Determining the Plausibility of Answers to Questions

Troy Smith, Thomas M. Repede, and Steven L. Lytinen
DePaul University
School of Computer Science, Telecommunications and Information Systems
243 S. Wabash Ave.
Chicago, IL 60604
{troy@smith.net, trepede@comcast.net, lytinen@cs.depaul.edu}

June 10, 2005

Abstract

This paper describes AnswerOK, a system which determines whether the answer to a question is plausible. While not a question-answering system itself, it could be used by question-answering systems to check the plausibility of possible answers found in a large corpus (i.e., the TREC Q/A task). We explain the strategies used by AnswerOK, and present results which suggest that AnswerOK would likely improve the performance of many TREC Q/A systems.

1 Introduction

During the last several years, interest in computer systems that can perform question answering has grown, in large part due to the Text Retrival Conference's Question Answering track (Voorhees and Tice, 2000; Voorhees, 2003). While recent TREC Q/A competitions have included more complex types of questions (e.g., questions which require a list of answers, potentially assembled from several different documents), generally the types of questions have been limited to those which can be answered with short answers, or so-called "factoid" questions. Here is an example:

Q: What city is Disneyland in?
Acceptable answers: Anaheim, Paris, or Tokyo

The correct answer to a factoid question is typically a single noun phrase. Although these questions are somewhat limited, the performance of TREC Q/A participant systems shows the complexity of retrieving an answer to a factoid question from a large corpus of text. For example, performance of the top 15 groups on factoid questions in TREC 2003 ranged from a maximum of 70% accuracy to a minimum of 14.5%, with mean performance of about 28.5% (Voorhees, 2003).

TREC data is available from the Linguistic Data Consortium (`http://www.ldc.penn.edu`). The TREC-2003 data includes the answers produced by all 25 participating groups to each of the 380 factoid questions in the testset, although the data does not identify which group produced which answer. Upon examining this data, one finds that many systems produce answers to questions which indicate a lack of "common sense". For example:

Q: What city is Disneyland in?
A: visit www

Q: How far is it from Earth to Mars?
A: one scientist

Q: How did Patsy Kline die?
A: Loretta Lynn

This paper describes a system, called AnswerOK, which attempts to judge the plausibility of an answer to a question. It is not a stand-alone TREC Q/A system; rather, AnswerOK could serve as a back end to such a system, determining whether or not an answer found in the corpus by the system is plausible. In theory, if our system was integrated with an existing TREC Q/A system, then it could reject implausible answers and request that the Q/A system try again to find a different (and hopefully plausible) answer.

AnswerOK uses a mixture of hand-crafted rules based on a small set of question types, plus a neural net which is trained on the TREC 2003 Q/A factoid corpus. The system utilizes two knowledge sources in order to help it determine if an answer to a question is plausible: WordNet (Miller, 2000), and the Google Web search engine (`www.google.com`). In this paper, we describe the process by which we developed our system, and then we present the results of testing. The results indicate that our system would increase the accuracy of the average TREC Q/A system.

2 Plausible Answers

AnswerOK is meant to distinguish between plausible and implausible answers, as opposed to correct vs. incorrect answers. So, for example:

Q: What city is Disneyland in?
A1: Hollywood
A2: Anaheim

Our system is not intended to be able to discrimate between Hollywood (a plausible but incorrect answer) and Anaheim (a correct answer). This is because our knowledge bases, and the way that we utilize them, give us general knowledge about certain objects (e.g, WordNet tells us via its *hypernym* links that Hollywood and Anaheim are both cities), but not specific information about locations of objects or other relations between objects (although the WordNet glosses sometimes indicate other kinds of relationships between words).

2.1 Building a corpus of plausible answers

To construct AnswerOK, we used a subset of the 2003 TREC Q/A factoid questions, along with all answers generated by the 25 participating groups for those questions. The subset consisted of 246 questions, and 5095 answers to these questions. This served as our "training set": as we constructed the hand-crafted portion of AnswerOK, we restricted ourselves to looking only at this corpus, so that once we had completed development, we could use a completely fresh set of data to test our system's performance. This corpus was also used to train the neural network portion of the system, to be described later in this paper.

The TREC dataset indicates which answers are correct; of the 5095 answers generated by the 25 groups, only about 15% of the answers were correct. However, since we were interested in judging plausibility of answers rather than correctness, we manually inspected the 5095 answers and determined which answers were plausible and which were implausible. This determination was based on our own intuitions, but was also quite simple; for example, for a question that starts with "Which city ...", we deemed any answer that provided a city name to be plausible. Specific knowledge about people and other objects was not used to restrict answers. For example, in response to the question, "How old was Babe Ruth when he died?" one group produced the answer "1939". Although knowledge about how long people live would elimiate this answer, we deemed it to be plausible, because it was a possible age (although not an age for humans).

