Machine-oriented NMT Adaptation for Zero-shot NLP tasks: Comparing the Usefulness of Close and Distant Languages

Amirhossein Tebbifakhr
FBK, Trento, Italy
University of Trento, Italy
atebbifakhr@fbk.eu

Matteo Negri
FBK, Trento, Italy
negri@fbk.eu

Marco Turchi
FBK, Trento, Italy
turchi@fbk.eu

Abstract

Neural Machine Translation (NMT) models are typically trained by considering humans as end-users and maximizing human-oriented objectives. However, in some scenarios, their output is consumed by automatic NLP components rather than by humans. In these scenarios, translations' quality is measured in terms of their "fitness for purpose" (i.e. maximizing performance of external NLP tools) rather than in terms of standard human fluency/adequacy criteria. Recently, reinforcement learning techniques exploiting the feedback from downstream NLP tools have been proposed for "machine-oriented" NMT adaptation. In this work, we tackle the problem in a multilingual setting where a single NMT model translates from multiple languages for downstream automatic processing in the target language. Knowledge sharing across close and distant languages allows to apply our machine-oriented approach in the zero-shot setting where no labeled data for the test language is seen at training time. Moreover, we incorporate multilingual BERT in the source side of our NMT system to benefit from the knowledge embedded in this model. Our experiments show coherent performance gains, for different language directions over both *i)* "generic" NMT models (trained for human consumption), and *ii)* fine-tuned multilingual BERT. This gain for zero-shot language directions (e.g. Spanish–English) is higher when the models are fine-tuned on a closely-related source language (Italian) than a distant one (German).

1 Introduction

With the rapid growth of cloud computing, there are plenty of online services for a variety of natural language processing (NLP) tasks such as document classification, sentiment analysis, and spam detection. However, building them from scratch typically requires a massive amount of labeled data, which is not always publicly available and, for many tasks, is limited to high-resource languages like English. A possible solution to leverage these services in low-resource settings is using Neural Machine Translation (NMT) in the so-called "translation-based" approach, where a text in the low-resource language is first translated into a high-resource one for which dedicated NLP tools exist. Then, the translated text is processed by these downstream tools and, finally, the results are propagated back to the source language.

Although the translation-based approach shows promising results in low-resource settings (Conneau et al., 2018), it still has drawbacks. First, the output quality of current NMT models is not perfect yet (Koehn and Knowles, 2017). Second, even a good translation can alter some traits in the text, which are essential for the downstream NLP tool. This, for instance, is typical for sentiment traits, whose loss can result in final performance drops in sentiment classification tasks (Mohammad et al., 2016). Finally, state-of-the-art NMT models are trained considering humans as end-users and hence optimized to maximize human-oriented objectives like fluency and semantic equivalence of the translation with respect to the source sentence. However, these objectives are not necessarily the optimal ones to exploit an NLP tool at its best. Machines, in fact, are still worse than humans in handling ambiguous or overly complex sentences. This observation calls for strategies that are alternative to the human-oriented enhancement

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 36–46
Barcelona, Spain (Online), December 13, 2020

of NMT. Rather, models should be adapted in a *machine-oriented* way that is optimal for automatic processing of their output.

Traditionally, NMT models are trained using parallel corpora, consisting of sentences in the source language and their human translations in the target language. Recently, Tebbifakhr et al. (2019) proposed Machine-Oriented Reinforce (MO-Reinforce), a method based on Reinforcement Learning to pursue machine-oriented objectives for sentence-level classification tasks. In a nutshell: given the output of the downstream classifier (i.e. a probability distribution over the labels), MO-Reinforce considers the probability given to the true class as the collected reward from the downstream classifier. By maximizing the expected value of the collected reward, MO-Reinforce adapts the NMT model's behavior to generate outputs that are easier to label by the downstream classifier.

