Student's *t*-Distribution: On Measuring the Inter-Rater Reliability When the Observations are Scarce

Serge Gladkoff [1*], **Lifeng Han**[2*], and **Goran Nenadic**[2]

[1] Logrus Global, Translation & Localization
[2] The University of Manchester, UK

`lifeng.han, g.nenadic @ manchester.ac.uk`
`serge.gladkoff @ logrusglobal.com`

[*] *co-first authors*

Abstract

In natural language processing (NLP) we always rely on human judgement as the golden quality evaluation method. However, there has been an ongoing debate on how to better evaluate inter-rater reliability (IRR) levels for certain evaluation tasks, such as translation quality evaluation (TQE), especially when the data samples (observations) are very scarce. In reality, practitioners need to be able to assess the reliability of human MT quality evaluation based on one, two, or maximum three human linguists' judgements. In this work, we first introduce the little-known method to estimate the confidence interval for the measurement value when only one data (evaluation) point is available. This leads to our example with two human-generated observational scores, for which we describe "Student's *t*-Distribution", and explain how to use it to measure the IRR score using only these two data points, and calculate the confidence interval (CI) of the quality evaluation. We give a quantitative analysis of how the evaluation confidence can be greatly improved by introducing more observations, even if only one extra observation. We encourage practitioners and researchers to report their IRR scores and confidence intervals in all evaluations, e.g. using Student's *t*-Distribution method whenever possible; thus making the NLP evaluation more meaningful, transparent, and trustworthy.

1 Introduction

Human evaluations have been always the gold standard to judge the quality of natural language processing (NLP) system's outputs (Han et al., 2021; Freitag et al., 2021; Gladkoff and Han, 2022). This applies to many sub-tasks including machine translation (MT) (Han et al., 2020; Han, 2022a; Charalampidou and Gladkoff, 2022; MILAD, 2022), text summarisation (Bhandari et al., 2020; Latif et al., 2009), question answering (Aldahi et al., 2020), information extraction (Wu et al., 2022; Nenadic et al., 2004), and prediction (Yang et al., 2009), as well as domain applications such as social media, biomedical and clinical domains knowledge representation (Milošević et al., 2019; Yang et al., 2021; Krauthammer and Nenadic, 2004). Nonetheless, human evaluations have been subject to criticisms and debates about their reliability, particularly when conducted without strictly defined procedures. (Han, 2022b; Han and Gladkoff, 2022; Graham et al., 2017). Despite the inclusion of factors such as quality controls and clear guidelines, human evaluation results can vary greatly among different individuals due to subjective judgements influenced by factors such as backgrounds, personalities, cultures, and so on.

Naturally, the confidence levels of human evaluation become the key point to the validity of such work. There have been some efforts made on how to measure the confidence level of human evaluations from a statistical point of view, such as very recent work using Monte Carlo Sampling Simulations by Gladkoff et al. (2022). However, this kind of statistical measurement still needs a good amount of data points, or *observations*, to be based upon. When there are a limited amount of observations obtained from the experiments, *how to measure the confidence level properly* is still a challenging question. One of the solutions to address this is to calculate the inter-rater agreement level and inter-rater reliability (IRR) scores. There are some historical IRR measurement metrics including Cohen's Kappa (Cohen, 1960) and Krippendorff's Alpha (Krippendorff, 1987, 2011). How-

ever, as the last issue with statistical sampling, both Cohen's Kappa and Krippendorff's Alpha need a certain amount of samples data for probability calculation, which becomes troublesome when the observations are really *scarce*, e.g. only one, two, or a few data points. In addition, there are existing criticisms regarding the undesired prediction of Cohen's Kappa, e.g. Delgado and Tibau (2019) gave examples about how Kappa produced better scores for worse classifiers.

In this study, we examine scenarios in which observations from human evaluations are extremely scarce (e.g., limited to one or two values) and explore potential solutions. This endeavor is motivated by the realities of the translation and localization industry. Practitioners often need to determine the reliability of human machine translation (MT) quality evaluations based on the judgments of a single linguist, or at most, two to three linguist evaluations.

