Prompt-based Explainable Quality Estimation for English-Malayalam

Archchana Sindhujan, **Diptesh Kanojia**, **Constantin Orăsan**

Institute for People-Centred AI and Centre for Translation Studies,
University of Surrey, United Kingdom
{a.sindhujan, d.kanojia, c.orasan}@surrey.ac.uk

Abstract

This project aimed to curate data for the English-Malayalam language pair for the tasks of Quality Estimation (QE) and Automatic Post-Editing (APE) of Machine Translation. Whilst the primary aim of the project was to create a dataset for a low-resource language pair, we plan to use this dataset to investigate different zero-shot and few-shot prompting strategies, including chain-of-thought, towards a unified explainable QE-APE framework.

1 Introduction

This project is a one-year-long initiative funded by the European Association for Machine Translation (EAMT)[1]. The primary focus of our project was to create novel Quality Estimation (QE) and Automatic Post-editing (APE) datasets for the English (En) - Malayalam (Ml) language pair. QE refers to the task of automatically predicting the quality of machine-translated output without reference translations, while APE aims to automatically correct errors in machine translations. The Malayalam language is a low-resource language with over 38 million speakers across the world. Despite its presence with 86,553 Wikipedia articles[2] on the web, there was no available data for evaluating the quality of translation from English to Malayalam. For English to low-resource Indic language pairs, QE data exists for English to {Hindi, Marathi, and Gujarati} where target languages belong to the Indo-Aryan language family. However, from the Dravidian language family, QE data is only available for English-Tamil and English-Telugu (Blain et al., 2023). Our project addresses this gap by introducing a novel English-Malayalam QE dataset, expanding QE research within the Dravidian language family. The dataset comprises three direct assessment (DA) scores assigned by human annotators, along with human post-edited translations for the APE task. The manual post-editing process further facilitates the creation of word-level QE data[3], enriching its usability for fine-grained evaluation. Additionally, the project aimed at a comprehensive evaluation of multiple large language models (LLMs) for QE of low-resource language pairs.

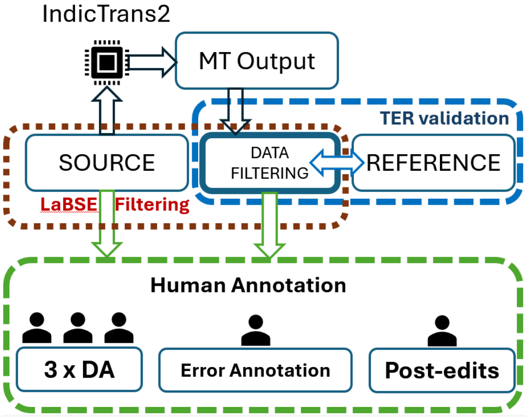

Figure 1: Our data curation workflow for QE-APE

2 Project Progress & Impact

Figure 1 depicts our workflow as described below. Our workflow consisted of- filtering high-quality parallel data, machine translation, TER-based validation, and human annotation for QE & APE.

2.1 Data Curation

We perform initial data curation leveraging data filtration techniques and iterative feedback to guidelines for annotation. For an initial comparison with existing references, we obtain a parallel corpus

[1] This is an individual project carried out at the University of Surrey

[2] en.wikipedia.org/wiki/Malayalam_Wikipedia

[3] github.com/WMT-QE-Task

for En-Ml translations via the Anuvaad parallel corpus[4], which provides domain-specific parallel data. We selected data instances from the finance, legal, and news domains to curate an initial larger set of instances. We filter out high-quality parallel data leveraging Language-agnostic BERT Sentence Embedding (LaBSE) scores with a high threshold (0.8) for contextual accuracy.

Post data filtration, we performed the translation using IndicTrans2 (Gala et al., 2023) model[5], the first fully open-source Transformer-based multilingual NMT model that supports translations across 22 Indic languages. The model adopts script unification wherever feasible to leverage transfer learning by lexical sharing between languages, which minimizes subword vocabulary fragmentation and enables the use of a smaller subword vocabulary.

To assess the translation quality, we compute the Translation Edit Rate (TER) by comparing the outputs of IndicTrans2 with the corresponding references from Anuvaad. TER acts as a reliable early indicator of translation quality and helps us manage DA score distribution. To validate this translation quality, a random sample of 25 translations was manually reviewed by a native Malayalam speaker fluent in English, providing early insights on common errors. For the final stage, we select $8,000$ segments, ensuring a balanced TER distribution. Our approach ensures that the curated dataset is well-distributed in terms of DA, suitable for segment-level computational modelling.

2.2 Annotation and Human Post-edits

After data curation, we shared segments with source and MT output, for DA score annotation with the annotation agency *TechLiebe*. First, a sample of 500 data instances was shared. At this step, we determine any deviations from the annotation guidelines and provide early feedback, then iteratively over weekly meetings. Each segment was evaluated and assigned a DA score by *three native speakers of Malayalam*, who are also fluent in English. Additionally, annotators were asked to provide a brief description of the identified errors. These *error descriptions* will act as ‘weak error explanations’, and will support the implementation of an explainable QE approach. After reviewing the 500 annotated samples and updating our annotation guidelines addressing the identified issues, we initiated the DA annotation in two batches, each containing 3750 segments. In weekly meetings, validation of random samples from the annotated data was performed. Any discrepancies observed were conveyed to all three annotators. Updated annotations were then re-evaluated in subsequent meetings to ensure alignment and consistency.

We started the post-editing process in parallel to the DA score annotation process with the help of an evaluator who was not involved in the DA annotation. The objective of post-editing is to make minimal edits to the translated output to convey the meaning of the source sentences. Initially, we shared a sample of 500 instances, validated the edits, and refined the post-editing guidelines before proceeding with two larger batches of 3750 segments each. By enhancing both translation evaluation and correction, this dataset aims to improve the performance of QE and APE of En-Ml.

2.3 Conclusion and Future Work

The annotation process progressed as planned, but validation and iterative corrections during weekly reviews required more time than expected. Prioritizing quality over speed ensured accuracy and consistency. Despite these challenges, our rigorous approach has guaranteed a high-quality English-Malayalam QE and APE dataset, with the majority of annotations completed and public release planned soon.

In future, we would like to perform synthetic reasoning generation based on error descriptions provided by human annotators collected with our dataset, leveraging LLMs to identify the penalization of DA score more accurately, further leading to improved QE and APE.

References

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. pages 629–653, Singapore.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. IndicTrans2: Towards high-quality and accessible Machine Translation models for all 22 Scheduled Indian Languages. *arXiv preprint arXiv:2305.16307*.

[4] github.com/project-anuvaad/anuvaad-parallel-corpus
[5] github.com/AI4Bharat/IndicTrans2