

# A Classification Scheme and Corpus for Reuse in Question Answering

Abraham Ittycheriah<sup>†</sup>, Andrew Latto<sup>‡</sup>, Marc Light<sup>°</sup>, Nancy McCracken<sup>◇</sup>

<sup>†</sup> IBM TJ Watson Research Center, NY, abei@us.ibm.com

<sup>‡</sup> Metacarta, PO Box 397207, Cambridge, MA 02139, andy@metacarta.com

<sup>°</sup> The MITRE Corporation, 202 Burlington Rd., Bedford, MA 01730, light@mitre.org

<sup>◇</sup>Center for Natural Language Proc., Syracuse University, Syracuse, NY 13244-4100, njm@ecs.syr.edu

June 4, 2002

Please send info on affiliation: IBM, MetaCarta, MITRE, Syracuse University

## 1 Introduction

Current information retrieval systems and search engines help users to find documents that are relevant to their needs, but leave it to the user to extract the useful information in those documents. In particular, users often have questions for which they would like answers, not documents. Question answering systems aim to allow users to ask questions such as “Which terrorist organizations have issued threats on U.S. embassies this year?” and to receive succinct answers. Many initial end-to-end systems have been built and have been evaluated as part of the TREC Question Answering Track (Vorhees 1999, 2000, 2001). The best of these systems are able to answer factual questions, similar in style to trivia questions, 70% of the time, while searching a million newswire texts.

The ARDA Advanced Question Answering for Intelligence Analysts Program (AQUAINT) aims to push this state of the art into new realms of question types, document types, media types, etc. One aspect of an advanced question answering system would be that it would accumulate questions, answers, and other auxiliary information derived in the process. This information could then be “reused” to enable the system to better answer future questions. In this way, a system could duplicate a human’s ability to gain knowledge in an area as she or he answers questions.

The research presented here was inspired by this vision of a system improving with respect to a topic as it processes questions. However, it was not immediately clear how to implement this vision in a general way. The next step is to analyze instances of “reuse”:

- What types of “reuse” existed?
- How often do the different types appear?
- How could a system capitalize on different reuse opportunities?

This paper describes a corpus of questions, the Reuse Corpus, that the authors collected. (The Reuse Corpus is freely available from the authors.) Using this corpus, we found twelve categories of reuse. In addition, we explored (without actual implementations) different ways systems could operationalize reuse.

Before going any further, let us look at an example. Consider a scenario where a number of users are asking questions about anthrax, a topic that was previously unexplored by the user base and the QA system's developers. A sequence of questions and document excerpts might be:

*Q1 What are some medicines that treat anthrax?*

*D1 The FDA has approved Cipro (ciprofloxacin), tetracyclines including doxycycline, and penicillins to treat anthrax. (from www.medlineplus.com)*

*Q2 What are some of the side effects of anthrax medicines?*

*D2 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. (from www.medlineplus.com)*

*Q3 Who manufactures anthrax medicine?*

*D3 Cipro is produced in the U.S. by the German pharmaceutical company Bayer AG. (constructed)*

This sequence might be from a single user or from different users. The important feature of these question-passage pairs is that Q2 and Q3 should be easier to answer if the system can make use of the answer to the previous questions. For example, knowing that Cipro is a medicine and that it is used to treat anthrax should help the system find information about the side effects of anthrax medicines. Note that D2 and D3 do not mention anthrax. A similar sequence of questions and documents is below.

*Q4 What is anthrax?*

*D4 Anthrax is an infectious disease caused by the spore-forming bacteria called Bacillus anthracis. Infection in humans most often involves the skin (cutaneous anthrax), ... (from www.medlineplus.com)*

*Q5 Where can anthrax be obtained?*

*D5 Los Alamos National labs has stored samples of B. anthracis spores. (constructed)*

*Q6 Which countries have anthrax cultures?*

*D6 In this NOVA episode, they mentioned the American Type Culture Collection (ATCC), which had made the embarrassing mistake of selling B. anthracis cultures to Iraq. (slightly modified email message)*

Again answers to the initial questions provide crucial information for answering later questions. Here the crucial information is that *anthrax*, *anthrax culture*, and *Bacillus anthracis* are close to equivalent.<sup>1</sup>

It is useful at this point to contrast reuse with two topics that have recently received attention in the question answering community: discourse processing and redundant questions.

