Creating a Research Collection of *Question Answer Sentence Pairs* with Amazon's Mechanical Turk

Michael Kaisser, John B. Lowe

University of Edinburgh / Powerset Inc. , Powerset Inc.
m.kaisser@sms.ed.ac.uk, jblowe@powerset.com

Abstract

Each year NIST releases a set of *question, document id, answer*-triples for the factoid questions used in the TREC Question Answering track. While this resource is widely used and proved itself useful for many purposes, it also is too coarse a grain-size for a lot of other purposes. In this paper we describe how we have used Amazon's Mechanical Turk to have multiple subjects read the documents and identify the *sentences themselves* which contain the answer. For most of the 1911 questions in the test sets from 2002 to 2006 and each of the documents said to contain an answer, the Question-Answer Sentence Pairs (QASP) corpus introduced in this paper contains the identified answer sentences. We believe that this corpus, which we will make available to the public, can further stimulate research in QA, especially linguistically motivated research, where matching the question to the answer sentence by either syntactic or semantic means is a central concern.

1. Motivation

Since 1998, one of the sub-tracks in NIST's Text Retrieval Conference (TREC) has been the Question Answering (QA) track (see, for example, (Voorhees and Dang, 2006)). For each question in a given set of questions, participants' systems are expected to return an *answer, document id*-pair. These answers must be found in a provided document collection, but external sources (e.g. the Web) can be used to locate the answer as well. The document collection used from 2002 to 2006 was the *The AQUAINT Corpus of English News Text* (Graff, 2002).

At the end of each year's evaluation period, TREC releases a judgment file consisting of all *question id, document id, answer, judgment* quadruples returned by the participant's systems, e.g.:

```
1395 NYT19990326.0303 1 Nicole Kidman
```

Here, Question 1395 (*Who is Tom Cruise married to?*) has been answered with the string "Nicole Kidman". `NYT19990326.0303` is the identifier of one particular document in the AQUAINT corpus. The third column ("`1`"), in this case, indicates that the system returned the correct answer. This data has been used by researchers since then in a variety of ways; see e.g. (Echihabi et al., 2004; Monz, 2004).

But whenever the researcher wants to find the exact evidence for the answer provided, he or she has to look for it him/herself: no resource has been available that lists the *sentences* in these documents that provide evidence for the given answer. This is because NIST has not asked participants to provide such detail and, given the additional cost of doing so, NIST has not provided it themselves.

To address this gap, we collected the answer sentences for most *question id, document id, correct answer* triples for the years 2002 to 2006. There are 8,107 such triples in total that have been published by NIST during this period (counting only those that point to supporting documents). In addition, we identified the answer sentences for the *question id, document* pairs collected in (Lin and Katz, 2005). In that paper the authors attempted to locate *every* document in the AQUAINT collection that contains the answer, whereas TREC publishes only incomplete lists based on the documents that the actual participating systems regarded as relevant.

We believe the corpus we have produced, comprising *question, answer sentence, answer, doc id* tuples will facilitate research in QA and related areas in several ways:

1. The community would be able to better understand the various relations between the question and answer sentences – e.g.
   (a) average degree of word overlap;
   (b) grammatical relations or transformations between question and answer;
   (c) lexical relations between question and answer;
   (d) frequency with which anaphora mediates between question and answer;
   (e) frequency with which other discourse phenomena mean that evidence for the answer is spread across multiple sentences.
2. The data can also be used to train various kinds of statistical classifiers with the aim to improve a QA systems' performance.

We expect our resource to be especially helpful for research that is linguistically motivated. A syntactically or semantically inspired QA system will almost certainly contain a (very central) processing step in its pipeline where it matches the question to a set of candidate sentences. Our data presents a large set of positive training or testing examples.

[1]The authors wish to thank the following colleagues for their comments and assistance in preparing this article and the accompanying resource: David Graff, Marti Hearst, and Bonnie Webber. Any errors or omissions that remain are our own.

