Leveraging Mandarin as a Pivot Language for Low-Resource Machine Translation between Cantonese and English

King Yiu Suen
Fano Labs
cyrus.suen@fano.ai

Rudolf Chow
Fano Labs
rudolf@fano.ai

Albert Y.S. Lam
Fano Labs
albert@fano.ai

Abstract

Cantonese, the second most prevalent Chinese dialect after Mandarin, has been relatively overlooked in machine translation (MT) due to a scarcity of bilingual resources. In this paper, we propose to leverage Mandarin, a high-resource language, as a pivot language for translating between Cantonese and English. Our method utilizes transfer learning from pre-trained Bidirectional and Auto-Regressive Transformer (BART) models to initialize auxiliary source-pivot and pivot-target MT models. The parameters of the trained auxiliary models are then used to initialize the source-target model. Based on our experiments, our proposed method outperforms several baseline initialization strategies, naive pivot translation, and two commercial translation systems in both translation directions.

1 Introduction

Cantonese is estimated to have 86.6 million native speakers (Eberhard et al., 2024), primarily spoken in Hong Kong, Macau, Guangdong, Guangxi, and various overseas Chinese communities (Wong et al., 2017). Originating from the same language family as Mandarin, Cantonese shares numerous vocabulary and grammatical similarities with its high-resource counterpart. However, despite these resemblances, the linguistic disparities between Cantonese and Mandarin are substantial enough to render them mutually unintelligible (Snow, 2004; Matthews and Yip, 2013). Consequently, the feasibility of leveraging Mandarin resources to process Cantonese text is widely questioned (Sio and Da Costa, 2019).

Despite the popularity of Cantonese, there has been limited effort for developing a quality translation system for Cantonese. As of the time of writing, Google Translate has yet to support Cantonese. In contrast to Mandarin, parallel corpora for Cantonese are extremely scarce, presenting significant challenges in training neural machine translation (NMT) models for Cantonese.

Researchers have proposed various strategies to address low-resource NMT. A technique that has been shown to be effective is to involve a pivot language. In pivot-based translation, the source sentence is first translated into a pivot language, which is then translated to the target language (De Gispert and Marino, 2006; Wu and Wang, 2007). Despite the simplicity, this method has a few disadvantages. Namely, it requires two translation models for decoding, equivalently doubling the number of parameters as well as the latency; errors from the source-pivot translation may also propagate into the final prediction. As such, studies have been investigating how to directly train a source-target model with the help of the pivot language, such as using the encoder from the source-pivot model and the decoder from the pivot-target model to initialize the source-target model (Kim et al., 2019; Zhang et al., 2022).

The choice of the pivot language is vital to the translation quality (Paul et al., 2009, 2013). Prior research typically selected the pivot language based on the relatedness between source and pivot languages, and the availability of bilingual language resources (Paul et al., 2009, 2013). For Cantonese-English translation, Mandarin is an obvious choice, as it is closely related to Cantonese and there is an abundance of Mandarin-English parallel corpora. However, to the best of our knowledge, no prior studies have examined using Mandarin as a pivot language for translating between Cantonese and English.

In this paper, we aim to bridge the research gap by providing empirical evidence that the usage of Mandarin can improve the translation performance between Cantonese and English. In particular, we use pre-trained Cantonese, Mandarin, and English Bidirectional and Auto-Regressive Transformer (BART; Lewis et al., 2020) models to initial-

Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 74–84

ize source-pivot and pivot-target translation models. The trained source-pivot and pivot-target translation models are then used to initialize the desired source-target translation model.

2 Background

In this section, we highlight some of the linguistic differences between Cantonese and Mandarin.

