The LIG system for the English-Czech Text Translation Task of IWSLT 2019

Loïc Vial *Benjamin Lecouteux* *Didier Schwab*
Hang Le *Laurent Besacier*

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{loic.vial, benjamin.lecouteux, didier.schwab,
thi-phuong-hang.le, laurent.besacier}@univ-grenoble-alpes.fr

Abstract

In this paper, we present our submission for the English to Czech Text Translation Task of IWSLT 2019. Our system aims to study how pre-trained language models, used as input embeddings, can improve a specialized machine translation system trained on few data.

Therefore, we implemented a Transformer-based encoder-decoder neural system which is able to use the output of a pre-trained language model as input embeddings, and we compared its performance under three configurations: 1) without any pre-trained language model (constrained), 2) using a language model trained on the monolingual parts of the allowed English-Czech data (constrained), and 3) using a language model trained on a large quantity of external monolingual data (unconstrained). We used BERT as external pre-trained language model (configuration 3), and BERT architecture for training our own language model (configuration 2).

Regarding the training data, we trained our MT system on a small quantity of parallel text: one set only consists of the provided MuST-C corpus, and the other set consists of the MuST-C corpus and the News Commentary corpus from WMT.

We observed that using the external pre-trained BERT improves the scores of our system by +0.8 to +1.5 of BLEU on our development set, and +0.97 to +1.94 of BLEU on the test set. However, using our own language model trained only on the allowed parallel data seems to improve the machine translation performances only when the system is trained on the smallest dataset.

1. Introduction

The recent advances in pre-trained Language Models [1, 2, 3, 4, 5] have shown that they could greatly improve many NLP tasks such as Natural Language Understanding, Question Answering, Natural Language Inference, Word Sense Disambiguation, etc.

With our submission, we would like to explore what these models can bring to a typical Transformer-based encoder-decoder Neural Machine Translation system. Unlike the works of [6] and [3] where the authors fine-tune the weights of the language models on the translation task, we propose to use the language models as input embeddings for our neural system.

We expect that the language model, because it is trained on a great quantity of monolingual data, will bring some additional information to a MT system trained on relatively few parallel data.

Therefore, we conducted experiments that compare our system with and without the information from a BERT pre-trained model [2]. In addition, we created our own BERT LM by training it on the allowed training data only, in order to see if the language model is still useful in a constrained setting.

For the training data, we used only the provided MuST-C [7] and News Commentary [8] from WMT, for a total of less than 400k parallel sentences.

2. System Description

2.1. Architecture

Our system relies mostly on the Transformer architecture [9]. More precisely, it consists of the following layers, as pictured in Figure 1:

- The input embeddings layer, which takes words in their vector form from either 1) a classical look-up table trained jointly with the model or 2) a pre-trained language model which remains fixed during the training.
- A linear layer, only if the embeddings come from a pre-trained language model, in order to resize their vectors to the desired size.
- Multiple Transformer encoder layers.
- The output embeddings layer (trained look-up table).
- Multiple Transformer decoder layers.
- A linear layer which resizes the decoder output to the output vocabulary size, followed by a softmax.

We implemented our system using PyTorch[1]. For the Transformer encoder and decoder layers, we used the implementation from OpenNMT[2]. The parameters used are the

[1] https://pytorch.org/
[2] https://github.com/OpenNMT/OpenNMT-py

same as the "base" model of [9]: 6 layers, 8 attention heads and a hidden feed-forward size of 2048, except for the dropout rate that we set to 0.3 to improve the robustness of our model.

It is to be noted that, as in [9], we share the weight matrix between the output embeddings and the last linear layer. However, we do not share the vocabulary nor the matrices between the input and the output languages.

Also, for the input embeddings, if they come from a look-up table, we add sinusoidal positional encoding to the vectors as in [9]. We do not need it when using a language model because the positions are already encoded.

Finally, for the size of the embeddings, which is the same as the input and output of the Transformer layers, we tried two different parameters: 512 and 1024.

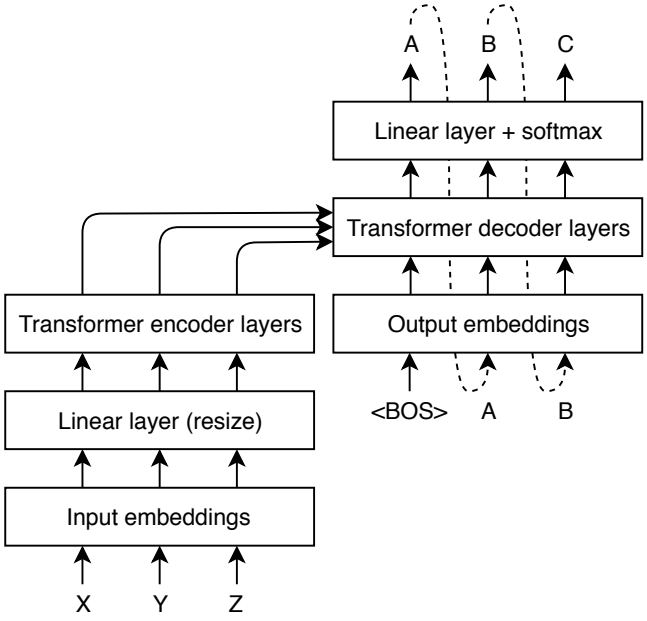

Figure 1: Architecture of our neural MT system.

