Contextual Question Answering for the Health Domain

Wilson Wong, John Thangarajah, Lin Padgham
School of Computer Science and Information Technology
RMIT University, Melbourne VIC 3000 Australia
{wilson.wong,john.thangarajah,lin.padgham}@rmit.edu.au

Studies have shown that natural language interfaces such as question answering and conversational systems allow information to be accessed and understood more easily by users who are unfamiliar with the nuances of the delivery mechanisms (e.g., keyword-based search engines) or have limited literacy in certain domains (e.g., unable to comprehend health-related content due to terminology barrier). In particular, the increasing use of the Web for health information prompts us to re-examine our existing delivery mechanisms. We present enquireMe, which is a contextual question answering system that provides lay users with the ability to obtain responses about a wide range of health topics by vaguely expressing at the start and gradually refining their information needs over the course of an interaction using natural language. enquireMe allows the users to engage in 'conversations' about their health concerns, a process that can be therapeutic in itself. The system uses community-driven question-answer pairs from the Web together with a decay model to deliver the top scoring answers as responses to the users' unrestricted inputs. We evaluated enquireMe using benchmark data from WebMD and TREC to assess the accuracy of system generated answers. Despite the absence of complex knowledge acquisition and deep language processing, enquireMe is comparable with the state of the art question answering systems such as START as well as those interactive systems from TREC.

INTRODUCTION

The World Wide Web has revolutionised the ways we access information, and health information is no exception. In this paper, we describe a Web-based system that supports interactions with lay users in a natural manner to discuss and obtain information about their health concerns. There is a huge range of health-related services available on the Web, ranging from live chat with health professionals (e.g., `AskTheDoctor.com`) and self-help communities for sharing patient experiences (e.g., `PatientsLikeMe.com`) (Bennett et al., 2010) to health search engines (e.g., `healia.com`). Although this increasing reliance on information from the Web for health-related purposes is highly contentious (Robertson and Harrison, 2009) and has been greatly criticised by health professionals (Ryan and Wilson, 2008), it is clearly very popular and according to some, does also have benefits (McDaid and Park, 2011). Potential benefits include the fact that online systems can reach and raise awareness of health-related issues in hard to reach groups such as teenagers, as well as the ability to provide information for those with minor health problems that may be managed without the need for medical consultations (McDaid and Park, 2011), thus freeing up the time of health professionals to deal with people requiring attention. Thus, the focus of this paper is on the delivery mechanism rather than the source of information.

Despite the differences between the existing delivery mechanisms, the more popular ones share a common trait, which is allowing lay users to access and share information via natural language. Research has shown that *"people prefer natural expression of queries over keywords"* (Hearst, 2011) for obvious reasons. For example, if a searcher needs to find the possible medical conditions given a long list of symptoms (e.g., cough, red eye, headache), a conventional keyword query could fail. However, given the ability to express the symptoms in natural language queries over several connected attempts, such a system could iteratively refine and provide a short list of medical conditions through this process. In short, although the searchers may lack the familiarity with the delivery mechanism or the specialised vocabulary of the domain, they still possess the same basic vocabulary as other people in the same cognitive situations (Hearst, 2011). The paper by Zhou (2007) discusses in depth the other benefits of natural language interfaces including less clutter as compared to menu-driven interfaces, accessibility for the visually impaired (requires a speech layer on top) and efficiency in task accomplishment, to name a few.

Given the potentials of natural language interfaces, it would be interesting to see if the state-of-the-art online delivery mechanisms are adequate or appropriate for health information. Our review of the literature revealed that current systems are lacking in real time interactivity and concise responses. When we consider the difficulty of expressing the multi-faceted nature of health information needs as isolated queries and the fact that *"information seeking...is a highly contextual activity"* (Ruthven, 2011), the need for interactivity is obvious. The inability to refine irrelevant or large result sets can lead to information overload, or worse still, create anxiety amongst the users (White and Horvitz, 2010). In this research, we consider question answering as a more effective mechanism in terms of delivering concise responses in real-time to health questions over the Web. However, the state of the art systems such as the health-related HONqa (Cruchet et al., 2009b,a) and MedQA (Yu et al., 2005) as well as the cross-domain START (Katz and Lin, 2002) and QuALiM (Kaisser, 2008) remain focused on providing one-off responses (i.e., no interactivity), which may or may not be concise, to individual factoid wh-questions (i.e., restricted input type). START, for example, produces paragraphs and bulleted items from Wikipedia and other online dictionaries as answers. Also, the need to always interact with systems such as START using wh-questions prevents the users from expressing their information needs more freely in ways that they are familiar with. In general, while current question answering systems may have progressed far in terms of performance (e.g., about 85% precision in factoid, cross-domain questions), they remain restricted in terms of interactivity and the types of input supported. Both of these characteristics together with concise, more colloquial responses are crucial in our attempt to develop a system that engages lay users on their health information needs.

The main focus of this paper is to address the need for interactivity in delivering answers to natural language questions over the Web in domains such as health. To design and develop an interactive question answering system that supports beyond wh-question inputs as well as potentially produces more concise, 'natural' responses, we have identified the following challenges that require addressing:

- Answer generation: Current approaches typically rely on templates (for knowledge based approaches) and snippets extracted from unstructured Web text (for document retrieval based approaches) for producing answers, which avoid more complex natural language generation. The rigidity of the use of templates, and the non-colloquial and verbose nature of Web documents (Inoue et al., 2011), however, cause system responses to be predictable, unnatural and verbose. In particular, the 'naturalness' aspect of system responses is crucial in any conversational system (Inoue et al., 2011).
- Input type restriction: Due to the reliance on deep syntactic and semantic analysis in existing systems for recognising answer types from inputs, resolving pronouns, and extracting 'knowledge' (e.g., binary predicates that capture verbs and their arguments, nouns and their modifiers), the types of input that they support are greatly restricted. This restriction has the potential to create stilted interaction where the users become fixated on trying to conform to the system's input restriction as opposed to their information needs (Allen et al., 2001).
- Context awareness: The inputs that arrive in a sequence during contextual question answering are evolving and related expressions of some common informations needs (Sun and Chai, 2007). In other words, input sequences are not random and the previous inputs in a sequence play an important role in determining the most appropriate

system response for the current input. To maintain a coherent and focused dialogue, the system must have some context awareness. The challenge is to manage contextual information and resolve pronouns without the need for annotated corpora and domain knowledge typically associated with discourse processing (Mitkov, 2001).

In this paper, we describe in detail an automated interactive facility called *enquireMe* (Wong et al., 2011) where a lay person can ask questions and obtain concise responses about any health-related issues via a natural, potentially helpful conversation in real time. The main feature that differentiates enquireMe from existing (contextual) question answering systems is the innovative use of disjointed question-answer pairs from community-driven websites. Unlike existing systems that rely on templates and text snippets from Web documents to produce answers, enquireMe uses the answer component of question-answer pairs for its 'naturalness'. The need for deep natural language processing, which is common in systems such as START, is avoided in enquireMe through its information retrieval-based approach that uses keyphrase extraction and context management for input processing and answer finding. Context, in this system, loosely refers to the *interaction context* as described by Ruthven (2011), which is the weighted words or phrases that are maintained over the course of an interaction. The system's context management allows the users to vaguely express and then gradually build up over time their health information needs. Overall, this system addresses the above mentioned three problems in the following ways:

(1) The system produces natural and concise responses by using the answer component of question-answer pairs (which are crafted by large communities of human contributors).
(2) The system processes unrestricted natural language inputs because no deep natural language processing is performed. Instead, the system uses the overlap of keyphrases between the inputs and the question-answer pairs as well as other criteria such as user votes and the content-bearing aspect of the keyphrases to find the best answer.
(3) The system collects words and phrases, and progressively alters their weights throughout the course of an interaction with the end-users, using a decay model which biases the weighting process based on part of speech information.

