CANTONMT: Cantonese to English NMT Platform with Fine-Tuned Models using Synthetic Back-Translation Data

Kung Yin Hong, Lifeng Han, Riza Batista-Navarro, Goran Nenadic
Department of Computer Science, The University of Manchester
Oxford Rd, Manchester M13 9PL, United Kingdom
kenrick.kung@gmail.com
{lifeng.han, riza.batista, g.nenadic}@manchester.ac.uk

Abstract

Neural Machine Translation (NMT) for low-resource languages is still a challenging task in front of NLP researchers. In this work, we deploy a standard data augmentation methodology by back-translation to a new language translation direction Cantonese-to-English. We present the models we fine-tuned using the limited amount of real data and the synthetic data we generated using back-translation including OpusMT, NLLB, and mBART. We carried out automatic evaluation using a range of different metrics including lexical-based and embedding-based. Furthermore. we create a user-friendly interface for the models we included in this CANTONMT research project and make it available to facilitate Cantonese-to-English MT research. Researchers can add more models into this platform via our open-source CANTONMT toolkit https://github.com/kenrickkung/CantoneseTranslation.

1 Introduction

Cantonese is one of the most popular dialects of Chinese languages, after the standard language Mandarin (the current official language in China, originally from the Beijing area), originally from the capital of Guangdong province, Guangzhou (a.k.a. Canton) in China. The population of Guangdong province is 126.84 million in 2021. In addition, Cantonese is also the native language in Hong Kong and Macau regions which have populations of 7,503,100 and 704,149 in 2023, according to HK Census and Statistics Department https://www.censtatd.gov.hk/en/ and Macrotrends Global Population statistics https://www.macrotrends.net/global-metrics/countries/MAC/macao. Furthermore, because of the economic growth in Guangdong, HK, and Macau, many people from other Chinese provinces also learned to speak Cantonese for job purposes and cultural influences. There is also a big amount of population worldwide outside of China speaking Cantonese. In the era of the fast development of natural language processing (NLP), many machine translation (MT) models have been proposed for the majority of languages worldwide. However, *low-resource language MT* still challenges researchers. Cantonese translation using MT, among one of them, is yet to be developed without much attention being paid so far.

In this work, we aim at investigating one of the popular methods, i.e. synthetic data augmentation via back-translation and model fine-tuning, on Cantonese-to-English neural MT (NMT), a new language pair. We select several models for comparisons including both smaller and larger language models. We compare the system performance using a range of evaluation metrics. Furthermore, we open source our toolkit and create a web-based user-friendly platform called **CantonMT** to facilitate the research on Cantonese-English translation and beyond. [1]

In the next section (Section 2), we survey related work on Cantonese-English MT, data augmentation for MT, and available demos/engines. Section 3 introduces our methodology and framework. Section 4 explains the web-based CANTONMT platform. Section 5 concludes this work with discussions.

2 Related Work

Research work focusing on Cantonese-English MT has not gained much attention to date. Earliest works include (Wu et al., 2006) where Example-based and Rule-based MT were investigated. In recent years, there was the project plan on Cantonese-English Translation at HKU by (Wing, 2020) where the authors listed models to be investigated including RBMT, EBMT, SMT, GRU, Transformer mod-

[1] a video demo of CANTONMT is available at https://youtu.be/s8P5fJjS7Ls

els. In a looser connection, regardless of English as the target language, there has been some work on Cantonese-related MT. These include dialectal translation between Cantonese and Mandarin Chinese by Zhang (1998), Yi Mak and Lee (2022), and Liu (2022).

Data augmentation via Backtranslation has been one of the standard practices to generate a synthetic corpus for assisting the MT performances of low-resource language pairs. This has been popular for both SMT and NMT (Sugiyama and Yoshinaga, 2019; Graça et al., 2019; Edunov et al., 2020; Nguyen et al., 2021; Pham et al., 2023). However, to the best of our knowledge, none of these works focused on Cantonese-to-English translation.

