Identifying Annotator Bias: A new IRT-based method for bias identification

Supplementary material

Introduction

This file serves as the supplementary material to the paper *Identifying Annotator Bias: A new IRT-based method for bias identification*.

Here, we complete the example that we adopted to present our IRT-based method for identifying annotator bias. The example is based on the QG-STEC evaluation dataset, specifically, the annotation performed by Judge 1 and Judge 3. The evaluation was carried out based on four criteria: ambiguity, variety, fluency and relevance.[1] Because of space constraints, in the paper we have only presented the analyses for the ambiguity and variety criteria. In this file we present the analysis for the the fluency and relevance criteria.

All the information about the statistical software used to carry out the analysis here presented, can be found in the Section 2.3 of the paper.

The Fluency criterion

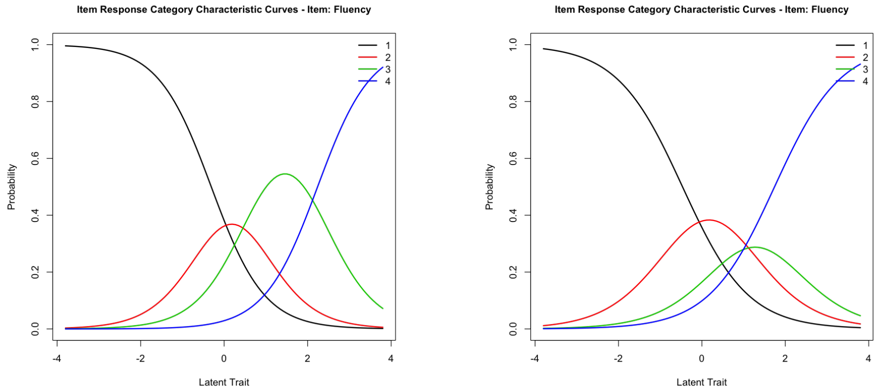
For the fluency criterion the judges reach a Conger's κ value of 0.51.

The IRCCC graphs analysis

In the case of the fluency criterion, the frequency analyses[2] suggest that Judge 1 prefers to use scores 1 and 3, whereas Judge 3 prefers to use scores 2 and 4. This is illustrated in Figure 1, which clearly shows the phenomenon in the latent trait interval that is approximately between -2 and 2. More specifically, for Judge 1 we can see that:

[1] For more details about the dataset and the evaluation, we refer to Section 2 of the paper.

[2] For more details we refer to Section 2.1 of the paper.



(a)

(b)

Figure 1: IRCCC for the fluency criterion for Judge 1 (a) and Judge 3 (b). The graphs show that the main sources of disagreement can be found in the latent trait interval [(±)1, (±)2.2] mainly relative to score 3 (green line) and score 4 (blue line).

- The probability of selecting the score 3 (green line) is higher than selecting the score 4 (blue line).
- At the same time, the probability of selecting the score 2 (red line) is slightly higher than selecting the score 3 (green line) in the latent trait interval that is approximately between -2 and 0.
- On the other hand, from level 0 to level 2 of the latent trait, the probability of selecting the score 3 (green line) is higher than selecting the score 2 (red line).

Conversely, for Judge 3:

- the probability of selecting the score 3 (green line) is slightly higher than selecting the score 4 (blue line) up to the latent trait level of approximately 1.
- For values of the latent trait level higher than 1, the probability of selecting score 4 increases dramatically.
- At the same time, the probability of selecting the score 2 (red line) is higher than selecting the scores 3 (green line) and 4 (blue line) in the latent trait interval that is approximately between -3 and 1.

The extremity parameter analysis

The extremity parameters for the fluency criterion show that:

- Judge 1 has a 50% chance of selecting score 1 with a latent trait level of -0.309, whereas Judge 3 has a 50% chance of selecting score 1 with a latent trait level of -0.464.

- At the same time, Judge 1 has a 50% chance of selecting the scores 1 and 2 with a latent trait level of 0.677 whereas Judge 3 has a 50% chance of selecting the scores 1 and 2 with a latent trait level 0.809. In this part of the latent trait, the judges show a similar annotation trend.

- A very interesting divergence can be found for the high scores. Here, Judge 1 has a 50% chance of selecting the scores 1, 2 and 3 with a latent trait level of 2.238 whereas Judge 3 has a 50% chance of selecting the scores 1, 2 and 3 with a latent trait level of 1.742.

The analysis of the extremity parameters suggests that the main divergence between Judge 1 and Judge 3 takes place in the latent trait that is in the interval [1.742; 2.238] (let's denote it as I_f^1). Based on the evaluation guidelines provided for the QG-STEC task B for the fluency criterion, and under the assumption that one question can be more fluent than another one, we can draw the following conclusions:

The questions that fall in I_f^1 can be interpreted as questions that present clear grammatical errors, and so for which the fluency is problematic. A couple of examples from the dataset are:

> “Was the British information on Dean Thomas was left in the US version?” and “To what is dating of prehistoric materials particularly crucial? ”.

In I_f^1, Judge 3 shows more strict annotation behavior than Judge 1, tending to give the score 4 whereas Judge 1 scores 3 or 2.

The same strict behavior of Judge 3 can be detected in the latent trait interval [-0.464, -0.309] (let's denote this as I_f^2). The range I_f^2 contains questions which are slightly lacking fluency, or questions with minor grammatical errors. A couple of examples from the dataset are:

> “The son purchased which company?” and “What apply to a range of social care professionals?”.