Otherwise reasonable answers which contained extraneous words were judged by us to be implausible. For example, the answer "Impressionist Paris" in response to the question "In which city is the River Seine?" was deemed implausible, because of the extra word "Impressionist" in the response.

After manually inspecting all answers, we judged 2224 out the 5095 answers in the training set, or about 43.7%, to be plausible.

3 AnswerOK – General Design

3.1 Question Groups

The first step used by AnswerOK to determine if an answer is plausible is to determine the type of question asked. The general approach of categorizing questions in order to help determine an expected type of answer is common in TREC Q/A systems (e.g., Moldovan *et al.*, 2000). In AnswerOK, question categorization is very coarse, and is based simply upon the interrogative pronoun which begins the question (Who, What, Where, etc.) with one special group: HowMany. HowMany questions are differentiated from How questions in knowing that the answer should be a number. The questions are then reduced to a somewhat finer degree of granularity during further processing of some of these question types, as is described below.

Our question categorization is in contrast to the very fine-grained categorizations used by some TREC Q/A systems such as Webclopedia (Hovy *et al.*, 2002). We found that finer-grained categorizations added little benefit to the AnswerOK task of judging answer plausibility, as opposed to the more difficult TREC task of extracting an answer from a large corpus.

3.2 Question tagging

To facilitate further processing of questions based on their question type, AnswerOK uses a combination of taggers to provide further information about the words in the question. First, we use two named entity recognizers: GATE (Cunningham *et al.*, 2002) and Lingpipe (Bagga and Baldwin, 1998). We use the union of the results of the two systems: if a phrase in a question is recognized by either system as a named entity, then the phrase is replaced in the question by its named entity tag.

Our experience with the named entity recognizers is that they perform much better on declarative sentences rather than questions. As a result, AnswerOK rewrites each question with its candidate answer embedded in the appropriate place before passing the question to the named entity recognizers, in order to give them enough context. Figure b1 details how questions are automatically rewritten for the question groups that use named entity recognition.

These rewritten questions are usually grammatically correct. However, even when the transformed question is ungrammatical, we have found that the named entity recognizers are still able to correctly identify the entities. Manually fixing these questions and resubmitting to the named entity recognizers did not have an impact.

Next, AnswerOK uses part of speech taggers on the questions. Again we use two tools for this task: LT-Chunk (Mikheev, 2000) and tree-tagger (Schmid, 1997). Each tagger's results for a question are attached to the question to be used by the question type's algorithm. Wordnet and Google are also used by some of question types. These uses are detailed below.

3.3 Answer types

Since AnswerOK is a back-end system which determines the plasibility of an answer to a question, system input includes both a question and a candidate answer. To determine plausibility for many of the question types, AnswerOK determines an *answer type*, or a rough categorization of the candidate answer. Our set of answer types includes the following:

What: What→ answer
What was the Hunchback of Notre Dame's real name?
Quasimodo was the Hunchback of Notre Dame's real name?

Who: Who → answer
Who created the literary character Phineas Fogg? Jules Verne created the literary character Phineas Fogg?

Which: Which and all words up to first verb → answer
In which city is the river Seine?
In Paris is the river Seine?

Where: Where and all words up to the first noun.
If an 'is' verb is present then replace with: answer is
Where is Mount Olympus?
Greece is Mount Olympus?
Otherwise replace with answer is where
Where did Roger Williams, pianist, grew up?
Utah is where Roger Williams, pianist, grew up?

Figure 1: Questions rewritten as answers prior to named entity recognition

cause_of_death	month
city	number
country	percent
date	proper_name
day	state
letter	time_of_day
location	time_unit
measure_unit	year
money	

The system uses a combination of tools to determine answer type. First, WordNet is used to see if the words in the answer can be located in WordNet's synonym set (or *synset*) hierarchy. Only the nouns in the answer are used for this purpose. If the answer contains more than one noun, then synsets are located for each. Then, a check is performed to see if there is a hypernym path connecting the synset(s) with any of the possible answer types listed above. For example, if the answer contains the word "Chicago", there is a hypernym path which connects sense 1 of "Chicago" with the WordNet "location" synset, via "city", "municipality", "urban area", "geographical area", and "region".

AnswerOK also uses named entity recognition to determine answer type. As with processing of questions, both GATE and Lingpipe are used. If either recognizes a phrase as a named entity, then the named entity tag determines answer type.

4 AnswerOK – Processing of Question Types

4.1 "What" Questions

"What" questions make up the largest question group in the TREC-2003 testset, accounting for 55% of the 380 questions. "What" questions also pose a challenge in determining if an answer is plausible. Unlike Who or When questions, for which an appropriate answer type can be determined just by the question category, a What question needs to be further analyzed. Specifically, AnswerOK identifies one or two key noun groups (or named entity tags) in the question, which are then compared with the answer to determine plausbility.

