CUNI-MTIR at COVID-19 MLIA @ Eval Task 3

Shadi Saleh, Hashem Sellat, Hadi Abdi Khojasteh, Pavel Pecina

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague
{saleh,khojasteh,sellat,pecina}@ufal.mff.cuni.cz

Abstract. This paper describes the participation of our team (CUNI-MTIR) in the COVID-19 MILA Machine Translation (MT) task. We present our implementation of four systems (English into French, German, Swedish and Spanish) in both constrained and unconstrained settings. We employ the Marian implementation of the Transformer model to train the constrained systems in the given training data (MLIA MT parallel data), while in the unconstrained systems, we use external medical-domain training data for the base models and then fine-tune those models using MLIA MT data.

1 Introduction

Covid-19 MultiLingual Information Access (MLIA) Eval is a community evaluation effort aims at developing resources and tools to improve information access for the public regarding the coronavirus emerging pandemic (2).

The machine translation task is relatively challenging because there are many terms appeared during the pandemic with no prior existence or they were used before in different context; thus, there was notably a lack of parallel data that contain these terms which might result in bad performance when translating COVID-19 related text. Also, people from different levels of medical background want to access information related to the COVID-19 pandemic either by search or translating text from a language they do not understand. Policymakers, doctors and medical practitioners nowadays want to stay informed about the measures related to the virus or the latest findings in its behaviour, potential treatment or vaccination; thus, there is now a demand for effective information access systems in the medical domain like it was never before.

We took part in the MLIA MT task (task 3) and submitted in the first round two systems (constrained and unconstrained) for four language pairs (English to German, French, Spanish and Swedish). We present in this work our approach to developing Neural Machine Translation (NMT) systems based on the Transformer model.

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

2 Task and Data Specification

2.1 COVID19-MILA

The purpose of the task is to develop MT systems that are focused on the COVID-19 related text, which leads to speed up the creation of multilingual information access systems and (language) resources for COVID-19 as well as openly share these systems and resources as much as possible.

In the machine translation (MT) task, the aim is to translate Covid-19 related text from English into six European languages. Participants were asked to submit at least one constrained system using the MLIA MT data (the data that is provided by the organisers) and optionally submit unconstrained systems with the freedom of use any data sources they choose (2).

2.2 MLIA MT Data

The COVID19-MILA provides data for training, validation and testing for all languages pairs. Statistics of the provided data in terms of the number of sentences, the number of tokens in both source and target languages are shown in Table 1. In addition to that, the test set includes 2000 sentences to be translated into each of the four languages.

data set	pair	#sentences	#tokens in src	#tokens in tgt
training	English - French	1,004,715	15,075,764	17,680,508
	English - German	926,147	14,591,527	13,705,229
	English - Spanish	1,028,287	15,055,063	17,388,950
	English - Swedish	806,925	14,535,593	13,163,709
validation	English - French	728	17,006	18,828
	English - German	528	8,172	7,619
	English - Spanish	2,473	48,868	56,235
	English - Swedish	723	11,366	10,028

Table 1: Statistics of the MLIA MT data in four language pairs, in terms of the number of sentences, number of tokens in the source language (English) and the number of tokens in the target language

2.3 Khresmoi Medical Data

The parallel data that is used in the unconstrained system is taken from the UFAL Medical Corpus[1] which was assembled during the course of several EU projects aiming at more reliable machine translation of medical texts and used for the purposes of WMT Biomedical Translation Task (1). It mainly includes the

[1] `http://ufal.mff.cuni.cz/ufal_medical_corpus`

EMEA corpus (8), UMLS metathesaurus (3), titles from Wikipedia articles in the medical categories mapped to other languages using Wikipedia Interlingual links, medical domain patent applications (6; 10), and various web-crawled data.

All the data is tokenised using Khresmoi tokeniser and then encoded using Byte-Pair-Encoding (BPE) with 32K merges(7).

3 Round 1

In round1, we participate in building systems for four language pairs (English into French, German, Spanish and Swedish) and we submit two systems for each pair (constrained and unconstrained systems). Neural Machine Translation (NMT) has been achieving state-of-the-art performance in the machine translation task for almost a decade now(4). Our systems are based on the Marian implementation of the Transformer model (5), with the same parameters reported by (9).

3.1 Constrained Systems

In the constrained systems, we strict our training data to the provided parallel data given by the MLIA organisers (including training and tuning sets). We stop training the NMT models after 5 iterations in which the models did not bring any improvement when validating its performance on the given validation sets. Then, we use the intermediate model that gives the best BLEU score on the validation set. This model is later used to translate the test sets from English into the target language.

