Abstract
This paper describes the Air Force Research Laboratory (AFRL) machine translation systems and the improvements that were developed during the WMT18 evaluation campaign. This year, we examined the developments and additions to popular neural machine translation toolkits and measure improvements in performance on the Russian–English language pair.

1 Introduction
As part of the 2018 Conference on Machine Translation (Bojar et al., 2018) news-translation shared task, the AFRL human language technology team participated in the Russian–English portion of the competition. We largely employed our strategies from last year (Gwinnup et al., 2017), but adapted them to the past year’s developments, including the University of Edinburgh’s “bi-deep” (Miceli Barone et al., 2017; Sennrich et al., 2017) and Google’s transformer (Vaswani et al., 2017) architectures. For Russian–English we again submitted an entry comprising our best systems trained with Marian (Junczys-Dowmunt et al., 2018), OpenNMT (Klein et al., 2017), and Moses (Koehn et al., 2007) combined using the Jane system combination method (Freitag et al., 2014).

2 Data and Preprocessing
We used and preprocess data as outlined in Gwinnup et al. (2017). For some systems, we included the Russian–English portion of the Paracrawl\(^1\) corpus despite the noisy nature of the data. For all systems trained, we applied byte-pair encoding (BPE) (Sennrich et al., 2016) to address the vocabulary-size problem.

3 MT Systems
This year, we focused system-building efforts on the Marian, OpenNMT, and Moses toolkits, having explored a variety of parameters, data, and conditions.

3.1 Marian
We spent most of our effort investigating variations in our experimental setup with the Marian toolkit, varying training corpora, network architecture and validation metrics.

In order to facilitate ease of ensembling of models and to reduce variables while comparing the effects of settings with our Marian systems we held constant the following settings:

- We trained a joint BPE model with 49500 splits.
- We held the vocabulary size constant during training to 90k entries each for source and target.
- We held the word embedding dimensionality to 512 for all models.
- We used 1024 units in the hidden layer (where appropriate).
- We exclusively used newstest2014 as the validation set.

We experimented with building both bi-deep and transformer models - we used the same network settings with each to again provide a basis for comparison between other conditions.

For the bi-deep systems we used the following parameters:

- Alternating encoder
- Encoder cell depth of 2

\(^1\)http://www.paracrawl.eu
• Encoder layer depth of 4
• Decoder cell base depth of 4
• Decoder cell 'high depth' of 2
• Decoder layer depth of 4
• Layer normalization
• Tied embeddings for source, target and output layers
• Skip-connections

For the transformer models we used the following parameters:
• 6 layer encoder
• 6 layer decoder
• 8 transformer heads
• Tied embeddings for source, target and output layers
• Layer normalization
• Label smoothing
• Learning rate warm-up and cool-down

3.1.1 Validation Metric Choice
We experimented with varying the metric used during training to determine if using an alternate metric yielded improvements. Based on comments from previous years’ efforts, we employed BEER 2.0 (Stanojević and Sima’an, 2014) as an alternate validation metric. BEER is a trained machine translation evaluation metric with high correlation with human judgment both on sentence and corpus level. Use of this metric is motivated by the human evaluation portion of the WMT news translation task.

To compare this effect, we trained three bi-deep systems on the parallel corpus used in our WMT17 submission. These systems are trained with our common parameters outlined above, only varying the choice of validation metric: cross-entropy, BLEU, and BEER. The results of this comparison are shown in Table 1. We noted that cross-entropy and BLEU as validation metrics produce similar BLEU scores for the available test sets, but the use of BEER as a validation metric yielded an increase of between +0.7 and +1.5 BLEU when decoding the test sets.

3.1.2 Pretrained Word Embeddings
Settling on the choice of BEER as a validation metric, we then investigated the use of pretrained word embeddings (Neishi et al., 2017) in order to boost translation performance. We took the Russian and English monolingual CommonCrawl (Smith et al., 2013) data provided by the organizers and applied tokenization and BPE with our common, joint model. We then used word2vec (Mikolov et al., 2013) to train word embeddings with 512 dimensions on each of the prepared corpora. These embeddings were then used during model training. We did not fix these word embeddings while training.

For comparison purposes, we trained a bi-deep model on the WMT18 provided training data, using our common criteria with BEER as a validation metric (as outlined in Section 3.1.1). The results of this comparison are shown in Table 2. We noted an over +1.0 BLEU improvement across all available test sets solely from the use of these pretrained word embeddings.

3.1.3 Training Corpus Choice
The last major comparison for our Marian systems involved the choice of training corpora. For various training runs, we used the corpus from our WMT17 system, which included backtranslated data generated by a Marian ‘Amun’ system as described in Gwinnup et al. (2017). For others, we used the entirety of the WMT18 preprocessed data provided by the organizers. We trained bi-deep systems with pretrained word embeddings, with BEER as a validation metric, for both the WMT18 provided data and the concatenation of both the WMT17 and WMT18 corpora described earlier.

The results of this comparison are shown in Table 3. We noted there is between a +0.7 and +1.5 BLEU increase for test sets not used for validation purposes (newstest2014 showed an increase of +2.1 BLEU, but this may be due to the models overfitting on the validation set.)