We start from one observation and introduce a confidence estimation method borrowed from Abbott and Rosenblatt (1962); Furnival et al. (1989) which was applied to forest study by Valentine et al. (1991). Then we discuss how this one observation-based confidence estimation is problematic and not much reliable. Following this, we bring an example of MT evaluation where two observations are obtained from the human assessment. In this case study, we introduce how to apply **Student's *t*-Distribution** to measure IRR with detailed formula interpretation and guidance. We also further give instructions on how to measure confidence intervals (CIs) using this method. We discuss the much improvement achieved by using two observational data points and Student's *t*-Distribution regarding narrowed-down CIs. Finally, we suggest that researchers also apply Student's *t*-Distribution to other NLP tasks and even beyond, i.e. outside of NLP tasks.

The rest of the paper is organised as below: Section 2 surveys the related work to ours on measuring IRR and confidence intervals from different fields, Section 3 presents a case study with a single observation, Section 4 follows up with two and more observations where we introduce Student's *t*-Distribution, and Section 5 concludes the paper with discussion and future work.

2 Related Work

Regarding agreement measurement, Cohen's Kappa metric was defined by Cohen (1960) as below:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where p_o is used to represent the ratio (proportion of units) when the raters agree with each other, while p_e is the agreement expected by chance. In the expression of frequencies, the Kappa value can be calculated by:

$$\kappa = \frac{f_o - f_e}{N - f_e} \quad (2)$$

In other words, the Kappa value reflects the agreement level (or proportion of agreement) after deducting the chance agreement. In the perfect situation, when the raters all agree with each other, i.e. the p_o value equals 1, the Kappa value will be 1 (Cantor, 1996). However, if the raters totally disagree with each other, i.e. the value of p_o is almost the same value of agreement by chance p_e, the Kappa value will be close or equal to 0. However, Kappa's value can be a negative number, when the agreement exhibition level is even smaller than by chance, e.g. using the above equations when the value of $p_o - p_e$ or $f_o - f_e$ is negative. As mentioned in the earlier section, the Kappa value requires a certain number of observations to properly estimate the metric scores.

Looking into the IRR measurement in crowdsourcing human evaluation domain, Wong et al. (2021) argued that the traditional Krippendorff's alpha or Cohen's kappa threshold values, e.g. above 0.6, are not ideal due to the ignorance of cultural and individual differences from crowdsource workers. They proposed a cross-replication reliability method based on Cohen's kappa and tested the methods on human judgements of facial expressions using a large amount of 4 million data points.

From NLP and MT field, Alekseeva et al. (2021); Gladkoff et al. (2022) applied Monte Carlo Simulation Analysis method to generate more samples for statistically estimating the confidence intervals of judgements when the samples presented for the human evaluation are small. Their experimental outputs demonstrate that not less

than 100, and ideally 200 segments are necessary for the test set to produce an unbiased, statistically significant quality evaluation of the MT system output.

Outside of NLP fields, there are also some efforts made to address similar issues in measuring reliability and confidence intervals. For instance, Hallgren (2012) gave a tutorial on measuring IRR for the psychology domain when multiple coders are involved using case studies using commonly used Cohen's kappa and intra-class correlation (ICC). Similarly in the educational and psychological domain, Walker and Göçer Şahin (2020) carried out a study on applying differential item functioning (DIF) analysis to measure IRR, in comparison to the inter-class correlation coefficient and Cohen's kappa statistics.

From animal behaviour studies, Harvey (2021) raised the issues on inter-rater and intra-rater reliability and made a discussion on Cohen's Kappa and Krippendorff's Alpha values. From the sociological domain, Belur et al. (2021) reported a systematic survey on reporting IRR values from crime studies on multiple coders. They made further discussion on how human factors affect decision-making and how important it is to report accuracy, precision and reliability from screening/coding.

There is some existing business software integrating IRR into their statistical tools such as SPSS that has been used in different sectors including medical training assessment (Beck et al., 2016). The IBM SPSS uses interclass correlation coefficient (ICC) to measure the IRR values among different groups of raters [1].

However, to the best of our knowledge, there is no existing work on applying Student's *t*-Distribution for measuring IRR in NLP applications, especially in translation quality evaluation (TQE) field.