The part of speech tagger results are used to determine the key noun groups in the question. The algorithm for finding these noun groups is detailed below (key noun groups are in [...]).

1. Find noun phrases on either side of a "be" verb
   e.g., What [city] is [Disneyland] in?

2. For other verbs, identify the noun phrase(s) between "what" and the verb.
   e.g., What [book] did Rachel Carson write in 1962?
   e.g., What [flavor filling] did the original Twinkies have?

After one or more key nouns are identified in the question, these are compared with the nouns/named entity tags in the answer to determine its plausibility. If more than one key noun is extracted from the question, then they are each compared with the answer; any resulting match indicates that the answer is plausible. For each key noun in the question, AnswerOK retrieves the definition of the root form of the noun from Wordnet. Then, if any of the following conditions are true, we assume that there is a connection between the question and answer, and therefore determine that the answer is plausible.

1. The WordNet gloss for the question noun contains a noun in the answer.

2. The WordNet gloss for a noun in the answer contains one of the key nouns in the question.

3. The question noun is of type "number" and the answer has a number named entity.

4. The question noun is of type "person" and the answer has a person named entity.

5. The same named entity appears in the question and the answer.

6. There is a hypernym path between the answer and one of the question nouns

7. There is a hypernym path between one of the question nouns and the answer

If any of the first 5 conditions above are matched, the confidence score (which is used to combine the rule-based results with the neural network) is 1.0. A match involving hypernyms yields a confidence score of 0.5.

If none of the above conditions are true, then AnswerOK builds a query to send to Google. There are four types of queries that are currently constructed. Each is constructed by combining one of the key nouns from the question with a noun from the answer. If any of the queries returns documents, then AnswerOK deems the answer to be plausible, but with a lower confidence score of 0.25. The queries are as follows:

1. Location queries

   A location query is constructed if the question contains one of the following key nouns: city, state, island, province, college, museum, country or county. If so, then a query of the form "question-noun of answer-noun" is sent to Google (in quotes to indicate that Google should search for the phrase). For example:

   > Q: What city was Albert Einstein born in?
   > A: Ulm
   > Question nouns: City, Albert Einstein
   > Google Queries: "city of Ulm", "Albert Einstein of Ulm"

   This query is particularly useful if WordNet or the named entity recognizers are not successful in identifying an answer as a location. In this case, Google can verify that "Ulm" is a city by retrieving documents from the query "city of ulm."

2. Business queries

   If the key noun in the question is a type of business, then AnswerOK quries Google with the candidate answer, but restricts the search to `www.hoovers.com`. This restriction helps to constrain the search to a local domain of knowledge and reduce false matches. For example:

   > Q: What company makes Bentley cars?
   > A: Rolls Royce
   > Question noun: company
   > Google Query: site:hoovers.com "Rolls Royce"

3. Music queries

   If the question asks about a band or a singer, then AnswerOK asks Google to search for the candidate answer, this time on `www.artistdirect.com`. The key noun from the question is also added to the query. For example:

   > Q: What band was Jerry Garcia with?
   > A: Grateful Dead
   > Question nouns: band, Jerry Garcia
   > Google Query: site:artistdiret.com "Grateful Dead" band

4. Names

   This type of query is contructed when analysis of the question indicates that the answer should be a name. For these instances the Google query is contructed as "name is [answer]" (again in quotes, to indicate a phrase). For example:

   > Q: What is the name given to a collection of poetry?
   > A: compendium
   > Question noun: name
   > Google Query: "name is compendium"

5. Generic queries

   This is the default type of query done if the analysis of the question cannot determine the query type. The query is constructed using the key nouns of the question. For example:

   > Q: What book did Rachel Carson write in 1962? A: Silent Sprint Question key nouns: Rachel Carson book Google Query: "Rachel Carson" book

   This type of query though suffers from finding the correct answer and not reasonable answers. For instance the Bible is a book but would probably not be returned in the results of this query.

4.2 "Which" Questions

Which questions are handled in exactly the same way as What questions.

4.3 "How" Questions

The how questions present a challenge in that there is not a straightforward way to determine if an answer is plausible. Due to this, AnswerOK further subclassifies "how" questions as follows:

1. If the question begins with "how late", then the answer type should be time_of_day
2. If the question begins with "how often", then the answer type should be a time_unit
3. If the question begins with "how old", then the answer type should be number
4. If "how" is followed by an adverb or an adjective, then the answer type should be a measure_unit
5. If a form of the verb "die" is in the question, then the answer type should be cause_of_death

These rules matched 89% of the training set questions.

