*electronics*

MDPI

*Article*

# WCC-EC 2.0: Enhancing Neural Machine Translation with a 1.6M+ Web-Crawled English-Chinese Parallel Corpus

**Jinyi Zhang [1,2,*], Ke Su [1], Ye Tian [3] and Tadahiro Matsumoto [2]**

[1] School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China; suke1103@outlook.com
[2] Faculty of Engineering, Gifu University, Gifu 501-1193, Japan; tad@gifu-u.ac.jp
[3] Zhuzhou CRRC Times Electric Co., Ltd., Zhuzhou 412001, China; tianye@csrzic.com
* Correspondence: zhangjinyi@sylu.edu.cn

**Abstract:** This research introduces WCC-EC 2.0 (Web-Crawled Corpus—English and Chinese), a comprehensive parallel corpus designed for enhancing Neural Machine Translation (NMT), featuring over 1.6 million English-Chinese sentence pairs meticulously gathered via web crawling. This corpus, extracted through an advanced web crawler, showcases the vast linguistic diversity and richness of English and Chinese, uniquely spanning the rarely covered news and music domains. Our methodical approach in web crawling and corpus assembly, coupled with rigorous experiments and manual evaluations, demonstrated its superiority by achieving high BLEU scores, marking significant strides in translation accuracy and model resilience. Its inclusion of these specific areas adds significant value, providing a unique dataset that enriches the scope for NMT research and development. With the rise of NMT technology, WCC-EC 2.0 emerges not only as an invaluable resource for researchers and developers, but also as a pivotal tool for improving translation accuracy, training more resilient models, and promoting interlingual communication.

**Keywords:** neural machine translation; sentence alignment; corpus construction; English-Chinese parallel corpus; web crawling techniques



**Citation:** Zhang, J.; Su, K.; Tian, Y.; Matsumoto, T. WCC-EC 2.0: Enhancing Neural Machine Translation with a 1.6M+ Web-Crawled English-Chinese Parallel Corpus. *Electronics* **2024**, *13*, 1381. https://doi.org/10.3390/electronics13071381

Academic Editor: Arkaitz Zubiaga

Received: 29 February 2024
Revised: 30 March 2024
Accepted: 1 April 2024
Published: 5 April 2024



## 1. Introduction

In recent years, Neural Machine Translation (NMT) [1–4] has become the leading technology in the machine translation field, providing many improvements over older methods. The core of NMT was an encoder-decoder neural network model, improved by an attention mechanism that skillfully captured the complex relationships and nuances between languages. This feature greatly increased the accuracy and smoothness of translations. As a result, NMT was better than traditional statistical methods at translating long sentences, words with multiple meanings, and sentences with complicated grammar. An important feature of NMT was its end-to-end training process, which allowed for direct learning from the source to the target language without needing the multiple stages typical of older methods. This unified approach made the model's design and training simpler and greatly improved the efficiency of the translation system.

However, NMT also has its weaknesses. It depends heavily on large, high-quality parallel corpora, which is a challenge for languages with fewer resources. NMT models are also sensitive to mistakes or unclear parts in the input, which could lead to less accurate translations. They could struggle with specialized terms and specific contexts. On the other hand, Large Language Models (LLMs), such as ChatGPT [5] and GPT-4 [6], trained on large text collections, are good at learning a wide range of language knowledge and understanding context. This broad training gave LLMs an advantage in creating translations that are coherent and make sense, especially with complex or unclear text. Their ability to translate is impressive, and several studies [7–10] have highlighted their effectiveness. In fact, some LLMs have produced translations that are as good as the best systems from

the Workshop on Machine Translation (WMT) [11]. Researchers have been actively looking into how these LLMs perform with both widely spoken and low-resource languages [9,12,13], multilingual translation [8], and document-level translation [14,15]. There has also been a trend towards making LLMs better at translating through fine-tuning [16,17] and developing better strategies for prompting [18,19]. Moreover, NMT still needs improvements for some language structures and low-resource languages [20]. It is vital to find effective methods to make translation models perform better across different areas of language processing.

The quality and size of a corpus greatly affects how well NMT and LLMs work. Creating large, high-quality corpora is still a key focus of research and practical use of NMT. In this situation, having large and high-quality English-Chinese corpora that are specific to certain fields is very valuable. They give researchers more diverse data, which is crucial for advancing NMT.

WCC-EC 1.0 [21] contains approximately 340,000 pairs of English-Chinese news data. This corpus is useful because it covers many subjects and has formal language features, but it is not perfect. To make up for its lack of casual language and everyday expressions, this study added about 1.3 million pairs of lyrics data to WCC-EC 1.0, creating the WCC-EC 2.0 parallel corpus that included both news and music content. There were several reasons for adding the lyrics:

- Firstly, there was a clear gap in the creation of bilingual English-Chinese corpora for music, a gap not addressed by major projects such as WMT and OPUS [22], even though there were corpora with tens of millions of pairs. Furthermore, lyrics tended to be more conversational, full of spoken phrases, slang, and common expressions that are closer to everyday language. The common use of words with multiple meanings in spoken language gives NMT more context to work with, helping it deal with uncertainties and improve translation accuracy.
- The scaling laws of language modeling suggests that the effectiveness of a model depended on how much data it has. In recent years, increasing the amount of data has been a common way to make language models better. However, the Epoch AI Research team [23] predicted that the stock of high-quality language data would run out by 2026, and even the stock of lower-quality data would start to run out between 2030 and 2050. Therefore, combining a music domain parallel corpus with WCC-EC 1.0 to create a Domain-Specific Parallel Corpus—WCC-EC 2.0 is very important.

The main contributions of this paper are as follows:

- We introduced WCC-EC 2.0, an expanded parallel corpus that combins approximately 1.3 million pairs of lyrics data with existing news data, significantly enhancing the diversity of languages represented and augmenting the corpus's applicability for NMT research. Additionally, we have filled a notable gap by including texts from the music domain, which are currently scarce. This corpus is freely available for download for non-commercial research purposes, providing a valuable resource for the NMT community.
- We set up a strong human evaluation system that goes well with automatic measures, such as BLEU scores, to give a more detailed look at the quality of translations. This system makes it possible to judge translations based on how natural they sound, how complete they are, and how well they use everyday language, giving a deeper insight into the performance of WCC-EC 2.0.

This study mainly focuses on how WCC-EC 2.0 was built, how its quality was checked, and the evaluations of both the lyrics and news data. The paper is organized as follows. Section 2 discussed how parallel corpora are built and aligned. Section 3 described the complete process of building the corpus in this study, including the problems faced and the solutions found. Section 4 explained the experimental steps and gives a detailed analysis of the results. Section 5 ended with a review of WCC-EC 2.0 and looked at possible directions for future work.

