CantonMT: Cantonese-English Neural Machine Translation Looking into Evaluations

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
*corresponding author

Abstract

Cantonese-English is a low-resource language pair for machine translation (MT) studies, despite the vast amount of English content publicly available online and the large amount of native Cantonese speakers. Based on our previous work on CantonMT from Hong et al. (2024), where we created the open-source fine-tuned systems for Cantonese-English Neural MT (NMT) using base-models NLLB, OpusMT, and mBART and corpus collections and creation, in this paper, we report our extended experiments on model training and comparisons. In particular, we incorporated human-based evaluations using native Cantonese speakers who are also fluent in the English language. We designed a modified version of the HOPE metric from Gladkoff and Han (2022) for the categorised error analysis and serenity-level statistics (naming **HOPES**). The models selected for human evaluations are NLLB-mBART fine-tuned and two translators from commercial companies: Bing and GPT4. Further analysis of fine-tuned systems and human-evaluation insights can shed some light on Cantonese-English NMT and its future development. The open-source CantonMT toolkit and analytics will be accessible via the GitHub page (at https://github.com/kenrickkung/CantoneseTranslation).

1 Introduction

Cantonese is a Sinitic language spoken in Hong Kong, Macau, and the Guangdong region of southern PRC, it is the second most spoken Sinitic language, after Mandarin Chinese (Wiedenhof, 2015). With a substantial 80 million native speakers (Eberhard et al., 2023), Cantonese is still an under-researched area in the spectrum of Natural Language Processing, as demonstrated in ACL Anthology, where only 47 papers are related to Cantonese, compared with 2355 for (Mandarin) Chinese (Xiang et al., 2022).

Despite having the second most speakers in the family of Sinitic languages, most State-of-the-art commercial translators either do not support Cantonese or have below-par translation quality when translated to English. This leads to scenarios where individuals seeking Cantonese resources face challenges, particularly in casual forums where tones are often very similar to spoken language.

We believe that Cantonese is a unique language that captures the rich cultural history of Hong Kong, Macau, and the Guangdong province of China. Two major challenges when dealing with Cantonese translations are Colloquialism and Multilingualism. Colloquialism, a linguistic style used for informal and casual conversation, often occurs in Cantonese and includes non-standard spelling, slang, and neologisms. As for Multilingualism, Hong Kong was once a British colony and has a rich Chinese cultural influence; code-switching [1] happens often in day-to-day conversation; and words can also be loaned from English through phonetic transliteration (Bauer, 2006).

[1] the act of using multiple languages together

Therefore, following the trend of language diversity and inclusion in NLP, we have set out the aim to develop a translation system that can translate texts from Cantonese to English and reach comparable results against commercial translators, as reported in our CANTONMT1.0 (Hong et al., 2024).

As an extended investigation of our first milestone, regarding the Evaluation Strategy, the models developed are evaluated through a range of metrics, including lexicon-based word surface matching (SacreBLEU and hLEPOR) and those based on embedding spaces (COMET and BERTscore). Following these metrics, the top-performing model is chosen for comparison with the two top-performing commercial translation tools. We designed the HOPES (standing for "Simplified HOPE") human evaluation framework, which we modified based on HOPE, a human-centric post-editing based metric by Gladkoff and Han (2022).

2 Background and Related Works

2.1 Large Language Models

With the rise of LLMs, there are dozens of pre-trained models which are capable on MT tasks with none or few fine-tuning. In our investigation, there are 3 models chosen for further fine-tuning with our dataset, the reason behind choosing these models can be found at CANTONMT1.0. Here is a brief introduction of each model, which could help readers understand the difference with depth.

2.1.1 Opus-MT

Opus-MT (Tiedemann and Thottingal, 2020), developed by Helsinki-NLP, is a Transformer-based NMT, which is using Marian-NMT [2] as the framework for the model training. The model family is trained with a publicly available parallel corpus collected in OPUS[3]. The model is specifically trained for MT task, and should not be classified as a general purpose LLM. Two specific models are used in this project, *Opus-mt-zh-en* and *Opus-mt-en-zh*, which are models that translate Chinese to English and English to Chinese. The forward model (Chinese to English) has around 77M parameters, which is considered quite a small model when compared to LLMs.

2.1.2 mBART

mBART (Liu et al., 2020), a multilingual Seq2Seq denoising auto-encoder. It is trained with the BART (Lewis et al., 2020) objectives with a multilingual corpus. The pre-training of mBART is trained by corrupting text with a noising function and also learning a model to reconstruct the original text. It uses the CC25 Corpus which contains 25 languages and follows the standard Transformer architecture with 12 layers of encoders and 12 layers of decoders.

