On the use of BERT for Neural Machine Translation

Stéphane Clinchant
NAVER LABS Europe, France
stephane.clinchant@naverlabs.com

Kweon Woo Jung
NAVER Corp.,
South Korea
kweonwoo.jung@navercorp.com

Vassilina Nikoulina
NAVER LABS Europe, France
vassilina.nikoulina@naverlabs.com

Abstract

Exploiting large pretrained models for various NMT tasks have gained a lot of visibility recently. In this work we study how BERT pretrained models could be exploited for supervised Neural Machine Translation. We compare various ways to integrate pretrained BERT model with NMT model and study the impact of the monolingual data used for BERT training on the final translation quality. We use WMT-14 English-German, IWSLT15 English-German and IWSLT14 English-Russian datasets for these experiments. In addition to standard task test set evaluation, we perform evaluation on out-of-domain test sets and noise injected test sets, in order to assess how BERT pretrained representations affect model robustness.

1 Introduction

Pretrained Language Models (LM) such as ELMO and BERT (Peters et al., 2018; Devlin et al., 2018) have turned out to significantly improve the quality of several Natural Language Processing (NLP) tasks by transferring the prior knowledge learned from data-rich monolingual corpora to data-poor NLP tasks such as question answering, bio-medical information extraction and standard benchmarks (Wang et al., 2018; Lee et al., 2019). In addition, it was shown that these representations contain syntactic and semantic information in different layers of the network (Tenney et al., 2019). Therefore, using such pretrained LMs for Neural Machine Translation (NMT) is appealing, and has been recently tried by several people (Lample and Conneau, 2019; Edunov et al., 2019; Song et al., 2019).

Unfortunately, the results of the above-mentioned works are not directly comparable to each other as they used different methods, datasets and tasks. Furthermore, pretrained LMs have mostly shown improvements in low-resource or unsupervised NMT settings, and has been little studied in standard supervised scenario with reasonable amount of data available.

Current state of the art NMT models rely on the Transformer model (Vaswani et al., 2017), a feed-forward network relying on attention mechanism, which has surpassed prior state of the art architecture based on recurrent neural nets (Bahdanau et al., 2014; Sutskever et al., 2014). Beyond machine translation, the transformer models have been reused to learn bi-directional language models on large text corpora. The BERT model (Devlin et al., 2018) consists in a transformer model aiming at solving a masked language modelling task, namely correctly predicting a masked word from its context, and a next sentence prediction task to decide whether two sentences are consecutive or not. In this work, we study how pretrained BERT models can be exploited for *transformer-based* NMT, thus exploiting the fact that they rely on the same architecture.

The objective of this work is twofold. On one hand, we wish to perform systematic comparisons of different BERT+NMT architectures for standard supervised NMT. In addition, we argue that the benefits of using pretrained representations has been overlooked in previous studies and should be assessed beyond BLEU scores on in-domain datasets. In fact, LMs trained on huge datasets have the potentials of being more robust in general and improve the performance for domain adaptation in MT.

In this study, we compare different ways to train and reuse BERT for NMT. For instance, we show that BERT can be trained only with a masked LM task on the NMT source corpora and yield significant improvement over the baseline. In addition, the models robustness is analyzed thanks to synthetic noise.

The paper is organized as follows. In section 2, we review relevant state of the art. Section 3 enumerates different models we experiment with. Finally section 4 and 5 present our results before discussing the main contributions of this work in section 6.

2 Related Works

The seminal work of (Bengio et al., 2003; Collobert and Weston, 2008) were one of the first to show that neural nets could learn word representations useful in a variety of NLP tasks, paving the way for the word embedding era thanks to *word2vec* (Mikolov et al., 2013) and its variants (Pennington et al., 2014; Levy and Goldberg, 2014).

With the recent advances and boost in performance of neural nets, ELMO (Peters et al., 2018) employed a Bi-LSTM network for language modelling and proposed to combine the different network layers to obtain effective word representations. Shortly after the publication of ELMO, the BERT model (Devlin et al., 2018) was shown to have outstanding performance in various NLP tasks. Furthermore, the BERT model was refined in (Baevski et al., 2019) where the transformer self-attention mechanism is replaced by two directional self-attention blocks: a left-to-right and right-to-left blocks are combined to predict the masked tokens.

