IJE TRANSACTIONS C: Aspects Vol. 25, No. 3, (September 2012) 187-196

International Journal of Engineering

Journal Homepage: www.ije.ir

RESEARCH NOTE

Boosting Passage Retrieval through Reuse in Question Answering

M. Mansoori [a], H. Hassanpour* [b]

[a] Department of Electrical and Computer Engineering, Babol University of Technology, P.O. Box 484, Babol, Iran
[b] Department of Computer Engineering & IT, Shahrood University of Technology, P.O. Box 316, Shahrood, Iran

PAPER INFO

Paper history:
Received 13 March 2012
Received in revised form 12 April 2012
Accepted 17 May 2012

Keywords:
Question Answering
Information Retrieval
Reuse
Passage Retrieval
Discourse Processing

ABSTRACT

Question Answering (QA) is an emerging important field in Information Retrieval. In a QA system the archive of previous questions asked from the system makes a collection full of useful factual nuggets. This paper makes an initial attempt to investigate the reuse of facts contained in the archive of previous questions to help and gain performance in answering future related factoid questions. It models the role of facts in questions through discourse transition of user question answering process, and presents approaches to identify and extract these facts with the help of lexical semantic resources. Strategies to implement the reuse of facts to boost query generation in the passage retrieval stage of a QA system as well as ideas on system evaluation are discussed.

doi: *10.5829/idosi.ije.2012.25.03c.02*

1. INTRODUCTION

Question Answering (QA) is the task of providing the user with one or more answers rather than whole documents to a question posed in natural language [1]. QA is a branch of Information Retrieval (IR) and Information Extraction (IE), and has been researched widely since the TREC[2] QA track, a significant activity of QA evaluation campaign in 1999.

To process and answer a question, a QA system typically includes an initial document/passage retrieval step to retrieve candidate documents/passages that may potentially contain answers. Document/passage retrieval is a very important step in the QA framework. The accuracy of final answer will depend to some extent on the quality of passages retrieved either directly or from the documents in corpus.

The input to the document/passage retrieval stage is a query of keywords that represent the current question. In many cases the basic query is further processed and expanded to improve the quality of document/passage retrieval. Query expansion introduces lexical paraphrases of the original keywords contained in the query.

Our research in boosting the passage retrieval step enriches the keyword query representing user factoid question with feature different from merely lexical paraphrases of question keywords. We introduce into the query additional new keyword that could potentially appear in the answer passages to user queried question. Specifically, an attempt is made to search in the archive of the previous questions for keywords representing facts in the domain of the user question topic that are semantically related to the answer sought after by the user question. If successful, the extracted fact is reused as a new additional keyword feature in the query formulation process of the QA system. We target facts that are entities and present in the archived questions with a potential role in the final answer passage to the user queried question. A fact is either an answer entity to the user question, or otherwise has a strong semantic relationship to the answer entity.

When people ask questions, knowing that questions are expected to be short and concise utterances of few words aiming at the intention in mind, they frame the questions with precision. Consequently any useful salient fact that may exist within the question body has an important role in the domain of the question topic. Because these facts originate from the human knowledge of the topic domain, they constitute valuable information of the domain that can be useful in answering other questions in the domain.

*Corresponding Author Email: h_hassanpour@yahoo.com (H. Hassanpour)

[2] http://trec.nist.gov

Consider the sequence of questions below. The questioner of the proceeding question Q1 states the fact that *sweeteners* are used for making ice cream. This fact when extracted from Q1 can be useful for answering the current user question Q2.

Q1 : Which sweeteners are used in ice cream?
Q2: What ingredients are used in ice cream?

This kind of QA dialogue can occur both when in a QA service multiple users ask questions and also a single user asks a series of questions about the same topic. In a multi-user QA service users have different skills and knowledge on a specific topic and target their questions accordingly. Some users ask basic questions such as question Q2 above that by itself does not carry any useful facts other than the point that *ingredients* are used for making ice cream. Other users more knowledgeable about the topic ask questions that are more specific and detailed, such as question Q1, with varying level of facts embedded in them. In stating the more specific questions the user needs to specify additional facts relating to the domain of the question topic so that the question is properly focused on user's specific information need. The sources of these facts are either from user's previous knowledge on the topic, or the user may have acquired it in an earlier interaction with the QA system. It is these more specific questions, i.e. Q1, that when posed to the QA system, provide the facts that can be useful for answering later questions, i.e. Q2, on the same topic.

Although the above scenario occurs more frequently in multiple user sessions compared to single user sessions, it is also feasible in the latter. A single user conducting a dialogue in an information seeking session who is familiar with the topic and also for the purpose of clarifications of previous answers, may also pose questions ordered like the sequence above, therefore, making opportunities for the reuse of facts.

In the rest of the paper in Section 2 on related work we review query expansion techniques in QA and the topic of reuse in general in open domain QA. Next, in Section 3 we show the various kinds of facts we target, and we then set out to formalize the role and relatedness of facts in the discourse, transitioning from current question to the follow-up question. Next, in Section 4 strategies to implement reuse of facts are discussed. Finally, in Section 5 we present ideas for evaluating system performance and end with our conclusion in Section 6.

