IJE TRANSACTIONS C: Aspects Vol. 26, No. 3, (March 2013) 245-256

International Journal of Engineering

Journal Homepage: www.ije.ir

Investigating Embedded Question Reuse in Question Answering

M. Mansoori [a*], H. Hassanpour

[a] Department of Electrical and Computer Engineering, Babol University of Technology, Babol, Iran
[b] School of Information Technology & Computer Engineering, Shahrood University of Technology, Shahrood, Iran

PAPER INFO

Paper history:
Received 30 August 2012
Received in revised form 17 October 2012
Accepted 18 October 2012

Keywords:
Question Answering
Information Retrieval
Reuse
Noun Phrase Matching

ABSTRACT

The investigation presented in this paper is a novel method in question answering (QA) that enables a QA system to gain performance through reuse of information in the answer to one question to answer another related question. Our analysis shows that a pair of question in a general open domain QA can have embedding relation through their mentions of noun phrase expressions. We present methods for recognition of embedding property between a pair of questions by focusing on the techniques applied to match noun phrase mentions in the questions. We then take advantage from the discovery of embedding relationship and extract referent named entities corresponding to the noun phrase expression that are present in the answer of one question. The named entities will then be used as significant terms in the query generation phase of the QA system to retrieve more pertinent answers. Finally, we discuss on data set resources and system evaluation.

doi: *10.5829/idosi.ije.2013.26.03c.04*

1. INTRODUCTION

Current search engines return a ranked list of documents to a user query but leave it to the user to extract the answers. Question Answering (QA) [1] is a more challenging task in that it allows the user to ask questions in everyday natural language and produces exact answers freeing the user from the task of document lookup. QA is a paradigm in the field of Information Retrieval (IR) and Information Extraction (IE).

Many of the questions in the archive of a QA system follow a common context and are closely related. With related questions there are opportunities for the reuse of information in one question or its answer to answer other questions. One of the many relationships between a pair of questions that provides for reuse opportunity is that using the answer of one question could greatly enhance in finding better answers for the other question [2]. In this paper, we focus on analyzing this type of relationship and the reuse benefits which it brings to a QA system.

Another form of the relationship between a pair of questions that also benefits reuse in QA but regards to a different sub-category of reuse relationships is that one of the question contains facts semantically related to the answer sought after by the user question. This was the subject of another paper [3] and in that we investigated this sub-category of reuse by demonstrating the patterns among the questions in the discourse that give rise to it and the mechanism to implement this reuse facility in a QA system.

Assume a scenario where a number of users are asking questions about the *capital of China*. A sequence of questions and document excerpts are displayed in Table 1.

Viewing the examples in Table 1, it is simple to notice **Q1** as being a simple base question, while Q2 and Q3 are more complex questions that are build up on **Q1**. Also, note that while the noun phrase "***the capital of China***" is used to refer to the topic of interest in all questions, the answer passages do not use that phrase and only mention the referent "***Beijing***". In terms of QA processing this means that relying on the keywords of the noun phrase "***the capital of China***" in the passage retrieval phase of a QA system to retrieve answers for Q2 and Q3 does not help much. The important feature of these question-answer passage pairs is that Q2 and Q3 should be easier to answer if the system could make use of the answer to the base question **Q1**. For example, knowing that "***Beijing***" is a city, and that it is "***the capital of China***" should help

*Corresponding Author Email: Mansoori@nit.ac.ir (M. Mansoori)

the system find more pertinent information about *the population* and *the average temperature* asked in Q2 and Q3, respectively.

The questions in each of the pair (**Q1**, Q2) or (**Q1**, Q3) are related in such a way that if Q2 or Q3 is posed to a human question answerer with no prior knowledge of "***the capital of China***"**,** the person would be tempted to initially figure out the referent city of the phrase "***the capital of China***" in these questions, in effect answering to **Q1**, and use the answer to search for an answer to the original questions Q2 or Q3. We call such a relationship between questions **Q1** and Q2 or between **Q1** and Q3 an embedded relation by the virtue of question **Q1** being implicitly embedded in Q2 or in Q3, and designate **Q1** as the base question (to highlight, the labels for the base question and its answer are bolded) and Q2 or Q3 as the embedding question.

Our work in this paper deals with factoid and list type questions only. The answer to many factoid and list questions are named entities. The basic idea of the paper centers on the existence of an embedded relationship between two questions, one question being the current user question as the embedding question, and the other a previous question as the base question.

The embedded relationship between the base and embedding question is established through the presence of semantically equivalent noun phrases in the two questions. Once the presence of this relationship is recognized, then the named entities present in the answer passages of the base question can help find better answers to the user question.

By extracting and using these named entities as additional significant terms to boost the query generation of the passage retrieval stage for the embedding user question, it is expected that more pertinent answers will result. We deal with a particular basic construction of noun phrases and refer to it as BasicNP as explained in a later section.

In the rest of the paper in Section 2 on related work we review the few areas of research in reuse in QA. In Section 3 we investigate embedded questions in discourse and its performance benefit to the QA system. We then focus on noun phrase as the linkage between a base and embedding question and follow up on the problem of noun phrase matching and the challenge that linguistic variability imposes on this process.

In Section 4 we present detailed methods and techniques to design and implement the reuse by discussing embedded relation recognition and the extraction of information to build a repository that embodies a lexicon of basic noun phrases and their referent entities. We also point to the issue of interpretation of the semantic relation in noun-noun compounds as part of noun phrase matching. We finally discuss on ideas for the embedded reuse corpus generation and system evaluation in Section 5 and present our conclusion in Section 6.

TABLE 1. A sequence of questions and document excerpts

Q1: *What is* ***the capital of china****?*
D1**:** Beijing *is a metropolis in northern* **China** *and the* **capital** *of the People's Republic of* **China.**
Q2: *What is the population of* ***the capital of China****?*
D2: The population of ***Beijing*** *on December 21st 2011 is approximately 19,872,174. (Extrapolated from a population of 17,430,000 in 2007 and a population of 19,612,368 on June 28th 2011).*
Q3: *What is the average temperature of* ***the capital of China*** *in December?*
D3: The average low temperature of ***Beijing*** *in December is -6 ^{0}C with the average high of 4 ^{0}C.*

2. RELATED WORK

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.

