taxonomy, a three step analysis was adopted for designing the question templates. In the beginning, the draft version of the question generic templates was created by placing the questions with essentially identical structures ('Did he have a pupil dilatory response?', 'Did the patient have a picc line yesterday?', 'Does he have inner spleen lacerations?') into a single class ('Did the patient (have) X, T?') (X stands for the retrieval expression, and T stands for the time expression), which was done by a computational linguist. Subsequently, the templates and annotations were revised by another computation linguist and a doctor. Finally, the outcome of this iterative process was reviewed and finalized by these three annotators to achieve 100% annotation agreement (the annotation statistics are presented in Table 2).

The three most frequent templates are: 'Did the patient (have) X, T?' (122 questions, 31.6% of the 'Specific note' question), 'What was the value of X, T?' (99 questions, 25.6% of the 'Specific note'

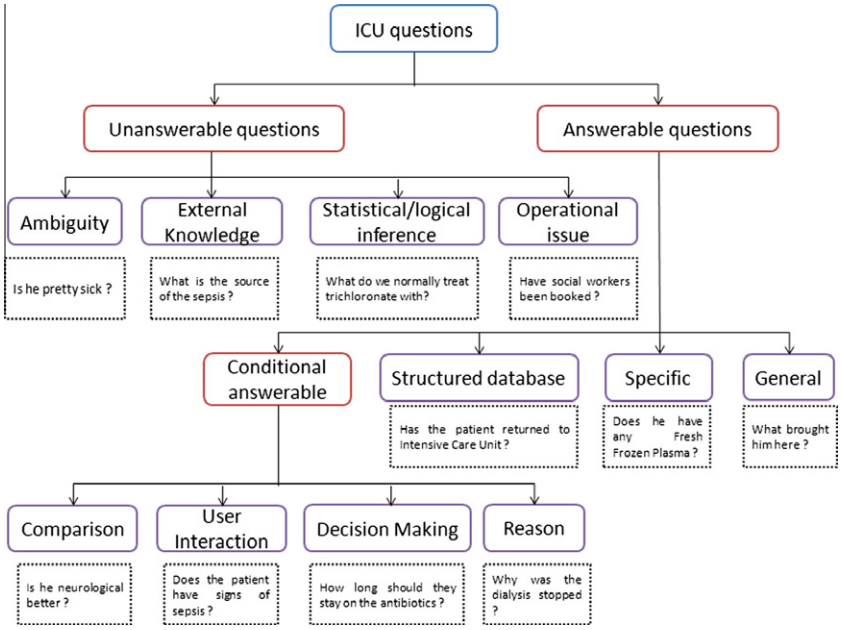


Fig. 2. A question taxonomy based on answering strategies.

question), and 'What is the description of X, T' (74 questions, 19.2% of the 'Specific note' question).

2.3. Multilayer question classification

A machine learning approach (SVM) was adopted as the standardized classification approach to automatically classify a question according to the question taxonomy and the generic question templates. As well, two fold cross validation was chosen as the evaluation mechanism by calculating the overall accuracy and individual *F*-score. Learning feature selection is one of the most crucial issues to impact the performance of a machine learner. The features should be general enough to support the variation of different questions that belong to one category, and strict enough to capture the differences between questions from different categories. The feature sets which were chosen here involve five major feature sets which can be extended to create eleven feature sets if necessary:

I. Unigram: each token in a question. This feature was extended to a new feature set, called 'Lemmatized unigram' in which each token was converted into its lemma form by using GENIA tagger [43].
II. Bigram: a group of two continuing tokens in a question. Similar to the 'Lemmatized unigram', 'Lemmatized bigram' is an extended version of bigram.
III. Interrogative word: the first token or first two tokens in a question which commonly represent the answer type.
IV. SNOMED [44] category: the SNOMED top category of each medical term in the question. The top category is generated by a "Text to SNOMED CT" (TTSCT) conversion process [45] and indicates the medical category of each question. Since the SNOMED concept exists in most of the questions, the SNOMED-presence feature was not considered here.
V. Predicate argument structure (PAS): a predicate-argument relation between two words, which can be used to convey the meaning of a question. Two types of PASs were considered, namely the verb and its subject (ARG1), and the verb and its object (ARG2), which were generated by the Enju parser [46]. Taking a question 'Does the patient need an X-ray?' as an example, three predicate argument structures can be found: ① ARG1, need, patient. ② ARG2, need, X-ray. ③ ARG2, does, need. Furthermore, this feature set was extended to another three feature sets by involving generalization and semantization:
   (i) Lemmatized PAS: the verbs were converted into its lemma form, as well the pronouns were changed to a canonical form, such as 'ARG1, was, patient' → 'ARG1, be, PT'.
   (ii) Lemmatized PAS with SNOMED category: based on the lemmatized PAS, the noun is replaced by its SNOMED category, such as 'ARG2, need, x-ray' → 'ARG2, need, Procedure', 'ARG2, need, Physical force'.
   (iii) Lemmatized PAS with SNOMED-presence: based on the PAS, if the noun is identified as a SNOMED concept, then it would be replaced by 'SCT', such as 'ARG2, need, x-ray' → 'ARG2, need, SCT'.

The purpose of this feature set is to generalize the predicate argument structure feature (PAS). The medical term can belong to multiple categories, as well its semantic meaning can be represented by different concept IDs in other categories. This polysemy breaks the strength of the statistical distribution of a feature weakening its ability to assist in the classification. Moreover, the PAS with SNOMED concept ID is more specific than the PAS with SNOMED category, which makes the feature set too diverse. Since our question collection is not large, a generalized feature set should be more helpful. Thus, the PAS feature was generalized in this order: original PAS → lemmatized PAS → lemmatized PAS with SNOMED category → lemmatized PAS with SNOMED-presence.

Besides, this general learning model, task dependent strategies were designed to improve the three subtasks below.

2.3.1. Unanswerable question filter

The unbalanced distribution (37 vs. 588) between answerable questions and unanswerable questions reflected that the machine learning approach may not be optimal for the 'unanswerable question filter'. Thus, a rule based approach was investigated to overcome this issue. A maximum of five generalization levels was used in the rules to capture the unanswerable questions, such as

Table 2
Annotation statistics.[a]

Class	ICU questions
Frequency	595

Class	Unanswerable	Answerable
Kappa	0.872	0.872
Frequency	37	558

	'Unanswerable' subclasses				'Answerable' subclasses			
	Ambiguity	External knowledge	Operational	Statistical/logical inference	Conditional answerable	Structured database	Specific	General
Kappa	0.328	0.855	0.956	1.000	0.983	1.000	0.973	1.000
Frequency	9	11	13	4	126	17	386	29

	'Conditional answerable' subclasses			
	Comparison	Decision making	User interaction	Reason
Kappa	0.988	1.000	0.898	1.000
Frequency	46	46	15	19

Templates for 'Specific' subclass

Template (X: Retrieval expression, T: Time expression)		Frequency
1.	Did the patient (have) X, T?	122
2.	Does the patient's X have (been) Z? * Z: Reference Constraint	10
3.	How often did the patient (have) X?	3
4.	What has grown in X, T?	3
5.	What is the color of X, T?	3
6.	What is the trend of X, T?	7
7.	What was the description of X, T?	75
8.	What was the treatment for X, T?	2
9.	What was the value of X, T?	98
10.	What was the value of X, T? X vs. W? * W: Reference value	5
11.	When was the last time for the patient (having) X?	21
12.	When was the last time for the patient (having) X? And how long?	7
13.	Where is the location of X, T?	6
14.	Who was/has X, T? And how many of them?	19
15.	Who was the patient's X?	3
16.	Other	2

[a] Please see Appendix A for the definition of each question class, and Appendix B for more examples for different templates.