2.1 Vocabulary

While there is a considerable lexical overlap between Cantonese and Mandarin, it is estimated that approximately one-third of the vocabularies used in regular Cantonese speeches are absent in Mandarin (Snow, 2004). For example, "umbrella" is "遮" (*ze1*) in Cantonese but "雨傘" (*yǔ sǎn*) in Mandarin (Sio and Da Costa, 2019). In the cases where Cantonese and Mandarin share the same lexical items, it is almost always written with the same character (Snow, 2004). However, even the same Cantonese and Mandarin characters can be used differently (Snow, 2004). For example, "話" (*waa6*) in Cantonese is often used as a verb, meaning "to say". In Mandarin, the same character functions as a noun, meaning "speech". Moreover, there are characters that are unique to Cantonese, such as "冇" (*mou5*; to not have) and "咁" (*gam3*; so). Finally, a number of Cantonese words do not have a standardized written form (Matthews and Yip, 2013). For example, "to give" can be written as "比", "俾", "畀" or "被" in Cantonese (Bauer, 2018), which are all pronounced as "*bei2*".

2.2 Grammar

The differences in grammar between Cantonese and Mandarin are often very subtle. For example, in Cantonese, the noun representing the agent of the action must be present in indirect passive construction (Matthews and Yip, 2013), so "I was scolded" in Cantonese would be "我俾人鬧" (*ngo5 bei2 jan4 naau6*; I + by + person + scolded). In contrast, the agent can be omitted in Mandarin, so the sentence can either be "我被罵" (*wǒ bèi mà*; I + by + scolded) or "我被人罵" (*wǒ bèi rén mà*; I + by + person + scolded). Readers can refer to Snow (2004, p. 47) for more examples of grammatical differences.

2.3 Pronunciation

The pronunciation of the same character often varies between Cantonese and Mandarin. In numerous instances, the characters in Cantonese sound completely different from their Mandarin equivalents. For example, "學習" (to study) is pronounced as "*hok6 zaap6*" in Cantonese but "*xué xí*" in Mandarin, which are substantially different phonetically (Snow, 2004). Although these pronunciation differences are a primary reason why the two languages are not mutually intelligible when spoken, they typically do not impact the written form of the languages, making this issue irrelevant for translation.

2.4 Writing System

There are two written forms of Chinese: traditional and simplified Simplified Chinese, as it name suggests, is a simplified version of traditional Chinese. Simplified characters requires fewer strokes than their traditional counterparts. Cantonese and Mandarin do not inherently dictate which character set is used. Both spoken forms of Chinese can be written in traditional and simplified Chinese characters, although regional preferences exist. Traditional characters are predominantly used in Hong Kong, Macau and Taiwan, while Mainland China, Malaysia and Singapore favor simplified characters.

In our experiments, all simplified characters are converted to traditional characters using OpenCC[1] for smoother transfer learning.

3 Related Work

3.1 Machine Translation for Cantonese

In this section, we review the existing literature on Cantonese MT.

For Cantonese-Mandarin MT, Mak and Lee (2021) examined the feasibility of mining semantically similar sentences from articles on the same subject in Mandarin Wikipedia and Cantonese Wikipedia. Liu (2022) conducted a comparative analysis on the translation performance of Long Short-Term Memory (LSTM) and Transformer model architectures, alongside word-based and byte-pair encoding tokenization methods. Kwok et al. (2023) fine-tuned a pre-trained Mandarin BART using a parallel corpus of 130k sentence pairs from various online resources.

The earliest attempt on Cantonese-English MT was done by Wu and Liu (1999). They developed a statistical MT system that employed a combination of example-based and rule-based methods grounded on a bilingual knowledge base. A more

[1] https://github.com/yichen0831/opencc-python

recent effort by Hong et al. (2024) employed back-translation to synthesize a parallel corpus containing 200k sentence pairs. No prior research has studied the use of Mandarin as a pivot language for translating Cantonese to another language.

3.2 Pivot-based Machine Translation

In this section, we review existing approaches to leverage a pivot language in low-resource MT.

The naive approach is to independently train two auxiliary MT models, one for source-pivot and one for pivot-target, decoding twice via the pivot language (De Gispert and Marino, 2006; Wu and Wang, 2007). To reduce prediction errors, one can translate the top-n pivot-language sentences into target language, and then select the highest scoring sentence among the n target-language sentences (Utiyama and Isahara, 2007; R. Costa-jussà et al., 2011). A drawback of this strategy is that the translation speed is n times slower than the naive approach.