3 On Single Judgement

When observational data is very scarce, more than half a century ago, Abbott and Rosenblatt (1962) proved the possibility of measuring confidence intervals on *a single data point* from a mathematical point of view, and the later work from Furnival et al. (1989) further elaborated Abbott and Rosenblatt's formula with a more narrowed interval generation. We name it the **ARF** Interval by taking

the initial letters of their names. [2]

This method may appear statistically counter-intuitive, but it is certainly worth mentioning here, particularly as production decisions are frequently based on a single quality measurement. An intriguing paradox arises: while many statisticians would argue it's impossible to determine a confidence interval from one measurement, project managers often rely on a single TQE (Translation Quality Evaluation) value to make their decisions. In actuality, conclusions about the reliability of a single measurement can be made, but they require supplemental information, e.g., for translation indusfy, known vendor's past performance. Within the ARF interval calculation method, this additional data is also derived from an experimentally-based prior knowledge or theoretically-based value that, while external to the measurement, arises from the project context. Interestingly, project managers who use a single measurement's value to make their decisions apply a similar intuition. Consequently, it's fascinating to explore what mathematical principles can elucidate within this context.

The width of the confidence intervals reflects the uncertainty of the experiments, i.e., the wider it is, the less knowledge is available about the setup.

A relatively narrowed confidence interval indicates the controlled situations, for instance, the normal distribution in the following formula of standardised transformation:

$$Z = \frac{y - \mu}{\sigma} \quad (3)$$

of which, y, μ, and σ are the variables of the *variate*, *mean value*, and *standard deviation*. The parameter z represents the *standardised variate*.

For the situation with one observation, let $\hat{\mu}$ be the independent and fixed value that is known before and outside of the measurement, and y be the experimental measurement value. Furnival et al. (1989) gives the following calculation intervals:

$$ARF = \frac{y + \hat{\mu}}{2} \pm k|y - \hat{\mu}| \quad (4)$$

This ARF interval contains the probability of μ that is larger than or equal to $1 - \alpha$, and α meets the following equation with k:

$$k = \frac{1 - \alpha + \sqrt{1 - 2\alpha}}{2\alpha} \ , \ 0 < \alpha \leq 0.5 \quad (5)$$

The pair value of (k, α) was given by Furnival et al. (1989) as in Table 1.

3.1 Case Study using ARF Intervals

Let's have a case study on using ARF intervals for MT evaluation. Assuming that a translation vendor has been evaluated earlier on certain lines of projects and the average result was a score of 96.3 ($\hat{\mu}$). The next translation quality measurement produced by another vendor is a lower rating of 85.2 (y). How reliable is this measurement by itself purely from the statistical point of view? and what does it tell us? If we assume the quality measurements are distributed normally, it is logical to take the average value of the prior history evaluations as the predicted value for future outcomes. Below we give two practices using ARF intervals.

1) If we construct a 75% confidence interval for the true quality rating, we need to use k=1.8 from the instruction by Table 1, and the corresponding α value is 0.25. Using the ARF interval formula, it gives:

$$ARF = \frac{96.3 + 85.2}{2} \pm 1.8 \times |96.3 - 85.2| \quad (6)$$

which is 90.75±19.98. Therefore, the ARF interval for the true value of quality rating is [70.77, 100]. As we can see from this example, the 75% confidence interval is almost half of the measured value itself, i.e. the maximum deviation of 19.98 is 22% of the measurement result (90.75). Although the mathematical precision of the single quality measurement is limited to this level, it can be beneficial to define these limits.

2) Similarly, for an 80% CI ($\alpha = 0.2$) for the true quality rating, the corresponding k value from the Table 1 is 2.31 and the above formula gives the following ARF value:

$$ARF = \frac{96.3 + 85.2}{2} \pm 2.31 \times |96.3 - 85.2|$$
$$(7)$$

which is 90.75±25.64. In other words, the interval for the true value of quality rating is [65.1, 100].

From this, we can see that with 80% CI, the maximum deviation of 25.64 is 25% of the measurement result (90.75).

From these two case studies, what is probably more interesting in the context of translation quality evaluation is that "the middle of the CI lies halfway between an earlier average result and the recent lower measurement". We can spell a good rule of thumb: *if the single measurement deviates from the average, the true value is likely halfway between the average and the new measurement*. Knowledge of this would help not to overreact to unusually low single scores newly generated.