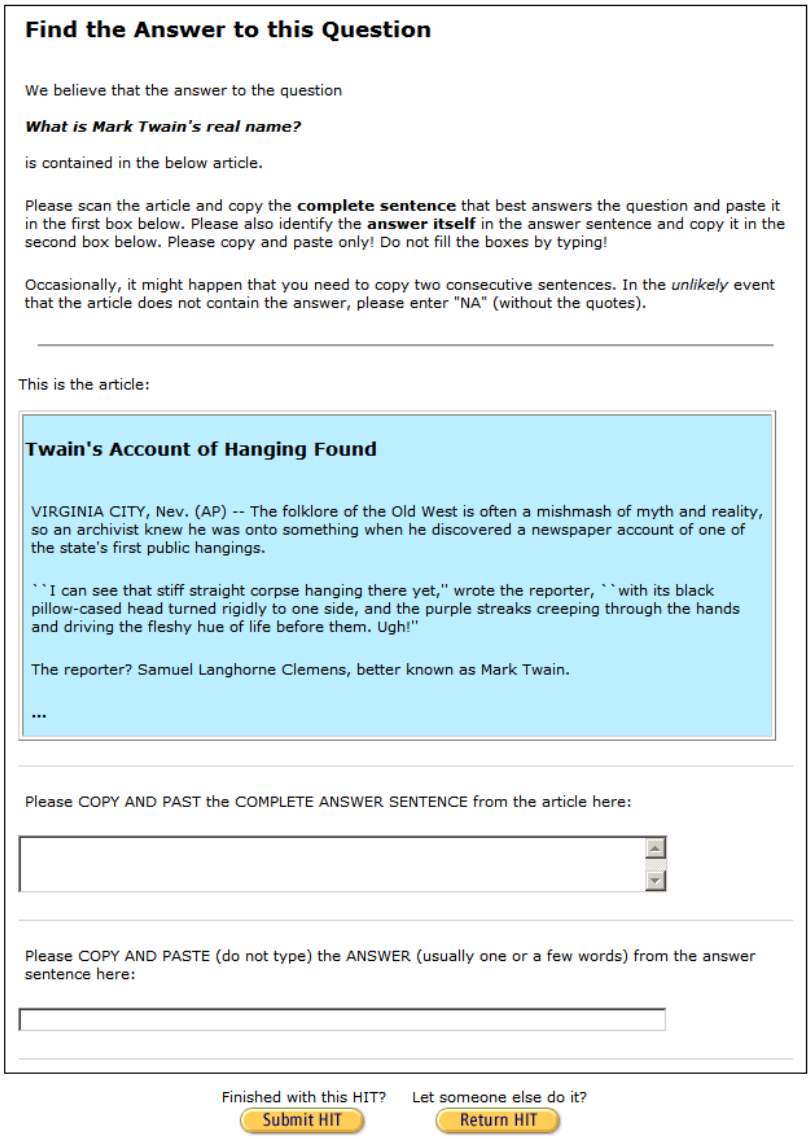

Figure 1: Example HIT, as shown to the subjects. (For this screenshot the text of the article was shorted from the original.)

2. Use of Mechanical Turk

We employed Amazon's Mechanical Turk (AMT)[2] to locate the answer sentence for a TREC question in each of the documents that NIST judged as relevant. Amazon promotes this web-service as "artificial artificial intelligence" and it is used in a wide variety of applications, from mapping utility poles to searching for missing persons. Subjects (called "turkers" in the lingo) are invited for a small reward to work on "Human Intelligence tasks" (HITs), generated from an XML description of the task created by the investigator or sponsor (called a "requester"). HITs can display a wide range of content (e.g. text and graphics) and provide many different input options, e.g. radio buttons, check boxes or input fields for free text. In our case, turkers were used to look at a question and then select the sentence from a given AQUAINT corpus document that best answered it. A screenshot of one of our HITs can be seen in Figure 1. Every HIT was completed by three different turkers before it was removed from the HIT-list. This enabled us to check inter-annotator agreement and thus have a measure for the plausibility of every collected answer sentence individually, as well as to evaluate the reliability of the complete collection. The actual execution of the AMT experiment cost USD 655.31 (including 10% fees paid to Amazon; turkers received USD 0.02 for each completed HIT).