2. RELATED WORK

In QA, the vocabulary mismatch problem between the question keywords that a basic query is essentially composed of, and the potential answer passages is commonly addressed through query expansion. Syntactic, semantic, and corpus based frequency information is used to expand queries to bridge this gap.

Approaches to query expansion fall mainly into two categories. The first is based on certain external semantic knowledge resources such as WordNet to expand the query by adding words with similar semantic concept as the original query words [2]. The second approach called blind relevance feedback analyses the top N retrieved documents from the initial query run and adds new term to the query based on the standard Rocchio term weighting method [3]. There has also been more recent efforts to leverage concepts stored in Wikipedia to expand queries. Yajie et al. [4] expand queries by retrieving and ranking concepts in Wikipedia relevant to query words in question and selecting the high-quality concepts serving as additional query features.

Our query expansion takes a different approach by leveraging the archive of previous questions and identifies and extracts words in the question archive that are semantically related to the expected answer type of the user queried question. The extracted word is then introduced in the query formulation of the QA as additional query feature.

The other aspect of our research concerns reuse in QA. The problem of reuse in QA as basis for improving performance has not been fully investigated either as a defined task in the standard QA track or in individual QA systems. The preliminary study by Light et al. [5] that resulted in collecting and analyzing a corpus of questions and answers to find and classify reuse possibilities is one of the first attempts that lays the foundation for much needed work. In that study several categories and sub-categories of reuse in QA were identified. Our research in this paper on the reuse of fact in questions to help answer future questions picks up from one of the sub-categories discovered in their work.

The few areas of research undertaken on the problem of reuse in QA include the forms of reuse different from our work of finding facts in the previous questions and reusing them. One major area of reuse in QA has to do with question similarity which tries to recognize that the same question, in different words, has been asked and answered before. When a previous question similar to the user question is identified its cached answer can be reuse to answer the user question. This would avoid the lag time of normal QA processing pipeline thereby improving its performance. Question similarity was first conducted using FAQ data [6] and further extended to the community-based QA data [7]. Question similarity reuse was also pursued in TREC-9 QA track termed as redundant question [8].

Another form of reuse of previous questions and answers involves the issue of question recommendation.

Given a question as query, the recommending system retrieves and ranks other questions in the previous cached data base according to their likelihood of being good recommendations of the queried question, therefore, providing alternative aspects around user's original interest [9].

As the review of related work of reuse in QA indicates employing reuse for performance, benefits have been limited to mainly redundant or similar questions. One factor that stands out as to the reason for the limited undertaking of the reuse issues as performance factor in QA research is the range of questions and the set of fixed documents that make the basis for QA system evaluation at the standard QA tracks including TREC QA. We point out the TREC QA track here because the origin of modern QA is believed to have its roots in the TREC conferences starting with TREC-8 and the role that it played in the many important achievements and advancements that followed up in the field of QA. However, because of its dependence on the single shot questions usually collected and produced by the assessors which are in sharp contrast to the questions in real user dialogue [10] where many questions might relate in different ways to each other providing reuse opportunities, the reusability issue was ignored.

3. INTERPRETING FACTS IN QUESTIONS

Let's take the earlier example of questions on the topic *ice cream* and provide some answers and then put them into a sequence representing a short dialogue that a user would conduct to find facts about *ice cream* as depicted in Table 1.

TABLE 1. A dialogue on facts about ice cream.

Q2: What ingredients are used in ice cream?
A2: Ice cream must contain at least 10% milk fat, and at least 20% total milk solids, and may contain safe and suitable sweeteners, emulsifiers and stabilizers, and flavoring materials.
Q1 : Which sweeteners are used in ice cream?
A1: Sweeteners in conventional ice cream compositions include carbohydrates such as sucrose, corn syrup, high fructose corn sweeteners (HFCS) and, in some cases, maltodextrins.

To formulate the follow-up question Q1, the user takes into account the answer to the first question Q2 with the focus of *ingredients* which includes amongst other, the ingredient *sweeteners.* The user then proceeds to further explore the specific *sweeteners* used in making ice cream by issuing question Q1. As it is shown with underlined words, user's curiosity about asking the type of *sweeteners* comes directly from the entity *sweeteners* present in the answer to the previous question Q2. We observe that the designated fact, *sweeteners*, from the answer of the previous question is an object of type *ingredient*, the focus of Q2, which also occurs in the focus of the next question Q1, indicating that the user intends to probe the system for more detail answers. We also note that the relation between the focus of Q2, *ingredient*, and the target fact appearing in the focus of Q1, *sweeteners*, is of an IS-A type.