In CANTONMT, a specific version of the model is used (*mbart-large-50-many-to-many-mmt*) which supports 50 languages, including (Mandarin) Chinese. However, it does not support Cantonese as a language. The model is also fine-tuned for multilingual translation and is introduced by Tang et al. (2020) which has added 25 additional languages without hurting the performance of the model. The model has a total of 610M parameters, a massive increase compared to the previous Opus model.

2.1.3 NLLB

No Language Left Behind (NLLB) (NLLB-Team et al., 2022), to the best of our knowledge, is the only publicly available LLM which contains the language Cantonese (Lang-Code: yue_Hant). It is trained upon the FLORES-200 dataset which contains 200 languages and serves as a high-quality benchmark dataset. The model architecture is also based on the Transformer encoder-decoder architecture (Vaswani et al., 2017).

In CANTONMT, a distilled version of NLLB (*nllb-200-distilled-600M*) is used since based on our available computation power, there is no chance of fine-tuning a larger model. The model is already fine-tuned on MT task, and the language pair in focus is Cantonese-English.

2.2 Back-Translation

Data Augmentation via Back translation is a technique used by MT researchers when tackling low-resource languages. Typically, since not enough data is available, the model may not be able to learn the translation of the language thoroughly and, thus might harm the performance of MT. This technique has been one of the standards for leveraging monolingual corpora since SMT (Bojar and Tamchyna,

[2] https://marian-nmt.github.io/

[3] http://opus.nlpl.eu/

2011), and is still being used with NMT (Sennrich et al., 2016).

The approach uses a model, which translates target language text to the source language (back model), for translating a monolingual corpus in the target language to the source language. This creates a synthetic parallel corpus (Silver Standard), which is different from human annotated parallel corpus (Gold Standard). In theory, with more data, the model can be performing better.

2.3 MT on Cantonese

2.3.1 Commercial Translators

A survey has been conducted on four different commercial MT software, including Google, Bing, Baidu, and DeepL.

For Google[4] and DeepL[5], despite being the most popular software used for translation in daily lives, they do not support Cantonese as an option, but only (Mandarin) Chinese. Therefore, no further investigations are being made on the platforms. For Bing[6] and Baidu[7], there are native Cantonese support in translation and therefore are chosen as a state-of-the-art comparison in the following sections.

With the rise of LLMs, there are also questions on whether or not this kind of model with prompting can give better results when compared with a more traditional approach with fine-tuning on LLMs. In this project, Generative Pre-trained Transformers(GPT)-4 (OpenAI, 2024) are being investigated with specific prompting to compare against our model. The implementation of GPT-4 that we used is Cantonese Companion, which was custom-made for translation to Cantonese by a community builder.[8] However, it should be noted that we do not know how much data was used for this community-trained Cantonese Companion and the training was not transparent, in addition to its dependence on the commercial platform.

2.3.2 Research Models and Toolkits

Research work focusing on Cantonese-English MT has not gained much attention up to date unfortunately. Some typical literature work we found includes example-based MT by Wu et al. (2006); RNN-based model by Wing (2020); BiLSTM model by Liu (2022); Transformer-models by Yi Mak and Lee (2022). In addition, TransCan[9] is a NMT model which translates English to Cantonese and is trained based on bart-base-Chinese and BART with additional linear projection to connect them.

3 Review CantonMT Methodology

3.1 Datasets and Preprocessing

Since Cantonese-English parallel corpora are not readily available, combinations of different datasets are used for the initial training of baseline models. Furthermore, to aid the back-translation strategy in the latter part of the project, monolingual corpora for both Cantonese and English are required, and therefore, they will be discussed in the following section.

3.1.1 Parallel Corpus

To fine-tune different baseline models, a parallel corpus is required to train the model to translate Cantonese to English at a reasonable level. In the end, three different parallel corpora are found between different timestamps of the investigation. Therefore, the latter two are used for training only, while the former are used for training and evaluation.

Words.hk Corpus Words.hk[10] is an open Cantonese-English dictionary publicly available for people to download. We used the full dataset from their website, which contains different Cantonese words and some example sentences with their English translation. An example of the word "投資/touzi" in the dictionary is given in Figure 1.