① Using the lemmatized the token. ② Normalizing the pronoun. ③ Using the synonym and antonym. ④ Using the SNOMED category to replace the medical terminology. ⑤ Using the predicate argument. For example:

Question 1: Will the parents be having a family conference soon?
Question 2: Will the family have a meeting?
PAS1 = "(V_ARG1, having, parents), (V_ARG2, having, conference), (A_ARG1, be, parents), (A_ARG2, be, having), (A_ARG1, will, parents), (A_ARG2, will, having)".
PAS2 = "(A_ARG1, will, family), (V_ARG2, have, meeting)".
Lemmatized PAS 1 = "(V_ARG1, have, parent), (V_ARG2, have, conference), (A_ARG1, be, parent), (A_ARG2, be, have), (A_ARG1, will, parent), (A_ARG2, will, have)".
Lemmatized PAS 2 = "(A_ARG1, will, family), (V_ARG2, have, meeting)".
Lemmatized PAS 1 with SNOMED category = "(V_ARG1, have, Social context), (V_ARG2, have, conference), (A_ARG1, be, Social context), (A_ARG2, be, have), (A_ARG1, will, Social context), (A_ARG2, will, have)".
Lemmatized PAS 2 with SNOMED category = "(A_ARG1, will, Social context), (V_ARG2, have, meeting)".
Rule = r'A_ARG1, will, Social context.

In the end, 29 rules were crafted to capture the 37 unanswerable questions.

2.3.2. Answerable question classification

As a limited similarity can be discovered among the small class in answerable questions (e.g. 'Comparison', 'Decision Making', 'Structured', 'General', 'Reason') by exploring the above surface level features, a specific feature set was investigated to extend indicative elements by exploring the synonym and antonyms in WordNet. For example, the 'Comparison' question can be easily indicated by the adjective or adverb, like 'Was this patient negative to this treatment?', 'Did they get any worse last night?', etc. On the other hand, the 'Decision Making' question can also be easily identified by the subject and its verb, such as 'Did the patient need Morphine?' and 'Are you going to preemptively correct the International Normalization Ratio?'. The noun phrase in the 'General' question also represents its class, e.g. 'What is the history of the patient?', 'What is wrong with the patient?'. Similarly, the subject and its verb in the 'Structured' question also differ from the other classes, such as 'When did she arrive' and 'Did any new patients get admitted last night?'. Thus, when the SVM was learning the training set, these indicative elements were automatically extracted from the training set and used for building the learning model. Meanwhile these elements were extended by using the synonym and antonyms in the WordNet. When the test set was used, this new resource was adopted as a feature set to assist in the predicate generalization. This step thereby is an automation of the knowledge discovery and knowledge reuse (KD–KR) process that derives the workflow in our works.

2.3.3. Generic templates classification

As an alternative, both the answer type and the subject of a question were considered in the design of the generic templates. For instance, if an answer requests a feature of polarity, then its question goes to the 'Yes/No' class (template started with 'Do/Did'), if an answer requests the commentary about something, then its question goes to the 'What' class., etc. Next, these classes were subdivided into more detailed templates, e.g. for the questions that belong to the 'Yes/No' template, if its subject is a patient, then it goes to template 1 (Did the patient (have) X, T?), if its subject is something related to a patient, then it goes to template 2 (Does the patient's X have (been) Z?), for the 'what' class, if the question asks for the commentary of observational entities it goes to templates 6, 9 and 10. Hence, the question subject could be a beneficial feature for the template classification task. Similar to WordNet feature generation in the answerable question classification, the automatic KD–KR process was also used to improve the performance of the template classifier by discovering the subject of the training questions and reusing them for the learning and prediction functions. It is worth pointing out that the observational entity was extended by SNOMED CT concept coding and our in-house measurement list.

3. Theory

The multilayer classification model was created to fit into a clinical QA system, which is displayed in Fig. 3.

In the first instance, questions are sent to a pre-processing engine for proof reading which involves white space tokenization, abbreviation expansion, acronym expansion and misspelling correction. After that, an 'unanswerable question filter' which is a binary classifier is used to filter the unanswerable questions from the set of proofed questions. Next, the answerable questions are processed by an 'answerable question taxonomy classifier' (multiclass classifier) to separate them into seven subclasses of: 'Comparison', 'Decision making', 'User interaction', 'Reason', 'Structured database', 'Specific note', and 'General note'. Finally, the 'Specific note' question is classified into one of the generic question templates which will assist the document retrieval engine and answer extraction engine. The development of the 'unanswerable question filter', 'answerable question taxonomy classifier', and 'generic question template classifier' was introduced in the previous section.

3.1. Evaluation metrics

3.1.1. Annotation agreement

In order to evaluate reliability of the dual annotation (taxonomy annotation), Cohen's Kappa [47,48] was used to measure the agreement between the two annotators. Kappa is calculated by the formulation of $\kappa = \frac{Pr(o)-Pr(e)}{1-Pr(e)}$, in which $Pr(o)$ is the observed annotator agreement, and $Pr(e)$ is the expected annotator agreement by chance.

3.1.2. Question classification

Two fold cross validation was chosen as the evaluation mechanism by calculating the precision, recall and the *F*-score for individual classes, as well as the overall accuracy of each classifier by using the following formulas:

1. *precision = true positive/(true positive + false positive)*,
2. *recall = true positive/(true positive + false negative)*,
3. *F-score = 2 ∗ (precision ∗ recall)/(precision + recall)*,
4. *overall accuracy = correctly classified instances/total instances.*

4. Results

4.1. Annotation reliability and statistics

The statistics of the question taxonomy and generic question templates are presented in Table 2. There were two templates which contained only one question each. In order to perform two fold cross validation by computational methods, these two templates were combined as one single class, called 'Other'.

