# LIMSI @ MLIA

Sadaf Abdul Rauf, François Yvon

# LIMSI @ MLIA

Sadaf Abdul Rauf and François Yvon

Univ. Paris-Saclay, & CNRS, LIMSI
https://www.limsi.fr/fr/

**Abstract.** This report describes LIMSI's submissions to the Round 1 of translation shared task at MLIA . We have submitted systems using various combinations of in-domain and out-of-domain corpora fine tuned on the MLIA COVID data, in order to develop resource-heavy systems for the translation of COVID mlia data from English into French, using two pre-processing pipelines and pre-trained representations. Our experiments demonstrate the effectiveness of in-domain and task specific corpus, even if in small amounts.

## 1 Introduction

In Covid-19 MLIA @ Eval initiative, the Machine Translation (MT) task aims to organize a community evaluation effort for facilitating creation and dissemination of resources and tools for MT systems on Covid-19 related sentences. This domain falls under the category of biomedical domain, which is gaining interest owing to the unequivocal significance of medical scientific texts. The vast majority of these texts are published in English and Biomedical MT aims to also make them available in multiple languages. This is a rather challenging task, due to the scope of this domain, and the corresponding large and open vocabulary, including terms and non-lexical forms (for dates, biomedical entities, measures, etc). The quality of the resulting MT output thus varies depending on the amount of biomedical (in-domain) resources available for each target language.

We participated in round1 of MLIA translation task(3) for English to French direction. English-French is a reasonably resourced language pair with respect to Biomedical parallel corpora, allowing us to train our Neural Machine Translation (NMT) [11] with in-domain corpora as well as large out-of-domain data that exist for this language pair. Following [1], our main focus was to develop strong baselines by making the best of auxiliary resources. Two pre-prossessing pipelines, one using the standard Moses tools[1] and subword-nmt [10] and other using HuggingFace BERT API were developed. All systems are based on the transformer architecture [12], or and on the related BERT-fused transformer model of Zhu et al. [13]. Our systems are presented in Section 3.1.



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

[1] http://www.statmt.org/moses/

## 2 Translation from English into French

### 2.1 Data sources

We trained our baseline systems on a collection of biomedical corpora (in domain) and out-of-domain parallel corpus. Table 1 details the corpora used in training.

| Training | | | |
|---|---|---|---|
| **Corpus** | **Words (M)** | | **Sentences** |
| | **English** | **French** | |
| In-domain | 239379673 | 266842261 | 8139198 |
| Out-of-domain | 1138581981 | 1292055231 | 35762532 |
| MLIA | 19409220 | 22579335 | 1004174 |
| Mlia-dev | 17006 | 18828 | 728 |
| Mlia-self | 19704 | 22916 | 1000 |

**Table 1.** Data sources for the English-French mlia task (before tokenization)

We gathered indomain parallel corpora available for English-French in the biomedical domain. These included the biomedical texts provided by the WMT'20 organizers: Edp, Medline abstracts and titles [5], Scielo [7] and the Ufal Medical corpus[2] consisting of Cesta, Ecdc, Emea (OpenSubtitles), PatTR Medical and OpenSubtitles. In addition, we used the Cochrane bilingual parallel corpus [4][3] and the Taus Corona Crisis corpus.[4]. For out-of-domain corpus, we used the parallel corpora provided by the WMT14 campaign, which included Gigafr-en, Common crawl, Europarl, News Commentary and the UN corpora.

For development purposes, we used the tuning corpus provided by MLIA organisers and created a self-test corpus by selecting 1000 sentences randomly from the training corpus. The self-test corpus was not part of our mlia training corpus used to build the systems.

### 2.2 Pre and post-processing

The document level corpora were first retrieved from xml, split[5] into sentences and sentence aligned using Microsoft bilingual aligner [6]: these include Cochrane, Scielo and some unaligned documents from Edp. All train, development and test corpora were cleaned by removing instances of empty lines, URLs and lines containing more than 60% non-alphabetic forms.

