Agreement is overrated: A plea for correlation to assess human evaluation reliability

Jacopo Amidei and **Paul Piwek** and **Alistair Willis**
School of Computing and Communications
The Open University
Milton Keynes, UK
{jacopo.amidei, paul.piwek, alistair.willis}@open.ac.uk

Abstract

Inter-Annotator Agreement (IAA) is used as a means of assessing the quality of NLG evaluation data, in particular, its reliability. According to existing scales of IAA interpretation – see, for example, Lommel et al. (2014), Liu et al. (2016), Sedoc et al. (2018) and Amidei et al. (2018a) – most data collected for NLG evaluation fail the reliability test. We confirmed this trend by analysing papers published over the last 10 years in NLG-specific conferences (in total 135 papers that included some sort of human evaluation study). Following Sampson and Babarczy (2008), Lommel et al. (2014), Joshi et al. (2016) and Amidei et al. (2018b), such phenomena can be explained in terms of irreducible human language variability. Using three case studies, we show the limits of considering IAA as the only criterion for checking evaluation reliability. Given human language variability, we propose that for human evaluation of NLG, correlation coefficients and agreement coefficients should be used together to obtain a better assessment of the evaluation data reliability. This is illustrated using the three case studies.

1 Introduction

Data reliability plays a pivotal role in human annotation efforts. Krippendorff (1980) delineates three types of reliability, which are *stability*, *accuracy* and *reproducibility*.

Stability or intra-coder agreement is generally measured by the test-retest strategy, which is based on the resubmission, after some time, of some items to the original annotators. That is, annotators are asked to re-assess the same items after some time has elapsed. Comparing the annotations of the same items provides a measure of the annotator's consistency.

Accuracy is measured calculating the deviations from a given gold standard.

Reproducibility is a measure of the extent to which different annotators arrive at the same annotation when working independently. If different annotators[1], when independently performing the annotation task, consistently make the same annotation decision, then we have strong support for the belief that the phenomena to be annotated are well understood and shared across the annotators. The reproducibility of the annotation is dependent on a well-defined coding scheme and clear annotation guidelines. With these, different annotators can perform the same annotation task reaching equivalent (or very similar) results. As shown in Artstein (2017), Finlayson and Erjavec (2017), Hovy and Lavid (2010) and Pustejovsky and Stubbs (2013), where general rules for annotation design are developed, this idea of reliability as reproducibility has become the predominant reliability concept used in any Computational Linguistics (CL) annotation task. Accordingly, guidelines and good practice descriptions for applying IAA in CL annotation tasks have been developed (for example, Lombard et al. (2002); Artstein and Poesio (2008); LeBreton and Senter (2008); Kottner et al. (2011)). An assumption behind such good practice is the existence of a gold standard, which although true in many annotation tasks, for example speech tagging, may not always be the case. Such an assumption falls short in the case of NLG. Indeed, in the case of NLG, where the existence of a gold standard is mostly not available – for example, criteria such as ambiguity, relevance, usefulness or overall quality – the concept of reliability as reproducibility can hide some pitfalls.

For this paper we analysed papers published in

[1] In Section 4 we will refer to annotators as judges. We chose such terminology to emphasize the evaluation aim of the annotation.

NLG specialist conferences over the last ten years (135 in total) in order to check how IAA is used in the human evaluation of NLG systems. In order to carry out our survey, we selected the papers from the Special Interest Group on Natural Language Generation (SIGGEN) webpage hosted by the ACL Anthology website[2]. We examined the papers with a publication date between the years 2008 to 2018. To select the papers, we decided to use the following criteria: 1) the paper should include a study with human annotators or judges; 2) the study should be an evaluation study (we did not take into account other tasks involving human annotation) 3) the study should allow for measurement of the IAA (for example, we did not take into account papers in which the human evaluation was done with open questions, nor papers whose human evaluation consisted of an author manually inspecting outputs. Likewise, we did not take into account papers that use extrinsic evaluation methodology. However, we did include papers whose extrinsic evaluation methodology was followed by a survey which allows the study of the IAA, for example, surveys done with Likert scale questions.[3]

Our analysis highlights that there is little use of reliability studies in the evaluation phase and a lack of common practice in the use of IAA. More interestingly, we confirm a trend already suggested in Craggs and Wood (2005), Lommel et al. (2014), Liu et al. (2016), Sedoc et al. (2018), Amidei et al. (2018a) and the supplementary material of Reiter (2018): according to existing scales of IAA interpretation – see for example Table 1 and Table 2 – most data collected for NLG evaluation fail the reliability test.

