Multilingual Neural Machine Translation with Task-Specific Attention

Graeme Blackwood **Miguel Ballesteros** **Todd Ward**
IBM Research AI
1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA
{blackwood,toddward}@us.ibm.com, miguel.ballesteros@ibm.com

Abstract

Multilingual machine translation addresses the task of translating between multiple source and target languages. We propose task-specific attention models, a simple but effective technique for improving the quality of sequence-to-sequence neural multilingual translation. Our approach seeks to retain as much of the parameter sharing generalization of NMT models as possible, while still allowing for language-specific specialization of the attention model to a particular language-pair or task. Our experiments on four languages of the Europarl corpus show that using a target-specific model of attention provides consistent gains in translation quality for all possible translation directions, compared to a model in which all parameters are shared. We observe improved translation quality even in the (extreme) low-resource zero-shot translation directions for which the model never saw explicitly paired parallel data.

1 Introduction and Motivation

Multilingual machine translation is the task of building a system capable of translating between more than just a single source and target language. The approach is motivated by the idea that learning to translate between one pair of languages can help to improve the translation quality of related language pairs. Multilingual MT can be divided into three main types according to the support for source and target languages: (i) single source, multiple target (ii) multiple source, single target, and (iii) multiple source, multiple target. Our work focuses on improving the quality of fully n-way multilingual NMT models that can translate between multiple source and target languages.

Multinational organizations and companies increasingly publish content in a variety of languages. In addition to exploiting multiple sources of parallel data and leveraging similarities across different languages, multilingual translation can considerably simplify deployment. Directly supporting all possible translation directions for a set of n languages would require $n(n-1)$ sets of parallel data and trained models. This is an expensive proposal, from both a data and deployment perspective. The number of required systems can be reduced to n if translation by bridging is used. In contrast, a fully n-way multilingual system can be trained to translate between all known language pairs using a single model. It is even possible to translate between language pairs that were never explicitly paired in the parallel training data, so called zero-shot translation (Johnson et al., 2017).

Neural machine translation (NMT) (Bahdanau et al., 2014; Sutskever et al., 2014) has recently become the standard model of translation due to the high levels of fluency and accuracy that it can achieve (Koehn and Knowles, 2017). NMT is an excellent choice for multilingual translation since the neural architecture is language-agnostic and capable of capturing translation properties, such as long-distance re-ordering, between even highly dissimilar languages. One of the main advantages of NMT is that the model is able to learn useful linguistic representations for many languages, and to share parameters and leverage similarities in the abstractions learned by the embeddings and hidden layers of the model. Exploiting multilingual data and representations is of particular interest in improving the quality of translation for low-resource (or even zero-resource) language pairs.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3112–3122
Santa Fe, New Mexico, USA, August 20-26, 2018.

In this paper, we describe a simple but effective extension to the attentional model of neural machine translation that improves the quality of multilingual NMT. We seek to retain as much of the parameter sharing of NMT models as possible, while allowing for language-specific specialization of the attention model to a particular task. Our experiments on the Europarl corpus show that multilingual translation quality is improved for all tested language pairs, with the largest improvements in the (extreme low-resource) zero-shot directions.

The architecture of our sequence-to-sequence neural machine translation system is described in Section 2. An overview of related work in multilingual machine translation and multi-task learning is provided in Section 3. In Section 4, we introduce three task-specific attention model variants that can be used to improve the quality of multilingual NMT. Our experimental framework, implementation and results are described in Section 5, followed by an analysis and example translations with attention plots in Section 6. We conclude with a discussion of possible future work in Section 7.

2 Sequence to Sequence Translation with Attention

Our NMT system is based on the sequence-to-sequence model of translation described by Bahdanau et al. (2014), consisting of a recurrent neural network encoder and decoder with an attention mechanism.

The encoder uses bi-directional gated recurrent units (GRU) (Cho et al., 2014) to encode a source sequence $\mathbf{x} = (x_1, \ldots, x_l)$, where x_i is the embedding vector for the i-th word of the source sentence and l its length. The encoded form of the complete sentence is defined by the sequence of hidden states $\mathbf{h} = (h_1, \ldots, h_l)$, where each h_i is computed as

$$h_i = \begin{bmatrix} \overleftarrow{h}_i \\ \overrightarrow{h}_i \end{bmatrix} = \begin{bmatrix} \overleftarrow{f}(x_i, \overleftarrow{h}_{i+1}) \\ \overrightarrow{f}(x_i, \overrightarrow{h}_{i-1}) \end{bmatrix}, \tag{1}$$

and $\overleftarrow{f}$ and $\overrightarrow{f}$ denote the operations of the right-to-left and left-to-right GRU cells, respectively.