## 2. Related Works

The importance of a corpus in NMT was crucial and cannot be overstated. As the fundamental base for training NMT models, a large bilingual corpus allowed the model to learn the language patterns and links between the source and target languages. This learning was key to achieving high-quality translations.

### *2.1. Corpus Construction*

The scale of a corpus was directly correlated with the performance of models in NMT. A substantial bilingual dataset aided the models in better understanding the correspondences between different languages, thereby enhancing translation quality. Over recent years, the development of WMT has been pivotal. Its annual dataset releases provided standardized benchmarks for researchers. Tiedemann [22] presented OPUS, an extensive freely available parallel corpus encompassing over 200 languages with tools for exploration and integration, enhancing research and development in linguistic studies. Initiatives such as Mackenzie et al.'s [24] creation of the CC-News-En corpus from the Common Crawl Foundation data mitigated the shortage of journalism corpora to an extent. To clarify corpus evaluation, Lefer's [25] chapter on Parallel Corpora in "A Practical Handbook of Corpus Linguistics" outlined the main features of parallel corpora. It also explored methods of analysis, including the combined use of parallel and comparable corpora, and addressed challenges in corpus design and analysis. ParaCrawl [26], the renowned parallel corpus developed by Marta Bañón et al., which leverages open-source software for web crawling to assemble the largest publicly available parallel corpus as of 2020, served as an inspiration for this research. Ziemski et al. [27] detailed the establishment of the official United Nations Parallel Corpus, marking the first parallel corpus from UN documents for its six official languages, accessible under a liberal license. Liu et al. [28] developed a pipeline for acquiring and processing an English-Chinese parallel corpus from the New England Journal of Medicine, demonstrating significant translation quality improvements with targeted training data. Liu et al. [29] introduced DuRecDial 2.0, a bilingual parallel dialog dataset for English and Chinese, aimed at advancing monolingual, multilingual, and cross-lingual conversational recommendation systems, showcasing the benefits of incorporating additional English data for Chinese conversational recommendations. Furthermore, Zhang et al. [30,31] made significant contributions by developing the WCC-JC Japanese-Chinese translation corpus and the manually aligned WCC-JC 2.0, a large-scale Japanese-Chinese parallel corpus, through web crawling, providing considerable support for Japanese-Chinese translation research.

Increases in corpus data volume have significantly contributed to the field of NMT. Sugiyama et al. [32] introduced context-aware neural machine translation (CAMT), which involved generating a large-scale pseudo-parallel corpus through back-translating monolingual data, supported by a substantial amount of parallel corpora. In 2023, Li et al. [33] enhanced traditional back-translation techniques with their novel approach, instruction back-translation. This method involved fine-tuning a language model with a minimal dataset and then using it to generate instructional cues for web documents, selecting only high-quality examples for training. To augment low-resource corpora, Morita et al. [34] proposed a method of hybrid and dynamic hybrid sampling, merging optimal and random sampling. Experimental results showed that dynamic hybrid sampling consistently outperformed previous optimal sampling methods. Zhang and Matsumoto [35] developed a strategy to expand the parallel corpus available for Japanese-Chinese NMT system, aiming to significantly elevate the translation quality in situations where bilingual corpora are limited. Zhang et al. [21] introduced a technique for corpus extension, involving dividing long sentences into shorter segments, recombining them, and then back-translating. This method not only expanded the corpus but also improved the quality of the predictions. Additionally, we have conducted an extensive review and comparative analysis of various corpora in recent years. The basic profiles of these corpora are concisely presented in Table 1, providing a clear understanding of their characteristics and differences.

**Table 1.** Basic information of some parallel corpora, where EU, AR, ZH, EN, FR, RU, and ES, respectively, represent the European Union, Arabic, Chinese, English, French, Russian, and Spanish.

| Corpus | Quantity | Language Pairs | Source Field | Applications | Manual Evaluation |
|---|---|---|---|---|---|
| ParaCrawl [26] | 223 million | 23 EU languages with EN | General | NMT | No |
| OPUS [22] | 3.2 billion | Over 200 | General | NMT | No |
| UN Corpus [27] | Over 10 million | AR, ZH, EN, FR, RU, ES | News | NMT | No |
| NEJM-enzh [28] | 100,000 | ZH & EN | Biomedical | NMT | No |
| DuRecDial 2.0 [29] | Over 8000 | ZH & EN | Dialogue | Conversation | No |
| WCC-EC 2.0 | Over 1.6 million | ZH & EN | News & Lyrics | NMT | Yes |

### *2.2. Text Alignment*

Discussing corpora inevitably brings up the topic of text alignment. Historically, methods for word alignment in parallel texts through unsupervised learning were common. Another approach involved using pre-trained contextual word embeddings from multilingual language models. Dou et al. [36] combined these methods by fine-tuning pre-trained models and improving alignment quality with specific goals, achieving stable performance across different language pairs. For aligning texts involving Chinese, embedding techniques such as Word2Vec [37] and FastText [38] proved effective in calculating sentence vectors for alignment.

Li et al. [39] observed that although word-based models are more susceptible to data sparsity and out-of-vocabulary words, they generally performed better across tasks, regardless of Chinese word segmentation. This observation provided a strong argument for re-evaluating the necessity of Chinese word segmentation in deep learning models. Jiang et al. [40] proposed a novel neural Conditional Random Field (CRF) alignment model, which not only took advantage of the sequential nature of sentences in parallel documents but also used a neural sentence pair model to capture semantic similarities, outperforming previous methods in monolingual sentence alignment tasks.

Web crawling has become a popular method for corpus acquisition, though the resulting corpus often contains considerable noise and impurities. Zhang et al. [41] utilized the multilingual capabilities of BERT for sentence alignment and employed the Generative Pre-Training (GPT) language model as a domain filter to achieve data domain balance. Cao et al.'s [42] analysis of BERT identified systematic issues, such as misalignments in open class lexemes and word pairs across different character sets, which were corrected through a series of alignment procedures.

Overall, these diverse investigations not only laid the foundation for the development of NMT, but also offered valuable insights for overcoming challenges in corpus construction and text alignment. In-depth research in these areas has improved our understanding of NMT mechanisms and continuously enhanced translation quality. Future efforts can explore innovative approaches in corpus construction, alignment methods, and data expansion, further propelling the field of NMT forward.