Although NMT models adapted with MO-Reinforce show promising results compared to the "generic" ones trained by only pursuing human-oriented objectives, they still need a small amount of labeled data in the source language to compute the reward that may not be available for some languages. To address this problem, we exploit multilingual NMT models (Johnson et al., 2017), the MO-Reinforce algorithm and a small quantity of data in closely-related languages. Starting from a multilingual NMT system used to translate texts from n low-resource languages into a resource-rich one, the MO-Reinforce algorithm is run in three different conditions by using source data in: *i)* the same language of the test set (tuning on Italian - testing on Italian); *ii)* a different language, but closely related to the one of the test set, (Italian - Spanish), and *iii)* a different and distant language (German - Spanish). The main goal of these experiments is to show that MO-Reinforce leveraging a multilingual NMT and data in a closely-related language is able to overcome the lack of source labelled data in a specific task.

Moreover, recently, multilingual BERT (Devlin et al., 2019) has shown good performance when fine-tuned for downstream tasks. Multilingual BERT is a pre-trained model built on the union of unlabeled data for more than 100 languages. The availability of unlabeled text in significant quantities in different languages helps this model to extract valuable knowledge about the languages resulting in good performance for different tasks. To strengthen the capability of the NMT system to represent the source sentence, we try different approaches to incorporate multilingual BERT in the NMT system's encoder. Our goal, in this case, is to show that BERT-based NMT systems can benefit from the knowledge embedded in BERT, particularly in zero-shot language directions.

We focus on a 4-class sentence classification task (news classification), which is harder compared to the binary task (polarity detection) chosen in (Tebbifakhr et al., 2019). We evaluate the translation-based approach from German, Italian, and Spanish (simulating low-resource language settings) into English. It is important to remark that, although our source languages are not low-resource ones, their choice is motivated by the availability of standard benchmark for a comparative evaluation in a simulated low-resource scenario. The results show that:

- Closely-related languages can help to cope with the lack of annotated data for specific NLP tasks (in this case for document classification into four domains).
- MO-Reinforce is able to take advantage of these data to outperform the classification performance of the generic NMT system and Multilingual BERT.
- Although the addition of Multilingual BERT does not yield improvements in translation quality, its capability of generating good source-sentence representations helps MO-Reinforce to achieve better performance.

2 Related Works

Reinforcement Learning methods have been mainly proposed to address the *exposure bias* problem inside sequence-to-sequence models, which refers to the discrepancy between training and inference time in NMT systems. During training, in fact, the model is exposed to the reference translations, while at inference time the model generates the translation based on its own (typically sub-optimal) predictions, at the risk of cumulative errors at each step. In (Ranzato et al., 2016), the authors proposed a gradual

shifting from token-level maximum likelihood to sentence-level BLEU score to expose the model to its prediction instead of the reference translation. Shen et al. (2016) extended this idea by adopting minimum risk training (Goel and Byrne, 2000) to directly optimize task-specific metrics like BLEU or TER in NMT. Bahdanau et al. (2017) optimized the policy using the actor-critic algorithm. In another line of research, in situations where the reference translation is not available, (Kreutzer et al., 2017) proposed bandit structured prediction, which describes a stochastic optimization framework to leverage "weak" feedbacks collected from the user (e.g. Likert scores about output quality). A common trait of all the above-mentioned works is that they all consider *humans* as the end-users of NMT system's output, which should hence adhere to the human criteria of fluency and adequacy. Tebbifakhr et al. (2019) recently proposed a paradigm shift by considering *machines* as the final consumers on machine-translated text, which should hence maximize "fitness for purpose" criteria (i.e. providing easy-to-process input to downstream NLP components). To this aim, they adopted the REINFORCE (Williams, 1992) approach to leverage the feedback from a downstream task (e.g. classification accuracy in a polarity detection task) to update the agent's policy (the probability of taking a certain action α when in state s). This approach was extended in (Tebbifakhr et al., 2020) to address different NLP tasks in parallel, with a single NMT engine using the same policy. Both works showed that leveraging the downstream classifier's feedback adapts the NMT system to output translations that are easier to be classified by the downstream tool. None of them, however, explored the application of the approach in zero-shot settings, nor focused on how language closeness/distance affects final performance as done in this paper.

