On Zero-shot Cross-lingual Transfer of Multilingual Neural Machine Translation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Transferring representations from large-scale supervised tasks to downstream tasks have shown outstanding results in Machine Learning in both Computer Vision and *natural language processing* (NLP). One particular example can be sequence-to-sequence models for Machine Translation (Neural Machine Translation - NMT). It is because, once trained in a multilingual setup, NMT systems can translate between multiple languages and are also capable of performing *zero-shot translation* between unseen source-target pairs at test time. In this paper, we first investigate if we can extend the zero-shot transfer capability of multilingual NMT systems to cross-lingual NLP tasks (tasks other than MT, e.g. sentiment classification and natural language inference). We demonstrate a simple framework by reusing the encoder from a multilingual NMT system, a multilingual Encoder-Classifier, achieves remarkable zero-shot cross-lingual classification performance, almost out-of-the-box on three downstream benchmark tasks - Amazon Reviews, *Stanford sentiment treebank* (SST) and *Stanford natural language inference* (SNLI). In order to understand the underlying factors contributing to this finding, we conducted a series of analyses on the effect of the shared vocabulary, the training data type for NMT models, classifier complexity, encoder representation power, and model generalization on zero-shot performance. Our results provide strong evidence that the representations learned from multilingual NMT systems are widely applicable across languages and tasks, and the high, out-of-the-box classification performance is correlated with the generalization capability of such systems.

1 Introduction

Transfer learning has been shown to work well in Computer Vision where pre-trained components from a model trained on ImageNet [1] are used to initialize models for other tasks [2]. In most cases, the other tasks are related to and share architectural components with the ImageNet task, enabling the use of such pre-trained models for feature extraction. With this transfer capability, improvements have been obtained on other image classification datasets and on other tasks such as object detection, action recognition, image segmentation, etc [3]. Analogously, we propose a method to transfer a pre-trained component - the multilingual encoder from an NMT system - to other NLP tasks.

In NLP, initializing word embeddings with pre-trained word representations obtained from Word2Vec [4] or GloVe [5] has become a common way of transferring information from large unlabeled data to downstream tasks. Recent work has further shown that we can improve over this approach significantly by considering representations in context, i.e. modeled depending on the sentences that contain them, either by taking the outputs of an encoder in MT [6] or by obtaining representations from the internal states of a bi-directional *language model* (LM) [7]. There has also been successful recent work in transferring sentence representations from resource-rich tasks to

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

improve resource-poor tasks [8], however, most of the above transfer learning examples have focused on transferring knowledge across tasks for a single language, in English.

Zero-shot classification over the languages is one of the most interesting cross-lingual or multilingual NLP tasks, the task of transferring knowledge from one language to another, without any training data in the target language. This serves as a good test bed for evaluating various transfer learning approaches in a multilingual setup. For cross-lingual NLP, the most widely studied approach is to use multilingual embeddings as features in neural network models, and recent research has shown that representations learned in context are more effective [6, 7]. On the other hand, recent progress in multilingual NMT provides a compelling opportunity for obtaining contextualized multilingual representations, as multilingual NMT systems are capable of generalizing to an unseen language direction, i.e. zero-shot translation, and there is also evidence that the encoder of a multilingual NMT system learns language agnostic, universal interlingua representations, which can be further exploited [9].

In this paper, we explore the zero-shot classification performance of representations obtained from a multilingual NMT system. We first show that, by simply reusing the encoder of a multilingual NMT system, remarkably high zero-shot cross-lingual classification performance can be reached (i.e. classification task in a language that the classifier is not trained on). Next, we provide upper bound systems for zero-shot classification by bridging methods, where test data is translated into English (language that the classifier is trained on) and employ English classifiers on them. We demonstrate that, multilingual NMT representations achieve surprisingly close zero-shot classification performance compared to the provided bridging upper bounds on three different tasks - Amazon Reviews, SST, and SNLI. Finally, we carefully analyze how and why cross-lingual knowledge transfer works in the zero-shot setup, and study the effect of various factors on high zero-shot classification performance.

2 Proposed Method

We propose a multilingual *Encoder-Classifier* model, where the *Encoder*, leveraging the representations learned by a multilingual NMT model, converts an input sequence $\boldsymbol{x}$ into a set of vectors $\mathbf{C}$, and the *Classifier* predicts a class label y given the encoding of the input sequence, $\mathbf{C}$.