One factor that stands out as the reason for the limited undertaking of the reuse issues as performance factor in QA research has to do with the goals and objectives that is set forth by the standard QA tracks including TREC QA track. 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. But many of these achievements were centered on single, factoid question category. The idea to bring in context into QA which sets challenges for follow-up question processing and potential reusability first made its way into QA by the context task in 2001 [4]. But the concerns of the context task included tracking the discourse objects within and across questions through referential links and ellipses. From TREC2004 [5] questions were grouped into different series and each series was based on topic or target and questions in the series ask for some information about the topic. Even in these later tracks the role of context and the opportunities for different categories of reuse that it can provide was not fully addressed.

The preliminary study by Light et al. [2] 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. The varieties of reuse types discovered in their work in the question corpus that was built from questions posed by human subjects in the experiment and the query log of a search engine shows the strong

interdependence between questions that realistic context questions provide as compared to the questions made artificially from documents such as newswire or newspaper sources that is the normal practice in TREC QA tracks. Looking at the richness of the reuse relations annotated in the corpus in their work, we believe there is a strong potential in terms of performance benefit far behind the FAQ style similar question reuse that can be gained for QA systems. Our research in this paper on the embedded question reuse 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. 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 reused to answer the user question. 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].

Fleischman et al. [9] used lexico-syntactic patterns to extract highly precise relational information from text collection offline creating a data repository that is used to retrieve answers to questions directly. They extract concept-instance pairs of person name-title such as (Bill Clinton, the president of the USA) and deposit them in the repository to reuse them to answer related questions such as (Who is the president of the USA?). Their answer repository based approach to QA is similar to a part of our work where we also generate a repository, although online, of named entities that are answers to base questions but we use these entities to answer other related embedding questions. We also use the Web as the text collection to take advantage of its vast size and redundancy of information.

In Mansoori and Hassanpour's research [3], another sub-category of reuse to boost the passage retrieval stage of QA was introduced. In that work, the reuse of facts contained in the archive of previous questions to help and gain performance in answering future questions was investigated. The reuse of facts discussed in that paper integrated with the reuse facility discussed here can even further boost the overall performance of a QA system.

3. EMBEDDED QUESTIONS

3. 1. Embedded Questions in Discourse

Considering the different information needs of users on a common topic in a multi-user question answering service, occurrenc of two questions posed by two different users that display the embedding relationship should be relatively common. Single users involved in a session with the QA system to investigate on a topic of interest, conduct the session by issuing a sequence of correlated questions in a cohesive manner. Here, it is also very likely that some pairs display the embedding property as the discourse analysis below suggests.

In an effort to model discourse in Context Question Answering, Chai et al. [10] recognized discourse transition as one of the elements that defines the discourse status. The discourse transitions which determine how discourse roles are changed from one question to the next as the user QA interaction proceeds has an informational transition component which mainly centers around the topics of questions and how these topics evolve during the discourse. The informational transition which can help to demonstrate our focus of embedded relationship between questions is further categorized into three types: Topic Extension, Topic Exploration, and Topic Shift.

The examples presented below demonstrate the information transition in a QA discourse involving a base question (**Q4**) followed up by several embedding questions and their corresponding document excerpts. These instances further show the fact that some of the named entities present in the answer to the base question (**D4**) are repeated in D5 to D8 signifying their usefulness for the passage retrieval stage of Q5-Q8.

*Q4: What are **some medicines that treat anthrax**?* *D4: The FDA has approved **Cipro (ciprofloxacin)**, **tetracyclines** including **doxycycline**, and **penicillins** to treat anthrax.*
*Q5: What are some of the side effects of **anthrax medicines**?* *D5: The Physician's Desk Reference reports that of 2,799 patients who took **Cipro** during clinical investigations, 16.5 percent had adverse reactions that were possibly or probably related to the drug. The most frequently reported reactions; nausea, diarrhea, vomiting, abdominal discomfort, headache, restlessness, rashes.*
*Q6: Who manufactures **anthrax medicine**?* *D6: **Cipro** is produced in the U.S. by the German pharmaceutical company Bayer AG.*
*Q7: When did the FDA approve the **anthrax medicine** from Bayer AG.* *D7: **Ciprofloxacin** was first patented in 1983 by Bayer A.G. and subsequently approved by the U.S. Food and Drug Administration (FDA) in 1987.*
*Q8: What **anthrax medicines** is approved by FDA?* *D8: In August 2000, the U.S. Food and Drug Administration (FDA) approved **ciprofloxacin** hydrochloride (**Cipro**; Bayer; hereafter, **ciprofloxacin**) for management of postexposure inhalational anthrax.*

All these questions except Q8 share the main topic of "***anthrax medicine***". In Topic Exploration a question concerns a similar topic as that of a previous question but with a different focus or aspect of the topic. Question pairs (**Q4**, Q5) and (**Q4**,Q6) display this category of information transition in the discourse. Both Q5 and Q6 explore other aspects of the topic "***anthrax medicines***"*;* while **Q4** asks about the name of "***anthrax medicines***", Q5 explores the "*side effects*" aspect and Q6 the "*manufacturer*" aspect of the topic "***anthrax medicine***". In Topic Extension, two questions share similar topic but with different extensions. With question pair (**Q4**, Q7) both questions share similar main topic but Q7 extends it by adding the participant "*Bayer AG*", and Q7 also explores the "*approval*" date aspect of the topic. Transition in question pair (**Q4**, Q8) shows a shift of topic from "***anthrax medicine***" in **Q4** to "*FDA approval*" in Q8 making the topic of **Q4** to become the focus of Q8, indicating a further probing of the topic "***anthrax medicines***".

These examples demonstrate user behavior from a formal perspective of QA discourse analysis and emphasize the central idea presented in this paper on the reuse opportunity that the embedded relationship provides.

