Machine Translation Project

- Automatic Post-Editing, WMT 2023 -

Ionescu Andrei, 507

andreei.ionescu@gmail.com

andrei.ionescu8@s.unibuc.ro

Abstract

The aim of this paper is to tackle the task of automatic post-editing, in the context of the 9^{th} round of the APE shared task. The proposed challenge is to be able to create a machine translation model that can aid another hidden machine translation model by correcting its mistakes and reducing the human effort needed to perform such manual and laborious interventions. The approach highlighted in this work consists of leveraging an idea from the previous 2022 edition, by encoding the **SRC** and the **MT** inputs using two different encoders aiding in such a way a decoder transformer model to focus solely on providing corrections in an autoregressive manner.

1 Dataset

The dataset for this task targets the English to Marathi translation, due to the large usage of the Indo-Arayn language and yet limited corpus availability for neural machine tasks. By posing such a challenge, proposed solutions would need to find more robust solutions that can work in challenging situations. Concretely, the dataset is made up of two parts: human annotated data and synthetic resources from various domains. The human annotated data is a bit scarce, having around 19.000 samples while the synthetic subset has near 2.5 million examples obtained by applying the Indic-Trans En-X Model (Ramesh et al., 2022) over the *Anuvaad* (Gowtham Ramesh, 2024) parallel corpus. In the course of this work, the training set is considered to be the concatenation of these two subsets along with the removal of approx. 1M entries from the synthetic examples that have an empty MT entry. On the other hand, the given test set contains 1.000 examples.

2 Proposed Approach

Considering that the training dataset is made up of triplets of the form (SRC, MT, PE), where SRC represents an english text, MT is the translation performed by the black-box model over the SRC entry to Marathi language and PE is either a human annotated text or a synthetic back-translation, I could leverage different experts for each language and train a decoder on top to adjust the given translation. Concretely, I considered two different pretrained RoBERTa models:

- **RoBERTa Base** as an english expert (Liu et al., 2019)
- **MahaRoBERTa** as a marathi expert (Joshi, 2022)

These two models would be able to encode the SRC and the MT texts using rich pretrained representations. Then a Transformer (Vaswani et al., 2023) decoder may leverage these representations through the use of a guiding cross-attention mechanism. By taking inspiration from (Libovický, Helcl, & Mareček, 2018), I considered adopting a Serial Cross-Attention approach by first guiding using the SRC encoding and then the MT encoding from the pretrained RoBERTa models. Moreover, seeing that a similar approach was also tried by (Correia & Martins, 2019), I tied the input embedding of the decoder by leveraging those from the **MahaRoBERTa** as they work in the same language space. The three components are trained end-to-end by considering a causal language modeling objective where the post-edits have to be predicting by the decoder that is guided by the two pretrained encoders. The Transformer decoder is kept small due to insufficient computational resources and has only three layers with four attention heads. For the training procedure, Adam is used with a learning rate of $1e-4$ and $\beta = (0.9, 0.998)$ with mini-batches of size 4 for 90.000 steps. The architecture used can be seen in Figure 1.

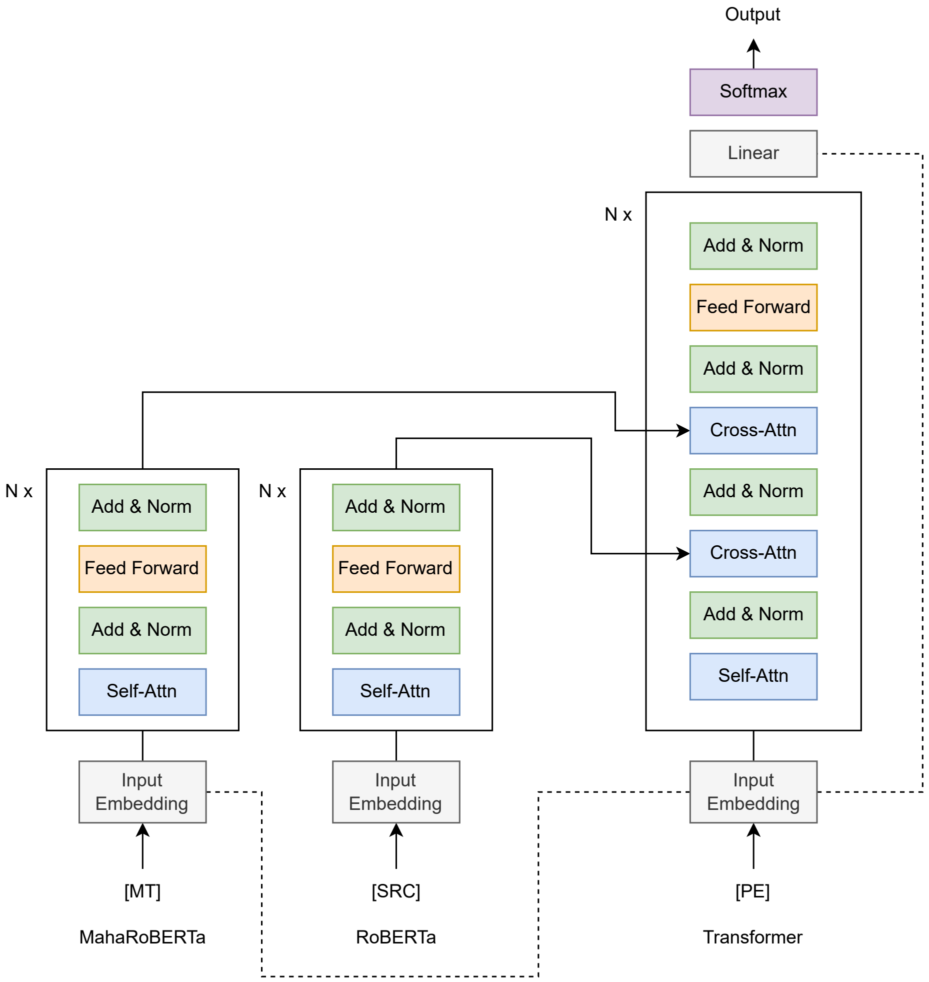

Figure 1: Encoder-Decoder Multi-Source Transformer
Implemented using OpenNMT (Klein, Kim, Deng, Senellart, & Rush, 2017) and HuggingFace.

