MITRA-zh: An efficient, open machine translation solution for Buddhist Chinese

Sebastian Nehrdich[1], **Marcus Bingenheimer**[2], **Justin Brody**[3], **and Kurt Keutzer**[1]

[1]University of California, Berkeley , Berkeley Artificial Intelligence Research (BAIR)
[2]Temple University, Philadelphia,
[3]Franklin and Marshall College, Lancaster

Abstract

Buddhist Classical Chinese is a challenging low-resource language that has not yet received much dedicated attention in NLP research. Standard commercial machine translation software performs poorly on this idiom. In order to address this gap, we present a novel dataset of 209,454 bitext pairs for the training and 2.300 manually curated and corrected bitext pairs for the evaluation of machine translation models. We finetune a number of encoder-decoder models on this dataset and compare their performance against commercial models. We show that our best fine-tuned model outperforms the currently available commercial solutions by a considerable margin while being much more cost-efficient and faster in deployment. This is especially important for digital humanities, where large amounts of data need to be processed efficiently for corpus-level operations such as topic modeling or semantic search. We also show that the commercial chat system GPT4 is surprisingly strong on this task, at times reaching comparable performance to our finetuned model and clearly outperforming standard machine translation providers. We provide a limited case study where we examine the performance of selected different machine translation models on a number of Buddhist Chinese passages in order to demonstrate what level of quality these models reach at the moment.

1 Introduction

Regarding the languages of the Buddhist tradition, there is a striking gap between the amount of available material in their ancient source languages Pāli, Sanskrit, Buddhist Classical Chinese, and Tibetan, and the number of available translations into Western languages, leaving the majority of texts inaccessible to a wider audience. In the case of Buddhist Chinese, only about 10% of the digitally available material has been translated into Western languages over the last two centuries. Machine translation (MT) therefore holds a great promise to help in the process of translating these texts, which is proceeding at a slow pace so far. Translators of Buddhist Chinese texts into Western languages generally do not work with MT tools yet, as their performance is rather poor. For common tasks in digital humanities such as topic modeling or semantic comparison, MT models also hold great promise as they make it possible to apply models trained primarily on English to this material. On the technical level, recent years have brought substantial advances in the performance of data-efficient MT systems, and their high level of reliability for language directions with good resources such as French or German to English has led to their wide adoption. In the case of low-resource languages with only limited data resources, the training of stable and usable MT systems remains difficult. This paper deals with the challenging situation of Buddhist Classical Chinese, which has been generally neglected in the compilation of openly available large parallel datasets such as OPUS[1] or the training of multilingual MT models such as NLLB (Team et al., 2022). We make the following contributions to this problem:

1. The first description of a dedicated Buddhist Classical Chinese to English parallel dataset with a total number of 209,454 bitext pairs covering a big variety of genres, including a manually curated and post-corrected evaluation dataset.

2. Two augmentation strategies using domain-specific feature engineering to

[1]https://opus.nlpl.eu/

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 266–277

increase the performance of encoder-decoder models on this task.

3. Fine-tuning and evaluation of a variety of different openly available MT models as well as commercial providers on this dataset, giving the first thorough assessment of the quality of different currently available solutions on this language. We make the best-performing fine-tuned model available publicly.

4. Analysis of the behavior of these translation systems on different domains of Buddhist Chinese using standard MT metrics, followed by a more careful analysis using in-domain knowledge of Buddhist Classical Chinese.

In section 2 we give an overview of the relevant literature. Section 3 describes the datasets as well as the two augmentation strategies. In section 4 we show the evaluation results of the different models. In section 5 we examine the performance on different domains and conduct the case study.

1.1 Buddhist Chinese

For the purpose of this paper, we take "Buddhist Chinese" to be the language of pre-modern Chinese Buddhist texts, both those translated from Indian originals and Buddhist texts composed directly in Chinese. Buddhist Chinese is a subset of Classical Chinese characterized morphologically by its frequent use of polysyllabic terms (usually translations of Indian words). Syntactically, translated texts in Buddhist Chinese often retain traces of Indian syntax and other grammatical features. Structurally, some of the texts mix prose and verse in a way that is characteristic for Indian literature, but highly unusual for non-Buddhist Classical Chinese (at least in the first millennium). In addition, compared to Classical Chinese in general, Buddhist Classical Chinese often preserves vernacular elements and has been used to study the development of Middle Chinese in the first millennium (Anderl, 2017). We estimate that there are between 8,000 and 10,000 extant Buddhist Chinese texts of which about 6000 have been digitized as full text. These texts were translated or authored by Chinese, Indian, Korean, and Japanese writers between the second and the nineteenth century. As the holy scriptures of Chinese, Korean, Japanese, and Vietnamese Buddhists, these texts are a highly significant part of the East Asian cultural heritage and are still widely used and studied.

