eTranslation's Submissions to the WMT 2021 News Translation Task

Csaba Oravecz[†] **Katina Bontcheva**[†] **David Kolovratník**[†]
Bhavani Bhaskar[†] **Michael Jellinghaus**[⋆] **Andreas Eisele**[⋆]
DG Translation – DG CNECT, European Commission
[†]firstname.lastname@ext.ec.europa.eu
[⋆]firstname.lastname@ec.europa.eu

Abstract

The paper describes the 3 NMT models submitted by the eTranslation team to the WMT 2021 news translation shared task. We developed systems in language pairs that are actively used in the European Commission's eTranslation service. In the WMT news task, recent years have seen a steady increase in the need for computational resources to train deep and complex architectures to produce competitive systems. We took a different approach and explored alternative strategies focusing on data selection and filtering to improve the performance of baseline systems. In the domain constrained task for the French–German language pair our approach resulted in the best system by a significant margin in BLEU. For the other two systems (English–German and English-Czech[1]) we tried to build competitive models using standard best practices.

1 Introduction

The eTranslation team is behind the translation services of the European Commission's eTranslation project[2]. This is a building block of the Connecting Europe Facility (CEF), with the aim of supporting European and national public administrations' information exchange across language barriers in the EU. The project is described in more details in (Oravecz et al., 2019).

The team's participation in the WMT shared tasks has provided valuable insights to improve the quality of our production systems and allowed us to explore languages and domains beyond the formal language of EU institutions, leading to a continuous extension of the eTranslation service and helping in the search for the right balance between the use of resources in production environments and the best possible performance of models.

This year the team participated in the news translation shared task with 3 different language pairs: English → German, English → Czech and French → German. The selection was motivated by the fact that these language pairs can all be considered as high or medium resource, which is the main scenario in the eTranslation service, while the constrained domain in Fr→De offered a good opportunity to focus on and experiment with data selection and filtering techniques, which is a more viable alternative in our environment than the resource demanding (brute-force) increase in model complexity.

2 Data Preparation

Here we briefly describe the base data sets, the general selection and filtering methods we applied to prepare these initial data sets used to train the first models. Further data selection and augmentation methods to improve the quality of baseline models are described in Section 3.2. For all models we only used the provided parallel and monolingual data, so our 3 submissions fall into the constrained category.

2.1 Base Data Selection and Filtering

As a first baseline approach, we tried to make use of all provided original parallel (OP) data to build the first models for reference or back-translation. Since these data sets were fairly similar to those from last year we followed the same practice and trained baseline models from all OP data. There was, however, a significant increase in the ParaCrawl data, which for En→De for example, doubled its size. As it turned out, the increase in size did not necessarily mean a better translation model trained from the full data set so we explored different subsets based on scoring by both source and target language models (see Section 4.1 for the details of these experiments).

The domain distribution of the data sets was not

[1]Due to returning problems of resource availability, the En→Cs experiments did not finish until the submission deadline so we could finally only submit last year's system.

[2]https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 172–179
November 10–11, 2021.

Data set	En→De	Fr→De	En→Cs
Europarl v10	1.77M	1.79M	0.62M
Common Crawl	2.16M	0.56M	0.11M
News Commentary v16	0.38M	0.29M	0.25M^{v15}
Tilde Rapid corpus	0.99M	–	0.28M
Wiki Titles v3	1.31M	0.52M	0.32M^{v2}
ParaCrawl v7.1	79.2M	6.30M	4.90M$^{v5.1}$
WikiMatrix	5.46M	2.80M	1.92M
CzEng 2.0	–	–	41.6M
Total:	91.27M	12.26M	50.0M

Table 1: Number of segments in the filtered parallel data used for baseline models.

uniform across language pairs, which had some influence on some of the workflows but the basic procedure of data cleaning was similar in all cases. As a general clean-up, we performed the following steps on the parallel data:

- language identification with FastText[3] (Joulin et al., 2016),
- segment deduplication with masked numerals, i.e. we deleted duplicate segments regardless of differences in numerals,
- deletion of segments where source/target token ratio exceeds 1:3 (or 3:1),
- deletion of segments longer than 100-150 tokens (depending on language pair),
- exclusion of segments where the ratio between the number of characters and the number of words was below 1.5 or above 40,
- exclusion of segments without a minimum number of alphabetic characters (2–5 depending on the data set).

These filtering steps led to an average reduction of about 15-20% of the training data with the number of segments as shown in Table 1.