2.1 Multilingual Representations Using NMT

Although there has been a large body of work in building multilingual NMT models which can translate between multiple languages at the same time [9–12], zero-shot capabilities of such multilingual representations have only been tested for MT [9]. We propose a simple yet effective solution - reuse the encoder of a multilingual NMT model to initialize the encoder for other NLP tasks. To be able to achieve promising zero-shot classification performance, we consider two factors: (1) The ability to encode multiple source languages with the same encoder and (2) The ability to learn language agnostic representations of the source sequence. Based on the literature, both requirements can be satisfied by training a multilingual NMT model having a shared encoder [9, 13] that jointly maps multiple languages into a shared representation and a separate decoder (each having a separate attention mechanism) for each target language [11] in a multi-task framework [14]. After training such a multi-task multilingual NMT model, the decoder and the corresponding attention mechanisms (which are target-language specific) are discarded, while the multilingual encoder is used to initialize the encoder of our proposed mutlitlingual *Encoder-Classifier* model.

2.2 Multilingual Encoder-Classifier

Encoder. In order to leverage pre-trained multilingual representations introduced in Section 2.1, our encoder strictly follows the structure of a regular *recurrent neural network* (RNN) based NMT encoder [15] with a stacked layout [16]. Given an input sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_{T_x})$ of length T_x, our encoder contextualizes or encodes the input sequence into a set of vectors $\mathbf{C}$ by first applying a bi-directional RNN [17], followed by a stack of uni-directional RNNs. The hidden states of the final layer RNN, h_i^l, form the set $\mathbf{C} = \{h_i^l\}_{i=1}^{T_x}$ of context vectors which will be used by the classifier, where l denotes the number of RNN layers in the stacked encoder.

Classifier. The task of the classifier is to predict a class label y given the context set $\mathbf{C}$. To ease this classification task given a variable length input set $\mathbf{C}$, a common approach in the literature is to extract a single sentence vector $\boldsymbol{q}$ by making use of pooling over time [18]. Further, to increase the modeling capacity, the pooling operation can be parameterized using pre- and post-pooling networks. Formally, given the context set $\mathbf{C}$, we extract a sentence vector $\boldsymbol{q}$ in three steps, using three networks, (1) pre-pooling feed-forward network f_{pre}, (2) pooling network f_{pool} and (3) post-pooling feed-forward network f_{post}, which is defined as $\boldsymbol{q} = f_{post}(f_{pool}(f_{pre}(\mathbf{C})))$. Finally, given the sentence vector $\boldsymbol{q}$, a class label y is predicted by employing a softmax function.

3 Experimental Design

3.1 Corpora

We evaluate the proposed method on three common NLP tasks: Amazon Reviews, SST and SNLI. We utilize parallel data to train our multilingual NMT system, as detailed below.

Machine Translation. For the MT task, we use the WMT 2014 En↔Fr parallel corpus. The dataset contains 36 million En→Fr sentence pairs. We swapped the source and target sentences to obtain parallel data for the Fr→En translation task. We use these two datasets (72 million sentence pairs) to train a single multilingual NMT model to learn both these translation directions simultaneously. We generated a shared sub-word vocabulary [19, 20] of 32K units from all source and target training data. We use this sub-word vocabulary for all of our experiments below.

Amazon Reviews. The Amazon Reviews dataset [21] is a multilingual sentiment classification dataset, providing data for four languages - *English* (En), *French* (Fr), *German* (De), and Japanese. We use the English and French datasets in our experiments. The dataset contains 6,000 documents in the train and test portions for each language. Each review consists of a category label, a title, a review, and a star rating (5-point scale). We only use the review text in our experiments. Following [21], we mapped the reviews with lower scores (1 and 2) to negative examples and the reviews with higher scores (4 and 5) to positive examples, thereby turning it into a binary classification problem. Reviews with score 3 are dropped. We split the training dataset into 10% for development and the rest for training, and we truncate each example and keep the first 200 words in the review. Note that, since the data for each language was obtained by crawling different product pages, the data is not aligned across languages.

SST. The sentiment classification task proposed in [22] is also a binary classification problem where each sentence and phrase is associated with either a positive or a negative sentiment. We ignore phrase-level annotations and sentence-level neutral examples in our experiments. The dataset contains 6,920 examples for training, 872 examples for development, and 1,821 examples for testing. Since SST does not provide a multilingual test set, we used the public translation engine Google Translate[1] to translate the SST test set to French. Previous work by Agić and Schluter [23] has shown that replacing the human translated test set with a synthetic set (obtained by using Google Translate) produces only a small difference of around 1% absolute accuracy on their human-translated French SNLI test set. Therefore, the performance measured on our 'pseudo' French SST test set is expected to be a good indicator of zero-shot performance.