Pre-training a neural network or parts of it with existing models is a common approach in several NLP tasks and it allows developers to speed up the training, to leverage different types of training data and to improve the overall performance of the learning system. Among various solutions, *word2vec* (Mikolov et al., 2013) and its variants (Pennington et al., 2014; Levy and Goldberg, 2014) have been the first resources used to pre-training the embeddings in an NMT system. They provide embedded vectors of individual words and have been widely used in NLP.

Recently, pre-trained Language Models (LM) showed better performance when fine-tuned for downstream tasks. ELMo (Peters et al., 2018) is among the first pre-trained LMs, which is based on Bi-LSTM architecture trained on monolingual data. The authors showed that combining the representations from different layers obtains contextual-aware word representations that can be used for other NLP tasks. Right after ELMo, BERT (Devlin et al., 2019) was proposed based on the encoder of Transformer (Vaswani et al., 2017). This model was trained on unlabeled data using two loss functions: *i)* Masked Language Model (MLM) and *ii)* Next Sentence Prediction (NPS). This pre-trained model showed outstanding performance when fine-tuned for a variety of NLP tasks. There are many variants of BERT proposed after, among them: Conneau and Lample (2019) add cross-lingual data and only use MLM loss function, and Yang et al. (2019) train the model on the permuted data.

Specifically for NMT, different attempts have been done to integrate pre-trained LMs in the sequence-to-sequence model. Among others, ELMo was used for initializing the embedding layer in the NMT system (Edunov et al., 2019). Clinchant et al. (2019) used BERT for initializing the encoder or embedding layer of the NMT systems. Then the BERT models are fixed or fine-tuned along with other variables of the model. In (Zhu et al., 2020), a method was proposed to fuse the representations obtained from BERT with each layer of the encoder and decoder in the NMT model through attention mechanisms. We take a similar approach to (Clinchant et al., 2019) in initializing source embedding or the encoder of the NMT system while training a generic NMT systems. Here, this is done for the first time in a machine-oriented setting and in zero-shot conditions.

3 Background and Methodology

3.1 Neural Machine Translation

State-of-the-art NMT models are based on the encoder-decoder architecture (Bahdanau et al., 2015; Vaswani et al., 2017). In this architecture, the encoder encodes the sentence in the source language into vector representations. Then, the decoder autoregressively decodes these representations into a sentence in the target language, emitting a token at each time step until the end-of-sentence token is generated.

More formally, at time step i the NMT model generates a probability distribution $p_\theta(.|\mathbf{y}_{\{0..i-1\}}, \mathbf{x})$ based on the source sentence $\mathbf{x}$ and the already generated translation prefix $\mathbf{y}_{\{0..i-1\}}$, where θ is the model's parameter set. So, for a given translation pair $(\mathbf{x}, \mathbf{y})$ the probability of generating reference translation $\mathbf{y}$ for the given source sentence $\mathbf{x}$ can be computed as follows:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} p_\theta(\mathbf{y}_i|\mathbf{y}_{\{0..i-1\}}, \mathbf{x}) \tag{1}$$

where N is the length of $\mathbf{y}$. These models are usually trained by maximizing the likelihood of a given parallel corpus containing S translation pairs $\{\mathbf{x}^s, \mathbf{y}^s\}_{s=1}^{S}$. The Maximum Likelihood Estimation (MLE) objective function can be written as follows:

$$\begin{aligned} \mathcal{L}_{MLE} &= \sum_{s=1}^{S} \log P(\mathbf{y}^s|\mathbf{x}^s) \\ &= \sum_{s=1}^{S} \sum_{i=1}^{N^s} \log p_\theta(\mathbf{y}_i^s|\mathbf{y}_{\{0..i-1\}}^s, \mathbf{x}) \end{aligned} \tag{2}$$

The parameters of the model can be optimized by applying stochastic gradient descent to maximize MLE objective function. As mentioned in §1, this approach indirectly maximizes the *human-oriented* translation criteria embedded in the parallel corpora used for training.