4.4 "How Many" Questions

HowMany questions are a special case of a How question. A plausible answer must contain of a number, which may or may not be followed by a descriptive (a noun or noun group). The first step is to find the key noun in the question. The descriptive is assumed to be the noun which directly follows "how many"; for example:

How many [chromosomes] does a human zygote have?

If the noun group immediately following "how many" contains more than a single noun, then several descriptives are allowed to follow a number in the answer. For example:

How many [official languages] does Switzerland have?

In this case, the answers "3", "3 languages", and "3 official languages" are all judged to be plausible.

4.5 Other Question Types

AnswerOK uses simpler strategies for determining the plausibility of answers to other types of questions. For "Who" questions, the answer is run through the named entity recognizers. If all the words in the answer are identified as part of a person named entity, then the answer is marked as plausible with confidence 1.0. If some but not all of the words in the answer are tagged as such, then the answer is deemed plausible but the confidence score is reduced. If there are no person entities in the answer, it is assumed to be implausible.

"When" questions are also handled in a relatively straightforward fashion. If the answer's named entities include time_unit or time_of_day, then the answer is assumed to be plausible. While this approach correctly identifies the vast majority of the plausible answers, it is overly general. For example, one question in the TREC-2003 corpus is, "When is Jennifer Lopez's birthday?" The TREC-2003 participants generated the following answers:

1. `1994`
2. `24 Jul 70`
3. `Sunday`
4. `July`
5. `tomorrow`

All five answers are a time_unit and therefore are judged to be plausible by our algorithm. However, when we manually tagged the corpus, our human tagger only judged answer 2 to be plausible, since it is the only date. Further distinguishing between different kinds of times is one area in which AnswerOK could be improved on greatly.

"Where" questions are also handled chiefly by the named entity recognizers. If any of the named entities in an answer are a city, state, country or location, then the answer is deemed plausible. Otherwise the answer is determined to be implausible.

5 Results

As we developed AnswerOK, we only allowed ourselves to look at the training set that we had developed from the TREC-2003 corpus. For testing purposes, we used a subset of the TREC-2002 questions, specifically the first 108 questions. We did not look at the TREC 2002 questions until after we had completed development of AnswerOK based on the 2003 corpus.

5.1 TREC-2003 training set

After development of AnswerOK was complete, we ran the system on the training set. The following table shows the results, broken down by question type.

	How Many	How	What	When	Where	Which	Who	Total
Number of Answers	658	806	2847	475	21	155	133	5095
Correct	630	621	1905	354	18	120	121	3769
% Correct	95.7%	77.0%	66.9%	74.5%	85.7%	77.4%	91.0%	74.0%
Incorrect	28	185	942	121	3	35	12	1326
% Incorrect	4.3%	23.0%	33.1%	25.5%	14.3%	22.6%	9.0%	26.0%
Judged plausible	454	451	1623	447	18	109	86	3188
% Plausible	69.0%	56.0%	57.0%	94.1%	85.7%	70.3%	64.7%	62.6%
Judged implausible	204	355	1224	28	3	46	47	1907
% Implausible	31.0%	44.0%	43.0%	5.9%	14.3%	29.7%	35.3%	37.4%
Plausible/Correct	430	276	835	328	17	76	82	2044
% Plausible/Correct	99.1%	96.5%	84.4%	99.4%	100.0%	97.4%	91.1%	91.9%
Plausible/Incorrect	4	10	154	2	0	2	8	180
% Plausible/Incorrect	0.9%	3.5%	15.6%	0.6%	0.0%	2.6%	8.9%	8.1%
Implausible/Correct	200	345	1070	26	2	44	39	1726
% Implausible/Correct	89.3%	66.3%	57.6%	17.9%	50.0%	57.1%	90.7%	60.1%
Implausible/Incorrect	24	175	788	119	2	33	4	1145
% Implausible/Incorrect	10.7%	33.7%	42.4%	82.1%	50.0%	42.9%	9.3%	39.9%

5.2 TREC-2002 test set

After system development was completed, we ran AnswerOK on the TREC 2002 factoid questions. Here are the results.

