The AFRL WMT19 Systems: Old Favorites and New Tricks

Jeremy Gwinnup, Grant Erdmann, Timothy Anderson
Air Force Research Laboratory
{jeremy.gwinnup.1, grant.erdmann, timothy.anderson.20}@us.af.mil

Abstract

This paper describes the Air Force Research Laboratory (AFRL) machine translation systems and the improvements that were developed during the WMT19 evaluation campaign. This year, we refine our approach to training popular neural machine translation toolkits, experiment with a new domain adaptation technique and again measure improvements in performance on the Russian–English language pair.

1 Introduction

As part of the 2019 Conference on Machine Translation (Bojar et al., 2019) news-translation shared task, the AFRL Human Language Technology team participated in the Russian–English portion of the competition. We build on our strategies from last year (Gwinnup et al., 2018), adding additional language ID based data processing and optimizing subword segmentation strategies. For Russian–English we again submitted an entry comprising our best systems trained with Marian (Junczys-Dowmunt et al., 2018), Sockeye (Hieber et al., 2017) with Elastic Weight Consolidation (EWC) (Thompson et al., 2019), OpenNMT (Klein et al., 2018), and Moses (Koehn et al., 2007) combined using the Jane system combination method (Freitag et al., 2014).

2 Data and Preprocessing

2.1 Data Preparation

We used and preprocess data as outlined in Gwinnup et al. (2018). For all systems trained, we applied either byte-pair encoding (BPE) (Sennrich et al., 2016) or SentencePiece (Kudo and Richardson, 2018) subword strategies to address the vocabulary-size problem.

For this year, we also employed a language ID filtering step for the BPE-based systems. Using the pre-built language ID model developed by the authors of fastText (Joulin et al., 2016a,b), we developed a utility that examined the source and target sentence pairs and discarded that pair if either side fell below 0.8[1] probability of the desired language. We applied this filtering to all provided parallel corpora, removing 33.7% of lines. This process was particularly effective when used to filter the Paracrawl corpus where 57.1% of lines were removed. Pre and post-filtering line counts for various corpora are shown in Table 1.

Corpus	Total	Retained
CommonCrawl	723,256	655,069
newscommentary	290,866	264,089
Yandex	1,000,000	901,307
ParaCrawl	12,061,155	5,173,675
UN2016	11,365,709	9,871,406
Total Lines	25,440,968	16,865,546

Table 1: Training corpus total and retained lines after fastText filtering

testset	wmt18preproc	wmt19filt
newstest2014	33.0	34.1
newstest2015	28.6	29.6
newstest2016	28.4	29.4
newstest2017	30.8	31.8
newstest2018	26.9	27.9

Table 2: Test set comparison for non-filtered WMT18 training corpus and filtered WMT19 training corpus measured by SacreBLEU.

A comparison with the organizer-provided parallel training data used in our WMT18 system

[1] We chose this value arbitrarily; future work will explore varying this threshold.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 203–208

(which is largely the same as the provided parallel data for WMT19 in the Russian–English language pair) on baseline Marian transformer systems with identical training conditions show that aggressive language ID based filtering yields an approximate +1 BLEU point improvement as measured by SacreBLEU (Post, 2018). These results are shown in Table 2.

2.2 Exploration of Byte-Pair Encoding Merge Sizes

One of the problems faced when addressing the closed-vocabulary problem is the granularity of the subword units either produced by SentencePiece or BPE. To that end, we examined varying the number of BPE merge operations in order to determine an optimal setting to maximize performance for the Russian–English language pair.

For the OpenNMT-based systems, a vocabulary size of 32k entries was employed during training of a SentencePiece segmentation model[2]. This vocabulary size was determined empirically from the training data.

Alternatively, for the BPE-based systems, we systematically examined varying sizes of BPE merge operations and vocabulary sizes in 10k increments from 30k to 80k. Results in Table 3 show that 40k BPE merge operations perform best across all test sets decoded for this language pair. All subsequent Marian experiments in this work utilize this 40k BPE training corpus.

3 MT Systems

This year, we focused system-building efforts on the Marian, Sockeye, OpenNMT, and Moses toolkits, having explored a variety of parameters, data, and conditions. While most of our experimentation builds off of previous years' efforts, we did examine domain adaptation via continued training, including Elastic Weight Consolidation (EWC) (Thompson et al., 2019).