As we will discuss in the related work section, none of the working systems that we have reviewed combines the use of question-answer pairs, shallow natural language processing and context management into a single approach as described in points (1)-(3) above for interactive question answering. In addition to health, enquireMe is also applicable to other domains due to the domain independent nature of the context management and answer generation techniques used. The use of enquireMe in the ICT customer support domain is currently being tested with the collaboration of an industry partner - a large telecommunication company in Australia, where internal expert-data are used to determine the effectiveness of enquireMe in a task-based evaluation. We have also tested enquireMe using a very small number of question-answer pairs in the astronomy, animal, and arts & culture domains from the TREC dataset as reported in the experiments section.

This paper is structured as follow: In the related work section, we briefly discuss existing work on domain independent and health question answering, as well as the increasingly popular approach of mining community-driven resources for non-contextual question answering. We then discuss in the system architecture section our approach of utilising question-answer pairs from community-driven websites to respond to unrestricted inputs by users. In the experiments section, we discuss the results from our initial experiments comparing enquireMe's ability against three existing question answering systems in responding to both health-related as well as cross-domain questions using benchmark data from WebMD and TREC. We conclude this paper with the strengths and limitations of our approach as well as indicating some areas for future work.

RELATED WORK

In this section, we look at several state of the art health and domain independent question answering systems, and the approach of mining community-driven resources for question answering in general. We will discuss why the techniques used by these existing systems, despite their level of sophistication, are unsuitable for our purpose of developing a contextual question answering for health. We indicate ways in which our use of question-answer pairs for contextual question answering is novel.

Question Answering Systems

Cruchet et al. (2009b) developed a health question answering system called HONqa[1] based on supervised learning. The system requires the manual compilation and classification of pairs of questions and their expected responses for training SVM classifiers. The questions are partitioned according to medical types (e.g., symptom, treatment) and the type of excepted answers (e.g., definitional, causal, true/false). During operation, the classifiers are used to analyse the input question. Depending on the types of expected response, different modules will be used to query and process the search results from different commercial search engines to retrieve the answers.

Lee et al. (2006) have developed a medical question answering system MedQA[2] that uses supervised learning to classify questions based on a hierarchical evidence taxonomy created by physicians (Yu et al., 2005). The taxonomy comprises categories such as clinical vs non-clinical, general vs specific, evidence vs no evidence, and intervention vs no intervention. MedQA operates by first extracting noun phrases from inputs and using them to retrieve relevant documents from Medline, all using off-the-shelf tools (e.g., Apache Lucene for document retrieval). Next, a set of automatically extracted lexico-syntactic patterns are used to identify definitional sentences from the documents. The sentences are then clustered and the most representative ones from all clusters are selected and displayed.

Athenikos et al. (2009) put forward a framework for logic-based medical question answering called LOQAS-Med, which is intended to provide direct answers to medical questions based on explicitly-stated facts as well as to derive hypothesis supporting or denying evidences through inference. In the paper, the authors reported their attempt at manually constructing question and answer patterns as semantic triples (i.e., `<subject,predicate,object>`). During operation, the patterns are used to identify the semantic types (e.g., symptom, disease) of the arguments, which are the subject and object, and the semantic relations (e.g., cause-effect), in the form of predicates, between the arguments from the parsed input questions.

For domain-independent question answering, START[3] is one of the earliest systems publicly accessible on the Web (Katz, 1997). The main idea behind START is the annotation of textual content on the Web using triples in the form of `<object,attribute,value>`. These triples are stored in a knowledge base with pointers back to the actual text segment. To look for answers, the parse trees of inputs are matched against the triples in the knowledge base. The text segment referred to by the corresponding triple is then used to generate the answer. To cater for differences in surface syntax, manually-defined rules are used. Since its conception in the 90's,

[1] `http://services.hon.ch/cgi-bin/QA10/qa.pl`
[2] `http://askhermes.org/MedQA`
[3] `http://start.csail.mit.edu`

the system has been loaded with huge amounts of triples (i.e., annotations) that points to information on the Web about geography (e.g., cities, countries, lakes, coordinates, weather, maps, demographics), politics and economic systems, arts and entertainment (e.g., titles, actors, directors), history and culture (e.g., birth dates, biographies), and science and technology (e.g., astronomy, health). The amount of information indexed by START, however, was not made available to public.

QuALiM[4], on the other hand, is a domain-independent system (Kaisser, 2008) that relies heavily on its pattern base (e.g., `when did NP Verb NP|PP?`) for analysing questions and finding answers. Many of the patterns are decorated with descriptions of the potential answers (e.g., expecting a named entity `Date` for `when did NP Verb NP|PP?`). When posed with a question, the system queries search engines for passages using the words that match the different parts of speech in the patterns. The passages are then parsed and named-entity recognition is used to identify the answers matching the semantic type specified in the patterns. As an extension, QuALiM has started supplementing the basic answers in the form of named entities with relevant passages from Wikipedia.

The work that has the closest resemblance to enquireMe is a speech-based question answering system by Mishra and Bangalore (2010). Note that this system does not support interactivity. It uses question-answer pairs collected from the Web to answer questions using a retrieval and ranking model based on the tf-idf metric and the Levenshtein edit distance for string comparison. With this system aside, the prominent health as well as domain independent systems discussed in this section are relatively rigid in terms of the types of input supported and the sources of answers. These systems require their inputs to be fully parsed and 'understood' in order to identify the types of input and potential answer as well as the relationships between the different constituents in the inputs. For this to happen, the users have to ensure that their inputs are wh-questions and relatively well-formed. The stilted interactions that may result from this constraint make such an approach unsuitable for contextual question answering. Moreover, the approach based on document retrieval, clustering and summarisation that these systems used to locate and generate answers cannot be easily extended to create 'natural' system outputs. It is these shortcomings which enquireMe aims to address.

Mining Community-Driven Resources for Question Answering

Despite the increasing amount of work on mining community-driven sites for question answering, they remain confined to investigating isolated problems in the field (e.g., answer retrieval, answer generation, query expansion). Miao and Li (2010), for instance, investigated the generation of topic words to enrich the representation of input questions posed to community-driven websites such as Wikipedia and Yahoo! Answers. These topic words are used as queries in place of the actual questions for these websites to improve the relevance of results. The authors proposed the deep and broad mining of Wikipedia and Yahoo! Answers for extracting topic words. These techniques essentially extract related keywords from various forms of information such as titles, contents and links in Wikipedia articles, and chosen answers and subjects from Yahoo! Answers, as the topic words. Ye et al. (2009), on the other hand, looked at ways for creating summaries of varying length, which can be used for definitional question answering, using the different sections of Wikipedia articles (e.g., infobox, outline). The technique also uses the anchored text (i.e., links) in articles to explore potential associations between different sentences. Buscaldi and Rosso (2006) employed Wikipedia

[4] `http://demos.inf.ed.ac.uk:8080/qualim`

for validating answers and reformulating questions for a question answering system called QUASAR. The techniques can only cope with questions involving names and definitions. For the first task, hand-crafted patterns are used to extract names from the candidate answers generated by QUASAR, and the presence of Wikipedia articles matching the names is used as an indicator of the validity of the corresponding answers. The second task deals with questions (e.g., *"Which fruits have vitamin C?"*) that do not match any of the existing patterns. The category words (e.g., *"fruit"*) are first extracted from the questions using part-of-speech information and other heuristics such as the capitalisation of words. The corresponding category page on Wikipedia (e.g., `Category:Fruit`) is extracted and the titles of all articles (e.g., *"mangoes"*, *"apples"*) listed in the page are identified. These titles are used to construct a new query for QUASAR.

