THE MISSING INGREDIENT FOR ZERO-SHOT NEURAL MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual Neural Machine Translation (NMT) systems are capable of translating between multiple source and target languages within a single system. An important indicator of generalization within these systems is the quality of zero-shot translation - translating between language pairs that the system has never seen during training. However, until now, the zero-shot performance of multilingual models has lagged far behind the quality that can be achieved by using a two step translation process that pivots through an intermediate language (usually English). In this work, we diagnose why multilingual models under-perform in zero shot settings. We propose explicit language invariance losses that guide an NMT encoder towards learning language agnostic representations. Our proposed strategies significantly improve zero-shot translation performance on WMT English-French-German and on the IWSLT 2017 shared task, and for the first time, match the performance of pivoting approaches while maintaining performance on supervised directions.

1 INTRODUCTION

In recent years, the emergence of sequence to sequence models has revolutionized machine translation. Neural models have reduced the need for pipelined components, in addition to significantly improving translation quality compared to their phrase based counterparts (Sutskever et al., 2014; Wu et al., 2016). These models naturally decompose into an encoder and a decoder with a presumed separation of roles: The encoder encodes text in the source language into an intermediate latent representation, and the decoder generates the target language text conditioned on the encoder representation. This framework allows us to easily extend translation to a multilingual setting, wherein a single system is able to translate between multiple languages (Dong et al., 2015; Luong et al., 2015a).

Multilingual NMT models have often been shown to improve translation quality over bilingual models, especially when evaluated on low resource language pairs (Firat et al., 2016a; Gu et al., 2018). Most strategies for training multilingual NMT models rely on some form of parameter sharing, and often differ only in terms of the architecture and the specific weights that are tied. They allow specialization in either the encoder or the decoder, but tend to share parameters at their interface. An underlying assumption of these parameter sharing strategies is that the model will automatically learn some kind of shared universally useful representation, or *interlingua*, resulting in a single model that can translate between multiple languages.

The existence of such a universal shared representation should naturally entail reasonable performance on zero-shot translation, where a model is evaluated on language pairs it has never seen together during training. Apart from potential practical benefits like reduced latency costs, zero-shot translation performance is a strong indicator of generalization. Enabling zero-shot translation with sufficient quality can significantly simplify translation systems, and pave the way towards a single multilingual model capable of translating between any two languages directly. However, despite being a problem of interest for a lot of recent research, the quality of zero-shot translation has lagged behind pivoting through a common language by 8-10 BLEU points (Firat et al., 2016b; Johnson et al., 2016; Ha et al., 2017; Lu et al., 2018). In this paper we ask the question, *What is the missing ingredient that will allow us to bridge this gap?*

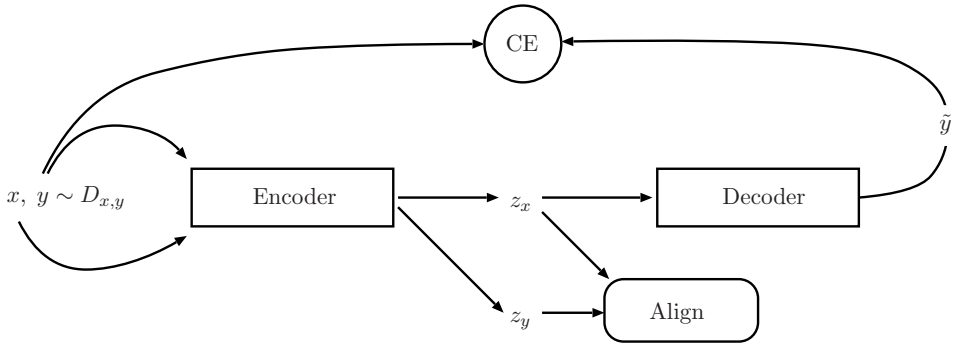

Figure 1: The proposed multilingual NMT model along with the two training objectives. CE stands for the cross-entropy loss associated with maximum likelihood estimation for translation between English and other languages. Align represents the source language invariance loss that we impose on the representations of the encoder. While training on the translation objective, training samples (x, y) are drawn from the set of parallel sentences, $D_{x,y}$. For the invariance losses, (x, y) could be drawn from $D_{x,y}$ for the cosine loss, or independent data distributions for the adversarial loss. Both losses are minimized simultaneously. Since we have supervised data only to and from English, one of x or y is always in English.