Given the encoded representation of the source sentence $\mathbf{h}$, the decoder produces the target translation $\mathbf{y}$ by computing the sequence $\mathbf{y} = (y_1, \ldots, y_m)$, where m is the target sentence length. At each time-step t, the probability of target word y_t is computed as

$$p(y_t|\mathbf{h}, y_{t-1}, \ldots, y_1) = g(s_t, y_{t-1}, H_t), \tag{2}$$

where g is a feed-forward network over the current decoder hidden state s_t, the word embedding of the previously predicted target word y_{t-1}, and a context vector H_t, followed by a softmax to predict the probability distribution over the output vocabulary. The decoder states are updated according to

$$s_t = q(s_{t-1}, y_{t-1}, H_t), \tag{3}$$

where q implements the conditional GRU with attention of Sennrich et al. (2017) and H_t is an attention-weighted representation of the encoded source sequence $\mathbf{h}$. The attention-weighted source representation at time-step t is defined as the weighted sum

$$H_t = \begin{bmatrix} \sum_{i=1}^{l}(\alpha_{t,i} \cdot \overleftarrow{h}_i) \\ \sum_{i=1}^{l}(\alpha_{t,i} \cdot \overrightarrow{h}_i) \end{bmatrix}, \tag{4}$$

where the normalized weights $\alpha_{t,i}$ indicate the relative importance of the vector representation of each source position h_i when producing the next target word at time-step t. The weights are computed using a two-layer feed-forward network r:

$$\alpha_{t,i} = \frac{\exp\{r(s_{t-1}, h_i, y_{t-1})\}}{\sum_k \exp\{r(s_{t-1}, h_k, y_{t-1})\}}. \tag{5}$$

3 Related Work

Early phrase-based approaches to multilingual MT focused on multi-source translation. Och and Ney (2002) used a simple product or max rule to select at the sentence-level the single best hypothesis with the highest translation score from multiple decoders. There is no sharing of parameters. Consensus network decoding (Matusov et al., 2006) can also be used to combine the word-level output of translations from multiple source languages. Such system combination techniques allow sharing of the words in the candidate translations, but still require training individual models for each language pair of interest.

Training multilingual MT systems capable of translating between multiple languages can be considered an instance of multi-task learning (Caruana, 1997), which is the idea of solving synergistic tasks while maximizing the number of shared parameters. Sharing parameters may be useful when attempting to solve different tasks, since we can minimize representation bias by learning a more regularized representation (Baxter, 2000). The flexibility nature of neural architectures allows for selection of the components of the model that are to be shared and those which are not.

Sequence to sequence models with attention are no exception. Each set of parameters provides different levels of generalization (Reimers and Gurevych, 2017), which is evidenced in the synergistic task of training multilingual translation models. For example, Dong et al. (2015) jointly train decoders while the rest of the parameters are task-specific; Zoph and Knight (2016) jointly train the encoders while the rest of the parameters are task-specific, and Johnson et al. (2017) train both encoders and decoders jointly with language-specific tokens to guide learning as in (Ammar et al., 2016). These latter approaches are the ones that we build on. We augment our decoder with a task-specific attention mechanism intended to better capture word order and language-specific nuances while continuing to share the rest of the model parameters (including token embeddings).

4 Task-Specific Attention Models

Fully n-way multilingual NMT systems need to support multiple source and target languages in the encoder and decoder GRUs. Our work focuses on improving the use of attention in the decoder. At each time-step t, the decoder computes the attention model weights $\alpha_{t,i}$ of Equation (5) in order to quantify the relative importance of the encoder states corresponding to each source position. To the extent that attention can be considered an analogue of word alignment, we would expect to see different patterns of attention for language-pairs with different word order and word-level alignments. Multilingual models restricted to a single shared attention will struggle when decoding multiple languages with different word orders, since the parts of the source sentence that should be attended to depend on the source and target languages of the translation task.

