Pretrained Language Models and Backtranslation for English-Basque Biomedical Neural Machine Translation

Inigo Jauregi Unanue
University of Technology Sydney
RoZetta Technology
inigo.jauregi
@rozettatechnology.com

Massimo Piccardi
University of Technology Sydney
massimo.piccardi@uts.edu.au

Abstract

This paper describes the machine translation systems proposed by the University of Technology Sydney Natural Language Processing (UTS_NLP) team for the WMT20 English-Basque biomedical translation tasks. Due to the limited parallel corpora available, we have opted to train a BERT-fused NMT model that leverages the use of pretrained language models. Furthermore, we have augmented the training corpus by backtranslating monolingual data. Our experiments show that NMT models in low-resource scenarios can benefit from combining these two training techniques, with improvements of up to 6.16 BLEU percentage points in the case of biomedical abstract translations.

1 Introduction

Nowadays, most of the literature and scientific terminology produced in the biomedical field is in English, which limits the access to this information by non-English speaking researchers, doctors and patients. Thus, it would be very useful to avail of machine translation systems that can effectively translate this information into other languages, so that more people can be able to access it and benefit from it.

However, many of the world languages lack sufficient parallel corpora to properly train machine translation systems in this domain. State-of-the-art neural machine translation (NMT) models (Sutskever et al., 2014; Bahdanau et al., 2015) suffer from overfitting when trained on insufficient data, and thus fail to generate accurate translations (Koehn and Knowles, 2017).

In this paper we address this problem for a low-resource language, Basque. We have taken part in the WMT20 Biomedical Translation challenge, which has released two interesting shared tasks involving this language, namely, the English-to-Basque translation of biomedical article abstracts and the English-to-Basque translation of medical terminology. In order to overcome the issue of having limited supervised training data, we have decided to apply two promising ideas proposed in the literature. First, we have applied transfer learning by training a *BERT-fused* NMT model (Zhu et al., 2020) that uses source-language contextual embeddings inferred by a pretrained language model (LM) as additional input features, both in the encoder and in the decoder. Second, we have augmented the training corpus using backtranslation (Sennrich et al., 2016; Burlot and Yvon, 2018). For this, a BERT-fused NMT model has been trained in the opposite translation direction (Basque → English) to translate sentences from large monolingual corpora (e.g. Wikipedia, medical texts).

The experiments have shown that an NMT baseline can greatly benefit from combining these training techniques. The three best performing systems in both tasks (terminology and abstracts) have been submitted to the WMT20 biomedical translation shared task under the UTS_NLP team name.

2 Related Work

2.1 Pretrained LMs

Pretrained LMs have been one of the most remarkable advancements in transfer learning for NLP in recent years. They are large neural networks that are trained over massive datasets (millions of sentences) in an unsupervised manner, and can effectively learn the regularities/patterns of a language. Then, such general networks can be applied to efficiently train smaller networks for downstream tasks, using much smaller annotated datasets. ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2018) are some examples of pretrained LMs that have achieved state-of-the-art results in various natural language understanding tasks such as, among others, sentiment analysis, paraphrase detection, and question answering.

In NMT, various recent works have proposed incorporating pretrained LMs into the standard encoder-decoder architecture. Lample and Conneau (2019) have proposed pretraining an LM with a novel "cross-lingual LM objective" that uses parallel training data to predict masked words of either language. Edunov et al. (2019) have replaced standard word embeddings with contextual word embeddings learned by a pretrained LM. Their experiments have showed that contextual embeddings are more effective in the encoder and when fixed ("ELMo-style augmentation"). Conversely, Clinchant et al. (2019) have initialized the encoder of an NMT model with the weights of a pretrained LM network, and fine-tuned them ("GPT-2/BERT-style augmentation"), reaching similar performance, but faster convergence. Yang et al. (2020) have added the contextual embeddings of a pretrained LM as additional input features to the standard embeddings, and incorporated a dynamic switch to let the model learn how to weigh each input. Finally, the BERT-fused NMT model (Zhu et al., 2020) follows a similar idea, by adapting the architecture of the transformer network in order to have an extra self-attention layer that learns to weigh the contextual embeddings of the LM. This attention-layer is seamlessly added to both the encoder and the decoder. Given the improvement in performance achieved by the BERT-fused NMT on several datasets and the well-supported code by the authors (built on top of *fairseq*), we have decided to adopt this model in our experiments.