Multilingual SNLI. Natural language inference is a task that aims to determine whether a natural language hypothesis $\boldsymbol{h}$ can justifiably be inferred from a natural language premise $\boldsymbol{p}$. SNLI [24] is one of the largest datasets for a natural language inference task in English and contains multiple sentence pairs with a sentence-level entailment label. Each pair of sentences can have one of three labels - *entailment*, *contradiction*, and *neutral*, which are annotated by multiple humans. The dataset contains 550K training, 10K validation, and 10K testing examples. To enable research on multilingual SNLI, Agić and Schluter [23] chose a subset of the SNLI test set (1,332 sentences) and professionally translated it into four major languages - Arabic, French, Russian, and Spanish. We use the French test set for evaluation in Section 4 and 5.

[1] `https://translate.google.com` as of October, 2017.

3.2 Model and Training Details

Here, we first describe the model and training details of the base multilingual NMT model whose encoder is reused in all other tasks. Then we provide details about the task-specific classifiers. For each task, we provide the specifics of f_{pre}, f_{pool} and f_{post} nets that build the task-specific classifier.

All the models in our experiments are trained using the Adam optimizer [25] with label smoothing [26]. Unless otherwise stated below, layer normalization [27] is applied to all LSTM gates and feed-forward layer inputs. We apply L2 regularization to the model weights and dropout to layer activations and sub-word embeddings. Hyper-parameters, such as mixing ratio λ of L2 regularization, dropout rates, label smoothing uncertainty, batch sizes, learning rate of optimizers and initialization ranges of weights are tuned on the development sets provided for each task separately.

NMT Models. Our multilingual NMT model consists of a shared multilingual encoder and two decoders, one for English and the other for French. The multilingual encoder uses one bi-directional LSTM, followed by three stacked layers of uni-directional LSTMs in the encoder. Each decoder consists of four stacked LSTM layers, with the first LSTM layers intertwined with additive attention networks [15] to learn a source-target alignment function. All uni-directional LSTMs are equipped with residual connections [28] to ease the optimization both in the encoder and the decoders. LSTM hidden units and the shared source-target embedding dimensions are set to 512.

Similar to [11], the multilingual NMT model is trained in a multi-task learning setup, where each decoder is augmented with a task-specific loss, minimizing the negative conditional log-likelihood of the target sequence given the source sequence. During training, mini-batches of En$\rightarrow$Fr and Fr$\rightarrow$En examples are interleaved. We picked the best model based on the best average development set BLEU score on both of the language pairs.

Amazon Reviews and SST. The multilingual *Encoder-Classifier* model here uses the encoder defined previously. With regards to the classifier, the pre- and post-pooling networks (f_{pre}, f_{post}) are both one-layer feed forward networks to cast the dimension size from 512 to 128 and from 128 to 32, respectively. We used max-pooling operator for the f_{pool} network to pool activation over time.

Multilingual SNLI. We extended the proposed multilingual *Encoder-Classifier* model to a multi-source model [29] since SNLI is an inference task of relations between two input sentences, “premise" and “hypothesis". For the two sources, we use two separate encoders, which are initialized with the same pre-trained multilingual NMT encoder, to obtain their representations. Following our notation, the encoder outputs are processed using f_{pre}, f_{pool} and f_{post} nets, again with two separate network blocks. Specifically, f_{pre} consists of a co-attention layer [30] followed by a two-layer feed-forward neural network with residual connections. We use max pooling over time for f_{pool} and again a two-layer feed-forward neural network with residual connections as f_{post}. After processing two sentence encodings using two network blocks, we obtain two vectors representing premise $\boldsymbol{h}_{premise}$ and hypothesis $\boldsymbol{h}_{hypothesis}$. Following [31], we compute two types of relational vectors with $\boldsymbol{h}_{-} = |\boldsymbol{h}_{premise} - \boldsymbol{h}_{hypothesis}|$, and $\boldsymbol{h}_{\times} = \boldsymbol{h}_{premise} \odot \boldsymbol{h}_{hypothesis}$, where $\odot$ denotes the element-wise multiplication between two vectors. The final relation vector is obtained by concatenating $\boldsymbol{h}_{-}$ and $\boldsymbol{h}_{\times}$. For both “premise" and “hypothesis" feed-forward networks, we used 512 hidden dimensions.

