AIST AIRC Systems for the WMT 2024 Shared Tasks

Matīss Rikters[1]
[1]Artificial Intelligence Research Center (AIRC)
National Institute of Advanced Industrial Science and Technology
matiss.rikters@aist.go.jp

Makoto Miwa[1,2]
[2]Toyota Technological Institute, Japan
makoto-miwa@toyota-ti.ac.jp

Abstract

This paper describes the development process of NMT systems that were submitted to the WMT 2024 General Translation and Biomedical shared tasks by the team of AIST AIRC. At WMT 2024 AIST AIRC participated in the General Machine Translation shared task and the Biomedical Translation task. We trained constrained track models for translation between English, German, and Japanese. Before training our models, we first filtered the parallel data, then performed iterative back-translation and additional filtering. We experimented with training baseline Transformer models, Mega models, and fine-tuning open-source T5 and Gemma model checkpoints using the filtered parallel data. Our primary submissions contain translations from ensembles of two Mega model checkpoints and our contrastive submissions are generated by our fine-tuned T5 model checkpoints.

1 Introduction

We describe the machine translation (MT) systems submitted to the WMT 2024 General Translation and Biomedical Translation tasks developed by the team of AIST AIRC. We experimented with data quality control by filtering out noisy examples from parallel and monolingual data sets before training, and corpora selection. We also compared several modeling approaches by contrasting our previous year's best constrained submission (Rikters and Miwa, 2023) – the Mega model (Ma et al., 2023) to open track approaches of fine-tuning T5 (Raffel et al., 2020) and Gemma (Mesnard et al., 2024) model open-source checkpoints. When fine-tuning T5 and Gemma models, we experimented with adding named entity (NE) annotations (Rikters and Miwa, 2024) to improve rare word translation, since struggling to correctly translate less common NEs was one of the most common errors identified in human evaluations of our WMT 2023 submissions.

2 Data

In the General Translation task we only participated in the constrained track, so our data selection was limited to only the parallel corpora provided by the shared task organizers, which for German and Japanese was unchanged from the previous year. For the Biomedical Translation task we used a combination of General Translation task data and Biomedical Translation task data.

All parallel training data and monolingual data for back-translation were filtered before starting any training, which has been proven very effective in previous WMT shared tasks (Pinnis et al., 2018). The filtering process we used is detailed by Rikters (2018). We did not perform any parallel data distillation for our submissions this year.

For the system development process in the General Translation task, we selected News Test sets from the WMT 2022 shared task as development data and test sets from WMT 2023 as evaluation data. Statistics of the data we used are shown in Table 1. For the Biomedical Translation task we used the same combination of 2022 and 2023 development / evaluation data sets.

2.1 Data Selection

To not overwhelm the full combined training data set with lower-quality web-crawled data, we 1) limited the English-German Paracrawl to 50 million parallel sentences; and 2) up-scaled all data from other sources to match the amount of the Paracrawl data after filtering by doubling for English-German and tripling for English-Japanese.

2.2 Filtering

Even though all training data need not always be perfect and methods like back-translation intentionally generate somewhat noisy additional training data, some types of noise are more harmful than others. Since most training corpora are produced

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 286–291
November 15–16, 2024.

Corpus / Filtering		EN-DE	EN-JA
All other	Before	16,752,302	8,076,155
	After	13,737,028	7,076,869
Paracrawl	Before	50,000,000	21,891,738
	After	44,533,635	21,088,689
Combined		72,007,691	42,319,296
Medline		45,796	-
UFAL Medical		3,036,581	-

	Monolingual	
Corpus / Filtering	Before	After
DE	43,613,631	37,110,981
JA	22,193,545	21,558,123
EN	47,333,840	36,756,542

Table 1: Training data statistics for all other parallel data without Paracrawl, a subset of Paracrawl, combined development and evaluation data from the past WMT shared tasks, and monolingual data. Sentence counts are listed before and after filtering.

partially or fully automatically, errors such as misalignments between source and target sentences or direct copies of source to target can occur, as well as some amounts of third language data in seemingly bilingual data sets.

To avoid such problems, we used data cleaning and pre-processing methods described by Rikters (2018). The filtering part includes the following filters: 1) unique parallel sentence filter; 2) equal source-target filter; 3) multiple sources - one target and multiple targets - one source filters; 4) non-alphabetical filters; 5) repeating token filter; and 6) correct language filter. We also perform pre-processing consisting of the standard Moses (Koehn et al., 2007) scripts for punctuation normalization, cleaning, and Sentencepiece (Kudo and Richardson, 2018) for splitting into subword units for training MEGA models, and the default tokenizers for T5 and Gemma. The filters were applied to the given parallel sentences, monolingual news sentences before performing back-translation, and both sets of synthetic parallel sentences resulted from back-translating the monolingual news.

