LIMSI @ MLIA Round1 (Task 3)

Sadaf Abdul Rauf and François Yvon

Université Paris-Saclay, LISN-CNRS, France

MLIA Round1, January 14, 2021

COVID-19 MLIA Machine Translation Task

We built systems only for the English-French language pair.

Our Focus

Develop strong baselines by making the best of auxiliary resources:

- In-domain corpora
- Out-of-domain corpora
- Pre-trained multilingual contextual embeddings

Data Sources

In-domain:

- WMT'20 (Edp, Medline abstracts and titles, Scielo, Ufal Medical corpus)
- Cochrane Ive et al. (2016)[a] and the Taus Corona Crisis corpus.[b].

[a] https://github.com/fyvo/CochraneTranslations/
[b] https://md.taus.net/corona

Out-of-domain:

- WMT'14 corpus (Gigaword, Common crawl, Europarl, News Commentary and the UN corpora.)

Data Sources

Training			
Corpus	**Words (M)**		**Sentences**
	English	**French**	
In-domain	239	267	8.1M
Out-of-domain	1139	1292	35M
MLIA	19	23	1.0M
Mlia-dev	17K	18K	728
Mlia-self	20K	23K	1000

Table: Data sources for the English-French MLIA task (before tokenization)

Pre and post-processing and Tokenisation piplelines

Hugging face(H*)

1. Tokenize the French and English texts using HuggingFace API
2. BPE units are mapped to pre-trained encodings generated according to Devlin et al. (2019) as input to the translation system
3. MT output is a sequence of multilingual BPE units
4. These are reaccentuated and recased, before a final detokenization.

Moses(M*)

1. Tokenize the French and English texts using Moses scripts
2. Compute a joint Byte-pair Encoding (BPE) inventory of 32K units with subword-nmt
3. Generate the translation
4. Detokenize and truecase the output, again with Moses scripts .

Details in (Abdul-Rauf et al., 2020)

Translation framework

Architecture

We mostly used two architectures to build our systems:

- Transformer models (Vaswani et al. (2017))
- BERT-fused Transformer models (Zhu et al. (2020))

All systems use Facebook's seq-2-seq library fairseq (Ott et al. (2019)) with identical hyper-parameters settings for all models (borrowed from `transformer_iwslt_de_en`).

Fine-tuning

Fine-tuning framework

The fine-tuning process starts from corresponding models trained to convergence, based on BLEU score on development set.

The best checkpoint is then further fine-tuned using the fine-tuning corpus (either MLIA or the in-domain corpus as per the experiment).

Systems Submitted

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: `M*` prefixed systems use the Moses tokenisation pipeline, while `B*` use HuggingFace's pipeline.

Constrained System

ID	Name	Detail	Dev	Self-test	BLEU	ChrF
		Constrained system				
M1	trans	mlia	34.4	57.7	43.5	0.660

- Transformer big architecture
- Moses tokenisation with 32K BPE vocabulary units

Unconstrained Systems

ID	Name	Detail	Dev	Self-test	BLEU	ChrF
		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

BioMed BPE

BPE codes learned from all the biomedical data .

$\Rightarrow$ Segmentation of out-of-domain corpus done using BioMed BPE.

- *M2:* All in-domain corpus (sans mlia) fine-tuned on MLIA corpus.
- *M3 & B4:* Initial Parameters learned from huge out-of-domain corpus and later fine-tuned on in-domain corpus containing MLIA (x2).
- *M5:* MLIA corpus segmented with BioMed BPE (contrast with M1).

Systems Submitted

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: `M*` prefixed systems use the Moses tokenisation pipeline, while `B*` use HuggingFace's pipeline.

Observations

- Best model by transfer learning from in-domain corpus
- Training models on significantly fewer in-domain data with domain adapted units results in comparable performance with models that are trained on huge but less domain specific data.

Future Plans

- Multilingual Machine Translation
- Transfer learning

Thank you

Thank you :-)

References I

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

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), doi:10.18653/v1/N19-1423, URL https://www.aclweb.org/anthology/N19-1423

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

References II

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), doi:10.18653/v1/N19-4009, URL https://www.aclweb.org/anthology/N19-4009

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

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