2.2 Backtranslation

In NMT, backtranslation (Sennrich et al., 2016; Burlot and Yvon, 2018) has become a common approach to alleviate the problem of having limited parallel data for training. It consists of first training a *target* → *source* NMT model with the available parallel corpus. Then, this model is used to translate a large number of sentences from monolingual corpora in the target language, which are usually more available than parallel corpora, to the source language. The resulting "silver corpus" is used as additional training data for the *source* → *target* NMT model, and can often help to boost the fluency of the generated translations.

However, the most effective way of using backtranslation is still an open research question. Poncelas et al. (2018) have explored different combinations of backtranslated and human-translated datasets, and have found that in their low-resource scenario a 2:1 backtranslated-to-human-translated sentences ratio is optimal; beyond that, increasing the size of the backtranslated data deteriorates the performance. Edunov et al. (2018) have shown that backtranslating by either sampling from the model or adding noise to a standard beam search can improve the final translation accuracy substantially. Burlot and Yvon (2018) have generated more natural *pseudo-source* sentences by training a generative adversarial network (GAN). Finally, Soto et al. (2020) have found that combining backtranslated data from different sources (i.e. out-of-domain data, in-domain data) and different models (i.e. rule-based, SMT, NMT) can also improve the accuracy of the final translations. In our work, we have explored using a BERT-fused NMT model for backtranslating monolingual data, expecting that the transfer learning achieved from using pretrained LM in Basque will produce better quality pseudo-source sentences.

3 Resources

All the experiments have been carried out using only the parallel and monolingual training data recommended by the organisers on the shared task website. Table 1 summarizes all the data used in our experiments.

3.1 Parallel Data

The medical terminology translation task consists on translating ICD-10 (International Clasification of Diseases) code descriptions from English to Basque. The descriptions are relatively short sentences (8 tokens on average). The organizers have provided an in-domain parallel corpus for training and validation. The blind test set contains 2,000 sentences in English.

The abstract translation task involves translating sentences of abstracts from biomedical scientific research papers. The sentence in this task are longer compared to those in the terminology task (24 tokens on average). However, for this task the organizers have not provided any in-domain parallel data for training or validation, only 375 English sentences for blind testing. Consequently, we have decided to form a small validation set from

	train	dev	test
in-domain (IO)			
ICD-10	25,900	2,000	2,000
abstracts	-	50	375
out-of-domain (OOD)			
EhuHac	550,000	-	-
QED	16,000	-	-
TED talks	5,623	-	-

(a) Parallel data.

general	
wikipedia	1.5M
biomedical	
wikipedia biomedical	8,000
hospital notes	2,000
Snomed CT	50,000

(b) Monolingual data in Basque.

Table 1: Number of sentences in each dataset.

the English-Italian biomedical abstract translation dataset provided on the website. We have selected 50 sentences in English from that dataset and translated them manually into Basque. In this way, we have managed to assemble a small, yet high-quality validation dataset in a domain similar to that of the actual task, and used it for selecting the best models for submission.

Finally, we have used out-of-domain (OOD) parallel corpora to compensate for the lack of in-domain training data. From the data provided by the organizers, we have used: the *EhuHac* dataset, which consists of translations of 136 fiction books; the *QED* dataset, which are translations of subtitles for educational videos and lectures; and the *TED talks* dataset, containing transcripts of TED talk videos.

3.2 Monolingual Data

The monolingual data have been grouped in two categories. First, we have the *general* domain texts, which include 1.5M sentences from the Basque Wikipedia. Second, we have the *biomedical* domain texts, which include a group of medical articles from the Basque Wikipedia and hospital notes written by doctors. We have applied backtranslation to generate pseudo-parallel datasets from these monolingual data. A BERT-fused NMT model has been trained over the available OOD parallel data in the reverse translation direction, and applied to translate Basque sentences to English. For more details on the training of the BERT-fused NMT model, please see Section 4.

Additionally, we have included as part of the biomedical monolingual data the subset of Basque SNOMED CT terms provided by the organizers, which have been automatically translated from English using a rule-based machine translation system. Using the IDs of the terms, we have been able to match them with the original English terms and include them as additional training data.

3.3 Pretrained BERT Models

We have explored using several different pretrained BERT LMs to include them in our BERT-fused NMT model. The pretrained BERT models have been downloaded from Hugging Face[1] :

- **bert-base-uncased** : Original BERT LM model proposed by Devlin et al. (2019). Pretrained on the BookCorpus (800M words) (Zhu et al., 2015) and the English Wikipedia (2,500M words).