## 3. Construction of the WCC-EC 2.0

Our corpus, featuring texts from both the news and lyrics domains, was originally established as WCC-EC 1.0, comprising about 340,000 English-Chinese news data pairs. This section discusses the construction process of WCC-EC 2.0, highlighting the main challenges we encountered during the calibration phase and the strategies we implemented to overcome them.

### *3.1. Web Crawling*

For the English-Chinese bilingual news texts in WCC-EC 1.0, we selected the KEKE English website (http://www.kekenet.com (accessed on 15 February 2024)) as our primary data source. This website provides a broad spectrum of news content across various fields, such as campus life, entertainment, international affairs, economics, sports, social issues,

and factual reporting, ensuring the timeliness and relevance of our data. We utilized the crawler frameworks BeautifulSoup (https://www.crummy.com/software/BeautifulSoup/ (accessed on 15 February 2024)) and Scrapy (https://scrapy.org/ (accessed on 15 February 2024)) to navigate and download the bilingual articles, successfully constructing a diverse corpus of bilingual news articles.

The English-Chinese bilingual lyrics corpus was acquired with LrcHelper (https://github.com/ludoux/LrcHelper (accessed on 15 February 2024)), a tool designed for downloading lyrics from NetEase Cloud Music based on song list IDs. Through LrcHelper, we downloaded approximately 230,000 LyRiCs (LRC) files over nearly 15 days, covering around 2200 English song lists. A comprehensive de-duplication process was carried out on both the news and lyrics texts to ensure the uniqueness and quality of our corpus.

### *3.2. Extraction of Bilingual Sentences*

In constructing the news corpus, we began with segment alignment, followed by sentence splitting, and introduced the concept of a translation matrix during sentence alignment. A detailed methodology is documented in WCC-EC 1.0 [21], which extensively described the process of acquiring news texts, data augmentation methods, and corpus quality measurement, leading to the creation of a high-quality news parallel corpus.

For version 2.0, the lyrics texts were updated to LRC files, containing a significant amount of extraneous non-lyrical content. Extracting usable data without disrupting the original alignment of the lyrics required the development of a sophisticated proofreading program. This program, designed around the temporal alignment of English lyrics and their Chinese translations within a one-second interval, ensured the efficient extraction of bilingual text. Despite these efforts, the data still contained considerable noise. We applied regular expressions to remove annotations, primarily those within “【】” and “[ ]”, and eliminated special symbols such as “/n” and any non-standardized Chinese symbols, yielding more than 2 million sentence pairs. As depicted in Figure 1, the left side presents a sample of the lyric data, while the right side shows a sample of the news text before data cleaning.


**Figure 1.** (**Left**): NetEase Cloud Music lyrics text; (**Right**): KEKE News text (all dummy data).

### *3.3. Text Alignment*

In the process of aligning news texts, we used the Sentence BERT (SBERT) model [43] to calculate the similarity between different sentences, thereby identifying pairs of sentences with a similarity exceeding a predefined threshold. In the sentence embedding model, a higher similarity value indicates a closer semantic resemblance between two sentences. To improve the performance of the SBERT model, we used an intra-paragraph matching method that took paragraph information into account to enhance the accuracy of matching parallel sentences [21]. To validate this method, we randomly selected 1000 sentence pairs and conducted manual verification. The alignment rate increased from 92.1% to 95.2% as a result.

The lyrics text inherently had high alignment. However, the NetEase Cloud Music platform did not have strict regulations for lyric uploads, leading to complex redundant noise data that could affect the corpus's alignment rate. Therefore, data cleaning was essential before conducting alignment experiments. This involved resolving issues such as severe mixed-language lyrics and errors leading to bilingual content being exclusively in English or Chinese. Considering English lyrics often contained profanity and the text needed to conform to moral and legal standards, we systematically removed sentences with uncivilized language based on predefined keywords. Finally, we randomly selected a sample set of 2000 pairs of data for manual test alignment experiments, showing an impressive alignment rate of approximately 98.4%, indicative of a high level of alignment.

Ultimately, the lyrics text and news text were combined to form the WCC-EC 2.0 corpus.

### *3.4. Corpus Segmentation*

In the final phase of corpus construction, selecting sentence pairs for the corpus's validation and test sets was crucial. We adopted the conventional NMT parallel corpus construction method for this task. Given the corpus's composition of lyrics and news articles, we conducted proportional random sampling to ensure that the 2000 selected sentence pairs accurately represented the proportional distribution of lyrics and news articles throughout the entire corpus. Moreover, to maintain data quality, we exclusively chose sentence pairs containing sentences with at least 10 characters. As a result, a total of 4000 sentence pairs were sampled, with 2000 pairs each forming the development set and the test set, respectively. The remaining data were designated as the training set.

### *3.5. Summary of the Corpus Construction*

Figure 2 illustrates the comprehensive process of constructing the WCC-EC 2.0, encompassing four key phases: (1) web crawling; (2) data extraction; (3) text alignment; and (4) corpus segmentation. Detailed descriptions of these stages are provided in Sections 3.1–3.4, covering the various phases involved in building the corpus.

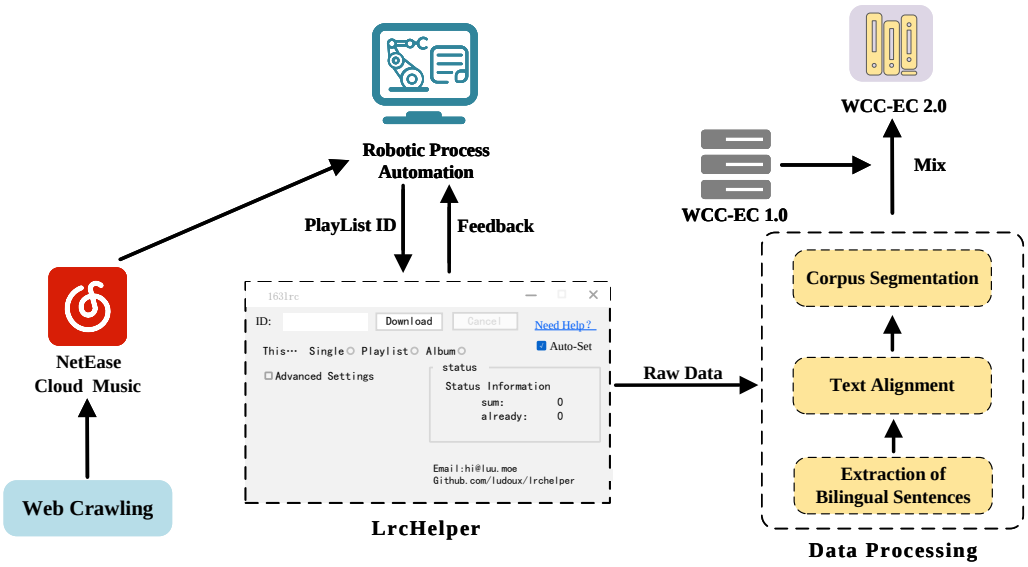


**Figure 2.** Process of crawling WCC-EC 2.0.

Table 2 presents the dataset breakdown across different corpora: WCC-EC 2.0, denoted as WCC-EC 2.0-Full; the lyrics segment of WCC-EC 2.0, referred to as WCC-EC 2.0-Lyrics; and the news segment of WCC-EC 2.0, identified as WCC-EC 2.0-News (also known as WCC-EC 1.0) [21]. Additionally, it includes the V16 Dataset (https://data.statmt.org/news-commentary/v16/training/news-commentary-v16.en-zh.tsv.gz (accessed on 15 February 2024)).