3.1.4 Fine Tuning
We briefly examined fine-tuning (or continued training) (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016) late into the evaluation period. A fine-tuning corpus was constructed from the concatenation of all of the news task testsets from 2013 to 2017. A bi-deep model trained on both the WMT18 preprocessed data and the data used from our WMT17 system, pretrained word embeddings
Table 1: Comparison between using cross-entropy, BLEU and BEER as validation metrics with Marian systems. Scores for various WMT test sets measured in cased BLEU.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cross-entropy-valid</td>
<td>24.60</td>
<td>30.32</td>
<td>26.98</td>
<td>26.78</td>
<td>27.71</td>
</tr>
<tr>
<td>bleu-valid</td>
<td>24.74</td>
<td>30.23</td>
<td>26.63</td>
<td>26.94</td>
<td>27.42</td>
</tr>
<tr>
<td>beer-valid</td>
<td>25.43</td>
<td>31.51</td>
<td>28.09</td>
<td>28.10</td>
<td>28.74</td>
</tr>
</tbody>
</table>

Table 2: Comparison on using pretrained word embeddings with Marian systems. Scores for various WMT test sets measured in cased BLEU.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>newstest2013</th>
<th>newstest2014</th>
<th>newstest2015</th>
<th>newstest 2016</th>
<th>newstest2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>wmt17backtrans</td>
<td>27.75</td>
<td>33.83</td>
<td>31.07</td>
<td>30.24</td>
<td>21.39</td>
</tr>
<tr>
<td>wmt18preproc</td>
<td>26.72</td>
<td>32.59</td>
<td>28.69</td>
<td>28.41</td>
<td>31.56</td>
</tr>
<tr>
<td>wmt17/18 concat</td>
<td>28.03</td>
<td>34.70</td>
<td>30.21</td>
<td>29.67</td>
<td>32.21</td>
</tr>
</tbody>
</table>

Table 3: Comparison of different training corpora conditions. Scores for various WMT test sets measured in cased BLEU.

and validated with BEER was chosen as a starting point. We use the fine-tuning corpus to continue training for only two epochs. The results of this comparison are shown in Table 4. A gain of almost +3 BLEU is observed, showing promise with this technique, however concerns arise over possible overfitting to the fine-tuning corpus.

Table 4: Standard and Fine-tune results for newstest2018 measured in cased BLEU and BEER.

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>BEER</th>
</tr>
</thead>
<tbody>
<tr>
<td>general</td>
<td>27.05</td>
<td>0.575</td>
</tr>
<tr>
<td>fine-tuned</td>
<td>30.02</td>
<td>0.597</td>
</tr>
</tbody>
</table>

3.2 OpenNMT

Our OpenNMT system trained on the provided parallel data excepting paracrawl and the backtranslated corpus we employed for our WMT17 system. This system uses a standard RNN architecture and was fine-tuned with the other available news task test sets.

All systems used 1000 hidden units and 600 unit word embeddings.

3.3 Moses

In order to provide diversity for system combination, we trained a phrase-based Moses (Koehn et al., 2007) system with the same data as the Marian system outlined in Section 3.1. This system employed a hierarchical reordering model (Galley and Manning, 2008) and 5-gram operation sequence model (Durrani et al., 2011). The 5-gram English language model was trained with KenLM on the constrained monolingual corpus from our WMT15 (Gwinnup et al., 2015) efforts. The BPE model used was applied to both the parallel training data and the language modeling corpus. System weights were tuned with the Drem (Erdmann and Gwinnup, 2015) optimizer using the “Expected Corpus BLEU” (ECB) metric.

3.1.5 Marian Submission System

We ultimately employed an ensemble system of 5 bi-deep models and 6 transformer models trained in varying conditions (with the exception of the finetuned system in Section 3.1.4) outlined above as the Marian contribution to our submission system. This system also employed a R2L transformer model performing rescoring on the n-best lists generated during the decoding step.
System Combination

Jane System combination (Freitag et al., 2014) was employed to combine outputs from the best systems from each approach outlined above. Individual component system and final combination scores are shown in Table 5. The final system combination output comprised our entry to the Russian–English portion of the WMT18 news task evaluation.

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>BEER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marian</td>
<td>29.42</td>
<td>0.592</td>
</tr>
<tr>
<td>OpenNMT</td>
<td>28.88</td>
<td>0.580</td>
</tr>
<tr>
<td>Moses</td>
<td>24.25</td>
<td>0.565</td>
</tr>
<tr>
<td>Syscomb</td>
<td>30.01</td>
<td>0.597</td>
</tr>
</tbody>
</table>

Table 5: System combination and input system scores measured in BLEU and BEER on the newstest2018 test set.

5 Conclusion

We presented a series of improvements to our Russian–English systems focusing on improvements to neural machine translation toolkits. We again combined the best of several approaches via system combination creating a composite submission exhibiting the best of all contributing approaches.

References


Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Katherine Young, Christina May, Michael Kazi, Elizabeth Salesky, and Brian Thompson. 2015. The afrl-mill wmt15 system: There’s more than one way to decode it! In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 112–119, Lisbon, Portugal. Association for Computational Linguistics.