3.2 Multilingual Machine-Oriented REINFORCE

Tebbifakhr et al. (2019) proposed an approach based on Reinforce (Williams, 1992) that, instead of maximizing the likelihood of the training data, maximizes the expected value of the reward on the output of the NMT system. Formally, the NMT model defines an agent that chooses an action, i.e. generating a translation candidate $\hat{\mathbf{y}}$, and gets a reward $\Delta(\hat{\mathbf{y}})$ according to the action taken. This reward is external to the NMT system and can be collected either from humans (Ranzato et al., 2016; Kreutzer et al., 2017) or, as in MO-Reinforce, from a downstream NLP tool (Tebbifakhr et al., 2019). This objective function can be written as follows:

$$\begin{aligned} \mathcal{L}_{RL} &= \sum_{s=1}^{S} E_{\hat{\mathbf{y}} \sim P(.|\mathbf{x}^{(s)})} \Delta(\hat{\mathbf{y}}) \\ &= \sum_{s=1}^{S} \sum_{\hat{\mathbf{y}} \in \mathbf{Y}} P(\hat{\mathbf{y}}|\mathbf{x}^{(s)}) \Delta(\hat{\mathbf{y}}) \end{aligned} \tag{3}$$

where $\mathbf{Y}$ is the set containing all the possible translations. Since the size of this set is exponentially large, the expected value is usually estimated by sampling one or few candidates from $\mathbf{Y}$. In (Ranzato et al., 2016), the expected value is estimated by sampling only one candidate using multinomial sampling:

$$\hat{\mathcal{L}_{RL}} = \sum_{s=1}^{S} P(\hat{\mathbf{y}}|\mathbf{x}^s)\Delta(\hat{\mathbf{y}}), \hat{\mathbf{y}} \sim P(.|\mathbf{x}^s) \tag{4}$$

In MO-Reinforce, the reward is computed as the probability given to the true class by the downstream classifier. The maximum value of this reward is 1 when the downstream classifier assigns the correct label to the translation candidate with total confidence. Also, to increase the contribution of the reward, MO-Reinforce exploits a sampling strategy (Algorithm 1) where: *i)* K translation candidates are sampled from the output probability distribution of the NMT system, *ii)* the reward is computed for each of them, and *iii)* the one with the highest reward is chosen as final candidate. Although MO-Reinforce shows promising results in adapting NMT models to pursue machine-oriented objectives, it still needs a small amount of labeled data for computing the reward. These data, however, are not always available in the low-resource settings for which it is proposed. To tackle this problem, we extend MO-Reinforce to

Algorithm 1 Machine-Oriented Reinforce

Input: $\mathbf{x}^{(s)}$ s-th source sentence in training data, K number of sampled candidates
Output: sampled candidate $\hat{\mathbf{y}}^{(s)}$
$\mathbf{C} = \emptyset$ {Candidates set}
for $k = 1$ **to** K **do**
  $\mathbf{y}_0 = BOS$ {Beginning-Of-Sentence token}
  $i = 0$
  repeat
    $i = i + 1$
    $\mathbf{y}_i \sim p_\theta(.|\mathbf{x}^s, \mathbf{y}_{\{0..i-1\}})$
    $\mathbf{y}_{\{0..i\}} = \mathbf{y}_{\{0..i-1\}} + \mathbf{y}_i$
  until $\mathbf{y}_i$ is EOS {End-Of-Sentence token}
  $\mathbf{y} = \mathbf{y}_{\{1..i-1\}}$
  $\mathbf{r} = \Delta(\mathbf{y})$ {Reward from the classifier}
  $\mathbf{C} = \mathbf{C} \cup (\mathbf{y}, \mathbf{r})$
end for
$\hat{\mathbf{y}}^{(s)} = \max_{\mathbf{r}}(\mathbf{C})$ {Candidate with maximum reward}