2.1.1 Monolingual data

To build language models or create synthetic parallel text from monolingual data, we generally selected recent target language News Crawl data sets filtered according to the above steps (where applicable) with some minor adjustments. For En→De, we used the 2016–2020 German News Crawl data but as in the previous years excluded the 2018 set due to the high number of garbage segments with scrambled tokens, we set a threshold on the maximum length of a token (40) and the minimum ratio of letters to digits in a segment (4), and reduced the maximum segment length to 80 tokens, resulting in a 167M segment monolingual German data set. A similar procedure applied to the 2016–2020 English NewsCrawl corpus resulted in a monolingual English data set of 133M segments.

To create domain specific back-translation data for Fr→De we used the same data as for En→De, but due to the document based filtering method (see Section 3.2.2) the versions with document boundaries were used.

2.1.2 Development and test data

Development and test data sets were selected from the development suites provided. For En→De, we used the 2019 test set as validation set in the trainings and the 2020 test set as the test set to evaluate the trained models[4]. These data sets already contained only source original segments. We also extracted a source original subset from the full En→De development set, which was used in fine tuning of the final En→De models (see Section 3.2.3).

For Fr→De, the development set was shuffled and split into 3000 segment pairs for validation set and the rest (1813 segment pairs) for a general test set. To get an indication of the effect of data selection as described in Section 3.2.2, it was necessary to create a domain specific custom test set as well. The Fr→De 2008–14 development sets were filtered using a pattern based approach based on a

[3] https://fasttext.cc/docs/en/language-identification.html

[4] The reverse direction was used for the back-translation engines.

small list of 50 manually selected domain specific keywords[5], as well as scored and ranked by a target language model built from selected monolingual data (see Section 3.2.2). These two candidate lists were then manually revised and filtered to result in a 2k domain specific test set. These segments were removed from the training data.

2.2 Pre- and Postprocessing

Similarly to previous years (Oravecz et al., 2019, 2020) we opted for the simplest possible workflow leaving out the standard pre- and postprocessing steps of truecasing, or (de)tokenization, and simply used SentencePiece (Kudo, 2018), which allows raw text input/output within the Marian toolkit (Junczys-Dowmunt et al., 2018)[6] in the experiments. In some language pairs some simple normalization steps were applied in post-processing, which are described in the language pair specific result sections.

3 Trainings

In competitive systems big transformer architectures have become the norm in recent years (Barrault et al., 2020). We can in general see a significant increase in the need for computational resources to train deeper and more complex architectures up to 40–50 encoder layers (Wu et al., 2020b; Zhang et al., 2020; Wu et al., 2020a). Our resource environment does not allow us to fully follow this trend, limiting the complexity of the models as well as the scope of the experiments. Similarly to previous years, in all experiments we used Marian, as the core tool of our standard NMT framework in the eTranslation service. All trainings were run as multi-GPU trainings on 2 or 4 NVIDIA V100 GPUs with 16GB RAM, while for one training we were able to use a server with 8 32GB V100 GPUs.[7] Base transformers were typically trained for 20-30 epochs, whereas big transfomers were generally trained for 4–9 epochs for very high resource setups (>400M segments) and 20–25 epochs for medium resource.

[5]For example: Abwicklung, Betrug, Finanzbeitrag, Kapital etc.

[6]We did not change the default settings for Marian's built-in SentencePiece: unigram model, built-in normalization and no subword regularization.

[7]Access to high capacity resources at an affordable price has been especially challenging for us this year. In a race where computational power plays a crucial role (particularly in high resource settings) this might lead to an inherent disadvantage, which can be difficult to handle.

3.1 NMT Models

We only used base transformer models (Vaswani et al., 2017) for the first baseline models and for models used for back-translation to gain time and efficiency in back-translating large amounts of target monolingual data. For more competitive systems we switched to big transformer architectures, which resulted in significant improvements but at the same time the rise in computing costs and training time was also substantial. Due to the limitations of available resources we could build only one set of a 2–4 member ensemble from big transformers as our submission systems for En→De and Fr→De; again a high cost for a relatively smaller scale improvement. Our training settings have not changed from last year's setup: for most of the hyperparameters we used the default settings for the base transformer architecture in Marian[8] with dynamic batching and tying all embeddings. To save time and resources, we stopped the trainings if sentence-wise normalized cross-entropy on the validation set did not improve in 5 consecutive validation steps. In the big transformer experiments, also following recommended settings for Marian, we doubled the filter size and the number of heads, decreased the learning rate from 0.0003 to 0.0002 and halved the update value for `--lr-warmup` and `--lr-decay-inv-sqrt`.

Following common ranges of subword vocabulary sizes, we set a 36k joint SentencePiece vocabulary in En→De and En→Cs, and 30k in Fr→De.

