Low-Resource Neural Machine Translation: A Case Study of Cantonese

Evelyn Kai-Yan Liu
Uppsala University
kaiyan.liu.7557@student.uu.se

Abstract

The development of Natural Language Processing (NLP) applications for Cantonese, a language with over 85 million speakers, is lagging compared to other languages with a similar number of speakers. In this paper, we present, to our best knowledge, the first benchmark of multiple neural machine translation (NMT) systems from Mandarin Chinese to Cantonese. Additionally, we performed parallel sentence mining (PSM) as data augmentation for the extremely low resource language pair and increased the number of sentence pairs from 1,002 to 35,877. Results show that with PSM, the best performing model – bidirectional LSTM with Byte-Pair Encoding (BPE) – scored 11.98 BLEU better than the vanilla baseline and 9.93 BLEU higher than our strong baseline. Our unsupervised NMT (UNMT) results also refuted previous assumption (Rubino et al., 2020) that the poor performance was related to the lack of linguistic similarities between the target and source languages, particularly in the case of Cantonese and Mandarin. In the process of building the NMT system, we also created the first large-scale parallel training and evaluation datasets of the language pair. Codes and datasets are publicly available at https://github.com/evelynkyl/yue_nmt.

1 Introduction

There are over 85 million Cantonese speakers around the globe, and it is the de facto spoken language in Hong Kong, Macau, and the Canton region in China (Wong et al., 2017; Eberhard et al., 2021). The language is also deemed the most influential and well-known variety of Chinese languages after Mandarin (Matthews and Yip, 2013); nevertheless, Cantonese has rather limited linguistic resources. While there are varying sizes of Cantonese-English corpora, such as Hong Kong Hansards (Legislative Council of the Hong Kong Special Administrative Region, 2022) and Hong Kong Laws Parallel Text (Ma, 2000), the latter of which contains nearly 3 million parallel sentences between the two languages, the same cannot be said for the pair of Cantonese and Mandarin. English and Cantonese share very few linguistic features, and are considered distant languages. On the contrary, Cantonese and Mandarin are typologically similar in that they share more linguistic features such as grammatical structures and basic lexical items than Cantonese does with English (Wong and Lee, 2018). As such, our work aims to take advantage of the typological similarities between the two languages and investigate whether the similarities would enable decent translation quality despite having a limited amount of training data.

The existing Cantonese (Hong Kong variant) - Mandarin corpora are quite small and mostly in the domain of conversational transcripts and social media (Luke and Wong, 2015; Wong et al., 2017). This can be further demonstrated by the dependency treebank built by Wong et al. (2017), which consists of only 13,918 words/tokens, as compared to 285,000 in Mandarin in Universal Dependencies (UD; Nivre et al., 2020). Most state-of-the-art (SoTA) deep learning algorithms require a large amount of data to perform well. It holds true especially for more complex tasks, such as machine translation (MT), question answering, and neural text generation (Koehn and Knowles, 2017; Puri et al., 2020; Malandrakis et al., 2019). As a consequence, most of these complex tasks are not commonly applied to Cantonese.

Language, however, is the core of one's cultural identity (Coupland, 2007). In light of that, the main goal of this paper is to benchmark different Mandarin to Cantonese NMT approaches to pave the way for future research on Cantonese NMT systems. The contributions of the paper include providing the first baseline of Cantonese NMT and the first large training and evaluation parallel dataset of the language pair. Our hypothesis is that creating an MT system with a high-resource, typolog-

Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 28–40
October 16, 2022. ©2022 Association for Computational Linguistics

ically close language might produce decent translation outputs. If that is the case, the limited resources of Cantonese can be improved by utilizing the MT system, hence enabling implementations of NLP systems with better performance for the low-resource language.

2 Linguistic Considerations of Cantonese and Mandarin

Most Chinese texts encountered in NLP is in Mandarin, largely due to its high availability in linguistic resources. Nearly all Cantonese speakers can read and write in written Mandarin, and it is conventionally preferred in academic and legal settings to write in written Mandarin to convey a sense of formality (Snow, 2004). As a consequence, there is very little Cantonese text data available, which, in turn, makes Cantonese seldom included in a majority of the NLP research and systems (Lee et al., 2022).