the multilingual setting. We train a multilingual NMT model (Ha et al., 2016), which translates from different low-resource languages (S_1, S_2,...S_n) to a high-resource one (T). This model is trained on the union of (S_i, T) parallel corpora in which the source language differs. This unification of corpora results in learning a language-agnostic representation on the encoder side and enables knowledge transfer from language for which labeled data exist to the (zero-shot) language without labeled data (Eriguchi et al., 2018). The fact that multilingual NMT results in a single model covering multiple languages (as opposed to relying on dedicated models for each language pair) represents an architectural advantage that makes it scalable, easy to maintain and, in turn, particularly appealing for real-world applications.

3.3 BERT-Based NMT

Recently, the pre-trained BERT has shown outstanding results when it is fine-tuned for a downstream task. This superiority comes from the fact that this model has been trained on a huge amount of unlabeled data, which helps it to learn valuable knowledge about the language. The multilingual version of BERT has been trained on the union of the unlabeled data from more than 100 different languages. This multilingual information motivated us in incorporating the multilingual BERT in our NMT system to take advantage of its embedded knowledge. In our setting, where the NMT system serves the downstream tool having a better representation of the input can be beneficial to generate a better and more useful translation, in particular in zero-shot languages.

We employ two different approaches to incorporating BERT in our NMT system based on Transformer (Vaswani et al., 2017). In the standard Transformer, all the variables of the model are randomly initialized and then trained. In our implementation of this model, we use Byte-Pair Encoding (BPE) (Sennrich et al., 2016) to extract the vocabulary from the source and target side of the parallel data. The following paragraphs explains the details of each BERT-based NMT implementations.

BERT Encoder The first approach to incorporating BERT in our NMT system is initializing the encoder of the NMT system using the weights of the multilingual BERT. In this approach instead of using BPE, we tokenize the input sentence to sub-words using the BERT tokenizer and add special tokens **[CLS]** and **[SEP]** to the beginning and the end of the sentence. Then the tokenized sentence is encoded with multilingual BERT and the encoded representations are passed to the decoder of the NMT system.

BERT Embedding In the second approach, we use the output of multilingual BERT as contextualized embeddings of the source sentence. Then these embeddings are passed to the encoder of the NMT system to encode the source sentence. Finally, the output of the encoder is passed to the decoder of the NMT system.

The next Section will explain how the BERT-based NMT systems and MO-Reinforce are used in our experiments to address the lack of source data in a specific language.

4 Experiments

Experimental Settings We pre-train the two NMT systems described in § 3.3 and the standard Transformer using the Maximum-Likelihood Estimation on the parallel corpora for the human-oriented translation task. Their translation performance are evaluated in § 5.1. The outputs of these NMT systems are then passed to the downstream classifier and its classification performance is evaluated in § 5.2. We compare the performance of these approaches (different NMT + Downstream classifier) with the multilingual BERT trained on English data. In this set of experiments, the NMT systems and the BERT classifier are not tuned on any kind of source language classification training data (e.g. using the MO-Reinforce algorithm for the NMT systems), so we consider this setting a zero-shot scenario.

For training the NMT systems using multilingual BERT, we freeze the variables of BERT. We use six layers of the encoder (if any) and six layers of the decoder. We keep the hyper-parameters of the model similar to the original settings (Vaswani et al., 2017). We train each model with the effective batch size equal to 25K tokens.

We then consider the condition when a minimum amount of downstream labeled data is available for a closely-related language to Spanish (Italian) and a distant one (German). We use these data to adapt each NMT system using the Multilingual MO-Reinforce approach described in § 3.2. We evaluate the downstream classifier's performance on the output of each adapted NMT system. We compare the Multilingual MO-Reinforce approach with the multilingual BERT fined-tuned for the downstream task using the same limited amount of the labeled data (see § 5.3). Similar to (Tebbifakhr et al., 2020), we adapt the NMT systems using MO-Reinforce by disabling and enabling the dropout while generating the translation candidates. We keep the parameter K in MO-Reinforce equal to 5, and adapt each NMT system for 50 epochs and choose the best checkpoint based on the performance on the development set. For simulating the downstream classifier we use English BERT fine-tuned for the downstream task using the English labeled data.