In a dialogue conducted in this manner, to continue with the discourse the user picks an entity or a concept as a fact, such as *sweeteners*, from the answer of the previous question and uses it to formulate the next question. In this example we targeted the fact that is the entity or concept from the answer text of the previous question which has an IS-A type semantic relation with the focus of its source question. When this fact appears in the next question, i.e. *sweetener* in Q1, we have a pair of closely connected questions with lexical cohesion and if we were to reverse the order of the questions to frame it as our original problem, that is Q1 is issued first followed by Q2 as:

Q1 : Which sweeteners are used in ice cream? *Q2: What ingredients are used in ice cream?*

to help answer question Q2, the fact present in Q1 can be reused. In general with the help of lexical semantic resources, we are able to target facts in previous questions having other semantic relations in addition to IS-A type to the focus of the current user question.

In this example the relation that connects the two questions through the fact in one question was determined to be of an IS-A type. In general for two questions with common context. the fact relatedness can be established through several means. First the fact in a previous question is an answer entity to the user factoid question as *sweetener* in Q1 is to Q2. Clearly, this type of fact in a question can be extracted by the answer processing module of the QA system. In an answer typing QA system and with respect to its question and answer ontology, this type of fact would have an IS-A type relation with the focus or expected answer type of the user question.

Secondly, if however the answer processing module does not return a satisfactory result meaning that an IS-A type fact relatedness was not established between the focus of the user question and the fact in the previous question, the fact relatedness between the question pair can alternatively be examined through other representative semantic relations in a semantic network such as WordNet. Finally a semantic network such as WordNet does not cover many hard to classify relations between concepts, and if there is no coverage detected

for the question pair in WordNet, as a last attempt we propose to use richer lexical resources such as a thesaurus. A thesaurus can help establish relatedness between concepts by virtue of their frequent association or situational relation [11].

In the following section we will investigate various types of fact relatedness between questions of common context through a systematic modeling of QA discourse.

3. 1. Question-fact Relatedness The brief dialogue in Table 1 on topic of *ice cream* is an example of user issuing questions in an orderly manner to gradually explore various aspects of a topic. This interaction can be put into a wider setting of an area of QA known as Context Question Answering (CQA) in which the user carries on a dialogue of question/answer on various aspects of a topic with the QA system. In CQA the capability to interpret and answer questions based on context is important. While CQA concerns itself with several issues including tracking focus, anaphora and coreference resolution, and ellipses we are interested here on the issue that given the previous question and its answer, how a user proceeds with tracking focus and other aspects of topic in the follow-up question. For example, referring to the topic of ice cream and viewing the steps required for making ice cream as a procedural discourse, the user certainly goes through a discourse transition in which at each step of the transition new knowledge is gained incrementally on the topic through issuing questions and receiving their answers. In this process the user formulates the follow-up question by taking into account the new knowledge gained from the proceeding answers.

As pointed out, the follow-up question interpretation has its foundation in CQA. One key issue of research in this area is discourse modeling in which the discourse role of entities including the topics or focus of a question and discourse transitions are investigated from one question to the next in the course of the progress of user information needs. Chai et al. [12] propose that the discourse transition which determines how context will be used in interpreting and answering questions as the discourse proceeds from one question to the next, consists of intentional, informational, and presentational transition components. Here, we focus on informational transition which is mainly concerned about how the topic of a question evolves and we can apply that to track the role of fact in the follow-up question.

Chai et al. [12] categorize information transitions in a context into three types: Topic Extension, Topic Exploration, and Topic Shift. In Topic Extension a question concerns a similar topic as that of a previous question, but with different participant or constraints, in Topic Exploration two questions concern the same topic, but with different focus or aspects of the topic, and in Topic Shift two consecutive questions ask about different topics.

The question sequence discussed earlier and repeated in Table 2 is an example of Topic Exploration.

TABLE 2. A dialogue with IS-A fact relatedness**.**

Q2: What ingredients are used in ice cream?
A2: Ice cream must contain at least 10% milk fat, and at least 20% total milk solids, and may contain safe and suitable sweeteners, emulsifiers and stabilizers, and flavoring materials.
Q1 : Which sweeteners are used in ice cream?

In this example both questions are about the topic *ice cream* but with different focus (i.e., asking about different aspects of the topic). In question Q1 the user aims to explore further about the ingredients *sweeteners* learned from the answer to Q2. This example shows that the keywords in the focus of the follow-up question Q1 has an IS-A relation to the focus of the previous question Q2, in other words is an entitiy present in the answer of Q2. If the questions were presented to the QA system in reverse order similar to the sequence of the original reuse problem as repeated below:

Q1 : Which sweeteners are used in ice cream?
Q2: What ingredients are used in ice cream?

an element of Q2's answer is present in Q1 text.

In the second category of information transition, the elements comprising the facts are used to extend the topic. In Table 3 the entity in the answer of Q5, *Wilson*, is used as a fact for Topic Extension by adding it as a constraint to the topic of the follow-up question. Both questions share the main topic of *Roger Federer*,

TABLE 3. Another dialogue with IS-A fact relatedness.

