Collecting Reliable Human Judgements on Machine-Generated Language: The Case of the QG-STEC Data *

Keith Godwin[†] and **Paul Piwek**[‡]
The Open University, UK

Abstract

Question generation (QG) is the problem of automatically generating questions from inputs such as declarative sentences. The Shared Evaluation Task Challenge (QG-STEC) Task B that took place in 2010 evaluated several state-of-the-art QG systems. However, analysis of the evaluation results was affected by low inter-rater reliability. We adapted Nonaka & Takeuchi's knowledge creation cycle to the task of improving the evaluation annotation guidelines with a preliminary test showing clearly improved inter-rater reliability.

1 Introduction

Since 2008, researchers from Discourse Analysis, Dialogue Modelling, Formal Semantics, Intelligent Tutoring Systems, NLG, NLU and Psycholinguistics have met at a series of QG workshops (Piwek and Boyer, 2012). These workshops bring together different researchers working on QG activities and collectively are of great value to the QG community.

One such activity was the Shared Task Evaluation Challenge Task B that took place in 2010 (Rus et al., 2012). The challenge was to generate specific questions from single sentences. These questions were evaluated independently by human judges. The average scores of the annotations were used to rank participating QG-STEC systems on these criteria. Of particular interest were the criteria relating to relevance of the generated questions and their grammaticality and fluency. Ideally, when a system generates a question from a sentence, the question should be about the information in that sentence (i.e., be relevant) and it should be fluent and grammatical. Our assumption is that ordinary speakers of English are reasonably in agreement with each other when they make such judgements.

However, in practice, we found low inter-rater reliability (IRR) for the task results. We established this using Krippendorff's α, see Table 6. For four evaluation criteria, α was well below 0.4, with only one criterion achieving an α of 0.409. This does not meet Krippendorff's requirement of an α of at least 0.8, if one wants to draw any conclusions from the results. Nor does it meet the requirement that tentative conclusions are only permitted for $0.67 < \alpha < 0.8$.

It is common practice when evaluating statistical NLP to create an annotation manual. The manual must systematise the annotation process, making it as unambiguous as possible. It should contain a scheme and a set of guidelines. The scheme represents the theoretical backbone of the evaluation process. The guidelines that supplement the scheme provide additional information, often with examples, making clear the scheme usage (Palmer and Xue, 2010). In the original evaluation, the guidelines were minimal.

As the QG-STEC IRR reliability scores show, it seems that judges interpret an annotation scheme for these criteria very differently, when they use the scheme independently, with minimal guidelines.

*We would like to thank Alistair Willis and Brian Plüss for helpful feedback on the work reported in this paper.

[†] keith.godwin@open.ac.uk
[‡] paul.piwek@open.ac.uk

Proceedings of The 9th International Natural Language Generation conference, pages 212–216,
Edinburgh, UK, September 5-8 2016.

Rank	Description
1	The question is completely relevant to the input sentence.
2	The question relates mostly to the input sentence.
3	The question is only slightly related to the input sentence.
4	The question is totally unrelated to the input sentence.

Table 1: Relevance. Questions should be relevant to the input sentence. This criterion measures how well the question can be answered based on what the input sentence says.

Rank	Description
1	The question is grammatically correct and idiomatic/natural.
2	The question is grammatically correct but does not read as fluently as we would like.
3	There are some grammatical errors in the question.
4	The question is grammatically unacceptable.

Table 2: Syntactic correctness and fluency. The syntactic correctness is rated to ensure systems can generate sensible output. In addition, those questions which read fluently are ranked higher.

Rank	Description
1	The question is unambiguous.
2	The question could provide more information.
3	The question is clearly ambiguous when asked out of the blue.

Table 3: Ambiguity. The question should make sense when asked more or less out of the blue. Typically, an unambiguous question will have one very clear answer.

Rank	Description
1	The question is of the target question type.
2	The type of the generated question and the target question type are different.

Table 4: Question Type. Questions should be of the specified target question type. E.g. who, what, where, when etc..

Rank	Description
1	The two questions are different in content.
2	Both ask the same question, but there are grammatical and/or lexical differences.
3	The two questions are identical.

Table 5: Variety. Pairs of questions in answer to a single input are evaluated on how different they are from each other. This rewards those systems which are capable of generating a range of different questions for the same input.