Translation from Buddhist Chinese into various European languages began slowly in the 19th century. Currently, the largest bibliography (Bingenheimer, 2023) lists 1452 translations of 650 texts. I.e. over the course of some 200 years c. 10% of the digitally available published corpus has been translated into Western languages. The picture looks very different for Korean and Japanese. In Japanese, there is a translation of the Taishō canon in 355 vols. (Kokuyaku issaikyō 国訳一切経, 1930-1988) and a number of other large translation collections. A full translation into modern Korean exists of an influential 13th-century edition of the Buddhist canon (Dongguk yŏkkyong wŏn 東國譯經院, 1964-2001), which has been fully digitized[2].

2 Related Work

Along with natural language processing in general, the field of neural MT was revolutionized by the introduction of the transformer architecture in 2017 (Vaswani et al., 2017). These networks introduced a way of processing language that emphasized determining which parts of a sentence should be *attended to*. By processing a number of tokens simultaneously (rather than sequentially), transformers are capable of learning relations between more distant parts of a sentence than had been possible in widely used models like LSTMs. The resulting revolution in NLP is ongoing, with current iterations of models like OpenAI's ChatGPT hardly needing any introduction.

Transformers come in 3 broad flavors: encoder-only, decoder-only, and encoder-decoder. Decoder-only models such as GPT are trained to predict the next token in a partially completed sequence, while encoder-decoder models learn to encode inputs and then decode them appropriately.

In this paper, we will use decoder-only and encoder-decoder variants, with mBART50 (Tang et al., 2021), WMT21 (Tran et al., 2021)

[2] https://kabc.dongguk.edu

and NLLB (Team et al., 2022) being encoder-decoder models while LlaMA2 is decoder-only.

The first of these, mBART50, utilizes monolingual pretraining with a denoising objective. NLLB on the other hand is trained directly on a massive multilingual parallel dataset without a pretraining stage. The WMT21 model we use consists of dense English to many and many to English models for translating between various languages (including Chinese). The open-source LlaMA2 (Touvron et al., 2023) is a decoder-only models which is pretrained on massive monolingual datasets consisting primarily of English data.

We decided to work with mBART and NLLB since both models have shown significant jumps in performance on low-resource languages when compared to randomly initialized transformer configurations. We will not use mT5 (Xue et al., 2021), which follows a similar pretraining objective as mBART, in our evaluation scheme since mBART has shown to be of equal or slightly better performance on low-resource translation tasks (Lee et al., 2022). The first and to our knowledge only dedicated publication on the problem of Buddhist Classical Chinese MT is (Li et al., 2022). Unfortunately, they have not made their models or evaluation data available, making it impossible to compare their findings in this paper.

3 Dataset

We collect a total number of 209,454 bitext pairs/5,738,025 characters from a variety of texts of the Taishō canon that have been translated into English. Due to the mechanics of the tokenizers of common language models, a single Chinese character is roughly equal to one token. The detailed genre distribution is given at table **??**. The composition dates of the Chinese texts in this dataset range from about ca. 150 to 1600 CE, while the English translations have been composed between ca. 1900 and 2020 CE. The training dataset covers a wide variety of different Buddhist Chinese domains: early and Mahāyāna sūtras, canon law, philosophical treatises, commentaries, ritual texts, etc. About half of the texts are translations of Indic Buddhist sources, others were composed directly in Buddhist Classical Chinese. Due to the diachronic spread and the variety of genres, the language of the Chinese Buddhist corpus is quite varied. Like there was no standard glossary to translate Indian terms into Chinese, English translations from Chinese were never standardized. Thus for any one Indic Buddhist term, we usually have a variety of different renderings in Chinese and English (Sanskrit āyatana, Buddhist Chinese: 處, 界, 入, English: sphere, field, sense organ, sense object, stage, level, base of cognition, sense sphere etc.). These circumstances do not only create challenges for the training of a MT system but also for the automatic evaluation of their performance, as in many cases, multiple different translations for a given Buddhist Chinese term are valid.