Discourse processing was explored in the TREC-10 QA context subtrack. Some primary concerns of discourse processing include tracking the focus of a discourse, resolving anaphora and coreference in general, and ellipses. Such information is useful for reuse but does not subsume

---

<sup>1</sup>We are indebted to Alexander Morgan for examples Q1 through Q6.

it. For example, *cipro* and *penicillin* are not coreferential with *anthrax medicines* in the examples above. However, the TREC-10 QA context subtrack data did contain a number of examples of reuse corresponding annotations are included in the annotation file for the Reuse Corpus.

Redundant questions were explored in TREC-9, a number of questions were reformulated and included in the test set. Certainly an important form of reuse is to recognize that the same question, in different words, has been asked and answered before. However there are many other forms of reuse as will be made clear in section 3.

The goals of the research described in this paper were to:

- introduce and demarcate the topic of reuse in question answering,
- develop a set of categories of reuse,
- find examples of these categories of reuse,
- make recommendations with respect to future work on reuse.

(Note that the development of the categories and the search for examples of reuse were interleaved processes.)

The rest of the paper is structured as follows. The next section will describe how the Reuse Corpus was collected. Then the 12 reuse categories are described. Next strategies for implementing reuse are discussed along with methods for evaluating system performance. We then summarize and present ideas for future work. An appendix is also included with specific recommendations for how to proceed with a 6-8 week workshop on the topic of reuse.

## 2 Constructing the Reuse Corpus

The Reuse Corpus consists of two types of question sets, where a question set contains questions about a single topic. One type is single user sets where a single user asked and ultimately answered a sequence of questions while performing a particular information gathering task. One example is finding information on calcium needed in a woman's diet. Note that such question sets are really question answer sequences. The second type of question set contains questions from multiple users. Such question sets were collected by doing keyword searches on web search engine logs. Continuing the example above, a search for *calcium* was performed on questions collected from a web search engine log.

The collection of data for the Reuse Corpus occurred at three different sites. IBM and Syracuse University collected single user question sets while MITRE collected multiple users sets.

At IBM twenty-one topics were chosen to represent current newsworthy events. The question sequences were generated by an independent person and one of the co-authors. The directions for generating these questions were to write questions whose answers might enable writing a research report on the topic. The 147 questions were generated in this manner and then answers were searched for on the internet using a common search engine. Table 2 lists the categories and the number of questions in each category.

At the Syracuse University Center for Natural Language Processing research, the staff (the subjects) assisted in generating question sequences. The goal was for each of the 10 subjects to generate a sequence of 10 questions and answers on an information gathering task of interest to them.

The subjects were primarily graduate students at Syracuse University, two from Computer Science and five from Information Studies. The remaining three subjects were full-time research staff.

Table 1: Topics and the number of questions in each topic for the IBM segment of the reuse corpus.

Topic	Question Count
Challenger Space Craft	8
Choosing between a macintosh and a pc	5
Columbine High School shooting	6
Dave Koresh and Waco	9
Death Penalty	8
Ebay business model	7
End of Apartheid in 1993	11
Global warming	5
Human cloning	4
Martha Stewart Corporation	7
NAFTA	6
Pathfinder lands on Mars	8
Terrorism in the 21st Century	4
The Brady Bill	7
The collapse of the Soviet Union	9
The fall of Enron	10
The Israeli-Palestinian conflict	6
Timothy McVeigh	7
Who are the Taliban	9
World Trade Center - Structural integrity	5
World Wide Web	6
Total	147

The problem was stated to the subjects in terms of researching a topic of interest or in accomplishing some task. In the absence of a question answering system, each person emulated such a system by using a search engine, or other web searching techniques such as following links, to first find documents that contained the answer to the question and then to find the answer in a document. The URL was recorded for each answer document, and each answer where feasible. The subjects were also instructed not to generate the question sequence in advance, but to generate questions that arose in their minds either from their own knowledge or from reading answer documents from earlier questions. The subjects worked in a group in a large shared office so that questions and issues could easily be settled among the group.