Data For pre-training the NMT systems, we use the parallel corpora reported in Table 1, and we evaluate the Spanish and Italian translation performance of the NMT systems on the Ubuntu parallel corpus (Tiedemann, 2012). For the Transformer model, we tokenize and encode each side of the parallel corpora with 32K byte-pair encoding rules. For the other Bert-based NMT systems, on the source side, we use the BERT encoder setting to split the sentences to the tokens.

We evaluate our translation-based classification approaches on a multilingual document classification task where Spanish and Italian news documents have to be automatically annotated with domain labels. Our classification data consists of the first sentence of each document that, according to (Bell, 1991) is a good proxy to determine the domain of news texts. The data used (Schwenk and Li, 2018) cover 4 domains: Corporate/Industrial, Economics, Government/Social, and Markets. The training, development, and test sets for each language respectively contain 10K, 1K, and 4K documents, equally distributed in the 4 classes. For the English downstream classifier we use whole 10K documents while, to simulate the low-resource setting, we sample 100 documents for each class from the Italian training set. In addition, we collect the same amount of data also for German. This data is used to fine-tune the Spanish system on a distant language and compare downstream performance results achieved by our approach in the two fine-tuning conditions (close – Es-It – vs distant – Es-De – languages).

Evaluation metrics We evaluate the translation performance of the NMT systems using BLEU Score (Papineni et al., 2002) and the classification performance with macro average F1 Score.

5 Results

5.1 The NMT systems' translation performance

We start the evaluation by comparing the translation performance of the three different NMT systems. The performance of the NMT systems in terms of BLEU score is reported in Table 2. As shown, *BERT*

	Europarl	JRC	Wikipedia	ECB	TED	KDE	News11	News	Total
Es-En	2M	0.8M	1.8M	0.1M	0.2M	0.2M	0.3M	0.2M	5.6M
It-En	2M	0.8M	1M	0.2M	0.2M	0.3M	0.04M	0.02M	4.56M
De-En	2M	0.7M	2.5M	0.1M	0.1M	0.3M	0.2M	0.2M	6.1M

Table 1: Number of sentences in the parallel corpora used for training the generic NMT systems.

	BERT Encoder	BERT Embedding	Transformer
Italian	20.19	21.88	**25.56**
German	17.18	19.04	**21.86**
Spanish	26.15	28.12	**32.02**

Table 2: Translation performance of the different NMT systems in terms of BLEU score.

Encoder has lower performance compared to *Transformer* (-5.37 in Italian, -4.68 in German, and -5.87 in Spanish). This is due to the fact that the encoder of the NMT system is the fixed multilingual BERT and only the weights on the decoder side (embeddings, decoder weights, and linear projection to vocabulary) are trained in this setting. *BERT Embedding* outperforms *BERT Encoder* in translation task (+1.69 in Italian, +1.86 in German, +1.97 in Spanish). This improvement was expected, because in this setting the weights in the encoder are also trained along with the parameters on the decoder side. However, the performance of this systems is still lower than *Transformer* (-3.68 in Italian, -2.82 in German, and -3.90 in Spanish). This is because multilingual BERT is trained only on monolingual data. Compared to *Transformer*, in which all the weights are trained on parallel data, its output representations are hence less effective for translation tasks. These results are mainly in line with those reported in (Clinchant et al., 2019) when using BERT as fixed encoder. However, while they showed improvements using BERT as embedding matrix on the encoder side in some settings, this is not visible in our experiments.