Existing platforms or off-the-shelf demos for Cantonese-to-English MT are very Scarce. Popular MT engines from commercial IT companies do not include this language pair including Google Translator https://translate.google.com and Bing Translator https://www.bing.com/translator. Both of them only included simplified and traditional characters of Mandarin Chinese. The available translators from commercial IT companies include Baidu Translator (Fanyi) https://fanyi.baidu.com/, an IT company from China that includes several Chinese dialectal languages [2]. In the opposite direction, there are open-source tools on English-to-Cantonese MT from TransCan https://github.com/ayaka14732/TransCan.

3 Experimental Work

We introduce the methodology of CantonMT, experimental evaluations using 38K words.hk real bilingual corpus, and extended work when we acquired more real bilingual dictionary data from Wenlin.com.

3.1 Methodology and Framework

The methodology of this work is presented in Figure 1, which includes the following steps:

- DataPrep: data collection and pre-processing
- ModelFineTunePhase1: model selection for initial translator fine-tuning (ft, v1)
- SynDataGenerate: synthetic data generation using the initial translator and cleaned data
- ModelFineTunePhase2: second step MT fine-tuning using real and synthetic data (ft-syn, v2)
- ModelEval: model evaluation using both embedding-based metrics (BERTscore and COMET) and lexical metrics (SacreBLEU and hLEPOR)

For data collection, we scraped the data from the public Hong Kong forum LIHKG https://lihkg.com, which was launched in 2016 and has multiple categories including sports, entertainment, hot topic, gossip, current affairs, etc. We extracted more than 1 million sentences from this website, however, the raw data is with a lot of noises that need to be cleaned, with an example shown in Figure 6 (Appendix Section A). We did data cleaning to reduce noisy strings as well as data *anonymisation* by removing the user IDs from the text. We also filtered out the sentences that were too short with less than 10 Chinese characters. In the end, we prepared 200K clean monolingual Cantonese sentences for parallel synthetic data generation purposes. We shuffled the data for model training.

In the model fine-tuning phase 1, we aim to train a set of reasonable Cantonese-English MT models for synthetic data generation and model comparisons. The baseline models we selected are Opus-MT, NLLB, and mBART with the following rationale: 1) How much does the model size impact the fine-tuning performances? For this, we use Opus-MT which is a much smaller model trained on Opus corpus using MarianMT framework and NLLB-200, a very large language model pre-trained on 200+ languages from Mata-AI. 2) How much does it matter if the pre-trained translation models have Cantonese in their pre-training? For this, we add mBART (mbart-large-50-many-to-many-mmt) which is another LLM but without Cantonese in his pre-training, vs NLLB which has Cantonese. Because the full-size NLLB is too large, we used the distilled model "nllb-200-distilled-600M".

We fine-tuned these models using the available bilingual data from a bilingual Cantonese-English dictionary "Yue-Dian" from https://words.hk, which is in total 44K in size. We divided this data into (training, development, and testing) sets with the sizes (38K, 3K, and 3K), in light that the standard WMT –workshop of machine translation – shared task uses around 3K sentences for testing sets.