[2] https://ufal.mff.cuni.cz/ufal_medical_corpus
[3] https://github.com/fyvo/CochraneTranslations/
[4] https://md.taus.net/corona
[5] https://github.com/berkmancenter/mediacloud-sentence-splitter

For tokenization into words and subwords units, two pipelines were considered. The first one is set up as follows (a) tokenize the French and English texts using Moses scripts[6]; (b) compute a joint Byte-pair Encoding (BPE) inventory of 32K units with subword-nmt;[7] (c) generate the translation; (d) detokenize and truecase the output, again with Moses scripts. Systems based on this pipeline are prefixed `M*`. The second one is slightly more complex as it heavily relies on the HuggingFace API[8] for accessing pre-trained BERT models. The corresponding systems are prefixed with `H*` or `B*`and comprise the following steps: (a) a simple tokenization script, (b) a multilingual segmenter mapping BPE units to pre-trained encodings generated according to [3] as input to the translation system (step (c)). In that case, the MT output is also a sequence of multilingual BPE units that further needs (d) to be reaccentuated and recased, before a final (e) detokenization. Step (d) is non-trivial and is performed by a monolingual translation system trained to convert HuggingFace BPE units into Moses BPE units,[9] which can then be properly reassembled and detokenized as for the Moses pipeline.

## 3 Round 1

### 3.1 Translation Framework

All systems use Facebook's seq-2-seq library fairseq [8]. We mostly used two architectures to build our systems: basic and big Transformer models [12] as well as BERT-fused transformer models [13]. For transformer small, the parameters settings are borrowed from `transformer_iwslt_de_en`.[10] We used memory efficient FP16 optimizer. The ReLU activation function was used in all 6 encoder and 6 decoder layers, 1024 hidden layer size and batch size of 4K. Training was optimized using Adam and a learning rate of 0.0005 was fixed for all experiments. Systems based on transformer big borrowed settings from `transformer_vaswani_wmt_en_fr_big`.

For the BERT-based models, we relied on BERT-NMT.[11] This allowed us to build the BERT-fused models using the same architecture and parameters as the baseline transformer models and to establish fair comparisons. In BERT-fused NMT model, the contextual representations are first computed by the BERT model for each token (in the source and target), these are then combined at each encoder and decoder layer using the attention mechanism. Full details are in Zhu et al. [13].

Given the large size of our training data in the unconstrained settings, the "lazy" output data set implementation was used to enable data loading in the RAM. Systems were trained until convergence based on the BLEU score on the development sets. Evaluation is performed using sacrebleu [9]. Scores are chosen based on the best score on the development set and the corresponding scores for that checkpoint are reported on the mlia self-test set and the scores provided by the organisers.

[6] `http://www.statmt.org/moses/`
[7] `https://github.com/rsennrich/subword-nmt`
[8] `https://Huggingface.co/transformers/model_doc/bert.html`
[9] This process is not completely error prone, and yields a BLEU score of 98.2 on Medline 18 test set.
[10] `https://fairseq.readthedocs.io/en/latest/models.html`
[11] `https://github.com/bert-nmt/bert-nmt`

| ID | Name | Detail | Dev | Self-test | BLEU | ChrF |
|---|---|---|---|---|---|---|
| | | **Constrained system** | | | | |
| M1 | trans | mlia | 34.4 | 57.7 | 43.5 | 0.660 |
| | | **Unconstrained systems** | | | | |
| M2 | indom | indom-ftmlia | 36.8 | 56.5 | 51.2 | 0.721 |
| M3 | trans | outdom-ms-ftindom | 33.8 | 45.5 | 49.3 | 0.710 |
| B4 | bert | outdom-hg-ftindom | 39.8 | 56.0 | 49.3 | 0.703 |
| M5 | mlia | biobpe | 36.4 | 58.5 | 48.5 | 0.705 |

**Table 2.** Constrained and unconstrained systems submitted in round 1. `M*` prefixed systems were built using the moses tokenisation pipeline, while `B*` used HuggingFace pipeline.

### 3.2 Constrained Systems

Constrained systems were built using only the training and development data provided by the MLIA organisers. BPE using 32K vocabulary units were learned only on the MLIA training and development corpus. The system submitted was built using transformer big, with settings being borrowed from `transformer_vaswani_wmt_en_fr_big`. It ranked third in evaluation with BLEU score of 43.5 and chrF score of 0.660 [2].

### 3.3 Unconstrained Systems

We experimented with several data combinations and frameworks and submitted four unconstrained systems. We will present the system in order of their rank according to round 1 evaluation scores. Our first unconstrained system `M2` was built using all the in domain biomedical corpora listed in Table 1. The system was trained until convergence and then fine-tuned by continuing training on the MLIA data. As per the scores from evaluation campaign, this is the highest scoring system amongst the ones we submitted.