For Amazon Reviews, SST and SNLI tasks, we picked the best model based on the highest development set accuracy.

4 Zero-Shot Classification Results

In this section, we explore the zero-shot classification task in French for our systems. We assume that we do not have any French training data for all the three tasks and test how well our proposed method can generalize to the unseen French language without any further training. A reasonable upper bound to which zero-shot performance should be compared to is *bridging* - translating a French test text to English and then applying the English classifier on the translated text. If we assume the translation to be perfect, we should expect this approach to perform as well as the English classifier, hence constituting an upper bound.

Table 1: Zero-Shot performance on all French test sets.

Model	Amazon (Fr)		SST (Fr)		SNLI (Fr)	
	Bridged	Zero-Shot	Bridged	Zero-Shot	Bridged	Zero-Shot
Encoder-Classifier	73.30	51.53	79.63	59.47	74.41	37.62
+ Pre-trained Encoder	79.23	75.78	84.18	81.05	80.65	72.35
+ Freeze Encoder	83.10	81.32	84.51	83.14	81.26	73.88

Table 2: Comparison of our best zero-shot result on the French SNLI test set to other baselines. See text for details.

Model	SNLI (Fr)
Our best zero-shot *Encoder-Classifier*	**73.88**
INVERT [32]	62.60
BiCVM [33]	59.03
RANDOM [34]	63.21
RATIO [34]	58.64

The Amazon Reviews and SNLI tasks have a French test set available, and we evaluate the performance of the bridged and zero-shot systems on each French set. However, the SST dataset does not have a French test set, hence the 'pseudo French' test set described in Section 3.1 is used to evaluate the zero-shot performance. The bridged system in the SST column reports the classification performance of the English classifier on the original English test set, as a high quality proxy for the SST bridged system. We do this since translating the 'pseudo French' back to English will result in two distinct translation steps and hence more errors.

Table 1 summarizes all of our zero-shot results for French classification on the three tasks. It can be seen that just by using the pre-trained NMT encoder, the zero-shot performance increases drastically from almost random to within 10% of the bridged system. Freezing the encoder further pushes this performance closer to the bridged system. On the Amazon Review task, our zero-shot system is within 2% of the best bridged system. On the SST task, our zero-shot system obtains an accuracy of 83.14%, which is within 1.5% of the bridged equivalent (in this case the English system).

Finally, on SNLI, we compare our best zero-shot system with bilingual and multilingual embedding based methods evaluated on the same French test set in [23]. As illustrated in Table 2, our best zero-shot system obtains the highest accuracy of 73.88%. INVERT [32] uses inverted indexing over a parallel corpus to obtain crosslingual word representations. BiCVM [33] learns bilingual compositional representations from sentence-aligned parallel corpora. In RANDOM [34], bilingual embeddings are trained on top of parallel sentences with randomly shuffled tokens using skip-gram with negative sampling, and RATIO is similar to RANDOM with the one difference being that the tokens in the parallel sentences are not randomly shuffled. Our system significantly outperforms all methods listed in the second column by 10.66% to 15.24% and demonstrates the effectiveness of our proposed approach.

5 Analyses

In this section, we try to analyze why our simple multilingual *Encoder-Classifier* system is effective at zero-shot classification. We perform a series of experiments to better understand this phenomenon. In particular, we study (1) the effect of shared sub-word vocabulary, (2) the amount of multilingual training data to measure the influence of multilinguality, (3) encoder/classifier capacity to measure the influence of representation power, and (4) model behavior on different training phases to assess the relation between generalization performance on English and zero-shot performance on French.

Effect of Shared Sub-Word Vocabulary. As mentioned in Section 3.2, we use a shared sub-word vocabulary which can encode both English and French text in all of our models. In this subsection, we analyze how much using a shared sub-word vocabulary can help the model generalize to a new language. To verify the effectiveness of just the sub-word vocabulary on generalization, we picked the German test set from the Amazon Review task. Since German shares many sub-words with English and French, the *out-of-vocabulary* (OOV) rate for the German test set using our vocabulary is just 0.078%. We design this experiment as a control to understand the effect of having a shared sub-word

Table 3: Results of the control experiment on zero-shot performance on the Amazon German test set.