[2] All these websites were last visited 2024 March 4th

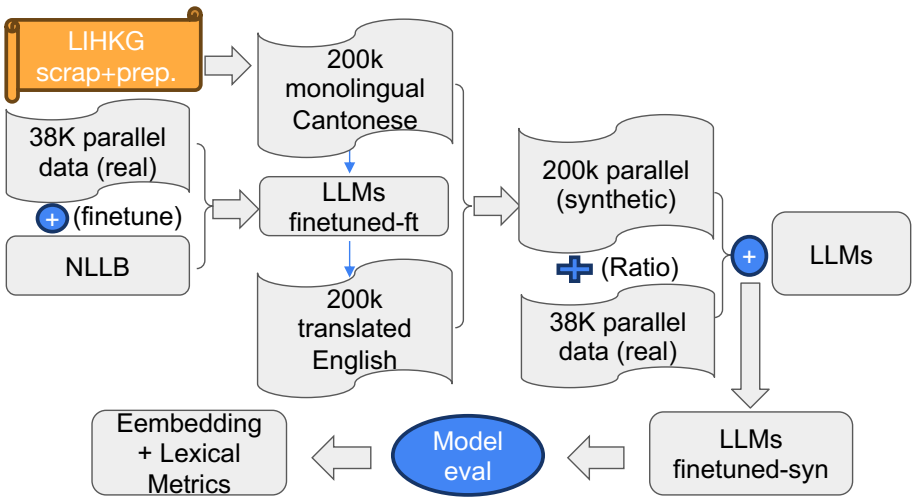

Figure 1: CANTONMT Pipeline: data collection and preprocessing, synthetic data generation, model fine-tuning, model evaluation

In Step 3 for synthetic data generation, we used Step 2 initially fine-tuned LLMs (LLM-ft-v1) to translate the collected monolingual Cantonese text we extracted from Step 1. In this way, we get 200K back-translated English sentences and these synthetic sentences together with the Cantonese sentences create the 200K synthetic parallel corpus we generated. From now on, we call the synthetic parallel corpus 200K-ParaSyn.

In Step 4, we apply different ratios on the real parallel data we have and the 200K-ParaSyn for LLM fine-tuning. We also test the influence of model switches, i.e. using different types of LLMs for LLM-ft (Phase 1) and LLM-syn (Phase 2).

In the last step, we deploy the fine-tuned LLMs in Phase 2 (LLM-syn) on the same test data and compare the results with LLM-ft (Phase 1) and baseline models without fine-tuning. We also report comparisons with available translation engines from IT companies, such as Baidu Translator, Bing Translator, and GPT4. For GPT4, there is an available fine-tuned version towards translation to Cantonese by "community builder" [3] called Cantonese Companion.

We used a range of different evaluation metrics including lexical-based SacreBLEU (Post, 2018), hLEPOR (Han et al., 2013a, 2021), and embedding-based BERTscore (Zhang* et al., 2020) and COMET (Rei et al., 2020). hLEPOR has reported much higher correlation scores to the human evaluation than BLEU and other lexical-based metrics on the WMT shared task data (Han et al., 2013b). However, recent WMT metrics task findings have demonstrated the advantages of neural metrics based on embedding space similarities (Freitag et al., 2022).

3.2 Evaluations of CANTONMT

The Learning curves of three base models during training using the 38K real data are shown in Figures 2 from left to right: mBART, NLLB-200, and Opus-MT. We used three epochs for mBART because it was too large for our available computational resources. From the learning curves, we can see that NLLB-200 has a peak score at Epoch 3 then there is a dramatic drop till Epoch 6 and grows back at Epoch 10. Instead, the Opus-MT model has a steady increasing of the SacreBLEU score with more epochs though there are litter drops in between.

The automatic evaluation scores from CANTONMT models and other commercial engines are listed in Figure 3. There are some interesting findings from the evaluation outcomes.

- LLM-ft vs -LLM-ft-syn: 1) NLLB-syn-1:1 has slightly better scores than NLLB-bl on all metrics, but more ratio of synthetic data will decrease the scores such as 1to3 and 1to5 with around 1 absolute SacreBLEU point. 2) Similarly, mBART-syn-1:1 also outperforms mBART-ft but more ratios of synthetic data will reduce the evaluation scores such as 1to3. 3) Surprisingly, the synthetic model for Opus-mt does not win the Opus-ft-bl, which indicates the quality of the generated synthetic data matters.