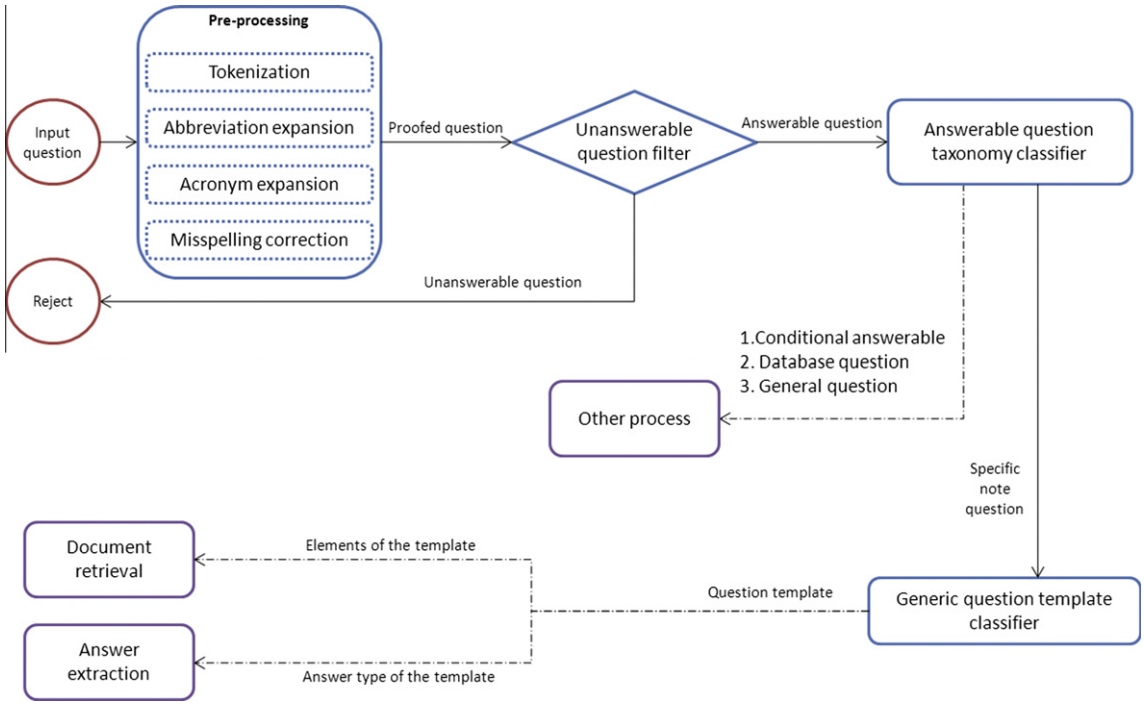


Fig. 3. Multilayer question classification model.

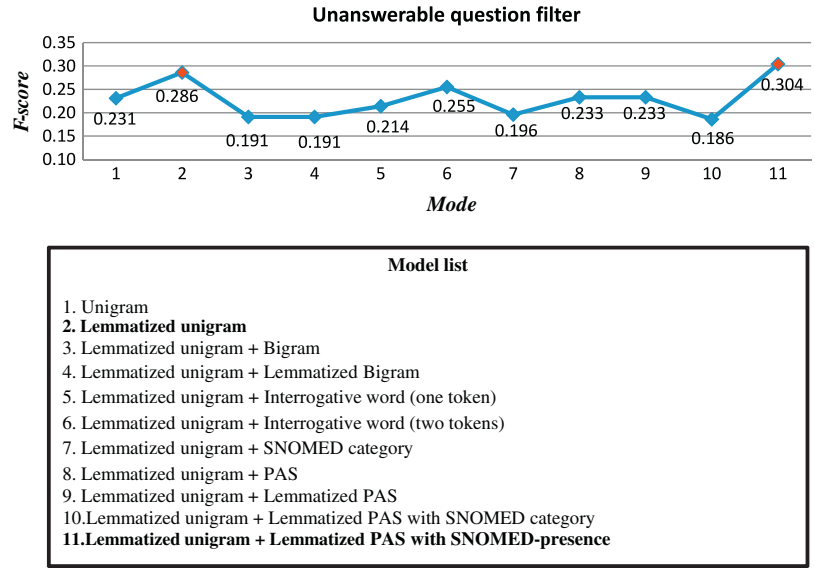


Fig. 4. Unanswerable question filter performance for 11 different learning models.

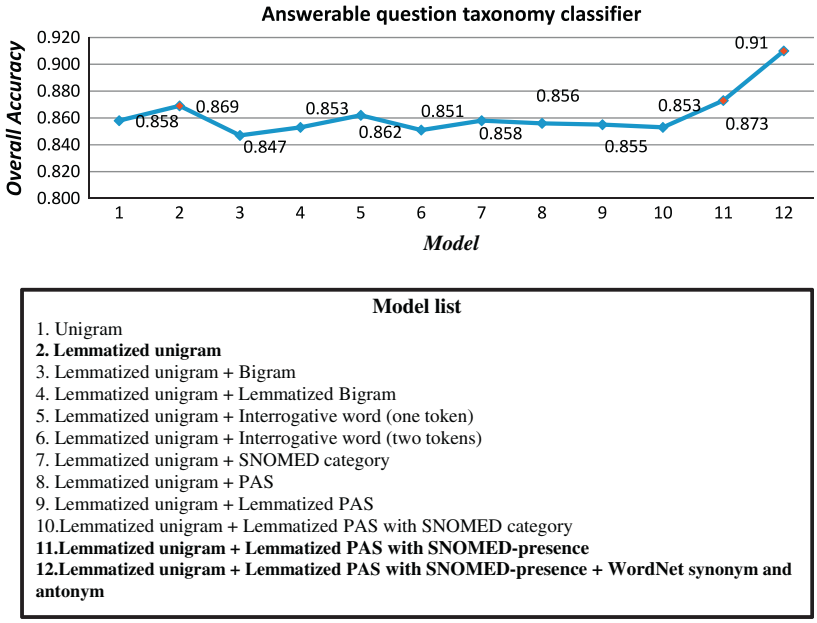


Fig. 5. Answerable question taxonomy classifier performance for 12 different learning models.

4.2. Question classification

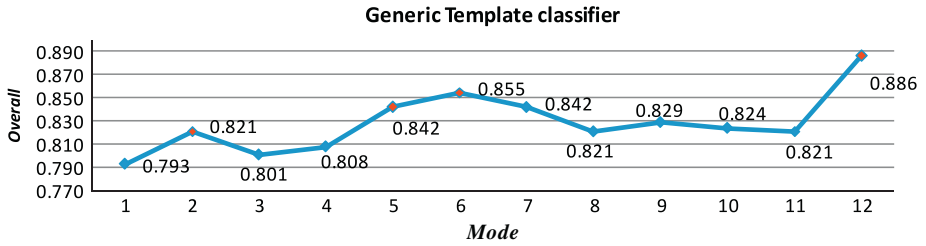
Three SVM learners were developed in this classification task, namely 'Unanswerable question filter', 'Answerable question taxonomy classifier' and 'Template classifier' which is presented in Fig. 3. In order to discover the best feature sets, a selective incremental method was used. If the performance benefited by one feature set, then this feature set was retained, otherwise, it was dropped. The outcomes of three series of experiments are displayed in Figs. 4–6 respectively, as determined by two fold cross-validation.

4.2.1. Unanswerable question filter

From Fig. 4, it can be seen that the *F*-score for identifying unanswerable questions cannot be improved effectively by various feature sets in the SVM learner. The *F*-score for unanswerable questions was relatively low, with the highest score 30.4% by using 'Lemmatized unigram' and 'Lemmatized PAS with SNOMED-presence' as learning features. Although the rule based approach (introduced in Section 2.3.1) can achieve an *F*-score of 100%, its robustness is doubtful, which will be discussed in next section.