3.1 Marian

As with last year's efforts, we train multiple Marian (Junczys-Dowmunt et al., 2018) models with both University of Edinburgh's "bi-deep" (Miceli Barone et al., 2017; Sennrich et al., 2017) and Google's transformer (Vaswani et al., 2017) architectures. Network hyperparameters are the same as detailed in Gwinnup et al. (2018). We again use `newstest2014` as the validation set during training.

Utilizing the best-performing BPE parameters from Section 2.2, we first trained a baseline system in each of the two network architectures, noting the Transformer system's better performance of +0.82 BLEU on average across decoded test sets. An additional six distinct transformer models were then independently[3] trained for use in ensemble decoding. We then ensemble decoded test sets with all eight models.

Marian typically assigns each model used in ensemble decoding a feature weight of 1.0; thus each model contributes equally to the decoding process. Borrowing from our Moses training approach, we utilize a multi-iteration decode and optimize feature weights using the "Expected Corpus BLEU" (ECB) metric with the Drem optimizer (Erdmann and Gwinnup, 2015). We experimented using `newstest2014` and `newstest2017` as tuning sets – 2017 did not help performance, but using 2014 did improve performance by up to +0.9 BLEU[4] over the non-tuned ensemble.

Scores for all the above-mentioned systems are shown in Table 4. The best-performing ensemble (ensemble tune14) was used in system combination.

3.2 Sockeye

For our Sockeye (Hieber et al., 2017) systems, we experimented with continued training (Luong and Manning, 2015; Sennrich et al., 2015) – a means to specialize a model in a new domain after a period of training on a general domain. One downside of utilizing continued training is the model adapts "too-well" to the new domain at the expense of performance in the original domain (Freitag and Al-Onaizan, 2016). One method to mitigate this performance drop is to prevent certain parameters of the network from changing with Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017). Thompson et al. (2019) conveniently provides an implementation of this technique in Sockeye.

That work illustrated a use case where the original domain is news articles, while the new domain is text of patent applications – a marked dif-

[2] SentencePiece was used in part to provide diversity between our OpenNMT and other systems trained with BPE data.

[3] Identical training data and starting parameters except for random seed.

[4] This may be due to the choice of newstest2014 for validation during training.

System	newstest2014	newstest2015	newstest2016	newstest2017	newstest2018
bpe30k	33.7	28.9	28.7	31.4	27.6
bpe40k	**34.1**	**29.6**	**29.4**	**31.8**	**27.9**
bpe50k	33.9	29.2	29.1	31.6	27.8
bpe60k	33.4	29.1	28.7	31.3	27.6
bpe70k	33.0	28.8	28.8	31.2	26.9
bpe80k	32.6	28.7	28.2	31.1	26.9

Table 3: Cased, detokenized BLEU for various test sets and BPE merge-value treatments. Best scores for each test set are denoted with bold text.

System	newstest2014	newstest2015	newstest2016	newstest2017	newstest2018
single bi-deep	32.7	29.0	28.7	31.3	27.0
single transformer	34.1	29.6	29.4	31.8	27.9
untuned ensemble	36.2	**31.6**	30.5	34.2	29.7
ensemble tune17	35.3	31.1	30.2	34.2	29.7
ensemble tune14	**37.1**	31.3	**31.2**	**34.5**	**30.5**

Table 4: Test set comparison for baseline bi-deep, transformer, untuned and tuned ensembles for various test sets measured in cased, detokenized BLEU. Best scores for each test set are denoted with bold text.

ference in style and content. Here, we created a news subdomain corpus from the `newstest2014` through `newstest2017` test sets. The intuition is that more current events will be discussed in these test sets than the remainder of the provided training corpora, allowing better adaptation of new events in the newest test sets (`newstest2018` and `newstest2019`.)

We first trained a baseline transformer system using the best-performing BPE parameters from Section 2.2, 512-dimension word embeddings, 6 layer encoder and decoder, 8 attention heads, label smoothing and transformer attention dropout of 0.1. We then continue-train a model on the adaptation set described above. We also followed the Sockeye EWC training procedure, producing a model more resilient to overfitting due to continued training. Results for these systems are shown in Table 5.

We see that the baseline Sockeye transformer model performs similarly to the baseline single-model Marian transformer system shown in Table 4. The continued-training system (con't train) system predictably overfit on `newstest2014` as expected, since that test set is a part of the adaptation set. Likewise, performance on the out-of-domain `newstest2018` also dropped as a result of overfitting. The best-performing EWC system[5] actually improved performance on 2018 with less-pronounced overfitting on 2014.