[3] https://chat.openai.com/share/7ee588af-dc48-4406-95f4-0471e1fb70a8

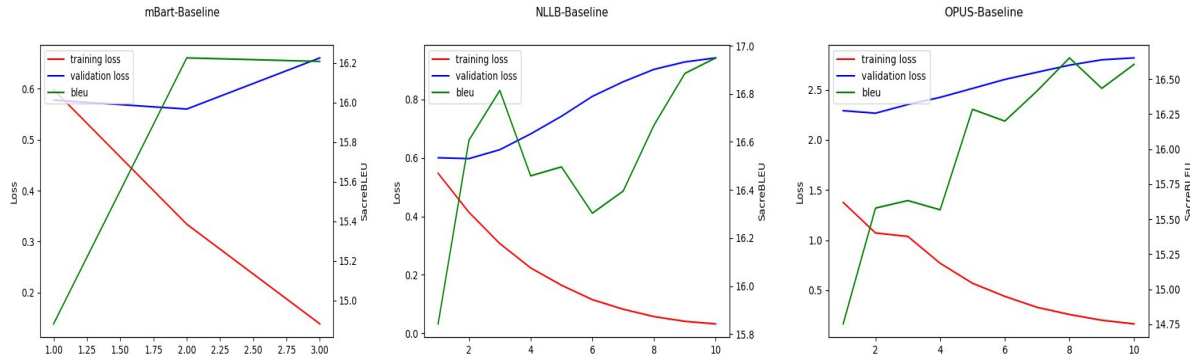

Figure 2: Learning Curve of Model Training using Real Data

- Model Switch Matters: 1) the NLLB fine-tuned model using synthetic data from mBART (top-secon slot) produced higher scores than using the synthetic data generated from its own (top-first slot). 2) mBART fine-tuned using NLLB-generated synthetic data also outperforms mBART fine-tuning using only bilingual real data. 3) In a similar situation, Opus-MT perform differently in comparison to the other two models.

- MT from IT companies: 1) GPT4-finetuned produced the highest evaluation scores but the free version of GPTs restricts the input number of strings; furthermore, it is much more unclear how GPT4 performed the MT; in addition, there are risks on data privacy when users choose to use engines from commercial companies. In contract, CANTONMT is open-sourcely free and researchers can continue to fine-tune it with their data, or insert more models, fully confidential for users. 2) Bing and Baidu translators produced similar evaluation scores to the best system from CANTONMT, though Bing produced slightly higher scores than Baidu, especially on lexical-based metrics SacreBLEU and hLEPOR.

- Comparing to Model Deployment without Finetuning: in Cluster 3 (bottom) of the score table, model deployment without fine-tuning has much lower scores; these scores show that the fine-tuning and synthetic data augmentation make a large ratio of score increase of around 50% for all models using SecreBLEU.

3.3 Adding more real data

In the extension of our work, we managed to fine-tune the baseline models using more real data from another source called Wenlin https://wenlin.com/ where we obtained another 14.5K parallel Cantonese English dictionary corpus. We are curious about the model performances using more real data in addition to the 38K training corpus from words.hk. We listed the comparison scores in the second cluster of Figure 1 where it shows that the newly fine-tuned NLLB-200 using 52.5K data (38+14.5K=52.5K) produced higher scores on all metrics in comparison to 38K trained model; mBART fine-tuned using 52.5K gets better scores on three metrics except for SacreBLEU; Opus-MT surprisingly did not get any increase across metrics. Nevertheless, these outcomes demonstrated the possibility of improving model performances with more available real data, at least for NLLB and mBART models. Furthermore, **data quality matters**: simply adding 14.5K real data NLLB fine-tuned produced higher scores (underlined scores) than the best synthetic system that used 38x2=76K data.

4 CANTONMT Platform

To further facilitate Cantonese-English MT research and for users to easily access freely available fine-tuned models, we developed a user-friendly interface for the CantonMT platform. Users can choose different models and translation directions (Cantonese⇔English) via the interface (Figure 4). The CantonMT web application contains two main parts, Interface and Server.