Q5: What tennis racquet did Roger Federer use in U.S. Open?
A5: Federer currently plays with a customized Wilson tennis racquet which is characterized by its smaller hitting area of 90 square inches, heavy strung weight of 12.5 ounces (350 g), and thin beam of 17 millimeters.
Q6:What is the name of the Wilson tennis racquet Roger Federer use?

the topic in Q6 is constrained by the fact *Wilson*, the maker of the tennis racquet. The questions also have a shift in focus with Q5 asking for the maker of the tennis racquet and Q6 for the model name of tennis racquet. The two questions are related in that Q6 contains the named entity, *Wilson*, that counts as an answer entity to question Q5 with an expected answer type of *organization*, maker of the tennis racquet. When we

reorder the questions to set it up for the reuse problem as follows:

Q6:What is the name of the Wilson tennis racquet Roger Federer use?

Q5: What tennis racquet did Roger Federer use in U.S. Open?

The QA system upon recognition of the expected answer type of *organization* sought in Q5 should be able to find the answer entity *Wilson* in Q6.

TABLE 4. A dialogue with situational fact relatedness [5].

Q7 : What retirement plan did Enron's employee have?
A7: Enron Corp. established the Enron ESOP effective November 1, 1986. The Plan document and summary plan description state that the primary purpose of the Plan was to enable participants to acquire stock ownership interests in Enron.
Q8 : What happened to Enron's employees stock?

The dialogue in Table 4 [5] is another case of information transition in questions involving both Topic Extension and Topic Exploration. Q7 asks about the employment plan of *Enron's employee* and is followed by Q8 which has its topic extended with additional constrain of *stock* option. The *stock* option is the fact that is reported in Q7's answer. Both questions have a common topic of *Enron's employee* and the answer entity *stock* in Q7's answer appears in Q8 to extend its topic to the *stock* option of *Enron's employees*. The questions are reordered to set them up for the reuse problem as follows:

Q8 : What happened to Enron's employees stock?

Q7 : What retirement plan did Enron's employees have?

In a QA system based on lexico-semantic resources such as WordNet the fact *stock* present in Q8 does not constitute an answer for user question Q7. WordNet lists *401-k plan, IRA, and Keogh plan* but not the *stock* option as hyponyms (IS-A relation) of the focus of Q7, *retirement plan*. Moreover, if an open domain QA system uses its own internal question ontology it is unlikely that it would extract answer entity *stock* in Q8 as an answer to Q7. This is an example of a situational relation between the keywords *retirement plan* and *stock* and a thesaurus with richer grouping of words can leverage better results.

In the dialogue displayed in Table 5, both questions are about the topic *NAFTA* program. The Information transition in this example is of combined Topic Exploration and Topic Extension, but the target fact, *states*, that is used as a constraint to extend the topic in Q10, is not a direct answer to Q9 with expected answer

TABLE 5. A dialogue with MEMBER-OF fact relatedness.

Q9: What countries are involved in NAFTA?
A9: The Government of Canada, the Government of the United Mexican States and the Government of the United States of America.
Q10: What is the impact of NAFTA on the states?

type of *country*, but rather it has a Holonymy or MEMBER-OF semantic relation to it. When the questions are reordered as shown below, upon recognition of this semantic relation the keyword *states* can be extracted as a fact related to Q9's focus.

Q10: What is the impact of NAFTA on the states?

Q9: What countries are involved in NAFTA?

The several examples of question pairs that are candidate for reuse discussed so far included user questions with the stem word *what* and *which*. In the followings, user questions with the stem words *where, who,* and *when* are covered. As can be seen the same analysis of reuse of facts applies to these sequences as well.

Q11: Who is the architect of Taj Mahal in India?
Q12: Where is Taj Mahal?

Q13: How old is Bill Gates, the chairman of Microsoft?

Q14: Who is the chairman of Microsoft?

Q15: What is the name of volcano that destroyed the ancient city of Pompeii in 1840?

Q16: When was the city of Pompeii destroyed by volcano?

The discussion on reuse of facts covered up to this point included question pairs of type factoid or list in which the focus or expected answer type of a question can be determined with considerable success. In our analysis of reuse for user factoid questions we used the focus or expected answer type to connect the question to semantically related facts in a previous question. With non-factoid user questions it is more difficult to determine with considerable success the expected answer type of the question. Table 6 demonstrates a short dialogue with non-factoid questions in which the fact in one question can be reused to help answer another question:

TABLE 6. A dialogue with non-factoid user question.

Q11: How do you make ice cream?
A11: There are several small machines that easily make ice cream by putting the ice cream mixture into a pre-frozen tub and churning the mixture until frozen. Liquid Nitrogen can also be used instead to freeze the mixture.
Q12: How is Liquid Nitrogen used to make ice cream?

In the this dialogue, the fact *Liquid Nitrogen,* in A11 is used for Topic Extension to provide a participant shift to the topic of Q12. Both questions are about making *ice cream* and both are procedure oriented questions (wh-word: how). Question Q11 asks about the general procedure for making ice cream involving the roles of all the parts in ice cream maker machine and all the ingredients used in it while in Q12 there is a shift of participant in topic focusing on the role of *Liquid Nitrogen* in the process. As we can see the entity *Liquid Nitrogen* in A11 is used to extend the topic in Q12, therefore we have an entity of fact in Q12 that is transferred from the answer A11.

