applied sciences

MDPI

Article

WCC-JC: A Web-Crawled Corpus for Japanese-Chinese Neural Machine Translation

Jinyi Zhang [1,*], Ye Tian [2], Jiannan Mao [3], Mei Han [4] and Tadahiro Matsumoto [3]

[1] School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China
[2] Zhuzhou CRRC Times Electric Co., Ltd., Zhuzhou 412001, China; tianye@csrzic.com
[3] Faculty of Engineering, Gifu University, Gifu 501-1193, Japan; mao@mat.info.gifu-u.ac.jp (J.M.); tad@gifu-u.ac.jp (T.M.)
[4] School of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou 412007, China; hanmei@hut.edu.cn
* Correspondence: zhangjinyi@sylu.edu.cn

Featured Application: This research crawled a bilingual Japanese-Chinese corpus of a certain size through websites. As a necessary resource for Japanese-Chinese neural machine translation (NMT), it is beneficial for researchers to promote the progress of Japanese-Chinese language-related natural language processing research. Specifically, topics include comparative analysis of grammar, comparative studies of Chinese and Japanese languages, compilation of dictionaries, etc. This will have great significance and contribution to the cultural exchange and industrial cooperation between China and Japan. It also has important theoretical significance and application value to the industrialization of Japanese-Chinese machine translation. In addition, the application of this research will be of great significance in strengthening civil communication and enhancing mutual understanding between China and Japan, as the current Chinese and Japanese relations are not well perceived by the citizens of both countries. We hope that the construction and pathways of the Japanese-Chinese bilingual corpus in this research will help to solve the problem of language barriers in Japanese-Chinese people-to-people communication and mutual understanding. We offer the WCC-JC as a free download under the premise that it is intended for research purposes only.

Abstract: Currently, there are only a limited number of Japanese-Chinese bilingual corpora of a sufficient amount that can be used as training data for neural machine translation (NMT). In particular, there are few corpora that include spoken language such as daily conversation. In this research, we attempt to construct a Japanese-Chinese bilingual corpus of a certain scale by crawling the subtitle data of movies and TV series from the websites. We calculated the BLEU scores of the constructed WCC-JC (Web Crawled Corpus—Japanese and Chinese) and the other compared corpora. We also manually evaluated the translation results using the translation model trained on the WCC-JC to confirm the quality and effectiveness.

Keywords: Japanese-Chinese bilingual corpus; neural machine translation; corpus construction; manual evaluation

Citation: 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. https://doi.org/10.3390/app12126002

Received: 6 May 2022
Accepted: 10 June 2022
Published: 13 June 2022

1. Introduction

In recent years, Japan has been the second largest trading partner of China, and China is the largest trading partner of Japan. As an important contributor to cultural exchange and industrial cooperation, China and Japan have been paying attention to each other. However, the language barrier problem has become a serious challenge that limits China and Japan from strengthening their communications. Both countries expect a high-quality Chinese-Japanese machine translation system. Therefore, the establishment of a high-quality Chinese-Japanese machine translation system is helpful for the language

barrier arising from Chinese-Japanese exchange activities. It is of great significance to the development of both Japan and China in terms of technology and economy.

Machine translation is an important field of artificial intelligence research which translates source language into target language. It is one of the most effective methods for solving language barriers. After years of development, NMT is a new machine translation model with great potential that has exhibited superior translation results to the traditional frameworks for various language pairs. NMT is capable of training large-scale parallel corpora, although the amount of training data greatly affects the quality of translation results. NMT has a great potential for industrialization, as well as significant research value, making it the most advanced hotspot in machine translation research today.