Nevertheless, Cantonese and Mandarin are typologically similar languages, where they have a similar grammatical system and share basic lexical items such as times, numbers, and personal pronouns that are identical in orthography (Zhang, 1998). Even though the two languages are closely related, there are, indeed, a plethora of linguistic differences. The most notable one is the phonological systems, in which their similarities are minimal in terms of sound inventory and intonation (Zhang, 1998; Tang and Van Heuven, 2009). On the aspects of syntactic structure, the main difference between the two languages is their word orders, where Cantonese allows a more flexible word ordering compared to Mandarin (Ding and Féry, 2014). Furthermore, there are distinctive grammatical features in Cantonese that do not exist in Mandarin (Zhang, 1998), including, but not limited to, post-verbal elements, structural particles, directional verbs, definiteness, and aspect markers. In terms of lexical dissimilarity, there are seven to eight thousand distinct words and expressions in Cantonese that are written in a different character from any Mandarin words, or that carry a different meaning from the Mandarin words of similar forms (Zhang, 1998). These distinct words attribute almost one third of the total vocabulary in Cantonese, and half of them are commonly used in daily conversation among Cantonese speakers.

Take a parallel sentence pair from the UD data as an example to illustrate the differences and similarities between the two languages. Sentence 1 denotes its expression in Cantonese, while Sentence 2 refers to the sentence in Mandarin.

(1) 嗰**時**啲 CD舖 仲多過
That **time**'s CD shops even more than
而家啲 **七十一**。
now's **7-11**.

"There were more CD shops at that time than the 7-11 (convenience stores) we have now."

(2) 那**時**候 唱片店 比現在
At that **time** CD shops compared to now
七十一 還要多。
7-11 even more.

"There were more CD shops at that time than the 7-11 (convenience stores) we have now."

As can be observed, the lexical tokens between the two sentences are quite different, with only four words (in bold) in overlap and three of them being a numerical item and one being a timing word. On the contrary, the syntactic structures are roughly similar, with word order differences such as the placement of time, subject, and comparison expression.

The distinctive lexical, syntactic, and phonological differences result in the language pair being mutually unintelligible (Zhang, 1998). Consequently, transforming from Mandarin to Cantonese should be treated as a translation task.

3 Related Work

3.1 Cantonese Parallel Corpus

Wong et al. (2017) constructed a parallel corpus of Cantonese and Mandarin in Standard Traditional Chinese scripts. This corpus is the first, albeit small (1,002 sentence pairs), Cantonese-Mandarin parallel corpus. It is created by transcribing television programs in Hong Kong as Cantonese data and using the original subtitles of the programs in Mandarin (Wong et al., 2017).

3.2 Parallel Sentence Mining (PSM)

PSM, sometimes referred to as bitext mining, identifies sentence pairs that are, or are close to, translations of one another (Feng et al., 2020). It makes use of two comparable corpora, which contain non-translated bilingual documents that are aligned on topics but not at the sentence-level (Rapp et al., 2016). PSM has been commonly applied to MT for lower resource languages as a data augmentation to improve the performance of an MT system (Stefanescu et al., 2012; Uszkoreit et al., 2010; Munteanu and Marcu, 2005). It can also be applied to a larger-scale scenario that contains a multilingual machine translation system with thousands of language directions (Fan et al., 2021). Hence, PSM enables one to source high quality parallel sentences effectively and efficiently and is the most useful in multilingual research, especially in a low resource setting. Moreover, mining sentences from comparable corpora overcomes some of the limitations that exists in parallel datasets (Zweigenbaum et al., 2018). In particular, large parallel corpora typically cover only a subset of the variety of language pairs, and they are often in very specific domains and genres. Furthermore, most of the parallel sentences are constructed by using human translations; therefore, these translations are likely to contain translation biases such as calques and other phenomena (Zweigenbaum et al., 2018). Contrarily, comparable corpora often display more variety and are generally original texts instead of translations. As such, it holds more promises as a complement to parallel corpora to aid in terms of variety and quantity of the data.

The goal of PSM is to find semantically similar sentences by calculating multilingual sentence embeddings, followed by finding the K-nearest neighbor sentences for all sentences in both directions, and finally, calculating all possible sentence combinations (Feng et al., 2020). The higher score a sentence pair has, the better it could serve as a translation pair (Reimers and Gurevych, 2020). Generally, scores higher than 1 indicate that it is of quality. Reimers and Gurevych (2020) reported that using LaBSE (Feng et al., 2020) as the mining model returned the best results in their experiments.