Typically when the IRR is low this can be attributed to the complexity of the phenomena being annotated. Capturing complex phenomena requires complex theory which in turn requires complex instructions (Hovy and Lavid, 2010). Either the scheme does not accurately represent the theory behind identifying the phenomena, or the guidelines to the scheme were insufficient to explain it to the breadth of audience using the scheme, or the the annotators did not receive appropriate training. For this research we assumed the scheme was sound and our goal was to improve the guidelines without modifying the scheme. Training length and intensity would be addressed once we had an appropriate set of guidelines.

The scheme criteria used by evaluators in the QG-STEC are described in Tables 1-5. The criteria defined by these tables were applied to each of the generated questions independently during evaluation. The ranges of Rank vary, but 1 is always the highest score.

As a first step towards remedying guidelines, we used a set of judges to iteratively and collaboratively train using the guidelines accompanying the scheme, until we were satisfied that they had reached a common understanding of the scheme. This allowed us to 'debug' the guidelines whilst the judges produced improved guidelines (see Section 2).

Our next step would be to use the scheme with the revised guidelines and a new set of judges to annotate the QG-STEC data. This would allow us to find out whether the new guidelines facilitate IRR. However, this is work in progress and in advance of that, we decided to find out a possible upper-bound on IRR that could be achieved with these new guidelines. To do so, we got our current judges to independently annotate the QG-STEC data. The results, see Table 6, are very encouraging.

2 Annotation Method

The problem we identified in Section 1 is that if the judges disagree significantly (and thus have internalised their own version of the annotation scheme, which isn't documented, and therefore isn't repeatable or open for critical analysis) then the analysis will suffer. We defined a significant difference as a disagreement greater than one rank, therefore we

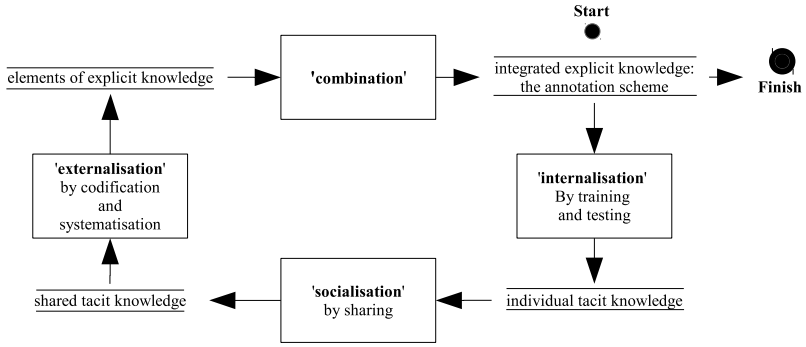

Figure 1: The knowledge creation cycle of collaborative training.

kept training until the judges mostly agreed to within one rank.

This process is shown in Figure 1 where we describe it using a modified version of the Knowledge Creation Cycle of Nonaka and Takeuchi (1995). The main difference being the much shorter time between iterations of the cycle in our method. The training began at the start position with the existing annotation scheme and minimal guidelines. This was the initial version of the integrated explicit knowledge that existed at the start of training. The four stages of the cycle are detailed as follows:

i) INTERNALISATION: The judges read through the annotation scheme and guidelines. Each judge was given a training set of nine input sentences with a series of generated questions (approximately 40) to annotate, simulating the evaluation activity. The input sentences used for training were disjoint from the QG-STEC data, but similar in nature: selected at random from The Guardian Newspaper in an attempt to interest the annotators and keep them motivated. The generated questions were mostly created using the question generator developed by Heilman (2011), to provide realistic examples. For each iteration through this stage a new training set was provided. Upon completion each judge would have internalised the annotation scheme and guidelines to the best of their ability and would have developed additional tacit knowledge based on their experience with the simulated evaluation process. The results were compared and any differences greater than one rank apart were marked for discussion during the Socialisation stage.

ii) SOCIALISATION: Motivated by the marked results above, the judges discussed how they reached their individual evaluation, sharing and discussing their tacit knowledge.

iii) EXTERNALISATION: The judges were encouraged to think about a way to generalise describing this process by codification and systematisation. When the judges reached a consensus, they moved onto the next stage.

iv) COMBINATION: The annotation guidelines were updated to reflect the changes developed in this iteration of the training cycle, ready for the next iteration. This cycle repeated until a sufficient degree of agreement was reached, as described above.