In Johnson et al. (2016), it was hinted that the extent of separation between language representations was negatively correlated with zero-shot translation performance. This is supported by theoretical and empirical observations in domain adaptation literature, where the extent of subspace alignment between the source and target domains is strongly associated with transfer performance (Ben-David et al., 2007; 2010; Ganin et al., 2016). Zero-shot translation is a special case of domain adaptation in multilingual models, where English is the source domain and other languages collectively form the target domain. Following this thread of domain adaptation and subspace alignment, we hypothesize that aligning encoder representations of different languages with that of English might be the missing ingredient to improving zero-shot translation performance.

In this work, we develop auxiliary losses that can be applied to multilingual translation models during training, or as a fine-tuning step on a pre-trained model, to force encoder representations of different languages to align with English in a shared subspace. Our experiments demonstrate significant improvements on zero-shot translation performance and, for the first time, match the performance of pivoting approaches on WMT English-French-German (en-fr-de) and the IWSLT 2017 shared task, in all zero shot directions, without any meaningful regression in the supervised directions.

We further analyze the model's representations in order to understand the effect of our explicit alignment losses. Our analysis reveals that tying weights in the encoder, by itself, is not sufficient to ensure shared representations. As a result, standard multilingual models overfit to the supervised directions, and enter a failure mode when translating between zero-shot languages. Explicit alignment losses incentivize the model to use shared representations, resulting in better generalization.

2 Alignment of Latent Representations

2.1 Multilingual Neural Machine Translation

Let $\mathbf{x} = (x_1, x_2 ... x_m)$ be a sentence in the source language and $\mathbf{y} = (y_1, y_2, ... y_n)$ be its translation in the target language. For machine translation, our objective is to learn a model, $p(\mathbf{y}|\mathbf{x}; \theta)$. In modern NMT, we use sequence-to-sequence models supplemented with an attention mechanism (Bahdanau et al., 2015) to learn this distribution. These sequence-to-sequence models consist of an encoder, $Enc(\mathbf{x}) = \mathbf{z} = (z_1, z_2, ... z_m)$ parameterized with θ_{enc}, and a decoder that learns to map from the latent representation $\mathbf{z}$ to $\mathbf{y}$ by modeling $p(\mathbf{y}|\mathbf{z}; \theta_{dec})$, again parameterized with θ_{dec}. This model is trained to maximize the likelihood of the available parallel data, $D_{x,y}$.

$$L_{CE}(\theta_{enc}, \theta_{dec}) = \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}\sim D_{x,y}}[-\log p(\mathbf{y}|\mathbf{x})] \quad (1)$$

In multilingual training we jointly train a single model(Lee et al., 2016) to translate from many possible source languages to many potential target languages. When only the decoder is informed about the desired target language, a special token to indicate the target language, $< tl >$, is input to the first step of the decoder. In this case, $D_{x,y}$ is the union of all the parallel data for each of the supervised translation directions. Note that either the source or the target is always English.

2.2 Explicit Alignment of Encoder Representations

For zero-shot translation to work, the encoder needs to produce language invariant feature representations of a sentence. Previous works learn these transferable features by using a weight sharing constraint and tying the weights of the encoders, the decoders, or the attentions across some or all languages (Dong et al., 2015; Johnson et al., 2016; Lu et al., 2018; Firat et al., 2016a). They argue that sharing these layers across languages causes sentences that are translations of each other to cluster together in a common representation space. However, when a model is trained on just the end-to-end translation objective, there is no explicit incentive for the model to discover language invariant representations; given enough capacity, it is possible for the model to partition its intrinsic dimensions and overfit to the supervised translation directions. This would result in intermediate encoder representations that are specific to individual languages.

We now explore two classes of regularizers, Ω, that explicitly force the model to make the representations in all other languages similar to their English counterparts. We align the encoder representations of every language with English, since it is the only language that gets translated into all other languages during supervised training. Thus, English representations now form an implicit pivot in the latent space. The loss function we then minimize is:

$$L = L_{CE} + \lambda\Omega \quad (2)$$

where L_{CE} is the cross-entropy loss and λ is a hyper-parameter that controls the contribution of the alignment loss Ω.