SYSTEM ARCHITECTURE

In this section, we will discuss the three main components of our enquireMe system: (1) the extraction of question-answer (QA) pairs from community-driven question answering websites such as Yahoo! Answers and Answers.com, (2) the extraction of weighted keyphrases from user inputs, and (3) the scoring and ranking of QA pairs based on the overlapping of keyphrases and several other criteria. Each of these components play a role in addressing the three challenges described in the introduction section. The colloquial nature of the QA pairs allows the system to generate more natural responses as compared to extracting sentences or paragraphs from other types of Web content. The decaying and reinforcement of the weights of keyphrases by the system provides a systematic way of maintaining interaction context, which is crucial to a contextual question answering system. The use of keyphrase overlaps and other criteria such as user votes allows enquireMe to locate answers using a simple and flexible approach that does not attempt deep language processing, and hence allows for unrestricted inputs.

Before moving into the details, we first describe how these components interact and work together as a system. Firstly, the QA pairs are extracted offline and are kept in a local storage for use by enquireMe. Whenever the system is posed with an input, it extracts keyphrases from that input and assigns weights to them that represent the amount of 'content' that they carry. The more content a keyphrase carries, the more discriminating power it has for representing the input that it appears in. The weights of these phrases fluctuate over time depending on their recurrence over the course of an interaction. These weighted keyphrases and a few other criteria (e.g., user votes) are then used to retrieve and score candidate QA pairs. The answer component of the top-scoring QA pair is finally used as the system's response to the user's input. The interface that displays the system's response also allows the user to `like` or `dislike` the answer that he or she receives. This feedback is stored by enquireMe and used to influence its scoring and ranking process.

QA Pair Extraction

Insert figure [component1.eps] here

FIG. 1. The process of extracting QA pairs from community-driven question answering websites.

In the first component, the QA pairs used in the version of enquireMe evaluated for this paper are extracted from Yahoo! Answers. The querying and extraction of data from Yahoo! Answers are performed using the API[5] provided by Yahoo!. A QA pair is simply the pairing of a question posted by a human together with the possible answers contributed by other volunteers. The process of posting questions and contributing answers together with other quality control activities are the backbone of these services. Our choice of Yahoo! Answers is motivated simply by the availability of APIs to ease the implementation process. We could equally well obtain the QA pairs from any community-driven question answering website such as Answers.com. Many of these websites, however, do not permit the non-commercial, automated extraction of data from their sites, and hence, prevent their use in this kind of work. We implemented a GUI that allows the administrator of enquireMe to easily provide *seed concepts*, in the form of phrases or words, to extract QA pairs from Yahoo! Answers. Figure 1 shows an overview of the process of extracting QA pairs from the Web and storing them in a database. As a way to broaden the coverage of the QA pairs to better cope with conversations about related topics, we have also included an automatic keyphrases extraction step. This step identifies the keyphrases, which can either be a phrase or a word, in the questions and answers of the QA pairs to trigger an additional iteration of QA pair extraction. For example, given the QA pairs that correspond to the seed concept *"neck pain"*, the keyphrase extraction step will identify other related concepts such as *"whiplash"*, *"back pain"*, etc. In addition to the questions and answers, we also record three other forms of metadata about each pair (if available), namely, the category, the date of posting and answering, and the URL to the source. These metadata are not used in this version of enquireMe. They can however be taken into consideration during the scoring and ranking process to improve the quality of the responses, and we do expect to use them in future versions of the system. The category column for instance can be used to eliminate candidate answers if their categories are only distantly related to the current conversational theme.

User Input Analysis

In this section, we look at how phrases and words are extracted from user inputs and assigned with weights that reflect their content bearing property. A description of the pronoun resolution technique used by enquireMe is also provided.

Insert figure [components_2_3.eps] here

FIG. 2. The process of retrieving, scoring and ranking QA pairs using context information derived from user utterances to determine the best response for a user input.

In the second component, phrases are extracted from user inputs and then weighted as shown in Figure 2 to determine the keyphrases. A user input, in this work, can either refer to a natural language question (e.g., *"What is whiplash?"* or simply a statement (e.g., *"My neck hurts."*). The input is first analysed for phrases to produce the set $X = \{x_1, ...\}$. Attached to each phrase or word $x_i \in X$ is a weight denoted as w_{x_i}. To achieve this, the FastTag part-of-speech tagger[6] is used to decorate the input with part of speech. This information is then used to chunk nouns and adjectives to create noun phrases. A simple regular expression $\texttt{(Adj)}^*$ $\texttt{(Noun)}^+$ is used for this purpose. Our choice of this rather simplistic approach to phrase extraction is motivated solely by speed.

Weight derivation: Next, keyphrases are identified by assigning weights to the

[5] http://developer.yahoo.com/answers/

[6] `http://markwatson.com/opensource`

phrases and words extracted from the inputs. In this work, we consider a keyphrase as a word or phrase that is content-bearing, or in other words, not a function word. The weights, which represent the content-bearingness of words, are assigned to phrases or words based on how accurate their occurrences can be modelled using the Poisson distribution (Church and Gale, 1995). This established approach can be explained as such: if we assume that a document is merely a bag of words with no interesting structure, then the words in that document are essentially the result of a random (i.e., Poisson) process. In this sense, we should be able to quite accurately describe the occurrences of randomly occurring words in a document using the Poisson distribution $Pr(k;\theta)$ where k is the number of times a word occurs in a document and $\theta = f/N$ is the known average rate with f as the number of times a word occurs in a collection of N documents. On the contrary, if a word cannot be predicted with relative accuracy using Poisson, then that word's occurrence is not random (i.e., an occurrence with a purpose). To determine this inability of Poisson to predict the occurrences for non-random or content-bearing words, we compute the probability of a document having at least one instance (i.e., $k > 0$) of the word, which is $Pr(k > 0;\theta)$ or $1 - Pr(k = 0;\theta)$ where $Pr(k = 0;\theta) = \exp(-\theta)$. The inability to accurately predict content-bearing keyphrases using Poisson, manifesting as large ratios between Poisson predictions and actual observations, has been demonstrated to be robust across different collections (Church and Gale, 1995).