3.3 Low-resource NMT

While NMT has demonstrated its performance in resource-rich language pairs, research has shown that the same performance does not apply in limited data situations (Koehn and Knowles, 2017; Sennrich and Zhang, 2019). As reported by Gu et al. (2018), NMT systems cannot achieve reasonable translation results if the corpus has less than 13K parallel sentences. As such, to improve the quality of low-resource NMT models, researchers have proposed a plethora of methods, which can be categorized into two groups:

1. **Monolingual data**. Exploiting data from the target language is low-cost and effective. Approaches range from back translation which takes advantage of the target-side monolingual corpus (Sennrich et al., 2016a), bilingual text mining (Feng et al., 2020), joint training in both translation directions (Zheng et al., 2019), as well as language models pretraining (Conneau and Lample, 2019; Lewis et al., 2019).
2. **Auxiliary languages' data**. Leveraging other language pairs' corpora for pre-training or joint representation learning has shown great success even with extremely low-resource language pairs (Zoph et al., 2016; Kocmi and Bojar, 2018). There are several methods of leveraging multilingual data for low-resource NMT, including transfer learning (Conneau et al., 2019; Chronopoulou et al., 2020), multilingual training (Gu et al., 2018; Zhang et al., 2020), as well as pivot translation (Wu and Wang, 2009; Wang et al., 2021).

3.3.1 Unsupervised NMT

There has been tremendous progress in using unsupervised NMT (UNMT) as opposed to supervised NMT in recent years (Artetxe et al., 2018). While a UNMT model performs well when trained on a large, high quality, and comparable dataset, it does not perform well for languages with lesser availability of data (Chronopoulou et al., 2020). To solve this issue, Chronopoulou et al. (2020) proposed pre-training a monolingual LM (MonoLM) on a high-resource language, then fine-tuning the LM on the language pair, followed by an initialization of a UNMT model. They also introduced a new vocabulary extension approach that enables fine-tuning a pre-trained LM to any unseen language. The results showed that their approaches outperformed XLM (Conneau and Lample, 2019), a SoTA cross-lingual language model pre-training framework, on several language pairs. Furthermore, they added residual adapters (Rebuffi et al., 2018) to the layer of each of the pre-trained MonoLM. Residual adapters are feed-forward networks that prevent catastrophic forgetting of the model (Bapna and Firat, 2019). Chronopoulou et al. (2020) reported that adapters enable fine-tuning parameters in a more time-saving and cost-efficient manner with little to no cost on the performance as compared to the originally proposed model.

4 Methodology and Experimental Setup

We implemented the following NMT systems on the direction of Mandarin to Cantonese in the experiment:

1. Word-based bidirectional LSTM model with general attention mechanism as the baseline ($BiLSTM$),
2. Word-based (1) + fine-tuning as a strong baseline ($BiLSTM_t$),
3. Word-based (2) + PSM ($BiLSTM_t$ + PSM),
4. BPE-based fine-tuned BiLSTM + PSM ($BiLSTM_{bpe+t}$ + PSM),
5. Word-based Transformer ($Trans_w$ + PSM),
6. BPE-based Transformer ($Trans_{bpe}$ + PSM),
7. Unsupervised NMT via language model pretraining and transfer learning with adapters ($RELM_{adap}$ + PSM)

4.1 Data, Bitext mining, and Prepossessing

Original Dataset This paper used the Cantonese-Mandarin parallel corpus by Wong et al. (2017) in Universal Dependencies (Nivre et al., 2020) as the foundation, which we refer to as UD in this paper. It consists of 1,002 sentence pairs (see Section 3.1).

Data Augmentation: Bitext Mining Considering the small size of the corpus, we used a data augmentation technique by mining sentence pairs. The Cantonese and Mandarin Wikipedia sites were extracted to perform the mining.[1] The bitext mining was performed via the SoTA LaBSE (Feng et al., 2020) to select pairs of semantically similar sentences following the scripts from Reimers and Gurevych (2019). Feng et al. (2020) suggested that sentences with a score of 1 are of quality, and 1.2 of high quality. However, we performed a qualitative review on a subset of the results and observed that sentences that scored 1.1286 are already of high quality and are semantically similar to each other. As such, we set the score threshold to 1.1286. After filtering out sentences below the threshold, we found 34,873 sentence pairs with equal to or over 1.1286 score. Having performed bitext mining, our total number of sentence pairs for training and evaluation has increased from 1,002 to 35,877. The increase in data size enables us to train a NMT system that is possible to perform well, since NMT systems are not able to achieve decent results when the training data has less than 13K pairs (Gu et al., 2018).

