The AFRL WMT20 News-Translation Systems

Jeremy Gwinnup and **Timothy Anderson**
Air Force Research Laboratory
{jeremy.gwinnup.1, timothy.anderson.20}@us.af.mil

Abstract

This report summarizes the Air Force Research Laboratory (AFRL) machine translation (MT) systems submitted to the news-translation task as part of the 2020 Conference on Machine Translation (WMT20) evaluation campaign. This year we largely repurpose strategies from previous years' efforts with larger datasets and also train models with precomputed word alignments under various settings in an effort to improve translation quality.

1 Introduction

As part of the 2020 Conference on Machine Translation (wmt, 2020) 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 including language-based filtering of training corpora with fastText (Joulin et al., 2016b,a), employing transformer-based (Vaswani et al., 2017) translation models and once again utlitizing system combination to fuse outputs from OpenNMT (Klein et al., 2018), Marian (Junczys-Dowmunt et al., 2018) and Moses (Koehn et al., 2007) systems. We also examine the effects of training Marian models with externally generated word alignments as described in (Alkhouli et al., 2018).

2 Data processing

For purposes of training our systems, we use the following parallel corpora: Commoncrawl (Smith et al., 2013), Yandex[1], UN v1.0 (Ziemski et al., 2016), Paracrawl[2](Esplà et al., 2019), Wikimatrix (Schwenk et al., 2019), and backtranslated data from our WMT17 system (Gwinnup et al., 2017) as well as Edinburgh's WMT17 system (Sennrich et al., 2017) yielding a raw corpus of over 76.3 million lines.

We prepare training corpora in a similar manner described in (Gwinnup et al., 2018), however this year, we utilize SentencePiece (Kudo and Richardson, 2018) with a 46k-entry vocabulary[3] for processing subword units instead of byte-pair encoding (BPE) (Sennrich et al., 2016).

2.1 Language-ID based data filtering

As with last year's efforts, we again employ fastText (Joulin et al., 2016b,a) to filter the various parallel corpora with a utility examining the source and target sentence pairs, discarding pairs where either (or both) sentence in the pair falls below a threshold score of 0.8. We wished to explore different threshold values, but our team did not have access to the majority of our computational assets due to the COVID-19 pandemic, limiting the bandwidth available for experiments.

We show the results of language-ID based filtering in Table 1. On average, 76.79% of the original training data is retained, with our WMT17 backtranslated data retaining the largest percentage of lines at 93.22% - this is interesting since that data originated as English and was translated to Russian with a very shallow Amun (Hoang et al., 2018) model. Again, Paracrawl yielded the least percentage of retained lines at 42.90%, but is understandable due to the "raw" nature of this particular release.

2.2 Guided Alignment

Inspired by the results in (Alkhouli et al., 2018), we've examined effects of using precomputed word alignments as a guide during training; Marian has a facility to train in this manner. Alignments were generated using Fastalign (Dyer et al., 2013)

[1] https://translate.yandex.ru/corpus?lang=en

[2] Version 1 Russian–English parallel data

[3] This vocabulary size performed best in empirical testing in our WMT19 submission.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 207–212
Online, November 19–20, 2020.

corpus	unfiltered lines	filtered lines	percent remain
commoncrawl	723,256	655,069	90.57%
news-commentary-v15	319,242	286,947	89.88%
yandex	1,000,000	901,318	90.13%
un-2016	11,365,709	9,871,406	86.85%
paracrawl	12,061,155	5,173,675	42.90%
wikimatrix	5,203,872	4,287,881	82.40%
wmt17-afrl-bt	8,921,942	8,317,107	93.22%
wmt17-uedin-bt	36,772,770	29,074,022	79.06%
Total	76,367,946	58,567,425	76.69%

Table 1: Results of language-id based Russian–English corpus filtering with threshold of 0.8

on both "plain" and SentencePiece-processed data; MGIZA (Gao and Vogel, 2008) alignments were only generated for the word-based data. In order to generate these alignments, the language-id filtered corpus described in the previous section was further processed using Moses's clean-corpus-n-ratio script as well as escaping various characters and entities (such as ' replaced with &amp;) yielding a final corpus of 49,866,140 lines. Additionally, a 46k entry SentencePiece model is built on this corpus with user-defined vocabularies for the tokens escaped during processing.

We use a Procrustes alignment projection script[4] to effectively map alignments generated on whole word tokens to the equivalent series of subword tokens in the SentencePiece processed data. Comparisons are drawn between Marian models trained on these various conditions in Section 3.2.

3 Machine Translation Systems

This year, we focused system-building efforts on the OpenNMT, Marian, and Moses toolkits. While most of our experimentation builds off of previous years' efforts, this year we examine the effects of "guided-alignment" training with the Marian toolkit in an attempt to improve translation quality.

3.1 Open-NMT

The OpenNMT system trained for this task used the the configuration for a large transformer network.

We used the following network hyperparameters:

- 1024 embedding size
- 4096 hidden units
- 12 layer encoder
- 12 layer decoder
- 16 transformer heads
- dropout 0.3
- attention dropout 0.1
- Tied embeddings for source, target and output layers
- Layer normalization
- Label smoothing
- Learning rate warm-up

The corpus was processed with SentencePiece using a model with a vocabulary size of 40K trained on the ru-en corpus. The network was trained for 10 epochs of this training data using a batch size of 1562, with an effective batch size of 24,992 using the lazy Adam (Kingma and Ba, 2015) optimizer. The final system was an average of the last 8 checkpoints of the training. Checkpoints were saved every 5000 steps. The system was then tuned with one epoch of newstest data from years 2014-2017.

