Lingua Custodia @ Covid-19 MLIA - Eval Initiative

Raheel Qader

Lingua Custodia , Paris, France

Abstract. This paper describes the participation of Lingua Custodia in the Covid-19 MLIA @ Eval Initiative. We propose a multilingual machine translation engine that is able to translate English to French, German, Spanish, Italian and Swedish. We also compare our model to single language models. The results show that our model performs among the top participating systems.

1 Introduction

The wide-spread of covid-19 has caused major health and economic problems around the world. The sudden appearance of this virus has lead to difficulties in communication between nations as most current Machine Translation (MT) engines do not recognize covid-19 related terminology, and thus, not able to properly translate such text. The idea of the Covid-19 MLIA - Eval Initiative is to accelerate the creation of necessary resources and tools in order to improve the quality of current MT systems in the context of covid-19 [1].

The initiative is basically a challenge of 3 rounds. At each round, the organizers release training, development, and test sets and participants have to develop MT models using this data only (called constrained MT) or they can opt to use additional data (called unconstrained MT). The first round of the evaluation initiative addresses 6 language pairs: English to German, English to French, English to Spanish, English to Italian, English to Modern Greek and English to Swedish.

This paper describes the participation of Lingua Custodia in this initiative. We participate in all but the English to Modern Greek task. The reason for leaving out this latter task was the time constraint and the fact the Greek has a very different writing script than the other languages. Since the source language is always limited to English, we experiment with multilingual machine translation approach.

The rest of the paper describes the data processing, the proposed MT architecture and the conducted experiments.

Covid-19 MLIA @ Eval Initiative, http://eval.covid19-mlia.eu/.

2 Data

This section gives details of the data provided by the organizers and the pre-processing done by our team in order to prepare the data for training the 5 engines.

2.1 Data of Round 1

As stated earlier, in the first round we participate in the English to German, French, Spanish, Italian and Swedish tasks. Table 1 shows the statistics of the data used for the six language directions. The French and Spanish directions have the largest training data with almost one million sentences each while the Greek and Swedish direction have the fewest. The validation sets vary significantly from one direction to another, with the German having the smallest set (528 sentences) and the Greek direction having the largest set (3.9K sentences). All language directions have a test set of 2K sentences.

Table 1. Corpora statistics. $|S|$ stands for number of sentences, $|T|$ for number of tokens and $|V|$ for size of the vocabulary. M denotes millions and K thousands.

		German		**French**		**Spanish**		**Italian**		**Modern Greek**		**Swedish**	
		En	**De**	**En**	**Fr**	**En**	**Es**	**En**	**It**	**En**	**El**	**En**	**Sv**
Train	$\|S\|$	927K		1M		1M		900K		834K		807K	
	$\|T\|$	17.3M	16.1M	19.4M	22.6M	19.5M	22.3M	16.7M	18.2M	15.0M	16.4M	14.5M	13.2M
	$\|V\|$	372.2K	581.6K	401.0K	438.9K	404.4K	458.0K	347.7K	416.0K	305.7K	407.5K	298.2K	452.0K
Validation	$\|S\|$	528		728		2.5K		3.7K		3.9K		723	
	$\|T\|$	8.2K	7.6K	17.0K	18.8K	48.9K	56.2K	78.2K	84.0K	73.0K	72.7K	11.4K	10.0K
	$\|V\|$	2.4K	2.6K	4.1K	4.5K	9.7K	10.6K	12.4K	14.9K	10.3K	14.5K	2.6K	2.8K
Test	$\|S\|$	2000		2000		2000		2000		2000		2000	
	$\|T\|$	34.9K	33.2K	33.2K	35.8K	32.6K	34.3K	33.7K	34.2K	42.6K	44.3K	35.3K	30.6K
	$\|V\|$	7.8K	9.6K	6.7K	7.7K	6.7K	7.9K	8.6K	10.4K	9.5K	12.5K	7.1K	8.2K

3 Machine translation models

In the first round of the challenge we experimented with several techniques to train our model. This include training models for a specific language pairs, and models that can translate from English to several languages. In addition to that, we performed few steps of pre-possessing on data provided data.