From the data, only the sentence after the tag *eng* has been used in this case, the sentence, "*She invested $1 million in renovating the shop*", has been extracted and also its corresponding Cantonese translation which is the sentence after the tag *yue*. Data pre-processing has also been done, including removing hashtags and space since there is quite a lot in the dataset, potentially affecting data quality. In ad-

[4] https://translate.google.com/

[5] https://www.deepl.com/translator

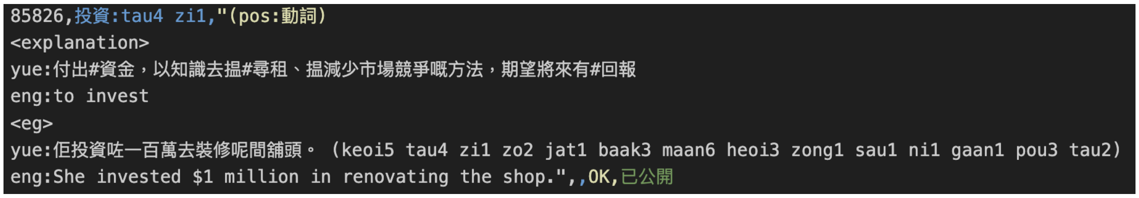
[6] https://www.bing.com/translator

[7] https://fanyi.baidu.com

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

[9] https://github.com/ayaka14732/TransCan

[10] https://words.hk



```
85826,投資:tau4 zi1,"(pos:動詞)
<explanation>
yue:付出#資金，以知識去搵#尋租、搵減少市場競爭嘅方法，期望將來有#回報
eng:to invest
<eg>
yue:佢投資咗一百萬去裝修呢間舖頭。 (keoi5 tau4 zi1 zo2 jat1 baak3 maan6 heoi3 zong1 sau1 ni1 gaan1 pou3 tau2)
eng:She invested $1 million in renovating the shop.",,OK,已公開
```

Figure 1: Sample Data Format for Words.hk

dition, there are sentences with multiple translations; in that case, the first translation has been taken. In the end, 44K sentences have been extracted from the dataset. A graph of the frequencies of the length of the Cantonese sentence has been plotted in Figure 2. It is noticed that despite the effort only to keep sentences and no definitions, there are still quite a lot of short sentences in the dataset. Since for short sentences, it could be straightforward for the model to translate and, therefore, may lead to a bias in the evaluation, we have decided to split the dataset into short sentences and long sentences, where short sentences are sentences that have ten characters or less. In the end, there are 19.4K short sentences and 24.6K long sentences. Since data are already very scarce, we have decided not to opt into the standard train-dev-test split of 8/1/1 or 7/2/1 and instead went for the approach of a 3K dev set and 3K test set. The reason behind this is based on that the standard practice for Workshop of Machine Translation (WMT)[11] shared task uses around 3K sentences for test sets when comparing different MT systems.

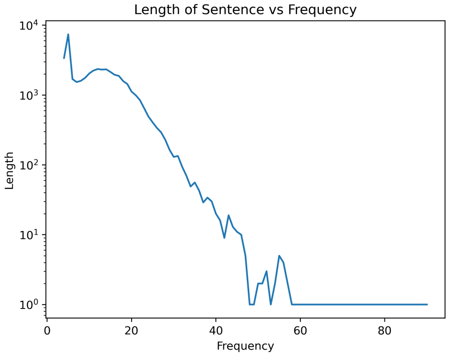


Figure 2: Words.hk - Sentence Length

Wenlin Corpus Wenlin Institute [12] creates software and dictionaries for learning the Chinese language, and there is a dictionary, ABC Cantonese-English Comprehensive Dictionary, which is readily available for registered users to use for research purposes. The process to obtain the dataset, however, is not straightforward. It involves first getting a list of URLs which store the data, and after that, it requires web scraping; at the end, an XML file is obtained, which includes all the sentences and other content.

Extracting is required to convert an XML file to a parallel corpus after obtaining an XML file. Based on initial inspection, the sentence should be inside the tag *WL*; therefore, regular expression techniques are used to extract those sentences. After that, similar pre-processing as Words.hk has been done to obtain the training set and 14.5K parallel sentences are extracted.

Opus Corpora (Tiedemann and Nygaard, 2004) is a collection of translated documents collected from the internet. The corpus is already aligned, and therefore, no pre-processing is required. It can be easily downloaded via their website [13]. An additional 9.6K parallel sentences are added to the final training set.

3.1.2 Monolingual Corpus

To aid the process of back-translation, a monolingual corpus from both the source and target language is required to investigate the *iterative back-translation* approach.