2.2. Training and development corpora

Due to time constraints, we limited our training to only two English-Czech corpora that we considered of good quality and relevant for the task: the provided MuST-C [7] and the News Commentary corpus provided by WMT [8]. MuST-C is a speech translation corpus of TED talks, similar to the test data of the task, and we added the News Commentary corpus, which consists of political and economic commentaries, because it was the second smallest corpus provided by WMT, after Common Crawl, and we estimated that its quality was better than Common Crawl.

The training data of MuST-C contains 128 179 sentences, and the News Commentary corpus contains 246 513 sentences. We conducted two sets of experiments: one using only the MuST-C, and the other using both corpora, hence cumulating 374 692 training sentences.

For the development set used in both settings, we used the development and test corpora from MuST-C, which corresponds to 3 928 sentences.

2.3. Preprocessing

We preprocessed every corpus using the standard scripts from the Moses repository.[3] In particular, we normalized punctuation characters, removed non-printing characters, and tokenized the data. Finally, we removed sentences with more than 80 words and those with a source-target word ratio greater than 1.5.

Corpus	Original sentence count	Cleaned sentence count
MuST-C	128 179	112 993
MuST-C + News	374 692	344 973
Dev	3 928	3 507

Table 1: Corpora statistics before and after cleaning.

Table 1 summarize the corpus lengths before and after the preprocessing phase.

2.4. Vocabulary

2.4.1. English side

There are tree cases for the input English vocabulary:

1. For the case where we used BERT external pre-trained language model, we use the same vocabulary as their model named "bert-base-cased" which consists of 30 000 subwords.
2. For the case where we trained our own BERT constrained model, we used a BPE vocabulary of size 30 000 trained on all allowed corpora for the task, which consists of MuST-C and 6 other corpora from WMT[4].
3. For the case where we do not use any language model, we trained a BPE vocabulary of size 30 000, but only on MuST-C and News Commentary.

2.4.2. Czech side

For the output Czech vocabulary, we used the same in every configuration: we learned a BPE vocabulary of size 14 000 on the Czech side of the MuST-C and the News Commentary corpora. For BPE learning, we used the tool subword-nmt[5].

2.5. Language model pre-training

In order to both be able to explore how much a pre-trained language model can improve a NMT system, and submit a system constrained in terms of training data, we trained our own language model restricted to the allowed data.

[3] https://github.com/moses-smt/mosesdecoder

[4] http://www.statmt.org/wmt19/translation-task.html

[5] https://github.com/rsennrich/subword-nmt

We used the English side of the corpora listed in Table 2 for the pre-training data, and we extracted 0.5% of the sentences for the validation and test sets (approximately 314 000 sentences). Comparing to the BERT external pre-trained model "bert-base-cased" provided by the authors, which is trained on a corpus set that contains more than 3 billions words, we have 708 622 867 words in total which amounts to approximately 20%.

Corpus	Sentence count	Word count
MuST-C	128 179	2 413 793
News Commentary	246 513	5 168 469
Common Crawl	161 838	3 348 584
Wiki Titles	362 015	897 564
Europarl	654 323	15 628 367
ParaCrawl	2 981 949	48 918 150
CzEng	58 315 645	632 195 134
Total	62 850 462	708 622 867

Table 2: Corpora statistics for the pre-training of the language model $\text{BERT}_{\text{constr}}$.

We used the XLM[6] tool with the Masked Language Model (MLM) objective, and with the following parameters: 6 layers, 8 attention heads and an embeddings size of 512, for a total of 34.78M parameters. In constrast, the original BERT model "bert-base-cased" has 12 layers, 12 attention heads and an embeddings size of 768, for a total of 110M parameters. We chose to reduce these parameters because we had less training data.

For the optimizer, we used Adam, with a learning rate equals to 0.0001, warmup steps=30K, β_1=0.9, β_2=0.999, weight decay=0.01 and ϵ=000001.

We trained for 1016 epochs. The validation/test MLM accuracy was 53.82%/53.97% and the validation/test perplexity was 11.07/10.90.

3. Experiments

3.1. Training process

We trained 9 different systems by making the following parameters vary:

1. The training data: either **MuST-C** or **MuST-C + News Commentary**.
2. The input language model, either **None**, $\textbf{BERT}_{\textbf{extern}}$ (external pretrained model "bert-base-cased") or $\textbf{BERT}_{\textbf{constr}}$ (constrained on allowed data only).
3. The embeddings size, either **512** or **1024** (only 512 when the training data is only MuST-C).

We applied label smoothing with a parameter of 0.1 to the cross entropy criterion (as in [9]). We trained on batches of sentences of size 100, on a single NVIDIA GTX 1080 Ti. We evaluated the quality of our system in terms of BLEU [10] on the development corpus at the end of every epoch, and we kept the best on a total of 250 epochs.