4.2.2. Taxonomy classification

Unlike the experiments in the 'Unanswerable question filter', the performance of the classifier profited from the feature sets (see Fig. 5). By investigating the initial eleven feature sets, the best overall accuracy (87.3%) was obtained by using 'Lemmatized unigram' and 'Lemmatized PAS with SNOMED-presence'. It is noticeable that only three of the feature models (2, 5 and 11) in the initial eleven feature sets perform better than the baseline unigram



Model list

1. Unigram
2. **Lemmatized unigram**
3. Lemmatized unigram + Bigram
4. Lemmatized unigram + Lemmatized Bigram
5. **Lemmatized unigram + Interrogative word (one token)**
6. **Lemmatized unigram + Interrogative word (one token) + Interrogative word (two tokens)**
7. Lemmatized unigram + Interrogative word (one token) + Interrogative word (two tokens) + Interrogative word (one token) + Interrogative word (two tokens)+ SNOMED category
8. Lemmatized unigram + Interrogative word (one token) + Interrogative word (two tokens) + PAS
9. Lemmatized unigram + Interrogative word (one token) + Interrogative word (two tokens) + Lemmatized PAS
10. Lemmatized unigram + Interrogative word (one token) + Interrogative word (two tokens) + Lemmatized PAS with SNOMED category
11. Lemmatized unigram + Interrogative word (one token) + Interrogative word (two tokens) + Lemmatized PAS with SNOMED-presence
12. **Lemmatized unigram + Interrogative word (one token) + Interrogative word (two tokens) + Subject + Observational entity**

Fig. 6. Generic template classifier performance for 12 different feature models.

Table 3
Answerable question taxonomy classifier scores of three models.

Features	Class	Precision	Recall	*F*-score	Overall accuracy
Lemmatized unigram	Comparison	0.813	0.578	0.675	0.869
	Decision making	0.947	0.783	0.857	
	User interaction	0.400	0.533	0.457	
(Model 2)	Reason	0.864	1.000	0.927	
	Structured database	0.889	0.471	0.615	
	General note	0.800	0.414	0.546	
	Specific note	0.891	0.972	0.929	
1. Lemmatized unigram	Comparison	0.879	0.644	0.744	0.873
	Decision making	0.921	0.761	0.833	
2. Lemmatized PAS with SNOMED-presence	User interaction	0.700	0.467	0.560	
	Reason	0.895	0.895	0.895	
	Structured database	0.600	0.353	0.444	
(Model 11)	General note	0.667	0.345	0.455	
	Specific note	0.884	0.990	0.934	
1. Lemmatized unigram	Comparison	0.900	0.800	0.847	0.910
	Decision making	0.907	0.848	0.876	
2. Lemmatized PAS with SNOMED-presence	User interaction	0.700	0.467	0.560	
	Reason	0.950	1.000	0.974	
3. WordNet synonym and antonym	Structured database	0.769	0.588	0.667	
(Model 12)	General note	0.833	0.690	0.755	
	Specific note	0.924	0.974	0.948	

model indicating that this task is more difficult than expected. However, this accuracy value was boosted to 91.0% (model 12) by integrating a WordNet synonym and antonym feature into the best initial learning feature model (model 11). The detailed scores for each taxonomy class in three beneficial feature models (models 2, 11 and 12) are presented in Table 3.

4.2.3. Generic templates classification

The experiment statistics of 'Template classifier' are displayed in Fig. 6. The best overall accuracy (88.6%) was achieved by adopting 'Lemmatized unigram', 'Interrogative word (one token)' and 'Interrogative word (two tokens)', and 'Subject'. In comparison with the 'Answerable question taxonomy classifier', the PAS based feature sets did not offer any contribution to the system. This gives some confidence that the design of the templates is effective. It is also encouraging that there is only a 2% drop in overall accuracy between the Template and the Answerable Question classification process. The detailed scores for each template class in model 2, 6, and 12 were presented in Tables 4–6 respectively.

5. Discussion

5.1. Annotation reliability and statistics

A previous study [49] has argued that "Kappa > 0.8 as good reliability, with $0.67 < K < 0.8$ allowing tentative conclusions to be drawn". More recent studies [48,50] begin to adopt a much more stringent threshold of kappa statistic of 0.8 or even as high as 0.9 as the minimum acceptable level of reliability. From Table 2, it is shown that a very good kappa statistic was obtained in dual

Table 4
Generic template classifier performance of model 2.

Features	Class (X: Retrieval expression, T: Time expression)	Precision	Recall	*F*-score	Overall accuracy
Lemmatized unigram(Model 2)	1. Did the patient (have) X, T?	0.851	0.934	0.891	0.821
	2. Does the patient's X have (been) Z?	1.000	0.091	0.167	
	3. How often did the patient (have) X?	0.000	0.000	0.000	
	4. What has grown in X, T?	1.000	0.667	0.800	
	5. What is the color of X, T?	1.000	1.000	1.000	
	6. What is the trend of X, T?	1.000	0.571	0.727	
	7. What was the description of X, T?	0.733	0.851	0.788	
	8. What was the treatment for X, T?	1.000	1.000	1.000	
	9. What was the value of X, T?	0.812	0.828	0.820	
	10. What was the value of X, T? X vs. Z?	0.750	0.600	0.667	
	11. When was the last time for the patient (having) X?	1.000	0.857	0.923	
	12. When was the last time for the patient (having) X? And how long?	0.833	0.714	0.769	
	13. Where is the location of X, T?	1.000	0.500	0.667	
	14. Who was/has X, T? and how many of them?	0.833	0.790	0.811	
	15. Who was the patient's X?	1.000	1.000	1.000	
	16. Other	0.000	0.000	0.000	

Table 5
Generic template classifier performance of model 6.

Features	Class (X: Retrieval expression, T: Time expression)	Precision	Recall	*F*-score	Overall accuracy
1. Lemmatized unigram	1. Did the patient (have) X, T?	0.836	0.959	0.893	0.855
	2. Does the patient's X have (been) Z?	1.000	0.009	0.167	
	3. How often did the patient (have) X?	0.000	0.000	0.000	
2. Interrogative word (first token)	4. What has grown in X, T?	1.000	0.667	0.800	
	5. What is the color of X, T?	1.000	1.000	1.000	
	6. What is the trend of X, T?	1.000	0.571	0.727	
3. Interrogative word (first two token)	7. What was the description of X, T?	0.791	0.919	0.850	
	8. What was the treatment for X, T?	1.000	1.000	1.000	
	9. What was the value of X, T?	0.893	0.838	0.865	
	10. What was the value of X, T? X vs. Z?	0.750	0.600	0.667	
(Model 6)	11. When was the last time for the patient (having) X?	1.000	0.905	0.950	
	12. When was the last time for the patient (having) X? And how long?	0.833	0.714	0.769	
	13. Where is the location of X, T?	1.000	0.833	0.909	
	14. Who was/has X, T? and how many of them?	0.889	0.842	0.865	
	15. Who was the patient's X?	1.000	1.000	1.000	
	16. Other	0.000	0.000	0.000	

Table 6
Best generic template classifier performance (model 12).