3.2 Unconstrained Systems

In the unconstrained systems, participants are allowed to use external data resources (data that is not provided by the organisers) to train/tune their systems. We build our unconstrained systems as follow:

- Train NMT systems for all pairs using the Khresmoi data (Section 2.3), and apply early-stop of training when there is no improvement on the validation set in three consecutive iterations.
- Fine-tune the systems by continuing training using the MLIA data, and then choose the model that gives the best BLEU score on the validation set.

3.3 Results

Results of our submitted systems are shown in Table 2. We report the scores of both BLEU (bilingual evaluation understudy) and ChrF (the percentage of n-grams in the translation candidate which have a counterpart in the reference) in percentages.

		EN-ES	EN-DE	EN-FR	EN-SV
Constrained	BLEU	**32.9**	19.7	**34.9**	**25.1**
	ChrF	**59.1**	49.4	**60.5**	**54.1**
Unconstrained	BLEU	32.1	**20.0**	33.0	24.0
	ChrF	58.2	**49.9**	59.0	51.4

Table 2: Results of our submitted systems on the test set in terms of BLEU and ChrF in percentages

Results show that training the models directly using the MLIA parallel data results in better performance than fine-tuning them later using the same data, except for English-German pair where we notice a small improvement in the unconstrained system. This can be explained because the Khresmoi data, which is used to train the base models before fine-tuning them, is more general with respect to the COVID-19 domain data, contains 10 million sentences for each pair and MLIA parallel data contains less than 1 million sentences. When the MLIA data is used to fine-tune the models, it could not influence the weights of the models enough to favour the new expressions and vocabularies that appear in the fine-tuning data.

4 Conclusions

In round 1, we submitted four NMT models for each subtask (constrained and unconstrained). All our models were based on the Transformer model. The constrained models were trained using the provided training data, and the model that gave the best BLEU score on the validation set was chosen as a final model for testing. In the unconstrained systems, we trained base models using medical domain data; and then fine-tuned those models using the MLIA parallel data.

Our results in the first round show that training NMT models on significantly fewer data can results in better performance if the domain is very specific (COVID-19 related text), and can outperform models that are trained on more (but less domain-specific) data.

References

[1] Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., Tamchyna, A.: Findings of the 2014 Workshop on Statistical Machine Translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. pp. 12–58. ACL, Baltimore, Maryland, USA (June 2014)

[2] Casacuberta, F., Ceausu, A., Choukri, K., Declerck, T., Deligiannis, M., Di Nunzio, G.M., Domingo, M., Eskevich, M., Ferro, N., García-Martínez,

M., Grouin, C., Herranz, M., Jacquet, G., Papavassiliou, V., Piperidis, S., Prokopidis, P., Zweigenbaum, P.: The Covid-19 MLIA @ Eval initiative: Developing multilingual information access systems and resources for Covid-19. https://bitbucket.org/covid19-mlia/organizers-overall/src/master/report/ (2021)

[3] Humphreys, B., Lindberg, D., Schoolman, H., Barnett, O.: The Unified Medical Language System. Journal of the American Medical Informatics Association **5**(1), 1–11 (1998)

[4] Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On Using Very Large Target Vocabulary for Neural Machine Translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1–10. ACL, Beijing, China (2015)

[5] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (July 2018), http://www.aclweb.org/anthology/P18-4020

[6] Pouliquen, B., Mazenc, C.: COPPA, CLIR and TAPTA: Three Tools to Assist in Overcoming the Patent Barrier at WIPO. In: Proceedings of the Thirteenth Machine Translation Summit. pp. 24–30. Asia-Pacific Association for Machine Translation, Xiamen, China (2011)

[7] Sennrich, R., Haddow, B., Birch, A.: Improving Neural Machine Translation Models with Monolingual Data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 86–96. ACL, Berlin, Germany (2016)

[8] Tiedemann, J.: News from OPUS: A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: Recent advances in natural language processing. vol. 5, pp. 237–248. John Benjamins, Borovets, Bulgaria (2009)

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.a.N., Kaiser, L.u., Polosukhin, I.: Attention is All You Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., t, R.G. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)

[10] Wäschle, K., Riezler, S.: Analyzing Parallelism and Domain Similarities in the MAREC patent corpus. In: Multidisciplinary Information Retrieval, Lecture Notes in Computer Science, vol. 7356, pp. 12–27. Springer (2012)