Following Craggs and Wood (2005), Sampson and Babarczy (2008), Lommel et al. (2014), Joshi et al. (2016) and Amidei et al. (2018b) such phenomena can be explained with variability in language interpretation and quality judgement, particularly for semantic or pragmatic language aspects – such as for instance concepts such as text usability, fluency, comprehensibility etc. Human language processing and understanding are fundamental aspects of the human language. Given their subjectivity, they are exposed to high variability. As noted in Craggs and Wood (2005), Sampson and Babarczy (2008), Lommel et al. (2014), Joshi et al. (2016) and Amidei et al. (2018b) annotators diverge in language annotation tasks due to a range of ineliminable factors such as background knowledge, preconceptions about language and general educational level. Such divergence or variability is what makes human language so broad in its use, interpretation and understanding. For this reason, this divergence and variability should not be eliminated from NLG generation tasks. If evaluation results have to inform generation system developers of the extent to which they can improve the communicative power of their systems, levelling the human language interpretation and use divergences is in danger of biasing system developers towards ignoring important aspects of human language. However, the concept of reliability as reproducibility goes in the direction of levelling human languages divergences. This raises the need of a better understanding of reliability of human evaluations.

Given the human language variability that NLG systems have to take into account, we propose the use of correlation coefficients[4] alongside the Kappa statistic[5] in order to obtain a more faithful picture of the evaluation reliability.

2 Related work

The use of IAA in corpus annotation tasks has been widely studied. To our knowledge, less attention has been paid to the use of IAA in human evaluation for NLG systems. Our paper tries to fill this gap.

Regarding the use of IAA in corpus annotation tasks, and more specifically the task of linguistic annotation, we refer to Palmer and Xue (2005) and Pustejovsky and Stubbs (2013). Both provide extensive theoretical descriptions of how to perform

[2] https://aclweb.org/anthology/venues/inlg/

[3] Further information about the paper selection can be found in the supplementary materials, at: https://bit.ly/2lKL516.

[4] Some examples of correlation coefficients are Kendall's τ, Pearson's r, Spearman's ρ and Goodman and Kruskall's Gamma.

[5] In CL, since Carletta (1996)'s paper, the standard measure to calculate human agreement in annotation efforts is some variation of the $kappa$ coefficient of agreement, which Carletta collectively refers to as the name of *Kappa statistic*. Following the notation used in Carletta (1996), the Kappa statistic K can be expressed in the following general formulation: $K = P(A) - P(E)/1 - P(E)$ where $P(A)$ is the proportion of times the annotators agree, whereas $P(E)$ is the proportion of times the annotators would be expected to agree by chance. Some example of the Kappa statistic are Cohen's κ (Cohen, 1960) and Fleiss' κ (Fleiss, 1971). Krippendorff's α coefficient (Krippendorff, 1980) is expressed in a similar way but in terms of disagreement.

an annotation task. For human evaluation of NLG systems, we refer to Krahmer and Theune (2010) and Gatt and Krahmer (2018). Both devote an entire section to evaluation. In particular, Section 7 of Gatt and Krahmer's paper gives a helpful description of the methodologies used in NLG for the purpose of evaluation, alongside examples and a discussion of the relevant problems.

A very helpful survey paper for understanding IAA in CL is presented by Artstein and Poesio (2008), who give a deep analysis of the kappa agreement measures. The authors discuss the mathematics and interpretation of these coefficients and their use in several computational linguistic tasks.

Regarding the basic statistics concepts and statistical analysis we refer to Witte and Witte (2017). More specifically, regarding the use of correlation coefficients in annotation tasks we refer to Stemler and Tsai (2008), LeBreton and Senter (2008) and Gisev et al. (2013). A brief introduction to some of the statistical concepts used in this paper, as well as a complete list of the papers we examined, can be found at `https://bit.ly/2lKL516`.

3 10 years of IAA in evaluation of NLG systems

The main findings of our analysis are: (1) little use of reliability studies in the evaluation phase, (2) shortcomings and oversights in reporting the IAA studies, and consequent lack of a common practice in the use of IAA, (3) generally a low value of IAA.

Point 1: Use of reliability studies

The first thing that stands out in our analysis is the small number of papers which compute IAA in order to validate the evaluation results. Indeed, of the 135 papers in our study, just 18% (24 papers) report information about the IAA. Among these, four papers use two different coefficients to measure the IAA. The other 20 use just one coefficient. In 67% of the papers (16 papers) the IAA was reported in papers published between the 2016 and the 2018. This fact shows an improving trend, in reporting the IAA values, in the area.

Point 1 underlines a shortcoming of NLG human evaluation tasks. When human evaluations are performed, it is good practice to verify the reliability of the evaluations. Without a reliability study there are no solid reasons to accept the conclusions from an evaluation. In Section 4 we suggest that to assess reliability, correlation should play a central role.

Point 2: Reporting IAA studies

With regard to point 2, much has been said in previous works. Because presenting a detailed report of those works is beyond the scope of this paper, we refer to Krippendorff (1980), Lombard et al. (2002), Artstein and Poesio (2008), LeBreton and Senter (2008), Kottner et al. (2011) and Artstein (2017), where guidelines and good practice descriptions for applying IAA have been developed. Based on our research, the following shortcomings have been identified. Papers often:

- do not report the names of the coefficients used;
- do not report sufficient details about the experiments used to collect the data;
- use a coefficient that is not suitable for the data collected;
- do not report the number of items on which the IAA study is performed;
- do not report whether the annotators were performing the evaluation independently or not;
- do not report the scale used to interpret the IAA values, and when reported do not discuss the results accurately.