Our task-specific attention models can be used to address this issue. We train and empirically evaluate three task-specific attention model variants: *target-specific* attention, *source-specific* attention, and *paired* attention which is associated with a particular language pair (i.e. translation direction). Each of our models introduces conditioning on the weights and biases used to compute the attention coefficients. They differ only in the choice of the key used for conditioning attention:

- *Target-specific*: separate attention weights and bias for each target language
- *Source-specific*: separate attention weights and bias for each source language
- *Paired*: separate attention weights and bias for each source + target pair

Our multilingual NMT system follows Johnson et al. (2017) in using special tokens to indicate the desired target language. These tokens can be considered to define the 'task' from a multi-task learning perspective as shown in (Ammar et al., 2016; Kann et al., 2018). Since the encoder states are composed from bi-directional GRUs, we augment the source side of our parallel training data with both prefix and suffix task tokens. This ensures that the task token is not attenuated in the left-to-right encoder GRU. Such tokens are sufficient to ensure that the multilingual decoder produces words in the correct target language[1]. A French source sentence that is to be translated into English would be augmented as follows:

[1]We tried explicit softmax decoding constraints on the target vocabulary, but found them to be unnecessary.

target-specific	<ToEn>, <ToFr>, <ToEs>, <ToDe>
source-specific	<FromEn>, <FromFr>, <FromEs>, <FromDe>
paired	<FrEn>, <EnFr>, <EsEn>, <EnEs>, <DeEn>, <EnDe>

Table 1: Valid source-side task tokens for task-specific attention model training and decoding.

target-specific	<ToFr> w_1 w_2 ... w_l <ToFr>
source-specific	<FromFr> <ToEn> w_1 w_2 ... w_l <ToEn>
paired	<FrEn> w_1 w_2 ... w_l <FrEn>

Table 2: Source side training data augmented with task-specific attention model tokens.

<ToEn> Guide des industries canadiennes : <ToEn>

During training and decoding, we dynamically construct the computation graph using the parameters for each task. Table 1 defines the valid task tokens for target-specific, source-specific and paired attention models, given the four languages supported by our multilingual model. See Section 5 for full details of our experimental framework and supported language pairs.

The augmented source sides of the parallel training data for our three model variants take the abstract forms shown in Table 2. The target-specific and paired attention models simply add the desired target language or language pair prefix and suffix tokens. For the source-specific attention model, we want to continue to allow for run-time selection of the desired target language so we introduce a second prefix token to indicate the selection of the specific attention parameters. The first token for the source-specific models is used only to select the parameters and stripped from the source sentence before using it in training or decoding. This allows us to dynamically select the source-specific attention parameters associated with translation from French, while (i) still allowing run-time selection of the target language, (ii) supporting zero-shot directions, and (iii) ensuring that the multilingual model sees exactly the same sequence of tokens as the target-specific model.

Note that only the target-specific and source-specific model variants enable zero-shot translation. The paired form of attention model trains separate attention weights and biases for each of the translation directions observed in the training data. It would be possible to use a separate prefix token for the attention key (in a similar manner to that of source-specific attention), but there is no explicit set of attention parameters that should be used for the zero-shot directions. The prefix token could specify that the attention parameters associated with either the source or target language of the zero-shot direction should be used.

All three of our task-specific attention model variants still share most of their encoder and decoder parameters. The task-specific attention models require only a very small increase in the total number of parameters. For hidden state dimensionality d, we add one additional set of attention weights ($d \times d$ parameters) and bias (d parameters) for each supported task. For the neural network topology used in our experiments (see Section 5, below), a target-specific attention model with support for four distinct target languages (i.e. tasks) requires only a 1.2% increase in the total number of model parameters, compared to the shared-attention version of the model.

5 Experiments

We evaluate the quality of our multilingual translation models using training data from the Europarl Corpus[2], Release V7. Our experiments focus on three primary language pairs: French-English, Spanish-English and German-English. We include German with SOV-type word order to contrast the SVO-type word order of the other languages. The source and target sides of the parallel training data are processed with the standard tokenizer included in the Moses SMT Toolkit (Koehn et al., 2007). For

[2] http://www.statmt.org/europarl/

	Sentence Pairs	Source Tokens	Target Tokens
French-English	2.01m	62.6m	56.3m
Spanish-English	1.97m	57.1m	55.0m
German-English	1.92m	51.0m	53.6m
Merged	11.79m	335.6m	335.6m

Table 3: Tokenized corpus statistics for training single-data baselines and n-way multilingual models.