3. 2. Embedded Question Analysis Let's see how the performance of a QA system could benefit by taking advantage of embedded relationship between questions. Two of the parameters involved in performance evaluation of a QA system pertains to precision and recall. Let's use the scenarios presented earlier in **Q4** to Q8. When a human question answerer is asked either of the questions Q5, Q6, Q7, or Q8 without having prior knowledge of the answer to question **Q4**, the person would initially want to throw some light on the phrase "***anthrax medicines***" in these questions by finding out what the possible *medicines* for *anthrax* are, in effect indirectly generating an implicit question such as "*what are some medicines that treat anthrax?*". To answer this self-generated question, assume that the person searches in the disease-treatment section of a family medical guide book for the *medicines* that *treat anthrax* and is able to find the medicine, *Cipro.* Having found *Cipro,* the person would then use this result to look up the answer to either of Q5, Q6, Q7, or Q8 question. Let's take Q5 and follow up the process. The person then would look for the *side effects* of *Cipro* by looking, for example, in the drug glossary section of the guide. Although it is quite possible that the disease-treatment sections of many of these guide books cover passages that mention disease-medicine-side effects, whereby directly pointing out the answer to Q5, in this case "*the side effects of anthrax medicines*", it is expected that the extend of drug-side effects coverage would certainly be higher in the drug glossary sections in many of these guides. Therefore, more precision is achieved by extracting passages from the section of the guide specifically attributed to drug-side effects than in the disease-treatment section. With regard to QA question processing, the preprocessing in effect translates Q5 into the equivalent question:

> *What are some of the side effects of* ***[Cipro, ciprofloxacin, tetracyclines, doxycycline, penicillins]?***

Furthermore, since the drugs mentioned have application in other disease treatment, the precision can further be improved if both of the search arguments explained above (drug glossary and disease-medicine glossary) are used in an AND manner in the retrieval process, in effect generating the following question implicitly:

> *What are some of the side effects of* ***[anthrax medicine, Cipro, ciprofloxacin, tetracyclines, doxycycline, penicillins]?***

3. 3. Recognition of Embedded Question Relationship To figure out the linkage between a pair of questions candidate to have embedded relation additional instances are presented below:

> ***Q11****: What are the* ***dietary sources of calcium****?*
> *Q12: How do the* ***calcium sources*** *fit into the overall diet?*

> ***Q13****: Who is* ***the mayor of Chicago****?*
> *Q14: What is the salary of* ***the mayor of Chicago****?*

> ***Q9****: How much* ***calcium should an adult woman get in her diet****?*
> *Q10: How much vitamin D do you need in order to absorb* ***the recommended calcium****?*

Looking at these and the previous examples it shows that the embedded relation between a pair of questions involves a pair of noun phrases. The *capital of China, anthrax medicine, recommended calcium amount, and calcium sources* are all normalized forms of a noun phrase that relate the two questions in their respective pair. Each pair in the preceding examples consists of a base question followed up with an embedding question that has the base question implicitly embedded. Therefore, to recognize the existence of an embedded relation between a pair of questions the two noun phrases (NP) in their respective questions must be matched. For example the NPs "***some medicines that treat anthrax***" in the base question **Q4** and "***anthrax medicine***" in Q5 or Q6 need to be matched in order to establish the existence of an embedded relationship between the respective pairs.

A noun phrase in English describes a concept and is the grammatical unit that the topic and focus of questions is built from. In effect noun phrases in questions tell us what a question is all about. Syntactically, a noun phrase is composed of a head and

a modifier. Whenever a head alone is not precise enough to describe a concept it is modified by another noun, adjective, or propositional phrases.

Based on the observation on the usage of noun phrase patterns in large text collections, Girju et al. [11] identified five most frequently used NP level constructions. In this work we will investigate on three of these NP level constructions; (1) compound nominals or noun-noun compounds consisting of two consecutive nouns (e.g. *anthrax medicine*) with the first noun as the modifier and the second as head, (2) adjective clauses where the head noun is modified by a relative clause (e.g. *medicines that treat anthrax*), and (3) genitives which include the of-genitives (e.g. *the capital of China*) where the modifier is syntactically marked by preposition 'of' and follows the head noun and s-genitives (e.g. *China's capital*) in which the modifier is morphologically linked to the possessive clitic's and proceeded the head noun. We also restrict our treatment of these noun phrases to their basic constructions and call them BaseNPs. We define a BaseNP as follows: noun-noun compounds, s-genitives, and of-genitives are limited to two one-word nouns with possible predetermines; of-genitive can have a determiner before the second noun (e.g. *chairman of the board*); the head and the relative clause of the adjective clause is also limited to one-word nouns.

3. 3. 1. BasicNP Matching As discussed in the previous section, the linkage between a base and the corresponding embedded question in an embedded relationship involves a pair of BasicNP noun phrases. In order to establish the embedded relationship between the two questions, the pair must be semantically equivalent. Furthermore, the embedding question includes a BasicNP whose constituent head and modifier form the focus and the topic of the base question, respectively.

Our focus of BasicNP matching centers on the issue of linguistic variability. Linguistic variations in natural languages allow the same semantic information to be expressed syntactically in many different ways but all with the same meaning. Specifically a BasicNP concept can be expressed syntactically in several ways. For example, the BasicNPs "*sources of calcium*" and "*calcium sources*" or "*anthrax medicine*" and "*medicines that treat anthrax*" are all semantically equivalent while using different surface forms. In the following sections, we will explore these possibilities further in BasicNPs:

I) In English the two constructions of the genitives, the s-genitive and the of-genitive, in special cases are interchangeable; that is the s-genitive ('N1's N2) can be substituted with the of-genitive ('N2 of N1') or vice versa. In **Q15** and Q16 the two constructions of genitives are equivalent which would qualify the question pair for having embedded relationship.

> **Q15**: What is ***the capital of China***?
> Q16: What is the population of ***China's capital***?

II) There is also a strong tendency in English to use nouns as premodifier in order to avoid the post modifying of-genitive. This means that for some BasicNPs the constructions ('N2 of N1') and (N1 N2) can substitute each other keeping the semantics unaltered. In **Q17** and Q18 the two constructions are equivalent and embedded relation holds between the question pair.