2.3 Back-translation

Increasing the amount of in-domain training data with synthetic back-translated corpora (Sennrich et al., 2016) is a common practice in cases with considerable amounts of in-domain monolingual data. However, since the task recently shifted from 'news' to 'general' text translation, the definition of what would be considered in-domain data became less clear. Furthermore, for the constrained track the selection of provided monolingual data from the organizers is still limited to news and web-crawled data. No other monolingual data that would be considered more similar to what the 'general' test sets may include, such as user generated (social media), conversational, and e-commerce data are provided in the task. For our experiments we continued to assume that a significant portion of the test data would still be from the news domain. Therefore, we chose to only use the provided monolingual News crawl, News discussions, and News Commentary corpora for back-translation.

2.4 Post-processing

In post-processing of the model output we aimed to mitigate some of the most commonly noticeable mistakes that the models were generating. We mainly noticed two often occurring problems in output from all models: 1) difficulties in translating emoji symbols; and 2) occasional repetitions of words or phrases.

While all English and German alphabet letters and even Japanese characters are covered in the large training data corpora, the Unicode emoji were mostly formed and clearly defined only in the past decade, and new emoji are still added every year or two with the next release planned for late 2024[1]. Emoji are also not often present in MT training data, therefore full emoji coverage is absent from model vocabularies, which leads to occasional <*unk*> tokens being generated as output if emoji were present in the input. In order to keep using the models without re-training, we replaced any <*unk*> tokens in the output using a dictionary of any emojis appearing in the input.

Furthermore, the occasional hiccuping or hallucinating of models on less common input sequences seems to still be present, sometimes generating repetitions of tokens or phrases. We replaced any consecutive repeating n-grams with a single n-gram. The same was applied to repeating n-grams that have a preposition between them, i.e., *the victim of the victim*.

Both post-processing approaches gave BLEU score improvements of around 0.1 - 0.2.

[1] https://emojipedia.org/unicode-16.0

3 Model Configurations

While preparing our submissions we experimented with three main model types between the constrained and open system tracks. For our primary submission we chose the constrained Mega models similar to our last year's primary submission (Rikters and Miwa, 2023), and for contrastive submissions we used T5 models (Raffel et al., 2020) fine-tuned on NE-annotated General Translation task data, and Gemma models (Mesnard et al., 2024) tuned on General Translation task data.

3.1 Mega

Ma et al. (2023) proposed a moving average equipped gated attention mechanism (MEGA) - a single-head gated attention mechanism equipped with exponential moving average to incorporate inductive bias of position-aware local dependencies into the position-agnostic attention mechanism. Compared to the Transformer model, MEGA has a single-head gated attention mechanism instead of multi-head attention, which enables gains in efficiency while not sacrificing on performance.

For training our Mega models we used the implementation[2] provided by the authors, which is based on FairSeq (Ott et al., 2019).

3.2 T5

We experiment with multi-task training and fine-tuning the T5 model (Raffel et al., 2020) for translation between English → German, as well as its multilingual counterpart mT5 (Xue et al., 2021) for English → Japanese translation. We compare the results with non-modified versions of T5, Flan-T5, and the multilingual mT5.

We combine and shuffle all training data for the tasks, and experiment fine-tuning the large versions (1B parameters) of the T5 models using a random subset of 10M parallel sentences. We base this choice on observations from preliminary experiments where the small versions of T5 models often converged before reaching 1M examples and base models converged before seeing 10M, since the pre-trained checkpoints are already quite capable as is.

We used the Adafactor optimizer (Shazeer and Stern, 2018) with FP16 training, effective batch sizes of 256 or 512 depending on the model size, evaluation every 1000 steps, and early stopping set to 10 checkpoints of evaluation loss not improving. We set learning rate to 0.0001, weight decay to 0.01, and train each model on a single machine with eight NVIDIA A100 GPUs.

3.3 Gemma

We experimented with adapting 7B and 9B parameter sizes of the 1.1 and 2 version Gemma models (Mesnard et al., 2024) using the in-domain data provided for the General Translation shared task. We used the same random subset of 10M training examples as we did for training T5 models.

4 Results

4.1 General Translation Task

We include the official preliminary automatic ranking results provided by the organizers in Tables 2 and 3. Our primary submissions rank 2nd and 4th among the constrained track (with a white background) for EN-DE and EN-JA respectively. Sadly, they were both not selected for human evaluation by the task organizers due to a large number of submissions and budget constraints this year. References had also not been released as of writing the final submission, therefore, additional metrics or manual assessment of the translations could not be performed.

4.2 Biomedical Translation Task

For the Biomedical Translation task we compared our best models trained for the General Translation task with ones fine-tuned on the biomedical training data, as well as dedicated models trained on the biomedical data from the start. Table 4 shows our preliminary results from developing Mega models for the English↔German tracks of the Biomedical Translation task. We only used different configurations of the MEGA models and compared them with the baseline model submitted to the general translation task. Our best configuration was an ensemble of three separate model checkpoints trained on a mixture of biomedical training data and general data, and fine-tuned on biomedical data.