**Table 2.** Overview of the English-Chinese parallel corpora.

| Contents | Number of English and Chinese Bilingual Sentences | | | |
|---|---|---|---|---|
| | WCC-EC 2.0-Full | WCC-EC 2.0-Lyrics | WCC-EC 2.0-News | V16 Dataset |
| Training set | 1,662,908 | 1,323,651 | 339,257 | 318,235 |
| Development set | 2000 | 2000 | 2000 | 2000 |
| Test set | 2000 | 2000 | 2000 | 2000 |

## 4. Experiment and Evaluation

To assess the corpus's quality, we executed the following experiments, with detailed descriptions of the datasets employed provided in Section 4.1. The experimental framework is expounded upon in Section 4.2. In Section 4.3, the evaluation of WCC-EC 2.0 quality is presented in comparison.

### *4.1. Dataset*

We employed two datasets for our experiments. The first dataset is derived from News Commentary v16, a news dataset provided by WMT2022 (https://www.statmt.org/wmt22 (accessed on 15 February 2024)), containing about 313,000 English-Chinese sentence pairs. From this dataset, we randomly extracted 4000 statements, allocating 2000 for the development set and the remaining 2000 for the test set.

Furthermore, by leveraging our self-developed WCC-EC 2.0, which consists of approximately 1.6 million Chinese and English sentence pairs, we ensured the independence of WCC-EC 2.0-Full, WCC-EC 2.0-Lyrics, and WCC-EC 2.0-News during the experiments. To achieve this, we randomly selected 4000 sentence pairs from both WCC-EC 2.0-Lyrics and WCC-EC 2.0-News, allocating 2000 pairs as the test set and the other 2000 as the development set for each segment. As a result, both the test set and the development set comprised a total of 4000 utterances each. Following this, we randomly chose 2000 sentence pairs from these 4000 in the development set as the test set for WCC-EC 2.0-Full,

and another 2000 from the 4000 in the test set as the validation set. This method ensured data independence while maximizing the use of the available data.

### *4.2. Setting Up the NMT Framework*

For our subsequent experiments, we chose the fairseq framework (https://github.com/facebookresearch/fairseq (accessed on 15 February 2024)) and employed its Transformer pre-trained model. This model is equipped with 6 encoder and 6 decoder layers, featuring 8 encoder attention heads and a word vector size of 512. The architecture of Transformers is shown in Figure 3. We adhered to other hyperparameter configurations as well, including a dropout rate of 0.3, beta values of 0.9 and 0.98, setting the learning rate at $1 \times 10^{-7}$, a token limit of 4096, the batch size of 128, and a maximum update cap of 200,000 steps. For preprocessing subwords, we implemented the BPE algorithm (https://github.com/rsennrich/subword-nmt (accessed on 15 February 2024)) with a vocabulary size of 32,000. The configuration of these hyperparameters reflects the typical setup of deep learning models in machine translation tasks. It ensures both the depth and complexity of the model to capture the intricate relationships between languages, while effectively preventing overfitting and enhancing training stability through the adjustment of training parameters, thereby guaranteeing the quality and efficiency of machine translation.

During the prediction phase, a beam size of 8 was used to produce the translation outputs. Considering the absence of spaces in Chinese text, Jieba (https://github.com/fxsjy/jieba (accessed on 15 February 2024)) was utilized for sentence segmentation, and Moses (https://github.com/moses-smt/mosesdecoder (accessed on 15 February 2024)) was applied for punctuation and case modifications. The quality of machine translation was evaluated using the BLEU (Bilingual Evaluation Understudy) metric [44], a recognized standard, calculated with the "fairseq-score" command following word segmentation.

The specific procedure integrated into our methodology is outlined as follows:

1. Word segmentation and tokenization: Initially, Jieba was applied to segment Chinese sentences, followed by the use of Moses for tokenizing both Chinese and English sentences. This step transformed the original continuous text into individual words or tokens, allowing the model to better comprehend the structure and semantics of the text.
2. BPE: We utilized the subword-nmt tool for encoding the bilingual files with Byte-Pair Encoding. BPE was a method that encoded common word combinations or phrases into single tokens, which helped the model better understand and process low-frequency vocabulary, thus enhancing the accuracy of translation.
3. Length limit: The clean-corpus-n.perl tool in Moses was employed to clean the data and eliminate sentences exceeding 256 words. This step could reduce the computational burden on the model and avoid potential interference from overly long sentences during training. This step might have decreased the size of the dataset, but we believe it could improve the training efficiency and translation quality of the model.
4. Generate input text: We generated the vocabulary and binaries needed for model training using the "fairseq-preprocess" preprocessing function in fairseq.

By meticulously following these steps, we aimed to refine our model's training and prediction processes, enhancing the overall accuracy and quality of our NMT system.

