Samsung R&D Institute participation in WMT 2022 General MT Task

Adam Dobrowolski[1], **Mateusz Klimaszewski**[*2], **Adam Myśliwy**[1], **Marcin Szymański**[1], **Jakub Kowalski**[1], **Kornelia Szypuła**[1], **Paweł Przewłocki**[1], **Paweł Przybysz**[1]

[1]Samsung R&D Institute, Warsaw, Poland

[2]Warsaw University of Technology, Warsaw, Poland

{a.dobrowols2, a.mysliwy, m.szymanski, j.kowalski5, k.szypula, p.przybysz}@samsung.com

mateusz.klimaszewski.dokt@pw.edu.pl, p.przewlocki@partner.samsung.com

Abstract

This paper presents the system description of Samsung R&D Institute Poland participation in WMT 2022 for General MT solution for medium and low resource languages: Russian and Croatian. Our approach combines iterative noised/tagged back-translation and iterative distillation. We investigated different monolingual resources and compared their influence on final translations. We used available BERT-like models for text classification and for extracting domains of texts. Then we prepared an ensemble of NMT models adapted to multiple domains. Finally we attempted to predict ensemble weight vectors from the BERT-based domain classifications for individual sentences. Our final trained models reached quality comparable to best online translators using only limited constrained resources during training.

1 Introduction

Samsung R&D Institute Poland (SRPOL) participated in the WMT 2022 General MT task for three translation directions: EN→RU, RU→EN and EN→HR. All our systems were built using only constrained datasets. In contrast to previous years, where the task focused on news translation, this year's task was domain-independent. However, MT models benefit a lot from domain adaptation. Therefore, we decided to prepare an ensemble of NMT models adapted to multiple domains to benefit from domain adaptation and improve generalization. We prepared a news profiled model but also a general-purpose one. Additionally, we worked on medical and legal domains; however, there was very limited in-domain data in the constraint path for this domains and we had to extract pseudo in-domain data from monolingual corpora.

Our system was implemented using Marian framework. The core of the submitted solution is iterative back-translation and iterative distillation combined with finetuning and ensembling. Besides, we used BERT models for data filtering to prepare corpora for training domain-adapted models. Finally, we created dynamic ensemble weighting to choose the best combination of single models in the final translations. All techniques combined allowed us to improve baseline models by 3-6 BLEU (Papineni et al., 2002) and reach the quality comparable with online translators (measured by BLEU).

2 System overview

2.1 MT model

Our models were trained with the Marian NMT (Junczys-Dowmunt et al., 2018) toolkit. We used Marian for training, back-translation, noise generation, language models and data filtering.

The training was performed on a *transformer-big* model (embedding dimension of 1024 and a feed-forward layer dimension of 4096) (Vaswani et al., 2017). We experimented with different sizes of models and different configurations of encoder-decoder layers, but we achieved no significant improvement over the default *transformer-big* configuration. Most models had a setup of either 7-5 or 8-4 encoder-decoder layers. Best single models were trained with FF layer dimension 6144, but the improvement was marginal – 0.1 BLEU better than the default dimension of 4096.

Our training used batches of size 256GB (8xGPU, 32GB workspace). The optimizer was Adam (Kingma and Ba, 2015) with a learning rate of 0.0003 and linear warm-up for the initial 40 000 updates with subsequent inverted squared decay. A few initial EN↔RU training were regularized with dropout 0.1, but the following did not use any dropout. All training for EN→HR had the dropout set to 0.1.

2.2 Iterative training process

Iterative back-translation (Hoang et al., 2018) is a known technique of improving performance of

[*]Work done while at Samsung R&D Institute Poland.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 251 - 259

MT models. Iterative distillation approach applied by NiuTrans (Zhou et al., 2021) allowed them to achieve impressive results in WMT21. During our work we combined both techniques in parallel during each iteration.

First baseline models were trained using only provided parallel corpora. Further training iterations were enriched with back-translation (iterative back-translation). With each iteration we used new back-translation prepared by best ensembles translating from target to source.

After a few iterations of iterative back-translation we started iterative distillation. Training corpus was enriched with corpora distilled from best ensembles. (→ 3.3). As a result the whole corpus consisted of parallel part, back-translated part and distilled part.

After training iteration converged we finalized the iteration with additional tuning using parallel corpora or specialized tuning corpora (→ 3.4). After the tuning we selected a new best ensemble containing the new trained model. The best ensemble was chosen by selecting the best performing on Flores devtest (Goyal et al., 2022) and Newstest 2021. With this new ensemble we prepared new back-translation and a new distilled corpus for next iterations.