English Corpus - There are many English monolingual corpora available, and in this project, the dataset we have decided to use is from the WMT 2012 News Collection (Callison-Burch et al., 2012). It can be downloaded on the WMT website and contains 434K sentences, which is more than required for the back-translation.

[11] https://www2.statmt.org/wmt24
[12] https://wenlin.com
[13] https://opus.nlpl.eu/

Cantonese Corpus - However, for the Cantonese corpus, it is difficult to find an existing monolingual corpus. There is a Hong Kong Cantonese Corpus (HKCanCor) available (Lee et al., 2022). However, this is based on spontaneous speech and radio programs from the late 1990s and, therefore, might be outdated and there is the language evolution factors with time passing by. Another reason for not choosing the data is that it only consists of 10K sentences, which is insufficient for back-translation purposes.

Based on findings from Liang et al. (2021), there should be abundant data on social media, including Facebook, YouTube, Instagram and different local forums. Since it will be hard to filter out Hong Kong users who use Cantonese in their social media comments, we have decided to turn to local forums. There are few mainstream ones which have an abundance of data, including Baby-Kingdom[14], DiscussHK[15], and LIHKG[16].

In the end, based on tools available online, we have decided to collect data from LIHKG. It is an online forum platform that was launched in 2016 and has multiple categories, including sports, entertainment, hot topics, gossip, current affairs, etc. There is a scraper readily available online from Ho and Or (2020), which we have used to scrape the data from LIHKG. Data is scraped in CSV format, where an example can be seen in Figure 3 (profile ID masked).

Overall, 29K posts have been scraped, and only the text part has been used as the monolingual data. Some more **pre-processing** has been done to the data, including stripping all the links in the data and filtering out all the sentences shorter than 10 Chinese characters. In the end, 1.1M sentences have been scraped, which is more than enough for our investigation. We **shuffled** the dataset so that it can be used by the research community for free, as long as they sign a user agreement form for non-commercial usage.

3.2 Model Trainings

The model fine-tuning methodology of CantonMT is presented in Figure 4, which includes the following steps:

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

Detailed techniques on each step was explained in CantonMT1.0 by Hong et al. (2024). We also report comparisons with commerically available translation engines such as the Baidu Translator, Bing Translator and GPT4. The implementation of GPT-4 that we used is Cantonese Companion, which was custom-made for translation to Cantonese by a community builder.[17]

3.3 Automatic Evaluations

We used a range of different evaluation metrics including the lexical-based SacreBLEU (Post, 2018) and hLEPOR (Han et al., 2013a, 2021), and the 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).

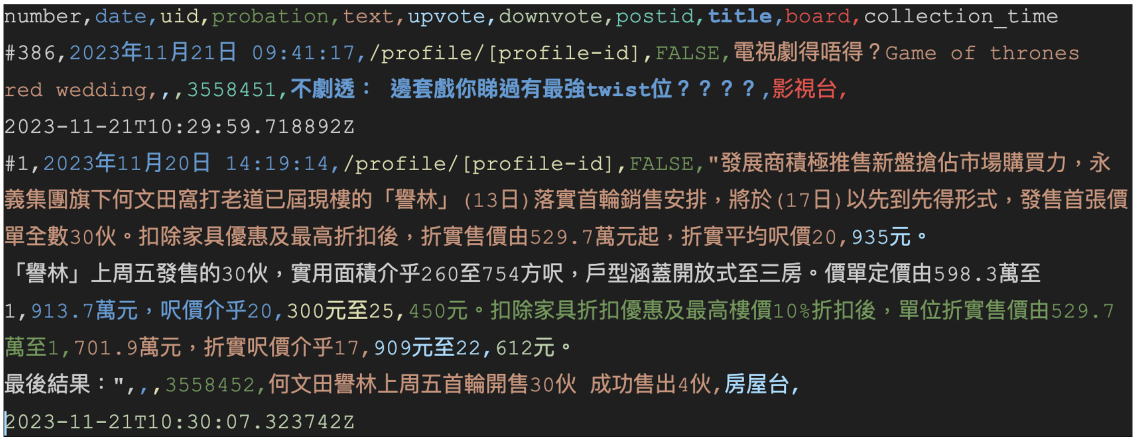
The automatic evaluation scores from CantonMT models and other commercial engines are listed in Table 1. From the automatic evaluation metrics, the results demonstrated that the model-switch fine-tuned NLLB-mBART using 1:1 ratio of synthetic and real data achieved relatively higher scores than other fine-tuning models. Thus, we selected this model into the human evaluation loop, together with Bing and GPT4-ft.