3.2 Improving Baseline Models

In this section we briefly describe the methods we experimented with to improve the baseline models, such as selecting and filtering domain specific monolingual corpora to build additional synthetic data sets with back-translation (Sennrich et al., 2016), using development data (where available) or language model scored subsets of original parallel data to continue the training of already converged models and building ensembles of deep models originally trained from different seeds. Evaluation scores are reported in Section 4.

3.2.1 Filtering ParaCrawl

Training the En→De baseline model from the original parallel (OP) data (Table 1) we noticed that the model performed only as well (32.8 BLEU

[8]See eg. `https://github.com/marian-nmt/marian-examples/tree/master/transformer`.

on the 2020 test set) as our comparable model from last year despite having about twice as much ParaCrawl data while the other datasets remained basically very similar. This suggested that the v7.1 ParaCrawl (PC) data might have been noisier or contained more out of (news) domain data than expected. This was confirmed by training an alternative baseline excluding the whole ParaCrawl data set, which in the end resulted in a better score (33.3). To find a more beneficial subset of the PC data we first experimented with the stock Bicleaner filtering (Ramírez-Sánchez et al., 2020), setting higher thresholds of 0.65 and 0.75, which filtered the PC data to 51M and 26M segments, respectively. Adding either of these subsets to the other OP data sets did not lead to a significant increase (33.4 in both setups), however, we used the 51M segment subset instead of the full PC data in some further filtering experiments (see Section 3.2.3).

As a second filtering method we trained transformer language models (LM) with Marian from the filtered monolingual English and German data sets, scored both sides of the ParaCrawl data and ranked the segments (by simply averaging the scores). We experimented with models trained by adding the top 10, 20 and 30M highest scoring PC segments to the other OP data and found the 20M segment subset to produce the best baseline score (35.2), therefore we selected this data set (non ParaCrawl OP data plus the 20M segment LM scored ParaCrawl subset) as the initial parallel data for more complex models as well as for back-translation.[9]

3.2.2 Synthetic Data

Back-translation (BT) is the most used data augmentation technique in neural machine translation, but one which can introduce a wide range of scenarios in the search for finding the most optimal setup in the amount of synthetic data, the ratio of bitext to back-translation data or in the methods to generate the synthetic source (Edunov et al., 2018; Hoang et al., 2018). Tagged back-translation (Caswell et al., 2019) has been proposed as a simple and efficient alternative to noising techniques, arguing that it is the indication of the data being synthetic that is relevant for the model. This has been confirmed in our experiments in previous years, therefore we tried to use this technique in our workflows.

In the En→De system, we trained the reverse engine as a base transformer from the best baseline data setup mentioned above. After the convergence of this model we continued the training with a 30M segment subset of the OP data created by language model scoring (with the same models as for ParaCrawl). This gave an additional small increase in BLEU (0.4). With this model we back-translated an aggressively sentence segmented version of the filtered German monolingual data (see Section 2.1), which increased the size of the training set from the initial 167M segments to 219M. Our first intention was to build strong sentence based models and postprocess their output with dedicated sentence-to-document methods (which we describe in Section 3.2.5), so we tried to build one sentence per segment back-translated data sets by splitting up segments containing several sentences.

To train the submission ready systems we upsampled the best baseline OP data set to a 1:1 ratio with the BT data (Ng et al., 2019; Junczys-Dowmunt, 2019). This setup was a one shot configuration, we had no time and resources to experiment with other OP-BT combinations.

The task in the Fr→De language pair was domain specific, which offered us the opportunity to follow suit with the more recent shift from model centric approaches to data centric ones and focus on methods for finding the optimal subsets of the provided data which help improve performance in the selected domain. Therefore we tried to tune our models towards the domain by making use of guided topic modeling[10]. We created financial seed word lists by manually selecting 40 and 175 domain specific tokens from the top of a raw frequency list from a few million German News Crawl segments, and then we clustered the documents in the 2016, 2017, 2019 and 2020 German News Crawl data set into different topics guided by the selected seed word list.[11] By selecting the documents clustered into the seed word list induced topic we finally collected ca. 12M German News Crawl segments derived from two topic modelling runs based on one or the other list. These segments overlapped to a great extent. We back-translated both selections then cleaned up the back-translated data the way

[9] Clearly, there are other data selection combinations possible, for example, by taking only the 0.65 threshold Bicleaner subset as the base data for the LM based filtering, however, we did not have the time and resources to explore more scenarios for this language pair.