Algorithm 1 Iterative training process

procedure ITERATEDTRAININGS
  $M_{enru} \leftarrow$train$(bitext_{enru})$
  $M_{ruen} \leftarrow$train$(bitext_{ruen})$
  while $models$ not converged **do**

    $bktr \leftarrow$translate$(mono_{en}, M_{enru})$
    $dist \leftarrow$distill$(bitext_{ruen}, M_{ruen})$
    $corpus = bktr + dist + bitext_{ruen}$
    $model_{ruen} \leftarrow$train$(corpus)$
    $model_{ruen} \leftarrow$tune$(tuning_corpus_{ruen})$
    $M_{ruen} \leftarrow$getBestEns$(models_{ruen})$

    $bktr \leftarrow$translate$(mono_{ru}, M_{ruen})$
    $dist \leftarrow$distill$(bitext_{enru}, M_{enru})$
    $corpus = bktr + dist + bitext_{enru}$
    $model_{enru} \leftarrow$train$(corpus)$
    $model_{enru} \leftarrow$tune$(tuning_corpus_{enru})$
    $M_{enru} \leftarrow$getBestEns$(models_{enru})$

  end while
end procedure

2.3 Domain adaptation

WMT 2022, for the first time, allowed the usage of pre-trained masked language models (MLM; exclusively in BERT-based architecture). We leveraged them to extract domain-specific subsets of mono and parallel corpora to fine-tune our NMT models in two chosen domains: *legal* and *medical*. We divide our approach into three steps: 1) Rule-based seed extraction, 2) Iterative Classifier training 3) Domain corpora extraction. Domain adaptation was performed only for the EN↔RU language pair. Finally, we used corpora described in Section 2.3.4, to adapt to the competition test sets.

2.3.1 Rule-based seed extraction

Our work focuses on two non-news domains: *medical* and *law*. We prepared initial monolingual (EN) seed corpora based on handcrafted rules and manual filtering. The datasets were too small to perform fine-tuning of MLM; therefore, we added an intermediate step. We encoded the sentences using general-purpose BERT (Devlin et al., 2019) and applied a K Nearest Neighbours (KNN) classifier to filter the extended version of the initial corpora. The extended version was extracted using the same rules but without manual filtering.

2.3.2 Iterative Classifier training

We base our approach on tri-training (Zhou and Li, 2005; Ruder and Plank, 2018). Rule-based extracted seed serves as the training data, and the manually filtered examples are the test set. In contrast to the original tri-training, we enlarge our training dataset after training the three classifiers instead of continuously adding new examples during training (we call this an iteration). Due to time constraints, we performed two such iterations per domain. The classifiers are fine-tuned BERT models, yet domain-specific ones: Lee et al. (2019) for the *medical* domain and Chalkidis et al. (2020) – *legal*.

2.3.3 Domain corpora extraction

With the final ensemble of classifiers, we scored a subset of monolingual, English data (CommonCrawl) and parallel corpora, which was not used during the ensemble training. We raised a threshold for the classifiers to 0.9 and included a sentence to a domain using unanimous voting. The resulting monolingual/parallel corpora size is presented in Table 1.

	Domain	
Corpora	Medical	Legal
Monolingual	93.3	184.6
Parallel	5.1	135.5

Table 1: Size of extracted domain-specific corpora (in thousands)

2.3.4 Test set adaptation

The last step of domain adaptation was the WMT 2022 test set adaptation. Our main intention was to prepare a corpus based on sentences similar to those present in the competition test set. To achieve this goal, we used the KNN algorithm. The first step was creating a dataset consisting of sentence embeddings from the WMT 2022 test set and all constrained corpora. Embeddings were acquired using the BERT base model (cased) (Devlin et al., 2019). Afterwards, we applied the k nearest neighbours search. The parameters were selected empirically: the number of nearest neighbours was set to 20, and we chose the Euclidean distance metric. Finally, the candidates were picked by finding neighbours whose distance to a given sentence from WMT 2022 test set was lower than 1.2.