- **bert-pubmed**: Pretrained over biomedical articles and journals collected from PubMed. There is no clear description of the amount of data used for pretraining. Hugging Face model name: `monologg/biobert_v1.1_pubmed`.

- **bert-mimic-pubmed**: Pretrained over biomedical articles and journals collected from PubMed and electronic health records of intensive care unit patients from MIMIC-III (Johnson et al., 2016). There is no clear description of the amount of data used for pretraining. Hugging Face model name: `adamlin/NCBI_BERT_pubmed_mimic_uncased_base_transformers`

- **bert-discharge-summaries**: Pretrained model proposed by Alsentzer et al. (2019) trained on all discharge summaries from MIMIC-III. Hugging Face model name: `emilyalsentzer/Bio_Discharge_Summary_BERT`.

Pretrained LM for backtranslation:

- **berteus-base-cased**: Pretrained LM in Basque (Agerri et al., 2020) trained over the Basque Media Corpus (BMC) (224M words). Hugging Face model name: `ixa-ehu/berteus-base-cased`.

[1] https://huggingface.co/models

4 Training and Hyperparameter Tuning

We have trained a BERT-fused NMT model (Zhu et al., 2020) with the open-source code provided by the authors[2], which is built on top of *fairseq*[3]. Following the authors recommendation, as a *warmup* step, first a standard transformer-based (Vaswani et al., 2017) NMT model has been trained over the training data. Then, this model has been used as both a baseline and to initialize the weights that the BERT-fused NMT model has in common. We have used the `transformer_iwslt_de_en` architecture as the NMT model, which consists of a 6-layer transformer network as the encoder and the decoder, with the embedding dimension set to 512 and the hidden layer dimension to 1024. Additionally, we have used the following training hyperparameters: dropout 0.1, label-smoothing 0.1, `inverse_sqrt` learning scheduler, warmup updates $4,000$, warmup initialization learning rate $1e^{-7}$, minimum learning rate $1e^{-9}$, weight decay 0.0001, BERT encoder dropout 0.5 and the Adam optimizer (Kingma and Ba, 2015). The learning rate $[0.0002, 0.00002]$ and the number of tokens per batch $[1024, 4048]$ have been tuned using the validation set. All the datasets have been lowercased and tokenized using the *moses tokenizer*. Additionally, we have learned subword units using Byte Pair Encoding (BPE) (Sennrich et al., 2015) with $10,000$ merge operations in order to reduce the vocabulary size and handle unknown words.

In the terminology translation task we have only used the in-domain ICD-10 code description data to train our models, because adding additional OOD parallel data or backtranslated data was degrading the performance of the model. This is probably likely due to the specific and structured language used in the code descriptions, which is very different from the rest of the available texts. The baseline NMT was warmed up for 50 epochs and the best model over the validation was selected. Then, the BERT-fused models was tuned for 10 more epochs.

In the abstract translation task, due to the fact that in-domain parallel data were not available, we have explored training the model with different combinations of the OOD parallel data, the ICD-10 training data and the backtranslated data. The ICD-10 data and the backtranslated biomedical data have been upsampled x5 and x10, respectively. In this task, the baseline NMT was warmed up for 30 epochs, as the training data are much larger (longer training times) and because we have seen no noticeable improvement after the 30th epoch. Like in the previous task, the BERT-fused models have been tuned for another 10 epochs over the best baseline model.

Evaluation of the models has been carried out using the standard BLEU metric (Papineni et al., 2002). In the case of the terminology translation task, we have also used a case-insensitive strict accuracy metric, in which an ICD-10 code description is considered correct only if it is a complete string match with the reference (no partial scores).

5 Results

5.1 Terminology Translation

In the terminology translation task (Table 3a) all the models have achieved high numerical results over the validation set ($> 73\%$ accuracy and > 88.7 BLEU). In terms of translation scores, one could say that this was an easy task and that it is almost solved. However, we would like to argue that this is not the case. Compared to other translation tasks (e.g. abstracts, news, TED talks), the space of correct translations is much smaller in the ICD-10 task since even a single-word mistake (e.g. *abscess of bursa, **right** shoulder* VS *abscess of bursa, **left** shoulder*) may result in a misunderstanding with serious consequences. Therefore, there is still margin for improvement.

On the other hand, we have observed that the BERT-fused NMT models have consistently outperformed the baseline, on average by $+1.61$ percentage points (pp) of accuracy and by $+0.7$ pp of BLEU. All pretrained BERT LMs have achieved comparable results, yet surprisingly the *bert-base-uncased* model has proved the best, despite being the only LM that had not been pretrained on biomedical data.