Example of weight derivation: Using x_1 =*"whiplash"* and x_2 =*"any"* from our collection of about $N = 36,419$ Wikipedia articles as an example, the word *"whiplash"* occurs $f_{x_1} = 37$ times in $n_{x_1} = 12$ documents while *"any"* appears $f_{x_2} = 1169$ times in $n_{x_2} = 824$ documents. Keywords, such as *"whiplash"*, are important in the retrieval of QA pairs because they can be used to pick out very specific subsets of a collection. The occurrences of non-keywords such as *"any"*, however, behave almost like chance (i.e., Poisson). Based on Poisson, the probability of encountering at least one of the 37 and 1169 occurrences of *"whiplash"* and *"any"* is $0.001 = 1 - exp(-37/36419)$ and $0.03 = 1 - exp(-1169/36419)$, respectively. According to observations, the probability of encountering *"whiplash"* in a document is $0.0003 = 12/36419$, and the word *"any"* has a higher probability at $0.0226 = 824/36419$. By measuring the predictions' deviation from the actual observations, we can see that the word *"whiplash"* has a large *prediction deviation ratio* at $3.33 = 0.001/0.0003$ as compared to the word *"any"* at only $1.33 = 0.03/0.0226$. In other words, the occurrence of the word *"any"* can be predicted much more accurately as compared to *"whiplash"*.

Weight assignment: This deviation ratio based on Poisson, which has been successfully applied to identifying important terms to describe Web services (Liu and Wong, 2009), is used in this work to assign weights to the phrases and words extracted from an input to reflect their content-bearing properties. This deviation ratio ρ in its current form is less suitable to be used as weight since ρ can, theoretically, reach a very large positive number. In order to restrict the weights w_{x_i} of phrases or words x_i to within a known bound, we convert ρ into w_{x_i} as follows:

$$w_{x_i} = \exp(-1/\rho) \tag{1}$$

where $\rho = \hat{p}_{x_i}/p_{x_i}$ if $\hat{p}_{x_i} > p_{x_i}$ or $\rho = p_{x_i}/\hat{p}_{x_i}$ otherwise, $\hat{p}_{x_i} = Pr_{x_i}(0;\theta_{x_i})$ and $p_{x_i} = n_{x_i}/N$. In this work, the English Wikipedia dump from 16 September 2010 is used to obtain the word frequency f, document frequency n and the collection size N for computing the deviation ratio. Since the constraint discussed above ensures that $\rho \geq 1$, we are able to force w_{x_i} to remain within $\exp(-1) \leq w_{x_i} < 1$. Intuitively, the larger the deviation ratio ρ, the more difficult it is for us to model the phrase or word using Poisson. This in turn translates to the increase in likelihood that the phrase or word is content-bearing. A very important point to note is that *this principled approach eliminates the need to remove stopwords and to assign weights arbitrarily*. Using this approach, content bearing verbs, noun phrases, adverbs and adjectives are systematically weighted higher than other parts of speech such as interjection, determiners and conjunctions.

Pronoun resolution: In addition to keyphrase weighting, a simple pronoun resolution mechanism is included during user input analysis. The technique that we employ resolves pronouns encountered in subsequent inputs to the most prominent nouns extracted from previous inputs, which shares some resemblance with the *backward looking center* approach in *Centering Theory* (Mitkov, 2001). The algorithm works in this way: For each pronoun detected during input analysis, we iterate through the set of previously weighted phrases and words (i.e., contexts). The context word with the highest weight that also happens to be a noun as well as a seed concept is considered as the antecedent referred to by the pronoun. This approach is limited by the coverage of the list of seed concepts that the administrator of the system provides during QA pair extraction. This restriction, however, makes sense in that pronouns cannot be resolved to concepts that the system has no knowledge of (i.e., unavailability of QA pairs about certain concepts). More advanced resolution techniques from the literature (Mitkov, 2001) can potentially be used in the future.

QA Pair Retrieval and Ranking

In the third component, the weighted keyphrases in X, and a context set which is essentially the keyphrases extracted from previous inputs, are used to retrieve candidate QA pairs for further processing.

Context management and weight revision: Context, in this work, is essentially a set of weighted keyphrases, denoted as Y, which is maintained during the course of an interactive question answering session. Context management is the process of adding new keyphrases from recent inputs into set Y and revising the weights of all elements in Y to reflect their recurrence. The elements in the context set Y are used to retrieve and score candidate question-answer pairs in the subsequent steps.

Similar to X, each element y_i in the context $Y = \{y_1, ...\}$ is accompanied by a weight denoted as w_{y_i}. This set Y contains all previous keyphrases collected over the course of a conversation. If the current input is the 5-th utterance by the user, then Y would contain all the keyphrases extracted from the previous four inputs and X would contain only the keyphrases from the most recent 5-th input. At each turn of a conversation where the user provides an input, the keyphrases in X are assimilated into set Y to create a revised set Y'. We adapted the general exponential decay to be word-class sensitive to compute the weight $w_{y'_i}$ for each $y'_i \in Y'$ as shown below in Equation 2.

$$w_{y'_i} = \begin{cases} w_y \exp(-t\lambda\alpha) & \text{if}(y'_i \notin X \wedge y'_i \in Y \\ & \wedge y'_i = y \wedge y \in Y) \\ w_x & \text{if}(y'_i \in X \wedge y'_i \notin Y \\ & \wedge y'_i = x \wedge x \in X) \\ w_x + w_y \exp(-t\lambda\alpha) & \text{if}(y'_i \in X \wedge y'_i \in Y \\ & \wedge y'_i = x = y \\ & \wedge x \in X \wedge y \in Y) \end{cases} \tag{2}$$

In Equation 2, t is the t-th turn of the user starting from $t = 1$. As for the decay factor λ, the larger the value, the faster the decay, which contributes to weights that approach zero. λ is set to 0.8 during our experiments. Intuitively, if the keyphrase a_i appears only in the past context (i.e., set Y), then it should have less influence on which answer to be used as the response as compared to those keyphrases appearing in both X and Y or just X. α, on the other hand, augments the decay factor depending on the parts of speech of the phrases. The α value has been set to the following empirically: nouns is set to 0.25, adjectives and adverbs to 0.75, verbs to 1.25, and others to 2.50. In other words, we value nouns more followed by adjectives and adverbs and so on for their differing roles when it comes to discriminating and zooming in on relevant QA pairs for candidate answers.

Example of weight revision: As an example, assume $X = \{$*"whiplash"*$\}$ as the only keyphrase extracted from the first user input at turn $t = 1$. Using Equation 1, the keyphrase would be assigned an initial weight of $\exp(-1/\rho_{whiplash})$. The revised set Y' at $t = 1$ is equal to X since $Y = \emptyset$. In the next user turn $t = 2$, assuming the new input set as $X = \{$*"treatment"*$\}$, the context set would be $Y = \{$*"whiplash"*$\}$. These two sets are combined into $Y' = \{$*"whiplash","treatment"*$\}$ and the weights of its elements are revised according to Equation 2. Since *"whiplash"* occurred in Y and not X at turn $t = 2$, its weight is revised to become $\exp(-1/\rho_{whiplash}) \times \exp(-2 \times 0.8 \times 0.25) = 0.67 \exp(-1/\rho_{whiplash})$. As for *"treatment"*, the noun is assigned with the initial weight of $\exp(-1/\rho_{treatment})$ using Equation 1, since the word appears in X and not in Y at turn $t = 2$ as defined in Equation 2. This process of extracting words to form set X, carrying forward context set Y from previous turns, and revising the weights in Y' takes place for every input provided at every turn t.