3 Results

The training data was split into two subsets with the validation subset containing 800 examples and the training subset containing the rest. As it can be seen from the following figures, while the network was able to minimize the cross-entropy loss towards a certain point, the validation metrics TER and chrF have not seen improvements with more data. Also BLEU was computed but it always gave results close to zero which might indicate that there might be an issue somewhere in the evaluation step or just that the model does not perform well at all by following this training scheme and should be further adjusted.

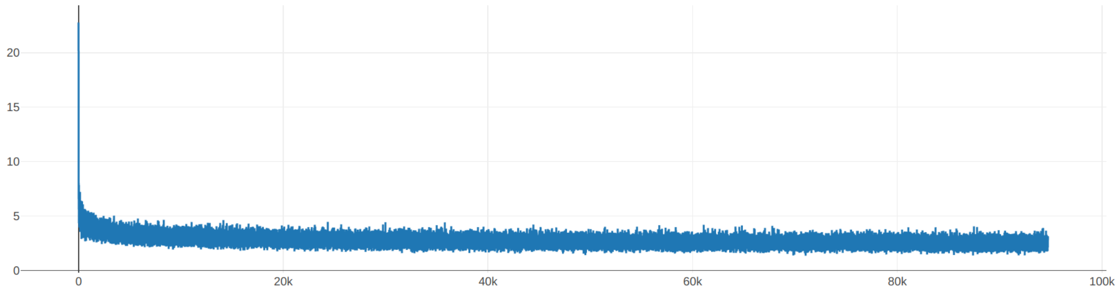

Figure 2: Training Loss

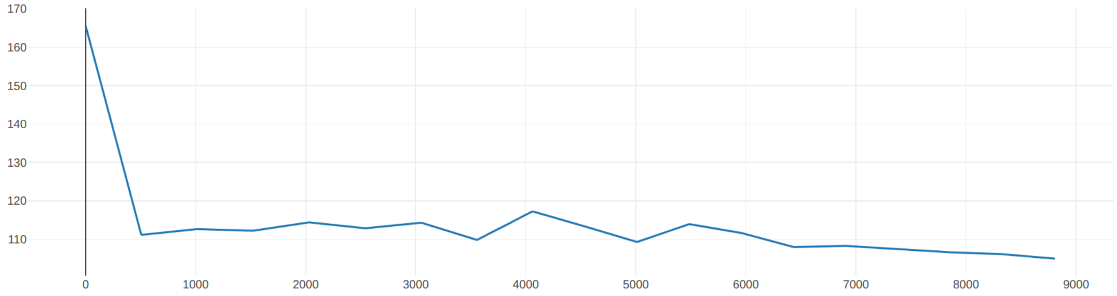

Figure 3: TER Score Computed Across the Validation Subset

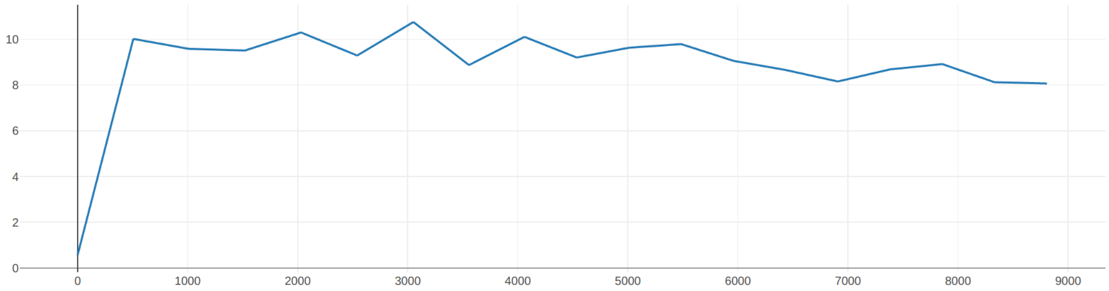

Figure 4: chrF Score Computed Across the Validation Subset

To conclude, the experiments tried in this work have not yielded success, possibly due to underlying issues that need further investigation, but with possible further refinements this might actually be a viable solution as it was proved empirically by other authors as mentioned above.

References

Correia, G. M., & Martins, A. F. T. (2019). *A simple and effective approach to automatic post-editing with transfer learning.*

Gowtham Ramesh, S. D. (2024, Jan). Retrieved from https://github.com/project-anuvaad/anuvaad-parallel-corpus

Joshi, R. (2022). *L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources.*

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017, July). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, system demonstrations* (pp. 67–72). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-4012

Libovický, J., Helcl, J., & Mareček, D. (2018, Jan). Input combination strategies for multi-source transformer decoder. *Edinburgh Research Explorer (University of Edinburgh).* DOI: https://doi.org/10.18653/v1/w18-6326

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. Retrieved from http://arxiv.org/abs/1907.11692

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., ... Khapra, M. S. (2022, 02). Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, *10*, 145-162. Retrieved from https://doi.org/10.1162/tacl_a_00452 DOI: 10.1162/tacl$_a$0452

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2023). *Attention is all you need.*