4.1 User Interface

To test out the User Interface and different models for translation, users can just choose different model types and source languages, which dynamically capture the available model from the server, and allow users to select different training methods

model names	metrics			
	SacreBLEU	hLEPOR	BERTscore	COMET
nllb-forward-bl	16.51166348	0.5651	0.9247503553	0.7376299099
nllb-forward-syn-h:h	15.77513853	0.5616	0.9234894325	0.7341771399
nllb-forward-syn-1:1	**16.59010963**	0.5686	**0.9249665601**	**0.7409206822**
nllb-forward-syn-1:1-10E	16.52034388	**0.5689**	0.9247346256	0.7379801409
nllb-forward-syn-1:3	15.91747159	0.5626	0.9240035192	0.7375647476
nllb-forward-syn-1:5	15.80736185	0.562	0.9237049501	0.7386101493
nllb-forward-syn-1:1-mbart	**16.80769992**	**0.571**	**0.9255628745**	**0.7424624978**
nllb-forward-syn-1:3-mbart	15.86211866	0.5617	0.9245674323	0.7383955538
nllb-forward-syn-1:1-opus	16.55372052	0.5704	0.9253989744	0.7416294167
nllb-forward-syn-1:3-opus	15.9347908	0.5651	0.9242393556	0.7373770117
mbart-forward-bl	15.75132087	0.5623	0.9227211874	0.731434104
mbart-forward-syn-1:1-nllb	**16.03575636**	**0.5681**	**0.924124781**	**0.7379750162**
mbart-forward-syn-1:3-nllb	15.32599502	0.5584	0.9224773548	0.7319351509
opus-forward-bl-10E	**15.06020511**	**0.5581**	**0.9218569376**	**0.7193340472**
opus-forward-syn-1:1-10E-nllb	13.06228317	0.5409	0.9164279981	0.6896684725
opus-forward-syn-1:3-10E-nllb	13.36655432	0.5442	0.9167215968	0.6957020973
baidu	16.56685314	0.5654	0.9242970145	0.7400669999
bing	17.10978564	0.5735	0.9258104091	0.7473919678
gpt4-ft(CantoneseCompanion)	**19.16223506**	**0.5917**	**0.9359668875**	**0.8050356872**
nllb-forward-bl-plus-wenlin14.5k	*16.66623544*	*0.5828*	*0.9260226811*	*0.7495894532*
mbart-forward-bl-plus-wenlin14.5k	15.24035868	*0.5734*	*0.9238412236*	*0.7411118948*
opus-forward-bl-plus-wenlin14.5k	13.01720661	0.5473	0.9157281907	0.6881661525
nllb-200-deploy-no-finetune	11.18274947	0.4925	0.9128870283	0.6863499108
opus-deploy-no-finetune	10.4034635	0.4773	0.9081926861	0.6583570215
mbart-deploy-no-finetune	8.315683152	0.4387	0.9004590699	0.6272943796

Figure 3: Automatic Evaluation Scores from Different Models in CANTONMT. bl: bilingual real data; syn: synthetic data; h:h - half and half; 1:1/3/5 - 100% real + 100/300/500% synthetic; 10E: 10 epochs (default: 3); top-down second slot: model switch: model type using NLLB but synthetic data from other models (mBART and OpusMT); top-down third slot: including model switch for mBART fine-tuning using synthetic data generated from NLLB; similarly top-down forth slot: including model switch for OpusMT fine-tuning using synthetic data from NLLB. Bottom slot of Cluster 1: Bing/Baidu Translator and GPT4-finetuned Cantonese Companion; **bold** case is the best score of the same slot among the same model categories. Cluster 2: bilingual fine-tuned models using 38K words.hk data plus 14.5k Wenlin data; *italic* indicates the number outperforms the same model fine-tuned with less data 38K.

for the model. One can then type the sentence in the input box and click the translate button for the translation output from the model. The application layout is quite modular in case different model types or languages are added to the system, which could potentially be used as a base framework for different translation systems and simply add more languages to the input and output if one wishes to expand the implementations. This web application has taken a template (Wrigley, 2023) for an AI Code translator and modified it to fit the need, which is developed in TypeScript with the Next.js framework. The reason for choosing this framework is that it provides a very modern and minimalistic approach to web development.