Another possibility is to use the pivot language to generate synthetic parallel data. This can be achieved by translating pivot-language sentences in pivot-target parallel corpora into source language (Bertoldi et al., 2008), translating pivot-language sentences in source-pivot parallel corpora into target language (De Gispert and Marino, 2006), or translating pivot monolingual data into source and target languages (Currey and Heafield, 2019).

Finally, one can combine pivoting with transfer learning (Kim et al., 2019). The high-resource source-pivot and pivot-target auxiliary models are first trained independently. Subsequently, a source-target model is initialized with the encoder from the source-pivot model, and the decoder from the pivot-target model. The source-target model is then fine-tuned with source-target data.

4 Proposed Method

Our proposed method is largely based on the transfer learning approach by Kim et al. (2019). A limitation of their approach is the requirement for a large amount of source-pivot and pivot-target parallel data to train the auxiliary models. However, given the scarcity of Cantonese parallel data, it is challenging to train a robust source-pivot model entirely from scratch. To address this issue, we also transfer parameters from pre-trained BART models to the source-pivot and pivot-target models, leveraging the data efficiency of pre-trained language models. We will illustrate our method in terms of Cantonese to English translation, but as our experiments will demonstrate, the method is equally effective in the reverse direction. The core steps of our method are as follows (Figure 1):

1. Pre-train the Cantonese (Yue), Mandarin (Zh) and English (En) BART models with monolingual corpora.

2. (a) Initialize the Yue-Zh model with the encoder from Yue BART and the decoder from Zh BART. Similarly, initialize the Mandarin-English model with the encoder from Zh BART and the decoder from En BART.
   (b) Continue training the Yue-Zh model with Yue-Zh parallel corpora, and the Zh-En model with Zh-En parallel corpora.

3. (a) Initialize the Yue-En model with the encoder from the trained Yue-Zh model and the decoder from the trained Zh-En model.
   (b) Continue training the Yue-En model with Yue-En parallel corpora.

Instead of training our own BART models from scratch, we use the base version of Zh BART model released by Shao et al. (2021) and the base version of En BART model released by Lewis et al. (2020). Both models have the same Transformer architecture (6 encoder and 6 decoder layers, with 12 attention heads and a hidden size of 768). Since there is no publicly available Yue BART model, we continue the pre-training of Zh BART with additional Yue monolingual data, leveraging the shared Chinese character system between Yue and Zh. The vocabularies of Zh BART already contain Cantonese characters, but the original pre-training materials are predominately in Mandarin. Considering the linguistic differences described in Section 2, we believe that this additional pre-training is warranted. Following Lewis et al. (2020), we pre-train Yue BART with the text infilling task: for each sentence, a random number of text spans are sampled, with span lengths drawn from a Poisson distribution ($\lambda = 3$). Each span is replaced with a single `[MASK]` token. The model is trained to reconstruct the original text without knowing how many tokens are missing for each span.