System	newstest2014	newstest2018
baseline	33.4	27.6
con't train	89.3	24.3
best EWC	48.5	29.5

Table 5: Sockeye system scores for `newstest2014` (in-domain) and `newstest2018` (out-of-domain) test sets for various training conditions measured in SacreBLEU.

For system combination outlined later in Section 4, we decoded test sets with an ensemble of the four highest-scoring model checkpoints from the best EWC training run.

3.3 OpenNMT-T

Our first Open-NMT system was trained using the Transformer architecture with the default "TransformerBig" settings as described in Vaswani et al. (2017): 6 layers of 1024 units, 16 attention heads. Dropout rates of 0.3 for layers and 0.1 for attention heads and relu's. Training data for this system utilized the training corpus from our WMT17 Russian–English system (Gwinnup et al., 2017) consisting of provided parallel and backtranslated

[5]EWC applied with weight-decay of 0.001 and learning-rate of 0.00001

data. This data was then processed with a joint 32k word vocabulary SentencePiece model.

3.4 OpenNMT-G

For our second OpenNMT system, we first trained language-specific, 32k word vocabularies using SentencePiece. WMT news test data from all years except 2014 and 2017 were used to train SentencePiece. These data, with the addition of the language ID filtered ParaCrawl corpus outlined in Section 2.1, were used for training the system. WMT news test data from 2014 was used for validation. OpenNMT-tf was used to create the system, using the stock "Transformer" model.

3.5 Moses

As in previous years, we trained a phrase-based Moses (Koehn et al., 2007) system with the same data as the Marian system outlined in Section 3.1 in order to provide diversity for system combination. 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 all permissable monolingual English news-crawl data. 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.

4 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 6 for cased, detokenized BLEU and BEER 2.0 (Stanojević and Sima'an, 2014) .

5 Submission Systems

We submitted the final 5-system combination outlined in Section 4 and the four-checkpoint EWC ensemble detailed in Section 3.2 to the Russian–English portion of the WMT19 news task evaluation. Selected `newstest2019` automatic scores from the WMT Evaluation Matrix[6] are shown in Table 7.

[6] `http://matrix.statmt.org`

System	BLEU	BEER
1. Marian	30.47	0.5995
2. Sockeye EWC	29.43	0.5968
3. OpenNMT-T	26.22	0.5737
4. OpenNMT-G	30.05	0.6017
5. Moses	27.33	0.5836
Syscomb-5	32.12	0.6072

Table 6: System combination and input system scores measured in cased, detokenized BLEU and BEER on the `newstest2018` test set.

System	BLEU	BEER
afrl-syscomb19	37.2	0.627
afrl-ewc	34.3	0.613

Table 7: Final submission system scores measured in cased BLEU and BEER on the `newstest2019` test set.

6 Conclusion

We presented a series of improvements to our Russian–English systems, including improved preprocessing and domain adaptation. Clever remixing of older techniques from the phrase-based MT era enabled improvements in ensembled neural decoding. Lastly, we performed system combination to leverage benefits from these new techniques and favorite approaches from previous years.

References

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Christof Monz, Mathias Müller, and Matt Post. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pages 1045–1054, Portland, Oregon.

Grant Erdmann and Jeremy Gwinnup. 2015. Drem: The AFRL submission to the WMT15 tuning task.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 12 June 2019. Originator reference number RH-19-119921. Case number 88ABW-2019-2969.

In *Proc. of the Tenth Workshop on Statistical Machine Translation*, pages 422–427, Lisbon, Portugal.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast Domain Adaptation for Neural Machine Translation. *CoRR*, abs/1612.06897.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open source machine translation system combination. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, Gothenburg, Sweden.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, and Katherine Young. 2018. The AFRL WMT18 systems: Ensembling, continuation and combination. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 394–398. Association for Computational Linguistics.

Jeremy Gwinnup, Timothy Anderson, Grant Erdmann, Katherine Young, Michaeel Kazi, Elizabeth Salesky, Brian Thompson, and Jonathan Taylor. 2017. The AFRL-MITLL WMT17 systems: Old, new, borrowed, BLEU. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 303–309, Copenhagen, Denmark. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 177–184. Association for Machine Translation in the Americas.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 99–107. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

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

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, page to appear.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.