4.2 Server

A Diagram outlining the modules can be seen in Figure 5 to understand the general structure of the server. Users can easily run the server on their local machines and it is well-documented in the Readme file in the open-source platform. The server has two main functionalities, where the first one will output the list of model paths given the model type and source languages. The second one provides the translation, where one could provide the details of the model and also the sentence in the language specified, and the server would respond with the translated sentence using the model output.

During the implementation, due to a lack of memory space, the server crashed multiple times

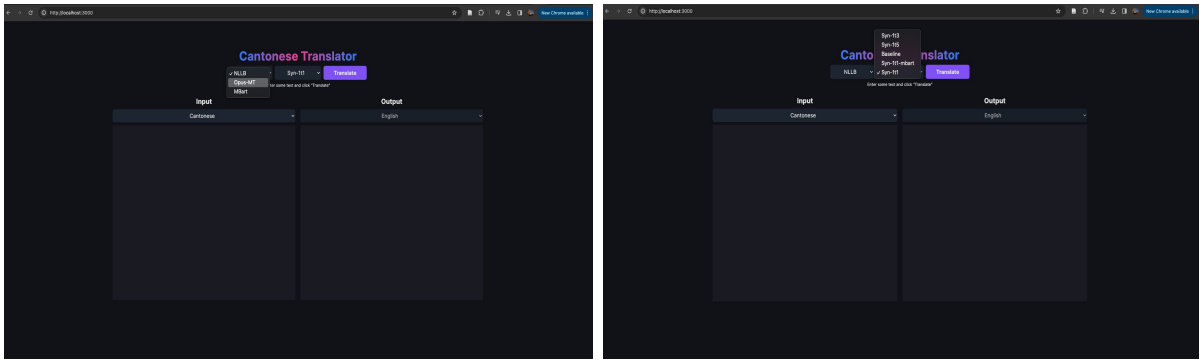

Figure 4: CANTONMT Platform with options of model types, training categories, and translating directions. Frontend: TypeScript with Next.js. Backend: Python - Flask

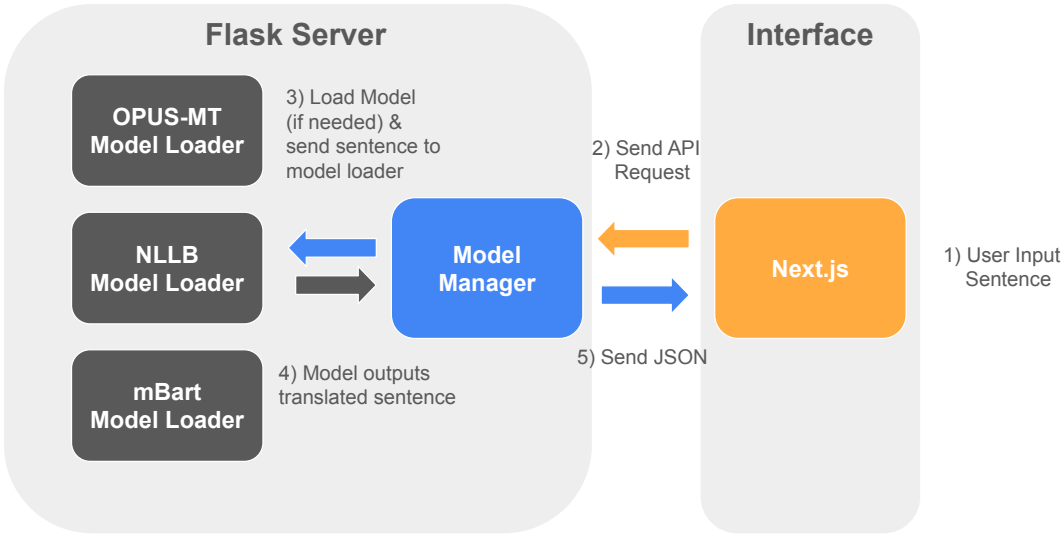

Figure 5: CANTONMT Server and Interface Flowchart diagram.

on our local machine. To account for the potential crash, a *model manager* was produced, which implements an LRU cache for the different model loaders, where the least recently used model will be deleted from memory if it exceeds the limit of the number of models. The server is built entirely on the *Python Flask* library. The reason for choosing this framework is that the models can be run on the Python Transformers library, which provides a seamless implementation without much additional effort.