The resulting questions were usually just one sentence without complex structure, but that could have been influenced by the fact that they were using a search engine. Many questions had simple factual answers, but others asked about more complex issues. And some question sequences consisted almost entirely of questions that were too general for the short answer format, e.g., *How do I build a manipulable arm for a robot?*. The more general sequences were eliminated to obtain 74 total questions on 8 single user topics. The topics were: heel pain, calcium, ice cream, poetry, Saskatoon (Saskatchewan, Canada), special education, and Washington, D.C.

The multiple users sets were collected at MITRE. The topics included all of Syracuse's single user topics and a number of additional topics: Tiger Woods, digital cameras, Tupac Shakur, the Oscars, the Stanley Cup, the 2000 Summer Olympics. We looked at these additional topics because we wanted more focus on people and events. The digital camera topic was picked because it was of interest to one of the authors.

The questions came from a popular web keyword search engine during the summer of 2000. Over many weeks, queries to the search engine were logged. A

regular expression like the one below was used to find questions in the log: (who|whom|what|when|where|why|how|which|^is|^name|^show|^find me|)

Over a million questions were extracted. This question corpus was then searched using relevant keywords for the topics. In most instances the topic names themselves were sufficient keywords. There were over 2000 questions for the 15 topics. Peregrine Falcons had the least with 8 questions and poetry had the most with 631. The average was 154. The Reuse Corpus only contains a small fraction of these question due to copyright restrictions. The questions that are included are rephrasings of the originals.

The questions from these 3 sources were then placed in a standard format and combined to form the Reuse Corpus. For each question, the question, the topic, and if available, an answer and the URL's of relevant web pages are specified in the standard format. Each question was given a unique identifier and this identifier is given in the examples below prefixed by *I*:

### 3 Reuse Categories

To review: a reuse opportunity occurs when the fact that a question answering system has processed one question can allow the system to provide improved subsequent performance. This possibility is dependent on the relationship between the original and the subsequent questions asked of the system. We have attempted to provide a classification of the possible relationships between questions that provide reuse opportunities.

We first break these relationships down into three broad clusters. The first cluster consists of those reuse opportunities that only require the system to track the questions it is asked. The second cluster involves those questions in which the reuse opportunity is apparent merely from looking at the answers to the questions. A third cluster consists of those reuse categories where both the questions and answers must be examined in order to determine that a possibility for reuse exists. Each of these three clusters is then further broken down into categories.

#### 3.1 Data Annotation Strategy

Instances of reuse categories were annotated for the Reuse Corpus. We both analyzed a portion of the corpus one question at a time, looking for instances of any sort of reuse opportunity and we scanned the questions with a particular category of reuse in mind. Although we found a good number of instances of the different categories, we are far from a complete analysis of the corpus.

We annotated the reuse “tokens” using the following format.

```
c:<userid>:<reuseTypeLabel>:<questionNumber,...>:comment
```

Each line in the file lists one instance of potential reuse, and the questions that are involved in that particular reuse instance. A single question and answer pair can participate in multiple reuse types. There are currently eleven type labels: `sameQuestion`, `subSuperQuestions`, `embeddedQuestions`, `questionFacts`, `questionClarifications`, `sameDocuments`, `inductiveAnswers`, `answerLimits`, `sameFacts`, `topicCharacterizations`, and `evidences`. They are described in detail below. After the last question number, there is an optional comment field on which we imposed no structure.