More specifically, between the papers that report the IAA, 37% of the papers (9 works) use a IAA coefficient that is not suitable for the data collected. For example, the use of Fleiss' κ coefficient for data whose level of measurement is interval. Related to this point, we note that often the researchers do not report in sufficient detail the experiment used to collect the data, which can also give information about the data level of measurement – that is whether the data are nominal, ordinal, intervals or ratios. Across the papers we studied, such information had to be deduced from the statistic used for analysing the data.[6]

[6] We note that this is an imperfect, although sometimes the only possible, way to deduce the data level of measurement. Indeed, researchers can use the wrong statistic to analyse the data, which results in a distorted image of the data level of measurement.

From Table 3 we can see that although discouraged by previous work – see for example Krippendorff (1980), Craggs and Wood (2005) and Artstein and Poesio (2008) – percent agreement is the coefficient used the most. Indeed it is applied in 25% of the works (7 papers). It is followed by Krippendorff's α (Krippendorff, 1980) and Fleiss's κ (Fleiss, 1971). Both coefficients were used in 5 papers each. Three papers do not report the name of the Kappa statistic used. Because each metric is different, reporting the exact coefficient used in the analysis would help the readers to better understand the data reliability and the evaluation results.

Few papers discuss the interpretation of the IAA for their evaluation. Between the papers that report the IAA, just 20% of the papers (5 works) make implicit or explicit reference to the interpretation scales used. The IAA interpretation scales reported by these papers are the Krippendorff scale (Krippendorff, 1980, see Table 1) and the Landis and Koch scale (Landis and Koch, 1977, see Table 2).

IAA value	IAA interpretation
IAA < 0.67	Discard
$0.67 \leq$ IAA < 0.8	Tentative
$0.8 \leq$ IAA ≤ 1	Good

Table 1: Krippendorff (1980).

IAA value	IAA interpretation
IAA < 0	Poor
$0 \leq$ IAA ≤ 0.2	Slight
$0.2 <$ IAA ≤ 0.4	Fair
$0.4 <$ IAA ≤ 0.6	Moderate
$0.6 <$ IAA ≤ 0.8	Substantial
$0.8 <$ IAA ≤ 1	Almost Perfect

Table 2: Landis and Koch (1977).

In almost every paper we analysed, the number of items used for the IAA studies was not reported. Likewise, there were few cases in which it was reported whether or not the annotators worked independently.

Finally, we also note that the terminology used is not shared across the analysed papers. Some examples are: reliability, agreement, inter-evaluator agreement, pair-wise agreement, inter-annotator agreement, inter-assessor agreement, inter-rater reliability and inter-coder agreement.

Point 3: Low IAA values

Table 3 shows a tendency also found in other work; see for example Craggs and Wood (2005), Lommel et al. (2014), Liu et al. (2016), Sedoc et al. (2018) and Amidei et al. (2018a) and the supplementary material of Reiter (2018).[7] The trend is that in human evaluation of NLG systems the IAA values reached are relatively low. Following the Krippendorff scale of IAA interpretation (Krippendorff, 1980) – which considers the threshold 0.67 as the minimum to be reached in order to get a reliable set of data (see Table 1) – the majority of the evaluations should be discarded. The problem of how to interpret IAA values is an intriguing and complicated one. Artstein and Poesio (2008) describe this as "the most serious problem with current practice in reliability testing". As noted by Krippendorff (1980, 2004), Craggs and Wood (2005) and Hovy and Lavid (2010), the choice of IAA interpretation scale is arbitrary and task-dependent. The reduction of a statistical test interpretation to a simple number, whilst common, can be arbitrary and accordingly give us little information[8]. For example, Artstein (2017) show that a single label is not sufficient to give a deep understanding of the reliability of an annotation. In this paper, we do not face the problem of how to interpret IAA, rather, we try to tackle the prob-

Coefficient	# used	Average	Min.	Max.
Percent agreement	7	0.69	0.44	0.94
Cohen's κ	4	0.40	0.10	0.88
Krippendorff's α	5	0.62	0.37	0.90
Fleiss's κ	5	0.53	0.29	0.78
Pearson's r	2	0.42	0.20	0.71
Kendall's W	1	0.61	0.47	0.76
Weighted κ	1	0.07	0.07	0.07
κ no better specified	3	0.57	0.32	0.77

Table 3: Average, minimal and maximum IAA value per coefficient. *# used* means the number of times that a coefficient was used in total across the papers. In each paper each coefficient was used to measure the annotator's agreement about one or more questions or criteria.

[7] We note that for the κ coefficients which are "no better specified" the average measure is not appropriate. Indeed, they could be different κ coefficients. However, we chose to report the average for uniformity reasons. It worth saying that such a choice does not affect the theoretical point here presented.

[8] Lately, this point has been raised also for the $p - value$. See for example the special issue *Statistical Inference in the 21st Century: A World Beyond* $p < 0.05$ (Wasserstein et al., 2019) published by The American Statistician.

lem of data reliability by suggesting that correlation coefficients and agreement coefficients should be used together to obtain a better assessment of the evaluation data reliability.