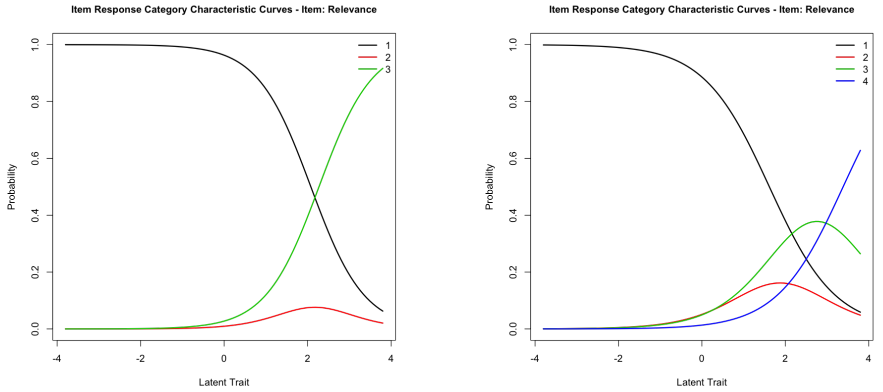
In I_f^2, Judge 3 tends to give scores higher than 1, whereas Judge 1 tends to give the score exactly 1.

The relevance criterion

The relevance criterion reaches a low Conger's κ value of 0.17. We can see from the actual scores frequency (provided in Section 2.1 of the paper) that the frequency of score 1 is much higher than that of the other scores. As explained in Artstein and Poesio (2008) and Gwet (2014), this gives rise to the *prevalence paradox*: a high degree of observed agreement is associated with a low agreement coefficient. Artstein and Poesio (2008) suggest that to address the prevalence paradox (if necessary) it may be best to also report the observed agreement.[3] In this case the observed agreement is 80%.

The IRCCC graphs analysis

The scores frequency analyses show for relevance that the Judges 1 and 3 tend to make most use of score 1.



(a)

(b)

Figure 2: IRCCC for the relevance criterion for Judge 1 (a) and Judge 3 (b). The graphs show as the small source of disagreement resides in the extreme part of the latent trait.

A first cursory look at the IRCCCs for relevance in Figure 2 suggest a substantial difference between the graph for Judge 1 (a) and Judge 3 (b). However, closer inspection suggests that the differences may not be that substantial: The curve for score 1 (black line) and score 2 (red line) are quite similar — the differences are:

- Judge 1 (Figure 2(a)) does not use the category 4.
- Graph 2(b) shows a slightly higher peak for score 2 (showing that Judge 3 uses score 2 slightly more than Judge 1).
- The curve for score 1, goes down more slowly for Judge 1 than for Judge 3 (showing that Judge 1 uses score 1 slightly more than Judge 3).

[3] It is important to remember that, although the observed agreement provides information about the Judges' agreement, it shouldn't be used as a measure of reliability, see for example Krippendorff (1980) and Artstein and Poesio (2008).

- The curve for the score 3 (green line) is quite similar for both graphs till the latent trait level reaches about 2. At this point the curve increases for Judge 1 — see Graph 2(a) — and decreases for Judge 3 — see Graph 2(b). For Judge 3, at that level of the latent trait, the score 3 (green line) gives way to the score 4 (blue line). In contrast, Judge 1 never selects score 4 for higher levels of the latent trait, preferring score 3 instead.

The small differences between the graphs, particularly for scores 1 and 2, explain the high level of observed agreement. The marked difference between use of scores 3 and 4 explain the low κ value.

The extremity parameter analysis

The extremity parameters for the relevance criterion show as the divergence in annotation between Judges 1 and 3 that can be found at high levels of the latent trait:

- Judge 1 has a 50% chance of selecting the score 1 with a latent trait level of 2.076, whereas Judge 3 has a 50% chance of selecting the score 1 with a latent trait level of 1.618.
- Judge 3 has a 50% chance of selecting the scores 1, 2 or 3 with a latent trait level of 3.387, whereas for Judge 1 such a level is not defined because they never use score 4.
- The Judges behave similarly regarding the 50% chance of selecting the scores 1 and 2. In this case, for Judge 1 the latent trait level is 2.271 and for Judge 3 the latent trait level is 2.132.

The extremity parameters suggest that the main difference between Judges 1 and 3 resides in the latent trait interval [1.618, 2.076] (let's denote it as I_r). Based on the evaluation guidelines provided for the QG-STEC task B for the relevance criterion, and under the assumption that, given an input text T, one question can be more relevant to T than another one, we can draw the following conclusions:

The questions that fall in I_r can be interpreted as questions that are somewhat related to the input sentence. A couple of examples from the dataset are:

> *Input sentence A*: Women tend to cover shorter urban journeys and therefore their driving is slower and accidents tend to be relatively minor.
>
> *question A*: What cover shorter urban journeys ?
>
> *Input sentence B*: In Philosopher's/Sorcerers Stone, information on Dean Thomas was left in the US version, but not the British,
>
> *question B*: who was left in the US version, but not the British?

In I_r, Judge 1 shows a more lenient annotation behaviour than Judge 3. Indeed, we can expect Judge 3 to give higher scores then Judge 1 for these and similar items.

Because Judge 1 does not use the score 4, for latent trait levels higher than 3.387, the model predicts a persistent difference in annotation behavior between the two Judges: Where Judge 1 uses score 3, Judge 3 uses score 4.[4]

References

R. Artstein and M. Poesio, *Inter-Coder Agreement for Computational Linguistics*, Computational Linguistics, 2008, Volume 34, Issue 4, pp 555-596.

K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*, Advanced Analytics, LLC, 2014.

K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, Sage Publications, Beverly Hills, CA, 1980.

[4] We have to keep in mind that IRT provides a probabilistic analysis based on a set of effective annotations. Indeed, the analysis we presented allows for an analysis of Judges' bias based on the actual annotations.