[14] https://www.baby-kingdom.com/forum.php

[15] https://www.discuss.com.hk/

[16] https://lihkg.com

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



```
number,date,uid,probation,text,upvote,downvote,postid,title,board,collection_time
#386,2023年11月21日 09:41:17,/profile/[profile-id],FALSE,電視劇得唔得？Game of thrones
red wedding,,,3558451,不劇透： 邊套戲你睇過有最強twist位？？？？,影視台,
2023-11-21T10:29:59.718892Z
#1,2023年11月20日 14:19:14,/profile/[profile-id],FALSE,"發展商積極推售新盤搶佔市場購買力，永
義集團旗下何文田窩打老道已屆現樓的「譽林」(13日)落實首輪銷售安排，將於(17日)以先到先得形式，發售首張價
單全數30伙。扣除家具優惠及最高折扣後，折實售價由529.7萬元起，折實平均呎價20,935元。
「譽林」上周五發售的30伙，實用面積介乎260至754方呎，戶型涵蓋開放式至三房。價單定價由598.3萬至
1,913.7萬元，呎價介乎20,300元至25,450元。扣除家具折扣優惠及最高樓價10%折扣後，單位折實售價由529.7
萬至1,701.9萬元，折實呎價介乎17,909元至22,612元。
最後結果：",,,3558452,何文田譽林上周五首輪開售30伙 成功售出4伙,房屋台,
2023-11-21T10:30:07.323742Z
```

Figure 3: LIHKG Data Example

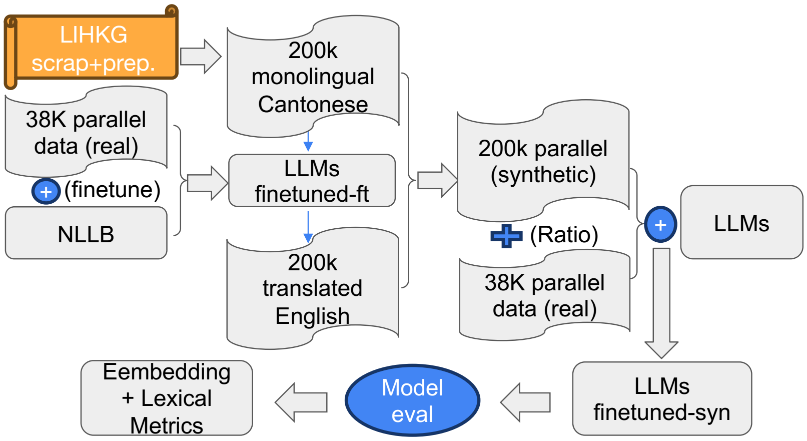


Figure 4: CANTONMT Pipeline: data collection and preprocessing, synthetic data generation, model fine-tuning, model evaluation Hong et al. (2024)

4 Human Evaluations

Even with four different automatic metrics, it is still hard to judge the model's performance based on those chosen metrics. Therefore, human evaluations are conducted to understand better the comparison with state-of-the-art models and the *different types* of errors that the trained models or deployed translators tend to make.

4.1 HOPES framework

With that in mind, we have borrowed the HOPE framework (Gladkoff and Han, 2022). The original HOPE framework includes eight detailed error types from industrial practice, already much simpler than MQM (Lommel et al., 2024). However, upon our review, some error types can be merged to make the human evaluation task more efficient and better match our data, where a modified framework, HOPE-Simplified (HOPES), is proposed. The merging procedure is shown in the below list.

1. **Merge Impact(IMP) and Mistranslation(MIS) as MIS**:
   The definitions of IMP and MIS are "The translation fails to convert main thoughts clearly" and "Translation distorts the meaning of the source and presents mistranslation or accuracy error" respectively. They overlap in accuracy and meaning preservation from the source sen-