5.2 Abstract Translation

In the abstract translation task (Table 3b) the overall performance of the models in terms of BLEU scores has been considerably lower. This is understandable, as we did not have access to any in-domain parallel data for training. The baseline model using only the OOD parallel data has achieved an 8.67 BLEU score.

Nevertheless, we have been able to improve this result by applying our backtranslation and pretrained LMs. Just adding the backtranslated sen-

[2] https://github.com/bert-nmt/bert-nmt

[3] https://github.com/pytorch/fairseq

Training Data	**Models**									
	baseline		bert-base-uncased		bert-pubmed		bert-mimic-pubmed		bert-discharge-summaries	
	Accuracy	BLEU	Accuracy	BLEU	Accuracy	BLEU	Accuracy	BLEU	Accuracy	BLEU
ICD-10 train	73.15	88.70	**74.93**	**89.49**	74.60	89.39	74.67	89.31	74.78	89.42

(a) Terminology translation. Average results of 3 independent runs.

Training Data	**Models**				
	baseline	bert-base-uncased	bert-pubmed	bert-mimic-pubmed	bert-discharge-summaries
OOD Parallel	8.67	9.36	9.92	9.84	9.55
+ backtranslated general	11.57	14.34	13.71	**14.83**	13.97
+ backtranslated biomedical and ICD-10 train/dev	13.91	14.25	12.39	13.42	11.87

(b) Abstract translation. Average results of 3 independent runs.

Table 2: Results over the validation sets.

tences from Wikipedia (*backtranslated general* in Table 2) to the training set has improved the baseline by $+1.90$ BLEU pp, and adding the biomedical domain backtranslated sentences (*backtranslated biomedical*) and the ICD-10 parallel data has achieved a further $+2.41$ BLEU pp.

Additionally, we have observed comparable improvements with the BERT-fused NMT models, which have again consistently outperformed the baseline. In this case, the best performing pretrained LM has been the *bert-mimic-pubmed* model (14.83 BLEU) trained using only the *backtranslated general* data. This model has achieved an incremental improvement of $+6.16$ BLEU pp with respect to the baseline trained only over the OOD parallel data. It is interesting to see that adding the *backtranslated biomedical* data to the BERT-fused NMT model has not resulted in any improvement, probably because the data are more "noisy" (not grammatically well-structured sentences) and have fewer samples than the *general* backtranslations ($\sim 60,000$ vs 1.5 M).

5.3 Results over the Blind Test Sets

Table 3 shows the results achieved by our best performing models over the test sets. The translations made by our models have been submitted "blindly" and the results have been computed by the organizers. Our proposed runs for the terminology translation task have performed similarly to the validation set, achieving over 73% accuracy. On the contrary, the systems submitted to the abstract translation task have underperformed compared to the results in the validation set. We speculate this is likely due to the domain differences between our validation data and the test data. Even though both datasets are composed of translations of biomedical abstracts, they are probably coming from different databases and may have significantly different writing styles.

Model	Accuracy
bert-mimic-pubmed (run 1)	73.00
bert-discharge-summaries (run 2)	73.00
bert-base-uncased (run 3)	73.00

(a) Terminology translation.

Model	BLEU
bert-mimic-pubmed (run 1)	5.30
bert-pubmed (run 2)	**5.49**
baseline (run 3)	5.28

(b) Abstract translation.

Table 3: Official results over the blind test sets.

6 Conclusion

This work has described the translation systems submitted by the UTS_NLP team to the WMT20 Biomedical Translation shared task. The proposed systems are BERT-fused NMT models trained on a combination of in-domain parallel data, out-of-domain parallel data and backtranslations of monolingual data. The experiments have shown that combining pretrained BERT LMs and backtranslations during training has contributed to achieve considerable accuracy improvements with respect to a standard transformer-based NMT model trained only on the parallel data. Nevertheless, the official test results have shown that the performance of the systems can significantly drop if the translation domain is different. Therefore, there is still significant work to do to improve the domain adaption of these models.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the Language Resources and Evaluation Conference*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the Clinical Natural Language Processing Workshop*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation*.

Stéphane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of bert for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference for Learning Representations*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proceedings of the Conference on Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

A Poncelas, D Shterionov, A Way, GM de Buy Wenniger, and P Passban. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/ language understanding paper. pdf*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *Proceedings of the Association for Computational Linguistics*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the Association for Computational Linguistics*.

Xabier Soto, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of the Association for Computational Linguistics*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. 2020. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. In *Proceedings of the International Conference on Learning Representations*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.