Table 1 shows inter-annotator agreement when computing the similarity of responses by using strict string equality. One problem we encountered was that different browsers and/or operating systems use different copy-and-paste implementations. So even if two users intend to select exactly the same sentence, some implementations automatically include the closing punctuation mark while others do not. The same holds for opening/closing quotes and brackets. Table 2 shows inter-annotator agreement when allowing an Levenstein edit distance of 5, which treats sentences with minor differences as similar.

We consider the inter-annotator agreement reported in Table 2 as satisfactory. Reasons why agreement is not better

[2][http://www.mechanicalturk.com]

Tag	Question	Answer Sentence
A	When did the shootings at Columbine happen?	The Columbine High School shootings April 20 also had an effect on ...
C	What is the capital of Kentucky?	The capital, Frankfort, is about 15 miles down river.
D	When was the internal combustion engine invented?	The first internal-combustion engine was built in 1867, but ...
1	How tall is Mount McKinley?	Together, they climbed Mount McKinley (20,320 feet), the highest peak in the United States.

Table 3: Examples to illustrate the tags used in the corpus: The first sentence gives only an inexact answer ("April 20" instead of "April 20, 1999"). The second sentence gives the correct answer, but does not mention "Kentucky". Most likely Kentucky is mentioned in a preceding sentence. Whether the third sentence answer the question is somewhat doubtful. The final sentence clearly answers the question.

Three	3577	44.1%
Two	3248	40.1%
None	1282	15.8%

Table 1: Inter-annotator agreement for the 8107 TREC 2002-2006 QAS-pairs when using strict string equality. The table shows how often all three turkers selected the same sentence (and the same answer), how often two turkers made the same selection, and how often none of the turkers agreed.

Three	4345	53.6%
Two	2907	35.9%
None	855	10.5%

Table 2: Inter-annotator agreement for the 8107 TREC 2002-2006 QAS-pairs when allowing a Levenstein edit distance of 5.

are, for example:

1. Turkers selected different sentences from a document which indeed includes more than one sentence that answers the question
2. Sometimes it is not obvious for turkers to decide where the selection boundaries should be.
3. Some selection made were suboptimal or simply wrong.

The second point can be illustrated with the example shown in Figure 1. We see in our data that, for the given text, two turkers selected the passage "The reporter? Samuel Langhorne Clemens, better known as Mark Twain." while one selected the shorter "Samuel Langhorne Clemens, better known as Mark Twain."

3. Post Processing the Data

As noted before in the literature, the task to build a high quality research collection for QA, might it contain documents, answer sentences or answers, is not always straightforward (Voorhees and Tice, 2000; Lin and Katz, 2005). The most important issue here, beside the quantity of data involved, is that human judges tend to disagree about what constitutes a valid answer, answer sentence or supporting document.

In order to increase the quality of our data, we decided to let a second set of subjects check the results of the turkers. For this second round we did not employ AMT, instead the subjects consisted of PhD students at the University of Edinburgh's School of Informatics. As a starting point the students received a file with all the judgments from round one, which included all sentences selected by the turkers. Each sentence was tagged to indicate how many turkers (one, two or three) had selected it. By default sentences which were tagged as *two* or *three* received an additional tag indicating that the sentence should become part of the final collection, whereas sentences selected by only one turker did not have this tag. The students task then was to look at all sentences and add or remove the tag indicating that sentence should belong to the final selection if they thought that the turkers had made a mistake. In this final phase, only one student looked at each sentence to make the final decision.