2.2.1 Unsupervised: Adversarial Representation Alignment

Here we view zero-shot translation through the lens of domain adaptation, wherein English is the source domain and the other languages together constitute the target domain. Ben-David et al. (2007) and Mansour et al. (2009) have shown that target risk can be bounded by the source risk plus a discrepancy metric between the source and target feature distribution. Treating the encoder as a deterministic feature extractor, the source distribution is $Enc(\mathbf{x_{en}})p(\mathbf{x_{en}})$ and the target distribution is $Enc(\mathbf{x_t})p(\mathbf{x_t})$. To enable zero-shot translation, our objective then is to minimize the discrepancy between these distributions by explicitly optimizing the following domain adversarial loss (Ganin et al., 2016):

$$\Omega_{adv}(\theta_{disc}) = -\mathbb{E}_{\mathbf{x_{en}}\sim D_{En}}[-\log Disc(Enc(\mathbf{x_{en}}))] + \mathbb{E}_{\mathbf{x_t}\sim D_T}[-\log(1 - Disc(Enc(\mathbf{x_t})))] \quad (3)$$

where $Disc$ is the discriminator and is parametrized by θ_{disc}. D_{En} are English sentences and D_T are the sentences of all the other languages. Note that, unlike Artetxe et al. (2018); Yang et al. (2018), who also train the encoder adversarially with a language detecting discriminator, we are trying to align the distribution of encoder representations of all other languages to that of English and vice-versa. Our discriminator is just a binary predictor, independent of how many languages we are jointly training on.

Architecturally, the discriminator is a feed-forward network that acts on the temporally max-pooled representation of the encoder output. We also experimented with a discriminator that made independent predictions for the encoder representation, z_i, at each time-step i, but found the pooling based approach to work better. More involved discriminators that consider the sequential nature of the encoder representations may be more effective, but we do not explore them in this work.

2.2.2 SUPERVISED: ALIGNMENT OF KNOWN PARALLEL DATA

While adversarial approaches have the benefit of not needing parallel data, they only align the marginal distributions of the encoder's representations. Further, adversarial approaches are hard to optimize and are often susceptible to mode collapse, especially when the distribution to be modeled is multi-modal. Even if the discriminator is fully confused, there are no guarantees that the two learned distributions will be identical (Arora & Zhang, 2017).

To resolve these potential issues, we attempt to make use of the available parallel data, and enforce an instance level correspondence between the pairs $(\mathbf{x}, \mathbf{y}) \in D_{x,y}$, rather than just aligning the marginal distributions of $Enc(\mathbf{x})p(\mathbf{x})$ and $Enc(\mathbf{y})p(\mathbf{y})$ as in the case of domain-adversarial training. Previous work on multi-modal and multi-view representation learning has shown that, when given paired data, transferable representations can be learned by improving some measure of similarity between the corresponding views from each mode. Various similarity measures have been proposed such as Euclidean distance (Ham et al., 2005), cosine distance (Frome et al., 2013), correlation (Andrew et al., 2013) etc. In our case, the different views correspond to equivalent sentences in different languages.

Note that $Enc(\mathbf{x})$ and $Enc(\mathbf{y})$ are actually a pair of sequences, and to compare them we would ideally have access to the word level correspondences between the two sentences. In the absence of this information, we make a bag-of-words assumption and align the pooled representation similar to Gouws et al. (2015b); Coulmance et al. (2016). Empirically, we find that max pooling and minimizing the cosine distance between the representations of parallel sentences similar to works well. We now minimize the distance function:

$$\Omega_{sim} = -E_{\mathbf{x},\mathbf{y}\sim D_{x,y}}[sim(Enc(\mathbf{x}), Enc(\mathbf{y}))] \tag{4}$$

3 EXPERIMENTS

A multilingual model with a single encoder and a single decoder similar to Johnson et al. (2016) is our baseline. This setup maximally enforces the parameter sharing constraint that previous works rely on to promote cross-lingual transfer. We first train our model solely on the translation loss until convergence, on all languages to and from English. This is our baseline multilingual model. We then fine-tune this model with the proposed alignment losses, in conjunction with the translation objective. We then compare the performance of the baseline model against the aligned models on both the supervised and the zero-shot translation directions. We also compare our zero-shot performance against the pivoting performance using the baseline model.