Model	Amazon (De)
Zero-shot *Encoder-Classifier*	52.33
+ Pre-trained Encoder	52.98
+ Freeze Encoder	57.72

Table 4: Effect of MT data over our proposed multilingual *Encoder-Classifier* on the SNLI tasks. The results of SNLI (Fr) shows the zero-shot performance of our system.

Parallel data type for NMT	SNLI (En)	SNLI (Fr)
Symmetric data (full)	84.13	73.88
Symmetric data (half)	80.79	66.72
Asymmetric data (half)	81.15	67.63

vocabulary which can encode the language but for which no translation data was seen while training the multilingual NMT encoder.

From Table 3, we can see that despite the very low OOV rate, the ability of our system to perform zero-shot classification on German is close to random, i.e. around 50% accuracy. The third row in the table shows the small deviation of 7% over random, which is likely obtained from common sub-words having similar meaning across languages. This control experiment suggests that although having a shared sub-word vocabulary is necessary, we still need to train the NMT system on parallel data from the language of interest so that the system can perform zero-shot classification.

Effect of Translation Data. We explore two dimensions that could affect zero-shot performance related to our training data in the multilingual NMT model. First, we investigate the effect of using symmetric training data to train both directions in the multilingual NMT system. We conduct an experiment where we take half of the sentences from the En→Fr training set and use the swapped version of the other half of the sentences for training the model. Second, we want to see the effect of training data size, so we run an experiment where we use only half of the training set in a symmetric fashion. From Table 4, we can see that halving the training data size significantly lowers the zero-shot accuracy on the French SNLI test set by 7.16%. However, both the symmetric and asymmetric versions of the data perform comparably on both tasks. This shows that the multilingual NMT system is able to learn an effective interlingua even without the need of symmetric data across the language pairs involved.

Effect of Encoder/Classifier Capacity. We study the effect of the capacity of the two parts of our model on the final accuracies. Specifically, we experimented with two variants of the classifier - a simple linear classifier where we set f_{pre} and f_{post} networks to identity[2] and a complex classifier (details provided in Section 3.2). Next, we experimented with only reusing different parts of the multilingual encoder in a bottom-up fashion. Table 5 summarizes all of our experiments with respect to model capacity. As expected, going from a simple linear classifier to a complex classifier significantly improves both English and zero-shot French performance on the SNLI tasks, while even a simple linear classifier can achieve significant zero-shot performance when provided with rich enough encodings (49.66 to 61.61 accuracy). However, changing the encoder capacity tells an interesting story. As we selectively reuse parts of the encoder from the embedding layer to the top, we notice that the English performance only increases by about 2% whereas the zero-shot performance increases by about 18% in the complex classifier. This means that the additional layers in the encoder are essential for the proposed system to model a language agnostic representation (interlingua) which enables it to perform better zero-shot classification. Moreover, it should be noted that best zero-shot performance is obtained by using the complex classifier and up to layer 3 of the encoder. Although this gap is not big enough to be significant, we hypothesize that top layer of the encoder could be very specific to the MT task and hence might not be best suited for zero-shot classification.

Effect of Early vs Late Phases of the Training. Figure 1 shows that as the number of training steps increases, the test accuracy goes up whereas the test loss on the SNLI task increases slightly,

[2]We empirically found that for simple classifiers using mean pooling for f_{pool} performs considerably better over max-pooling (67.26 vs 61.19 test accuracies respectively) on the SNLI task.

Table 5: Zero-shot analyses of classifier network model capacity. The SNLI (Fr) results report the zero-shot performance.

Encoder components	Simpler classifier		Complex classifier	
	SNLI (En)	SNLI (Fr)	SNLI (En)	SNLI (Fr)
Embeddings only	65.18	49.66	82.43	56.66
+ bi-directional layer 1	67.99	58.19	83.40	64.74
+ layer 2	67.00	61.01	83.63	72.81
+ layer 3	67.26	60.55	84.17	74.33
+ layer 4	67.26	61.61	84.41	74.11

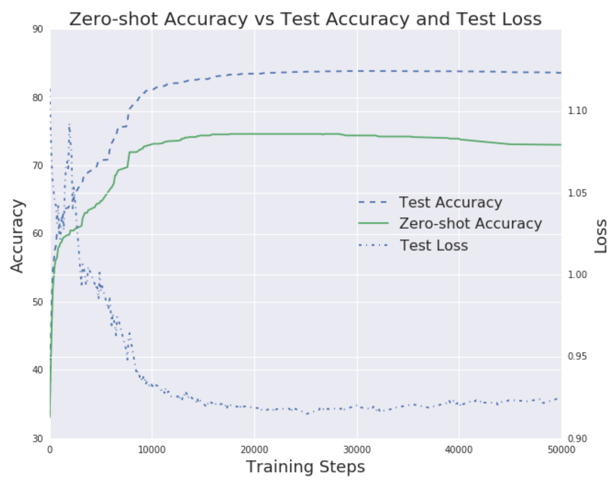

Figure 1: Correlation between test-loss, test-accuracy (the English SNLI) and zero-shot accuracy (the French SNLI test set).