When the questions are reordered as follows and posed to the QA system, the keyword *Liquid Nitrogen* in Q12 is a fact that can play a significant role as an additional query term in the passage retrieval stage of question Q11. In a similar fashion as for user factoid questions, to reuse facts in previous question to help answer user non-factoid question the two questions need to be connected semantically through that fact. Our analysis effort in this work is limited to user factoid questions only.

Q12: How is Liquid Nitrogen used to make ice cream?
Q11: How do you make ice cream?

4. STRATEGIES FOR IMPLEMENTING REUSE

To begin with, the reuse mechanism can be added as a modular component to the baseline QA system and enabled optionally. In a QA system with reuse mechanism enabled there will be an initial attempt to extract fact from the archive of previous questions to help answer current user question. In the following sections we will outline the overall strategy and issues of the design process.

4. 1. Previous Question Archive Previous questions make up the data base for searching the facts. An archive will be made of previous questions asked and newly entered questions will also be added to this archive. For QA systems that use Named Entity (NE) tags in text corpus to locate answer entities, the original archive can be processed offline to add NE labels and new questions will also be tagged for named entities when they are entered into the archive.

One design issue of concern here is that of content efficiency of the archive. Notably, only those questions that contain useful facts should be added to the archive. As will be outlined in Section 4.3, the main use of the keyword corresponding to the extracted fact is to pad the query generated for the user question in the passage retrieval stage. However, when padded to the query, factual concepts lacking sufficient specificity cannot be effective in retrieving more related passages and may even cause noise in the process. In assessing the informative role of WordNet in open domain QA, Pasca et al. [13] experimented with specificity of keywords in query formation and concluded that enabling the specificity option increased TREC-8 correctly answered questions by 11%. The specificity measure for a concept in their experiment was defined as the number of hyponyms of the concept excluding hyponyms that are proper nouns and those with the same heads. Concepts with the hyponym count of less than the threshold of 10 were picked to form the keyword query for passage retrieval. With this perspective of keyword specificity only questions having at least one concept with sufficient specificity should be entered in the previous question archive.

4. 2. Generating the Base Set for Extraction The base set consists of the questions in the archive topically related to user question and will be used to extract facts from. To be considered as related, each question in the base set should have at least one overlapping topic with the user question. Figure 1 illustrates the procedure to generate the base set.

Questions normally have a main topic and a questioner asks a question by focusing on a particular aspect of the main topic. The keywords that constitute the main topic are the ones that almost appear unchanged in an answer passage in a QA process. For example, consider the question *"which ingredients are used in ice cream"*. Here the main topic is *ice cream* and the question is focusing on the *ingredients* aspect of ice cream.

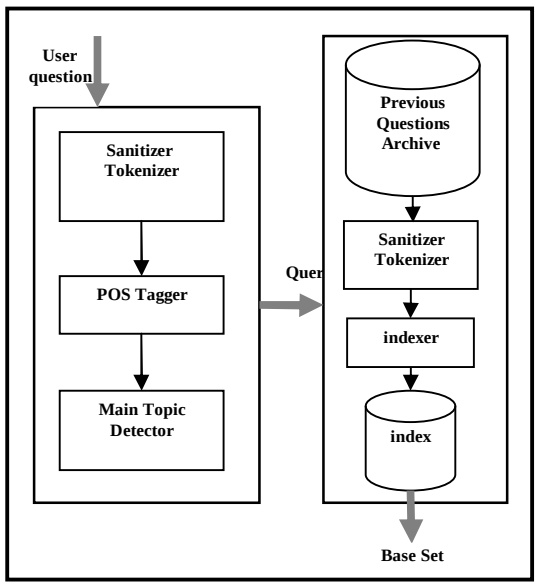

Figure 1. Base set of previous questions

For most questions with wh-word including *who*, *where*, *when*, *which*, and *what* the main topic keywords consist of nouns or proper nouns that appear after the main verb of question such as topic *ice cream* in the above example. Sometimes in questions with *which* and *what* wh-word the focus keyword may also appears after the main verb and in that case this keyword should not be included as the main topic keywords. An example would be *"what are the ingredients used in ice cream"* in which the keyword *ingredients* appear after the verb and is the focus and not the main topic and should not be considered as a main topic keyword. Therefore for questions with wh-word *what* and *which* if there is no noun between the wh-word and the verb, then the first noun occurring after the verb is the focus and is not a part of the main topic. The nouns including proper nouns that appear after this focus word make up the keywords of the main topic. The keywords that make up the main topic of a question can be extracted using a POS tagger. Using this approach the keywords of the main topic of user question can be identified and used to form a query to retrieve the related base set of questions from the question archive.