Point 3 also reveals a big issue in the area. Indeed, the main purpose of IAA is to check the reliability of the annotated data. Following the existing scales of IAA interpretation, for example those of Krippendorff (1980) and Landis and Koch (1977), the majority of the evaluations should be discarded because they are unreliable. However, Sampson and Babarczy (2008), Lommel et al. (2014), Joshi et al. (2016) and Amidei et al. (2018b) suggest that a low level of IAA can be explained with human language variability. Arguably such a property, which must be preserved by NLG systems, makes strict agreement unsuitable for testing the reliability of human evaluations. This raises the problem of how to improve the analysis of reliability of human evaluation datasets. In Section 4 we argue that to assess reliability, correlation coefficients should play a central role.

4 The use of correlation coefficients for NLG human evaluation tasks

Correlation coefficients are generally considered inappropriate for measuring the reliability of annotated data; see for example, Lombard et al. (2002), Krippendorff (2004) and Artstein and Poesio (2008). The main concern about the use of correlation coefficient for reliability studies is well expressed by the following quotation:

> [Correlation coefficients, for example Pearson's r] measure the extent to which two logically separate interval variables, say X and Y, covary in a linear relationship of the form $Y = a + bX$. They indicate the degree to which the values of one variable predict the values of the other. Agreement coefficients, in contrast, must measure the extent to which $Y = X$. (Krippendorff, 1980, p. 244)

Indeed, the rationale behind agreement coefficients, such as the Kappa statistic, is to catch the extent to which judges rank a given item equally. When judges rank a given item in the same way, it is assumed that the judges share the same interpretation and understanding of the schema and guideline used in the annotation task. When this happens, given the fact that the annotation is reached with judges that work independently, the concept of reliability as reproducibility suggests that the same annotation can be reached with other judges. This makes the annotation repeatable and consequently reliable.

Although such a concept of reliability as reproducibility is well-founded in cases where the phenomenon under investigation has some objective meaning, for example in the case of many CL annotation tasks where the gold standard is available, it falls short in the case of NLG evaluation tasks. As we argued in the introduction, in the case of NLG the concept of IAA as reproducibility can hide some pitfalls. For evaluation tasks that aim to evaluate semantic or pragmatic language aspects – such as for instance concepts such as text usability, fluency, comprehensibility etc. – two people can entertain different, although equally valid, opinions. In cases such as these, given the variability of human language – specifically variability in language interpretation and quality judgement – expecting judges to always arrive at exactly the same judgement may be both unrealistic and over-constrained. Variation in language interpretation and use makes strict agreement unsuitable for measuring human evaluation reliability.

It is arguable that, from an evaluation point of view, what is important, more than the fact that judges have the same interpretation of the phenomena studied, is to know whether the judges are consistent relative to each other. A possible first step to test this is checking judges' relative consistency, that is checking whether the judges follow a systematic pattern in their assessments.

A feasible strategy to frame this problem is the following. Expecting judges to always arrive at exactly the same judgement may be unrealistic. For instance, one judge may be stricter than another one. However, in such situations the judgements would still covary. In other words, we can ask: Is it possible to predict J_a's judgements based on J_b's judgements, where J_a and J_b are two judges who are judging the same set of sentences?

Correlation coefficients can be used to answer this question. Such coefficients measure to what extent a variable changes, in a way not expected by chance alone, in relation to the change of another variable. That is, they measure the covariation of two variables. The change can be either in the same (positive correlation) or in the opposite (negative correlation) direction. In the presence of correlation, given a judges' annotation, it is mostly possible to predict the annotation of another judge. Correlation coefficients, measuring the judges' relative covariance, can give an insight into to the extent different judges are consistent relative to each other when annotating the data, even when their individual interpretations of the phenomena are not identical but following a consistent pattern, see for example (Stemler and Tsai, 2008, page 38) and (Gisev et al., 2013, page 331).

To test such an interpretation of correlation coefficients, we use data collected in a previous pilot study, and extend the analysis to two publicly available datasets with human evaluation: the QG-STEC[9] (Rus et al., 2010) and the Flickr-8k (Elliott and Keller, 2014) [10].

The pilot study was an attempt to define annotation guidelines for an Atumatic Question Generation task[11]. The methodology we used was that of refining the criteria chosen through several iterations of discussions and pilot evaluations. During these iterations we noticed that regardless of how many changes we made, there remained a divergence in the judgements that we could not reduce by modifying the guidelines. Nevertheless, we realized that such divergences showed an interesting degree of consistency, due to the fact that the judges were consistent in following their interpretation of the criteria in play. The pilot study we use in this paper, although consisting only of ten items, helps to formalize the problem and makes it clear from a visual point of view. Indeed, the use of ten items allows a clear visualization of the data. Although judgements are different in values, they show a clear pattern – see Figure 5 and Figure 1. Once we test the use of correlation coefficients in the pilot study we scale the experiment by the use of larger datasets, QG-STEC and Flickr-8k, that allow more stronger statistical conclusions.