Criteria for scoring and ranking QA pairs: Next, the system performs a simple structured query to the database (e.g., SQL) to obtain a set of all QA pairs $QA = \{qa_1, ..., qa_n\}$ containing at least one keyphrase from Y' in the questions. The QA set is then passed to the scoring function described in Algorithm 1. Each pair $qa_i = (q_i, a_i)$ is a tuple of question q_i and answer a_i. Each pair qa_i is assigned a score s_{qa_i} based on the following four criteria, where the details are outlined in Lines 3-28 in Algorithm 1. These criteria are:

- The more keyphrases from Y' that appear in the QA pairs, the higher the pairs' scores will be. Since the elements of Y' are unique, the weight of each keyphrase will contribute to a QA pair's score at most once. In other words, even though a QA pair qa_i contains multiple occurrences of $y'_j \in Y'$, the corresponding $w_{y'_j}$ will only be considered once during the computation of s_{qa_i}.
- Keyphrase matches in the questions are scored more than matches in the answers. This bias is implemented as two constants β_q and β_a, which are set to 0.7 and 0.3, respectively, where these two numbers are determined empirically. This and the previous criteria are implemented as Lines 7-19 in Algorithm 1.
- Pairs that have been used previously as responses are penalised. This criterion is implemented as Lines 20-24 in Algorithm 1.
- The higher the vote of a QA pair by the users, the more we score the corresponding answer. This criterion is implemented as Lines 25-26 in Algorithm 1. The set Y' at each turn of a particular conversation is used to identify that conversation at that point in time, which we refer to as a *conversational context* C. Every time a user clicks on the `like` or `dislike` button of an answer a_i of qa_i at turn t of a conversation, a record $r = (qa_i, C, v)$ is created comprising the context at that point in time C

together with qa_i and the user's vote v. The vote v is either initialised to 0.1 if liked, or -0.1 if disliked. If a record for the same C and qa_i already exists, the corresponding vote v is incremented or decremented accordingly by 0.1 and the record is updated. The collection of all records kept by enquireMe is denoted as R.

Algorithm 1 The scoring of QA pairs.

```
**input**: turn $t$, input question $iq$, sets $Y'$ and $QA$.
Initialise all scores $s_{qa_i}$ for all pairs $qa_i \in QA$ to $0$.
**for** each $qa_i = (q_i, a_i) \in QA$ **do**
    Set $s_m$, $s_r$ and $s_v$ to $0$, which are the keyphrase matching score, the reuse score and the user vote.
    Set $c_q$ and $c_a$ to $0$, which are the sums of the weight of phrases that appear in $q_i$ and in $a_i$, respectively.
    Set $o_q$, $o_a$ to $0$, which are numbers of phrases that appear in $q_i$ and in $a_i$, respectively.
    **for** each $y'_j \in Y'$ **do**
        **if** $y'_j$ occurs in $q_i$ **then**
            $c_q \leftarrow c_q + w_{y'_j}$
            $o_q \leftarrow o_q + 1$
        **end if**
        **if** $y'_j$ occurs in $a_i$ **then**
            $c_a \leftarrow c_a + w_{y'_j}$
            $o_a \leftarrow o_a + 1$
        **end if**
    **end for**
    $c_q \leftarrow c_q \times o_q/|Y'| \times \beta_q$
    $c_a \leftarrow c_a \times o_a/|Y'| \times \beta_a$
    $s_m := \exp(-1/(c_q + c_a))$
    **if** $a_i$ was used as responses previously **then**
        $s_r := 0.5$
    **else**
        $s_r := 1$
    **end if**
    Find all records of voting by users for $qa_i$ from $R$, and pick the tuple $r \in R$ that contains vote $v$ with the conversational context $C$ that maximises the overlap $|C \cap Y'|$.
    $s_v := \exp(-1/\exp(v)) \times ((1 + |C \cap Y'|)/(1 + |C|))$
    $s_{qa_i} := (0.6 \times s_m) + (0.2 \times s_r) \times (0.2 \times s_v)$
**end for**
**output**: $a_i$ of $qa_i \in QA$ with the highest $s_{qa_i}$
```

EXPERIMENTS

In this section, we discuss the performance of our contextual question answering system enquireMe against three other systems, namely, HONqa (health-specific), START and QuALiM, in terms of the accuracy of the answers generated. These experiments provide insights into the upper and lower bounds for investigating the performance of enquireMe in the context of the state of the art in the field, despite the absence of directly comparable contextual question answering systems in health.

Experimental Setup

We prepared three sets of data for testing the four systems. The first set (S1) is benchmark data in the health domain obtained from the website WebMD, similar to the one used by Olvera-Lobo and Gutierrez-Artacho (2011), which comprises 150 definitional questions in the form of *"What is X"*. The second set (S2) extends the *'one-off'* questions in the first set to include follow-up questions to simulate an interactive environment for contextual question answering. This second set contains a total of 274 questions about the first 99 medical concepts (those that starts with `A` up till `N`) from the complete set of 150 concepts. Each of the *"What is X?"* questions for these 99 medical concepts is followed by one or two additional questions. If the concept X is a medical condition (i.e., disease or disorder), which accounts for 77.78%, two more questions *"What causes it?"* and *"What are its treatments?"* are added as shown in Figure 3.

Insert figure [lungcancer.eps] here

FIG. 3. Interactive questions from set S2 for the medical concept *"lung cancer"*.

For the remaining concepts about medical procedures (e.g., *"circumcision"*), treatment options (e.g., *"chemotherapy"*), medical devices or instruments (e.g., *"cochlear implant"*) and drugs (e.g., *"ephedra"*), which account for 21.21% of the 99 concepts in set S2, the additional question of *"What are its uses"* is added as shown in Figure 4.

Insert figure [ephedra.eps] here

FIG. 4. Interactive questions from set S2 for the medical concept *"ephedra"*.

The remaining one medical concept *"abortion"* does not fit into any of the two categories and as such, was not expanded with any interactive questions.

The third set (S3) contains 41 cross-domain questions from the TREQ 2004 QA track (Voorhees, 2004) which are grouped into 11 series about 11 different targets in various domains, where each question in a series asks for some information about the target. These 11 targets are *"Hale Bopp comet"*, *"Agouti"*, *"prions"*, *"Horus"*, *"Jar Jar Binks"*, *"cataract"*, *"Concorde"*, *"Tale of Genji"*, *"quark"*, *"boll weevil"* and *"space shuttle"*. The only configuration required for enquireMe for this experiment is the automatic extraction of QA pairs from Yahoo! Answers using the concepts that the questions in the three sets pertain to as seed keyphrases (i.e., 150 concepts from sets S1 and S2, and 11 concepts from S3), using the approach described in the question-answer pair extraction section. As for the START system, an online interface is available at `http://start.csail.mit.edu`[7]. Since this online interface is a front-end to a server-side application, no configurations can be made with regard to START for this experiment. Moreover, no up-to-date information about the system's knowledge base is made available. As such, we cannot be sure if START contains the knowledge necessary for answering questions about all the 161 concepts used in this experiment. However, considering that START indexes information on the Medline website for the health domain and Wikipedia for general knowledge, and the fact that it has been live on the Web since 1993, we would assume that START has the knowledge to handle most if not all of the questions in the three sets.