Table 5 lists the preliminary official results of the Biomedical Translation task provided by the task organizers. According to the BLEU scores, our models seem to be ranked 2nd in both translation directions, overtaken only by the submissions from Unbabel, which are 70B parameter large language models. Similarly to the General Translation task, references for these had also not been released as of writing the final submission, therefore, additional

[2] https://github.com/facebookresearch/mega

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation
IOL-Research	2.3	1.6	0.692	✓
Llama3-70B §	2.5	1.7	0.686	✓
Aya23	2.7	1.8	0.680	✓
IKUN	3.0	1.8	0.668	✓
IKUN-C	3.8	2.0	0.641	✓
CUNI-NL	4.2	2.1	0.624	
AIST-AIRC	7.2	3.3	0.551	
Occiglot	8.2	3.8	0.539	
MSLC	11.9	4.4	0.390	
TSU-HITs	13.3	5.6	0.395	
CycleL2	27.0	11.5	0.091	
CycleL	27.0	11.5	0.091	

Table 2: Preliminary WMT24 General MT automatic ranking for English→German (excluding closed systems).

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑	Human evaluation
Team-J	1.9	2.9	0.740	✓
NTTSU	1.9	2.6	0.731	✓
IOL-Research	2.3	3.1	0.724	✓
Aya23	2.3	3.1	0.719	✓
Llama3-70B §	2.6	3.5	0.714	✓
IKUN	3.1	3.7	0.696	
IKUN-C	3.9	4.3	0.669	✓
AIST-AIRC	6.6	6.5	0.583	
CycleL	24.0	22.4	0.101	

Table 3: Preliminary WMT24 General MT automatic ranking for English→Japanese (excluding closed systems).

metrics or manual assessment of the translations could not be performed.

5 Conclusion

In this paper we described the development process of the AIST AIRC's NMT systems that were submitted for the WMT 2024 shared tasks on general domain text translation and biomedical translation. We compared training MEGA models to fine-tuning T5 and Gemma model architectures in search of the best decoding approach for improving upon output quality. Our results showed that the MEGA model architecture remains highly competitive even in the modern world of large language models, and fine-tuning LLMs with NE-annotated data does not necessarily lead to higher automatic evaluation scores. Especially in the Biomedical Translation task our 100M parameter models demonstrated high competitiveness with the leading 70B parameter models, falling only 0.42 BLEU points behind for EN→DE.

In total, output from four primary systems was submitted to the two shared tasks by AIST AIRC for the English↔German and English→Japanese translation directions.

In future work, we plan to experiment with incorporating document-level training data and modeling longer sequences with appropriate available training data. In terms of data, we intend to increase vocabulary coverage by adding all known unicode emoji symbols to the vocabulary even if they are not present in the training data, as well as additionally sample Paracrawl data where emoji are present.

Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Configuration	EN→DE	DE→EN
Baseline General model	27.23	35.00
General BT model	26.47	33.90
Bio trained/adapted	31.33	40.21
Bio-Baseline ensemble	30.95	39.14
Bio-best-last	31.33	40.14
Bio-ens-15	31.23	40.12
Bio-ens-14	31.21	39.80
Bio-ens-14-15	31.44	40.17
Bio-ens-14-15-2	**31.47**	**40.45**

Table 4: Biomedical task development BLEU score results evaluated on the 2023 Biomedical Translation task test set. The top 3 rows are single model results from the baseline model of the General Translation task, the model after back-translation (BT), and the models specifically trained and adapted on the biomedical (Bio) task data. All remaining rows are combinations of ensembles consisting of best, last, and other checkpoints from the baseline and biomedical specific models.

System Name	EN→DE	DE→EN
ADAPT	30.16	36.93
AIST-AIRC	33.80	45.92
DCUGenNLP	16.46	32.60
HW-TSC	28.77	45.79
Unbabel	34.22	49.05

Table 5: Preliminary WMT24 Biomedical Translation Task BLEU score results.

Ethics Statement

Our work fully complies with the ACL Code of Ethics[3]. We use only publicly available datasets and relatively low compute amounts while conducting our experiments to enable reproducibility. We do not perform any studies on other humans or animals in this research.

References

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 Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2023. Mega: Moving average equipped gated attention. In The Eleventh International Conference on Learning Representations.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

[3] https://www.aclweb.org/portal/content/acl-code-ethics

Mārcis Pinnis, Matīss Rikters, and Rihards Krišlauks. 2018. Tilde's machine translation systems for WMT 2018. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 473–481, Belgium, Brussels. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

Matīss Rikters. 2018. Impact of Corpora Quality on Neural Machine Translation. In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018), Tartu, Estonia.

Matiss Rikters and Makoto Miwa. 2023. AIST AIRC submissions to the WMT23 shared task. In Proceedings of the Eighth Conference on Machine Translation, pages 155–161, Singapore. Association for Computational Linguistics.

Matiss Rikters and Makoto Miwa. 2024. Entity-aware multi-task training helps rare word machine translation. In Proceedings of the 17th International Natural Language Generation Conference, pages 47–54, Tokyo, Japan. 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.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In International Conference on Machine Learning, pages 4596–4604. PMLR.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.