Humanistic Buddhism Corpus: A Challenging Domain-Specific Dataset of English Translations for Classical and Modern Chinese

Youheng Wong, Natalie Parde, Erdem Koyuncu
University of Illinois Chicago
Chicago, IL 60607
{wwong31, parde, ekoyuncu}@uic.edu

Abstract

We introduce the Humanistic Buddhism Corpus (HBC), a dataset containing over 80,000 Chinese-English parallel phrases extracted and translated from publications in the domain of Buddhism. HBC is one of the largest free domain-specific datasets that is publicly available for research, containing text from both classical and modern Chinese. Moreover, since HBC originates from religious texts, many phrases in the dataset contain metaphors and symbolism, and are subject to multiple interpretations. Compared to existing machine translation datasets, HBC presents difficult unique challenges. In this paper, we describe HBC in detail. We evaluate HBC within a machine translation setting, validating its use by establishing performance benchmarks using a Transformer model with different transfer learning setups.

Keywords: Neural Machine Translation, Classical Chinese, Domain-Specific, Religion, Buddhism

1. Introduction

In recent years, Neural Machine Translation (NMT) approaches have become commonplace, especially after the introduction of the Transformer model (Vaswani et al., 2017). However, it is widely recognized that these methods tend to deliver less accurate translations when applied to data outside their specific domain (Koehn and Knowles, 2017). In medicine, law, or religion, all of which are characterized by technical jargon and limited parallel data available for training, universal machine translators provide inaccurate translations or miss out on stylistic nuances. Moreover, while remarkable results have been achieved for more general translation between similar languages (e.g., both source and target languages are Latin-derived) (Sun et al., 2021; Mueller et al., 2020), translation between distant languages, such as Chinese to English, still has plenty of room for improvement (Kim et al., 2020).

By introducing the Humanistic Buddhism Corpus (HBC), we attempt to alleviate these problems by offering a domain-specific dataset in a challenging field: religious documents for distant languages. In particular, the HBC includes modern and classical Chinese text concerning Humanistic Buddhism. Classical Chinese scriptural texts and Buddhist proverbs offer rich and challenging domain-specific terminology that increases the complexity of their automated translation. The main contributions of this paper include:

- We introduce a high-quality, free, and extensive domain-specific dataset comprising 80,000+ parallel Chinese-English phrases for NMT research. The corpus contains both classical and modern Chinese religious texts.
- We evaluate our corpus using various experimental setups. We observe that many pre-trained models, although achieving high BLEU or COMET scores, cannot always effectively capture the nuances associated with this challenging domain.
- We qualitatively show that a Transformer model trained using our corpus combined with ordinary Chinese text can provide more accurate and fluent translations despite achieving a quantitatively lower BLEU or COMET score than pre-trained models.
- We also provide benchmark code and models for training and testing the corpus to accelerate future research.

1.1. Buddhist Translation History

Buddhism has a remarkable and well-documented translation history from its original ancient Indian language to classical Chinese. Unlike other religions with only one primary scripture (i.e., the Bible or the Koran), a typical collection of Chinese Buddhist scriptures (Tripitaka or Canons) contains more than 70 million Chinese characters from over 5,000 texts. These scriptures are written in ancient Chinese styles that span over 2,000 years. They have various transliterations from their original Sanskrit or Pali text, and multiple translators

LREC-COLING 2024, pages 8406–8417
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

might translate the same Sanskrit or Pali terminology into different Chinese words. Furthermore, since the Chinese writing styles have also evolved over this long timespan, the writing styles for these Buddhist texts vary significantly, making translating Chinese to English even more challenging.

Since Buddhism propagated from India to China over 2,000 years ago, numerous Buddhist proverbs, concepts, and vocabulary have been integrated into Chinese literature throughout history, and many are still used today. Buddhist literature significantly influences Chinese culture (Guang, 2013). For example, the "Platform Sutra" is considered one of the 7 Chinese classics. Numerous new words were introduced into Chinese from Buddhism, and many hybrid terminologies also have Buddhist origin (Zhu, 2003). Therefore, when one translates Chinese texts into any language, encounters with Buddhist-influenced contexts are inevitable.

1.2. Corpus Source and Naming Convention

The HBC includes the most common Buddhist scriptures, known as sutras, which are discourses taught by Sakyamuni Buddha in India over 2,500 years ago. In addition, the corpus mainly features publications authored by one person, Venerable Master Hsing Yun. He was the founder of the Fo Guang Shan (FGS) International Buddhist Order and was an advocate of Humanistic Buddhism throughout his entire life. Humanistic Buddhism mainly focuses on applying complicated Buddhist teachings in everyday life. Therefore, even though some of his works collected poems and proverbs from ancient Chinese scholars, Hsing Yun composed his writings in easy-to-understand, everyday Chinese so everyone could comprehend them.

The *Complete Works of Venerable Master Hsing Yun* comprises 365 volumes in the first edition (Hsing Yun, 2017) and 395 volumes in the second edition. This collection is accessible online at `https://books.masterhsingyun.org/`, and most English translations of his work are also obtainable at various websites (BLP, 2023; FGSITC, 2020). With the massive collection of his writing in modern, everyday Chinese, his magnanimity in making his complete collection available online, and the extensive quantity of his works already translated into English, this work was an ideal starting point for the development of the HBC and thus inspired its name (the *Humanistic Buddhism Corpus*). We plan to expand and refine the HBC periodically in the years to come as more translations of Hsing Yun's work are published.