Features	Class (X: Retrieval expression, T: Time expression)	Precision	Recall	*F*-score	Overall accuracy
1. Lemmatized unigram	1. Did the patient (have) X, T?	0.892	0.943	0.916	0.886
	2. Does the patient's X have (been) Z?	0.857	0.546	0.667	
2. Interrogative word (first token)	3. How often did the patient (have) X?	0.000	0.000	0.000	
	4. What has grown in X, T?	1.000	0.667	0.800	
3. Interrogative word (first two token)	5. What is the color of X, T?	1.000	1.000	1.000	
	6. What is the trend of X, T?	0.833	0.714	0.769	
4. The Subject of the question	7. What was the description of X, T?	0.852	0.932	0.890	
	8. What was the treatment for X, T?	1.000	1.000	1.000	
5. Observational entity	9. What was the value of X, T?	0.908	0.899	0.904	
	10. What was the value of X, T? X vs. Z?	0.750	0.600	0.667	
	11. When was the last time for the patient (having) X?	1.000	0.905	0.950	
	12. When was the last time for the patient (having) X? And how long?	0.714	0.714	0.714	
(Model 12)	13. Where is the location of X, T?	1.000	0.833	0.909	
	14. Who was/has X, T? and how many of them?	0.850	0.895	0.872	
	15. Who was the patient's X?	1.000	1.000	1.000	
	16. Other	0.000	0.000	0.000	

annotation of the taxonomy classes, except the 'Ambiguous' class. This low reliability in the 'Ambiguous' class was induced because one of the annotators (the doctor) was more concerned about whether the question was asked precisely or not. For example, 'What is the patient condition?' and 'What is the status of a patient?' were labeled as unanswerable, since the doctor believed that the answer targets were unspecified. In other words, the answer can be various according to user's preferences. However, these questions were considered answerable by the other annotator (the computational linguist), which was based on the idea of that the answer target can be restricted by interacting with a user and allowing them to choose each point from a predefined list.

5.2. Question classification

By observing the three performance diagrams (Figs. 4–6), there is no doubt that the 'Lemmatized unigram' feature set brought more contribution than the normal 'Unigram' feature set, which suggested that the variation of different questions which belong to one category was smoothed by this surface level generalization.

Thus, this feature set was chosen as a competitive baseline. In contrast, the performances of these three systems were decreased when introducing the bigram level feature, namely, the 'Bigram' and the 'Lemmatized Bigram', by which the diversity in each category was increased.

On the other hand, for identifying unanswerable questions and classifying answerable questions, the reliability was improved by exploring the generalized semantic feature set ('Lemmatized PAS with SNOMED-presence'). Due to the complexity of the SNOMED hierarchy, the 'Lemmatized PAS with SNOMED category' feature set did not benefit these two classifiers, for example, many medical terms belong to multiple SNOMED categories which are used across different classification classes. The 'Template classifier' is more concerned about the syntax structures and answer types of the questions. Consequently, the overall accuracy was not increase by the generalized semantic feature set. Furthermore, a significant improvement (more than 3.0%) was brought about by adding the interrogative word as a feature in the 'Template classifier' in which the answer type is a main classification criterion.

5.2.1. Unanswerable question filter

A relatively stable but poor unanswerable *F*-score is displayed in Fig. 4 by exploring different feature sets through the SVM learner, which reflected that the machine learning approach is not optimal for this task. Thus, a rule based approach was investigated to overcome this issue. In the end, 29 rules were crafted to capture these 37 unanswerable questions based on our generalization strategies which achieved 100% *F*-score. However, only 16 rules were acceptable which were generalized by at least three levels of generalization, namely token lemmatization, synonym and antonym, and predicate argument. These rules cover 24 questions (all the 'Operational issue' questions, half of the 'External knowledge' questions and one third of the 'Ambiguous' questions), since common PASs can be found throughout these questions. For the remaining 13 questions, the common PASs cannot be extracted due to the failure of the parsing or uniqueness of the questions, thus, 13 rules were created with simple generalization. Through experiments, only the 'Operational issue' questions can be well covered while poor performance is obtained for the unseen 'Ambiguous' questions, 'External knowledge' questions and 'Statistical/logical inference' questions. This rule based method could be a temporary solution for the situation of a shortage of Unanswerable Questions. Once more Unanswerable Questions were available, the rule based method should be reverted to a statistical model.

5.2.2. Answerable question taxonomy classifier performance

From the graph in Fig. 5 and the detailed statistics in Table 3, it is shown that the general feature does not perform well at identifying the small classes ('Comparison', 'Decision Making', 'User Interaction', 'Structured database' and 'General note'), while the larger class ('Specific note') can achieve a much better *F*-score (see the class frequency in Table 2). By adopting an extra feature set (WordNet synonym and antonym), the overall accuracy was boosted to 91.0%. Meanwhile, the *F*-scores for these small classes were increased by nearly 13.0%. However, drawbacks also exist in this approach, for example:

1. Due to the nature of WordNet which is an open domain knowledge resource, in some cases WordNet cannot provide proper synonyms or antonyms for medical terms, such as 'admit → acknowledge, accept, allow, etc.' and 'transfer → shift, change, etc.'
2. The indicative elements may not be unique. For instance, a few questions in the general question set classified as database questions, due to the similar verb argument, like 'Why did he come to Intensive Care Unit?', 'What happened before he came to hospital?', etc.
3. This specific feature is not applicable. Such as, there is no adjective in the comparison question ('Are his kidneys responding to the fluid bolus?'). The worst case is where there is no indicative element discovered, as for the 'User interaction' questions, which leads to no improvement for this class.

5.2.3. Template classifier performance

Similar to the 'answerable question classification', the overall accuracy was improved by integrating the extra feature set ('Subject'). In comparison with the small class boosting approach used in the answerable question taxonomy classifier, the large classes (templates 1, 7 and 9) (see the class frequency in Table 2 and the class performance in Table 5) are improved in the template classifier. The main reason is that the *F*-scores for these three large classes are relatively low. If the performances of these classes can be increased, the overall accuracy can be boosted (the detailed *F*-scores for each template in the initial best feature model were presented in Table 5). Model 12 revealed that the false positives in these three large classes can be significantly reduced by introducing the 'Subject' feature set (the detailed *F*-scores for each template in model 12 were presented in Table 6). As a result, the average *F*-score for these three large classes was increased to approximately 90.0% due to the improvement of their precisions. Meanwhile, these false positives in the large classes would be successfully classified into the small classes to increase their recall. Finally, an overall accuracy of 88.6% was achieved in the multi-classification task. Meanwhile, a zero *F*-score was obtained for the templates 3 and 16. This occurs because the questions in template 16 are totally different from each other which should belong to two individual templates (mentioned in the first part of Section 4). On the other hand, the three questions in template 3 were not separated symmetrically in the two fold cross validation, for example, 'How many times has she opened her bowels in last 24 h?' and 'How often is she coughing?' was distributed in the first fold while 'How often were they bleeding?' was allocated in the other fold. However, if an extra question with an interrogative word of 'How often' was added into the last fold, a better *F*-score (80.0%) can be achieved for template 3.