5.2 The classification performance on the output of the NMT systems

We continue the evaluation, by using these three NMT systems in our translation-based classification approach for the downstream task. All the NMT systems are not aware of the downstream task and are not fine-tuned on any task-specific data in the source and target languages. For these reasons, we consider these experiments as in zero-shot conditions. Table 3 shows the downstream classification task's performance using the different NMT systems. We also compare them with the multilingual BERT fine-tuned for the downstream task using the English labeled data. As shown in the table, except for the *BERT Encoder* in German (-0.9 F1 Score), the translation-based classification approach in all the other settings has better performance in the zero-shot settings than the multilingual BERT. Among the BERT-based NMT systems, except Spanish in which the results are similar (85.9 for *BERT Encoder* and 85.8 for *BERT Embedding*), *BERT Embedding* outperforms *BERT Encoder* in the other two languages (+0.8 in Italian and +2.8 in German). These results are in line with BERT-based NMT systems' translation performance reported in Table 2. However, when comparing the BERT-based NMT systems with the *Transformer*, the translation gap showed in Table 2 in favour of Transformer is minimized for Italian and German in terms of classification performance and overturned for Spanish (+1.6 F1 Score). This analysis shows that the translation quality in terms of human-oriented scores (like the BLEU score) does not have a high correlation with the performance of the downstream task in the translation-based classification scenario.

5.3 MO-Reinforce adaptation

Table 4 reports the performance of Multilingual BERT and the three NMT systems when testing on Italian and Spanish. The systems are fine-tuned using the Italian and German labeled data. This set of experiments covers the conditions when the NMT system is fine-tuned by MO-Reinforce on the data belonging to a) the same language of the test set (Italian - Italian, second column); b) a different language,

	Italian	German	Spanish
Multilingual BERT	70.2	87.0	80.6
BERT Encoder	74.5	86.1	85.9
BERT Embedding	75.3	88.9	85.8
Transformer	76.1	88.8	84.2

Table 3: Document classification performance in terms of F1 Score using only English labeled data.

	Fine-tuned on labeled data in Italian		*Fine-tuned on labeled data in German*
	Italian	Spanish	Spanish
Multilingual BERT	82.9	80.7	73.2
BERT Encoder	76.7	85.3	83.3
BERT Embedding	76.8	**86.7**	**86.3**
Transformer	77.8	86.1	84.3
	(enabled dropout)		
BERT Encoder	83.6	83.4	76.8
BERT Embedding	**84.0**	82.7	76.0
Transformer	82.5	82.3	75.7

Table 4: Document classification performance in terms of F1 Score by fine-tuning on labeled data in close and distant languages.

but closely-related (Italian - Spanish, third column) and c) a different and distant language (German - Spanish, fourth column), the more extreme case. The NMT systems do not use the dropout during the MO-Reinforce fine-tuning in the top part of the table, while it is enabled in the bottom experiments.

In the top part of Table 4, comparing the performance of all the systems tested on Spanish (third and fourth columns), fine-tuning on Italian and German shows coherent performance gains when the systems are trained on the closely-related language (Italian) over the distant language (German). The NMT systems adapted using MO-Reinforce, except for *BERT Encoder*, have performance improvement over the generic NMT systems (Table 3), showing that MO-Reinforce can alter the output of the translation system to benefit more from the knowledge embedded in the downstream classifier. Also, MO-Reinforce outperforms multilingual BERT in both cases (fine-tuning on close and distant languages), thanks to the language-agnostic representation of the NMT encoder , which is not the case in multilingual BERT trained only on monolingual data. The best result for Spanish is obtained when *BERT Embedding* is fine-tuned on Italian labeled data. It confirms our hypothesis that the better representation of the language by multilingual BERT trained on huge amount of data can be beneficial compared to *Transformer*, which has seen only the limited parallel data. When testing on Italian, as expected, all the systems have an increase in performance compared to the results in Table 3. The best result is obtained by multilingual BERT (82.19 F1 Score) showing that it is able to better leverage the labeled data. Our intuition is that the NMT systems without enabling the dropout during fine-tuning do not properly explore the searching space. On one side, this allows the system to better transfer the source-language data knowledge to the other languages, but, on the other, it limits the learning capability on the language for which the labelled data is available.