4.1 Datasets analysis

Following Siegel and Castellan (1988) and Singh (2007) we use Goodman and Kruskal's Gamma as a correlation coefficient (Goodman and Kruskal, 1954) and Fleiss' κ (Fleiss, 1971) to measure the IAA. Goodman and Kruskals Gamma is the most adequate coefficient for ordinal data with many ties which is exactly our case[12]. Fleiss' κ is a measure for nominal or ordinal data annotated from two or more judges. To measure Fleiss' κ we used the implementation supplied in the nltk library.[13] Goodman and Kruskal's Gamma was measured with the *GoodmanKruskalGamma* function supplied by the *R* software.[14] In order to interpret the values obtained in the analysis, we use the Krippendorff scale of interpretation for IAA (Krippendorff, 1980) (see Table 1) and the interpretation for non-parametric correlation coefficient introduced in Rosenthal (1996) (see Table 4). Since Carletta (1996), the Krippendorff scale of interpretation has become the standard for CL annotation tasks.

For a nonparametric correlation coefficient we chose the Rosenthal (1996) scale because it extends Cohen's popular scale (Cohen, 1988). More precisely, it allows a more fine grained value distinction for the interval $[0.50, 1]$ – in particular, Rosenthal's scale specifies Cohen's "large" interval $[0.50, 1]$ into the two intervals "large" $[0.50, 0.7]$ and "very large" $[0.70, 1]$. Because Goodman and Kruskal's Gamma tends to give higher values than other correlation coefficients, such a choice allows a finer-grained analysis.

Correlation value	Value interpretation
[0, - 0.1] \ [0, 0.1]	Negligible
[-0.1, - 0.3] \ [0.1, 0.3]	Small
[-0.3, - 0.5] \ [0.3, 0.5]	Medium
[-0.5, - 0.7] \ [0.5, 0.7]	Large
[-0.7, - 1] \ [0.7, 1]	Very large

Table 4: Rosenthal (1996).

[9] The dataset is available at: https://github.com/Keith-Godwin/QG-STEC-plus/blob/master/Export-Subsets.zip.

[10] The dataset is available at: https://github.com/elliottd/compareImageDescriptionMeasures. For the original dataset detail we refer to (Hodosh et al., 2013).

[11] Given a text T as an input, the task was to generate a question which can be used, for example, to verify the respondents knowledge about T.

[12] Ties data are data with value repetition. In the case of the pilot dataset we use categorical questions (yes/no) and rating graphical scale. Likewise, the QG-STEC evaluation was performed with rating graphical scale, see Rus et al. (2012).

[13] The documentation for the agreement metric can be found at: https://www.nltk.org/_modules/nltk/metrics/agreement.html.

[14] The documentation for this function can be found at: https://www.rdocumentation.org/packages/DescTools/versions/0.99.19/topics/GoodmanKruskalGamma.

Interpretation of correlation coefficients case studies

Pilot: The pilot dataset was created taking random input paragraphs and questions from the SQuAD dataset (Rajpurkar et al., 2016). Seven judges engaged in the annotation task. Out of the seven judges, three were native speakers of English. The other four were proficient in English. The criteria measured were: i) Pertinence, ii) Grammaticality, iii) Comprehensibility, and iv) Fluency. The Pertinence criterion is ranked on a scale from 0 to 3, whereas the other criteria use a binary scale. Further detail about the dataset can be found in Amidei et al. (2018b)[15].

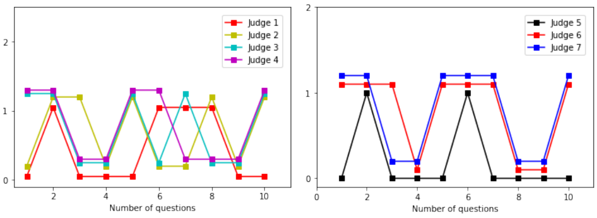
Table 5 reports the result for the pilot study. As we can note the native English speakers get the better IAA value and correlation results. Quite interestingly, although IAA value is below 0.4, for the fluency criterion they get a perfect correlation, which is 0.73 bigger than the correlation reached from non-native English speakers. This can be an indication that native judges have a different but strong interpretation of the concept of fluency[16]. Figure 1, which depicts the evaluation of question fluency, can help to better understand this phenomenon. Judge 5 systematically ranks with a value that is equal or less than the value given by judges 6 and 7. In contrast, the ranks provided by non-native English speakers lack systematicity.

Criteria	Coefficients	Dataset		
		All	Native	Non-native
Syntactic	Fleiss' kappa	0.36	0.59	0.21
	Goodman and Kruskal's Gamma	0.56	0.86	0.37
Comprehensibility	Fleiss' kappa	0.55	0.63	0.48
	Goodman and Kruskal's Gamma	0.92	1	0.91
Fluency	Fleiss' kappa	0.30	0.39	0.17
	Goodman and Kruskal's Gamma	0.56	1	0.27
Pertinence	Fleiss' kappa	0.20	0.22	0.07
	Goodman and Kruskal's Gamma	0.57	0.47	0.48

Average	Fleiss' kappa	0.35	0.45	0.23
	Goodman and Kruskal's Gamma	0.65	0.83	0.50

Table 5: Results of Fleiss' κ and Goodman and Kruskal's Gamma in the pilot dataset. *All*, *Native* and *Non-native* indicate the measure performed respectively over the seven judges, over the three English native speaker judges and over the four no English native speaker judges.