Final Datasets The newly complied data served as a synthetic parallel dataset to augment the UD dataset and alleviated the lack of sufficient training data. We refer to the combination of the two datasets as UD + Bitext, which was used to train all experimental models except the baselines. UD and UD + Bitext sets were both used for training and evaluation. Table 1 shows the distribution of the datasets used in the experiments.

No. of sentences	**UD**	**UD + Bitext**
Training	801	24,396
Validation	100	5,382
Test	101	6,099
Total	1,002	35,877

Table 1: Ratio of the datasets in the experiment (all randomly divided)

Preprocessing No word segmentation is done for the UD dataset as it is already tokenized. The mined parallel sentences were tokenized using Jieba.[2] We removed blank lines but did not normalize punctuation or non-Chinese characters. We used Byte Pair Encoding (BPE) preprocessing for $BiLSTM_{bpe+t}$ + PSM, $Trans_{bpe+t}$ + PSM, and $RELM_{adap}$ + PSM while word-based preprocessing was used for the baselines, $BiLSTM_t$ + PSM, and $Trans_w$ + PSM. We used fastBPE for $RELM_{adap}$ + PSM since the pre-trained model used this technique and subword NMT (Sennrich et al., 2016b) for the rest of the models with BPE representation. We trained the BPE tokenizers on our datasets with a maximum number of 8K BPE tokens in the vocabulary for models with these word representations.

4.2 Experiments

We trained (i) a BiLSTM model with attention and (ii) a Transformer model and compare them with (iii) an unsupervised NMT (UNMT) model using the RELM framework. Both (i) and (ii) were trained using Adam optimizer (Kingma and Ba, 2014) and cross-entropy loss function. We conducted the supervised NMT (SNMT) experiments

[1] https://dumps.wikimedia.org/backup-index.html

[2] https://github.com/fxsjy/jieba

using JoeyNMT (Kreutzer et al., 2019), while the UNMT is trained via PyTorch (Paszke et al., 2019). More details about the training parameters in the experiments can be found in Appendix A.

4.2.1 BiLSTM-based NMT

NMT Baselines We trained two word-based BiLSTM models with a learning rate of 3e-04 as our baselines. The vanilla baseline was implemented without exhaustive fine-tuning. Considering the sensitivity of under-resourced NMT to hyperparameters tuning (Sennrich and Zhang, 2019), it is crucial to optimize the model. Hence, a strong baseline was implemented following the parameter settings in Sennrich and Zhang (2019). We used 1 layer of encoder with 64 embedding dimension and 128 hidden units, and a batch size of 64. For regularization, we applied 0.2 drop-out and 0.3 hidden drop-out. A beam size of 5 was used for decoding.

BiLSTM with augmented data After parallel sentence mining, we extended our baselines to examine the effectiveness of data augmentation. We trained two models of different encoding schemes with the augmented data (approach 3 and 4) using the same model architecture (1 layer encoder of BiLSTM). The training parameters of these models were adjusted based on the strong baseline in consideration of the increased size in training data. Both approaches were trained on 3e-04 learning rate, an embedding dimension of 128, a hidden size of 256, 0.25 drop-out and 0.3 hidden drop-out, a batch size of 64, as well as a beam size of 10.

4.2.2 Transformer-based NMT

With the additional data from parallel sentence mining, it increases the chance of having a better performing Transformer NMT. Thus, we implemented two exhaustively tuned Transformer-based models on both word-level and BPE-level. The models were trained with identical parameters, including a learning rate of 2e-04, a batch size of 10, 2 layers of encoders with 4 attention heads, 0.1 drop-out rate, and a beam size of 5. The only differences are the embedding dimension and hidden size, where we used 64 each for the BPE-level model and 128 each for the word-level one.

4.2.3 UNMT via Transfer Learning

As mentioned in Section 3.3.1, researchers have reported success on transferring a pre-trained monolingual LM to a UNMT model even with some resource-poor language pairs (Chronopoulou et al., 2020). In light of that, we trained a UNMT system using the RELM framework (Chronopoulou et al., 2020) using the UD + Bitext dataset for monolingual model. For monolingual LM pre-training, we used 385,486 sentences (Mandarin) as the training data. Then, we fine-tuned part of the LM on the target language using only adapters with the same amount of parallel sentences. Finally, we trained a Transformer-based UNMT model by initializing the encoder and decoder with the fine-tuned model plus the adapters in both translation directions. We followed the default parameters of RELM for model training, with a learning rate of 1e-04, a batch size of 32, 512 embedding dimension and hidden size, 3 layers and 4 heads, a hidden and non-hidden drop-out rate of 0.1, A multilayer perceptron (MLP) attention, along with a beam size of 5.