Table 6: Effect of parameter smoothing on the English SNLI test set and zero-shot performance on the French SNLI test set.

Smoothing Range (steps)	SNLI (En)	SNLI (Fr)
1	84.41	74.11
400	84.62	75.02
1K	84.67	75.48
20K	84.65	75.93
35K	84.46	75.63

hinting at over-fitting on the English task. As expected, choosing checkpoints which are before the onset of the over-fitting seems to benefit zero-shot performance on the French SNLI test set. This suggests that over-training on the English task might hurt the ability of the model to generalize to a new language and also motivated us to conduct the next set of analysis.[3]

Effect of Parameter Smoothing. Parameter smoothing (checkpoint averaging [35]) is a technique which aims to smooth point estimates of the learned parameters by averaging n steps from the training run and using it for inference. This is aimed at improving generalization and being less susceptible to the effects of over-fitting at inference. We hypothesize that a system with enhanced generalization might be better suited for zero-shot classification since it is a measure of the ability of the model to generalize to a new task. Table 6 validates our hypothesis by showing that although the average of 20k steps only improves the English SNLI score by 0.24%, it improves the corresponding French zero-shot score by 1.82%.

[3]We observe that test loss better correlates with zero-shot accuracy than test accuracy.

6 Related Work

Word and Sentence Representations. Pre-trained word representations, which leverage large scale unlabeled data [4, 5], have been shown to be a key ingredient in many standard NLP tasks. The tasks include sentiment analysis [22], entailment [24], summarization [36], question answering [37], and semantic role labeling [38]. However, these representations are usually learned from unsupervised data sources which are often unrelated to the downstream task.

Contextualized Representations. Several studies have overcome the fact that these representations are context-independent by proposing contextualized word embeddings. Representations obtained from an LM have been shown to obtain effective contextualized word representations [7, 39]. There has also been work in enriching these word representations using sub-word information [40, 41]. MT naturally lends itself as a suitable task for obtaining contextualized embeddings since the encoder has to encode units in context so as to decode them into another language. Hill et al. [42] show the effectiveness of representations obtained from an NMT model in semantic similarity tasks. They further report that the representations obtained from the NMT model are better than those obtained from LMs. McCann et al. [6] showed that using the representations obtained from the encoder of an NMT system as context vectors in downstream NLP tasks significantly improves performance over using only unsupervised word or character n-gram vectors. To learn multilingual representations over multiple languages, Yu et al. [43] combined similarity constraints with a sequence-to-sequence model and reported its effectiveness on cross-lingual and zero-shot document classification tasks.

Finally, there has been a large body of work on obtaining transferable sentence representations. Conneau et al. [8] obtain representations from the supervised SNLI task and show that these are effective for transferring to other tasks. Their method outperforms other similar approaches to obtain representations like FastSent [44] and SkipThought [45]. Arora et al. [46] show that a simple average of word embeddings approach is competitive with more complex methods like SkipThought representations.

Cross-lingual or Multilingual Representations. Previous approaches to cross-lingual or multilingual representations have fallen into three categories. Obtaining representations from *word level alignments* - bilingual dictionaries or automatically generated word alignments - is the most popular approach [4, 47, 48]. The second category of methods try to leverage *document level alignment* like parallel Wikipedia articles to generate cross-lingual representations [32, 34]. The final category of methods use *sentence level alignments* in the form of parallel translation data to obtain cross-lingual representations. Hermann and Blunsom [33] propose a deep neural model named BiCVM which compares two sentence representations at the final layer and forces them into the same intermediate sentence representation. BilBOWA [49] is a simpler model which extends skip-gram with negative sampling [4] to optimize each word's similarity with its context in both the current language and the other parallel language. Luong et al. [50] also propose obtaining cross-lingual representations using a similar approach. Ammar et al. [51] propose two algorithms, multiCluster and multiCCA, for learning multilingual representations from a set of bilingual lexical data.