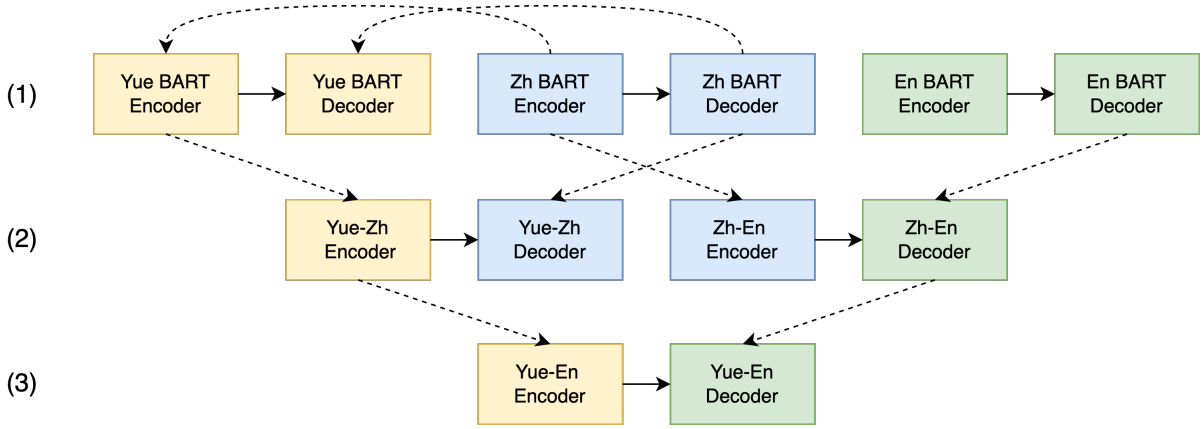

Figure 1: Our proposed pivot-based transfer learning method. Step 1 is the pre-training of BART models. Step 2 is the training of source-pivot and pivot-target models. Step 3 is the training of the source-target model. Solid lines represent translation directions. Dashed lines represent parameter initialization.

Corpus Type	Language	Size
Monolingual	Yue	10.0M
Parallel	Zh-En	17.1M
Parallel	Yue-Zh	31.0K
Parallel	Yue-En	61.7K

Table 1: Number of sentences in monolingual and parallel datasets.

5 Data

In this section, we describe the data used for pre-training and fine-tuning. Table 1 provides a summary of the sizes of the datasets.

5.1 Monolingual Datasets

Our continued pre-training data for Yue BART are composed of web-scraped data from online forums in Hong Kong. To cover a variety of domains, we scraped data from three forums: LIHKG[2], Baby Kingdom[3], and HKEPC[4]. LIHKG, often referred to as the "Reddit of Hong Kong" (Au, 2022), is a multi-category forum with topics including current affairs, gossips, sports, finance and entertainment. Baby Kingdom mainly targets local parents looking for parenting advice. HKEPC contains discussions on the latest technology and reviews on computer products. The text is split into sentences based on punctuation marks. Sentences that contains URL or have fewer than five Chinese characters are removed. This amounts to 10M sentences after pre-processing.

[2] https://lihkg.com
[3] https://baby-kingdom.com
[4] https://hkepc.com

5.2 Parallel Datasets

5.2.1 Mandarin-English

For Mandarin-English data, we use publicly available corpora: News Commentary v18.1 from WMT24 competition[5], UN Parallel Corpus v1.0 (Ziemski et al., 2016) and, WikiMatrix (Schwenk et al., 2019). For WikiMatrix, we filter sentence pairs with alignment quality below the score of 1.04, the same threshold used in Schwenk et al. (2019).

5.2.2 Cantonese-Mandarin

The sources of our Cantonese-Mandarin parallel data include story books[6], language learning websites[7,8], TED talks[9], a previous linguistic study (Wong et al., 2017)[10,11] and a dictionary[12].

5.2.3 Cantonese-English

The sources of our Cantonese-English parallel data include a language learning website[13] and two dictionaries[14,15].

[5] https://www.statmt.org/wmt24/translation-task.html
[6] https://global-asp.github.io/storybooks-hongkong
[7] https://tatoeba.org/
[8] https://www.ilc.cuhk.edu.hk/workshop/Chinese/Cantonese/OnlineTutorial/intro.aspx
[9] https://opus.nlpl.eu/TED2020/zh&zh-tw/v1/TED2020
[10] https://github.com/UniversalDependencies/UD_Cantonese-HK
[11] https://github.com/UniversalDependencies/UD_Chinese-HK
[12] https://kaifangcidian.com/han/yue/
[13] https://opus.nlpl.eu/Tatoeba/yue&en/v2023-04-12/Tatoeba
[14] https://wenlin.com/
[15] https://words.hk/

6 Experiment

In this section, we outline the baselines chosen for comparison. These baselines are selected to address the following research questions:

1. Can continued pre-training on Cantonese monolingual data improve translation performance?
2. Does transfer learning from source-pivot and pivot-target auxiliary models yield better translation performance than transfer learning from BART models?
3. Can direct training of a source-target model mitigate the issue of error propagation associated with naive pivot translation?