4.3 Evaluation

4.3.1 Datasets

We used two datasets for evaluation, including (i) UD, and (ii) UD + Bitext. They were used as input to the translation systems for evaluating the quality of the NMT models aside from the automatic metric. It allows us to perform a qualitative investigation on the translation outputs of the proposed Mandarin-Cantonese NMT systems.

4.3.2 Methods

The automated evaluation metric used in this paper is detokenized SacreBLEU scores (Post, 2018). We report test set scores on the checkpoints with the highest BLEU score in the validation set. In addition, we performed manual evaluation on a subset of the evaluation data to get a better sense of the translation quality. The SacreBLEU results are reported and discussed below.

5 Results

Table 2 reports the primary results of our experiments. Having such a limited amount ($\sim$1K sentence pairs) of data, as expected, completely fails to train a vanilla BiLSTM translation model. Applying training tricks and exhaustive hyperparameter tuning, as suggested by Sennrich and Zhang (2019), has led to an improved result (+2.05 BLEU). However, the score and quality is too low for the translation outputs to be comprehensible.

Among all the models in the experiment, the data-augmented BiLSTM models are the best-performing, with the word-level model scoring

Architecture	Model	SacreBLEU
BiLSTM	Word level vanilla NMT, baseline 1	1.24
	Word level, $BiLSTM_t$, baseline 2	3.29
	Word level, $BiLSTM_t$ + PSM	12.37
	BPE level, $BiLSTM_{t+bpe}$ + PSM	**13.22**
Transformer	Word level + PSM ($Trans_{word}$)	3.56
	BPE level + PSM ($Trans_{bpe}$)	11.66
UNMT	$RELM_{adap}$ + PSM	1.85

Table 2: Experimental results on the Mandarin-Cantonese translation direction. PSM refers to the parallel sentence mining technique to increase data size. The highest score is in bold.

12.37 BLEU and the BPE-level one scoring 13.22 points. Word-level MT models are typically slower to converge, and thus, require more training to have on-par performance with their BPE-level counterparts (Sennrich et al., 2016b; Wu et al., 2016). Given that the two models were trained on an identical number of epochs, it is reasonable that the BPE-level one, which converges faster, performed better. The Transformer-based models are outperformed by the BiLSTM models. It is not surprising given the limited data in the experiments. The UNMT system with pre-trained LM scored 1.85 on SacreBLEU (+0.61 points compared to the vanilla baseline) and is the second-worst performing model in the experiment. The strong baseline (model 2, fine-tuned vanilla NMT) outperforms it by 1.44 BLEU even with only 1K sentence pairs.

6 Analysis

Effect of corpus size Bitext mining improved the model performance substantially (+9.08 BLEU, $BiLSTM_t$ + PSM compared to $BiLSTM_t$) with merely some minor changes in the training parameters in view of the increased data size. It shows that this technique is successful in assisting model learning and thus improving its performance by increasing the size of the training data.

6.1 Out-of-vocabulary (OOV)

Upon careful examination of the translations, we observed that OOV is a critical issue for both BiLSTM and Transformer models. OOV occurs when the translation output contains unknown tokens (UNK), which are unseen words or rare words whose occurrences are less frequent than other words in the vocabulary in the training data. The issue is a major challenge for any language in a low-resource scenario (Liu and Kirchhoff, 2018). In a low-resource setting, the dictionary created from the selected training data is not able to cover all the possible words and characters in the language. Consequently, when evaluated on an independent test set, it is highly likely that many terms that were not covered in the training data have then become unknown tokens. Given that our limited size of training data, OOV is a severe problem that negatively impacts model performance. Table 3 shows examples of translations generated by BiLSTM and Transformer models with word and Byte-Pair Encoding (BPE) representations.