Here we combined the best of both worlds by learning contextualized representations which are multilingual in nature and explored its performance in the zero-shot classification tasks. We demonstrated that using the encoder from a multilingual NMT system as a pre-trained component in other downstream NLP tasks allows us to conduct cross-lingual transfer learning for an unseen language, i.e. French and supported our findings with further analysis.

7 Conclusion

In this paper, we have demonstrated a simple yet effective approach to perform zero-shot cross-lingual transfer learning using representations from a multilingual NMT model. Our proposed approach of reusing the encoder from a multilingual NMT system as a pre-trained component enables us to perform surprisingly competitive zero-shot classification on an unseen language and outperforms cross-lingual embedding base methods. Finally, we end with a series of analyses which shed light on the factors that contribute to the zero-shot phenomenon. We hope that these results showcase the efficacy of multilingual NMT to learn transferable contextualized and linguistically generalized representations for many downstream tasks.

References

[1] Krizhevsky, A., I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[2] Yosinski, J., J. Clune, Y. Bengio, et al. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

[3] Huh, M., P. Agrawal, A. A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[4] Mikolov, T., I. Sutskever, K. Chen, et al. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[5] Pennington, J., R. Socher, C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, Doha, Qatar, 2014.

[6] McCann, B., J. Bradbury, C. Xiong, et al. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc., 2017.

[7] Peters, M., M. Neumann, M. Iyyer, et al. 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. Association for Computational Linguistics, 2018.

[8] Conneau, A., D. Kiela, H. Schwenk, et al. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics, 2017.

[9] Johnson, M., M. Schuster, Q. Le, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[10] Luong, M., Q. V. Le, I. Sutskever, et al. Multi-task sequence to sequence learning. In *Proceedings of International Conference on Learning Representations*. 2016.

[11] Dong, D., H. Wu, W. He, et al. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732. Association for Computational Linguistics, 2015.

[12] Firat, O., K. Cho, Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL: HLT*, pages 866–875. 2016.

[13] Lee, J., K. Cho, T. Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017.

[14] Caruana, R. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997.

[15] Bahdanau, D., K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations*. 2016.

[16] Wu, Y., M. Schuster, Z. Chen, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv: 1609.08144*, 2016.

[17] Schuster, M., K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[18] Collobert, R., J. Weston, L. Bottou, et al. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[19] Sennrich, R., B. Haddow, A. Birch. 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. Association for Computational Linguistics, 2016.

[20] Schuster, M., K. Nakajima. Japanese and korean voice search. In *Proceedings of 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5149–5152. IEEE, 2012.

[21] Prettenhofer, P., B. Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127. Association for Computational Linguistics, Stroudsburg, PA, USA, 2010.

[22] Socher, R., A. Perelygin, J. Wu, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA, 2013.

[23] Agić, Ž., N. Schluter. Baselines and Test Data for Cross-Lingual Inference. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[24] Bowman, S. R., G. Angeli, C. Potts, et al. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics, 2015.

[25] Kingma, D. P., J. Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*. 2014.

[26] Szegedy, C., V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826. IEEE Computer Society, 2016.

[27] Ba, L. J., R. Kiros, G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[28] He, K., X. Zhang, S. Ren, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society, 2016.

[29] Zoph, B., K. Knight. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34. Association for Computational Linguistics, 2016.

[30] Lu, J., J. Yang, D. Batra, et al. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29*, pages 289–297. Curran Associates, Inc., 2016.

[31] Tai, K. S., R. Socher, C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566. Association for Computational Linguistics, 2015.

[32] Søgaard, A., v. Agić, H. Martínez Alonso, et al. Inverted indexing for cross-lingual nlp. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1713–1722. Association for Computational Linguistics, 2015.

[33] Hermann, K. M., P. Blunsom. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68. Association for Computational Linguistics, 2014.

[34] Vulić, I., M.-F. Moens. Bilingual distributed word representations from document-aligned comparable data. *Artificial Intelligence Research*, 55(1):953–994, 2016.

[35] Junczys-Dowmunt, M., T. Dwojak, R. Sennrich. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325. 2016.

[36] Nallapati, R., B. Xiang, B. Zhou. Sequence-to-sequence RNNs for text summarization. In *Proceedings of International Conference on Learning Representations Workshop*. 2016.