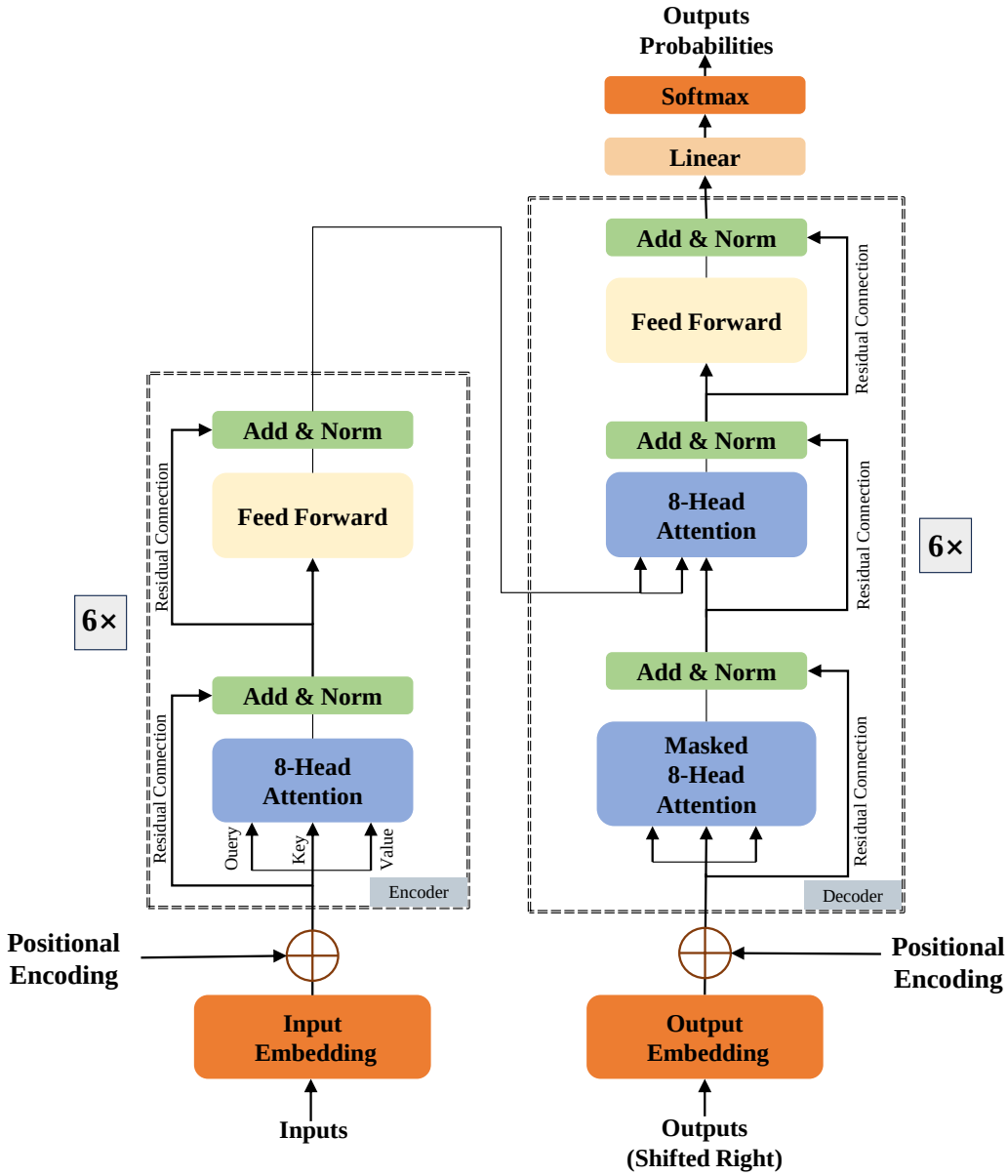


**Figure 3.** The architecture diagram of the transformers used in the experiment.

### *4.3. Evaluation*

#### 4.3.1. Machine Translation Performance and Analysis

We used BLEU scores, calculated with the fairseq-generate command, to evaluate our model's performance. As shown in Tables 3 and 4, the test data from the "V16 Dataset" is referred to as "V". Similarly, the test set from "WCC-EC 2.0-Full" is labeled as "W". Additionally, test data from the "WCC-EC 2.0-Lyrics" are tagged as "WL". The test set from the "WCC-EC 2.0-News" is abbreviated as "WN".

**Table 3.** English→Chinese translation results (BLEU scores), bold numbers indicate the best scores.

| Training Data | Test Data | | | | Average BLEU |
|---|---|---|---|---|---|
| | V | W | WL | WN | |
| V16 Dataset | 11.40 | 1.99 | 2.70 | 3.69 | 4.94 |
| WCC-EC 2.0-Full | 15.18 | **19.16** | 12.96 | 12.57 | **14.97** |
| WCC-EC 2.0-Lyrics | 2.11 | 5.04 | **16.46** | 1.87 | 6.37 |
| WCC-EC 2.0-News | **16.68** | 15.73 | 7.52 | **15.14** | 13.77 |

**Table 4.** Chinese→English translation results (BLEU scores), bold numbers indicate the best scores.

| Training Data | Test Data | | | | Average BLEU |
|---|---|---|---|---|---|
| | V | W | WL | WN | |
| V16 Dataset | 12.80 | 1.88 | 1.25 | 1.28 | 4.31 |
| WCC-EC 2.0-Full | **20.77** | **24.72** | 21.85 | 16.89 | **21.06** |
| WCC-EC 2.0-Lyrics | 2.02 | 7.25 | **23.42** | 1.74 | 8.61 |
| WCC-EC 2.0-News | 18.77 | 17.77 | 9.35 | **17.80** | 15.93 |

Due to the experimental English-Chinese dataset being categorized as low-resource domain material, the BLEU scores observed in our experiments were relatively low. In the English→Chinese translation task (referenced in Table 3), the translation results of version 16 of the WCC were relatively poor. However, WCC-EC 2.0, despite being a low-resource dataset, has a significantly large volume of data, which contributed to better translation outcomes compared to version 16. However, a decline in translation quality was noted in sections of the corpus related to news and music, with the news segment particularly experiencing a more marked decrease in quality. Surprisingly, the BLEU score for the model trained on WCC-EC 2.0-News in the "V" test set was higher than its score in its own test set. Although WCC-EC 2.0-Full did not achieve the lead across all test sets, the gap with other corpora was minimal. This can be attributed to our strict deduplication process, ensuring that the training set did not contain test sets from other corpora. Notably, the average result of WCC-EC 2.0-Full was the best among all the corpora considered, demonstrating significant generalizability.

In the Chinese→English translation experiment (as detailed in Table 4), WCC-EC 2.0-Full achieved the highest BLEU score of 24.72 on the W test set. This score was not only relatively high but also significantly surpassed the performance of V16, which scored 12.8 on the "V" test set. The superior performance of WCC-EC 2.0-Full, developed by our team, underscored its effectiveness in Chinese→English translation tasks. These results robustly demonstrated the utility of the WCC-EC 2.0 and its potential to enhance the accuracy of machine translation. Considering the BLEU score results as a whole, we believed the lower scores were particularly due to the inclusion of colloquial texts from the music domain. These texts tended to be shorter in length and contained less knowledge and contextual information, which likely led to lower BLEU scores.