Category	Translated	Total	(%)
T01-0151 Āgama	516.570	2.861.382	18.1
T0152-0219 Past Lives	280.605	2.695.950	10.4
T0220-0261 Perfection of Wisdom	39.477	6.896.505	0.6
T0262-0277 Lotus Sūtra	133.955	587.509	22.8
T0278-0309 Flower Garland	1.052.629	2.262.554	46.5
T0310-0373 Treasure Trove	50.599	2.070.942	2.4
T0374-0396 Great Final Nirvāṇa	126.955	1.207.887	10.5
T0397-0424 Great Collection	37.109	1.566.096	2.4
T0425-0847 Sūtra	173.345	5.459.878	3.2
T0848-1420 Tantra	197.859	5.058.930	3.9
T1421-1504 Vinaya	122.236	5.264.857	2.3
T1505-1535 Sūtra Commentaries	352.091	2.046.519	17.2
T1536-1563 Abhidharma	503.450	5.125.379	9.8
T1564-1578 Madhyamika	26.720	417.713	6.4
T1579-1627 Yogācāra	331.903	2.506.840	13.2
T1628-1692 Collection of Treatises	257.920	1.202.910	21.4
T1693-1803 Chinese Commentaries	10.351	11.612.367	0.1
T1851-2025 Chinese Sectarian Writings	153.942	7.357.158	2.1
T2026-2120 History/Biography	599.823	6.235.952	9.6
T2121-2136 Encyclopedias/Dictionaries	1.350.386	2.255.979	59.9

Table 1: Distribution of the training data in the dataset according to different categories. "Translated" indicates the number of translated characters, "Total" indicates the total number of characters in the given Taishō section. The last column indicates how much of a section is available in translation. We only give Taishō sections that actually have translations into English.

As a first step, the English translations have been digitized and optical character recognition has been applied when necessary in order to obtain machine-readable unicode text. The translations have then been aligned with their Chinese counterparts as contained in the Chinese Buddhist Electronic Text Association (CBETA) corpus.[3] Many CBETA texts are based on an early 20th-century canonical edition, the "Taishō Canon". The texts have been thoroughly proofed against their original print editions. In some cases, punctuation has been added which increases

[3] http://cbeta.org/

intelligibility for humans (at least). Since on average the language models that we are evaluating have been exposed to more training data in simplified CJKV characters than in traditional CJKV, we convert the characters in the dataset to simplified characters during preprocessing.

The alignment was performed with vecalign (Thompson and Koehn, 2019). The embedding model used for the alignment process is a modified version of the multilingual sentence embedding model LaBSE (Feng et al., 2022). This model was further finetuned on a small corpus of gold-quality Buddhist Chinese to English bitext pairs. In order to reduce the influence of misaligned sentences, we use a rule-based scheme to remove sentences where the aligned English section is either much shorter or much longer than the Chinese counterpart. We also exclude samples from the training process where the Chinese part is shorter than four characters, assuming that NMT models do not learn well from very short samples.

We manually curated an evaluation dataset with a total size of 2,300 bitexts from a variety of texts from different genres of the Buddhist Chinese canon. Passages of the following texts have been included: T0026 (447 bitexts), T0374 (518 bitexts), T0475 (185 bitexts), T1585 (307 bitexts), T1600 (784 bitexts), T1970 (234 bitexts), T2062 (66 bitexts). The alignment of the evaluation dataset was performed by vecalign and then manually post-corrected. Training and evaluation data can be made available on request.

3.1 Data Augmentation

One commonly used strategy to improve MT performance with low-resource languages is the generation of synthetic training data that can be used to augment the original dataset during the training process. One synthetic augmentation technique for NMT systems is backtranslation (Sennrich et al., 2016), in which a large corpus in the target language is translated into the source language with the help of a MT system, creating a dataset where one side is automatically produced, and this data is then included in the training of the model. In our case, backtranslation has not proven to be helpful. While a lot of English data in the target language exists that could potentially be used, this data is not in the desired target domain, and utilizing this data results in significantly lower performance.

We therefore propose two different strategies for augmentation of the Buddhist Classical Chinese dataset:

1. Creation of a synthetic Classical Chinese to English dataset by machine-translating the Modern Chinese sentences of the NiuTrans Classical Chinese to Modern Chinese dataset.[4] we use the multi-lingual Transformer model of the Meta-AI research group submitted to the WMT2021 shared task, wmt21-dense-24-wide-x-en with 4.7billion parameters (Tran et al., 2021), to translate the Modern Chinese into English. We decided to use this model as among the openly available translation models, this has shown the best performance for Modern Chinese to English translation for this domain. This generates a total number of 972,470 bitext pairs. We make this dataset available at https://github.com/dharmamitra/NiuTrans-Classical-Modern-English.

2. Prompting ChatGPT3.5 with Buddhist Chinese paragraphs together with their translation into Modern Korean and dictionary entries in order to create a synthetic Buddhist Chinese to English dataset.