Word-level systems For word-level BiLSTM and Transformer systems, we observed that the translation quality of the validation set is better than the test set, and they did not produce UNK tokens like the BPE-level models. They still, however, suffer from OOV. Due to the lack of UNK tokens, we are unable to measure the severity of this issue for word-level systems. The reason behind the absence of UNK tokens is word-based NMT models' inability to translate unseen words (Sennrich et al., 2016b); instead, they copy unknown words to the outputs, resulting in plenty of words copied directly from the training data of the source language. The quality of the translation from word-based models, as a result, is similar to the BPE-level one for the BiLSTM models. In contrast, the performance is significantly worse for the Transformer model. The result from the word-level Transformer model contains either single, irrelevant words or numerous duplicate words, making it uninterpretable. Referring to Table 3, the output sentence from this model is completely different from the reference sentence, either in terms of topic, sentence structure, or semantics. The output from word-level BiLSTM bears a closer resemblance to the reference text, albeit barely intelligible. It also copied many words from other sentences in the training data, as some words like **轉到** "turned" and **都係** "is also" are unrelated and thus should not be used in the sentence.

BPE-level systems BiLSTM with BPE representations has the highest number of UNK tokens compared to the rest of the experimental models

Model	Sentence from BPE-level model	Sentence from word-level model
Gold standard		
Original sentence	沙田區議員曾提出重建西林寺爲旅遊地	
Translation	District Councillors of Sai Tin had proposed to renovate Sai Lam Temple as a tourist attraction.	
BiLSTM		
Original output:	沙田<unk>曾經委任做旅遊<unk>	沙田都係之後轉到西林寺爲旅遊地
Translation	<unk> Sai Tin was assigned as a tourist <unk>.	Sai Tin then turned the Sai Lam Temple to a tourist attraction
Transformer		
Original output	而家都係沙田<unk>西林寺爲旅遊地	問題
Translation	Sai Lam Temple is still a tourist attraction in Sai Tin <unk>.	Problem

Table 3: Translation examples from the word-level and BPE-level models illustrating Out-of-Vocabulary (OOV) issue

(UNK to word ratio is 63.3% on the test set). In spite of that, the SacreBLEU of this model surpassed the rest, meaning that the accuracy of the non-UNK translated tokens is quite decent. For the BPE-level Transformer model, its occurrence frequency of UNK tokens is much lower than its BiLSTM counterpart (UNK to word ratio 38.5% as compared to 63.3%). Although $BiLSTM_{bpe+t}$ + PSM's BLEU score is higher than $Trans_{bpe}$ + PSM's, our analysis suggests the opposite in terms of translation quality. We found the translations by $Trans_{bpe}$ + PSM contains fewer UNK tokens and a closer semantic meaning to the reference sentences. These findings corroborate the UNK to word ratios reported above. Despite having a less severe OOV issue, the Transformer model still performs worse in terms of BLEU score, yet it intriguingly performs better on the aspects of translation quality. As shown in Table 3, the sentence output by the BPE-level Transformer model contains fewer UNK tokens, as well as a closer semantic meaning to the reference sentence. It is due to the fact that the Transformer model does not produce the exact words as the reference text, but a rephrased version; conversely, the BiLSTM model, as a sequence-to-sequence model, is more prone to direct-copying from the training text (Sutskever et al., 2014; Gu et al., 2016). Hence, its output would theoretically have more exact words. Since BLEU (Papineni et al., 2002) is concerned about the exact match in the translated text and the reference text, one of the plausible explanations of the above phenomenon is that the metric favors models that have a copying tendency.

Moreover, consistent with the findings of Artetxe et al. (2018), we observed that BPE is of scant help in terms of UNK tokens when the name entities or phrases are infrequent. Despite subword translations such as BPE being beneficial to OOV problems in general, such an advantage is hardly observed in this study. A likely explanation is that our source and target languages are both character-rich languages. While they can have over 50,000 characters in their languages, only a fraction of those are used regularly (Wang et al., 2020). Yet, many infrequently used characters can take up a considerable amount of vocabulary slots (Wang et al., 2020). As such, when two languages do not have many overlapping character sets, BPE might not be an optimal choice compared to other subword tokenization schemes such as Byte level BPE (BBPE; Wang et al., 2020) or unigram language modeling (Kudo, 2018). Future studies can explore the impacts of different subword tokenization techniques on this language pair to further increase the NMT performance.