Our baselines include different initialization strategies for the Yue-En model. In particular, we choose to initialize the Yue-En encoder using the encoder from one of the followings: Zh BART, Yue BART or Yue-Zh model, and the Yue-En decoder using the decoder from one of the followings: En BART or Zh-En model. These encoder and decoder initialization strategies are fully crossed, resulting in six conditions. Additionally, we include a baseline where all parameters in the Yue-En model are initialized randomly from $N(0, 0.02)$, following the configuration used by Lewis et al. (2020). Moreover, we include pivot translation as a baseline, where source sentences are decoded twice: first through the Yue-Zh model, and then through the Zh-En model. The encoders and decoders are Yue-Zh and Zh-En models are initialized from pre-trained BART models, and are trained on Yue-Zh and Zh-En parallel data respectively. Finally, we compared our proposed method to two existing translation platforms that support translation between Cantonese and English: Azure AI Translator[16] and Baidu Fanyi[17].

To examine whether the same approach would work for translation from English into Cantonese, we also repeat the experiments for English to Cantonese, using Mandarin as a pivot. The translation directions in the auxiliary models are adjusted accordingly.

For all training and inference, we use one NVIDIA GeForce RTX 3090 GPU with a batch of 64. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a constant learning rate of 3e-5. To speed up training, the models are trained in half (16-bit) precision. We use a maximum of 500K training steps for pre-training Yue BART, and 10 training epochs for fine-tuning source-pivot, pivot-target and source-target models. We randomly select 5% of the parallel corpora as validation sets. The final models are selected based on validation loss. For the source-target model, we additionally select 10% of the parallel corpora to use as the test set. For decoding, we use beam-search with a beam size of 4. The translation performance is measured by the BLEU score (Papineni et al., 2002).

[16] https://azure.microsoft.com/en-us/products/ai-services/ai-translator

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

Encoder	Decoder	BLEU
Random	Random	6.03
Zh BART	En BART	15.10
Zh BART	Zh-En	16.94
Yue BART	En BART	17.25
Yue BART	Zh-En	19.33
Yue-Zh	En BART	17.12
Yue-Zh	Zh-En	**19.64**
Pivot Translation		10.12
Azure AI Translator		17.50
Baidu Fanyi		17.21

Table 2: Experiment results for Yue-En translation.

7 Results

In this section, we present the BLEU score of our experiments. Examples of translation results are analyzed in Appendix A. For each example, our model's translation is compared to that from Azure AI Translator and Baidu Fanyi.

7.1 Cantonese to English

Table 2 presents the results of our experiment for Yue-En translation. Random initialization yielded the poorest translation performance, with a BLEU score of just 6.03. This outcome is expected, given that the model is trained on a low-resource parallel corpus. Pivot translation resulted in the second lowest BLEU score of 10.12, likely due to translation error propagation from the Yue-Zh phase to the Zh-En phase.

Initializing the Yue-En model with BART models significantly enhanced translation performance. Specifically, the Zh BART encoder + En BART decoder combination achieved a BLEU score of 15.10. Replacing the Zh BART encoder with the

Encoder	Decoder	BLEU
Random	Random	13.43
En BART	Zh BART	24.56
En-Zh	Zh BART	27.22
En BART	Yue BART	24.68
En-Zh	Yue BART	27.82
En BART	Zh-Yue	25.01
En-Zh	Zh-Yue	**28.22**
Pivot Translation		15.63
Azure AI Translator		15.70
Baidu Fanyi		17.91

Table 3: Experiment results for En-Yue translation.

Yue BART encoder further improved the BLEU score to 17.25, likely because the Yue BART encoder can more accurately encode Cantonese sentences compared to the Zh BART encoder. This result supports our hypothesis that reusing Mandarin resources without additional pre-training is suboptimal for Cantonese processing.

A similar pattern can be observed when comparing Zh BART encoder + Zh-En decoder (16.94) and Yue BART encoder + Zh-En decoder (19.33). Using an encoder that can interpret Cantonese greatly improve the translation performance. The replacement of En BART decoder with Zh-En decoder also results in higher BLEU scores. This is because the Zh-En decoder, unlike En BART decoder, is trained to interpret the outputs from a Chinese encoder, allowing a smoother transfer learning.