ChatGPT3.5 is a large language model created by OpenAI. Its most recent, more expensive version is GPT4. We utilize the digitally available complete translation of the Chinese Buddhist canon into Korean[5] in order to train a mBART (Lewis et al., 2020) Buddhist Chinese to Korean translation model. Then, to create additional data, we take random pseudo-paragraphs of up to 200 characters in length together with their Korean translation obtained via the mBART model and feed them to ChatGPT3.5, prompting it to translate the pseudo-paragraph into English, making use of the Korean translation. We also augment the prompt with dictionary entries obtained via the Digital Dictionary of Buddhism[6] (Muller, 2019) to ensure better translation of specific Bud-

[4] https://github.com/NiuTrans/Classical-Modern

[5] https://kabc.dongguk.edu/

[6] http://www.buddhism-dict.net/

dhist terminology. In order to avoid overgeneration of possible entries, we limit the retrieval to entries that are three characters or longer. We prompt ChatGPT to output the English translation together with the Chinese source sentences in a sentence-aligned format, thus generating as many sentence pairs as are needed to meet our desired augmentation target. In this way, we generate a total number of 436,945 synthetic Buddhist Chinese to English sentence pairs. We make this dataset available at https://github.com/dharmamitra/buddhist-chinese-agumentation.

4 Experiments

We evaluate the following models: Bing Translator[7], DeepL[8] and Google Translate[9] are commercial translation engines. We test ChatGPT3.5 and GPT4 are the commercial chat systems provided by OpenAI. We query the OpenAI models with a simple prompt: "*Translate the following Buddhist Chinese passage into English: <sentence> English:*". Transformer 600M serves as the baseline for the finetuned models, which is the NLLB600M model with randomly initialized weights, simulating the training of a transformer model with 600M parameters from scratch. mBART50 and mBART50-to-1 are two different versions of the multilingual BART model with a size of 611M parameters (Tang et al., 2021). Both are pretrained on a denoising task, while the latter is the many–to-one version that is finetuned on a many-to-one translation task, including Chinese, with English as the target language. No further information is provided during the prompting step. NLLB600M-3.3B is the massive multilingual model of Meta AI in different sizes, trained among other languages also on Chinese to English. WMT21 is the Meta AI's submission to WMT21 News Translation task (Tran et al., 2021). We use the wmt21-dense-24-wide-x-en version with 4.7B parameters, which was also trained on the Chinese to English task.

We fine-tuned the 7B parameter version of LlaMA2 available on HuggingFace, using QLoRA for fintetuning on the full parallel dataset. During training and inference, we used the prompt *"Below is some text in Classical Chinese. It is taken from the Buddhist literature. Write a translation of the text into English."* followed by labelled Chinese inputs and a label for the English translation.

[7] https://www.bing.com/translator
[8] https://www.deepl.com/translator
[9] https://translate.google.com/

Model	BLEU	chrF++
Bing Translator sent	4.1	25.5
Bing Translator par	4.4	27.9
Deepl sent	7.6	30.1
DeepL par	8.2	33.0
Google Translate sent	8.5	31.8
Google Translate par	8.9	35.1
ChatGPT3.5 sent	9.5	35.0
ChatGPT3.5 par	11.2	38.8
GPT4 sent	11.8	37.4
GPT4 par	12.8	40.3
Transformer 600M sent	4.4	31.0
Transformer 600M par	7.7	35.8
mBART50 sent-ft	11.2	35.4
mBART50 par-ft	11.6	37.5
mBART50-to-1 sent	3.3	22.3
mBART50-to-1 sent-ft	13.0	37.8
mBART50-to-1 par-ft	12.9	39.4
NLLB600M sent	2.0	19.0
NLLB600M sent-ft	12.6	37.0
NLLB600M par-ft	13.3	39.6
NLLB1.3B sent-ft	13.5	38.4
NLLB1.3B par-ft	13.8	40.0
NLLB3B sent-ft	14.6	39.5
NLLB3B par-ft	14.4	40.9
WMT21 sent	5.1	26.1
WMT21 sent-ft	14.2	38.7
WMT21 par-ft	14.4	41.0
WMT21+aug sent-ft	**15.2**	**39.9**
WMT21+aug par-ft	**15.1**	**41.7**
LlaMA2-ft sent	8.6	31.8
LlaMA2-ft par	8.8	32.8

Table 2: Main results on the MT task. Models finetuned with our parallel data are indicated with ft. Sent indicates evaluation on sentence-level, par indicates evaluation on paragraph-level.

We finetune all encoder-decoder models on sentence-level and on pseudo-paragraph-level as we assume that a larger context might help the models to arrive at better translation solutions. For the pseudeo-paragraph level, we concatenate adjecent sentences with a total length of up to 200 tokens. We decided on this number as the encoder-decoder model with the shortest context length, WMT21, only supports up to 200 tokens. We did a thorough hyperparameter search on a fixed holdout set to determine the optimal learning rate and number of training steps for each model.