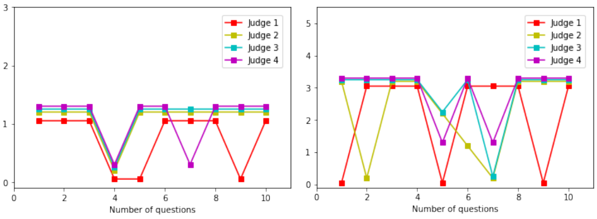
It is also worth noticing that for the cases of comprehensibility and pertinence, in the case of non-native English speakers, there is an interesting gap (more than 0.4) between IAA and correlation value. Figure 2 shows the annotators' ranks are different in value, which explain a medium/low, for comprehensibility, and very low, for pertinence, IAA. However, there is systematicity in the annotators' ranks – it is really clear in the case of compressibility, and less accentuated in the case of pertinence.



(a) Non-native English

(b) Native English

Figure 1: Plots of the evaluation of question fluency. Non-native English speakers (a) and native English speakers (b). For better readability, the scores are shifted upward slightly.



(a) Comprehensibility

(b) Pertinence

Figure 2: Plots of the evaluation of question comprehensibility (a) and question pertinence (b) in the case of non-native English speakers. For better readability, the scores were shifted upward.

The average score in Table 5 can be used to attempt a conclusion about the reliability of the dataset. Following the Krippendorff scale of interpretation (Krippendorff, 1980), the evaluation data should be discarded because the IAA is below the threshold of 0.67. However, following the scale of interpretation for non-parametric correlation coefficients introduced in Rosenthal (1996), the data reach a large correlation, and a very large correlation in the case of native English speakers.

Taking into account the interpretation we gave in the previous section, although the annotators use different values in the evaluation, their interpretations are constant with each other: they judgements covary systematically with each other. This interpretation suggests that the data are reliable.

[15] The evaluation guideline and the actual evaluation can be found via: `https://bit.ly/2lKL516`.

[16] It is worth noticing that in the case of No English native speaker, the Goodman Kruskal's Gamma measured on triple of judges reached the following values: 0.54, 0.29, 0.12, 0.12.

Flickr-8k: The Flickr-8K dataset contains quality judgements for 5,822 sentences[17] (Elliott and Keller, 2014)[18]. Each sentence was a description of an image. The annotation was carried out by 3 human experts who judged the sentence semantic correctness in a scale from 1 to 4.

Because we don't have the information about how the data were collected, in order to decide which kind of analysis to carry out on the Flickr-8k dataset we plot the distribution of the categories used by the judges. Figure 3 suggests that the data do not have a normal distribution, and so we opt for the use of nonparametric statistics. As in the previous case we used Goodman and Kruskal's Gamma and Fleiss' κ to carry out our analysis. For Goodman and Kruskal's Gamma, we report the average results of the pairwise measure between the annotators. This method is suggested by Siegel and Castellan (1988) for the case of Kendall τ correlation coefficient, which is a variant of the Goodman and Kruskal's Gamma.

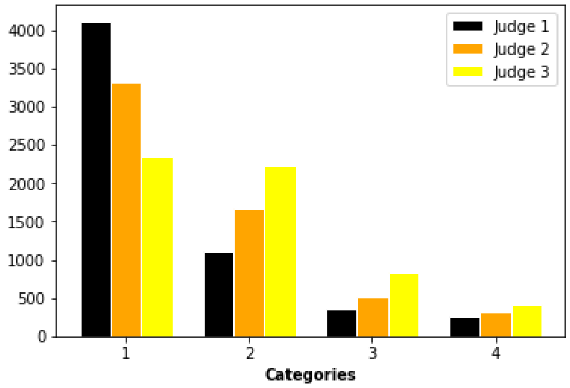


Figure 3: Distribution of the categories used by the judges in the Flickr-8k dataset.

The measurements give a Fleiss' value of 0.52 and a Gamma value of 0.98. Following the Krippendorff interpretation of IAA, the annotation has to be considered not reliable. However, the annotation achieves a very high correlation, which suggests a high relative consistency between the judges. Indeed, when they are in disagreement, judge 2 ranks systematically higher than judge 1, and judge 3 ranks systematically higher than judge 2. Although judges rank the items with different magnitude their judgement covary systematically. Also in this case, the correlation coefficient suggests the evaluation data are reliable and justifies a deeper analysis of the data quality. For example, following (Bayerl and Paul, 2007), the use of Generalizability Theory (Brennan, 2001), which allows a deeper analysis of the factors that influence annotation quality.

