PROMT Systems for Covid-19 MLIA Translation Task

Alexander Molchanov

PROMT LLC, 17E Uralskaya str. building 3, 199155, St. Petersburg, Russia
Alexander.Molchanov@promt.ru

Abstract. This paper describes the PROMT submissions for the Covid-19 MLIA Shared Translation Task. We participated in all language pairs provided by the organizers (English to Spanish, German, Italian, Greek, French and Swedish). All our submissions are MarianNMT transformer-based neural systems. We submit two types of systems: constrained (i.e. built using only the data provided by the organizers) and unconstrained. We ranked top in all directions except for the English-German.

Keywords: Covid-19, Machine Translation, Neural Machine Translation, NMT, Transformer.

1 Introduction

In this paper we describe the PROMT submissions for the Covid-19 MLIA Shared Translation Task. We participate in all language pairs provided by the organizers, both in the constrained and unconstrained tracks. We ranked top in all directions except for the English-German pair in the first round. The paper is organized as follows: the three main sections describe our submissions for Rounds 1, 2 and 3 of the competition respectively. Each main section has three subsections: data, systems and results. The data section describes what data we use to build our systems (including synthetic data for the constrained settings). The system section describes our system settings (architecture, platform etc.). The results section summarizes the results and sets possible objectives for future work. Finally, each subsection is divided into 'Constrained' and 'Unconstrained' subsections.

2 Round 1

This section describes our submissions for the first round of the Translation Task.

2.1 Data

Constrained We use all data provided by the organizers. We use the train data for training and the devsets as our validation sets during training. No filtering or preprocessing is applied. We concatenate all data using deduplication to one single multilingual corpus to build a 8k SentencePiece [1] model for subword segmentation. Due to

time constraints we decided to build a single multilingual model trained on all provided data. We only apply the SentencePiece segmentation to the data in terms of preprocessing and add a language-specific tag to the source side of the parallel sentence pairs (i.e. we add the '<it>' token to the beginning of the English sentence of the English-Italian sentence pair, whereas we add the '<sv>' token to the beginning of the English sentence of the English-Swedish sentence pair). We also remove all tokens that appear less than ten times in the combined deduplicated monolingual corpus from our vocabulary.

Unconstrained Our unconstrained systems are trained using all available data with GPL license (mainly from the OPUS [2] and statmt.org [3] websites) and private data mostly harvested from the Internet. We use our own implementation of BPE (described in detail in [4]) instead of SentencePiece as it shows better results according to our experiments, but we stick to SentencePiece in the constrained task as it seems to work better in low-resource settings.

2.2 Systems

Constrained We train a baseline transformer [5] multilingual system (with a single encoder and a single decoder) following the recipe [6] from the Marian [7] website. We use a shared vocabulary as we trained a single SentencePiece model on all monolingual data. We trained the model for 460k steps on a machine with two RTX2080 GPUs until it stopped to converge on the devset.

Unconstrained We basically train the baseline transformers following the same recipe as cited above. Our unconstrained models use our own BPE implementation instead of SentencePiece. The size of the BPE models and vocabularies varies from 8k to 16k. We do not use a shared vocabulary (and thus train separate BPE models) for the English-Greek pair as the two languages have different alphabets.

2.3 Results

Constrained We rank top [8] in all language pairs except for the English-German pair. We did not perform any human evaluation due to the time constraints.

Unconstrained Again, we rank top in all language pairs except for the English-German pair. We plan to tune our baseline systems for the second round.

3 Round 2

This section describes our submissions for the second round of the Translation Task.

3.1 Data

Constrained We use all parallel data provided by the organizers for both rounds. We use the train data for training and the devsets as our validation sets during training. No

filtering or preprocessing is applied. We update the 8k SentencePiece model for subword segmentation using all data from both rounds.

This year we decided to use synthetic data in addition to parallel data. We use the MEDYSIS corpora from the Multilingual Semantic Search task for Round 2. We noticed that this data is quite noisy so we used some simple heuristics to extract what 'looks like' sentences (i.e. strings that start with a capital letter and end with a punctuation mark). After extraction the data is additionally split into sentences using sentence-splitter from the Moses toolkit. We build an intermediate multilingual X-to-English model which we use to translate the extracted MEDYSIS data into English. These translations are then scored using our model from Round 1. We select the top-scored sentence pairs roughly the size of the corresponding language pair parallel corpus and use them as back-translations to train our final English-to-X model.

Unconstrained We submit the same systems as for Round 1.

3.2 Systems

Constrained The system architecture is basically the same as for Round 1. We train a baseline transformer multilingual system with a single encoder and a single decoder. We use a shared vocabulary. We train the model for approximately 1.5M steps on a machine with two RTX2080 GPUs until it stopped to converge on the devset. This year we additionally fine-tune the model on all language pairs separately. This gives us 1-2 additional BLEU points on average.

Unconstrained We submit the same systems as for Round 1.

3.3 Results

Constrained We rank top in English-Greek and show competitive results in other language pairs. Unfortunately we did not perform any human evaluation again due to the time constraints.

Unconstrained We show competitive results in all language pairs.

References

1. Kudo. T., Richardson. J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Computing Research Repository, arXiv:1808.06226 (2018).
2. Tiedemann, J: Parallel data, tools and interfaces in opus. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey (2012).
3. http://statmt.org/
4. Molchanov, A.: PROMT Systems for WMT 2019 Shared Translation Task. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 302–307, Florence, Italy (2019).

5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I.: Attention is all you need. In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA (2017).
6. https://github.com/marian-nmt/marian-examples/tree/master/wmt2017-transformer
7. Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A., and Birch, A.: Marian: Fast Neural Machine Translation in C++. In Proceedings of ACL 2018, System Demonstrations, pages 116–121, Melbourne, Australia (2018).
8. Casacuberta, F., Ceausu, A., Choukri, K., Deligiannis, M., Domingo, M., García-Martínez, M., Herranz, M., Papavassiliou, V., Piperidis, S., Prokopidis, P.: The Covid-19 MLIA @ Eval Initiative: Overview of the Machine Translation Task. https://bitbucket.org/covid19-mlia/organizers-task3/src/master/report/ (2021).