The University of Edinburgh's submission to the German-to-English and English-to-German Tracks in the WMT 2020 News Translation and Zero-shot Translation Robustness Tasks

Ulrich Germann
University of Edinburgh, UK
ugermann@inf.ed.ac.uk

Abstract

This paper describes the University of Edinburgh's Submission of German ↔ English systems to the WMT2020 Shared Tasks on News Translation and Zero-shot Robustness.

1 Introduction

This paper describes the University of Edinburgh's submission to the German-to-English and English-to-German tracks in the WMT 2020 News Translation and Zero-shot Robust translation tasks. We built our systems in three stages, loosely following the procedure by Junczys-Dowmunt (2018b). All translation models mentioned in this paper were trained with the Marian toolkit (Junczys-Dowmunt et al., 2018).

2 3-stage Build Process

We distinguish three types of training data provided for the tasks, shown in Table 1.

Table 1: Data available for training systems

corpus	sentence(pair)s
High-quality parallel data	
Europarl	ca. 1.79 M
Rapid	ca. 1.45 M
News Commentary	ca. 0.35 M
Crawled parallel data	
ParaCrawl 5.1	ca. 34.37 M
CommonCrawl	ca. 2.40 M
WikiMatrix	ca. 6.22 M
WikiTitles	ca. 1.38 M
Monolingual crawled news data	
German	ca. 327.69 M
English	ca. 233.50 M

2.1 Round 1: Training models for scoring crawled parallel data

In the first round, we trained base transformer models (Vaswani et al., 2017) with SentencePiece subword segmentation (Kudo and Richardson, 2018) with a joint vocabulary of 32K tokens on the high-quality parallel data. The joint vocabulary remains the same for all models. We applied a binomial sentence length model to remove from the parallel data sentence pairs with an unreasonable sentence length ratio. The model assumes that a pair of sentences of lengths K and L is produced by a series of K+L flips of a biased coin. The bias is based on the corpus-level ratio of tokens; for German and English, we determined that there are on average 1.0723 English tokens per German token, so the Null Hypothesis assumes that an English word is generated with a probability of 51.75%, and a German word with a probability of 48.25%. For each sentence pair, we determine the p-value of the Null Hypothesis; if it is less than 0.5%, the sentence pair is discarded. This sentence filtering filtered out less than 2% of the EuroParl data, ca 3.4% of the NewsCommentary data, and 11% of the Rapid corpus. The numbers given in Table 1 are after filtering.

2.2 Selecting crawled parallel data

We used these models to compute the length-normalized cross-entropy for each sentence pair in the available crawled parallel data in both translation directions, and combined these two entropies into the dual cross-entropy score (Junczys-Dowmunt, 2018a). To bias data selection towards the news domain, we also computed length-normalized cross-entropy for each sentence with a 5-gram language model[1] trained on the respective monolingual news data for the target side and added the two scores to obtain a single score for ranking candidates. The top n candidates from the crawled parallel data were pooled with the high-quality parallel data for the second round of training. We experimented with the top 15 M, top 20 M, and top 25 M candidates from the pool of crawled parallel data. We did not put effort into cleaning or filtering the data prior to scoring, as we assumed that poor candidates would be detected by the dual cross-entropy score.

[1] https://github.com/kpu/kenlm

Proceedings of the 5th Conference on Machine Translation (WMT), pages 197–201
Online, November 19–20, 2020.

For the German-to-English system, we unfortunately committed a serious blunder that we did not notice until shortly before submitting this system description: instead of sorting in descending order of ranking score, we accidentally sorted in descending order of provenance label first and (also lexicographically) ranking score second, so that for the translation direction German→English, the crawled data selection contained all of WikiMatrix data, none of the CommonCrawl data, and a selection of ParaCrawl data. The selection error rate in the Top-25M confiuration is ca 66% (i.e., 66% of that data should not have been selected, and we missed 66% of the data that we wanted to select). As we used Round2-models for back-translation of monolingual data, this error may have also tainted the training of the English-to-German system.

