Validating Automatic Post Editing With T5

Introduction

In machine translation, the outputs of a model are often lower quality than what we would like. One solution to this problem is to take the outputs of the model and edit them to be more accurate, more grammatically correct, or just more fluid and natural as sentences.

Automatic Post Editing is the process of automating this correction process and research like Vu & Haffari 2018[1], has shown that this can be done effectively by a neural network. The usage of transformer models for this task is now common[2][3].

In this write-up I seek to validate that this kind of automatic post editing is viable with small models like T5 and to investigate if a single small T5 could be used for both the initial translations as well as for the post-editing process.

To validate that this approach can provide meaningful benefits to small models like T5, I have run the experiment described below.

Methodology

I began by training a T5-small model on the dataset **billingsmoore/LotsawaHouse-bo-en** for 10 epochs with 10% of the data held out for testing. During training input texts were prefixed with "Translate Tibetan to English:". This model will be referred to as the "baseline model".

The baseline model was used to produce predicted translations for the training dataset. These predicted translation were then used as the input data for a second model.

The second model was also a T5-small, trained for 10 epochs on the predicted translations as inputs with the actual English translations used as true labels. The input texts for this model were prefixed with "Post-Edit Translation:". This model will be referred to as the "post-editing model".

The third model, also a T5-small, was trained for 10 epochs on both the original translation dataset and the synthetic dataset produced for the second model.

For all three models, the metrics tracked were BLEU, chrF, and TER scores. For BLEU and chrF, higher scores are better. For TER lower scores are better.

Results

The baseline model achieved the scores shown in the table below.

BLEU	chrF	TER (lower is better)
22.0519	36.481	82.2966

The post-editing model achieved the scores shown in the table below.

BLEU	chrF	TER (lower is better)

BLEU	chrF	TER (lower is better)
33.6962	44.8243	70.0194

Discussion

You can see above that the post-editing model improves substantially on the baseline model's scores which I take to be sufficient to validate that automatic post editing is viable with small models.

Citations

[1]Automatic Post-Editing of Machine Translation: A Neural Programmer-Interpreter Approach (Vu & Haffari, EMNLP 2018)

[2]Do Carmo, Félix, et al. "A review of the state-of-the-art in automatic post-editing." Machine Translation 35 (2021): 101-143.

[3]Ki, Dayeon, and Marine Carpuat. "Guiding large language models to post-edit machine translation with error annotations." arXiv preprint arXiv:2404.07851 (2024).