3.2 Marian

Our Marian systems also utilize the transformer architecture. We use the WMT14 newstest2014 test set for validation during training and the following network hyperparameters:

- 2048 hidden units

[4] https://bitbucket.org/ndnlp/procrustes/src/master

- 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

We first train a baseline system with the 58 million line corpus outlined in 2.1 and then train another baseline on the further-filtered 49 million line corpus outlined in 2.2. Using the word alignments generated earlier, we train systems utilizing alignments on subwords using fastalign (ga-spm-fastalign), alignments generated by projecting word-based fastalign alignments onto SentencePiece tokens (ga-procrusted-fastalign), and word-based MGIZA alignments onto SentencePiece tokens (ga-procrustes-mgiza). Results for decoding newstest2014 for each of these models are shown in Table 2.

system name	newstest2014
full-corpus baseline	39.81
ga-baseline	34.17
ga-spm-fastalign	33.06
ga-procrustes-fastalign	33.49
ga-procrustes-mgiza	31.92

Table 2: Experimental results for both baseline and guided-alignment systems decoding WMT14 testset measured in cased, detokenized BLEU.

We see that the best performing system is the one trained on the larger corpus, which is not surprising. We also see that while none of the guided-alignment based approaches we tried scored higher than the baseline on the smaller guided-alignment corpus, using the fastalign projected alignments performs better than the fastalign subword-based alignments by approximately 0.4 BLEU. We did experience issues getting MGIZA to successfully run on the 49 million line corpus, which may suggest additional processing of the training corpus is necessary to generate "correct" alignments using that approach. However, this specific MGIZA run provided the word alignments used in the Moses system described in the next section. This suggests more careful examination may be necessary before drawing conclusions as to the efficacy of using guided alignments to the Marian training process.

3.3 Moses

As in previous years, we trained a phrase-based Moses (Koehn et al., 2007) system with the guided-alignment data outlined in Section 2.2 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 (Heafield, 2011) on the constrained monolingual corpus from our WMT15 (Gwinnup et al., 2015) efforts. System weights were tuned with the Drem (Erdmann and Gwinnup, 2015) optimizer using the "Expected Corpus BLEU" (ECB) metric.

3.4 System Combination

Once again, Jane (Freitag et al., 2014) system combination was used to combine various systems, tuned on newstest2016. We were able to successfully combine variations of three and four input systems, with results discussed in the following section.

4 Experimental Results

Results of decoding our various MT systems on WMT test sets from 2014 through 2019 are shown in Table 3.

Marian-base is an ensemble of 5 transformer models trained with identical hyperparameters as outlined in Section 3.2, with the exception of the initial random seed and using the language-id filtered corpus described in Section 2.1. Individual model weights are trained via Drem (Erdmann and Gwinnup, 2015) as outlined in last year's system.

Marian-ga is an ensemble of the four guided-alignment models described in Section 3.2: ga-baseline, ga-spm-fastalign, ga-procrustes-fastalign and ga-procrustes-mgiza. Individual model weights are also trained with Drem.

Onmt-base is the baseline system described in Section 3.1 and onmt-tune is the system that was further finetuned on newstest 2014-2017; Scores on those test sets are not reported due to overfitting during the fine tuning process.

Variations of system combinations are also reported - again with the absence of onmt-tune due to concerns of overfitting as newstest2016 is the test set used for tuning the system combination process. Combinations of only two systems resulted in a segmentation fault during processing due to fragility in the combination process.

We entered System 8 as our primary submission due to its performance gain on newstest2019 in a general (non-finetuned) setting, with the intuition that this years test set would discuss similar topics or issues as last years, while the earlier sets may be dated. In contrast, we submit System 5 as an alternative due to finetuning adapting the model to the collection of recent test sets.

5 Conclusion

In addition to our "known-good" approaches with increased data to submit respectably-performing translation systems, we conducted several experiments with guided alignments. Although these systems didn't outperform our prior approaches, they did figure into our final system combination submitted to the evaluation.

The authors wish to thank David Chiang for his implementation of the Procrustes alignment-projection script. The authors would also like to thank Grant Erdmann, Emily Conway and Grace Smith for their assistance in human evaluation of MT output.

References

2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, 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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

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.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open source machine translation system combination. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

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, 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*, pages 303–309, Copenhagen, Denmark. Association for Computational Linguistics.

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 1 October 2020. Originator reference number RH-20-121361. Case number 88ABW-2020-3049.

		WMT newstest					
#	system name	2014	2015	2016	2017	2018	2019
1	marian-base	41.09	35.30	35.64	38.89	34.03	37.04
2	marian-ga	32.86	27.97	28.46	31.49	27.43	31.96
3	moses-base	35.47	30.85	30.69	33.62	28.16	32.38
4	onmt-base	36.87	32.58	32.48	35.50	30.76	38.26
5	onmt-tune	–	–	–	–	32.31	39.27
6	syscomb 1+2+4	40.91	35.74	36.07	39.4	33.95	38.29
7	syscomb 1+3+4	41.00	35.85	35.87	39.52	33.99	38.24
8	syscomb 1+2+3+4	40.89	35.97	36.12	39.15	33.75	39.21

Table 3: Experimental results for both input systems and system combination results decoding WMT testsets measured in cased, detokenized BLEU as scored by mteval-13a.pl.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Hieu Hoang, Tomasz Dwojak, Rihards Krislauks, Daniel Torregrosa, and Kenneth Heafield. 2018. Fast neural machine translation implementation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 116–121. Association for Computational Linguistics.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, New Orleans. 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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

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.

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.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.