Even though it is possible to measure the confidence levels, this ARF interval is very wide and the worse thing is that it can not be improved by the choice of α. As shown in Table 1, 0.5 is the narrowest option of choice for intervals. However, this value is considered not high enough to make a significant impact. To the right side of the table, the smaller value the α is, the wider the resulting intervals will be. Therefore, choosing α value between (0.2, 0.25) is probably the compromise to make when there is only one observation or judgement available. From this case study, our finding is that evaluations consisting of only a single measurement are not recommended as there will be a higher chance of bias as illustrated by our translation evaluation example. Such measurements have only rough and indicative values, so the data collection and analysis approaches must be invoked to improve the quality of measurement itself with the data science apparatus. This will lead to our next section when we recommend that a second quality measurement is very necessary, how to measure it in the new situation, and how much difference it will make.

4 On Observations of More Than One

Following the last section, we **call on more measurement points** for NLP evaluation tasks, especially in the language service provider sector where the single observation value is still very common in practice due to the cost concern. [3]

For instance, when a single translation quality measurement is not satisfactory for one of the parties, second quality measurement can be made to validate the first measurement. Then, how much improvement to the confidence interval can be ob-

---

[3]e.g. referring to the R&D report from Language Service Provider https://logrusglobal.com/

	value (α)						
Distribution	0.50	1/3	0.25	0.2	0.1	0.05	0.01
Normal (k)	0.05	1.26	1.8	2.31	4.79	9.66	48.39
Unknown (k)	0.5	1.87	2.91	3.94	8.97	18.99	99

Table 1: The Value Matching of (k, α) for both Normal Distribution and Unknown Ones by Furnival et al. (1989).

tained by introducing extra observational data? To answer this, the obvious problem, of course, is that at least 20-30 data points are required to calculate the mathematical variance for a normal distribution. In settings where the sample size is less than 30, and the standard deviation of the entire population is unknown, *Student's t-Distribution* can be used to evaluate standard deviation based only on the number of measurements between one and 30, e.g. 2, 3, etc.

4.1 On Student's *t*-Distribution

When the sample size (*aka* observations) is very small in comparison to the entire population, Student (1908) designed Student's *t*-distribution to measure the mean errors and the confidence intervals of estimation.

When there is one degree of freedom, the critical values for Student's *t*-Distribution are shown in Figure 1 including the confidence level, one tail, and two tail scores. [4] The full list of critical values with more degrees of freedom is shown in Figure 2.[5] There are many researchers who proposed different algorithms to calculate these critical values by hand and using computers, for instance, the work from Cheng and Fu (1983) and comparison studies by Blair and Higgins (1980).

The notation of Student's *t*-Distribution is defined as below if we use *T* as the random variable:

- $T \sim t_{df}$ where $df = n - 1$

where df is the degree of freedom and n is the number of observations. For instance, if we have a sample size $n = 2$, we calculate the $df = 2 - 1 = 1$ and write the distribution as $T \sim t_1$.

For the situation when the standard deviation is unknown, the error bound for the sample mean is defined as:

$$E = (t_{\alpha/2})(\frac{s}{\sqrt{n}}) \tag{8}$$

where $t_{\alpha/2}$ is the critical value of t-score with the area to the right equal to $\alpha/2$ (Figure 2), s is the standard deviation of observations (samples):

$$s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n - 1}} \tag{9}$$

where $\overline{x}$ is the mean value of n samples:

$$\overline{x} = (1/n)\sum x_i \tag{10}$$

The resulting confidence interval (CI) is then the following span:

$$CI = (\overline{x} - E, \overline{x} + E) \tag{11}$$

4.2 Deploying *t*-Distribution to IRR

Looking back to our MT evaluation experiments, from a practical industry project on language service, we have an example to demonstrate how to deploy Student's *t*-Distribution to measure IRR value. Assume we have used the Multidimensional Quality Metric (MQM) initialised by Lommel et al. (2014) [6] and professional translators for a translation evaluation project and got two numbers of overall quality scores: QS1=76.85 and QS2=81.99 on a scale from 0 to 100.[7] We can immediately see that the QS2 81.99 is 6.7% greater than the QS1 76.85, and oppositely QS1 76.85 is 6.3% less than QS2 81.99. Therefore QS2 agrees with 93.3% of QS1, and QS1 agrees with 93.7% of QS2. This is almost 95% agreement, so it looks good for most cases. However, if the PASS/FAIL threshold is 80, the difference may be crucial for

---

[4] https://people.richland.edu/james/lecture/m170/tbl-t.html
[5] https://www.stat.purdue.edu/~lfindsen/stat503/t-Dist.pdf

[6] open source project https://themqm.org/
[7] This is a real example from an industrial project on TQE called "Whale".