	How Many	How	What	When	Where	Which	Who	Total
Number of Answers	78	296	1605	362	205	144	429	3119
Correct	64	225	1007	241	168	103	312	2120
% Correct	82.1%	76.0%	62.7%	66.6%	82.0%	71.5%	72.7%	68.0%
Incorrect	14	71	598	121	37	41	117	999
% Incorrect	17.9%	24.0%	37.3%	33.4%	18.0%	28.5%	27.3%	32.0%
Judged plausible	40	160	831	289	151	70	318	1859
% Plausible	51.3%	54.1%	51.8%	79.8%	73.7%	48.6%	74.1%	59.6%
Judged implausible	38	136	714	73	54	74	111	1200
% Implausible	48.7%	45.9%	44.5%	20.2%	26.3%	51.4%	25.9%	38.5%
Plausible/Correct	35	104	406	200	118	41	233	1137
% Plausible/Correct	79.5%	87.4%	78.2%	86.2%	96.7%	77.4%	87.0%	84.0%
Plausible/Incorrect	9	15	113	32	4	12	32	217
% Plausible/Incorrect	20.5%	12.6%	21.8%	13.8%	3.3%	22.6%	12.1%	16.0%
Implausible/Correct	29	121	601	41	50	62	79	983
% Implausible/Correct	85.3%	68.4%	55.3%	31.5%	60.2%	68.1%	48.2%	55.7%
Implausible/Incorrect	5	56	485	89	33	29	85	782
% Implausible/Incorrect	14.7%	31.6%	44.7%	68.5%	39.8%	31.9%	51.8%	44.3%

5.3 Relating AnswerOK accuracy to the TREC Q/A task

Ideally AnswerOK would be used in conjunction with a question answering system that would generate a set of candidate answers. These answers would all be tested for plausibility, and the implausible answers could be eliminated, thereby increasing the chances that the overall system would retrieve the corect answer. However, since we did not have access to any such system, we could not directly test AnswerOK on the TREC Q/A task.

As another way to analyze whether AnswerOK could improve the pefor-mance of a TREC Q/A system, we did the following: we considered all the answers generated by the TREC 2002 participants for each question, and calculated the probability of choosing a plausible answer for each question. As a baseline by which to compare AnswerOK's performance, we randomly picked an answer for each question from all the answers generated by all TREC participants. Then, we (manually) judged the plausiblity of these answers. On average then, baseline performance picks a plausible answer to a question 43.1% of the time. We then calculated the same probability, only this time we only considered answers generated by all TREC participants that AnswerOK deemed to be plausible answers. If there remained more than one possible answer to a question, we again picked randomly. Using AnswerOK to eliminate some implausible answers improved performance: on average, our system with AnswerOK included picked a plausible answer to a question 52.9% of the time.

For the TREC 2002 test set there were eight questions that our system deemed that there were no plausible answers. Four of these questions truly had no valid answers in the reasonable answer set. Of the remaining four questions, there were a variety of errors that caused no reasonable answers to be found: question put into the wrong question group, query too general, query returned no matching results, and the how question subtype not found.

6 Neural Network

We tried to duplicate our algorithm in a neural network by generating all of the properties that we used to determine if an answer was plausible for all What type questions. These properties were generated for both the TREC 2002 and 2003 question sets. A neural network was trained using the 2847 answers from the 2003 data set. We trained four different networks, with various hidden node layouts ranging from a single hidden layer with 25 nodes to multiple layers with up to 75 nodes per layer.

The neural network performed as follows:

% correct, TREC-2003 training set

	Plausible	Implausible
Network1	70%	88%
Network2	67%	90%
Network3	68%	87%
Network4	69%	90%

% correct, TREC-2002 test set

Network1	43%	89%
Network2	36%	90%
Network3	32%	90%
Network4	35%	90%

The neural network systems showed themselves to be adept at identifying implausible answers. However, their accuracy on plausible answers was much lower than the manually developed system. It appears that all versions of the neural network suffer from overtraining on the training set. This would explain the drop off in correctly identifying plausible answers.

7 Hybrid System

The two systems described above complement each other, in that the manual system performs better at correctly identifying plausible answers, while the neural network performs better at identifying implausible answers. Therefore, our next step was to build a hybrid system, with the hope that the combined judgement of the rule-based system and the neural network would provide a better balance.

To combine the judgements of both systems, we needed rules for breaking a tie if the judgements differed. We used the following rules:

1. if the neural network judges an answer to be plausible, then the combined judgement is that the answer is plausible
2. if the rule-based system judges an answer to be plausible with confidence 0.5 or higher, then the combined judgement is that the answer is plausible
3. otherwise, the combined judgement is that the answer is implausible

The performace of all four hybrid systems is shown below.

Plausible	H1	H2	H3	H4
Correct	73.1%	69.6%	71.3%	71.3%
Incorrect	26.9%	30.4%	28.7%	28.7%
Implausible				
Correct	69.7%	69.8%	70.3%	70.3%
Incorrect	30.3%	30.2%	29.7%	29.7%

To compare these sytems against the purely rule-based version of AnswerOK, we also calculated the random probability of picking a plausible answer from the set of answers that were deemed plausible by each of the hybrid systems.

Baseline	43.1%
Rule-based alone	52.9%
Hybrid1	56.8%
Hybrid2	54.7%
Hybrid3	49.6%
Hybrid4	55.6%

8 Decision Trees

To verify that the hand coded AnswerOK system was reasonable we took the attributes that AnswerOK used and generated decision tress for the different question types. The initial thought was the results of the decision trees would be comparable to the AnswerOK results. The first step was to modify AnswerOK to output the properties used in the determination of the plausibilty of an answer. These properties where then fed into the J48 algorithm of Weka(Witten and Frank, 1999) to build a decision tree.