2. Related Works

2.1. Existing Datasets

A variety of Chinese-English translation datasets are currently available, although none fit the specific needs that HBC is designed to fill. The WMT news translation task has resulted in the publication of numerous Chinese-English parallel corpora for training and testing, especially in recent years (WMT21 and WMT22). However, these datasets require extensive filtering to be used for training (Wang et al., 2021; Tran et al., 2021; Zeng et al., 2021). Moreover, these extensive parallel data are intended for the news domain and perform poorly for out-of-domain texts (Koehn and Knowles, 2017).

The open source parallel corpus OPUS (Tiedemann, 2012) comprises a vast range of translated texts, including the datasets from the WMT that contain Chinese-English parallel data, such as Wiki Titles, UN Parallel Corpus, and TED talks. The OPUS collection also includes single-volume translated texts in multiple languages, such as the Bible-uedin corpus (Christodouloupoulos and Steedman, 2015) and the Tanzil datasets (Tiedemann, 2012), which have Chinese-English parallel data. However, most of the other collections within OPUS are in European languages and do not contain Chinese-English parallel texts.

The Linguistic Data Consortium (LDC) hosts many datasets for NMT. For example, the Chinese-English NIST datasets (Przybocki et al., 2009) available on LDC are frequently used by translation researchers. However, their datasets require membership and thus become expensive for non-member researchers to use them.

2.2. Available Buddhist Corpora

In the past decades, efforts from different schools of Buddhism have sought to make Buddhist scriptures available online. The Chinese efforts include the Chinese Buddhist Electronic Text Association (CBETA, 1998), digitizing the Buddhist canon since 1998. Fo Guang Shan started its Buddhist Electronic Texts (Fo Guang Shan, 2004) in 1995 and made them accessible online in 2004. The Sanskrit cannon (University of the West, 2003) and Pali texts (Vipassana Research Institute, 2010) are also available online, while the Tibetan texts are preserved in graphical format by the Buddhist Digital Resource Center (BDRC, 2016). The English translations for these documents are freely obtainable at various sites (BuddhaNet, 2023; BCBS, 1995). However, all of these resources are only available in one language—that is, the English translations are not directly mapped to the Chinese editions of a given document, and the same holds

true for other potential language pairings. The creators of these digital resources are incredibly supportive of making parallel corpora, but the resources as they currently stand are not usable in training neural machine translation models. Research on machine translation of Buddhist texts has been attempted, such as by Li et al. (2022). Unfortunately, the datasets they used are proprietary.

2.3. Other Religious Corpora

The Bible is the most extensively translated religious scripture in the world. The eBible includes the Bible text in 833 different languages across 75 language families (Akerman et al., 2023). The Bible-uedin is a collection of the translation of the Bible in 100 languages (Christodouloupoulos and Steedman, 2015). Another religious text, the Koran, is also widely translated. The Tanzil (Tiedemann, 2012) is a collection of Koran translations in 42 languages. Both the Bible and the Koran have Chinese-English parallel data readily available online. However, both of these religious scriptures are minuscule in size compared to the voluminous Buddhist collection of over 5,000 scriptures.

3. Humanistic Buddhism Corpus

The HBC contains 81,000 Chinese-English parallel phrases extracted from Buddhist publications. These publications include classical and modern Chinese, comprising translation of modern Chinese from 20 volumes of English books and 34 booklets, Buddhist scriptural text in classical Chinese from 9 sutras, proverbs and poems from 26 volumes of books, and subtitles in modern language from 7 DVDs. Table 1 provides the book titles in the HBC. Aside from the Buddhist scriptural texts known as sutras, most of the original Chinese texts are from the 365 volumes of the *Complete Works of Venerable Master Hsing Yun* (Hsing Yun, 2017). However, the "One-Liners," "Phrases," and "Poetry" from his collection, as well as *After Many Autumns* (Hsing Yun, 2011a), comprises phrases and poetry by over 100 ancient Buddhist scholars from the Tang dynasty (618-907) to the Qing dynasty (1644-1911). These scholars include the renowned Wang Xizhi, Du Fu, Bai Juyi, Pei Xiu, Su Shi, and Ouyang Xiu, emperors such as Shunzhi and Yongzheng, and numerous eminent Buddhist masters. Organizations within the Fo Guang Shan order publish the English translations. The subtitles are from Hsing Yun's oral lectures produced into DVDs by Beautiful Life Television.

The original texts are in Mandarin Chinese with Traditional Script (zh-Hant), and the translations are in American English (en-US). The average length of the phrases in the HBC is 18 Chinese characters, with 15% of the phrases being proverbs or lines from poems and verses, which are 4 to 7 Chinese characters. The format of the HBC contains lists of Chinese and English pairs, with an additional reference ID as the third element. The numbering for the reference ID is 0 to 99,999 for modern texts; classical Chinese phrases are numbered 100,000 to 199,999 for sutras, and poems and proverbs start at 200,000. Future researchers can use these reference numbers to filter the data.