## 3.2 Question-oriented Cluster

### 3.2.1 sameQuestions

The simplest reuse category, `sameQuestions`, occurs when multiple users of a question answering system ask the same question. The question may be asked in an identical form, or in a reformulation which can potentially be recognized by the system as having the same semantic content. One example from our annotated corpus is the pair of questions below:

<i>TOPIC:235: 2000SummerOlympicsMU</i> <i>Q: What countries have qualified in soccer for the 2000 Olympics?</i> <i>I:235</i>
<i>Q: What countries will compete in soccer at the Olympics in 2000?</i> <i>I:237</i>

Another example is the pair of questions

<i>Q: What is the moral status of human cloning?</i> <i>I: 84</i>
<i>Q: What are the ethical issues for human cloning?</i> <i>I: 87</i>

What we refer to as the `sameQuestions` reuse category is essentially the same as question reformulation in TREC-9.

### 3.2.2 subSuperQuestions

The next simplest opportunity for reuse occurs when the information requested by one question is a subset of the information asked for in another. If the question asking for the smaller quantity of information is encountered first, the system already has a partial answer for the question asking for more information. If the question asking for the larger quantity of information is asked first, and this reuse opportunity is recognized by the system, the performance benefit is even greater, since the only action required to answer the second question is to select from the information provided as the answer to the first question those components that comprise an answer to the second question.

An example of this reuse category is given by

<i>Q: How to wrap an achilles tendon to relieve pain?</i> <i>I: 328</i>
--

and

<i>Q: How do I treat a sore achilles tendon?</i> <i>I: 327</i>
---

A correct answer to the first question is a component of an answer to the second question, but does not provide a complete answer, since there may be other ways to treat a sore achilles tendon other than wrapping.

A second example is given by

<i>Q: Who are the Taliban ?</i> <i>I: 98</i>
---

and

*Q: What is the Taliban's religious background ?*

*I: 100*

since a description of the Taliban's religious background provides a partial, but not a complete, description of who the Taliban is.

### 3.2.3 embeddedQuestions

As was discussed in the section 1, questions like

*Q1 What are some medicines that treat anthrax?*

*Q2 What are some of the side effects of anthrax medicines?*

*Q3 Who manufactures anthrax medicine?*

contain an opportunity for reuse: knowing that cipro and penicillin are anthrax medicines allows a system to more effectively find answers to questions about the side effects and manufactures of these medicines.

We have named this kind of reuse **embeddedQuestions** because answers to earlier questions turn out to be crucial to answering later questions. If question Q2 were to be asked in isolation, a human question answerer is likely to first try to figure out what medicines treat anthrax, i.e., question Q1.

Let us consider another example. In question I:12, knowing how much calcium is recommended could play a role in finding out how much Vitamin D is needed, as in the following questions:

*Q: How much calcium should an adult woman get in her diet?*

*I: 11*

and

*Q: How much Vitamin D do you need in order to absorb the recommended calcium?*

*I: 12*

Similarly, in question I:18, it is useful to know what the dietary sources of calcium are before one tried to find out how they fit into an overall diet.

*Q: What are good dietary sources of calcium?*

*I: 15*

*Q: How do the calcium sources fit into the overall diet?*

*I: 18*

The relation between questions often involves a pair of noun phrases. For example, *anthrax medicines*, *recommended calcium amount*, and *calcium sources* are normalized forms of a noun phrase in their respective question pairs. The relation does not seem to simply be one of coreference. In contrast, strict identity coreference is exemplified by TREC-10 context questions:

<i>Q:What type of vessel was the modern Varyag?</i> <i>I: CTX7a</i>
<i>Q: In what facility was it constructed?</i> <i>I: CTX7b</i>
<i>Q:In what country is this facility located?</i> <i>I: CTX7c</i>
<i>Q:How many aircraft was it designed to carry?</i> <i>I: CTX7e</i>

### 3.2.4 questionFacts

Often the questions themselves contain information that can be useful for answering later questions. For example, the manufacture of Tiger Woods' ball can be extracted from I:2014 and this might be useful for answering I:2019.

<i>Q: What's the name of the Nike golf ball Tiger Woods uses?</i> <i>I: 2014</i>
---

and

<i>Q: What golf ball did tiger woods use in the us open?</i> <i>I: 2019</i>
--

Such information may also be less direct. For example, the quotes in the first question below indicate that *Gone with the Wind* is a title. This information might be useful when trying to find answers to the second question.

