
This version is available at https://strathprints.strath.ac.uk/3419/

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Unless otherwise explicitly stated on the manuscript, Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Please check the manuscript for details of any other licences that may have been applied. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (https://strathprints.strath.ac.uk/) and the content of this paper for research or private study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to the Strathprints administrator: strathprints@strath.ac.uk
Intra-Assessor Consistency in Question Answering

Ian Ruthven¹, Leif Azzopardi², Mark Baillie¹, Ralf Bierig¹, Emma Nicol¹, Simon Sweeney¹, Murat Yakici¹

¹Department of Computer and Information Sciences
University of Strathclyde
United Kingdom G1 1XH
ir, mb, ralf, emma, simon@cis.strath.ac.uk, leif@dcs.gla.ac.uk

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

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. 

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2007 ACM 978-1-59593-597-7/07/0007...$5.00.
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>
<thead>
<tr>
<th>question type</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>effect</td>
<td>0.81</td>
</tr>
<tr>
<td>evidence</td>
<td>0.86</td>
</tr>
<tr>
<td>position</td>
<td>0.54</td>
</tr>
<tr>
<td>relationship</td>
<td>0.83</td>
</tr>
<tr>
<td>transport</td>
<td>0.97</td>
</tr>
</tbody>
</table>

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