Monte Carlo Modelling of Confidence Intervals in Translation Quality Evaluation (TQE) and Post-Editing Distance (PED) Measurement

Alexandra Alekseeva[1α] **Serge Gladkoff**[2α] **Irina Sorokina**[2,3α] and **Lifeng Han**[4β]
[1] ROKO Labs
[2] Logrus Global LLC
[3] Tver State University
[4] ADAPT Centre, Dublin City University
alex_sandi@mail.ru, {serge.gladkoff, irina.sorokina}@logrusglobal.com
lifeng.han@adaptcentre.ie

Abstract: From both human translators (HT) and machine translation (MT) researchers' point of view, translation quality evaluation (TQE) is an essential task. This is especially the case, when language service providers (LSPs) face huge amount of request frequently from their clients and users to acquire high-quality translations. While automatic translation quality assessment (TQA) metrics and quality estimation (QE) tools are widely available and easy to access, human assessment from professional translators (HAP) are often chosen as the golden standard (Han et al., 2021). One challenge that comes to this point is this: *to avoid the overall text quality checking from both cost and efficiency perspectives, how to choose the confidence sample size of the translated text, so as to properly estimate the overall text or document translation quality*? This work carries out such an motivated research to correctly estimate the confidence intervals (Brown et al., 2001) regarding the sample size of translated text, e.g. the amount of words or sentences, that needs to be taken into account for confident evaluation of overall translation quality. The methodology we applied for this work is from Bernoulli Statistical Distribution Modelling (BSDM) and Monte Carlo Sampling Analysis (MCSA). [1]

Keywords: Translation Quality Evaluation, Quality Estimation, Post-editing Distance, Confidence Intervals, Monte Carlo Modeling, Bernoulli Statistics

We start the experimental investigation of translation quality evaluation (TQE) by assuming that the errors in translated text are evenly distributed, and no more than one error per translated sentence. This assumption is placed as a random seed of our statistical modelling. While this assumption may be questioned, in our experimental evaluations, it proves that the starting value of random seed does not affect the overall model learning and the solutions to be reached. With this even distribution assumption of each sentence regarding translation errors, having error probability p with value 1 and no error probability $1-p$ with value 0, each sentence represent a random variable in the modelling. This situation forms the Bernoulli statistical distribution. When the sample size n is significantly smaller than the overall population N, the standard derivation of sample measurement falls into the following formula:

$$\sigma = \sqrt{\frac{p \cdot (1-p)}{n}}$$

The confidence interval CI, using the Wald interval (Newcombe, 2012)), will be:

$$CI = p \pm \Delta$$

where Δ is the product of standard deviation and factor 1.96 (when confidence level 95% is chosen) (Agresti and Coull, 1998).

$$\Delta = 1.96 \cdot \sigma = 1.96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

When the sample size n is comparable to the population size N, e.g. in a smaller translation evaluation project, the standard deviation is calculated as bellow and the Δ value updates correspondingly:

[1α] authors in alphabetical order, [β]: corresponding author

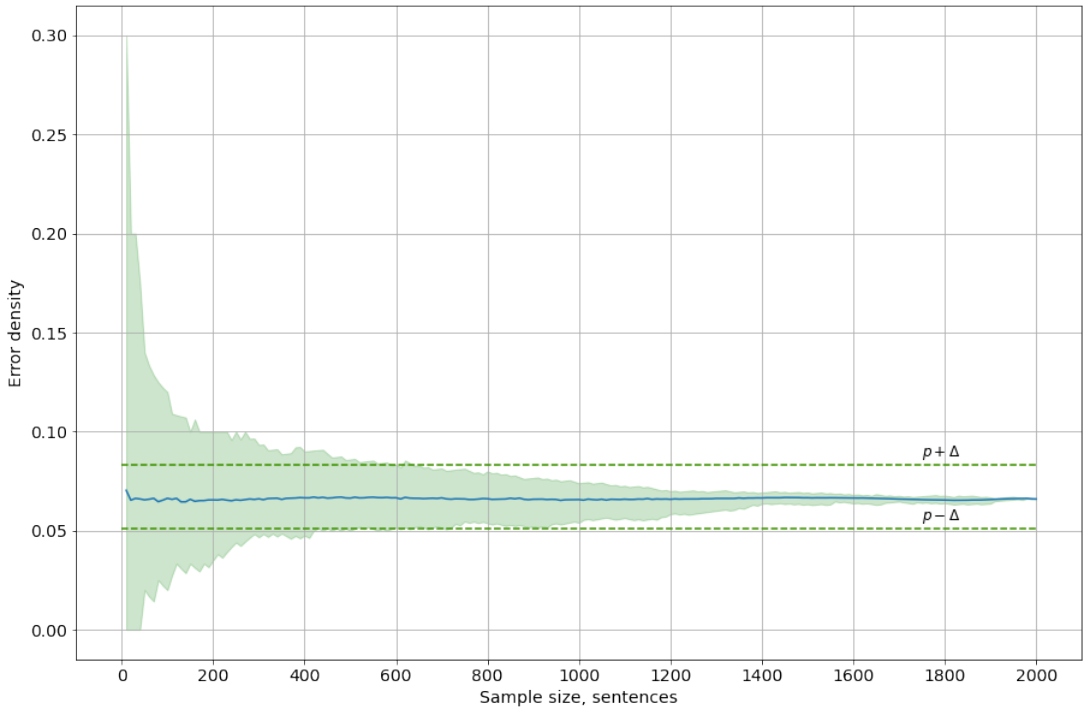

Fig. 1: Error density value with sample size variation from 100 to 2K sentences.

$$\sigma = \sqrt{\frac{p \cdot (1-p)}{n} \cdot \frac{N-n}{N-1}}$$

We carried out a case study using high quality translation with error density 0.07 (p value) and Δ value chosen as 0.02. Thus the confidence interval falls into 0.07 ± 0.02, i.e. from 0.05 to 0.09. Then the corresponding confident sample size n is inferred to be 625 sentences. To simulate the practical situation where the errors can come from different translators and different types, and span into a different weight across the translated text and documents, we applied the Monte Carlo sampling analysis, using a sample size of 2000 sentences and 95% confidence level (Fig. 1).