With respect to NMT, backtranslation (Sennrich et al., 2016a) is up to now one of the most effective ways to exploit large monolingual data. However, backtranslation has the drawback of being only applicable for target language data augmentation, while pretrained LMs can be used both for source and target language (independently (Edunov et al., 2019) or jointly (Lample and Conneau, 2019; Song et al., 2019)).

Lample and Conneau (2019) initializes the entire encoder and decoder with a pretrained MaskLM or Crosslingual MaskLM language models trained on multilingual corpora. Such initialization proved to be beneficial for unsupervised machine translation, but also for English-Romanian supervised MT, bringing additional improvements over standard backtranslation with MLM initialization.

Edunov et al. (2019) uses ELMO (Peters et al., 2018) language model to set the word embeddings layer in NMT model. In addition, the ELMO embedding are compared with the cloze-style BERT (Baevski et al., 2019) ones. The embedding network parameters are then either fixed, or fine-tuned. This work shows improvements on English-German and English-Turkish translation tasks when using pretrained language model for source word embedding initialization. However, the results are less clear when reusing embedding on the target language side.

Futhermore, Song et al. (2019) goes one step further and proposes Masked Sequence-to-Sequence pretraining method. Rather than masking a single token, it masks a sequence of token in the encoder and recovers them in the decoder. This model has shown new state of the art for unsupervised machine translation.

Our work is an attempt to perform systematic comparison on some of the aforementioned architectures that incorporate pretrained LM in NMT model, concentrating on BERT pretrained LM representations applied on supervised machine translation. However, we restrict ourselves to encoder part only, and leave the decoder initialization for future work.

Regarding robustness, several recent studies (Karpukhin et al., 2019; Vaibhav et al., 2019) have tackled robustness issues with data augmentation. In this work, we study whether the robustness problem can be addressed at the model level rather than at data level. Michel et al. (2019) address robustness problem with generative adversarial networks. This method, as well as data augmentation methods are complementary to our work and we believe that they address different issues of robustness.

3 Methods

Typical NMT model adopts the encoder-decoder architecture where the encoder forms contextualized word embedding from a source sentence and the decoder generates a target translation from left to right.

Pretrained LM, namely BERT, can inject prior knowledge on the encoder part of NMT, providing rich contextualized word embedding learned from large monolingual corpus. Moreover, pretrained LMs can be trained once, and reused for different language pairs[1].

[1] As opposed to backtranslation techniques which requires full NMT model retraining

In this study, we focus on reusing BERT models for the NMT encoder[2]. We will compare the following models:

- **Baseline:**. A *transformer-big* model with shared decoder input-output embedding parameters.
- **Embedding (Emb):** The baseline model where the embedding layer is replaced by the BERT parameters (thus having $6+6$ encoder layers). The model is then fine tuned similar to the ELMO setting from (Edunov et al., 2019)
- **Fine-Tuning (FT):** The baseline model with the encoder initialized by the BERT parameters as in Lample and Conneau (2019)
- **Freeze:** The baseline model with the encoder initialized by the BERT parameters and frozen. This means that the whole encoder has been trained in purely monolingual settings, and only parameters responsible for the translation belong to the attention and decoder models.

We exploit the fact that BERT uses the same architecture as NMT encoder which allows us to initialize NMT encoder with BERT pretrained parameters. BERT pretraining has two advantages over NMT training:

- it solves a simpler (monolingual) task of 'source sentence encoding', compared to NMT (bilingual task) which has to 'encode source sentence information', and 'translate into a target language'.
- it has a possibility to exploit much larger data, while NMT encoder is limited to source side of parallel corpus only.

Even though the role of NMT encoder may go beyond source sentence encoding (nothing prevents the model from encoding 'translation related' information at the encoder level), better initialization of encoder with BERT pretrained LM allows for faster NMT learning. Comparing settings where we freeze BERT parameters against fine-tuning BERT allows to shed some light on the capacity of the encoder/decoder model to learn 'translation-related' information.

[2]Similar approach can be applied on the target language but we leave it for future work.