3.1. Data Selection Techniques

Unlike online translations, which may be unedited or translated by amateurs, published works typically contain high-quality translations due to their thorough editing before publication. However, these high-quality translations are edited based on the fluency and adequacy of the content as a whole, which is often based on the surrounding paragraph, and the editors do not emphasize the accuracy of each sentence translation. Therefore, some original Chinese sentences are omitted during Chinese-English translation while other English sentences are added to make the contents more comprehensible for English readers who lack a Chinese cultural background. Additionally, it is common for Chinese sentences (especially those describing difficult-to-comprehend religious concepts) to be extremely long and separated by commas, such that one sentence could reasonably be translated into several English sentences. As a result, English translations of Chinese publications contain a high percentage of unmatched phrases on average (see the Exc. column in Table 1). The order of translations is also often different from that of their original works, increasing the difficulty of extracting parallel phrases using MT models.

Due to the difficulty of using automated methods for this process, we built the Humanistic Buddhism Corpus by manually matching, aligning, and scoring the Chinese-English phrases as part of a community effort collectively spanning 21 volunteer human translators over the past two years. The volunteers are all Buddhists and have received training regarding alignment and rating. All but one of the volunteers are native Chinese speakers, and most of them have at least 5 years of translation experience. The majority of matched phrases are complete English sentences matched to phrases in Chinese. In the case of dialogue and other complicated English sentences, the complete English sentences are split further into meaningful units. The poems and verses are divided on a line level.

The matched parallel phrases underwent manual review by an experienced human translator

#	Title	Vol.	Line	Char.	Exc.	Test
	(A) Sutras					
1	Amitabha Sutra (Sakyamuni Buddha, 2017)	1	91	2025	0.0%	No
2	Diamond Sutra (Sakyamuni Buddha, 2019a)	1	351	6194	0.0%	No
3	Repaying Parents Sutra (Sakyamuni Buddha, 2018)	1	270	3864	0.0%	No
4	Give Rise to the Bodhi Mind (Sheng'an, 2017)	1	652	8333	0.0%	No
5	Ksitigarbha Bodhisattva Sutra (Sakyamuni Buddha, 2023)	3	1010	20202	0.0%	No
6	Medicine Buddha Sutra (Sakyamuni Buddha, 2015)	1	293	6084	0.0%	No
7	Platform Sutra (Hsing Yun, 2011c)	10	2173	23981	0.0%	No
8	Samantabhadra Bodhisattva Ch. (Sakyamuni Buddha, 2019b)	1	453	6068	0.0%	No
9	Universal Gate Chapter (Hsing Yun, 2011c)	1	212	2429	0.0%	No
	(B) Books					
1	After Many Autumns (Hsing Yun, 2011a)	1	2604	22072	23.8%	No
2	Bells, Gongs, and Wooden Fish (Hsing Yun, 2012a)	1	1386	29878	14.6%	Yes
3	Biography of Sakyamuni Buddha (Hsing Yun, 2013)	1	6112	129331	10.4%	Yes
4	Bright Star (Fu, 2008)	1	5443	139403	18.3%	Yes
5	Buddha's Light Philosophy (Hsing Yun, 2010a)	1	2285	38702	21.9%	Yes
6	Buddha Dharma (Hsing Yun, 2019)	4	6023	126896	4.8%	Yes
7	Collection of VMHY - One-Liners (Hsing Yun, 2017)	6	1125	3936	7.0%	No
8	Collection of VMHY - Phrases (Hsing Yun, 2017)	6	1999	7996	7.1%	No
9	Collection of VMHY - Poetry (Hsing Yun, 2017)	1	561	3758	0.9%	No
10	Collection of VMHY - Wisdom 1, 3, 6, 7 (Hsing Yun, 2017)	4	2664	20736	13.4%	No
11	Epoch of Buddha's Light (Hsing Yun, 1999)	1	2102	37752	38.1%	Yes
12	Everlasting Light (Hsing Yun, 2002)	6	1570	19535	21.7%	No
13	For All Living Beings (Hsing Yun, 2010b)	1	1903	39320	32.2%	Yes
14	Four Insights (Hsing Yun, 2010d)	1	3135	61888	7.6%	Yes
15	Handing Down the Light (Fu, 1996)	1	2946	68555	14.6%	Yes
16	HB: Blueprint for Life (Hsing Yun, 2008)	1	2350	46225	6.2%	Yes
17	Humble Table, Wise Fare (Hsing Yun, 2011b)	2	478	5371	7.4%	No
18	Infinite Compassion (Hsing Yun, 2010c)	1	1161	20923	20.7%	Yes
19	Living Affinity (Hsing Yun, 2009)	1	582	12115	70.6%	Yes
20	Pearl of Wisdom (Prayers) (Hsing Yun, 2003)	2	4213	44664	29.5%	Yes
21	Rabbit's Horn (Hsing Yun, 2010d)	1	2706	49083	48.1%	Yes
22	Ten Paths to Happiness (Hsing Yun, 2014b)	1	3268	58118	5.4%	Yes
23	Universal Gate (Hsing Yun, 2011c)	1	1770	37844	25.9%	Yes
	(C) Booklets					
1	Buddhism in Every Step A1-13 (Hsing Yun, 2015a)	13	4114	74298	24.0%	Yes
2	Buddhism in Every Step B1-B2, B4-B8 (Hsing Yun, 2015b)	7	2410	48196	14.0%	Yes
3	Buddhism in Every Step C1-C7 (Hsing Yun, 2015c)	7	3425	71839	18.4%	Yes
4	Buddhism in Every Step D1-D2 (Hsing Yun, 2015d)	2	949	18505	8.7%	Yes
5	Buddhism in Every Step H2-H6 (Hsing Yun, 2015e)	5	878	19256	13.0%	Yes
	(D) DVDs					
1	Buddha Dharma (Hsing Yun, 2014c)	4	3082	49108	3.0%	Yes
2	Heart Sutra (Hsing Yun, 2014a)	3	2251	34135	6.2%	No