#### 4.3.2. Manual Evaluation Results and Analysis

To substantiate the validity of the WCC-EC 2.0, we employed the evaluation criteria defined by the Japan Patent Office (JPO) (https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html (accessed on 15 February 2024)). This criterion assesses the fidelity of translated content through a five-tier grading system, where a score of 5 represents the highest level of accuracy. The nuances of the JPO grading methodology are comprehensively delineated in Table 5.

**Table 5.** Details of the JPO evaluation criteria.

| Grade | Evaluation Criteria |
|---|---|
| 5 | All important content has been accurately conveyed. (100%) |
| 4 | Most of the key content has been accurately conveyed. (Approximately 80%∼100%) |
| 3 | More than half of the core content is accurately conveyed. (About 50%∼80%) |
| 2 | Some important content has been successfully conveyed. (Roughly 20%∼50%) |
| 1 | The minimum basic content has been transmitted correctly. (Up to about 20%) |

During the manual evaluation phase, we enlisted experts fluent in both Chinese and English for an in-depth review. The participants' profiles, outlined in Table 6, encompass a

wide range of professionals from researchers to professors. Importantly, the evaluators who are native Chinese speakers held at least a master's degree and had significant exposure to academic English. Their proficiency in English was further validated by high scores on English proficiency exams, including CET-6, a TOEFL score of at least 90, and an IELTS score of 6.5 or higher. This combination of linguistic skill and scholarly background guarantees that our evaluation process is thorough and reliable.

During our manual evaluation, we placed particular emphasis on the consistency and accuracy of the evaluation criteria to ensure objectivity and fairness in the assessment results. Through these measures, we successfully improved the quality of the assessment results presentation. Additionally, during the data analysis process, rigorous checks and filtering were applied to the evaluation data to eliminate potential biases and errors. These efforts not only enhanced the reliability of the evaluation results, but also further elevated the quality of their presentation.

We undertook a manual evaluation of the translations that garnered the highest BLEU scores on the "W" test set, detailed in Tables 3 and 4, for both the English → Chinese and Chinese → English directions. To achieve this, we formed three evaluator groups, designated as X, Y, and Z, with their particulars specified in Table 6. Each group was chosen based on their unique areas of expertise to ensure a balanced and thorough assessment.

The analysis, depicted in Figures 4 and 5, showed minimal variation in the average scores among the groups, with the greatest difference being only 0.15. Significantly, Team X tended to assign marginally higher scores, whereas Teams Y and Z exhibited similar scoring trends. The consistency in scores from the latter two groups might have indicated that their evaluations were more informative. Given that all groups' scores exceeded the 4.0 mark, it suggested that the critical components of the translations were effectively conveyed. Additionally, considering the colloquial and accessible nature of the language used, we believed that the manual evaluations might have carried a tendency to award higher scores. These results affirmed the effectiveness and value of the WCC-EC 2.0 in producing high-quality translations, while also suggesting that the approachable language style contributed to somewhat more generous scoring in the manual evaluations.