8.1 "What/Which" Questions

We grouped the What and Which questions together to feed into the J48 algorithm. Initially we used the following 14 attributes in the decision tree:

1. AnsIsInQuestion - True if answer is in the question
2. AnsIsLFHypen - True if answer is in any key noun's hypernyms
3. AnsMatchedLFOver - True if answer is in any of the key noun's wordnet glosses
4. LFAnsPartialMatch - True if there is a partial match between the answer and key nouns
5. LFEntMatchedAndEnt - True if the named entities of any key noun matched any of the answer's named entities.
6. LFIsAnsHypen - True if any key nouns are in the answer's hypernyms
7. LFIsNameAnsIsPersonOrg - True if the answer is a person or organization and any of the key nouns is "name"
8. LFIsNumberAnsIsNumber - True if the answer is a number and one of the key nouns is "number"
9. LFIsPercentAnsIsPercent - True if the answer is a percent and one of the key nouns is percent
10. LFIsWordAnsIsWord - True if the answer is a single word and one of the key nouns is "word"
11. LFMatchedAnswerOver - True if one of the key nouns is in the answer's wordnet gloss.
12. LfAnsMaxPartialMatchScore - True if the partial match score was the highest possible score
13. nounFollowsQuestion - True if a noun follows the question
14. verbFollowsQuestion - True if a verb follows the question

With these attributes the decision tree was able to determine an implausible answer with 95.2% accuracy. However, only 50.5% of the plausible answers were correctly classified. Looking at the cases where AnswerOK correctly classified an answer but the decision tree did not show that the matches involving the Google queries were very important. In the first attribute set there was no information regarding the query or web matches. To address this issue we added the following attributes to the original 14:

1. doQuery - True if a web query is to be done for a given answer
2. queryType - The type of query done: ofQuery, bandQuery, companyQuery or generic
3. specificQuery - True if the queryType was not generic

4. webmatch - True if there were enough web matches to warrant a plausbile answer.
5. AnsShouldBe - The type of answer that is being looked for: city,band,state, etc.

Adding this attribute increased the accuracy of plausible answer detection from 50.5% to 53.3%. The unplausible answer determination was 95.1%. However, looking at the tree it appeared that some overfitting might have been taking place. Looking at the resulting tree we removed both queryType and AnsShouldBe. Rerunning the data produced the same results so we concluded that removing these attributes does not affect the determination.

Looking at the decision tree with the queryType and AnsShouldBe attributes removed contained a single leaf, see below, which correctly identified 4000 non-plausible answers. However, if also contained 768 incorrectly classified plausible answers. This leaf accounted for 82% of the error in predicting plausible answers. The difference between AnswerOK and this leaf is that a webmatch in AnswerOK signifies a plausible answer. To help increase the plausible answer score we turned to using a cost sensitive classifier to build the decision tree. Specifically we increased the cost when a plausible answer was missclassified. With this cost based specification we increased the accuracy of plausible answer detection to 74.8%. The non-plausible determination though fell to 77.5%.

LFIsAnsHypen = 0
— LFIsPercentAnsIsPercent = 0
— — LFIsWordAnsIsWord = 0
— — — LFIsNameAnsIsPersonOrg = 0
— — — — LFEntMatchedAndEnt = 0: 0 (4000.0/768.0)

Looking at the results of this final tree, the web query and matching was the source of a large percentage of the errors. Specifically, some queries were too long or too vague to return appropriate answers. These errors were also seen in the results from AnswerOK. The ability to construct a better query for these question and answers would help reduce this error. Another source of errors for non-plausible answers was incorrect matching between key nouns and the answer. An example of this:

Q: What is the area of western Germany that is known for its rich coal deposits?
Key Nouns: area, Germany
Hypernyms for Germany: country

Since Germany's hypernyms were used, which included country, any answer that was a country would match and be deemed plausible. In this example the removal of western from Germany caused AnswerOK to incorrectly determine that a country was an acceptable answer. A future experiment could also include adjectives in the key nouns and also not allow the breaking up of key nouns. This might help in this example however, if might overall cause a decrease in performance.