Table 1: Humanistic Buddhism Corpus Collection

(the first author of this paper), who had over 25 years of Chinese-English Buddhist translation experience. This reviewer caught any misalignment of the parallel texts made by the volunteers. Non-optimal translations were then removed from the corpus. These removed phrases include cases when additional English content was added or untranslated Chinese was retained, as well as cases when the Chinese content was intertwined in several English sentences, making it difficult to split into sentence-based or even phrase-based parallel data. Light editing was performed on the re-

maining phrases, which improved the quality of phrase-based translations. Light editing included adding or removing pronouns or names, common in publication work where fluency was required when reading on a paragraph level instead of a phrase level.

We further filtered the data before tokenization through two simple steps. First, we filtered out long (>150 Chinese characters or English words) phrases. Next, we removed over-translated or under-translated phrases by filtering out phrases in which the number of words in one language (English words or Chinese characters, represented as E and C) was twice as many as in the other language: ($l_E > 2 * l_C$ or $l_C > 2 * l_E$). For proverbs and phrases with < 5 characters, the ratio of this filtering is changed to 1:4 or 4:1 (Zeng et al., 2021).

3.2. Corpus Collection

The collection of texts in the Humanistic Buddhism Corpus is listed in Table 1. Besides the number of lines and the number of Chinese characters included in the corpus, it indicates the percentage of the text manually excluded from the HBC due to either no matching of the original Chinese or translated English or other issues with the translation. Although samples from all the sutras, books, booklets, and DVDs are included in the training set, the table indicates whether or not samples from a book were also included in the validation and test sets. The "Test=No" in Table 1 indicates which sources are in classical Chinese, which comprises 20% of the corpus. The titles used in this table are shortened versions, and links to their complete bibliographic references are provided after the titles.

3.3. Corpus Location and License

The Humanistic Buddhism Corpus is publicly available at:

https://www.fgstranslation.org/hbc

This corpus is intended to be used in academic and research settings and not for use in for-profit industries. The corpus will be updated periodically to include more data and better translations. The Humanistic Buddhism Corpus is licensed under CC-BY-NC-SA, the Attribution-NonCommercial-ShareAlike of the Creative Commons 4.0 International License.

4. Experiments

4.1. Experiment Setup

4.1.1. Corpora Split

The HBC consists of 81,000 phrase pairs for training and testing after filtering. We split the data into 77,000 training, 2,000 validation, and 2,000 test instance subsets (an approximately 95%/5%/5% train/validation/test split, selected to maximize the amount of available training data). We first mix the phrases from books that are not sutras or proverbs ("Test=Yes" in Table 1) and split them randomly for train/validation/test sets, then add the phrases from sutras and proverbs to the train set and mix them randomly again. The validation and test sets do not contain sutras, poems, or proverbs, as those types of text contain phrases that are often quoted in other texts (and thus may be present in the training data).

The Datum2017 corpus from the China Workshop on Machine Translation (CWMT) from WMT21 was used for training some of the models in conjunction with the HBC. The Datum dataset contains 20 books with 50,000 paired sentences each, for a total of 1 million parallel sentences. After a similar filtering procedure as applied to HBC, we split the data into 600,000 training, 2,000 validation, and 2,000 test instance subsets (an approximately 98%/1%/1% train/validation/test split, selected such that the validation and test subsets are equivalent in size to those used with HBC). The training set was further subdivided into 200k and 600k instance subsets for assorted training. For some of the combined-corpora experiments, the HBC data were duplicated during training to add more weight to the Buddhist dataset.

4.1.2. Base Transformer Model

To establish a performance benchmark for HBC, we conducted experiments using the Base Transformer model with 6 encoders and 6 decoders originally designed by Vaswani et al. (2017). All experiments were conducted using one NVIDIA RTX A6000 GPU with 48 GB memory. The data is tokenized using the SentencePiece algorithm (Kudo and Richardson, 2018) before training on the Transformer model, with a total of 100,000 English tokens and 100,000 Chinese tokens.

4.1.3. Pre-Trained Models

To further emphasize the inherent and interesting challenges raised by the HBC, two pre-trained models were also applied to the WMT21 Datum2017 and HBC test sets: Argos Translate[1] and Google Translate.[2] Each model was executed "as is" on the test sets without further training or fine-tuning. Argos Translate trains a PyTorch Transformer model with OPUS (Tiedemann, 2012) as its primary data source. It is one of the limited open

[1] https://github.com/argosopentech/argos-translate

[2] https://pypi.org/project/googletrans/

Model	Dataset	Train Size	Val Size	BLEU	COMET
Base Transformer	Datum	200k	2k	40.04	0.792
Argos Translate	Opus	–	–	29.02	0.498
Google Translate	–	–	–	68.73	0.855

