Intra-Assessor Consistency in Question Answering

Ian Ruthven[1], Leif Azzopardi[2], Mark Baillie[1], Ralf Bierig[1], Emma Nicol[1], Simon Sweeney[1], Murat Yakici[1]

[1]Department of Computer and Information Sciences
University of Strathclyde
United Kingdom G1 1XH

[2]Department of Computing Science
University of Glasgow
United Kingdom G12 8QQ

ir, mb, ralf, emma, simon@cis.strath.ac.uk, leif@dcs.gla.ac.uk

ABSTRACT
In this paper we investigate the consistency of answer assessment in a complex question answering task examining features of assessor consistency, types of answers and question type.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval

General Terms
Experimentation, Human Factors

Keywords
Relevance, question answering, evaluation

1. INTRODUCTION
Question answering systems return textual fragments as answers to submitted questions. In 2006, the *ciqa* track of TREC, an optional sub-task of the main Question Answering track, investigated complex questions where the complexity arises from the relationships between 2 or more entities. For example, in question 32 - "*What financial relationships exist between drug companies and universities?*" - the relationship of interest is a financial relationship between the two entities *drug companies* and *universities*. These questions are seen as more complex than the simpler factoid type questions previously investigated in question answering [1], partly because the structure is more complicated – by relating concepts or entities – and also because the underlying information need may be more complex comprising of several sub-questions.

The process of judging answers is subject to the same diversity of opinions as judging documents for relevance: different people judging the same answers may have different opinions on the quality or accuracy of the answers [2]. Voorhees & Tice [2] indicate, however, that within the current TREC protocol for assessing answers, the current level of inter-assessor disagreement does not substantially alter the results of comparative evaluations of QA systems.

SIGIR'07, July 23–27, 2007, Ámsterdam, The Netherlands.

In this paper we describe an initial investigation, based on our participation in *ciqa* 2006, to investigate the variation *within* an assessor's own judgments rather than inter-assessor consistency. That is, how often do assessors agree with themselves?

2. *ciqa* AND ASSESSMENT
ciqa 2006 investigated five types of complex question, with each type of question having six examples used in the track. The question types are shown below

- What evidence is there for transport of [goods] from [entity] to [entity]?
- What [relationship]1 exist between [entity] and [entity]?
- What influence/effect do(es) [entity] have on/in [entity]?
- What is the position of [entity] with respect to [issue]?
- Is there evidence to support the involvement of [entity] in [event/entity]?

In *ciqa* each participating group was asked to submit two retrieval runs for each of 30 questions. A run consisted of the top 5 answers to each question ranked in decreasing order of perceived answer quality. Our research interest in *ciqa* was the human assessment of answers, not on developing novel QA systems. Consequently, our answers were obtained from manual searching of the document collection used in *ciqa*. Our 2 submitted runs were identical: two sets of identical answers for each question which were both assessed at the same time by the TREC assessors who posed the original questions.

We did not, at the time of submitting these runs, realize this would mean TREC having to assess our answers twice (and we apologize for this additional effort on their behalf). Nevertheless we can use this double assessment to investigate the consistency of answer assessment and the factors that lead to increased or decreased consistency. The aim is to better understand how to create questions for use in QA evaluations and how to use obtained answers to measure system performance.

In *ciqa* each answer consisted of a text fragment and was assessed according to the presence of *nuggets* – facts or concepts relevant to answering the question [1]. Nuggets can be marked *vital* if they must appear in a good answer or *okay* if they provide useful, if inessential, information [1].

In comparing the answers given to our first and second run we noted several cases where the same answer was rated differently in the two runs. That is, a nugget was identified in a sentence for one run but not in the same sentence submitted as part of the other run. In the remainder of this paper we examine the factors that lead to more or less consistency in identifying nuggets.