The actual training activity consisted of three iterations. The first iteration, which had 48 significant differences (evaluations different by more than one rank), was dominated by a discussion on the administration of the evaluation. Changes to the guidelines included correcting simple mistakes such as inappropriate wording in the guidelines or getting the rank order the wrong way round. E.g. general advise: 'Each criteria, defined below, is assigned a rank, with 1 being the greatest.'

The second iteration had 17 significant differences. The judges began to identify a number of key conceptual questions which should be answered during the process of making an evaluation. E.g. for ambiguity: 'One consideration when assessing this criterion is to ask the following question: Can more information be added from the input sentence to make the question more specific?'

The last iteration had three significant differences. At this point the training was deemed complete and our criterion for internalising the scheme had been

Criteria	QG-STEC	QG-STEC+
Relevance	0.25	0.806
Question Type	0.323	0.859
Correctness	0.409	0.838
Ambiguity	0.334	0.688
Variety	0.348	0.904

Table 6: Krippendorff's alpha IRR measure for original and re-evaluated data.

met. The judges were now having discussions that were constructed using the language and evaluation skill that had been collaboratively produced and recorded in the evaluation guidelines document.

3 Results

Table 6 compares the current results QG-STEC+[1] and those of the original QG-STEC. The IRR results of the QG-STEC are mostly rated Fair, using the Koch and Landis Scale. By contrast QG-STEC+ data are mostly rated Perfect.

4 Conclusion and Further Work

The purpose of the QG-STEC was to measure the quality of the automatically generated questions. We think of this quality in terms of the judgements of ordinary speakers of English. There isn't necessarily a gold standard: if most speakers of English deem a question fluent and relevant, the system has achieved its goal – even if an expert judges it to be flawed relative to some gold standard. For this reason, our main concern regarding the annotation scheme is reproducibility rather than accuracy. Following Artstein and Poesio (2008) we consider reproducibility 'the degree to which different coders achieve the same coding when working independently.'

If a question is given a particular rating by our judges, this should predict reliably how a new independent judge is going rate the question. Our current study has only revealed the upper-bound achievable, when using the judges that arrived at the revised guidelines. Future studies will need to prove the efficacy of these revised guidelines.

For now, one further check that can give us some confidence in the preliminary results, is to look at the distribution of judgements by our judges. See Figures 2 and 3. This allows us to rule out certain types of bias (e.g., the judges always agreeing to rate at a certain point on the scale).

[1] https://github.com/Keith-Godwin/QG-STEC-plus

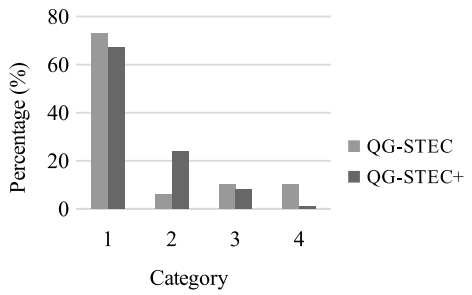

Figure 2: Distribution across categories for relevance

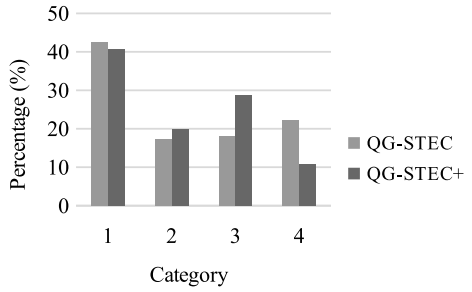

Figure 3: Distribution across categories for correctness

References

Ron Artstein and Massimo Poesio. 2008. Survey Article Inter-Coder Agreement for Computational Linguistics. *Association for Computational Linguistics*, 34(4):555 – 596.

Michael Heilman. 2011. *Automatic factual question generation from text.* Ph.D. thesis, Carnegie Mellon University.

Eduard Hovy and Julia Lavid. 2010. Towards a "Science" of Corpus Annotation : A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, 22(1):1–25.

Ikujir Nonaka and Hirotaka Takeuchi. 1995. *The knowledge-creating company*. Oxford University Press.

Martha Palmer and Nianwen Xue. 2010. Linguistic Annotation. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The handbook of computational linguistics and natural language processing*, chapter 10, pages 238–270. John Wiley & Sons.

Paul Piwek and KE Boyer. 2012. Varieties of question generation: introduction to this special issue. *Dialogue & Discourse*, 3(2):1–9.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2012. A detailed account of the First Question Generation Shared Task Evaluation challenge. *Dialogue & Discourse*, 3(2):177–204.