Conf. Level	50%	80%	90%	95%	98%	99%
One Tail	0.250	0.100	0.050	0.025	0.010	0.005
Two Tail	0.500	0.200	0.100	0.050	0.020	0.010
df = 1	1.000	3.078	6.314	12.706	31.821	63.657

Figure 1: Critical Values for Student's *t*-Distribution with One Degree of Freedom (from people.richland.edu)

the translator. Then, *what is the reliability of this evaluation result*?

The Sample Mean $\overline{x}$ of QS1 and QS2 is $(QS1 + QS2)/2 = 79.42$ for this sample of two measurements. The Sample Standard Deviation s for this sample of two values is $\sqrt{(QS1 - \overline{x})^2 + (QS2 - \overline{x})^2} = \sqrt{6.6049 \times 2}$ which is 3.6345.

The Confidence Interval depends on the desired Confidence Level, which, in turn, depends on the subject matter area of the content which was translated. For most fields, the confidence level should be at least 80%. The critical number $t_{\alpha/2}$ for that level (0.1 which is 20% divided by 2 for one tail of the graph, $\alpha = 0.2$) and two measurements (one degree of freedom, *df*=1) is 3.078, as shown in Figure 1 and 2.

Therefore, the margin of error for these measurements is:

$$E = \frac{3.078 \times 3.6345}{\sqrt{2}} = 7.91 \qquad (12)$$

This means that the confidence interval for these two measurements is 7.91×2=15.8, which indicates that with an 80% degree of confidence, the true quality score lies on this interval: [79.42 − 7.91, 79.42 + 7.91], or [71.51, 87.33].

As we can see from this result, given the second measurement, we can significantly improve the confidence interval as compared to the single measurement. The two different judgements (observations) that we obtained reduce variance from 25% of a single judgement (Section 3) to 9.96% (7.91/79.42), i.e. more than two times narrower with an 80% confidence setting.

However, as in the previous example, this confidence interval is still relatively large. Can we tell anything about the translator passing or failing the 89% PASS/FAIL threshold? The answer is that since the sample mean is below 80% and equals 79.42, the evaluation result is borderline FAIL.

Ideally, we need a third measurement or even more observations to further improve this interval, but in a production setting, the additional data points are rarely obtained by repeated evaluations, due to the cost and time constraints required for such a process.

The good news is that we already have reliable information for translation quality evaluation (TQE) purposes: this is a borderline FAIL, becasue the Sample Mean is lower than the threshold, and therefore more than half of possible values are below the PASS threshold. This is not bad for the measurement of such a subtle, almost intangible object as the human perception of quality. But you can only obtain a history of performance based on multiple evaluations for different content pieces and apply data science approaches.

5 Conclusion and Future Work

When it comes to evaluating translation quality, the ability to measure alone is not sufficient; we also need to know how reliable the measurement is. Automatic evaluation of quality quickly produces the same scores if repeated a number of times, which creates an illusion of precision. Unfortunately, the results of automatic quality measurement not only depend on the language pair, the Machine Learning system, the decoder, and the content type, but also vary from dataset to dataset, depending on the way the data have been cleaned and formatted. Given these factors, automatic measurement can be very fast and "reliable", but it may be (and often is) invalid, as well as inconsistent. Human translation quality evaluation (TQE) is currently the only way to obtain valid measurements of human perception of quality, and considered to be the golden standard of TQE. However, human measurement's inter-rater reliability (IRR) should be assessed, even if evaluation has been carried out correctly. Even if evaluators are experienced linguists, trained to do evalu-

ation according to proper system, and client specifications are clearly defined, the evaluators would still produce close but not the same evaluation results due to the very nature of human perception of quality, which is by definition the function of personal perception. This problem is exacerbated by the fact that in real life production setting there is no time or budget to validate the translation quality measurement even with the second reviewer, and even if there is a second reviewer, the low IRR of such measurement makes it difficult to confirm the first measurement.