3.1 EXPERIMENTAL SETUP

For our en$\leftrightarrow${fr, de} experiments, we train our models on the standard en$\rightarrow$fr (39M) and en$\rightarrow$de (4.5M) training datasets from WMT'14. We pre-process the data by applying the standard Moses pre-processing[1]. We swap the source and target to get parallel data for the fr$\rightarrow$en and de$\rightarrow$en directions. The resulting datasets are merged by oversampling the German portion to match the size of the French portion. This results in a total of 158M sentence pairs. We get word counts and apply 32k BPE (Sennrich et al., 2016) to obtain subwords. The target language $< tl >$ tokens are also added to the vocabulary. We use newstest-2012 as the dev set and newstest-2013 as the test set. Both of these sets are 3-way parallel and have 3003 and 3000 sentences respectively.

We run all our experiments with Transformers (Vaswani et al., 2017), using the TransformerBase config. We train our model with a learning rate of 1.0 and 4000 warmup steps. Input dropout is set to 0.1. We use synchronized training with 16 Tesla P100 GPUs and train the model for 500k steps. The model is instructed on which language to translate a given input sentence into, by feeding in a unique $< tl >$ token per target language. In our implementation, this token is pre-pended into the source sentence, but it could just as easily be fed into the decoder to the same effect.

[1]We use `normalize-punctuation.perl`, `remove-non-printing-char.perl`, and `tokenizer.perl`.

For the alignment experiments, we fine-tune a pre-trained multilingual model by jointly training on both the alignment and translation losses. For adversarial alignment, the discriminator is a feed-forward network with 3 hidden layers of dimension 2048 using the leaky ReLU($\alpha = 0.1$) non-linearity. λ was tuned to 1.0 for both the adversarial and the cosine alignment losses. Simple fine-tuning with SGD using a learning rate of 1e-4 works well and we do not need to train from scratch. We observe that the models converge within a few thousand updates.

3.2 RESULTS

	$de \rightarrow fr$	$fr \rightarrow de$	$en \rightarrow fr$	$en \rightarrow de$	$fr \rightarrow en$	$de \rightarrow en$
Direct translation	16.80 (zs)	12.03 (zs)	32.68	24.48	32.33	30.26
Pivot through English	**26.25**	**20.18**	-	-	-	-
adversarial	**26.00** (zs)	**20.39** (zs)	32.92	24.50	32.39	30.21
pool-cosine	**25.85** (zs)	**20.18** (zs)	32.94	24.51	32.36	30.32

Table 1: Zero-shot results with baseline and aligned models compared against pivoting. Zero-Shot results are marked zs. Pivoting through English is performed using the baseline multilingual model.

Our results, in Table 1, demonstrate that both our approaches to align representations result in large improvements in zero-shot translation quality for both directions, effectively closing the gap to the performance of the strong pivoting baseline. We didn't notice any significant differences between the performance of the two proposed alignment methods. Importantly, these improvements come at no cost to the quality in the supervised directions.

While both the proposed approaches aren't significantly different in terms of final quality, we noticed that the adversarial regularizer was very sensitive to the initialization scheme and the choice of hyper-parameters. In comparison, the cosine distance loss was relatively stable, with λ being the only hyper-parameter controlling the weight of the alignment loss with respect to the translation loss.

4 ANALYSIS: WHY ALIGNMENT WORKS

We further analyze the outputs of our baseline multilingual model in order to understand the effect of alignment on zero-shot performance. We identify the major effects that contribute to the poor zero-shot performance in multilingual models, and investigate how an explicit alignment loss resolves these pathologies.

4.1 CASCADED DECODER ERRORS

	en	de	fr
$de \rightarrow fr$	14%	25%	60%
$fr \rightarrow de$	12%	54%	34%
$de \rightarrow en \rightarrow fr$	5%	0%	95%
$fr \rightarrow en \rightarrow de$	6%	94%	0%
fr references	4%	0%	96%
de references	4%	96%	0%

Table 2: Percentage of sentences by language in reference translations and the sentences decoded using the baseline model (newstest2012)

While investigating the high variance of the zero-shot translation score during multilingual training in the absence of alignment, we found that a significant fraction of the examples were not getting translated into the desired target language at all. Instead, they were either translated to English or simply copied. This phenomenon is likely a consequence of the fact that at training time, German and French source sentences were always translated into English. Because of this, the model never learns to properly attribute the target language to the $< tl >$ token, and simply changing the $< tl >$

token at test time is not effective. We count the number of sentences in each language using an automatic language identification tool and report the results in Table 2.