Moreover, since the BERT models are trained to predict missing tokens from their context, their representations may also be more robust to missing tokens or noisy inputs. We perform extensive robustness study at section 4 verifying this hypothesis.

Finally, language models trained on huge datasets have the potentials of being more robust in general and improve the performance for domain adaptation in MT. We therefore compare BERT models trained on different datasets, and perform evaluation on related test sets in order to assess the capacity of pretrained LMs on domain adaptation.

4 WMT experiments

4.1 Preprocessing

We learn BPE (Sennrich et al., 2016b) model with 32K split operations on the concatenation of Wiki and News corpus. This model is used both for Pretrained LM subwords splitting and NMT source (English) side subwords splitting. German side of NMT has been processed with 32K BPE model learnt on target part of parallel corpus only. Please note, that this is different from standard settings for WMT En-De experiments, which usually uses joint BPE learning and shared source-target embeddings. We do not adopt standard settings since it contradicts our original motivation for using pretrained LM: English LM is learnt once and reused for different language pairs.

4.2 Training

BERT For pretraining BERT models, we use three different monolingual corpora of different sizes and different domains. Table 1 summarizes the statistics of these three monolingual corpora.

- *NMT-src*: source part of our parallel corpus that is used for NMT model training.
- *Wiki*: English wikipedia dump
- *News*: concatenation of 70M samples from "News Discussion", "News Crawl" and "Common Crawl" English monolingual datasets distributed by WMT-2019 shared task[3]. This resulted in total 210M samples.

The motivation of using NMT-src is to test whether the resulting NMT model is more robust

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

	Lines	Tokens
NMT-src	4.5M	104M
Wiki	72M	2086M
News	210M	3657M

Table 1: Monolingual (English) training data

after having being trained on the source corpora. The Wiki corpora is bigger than the NMT-src but could be classified as out-of-domain compared to news dataset. Finally, the news dataset is the biggest one and consists mostly of in-domain data.

In all of our experiments, we only consider using the masked LM task for BERT as the next sentence prediction tasks put restrictions on possible data to use. We closely follow the masked LM task described in (Devlin et al., 2018) with few adjustments optimized for downstream NMT training. We use frequency based sampling (Lample and Conneau, 2019) in choosing 15% of tokens to mask, instead of uniformly sampling. Instead of MASK token we used UNK token hoping that thus trained model will learn certain representation for unknowns that could be exploited by NMT model. Warm-up learning scheme described in (Vaswani et al., 2017) results in faster convergence than linear decaying learning rate. The batch size of 64000 tokens per batch is used, with maximum token length of 250, half the original value, as we input single sentence only. We do not use [CLS] token in the encoder side, as attention mechanism in NMT task can extract necessary information from token-level representations. The BERT model is equivalent to the encoder side of Transformer Big model. We train BERT model up to 200k iterations until the accuracy for masked LM on development saturates.

NMT For NMT system training, we use WMT-14 English-German dataset.

We use *Transformer-Big* as our baseline model. We share input embedding and output embedding parameters just before softmax on the decoder side. Warm up learning scheme is used with warm-up steps of 4000. We use batch size of 32000 tokens per batch. Dropout of 0.3 is applied to residual connections, and no dropout is applied in attention layers. We decode with beam size 4 with length penalty described in Wu et al. (2016). We conduct model selection with perplexity on development set. We average 5 checkpoints around lowest perplexity.

	Lines	Tok/line (en/de)
news14	3003	19.7/18.3
news18	2997	19.5/18.3
iwslt15	2385	16.4/15.4
OpenSub	5000	6.3/5.5
KDE	5000	8/7.7
wiki	5000	17.7/15.5

Table 2: In/Out of Domain test sets. news14 and news18 are test sets from WMT-14 and WMT-18 news translation shared task. iwslt: test set from IWSLT-15 MT Track[4]. Wiki is randomly 5K sampled from parallel Wikipedia distributed by OPUS[5], OpenSub, KDE and Wiki are randomly 5K sampled from parallel Wikipedia, Open Subtitles and KDE corpora distributed by OPUS[6]

4.3 Evaluation

We believe that the impact of pretrained LM in NMT model can not be measured by BLEU performance on in-domain test set only. Therefore we introduce additional evaluation that allows to measure the impact of LM pretraining on different out-of-domain tests. We also propose an evaluation procedure to evaluate the robustness to various types of noise for our models.