<i>Q: How many awards did "Gone with the Wind" win?</i>
---

<i>Q: Who starred in gone with the wind?</i>
--

### 3.2.5 questionClarifications

Question answering systems often have trouble answering vague and unclear questions. This difficulty is exacerbated by always attempting to answer questions in isolation, rather than making use of reuse opportunities by examining related questions. The content of these related questions, even if they are asked by different users of the question answering system, can provide useful data that clarifies for the system the intended meaning of the vague or unclear question. We refer to this category of reuse as **questionClarifications**.

Consider the question pair below.

<i>Q: Which machine is more user friendly, a macintosh or a pc?</i> <i>I: 77</i>
---

and

<i>Q: What machine comes with a better package of software, a macintosh or a pc?</i> <i>I: 78</i>
--

The question answering system may find the first of these questions difficult, as it may not understand what is meant by a *user friendly* machine. But if the system was also exposed to the second question, it can infer that the user asking for a *user friendly machine* may be asking about the quality of the software that comes with the machine, since this is a matter of concern to other users.

To take another example, consider the question

<p><i>Q: Who poses the biggest threat to the United States ?</i></p> <p><i>I: 80</i></p>
--

There are many categories of possible threats, and the question answering system may not understand what sort of threat is being discussed. In addition, the use of the word *who* may mislead a naive system into restricting its search to individual people, rather than countries and organizations, that pose a threat to the United States.

However a system that was able to recognize and utilize the `questionClarifications` reuse category and had already encountered the question

<p><i>Q: What countries are developing weapons of mass destruction ?</i></p> <p><i>I: 81</i></p>
--

could infer from this that weapons of mass destruction were an important category of threat to consider, and that countries as well as individuals should be considered as possible answers, when answering the previous question.

### 3.3 Answer-oriented clusters

#### 3.3.1 sameDocuments

The questions in this type of reuse are related because they could be answered by one document. These questions occur in both the question sequences and in the multi-user questions where questions in one topic area could be answered by a document discussing various aspects of this topic. This type of reuse has an obvious operational realization where a question answering system could save a core pool of documents (or knowledge base of information obtained by processing those documents) on particular topics so that answering future questions on that topic may be much more efficient.

In the Reuse Corpus, the occurrence of using the same documents to answer two or more questions was so ubiquitous that we only annotated a few of these instances of reuse.

In the question sequence topic of heel pain, there are three questions specifically about achilles tendonitis.

<p><i>Q: What is achilles tendonitis?</i></p> <p><i>I: 3</i></p> <p><i>A: Inflammation of the Achilles tendon.</i></p>
--

<p><i>Q: What are the causes of achilles tendonitis?</i></p> <p><i>I: 4</i></p> <p><i>A: Excessive running or jumping especially without proper stretching and strengthening are the most common causes of achilles tendonitis. Uphill running in particular can cause this condition.</i></p>
--

*Q: How can achilles tendonitis be treated?*

*I: 5*

*A: (Answer is too long to be given here.)*

It is quite common for a web document on medical conditions to define the condition, give the causes and symptoms, and give recommended treatments.

A series of questions about a particular event often provide `sameDocuments` examples. For example, it is common to find a news article that summarizes many facts about the event that can be the answers of many of the questions. Below are three of the four questions from the topic Columbine High School shooting.

*Q: Where is Columbine High School ?*

*I: 136*

*A: Littleton, CO*

*Q: When and what happened at Columbine ?*

*I: 137*

*A: Two students walked into the school at 11:15 a.m. and fired shots from a multi-gun arsenal and lobbed homemade bombs throughout the school*

*Q: How many people died ?*

*I: 139*

*A: Twelve students and a teacher*

The following excerpt from a news story from [www.thedailycamera.com](http://www.thedailycamera.com) answers these questions.