Further, we find that for a given sentence, all output tokens tend to be in the same language, and there is little to no code-switching. This was also observed by Johnson et al. (2016), where it was explained as a cascading effect in the decoder: Once the decoder starts emitting tokens in one language, the conditional distribution $p(y_i|y_{i-1}, ..., y_1)$ is heavily biased towards that particular language. With explicit alignment, we remove the target language information encoded into the source token representations. In the absence of this confounding information, the $< tl >$ target token gives us more control to set the translation direction.

4.2 Improved adaptation performance

	# examples	Pivot (baseline)	Zero-Shot (baseline)	Zero-Shot (adversarial)
$de \rightarrow fr$	1875/3003	19.71	19.22	19.93
$fr \rightarrow de$	1591/3003	24.33	21.63	23.87

Table 3: BLEU on subset of examples predicted in the right language by the direct translation using the baseline system (newstest2012)

Here we try to isolate the gains our system achieves due to improvements in the learning of transferable features, from those that can be attributed to decoding to the desired language. We discount the errors that could be attributed to incorrect language errors and inspect the translation quality on the subset of examples where the baseline model decodes in the right language. We re-evaluate the BLEU scores of all systems and show the results in Table 3. We find that the vanilla zero-shot translation system (Baseline) is much stronger than expected at first glance. It only lags the pivoting baseline by 0.5 BLEU points on French to German and by 2.7 BLEU points on German to French. We can now see that, even on this subset which was chosen to favor the baseline model, the representation alignment of our adapted model contributes to improving the quality of zero-shot translation by 0.7 and 2.2 BLEU points on French to German and German to French, respectively.

4.3 Improving the Language Invariance of Multilingual Encoders

We design a simple experiment to determine whether representations learned while training a multilingual translation model are truly cross-lingual. We probe our baseline and aligned multilingual models with 3-way aligned data to determine the extent to which their representations are functionally equivalent, during different stages in model training. Because source languages can have different sequence lengths and word orders for equivalent sentences, it is not possible to directly compare encoder output representations.

However, it is possible to directly compare the representations extracted by the decoder from the encoder outputs for each language. Suppose we want to compare representations of semantically equivalent English and German sentences when translating into French. At time-step i in the decoder, we use the model to predict $p(y_i|Enc(\mathbf{x_{en}}), y_{1:(i-1)})$ and $p(y_i|Enc(\mathbf{x_{de}}), y_{1:(i-1)})$. However, in the seq2seq with attention formulation, these problems reduce to predicting $p(y_i|c_i^{en}, y_{1:(i-1)})$ and $p(y_i|c_i^{de}, y_{1:(i-1)})$, where c_i^{en} and c_i^{de} are the attention context vectors extracted from $Enc(\mathbf{x_{en}})$ and $Enc(\mathbf{x_{de}})$, respectively. Given the same set of $y_{1:(i-1)}$, with teacher forcing, c_i^{en} and c_i^{de} should be identical if our encoder is truly language agnostic.

We use a randomly sampled set of 100 parallel en-de-fr sentences extracted from our dev set, newstest2012, to perform this analysis. For each set of aligned sentences, we obtain the sequence of aligned context vectors (c_i^{en}, c_i^{de}) and plot the mean cosine distances for our baseline training run, and the incremental runs with alignment losses in Figure 2. Our results indicate that the vanilla multilingual model learns to align encoder representations over the course of training. However, in the absence of an external incentive, the alignment process arrests as training progresses. Incrementally training with the alignment losses results in a more language-agnostic representation, which contributes to the improvements in zero-shot performance.