Finally, for the optimizer, we used Adam [11] with a fixed learning rate of 0.0001.

3.2. Results

We evaluated every best system on the development corpus at the end of the training, with beam search applied with a beam size of 12. The results on this development set and on the task's test set are in Table 3.

As we can see, in every case, using the $\text{BERT}_{\text{extern}}$ language model consistently improves the BLEU score comparing to using no language model, or using our $\text{BERT}_{\text{constr}}$ language model, by an absolute value ranging from 0.8 to 1.5 on the development set, and from 0.97 to 1.94 on the test set.

Using our $\text{BERT}_{\text{constr}}$ language model however, decreases the score comparing to using no language model, but only on the MuST-C + News dataset. When training on the MuST-C alone, using our language model adds 0.1 to the BLEU score on the development set, and 0.89 on the test set. We think that this bad performance, compared to $\text{BERT}_{\text{extern}}$, may be explained by one or several factors such as: not having enough training data, a suboptimal choice of hyperparameters or because we stopped the training of the LM too early.

Concerning the training data, having more is generally better, but knowing that the MuST-C only consists of 112 993 sentences, the final score obtained by the systems trained solely on this corpus is still considerable. We can also notice that using the MuST-C alone is where both language models are the more useful,

Finally, using an embeddings size of 1024 instead of 512 on the second dataset is useful and it gives us our best scores, but the difference is not really high (+0.2 on the dev set and +0.16 on the test set, when using $\text{Bert}_{\text{extern}}$).

3.3. Submission

For our submission, we provided the output of our 9 systems on the test set, with the same beam size of 12, but we added an extra detokenization step at the end using the script `detokenizer.perl` provided by Moses.

Due to a lack of time, we stopped the training of some systems on less than 250 epochs. In the case where we use MuSTC + News as training data and with the $\text{BERT}_{\text{constr}}$ language model, we stopped the training at epoch 68 with the embeddings size of 512, and epoch 55 with the embeddings size of 1024. The BLEU score on the development corpus obtained by these models, after the training complete, are respectively 22.4 (instead of 20.4) and 22.7 (instead of 21.7).

We submitted our best constrained system as our primary submission (the one obtaining 23.1 BLEU on the dev set) and all the others as constrastive.

[6] https://github.com/facebookresearch/XLM

Training data	Input LM	Embeddings size	BLEU (Dev)	BLEU (Test)	TER	BEER	Charac-TER	BLEU (ci)	TER (ci)
MuST-C	None	512	20.4	19.13	61.55	51.64	52.23	19.77	60.54
MuST-C	$\text{BERT}_{\text{constr}}$	512	20.5	20.02	60.64	51.78	51.26	20.68	59.53
MuST-C	$\text{BERT}_{\text{extern}}$	512	21.9	21.07	59.19	52.74	49.97	21.78	58.15
MuST-C + News	None	512	22.8	22.26	58.68	53.84	48.29	22.93	57.63
MuST-C + News	$\text{BERT}_{\text{constr}}$	512	20.4	20.09	60.90	51.91	50.66	20.77	59.79
MuST-C + News	$\text{BERT}_{\text{extern}}$	512	23.7	23.53	57.55	54.36	47.30	24.27	56.41
MuST-C + News	None	1024	23.1	**22.72**	58.51	53.94	48.27	23.47	57.41
MuST-C + News	$\text{BERT}_{\text{constr}}$	1024	21.6	21.35	59.68	52.83	49.60	22.05	58.52
MuST-C + News	$\text{BERT}_{\text{extern}}$	1024	23.9	**23.69**	57.14	54.58	47.06	24.41	56.02

Table 3: Results of our systems on the development and the test set after 250 epochs (except in two cases, see subsection 3.3). Beam size is 12. Data are tokenized and cased.

4. Conclusion

In our submission for the English-Czech Text Translation Task, we submitted a neural MT system based on the Transformer architecture, and we studied the impact of a pre-trained language model used as input embeddings.

We experimented on two training sets: one which consists of a specialized speech translation corpus only, and the other which includes also a news commentary corpus. We compared the performance of an external BERT model provided by the original authors (in unconstrained settings) and a constrained BERT model that we trained ourselves on the allowed data only.

Our results showed that our model really benefits from the external BERT model trained on more than 3 billions words, really improving the quality of the translation in every case, but our constrained BERT model trained on less than 1 billion words does not always give a useful information to the MT system. However, we believe that this could be due to a suboptimal choice of hyperparameters (different embeddings size, optimizer, etc.) or because we stopped the training too early.

5. References

[1] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. [Online]. Available: https://www.aclweb.org/anthology/N18-1202

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[3] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] P. Ramachandran, P. Liu, and Q. Le, "Unsupervised pretraining for sequence to sequence learning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 383–391. [Online]. Available: https://www.aclweb.org/anthology/D17-1039

[7] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)", year = 2019*, Minneapolis, MN, USA, 2019.

[8] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck,

P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri, "Findings of the 2019 conference on machine translation (wmt19)," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 1–61. [Online]. Available: http://www.aclweb.org/anthology/W19-5301

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

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference for Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980