Domain Besides standard WMT-14 news test set, models are evaluated on additional test sets given by Table 2. We include two in-domain (news) test sets, as well as additional out-of-domain test sets described in Table 2.

Noise robustness. For robustness evaluation, we introduce different type of noise to the standard *news14* test set:

Typos: Similar to Karpukhin et al. (2019), we add synthetic noise to the test set by randomly (1) swapping characters (chswap), (2) randomly inserting or deleting characters (chrand), (3) uppercasing words (*up*). These test sets translations are evaluated against the golden *news14* reference.

Unk: An unknown character is introduced at the beginning (noted *UNK.S*) or at the end of the sentence (noted *UNK.E*) before a punctuation symbol if any (this unknown character could be thought as as an unknown emoji, a character in different script, a rare unicode character). This token is introduced both for source and target sentence, and the evaluation is performed with the augmented-reference.

Intuitively, we expect the model to simply copy UNK token and proceed to the remaining tokens.

Interestingly, this simple test seems to produce poor translations, therefore puzzling the attention and decoding process a lot. Table 3 gives an example of such translations for baseline model[7].

Since the tasks are correlated, a better model might be better on noisy test sets as it behaves better in general. If we want to test that some models are indeed better, we need to disentangle this effect and show that the gain in performance is not just a random effect. A proper way would be to compute the BLEU correlation between the original test set and the noisy versions but it would require a larger set of models for an accurate correlation estimation.

Δ(chrF) : We propose to look at the *distribution of the difference of sentence charf between the noisy test set and the original test set*. Indeed, looking at BLEU delta may not provide enough information since it is corpus-level metric. Ideally, we would like to measure a number of sentences or a margin for which we observe an 'important decrease' in translation quality. According to Ma et al. (2018); Bojar et al. (2017), sentence level chrF achieves good correlation with human judgements for En-De news translations.

More formally, let s be a sentence from the standard news14 test set, n a noise operation, m a translation model and r the reference sentence[8]:

$$\Delta(\text{chrF})(m, n, s) = \text{chrF}(m(n(s)), r) - \text{chrF}(m(s), r) \quad (1)$$

In the analysis, we will report the distribution of **Δ(chrF)** and its mean value as a summary. If a model is good at dealing with noise, then the produced sentence will be similar to the one produced by the noise-free input sentence. Therefore, the Δ(chrF) will be closer to zero.

4.4 Results

Table 4 presents the results of our experiments. As expected, freezing the encoder with BERT parameters lead to a significant decrease in translation quality. However, other BERT+NMT architectures mostly improve over the baseline both on in-domain and out-of-domain test sets. We conclude, that the information encoded by BERT is useful but not sufficient to perform the translation task. We believe, that the role of the NMT encoder is to encode both information specific to source sentence, but also information specific to the target sentence (which is missing in BERT training).

Next, we observe that even NMTSrc.FT (NMT encoder is initialized with BERT trained on source part of parallel corpus) improves over the baseline. Note that this model uses the same amount of data as the baseline. BERT task is simpler compared to the task of the NMT encoder, but it is still related, thus BERT pretraining allows for a better initialization point for NMT model.

When using more data for BERT training (Wiki.FT and News.FT), we gain even more improvements over the baseline.

Finally, we observe comparable results for News.Emb and News.FT (the difference in BLEU doesn't exceed 0.3 points, being higher for News.FT on in-domain tests, and News.Emb for out-of-domain tests). Although News.FT configuration keeps the size of the model same as standard NMT system, News.Emb adds BERT parameters to NMT parameters which doubles the size of NMT encoder. Additional encoder layers introduced in News.Emb does not add significant value.

4.5 Robustness analysis

Table 5 reports BLEU scores for the noisy test sets (described in section 4.3). As expected, we observe an important drop in BLEU scores due to the introduced noise. We observe that most pretrained BERT models have better BLEU scores compared to baseline for all type of noise (except NMTSrc.FT which suffers more from unknown token introduction in the end of the sentence compared to the Baseline). However, these results are not enough to conclude, whether higher BLEU scores of BERT-augmented models are due to better robustness, or simply because these models are slightly better than the baseline in general.