2.3 Round 2: Big transformers for back-translation

We then trained big transformer models for back-translation of monolingual news data, using the "top" (see above for our blunder on German-to-English data selection) 25M candidates.

2.4 Back-translation of news data

We used single models to translate all of the news data for German and English, adding a bit of noise in the translation by adding Gumbel noise to the output layer, thus adding some randomness to the translation process.

2.5 Round 3: Training final models with back-translation.

In the final round of training, we trained big transformer models on a blend of back-translated data (75%), crawled parallel data (15%) and high-quality parallel data (10%). Due to the volume of training data available for this round, we replaced full shuffling of the data for each epoch by random selection: we loop over each data source, fully shuffling the latter two data sources (crawled and high-quality parallel) in each iteration, but shuffling the backtranslated news data only once and then randomly selecting only 10% of the data in each iteration. Reading separately from the three data sources, the data feeder randomly selects one data source at a time according to the aforementioned distribution of 75,15, and 10% and outputs the next sentence pair in the queue.

3 Training details

For training, we experimented with variations on learning rate, batch size, warmup, and norm clipping. Due to an apparent bug in the implementation of norm clipping in Marian,[2] the Marian authors do currently not recommend to use norm clipping with Marian.[3] However, we found that without it, training would occasionally fail due to exploding gradients. Norm clipping also allowed us to be more agressive with the learning rate,

Settings and BLEU scores[4] on the validation set (WMT19) are shown in Table 2. Effective batch size was influenced by the GPU memory allocated and the number of gradients accumulated before a parameter update ("optimizer delay"). In principle, doubling the optimizer delay should double the batch size, but we found this not to be the case in practice. An analysis after the fact revealed that this was due to interactions between automatically fitting batch sizes to available memory (`--maxi-batch-fit`), setting maxibatch size, and the optimizer delay parameter that are currently not documented well for the Marian toolkit and that we misunderstood.

For German-to-English, training on back-translated data did not lead to improvements in terms of BLEU on the validation set, so we ensembled the 4 listed round-2 models for our primary submission. We were able to boost the BLEU score of the raw translation output by 1.3 BLEU points with a simple post-processing step that simply adjusts quotations marks to German spelling conventions.

For English-to-German, we submitted an ensemble of the 8 round-2 models with the highest BLEU score with respect to the validation set (WMT19).

Here, too, training on back-translated news data did not lead to an improvement over the best Round-2 model, so that we did not use any round-3 model for the final submission. We were able to boost performance by increasing the batch size during training, which is in line with our findings from last year (Bawden et al., 2019), but the effect was much smaller this year. This may be due to the fact that the initial model (#10, English→German) was already trained with a fairly large batch size.

4 Results

Table 3 shows the overlap (as measured in BLEU) for all primary systems submissions to the News Translation Task as released by the workshop organizers. We notice a few things. First, our data selection blunder for the German-to-English system has not catastrophically harmed final performance. In fact, in terms of ranking with respect to BLEU, our German-to-English system does better than our English-to-German system.

[2] Gradients aren't normalized but norm clipping isn't adjusted for batch size.

[3] Personal communication with R. Grundkiewicz.

[4] All BLEU scores reported in this paper were computed with SacreBLEU (Post, 2018); BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a +version.1.4.14.

Table 2: Training details. See Section 3 for details.