3.1 Pre-processing

We first clean the training data of all language pairs by removing very long sentences. Instead of a specific tokenization step (e.g., using Moses [2]), we applied SentencePiece [3] for subword segmentation, which replaces the tokenization and detokenization steps as well. We force sentencepiece to split numbers character-by-character. This will reduce the total number of digit combinations (i.e., 10) that the model has to see, thus, it can generalize much easier on numbers. We use a unigram SentencePiece by creating a shared vocabulary between source and target sequences with 50K for single and 70K for multilingual models.

3.2 Model architectures

Our machine learning models are all based on the transformer architecture [5]. We use the Seq2SeqPy toolkit [4], which is a very lightweight toolkit with several sequence-to-sequence implementations including the transformer model.

Single-language models:
Single-language models are basically models that take a sequence in a specific language and translates it to a sequence in the target language. One needs as many single models as the number of language directions with such an architecture.
Multilingual models: on the contrary, these multilingual model can be trained such that a single model can translate between several language directions. Multilingual machine translation can be implemented in several ways. One approach is to add a token in the beginning of the source sequence in order to indicate the target language (e.g., 2fr, 2de, 2es). Another approach is to use source factors, i.e., to attach the embedding of language-specific id to the embedding of each token in the source sequence. In our experiments, for the sake of simplicity, we use the former approach.
Hyper-parameters
For both models we use the standard transformer architecture with 6 encoder and 6 decoder layers. The size of the embedding and hidden states are set to 512 while the size of the feed-forward layer is 2048 and we use 8 attention heads. The source and target embeddings are tie with the vocabulary projection layer. The batch size was set to 80 and source/target max lengths were capped at 120. We use Adam optimizer with learning rate of 0.0002, a warmup step of 5000, and label smoothing of 0.1. Finally, during inference, we use a beam size of 5. Our models are trained on 5 RTX 2080 Ti gpus.

4 Experiments

The challenge allows participants to participate in to constrained MT or unconstrained MT. The latter allows for using additional training data and pre-trained model.In the first round, we only participated in the constrained translation task. We used a multilingual MT to train the English to French, Spanish, German Italian, and Swedish models. For comparison reason, we also train single models for English to German and French models. Table 2 shows th results for single and multilingual models. The English to French and Spanish models achieve significantly higher scores than the English to German, Italian and Swedish models.

Table 2. Results on the constrained machine translation task. Systems are scores by BLEU and chrF.

	En-De		**En-Fr**		**En-Es**		**En-It**		**En-Sv**	
	BLEU	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**
Single model	26.7	0.556	48.9	0.703	-	-	-	-	-	-
Multilingual	29.5	0.584	49.0	0.705	47.6	0.698	28.4	0.572	30.4	0.589

Since the number of training samples are not that much different, this could be due to the fact that the data for these two language pairs is the cleanest. As for the difference between single and multilingual models, we can see that in the English to German direction, the multilingual achieves a much higher score while on the English to French side the difference is insignificant. Further experiments are needed to understand why this has happened, but previous studied have already shown that multilingual models doesn't bring much improvement to rich language pairs such as English and French.

5 Conclusions

In the paper, we described the participation of Lingua Custodia in the Covid-19 MLIA @ Eval Initiative. As our first attempt, we used a multilingual MT model and achieved promising results. In the rest of the challenge we plan to use more advanced techniques such as transfer learning from massive language models.

References

[1] Casacuberta, F., Ceausu, A., Choukri, K., Deligiannis, M., Domingo, M., García-Martínez, M., Herranz, M., Papavassiliou, V., Piperidis, S., Prokopidis, P., Roussis, D.: The Covid-19 MLIA @ Eval initiative: Overview of the machine translation task. `https://bitbucket.org/covid19-mlia/organizers-task3/src/master/report/` (2021)

[2] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, pp. 177–180, Association for Computational Linguistics (2007)

[3] Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)

[4] Qader, R., Portet, F., Labbé, C.: Seq2seqpy: A lightweight and customizable toolkit for neural sequence-to-sequence modeling. In: LREC 2020, pp. 7140–7144 (2020)

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30**, 5998–6008 (2017)