[7] Accessed in November 2011

Performance of Question Answering using Health Data from WebMD

Table 1 summarises the results achieved by the four systems using the different datasets. In the table, rows 1-3 show the results for the three systems HONqa, QuALiM and START using questions from set S1. These results were extracted from the survey paper by Olvera-Lobo and Gutierrez-Artacho (2011). We were unable to repeat the experiment due to the unavailability of QuALiM at `http://demos.inf.ed.ac.uk:8080/qualim` (last accessed 6 March 2012). The `total correct answers` in Table 1 is actually the sum of the values in the `total correct answers` and `total inexact answers` rows in the survey paper. In the survey paper, correct answers are defined as those that answered the question adequately with two criteria: (1) using less than 100 words and (2) did not contain irrelevant information. Inexact answers, on the other hand, are correct responses that did not meet the two criteria. In order to remove any biasness against HONqa, QuALiM and START, both correct and inexact answers in the survey are considered simply as correct in this paper. Moreover, determining if the answers contain irrelevant information or otherwise varies between assessors, and in this paper, we were unable to duplicate exactly the survey authors' interpretation of relevance. For instance, is the answer containing a line on the treatments of lung cancer irrelevant to the question of *"What is lung cancer?"*. As such, eliminating these two criteria allows us to judge the correctness of answers generated by the systems for the other datasets less subjectively.

TABLE 1. The performance of enquireMe against HONqa, QuALiM and START determined using three different datasets. Insert table [experimentrevised.eps] here

Row 5 shows the results from evaluating enquireMe using the 150 *"What is X?"* questions from set S1. The answers produced by enquireMe were assessed in the same way as the survey authors Olvera-Lobo and Gutierrez-Artacho (2011) would for rows 1-3 with two caveats. First, unlike the survey authors which recruited health professionals, we assessed the answers generated by enquireMe based strictly on the definitions provided on WebMD. Second, to remove any doubt that we are biased towards enquireMe, we only assessed the first answer produced by enquireMe for each question. In other words, if the first answer for any question is incorrect, then enquireMe has failed to respond correctly to that question. This is in contrast to the way the survey authors Olvera-Lobo and Gutierrez-Artacho (2011) evaluated HONqa, QuALiM and START where the top five answers produced were considered. Not all questions, however, have five or more answers. This explains the unequal numbers of answers that were assessed in the `average number of answers assessed per question` column for rows 1-3 in Table 1. As can be seen from the `precision` column in the table, enquireMe achieved the highest percentage at 94% (row 5), while the three systems HONqa, QuALiM and START fared between 55% to 89%. Due to the involvement of different assessors, we do not categorically claim that enquireMe outperforms the other three systems. However, these numbers are good indicators that enquireMe would likely achieve a high precision if all four systems were to be evaluated by the same assessors.

Performance of Contextual Question Answering using Extended WebMD Data

Row 6 shows the performance of enquireMe in an interactive setting using the 247 health-related questions in set S2. In this particular experiment, enquireMe achieved a precision of about 87% for contextual question answering in health. A point worth noting is that no comparisons can be made between enquireMe and the state-of-the-art using questions in set S2 because there is no publicly accessible, working contextual question answering systems on the Web for the health domain.

Performance of Contextual Question Answering using TREC Data

In rows 4 and 7, we reported the results from evaluating START and enquireMe using cross-domain questions from set S3. HONqa is not included because it specialises in health information, and QuALiM was not available online. The 41 questions in set S3 are about 11 randomly selected targets from the TREC 2004 collection, namely, *"Hale Bopp comet"*, *"Agouti"*, *"prions"*, *"Horus"*, *"Jar Jar Binks"*, *"cataract"*, *"Concorde"*, *"Tale of Genji"*, *"quark"*, *"boll weevil"* and *"space shuttle"*. During the experiment, the START system seems to be able to cope with successive, related (i.e., interactive) questions. For instance, we first asked *"What is cataract?"*, to which it provides a paragraph from Wikipedia as the answer. Subsequently, after pushing the browser's back button and posting the next question *"What causes it?"*, the START system managed to resolve *"it"* to *"cataract"* and provides unstructured as well as bulleted contents from the American Medical Association website, MedlinePlus and the Merriam-Webster Dictionary as answers. However, to be fair to START (since it was never publicised as a contextual question answering system), we replaced all pronouns in the 41 questions in set S3 with the target every time a question is posed to START. The interactive questions are posed as-is to enquireMe. To illustrate, the questions for the target *"Jar Jar Binks"* from this set are shown in Figure 5. When these questions were posted to START, the pronouns such as *"his"* and *"he"* were replaced manually with the target *"Jar Jar Binks"*. The precision values in rows 6 and 7 demonstrate that enquireMe is likely as able to cope with cross-domain contextual question answering as it does with health-related questions. START, on the other hand, fared poorly with the 41 cross-domain interactive questions (row 4). A possible cause to this is START's focus on well-formed inputs or information that can be easily structured into triples (e.g., *"What is X?"*, *"Who invented X?"*) as discussed in the related work section. As many of the questions from set S3 can be quite complex in their surface structure, START was less able to 'understand' such inputs properly.

Insert figure [jarjarbinks.eps] here

FIG. 5. The questions from set S3 for the target *"Jar Jar Binks"*.

Limitations and Discussions

Looking at the results in Table 1, enquireMe on set S1 (row 5) appears to be the best performing question answering system as compared to rows 1, 2 and 3. We, however, avoided making this conclusion considering that the responses generated by enquireMe were examined by assessors different from those who evaluated the other three systems in the survey paper by Olvera-Lobo and Gutierrez-Artacho (2011), even though we used the same set of questions from S1. There are three things that we have to keep in mind when interpreting the results reported in the table. For one, while there has been some work done on contextual question answering, much of this work remains in the realm of research. There are no actual working systems out there on the Web that we can compare enquireMe against. Second, there are no other datasets containing interactive questions besides the TREC 2004 data with which we can evaluate enquireMe. Since our focus is to tune enquireMe towards the health domain, the lack of any standard evaluation data for contextual question answering in health can be a setback in our evaluation effort. Third, all the systems included in our experiments use different sources of answers. We or even the authors of the survey paper Olvera-Lobo and Gutierrez-Artacho (2011) were unable to determine with absolute certainty if the performance reported was the result of more superior techniques or simply poor answer coverage.

Discussion about state of the art: From Table 1, we can observe that enquireMe performs reasonably well, above the 85% mark, in the context of contextual question answering, both in the health domain (i.e., using set S2 in row 6) as well as cross-domains (i.e., using set S3 in row 7). If we were to look at cross-task comparison in the health domain, the performance of enquireMe in the more challenging task of contextual question answering (row 6 at 86.86%) is not too far off from START's performance in non-contextual question answering (row 3 at 88.46%). This result is promising considering the complexity behind the START system, which represents the state of the art in the field, as described in the related work section, and the fact that contextual question answering faces additional challenges such as managing context and resolving pronouns. The challenge of contextual question answering becomes more evident when we look at the precision drop faced by enquireMe from 94.00% in row 5 (i.e., question answering using 150 *"What is X?"* questions about 150 medical concepts) to 86.86% in row 6 (i.e., 274 interactive questions about 99 medical concepts). Moreover, a look into the literature shows that the achievement of some of the best performing state-of-the-art cross-domain contextual question answering systems from the TREQ 2004 QA track falls short of 80% (Voorhees, 2004). The submission by Language Computer Corp. produced the most number of correct answers for factoid questions at 77%. An interesting point to note about this TREC submission was that it achieved an accuracy of 83.9% for the first questions in their respective series and 74.4% for the remaining questions. This drop in performance when faced with interactive questions is a challenge that is also faced by enquireMe as discussed above.