UNMT The UNMT model performs considerably worse than the supervised MT models. The gap between the two approaches is very significant when we consider the identical data size. The BLEUs of the supervised approach are at least 9.81 higher than the UNMT model, whose score is only marginally better than the vanilla baseline. As such, for very low-resource language pairs, training an MT system with 36K synthetic parallel data is a better option. The majority of the translation output by the UNMT model are duplicates of some word, making the result unintelligible. Hence, we are not able to analyze it in-depth. Despite the success of Chronopoulou et al. (2020), our experimental results are in line with the previous work on UNMT for low-resource languages (Rubino et al., 2020). It is worth noting that even though our language pair (Cantonese and Mandarin) is highly similar typologically, the model performance is still similar to that of Rubino et al. (2020) in terms of BLEU. As such, in the case of Cantonese and Mandarin, we refuted their assumption that the poor performance was related to the lack of linguistic similarities between the target and source languages. We believe that the poor performance is largely tied to the amount of monolingual data in the LM pretraining step. It is also possible that although Cantonese and Mandarin are typologically close, the differences in word ordering or grammatical features made them linguistically less similar. However, compared to the language pairs in Rubino et al. (2020), our target and source languages share many more linguistic similarities. Hence, it is more likely that the poor performance is due to the limited data of our language pair. As a consequence, more training data is required to better aid the model to learn the language representations.

In addition, the language pair in this research differs greatly from the language pairs that performed well in the previous studies, such as English-French and German-English (Artetxe et al., 2018; Lample et al., 2018). Since both Cantonese and Mandarin are logographic languages, using a different subword representation method than the default BPE one might lead to a better-performing model.

7 Limitations

Translation systems are prone to making generalizations based on the frequency of gender-role, race, religion, and other stereotypes occurrences in the datasets. One typical example is “Man is to Programmer as Woman is to Homemaker” (Bolukbasi et al., 2016). The Cantonese-Mandarin UD parallel treebank used in this study was sourced from a television show, which might contain stereotypes in the dialogues. Besides, the bitext mined sentence pairs were sourced from the Wikipedia sites of Cantonese and Mandarin. Given that Wikipedia is an open-source community where everyone can contribute, its content could be vulnerable to social injustice and stereotypes as well. Their presence in the training data, if any, would reinforce the stereotypes in the translation system. One way to mitigate such potential issues is by treating it as a domain adaptation problem, as recommended by Saunders and Byrne (2020).

In terms of evaluation, the main automated metric in this study is SacreBLEU. Using only one metric, however, is not able to provide a full picture of the model performance and its translation quality. Although we used manual analysis along with SacreBLEU, having a non-matching based metric such as BERTScore (Zhang* et al., 2020) or SIMILE (Wieting et al., 2019) would be helpful in evaluating the contextual similarity between the input and the translation.

8 Conclusion and Future Work

In this paper, we presented the first benchmark of various NMT approaches for Cantonese. Due to the minimal amount of training data, the baseline models failed to produce intelligible results. We alleviated this issue by using parallel sentence mining as data augmentation and have increased the training data size from ∼1K to ∼36K. It resulted in a tremendous boost in performance (+9.08 BLEU) and produced higher-quality translations. Additionally, we provided a large parallel training and evaluation dataset of Cantonese and Mandarin for future research.

One of the interesting findings in this paper is that our Transformer MT systems performed worse than the BiLSTM systems in terms of SacreBLEU. This is reasonable given the large amount of data required by Transformer-based models and the limited amount of training data. What is more intriguing is that using varied word representations in an NMT system leads to very different results. We found that BPE-level models generally perform better. The BPE-level Transformer model produces more comprehensible translations despite having a lower BLEU score than the two BiLSTM models. We hypothesize that this is because of the evaluation metric’s (BLEU) architecture favoring models with a copying tendency. Besides the supervised models, we also implemented an unsupervised NMT with LM pre-training. It is, however, among the worst-performing models, in spite of the large amount of training data in comparison with the rest of the models.

Future work can be dedicated to different approaches to improve the performance of Mandarin-Cantonese NMT systems. While this study has investigated the direction of Mandarin to Cantonese as a way to alleviate the lower resource in Cantonese, our next step would include both translation directions as well. In addition, one could explore various approaches to mitigate the severe OOV issue, such as applying Jyutping romanization of the characters (Du and Way, 2017; Aqlan et al., 2019) or using BBPE (Wang et al., 2020) or unigram language modeling (Kudo, 2018) rather than BPE as the subword tokenization technique.