	Fr-En	Es-En	De-En	Merged
source (BPE)	35.3k	37.2k	43.1k	80.0k
target (BPE)	35.3k	35.2k	35.1k	80.0k

Table 4: Source and target vocabulary sizes for single-data baselines and n-way multilingual models.

training multilingual systems, we merge the parallel data for all available directions. The parallel data for each language pair is thus included twice, once in each direction.

In order to support experiments on many-to-many multilingual translation using a single model, we apply a jointly-learned set of 80k Byte-Pair Encoding (BPE) rules (Sennrich et al., 2015) obtained from the merged source and target sides of the training data for all three language pairs. This ensures that all experiments, including the single-language baselines, share exactly the same tokenization and sub-word vocabularies. Tokenized corpus statistics for the parallel training data are summarized in Table 3, while Table 4 compares the vocabulary sizes obtained from BPE processed data for the single-data baseline and fully n-way multilingual models.

The Europarl evaluation data set dev2006 is used as our validation set, while devtest2006 and test2007 are our blind test sets. We also evaluate the out-of-domain News Commentary test sets nc-dev2007 and nc-devtest2007 in order to demonstrate the robustness of our approach to different kinds of data. Case-sensitive single-reference BLEU scores (Papineni et al., 2002) are computed using the `multi-bleu.perl` script included with Moses. Reimers and Gurevych (2017) have shown that reporting a single metric score can sometimes be misleading for many neural architectures. For this reason, all BLEU scores reported in the tables are obtained by averaging the decoding results from five separate models initialized with distinct random seeds.

5.1 Implementation Details

Our sequence-to-sequence NMT model with task-specific attention is implemented in C++ using DyNet[3] (Neubig et al., 2017a). We use DyNet since the computation graph can be efficiently modified on a batch-by-batch basis during training and decoding, allowing for the runtime selection of attention weights and bias parameters according to the desired task.

We utilize the auto-batching feature (Neubig et al., 2017b) of DyNet for efficient matrix computations, but the parallel training data must still be separated into batches[4] such that each batch consists of sentences for a single task, e.g. `<ToEn>` for a batch of sentences that all use the same target-specific attention model for translating into English. The entire training data and batches are shuffled at the beginning of each epoch so that the order of tasks seen during training is random and that they occur in proportion to their distribution in the training data.

We use 256 dimensions for our source and target word embeddings, and 256 dimensions for the hidden states. A single recurrent layer is used for the encoder and decoder. The model parameters are optimized using an unbiased Adam[5] stochastic optimizer (Kingma and Ba, 2014) in order to minimize perplexity on a held-out validation set. The validation set for the single-data baseline systems is simply the dev2006 portion of the official Europarl evaluation data for that language pair, e.g. dev2006.fr and dev2006.en for

[3] `http://dynet.io/`

[4] Each batch contains a combined maximum of 5000 source and target tokens.

[5] We use a learning rate of 0.001.

	Fr-En			Es-En			De-En		
single-data	31.84	32.21	31.80	32.17	32.09	32.11	28.39	28.67	28.32
shared	29.99	30.34	30.12	30.99	30.87	30.78	26.30	26.52	26.06
target-specific	**30.50**	**31.00**	**30.68**	**31.58**	**31.62**	**31.63**	**26.96**	27.12	**26.85**
source-specific	30.47	30.87	30.36	31.47	31.54	31.51	26.75	**27.13**	26.70
paired	29.99	30.62	30.23	30.93	31.23	31.17	26.43	26.81	26.38

Table 5: BLEU scores for single-data baseline and task-specific attention model variants: xx-to-En

	En-Fr			En-Es			En-De		
single-data	32.24	32.44	32.87	32.60	32.60	33.18	22.15	22.55	22.36
shared	29.44	29.85	30.23	30.03	30.50	31.02	19.40	19.84	19.63
target-specific	**30.06**	**30.38**	**30.92**	**30.62**	**30.93**	**31.56**	**20.16**	**20.51**	**20.27**
source-specific	29.68	30.20	30.50	30.54	30.70	31.48	19.66	20.10	20.01
paired	29.19	29.68	30.07	30.17	30.40	31.07	19.62	20.13	19.97

Table 6: BLEU scores for single-data baseline and task-specific attention model variants: En-to-xx

French-to-English translation. For multilingual systems, we want the model to work well in all possible directions so we merge the dev2006 data for all directions of interest. Combining all six directions gives a validation set with a total of $6 \times 2000 = 12000$ sentences.