Table 2: BLEU and COMET Scores on the WMT21 Datum Test Set

Model	Dataset	Train Size	Val Size	BLEU	COMET
Base Transformer	Datum	200k	2k	3.03	0.502
Argos Translate	Opus	–	–	2.19	0.438
Google Translate	–	–	–	43.53	0.747
Base Transformer	HBC	77k	2k	8.73	0.585
Base Transformer	HBC	154k	2k	9.34	0.588
Base Transformer	Datum/HBC	200k/77k	1k/1k	11.75	0.651
Base Transformer	Datum/HBC	200k/154k	1k/1k	13.18	0.668
Base Transformer	Datum/HBC	600k/154k	1k/1k	12.81	0.670

Table 3: BLEU and COMET scores on the HBC Test Set

source pre-trained NMT models for Chinese to English pairs. Google Translate has a Python API that is available to all. The corresponding translation model is a hybrid architecture of a Transformer encoder and a Recurrent Neural Network (RNN) decoder. It is a proprietary model that has been trained using vast compute and data resources. However, it is not open source and its data sources are not available to outside institutions or for academic research.

4.1.4. Evaluation Methods

To evaluate the performance of these model conditions, we used both Bi-Lingual Evaluation Understudy (Papineni et al., 2002, BLEU) scores and Crosslingual Optimized Metric for Evaluation of Translation (Rei et al., 2020, COMET) scores. The SacreBLEU package (Post, 2018) was used to compute the BLEU score. The predicted translations from all models for the same test set were exported for quality evaluation so that non-expert readers could assess them. The training, validation, and test sets of the HBC are available online, as well as the predicted translations of the test set by each model, so that readers can qualitatively evaluate the translations in the test sets.

4.2. Quantitative Analysis

Table 2 provides benchmark results for evaluating our generic news test set, WMT21 Datum, using various models. The pre-trained Argos Translate provides a decent BLEU score of 29.02, whereas the Base Transformer trained on the Datum data and tested on the Datum data yields a 40.04 BLEU score. Google Translate scored best, with a BLEU

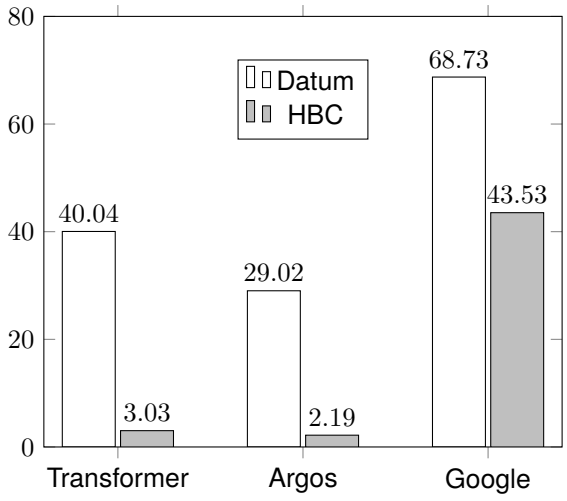

Figure 1: BLEU Scores for Datum and HBC

score of 56.30. We speculate that since Google Translate was trained on an enormous amount of data on a hybrid Transformer-RNN model, its score can reach 20 points higher than the Base Transformer, which, together with Argos Translate, used a pure Transformer architecture. This table also gives benchmark COMET scores, with 0.792, 0.498, and 0.855 for the Base Transformer, Argos Translate, and Google Translate, respectively.

Table 3 presents our benchmarking results using the Humanistic Buddhism Corpus in our transfer learning setup, leveraging the WMT21 Datum dataset for pre-training the Base Transformer model. We note that the purpose of these experiments was not to achieve state-of-the-art machine translation performance, but rather to provide a proof-of-concept for training models using

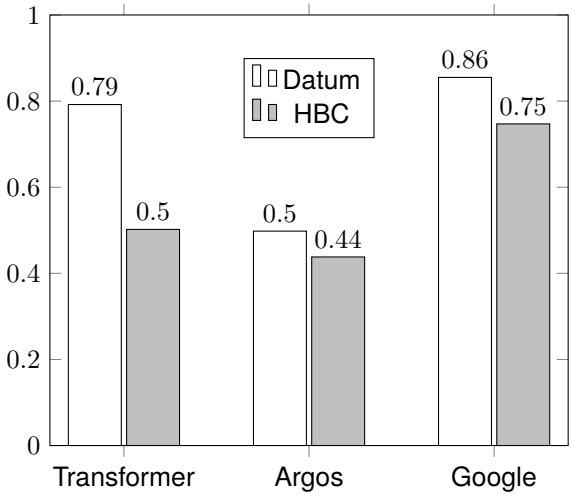

Figure 2: COMET Scores for Datum and HBC

our dataset as well as an initial starting point for other researchers. Although the Base Transformer training on the Datum dataset alone can produce a BLEU of 40.04 or a COMET of 0.792 when testing on the Datum test set, when the same model is tested on the HBC test set, the predicted translations are meaningless random words, likely due to the test data's domain mismatch. When training on the HBC data alone, the predicted translations are not ideal either, producing a BLEU of only 8.73 and a COMET of 0.585—qualitatively, this maps to about half of the translations being comprehensible with basic Buddhist terminologies, but the other half being mistranslated or not fluent in grammar. The Base Transformer likely struggles with HBC due to its scarcity. Even when we double the HBC training data via repetition, the BLEU score is only 9.34, and the COMET score is 0.588. This highlights the challenging nature of this task, even compared to other Chinese-English translation settings.