2.4 Dynamic ensemble weighting

For each given (expert-selected) collection of NMT models, two modes of ensemble translation were tested. In the standard mode, the entire test set is translated using the same "static" set of weights for ensemble components. Alternatively, we attempted to construct a regression model that would generate weights best suited to a given sentence type; we call this mode "dynamic". For this, we concatenated outputs from 3 BERT-based predictors, trained to classify sentences as belonging to legal, medical and news domain, respectively. The medical and legal predictors were as described in 2.3.2; the predictor for news domain was fine-tuned in the same way with the pretrained BERT model allenai/news-roberta-base. Because each predictor produced 6 strongly-correlated values, the resulting vectors underwent dimensionality reduction, before being passed as inputs to the weight regression model; the regression itself is a relatively simple affine transformation in the logit domain. We leveraged only English Bert models; therefore, in the RU→EN direction, we performed preliminary translation using some early ensemble and extracted the predictions from its English outputs; the Croatian task does not use weight optimization.

Because we could not perform a direct optimization of BLEU/chrF (Popović, 2015) with regard to ensemble weights (some sort of grid- or random-search would be possible, but was deemed too expensive), we settled on minimizing cross-entropy of reference translations. We experimented with two formulas for interpolation of probability distributions: in logarithmic-probability domain (more commonly found, e.g. in Marian), or in linear-probability domain.[1] However, because the minimization of cross-entropy in log-P domain will degenerate the ensemble to the single best model (it can be easily shown), we added the regularization parameter to optimization of this kind of ensembles. The regularization term penalizes the divergence from the uniform vector.

26k sentences were selected from the model training corpora as the training data, half of which was classified as news, the rest as legal, medical, or randomly sampled. Three validation sets were used: Flores, Newstest 2021 and training data held-out.

Static and dynamic weights were independently estimated using gradient-descent for a handful of different ensembles in each direction; the general observations on development sets were the following:

- for each of the directions, two different vectors/transforms seem to be optimal, depending on the development set (one for Flores, another for Newstest 2021 and held-outs)
- the impact of the interpolation model (log-P vs linear-domain) is moderate, usually with small advantage of log-P, except for EN→RU Flores where linear yields ca. +0.22 BLEU
- the impact of the dynamic weighting is minimal, giving 0.09 BLEU improvement on RU→EN direction, with 0.1–0.3 BLEU degradation on top EN→RU configurations.

For final submission, in RU→EN direction we used static ensembles as described; however, in EN→RU task we made a last-minute decision

[1] We added an in-house extension to Marian-NMT that implements this alternative ensemble interpolation (i.e. done in the linear-probability domain); a patch that facilitates running ensemble translations with a weight vector different for every sentence was also implemented.

to scrap automatically-derived weights and used expert-crafted ensembles (obviously, also static).

We conjecture that the reason for the limited benefits from the above experiments lies in the indirect optimization of BLEU through cross-entropy, as well as – in the dynamic approach – in small actual distinctiveness of domain-specific data.

3 English-Russian

All corpora were preprocessed by removing sentences of inappropriate languages, normalizing punctuation, replacing all Russian letters ë (yo) with e (ye), removing duplicate sentences.

3.1 Parallel corpora

During the training, we used all accessible English-Russian parallel data except UEDIN back-translated news corpus. This corpus was used only during the first training iteration before generating any new back-translated data. Later it was excluded from training because it was worsening the results. We filtered sentence pairs where the length ratio between source and target sentences exceeded 1.6. Paracrawl (Bañón et al., 2020) paragraphs consisting of more than one sentence were split into single sentences and appended to the original dataset.

We used our in-house rule-based filtering, but we did not detect improvement but worsened quality over not-filtered data. Similarly, inferior results were obtained by applying Cross-Entropy Filtering (Junczys-Dowmunt, 2018). Therefore, we used unfiltered data during most of the training process.

3.2 Monolingual corpora

We used the monolingual corpora in two ways: to train language models and to augment the parallel data with back-translated data. Back-translation (Sennrich et al., 2016) is a commonly used technique for improving machine translation, especially for low-resource languages (Edunov et al., 2018).

We chose three different sources of monolingual corpora and preprocessed them similarly to parallel data (with minimal preprocessing). The used corpora are:

- News crawl
- CommonCrawl
- News-CommonCrawl

All corpora were filtered by a language model trained on the same corpus leaving only sentences with a likelihood larger than 1e-5. Due to the poor quality of CommonCrawl, we used only lines/paragraphs containing three or more sentences, which we split into single sentences.

News-CommonCrawl is the same filtered CommonCrawl but additionally filtered by a fastText[2] model trained on 100k news sentences from News crawl and 100k sentences from CommonCrawl. Using this model, we selected sentences classified by fastText as news (Joulin et al., 2017).