”On April 20, 1999, two students walked non-chalantly into Columbine High School at 11:15 a.m. and fired shots from a multi-gun arsenal and lobbed homemade bombs throughout the school. At the end of the rampage, 12 students, one teacher and the gunmen were dead. The following articles, photos, audio clips and polls recount the events of that day as well as the recovery of the wounded and the remembrance of the victims. Below, are articles from the first five days after the shooting. At right, key stories are divided into certain issues surrounding the tragedy.”

### 3.3.2 inductiveAnswers

In this type of reuse, the knowledge of previous questions and answers allows the answer to new question to be induced. In some cases, parts of the answer can be induced and in others it may just be noted that the answer is similar to previous ones. In either case, the induced answer is to be considered a hypothesis that must be checked. Operationally, this may enable the system to find answers faster or to find better answers than a more direct answer lookup would find.

Consider the questions below.

*Q: What irons does tiger woods use?*

*I: 2011*

*A: Titleist*

*Q: what loft one wood does tiger woods use?*

*I: 2012*

*A: Titleist*

*Q: what putter does Tiger Woods use?*

*I: 2013*

*A: Titleist*

If a system has found the answer to any one of the questions to be the manufacturer Titleist, then it could guess that the manufacturer of the other types of golf clubs is the same. However, it would need to allow for the possibility that he does use golf clubs from more than one manufacturer.

### 3.3.3 answerLimits

Another category of reuse occurs when the answer to one question provides information that limits the possible answers to a second question. This can aid the system in searching for an answer to the second question, since the limitation may provide a corresponding limitation on the places in the corpus where it is necessary to search in order to answer the second question. An example is given below.

*Q: What is the weather like in Saskatoon, especially in the winter?*

*I:53*

*A: Generally not too much precipitation, and it can get very cold in the winter.*

*Q: What is there to do outdoors in Saskatoon?*

*I: 54*

The answer to the first question would provide the useful information that possible answers to the second question might well include snow skiing, but would be very unlikely to include water skiing. This example is also notable in that it shows that the more common-sense real-world knowledge is possessed by a question answering system, the more opportunities for reuse it will be able to recognize and make use of. Only a system that had some understanding of the appropriate temperatures for various real-world activities would be capable of recognizing this reuse opportunity.

This category has some overlap with the **subSuperQuestions** reuse category described above. Any time one question asks for a collection of data items, and a second, more specific, question asks for a single one of these data items, or a collection that is a subset of the first, the answer to the first question will provide a limit to the possible answers to the second question. So in some sense, any example of the **subSuperQuestions** category can also be considered to be an example of the **answerLimits** category. To reduce the amount of multiple categorization, we consider a reuse example to fall in the **answerLimits** category only if it is not apparent simply by examining the questions that the answer to the first question puts limitations on the possible answers to the second. Since **answerLimits** is in the answer-oriented cluster of reuse categories, we only classify a reuse opportunity as **answerLimits** if it only becomes apparent from the answer to the first question that this answer puts limits on the possible correct answers to the second question.

## 3.4 Question and Answer combination clusters

The following types of reuse require caching both questions and answers to previous questions.

### 3.4.1 sameFacts

This type of reuse occurs when the same fact is needed to answer two different questions, even though it is not apparent that the questions are not asking for the same information until at least

one of them has been answered. An example of this phenomenon is shown in the following pair of questions:

*Q: What medicines cure anthrax?*  
*A: Cipro, and other antibiotics*

*Q: What does Cipro cure?*  
*A: anthrax*

Simply examining the questions provides no evidence of any connection or reuse opportunity. But the single fact "Cipro cures anthrax" provides an answer to both questions, so a system that has answered one of these two questions, and recorded this fact, can answer the second question with no further processing.

### 3.4.2 topicCharacterizations

When a group of questions are on the same topic (determination of which is itself a research topic), there may be several lexical items that help to identify the topic. These items may additionally disambiguate the other question words. In the following example, the word *tennis* which characterizes the first question can assist in disambiguating which King is being mentioned in the subsequent question.