$M3$ and $B4$ are the systems that we built by first initialising the parameters from huge out-of-domain corpus and later fine tuning on in-domain corpus. The in-domain BPE codes learned from all the Biomedical data (other than mlia corpus) was used to segment the out-of-domain corpus. The initial systems were trained for 4 epochs on general domain WMT14 EN-FR corpora (see section 2.1 for corpus details). Model parameters were then freezed and used to initialize training for the in-domain system. The MLIA corpus was replicated two times to increase the probability distribution of COVID related sentences. $M3$ uses the moses tokenisation pipeline whereas $B4$ is based on the hugging face pipeline as already detailed in Section 2.2. The two systems have the same BLEU score and a very close ChrF score as per MLIA campaign evaluation score. However, this is not the case for the scores on development data and the self-test data where $B4$ outperforms $M3$.

The most interesting case is $M5$, which is the system built using only MLIA corpus, but the bpe units are computed using the BPE codes of all the in-domain data. This can be called a semi-constrained system as the only difference from $M1$ (the constrained

system) is that for $M1$, the bpe units were learned from MLIA corpus only, where as for $M5$ bpe codes learned from all in-domain corpora are used. Despite being trained on a very small amount of corpus as compared to other unconstrained systems, the scores are almost comparable to the other systems as per MLIA evaluation results. Whereas, this system is the best scoring system as per mlia-self test data.

Our results show the importance of in-domain corpus. System trained on only small amount of MLIA corpus, with bpes learned from in domain corpora is very close to systems trained on huge general domain data and fine tuned on in domain corpora. The system built using only in-domain data was ranked third, while the rest four ranked forth in evaluation [2].

### 3.4 Conclusion

Our experiments for the first round show that training models on significantly fewer data (which has learned domain specific codes) results in better performance if the domain is very specific, COVID-19 related text in this case, and can be comparable to models that are trained on huge but less domain specific data.

## References


[1] Abdul-Rauf, S., Rosales, J.C., Pham, M.Q., Yvon, F.: LIMSI @ WMT 2020. In: Conference on Machine Translation, Online, United States (Nov 2020), URL https://hal.archives-ouvertes.fr/hal-03013198

[2] 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)

[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://doi.org/10.18653/v1/N19-1423, URL https://www.aclweb.org/anthology/N19-1423

[4] Ive, J., Max, A., Yvon, F., Ravaud, P.: Diagnosing high-quality statistical machine translation using traces of post-edition operations. In: International Conference on Language Resources and Evaluation - Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (MT Eval 2016 2016), p. 8, Portorož, Slovenia (24/05 2016), URL http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MT%20Evaluation_Proceedings.pdf#page=65

[5] Jimeno Yepes, A., Névéol, A., Neves, M., Verspoor, K., Bojar, O., Boyer, A., Grozea, C., Haddow, B., Kittner, M., Lichtblau, Y., Pecina, P., Roller, R., Rosa, R.,

Siu, A., Thomas, P., Trescher, S.: Findings of the WMT 2017 biomedical translation shared task. In: Proceedings of the Second Conference on Machine Translation, pp. 234–247, Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017), https://doi.org/10.18653/v1/W17-4719, URL https://www.aclweb.org/anthology/W17-4719

[6] Moore, R.C.: Fast and accurate sentence alignment of bilingual corpora. In: Richardson, S.D. (ed.) Proc. AMTA'02, pp. 135–144, Lecture Notes in Computer Science 2499, Springer Verlag, Tiburon, CA, USA (2002), URL https://www.microsoft.com/en-us/research/publication/fast-and-accurate-sentence-alignment-of-bilingual-corpora/

[7] Neves, M., Yepes, A.J., Névéol, A.: The Scielo Corpus: a parallel corpus of scientific publications for biomedicine. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2942–2948, European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), URL https://www.aclweb.org/anthology/L16-1470

[8] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M.: fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 48–53, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://doi.org/10.18653/v1/N19-4009, URL https://www.aclweb.org/anthology/N19-4009

[9] Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 186–191, Association for Computational Linguistics, Belgium, Brussels (Oct 2018), https://doi.org/10.18653/v1/W18-6319, URL https://www.aclweb.org/anthology/W18-6319

[10] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715–1725, Berlin, Germany (Aug 2016), https://doi.org/10.18653/v1/P16-1162, URL https://www.aclweb.org/anthology/P16-1162

[11] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pp. 3104–3112 (2014), URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008, Curran Associates, Inc. (2017), URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[13] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., Liu, T.: Incorporating BERT into Neural Machine Translation. In: Proceedings of the International Conference on Learning Representations, ICLR (2020), URL https://openreview.net/forum?id=Hyl7ygStwB