Measuring Uncertainty in Translation Quality Evaluation (TQE)

Serge Gladkoff[1], Irina Sorokina[21], Lifeng Han[34*] and Alexandra Alekseeva[5],

{*serge.gladkoff, irina.sorokina*}*@logrusglobal.com, lifeng.han@*{*adaptcentre.ie, manchester.ac.uk*}

Institutes: [1]Logrus Global LLC [2]Tver State University [3]The Uni of Manchester [4]ADAPT Centre, DCU [5]ROKO Labs

LREC2022: 13th Edition of Language Resources and Evaluation Conference, Marseille, France, 20-25 June

Motivations

I. From both human translators (HT) and machine translation (MT) researchers' point of view, translation quality evaluation (TQE) is an essential task.

II. 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.

III. 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 [1].

IV. One challenge that comes to us: 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 translation quality?* [1]

Bernoulli Distribution on Errors

- Errors of certain type (category) either present in a sentence, or not.
- Errors are independent from each other.
- The probability of errors is the same.

When the sample size n is significantly smaller than the overall population N, the standard deviation of sample measurement falls into the following formula:

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

where p is the probability estimation of an event under study. The confidence interval CI, using the Wald interval, will be:

$$CI = p \pm \Delta$$

where Δ is the product of standard deviation and factor 1.96 (when confidence level 95 % is chosen):

$$\Delta = 1{,}96 \cdot \sigma = 1{,}96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

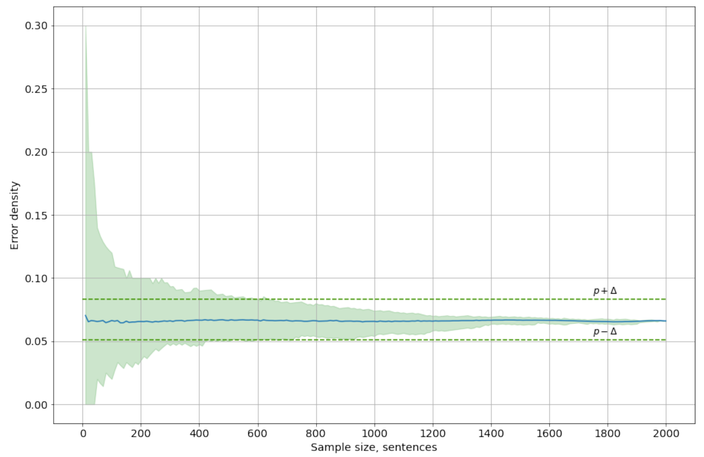

Figure 1 shows the Error density value with sample size variation from 100 to 2K sentences.

Confident Modelling

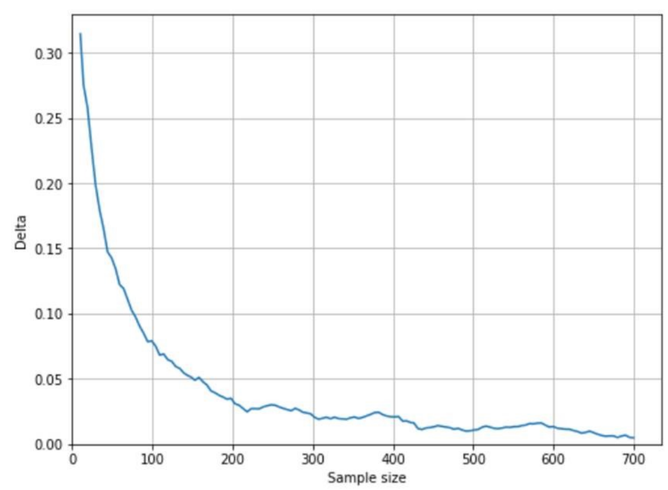

Figure 2 shows how the confident delta value changes with the variation of sample size.

Monte Carlo Simulation (MCS)

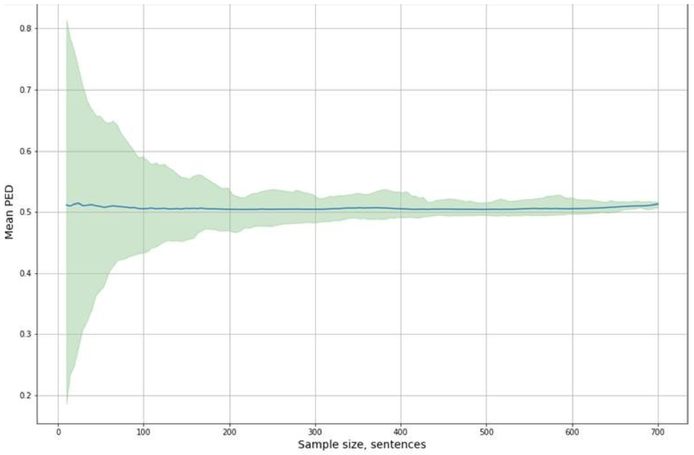

Figure 3 shows Mean post-editing distance (normalised) $PEDn$ value variation with increasing sample size using a 95 % confidence level.

Discussion and Summary

Our statistical modelling using MCS suggests that a sample size of less than 200 sentences can not reflect the overall material quality in a confident enough level on translation quality evaluation task.

Using MCS, we also reduced the suggested sample size from 10k words (around 625 sentences, from Bernoulli statistics to 4k for reliable estimation of overall translation quality.

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.

[1]Our data will be hosted as open-source repository on `https://github.com/lHan87/MCMC_TQE`. Corresponding author: LH*

Acknowledgement: Funding source: Logrus Global `https://logrusglobal.com/`. LH is partially funded by The Uni of Manchester and ADAPT Research Centre (DCU): The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.