Model Name	SacreBLEU	hLEPOR	BERTscore	COMET
nllb-forward-bl	16.5117	0.5651	0.9248	0.7376
nllb-forward-syn-h:h	15.7751	0.5616	0.9235	0.7342
nllb-forward-syn-1:1	**16.5901**	0.5686	**0.925**	**0.7409**
nllb-forward-syn-1:1-10E	16.5203	**0.5689**	0.9247	0.738
nllb-forward-syn-1:3	15.9175	0.5626	0.924	0.7376
nllb-forward-syn-1:5	15.8074	0.562	0.9237	0.7386
nllb-forward-syn-1:1-mbart	**16.8077**	**0.571**	**0.9256**	**0.7425**
nllb-forward-syn-1:3-mbart	15.8621	0.5617	0.9246	0.7384
nllb-forward-syn-1:1-opus	16.5537	0.5704	0.9254	0.7416
nllb-forward-syn-1:3-opus	15.9348	0.5651	0.9242	0.7374
mbart-forward-bl	15.7513	0.5623	0.9227	0.7314
mbart-forward-syn-1:1-nllb	**16.0358**	**0.5681**	**0.9241**	**0.738**
mbart-forward-syn-1:3-nllb	15.326	0.5584	0.9225	0.7319
opus-forward-bl-10E	**15.0602**	**0.5581**	**0.9219**	**0.7193**
opus-forward-syn-1:1-10E-nllb	13.0623	0.5409	0.9164	0.6897
opus-forward-syn-1:3-10E-nllb	13.3666	0.5442	0.9167	0.6957
baidu	16.5669	0.5654	0.9243	0.7401
bing	17.1098	0.5735	0.9258	0.7474
gpt4-ft(CantoneseCompanion)	**19.1622**	**0.5917**	**0.936**	**0.805**
nllb-forward-bl-plus-wenlin14.5k	*16.6662*	*0.5828*	*0.926*	*0.7496*
mbart-forward-bl-plus-wenlin14.5k	15.2404	0.5734	0.9238	0.7411
opus-forward-bl-plus-wenlin14.5k	13.0172	0.5473	0.9157	0.6882
nllb-200-deploy-no-finetune	11.1827	0.4925	0.9129	0.6863
opus-deploy-no-finetune	10.4035	0.4773	0.9082	0.6584
mbart-deploy-no-finetune	8.3157	0.4387	0.9005	0.6273
nllb-forward-all3corpus	*16.9986*	*0.583*	*0.927*	*0.7549*
nllb-forward-all3corpus-10E	16.1749	0.5728	0.9254	0.7508
mbart-forward-all3corpus	16.3204	0.5766	0.9253	0.7482
opus-forward-all3corpus-10E	14.4699	0.5621	0.9191	0.7074

Table 1: 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. Cluster 3: Deployed Model without fine-tuning Cluster 4: Finetuned with the previous 2 corpora and an additional 10K data from OPUS Corpora we managed to find in the end - it shows the evaluation improvement continues Hong et al. (2024)

tence, which both reflect the semantics error. Therefore, it is merged as Mistranslation(MIS), where the new definition is given as "perceived meaning differs from the actual meaning". Furthermore, the original data does not define the scoring mechanism in a specific way. For example, when the translation mistranslates a critical word, should it be given as a critical error since it distorts the meaning, or a minor error since there is only one mistake in the translation? With the newly defined MIS, the first case could be covered by that, and therefore, a minor error should be given.

2. **Merge Terminology(TRM) and Proper Name (PRN) as Terms(TRM)**:
The original definitions of TRM and PRN are "incorrect terminology, inconsistency on the translation of entities" and "a proper name is translated incorrectly" respectively. In our experimental data, the name is not popular, and proper names can be entity types if they appear in the test set. Therefore, the error types are merged as TRM, with the new definition of "Incorrect terminology", including proper names or inconsistency of translation of entities, where a higher score means there are more incorrect terms".

3. **Merge Style (STL), Proofreading (PRF), Required Adaptation Missing (RAM) into Style(STL).**
The original definitions of these three are "translation has poor style but is not necessarily ungrammatical or formally incorrect", "linguistic error which does not affect accuracy or meaning transfer but needs to be fixed", and "source contains error that has to be corrected or target market requires substantial adaptation of the source, which translator failed to make; impact on the end user suffers". These errors are all related to localisation and adaptation. We *summarise* the merged error type Style as "Translation has poor style, but is not necessarily ungrammatically or formally incorrect. It may also include linguistic error which does not affect meaning, but potentially makes the end user suffer".

Based on literature from Gladkoff et al. (2022) regarding evaluation uncertainty, less than 200 human evaluation sentences are insufficient to make a statistical significance. Therefore, 200 sentences from the test set are randomly sampled from the test set and used for human evaluation. Three different translation systems are chosen, including the best model from our training (NLLB-mBART), one of the commercial translators (Bing) and community-finetuned GPT4.