	cont.	**WMT19**	**transformer**	**batch**	**learning**	**clip**	**warmup**	**crawled**
run	**from**	**BLEU**	**type**	**tokens**	**rate**	**norm**	**update**	**data**
Round 1 German→ English								
1		29.6	basic	ca. 119K	0.0009	5	16K	—
Round 2 German→ English								
2		41.02	big	ca. 84K	0.0009	5	16K	20M
3		41.21	big	ca. 31K	0.0002	0	8K	20M
4		41.22	big	ca. 81K	0.0003	0	16K	20M
5		41.47	big		0.0003	0	16K	25M
Round 3 German→ English								
6		39.51	big	ca. 124K	0.0002	0	8K	25M
7	5	41.04	big	ca. 246K	0.0003	5	1K	25M
Round 1 English→ German								
1		TBD	basic	ca. 119K	0.0009	5	16K	—
Round 2 English→ German								
2		41.63	big	ca. 20K	0.0002	0	8K	20M
3		41.73	big	ca. 83K	0.0003	0	16K	25M
4		41.85	big	ca. 103K	0.0002	0	8K	20M
5		41.89	big	ca. 185K	0.0009	0	26K	20M
6		42.02	big	ca. 34K	0.0002	0	8K	20M
7		42.13	big	ca. 144K	0.0003	0	16K	25M
8		42.49	big	ca. 26K	0.0009	5	16K	15M
9		42.62	big	ca. 56K	0.0002	0	8K	20M
Round 3 English→ German								
10		31.23	big	ca. 120K	0.0003	0	8K	25M
11	10	41.46	big	ca. 120K	0.0002	5	–	25M
12	11	42.01	big	ca. 205K	0.0002	5	–	25M
13	12	42.61	big	ca. 334K	0.0002	0	–	25M
13	12	41.94	big	ca. 339K	0.0002	0	–	25M

Table 3: Overlap between references and submissions for German-to-English (top) and English-to-German (bottom), measured in BLEU, with "references" in the columns and "candidates" in the rows.

	REF1	Tohoku...	Huoshan...	OPPO	UEDIN	Online-B	Online-G	Online-A	PROMT...	Online-Z	REF2	Biomed...	zlabs-nlp	yolo
REF1	100.00	43.80	43.50	43.20	42.30	41.90	41.40	40.40	39.60	35.40	34.00	32.10	31.50	0.20
Tohoku...	43.80	100.00	77.00	71.80	70.90	63.90	69.30	66.60	64.10	54.20	42.20	47.40	47.10	0.20
Huoshan...	43.50	77.10	100.00	74.60	73.60	65.50	70.00	70.40	68.50	57.40	42.20	48.80	47.80	0.20
OPPO	43.20	71.90	74.60	100.00	71.10	64.50	66.80	63.90	63.50	55.10	41.90	48.10	47.00	0.30
UEDIN	42.30	70.90	73.60	71.10	100.00	64.90	69.60	67.40	67.60	57.00	40.80	50.10	50.20	0.20
Online-B	41.90	64.00	65.50	64.50	64.90	100.00	61.90	61.70	61.20	54.10	40.90	45.70	44.80	0.20
Online-G	41.40	69.30	70.00	66.80	69.70	61.90	100.00	64.70	63.90	53.80	39.80	46.20	45.50	0.20
Online-A	40.40	66.60	70.40	63.90	67.40	61.70	64.70	100.00	64.30	53.70	39.60	46.10	44.50	0.20
PROMT...	39.60	64.20	68.50	63.50	67.60	61.20	63.90	64.30	100.00	56.80	38.60	49.20	47.30	0.20
Online-Z	35.40	54.20	57.40	55.10	57.00	54.10	53.80	53.70	56.80	100.00	35.00	43.00	43.10	0.20
REF2	34.00	42.20	42.20	41.90	40.90	40.90	39.80	39.60	38.60	35.00	100.00	31.00	30.50	0.20
Biomed...	32.10	47.40	48.80	48.10	50.10	45.70	46.30	46.10	49.20	42.90	31.00	100.00	46.80	0.20
zlabs-nlp	31.50	47.10	47.80	47.00	50.20	44.70	45.50	44.40	47.30	43.10	30.50	46.80	100.00	0.30
yolo	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.30	100.00