*Q: Who did Billy Jean King beat in battle of the sexes?*  
*A: When tennis champion Billie Jean King accepted the half-humorous challenge Bobby Riggs threw down in 1973, most women (and a lot of men) watched the televised "Battle of the Sexes" with rapt attention.*

*Q: How many titles did King win?*

### 3.4.3 evidences

The idea of this category of reuse is that some complex questions benefit from related questions that provide evidence for or against particular answers to the question. For example, to answer I:2058 below

*Q: Is tupac still alive?*  
*I: 2058*

it would be useful to know the answer to all the following questions.

<i>Q: how tupac really died</i> <i>I: 2041</i>
<i>Q: where can i find a picture of tupacs death</i> <i>I: 2081</i>
<i>Q: what happened between tupac and suge knight</i> <i>I: 2063</i>
<i>Q: When did tupac die?</i> <i>I: 2070</i>
<i>Q: where can i find the newspaper entitled is tupac alive</i> <i>I: 2110</i>
<i>Q: who killed tupac shackur</i> <i>I: 2143</i>

If the system could provide pictures of Tupac’s death or name the killer or know a time, this would help answer negatively to question I:2058. In addition, such questions and answers would support the answer not just help find it. The relation between these questions is heavily dependent on world knowledge and thus operationalizing this form of reuse would initially require a restricted domain such as events of unnatural death. It is this event-specific knowledge that sets apart examples of the evidences reuse category from `subSuperQuestions` and `embeddedQuestions` categories.

### 3.5 Answer Revisions

When a particular type of reuse is detected and applied to a pair of questions, the possibility arises that earlier answers may need modification. This is not strictly a form of reuse but rather an action that might be taken when a reuse condition is detected. In the reuse corpus, an instance of this can be found in the following Q&A pairs.

<i>Q: How much calcium should an adult woman get in her diet?</i> <i>I: 11</i> <i>A: Adult women need 1000-1500mg per day (pre- and post-menopausal) and also need Vitamin D to absorb the calcium. But they should not take over 500mg at one meal.</i>
--

<i>Q: What are the factors that increase calcium excretion?</i> <i>I: 19</i> <i>A: high protein diet, caffeine, alcohol, sodium, low exercise. Note that the minimum daily requirements seem to assume some of these, so a lifestyle without these factors may require fewer mgs of calcium per day.</i>
--

Once the answer to the second question is found, the first answer should be modified to indicate further clarifications.

## 4 Strategies for implementing reuse

There are a variety of techniques that one can imagine to enable a question answering system to improve performance via the reuse categories we have identified. In this section we will outline a small number of them. However, how well such techniques work in practice would require actual implementation and experimentation.

The simplest reuse technique is to keep a cache of all questions previously asked of the system, and the answers provided by the system. This gives a potential performance benefit for questions that exemplify the reuse categories in the question-oriented cluster, namely `sameQuestion`, `subSuperQuestions`, `embeddedQuestions`, `questionFacts`, and `questionClarifications`. In order to achieve this enhanced performance, the question answering system must be sufficiently sophisticated to identify the relationship between the questions asked, and recognize that these questions are related in the way described in the given reuse category.

A question answering system has additional reuse potential when, in addition to caching the questions it has answered, it keeps a record of the answers it uses, and the documents used by the system to obtain these answers. This should make reuse that falls in the `sameDocuments` category fairly straightforward, and can potentially enable reuse in the more complex categories, such as `inductiveAnswers`, `answerLimits`, `sameFacts`, and `topicCharacterizations`.

A system might also

- keeps track of the semantic content of the answers it has already provided and
- updates this semantic content when perhaps as a result of a related question, it obtains additional information about the same topic.

It could then make use of reuse opportunities in the `answerRevisions` category.

In all cases, one of the most difficult parts of implementing reuse is the recognition by the system of the relationship between the various questions and answers it encounters, and the categorization of these relationships into the various reuse categories we have identified.