We used this opportunity to additionally ask the student to tag certain special cases. The following tags are included in the final version of the data set:

A indicates that the sentence does answer the question, but that the **a**nswer is inexact.

C indicates that the sentence does answer the question, but that some important information is missing in the sentence. This information can most likely be found in the remainder of the document. (C stands for ***C****ontext missing*)

D indicates that it is **d**oubtful whether the sentence answers the question.

1 indicates that the sentence indeed does answer the question.

Each sentence might be marked with more than one tag. Table 3 lists one example Question Answer Sentence Pair for each tag.

4. Results

Table 4 presents a numeric overview over the original data sets and of the data in our corpus. The first column shows

year	No. factoid questions (original)	No. supporting documents identified	No. factoid questions remaining	No. question-answer sentence pairs	mean no. pairs per question
2002	500	2,177	429	2,006	4.67
2003	413	1,764	354	1,448	4.09
2004	231	919	204	865	4.24
2005	363	1,599	319	1,456	4.56
2006	404	1,648	352	1,405	3.99
2002-2006	1,911	8,107	1,658	7,180	4.33
2002 (Lin)	109	1,822	97	1,650	17.0

Table 4: Quantitative overview of the data collected. The first column shows the origin of the data, usually the year TREC released the data set. The next column shows the number of question in the original data set. Column three gives the numbers of supporting documents identified by TREC. Column four lists the number of questions for which we were able to find at least one answer sentence. The fifths column shows how many sentences we could identify. The last column gives the average number of answer sentences found for each question.

the origin of the data, usually the year in which TREC released the data set. The next column shows the number of questions in the original data set. Column three gives the numbers of supporting documents identified by TREC. Column four lists the number of questions for which we were able to find at least one answer sentence. This number is lower than the number of questions in the original data set for three reasons: a) There are NIL questions in the question set, i.e. questions that do not have an answer in the document collection. b) For some non-NIL questions, TREC participants were unable to find the answer in the collection, although it exists. c) Our subjects were unable find a valid answer sentence in a document, judged as supportive in the original data set. The fifths column in the table shows how many sentences we could identify. There are three reasons why the number of sentences collected is lower than the number of document-ids in the original data set: a) The document itself might contain the answer, but no single text passage can be identified that answers the question. In such cases evidence from multiple passages would be needed to answer the question. b) Our subjects did not agree with TREC's judgment and decided that there is no answer in the document. c) There is a valid answer sentence in the document, but our subjects were not able to locate it. Finally, column six gives the average number of answer sentences we were able to identify for each question (i.e., column 4 divided by row 5).

5. Data Format

Our dataset comes in six files. Five files contain data based on TREC judgment files from 2002 to 2006. A sixth file is based on (Lin and Katz, 2005). Each line in the files shows the data for one Question Answer Sentence Pair. The data in each line is comma separated. There are six rows in each line:

1. The TREC question id.
2. The AQUAINT document id.
3. The question itself (in quotes).[3]
4. The answer sentence (in quotes).
5. The answer (in quotes).
6. A tag (e.g. `1`) or possibly a list of tags, separated by semicolons (e.g. `A;C`).

The answer given in row five is always a substring of the answer sentence in row four. Note that the data in rows three, four and five may contain commas itself.
Figure 2 illustrates the data in our corpus. (Line breaks were added for better readability.)

6. Conclusions

We described a corpus of *question, answer-id* pairs, which is based on TREC's QA track data. We believe that by making it available to the public, we can facilitate further research in the field. We furthermore described, how we have created this resource by using Amazon's Mechanical Turk. We think that, while certainly time and energy were required on the part of the researchers, there is an obvious attraction to using AMT for such experiments as it provides a large, inexpensive, motivated, and immediately available pool of subjects. Although we decided to have a second set of subjects check the data, we would not have been able to perform the complete experiment without utilizing Mechanical Turk.
The data itself can be found on the first author's home page, http://homepages.inf.ed.ac.uk/s0570760/. It is also linked to from TREC's page about *Additional QA Resources*, http://trec.nist.gov/data/qa/add_qaresources.html.