Another research direction is to train a multilingual NMT system (MNMT). With various source languages, the model is able to learn universal language representations from all the languages, thus enabling the systems to be language agnostic (Lee et al., 2017; Johnson et al., 2017; Feng et al., 2020). In our case, it may enable Cantonese to take advantage of the universal language representations in terms of linguistics and knowledge, hence allowing the system to perform well regardless the amount of available data (Gu et al., 2018).

References

Fares Aqlan, Xiaoping Fan, Abdullah Alqwbani, and Akram Al-Mansoub. 2019. Arabic–Chinese neural machine translation: Romanized Arabic as subword unit for Arabic-sourced translation. *IEEE Access*, 7:133122–133135.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

Picus Sizhi Ding and Caroline Féry. 2014. Word order, information structure and intonation of discontinuous nominal constructions in Cantonese/ordre des mots, structure de l'information et intonation des phrases nominales discontinues en cantonais. *Cahiers de Linguistique Asie Orientale*, 43(2):110–143.

Jinhua Du and Andy Way. 2017. Pinyin as subword unit for Chinese-sourced neural machine translation.

Eberhard, David M, and Gary F Simons. 2021. Ethnologue: Languages of the world. twenty-fourth edition.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. 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): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Jackson L Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. PyCantonese: Cantonese linguistics and NLP in python. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 6607–6611. European Language Resources Association.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Legislative Council of the Hong Kong Special Administrative Region. 2022. Hong Kong Hansard Database.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Angli Liu and Katrin Kirchhoff. 2018. Context models for oov word translation in low-resource languages. *arXiv preprint arXiv:1801.08660*.

Kang Kwong Luke and May LY Wong. 2015. The Hong Kong Cantonese corpus: design and uses. *Journal of Chinese Linguistics*, 25(2015):309–330.

Xiaoyi Ma. 2000. Hong Kong Laws Parallel Text.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. 2019. Controlled text generation for data augmentation in intelligent artificial agents. *arXiv preprint arXiv:1910.03487*.

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

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*.

Reinhard Rapp, Serge Sharoff, and Pierre Zweigenbaum. 2016. Recent advances in machine translation using comparable corpora. *Natural Language Engineering*, 22(4):501–516.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Raphael Rubino, Benjamin Marie, Raj Dabre, Atushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2020. Extremely low-resource neural machine translation for Asian languages. *Machine Translation*, 34(4):347–382.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

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

Dan Stefanescu, Radu Ion, and Sabine Hunsicker. 2012. Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 137–144.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Chaoju Tang and Vincent J Van Heuven. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, 119(5):709–732.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. *arXiv preprint arXiv:2107.04239*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. 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.

Tak-sum Wong and John Lee. 2018. Register-sensitive translation: a case study of mandarin and Cantonese (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 89–96, Boston, MA. Association for Machine Translation in the Americas.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 154–162.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

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*.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. 2019. Mirror-generative neural machine translation. In *International Conference on Learning Representations*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. A multilingual dataset for evaluating parallel sentence extraction from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

A Appendix

Table 4 lists the training hyperparameters used for the models in the experiments.

Experiments	Hyperparameters					
	Encoding	Learning rate	Batch size	Maximum epoch	Embedding dimension	Hidden size
$BiLSTM$	word	0.0003	256	600	128	128
$BiLSTM_t$	word	0.0003	64	800	64	128
$BiLSTM_t$ + PSM	word	0.0003	64	100	128	256
$BiLSTM_{bpe+t}$ + PSM	bpe	0.0003	64	100	128	256
$Trans_{word}$ + PSM	word	0.0002	10	300	128	128
$Trans_{bpe}$ + PSM	bpe	0.0002	10	300	64	64
$RELM_{adap}$ + PSM	bpe	0.0001	32	5000	512	512

Table 4: Hyperparameters of the experimental models in the study.

Experiments	Hyperparameters					
	Layer(s)	Head(s)	Drop-out	Hidden drop-out	Attention	Beam size
$BiLSTM$	2	0	0.2	0.2	MLP	10
$BiLSTM_t$	1	0	0.3	0.3	MLP	5
$BiLSTM_t$ + PSM	1	0	0.25	0.3	MLP	10
$BiLSTM_{bpe+t}$ + PSM	1	0	0.25	0.3	MLP	10
$Trans_{word}$ + PS	2	4	0.1	0	MLP	5
$Trans_{bpe}$ + PSM	2	4	0.1	0	MLP	5
$RELM_{adap}$ + PSM	3	4	0.1	0.1	MLP	5

Table 4: Hyperparameters of the experimental models in the study, continued.