Use of correlation coefficients, an application

QG-STEC: The QG-STEC dataset is composed of questions generated from four systems that participated in the QG-STEC (Rus et al., 2010) Task B, that is the task to generate a question from an input sentence. Each question is evaluated based on five criteria: Relevance (on a scale from 1-4), Question Type (on a scale from 1 to 2), Syntactic Correctness and Fluency (on a scale from 1-4), Ambiguity (on a scale from 1-3) and Variety (on a scale from 1-3). Six judges took part in the evaluation. They judged batches of sentences independently. Table 6 shows the batch of questions judged and independent judges for that batch.

Judges	Batches of question
J1 and J2	80
J1 and J3	67
J1 and J4	81
J1 and J5	7
J1 and J6	106
J2 and J5	158
J3 and J5	125
J4 and J5	142
J5 and J6	129

Table 6: Batches of question with independent judges assigned to them. For $i = 1, \ldots, 6$, J_i means judge i.

Table 7 shows the result of the analysis carried out for the QG-STEC dataset. Also in this case an interesting discrepancy between the IAA values and the correlation values is measured. The average result, shows that, although IAA values are low, annotators reach large, and in two cases, very large Gamma correlation values.[19]

As in the previous cases, the Gamma coefficient suggests that all the batches are annotated by judges that shows a relative consistency and suggest data reliability.

Each pair of annotators evaluated different batches of questions, which were generated from

[17] The dataset is available at: `https://github.com/elliottd/compareImageDescriptionMeasures`.

[18] For the original dataset detail we refer to (Hodosh et al., 2013).

[19] We note that in the case of judge 1 and 3 the Gamma value is low due to the negative, although perfect, correlation reached in the question type criteria. If the average was done with absolute value the average gamma value would be 0.72.

Criteria	Coefficients	Coefficients values for pair of judges							
		J1 and J2	J1 and J3	J1 and J4	J1 and J6	J2 and J5	J3 and J5	J4 and J5	J5 and J6
Ambiguity	Fleiss' kappa	0.06	0.17	0.29	0.08	0.45	0.14	0.17	0.23
	Goodman and Kruskal's Gamma	0.33	0.33	0.61	0.21	0.84	0.51	0.39	0.75
Correctness	Fleiss' kappa	0.31	0.31	0.31	-0.00013	0.20	0.32	0.28	0.13
	Goodman and Kruskal's Gamma	0.67	0.67	0.71	0.39	0.57	0.62	0.57	0.39
QuestionType	Fleiss' kappa	0.32	-0.05	0.2	0.34	1	0.14	0.15	0.52
	Goodman and Kruskal's Gamma	0.91	-1	1	0.89	1	0.45	0.80	0.95
Relevance	Fleiss' kappa	0.13	0.16	0.13	0.08	0.15	0.28	0.19	0.01
	Goodman and Kruskal's Gamma	0.41	0.63	0.67	0.76	0.58	0.79	0.63	1
Variety	Fleiss' kappa	0.35	0.89	0.10	0.08	0.52	0.36	0.29	0.35
	Goodman and Kruskal's Gamma	0.81	0.99	0.08	0.15	0.82	0.73	0.43	0.54
Average	Fleiss' kappa	0.23	0.29	0.20	0.11	0.46	0.24	0.21	0.24
	Goodman and Kruskal's Gamma	0.62	0.32	0.61	0.48	0.76	0.62	0.56	0.72

Table 7: Fleiss' κ and Goodman and Kruskal's Gamma values reached in the QG-STEC dataset. For $i = 1, \dots, 6$, J_i means judge i.

4 different systems. Consequently, given the variance in question quality, a deeper analysis is complicated. However, we can see that judge 5 gets good correlation in any batch, which is also the case for judge 2. This fact allows us to consider the batch they annotated together as the more reliable one. This is confirmed by the *k* and Gamma values reached.

We can also notice that regarding the variety criterion, it is arguable that judge 4 and judge 6 miss a sound interpretation of the variety criterion. Indeed, both of them get low correlation with judge 1. Judge 1, on the other hand, gets really high correlation with both judge 3 and judge 2. At the same time, judge 5 gets high correlation with both judge 2 and judge 3 and lower correlation with judge 4 and judge 6. This evidence suggests that, in the case of the variety criterion, care must be taken with the data collected by judge 4 and 6.

Following the same analysis we can notice that judge 2 may be inconsistent for the relevance criteria. Indeed, it is lower than the correlation value reached by judge 1 and judge 5.

5 Conclusion

Based on an analysis of papers published over the last 10 years in NLG-specific conferences (in total 135 papers), we presented a snapshot of the use of IAA in NLG human evaluation tasks. One of the main points that stands out is the low level of IAA reached, and how few reliability studies there are. From our study, the problem of human evaluation reliability stands up.