7. Acknowledgments

This work was supported by Powerset, Inc. and Microsoft Research through the MSR European PhD Scholarship Programme.

8. References

Abdessamad Echihabi, Ulf Hermjakob, Eduard Hovy, Daniel Marcu, Eric Melz, and Deepak Ravichandran. 2004. Multiple-Engine Question Answering in

[3]Here the data is slightly redundant, the question could of course be looked up in TREC's original question file, but we felt that including it increases human readability.

```
1395, NYT19990719.0343, "Who is Tom Cruise married to?",
"The movie is not a turn-on (it is really a horror film without gore),
and the sexual chemistry between its married stars, Tom Cruise and Ms.
Kidman, is tepid at best.", "Ms. Kidman", 1

1395, NYT19990706.0080, "Who is Tom Cruise married to?",
"Its allegedly sexually explicit nature and the star power of
real-life married couple Tom Cruise and Nicole Kidman have set the
rumor mills rumbling and led to widespread curiosity - and, in some
cases, apprehension.", "Nicole Kidman", 1

1395, NYT19990326.0303, "Who is Tom Cruise married to?",
"The drama is said to be about a pair of married psychiatrists (played
by the married Tom Cruise and Nicole Kidman) and their sexual lives,
but only a few Warner executives, Cruise and Kidman, and Pat Kingsley,
a top public relations executive, have seen the film.",
"Nicole Kidman", 1

1395, APW19990612.0066, "Who is Tom Cruise married to?",
"For example, the late Stanley Kubrick's secrecy-shrouded ''Eyes Wide
Shut'' is trading highly at H$68, but investors could take a bath if
the on-screen chemistry between married co-stars Nicole Kidman and Tom
Cruise fizzles.", "Nicole Kidman", 1

1395, APW19990712.0006, "Who is Tom Cruise married to?",
"A judge has given the green light to a lawsuit brought against the
Star newspaper after it reported Tom Cruise and Nicole Kidman had help
from sex therapists in filming ''Eyes Wide Shut.'' The actors, who are
married, are seeking unspecified damages.", "Nicole Kidman", 1

1395, APW19981029.0541, "Who is Tom Cruise married to?",
"Actor Tom Cruise and his wife Nicole Kidman accepted ''substantial''
libel damages on Thursday from a British newspaper that reported he
was gay and that their marriage was a sham to cover it up.",
"Nicole Kidman",1

1395, NYT19991101.0416, "Who is Tom Cruise married to?",
"One of the world's most lusted-after men, Tom Cruise, is married to
Nicole Kidman and her curls.", "Nicole Kidman", 1

1395, NYT19990706.0083, "Who is Tom Cruise married to?",
"Its allegedly sexually explicit nature and the star power of
real-life married couple Tom Cruise and Nicole Kidman have set the
rumor mills rumbling and led to widespread curiosity - and, in some
cases, apprehension.", "Nicole Kidman", 1
```

Figure 2: Eight Question-Answer Sentence Pairs, as contained in the corpus. (Line breaks were added for better readability.)

TextMap. In *The Proceedings of the 2003 Edition of the Text REtrieval Conference, TREC 2003*.

David Graff. 2002. The AQUAINT Corpus of English News Text.

Jimmy Lin and Boris Katz. 2005. Building a Reusable Test Collection for Question Answering. *Journal of the American Society for Information Science and Technology*.

Christof Monz. 2004. Minimal Span Weighting Retrieval for Question Answering. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering*.

Ellen M. Voorhees and Hoa Trang Dang. 2006. Overview of the TREC 2005 Question Answering Track. In *The Proceedings of the 2005 Edition of the Text REtrieval Conference, TREC 2005*.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a Question Answering Test Collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.