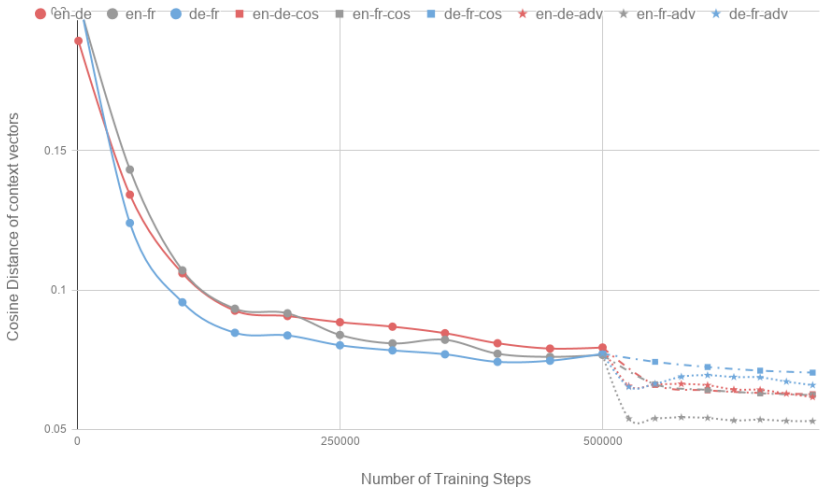

Figure 2: Average cosine distance between aligned context vectors for all combinations of English (en), German (de) and French (fr) as training progresses.

4.4 Scaling to More Languages

Given the good results on WMT en-fr-de, we now extend our experiments, to test the scalability of our approach to multiple languages. We work with the IWSLT-17 dataset which has transcripts of Ted talks in 5 languages: English (en), Dutch (nl), German (de), Italian (it), and Romanian (ro). The original dataset is multi-way parallel with approximately 220 thousand sentences per language, but for the sake of our experiments we only use the to/from English directions for training. The dev and test sets are also multi-way parallel and comprise around 900 and 1100 sentences per language pair respectively. We again use the transformer base architecture. We set the learning rate to 2.0 and the number of warmup steps to 8k. A dropout rate of 0.2 was applied to all connections of the transformer. We use the cosine loss with λ set to 0.001 because of how easy it is to tune.

	vanilla		cosine
	direct	pivot	direct
English to/from (8)	30.11	-	29.95
Non-English to/from (12)	16.73 (zs)	17.76	**17.72** (zs)
All directions (20)	22.2	22.81	22.72

Table 4: Average BLEU scores for IWSLT-2017; Zero-Shot results are marked (zs).

Our baseline model's scores on IWSLT-17 are suspiciously close to that of bridging, as seen in Table 4. We suspect this is because the data that we train on is multi-way parallel, and the English sentences are shared across the language pairs. This may be helping the model learn shared representations with the English sentences acting as pivots. Even so, we are able to gain 1 BLEU over the strong baseline system and demonstrate the applicability of our approach to larger groups of languages.

5 Related Work

5.1 Multilingual Translation

Multilingual NMT models were first proposed by Dong et al. (2015) and have since been explored in Firat et al. (2016a); Blackwood et al. (2018) and several other works. While zero-shot translation was the direct goal of Firat et al. (2016a), they were only able to achieve 'zero-resource translation', by using their pre-trained multi-way multilingual model to generate pseudo-parallel data for

fine-tuning. Johnson et al. (2016) were the first to show the possibility of zero-shot translation by proposing a model that shared all the components and used a token to indicate the target language. Platanios et al. (2018) propose a novel way to modulate the amount of sharing between languages, by using a parameter generator to generate the parameters for either the encoder or the decoder of the multilingual NMT system based on the source and target languages. They also report higher zero-shot translation scores with this approach.

5.2 Shared subspace learning

Learning coordinated representations with the use of parallel data has been explored thoroughly in the context of multi-view and multi-modal learning (Baltrušaitis et al., 2018). These often involve either auto-encoder like networks with a reconstruction objective, or paired feed-forward networks with a similarity based objective (Wang et al., 2015). This function used to encourage similarity may be Euclidean distance (Ham et al., 2005), cosine distance (Frome et al., 2013), partial order (Vendrov et al., 2015), correlation (Andrew et al., 2013), etc. More recently a vast number of adversarial approaches have been proposed to learn domain invariant representations, by ensuring that they are indistinguishable by a discriminator network (Ganin et al., 2016).

The use of aligned parallel data to learn shared representations is common in the field of cross-lingual or multilingual representations, where work falls into three main categories. Obtaining representations from *word level alignments* - bilingual dictionaries or automatically generated word alignments - is the most popular approach (Mikolov et al., 2013; Faruqui & Dyer, 2014; Zou et al., 2013). The second category of methods try to leverage *document level alignment*, like parallel Wikipedia articles, to generate cross-lingual representations (Søgaard et al., 2015; Vulić & Moens, 2016). The final category of methods often use *sentence level alignments*, in the form of parallel translation data, to obtain cross-lingual representations (Hermann & Blunsom, 2014; Gouws et al., 2015a; Mikolov et al., 2013; Luong et al., 2015b; Ammar et al., 2016). Recent work by Eriguchi et al. (2018) showed that the representations learned by a multilingual NMT system are widely applicable across tasks and languages.