During all training iterations, except the first, we back-translated monolingual data using the best ensembles of currently trained models. We used clean back-translation as well as noised (Edunov et al., 2018) and tagged back-translation (Caswell et al., 2019). We applied gumbel noise for noised back-translation, as implemented in Marian, changing the epsilon value from default 1e-5 to 1e-3.

3.3 Teacher-Student Knowledge Distillation

Distilled corpora were prepared by translating parallel corpora using best ensembles in the direction of training with a beam equal to eight and selecting two translations most similar to the original translation. Such corpus was added to the parallel corpus expanding it three times.

3.4 Tuning corpora - FLORES

Despite poor results of standard filtering, we experimented with modified filtering versions during further iterations. We finally found the following filtering by marian-scorer that applied to parallel corpora improved results in some of the final training iterations.

- Language model filtering - Using a language model trained on a monolingual corpus we filtered utterances for which the normalized likelihood of the target side was higher than 1e-5.
- Backward cross-entropy filtering - Using the backward translation model, we filtered only sentence pairs where target to source translation normalized likelihood was larger than 1e-2.

The filtering described above was not applied to the Wikititles corpus.

[2] https://fasttext.cc

3.5 Tuning corpora - NEWS

Models adapted for news were finetuned by two consecutive tuning iterations using the following corpora:

1. Paracrawl and News Commentary
2. News Commentary and all Newstests from WMT2012-20

3.6 Contextual corpus and decoding

The corpus used for contextual training translation was built of two parts:

- Parallel utterances from News Commentary containing 2-4 subsequent sentences.
- Sequence of 2-4 adjacent sentences from one paragraph of CommonCrawl monolingual corpus, back-translated sentence by sentence. The back-translated part was tagged.

During decoding, we translated a sentence four times:

- without a context
- with one preceding sentence
- with two preceding sentences
- with two preceding and one following sentence

From the four above translations, we chose the translation most similar to 3 others using Levenshtein distance (Levenshtein, 1965) as a similarity metric.

4 English-Croatian

We applied similar preprocessing as for Russian language. Additionally to all available EN-HR corpora from OPUS (Tiedemann, 2012) we added all available data for Serbian language to the training. We used custom validation set based on TED for first iterations and WMT22 dev set for last two iterations. We added directional tokens in front of each sentence that allowed to differentiate between Croatian and Serbian translation.

For back-translation we used news mono corpora and source language from all EN-HR parallel corpora as well. Additionally to the back-translated corpora we added EN-HR parallel data. We performed two iterations of back-translation. After training of first iteration with back-translated data we fine-tuned the model on all parallel EN-HR data. After training of the second iteration we fine-tuned the model on CCMatrix corpus (Schwenk et al., 2021). The back-translation was noised with gumbel noise.

After the above we started to apply knowledge distillation and fine-tuning the model on distilled data. We did only 2 iterations of distillation. First distillation was done on CCMatrix corpus and second on tuning corpus (created from DGT, QED, TedTalks, EuroPat, SETIMES, hrenWaC, TED2020 corpora). We experimented with different learning rates in order to find the best performing model after this step. Finally, we made an ensemble out of the best-performing models. Additionally, we found that a normalization value of 0.5 results in a better score.

5 Results

Results of training iterations for English to Russian are presented in Table 2. Table 3 presents results for the Russian to English direction. Finally, Table 4 presents results for the English to Croatian task. Abbreviations mean:

- BTN - noised back-translation
- BTT - tagged BT
- BTTN - tagged noised BT
- KD - training with distilled parallel corpus
- news / cc / ncc - back-translated corpus
  - News crawl
  - CommonCrawl
  - News-Commoncrawl

First iteration was trained using only constrained parallel corpora provided by organizers. Next iterations were trained on mixed parallel corpora combined with back-translated monolingual data (BT). Further iterations used also distilled forward translations (KD).

Tuning with domain adaptation corpora has improved slightly (0.1-0.2) some of single models but gave no noticeable improvement on final score of ensembles.

6 Conclusions

We confirmed that iterative knowledge distillation combined with iterative back-translation is sufficient to prepare high-quality translation models.