5 Discussion and Conclusion

In this work, we investigated the back-translation methodology for bilingual synthetic data generation for the sake of data augmentation for NMT, on a new language translation direction Cantonese to English. We tested both smaller-sized OpusMT and extra-large LLMs NLLB and mBART both using available bilingual real data and the larger synthetic data. The experiments show that all the fine-tuned models outperformed the baseline deployment models with large margins. Furthermore, the synthetic model nllb-syn-1:1-mbart produced higher scores using model switch method compared to without model switch. Last but not the least, the best performing fine-tuned models have similar (or even higher) evaluation scores than the current translators from IT companies of Baidu and Microsoft-Bing, although the fine-tuned GPT4 has the highest ones.

In the concern of **data privacy** such as sensitive data (e.g. clinical domain (Han et al., 2024)), CANTONMT can be fully controlled by users without interference from any third parties. We open-source our platform so that researchers can continue to integrate new models into the toolkit to promote Cantonese-English MT. We also plan to carry out human evaluations on the outputs from different systems to get more insights into the system errors.

Acknowledgements

We thank words.hk and wenlin.com for the data usage. LH and GN thank the support from Grant "Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease" (The project has been funded by the Nuffield Foundation, but the views expressed are those of the authors and not necessarily the Foundation. Visit www.nuffieldfoundation.org) and Grant EP/V047949/1 "Integrating hospital outpatient letters into the healthcare data space" (funder: UKRI/EPSRC).

References

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation.

Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013a. Language-independent model for machine translation evaluation with reinforced factors. In *Proceedings of Machine Translation Summit XIV: Posters*, Nice, France.

Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013b. A description of tunable machine translation evaluation systems in WMT13 metrics task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 414–421, Sofia, Bulgaria. Association for Computational Linguistics.

Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: An empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, 6:1211564.

Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model LaBSE. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1014–1023, Online. Association for Computational Linguistics.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq Joty, Thanh-Tung Nguyen, Kui Wu, and Ai Ti Aw. 2021. Cross-model back-translated distillation for unsupervised machine translation. In *International Conference on Machine Learning*, pages 8073–8083. PMLR.

Nghia Luan Pham, Van Vinh Nguyen, and Thang Viet Pham. 2023. A data augmentation method for english-vietnamese neural machine translation. *IEEE Access*, 11:28034–28044.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Liu Hey Wing. 2020. Machine translation models for cantonese-english translation project plan.

Mckay Wrigley. 2023. ai-code-translator.

Yan Wu, Xiukun Li, and Caesar Lun. 2006. A structural-based approach to Cantonese-English machine translation. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 11, Number 2, June 2006*, pages 137–158.

Hei Yi Mak and Tan Lee. 2022. Low-resource nmt: A case study on the written and spoken languages in hong kong. In *Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '21, page 81–87, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaoheng Zhang. 1998. Dialect MT: A case study between Cantonese and Mandarin. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Figure 6: Example text extracted from LIHKG website with lots noise before cleaning and anonymisation

	Opus	NLLB	mBart
Architecture	Transformer	Transformer	Transformer
Layers	12	24	24
Hidden Unit	512	1024	1024
Model Parameters	77,943,296	615,073,792	610,879,488
Release Year	2020	2022	2020

Figure 7: Parameters from deployed models.

A Appendix

Example raw text extracted from LIHKG website can be seen in Figure 6 before cleaning. The model parameters from OpusMT, extra-large NLLB and mBART are shown in Figure 7, which shows that NLLB and mBART have doubled the number of Transformer layers and almost 10 times more parameters than OpusMT.