When training the HBC using transfer learning, pre-training on just 200,000 lines of the Datum dataset, we qualitatively observed that the translations grew more fluent with few grammatical errors. Correspondingly, BLEU increased to 11.75 and COMET rose to 0.651. When changing the weight by doubling the HBC training data, the results increased to 13.18 and 0.668. This suggests that although Datum does not offer suitable performance when used in an entirely out-of-domain training context, it can still be productively leveraged in other capacities to support models designed for use with our challenging new dataset. However, the proportion of the transfer learning data (Datum) and the HBC data cannot be too high, as indicated by the slightly lowered BLEU of 12.81 when training on 600,000 lines of the Datum dataset with double the HBC training data. The in-domain and out-of-domain parallel data need to be balanced proportionally to produce optimal results. Dataset pruning methods (Gormez and Koyuncu, 2023) can also potentially identify the most significant samples during the training process.

The predicted translations by the two pre-trained models also indicate that the HBC is a challenging dataset. The Argos Translate model transparently showed the in-domain and out-of-domain discrepancy. Despite obtaining a BLEU of 29.02 and COMET of 0.498 on the Datum test set, Argos Translate performed especially poorly on the HBC test set, with a BLEU of 2.19 and COMET of 0.438, which contains incomprehensible translations. Even the powerful Google Translate underperformed on our challenging HBC test set, with BLEU score dropping by more than 25 and a COMET decrease of 0.11 compared with the more generic Datum test set. Lower scores suggest that significant room for enhancement exists in future research working with our HBC dataset.

Figure 1 illustrates the challenging aspect of this dataset. The BLEU scores of the various models testing on the Datum and the HBC test sets are shown in that chart. The white bars represent the Datum test set, and the gray bars represent the HBC test set. The 3.03 and 2.19 BLEU scores indicate that the predicted translations are incomprehensible, with numerous terminologies untranslated. In the case of Google Translate, the drop was over 25 BLEU. Figure 2 shows the COMET scores of these pre-trained and Base Transformer models. This chart also indicates the decrease in COMET scores for the HBC test set in all models, further emphasizing the complexity of the Humanistic Buddhism Corpus.

4.3. Translation Quality Analysis

The quantitative analysis of the predicted translations using BLEU and COMET scores provides a quick measurement of the quality of the translations. However, not only is BLEU score known to carry many limitations (Callison-Burch et al., 2006; Post, 2018; Mathur et al., 2020), automatic metrics in general present challenges as the de facto practice for evaluation processes (Gehrmann et al., 2023). Therefore, we randomly selected three phrases and their target and predicted translations from the HBC test set, and presented them for manual review by readers to analyze the translation quality. All predicted translations from the test set are also available online[3] for human evaluators to examine on a deeper level.

Tables 4-6 present sample translations. The first line is the Chinese phrase, followed by the target translation provided by the dataset. The third line

[3] https://www.fgstranslation.org/hbc

Source	甚至 佛教 講 「不思善、 不思惡」， shènzhì fójiào jiǎng bùsīshàn, bùsīè,
Target	There is another saying in the Buddhist tradition: ”give thought neither to what is wholesome nor unwholesome.”
Google Translate	Even Buddhism says ”don’t think good, don’t think evil”,
Transformer HBC77k	Moreover, Buddhism teaches us not to think of what is right or wrong;
Transformer Dat200k/HBC77k	In Buddhism, there is no thought of wholesome or unwholesome deeds.
Transformer Dat200k/HBC154k	Buddhism teaches us that we should not think of what is wholesome or what is unwholesome.

Table 4: Sample Translation 1

Source	菩薩 以 大智慧 照見 五蘊 皆空， 徹悟 諸法空性， púsà yǐ dàzhìhuì zhàojiàn wǔyùn jiēkōng, chèwù zhūfǎkōngxìng,
Target	With great wisdom, Bodhisattvas see through the emptiness of the five aggregates and thoroughly understand that all phenomena is empty.
Google Translate	With great wisdom, the Bodhisattva sees that the five aggregates are all empty, and fully understands the emptiness of all dharmas.
Transformer HBC77k	The great wisdom of Bodhisattvas is capable of understanding the emptiness of the five aggregates.
Transformer Dat200k/HBC77k	Bodhisattvas use great wisdom as the five aggregates to realize that all phenomena are empty.
Transformer Dat200k/HBC154k	Through the five aggregates of great wisdom, Bodhisattvas realize the emptiness of all phenomena.

Table 5: Sample Translation 2

is the predicted translation by Google Translate. The fourth line is our Base Transformer model, trained on the 77k HBC training set alone. The fifth line is the translation generated when training on both the Datum (200k) and HBC (77k) datasets jointly. The last line is the result from the weighted data setting, doubling the HBC size (154k) through duplication and training the model jointly with the doubled HBC dataset and the Datum dataset.

Sample Translation 1 in Table 4 is an excellent example of the variation of translations predicted by the Transformer models using the HBC dataset. All the translations are fluent and adequate. Despite the HBC’s 77k training size, the translation from training on HBC alone is correct, and to some editors it is preferred over Google Translate’s translation. This suggests that the HBC contains sufficient training data to provide fluent and adequate translations in some cases.