The system may be able to involve the user in an attempt to identify such relationships. This would involve an interaction with the user that is much more complicated than a straightforward sort of interaction where the user asks questions and the system responds with answers. A system might employ a more sophisticated interaction model where the user provides direct feedback about the answers and documents and perhaps even about the inner workings of the system. For example, the user may indicate whether she finds the answer provided by the system satisfactory. Maintaining a record of this information has obvious utility when the system identifies future reuse in the `sameQuestion` or `subSuperQuestions` categories. If the system provides access to several corpus documents in response to a question, there is additional information available in the form of which documents the user actually accesses, and in what order. The user can also be given the opportunity to provide the system with information concerning the perceived usefulness of these documents. This information can also be useful for reuse.

Possible modes of user feedback in a question answering system, and the ways in which this feedback can assist the system both in identifying and categorizing reuse opportunities, and obtaining maximum performance enhancement as a result of the reuse, is an important area for future research.

Regardless of method used to implement reuse, one will have to evaluate its effectiveness. A simple method is to compare its performance with that of the system with the reuse mechanisms turned off. Thus, one could perform user-centric evaluations such as in the TREC interactive track or the more traditional information retrieval evaluations like the TREC QA track evaluations. In either case, one would hope for an improvement in performance when the reuse mechanisms turned on when compared to the performance of the baseline system.

## 5 Future Work

This paper represents a start on collecting and analyzing data on reuse possibilities in question sets. However, more work of this kind is needed:

- We developed a corpus of about 200 questions and answers from single users. A corpus of many hundreds of questions was collected for multiple users. In order to do the analysis of the categories of reuse, we annotated various types of reuse. However, our goal was primarily to find examples of reuse and to investigate the different categories and we did not complete the annotation. Furthermore, as we annotated, our understanding of the categories of reuse evolved, and it would be interesting not only to complete the annotation but to re-annotate what was done to ensure that good consistent annotation.
- When we developed the corpus of questions, we did not have firm goals about the types of topics that QA systems would be particularly interested in. We ended up with many topics of a general scope and also topics about issues. We believe that topics that contained more people and events are also of great interest to QA systems, and it would be good to develop more topics along these lines.
- For the analysis for this workshop, we wanted to obtain documents of timely and topical interest, so we used the Web as a document source. However, the use of Web documents in the corpus poses some problems. One is the issue of intellectual property: we do not have the rights to include the actual documents in the corpus and we only used the URLs. The other is that this is not a static document set which can be used by future researchers for comparative experimentation. The TREC QA document collection, however, is such a static document collection. We feel that it would be very valuable to develop some question sequences that could be answered from documents in the TREC collection. One method would be to start with a current TREC question and to continue to develop a topic and sequence around that question.
- Finally, we have not collected enough annotations of reuse to estimate the distribution of the categories of reuse. Such a distribution would be one of the valuable measures of the importance of the various categories, particularly any categories that are either extremely dense in the data or extremely sparse.

## 6 Conclusion

In this paper we have

- introduced and demarcated the topic of reuse in question answering,
- developed a set of categories of reuse,
- found examples of these categories of reuse,
- discussed strategies for implementing reuse,
- outlines future work.

It is our hope that this paper combined with its corresponding question corpus and annotations will provide a starting point for further research in reuse. In the next section we make specific recommendations for research in the form of a six to eight week workshop involving six to ten researchers.

## 7 Recommendations for a full workshop

- **Collect and annotate a larger Reuse Corpus.** We feel that to have reasonable development and test sets, we need to annotate about 750 Q&A pairs for 50 topics, therefore requiring about 15 questions per topic. A style book for the annotation should be developed and an experienced annotator should perform the annotation.
- **Extend two functioning question answering system.** The two systems should be different in approach. For example, one could consist of statistical components and the other could be discrete and rule-based. Two component technologies that will have to be constructed are:
  - a question clustering module that recognizes that two questions are about the same topic and
  - a question set classifier that assigns reuse labels to question sets where the set might often contain two questions.
- **Focus on user interaction.** Mine the user for information relevant to identifying reuse categories and vetting information to be reused.
- **Evaluate the systems with and without the reuse modules.** Use both a TREC-style evaluation and a small user-centric evaluation.

## 8 Acknowledgements

The research this paper describes performed as part of a workshop funded by the ARDA Northeast Regional Research Center Program. The authors took part in this workshop and are grateful for the support.