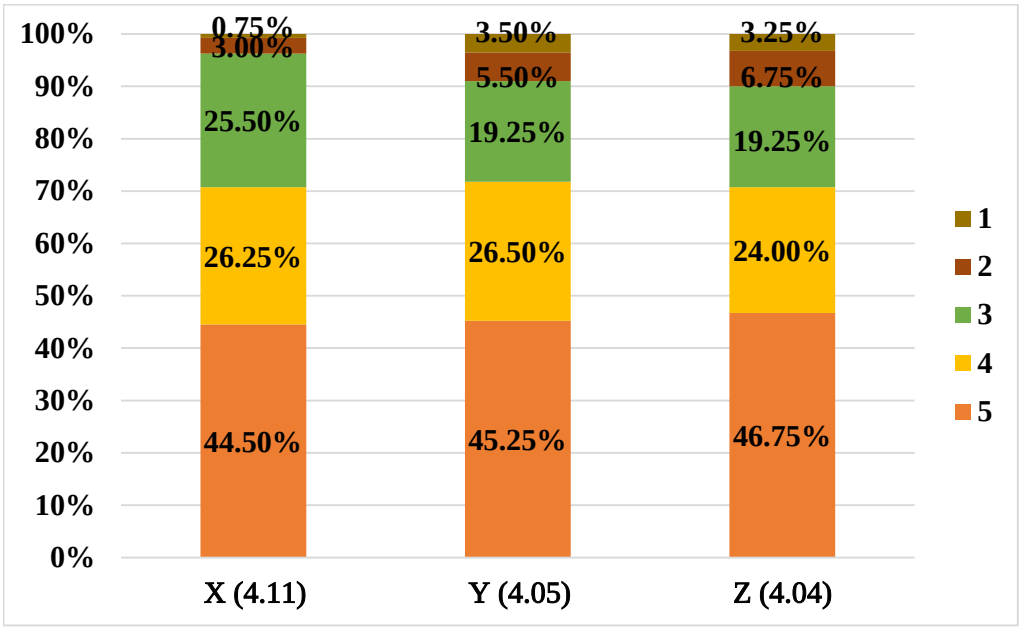


**Figure 4.** Manual evaluation results for the English → Chinese translation.

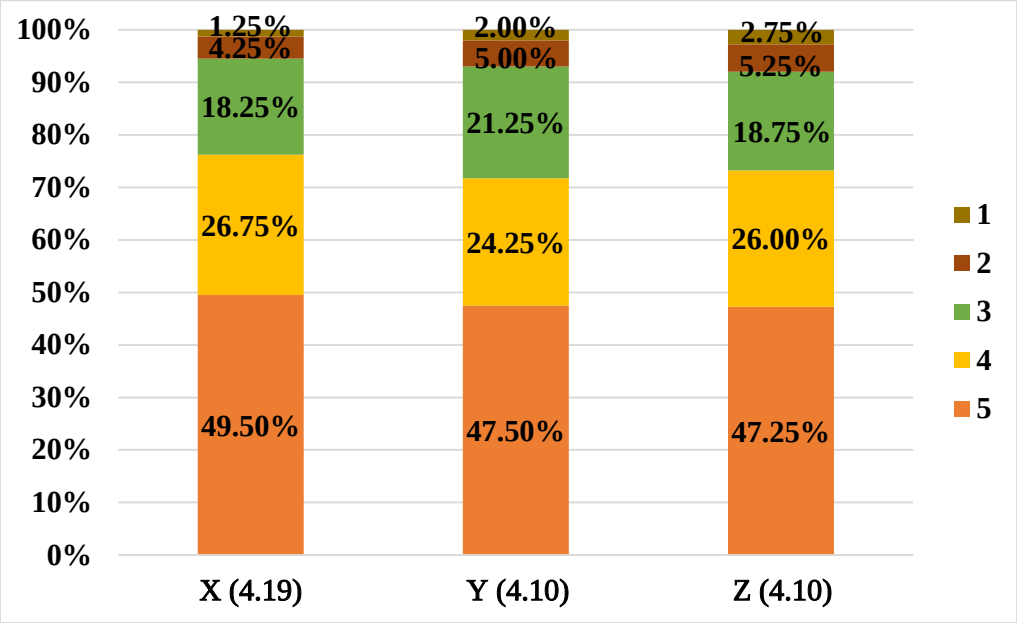


**Figure 5.** Manual evaluation results for the Chinese → English translation.

**Table 6.** Profiles of participants in the assessment.

| Attribute | Team X | Team Y | Team Z |
|---|---|---|---|
| **Age** | 34, 27 | 35, 31 | 30, 35 |
| **Gender** | Male, Male | Female, Male | Female, Male |
| **Occupation** | Lab Researcher, PhD Student | Asst. Prof., PhD Student | Media Staff, Asst. Prof. |
| **Language** | Excellent, Proficient | Excellent, Proficient | Proficient, Excellent |

## 5. Summary and Future Work

In this study, we outlined the detailed process used to develop the WCC-EC 2.0, a large bilingual English-Chinese corpus. With more than 1.6 million sentence pairs across the news and music sectors, this corpus was carefully put together and aligned. Notably, it achieves the highest average BLEU score across various test sets, complemented by an exceptional average manual evaluation score exceeding 4 points. Its large size and availability to the public make it one of the most comprehensive resources for NMT involving the English-Chinese language pair.

Given the online source of our data, we tackled potential copyright issues. The process of extracting and sharing content from the internet often involves the risk of copyright infringement, especially with materials that belong to others. To ensure we were following the law, we consulted with legal experts and confirmed that the WCC-EC 2.0 complied with Chinese copyright laws.

Our validation of the corpus included thorough English-Chinese translation experiments and detailed manual reviews, confirming its adaptability and textual authenticity. Future research will focus on employing data augmentation techniques to improve the corpus quality and open up new possibilities in NMT. Additionally, we plan to use the WCC-EC 2.0 for fine-tuning open-source LLMs such as BLOOM [45] and LLaMA 2 [46], thus boosting their translation performance. With the backing of the WCC-EC 2.0, we are set to make significant progress in English-Chinese translation. This advancement is expected to support cooperative projects in infrastructure development, trade dynamics, and policy coordination.

**Author Contributions:** Conceptualization, J.Z.; methodology, J.Z., Y.T. and T.M.; software, T.M.; validation, J.Z., K.S. and Y.T.; formal analysis, J.Z. and Y.T.; investigation, J.Z., Y.T., K.S. and T.M.; resources, J.Z., Y.T., K.S. and T.M.; data curation, J.Z., K.S. and Y.T.; writing—original draft preparation, K.S.; writing—review and editing, J.Z.; visualization, K.S.; supervision, J.Z. and T.M.; project administration, J.Z. and T.M.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the General Young Talents Project for Scientific Research grant of the Educational Department of Liaoning Province (Grant No. LJKZ0267) and the Research Support Program for Inviting High-Level Talents grant of Shenyang Ligong University (Grant No. 1010147001004). Jinyi Zhang is funded by the China Scholarship Council (No. 202208210120).

**Data Availability Statement:** The demo version of the WCC-EC 2.0, comprising 200,000 sentence pairs, along with the code utilized in this study, is available for public access on GitHub (https://github.com/zhang-jinyi/Web-Crawled-Corpus-for-English-Chinese-NMT (accessed on 15 February 2024)). For full access to the dataset, interested parties are encouraged to reach out via the email provided on the GitHub page, with the understanding that the data are intended solely for personal and research purposes.

**Acknowledgments:** This article benefits from the advanced capabilities of the OpenAI's GPT-4 model [6]. We acknowledge and value its significant contributions.

**Conflicts of Interest:** Among the authors, Ye Tian was employed by Zhuzhou CRRC Times Electric Co., Ltd., while the remaining authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; MIT Press: Cambridge, MA, USA, 2014; pp. 3104–3112.
2. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; ACL: Lisbon, Portugal, 2015; pp. 1412–1421. [CrossRef]
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; pp. 5998–6008.
4. Stahlberg, F. Neural Machine Translation: A Review. *J. Artif. Intell. Res.* **2020**, *69*, 343–418. [CrossRef]
5. OpenAI. OpenAI Blog: Introducing ChatGPT, 30 November 2022. Available online: https://openai.com/blog/chatgpt (accessed on 15 February 2024).
6. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774.
7. Sanz-Valdivieso, L.; López-Arroyo, B. Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation? In Proceedings of the International Conference Human-informed Translation and Interpreting Technology, Naples, Italy, 7–9 July 2023.
8. Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Kong, L.; Chen, J.; Li, L.; Huang, S. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *arXiv* **2023**, arXiv:2304.04675. https://doi.org/10.48550/arXiv.2304.04675.
9. Hendy, A.; Abdelrehim, M.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y.J.; Afify, M.; Awadalla, H.H. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv* **2023**, arXiv:2302.09210. https://doi.org/10.48550/arXiv.2302.09210.
10. Vilar, D.; Freitag, M.; Cherry, C.; Luo, J.; Ratnakar, V.; Foster, G.F. Prompting PaLM for Translation: Assessing Strategies and Performance. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Toronto, ON, Canada; pp. 15406–15427. [CrossRef]
11. Zhang, B.; Haddow, B.; Birch, A. Prompting Large Language Model for Machine Translation: A Case Study. In Proceedings of the International Conference on Machine Learning, ICML, Honolulu, HI, USA, 23–29 July 2023; Proceedings of Machine Learning Research (PMLR); Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; Volume 202, pp. 41092–41110.
12. Jiao, W.; Wang, W.; Huang, J.; Wang, X.; Tu, Z. Is ChatGPT A Good Translator? A Preliminary Study. *arXiv* **2023**, arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745.