5.2 Task-Specific Attention Model Decoding Results

Tables 5 and 6 show decoding results for the Foreign-to-English (xx-to-En) and English-to-Foreign (En-to-xx) directions, respectively. In all six translation directions, the single-data baseline systems obtain the highest overall BLEU scores. We expect these baseline models to perform well on their respective test sets since the translation direction (i.e. task) of the test set exactly matches the parallel data and validation set used to train the models.

The fully n-way multilingual system with a shared attention model shows degradations of up to -2.0 BLEU score, exhibiting a similar pattern to previously reported multilingual NMT results (Johnson et al., 2017). When the encoder and decoder both support multiple languages, a single shared attention model is harmful. Our multilingual NMT model with target-specific attention mitigates much of this degradation. We observe gains of between +0.5 and +0.9 BLEU, with respect to the single attention model shared across all language pairs.

Source-specific attention is also better than the shared attention model, but not as good as target-specific attention. Paired attention (i.e. a separate set of attention weights and biases for each of the six translation directions with explicitly paired parallel data) shows little change compared to the standard shared attention model. The paired attention model has more tasks than the other models (6 for paired attention vs. 4 for both target-specific and source-specific attention). It has therefore seen less data for each task and benefits from no sharing of attention-related parameters. The paired model probably lacks sufficient training data to learn a good separate attention for all tasks.

Table 7 shows BLEU scores on the out-of-domain News Commentary nc-dev2007 and nc-devtest2007 test sets. Our task-specific attention model again provides gains of between +0.6 to +1.2 BLEU showing that the technique is robust even for test sets less closely matched to the parallel training data.

5.3 Zero-Shot Decoding Results

Table 8 shows decoding results for the six zero-shot translation pairs, i.e. those directions (such as French-to-Spanish) never explicitly paired in the parallel data. Although the absolute numbers are lower (as expected for zero-shot pairs), our target-specific attention model gives gains of between +1.0 and +1.5 BLEU over the model that shares attention parameters across all languages. Source-specific attention does not work well for zero-shot translation.

	Fr-En		Es-En		De-En	
shared	23.20	21.56	30.00	28.31	20.60	18.40
target-specific	**23.82**	**22.45**	**30.95**	**29.47**	**21.65**	**19.40**
	En-Fr		En-Es		En-De	
shared	26.05	24.95	31.74	30.36	16.20	14.44
target-specific	**26.96**	**25.51**	**32.64**	**30.91**	**16.99**	**15.12**

Table 7: BLEU scores for single-data baseline and task-specific attention model variants on out-of-domain News Commentary testsets nc-dev2007 and nc-devtest2007.

	Fr-Es			Fr-De			
shared	13.64	13.63	13.75	7.82	8.04	7.84	
target-specific	**15.10**	**15.29**	**15.31**	**8.76**	**8.95**	**8.82**	From Fr
source-specific	12.27	12.35	12.31	6.99	7.01	6.82	
	Es-Fr			Es-De			
shared	13.43	13.33	13.48	7.57	7.82	7.57	
target-specific	**14.55**	**14.40**	**14.40**	**8.70**	**8.63**	**8.65**	From Es
source-specific	12.53	12.35	12.26	7.07	6.88	6.86	
	De-Fr			De-Es			
shared	10.50	10.12	10.31	9.86	10.06	10.02	
target-specific	**11.53**	**11.24**	**11.38**	**11.28**	**11.31**	**11.29**	From De
source-specific	9.64	9.59	9.41	8.76	8.83	8.78	

Table 8: BLEU scores for multilingual NMT 'zero-shot' translations comparing the baseline shared attention model to target-specific and source-specific attention models.

No single-data baseline exists for zero-shot translation since we have no parallel data for those pairs. However, we can pivot by data (i.e. find new parallel sentence pairs $x : z$ from existing data $x : y$ and $y : z$, with common y) or translate by bridging (translate from x to y, and then from y to z) for the zero-shot pairs. Pivoting by data is possible since there is a high degree of overlap amongst the various languages of the Europarl Corpus.

Although our multilingual NMT model has not seen explicitly paired data in the zero-shot directions, the encoder and decoder have seen many of the source or target sentences paired with other languages. Our zero-shot experiments are designed to test the hypothesis that the target-specific attention model is better than a model which shares attention parameters for all languages. Any benefit due to source or target data similarity applies equally to each of these multilingual models.