We investigate into another study on post-editing distance (PED) measurement for TQE, using confidence intervals regarding average PED scores depending on the sample size. Because PED score can be greater than the sentence length, we normalise the PED score to [0, 1] by the formula:

$$PEDn = 1 - tanh(c \cdot PED)$$

where c is a weighting parameter for PED value. The confident Δ value simulation according to sample size is shown in Fig. 2.

The estimated average sample size using $PEDn$ with confidence interval $\pm Delta$ is shown in Fig. 3, which indicates that the confidence interval starts to fall out of control when the sample size is less than 200 sentences.

In brief summary, this work investigates into confidence interval estimation for translation quality evaluation task, which has been an important role among language service providers and Informatics related fields, including machine translation (MT) and natural language processing (NLP). We used Bernoulli Statistical Distribution Modelling (BSDM) and Monte Carlo Sampling Analysis (MCSA), and gave concrete feed-backs and guidelines regarding practical situations when translation quality evaluation (TQE) is deployed.

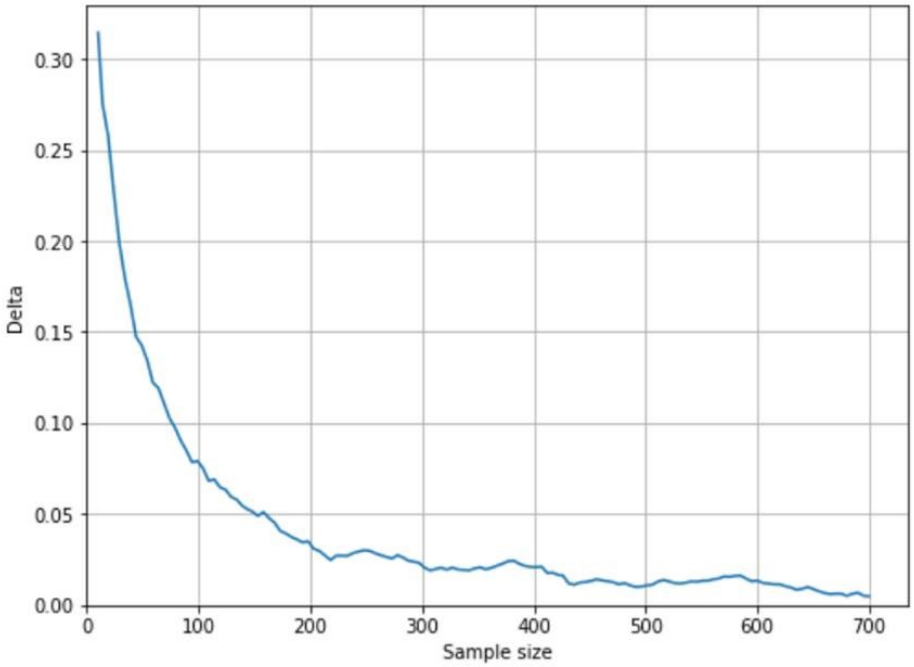

Fig. 2: Confident Delta value variation with sample size.

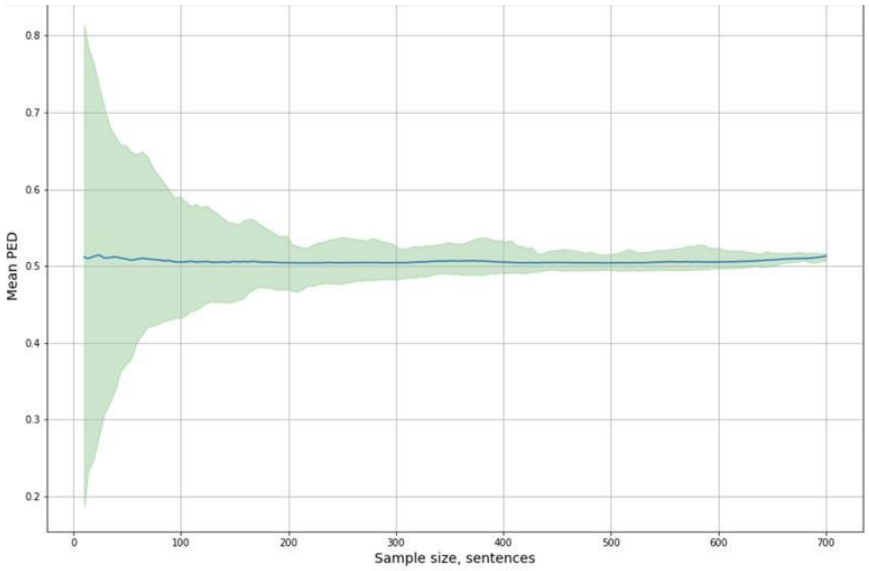

Fig. 3: Mean *PEDn* value variation with increasing sample size using a 95% confidence level.

References

Alan Agresti and Brent A. Coull. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2):101 – 133.

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, May. Association for Computational Linguistics.

Robert Gordon Newcombe. 2012. *Confidence Intervals for Proportions and Related Measures of Effect Size*. CRC Press. https://doi.org/10.1201/b12670.