This is why figure 1 reports the mean Δ(chrF) for several models. Δ(chrF) scores for UNK tests show that BERT models are not better than expected. However, for chswap, chrand, upper, the BERT models have a slightly lower Δ(chrF). Based on these results, we conclude that pretraining the encoder with a masked LM task does not really bring improvement in terms of robustness to unknowns. It seems that BERT does yield improvement for NMT as a better initialization for

[7] Output for (UNK.S+src) input is not an error, the model does produces an English sentence!

[8] In the case of UNK transformation, the reference is changed but we omit that to simplify the notation.

source sentence	”In home cooking, there is much to be discovered - with a few minor tweaks you can achieve good, if not sometimes better results,” said Proctor.
translation(src)	”Beim Kochen zu Hause gibt es viel zu entdecken - mit ein paar kleinen nderungen kann man gute, wenn nicht sogar manchmal bessere Ergebnisse erzielen”, sagte Proktor.
translation(UNK.S + src)	• ”In home cooking, there is much to be discovered - with a few minor tweaks you can achieve good, if not sometimes better results”, sagte Proktor.

Table 3: Example of a poor translation when adding unknown token to source sentences (translation done with a baseline transformer model)

	news14	news18	iwslt15	wiki	kde	OpenSub
Baseline	27.3	39.5	28.9	17.6	18.1	15.3
NMTsrc.FT	27.7	40.1	28.7	18.3	18.4	15.3
Wiki.FT	27.7	**40.6**	28.7	18.4	**19.0**	15.4
News.FT	**27.9**	40.2	29.1	18.8	17.9	15.7
News.Emb	27.7	39.9	**29.3**	**18.9**	18.2	**16.0**
News.Freeze	23.6	35.5	26.5	15.0	15.1	13.8

Table 4: FT: initialize NMT encoder with BERT and finetune; Freeze: fix NMT encoder parameters to BERT parameters; Emb: fix encoder embeddding layer with BERT contextual word embeddings.

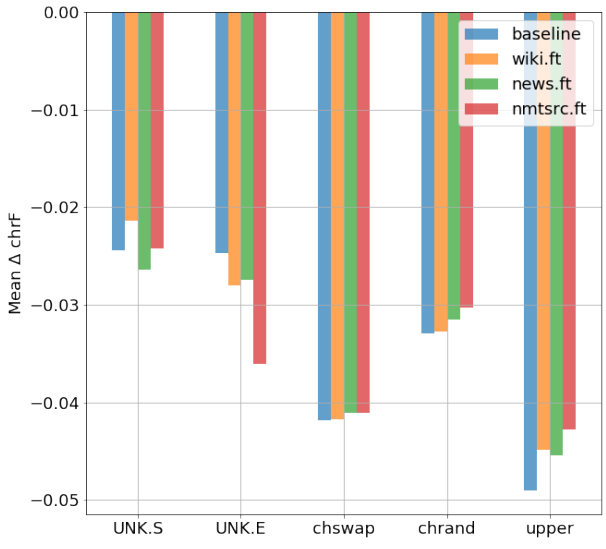

Figure 1: Mean **Δ(chrF)** for several noisy test set and models. For the UNK test, the BERT models are similar or worst than the basline. For the chrand, chswap, upper, the BERT models are slightly better.

NMT encoders but the full potential of masked LM task is not fully exploited for NMT.

5 IWSLT experiments

In order to explore the potential of masked LM encoder pretraining for NMT in lower resource settings, we train NMT models on English-German IWSLT 2015[9] and English-Russian IWSLT 2014[10] MT track datasets. These are pretty small datasets (compared to previous experiments) which contain around 200K parallel sentences each.

5.1 Experimental settings

In these experiments we (1) reuse pretrained BERT models from previous experiments or (2) train IWSLT BERT model. IWSLT BERT model is trained on the concatenation of all the data available at IWSLT 2014-2018 campaigns. After filtering out all the duplicates it contains around 780K sentences and 13.8M tokens.