[10] https://github.com/vi3k6i5/guidedlda

[11] The text was tokenized and we used a German stopword list but no lemmatization in creating the document-term matrices.

we cleaned up the OP data but removed additionally pairs of segments that contained more than 15 numeric characters or more than 15 non-decimal commas. We also used the two sets to train two domain specific language models to score and rank the original parallel data set.

After that we took the union of the filtered BTs and deduplicated it. This gave us ca. 15M BT segment pairs which was at almost 1:1 ratio with the OP data. We explored training with subsets of the BT data but this did not give any improvement so we decided to use it all. We also experimented with tagged and untagged BT data, of which somewhat unexpectedly the latter gave the better result. The reason might be that the BT data was more in-domain, while most of the OP data was out of (news) domain and the explicit OP vs. BT distinction might have presented a harmful signal to the model here.

3.2.3 Continued Trainings and Fine Tuning on Dev Sets

As last year, in the En→De system we followed a two-stage continued training process to improve performance as domain adaptation (Luong and Manning, 2015). We scored the non ParaCrawl OP plus the 0.65 threshold ParaCrawl subset (see Section 3.2.1) with the language models used for filtering the ParaCrawl data set (Section 3.2.1). Then we used the top 10, 20 and 30M subset to continue the training of the OP+BT converged models until the BLEU score on the test set increased (Junczys-Dowmunt, 2019); typically 2 epochs with an increase of 0.5 points. The second stage utilized the 2008–2019 development sets (34k segments) as fine tuning data in the experiments and for the final submission it was extended with the 2020 test set. We trained with reduced batch size and learning rate for 4 epochs on this set and then for additional 3 epochs we switched to a source original subset (16k) to reach the highest BLEU score. In the end this process gave only a minor improvement of 0.3 BLEU points.

For Fr→De, we experimented with fine-tuning the best converged models (see Section 4.2) by using different sets of in-domain data. We scored the OP data for domain, using the two different LMs as mentioned above. Then, we selected the top 1M segments of each scored set of OP data and intersected them. This gave us ca. 0.85M segment pairs. However, this approach was not successful. In the other setup, we selected the top 2M segments of each scored set of OP data and intersected them, which gave us ca. 1.75M segments. We fine-tuned with reduced batch size until the BLEU score increased, which gave us an increase of 0.8 points on the domain specific test set.

3.2.4 Ensembles

The En→De final submission consisted of a modest 4 model big transformer ensemble, trained with the same best configuration and workflow but with different seeds. This approach usually gives a small but steady improvement (about 0.5 BLEU points here) but for substantially high resource settings it also comes with large computational costs. It is not uncommon to use ensembling already for back-translation (Wu et al., 2020b) but for lack of time and resources we had to limit this technique to the submission setups.

The Fr→De ensemble was composed of 4 big transformer models – three of them trained on original parallel data and back-translated data in ratio 1:1. The 4th big transformer was one of the 3 big transformers, additionally fine-tuned for 7 epochs on the 1.75M OP data scored with the domain LMs. For lack of time it was only one experimental setup out of many other possible ones but proved to be better than our previous systems.

3.2.5 Methods Tested but not Selected for Submission Models

In the En→De system, this year we experimented with a two-stage translation process of using a strong sentence-level system at the first step and post-process its output with a dedicated sentence-to-document level model. Following the method proposed by Voita et al. (2019), we created a 100M segment synthetic dataset by round-trip translating the (filtered) 2019 and 2020 German News Crawl with document boundaries with the baseline sentence level (forward and reverse) systems, and then generating 1, 2, 3 and 4 sentence long "source German"–"target German" pairs from the round-trip translated segments and the sentences in the original News crawl documents. We trained a base transformer from this data set and used it as a second stage repair on the output of the best En→De sentence level system. Unfortunately, we observed a significant drop in BLEU (almost 5 points) and although this is somewhat consistent with what for example Ma et al. (2021) reports on automatic evaluation for this method, we did not want to take the risk of submitting a system with such a quality drop

on the automatic metric to manual evaluation.

4 Results

We submitted a constrained system for each of the 3 language pairs. For En→Cs, we ran out of time and had to reuse our last year submission. For the other language pairs, we provide the evaluation scores for models at important stages in the development, which reflect how the models got better as we tried various methods for improvement. All results are reported in detokenized BLEU.[12]

4.1 English→German

		Test sets	
System	Data	2020	2021
M1: Baseline	12M	33.3	–
M2: M1+PC	32M	35.2	–
M3: M2+BT^{bigT}	450M	36.7	–
M4: M3 tuned	450M+36k	37.5	–
M5: M4 ensemble	450M+36k	**38.0**	**29.6**

Table 2: Results for En→De models. The 2021 result is from the Ocelot submission.