> **Q17**: What are some ***sources of calcium***?
> Q18: How do the ***calcium sources*** fit into the overall diet?

III) The third category of the interchangeable BasicNPs involve the noun-noun compounds and constructions with adjective clauses in which the head noun is modified by a relative clause introduced by a relative pronoun/adverb (i.e. that, which, who). In **Q19** and Q20 the two constructions are equivalent and the question pair is candidate for embedded relationship.

> **Q19**: What are ***some medicines that*** **treat** ***anthrax***?
> Q20: Who does manufacture ***anthrax medicine***?

In contrast to of-genitive, s-genitive and noun-noun compound constructions which have nouns as their main constituents, the BasicNP construction with adjective clause includes a predicate that also contains a main verb with possible auxiliaries, e.g. *treat* in **Q19**. The verb of the clause acts as the backbone of the assertion being made and defines the semantic roles and the semantic relation between the two nouns in the adjective clause. The matching process of semantically equivalent BasicNPs cannot simply ignore the verb of the clause and base its decision solely on the matching of the corresponding nouns. Referring to the question **Q19** and Q20 this means that for the two corresponding BasicNPs in the two questions to match, in addition to the requirement that their corresponding nouns need to match, the verb of the clause in **Q19** also needs to match the interpretation of the act being performed between the two nouns in the noun-noun compound "*anthrax medicine*" of Q20.

The commonly accepted interpretation of the act in the noun-noun compound "*anthrax medicine*" is that of *treatment*, stemmed from the verb *treat* and therefore matches the verb in the clause of question **Q19**. Therefore, in this case the assertion being made in the relative clause of **Q19** is equivalent to the commonly accepted interpretation of the act in the corresponding noun-noun compound of "*anthrax medicine*". But what if in a coherent and valid question with similar

construction as **Q19** the verb indicating the assertion was other than the commonly accepted interpretation of the underlying semantic of the noun-noun compound. We therefore conclude that, as explained below, this semantic interpretation, preferably in the form of a verb, needs to be added as additional criteria for BasicNP matching when an adjective clause is involved. When human compound the noun *medicine* with a disease name such as *anthrax* resulting in *anthrax medicine*, our intuition always points to the general concept of "*treatment*". We tested this claim by issuing an exact phrase Google query of "*medicines* THAT * *anthrax*" where THAT stands for *that*, *which* and obtained the following phrases:

medicines that ***treat*** *anthrax*
medicines that ***will fight*** *anthrax*
medicines that ***tackle*** *anthrax*
medicines which ***may help*** *in anthrax*

All the verbs in the extracted phrases; "***treat***","***fight***", "***tackle***", "***help***"; are all semantically equivalent acts in the context of *medicines* and *disease names* as they all point to the general concept of diagnosing the disease "*anthrax*" albeit with different level of emphasis. Although in the case of *medicines* and *anthrax* as the extraction above shows all the extracted clauses are equivalent in meaning and we do not expect to see clauses such as "*medicines that cause anthrax*", for a noun-noun compound such as "*headache medicines*" some of the clauses extracted using the exact phrase Google query of "**medicines THAT * headache"** shows a different picture.

The verbs "***cause***", "***gives***", and "***provide***" obviously are not the most commonly accepted interpretation of the act being made by the utterance "*headache medicines*", although the respective questions using these clauses, e.g. "*What are some medicines that cause headaches*", are certainly coherent and valid. Again to emphasis, these examples demonstrate the fact that in BasicNP matching when a relative clause is involved, the verb indicating the underlying semantic relation of the nouns in noun-noun compound must also be considered in the matching process.

medicine which **relieved** the headache
medicines that **is useful** for headache
medicines that **could relieve** the headache
medicines *that* **would cure** his headache
medicines that **prevent** headache
medicines that **counteract** headache

medicines that **cause** headache
medicine that **gives** you a headache
medicines that **can provide** headache

4. STRATEGIES FOR IMPLEMENTING REUSE

To begin with, the embedded question reuse facility can be added as a modular component to the baseline QA system and enabled optionally. In a QA system with the reuse mechanism enabled there will be an initial stage to preprocess the user question to take advantage of the reuse opportunity for performance gain. If the preprocessing does not indicate a potential case of reuse, normal processing of the base line system will continue. Figure 1 shows the overall processing logic of the proposed approach. In the following sections we will first define some concepts used throughout the rest of the paper and then outline number of strategies and point to the important issues of design for reuse.

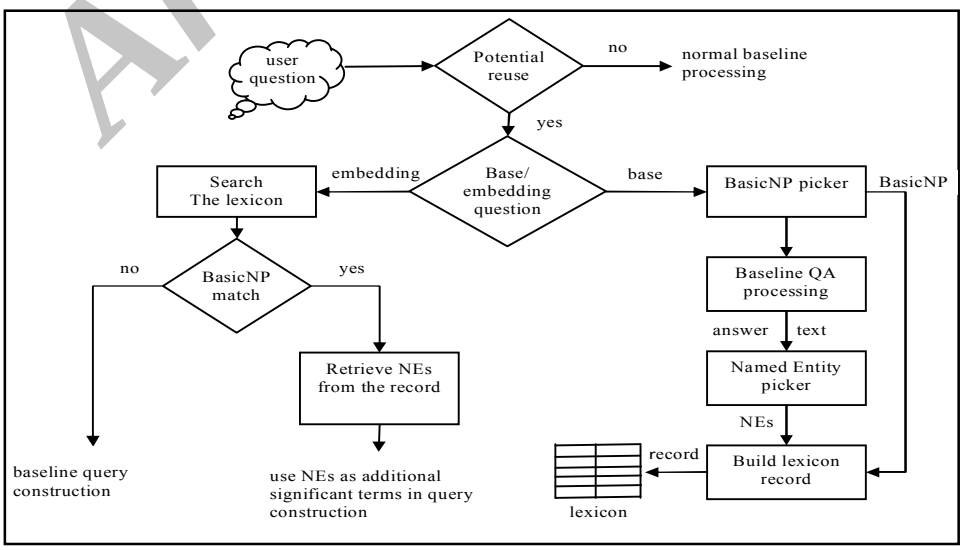

Figure 1. The overall processing of the proposed approach

4. 1. Frames, Records, and the Lexicon A frame is a structure used to capture the BasicNP constituents of a base question. The frame structure has a maximum of three slots namely h, m, and v representing the head, modifier, and the main verb of a clause within a BasicNP, respectively. A binary frame has two slots and is displayed as [h,m] and a ternary frame has all the three slots and it is displayed as [h,m,v]. In the following example the BasicNPs are represented with their equivalent frames:

BasicNP	Frame
capital of China	[capital,China]
medicines that treat anthrax	[medicine, anthrax, treat]

The frame structure normalizes different surface forms of semantically equivalent BasicNPs caused by linguistic variability into a standard structure suitable for BasicNP matching.

A record is defined as a higher level structure with two fields. The first field is a frame as described above that represents a BasicNP in a base question and the second field called the BasicNP referents groups the named entities referents extracted from the answer of the same base question. In the following example the frames are joined with their corresponding referents to make two records:

[capital, China]	Beijing
[medicine, anthrax, treat]	Ciprofloxacin, Tetracyclines, Doxycycline, Penicillins

And finally a lexicon is a collection of records that will be used for frame lookups to retrieve its corresponding named entities.

4. 2. Processing of Base Questions This section describes the reuse preprocessing required for the user base question. When this preprocessing is complete the question is sent to the baseline QA system processing pipeline for normal operation. For each submitted user question we need to identify whether the question is a base or an embedding question or otherwise not related to the reuse facility. If it is an unrelated question the question is sent to the baseline QA system with no further preprocessing.

When a question is recognized as a base question the BasicNP part of the question is extracted and its constituents (i.e. head, modifier, and the verb of the clause) are packaged into a frame. The frame which is really an abstraction representing the BasicNP in the base question is unified with the named entities extracted from the answer of the same base question obtained from the answer extraction phase of the baseline QA to form a record. This record is then entered into the lexicon.

The BasicNP frame entries in the BasicNP lexicon is used for looking up BasicNPs frames extracted from the embedding questions. When the lookup is successful the corresponding named entities of the matched frame entry is retrieved and used as additional significant terms in query generation of the baseline QA to answer the embedding question. The following sections describe these processes.

4. 2. 1. Recognition of Base Questions The base questions are simple trivia like questions with a BasicNP whose head is the focus and its modifier is the topic of the question (*e.g. What are some **medicines that treat anthrax**?*). The following general regular expression extraction templates cover variety of simple base questions (this work does not handle temporal base questions):

(what|which|where|who) (is|are) BasicNP

The linguistic patterns for BasicNP component of the above template need to be defined for the recognition process and also to be used for the extraction of the BasicNP constituents. A noun phrase in English from an abstract point of view has the general form:

NP → det pre* head post*

where *det* is the determiner and can consist of article, quantor, number, etc., *pre*, the premodifier is the adjective, noun, or coordinated phrase, *head* usually a noun, *post* (postmodifier) is the propositional phrase, relative clause, etc., and asterisk(*) denotes zero or more occurrences. As mentioned previously, in this work we are only concerned with basic noun phrases (termed as BasicNP) and we only consider single word nouns.

In Table 2 we define the regular expressions of the linguistic patterns in terms of POS and phrase labels. These patterns are used for the recognition of BasicNP part of a base question as defined earlier as well as the extraction of the BaiscNP constituents to construct frames for the lexicon explained in the next section.

As shown in Table 2, the linguistic patterns for three of the four categories of the BasicNP constructions can be realized with the POS tags of the words only. In contrast to these the adjective clause requires the capability of a shallow parser or a chunker (OpenNLP[2]) to demarcate the verb and noun phrases boundaries.

4. 2. 2. Construction of BasicNP lexicon The purpose of the BasicNP lexicon is to capture the BasicNP constituents of the base question into a frame and build lexicon entries by connecting the BasicNP frames to their corresponding referents. The referents are the named entities present in the answer passage of the base question.

[2] http://opennlp.sourceforge.net

TABLE 2. Recognition and extraction patterns

BasicNP	Example	Patterns
Of-genitive	Capital of China	(<NN>|<NNS>|<NNP>|<NNPS>) of (DT)? (<NN>|<NNS>|<NNP>|<NNPS>)
S-genitive	China's capital	(<NN>|<NNS>|<NNP>|<NNPS>) POS (<NN>|<NNS>|<NNP>|<NNPS>)
Noun-noun compound	Anthrax medicine	(<NN>|<NNS>|<NNP>|<NNPS>) (<NN>|<NNS>|<NNP>|<NNPS>)
Adjective clause	Medicines that treat anthrax	[NP some_DT medicines_NNS] [NP that_WDT] [VP treat_VBP] [NP anthrax_NN] ?_.

Of the four constructions of the BasicNPs three of them have only two nouns but the adjective clause construction also includes a main verb with possible auxiliaries. For each of these BasicNP constructions a frame is constructed using the constituents head, modifier, and in the case of adjective clause the main verb. The information to fill the slots comes from the extraction of words in BasicNP of base questions using the extraction patterns specified in Table 2.

For the first three categories of BasicNPs in Table 2 the extraction patterns specified in the first three rows of Table 2 are very straight forward. In each case the corresponding head noun; *capital, capital, medicine* in rows 1 to 3, respectively; and modifier noun; *China, China, anthrax* in rows 1 to 3, respectively; is extracted to build a new frame and use these nouns to fill the h and m slots of a two slots binary frame. To compensate for the morphological variations when searching the lexicon, the nouns in the frame slots can be reduced to their roots using a stemmer such as the Porter stemmer. The stemming is done for all the BasicNPs except for the noun-noun compounds as these nouns will be used to generate equivalent BasicNPs with relative clause from a text corpora as explained later. For the adjective clause BasicNPs we extract the verb between that_WDT (that,which,who) and the following NP including proposition, if any. We ignore and drop any adjectives or participles that falls between the verb and the preposition. Also, the modals and auxiliaries are allowed and ignored, but the passive *be* is kept. Finally, we convert the main verb to an infinitive using WordNet [12]. We also pick the noun in the first NP chunk that contains a noun as the head and the noun in the NP after the verb as the modifier and fill a three slots frame, a ternary frame, with the information extracted.