4. 3. Fact Extraction for Factoid Questions Answer to a factoid question normally consists of a NE item. once the expected answer type of user factoid question is identified, the QA system applies its answer processing procedure to search and rank answers from the base set of questions retrieved from the question archive. If the QA's answer processing procedure is partly or entirely based on finding answers with the matching named entity tags, the question archive as mentioned is pre-processed to contain the required NE tags. For the factoid questions type that use search patterns to extract answers, the search can also be accomplished on the retrieved base set using the patterns. The NE corresponding to the extracted answer, if any, in this case is the target fact sought. Figure 2 illustrates the fact extraction procedure.

If the extraction procedure above produces answers, the top ranked answers may be presented as the final answers to user. However if based on the feedback from the user or as a part of the normal operation of the QA system, the answer is preferred to be presented in a window of short text passage, each keyword of the top ranked answer entities can be used alternately as additional significant term to pad the keyword query in passage retrieval stage. The keyword query padded with the fact which in this case is an answer entity ensures retrieval of passages with much higher precision used for input to the answer processing stage.

If the extraction procedure outlined above is not successful in locating exact answers in the base set, the QA system can fall back to find fact that is semantically related to the answer and use its corresponding keyword in the keyword query generation. Consider the two sequences of questions below in which question Q13 and Q15 are used to extract facts from to answer user questions Q14 or Q16 respectively:

Q13: Which sweeteners are used in ice cream?
Q14: What ingredients are used in ice cream?

Q15 : How much corn syrup is used to make a pint of ice cream?
Q16: What ingredients are used in ice cream?

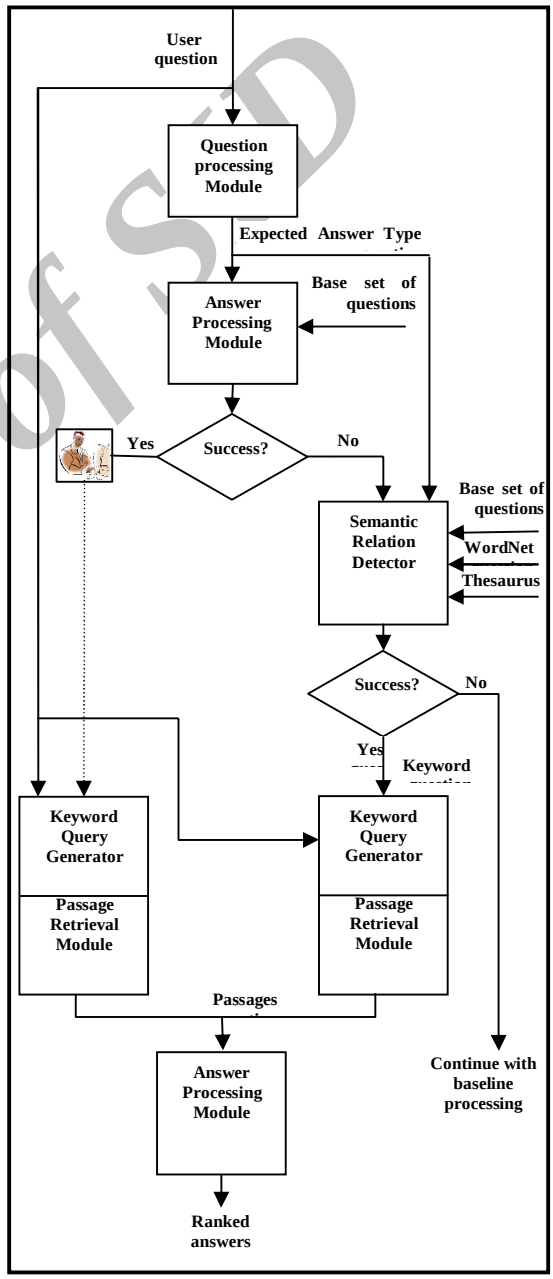

Figure 2. Fact extraction procedure

Since *corn syrup* in Q15 is a specialization of *sweeteners* in Q13 as coded in WordNet, it is more likely, compared to *sweeteners* in Q13, that it will be recognized and tagged by a name entity tagger to match the answer type *ingredients* of the user question. But even if the keyword *sweeteners* in Q13 may not be tagged as a NE to count as a candidate answer, it is still a valuable fact noting that it is semantically related to the focus *ingredient* of user question Q14. A lexical taxonomy resource such as WordNet would indicate that keyword *sweetener* is a specialization of *ingredients* making the keyword highly relevant to the answer passages of Q14. Therefore including this keyword in keyword query for Q14 certainly improves the passage retrieval results. As demonstrated in Section 3, an acceptable response to Q14 does contain the keyword *sweetener*.

We can extend this concept of semantic relatedness to include lexical relations other than IS-A type specialization used till this point. Let's take a closer look at the following pair of questions discussed earlier:

Q10: What is the impact of NAFTA on the states?
Q9: What countries are involved in NAFTA?