The sample phrase in Table 5 presents the unique challenges of the HBC because it contains multiple elements of Buddhism such as ”Bodhisattvas,” ”the five aggregates,” ”emptiness,” and ”phenomena/dharmas.” Among them are Buddhist metaphors and symbolism that are subject to multiple interpretations. For example, the Chinese word “法 (fǎ)” can be translated as “Dharma,” “dharmas,” “mental states,” or “phenomena.” Although the predicted translation from training on just the HBC is missing this terminology, it provides the most common translation of this term in the context of “諸法空性 (zhū fǎ kōng xìng)” when trained with the Datum dataset. Since the HBC’s data are from published works, “phenomena” is considered better than Google’s translation “dharmas.” All these translations are evident of the difficulty in translating abstract concepts.

Table 6 conveys common mistakes in translating the names of Buddhist figures. The names can be translated using their original Sanskrit or Pali, transliteration from the Chinese pronunciation, and even translating their meaning. In most

Source	淨飯大王 jìngfàndàwáng	心中 xīnzhōng	高興地 gāoxìngde	想著： xiǎngzhe:
Target	Filled with glee, King Suddhodana thought,			
Google Translate	King Jingfan happily thought in his heart:			
Transformer (HBC77k)	King Suddhodana explained to the king,			
Transformer (Dat200k/HBC77k)	King Suddhodana was pleased to think,			
Transformer (Dat200k/HBC154k)	King Suddhodana was delighted and thought,			

Table 6: Sample Translation 3

of the publications, the original Sanskrit or Pali names are preferred. Therefore, all the models that trained on the HBC produced the correct name in Sanskrit. However, Google Translate did not predict the name but provided the Chinese pronunciation of the meaning of the king's name. Google's translation is not considered incorrect, but it is inferior to using the original names when they are available. Therefore, training using the HBC provides valuable results that extend performance beyond what is currently available from huge pre-trained models.

5. Conclusion

In this work, we introduced a novel, challenging dataset of English translations for classical and modern Chinese: the Humanistic Buddhism Corpus. The dataset offers domain-specific paired Chinese-English language samples for a particularly challenging domain spanning religious books and booklets, proverbs and poems, and video subtitles all pertaining to Humanistic Buddhism. Translations were manually matched, aligned, and scored by 21 human translators followed by an additional final manual review. Overall, the dataset comprises 81,000 paired phrases, representing a substantial and important contribution to the field of neural machine translation. Furthermore, the HBC will be expanded gradually in the years to come, providing an invaluable large corpus to both the Buddhist community and Chinese-English NMT researchers.

We perform benchmarking experiments to establish a performance baseline for this challenging new dataset, achieving a maximum BLEU score of 13.18 and COMET score of 0.670 when leveraging a transfer learning setting pre-training on the WMT21 Datum general-domain Chinese-English dataset. This highlights the challenging nature of this task, offering opportunity for future research. We hope that HBC will inspire meaningful advances to machine translation within this and other complex, domain-specific translation settings.

6. Acknowledgements

Youheng Wong is grateful for the generosity of Venerable Master Hsing Yun for providing his works for this research, and thanks the Buddha's Light Publications, Fo Guang Shan International Translation Center, Gandha Samudra Culture Company, Beautiful Life Television, and Fo Guang Shan Institute of Humanistic Buddhism for providing English translations of the master's works.

The human translators involved in matching this parallel corpus are from the Buddha's Light International Association. They are Eve Lee, Avelyn Busch, Joyce Kueh, Helen Yau, Ann-C Lin, Sharon Chang, Jeremy Hsu, Athena Huang, Jia Liu, Amy Thomason, Bonnie Xie, George Sheu, Dean Isensee, Andrew Chou, Annie Chen, Vanessa Kwok, Eugene Lo, Yawei Yang, Felicia Hsieh, Yifan Wang, and Hui Wang. We thank them for their contributions.

The work of Erdem Koyuncu was supported in part by the Army Research Lab (ARL) under Grant W911NF2120272, and by the Army Research Office (ARO) under Grant W911NF2410049.

7. Bibliographical References

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The eBible Corpus: Data and Model Benchmarks for Bible Translation for Low-Resource Languages.

BCBS. 1995. Access to Insight.

BDRC. 2016. Buddhist Digital Resource Center.

BLP. 2023. Buddha's Light Publications.

BuddhaNet. 2023. Buddhist eLibrary.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in

machine translation. *Proceedings the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

CBETA. 1998. Chinese Buddhist Electronic Text Association.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

FGSITC. 2020. Fo Guang Shan International Translation Center.

Fo Guang Shan. 2004. Fo Guang Shan Buddhist Electronic Texts.

Zhiying Fu. 1996. *Handing Down the Light: The Biography of Venerable Master Hsing Yun*. Buddha's Light Publications.

Zhiying Fu. 2008. *Bright Star, Luminous Cloud: The Life of a Simple Monk*. Buddha's Light Publications.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166.

Alperen Gormez and Erdem Koyuncu. 2023. Dataset pruning using early exit networks. In *ICML Localized Learning Workshop*.

Xing Guang. 2013. Buddhist impact on chinese culture. *Asian Philosophy*, 23(4):305–322.

Hsing Yun. 1999. *Epoch of the Buddha's Light: A Letter to Members of the BLIA*. Buddha's Light Publications.