	REF1	Tohoku...	Tencent...	OPPO	Huoshan...	eTranslation	Online-B	UEDIN	Online-A	AFRL	REF2	PROMT...	Online-Z	zlabs-nlp	Online-G	Biomed...
REF1	100.00	38.70	38.60	38.50	38.10	37.90	36.40	36.30	36.10	34.10	32.50	31.80	29.60	28.10	27.20	25.20
Tohoku...	38.80	100.00	70.40	69.80	71.50	70.30	63.20	65.30	67.50	59.90	38.80	60.10	50.90	48.20	50.60	41.10
Tencent...	38.60	70.40	100.00	74.30	74.20	70.40	59.60	67.80	63.80	62.10	38.00	59.00	51.20	45.80	49.00	41.30
OPPO	38.60	69.80	74.30	100.00	75.20	72.90	60.60	68.00	63.60	63.70	38.00	58.00	50.10	45.00	48.40	40.10
Huoshan...	38.20	71.60	74.20	75.10	100.00	72.80	60.40	69.50	64.30	62.50	37.90	60.50	51.60	46.40	49.10	41.30
eTranslation	37.90	70.30	70.40	72.90	72.80	100.00	58.20	68.80	63.90	64.30	37.60	57.60	49.10	45.10	48.70	39.80
Online-B	36.50	63.20	59.50	60.50	60.40	58.10	100.00	57.00	62.40	53.60	36.60	56.10	48.50	46.00	47.80	41.80
UEDIN	36.30	65.30	67.80	68.00	69.40	68.80	57.10	100.00	62.50	61.70	35.50	60.90	51.50	48.70	50.10	42.00
Online-A	36.20	67.50	63.80	63.50	64.20	63.90	62.40	62.40	100.00	55.70	36.10	60.50	49.80	48.40	50.80	42.10
AFRL	34.10	59.90	62.20	63.80	62.60	64.40	53.70	61.70	55.80	100.00	34.20	52.60	46.60	43.60	45.50	39.70
REF2	32.50	38.80	37.90	38.00	37.90	37.60	36.50	35.40	36.10	34.20	100.00	31.60	28.70	27.40	27.00	24.50
PROMT...	31.80	60.20	59.00	58.00	60.50	57.50	56.20	60.90	60.50	52.50	31.70	100.00	50.60	47.70	51.60	43.20
Online-Z	29.60	51.00	51.10	50.10	51.60	49.10	48.50	51.50	49.80	46.60	28.70	50.60	100.00	42.00	43.20	38.00
zlabs-nlp	28.20	48.20	45.80	44.90	46.40	45.10	46.00	48.70	48.40	43.50	27.40	47.70	42.00	100.00	40.10	40.00
Online-G	27.20	50.70	49.00	48.50	49.10	48.70	47.90	50.20	50.80	45.40	27.00	51.60	43.20	40.10	100.00	37.20
WMTBiomedBaseline	25.20	41.10	41.30	40.20	41.30	39.80	41.80	42.00	42.10	39.70	24.50	43.20	38.00	40.00	37.30	100.00

Table 4: BLEU scores for the Zero-shot Robustness Task

test set	BLEU
en→de set 1	35.1
de→en set 1	38.8
de→en set 3	43.8

Second, the difference between the independently created human reference translations (REF1 and REF2) is larger than the difference between the top-performing automatic translations and either of the systems. We conjecture that is is due to the fact that individual translators will have individual translation styles, whereas automatically trained systems learn to emulate the "average" translator. However, this once again raises the question about the validity of BLEU as a measure of translation quality.

Third, we find the high overlap between the top-scoring automatic systems remarkable. This suggests that under the constrained conditions, independently systems do learn a very similar style of translation.

For the Zero-shot Robustness Task, we used the same systems as for for the News Translation Task. We report BLEU scores for the Zero-Shot Robustness Task in Table 4.

Acknowledgements

This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825303 (Bergamot) and 825627 (European Language Grid).

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (http://www.csd3.cam.ac.uk/), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

References

Bawden, Rachel, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. The university of Edinburgh's submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin. 2018a. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin. 2018b. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 425–430, Belgium, Brussels. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135, Melbourne, Australia. Association for Computational Linguistics.

Kudo, Taku 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. Code at `https://github.com/google/sentencepiece`.

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

Vaswani, Ashish, 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 5998–6008.