We considered various settings for IWSLT baseline. First, for source side of the dataset, we took 10K BPE merge operations, where BPE model was trained (1) either on the source side of NMT data only, or (2) on all monolingual English IWSLT data. Target side BPE uses 10K merge operations trained on the target side of the NMT dataset in all the IWSLST experiments. In our first set of experiments, BPE model learnt on source data only lead to similar translation performance as BPE model learnt on all IWSLT English data. Therefore, in what follows we report results only

[9]https://sites.google.com/site/iwsltevaluation2015/mt-track

[10]https://sites.google.com/site/iwsltevaluation2014/mt-track

Models	news14	+UNK.S	+UNK.E	+chswap	+chrand	+up
Baseline	27.3	24.8	24.4	24.2	24.7	23.5
NMTsrc.FT	27.7	24.9	22.9	24.4	25.2	**24.5**
Wiki.FT	27.7	**25.8**	**24.9**	24.4	24.9	24.4
News.FT	**27.9**	24.9	**24.9**	24.5	**25.3**	**24.5**
News.Emb	27.7	24.7	24.8	**24.6**	**25.3**	24.2

Table 5: Robustness tests. BLEU scores for clean and 'noisified' (with different noise type) *news14* testset.

for the latter (referred as $bpe10k$).

NMT model training on IWSLT datasets with Transformer Big architecture on IWSLT data has diverged both for en-de and en-ru dataset. Therefore we use Transformer Base ($tbase$) architecture as a baseline model for these experiments. IWSLT BERT model is also based on $tbase$ architecture described in Vaswani et al. (2017) and for the rest follows same training procedure as described in the section 4.

In order to explore the potential of single pretrained model for all language pairs/datasets we try to reuse Wiki and News pretrained BERT models from previous experiments for encoder initialization of NMT model. However, in the previous experiments, our pretrained BERT models used 32K BPE vocabulary and Transformer Big ($tbig$) architecture which means that we have to reuse the same settings for the encoder trained on IWSLT dataset. It has been shown by Ding et al. (2019), these are not optimal settings for IWSLT training because it leads to too many parameters for the amount of data available. Therefore, in order to reduce the amount of the parameters of the model, we also consider the case where we reduce the amount of the decoder layers from 6 to 3 ($tbig.dec3$).

5.2 Results

Table 6 reports the results of different sets of the experiments on IWSLT data. First, we observe that BERT pretrained model improves over the baseline, in any settings (BPE vocabulary, model architecture, dataset used for pretraining). In particular, it is interesting to mention that without pretraining, both $tbig.bpe32k$ and $tbig.bpe10k$ models diverge when trained on IWSLT. However, BERT pretraining gives a better initialization point, and allows to achieve very good performance both for en-de and en-ru. Thus, such pretraining can be an interesting technique in low-resource scenarios.

	en-de	en-ru
	Baseline	
$tbase.bpe10k$	25.9	9.6
$tbase.dec3.bpe10k$	26.4	16.3
	BERT+NMT	
IWSLT.FT.$tbase.bpe10k$	27.4	17.6
IWSLT.FT.$tbase.dec3.bpe10k$	27.2	**18.1**
Wiki.FT.$tbig.bpe32k$	26.9	17.6
Wiki.FT.$tbig.dec3.bpe32k$	**27.7**	17.8
News.FT.$tbig.bpe32k$	27.1	17.9
News.FT.$tbig.dec3.bpe32k$	27.6	17.9

Table 6: IWSLT dataset results. IWSLT.FT: encoder is initialised with BERT model trained on IWSLT data; $tbase$/$tbig$: transformer base/big architecture for NMT model; $dec3$: decoder layers reduced for 6 to 3; $bpe10k$/$bpe32k$: amount of BPE merge operations used for source language, learnt on the same dataset as BERT model (IWSLT or Wiki+News).

We do not observe big difference between IWSLT pretrained model and News/Wiki pretrained model. We therefore may assume that News/Wiki BERT model can be considered as "general" English pretrained encoder, and be used as a good starting point in any new model translating from English (no matter target language or domain).