The other field connected to each BasicNP frame in the lexicon is one or more named entities that exist in the answer passage retrieved by processing of the base question by the baseline QA system. Most QA systems retrieve answer passages from the text collection that contain entities(s) matching the expected answer type (EAT) of the question. These entities are marked by a named entity recognizer (NER) and used by the answer extraction phase of the baseline QA. These marked named entities are extracted and attached to the frame corresponding to the BasicNP of the base question being processed and entered as a record into the lexicon. Table 3 shows an example of a lexicon.

TABLE 3. Frame lexicon

BasicNP FRAME	BasicNP Referencs
[capital, China]	Beijing
[medicine, anthrax]	Ciprofloxacin, Tetracyclines, Doxycycline, Penicillins
[medicine, anthrax, treat]	Ciprofloxacin, Tetracyclines, Doxycycline, Penicillins
[types, bacteria]	cocci, bacilli, vibrios, spirochactes, staphylococci, streptococci
[rivers, Asia]	Irtysh, Han, Habur, Ganges
[animals, extinct]	Leopards, Rhinos, Gazelles, Pandas, Tigers, Komodos Dragons
[medicine, headache, cause]	Contraceptives, bronchodilators, alcohol, nitrates, carbonmonoxide
[medicine, headache, relieve]	acetaminophen, ibuprofen, ketoprofen, naproxen

4. 3. QA with Embedded Questions

4. 3. 1. Recognition of Embedding Questions An embedding question is a question that is linked to a base question and whose BasicNP is semantically equivalent to the BasicNP of the base question. The linkage between the two questions is such that, as shown in previous examples, the BasicNP part of the base question is more commonly embedded in the topic of the embedding question (e.g. *who does manufacture **anthrax medicine**?*) or sometimes in its focus (e.g. what ***anthrax medicine*** is approved by FDA?).

Questions normally have a main topic and a questioner asks a question by focusing on a particular aspect of the main topic. In the following examples annotated by a POS tagger the target BasicNP "*anthrax medicines*" is the topic of the question Q21 and the question is asking about the manufacturer of this topic. In Q22 the focus is on the target BasicNP "*anthrax medicine*" and the topic is the "*FDA*". In Q23 "*side effect*" is the focus with "*anthrax medicines*" forming the topic of the question and both the focus and topic have the syntax of a target BasicNP.

Q21: WP/who VBZ/manufactures NN/anthrax NNS/medicines./?
Q22: WP/what NN/anthrax NN/medicine VBZ/is VBN/approved IN/by DT/the NNP/FDA./?

Q23: WP/what VBP/are NN/side NNS/effects IN/of NN/anthrax NNS/medicines./?

As these examples show the target BasicNP can appear anywhere either in the focus or the topic or in both the focus and the topic. A question is considered an embedding question if 1) it contains a noun phrase matching the structure of a BasicNP and 2) its BasicNP can be matched in the lexicon. To find a BasicNP construction in the candidate embedding question the same syntactic regular expressions described above for BasicNP extraction in base questions can be used. When a BasicNP is present in the candidate embedding question, it has to fulfill the second requirement above by which it also needs to match a frame in the lexicon entries. This matching process for the BasicNPs of types noun-noun compounds or adjective clause are different from the genitives and are explained in the next sections.

4. 3. 2. Questions with Noun-noun Compound or Adjective Clause BasicNP The frames extracted from these group of embedding questions are binary frames [h,m] for the noun-noun compound and ternary frames [h,m,v] for adjective phrase BasicNPs. These frames are used as search frames to match against frames in the lexicon. When the search is successful the corresponding named entities referents of the located frame are retrieved from the lexicon and used in the query generation phase of the baseline QA processing. There are number of matching pair possibilities as depicted in Figure 2. Each pair along an arrow represents a candidate pair for the matching process.

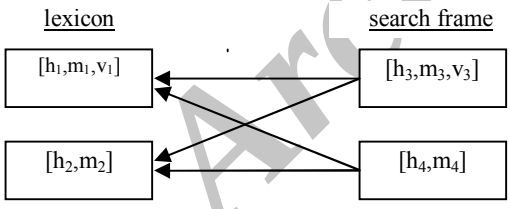

Figure 2. Frame matching pairs

Let's first consider the matching of an $[h_3,v_3,m_3]$ search frame. Before matching, the nouns and the verb are stemmed to their root forms. The $[h_3,m_3,v_3]$ search frame is initially matched against ternary $[h_1,m_1,v_1]$ frames in the lexicon. If the search is successful the corresponding named entities referents are retrieved from the corresponding referents field. However, if the search is not successful meaning that an equivalent adjective clause was not identified, binary frames $[h_2,m_2]$ in the lexicon entries whose stemmed h_2 and m_2 match with their counterparts h_3 and m_3 are alternatively considered. This is indicative of matching an adjective clause, e.g. "*medicines that treat anthrax*", with a noun-noun compound, *e.g.* "*anthrax medicine*". As explained in Section 3.2.1 on BasicNP matching, in addition to the matching of the corresponding h and m slots of the two frames, the matching process must also consider matching of the semantic relations between the two nouns of both frames and require their equivalence for a complete successful match of the two frames. This additional requirement necessitates the interpretation of the underlying semantic relation between the nouns in the noun-noun compounds of the binary frame $[h_2,m_2]$.