Hsing Yun. 2002. *The Everlasting Light: Dharma Thoughts of Master Hsing Yun*. Gandha Samudra Culture Company.

Hsing Yun. 2003. *Pearls of Wisdom: Prayers for Engaged Living*. Buddha's Light Publications.

Hsing Yun. 2008. *Humanistic Buddhism: A Blueprint for Life*. Buddha's Light Publications.

Hsing Yun. 2009. *Living Affinity*. Buddha's Light Publications.

Hsing Yun. 2010a. *The Buddha's Light Philosophy*. Buddha's Light Publications.

Hsing Yun. 2010b. *For All Living Beings: A Guide to Buddhist Practice*. Buddha's Light Publications.

Hsing Yun. 2010c. *Infinite Compassion, Endless Wisdom: The Practice of the Bodhisattva Path*. Buddha's Light Publications.

Hsing Yun. 2010d. *The Rabbit's Horn: A Commentary on the Platform Sutra*. Buddha's Light Publications.

Hsing Yun. 2011a. *After Many Autumns*. Buddha's Light Publications.

Hsing Yun. 2011b. *Humble Table, Wise Fare*. Buddha's Light Publications.

Hsing Yun. 2011c. *The Universal Gate: A Commentary on Avalokitesvara's Universal Gate Sutra*. Buddha's Light Publications.

Hsing Yun. 2012a. *Bells, Gongs and Wooden Fish: Voices for Buddhist Change*. Buddha's Light Publications.

Hsing Yun. 2012b. *Four Insights for Finding Fulfillment*. Buddha's Light Publications.

Hsing Yun. 2013. *Biography of Sakyamuni Buddha*. Buddha's Light Publications.

Hsing Yun. 2014a. *Heart Sutra - The Perspective on Life and the Universe*. Beautiful Life Television.

Hsing Yun. 2014b. *Ten Paths to Happiness*. Buddha's Light Publications.

Hsing Yun. 2014c. *Ven. Master Hsing Yun's Lectures - The True Meaning of the Buddha Dharma*. Beautiful Life Television.

Hsing Yun. 2015a. *Buddhism in Every Step - A*. Fo Guang Shan International Translation Center.

Hsing Yun. 2015b. *Buddhism in Every Step - B*. Fo Guang Shan International Translation Center.

Hsing Yun. 2015c. *Buddhism in Every Step - C*. Fo Guang Shan International Translation Center.

Hsing Yun. 2015d. *Buddhism in Every Step - D*. Fo Guang Shan International Translation Center.

Hsing Yun. 2015e. *Buddhism in Every Step - H*. Fo Guang Shan International Translation Center.

Hsing Yun. 2017. *The Complete Works of Venerable Master Hsing Yun*. Fo Guang Publications.

Hsing Yun. 2019. *Buddha Dharma: Pure And Simple*. Fo Guang Shan Institute of Humanistic Buddhism.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and Why is Unsupervised Neural Machine Translation Useless? *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020*, pages 35–44.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. *ACL 2017 - Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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. *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 228–232.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. *WMT 2018 - Proceedings of the Third Conference on Machine Translation*, 1:186–191.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The NIST 2008 metrics for machine translation challenge-overview, methodology, metrics, and results. *Machine Translation*, 23(2-3):71–103.

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

Sakyamuni Buddha. 2015. *The Medicine Buddha Sutra*. Fo Guang Shan International Translation Center.

Sakyamuni Buddha. 2017. *Amitabha Sutra as Discoursed by the Buddha*. Fo Guang Shan International Translation Center.

Sakyamuni Buddha. 2018. *Difficulty of Repaying the Profound Kindness of Parents Sutra*. Fo Guang Shan International Translation Center.

Sakyamuni Buddha. 2019a. *The Diamond Sutra*. Fo Guang Shan International Translation Center.

Sakyamuni Buddha. 2019b. *Flower Adornment Sutra's Practices and Vows of Samantabhadra Bodhisattva Chapter*. Fo Guang Shan International Translation Center.

Sakyamuni Buddha. 2023. *Original Vows of Ksitigarbha Bodhisattva Sutra*. Fo Guang Shan International Translation Center.

Sheng'an. 2017. *Inspiration to Give Rise to the Bodhi Mind*. Fo Guang Shan International Translation Center.

Haipeng Sun, Rui Wang, Masao Utiyama, Benjamin Marie, Kehai Chen, Eiichiro Sumita, and Tiejun Zhao. 2021. Unsupervised Neural Machine Translation for Similar and Distant Language Pairs : An Empirical Study. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(1):1–17.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI ' s WMT21 News Translation Task Submission. *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 205–215.

University of the West. 2003. Digital Sanskrit Buddhist Canon.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem.

Vipassana Research Institute. 2010. The Pali Tipitaka.

Longyue Wang, Mu Li, Fangxu Liu, Shuming Shi, Zhaopeng Tu, Xing Wang, Shuangzhi Wu, Jiali Zeng, and Wen Zhang. 2021. Tencent Translation System for the WMT21 News Translation Task. *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 216–224.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. WeChat Neural Machine Translation Systems for WMT21. *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 243–254.

Qingzhi Zhu. 2003. The Impact of Buddhism on the Development of Chinese Vocabulary (I). *Universal Buddhist Gate Journal*, 16(1):1–35.