Error analysis: Finally, we look at the causes of the incorrect answers reported during the experiments involving enquireMe on set S2 (row 6) and set S3 (row 7), as shown in Table 2. This helps us to identify the areas that require improvements. In row 6, 72.22% of the 13.14% incorrect answers are attributed to the absence of relevant QA pairs from the source. As for the cross-domain interactive questions in row 7, 80% of the 12.20% errors are caused by the lack of answers from the source. One way to address such errors is to diversify the sources of QA pairs. As for 19.44% of the 13.14% errors in row 6, they would likely be interpreted as inexact in the context of the survey paper (Olvera-Lobo and Gutierrez-Artacho, 2011). These answers are marked as incorrect in our assessment because, instead of succintly addressing the main concerns in the questions, they address different aspects such as treatments, causes and symptoms. The inexact answers generated by HONqa, QuALiM and START are still considered as correct in this paper. The remaining small percentage of incorrect answers at 8.33% of the 13.14% errors in row 6 highlight the importance of QA pair ranking and enquireMe's ability to 'float' the correct answers to the top. These 8.33% errors occurred because the correct answers were not ranked first although they did fall within the top 3. Similarly, the remaining 20% of the 12.20% incorrect answers in row 7 are attributed to the ranking of answers. For an immediate performance boost, we can simply assess the top 3 answers for correctness, an approach used by the survey authors Olvera-Lobo and Gutierrez-Artacho (2011). However, we avoided doing this since it was important for us to explore options to improve the ranking instead of short-term performance increase. Moreover, considering that recall is not a concern in this task, our longer term goal is to have a more fine-grained assessment of the ranking algorithm to discover ways to solely improve the precision. The breakdown

TABLE 2. The causes of incorrect answers generated by enquireMe based on questions in set S2 (row 6) and S3 (row 7). Insert table [experimenterroranalysis.eps] here

of the causes of errors that arise with the use enquireMe without interactivity is also shown in Table 2. The table shows that the main causes of incorrect answers in non-interactive question answering are the lack of answer from the source and inexact answers, which contribute to about 88.88% of the total 9 errors. This trend is similar to those of set S2 and S3 (i.e., interactive question answering using extended WebMD and TREC data). First and foremost, this shows that a question-answer pair collection with good coverage is crucial to the performance of both interactive and non-interactive question answering. The granularity of the question-answer pairs is also another factor. If the individual pairs describe too many aspects (e.g., treatment, symptom, cause) of a certain concept (e.g., lung cancer), then the tendency of having inexact answers will be higher.

Example Interaction

enquireMe was presented as a demo (Wong et al., 2011) at the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011) in Glasgow. As a result of this exposure, a number of anonymous interactions was recorded. In this section, we examine an actual interaction that took place in December 2011 between an anonymous Web user and enquireMe. Figure 6 shows the first four rounds of exchanges between the system (labelled as `SYSM`) and the user (labelled as `USER`). Note that this figure reads from bottom to top. The version of enquireMe that the anonymous Web user interacted with has close to $80,000$ QA pairs from Yahoo! Answers about 195 medical and health-related concepts which include headache, the 150 from WebMD as described in the experimental setup section, and more.

Insert figure [exampleconversation.eps] here

FIG. 6. An example interaction between a Web user and enquireMe about *"headache"* (reads from bottom to top).

The user started the conversation at time `13:31:00` by saying *"i'm having a bad headache with tight muscles around my neck."*. The system analysed this input and produced a set of keywords with weights as shown in Figure 7. A point worth noting is that in this prototype, nouns are labelled as `ENTITY`, verbs as `ACTION`, adverbs and adjectives as `DESCRIPTOR`, pronouns as `REFERENT` and everything else as `NULL`. The `weight` column shows the weight that is derived based on the deviation from the Poisson distribution, ρ, as described in the user input analysis section. The phrases and words extracted from this input were found to best match the QA pair with the question *"What is the best pain killer for a headache caused by tight neck and shoulder muscles?"* from Yahoo! Answers (`http://answers.yahoo.com/question/index?qid=20100117105008AAOcpQR`). The first sentence from the answer component of this QA pair was then produced as the system's response 8 seconds later at time `13:31:08`.

Insert figure [exampleconversation_weight1.eps] here

FIG. 7. The context phrases and words together with their weights for the input posted at time `13:31:00`.

At time 13:32:29, the user adds *"probably my sleeping position last night."*, which produces the weighted phrases and words shown in Figure 8. For this input, the QA pair with the question *"Please help! I slept very poorly last night and today I suffer nearly paralyzing neck pain on my left side!!?"* (http://answers.yahoo.com/question/index?qid=20110620092813AAfsscM) was found to be the best matching one, where the corresponding answer was used as the system's response. The following user input *"how can i get rid of it?"* at time 13:33:20 demonstrates the system's pronoun resolving ability where *"it"* was resolved to *"headache"*, which contributes to the reinforcement of the word's weight based on the decay model as shown in Figure 9.

Insert figure [exampleconversation_weight2.eps] here

FIG. 8. The context phrases and words together with their weights for the input posted at time 13:32:29.

Insert figure [exampleconversation_weight3.eps] here

FIG. 9. The context phrases and words together with their weights for the input posted at time 13:33:20.

Figure 10 shows the further reinforcement of the keyword *"headache"* due to its recurrence as well as the introduction of new relevant keywords *"pain"* and *"killer"*, all of which were used to search for the most relevant QA pair. This new input at time 13:34:27 was found to match the QA pair that has the question *"Having lower back pains, and neck pain, unable to get a good night sleep, now today I woke up with a headache,?"* (http://answers.yahoo.com/question/index?qid=20091019222107AA0uXUv). The first sentence of the answer of this QA pair which explains a remedy involving aspirin for headache was presented as the response by the system.

Insert figure [exampleconversation_weight4.eps] here

FIG. 10. The context phrases and words together with their weights for the input posted at time 13:34:27.

CONCLUSION AND FUTURE WORK

The way the users interpret, react to and benefit from health information accessed on the Web has a lot to do with the delivery mechanism. Many studies have shown that natural language interfaces such as question answering and conversational systems allow information to be accessed and understood easier by users who are unfamiliar with the nuances of the delivery mechanisms (e.g., keyword-based search engines), or have limited literacy in certain domains (e.g., unable to comprehend health-related content due to terminology barrier). In particular, the ability to clarify the different aspects of a piece of information and to incrementally represent one's information needs are highly desirable. For this reason, search interfaces are moving towards supporting more natural dialogue-like interaction to access opinions as well as answers to questions (Hearst, 2011). The long term goal of our research is to encapsulate and offer these capabilities via a health contextual question answering system.