In this paper, we first studied the typical production setting of gold standard human quality measurement, where TQE is performed by only one experienced, trained linguist, according to clearly defined customer specifications, producing a single measurement, and make conclusions about the reliability of such measurement. We then illustrated the results with the case of *Student's t-Distribution analysis of two measurements* made by two different reviewers.

From the first and second experiments, we can conclude that a single measurement has very low reliability and only has an indicative value. The confidence interval for one measurement is (as shown in Section 3) as wide as 25%, and therefore one evaluation cannot be taken as a basis for process decisions, more measurements are required. For instance, the second measurement can narrow down this interval and render it two times smaller, to around 10% (Section 4).

Yet we can say that the middle of the confidence interval lies halfway between the earlier average result and the lower recent measurement. *A good rule of thumb is born:* **if the single measurement deviates from the average, the true value sits, in all probability, halfway between the average and the new measurement**. Consequently, the recommendation is: please do not over-react on an unusually low new single score, take a middle ground between the older average and the new score, and think about it as the most probable result.

The second measurement may improve the confidence interval significantly but is rarely done unless during the arbitration. Therefore, it is more practical to obtain additional data points from other evaluations of different samples, in the course of the translation process.

Subsequent evaluations effectively are placed into two categories: (a) mostly PASS with only rare occasional FAIL, (b) all other cases (mostly FAIL, or many FAILs). This strategy is caused by the desire to ensure that a system is reliably well above the PASS/FAIL threshold and thus ensures quality results. Multiple evaluations also confirm the validity of quality measurements and allow the application of well-known maths of statistics of normal distribution.

However, it is worth noting that proper methods of data analysis are required to analyse *quality evaluation data-sets*, such as:

- Removal of outliers, which are caused by irrelevant causes.

- Evaluations made on very small or very large samples.

- Evaluations that are incorrect due to the improper application of metrics such as counting repeated errors, for example.

- Evaluations made by reviewers who were not trained, subjective, or had in mind different customer requirements.

It should be remembered, that data science only allows obtaining good results if you clean the data properly. Incorrect, biased, not properly calibrated, or imprecise conclusions and inferences may result from using uncleaned data.

Limitations

In this work, we discussed how to calculate confidence intervals and evaluation reliability when there are only one or two assessment scores from annotators, such as translation quality assessors. For the first case when there is only one new observation score, we assume there is a pre-estimated/expected score ready to use, i.e. for ARF interval. However, this might not be the case in some situations, or it might cost some time and money to get this value, for instance, for a newly established task without much prior knowledge. In the second case considered, we introduced Student's *t*-distribution method and gave two human judgement scores for estimation. This is expected to be helpful for the small number of observations; however, it does require some mathematical calculations using guided formulas and parameter tables, which might be not instantly convenient to translators or project managers who do

not have much statistical knowledge, and requires manual calculations from educated AI researchers anyway. For real world applications preliminary setup and additional clear and crisp guidance for practitioners may be required.

Ethical Statement

There are no ethical concerns in this work since it is only about introducing alternative methodologies for calculating the confidence and reliability of human evaluations.

Acknowledgements

References

JH Abbott and JI Rosenblatt. 1962. Two stage estimation with one observation on the first stage. *Annals of the Institute of Statistical Mathematics*, 14(1):229–235.

Alexandra Alekseeva, Serge Gladkoff, Irina Sorokina, and Lifeng Han. 2021. Monte carlo modelling of confidence intervals in translation quality evaluation (tqe) and post-editing dstance (ped) measurement. In *Metrics 2021: Workshop on Informetric and Scientometric Research (SIG-MET), 23-24 Oct 2021*. Association for Information Science and Technology.

Haifa Alrdahi, Uli Sattler, and Goran Nenadic. 2020. A text mining model for answering checklist questions automatically from parasitology literature. In *2020 International Conference on Computing and Information Technology (ICCIT-1441)*, pages 1–5.

Stefanie Beck, Bjarne Ruhnke, Malte Issleib, Anne Daubmann, Sigrid Harendza, and Christian Zöllner. 2016. Analyses of inter-rater reliability between professionals, medical students and trained school children as assessors of basic life support skills. *BMC medical education*, 16(1):1–8.

Jyoti Belur, Lisa Tompson, Amy Thornton, and Miranda Simon. 2021. Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, 50(2):837–865.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

R Clifford Blair and James J Higgins. 1980. A comparison of the power of wilcoxon's rank-sum statistic to that of student's t statistic under various nonnormal distributions. *Journal of Educational Statistics*, 5(4):309–335.