The BLEU score for the Yue-Zh encoder + En BART decoder (17.12) is higher than that of the two baselines that use Zh BART encoder. However, it is lower than other methods except random initialization and pivot translation.

The highest BLEU score of 19.64 was obtained when both auxiliary models were used for initialization (Yue-Zh encoder + Zh-En decoder). The score is very close to that of Yue BART encoder + Zh-En decoder (19.33), likely because the encoders from Yue BART and Yue-Zh share a similar latent space, but the Yue BART encoder is not as finely tuned as the Yue-Zh encoder to generate latent representations interpretable by the Zh-En decoder.

7.2 English to Cantonese

Table 3 presents the results of our experiment for En-Yue translation. Random initialization and pivot translation resulted in the lowest BLEU scores, at 13.43 and 15.63, respectively. Surprisingly, the two commercial translation systems, Azure AI Translator and Baidu Fanyi, performed only marginally better, achieving BLEU scores of 15.70 and 17.91 respectively. All other baselines exhibited significantly higher BLEU scores.

Moreover, models with En-Zh encoder have higher BLEU scores than their counterparts with En BART. Among the models utilizing the En-Zh encoder, the one employing the Zh-Yue decoder achieved the highest BLEU score of 28.22. This once again shows that pivot-based transfer learning can a provide improvement in performance.

8 Conclusion

In this paper, we experiment pivot-based transfer learning as a way to improve the quality of low-resource Cantonese-English translation. Our approach involves transferring the parameters of pre-trained Yue, Zh, and En BART models to auxiliary source-pivot and pivot-target NMT models. These auxiliary models are then fine-tuned with parallel data. Finally, the parameters of the source-pivot encoder and pivot-target decoder are transferred to the desired source-target model. Our results demonstrate significant improvements over randomly initialized models, demonstrating the benefit of transfer learning. Moreover, transferring from models that are trained for related tasks (MT in auxiliary models versus text infilling in BARTs) and languages (Cantonese versus Mandarin) can further enhance the translation performance. Additionally, by training a single source-target model, we reduce the problem of error propagation in naive pivot translation. Finally, our model also outperforms two existing commercial translation systems, Baidu Fanyi and Azure AI Translator. Examples of translation results reveal that our model is better than both understanding and generating Cantonese idioms.

Future work can explore the potential of using Mandarin to generate synthetic Yue-En parallel data by, for example, translating Mandarin sentences in Zh-En parallel data into Cantonese.

Limitations

A limitation of our method is that in order to allow a smooth transfer learning, there should be sufficient monolingual data for the low-resource language to train the initial BART model. Besides, the pivot language should be similar to the low-resource lan-

guage, so that the auxiliary translation model between the low-resource language and the pivot language can be trained even with limited data. Finally, the pivot language must be a high-resource language that has a large amount of parallel data with the target language. This is necessary to train the auxiliary translation model between the pivot language and the target language. Our method may not generalize well to other low-resource languages if they do not meet these conditions.

References

Yung Au. 2022. Protest, pandemic, & platformisation in hong kong: Towards cities of alternatives. *Digital Geography and Society*, 3:100043.

Robert S Bauer. 2018. Cantonese as written language in hong kong. *Global Chinese*, 4(1):103–142.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *Proceedings of the 5th International Workshop on Spoken Language Translation: Papers*, pages 143–149, Waikiki, Hawaii.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong. Association for Computational Linguistics.

Adrià De Gispert and Jose B Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, 27 edition. SIL International, Dallas, Texas.

Kung Yin Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. Cantonmt: Cantonese to english nmt platform with fine-tuned models using synthetic back-translation data. *arXiv preprint arXiv:2403.11346*.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Raptor Yick-Kan Kwok, Siu-Kei Au Yeung, Zongxi Li, and Kevin Hung. 2023. Cantonese to written chinese translation via huggingface translation pipeline. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, pages 77–84.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Hei Yi Mak and Tan Lee. 2021. 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*, pages 81–87.