In this paper, we report our system enquireMe that harnesses community-driven question-answer pairs on the Web for contextual question answering. More specifically, question-answer pairs, which are contributed by users from all over the World, are used by enquireMe to assist end-users during their interactive information seeking exercise regarding their personal health. Despite our focus on health-related information, our approach to contextual question answering is non domain-specific. The system uses a decay model combined with keyphrase extraction and weighting to systematically match and score the question-answer pairs, and deliver the top scoring answer as a response to the user's input. Our experiments showed that the performance of enquireMe is comparable with the state of the art question answering systems such as START as well as those interactive systems from TREC. This performance is promising considering the complexity of the state of the art question answering systems that were designed specifically to handle wh-questions.

There is additional work required on certain aspects for improving enquireMe. The more pressing one would be to have a mechanism as part of the context management to identify topics from inputs and deals with topic switch explicitly. The system in its current form implicitly assumes that all inputs from a single session pertain to one main topic (e.g., *"neck"* in the health domain). The effectiveness of enquireMe in dealing with information needs that move across topics (e.g., enquire about *"neck"* first and then move on to *"leg"*) or even other domains remains untested. The choice of phrase extraction will also have an impact on the performance of enquireMe. In the current version of the system, we use a lightweight part of speech tagger and regular expression patterns for identifying phrases. For this, we will experiment with different existing tools to determine the most suitable one that provides a good balance on accuracy and speed. Currently, different variants of the same word are not collapsed, causing diffusion of weights and improper decaying of certain keyphrases. String matching algorithms as well as relatedness or similarity measures will be used to aggregate similar phrases in different lexical forms. The rating of usefulness of the system generated responses will also help to provide positive feedback to the system. At the moment, only the first sentence of the answer component of the top-ranked QA pair is used as the system's response. Text summarisation may be considered for combining answers from the top n ranked answers.

In addition to the technical aspect of enquireMe, the coverage and quality of the question-answer pair collection can have great influence on the veracity of the answers generated by the system. Due to the emphasis on the trustworthiness and currency of health content, it is likely that the overall experience of interacting with enquireMe could be improved using expert-produced question-answer pairs from sources that are verified by authorities such as Health on the Net Foundation (`healthonnet.org`).

Acknowledgment

We thank the reviewers for their valuable comments. This work is partially supported by the Australian Research Council and Real Thing Entertainment Pty. Ltd. under Linkage grant number LP110100050.

REFERENCES

Allen, J., Ferguson, G., and Stent, A. (2001). An architecture for more realistic conversational systems. In *Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI)*, pages 1–8, Santa Fe, New Mexico, USA.

Athenikos, S., Han, H., and Brooks, A. (2009). A framework of a logic-based question-answering system for the medical domain (loqas-med). In *Proceedings of the ACM Symposium on Applied Computing (SAC)*, pages 847–851, Hawaii, USA.

Bennett, K., Reynolds, J., Christensen, H., and Griffiths, K. (2010). e-hub: An online self-help mental health service in the community. *The Medical Journal of Australia*, 192(11):S48–S52.

Buscaldi, D. and Rosso, P. (2006). Mining knowledge from wikipedia for the question answering task. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 727–730, Genoa, Italy.

Church, K. and Gale, W. (1995). Inverse document frequency (idf): A measure of deviations from poisson. In *Proceedings of the ACL 3rd Workshop on Very Large Corpora*, pages 121–130.

Cruchet, S., Gaudinat, A., Boyer, C., Morandi, A., and Marino, D. (2009a). Multilingual question/answering system applied to trusted health information. In *Proceedings of the 22nd International Conference of the European Federation for Medical Informatics (MIE)*, pages 1–5, Sarajevo.

Cruchet, S., Gaudinat, A., Rindflesch, T., and Boyer, C. (2009b). What about trust in the question answering world? In *Proceedings of the AMIA Annual Symposium*, pages 1–5, San Francisco, USA.

Hearst, M. (2011). Natural search user interfaces. *Communications of the ACM*, 54(11):60–67.

Inoue, M., Matsuda, T., and Yokoyama, S. (2011). Web resource selection for dialogue system generating natural responses. In *Proceedings of the 14th International Conference on Human-Computer Interaction (HCI)*, pages 571–575, Orlando, Florida, USA.

Kaisser, M. (2008). The qualim question answering demo: Supplementing answers with paragraphs drawn from wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 32–35, Ohio, USA.

Katz, B. (1997). Annotating the world wide web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet*, pages 136–159.

Katz, B. and Lin, J. (2002). Start and beyond. In *Proceedings of the 6th World Multiconference Systemics, Cybernetics and Informatics*, pages 1–8, Florida, USA.

Lee, M., Cimino, J., Zhu, H., Sable, C., Shanker, V., Ely, J., and Yu, H. (2006). Beyond information retrieval-medical question answering. In *Proceedings of the AMIA Annual Symposium*, pages 469–473, Washington DC, USA.

Liu, W. and Wong, W. (2009). Web service clustering using text mining techniques. *International Journal of Agent-Oriented Software Engineering*, 3(1):6–26.

McDaid, D. and Park, A. (2011). Online health: Untangling the web. Report Bupa Health Pulse 2010, London School of Economics.

Miao, Y. and Li, C. (2010). Mining wikipedia and yahoo! answers for question expansion in opinion qa. In *Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 367–374, Hyderabad, India.

Mishra, T. and Bangalore, S. (2010). Qme!: A speech-based question-answering system on mobile devices. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 55–63, Los Angeles, USA.

Mitkov, R. (2001). Outstanding issues in anaphora resolution. In *Computational Linguistics and Intelligent Text Processing*, pages 110–125. Springer.

Olvera-Lobo, M. and Gutierrez-Artacho, J. (2011). Open- vs. restricted-domain qa systems in the biomedical field. *Journal of Information Science*, 37(2):152–162.

Robertson, N. and Harrison, P. (2009). Whats wrong with me? concerns about online medical self-diagnosis. In *Proceedings of the Australia and New Zealand Marketing Academy Conference (ANZMAC)*, pages 1–9, Monash University, Australia.

Ruthven, I. (2011). Information retrieval in context. In Melucci, M. and Baeza-Yates, R., editors, *Advanced Topics in Information Retrieval*, pages 195–216. Springer Berlin Heidelberg.

Ryan, A. and Wilson, S. (2008). Internet healthcare: Do self-diagnosis sites do more harm than good? *Expert Opinion on Drug Safety*, 7(3):227–229.

Sun, M. and Chai, J. (2007). Discourse processing for context question answering based on linguistic knowledge. *Knowledge-based Systems*, 20(6):511–526.

Voorhees, E. (2004). Overview of the trec 2004 question answering track. In *Proceedings of the 13th Text Retrieval Conference (TREC)*, pages 52–62.

White, R. and Horvitz, E. (2010). Web to world: Predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search. In *Proceedings of the AMIA Annual Symposium*, pages 882–886, Washington DC, USA.

Wong, W., Thangarajah, J., and Padgham, L. (2011). Health conversational system based on contextual matching of community-driven question-answer pairs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM Demo)*, pages 2577–2580, Glasgow, UK.

Ye, S., Chua, T., and Lu, J. (2009). Summarizing definition from wikipedia. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 199–207, Singapore.

Yu, H., Sable, C., and Zhu, H. (2005). Classifying medical questions based on an evidence taxonomy. In *Proceedings of the AAAI Workshop on Question Answering in Restricted Domains*, pages 27–35, Pennsylvania, USA.

Zhou, L. (2007). Ontology learning: State of the art and open issues. *Information Technology and Management*, 8(3):241–252.