5.3. Issues for medical QA

According to the observations from previous studies [18–22] and our studies, clinicians often have questions about the care of their patient. There is no doubt that sophisticated IR systems could be utilized to assist users to answer questions [21,51–54], but the effectiveness of this method needs to be considered. IR systems, by themselves, are incapable of performing this kind of task, since they do not have the capability to convert questions into a set of search terms; as well they retrieve documents rather than answers. Thus, four major steps would be involved in the process of asking and answering clinical questions by using IR systems, e.g. 1. Recognize a question. 2. Reformulate the question into a set of search terms. 3. Search for relevant information. 4. Formulate the answer. Previous investigations [18,30,55] indicate the difficulties of question reformulation and the cost of time taken are two of the major barriers to have hindered search engines from answering questions. In contrast to finding answers through the IR system, a native QA system would be much more practical, from which, the answer would be returned automatically after a question is submitted. Question processing, being the first step of a QA system, not only classifies the answering method, but also identifies the answer type. Furthermore, the identifying keywords for retrieval expressions, temporal expressions, and constraint expressions are

also produced for subsequent processing. As the typical QA system is precision driven, if a question is unsuccessfully decoded, no answer can be returned. In order to solve this issue, a user interaction process would need to be adopted, for example, a request for user validation in every question processing stage, if the result is incorrect, the user can suspend the following processes and provide the correct information, or rewrite the initial question in a different way. A complementary strategy could be to have a robust document processing engine to automatically map terms to their synonyms and hence expand the set of candidate documents available for answer extraction.

Currently, the electronic patient record orientated QA system is under progressive development. The question processing component has been completed and utilizes this multilayer question classification model, as well as a keywords extraction model. Most of the document processing components have already been developed, which include the proof reading processing (tokenization, misspelling correction, abbreviation & acronym expansion, part of speech tagging and lemmatization), document structure analyzer, sentence boundary detector, scores and measures detector, temporal event detector, and SNOMED concept indexer. In the near future, the document processing engine will be completed by integrating new components and assembling them. At the same time, we are planning to build an answer corpus for our question set. Once this corpus is built, the study to assess the performance of the answer extractor can be conducted.

5.4. Limitations

The main limitation of our work is that the question set is restricted by small geographical location and small group of doctors, although we believe ICUs have the most progressive use of the latest technology to improve the efficiency of daily practice. If we want to reuse this multilayer question classification system for another hospital department, such as emergency unit and general surgery unit, this processing model may need to be adapted. A further analysis is needed to determine the generalizability of our categorization method.

6. Conclusion

In this paper, a comprehensive clinical question study is documented which offers three crucial contributions: ① The question corpus[1] which we collected can be studied for addressing different issues in the field, such as the need for a clinical question parser. To the best of our knowledge, no parser has been designed specifically for clinical questions; meanwhile the current parser did not perform well on our question corpus. ② The clinical question taxonomy and templates which we developed can be used for further QA studies, especially question processing. ③ The idea of introducing a KD–KR process to explore specific feature sets to improve the minimum classification was adopted. This study is a part of our larger plan, which is to design a model for a clinical QA system, which is an unexplored field. It is likely that this technology can significantly increase the efficiency of the work practices, and thereby improve the quality of medical care especially in the crucial care environment of the ICU.

Acknowledgments

We would like to thank Drs. Robert Herkes and Michael O'Lerie, and Angela Ryan, Chris Wise and other staffs in the Intensive Care Unit of Royal Prince Alfred Hospital for providing the observation opportunity and environment, as well as Dr. Stephen Crawshaw for his contribution to annotations and valuable feedback.

Appendix A. Question taxonomy description

The ICU question corpus can be separated into 'Answerable questions' and 'Unanswerable questions' in the first place. If the question cannot be answered by our system, then it will be classified into 'Unanswerable questions'. For the 'Unanswerable questions', four subclasses are involved:

a. Ambiguity: If a question requires knowledge not intrinsic to the words, that is it is highly ambiguous, then it is unanswerable. For example, 'Has anything been grown?', "grow" is used in many other contexts in clinical notes, so to answer this question would require word sense disambiguation. There needs to be knowledge that "grow" refers to cultures grown from tissues such as "sputum", "urine" and "blood".
b. External knowledge: If external knowledge (such as medical literature) is required to answer a question, then this question is unanswerable. Such as "What is the coliform sensitive to?" and "Is there any requirement to start Levosimendan?".
c. Statistical/logical inference: Questions that might be answerable from statistical inference or logical deduction over the content of a set patient records are also included in this category. "What do we normally treat trichloronate with?", "What do we test for gout aside from URIC acid?", etc.
d. Operational issue: Operational issues are managed by organizational members whose work is not expected to be recorded in the patient records, although they are frequently of importance to clinical staff. Such as 'Have social workers been booked?'

On the other hand, if the question belongs to any of the four answerable subclasses, then it is an answerable question. The four subclasses are:

a. Conditional answerable: The question is potentially answerable, if the answer cannot be directly extracted from the patient records (which means the fragments of the answer can be extracted, however, a sophisticated engine is required to analyze these fragments and assemble the answer) or the probability for this answer being written in the patient record is very low.
b. Structured database: If a question is talking about admissions and its answer can be found by simply applying an SQL query, then it is an answerable question, such as "Has the patient been readmitted to a hospital ward?".
c. Specific note: If the answer can be extracted from the patient records by using the information only in the questions, then this question is answerable. For instance: "What is his O2 requirement?" indicates "O2" volume is the main retrieval subject.
d. General note: If a question is general enough and can be answered by using one section of the patient record, then it is answerable. For instance: "What is the history of this patient?" can be answered by returning the "Past Medical History" section in the record.

Finally, the 'Conditional answerable' class can be further broken into four subclasses:

a. Comparison: If a question is about comparing, then this question is a comparison question, such as "Is her coagulopathy improving?", which needs a comparison engine which

[1] The question collection is available at http://hitrl.cs.usyd.edu.au/ICNS/question_collection/.

has the criteria for the decision making. In principle the QA system could find the values of the variable defined for use in the engine.

b. User interaction: This type of question requires references by user interactions, for instance, "What is the status for him?". Without any references, a coarse answer would be generated by returning all measurements from the patient's record. However, if the measurements could be selected by a clinician according to their preferences, then the answer can be more precise.
c. Decision making: If a question is about making a decision, then this question is classified as a 'Decision making' question, such as "Do you think he needs a higher Mean Arterial Pressure?", which needs a reasoning engine which has the criteria for the decision making.
d. Reason: This type of question is asking about a reason, such as "Why was the blood given?", "Why did antibiotics get ceased?" and "Why did we give the Lidocaine?". These reason questions tend to be centered around reasons for treatments which are rarely written explicitly in the patient record. However, sometimes it can be found in the records, such as "Antibiotics ceased as two drains removed". Consequently, there cannot be a high reliability in answering this question type and so 'Reason' questions are classified under the 'Conditional answerable' class.