4.1 Evaluation

We present the results in table 2. We evaluate using two different metrics: BLEU (Papineni et al., 2002) which uses word-level n-grams and chrF++ (Popović, 2017), which works with character-level n-grams. Since English translations of Buddhist Classical Chinese works frequently use borrowed terms from Sanskrit where different writing conventions might be applied, chrF++ seems a more appropriate choice as it considers similarity on character-level, and not just on word-level as is the case with BLEU. We do not use model-based metrics such as COMET or BERTscore as they have not been finetuned on the Buddhist domain and we can therefore not assume that they are appropriate for this scenario.

Regarding the commercial providers Bing, DeepL and Google Translate, their results are clearly inferior to those of ChatGPT3.5 and GPT4. The weak score of Bing Translate is especially remarkable in light of the fact that it was marketed to explicitly support Literary Chinese.[10] GPT4 in turn performs better than ChatGPT3.5 with a clear margin. All commercial systems perform better when the data is provided on pseudo-paragraph level instead of sentence level. The baseline model Transformer 600M struggles to reach usable performance. Also the openly available models that have been trained on the Chinese-to-English MT objective, mBART50-to-1, NLLB, and WMT21, perform badly without finetuning, being clearly inferior even when compared to DeepL and Google Translate. After finetuning on our dataset, they show a significant performance boost and clearly outperform the baseline Transformer 600M, showing that denoising pretraining in the case of mBART50 and transfer learning from Modern Chinese to English in cases of the other models is beneficial for this specific task. This is also confirmed by the fact that finetuned mBART50-to-1 performs better than finetuned mBART50 by 2.4 chrF++ score on sentence and 1.9 on paragraph level, which further proves that a model pretrained on the denoising objective benefits from further finetuning on the Modern Chinese to English translation task before being finetuned on our dataset. The fact that the NLLB models all perform bet-

[10] https://www.microsoft.com/en-us/translator/blog/2021/08/25/microsoft-translator-releases-literary-chinese-translation/

ter than mBART50 further supports the observation that transfer learning from Modern CHinese to English is helping. In the NLLB family, we see a clear improvement of performance with increasing model size. Noteworthy is the fact that while the smallest model NLLB600M benefits significantly from pseudo-paragraph-level training with an increase in BLEU of 0.7 and in CHRF of 2.6, the performance of the 3B version is not better in terms of BLEU, while better in terms of CHRF with an increase of 1.4. It is therefore safe to conclude that the increase of model performance by pseudo-paragraph level training decreases with model size for the NLLB family. The largest model that we finetune, WMT21 with 4.7B parameters, shows the best zero-shot performance of all openly available models. The finetuned version of this model shows almost identical performance with NLLB3B-ft on pseudo-paragraph level while being slightly inferior on sentence level. When we add the augmentation data to this model, we see a visible improvement of 1.2 chrF++ score on sentence level and 0.7 chrF score on the pseudo-paragraph level, leading to the highest performance of all models evaluated in this paper. Since the training with the augmentation data is very resource- and time-consuming, we could not evaluate its effects on the behavior of the other models. LlaMA2 does not yet competitive performance after finetuning on the dataset as it performs poorer than the finetuned mBART50 andNLLB600M models.

5 Analysis

Table 3 shows the performance of ChatGPT3.5, GPT4, and WMT21+aug-ft on different evaluation texts measured in BLEU and chrF++ on pseudo-paragraph level. WMT21+aug-ft outperforms the other models on T0026, T0374, T1585, and T1970. For T0026 and T1585, the difference to the second best-performing model GPT4 is significant. In the case of those texts where GPT4 performs better than WMT21+aug-ft, the difference is generally small, with T0475 being the only exception. ChatGPT3.5 performs worse than GPT4 on all texts, and, again with the exception of T0475, ChatGPT3.5 also performs worse than WMT21+aug-ft on all texts. The reason for GPTx output being comparatively strong on T0475-par might be because the text, the Vimalakīrtinirdeśa, is available online in a number of different versions, while most of the other texts in the evaluation set have only been translated from Chinese to English only once so far. It is known that GPTx is trained on large amounts of online data and therefore, memorization of the evaluation data is a possibility here. Compared to T0475, translations of the other texts are rather more recent and not as readily available online.