To test this hypothesis and similar to (Tebbifakhr et al., 2020), we repeat the adaption of the NMT systems using MO-Reinforce by enabling the dropout while generating the translation candidates. This adds some noise to the translation outputs that helps the system to avoid possible local optima and favours a deeper exploration of the probability space. The bottom part of Table 4 reports the performance of the translation-based approaches for the downstream classification task for NMT systems adapted by

MO-Reinforce on Italian and German labeled data using the dropout while generating the translation candidates. The first noticeable change in the result is the boost in the Italian language's performance. This difference is higher for *BERT Encoder* and *BERT Embedding* (83.6 and 84.0 F1 score respectively) outperforming the multilingual BERT (82.9 F1 Score). However, this improvement in Italian comes at the cost of lower performance in Spanish (zero-shot language). This drop is smaller for the systems fine-tuned on Italian confirming the advantage of using closely-related languages. Indeed, the more the NMT system becomes specialized on German (distant language), the lower the performance are on Spanish. However, even with this drop in performance on the zero-shot language, the performance of all the systems are higher than the multilingual BERT. This observation confirms that NMT systems have a more language-agnostic representation of the input text, which results in easier knowledge transfer between languages, in particular the closer ones.

6 Conclusion

In this paper, we proposed a multilingual extension of the MO-Reinforce algorithm able to work in zero-shot settings. Our solution takes advantage of a multilingual NMT model, which translates texts from different low-resource languages into English. To mitigate the lack of data in zero-shot languages, we also incorporated the multilingual BERT with different approaches in the NMT model. Our evaluation shows that using generic NMT systems in the translation-based approach works better than the multilingual BERT in zero-shot settings. Furthermore, the shared knowledge between the source languages allows MO-Reinforce to leverage the labeled data in one language to adapt the NMT model to pursue machine-oriented objectives in other languages, even in zero-shot settings. Our results show that data in closely-related languages can help to cope with the lack of task-specific resources and confirm the capability of MO-Reinforce to leverage and transfer information across languages. The best result in zero-shot settings is obtained with the NMT system by incorporating multilingual BERT as embeddings adapted on closely-related language. However, for the language with a small amount of data, the best approach uses the multilingual BERT as the embedding for the source side of the NMT system by enabling the dropout while generating the translation candidates. Our future works will consider fine-tuning the multilingual BERT variables along with other variables of the model in NMT systems, which can obtain a language-agnostic representation in the BERT model and be helpful in zero-shot settings.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, Conference Track Proceedings*, San Diego, California, USA, May.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations, Conference Track Proceedings*, Toulon, France, April.

A. Bell. 1991. *The Language of News Media*. Language in society. Blackwell.

Stephane 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*, pages 108–117, Hong Kong, November.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Vancouver, Canada, December.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. 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)*, pages 4052–4059, Minneapolis, Minnesota, June.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *CoRR*, abs/1809.04686.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Comput. Speech Lang.*, 14(2):115–135, April.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada, August.

Julia Kreutzer, Artem Sokolov, and Stefan Riezler. 2017. Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1503–1513, Vancouver, Canada, July.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55(1):95–130, January.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

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*, pages 1532–1543, Doha, Qatar, October.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237, New Orleans, Louisiana, USA, June.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, Conference Track Proceedings*, San Juan, Puerto Rico, May.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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, August.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany, August.

Amirhossein Tebbifakhr, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Machine translation for machines: the sentiment classification use case. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1368–1374, Hong Kong, China, November.

Amirhossein Tebbifakhr, Matteo Negri, and Marco Turchi. 2020. Automatic translation for multiple nlp tasks: a multi-task approach to machine-oriented nmt adaptation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 235–244, Virtual, November.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey, May.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Long Beach, California, USA, December.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Vancouver, Canada, December.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *8th International Conference on Learning Representations, Conference Track Proceedings*, Virtual, April.