There were a total of 4 annotators who are fluent English speakers and native Cantonese users annotated the translations for the 200 x 3 translations. Each translation is then evaluated by two annotators to measure the agreement level between them, and therefore, the results should be more accurate and reflect the performance of each system. It should also be noted that the results can also help us understand the general error types the models are making, which may be useful for future work.

4.2 Human Evaluation Outcomes

4.2.1 Text Degeneration

Upon first glance at the synthetic data and test set translations, some interesting phenomena are happening, described as *neural text degeneration* (Holtzman et al., 2020). Examples of text degeneration can be seen in Table 2. From the example, "handwritten" has been repeated multiple times, indicating the models generate repetitive and dull loops. This could be another point of future work to adopt some methods for minimising these situations.

4.2.2 Results

The results are then used to calculate inter-annotator agreement (IAA), via a quadratic-weighted Cohen's Kappa metric (Cohen, 1968), where the ratings are grouped into two individual raters. The results are shown in Table 3.

The results show that the annotators have a substantial agreement level in the category of mistranslation (Landis and Koch, 1977) and the overall rating, which is calculated by adding all 4 metrics together. For the other metrics, terminology and grammar have shown a moderate agreement between annotators. However, there seems to be a low agreement level for style, which suggests that the guidelines might need more refinement and detailed explanations, or more likely, translation style is very personal and should not be a major contributing factor to whether or not the translation is good or not.

Since the annotators have shown some kind of agreement, the results shown in Table 4 should have some indication of whether or not the translation is up-to-standard and can provide a better understanding of the models' performance. Another table can be seen in Table 5 for errors in individual models, where a major error is defined as a total score higher than 15 and a minor error is defined as lower than 15

Source Sentence	佢踢住對人字拖嘅行出。
Model Translation	He walked out with a pair of handwritten handwritten handwritten.

Table 2: Example of Text Degeneration

Metric	NLLB	Bing	GPT4
MIS	0.6671	0.6102	0.5700
TERM	0.5700	0.4775	0.3874
STYLE	0.1123	0.3490	0.0348
GRAM	0.4212	0.2899	0.2850
Overall	0.6230	0.6136	0.4935

Table 3: Cohen's Kappa for Different Models and Metrics

but excluding 0. Translations with no errors in all 4 categories are defined as No error.

The results have shown that fine-tuned GPT4 "CantoneseCompanion" is by far the best model for translation, where over half of the translations have shown no errors, and only 3% of translations have major errors according to the metric. Also, for the different metrics, GPT4 has shown similar performance except for **grammar**, which indicates that *error types are quite diverse for GPT4*.

Moreover, Bing performs better than the best model from NLLB, which is in line with the automatic metric. Nevertheless, both models have only around 25% translation, which is error-free. In the evaluation, it can be seen that there are quite a few cases for both models to translate the sentence literally, which leads to some slang not being correctly translated and, therefore, affects the quality of translation.

For our system, most errors stem from either *mistranslation* or *terminology*, which is often correlated since when a term is not correctly translated, it often causes meaning loss in the sentence. It can also be noticed that most of the sentences are often grammatically correct, which should be expected since the decoder part of the Transformers is trained with large amounts of English data and, therefore, should be well-versed in grammar knowledge.

The result here shows that additional effort will be needed to surpass one of the commercial translators, where there should be more effort put into improving the model's knowledge of *terminology and slang*. For example, having a knowledge graph and knowledge base to represent different terminology and slang (Zhao et al., 2020; Han et al., 2020) could potentially allow the model to understand more terminology in Cantonese. Further pre-training in Cantonese can potentially improve performance too.

5 Discussion and Conclusion

In this paper, we further investigated the system performances from CANTONMT, an open-sourced platform for Cantonese⇔English translation. We designed HOPES metric for human evaluation purposes, which is a simplified version of the HOPE framework by Gladkoff and Han (2022). The simplified HOPES metric has only four error types including mistranslation (MIS), term errors (TERM), style (STYLE), and grammatical errors (GRAM), while keeping the original error severity features from HOPE. The human evaluation result shows that NLLP-mBART fine-tuned model has average error score 12.58, vesus 8.3475 and 2.3575 from Bing and GPT4-ft. Regarding error severity levels, NLLB-mBART has fewer minor-errors than Bing, though more major-errors at this stage.

As we mentioned in CantonMT (Hong et al., 2024), in terms of concerns of **data privacy** such as handling of sensitive data (e.g., in clinical applications related to health analytics of patient data (Han et al., 2024)), CANTONMT can be fully controlled by users without interference from any third parties. We believe the performance of CantonMT models can be continuously improved with more high-quality real and synthetic data integrated for fine-tuning.