Using three case studies, we show the limitations of using the IAA as the only criterion for checking the reliability of an evaluation. Given the variability of human language, we suggest that in human evaluation of NLG, correlation coefficients and agreement coefficients, such as for example the Kappa statistic, can be used together to have a better picture of the evaluation data reliability. Agreement coefficients can be used both in pilot studies to improve annotation schemes and guidelines, and for data analysis to give a picture of how distant the annotators' interpretation of the phenomena is. Correlation coefficients can instead tell us to what extent annotators are consistent with each other. As we have seen in Section 4, a low agreement coefficient value can hide a consistent pattern in the annotation which is captured by high value for the correlation coefficient. Although judges have different opinions about the quality of a generated text, which is a result of the language variability, they entertain consistent relative interpretations. Consequently, their judgments may still be considered reliable, although ideally further investigation, for example test-retest annotation (Krippendorff, 1980) and where possible the use of internal coefficient as Cronbach's alpha (Cronbach, 1951), should be carried out. Regarding test-retest evaluation and Cronbach's alpha, it is important to note that they have to be considered in the evaluation design.

Our aim with this paper is to enhance, in the NLG community, awareness about the need to handle the problem of human evaluation reliability. This problem is much more relevant nowadays given the growing use of crowdsourced workers in the evaluation phase. Indeed, in our analysis, we found that of the 29 papers that used crowdsourced workers, 23 were published in the last three years.

Acknowledgments

We warmly thank the anonymous reviewers for their helpful suggestions.

References

J. Amidei, P. Piwek, and A. Willis. 2018a. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317.

J. Amidei, P. Piwek, and A. Willis. 2018b. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329.

R. Artstein. 2017. *Inter-annotator agreement*. in Handbook of Linguistic Annotation, (Eds.) Nancy Ide and James Pustejovsky, pages 297-313. Springer, Dordrecht.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

P. S. Bayerl and K. I. Paul. 2007. Identifying sources of disagreement: Generalizability theory in manual annotation studies. *Computational Linguistics*, 33(1):3–8.

RL Brennan. 2001. *Generalizability theory*. New York: Springer-Verlag.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

J. Cohen. 1988. *Statistical power analysis for the behavioral sciences*. West Publishing Company, USA.

R. Craggs and M. M. Wood. 2005. Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 3(3):289–296.

L. J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

D. Elliott and F. Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 452–457.

M. A. Finlayson and T. Erjavec. 2017. *Overview of Annotation Creation: Processes and Tools*. in Handbook of Linguistic Annotation, (Eds.) Nancy Ide and James Pustejovsky, pages 167-911. Springer, Dordrecht.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

A. Gatt and E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

N. Gisev, J. S. Bell, and T. F. Chen. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338.

L. A. Goodman and W. H. Kruskal. 1954. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

M. Hodosh, P. Young, and J. Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

E. Hovy and J Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22(1):13–36.

A. Joshi, P. Bhattacharyya, M. Carman, J. Saraswati, and R. Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.

J. Kottner, L. Audigé, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner. 2011. Guidelines for reporting reliability and agreement studies (grras) were proposed. *International journal of nursing studies*, 48(6):661–671.

E. Krahmer and M. Theune. 2010. *(Eds.) Empirical Methods in Natural Language Generation*. Springer-Verlag, Berlin Heidelberg.

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

K. Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

J. R. Landis and G. G.. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

J. M. LeBreton and J. L. Senter. 2008. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods*, 11(4):815–852.

C. W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

M. Lombard, J. Snyder-Duch, and C. C. Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research*, 28(4):587–604.

A. Lommel, M. Popović, and A. Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), 26-31 May. Reykjavik, Iceland.*

M. Palmer and N. Xue. 2005. Linguistic annotation. *Computational Linguistics*, 31(1):71–106.

J. Pustejovsky and A. Stubbs. 2013. *Natural Language Annotation for Machine Learning*, volume 1. Published by OReilly Media, Gravenstein Highway North, Sebastopol, CA.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

E. Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

J. A. Rosenthal. 1996. Qualitative descriptors of strength of association and effect size. *Journal of social service Research*, 21(4):37–59.

V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan. 2010. The first question generation shared task evaluation challenge. *In: Proceedings of the Sixth International Natural Language Generation Conference (INLG 2010), 7-9 Jul 2010, Trim Castle, Ireland.*

V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan. 2012. A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, 3(2):177–204.

G. Sampson and A. Babarczy. 2008. Definitional and human constraints on structural annotation of english. *Natural Language Engineering*, 14(4):471–494.

J. Sedoc, D. Ippolito, A. Kirubarajan, J. Thirani, L. Ungar, and C. Callison-Burch. 2018. Chateval: A tool for the systematic evaluation of chatbots. In *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*, pages 42–44.

S. Siegel and N. J. Jr. Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw-hill, New York.

K. Singh. 2007. *Quantitative social research methods*. Sage, New Delhi.

S. E. Stemler and J. Tsai. 2008. Best practices in interrater reliability: Three common approaches. *In, Osborne, J. W., (ed.) Best practices in quantitative methods*, pages 29–49. Sage, California.

R. L. Wasserstein, A. Schirm, and Lazar N. A. 2019. *Statistical Inference in the 21st Century: A World Beyond $p < 0.05$*, volume 73. Taylor & Francis.

R. S. Witte and J. S. Witte. 2017. *Statistics, Eleventh Edition*. John Wiley & Sons, Inc., LaVergne, Tennessee, USA.

A complete list of the papers we examined can be found at: `https://bit.ly/2lKL516`.