Alan B Cantor. 1996. Sample-size calculations for cohen's kappa. *Psychological methods*, 1(2):150.

Parthena Charalampidou and Serge Gladkoff. 2022. A case of application of a new human mt quality evaluation metric in the emt classroom. In *New Trends in Translation and Technology (NeTTT) Conference*, pages 161 – 165.

Smiley W Cheng and James C Fu. 1983. An algorithm to obtain the critical values of the t, $\chi 2$ and f distributions. *Statistics & Probability Letters*, 1(5):223–227.

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

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *arXiv e-prints*, page arXiv:2104.14478.

George M Furnival, Timothy G Gregoire, and Harry T Valentine. 1989. Confidence intervals and significance tests for a single trial. *Communications in Statistics-Theory and Methods*, 18(10):3749–3761.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. Measuring uncertainty in translation quality evaluation (TQE). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1454–1461, Marseille, France. European Language Resources Association.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Nat. Lang. Eng.*, 23(1):3–30.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Lifeng Han. 2022a. *An investigation into multi-word expressions in machine translation*. Ph.D. thesis, Dublin City University.

Lifeng Han. 2022b. An overview on machine translation evaluation. *arXiv preprint arXiv:2202.11027*.

Lifeng Han and Serge Gladkoff. 2022. Meta-evaluation of translation evaluation methods: a systematic up-to-date overview. In *Tutorial at LREC2022*, Marseille, France.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Naomi D Harvey. 2021. A simple guide to inter-rater, intra-rater and test-retest reliability for animal behaviour studies.

Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–526.

Klaus Krippendorff. 1987. Association, agreement, and equity. *Quality and Quantity*, 21(2):109–123.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Seemab Latif, Mary McGee Wood, and Goran Nenadic. 2009. Correlation between human assessment of essays and rouge evaluation of essays' summaries. In *2009 Eighth International Symposium on Natural Language Processing*, pages 122–127. IEEE.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

KHALED MILAD. 2022. Comparative evaluation of translation memory (tm) and machine translation (mt) systems in translation between arabic and english. In *New Trends in Translation and Technology (NeTTT) Conference*, pages 142–151.

Nikola Milošević, Dimitar Marinov, Abdullah Gök, and Goran Nenadić. 2019. From web crawled text to project descriptions: automatic summarizing of social innovation projects. In *International Conference on Applications of Natural Language to Information Systems*, pages 157–169. Springer.

Goran Nenadic, Sophia Ananiadou, and John Mc-Naught. 2004. Enhancing automatic term recognition through recognition of variation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 604–610.

CC Rodriguez. 1996. Confidence intervals from one observation. In *Maximum Entropy and Bayesian Methods: Cambridge, England, 1994 Proceedings of the Fourteenth International Workshop on Maximum Entropy and Bayesian Methods*, pages 175–182. Springer.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

Harry T Valentine, George M Furnival, and Timothy G Gregoire. 1991. Confidence intervals from single observations in forest research. *Forest science*, 37(1):370–373.

Cindy M Walker and Sakine Göçer Şahin. 2020. Using differential item functioning to test for interrater reliability in constructed response items. *Educational and Psychological Measurement*, 80(4):808–820.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.

Yuping Wu, Lifeng Han, Valerio Antonini, and Goran Nenadic. 2022. On cross-domain pre-trained language models for clinical text mining: How do they perform on data-constrained fine-tuning? In *arXiv:2210.12770 [cs.CL]*.

Hui Yang, Irena Spasic, John A Keane, and Goran Nenadic. 2009. A text mining approach to the prediction of disease status from clinical discharge summaries. *Journal of the American Medical Informatics Association*, 16(4):596–600.

Xi Yang, Chengkun Wu, Goran Nenadic, Wei Wang, and Kai Lu. 2021. Mining a stroke knowledge graph from literature. *BMC bioinformatics*, 22(10):1–19.

Appendix

A detailed Critical Value from the Student's *t*-Distribution is displayed in Figure 2 from Purdue University Statistics.