Text	Model	BLEU	chrF++
T0026	ChatGPT3.5	12.0	39.1
	GPT4	13.9	40.8
	WMT21+aug ft	**17.8**	**43.4**
T0374	ChatGPT3.5	12.0	39.0
	GPT4	13.6	40.4
	WMT21+aug ft	**15.6**	**42.4**
T0475	ChatGPT3.5	12.8	39.6
	GPT4	**13.6**	**41.0**
	WMT21+aug ft	11.9	38.0
T1585	ChatGPT3.5	9.5	38.1
	GPT4	12.2	40.7
	WMT21+aug ft	**19.9**	**48.9**
T1600	ChatGPT3.5	10.9	40.0
	GPT4	**12.3**	**41.1**
	WMT21+aug ft	12.2	39.8
T1970	ChatGPT3.5	9.6	37.9
	GPT4	11.4	39.1
	WMT21+aug ft	**12.0**	**40.3**
T2026	ChatGPT3.5	9.6	38.0
	GPT4	**11.6**	**39.4**
	WMT21+aug ft	11.0	37.9

Table 3: Performance on individual texts of the evaluation dataset. All results are calculated on pseudo-paragraph level.

In order to understand the nature of the mistakes that the different translation models produce, we analyzed several passages manually. We give the full samples in the appendix. The first paragraph is taken from T1585, the Cheng weishi lun, a core text of Sino-Indian Yogācāra philosophy. On this text, WMT21+aug-ft shows a generally superior quality, producing less serious mistakes, which mirrors the BLEU score results on the individual texts. It is noteworthy that while all three models on average use the right vocabulary to translate the philosophical terms in this paragraph, ChatGPT3.5

struggles significantly and GPT4 struggles somewhat to interpret the dense syntax of the Chinese. WMT21+aug-ft is doing visibly better, but certain points of confusion remain, i.e. rendering 非無 (here: "not nonexistent") as "neither nonexistent nor existent", which is not correct.

For T0026, an early Buddhist sūtra text, the BLEU and chrF++ score does not well align with our manual evaluation. Although the metric indicates a clear advantage for WMT21+aug-ft, many passages are actually rendered more accurately in the GPT4 output. It is possible that the metric was influenced by the tendency of WMT21+aug-ft to use Sanskrit terms for their Chinese equivalents, something that is common practice in Buddhist translation. In our example, the five great rivers of India (Jambudvīpa) are mentioned and while the GPTx models render 恒伽，搖尤那，舍牢浮，阿夷羅婆提，摩企 with at times misleading pinyin transcriptions (Hengqie/Hengqia, Shalao Fu/Sheloufu etc.), WMT has "Ganges, the Yamunā, the Śrāvastī, the Ajiravatī, and the Mahī". Śrāvastī is a mistake for Sarabhū here, but one can see how the Sanskrit terms in the output might influence the n-gram based BLEU and chrF++ scores, which compare it to the human reference translation that has similar terms.

T2062 is an early 17th-century biography of a Chinese monk. Next to T1970 (a 12th century Pure Land treatise) it is the text in our sample for which the linguistic markers of "Buddhist Chinese" are least evident. It is thus not that surprising that our domain-specific model does not produce significant differences to GPT4 for those two texts. The language of T2062 differs from that of the other evaluation texts in that it contains many named entities, esp. person and place name, which are often referenced in an abbreviated way. The syntax is exceedingly terse with almost no redundancy or repetition. Overall all models performed worst on T2062, with many passages translated wrongly to a degree that post-editing means retranslation. The example in the appendix is atypical in that it compares a relatively "easy" passage, which all models have managed to render reasonably well. As we have seen with T0026, WMT21+aug-ft tries to identify Sanskrit terms and render them as such (Jambudvīpa), but in this case unsuccessfully (ch. 荼毗, skr. kṣapita). An interesting passage that shows how the context understanding of ChatGPT3.5 is inferior to GPT4 and WMT21+aug-ft is 所聞種種，隨力不同 ("[all people] smelled something different, according to their powers [of insight]"). Although in itself its choices are reasonable, ChatGPT3.5 misses the subject with "Various sounds and scents were heard, depending on the strength of the fire." Note how the ambivalence of 聞 throws the model off. 聞 can indeed mean "to hear" or "to smell", but not both at the same time, in English "sounds and scents were heard" is nonsense.

Compared with the two large proprietary GPTx models, the domain-specific WMT21+aug-ft model shows at least approximately equal, and often better, BLEU and chrF++ scores. It needs to be remembered that inference on commercial GPTx models costs orders of magnitude more than the fine-tuned WMT21 model, which can be efficiently served even on a single consumer-grade GPU (Peng et al., 2023). Another problem when interacting with commercial models is the fact that their performance has been shown to differ significantly even in a relatively short amount of time, making their behavior unpredictable (Chen et al., 2023).