5.3 Unsupervised Neural Machine Translation

Parameter sharing based approaches have also been tried in the context of unsupervised NMT, where learning a shared latent space (Artetxe et al., 2017) was believed to improve translation quality. Some approaches explore applying adversarial losses on the encoder, to ensure that the representations are language agnostic. However, recent work has shown that enforcing a shared latent space is not important for unsupervised NMT (Lample et al., 2018), and the cycle consistency loss suffices by itself.

6 Conclusion

In this work we propose explicit alignment losses, as an additional constraint for multilingual NMT models, with the goal of improving zero-shot translation. We view the zero-shot NMT problem in the light of subspace alignment for domain adaptation, and propose simple approaches to achieve this. Our experiments demonstrate significantly improved zero-shot translation performance that are, for the first time, comparable to strong pivoting based approaches. Through careful analyses we show how our proposed alignment losses result in better representations, and thereby better zero-shot performance, while still maintaining performance on the supervised directions. Our proposed methods have been shown to work reliably on two public benchmarks datasets: WMT English-French-German and the IWSLT 2017 shared task.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. Massively multilingual word embeddings. *arxiv preprint arXiv:1602.01925*, 2016. URL http://arxiv.org/abs/1602.01925.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pp. 1247–1255, 2013.

Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *CoRR*, abs/1706.08224, 2017. URL http://arxiv.org/abs/1706.08224.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *CoRR*, abs/1710.11041, 2017. URL http://arxiv.org/abs/1710.11041.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy2ogebAW.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1409.0473.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. Multilingual neural machine translation with task-specific attention. *arXiv preprint arXiv:1806.03280*, 2018.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. Trans-gram, fast cross-lingual word-embeddings. *arXiv preprint arXiv:1601.02502*, 2016.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 1723–1732, 2015.

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

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *EACL*, pp. 462–471, Gothenburg, Sweden, April 2014. URL http://www.aclweb.org/anthology/E14-1049.

Orhan Firat, KyungHyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. *CoRR*, abs/1601.01073, 2016a. URL http://arxiv.org/abs/1601.01073.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 268–277, 2016b. URL http://aclweb.org/anthology/D/D16/D16-1026.pdf.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML - Volume 37*, pp. 748–756, 2015a. URL http://dl.acm.org/citation.cfm?id=3045118.3045199.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pp. 748–756, 2015b.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*, 2018.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Effective strategies in zero-shot neural machine translation. *arXiv preprint arXiv:1711.07893*, 2017.

Jihun Ham, Daniel D Lee, and Lawrence K Saul. Semisupervised alignment of manifolds. In *AISTATS*, pp. 120–127, 2005.

Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. In *ACL*, pp. 58–68, Baltimore, Maryland, June 2014. URL http://www.aclweb.org/anthology/P14-1006.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *CoRR*, abs/1804.07755, 2018. URL http://arxiv.org/abs/1804.07755.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*, 2016.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. A neural interlingua for multilingual machine translation. *arXiv preprint arXiv:1804.08198*, 2018.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *International Conference on Learning Representations*, 2015a.

Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of Workshop on Vector Space Modeling for NLP*, pp. 151–159, Denver, Colorado, June 2015b. URL http://www.aclweb.org/anthology/W15-1521.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 26*, pp. 3111–3119. 2013. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. Contextual parameter generation for universal neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, November 2018. URL https://arxiv.org/abs/1808.08493.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. Inverted indexing for cross-lingual nlp. In *ACL and IJCNLP*, pp. 1713–1722, Beijing, China, July 2015. URL http://www.aclweb.org/anthology/P15-1165.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Ivan Vulić and Marie-Francine Moens. Bilingual distributed word representations from document-aligned comparable data. *Artificial Intelligence Research*, 55(1):953–994, January 2016. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=3013558.3013583.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pp. 1083–1092, 2015.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing. *CoRR*, abs/1804.09057, 2018. URL http://arxiv.org/abs/1804.09057.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pp. 1393–1398, 2013. URL http://www.aclweb.org/anthology/D13-1141.