Iter	Corpus	Flores devtest	Newstest 2021
0	Parallel – baseline	30.1	26.8
1	BTN-news	**32.5**	**28.8**
	BTN-news, filtered bitext	31.9	28.3
2	KD BTN-news	33.7	29.5
	KD BTT-news	**34.0**	**29.6**
	KD BTT-cc	33.9	28.9
3	KD BTT-news	**34.1**	29.4
	KD BTT-news, tuned news	33.9	**29.9**
4	KD BTN-news	33.0	29.4
	KD BTTN-news	33.7	29.2
	KD BTT-news	34.0	29.6
	KD BTT-news, tuned news	33.6	**30.1**
	KD BTT-news + context	33.8	29.7
	KD BTT-cc	34.4	28.9
	KD BTT-cc + context	**34.5**	28.8
Best ensemble flores - SRPOL submission		34.8	30.7
Best constrained WMT2021			**29.3**

Table 2: Iterations and results of training for EN→RU direction.

Iter	Corpus	Flores devtest	Newstest 2021
0	Parallel – baseline	35.6	35.5
1	BTN-news	**37.0**	36.5
	KD + BTN-news	36.4	**37.9**
2	BTN-news	**36.6**	**37.1**
	BTN-news, filtered bitext	36.2	36.7
3	BTN-news	37.4	37.6
	BTN-ncc	**38.1**	36.7
	KD + BTN-ncc	37.8	37.9
	KD + BTTN-ncc	37.4	**39.0**
4	KD + BTT-ncc	37.0	38.8
	KD + BTN-ncc	38.0	38.1
	KD + BTTN-ncc	37.4	38.5
	KD + BTT-ncc tuned news	37.0	**40.2**
	KD + BTN-ncc tuned news	37.8	39.8
	KD + BTN-ncc + context	**38.1**	38.2
	KD + BTN-ncc + context tuned news	37.6	39.7
Best ensemble flores - SRPOL submission		**38.9**	40.8
Best ensemble news		38.3	**41.6**
Best constrained WMT2021			**41.8**

Table 3: Iterations and results of training for RU→EN direction.

Iter	Corpus	Flores devtest	WMT 22 devtest
0	Parallel – baseline	31.9	32.1
1	BTN	32.7	32.0
2	BTN	32.8	32.2
3	KD	33.5	33.4
4	KD	33.7	33.3
Best ensemble + normalization - SRPOL submission		33.6	33.7

Table 4: Iterations and results of training for EN→HR direction.

This method gives excellent results on low-resource and mid-resource languages. During the WMT 2022 General MT task, we reached one of the best results among constrained systems.

In our work, we compared different methods of back-translation: clean, noised, and tagged. Mostly, the tagged back-translation achieved the best results, but for some training iterations, noised back-translation's results were on-par or better.

We compared different sources of monolingual data used for back-translation: CommonCrawl and News crawl. The comparison suggests that the choice of the monolingual corpus has a significant influence on final results.

Our exploration of different filtering methods suggests that while using pre-filtered data (as provided in WMT 2022), it is sufficient to filter only target data, leaving source data unfiltered.

We presented a simple and effective method of adding contextual data to the training corpus, which gave a noticeable improvement.

We investigated a new method of dynamic ensemble weighting, but the results show no improvement over other methods.

Limitations

In our work we touched on a few aspects but did not have time to address them in more detail.

The research showed that tagged back-translation generally gives better results than other back-translation methods, but not always. It may be worth to investigate more deeply methods of different noising, different noise level and how it synergies on various parallel and monolingual corpora.

Almost all our training iterations were performed on very similar default transformer-big configurations. We haven't tested other configurations, larger or deeper models, different training parameters, what can improve the results.

We introduced very simple contextual translation method which can be improved in many ways.

We gained best results filtering data only on target size, leaving source data unfiltered. This issue looks worth to be investigated.

Ethics Statement

During our work we followed all the rules of ACL Ethics Policy

All our efforts aimed at conducting research for the benefit of society and to human well-being. During our research we used only fair and honest methods. We didn't hide any information needed to repeat our results. We didn't use any resources out of the data provided by organizers in constrained path.

Acknowledgements

We would like to thank all the people who helped us during the research with their consultations and resources. Special thanks to SRPOL's Linguistic Department, SRPOL's Management Staff, Samsung Research HQ in Suwon, Korea.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, 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.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. 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.

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*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and*

Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shuhan Zhou, Tao Zhou, Binghao Wei, Yingfeng Luo, Yongyu Mu, Zefan Zhou, Chenglong Wang, Xuanjun Zhou, Chuanhao Lv, Yi Jing, Laohu Wang, Jingnan Zhang, Canan Huang, Zhongxiang Yan, Chi Hu, Bei Li, Tong Xiao, and Jingbo Zhu. 2021. The NiuTrans machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 265–272, Online. Association for Computational Linguistics.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541.