6 Conclusion

For this paper we compiled a novel dataset for the training and evaluation of MT models for Buddhist Classical Chinese. We applied two methods of data augmentation and compared a number of different encoder-decoder models finetuned on this data against large commercial MT providers and commercial decoder-only chat models. The domain-specific evaluation as well as the BLEU/chrF++ scores show that with the help of the augmentation strategies, our much smaller and locally run model WMT21+aug-ft clearly outperforms the standard commercial providers as well as the commercial chat system ChatGPT3.5, while being on par with GPT4 and outperforming it on certain

domains. Significantly, in the case of texts in the Chinese canon that are originally translated from Indic sources, our finetuned model outperforms GPT4 and is therefore the currently best available solution. For Chinese-Chinese Buddhist texts, the performance of GPT4 is comparable to our models. This makes the finetuned model an ideal solution for semantic similarity tasks on corpus level, which are of central concern within the digital humanities and require cost-efficient and fast processing of large quantities of data.

The evaluation results show that WMT21+aug-ft as well as GPT4 have reached a level of maturity that for the first time in the history of the translation of Buddhist Chinese texts into Western languages, these tools can be of genuine help to translators.

We see a number of directions for further work:

First, the amount of digitized English translations aligned with their Buddhist Chinese counterparts is still very limited. Increasing the size of this dataset promises further improvements in performance especially when it comes to Chinese-Chinese material, for which there is less bitext in the current dataset.

Second, while we present two strategies for data augmentation in this paper that are clearly boosting the performance of encoder-decoder models, further refinements of these approaches, especially the prompting of commercial engines with the right prior data, promise further significant leaps in performance.

Third, while LlaMA2 has not shown competitive performance in our evaluation, the preliminary results are encouraging. By utilizing more monolingual data during the finetuning stage, we might see significant performance increases for smaller, open decoder-only models as well. Fourth, our limited manual examination of the output of the MT models has indicated that widely used evaluation methods such as BLEU or chrF++ do not always align well with human judgment, an observation that was made in other recent studies with a focus on large language models on MT as well (Wang et al., 2023). We therefore see a clear need for a thorough examination of alternative evaluation methods in future studies.

References

Christoph Anderl. 2017. Medieval Chinese Syntax, volume 2, pages 689–703. Brill.

Marcus Bingenheimer. 2023. Bibliography of translations (by human translators) from the chinese buddhist canon into western languages.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time?

Dongguk yŏkkyong wŏn 東國譯經院. 1964-2001. Hangul daejang kyong.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In Findings of the Association for Computational Linguistics: ACL 2022, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

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.

Denghao Li, Yuqiao Zeng, Jianzong Wang, Lingwei Kong, Zhangcheng Huang, Ning Cheng, Xiaoyang Qu, and Jing Xiao. 2022. Blur the linguistic boundary: Interpreting chinese buddhist sutra in english via neural machine translation. In 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), pages 228–232.

A. Charles Muller. 2019. The Digital Dictionary of Buddhism and CJKV-English Dictionary: A Brief History. In Digital Humanities and Buddhism: An Introduction, SISAP 2013, A Coruña, Spain, October 2-4, 2013, Proceedings, pages 143–156.

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.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation.

Maja Popović. 2017. chrF++: words helping character n-grams. In Proceedings of the Second Conference on Machine Translation, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai' s wmt21 news translation task submission. In Proc. of WMT.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. arXiv preprint arXiv:2304.02210.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

A Appendices

A.1 Case Study Full Examples

Double underline: serious mistakes. Single underline: less serious mistakes.

T1585: 復有迷謬唯識理者，或執外境如識非無、或執內識如境非有、或執諸識用別體同、或執離心無別心所

Human Reference: Some are deluded about this principle of consciousness only. Some grasp the external world as not nonexistent, like consciousness. Some grasp internal

consciousness as nonexistent, like the external world. Some grasp the various kinds of consciousness as being the same in substance but different in characteristics. Some believe that there are no separate mental functions apart from the mind.

ChatGPT3.5: There are those who hold mistaken views regarding the theory of the only consciousness. Some believe that external objects are not without consciousness, while others believe that internal consciousness is not the same as external objects. Some believe that various consciousnesses have different natures, while others believe that there is no separate mind apart from the mind.

GPT4: Furthermore, there are those who are confused and mistaken about the principles of consciousness-only, some clinging to the idea that external objects exist as consciousness and are not non-existent, some clinging to the idea that internal consciousness exists as objects and are not non-existent, some clinging to the idea that all consciousnesses function with the same substance, and some clinging to the idea that there is no separate mental object apart from the mind.

WMT21+aug-ft: Again, there are those who are deluded and mistaken about the principle of consciousness only. They may hold that external objects, like consciousness, are neither nonexistent nor existent, or they may hold that internal consciousness, like objects, is neither existent nor existent, or they may hold that the functions of the various forms of consciousness are different but the substance is the same, or they may hold that apart from mind there are no separate mental activities.