6 Analysis and Examples

Learning curves for the shared attention and target-specific attention model variants are shown in Figure 1. The plot includes five curves for each model variant, corresponding to different random initializations. Our target-specific attention model achieves much higher BLEU scores on the validation set in earlier epochs. There is also considerably less variance. The first few times the validation set is decoded with the shared attention model, the gap with respect to the target-specific attention model can be as large as 10 points BLEU. Our validation set contains sentences from all six translation directions so it is not surprising that a model with a single set of shared attention weights and biases performs poorly when applied to the 'wrong' task. Even after a full epoch of training (11.8m sentence pairs), the shared attention model continues to lag behind the target-specific attention model by almost one BLEU point.

The following example shows German-to-English translations obtained using a multilingual NMT system with a shared or target-specific attention model. This is an interesting example since the model must attend to the distant German source words "angesprochen worden ist" when producing the English trans-

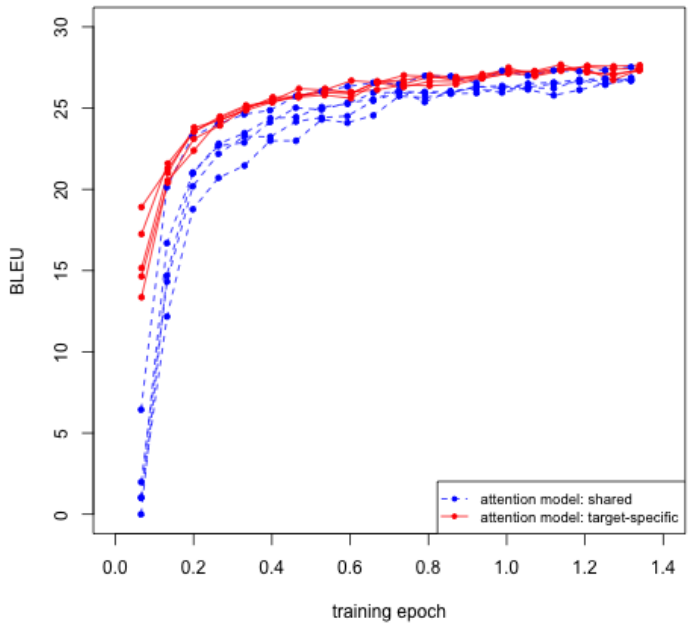


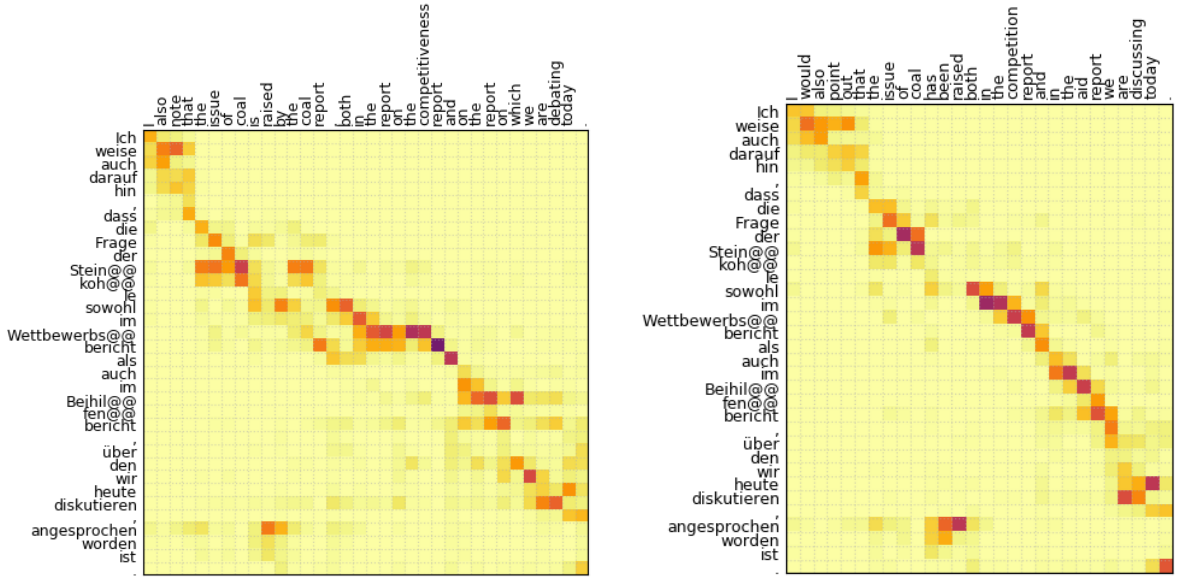
Figure 1: Multilingual validation set BLEU scores obtained during training for five different randomly initialized seeds using a shared or target-specific attention model.