3. FACTORS AFFECTING CONSISTENCY

For the 30 topics investigated we submitted a total of 137 answers (for a few questions we found less than 5 answers). For the 1st run 86 nuggets were identified (54 vital, 32 okay), for the 2nd run there were 89 nuggets (56 vital, 33 okay), a performance that appears relatively consistent. However, some 13% of these nuggets, across 12 topics, were judged differently between the two runs – judged to be present in one run but absent from the other. So even though the final nugget count appears consistent, different nuggets were identified in the same sentences in the two runs. In our case the difference in nugget count does not lead to a statistically significant difference between our two runs but the reasons for the lack of consistency could give clues on how to use nuggets to evaluate QA systems. In the following sections we examine some of the possible sources of inconsistency.

3.1 Assessor variation

Firstly we look at the variation in individual assessment behaviour. Following [2] we calculate a measure of consistency based on the *overlap* between the two sets of assessment, i.e. the number of nuggets in the intersection of our two runs divided by the number of nuggets in the union of the runs. We calculate this for the questions assessed by each assessor to get an individual overlap value. These overlap values ranged from 0.95 to 0.61 with a mean assessor overlap of 0.85, suggesting a high, but variable, level of consistency between an assessor's two assessments of the same data. This mean value is higher than most reported consistency levels *between* assessors but the range indicates that we should expect some level of inconsistency within individual assessments.

ciqa allowed research groups to pose questions to the assessors about the assessment process *before* the answers were assessed. We asked a range of questions to the assessors on their existing topical knowledge, confidence of assessing answers etc. but could find no correlation between the factors we investigated and their consistency in assessing answers. However, the overlap values presented above present a wide range of consistency values worthy of future investigation.

3.2 Nugget type

As noted previously nuggets could be classified as either vital or okay. *vital* nuggets contain essential information but a possible source of error is failing to be consistent on the less essential *okay* nuggets. Our 1st run missed 9 vital and 2 okay nuggets and the 2nd run missed 6 vital and 6 okay nuggets. This limited evidence does not suggest any particular trend towards greater consistency in spotting vital nuggets.

In addition there was roughly the same number of vital and okay nuggets, across all submitted runs, in the topics where we found at least one nugget. So it is not the case that vital/okay nuggets are easier to find on average for these topics.

We did note, however, a weak, negative correlation (-0.3657, p=0.001) between the total number of nuggets identified for a topic, across all runs, and the number of errors made. That is, the assessment was less consistent when fewer nuggets were identified in the submitted answers.

3.3 Question type

As Voorhees noted in [3], human variability is not the only source of variability; the questions asked may introduce variability into the evaluation process. In Table 1 we show the level of overlap between the two runs according to the question type. As can be seen for some question types – noticeably *position* questions – there is a lower level of consistency whereas for *transport* questions there is a higher level of consistency. Questions of each type are posed and assessed by different assessors so this difference is not due to the assessors themselves, although there may be some interaction effect between the assessor and question type. Rather, the variable rates of consistency appear to be a factor of the question set or answers produced.

Table 1. Mean overlap per question type

question type	overlap
effect	0.81
evidence	0.86
position	0.54
relationship	0.83
transport	0.97

A manual examination of the nuggets for the less consistent answer sets suggests that the nuggets that were missed, either *vital* or *okay*, represented more abstract information than the nuggets that were identified in both runs.

4. CONCLUSIONS

Our interest in this work was in factors that might cause variation in assessments within, rather than across, assessments of the same answer data. Our data is quite limited in size but our preliminary evidence suggests that type of nugget being identified does not lead to greater/less consistency in nugget detection. Assessor consistency on this evidence seems good but is variable and the factors that might lead to greater variability are worthy of further study. The strongest source of variation is the type of question posed, or perhaps more properly the nature of the answer to these questions. By investigating factors that lead to inconsistency in assessment we can better understand the assessment process and estimate confidence intervals within which to interpret the results of a question answering task.

5. REFERENCES

[1] Lin, J. and Demner-Fushman, D. *Will Pyramids of nuggets fall over?* Proceedings of the HLT/NAACL, 383-390, 2006.

[2] Voorhees, E., M. & Tice, D. M. *Building a question answering test collection*. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, 200-207.

[3] Voorhees, E. M. *Evaluating the evaluation: a case study using the TREC 2002 question answering track*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, 181-188.