T0026: 是時，晃昱諸天始生天者，不諳世間成敗，不見世間成敗，不知世間成敗，見大火已，皆恐怖毛竪而作是念：『火不來至此耶？火不來至此耶？』前生諸天諳世間成敗，見世間成敗，知世間成敗，見大火已，慰勞諸天曰：『莫得恐怖，火法齊彼，終不至此。』

Human Reference: Then the gods of radiance, who have just been born as gods and have never experienced, seen, or known the world's [cyclic] creation and destruction, on seeing that great fire will all be terrified, with their hair standing on end, and will think: "The fire will not reach up to here, will it? The fire will not reach up to here, will it?" Those who were born as gods at an earlier time and have experienced, seen, and known the world's [cyclic] creation and destruction will, on seeing that great fire, calm the distress of those gods, telling them, "Don't be afraid! That fire is of such a nature that it will not reach up to here."

GPT-3.5: At that time, the heavenly beings who were born in the beginning of the world were ignorant of the success or failure of the world. They did not see the success or failure of the world, nor did they know the success or failure of the world. When they saw the great fire, they were all terrified and had the following thought: "Has the fire reached here? Has the fire reached here?" The heavenly beings from previous lives who were knowledgeable about the success or failure of the world, who saw the success or failure of the world, and who knew the success or failure of the world, comforted the other heavenly beings, saying: "Do not be afraid. The fire will not reach here. It will be extinguished before it reaches here."

GPT-4: At that time, the newly born gods, who were not familiar with the success and failure of the world, who did not see the success and failure of the world, and who did not know the success and failure of the world, saw the great fire and were all terrified, their hairs standing on end, and thought: "Will the fire come here? Will the fire come here?" The gods who had been born earlier, who were familiar with the success and failure of the world, who saw the success and failure of the world, and who knew the success and failure of the world, saw the great fire and comforted the other gods, saying: "Do not be afraid, the law of fire is equal to that, it will never come here."

WMT21+aug-ft: At that time, the Ābhāsvara gods who had just been reborn among the gods did not study the formation and decay of the world, did not see the formation and decay of the world, and did not know the formation and decay of the world. But

when they saw the great fire, they were all terrified, their hair standing on end, and they thought, ‘Will the fire not come here? Will the fire not come here? ’ The gods who had been reborn previously studied the formation and decay of the world, saw the formation and decay of the world, and knew the formation and decay of the world. But when they saw the great fire, they comforted the gods, saying, ‘Do not be afraid! The nature of fire is equal to that. It will never come here.’

T2062 Zhenqing: 荼毗日天色霽明，淨無纖翳，舉火之際，忽有片雲，如蓋凝覆其上，灑微雨數點。烟焰起時，異香充塞，內自殿閣僧房，外自路人船子，所聞種種，隨力不同。火餘骨有三色，而鏘鏘有聲。

Human reference translation: On the day of his cremation, the weather had cleared up, and not a cloud was in the sky. But the moment the fire was lighted, a little cloud appeared suddenly. It grew denser just above the cremation site, like a canopy, and gently sprinkled a few drops of rain. When the fire arose, an extraordinary scent filled the air, and everybody, those in the monastery halls and monks’ quarters, as well as those outside, the people on the roads and in their boats, all smelled something different, according to their powers [of insight]. The bones remaining in the ashes were of three colors and made a clear sound [when tapped].

ChatGPT3.5: On the day of cremation, the sky cleared up, pure and without any blemish. Just as the fire was being lit, suddenly a patch of clouds appeared, covering and sprinkling a few drops of light rain. As the smoke and flames rose, a strange fragrance filled the air, coming from the inner halls and monk’s quarters as well as from passersby and boats on the road. Various sounds and scents were heard, depending on the strength of the fire. After the fire subsided, the bones were left with three different colors and made a clanging sound.

GPT4: On the day of the cremation, the sky was clear and bright, without a trace of cloud. At the moment the fire was lit, a patch of cloud suddenly appeared, covering the area like a canopy, sprinkling a few drops of light rain. When the smoke and flames rose, a strange fragrance filled the air, permeating from the inner temple halls and monks’ quarters to the outer passersby and boats. Various reactions were heard, differing according to one’s perception. The remaining bones after the fire had three colors, and they made a clinking sound.

WMT21+aug-ft: The day of Jambudvīpa The sky was clear and clean, without the slightest cloudiness. While the fire was being raised, a cloud suddenly appeared, like a canopy, covering the fire and sprinkling a few drops of rain. When the smoke and flames arose, they were filled with an extraordinary fragrance. From inside the palace and the monks’ quarters, and from outside the boats of the passers-by, the fragrance varied according to one’ s ability to smell it. The embers left by the fire were of three colors, and they emitted a sound when burned.