The interpretation of the underlying semantic relation in noun-noun compounds deals with the detection and semantic classification of the implicit relationship that holds between noun constituents in the compounds. This issue has been debated by many researchers in linguistics and various theories emerged as to the nature and extent of the implicit relationship. At one end Levi [13] proposed that the implied relationship is limited and can be stated in terms of a set of abstract relations (e.g. CAUSE, HAVE, MAKE, etc) while at the other end researchers such as Downing [14] claimed that the implied relationship is entirely unconstrained and cannot be exhausted by a finite listing of the relationships. In particular Levi [13] talks about the process by which a certain class of noun-noun compounds, known as complex nominals, are introduced into the language by elision of the predicates, e.g. "*medicines that treat anthrax*" → "*anthrax medicine*". Levi calls the predicate the Recoverably Deletable Predicates (RDPs). To the reader of the language, the structure of a noun-noun compound implicitly recalls the concept represented by the RDP through the cognitive process and understanding of the language.

One way to characterize this semantic relation is to consider the set of all possible paraphrasing verbs that can connect the two nouns [15]. Using verbs to represent the semantic properties of noun compounds is emphasized in many theories of noun compound interpretation and is appropriate in our effort of matching a [h,m,v] frame to a [h,m] frame since it will generate candidate verbs for the missing v slot of the [h,m] frame to match against the v slot of [h,m,v] frame. Nakov [15] built on the idea that the vast size of the text available on the Web make it a rich corpora to predict the semantic relation between nouns in noun-noun compounds and conclude that the semantics of a given noun–noun compound can be characterized by the set of all possible paraphrasing verbs that can connect the target nouns.

In the feasibility study performed on the method [16] Nakov collected two sets of paraphrasing verbs for a set of $noun_1$-$noun_2$ compounds referred to as *Levi-250 dataset*. One set is generated by a group of human subjects and the other set automatically extracted from

the Web using exact phrase queries such as "$noun_2$ THAT * $noun_1$", etc. against a search engine where THAT is a complementizer and can be *that, which, who*; and * stands for 0 or more (up to 8) instances of Google's star operator.

The study concludes that the verbs generated from the Web were generally good and had medium correlation with the verbs generated by the human subjects.

We propose this method to harvest paraphrasing verbs from the Web using exact phrase query "h_2 THAT * m_2" representing the underlying semantic relation of the nouns in a $[h_2,m_2]$ frame. Following the acquisition of text snippets from the search engine results some post-processing will be required to extract the paraphrasing verbs.

In the first step to help with POS tagging and shallow parsing of the snippets, the text before noun2 can be replaced with a fix phrase such as "*We look at the*" as suggested in [15] after which the snippet can be POS tagged and shallow parsed. From this point on the steps for extraction are very similar to the extraction of verbs from the base question with adjective clause as explained in Section 4.2.2. The extracted verbs are weighted with frequencies and from the set of these paraphrasing verbs the n top weighted verbs are selected.

For each verb of the final selection a new frame with h, m, and v slots is built. The h and m slots of the new frame is filled with the corresponding h and m values of the $[h_2,m_2]$ frame under consideration and its v slot is filled alternatively with the selected paraphrasing verb from the top selections of the Web harvest. The nouns and the verb values in the slots are stemmed and each newly generated frame is entered into the lexicon as a new member. All these frame point to the same named entities referents as their original $[h_2,m_2]$ frame. With the addition of these equivalent [h,m,v] frames to the lexicon, the original search frame $[h_3,m_3,v_3]$ can now be matched against these ternary frames.

The other lookup operation of the lexicon concerns matching of $[h_4,m_4]$ search frames to frames in the lexicon. These frames are initially stemmed and matched against stemmed binary $[h_2,m_2]$ frames in the lexicon and if the search result is not successful the same procedures to generate paraphrasing verbs as explained above for expanding a binary [h,m] frame to ternary [h,m,v] frames is followed and a set of top weighted verbs are collected. The verbs are used to fill the v slot of the binary search frame producing several equivalent ternary search frames.

These newly generated frames are stemmed and then used to look up equivalent ternary frames in the lexicon. When a match is found the same procedure to retrieve the named entities referents as explained above is performed and used for the query generation.

4. 3. 3. Questions with S-genitives and of-genitive BasicNP These questions are simpler to handle since there is no need to interpret the semantic relation between the nouns in the BasicNP as they are generally of a possessive or partitive nature. The binary frames corresponding to these BasicNP are stemmed and only matched against genitive frames in the lexicon. When the search is successful the corresponding named entities referent(s) is used in the query generation stage of the baseline QA system.

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 2 on related work, 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 [2]. Within the corpus, several instances of question sequence relating to reuse of embedded question sub-category are annotated.

To develop additional instances of question-answer set for the embedded reuse corpus, TREC QA document collections would be a valuable resource. The embedded reuse test collection requires two general types of questions. One set of questions should provide for what we called the base questions in the embedded relation. As pointed out these are trivia like questions that inquire about a BasicNP. The BasicNPs in these base questions help to populate the BasicNP lexicon. Much of the earlier work in the field of QA has centered around fact-based questions. Certainly selected questions in the data set of the earlier TREC QA track can be used with none or minor modification for the base set. These include TREC9 QA task [8] and TREC10 QA [4] main and list tasks data sets.

Questions within these data sets falling in the following general template that was used earlier for the recognition of base question :

(what|which|where|who) (is|are) BasicNP

are candidate for the base set. Below we list some of these base questions from TREC10 QA track. We also took the liberty to join the words if the head or the modifier consisted of more than one word. This will not cause any contradiction in noun phrase matching if we use the same joined words in the corresponding embedding question and disjoin them when presenting the question to the base line QA.

*What is **Australia's national_flower**?*
*Where is **John_Wayne airport**?*
*What is **the capital of Yugoslavia**?*
*What is **the population of Seattle**?*
*What is **the largest_city of the world**?*
*What is **the currency of Australia**?*
*Who was **the first_governor of Alaska**?*
*Who was **the first_female united_states_representative**?*
*What is **the Ohio_state bird**?*
*What is **Hawaii's state_flower**?*

The second set of questions required for evaluation involves the embedding questions. Once the base set is established, the embedding set can be generated using the BasicNP of each question as a topic. Multiple varying questions can then be generated on different aspects of each BasicNP topic. One can also experiment with different equivalent constructions of the same BasicNP by varying the BasicNP syntactically along the line of the four types of BasicNP constructions discussed in this paper. Once the embedding questions are generated the answers can be searched on the internet using a common search engine.

