The AFRL WMT18 Systems: Ensembling, Continuation and Combination

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

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

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 394–398
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64038

- 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

System	newstest2013	newstest2014	newstest2015	newstest 2016	newstest2017
cross-entropy-valid	24.60	30.32	26.98	26.78	27.71
bleu-valid	24.74	30.23	26.63	26.94	27.42
beer-valid	25.43	31.51	28.09	28.10	28.74

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.

System	newstest2013	newstest2014	newstest2015	newstest 2016	newstest2017
during training	25.63	31.20	–	26.68	29.60
pretrained	26.72	32.59	28.69	28.41	31.56

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

Corpus	newstest2013	newstest2014	newstest2015	newstest 2016	newstest2017
wmt17backtrans	27.75	33.83	31.07	30.24	21.39
wmt18preproc	26.72	32.59	28.69	28.41	31.56
wmt17/18 concat	28.03	34.70	30.21	29.67	32.21

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.

System	BLEU	BEER
general	27.05	0.575
fine-tuned	30.02	0.597

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

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.

3.2 OpenNMT

Our OpenNMT system trained on the provided parallel data excepting paracrawl and the back-translated 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.

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 5. The final system combination output comprised our entry to the Russian–English portion of the WMT18 news task evaluation.

System	BLEU	BEER
Marian	29.42	0.592
OpenNMT	28.88	0.580
Moses	24.25	0.565
Syscomb	30.01	0.597

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

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. 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. 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, Katherine Young, Christina May, Michaeel Kazi, Elizabeth Salesky, and Brian Thompson. 2015. The afrl-mitll 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.

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.

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*, Melbourne, Australia.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

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.

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.

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 16 Aug 2018. Originator reference number RH-18-118721. Case number 88ABW-2018-4047.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR) Workshop*.

Masato Neishi, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 99–109. Asian Federation of Natural Language Processing.

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. 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.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 1374–1383, Sofia, Bulgaria.

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.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.