In Table 2 we present the main stages of the development of the En→De systems. Model 1 was the initial baseline model and used only the original parallel data excluding ParaCrawl altogether. In Model 2 we added the language model filtered and scored top 20M subset from ParaCrawl (PC). For Model 3, we switched to the big transformer architecture and used the large aggressively segmented back-translation (BT) dataset with 1:1 upsampled original parallel data (OP). The next model (M4) was tuned for 3 additional epochs with the top 10M LM scored OP data and then with the development set, leading to a small but steady increase. Finally the system we submitted was an ensemble of four M4 models. Our primary system being a sentence-level model, we performed sentence segmentation as a preprocessing step and then simply remerged the sentence level hypotheses on the target side where needed. Finally, as in previous years, a post-processing step normalizing German punctuation and some space fixing around the % sign was run on the final output.

[12]sacreBLEU signatures: `BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+tok.13a+version.1.4.13`

4.2 French→German

Table 3 summarizes the results of the Fr→De experiments. The first baseline model (M1) was trained only on the original parallel data with news data upscaled 5 times (NewsCrawl, NewsCommentary), while in model 2 and 3 (M2, M3) we added the domain specific back-translated data set (as described in Section 3.2.2). Switching from base transformers (M1 to M3) to the big transformer architecture in model 4 (M4) led to a decent improvement. This setup was used for the models in the M5 three model ensemble. In the primary submission (M6) this was extended with a 4th big transformer. In M6, the 4 models were trained on the original parallel (OP) data and back-translated data (in ratio 1:1), and one of the models was additionally fine-tuned for 7 epochs on the 1.75M domain LM scored original parallel data subset (see Section 3.2.3).

4.3 English→Czech

Due to problems with computational resources, the En→Cs trainings had not finished until the submission deadline. Our primary submission presented in Table 4 is therefore a clone of the 2020 system (trained on OP plus BT data).

5 Conclusion

We presented the submissions of the eTranslation team to the WMT 2021 news translation shared task on 3 language pairs: English-German, French-German and English-Czech. Unlike in previous years, we had to face a few unexpected challenges with respect to resource availability, which inevitably affected some experiments we planned to carry out. We tried to put more emphasis on data selection, filtering and domain specific evaluation with custom test sets in the task where it seemed to be most rewarding and automatic evaluation results justified this approach.

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages

System	Data	Test sets		
		Dev	Domain	2021
M1: Baseline	13.6M	27.1	–	–
M2: M1+BT (tagged)	27.3M	29.1	–	–
M3: M1+BT (untagged)	27.3M	30.9	–	–
M4: M3 as big Tr.	27.3M	31.9	24.0	–
M5: M4 ensemble	27.3M	32.5	24.5	40.2
M6: M5+FT	27.3M	32.3	**25.0**	**40.6**

Table 3: Results for Fr→De models. The 2021 results are from the Ocelot submissions.

System	Data	Test sets	
		2020	2021
Submission	166M	**35.7**	**21.5**

Table 4: Results for En→Cs models. The 2021 result is from the Ocelot submission.

1–55, Online. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. 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*, pages 116–121. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 76–79.

Zhiyi Ma, Sergey Edunov, and Michael Auli. 2021. A comparison of approaches to document-level machine translation. *CoRR*, abs/2101.11040.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Csaba Oravecz, Katina Bontcheva, Adrien Lardilleux, László Tihanyi, and Andreas Eisele. 2019. eTranslation's submissions to the WMT 2019 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 320–326, Florence, Italy. Association for Computational Linguistics.

Csaba Oravecz, Katina Bontcheva, László Tihanyi, David Kolovratnik, Bhavani Bhaskar, Adrien Lardilleux, Szymon Klocek, and Andreas Eisele. 2020. eTranslation's submissions to the WMT 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 254–261, Online. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Liwei Wu, Xiao Pan, Zehui Lin, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2020a. The Volctrans machine translation system for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 305–312, Online. Association for Computational Linguistics.

Shuangzhi Wu, Xing Wang, Longyue Wang, Fangxu Liu, Jun Xie, Zhaopeng Tu, Shuming Shi, and Mu Li. 2020b. Tencent neural machine translation systems for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 313–319, Online. Association for Computational Linguistics.

Yuhao Zhang, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, Yongyu Mu, Jingnan Zhang, Xiaoqian Liu, Xuanjun Zhou, Yinqiao Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2020. The NiuTrans machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 338–345, Online. Association for Computational Linguistics.