Stephen Matthews and Virginia Yip. 2013. *Cantonese: A comprehensive grammar*. Routledge.

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

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):1–17.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 221–224, Boulder, Colorado. Association for Computational Linguistics.

Marta R. Costa-jussà, Carlos Henríquez, and Rafael E. Banchs. 2011. Enhancing scarce-resource language translation through pivot combinations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1361–1365, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Joanna Ut-Seong Sio and Luis Morgado Da Costa. 2019. Building the cantonese wordnet. In *Proceedings of the 10th Global WordNet Conference*, pages 206–215.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 266–275, Pisa, Italy. Linköping University Electronic Press.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Yan Wu and James Liu. 1999. A Cantonese-English machine translation system PolyU-MT-99. In *Proceedings of Machine Translation Summit VII*, pages 481–486, Singapore, Singapore.

Meng Zhang, Liangyou Li, and Qun Liu. 2022. Triangular transfer: Freezing the pivot for triangular machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 644–650, Dublin, Ireland. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Appendix

A.1 Cantonese to English Examples

Table 4 shows a few examples of Yue-En translation. These examples demonstrate that our model is better at understanding Cantonese idioms than existing commercial translation systems, but may potentially contain gender bias.

Example 1 The literal translation of "大佬" (*daai6 lou2*) is the eldest or elder brother, but it can also be used for addressing a person when one is annoyed. In this context, the latter translation is more appropriate. Only our model made the correct translation.

Example 2 The literal translation of "入廠" (*jap6 cong2*) is to enter a factory, but in Cantonese, it is often used to imply "to be hospitalized (for a surgery)". Once again, only our model was able to make the appropriate lexical transformation.

Example 3 "嫌三嫌四" (*jim4 saam1 jim4 sei3*) is an idiomatic expression, which means "to express discontent about something". Azure and Baidu completely failed to translate it.

Example 4 This example shows a common source of errors in the test set: mis-translation of third-person pronouns. In Cantonese, the third-person singular pronoun "佢" (*keoi5*) is gender-neutral, so it may refer to people of any gender. However, our model seems to have a masculine default. This is possibly because masculine pronouns are over-represented in the training corpus.

A.2 English to Cantonese Examples

Table 5 shows a few examples of En-Yue translation. These examples demonstrate that our model is better at generating Cantonese phrases as well.

Example 1 "Shopping mall" is "商場" ("soeng1 coeng4") in Cantonese and "購物中心" ("kau3 mat6 zung1 sam1") in Mandarin. Our model used the Cantonese phrase in its translation, while Azure and Baidu used the Mandarin counterpart, indicating that our model has a stronger understanding of Cantonese. Moreover, our model arguably produced a better translation than the reference sentence, as the counter word "個" (*go3*) in the reference sentence is unnecessary in this context.

Example 2 In this example, our model correctly used the Cantonese particle "丫" (*aa1*) to soften the force of requests, where Azure and Baidu failed

to do so. It also correctly translated “please” into “唔該” (*m4 goi1*) to make the request more polite. Even though the word ordering is different from the reference, the translation is perfectly acceptable. In contrast, Azure AI Translator rendered “please” as “請” (*ceng2*), which, while not incorrect, is generally reserved for more formal contexts in Cantonese. However, our model missed the possessive pronoun, “你” (*nei5*; your), in its translation.

Example 3 The translation by our model was exactly the same as the reference sentence. Both Azure and Baidu missed the demonstrative determiner, “個” (*go3*). Azure used a Mandarin vocabulary, “寬” (*kuān*), to represent the word “wide”, while the correct Cantonese translation is “闊” (*fut3*). Baidu even mis-translated “basketball court” into a nonsensical word.