Appendix B. Examples for generic question template

(X: Retrieval expression, T: Time expression, Z: Reference Constraint, W: Reference value)

Template instances	Template
Did he cough last night?	Did the patient (have) X, T?
Did he ever grow any organisms?	Did the patient (have) X, T?
Did she have Paracetamol this morning?	Did the patient (have) X, T?
Did this patient have family coming?	Did the patient (have) X, T?
Does he have any Fresh Frozen Plasma?	Did the patient (have) X, T?
Does she have plural infections?	Did the patient (have) X, T?
Has she had a chest X-ray yet?	Did the patient (have) X, T?
Has the patient sat out of bed?	Did the patient (have) X, T?
Was she hypertensive on admission?	Did the patient (have) X, T?
Is she wheeze this morning?	Did the patient (have) X, T?
Does his Computed Tomography have contrast?	Does the patient's X have (been) Z?
Has the chest drain been removed?	Does the patient's X have (been) Z?
Has the fistula closed?	Does the patient's X have (been) Z?
Has the fistula sealed off?	Does the patient's X have (been) Z?
Is his abdomen distended?	Does the patient's X have (been) Z?
Is his abdomen enlarged?	Does the patient's X have (been) Z?
Is his abdomen soft?	Does the patient's X have (been) Z?
Is the CO_2 level or the pH level being titrated?	Does the patient's X have (been) Z?
Is the drain still bubbling?	Does the patient's X have (been) Z?
Is there any consolidation on the lung X-ray?	Does the patient's X have (been) Z?
How many times has she opened her bowels in last 24 h?	How often did the patient (have) X?
How often is she coughing?	How often did the patient (have) X?
How often were they bleeding?	How often did the patient (have) X?
Has he grown anything in his sputum?	What has grown in X, T?
What has grown in the blood?	What has grown in X, T?
What has grown in the sputum?	What has grown in X, T?
What is the color of urine?	What is the color of X, T?
What is the color of sputum?	What is the color of X, T?
What is the color of stoma?	What is the color of X, T?
Are her inflammatory markers up?	What is the trend of X, T?
What is the trend in haemoglobin?	What is the trend of X, T?
What is the trend of Blood Pressure?	What is the trend of X, T?
What is the trend of his blood sugar after he got up?	What is the trend of X, T?
What is the trend of his PH?	What is the trend of X, T?
What microorganisms were cultured?	What is the trend of X, T?
What microorganisms were have been grown?	What is the trend of X, T?
How are her reflexes?	What was the description of X, T?
What are the patient's medications?	What was the description of X, T?
What did his Electro-Cardiogram show?	What was the description of X, T?
What is her hematological status?	What was the description of X, T?
What is his allergy?	What was the description of X, T?
What is his neurological state?	What was the description of X, T?
What was the preop echocardiogram result?	What was the description of X, T?
What is the description of sputum?	What was the description of X, T?
What is his emotional status?	What was the description of X, T?

Appendix B (*continued*)

Template instances	Template
What is the patient fluid status?	What was the description of X, T?
How many blood products did he receive during 24 h?	What was the value of X, T?
How many boluses of fluid have you given him?	What was the value of X, T?
How many fluid was given?	What was the value of X, T?
How much drain since last night?	What was the value of X, T?
How much insulin is he on?	What was the value of X, T?
What is his blood pressure?	What was the value of X, T?
What is his PO2 level?	What was the value of X, T?
What is his Pressure Support?	What was the value of X, T?
What is his Glasgow Coma Score?	What was the value of X, T?
What is his heart rate?	What was the value of X, T?
How did you respond to his wheeze?	What was the treatment for X, T?
How did you treat the wheeze?	What was the treatment for X, T?
Did her temperature fall below 38C?	What was the value of X, T? X vs. W?
Did the patient temperature exceed 38C in the last 48 h?	What was the value of X, T? X vs. W?
Did the patient temperature fall below 38C in last 48 h?	What was the value of X, T? X vs. W?
Has the Blood Sugar Level been greater than 10?	What was the value of X, T? X vs. W?
Whether the temperatures pass the 30c zone?	What was the value of X, T? X vs. W?
What day were the antibiotics started?	When was the last time for the patient (having) X?
When did the bowels last open?	When was the last time for the patient (having) X?
When was she extubated?	When was the last time for the patient (having) X?
When was she intubated?	When was the last time for the patient (having) X?
When was the blood product given?	When was the last time for the patient (having) X?
When was the patient most recently dialyzed?	When was the last time for the patient (having) X?
When was the patient most recently paced?	When was the last time for the patient (having) X?
When was she scanned?	When was the last time for the patient (having) X?
When was the drainage applied?	When was the last time for the patient (having) X?
How early has a patient eaten food?	When was the last time for the patient (having) X?
How long has he been on the octreotide?	When was the last time for the patient (having) X? And how long?
How long has she been on Caspofungin?	When was the last time for the patient (having) X? And how long?
How long was he ventilated?	When was the last time for the patient (having) X? And how long?
How many days since the central line was inserted?	When was the last time for the patient (having) X? And how long?
How many days since the bowel was opended?	When was the last time for the patient (having) X? And how long?
How old is his filter?	When was the last time for the patient (having) X? And how long?
How old is the line?	When was the last time for the patient (having) X? And how long?
What area is the suspected infection?	Where is the location of X, T?
Where are the chest drains located?	Where is the location of X, T?
What area has the skin been peeled off?	Where is the location of X, T?
Where are the pacing wires?	Where is the location of X, T?
Where is his line?	Where is the location of X, T?
Where was blood given?	Where is the location of X, T?
Find anyone who has a large amount of sputum?	Who was/has X, T? And how many of them?
Has there been a trauma admission?	Who was/has X, T? And how many of them?
Has there been a thrombotic event?	Who was/has X, T? And how many of them?
How many patients are ventilated?	Who was/has X, T? And how many of them?
How many patients were intubated overnight?	Who was/has X, T? And how many of them?
Who has a large amount of sputum?	Who was/has X, T? And how many of them?
Who has H1N1?	Who was/has X, T? And how many of them?
Who has had an adverse event?	Who was/has X, T? And how many of them?
Who has had multi-trauma?	Who was/has X, T? And how many of them?
Who is on NG feed?	Who was/has X, T? And how many of them?
Who is the patient's regular cardiologist?	Who was the patient's X?
Who is the patient's regular nephrologist?	Who was the patient's X?
Who was he intubated by?	Who was the patient's X?
When is Blood Sugar Level greater than 10?	Other
How has she been tolerating Non-invasive Ventilation?	Other

References

[1] Patrick J, Li M. Intelligent clinical notes system: an information retrieval and information extraction system for clinical notes. In: Proceedings of the 11th international conference on e-health networking, application and services. Sydney, Australia; December 16–18, 2009. p. 108–15.

[2] Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med 1985;103(4):596–9.

[3] Voorhees E. The TREC question answering track. Nat Lang Eng 2001;7(4):361–78.

[4] Li X, Roth D. Learning question classifiers. In: Proceedings of the 19th international conference on computational linguistics. Taipei, Taiwan, vol. 1; August 24–September 1, 2002. p. 1–7.

[5] Carlson A, Cumby C, Rosen J, Roth D (University of Illinois at Urbana-Champaign, College of Engineering, Department of Computer Science). The SNoW Learning Architecture; 1999 May. Report No.: UIUCDCS-R-99-2101.