8.2 "How" Questions

The handling of How questions differed from AnswerOK in that there was no "How Many" subtype. Instead the attributes from both "How Many" and "How" question types were used. The initial attribute set included:

1. answerTypeEntMatch - If the answerType matched any of the answer's named entities using tree tagger results
2. answerTypeEntMatchOrig - If the answerType matched any of the answer's named entities using LTChunk results
3. howMany - True if the question is a "How Many" question
4. howManyAnswer - True if the answer is determined to be a plausible "How Many" answer
5. hypenMatchOfDescriptive - True if descriptive matches the answer's hypernym using tree tagger results
6. hypenMatchOfDescriptiveOrig - True if descriptive matches the answer's hypernym using the LTChunk results
7. isNumber - True if the answer is a number
8. needNumberMatch - True if it is determined that we are looking for a number using tree tagger results.
9. needNumberMatchOrig - True if it is determined that we are looking for a number using the LTChunk results

These attributes built a tree that was 69.8% and 96.8% accurate at predicting non-plausible and plausbile answers respectively.

There are three pairs of attributes that try and determine the same data using different part of speech results. These six attributes were merged into three combined attributes. Doing this did not change the results dramatically.

To help better predict non-plausible answers we added an answerTypeDefined attribute. This attribute is true when an answer type is defined for a particular question. However, adding this attribute did increase the results to 70.2% and 97.2 although it was not significant.

An interesting note is that the last tree produced some non-intuitive decisions. Running the instances through a cost sensitive classifer created a more intuitive decision tree, with the same performance.

8.3 "When" Questions

The when question decision tree had the potential to offer the most insight into how to improve AnswerOK. AnswerOK was only able to correctly predict 17.8% of the non-plausible answers. The initial set of attributes included:

1. ae-cause_of_death - answer named entity contained cause of death
2. ae-city - answer named entity contained city
3. ae-country - answer named entity contained country
4. ae-date - answer named entity contained date
5. ae-day - answer named entity contained day
6. ae-letter - answer named entity contained letter
7. ae-location - answer named entity contained location
8. ae-measure_units - answer named entity contained measure_units
9. ae-month - answer named entity contained month
10. ae-number - answer named entity contained number
11. ae-state - answer named entity contained state
12. ae-time_of_day - answer named entity contained time_of_day
13. ae-time_unit - answer named entity contained time_unit
14. ae-year - answer named entity contained year

The first decision tree was able to correctly indentify 90.2% of the reasonable answers and 69.7% of the unreasonable answers. Compared to AnswerOK the reasonable answers dropped from 99.6% the unreasonable answers had a dramatic increase from 17.8% to 69.7%.

Looking at the incorrectly classified reasonable answers there were two errors that were very common. The first is that some month abbreviations were not being identified. Specifically Jun and Jul were not recognized as abbreviations of June and July respectively. Also, when the answer was recognized as a full date the ae-month attribute was set to 0, unless the answer specifically contained a written out month. In all cases where the answer contained a full date we set the ae-month attribute to true. With these two changes we increased the accuracy of reasonable answer predicition to 95.9%. The unreasonable answer prediction fell slightly to 68.2%. The slight decrease in predicting unreasonable answers is not unexpected, in the sense that the two changes made were targeted at increasing the accuracy of predicting reasonable answers. To target the unreasonable answers we first added these three attributes;

1. birthdayInQ - The word birthday was in the question

2. bornInQ - The word born was in the question

3. dayInQ - The word day was in the question

This changed increased the unreasonable answer accuracy to 71.4% although it reduced the reasonable answer prediction to 94.3%.

A final attribute, extraWordsFound, was added. This attribute was set to true if after removing any word associated with a date from the answer two or more words were still contained within the answer. This change increased the unreasonable answer prediction to 75.5% and dropped the reasonable answer prediction to 93.6%.

Looking at some of the errors for predicting reasonable answers we noticed that some answers were of the form: 700 BC. These answers did not have their ae-year attribute set to true. Correctly this problem increased the ability to detect reasonable answers to 94.7% while just barely reducing unreasonable answer determination to 75.2%.

8.4 "Who" Questions

Who questions mirrored the AnswerOK results since there was only a single attribute, personNE. This attribute was true when the answer was identified as a person. A potential future improvement would to be to add a web query which would help determine if the answer was a name.

8.5 Decision Tree Results

After building the decision trees on the 2003-TREC data we used the 2002-TREC data to test the decision trees. Comparing the results from the AnswerOK and the decision trees when ran over the TREC-2002 test set.

Decision Tree TREC-2002 test set

	How	What/ Which	When	Who	TOTAL
Reasonable Answers Correctly Classified	82.8%	62.8%	86.5%	88.0%	75.5%
Reasonable Answers Incorrectly Classified	17.2%	37.2%	13.5%	12.0%	24.5%
Unreasonable Answers Correctly Classified	68.2%	74.5%	94.4%	58.1%	73.7%
Unreasonable Answers Incorrectly Classified	31.8%	25.5%	5.6%	41.9%	26.3%

AnswerOK TREC-2002 test set

	How	What/ Which	When	Who	TOTAL
Reasonable Answers Correctly Classified	85.3%	78.1%	86.2%	88.0%	82.9%
Reasonable Answers Incorrectly Classified	14.7%	21.9%	13.8%	12.0%	17.1%
Unreasonable Answers Correctly Classified	71.1%	56.3%	31.5%	58.1%	56.4%
Unreasonable Answers Incorrectly Classified	28.9%	43.7%	68.5%	41.9%	43.6%

NOTE: in the AnswerOK data the where data was removed. This was due to the fact that the TREC-2003 data only contained a single Where question which was too small of a test set to create a decision tree.