Example 1	
Source	大佬！咁貴喎，梗係冇人買啦！ *daai6 lou2 gam3 gwai3 waa1 gang2 hai6 mou5 jan4 maai5 laa1*
Reference	Hey ! It costs too much! People surely won't buy it!
Our approach	Hey , it's so expensive! Of course nobody buys it!
Azure AI Translator	Big brother , it's so expensive, of course no one buys it!
Baidu Fanyi	Eldest brother ! It's so expensive, of course no one bought it!
Example 2	
Source	佢岩岩出院，又試要 入廠 。 *keoi5 ngaam4 ngaam4 ceot1 jyun2 jau6 si3 jiu3 jap6 cong2*
Reference	He had just left the hospital but then went back in again.
Our approach	He had just left the hospital, and went back to the hospital again.
Azure AI Translator	He had just been discharged from the hospital, and he tried to enter the factory again.
Baidu Fanyi	He has just been discharged and is going to the factory again.
Example 3	
Source	咪再 嫌三嫌四 啦，廉價酒店係咁架啦。 *mi1 zoi3 jim4 saam1 jim4 sei3 laa1 lim4 gaa3 zau2 dim3 hai6 gam3 gaa3 laa1*
Reference	Please stop bitching . This is what cheap hotels are like.
Our approach	Stop complaining about cheap hotels like that.
Azure AI Translator	Don't be suspicious , that's the case with budget hotels.
Baidu Fanyi	Don't be too picky about three or four anymore, cheap hotels are like this.
Example 4	
Source	佢 本來做老師，最近轉行去炒股票。 *keoi5 bun2 loi4 zou6 lou5 si1 zeoi3 gan6 zyun2 hang4 heoi3 caau2 gu2 piu3*
Reference	She was a teacher originally, but she quitted and is now playing with the stock market.
Our approach	He used to be a teacher, but recently he switched to invest in stocks.
Azure AI Translator	He used to be a teacher, but recently he turned to stock speculation.
Baidu Fanyi	He was originally a teacher, but recently he switched to trading stocks.

Table 4: Yue-En translation results of sentences selected from the test set. The light gray color is used to highlight the translations of a specific phrase in the source sentence.

Example 1	
Source	There are many shopping malls in Wong Tai Sin.
Reference	黃大仙有好多個商場。 *wong4 daai6 sin1 jau5 hou2 do1 go3 soeng1 coeng4*
Our approach	黃大仙有好多商場。 *wong4 daai6 sin1 jau5 hou2 do1 soeng1 coeng4* There are many shopping malls in Wong Tai Sin.
Azure AI Translator	黃大仙有許多購物中心。 *wong4 daai6 sin1 jau5 heoi2 do1 kau3 mat6 zung1 sam1* There are many shopping malls in Wong Tai Sin.
Baidu Fanyi	黃大仙有好多購物中心。 *wong4 daai6 sin1 jau5 hou2 do1 kau3 mat6 zung1 sam1* There are many shopping malls in Wong Tai Sin.
Example 2	
Source	Lend me your book, please.
Reference	唔該借你本書俾我丫。 *m4 goi1 ze3 nei5 bun2 syu1 bei2 ngo5 aa1*
Our approach	借本書俾我丫唔該。 *ze3 bun2 syu1 bei2 ngo5 aa1 m4 goi1* Lend me a book please.
Azure AI Translator	請將你既書借畀我。 *cing2 zoeng3 nei5 gei3 syu1 ze3 bei2 ngo5* Please lend me your book.
Baidu Fanyi	唔該將你書畀我。 *m4 goi1 zoeng3 nei5 di1 syu1 bei2 ngo5* Please give your book to me.
Example 3	
Source	The basketball court is ten or so meters wide.
Reference	個籃球場有十幾米闊。 *go3 laam4 kau4 coeng4 jau5 sap6 gei2 mai5 fut3*
Our approach	個籃球場有十幾米闊。 *go3 laam4 kau4 coeng4 jau5 sap6 gei2 mai5 fut3* The basketball court is ten or so meters wide.
Azure AI Translator	籃球場大約有十米寬。 *laam4 kau4 coeng4 daai6 joek3 jau5 sap6 mai5 fun1* Basketball court is about ten meters wide.
Baidu Fanyi	扣場約莫有十米闊。 *kau3 coeng4 joek3 mok6 jau5 sap6 mai5 fut3* Buckle court is about ten meters wide.

Table 5: En-Yue translation results of sentences selected from the test set.