Although the keyword *states* is not an answer entity for the user question Q9 with the focus of *country*, it is related to it in that a *state* is one of the constituent administrative district (part of) of a *country*. In WordNet this is indicated through the MEMBER-OF relationship. Having this relationship between the focus of user question Q9 and the keyword *state* in the previous question Q10 makes the keyword *state* very useful if it is included in the keyword query in the passage retrieval stage of Q9.

In WordNet MEMBER-OF lexical relation is a type of meronymy relation. In general we can extend the concept of semantic relatedness to include all the meronymy types. If W_h is the focus of user question and W_m a keyword in a related previous question, the meronymy types in WordNet are defined as follows [14]:

W_m #p→W_h indicates that W_m is a component part of W_h;

W_m #m→W_h indicates that W_m is a member of W_h; and

W_m #s→W_h indicates that W_m is the stuff that W_h is made from

There are also situations in which the focus of the user question and a keyword in the previous question are semantically related but a lexical semantic network such as WordNet does not have sufficient set of relations to relate the two words. The question pair in the following example discussed earlier demonstrates this case:

Q8 : What happened to Enron's employees stock?
Q7 : What retirement plan did Enron's employee have?

Obviously, we would like to relate *retirement plan* and *stock* to be able to use the keyword *stock* in keyword query for Q7. Relations such as hypernymy and meronymy are not appropriate to semantically relate (re*tirement plan*, *stock*). These words are related through frequent association in text similar to (*paper, pencil*) and this characteristic justifies their relatedness for question fact. Hirst et al. [11] refer to this kind of semantic relation as situational relation and conclude that many of such relations are hard to classify and prefer a thesaurus such as roget's thesaurus for the task. A thesaurus is a lexical resource conceptually similar to WordNet but different in design. A thesaurus has a unique system for classification of meaning, grouping of words to represent concepts, and particularly it has a vast set of semantic relations although they are not labeled explicitly as in WordNet. Roget's thesaurus is one of the wildly used thesauruses in NLP and a machine tractable version of that has been implemented by Jarmasz et al. [15].

5. RESOURCES AND EVALUATION

Two issues important for the development of the reuse mechanism for QA systems are resources and evaluation.

Resources include question-answer sets and collection of documents that contain the answers. As pointed out in Section 1, the original study on the general topic of reuse in QA has produced a corpus of question-answer sets exemplifying different categories of reuse and the URLs of supporting Web documents that contained the answers. This corpus is available from the authors of the study [5]. Within the corpus several instances of question sequence relating to reuse of fact sub-category are annotated.

To develop a more stable corpus and additional instances of question-answer set, TREC QA document collection would be a valuable source. In the context task of TREC 2001 [16] many question are grouped into different series with each series representing a context or topic. In TREC 2004 QA [17] collection question set has been divided into subsets. Each subset has a unique topic and set of questions on the topic. Using these data sets with the grouping of questions into topical series it is possible to make up sequence of questions along the line of sequences exemplified in Section 3 for additional reuse instances. Within each sequence of questions information transition of the types discussed in Section

3 including Topic Exploration and Topic extension can be used to connect the questions.

In our strategy for implementation we proposed two methods for reuse of fact to help answer user question. First when appropriate as a direct answer based on the user feedback, the fact is presented as a candidate answer to the user question. Here we expect to see performance enhancement in terms of the speed of system response to a user question. The alternative method was to use the fact to boost query expansion resulting in more relevant and targeted passages during passage retrieval stage. With query expansion we expect to see enhancement in several criteria of evaluation including relevance, correctness, and conciseness.

With the reuse mechanism configured as a modularized component into the QA system the evaluation task becomes fairly simple. Performance can be benchmarked by observing and assessing the above criteria both when the reuse mechanism is enabled and disabled.

6. CONCLUSION

In this paper, we showed that archive of previous questions in QA systems can serve as a valuable resource of facts that can be reused to help answer future related questions. We reviewed the research work in the area of reuse in QA and concluded that it lacked emphasis on the application of reuse in QA.

To demonstrate the presence of facts in questions that can be reused and the role they play in the discourse we used discourse transition model to observe how a user follows up with the next question. We focused on new facts introduced in the follow up question as the discourse advanced and using that we examined the role of facts in questions. We then examined various ways we can take advantage of facts in previous questions to help improve performance especially in the passage retrieval stage of a QA system through query expansion.