Critical Values for Student's *t*-Distribution.
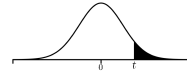
df	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850
21	0.859	1.323	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.819
22	0.858	1.321	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.792
23	0.858	1.319	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.768
24	0.857	1.318	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.745
25	0.856	1.316	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.725
26	0.856	1.315	1.706	1.822	1.967	2.056	2.162	2.479	2.779	3.707
27	0.855	1.314	1.703	1.819	1.963	2.052	2.158	2.473	2.771	3.690
28	0.855	1.313	1.701	1.817	1.960	2.048	2.154	2.467	2.763	3.674
29	0.854	1.311	1.699	1.814	1.957	2.045	2.150	2.462	2.756	3.659
30	0.854	1.310	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.646
31	0.853	1.309	1.696	1.810	1.952	2.040	2.144	2.453	2.744	3.633
32	0.853	1.309	1.694	1.808	1.950	2.037	2.141	2.449	2.738	3.622
33	0.853	1.308	1.692	1.806	1.948	2.035	2.138	2.445	2.733	3.611
34	0.852	1.307	1.691	1.805	1.946	2.032	2.136	2.441	2.728	3.601
35	0.852	1.306	1.690	1.803	1.944	2.030	2.133	2.438	2.724	3.591
36	0.852	1.306	1.688	1.802	1.942	2.028	2.131	2.434	2.719	3.582
37	0.851	1.305	1.687	1.800	1.940	2.026	2.129	2.431	2.715	3.574
38	0.851	1.304	1.686	1.799	1.939	2.024	2.127	2.429	2.712	3.566
39	0.851	1.304	1.685	1.798	1.937	2.023	2.125	2.426	2.708	3.558
40	0.851	1.303	1.684	1.796	1.936	2.021	2.123	2.423	2.704	3.551
41	0.850	1.303	1.683	1.795	1.934	2.020	2.121	2.421	2.701	3.544
42	0.850	1.302	1.682	1.794	1.933	2.018	2.120	2.418	2.698	3.538
43	0.850	1.302	1.681	1.793	1.932	2.017	2.118	2.416	2.695	3.532
44	0.850	1.301	1.680	1.792	1.931	2.015	2.116	2.414	2.692	3.526
45	0.850	1.301	1.679	1.791	1.929	2.014	2.115	2.412	2.690	3.520
46	0.850	1.300	1.679	1.790	1.928	2.013	2.114	2.410	2.687	3.515
47	0.849	1.300	1.678	1.789	1.927	2.012	2.112	2.408	2.685	3.510
48	0.849	1.299	1.677	1.789	1.926	2.011	2.111	2.407	2.682	3.505
49	0.849	1.299	1.677	1.788	1.925	2.010	2.110	2.405	2.680	3.500
50	0.849	1.299	1.676	1.787	1.924	2.009	2.109	2.403	2.678	3.496
60	0.848	1.296	1.671	1.781	1.917	2.000	2.099	2.390	2.660	3.460
70	0.847	1.294	1.667	1.776	1.912	1.994	2.093	2.381	2.648	3.435
80	0.846	1.292	1.664	1.773	1.908	1.990	2.088	2.374	2.639	3.416
90	0.846	1.291	1.662	1.771	1.905	1.987	2.084	2.368	2.632	3.402
100	0.845	1.290	1.660	1.769	1.902	1.984	2.081	2.364	2.626	3.390
120	0.845	1.289	1.658	1.766	1.899	1.980	2.076	2.358	2.617	3.373
140	0.844	1.288	1.656	1.763	1.896	1.977	2.073	2.353	2.611	3.361
180	0.844	1.286	1.653	1.761	1.893	1.973	2.069	2.347	2.603	3.345
200	0.843	1.286	1.653	1.760	1.892	1.972	2.067	2.345	2.601	3.340
500	0.842	1.283	1.648	1.754	1.885	1.965	2.059	2.334	2.586	3.310
1000	0.842	1.282	1.646	1.752	1.883	1.962	2.056	2.330	2.581	3.300
∞	0.842	1.282	1.645	1.751	1.881	1.960	2.054	2.326	2.576	3.291
	60%	80%	90%	92%	94%	95%	96%	98%	99%	99.9%

Upper Tail Probability: $\Pr(T > t)$

Confidence Level

Note: $t(\infty)_{\alpha/2} = Z_{\alpha/2}$ in our notation.

Figure 2: Critical Values List for Student's *t*-Distribution (from stat.purdue.edu)