6 Discussion

BERT pretraining has been very successful in NLP. With respect to MT, it was shown to provide better performance in Lample and Conneau (2019); Edunov et al. (2019) and allows to integrate large source monolingual data in NMT model as opposed to target monolingual data usually used for backtranslation.

In this experimental study, we have shown that:

- The next sentence prediction task in BERT is not necessary to improve performance - a *masked LM task* already is beneficial.
- It is beneficial to train BERT on the **source**

corpora, therefore supporting the claim that pretraining the encoder provide a better intialization for NMT encoders.

- Similar to Edunov et al. (2019), we observe that the impact of BERT pretraining is more important as the size of the training data decreases (WMT vs IWSLT).

- Information encoded by BERT is not sufficient to perform the translation: NMT encoder encodes both information specific to source sentence, and to the target sentence as well (cf the low performance of BERT frozen encoder).

- Pretraining the encoder enables us to train bigger models. In IWSLT, the transformer big models were diverging, but when the encoder is initialized with pretrained BERT the training became possible. For WMT14, training a 12 layer encoder from scratch was problematic, but News.Emb model (which contains 12 encoder layers) was trained and gave one of the best performances on WMT14.

- Finetuning BERT pretrained encoder is more convenient : it leads to similar performance compared to reusing BERT as embedding layers, with faster decoding speed.

- BERT pretrained models seem to be generally better on different noise and domain test sets. However, we didn't manage to obtain clear evidence that these models are more robust.

This experimental study was limited to a particular dataset, language pair and model architecture. However, many other combinations are possible. First, similar type of study needs to be performed with BERT pretrained model for NMT decoder. Also, the model can be extended to other scenarios with BERT models such as Baevski et al. (2019). In addition, the comparison with ELMO embeddings is also interesting as in Edunov et al. (2019). Using embedding mostly influenced by neighboring words seems to echo the recent results of convolutional self attention network (Yang et al., 2019). Using convolutional self attention network in BERT could bring additional benefit for the pretrained representations. Another direction could look at the impact of the number of layers in BERT for NMT.

Besides, one key question in this study was about the role of encoder in NMT as the roles of encoders and decoders are not clearly understood in current neural architectures. In the transformer architecture, the encoder probably computes some interlingual representations. In fact, nothing constraints the model in reconstructing or predicting anything about the source sentences. If that is the case, why would a monolingual encoder help for the NMT task?

One hypothesis is that encoders have a role of self encoding the sentences but also a translation effect by producing interlingual representations. In this case, a monolingual encoder could be a better starting point and could be seen as a regularizer of the whole encoders. Another hypothesis is that the regularization of transformers models is not really effective and simply using BERT models achieve this effect.

7 Conclusion

In this paper, we have compared different ways to use BERT language models for machine translation. In particular, we have argued that the benefit of using pretrained representations should not only be assessed in terms of BLEU score for the in-domain data but also in terms of generalization to new domains and in terms of robustness.

Our experiments show that fine-tuning the encoder leads to comparable results as reusing the encoder as an additional embedding layers. However, the former has an advantage of keeping the same model size as in standard NMT settings, while the latter adds additional parameters to the NMT model which increases significantly the model size and might be critical in certain scenarios.

For MT practioners, using BERT has also several practical advantages beyond BLEU score. BERT can be trained for one source language and further reused for several translation pairs, thus providing a better initialization point for the models and allowing for better performance.

With respect to robustness tests, the conclusion are less clear. Even if pretrained BERT models obtained better performance on noisy test sets, it seems that they are not more robust than expected and that the potential of masked LM tasks is not fully exploited for machine translation. An interesting future work will be to assess the robustness of models from Song et al. (2019).

References

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. *CoRR*, abs/1903.07785.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations. *CoRR*, abs/1905.10453.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. *CoRR*, abs/1903.09722.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. *CoRR*, abs/1902.01509.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2177–2185, Cambridge, MA, USA. MIT Press.

Qingsong Ma, Ondej Bojar, and Yvette Graham. 2018. Results of the wmt18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 682–701, Belgium, Brussels. Association for Computational Linguistics.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. *CoRR*, abs/1903.06620.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline.

Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. *CoRR*, abs/1902.09508.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019. Convolutional self-attention networks. *CoRR*, abs/1904.03107.