13. Robinson, N.R.; Ogayo, P.; Mortensen, D.R.; Neubig, G. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, 6–7 December 2023; Koehn, P., Haddon, B., Kocmi, T., Monz, C., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 392–418.
14. Karpinska, M.; Iyyer, M. Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, 6–7 December 2023; Koehn, P., Haddon, B., Kocmi, T., Monz, C., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 419–451.
15. Wang, L.; Lyu, C.; Ji, T.; Zhang, Z.; Yu, D.; Shi, S.; Tu, Z. Document-Level Machine Translation with Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 16646–16661.
16. Xu, H.; Kim, Y.J.; Sharaf, A.; Awadalla, H.H. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *arXiv* **2023**, arXiv:2309.11674. https://doi.org/10.48550/arXiv.2309.11674.
17. Yang, W.; Li, C.; Zhang, J.; Zong, C. BigTrans: Augmenting Large Language Models with Multilingual Translation Capability over 100 Languages. *arXiv* **2023**, arXiv:2305.18098. https://doi.org/10.48550/arXiv.2305.18098.
18. Bawden, R.; Yvon, F. Investigating the Translation Performance of a Large Multilingual Language Model: The Case of BLOOM. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12–15 June 2023; Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S.A., Aranberri, N., et al., Eds.; European Association for Machine Translation: Sheffield, UK, 2023; pp. 157–170.
19. Lu, H.; Huang, H.; Zhang, D.; Yang, H.; Lam, W.; Wei, F. Chain-of-Dictionary Prompting Elicits Translation in Large Language Models. *arXiv* **2023**, arXiv:2305.06575. https://doi.org/10.48550/arXiv.2305.06575
20. Ranathunga, S.; Lee, E.A.; Skenduli, M.P.; Shekhar, R.; Alam, M.; Kaur, R. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.* **2023**, *55*, 229:1–229:37. [CrossRef]
21. Zhang, J.; Guo, C.; Mao, J.; Guo, C.; Matsumoto, T. An Enhanced Method for Neural Machine Translation via Data Augmentation Based on the Self-Constructed English-Chinese Corpus, WCC-EC. *IEEE Access* **2023**, *11*, 112123–112132. [CrossRef]
22. Tiedemann, J. OPUS-parallel corpora for everyone. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products, EAMT 2016, Riga, Latvia, 30 May–1 June 2016.
23. Villalobos, P.; Sevilla, J.; Heim, L.; Besiroglu, T.; Hobbhahn, M.; Ho, A. Will we run out of data? An analysis of the limits of scaling datasetdatasets in Machine Learning. *arXiv* **2022**, arXiv:2211.04325.
24. Mackenzie, J.M.; Benham, R.; Petri, M.; Trippas, J.R.; Culpepper, J.S.; Moffat, A. CC-News-En: A Large English News Corpus. In Proceedings of the CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, 19–23 October 2020; d'Aquin, M., Dietze, S., Hauff, C., Curry, E., Cudré-Mauroux, P., Eds.; ACM: New York, NY, USA, 2020; pp. 3077–3084. [CrossRef]
25. Lefer, M.A. Parallel corpora. In *A Practical Handbook of Corpus Linguistics*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 257–282.
26. Bañón, M.; Chen, P.; Haddow, B.; Heafield, K.; Hoang, H.; Esplà-Gomis, M.; Forcada, M.L.; Kamran, A.; Kirefu, F.; Koehn, P.; et al. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 4555–4567. [CrossRef]
27. Ziemski, M.; Junczys-Dowmunt, M.; Pouliquen, B. The United Nations Parallel Corpus v1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, 23–28 May 2016; Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., et al., Eds.; European Language Resources Association (ELRA): Paris, France, 2016.
28. Liu, B.; Huang, L. NEJM-enzh: A Parallel Corpus for English-Chinese Translation in the Biomedical Domain. *arXiv* **2020**, arXiv:2005.09133. https://doi.org/10.48550/arXiv.2005.09133.
29. Liu, Z.; Wang, H.; Niu, Z.; Wu, H.; Che, W. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event/Punta Cana, Dominican Republic, 7–11 November, 2021; Moens, M., Huang, X., Specia, L., Yih, S.W., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 4335–4347. [CrossRef]
30. Zhang, J.; Tian, Y.; Mao, J.; Han, M.; Matsumoto, T. WCC-JC: A web-crawled corpus for Japanese-Chinese neural machine translation. *Appl. Sci.* **2022**, *12*, 6002. [CrossRef]
31. Zhang, J.; Tian, Y.; Mao, J.; Han, M.; Wen, F.; Guo, C.; Gao, Z.; Matsumoto, T. WCC-JC 2.0: A Web-Crawled and Manually Aligned Parallel Corpus for Japanese-Chinese Neural Machine Translation. *Electronics* **2023**, *12*, 1140. [CrossRef]
32. Sugiyama, A.; Yoshinaga, N. Data augmentation using back-translation for context-aware neural machine translation. In Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, 3 November 2019; Popescu-Belis, A., Loáiciga, S., Hardmeier, C., Xiong, D., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 35–44. [CrossRef]
33. Li, X.; Yu, P.; Zhou, C.; Schick, T.; Zettlemoyer, L.; Levy, O.; Weston, J.; Lewis, M. Self-Alignment with Instruction Backtranslation. *arXiv* **2023**, arXiv:2308.06259. https://doi.org/10.48550/arXiv.2308.06259

34. Morita, T.; Akiba, T.; Tsukada, H. Hybrid Sampling for Iterative Back-Translation of Neural Machine Translation. *IEICE Trans. Inf. Syst. (Jpn. Ed.)* **2023**, *106*, 298–306.
35. Zhang, J.; Matsumoto, T. Corpus Augmentation for Neural Machine Translation with Chinese–Japanese Parallel Corpora. *Appl. Sci.* **2019**, *9*, 2036. [CrossRef]
36. Dou, Z.Y.; Neubig, G. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv* **2021**, arXiv:2101.08231.
37. Church, K.W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162. [CrossRef]
38. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. *arXiv* **2016**, arXiv:1607.01759.
39. Li, X.; Meng, Y.; Sun, X.; Han, Q.; Yuan, A.; Li, J. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 28 July–2 August 2019; Volume 1: Long Papers; Korhonen, A., Traum, D.R., Màrquez, L., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 3242–3252. [CrossRef]
40. Jiang, C.; Maddela, M.; Lan, W.; Zhong, Y.; Xu, W. Neural CRF Model for Sentence Alignment in Text Simplification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 7943–7960. [CrossRef]
41. Zhang, B.; Nagesh, A.; Knight, K. Parallel Corpus Filtering via Pre-trained Language Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 8545–8554. [CrossRef]
42. Cao, S.; Kitaev, N.; Klein, D. Multilingual Alignment of Contextual Word Representations. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
43. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 2019, Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 3980–3990. [CrossRef]
44. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318. [CrossRef]
45. Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilic, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv* **2022**, arXiv:2211.05100. https://doi.org/10.48550/arXiv.2211.05100.
46. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. https://doi.org/10.48550/arXiv.2307.09288.