Looking at these results AnswerOK was better at predicting plausibily for How questions. The decision tree though was much better at answer When questions compared to AnswerOK. The What/Which questions though were about equal between the two systems. AnswerOK was better at determing plausible answers and the decision tree better at predicting unreasonable answers.

A simple hybrid would be to use the When decision tree and AnwerOK for all other questions. This would increase the Unreasonable Answers Correctly Classified to 60.9% while leaving the Reasonable Answers Correctly Classified at 82.9%. It would also be possible to create a hybrid similar to the neural network one described above to combine the What/Which results.

9 Future

There are a number of changes that can be made to the AnswerOK system that may improve performance. The query system for What questions could be enhanced in a number of ways. Another approach would be to handle the results returned in a different manner. As an example take the following question:

> What is the name of the volcano that destroyed the ancient city of Pompeii?
> Answer: Athens
> Answer Type: name volcano
> Current Query: “name is Athens”

This answer is determined to be reasonable because the phrase “name is Athens” is found in the query results. However, if the query was instead “Athens volcano name“ and we looked at how close Athens was to volcano to determine if a match occured we could conclude that Athens is not a reasonable answer. However, if the answer was St. Helens we would determine that St. Helens is indeed a volcano and would be deemed reasonable.

Another approach would be to have a list of nouns that should be ignored when followed by “of”. These nouns are typically nouns which represent a generic collection of items. Exaples of these nouns include name, kind and type. These nouns could be removed from the core noun list and the other core nouns could be used to determine if the answer is plausible. The nouns could also remain in the core noun list but not be allowed to match against wordnet output or Google search results.

The generic queries could also be enhanced by looking at the key nouns and determine what site has general information regarding the key noun. As an example:

Q: What book did Rachel Carson write in 1962?
A: Silent Sprint
Question key nouns: Rachel Carson book

For these types of questions we could identify what the answer should be, a book. Using Google we could do a simple query on book and use the first or first few sites returned to limit a search. For instance amazon.com was the first site returned when I searched for book. Now we can construct a query using a given answer and then limiting the search with the site: keyword to just amazon.com.

This approach would also work for tougher questions like:

Q: What are pennies made of?
A: copper
Question key nouns: pennies

Using just pennies as my search string the first site returned is www.pennies.org which contains a history of the American penny. Using this site to restrict the query would boast performance of this question significantly.

Another area where queries can help is in the Who section. Currently we do not do any queries for Who questions but we could build a query similar to the current name query. A few Who questions AnswerOK poorly predicated reasonable answers because names in the answer were not recognized by the named entity recognizers. In these cases it would be reasonble to expect a query of "name is [answer]" could be utilized to determine if the name is truely a name.

AnswerOK is also too lax in chosing a valid answer when the answer has extraneous words present. In this case AnswerOK does return a confidence score of 0.5 to illustrate this but it is possible to do more. Specificaly AnswerOK could return the words that are partially matching as the suggested correct answer and then mark the current answer as invalid.

10 References

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, May 1998,

Cunningham, H., Maynard, D, Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, July 2002.

Hovy, E., Hermjakob, U., and Ravichandran, D. (2002). A question/answer typology with surface text patterns. In *Proceedings of the DARPA Human Language Technology Conference (HLT)*, San Diego, CA.

Mikheev, A. (2000). Document centered approach to text normalization. In *Proceedings of SIGIR 2000*, p. 136-143.

Miller,, G. (1990). WordNet: An online lexical database. *International Journal of Lexicography, 3*(4), p. 235-312.

Moldovan, D., Harabagiu, S., Pasca M., Mihalcea R., Girju, R., Goodrum, R., and Rus, V. (2000). The structure and performance of an open-domain question answering System. in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, October 2000, Hong Kong.

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In Jones, D., and Somers, H. (eds), *New Methods in Language Processing Studies in Computational Linguistics*. UCL Press, London, p. 154-164.

Voorhees, E. (2003) Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text Retrieval Conference*. NIST Special Publication SP 500-255, p, 54-68.

Voorhees, E. and Tice, D. (2000). Building a question answering test collection. In *Proceedings of SIGIR-2000*, July, 2000, pp. 200-207.

Witten, H. Ian and Frank, Eibe. (1999). Data Mining: Practical Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.