7. REFERENCES

1. Hirschman, L. and Gaizauskas, R., "Natural language question answering: The view from here", ***Journal of Natural Language Engineering, Special Issue on Question Answering***, Vol. 7, No. 4, (2001), 275-300.
2. Yang, H., Chua, T.-S., Wang, S. and Koh, C.-K., "Structured use of external knowledge for event-based open domain question answering", ***26th annual international ACM SIGIR conference on Research and development in information retrieval***, Toronto, Canada, 2003, 33-40.
3. Monz, C., "From document retrieval to question answering", in ILLC Dissertation Series, (2003).
4. Yajie, M., Xin, S. and Chunping, L., "Improving Question Answering Based on Query Expansion with Wikipedia***", IEEE International Conference on Tools with Artificial Intelligence (ICTAI),*** France, Vol. 2, (2010), 233-240.
5. Light, M., Ittycheriah, A., Latto, A. and McCracken, N., "Reuse in question answering: a preliminary study", in Maybury, M.T. (Ed.), New Directions in Question Answering, AAAI Press/MIT Press, Menlo Park, CA, (2004), 169–181, (Chapter 13).
6. Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N. and Schoenberg, S., "Question answering from frequently asked question files: Experiences with the faq finder system", ***AI Magazine***, Vol. 18, No. 2, (1997), 57-66.
7. Jeon, J., Croft, W. and Lee, J., "Finding similar questions in large question and answer archives", ***ACM Conference on Information and Knowledge Management (CIKM)***, Bremen, Germany, (Oct. 31-Nov. 5, 2005), 2005, 84-90
8. Voorhees, E. M. and Harman, D., "Overview of the ninth text retrieval conference (TREC-9)", ***Text Retrieval Conference TREC-9***, Gaithersburg, Maryland, USA, (Nov. 13-16, 2000), 2000, 1–8.
9. Cao, Y., Duan H., C.-Y. Lin C.-Y., Yu, Y. and Hon, H.-W., "Recommending questions using the MDL-based tree cut model", ***International conference on World Wide Web (WWW),*** New York, NY, USA, 2008, 81–90.
10. Bernardi, R., Kirschner, M., "From artificial questions to real user interaction logs: Real challenges for Interactive Question Answering systems", In Proc. of Workshop on Web Logs and Question Answering (WLQA'10), Valletta, Malta, 2010.
11. Hirst, G. and St-Onge, D., "Lexical chains as representations of context for the detection and correction of malapropisms", in Fellbaum, C. (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, (1998), 305–332.
12. Chai, J. and Jing, R., "Discourse structure for context question answering", Workshop on Pragmatics of Question Answering, at HLT-NAACL, Boston, MA, USA, 2004, 23–30.
13. Pasca, M. and Harabagiu, S., "The informative role of Word-Net in open-domain question answering***", North American Chapter of the Association for Computational Linguistics (NAACL-01) Meet.,*** Carnegie Mellon Univ., Pittsburg, PA, USA, (June 2-7, 2001), 2001, 138-143.
14. Miller, G., "WordNet: A lexical database for English", ***Communications of the ACM***, Vol. 38, No. 11, (1995), 39-41.
15. Jarmasz, M. and Szpakowicz, S., "Roget's Thesaurus: a Lexical Resource to Treasure", NAACL WordNet and Other Lexical Resources workshop, Pittsburgh, PA, USA, 2001, 186 – 188.
16. Voorhees, E. M., "Overview of the tenth text retrieval conference (TREC-10)", ***Text Retrieval Conference TREC-10***, Gaithersburg, Maryland, USA, (Nov. 13-16, 2001), 2001, 1-15.
17. Voorhees, E. M., "Overview of the TREC 2004 question answering track", ***Text Retrieval Conference TREC 2004,*** Gaithersburg, Maryland, USA, (Nov. 16-19, 2004), 2004.

Boosting Passage Retrieval through Reuse in Question Answering

RESEARCH NOTE

M. Mansoori [a], H. Hassanpour[b]

[a] *Department of Electrical and Computer Engineering, Babol University of Technology, P.O. Box 484, Babol, Iran*
[b] *Department of Computer Engineering & IT, Shahrood University of Technology, P.O. Box 316, Shahrood, Iran*

P A P E R I N F O

Paper history:
Received 13 March 2012
Received in revised form 12 April 2012
Accepted 17 May 2012

Keywords:
Question Answering
Information Retrieval
Reuse
Passage Retrieval
Discourse Processing

چکیده

پرسش و پاسخ یکی از شاخه‌های مهم و در حال ظهور در حوزه بازیابی اطلاعات می‌باشد. در سیستم‌های پرسش و پاسخ آرشیو پرسش‌های قبلی ارائه شده به سیستم مجموعه‌ای حاوی از اطلاعات گران‌بها در باره واقعیت‌ها را در بر می‌گیرد. این مقاله برای اولین بار به بررسی استفاده مجدد از واقعیت‌های موجود در آرشیو پرسش‌ها پرداخته و از آنها برای پاسخ به پرسش‌های آتی و بهبود عملکرد سیستم بهره می‌گیرد. با استفاده از انتقال حالت در پرسش و پاسخ‌های کاربر، نقش واقعیت‌ها در پرسش‌ها مدل‌سازی و رویکردها برای شناسایی و استخراج این واقعیت‌ها با کمک منابع معنایی لغوی ارائه می‌گردد. راه‌کارهای پیاده‌سازی استفاده مجدد در ارتقاء تولید کلید واژه‌ها در مرحله بازیابی عبارات در سیستم‌های پرسش و پاسخ و همچنین رویکردهای ارزیابی سیستم مورد بحث قرار می‌گیرد.

doi: *10.5829/idosi.ije.2012.25.03c.02*