[37] Liu, X., Y. Shen, K. Duh, et al. Stochastic answer networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704. Association for Computational Linguistics, 2018.

[38] He, L., K. Lee, M. Lewis, et al. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483. Association for Computational Linguistics, 2017.

[39] Peters, M., W. Ammar, C. Bhagavatula, et al. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765. Association for Computational Linguistics, 2017.

[40] Wieting, J., M. Bansal, K. Gimpel, et al. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515. Association for Computational Linguistics, 2016.

[41] Bojanowski, P., E. Grave, A. Joulin, et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[42] Hill, F., K. Cho, S. Jean, et al. The representational geometry of word meanings acquired by neural machine translation models. *Machine Translation*, 31(1-2):3–18, 2017.

[43] Yu, K., H. Li, B. Oguz. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179. Association for Computational Linguistics, 2018.

[44] Hill, F., K. Cho, A. Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377. Association for Computational Linguistics, 2016.

[45] Kiros, R., Y. Zhu, R. R. Salakhutdinov, et al. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett, eds., *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015.

[46] Arora, S., Y. Liang, T. Ma. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of International Conference on Learning Representations*. 2017.

[47] Faruqui, M., C. Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471. Association for Computational Linguistics, 2014.

[48] Zou, W. Y., R. Socher, D. Cer, et al. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398. Association for Computational Linguistics, 2013.

[49] Gouws, S., Y. Bengio, G. Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In F. Bach, D. Blei, eds., *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37 of *Proceedings of Machine Learning Research*, pages 748–756. PMLR, Lille, France, 2015.

[50] Luong, T., H. Pham, C. D. Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of Workshop on Vector Space Modeling for NLP*, pages 151–159. Denver, Colorado, 2015.

[51] Ammar, W., G. Mulcaire, Y. Tsvetkov, et al. Massively multilingual word embeddings. *arxiv preprint arXiv:1602.01925*, 2016.

[52] Fernández, A. M., A. Esuli, F. Sebastiani. Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. *Artificial Intelligence Research*, 55:131–163, 2016.

A Supplementary Materials

A.1 Results on Transfer Learning

Here, we report the results of the proposed multilingual Encoder-Classifier for the three cross-lingual tasks - Amazon Reviews (English and French), SST, and SNLI, to investigate how effective the multilingual representations learned from the multilingual NMT model are. For each task, we first build a baseline system using the proposed *Encoder-Classifier* architecture described in Section 2 where the encoder parameters is initialized randomly and trained. Next, we experiment with using the pre-trained multilingual NMT encoder to initialize the system as described in Section 2.1. Finally, we perform an experiment where we freeze the encoder after initialization and only update the classifier component of the system.

Table 7 summarizes the accuracy of our proposed system for these three different approaches and the state-of-the-art results on all the tasks. The first row in the table shows the baseline accuracy of our system for all four datasets. The second row shows the result from initializing with a pre-trained multilingual NMT encoder. It can be seen that this provides a significant improvement in accuracy, an average of 4.63%, across all the tasks. This illustrates that the multilingual NMT encoder has successfully learned transferable contextualized representations that are leveraged by the classifier component of our proposed system. These results are in line with the results in [6] where the authors used the representations from the top NMT encoder layer as an additional input to the task-specific system. However, in our setup we reused all of the layers of the encoder as a single pre-trained component in the task-specific system. The third row shows the results from freezing the pre-trained encoder after initialization and only training the classifier component. For the Amazon English and French tasks, freezing the encoder after initialization significantly improves the performance further. We hypothesize that since the Amazon dataset is a document level classification task, the long input sequences are very different from the short sequences consumed by the NMT system, and hence freezing the encoder seems to have a positive effect. This hypothesis is also supported by the SNLI and SST results, which contain sentence-level input sequences, where we did not find any significant difference between freezing and not freezing the encoder.

Table 7: Transfer learning results of the classification accuracy on all the datasets. Amazon (En) and Amazon (Fr) are the English and French versions of the task, training the models on the data for each language. The state-of-the-art results are cited from [52] for both Amazon Reviews tasks and [6] for SST and SNLI.

Model	Amazon (En)	Amazon (Fr)	SST (En)	SNLI (En)
Proposed model: *Encoder-Classifier*	76.60	82.50	79.63	76.70
+ Pre-trained Encoder	80.70	83.18	84.18	84.42
+ Freeze Encoder	84.13	85.65	84.51	84.41
State-of-the-art Models	83.50	87.50	90.30	88.10