lation. The multilingual NMT system with a shared model of attention produces a deficient translation, containing multiple repetitions of the word 'report' and omitting the important content word 'aid':

source	Ich weise auch darauf hin , dass die Frage der Stein@@ koh@@ le sowohl im Wettbewerbs@@ bericht als auch im Beihil@@ fen@@ bericht , ber den wir heute diskutieren , angesprochen worden ist .
reference	I would also draw your attention to the fact that the coal issue is raised both in the competition report and in the subsidy report that we are discussing today .
shared	I also note that the issue of coal is raised by the coal report , both in the report on the competitiveness report and on the report on which we are debating today .
target-specific	I would also point out that the issue of coal has been raised both in the competition report and in the aid report we are discussing today .

The attention matrices computed during decoding for the shared and target-specific model variants are shown in Figure 2. Rows correspond to the words of the German source sentence and columns to the words of the English target translation. The target-specific attention model results in a sharper and less diffuse attention over the words of the source sentence, especially at the beginning and end of the sentence, and for the translation of the passive German construction "angesprochen worden ist" which requires long-distance re-ordering.

For this example, the shared attention model leads to diffuse alignments since a single set of attention weights and biases must be used to align the words of many target languages with disparate word orders. Our target-specific attention model leads to more accurate alignments since the decoder is conditioned for a single target language (as indicated by the 'task' token prefix). Given the target-specific set of attention weights and biases, and the initialization of the decoder state using the RNN encoded source, the decoder is able to more accurately attend to the correct source words when producing the words of the target sentence at each time-step.



(a) Attention model: shared

(b) Attention model: target-specific

Figure 2: Attention plots obtained while decoding a German-to-English sentence pair in the test set. Rows correspond to source words and columns to each target position. Target-specific attention results in less diffuse alignments.

7 Conclusions and Future Work

We have described a simple but effective technique for improving the quality of multilingual NMT. Our technique mitigates much of the loss that occurs when multiple source and target languages are supported in a system with a single encoder and decoder. It is particularly effective in the zero-shot (i.e. extreme low-resource) translation directions.

Our approach extends the use of target language prefix tokens described by Johnson et al. (2017) to select a task-specific set of attention weights and biases for each task of interest. For NMT, where attention can be considered an analogue of word alignment, the use of separate attention weights and biases for language pairs with different word orders leads to improved BLEU scores in our multilingual decoder. Our results show that multilingual NMT works best with a target-specific attention model, i.e. a distinct set of attention weights and bias parameters for each supported target language. Our improved model handles all possible translation directions with only a small increase in the total number of parameters, compared to the single-data baseline or standard shared attention model systems.

In future work, we plan to apply our multilingual techniques to true low-resource language pairs, with the goal of augmenting low-resource poor quality translation systems with knowledge obtained on languages with richer resources. Given our results for zero-shot translation, we expect our approach to work well. Our attention model variants can also be applied to multi-task learning frameworks, e.g. to improve the quality of translation using knowledge learned from a variety of other natural language processing tasks such as POS tagging and dependency parsing (Kiperwasser and Ballesteros, 2018). Recent work on learning and unsupervised induction of multilingual word representations (Upadhyay et al., 2016; Zhang et al., 2017) could also be used to improve multilingual translation since the models share a single vocabulary for all source and target languages of interest.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *TACL*, 4:431–444.

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

Jonathan Baxter. 2000. A model of inductive bias learning. *J. Artif. Intell. Res.*, 12:149–198.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. 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, Beijing, China, July. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, 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.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. 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 47–57. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *TACL*, 6:225–240.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *CoRR*, abs/1706.03872.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of the European Association for Computational Linguistics*.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017a. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Graham Neubig, Yoav Goldberg, and Chris Dyer. 2017b. On-the-fly operation batching in dynamic computation graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3974–3984.

Franz Josef Och and Hermann Ney. 2002. Statistical multi-source translation. In *Machine Translation Summit 2001*, pages 253–258, 02.

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 Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 338–348.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. *CoRR*, abs/1604.00425.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. 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, San Diego, California, June. Association for Computational Linguistics.