Metric	NLLB	Bing	GPT4
MIS	4.8025	2.9875	0.7025
TERM	3.62	2.1425	0.655
STYLE	3.01	2.3975	0.8425
GRAM	1.1475	0.82	0.1575
Overall	12.58	8.3475	2.3575

Table 4: Average Score for Different Models and Metrics on Error Types

Errors	NLLB	Bing	GPT4
No Error	81	119	242
Minor Error	183	206	144
Major Error	136	75	14

Table 5: Error Severity for different models (200 sentences x 2 annotators for each model)

Acknowledgements

We thank the reviewers for their valuable comments and feedback on our work. LH and GN are grateful for the grant "Integrating hospital outpatient letters into the healthcare data space" (EP/V047949/1; funder: UKRI/EPSRC).

References

Bauer, R. S. (2006). The stratification of English loanwords in Cantonese. *Journal of Chinese Linguistics*, 34(2):172–191.

Bojar, O. and Tamchyna, A. (2011). Improving translation model by monolingual data. In Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. F., editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Cohen, J. (1968). Weighted kappa - nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 70:213–20.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2023). *Ethnologue: Languages of the World*. SIL International, 26th edition.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. (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.

Gladkoff, S. and Han, L. (2022). HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Gladkoff, S., Sorokina, I., Han, L., and Alekseeva, A. (2022). Measuring uncertainty in translation quality evaluation (TQE). In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1454–1461, Marseille, France. European Language Resources Association.

Han, A. L.-F., Wong, D. F., Chao, L. S., He, L., Lu, Y., Xing, J., and Zeng, X. (2013a). Language-independent model for machine translation evaluation with reinforced factors. In *Proceedings of Machine Translation Summit XIV: Posters*, Nice, France.

Han, A. L.-F., Wong, D. F., Chao, L. S., Lu, Y., He, L., Wang, Y., and Zhou, J. (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.

Han, L., Gladkoff, S., Erofeev, G., Sorokina, I., Galiano, B., and Nenadic, G. (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.

Han, L., Jones, G., and Smeaton, A. (2020). AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Han, L., Sorokina, I., Erofeev, G., and Gladkoff, S. (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.

Ho, J. C. T. and Or, N. H. K. (2020). Lihkgr. https://github.com/justinchuntingho/LIHKGr. An application for scraping LIHKG.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hong, K. Y., Han, L., Batista-Navarro, R., and Nenadic, G. (2024). Cantonmt: Cantonese to english nmt platform with fine-tuned models using synthetic back-translation data. In *EAMT 2024*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Lee, J. L., Chen, L., Lam, C., Lau, C. M., and Tsui, T.-H. (2022). Pycantonese: Cantonese linguistics and nlp in python. In *Proceedings of The 13th Language Resources and Evaluation Conference*. European Language Resources Association.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liang, G., Zhao, J., Lau, H. Y. P., and Leung, C. W.-K. (2021). Using social media to analyze public concerns and policy responses to covid-19 in hong kong. *ACM Trans. Manage. Inf. Syst.*, 12(4).

Liu, E. K.-Y. (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.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Sparano, R. M., Foresi, M., Innis, J., Han, L., and Nenadic, G. (2024). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control.

NLLB-Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

OpenAI (2024). Gpt-4 technical report.

Post, M. (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.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (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.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In Erk, K. and Smith, N. A., editors, *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.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning.

Tiedemann, J. and Nygaard, L. (2004). The OPUS corpus - parallel and free: http://logos.uio.no/opus. In Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R., and Silva, R., editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Wiedenhof, J. (2015). *A Grammar of Mandarin*. John Benjamins, Amsterdam.

Wing, L. H. (2020). Machine translation models for cantonese-english translation project plan.

Wu, Y., Li, X., and Lun, C. (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.

Xiang, R., Tan, H., Li, J., Wan, M., and Wong, K.-F. (2022). When Cantonese NLP Meets Pre-training: Progress and Challenges. In Alonso, M. A. and Wei, Z., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 16–21. Association for Computational Linguistics.

Yi Mak, H. and Lee, T. (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.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert.

Zhao, Y., Xiang, L., Zhu, J., Zhang, J., Zhou, Y., and Zong, C. (2020). Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505, Barcelona, Spain (Online). International Committee on Computational Linguistics.