[6] Hovy E, Gerber L, Hermjakob U, Lin C, Ravichandran D. Toward semantics-based answer pinpointing. In: Proceedings of the DARPA human language technology conference (HLT). San Diego, CA; March, 2001.

[7] Voorhees E. The TREC-8 question answering track report. In: Proceedings of the 8th text retrieval conference; 1999. p. 77–82.

[8] Voorhees E. Overview of the TREC-9 question answering track. In: Proceedings of the 9th text retrieval conference; 2000. p. 71–80.

[9] Voorhees E. Overview of the TREC 2001 question answering track. In: Proceedings of the 10th text retrieval conference; 2001. p. 42–51.

[10] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press; 2000.

[11] Zhang D, Lee WS. Question classification using support vector machines. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval; 2003. p. 26–32.

[12] Moschitti A. Efficient convolution kernels for dependency and constituent syntactic trees. In: Proceedings of the 17th European conference on machine learning (ECML). Berlin, Germany; September, 2006. p. 318–29.

[13] Moschitti A, Quarteroni S, Basili R, Manandhar S. Exploiting syntactic and shallow semantic kernels for question answer classification. In: Proceedings of the 45th conference of the association for computational linguistics. Prague, Czech Republic; June, 2007. p. 776–83.

[14] Li X, Roth D. Learning question classifiers: the role of semantic information. Nat Lang Eng 2006;12(3):229–49.

[15] Fellbaum C, editor. WordNet: an electronic lexical database. The MIT Press; May, 1998.

[16] Huang Z, Marcus T, Qin Z. Question classification using head words and their hypernyms. In: Proceedings of the 2008 conference on empirical methods in natural language processing. Honolulu, Hawaii; October, 2008. p. 927–36.

[17] Huang Z, Marcus T, Celikyilmaz A. Investigation of question classifier in question answering. In: Proceedings of the 2009 conference on empirical methods in natural language processing. Singapore, vol. 2; August 6–7, 2009. p. 543–50.

[18] Currie LM, Graham M, Allen M, Bakken S, Patel VL, Cimino JJ. Clinical information needs in context: an observational study of clinicians while using a clinical information system. AMIA Annu Symp Proc 2003:190–4.

[19] Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. BMJ 1999;319(7206):358–61.

[20] Ely JW, Osheroff JA, Ebell MH, Gorman PN, Ebell MH, Chambliss ML, et al. A taxonomy of generic clinical questions: classification study. BMJ 2000;321(7258):429–32.

[21] Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, et al. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. BMJ 2002;324:710–3.

[22] Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. JAMIA 2005;12(2):217–24.

[23] Reynolds RD. A family practice article filing system. J Prim Health Care 1995;41(6):583–90.

[24] Yu H, Sable C. Being Erlang Shen: identifying answerable questions. In: Proceedings of international joint conference of artificial intelligence (IJCAI'05) workshop on knowledge and reasoning for answering questions; 2005.

[25] Yu H, Sable C, Zhu HR. Classifying medical questions based on an evidence taxonomy. In: American association for artificial intelligence (AAAI'05) workshop on question answering in restricted domains; 2005.

[26] Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc 1993;81(2):170–7.

[27] Fuhr N. Models for retrieval with probabilistic indexing. Inform Process Manage 1998;25(1):55–72.

[28] 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.

[29] D'Alessandro DM, Kreiter CD, Peterson MW. An evaluation of information seeking behaviors of general pediatricians. Pediatrics 2004;113:64–9.

[30] 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.

[31] Yu H, Cao YG. Automatically extracting information needs from ad hoc clinical questions. AMIA Annu Symp Proc 2008:96–100.

[32] Gao YG, Cimino JJ, Ely J, Yu H. Automatically extracting information needs from complex clinical questions. J Biomed Inform 2010;43:962–71.

[33] Terol MR, Martinez-Barco P, Palomar M. A knowledge based method for the medical question answering problem. Comput Biol Med Arch 2007;37(10):1511–21.

[34] Athenikos SJ, Han H, Brooks AD. A framework of a logic-based question-answering system for the medical domain (LOQAS-Med). In: Proceedings of the 2009 ACM symposium on applied computing (SAC). Honolulu, Hawaii, USA; March 9–12, 2009. p. 847–51.

[35] Nielsen RD, Masanz J, Ogren P, Ward W, Martin JH, Savova G, et al. An architecture for complex clinical question answering. In: Proceeding IHI'10 proceedings of the 1st ACM international health informatics symposium; 2010. p. 395–9.

[36] Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc 2006:359–63.

[37] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Comput Linguist 2007;33:63–103.

[38] Athenikos SJ, Han H. Biomedical question answering: a survey. J CMPB 2009;99(1):1–24.

[39] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312(7023):71–2.

[40] Gorman PN, Ash J, Wykoff L. Can primary care physicians' questions be answered using the medical journal literature. Bull Med Libr Assoc 1994;82(2):140–6.

[41] Straus S, Sackett D. Bringing evidence to the point of care. J Am Med Assoc 1999;281:1171–2.

[42] 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(6):541–7.

[43] Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, et al. Developing a robust part-of-speech tagger for biomedical text. Advances in informatics – 10th panhellenic conference on informatics; 2005. p. 382–92.

[44] SNOMED Clinical Terms User Guide 2010. International Health Terminology Standards Development Organisation. <http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Publications/doc_UserGuide_Current-en-US_INT_20100131.pdf> [accessed 01.06.11].

[45] Patrick J, Wang Y, Budd P. An automated system for conversion of clinical notes into SNOMED clinical terminology. In: Proc 5th Australasian symposium on ACSW frontiers, vol. 68; 2007. p. 219–26.

[46] Miyao Y, Tsujii J. Probabilistic disambiguation models for wide-coverage HPSG parsing. In: Proceedings of ACL-2005; 2005. p. 83–90.

[47] Cohen Jacob. A coefficient of agreement for nominal scales. Educ Psychol Measure 1960;20(1):37–46.

[48] Klaus Krippendorff. Content analysis: an introduction to its methodology. 2nd ed. Sage Publications; 2004.

[49] Carletta J. Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 1996;22(2):249–54.

[50] Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist 2008;34(4):555–96.

[51] Wildemuth BM, de Bliek R, Friedman CP, File DD. Medical students' personal knowledge, searching proficiency, and database use in problem solving. J Am Soc Inform Sci 1995;46:590–607.

[52] Hersh WR, Pentecost J, Hickam DH. A task-oriented approach to information retrieval evaluation. J Am Soc Inform Sci 1996;47:50–6.

[53] Friedman CP, Wildemuth BM, Muriuki M, et al. A comparison of hypertext and Boolean access to biomedical information. Proc AMIA Ann Fall Symp 1996:2–6.

[54] Hersh WR, Crabtree MK, Hickam DH, et al. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. J Am Med Inform Assoc 2002;9(3):283–93.

[55] Ely JW, Osheroff JA, Maviglia SM, Rosenbaum ME. Patient-care questions that physicians are unable to answer. J Am Med Inform Assoc 2007;14(4):407–14.