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 enhancement in several criteria of evaluation including speed, relevance, correctness, and conciseness both when the reuse mechanism is enabled and disabled.

6. CONCLUSION

In this paper we have demonstrated the fact that questions in QA scenarios are not asked in isolation but can be related through embedded relationship. This property provides the potential for performance gain of a QA system through the reuse of information about one question to answer the other question.

The linkage between the questions in an embedded relationship was identified to be a pair of basic noun phrases, termed as BasicNP. The BasicNP in the base question plays the role of a referencing expression and their referents are retrieved as part of the answer extraction of the base questions. These referent entities are then extracted from the answer passage and entered into a lexicon connecting with their corresponding frames representing BasicNP. The lexicon was then used to fetch the referent entities to help answer the embedding questions. These entities were used as significant terms in the query formulation phase of the embedding question resulting in more pertinent answers.

We presented techniques for the matching of different constructions of BasicNP including the noun-noun compound which required approaches to interpret the semantic relationship between the nouns in the compound. We presented paraphrasing verbs extracted from the Web to represent this semantic. Finally we presented ideas to generate the embedded reuse corpus for the purpose of reuse performance evaluation.

7. REFERENCES

1. Pasca, M. A. and Harabagiu, S. M., "High performance question/answering", in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, (2001), 366-374.
2. Light, M., Ittycheriah, A., Latto, A. and McCracken, N., "Reuse in question answering: A preliminary study", in New Directions in Question Answering: Papers from the 2003 AAAI Symposium, (2003), 78-86.
3. M. Mansoori and H. Hassanpour "Boosting passage retrieval through reuse in question answering", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 25, No. 3, (2012), 187-196.
4. Voorhees, E. M., "Overview of the tenth text retrieval conference (TREC-10)", ***International Speech Communication Association (INTERSPEECH'10), Brighton, UK***, (2001).
5. Voorhees, E. M., "Overview of the TREC 2004 robust retrieval track", in Proceedings of TREC , (2004).
6. Burke, R. D., Hammond, K. J., Kulyukin, V., Lytinen, S. L., 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.
7. Jeon, J., Croft, W. B. and Lee, J. H., "Finding similar questions in large question and answer archives", in Proceedings of the 14th ACM international conference on Information and knowledge management, ACM, (2005), 84-90.
8. Voorhees, E. M. and Harman, D., "Overview of the ninth text retrieval conference (TREC-9)", (2001).
9. Fleischman, M., Hovy, E. and Echihabi, A., "Offline strategies for online question answering: Answering questions before they are asked", in Proceedings of ACL. Vol. 3, (2003), 1-7.
10. Chai, J. Y. and Jin, R., "Discourse structure for context question answering", in Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004, (2004), 23-30.
11. Girju, R., Giuglea, A. M., Olteanu, M., Fortu, O., Bolohan, O., and Moldovan, D., "Support vector machines applied to the classification of semantic relations in nominalized noun phrases", in Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, Association for Computational Linguistics, (2004), 68-75.
12. Miller, G. A., "WordNet: a lexical database for English", ***Communications of the ACM***, Vol. 38, No. 11, (1995), 39-41.
13. Levi, J. N., "The syntax and semantics of complex nominals, Academic Press New York, (1978).
14. Downing, P., "On the creation and use of English compound nouns", ***Language***, (1977), 810-842.
15. Nakov, P. and Hearst, M., "Using verbs to characterize noun-noun relations", ***Artificial Intelligence: Methodology, Systems, and Applications***, (2006), 233-244.
16. Nakov, P., "Noun compound interpretation using paraphrasing verbs: Feasibility study", ***Artificial Intelligence: Methodology, Systems, and Applications***, (2008), 103-117.

Investigating Embedded Question Reuse in Question Answering

M. Mansoori [a], H. Hassanpour

[a] *Department of Electrical and Computer Engineering, Babol University of Technology, Babol, Iran*
[b] *School of Information Technology & Computer Engineering, Shahrood University of Technology, Shahrood, Iran*

P A P E R I N F O

Paper history:
Received 30 August 2012
Received in revised form 17 October 2012
Accepted 18 October 2012

Keywords:
Question Answering
Information Retrieval
Reuse
Noun Phrase Matching

چکیده

در این مقاله روش جدیدی برای ارتقاء عملکرد سیستم‌های پرس‌وجو (QA) با زمینه باز ارائه می‌گردد. با بهره‌گیری و استفاده مجدد از اطلاعات موجود در پرس‌وجوهای قبلی سیستم قادر خواهد بود پاسخ‌ها با دقت بیشتری را برای پرس‌وجوهای آتی با زمینه مشترک پرس‌وجو قبلی تولید نموده و به ارتقاء عملکرد دست یابد. این قابلیت استفاده مجدد بین زوج پرسش‌ها با زمینه مشترک هنگامی امکان‌پذیر است که یکی از پرسش‌ها از طریق یک عبارت گروه اسمی در پرسش دیگر بطور ضمنی نهفته باشد. جهت شناسایی این رابطه استفاده مجدد بین زوج پرسش‌ها روشهای تطبیق گروه‌های اسمی در زبان طبیعی مورد بررسی قرار گرفته و پس از تشخیص رابطه استفاده مجدد بین زوج پرسش موجودیت‌های نام‌دار در پاسخ پرسش جزئی‌تر استخراج و از آنها بعنوان کلیدواژه‌های مهم و مفید در مزخله تولید جستار (query) سیستم پرس‌وجو و بازیابی اطلاعات برای تولید پاسخ پرسش جامع‌تر استفاده میگردد. در پایان روش‌ها و منابع مورد نیاز برای تولید انباره پرس‌وجوها و ارزیابی عملکرد سیستم ارائه میگردد.

doi: *10.5829/idosi.ije.2013.26.03c.04*