In the field of machine translation, Chinese-Japanese machine translation is difficult due to the complex intertwining of the two languages. In order to obtain high-quality translations, a large amount of data in the Japanese-Chinese bilingual corpus is required. There are tens or hundreds of millions of sentence pairs, including English, and language pairs existing between European languages. It contain sentences from many different fields. However, there is a lack of Japanese-Chinese bilingual corpora that have been made public. For example, the Japanese Patent Office (JPO) Japanese-Chinese bilingual corpus has 130 million entries (about 26 GB) and 0.1 billion entries (about 1.4 GB), but research results using all of the data need to be submitted (https://alaginrc.nict.go.jp/jpo-outline.html (accessed on 30 June 2017)), and the only type of content in this corpus is patented. The ASPEC-JC corpus contains approximately 670k sentences [1]. This corpus contains only abstracts of scientific papers. The above two corpora are already have a very large number of sentences among the publicly available Japanese-Chinese corpora. However, there are still not many Japanese-Chinese bilingual corpora suitable for translating sentences from daily-use spoken language.

Currently, most of the research on NMT focuses on improving the translation quality of translation models. However, the lack of corpus with the basic properties remains an insurmountable problem. Although approaches such as back translation have been proposed to generate pseudo-data, they still cannot reach the essence of the problem. It is necessary to perform this basic research and lead the way by releasing the data. Professor Feifei Li of Stanford University had released the famous ImageNet dataset. This dataset led the field of computer vision and triggered the wave of AI that is sweeping the world today [2]. Professor Feifei Li said "One thing ImageNet changed in the field of AI is suddenly people realized the thankless work of making a dataset was at the core of AI research". This is exactly what our research was originally intended to do.

Most of the current research focuses on the translation of sentences that fall into the category of written language. On the other hand, research on spoken language is more ambiguous than written language, and it requires a greater understanding of context. For instance, the spoken language is usually shorter in length than written sentences. Even in spoken language, slang words and dialects may be overlooked by conventional translators. In addition, multi-modal translation is recognized as a future research trend, and a suitable bilingual corpus of spoken language is needed.

In this article, our research goal is to construct a Japanese-Chinese bilingual corpus by crawling data from websites. The corpus (WCC-JC) is constructed by crawling the data from movie, anime, and TV series subtitles. Then, the subtitles are mapped to Japanese and Chinese languages. The corpus also covers spoken language bilingualism, which has not been well researched in the existing Chinese-Japanese bilingual corpora. After the Japanese-Chinese bilingual corpus is completed, it is expected to attract researchers' attention as a crucial resource for Japanese-Chinese NMT and to promote the progress of NMT research.

In the remainder of this article, Section 2 presents the related works. Section 3 describes the construction of WCC-JC. Section 4 reports the experimental framework, the results obtained from Japanese-Chinese and Chinese-Japanese translation experiments, and the manual evaluations. Section 5 discusses the legitimacy of WCC-JC. Finally, Section 6 concludes with a discussion of the contributions of this article and the future works.

2. Related Works

The corpus was systematically introduced and analyzed in Stefanowitsch Anatol's book *Corpus linguistics: A guide to the methodology* [3]. Several examples from corpus linguistics are surveyed to show how they fit into the outlined methodology.

There are very few Japanese-Chinese bilingual corpora in the public domain, which only the previously mentioned JPO Japanese-Chinese bilingual corpus, ASPEC-JC [1], TED Talks (https://ted.com, https://github.com/ajinkyakulkarni14/TED-Multilingual-Parallel-Corpus (both accessed on 10 June 2020)), and others have.

The Japanese-Chinese bilingual corpus at Beijing Normal University contains about 80 full texts of novels, poems, essays, biographies, political commentaries, and legal treatises from the modern to the contemporary period [4]. However, due to copyright restrictions, this corpus was only partially available for research, and could not be trained and used in NMT.

Zhang et al. constructed a Japanese-Chinese parallel corpus by human translation, which was part of the NICT multilingual corpus. The quality was high and annotated, but there were only about 40,000 sentence pairs [5].

Koehn collected a corpus of parallel text in 11 languages from the proceedings of the European Parliament, which are published on the Web. This corpus had found widespread use in the NLP community [6].

Lavecchia et al. proposed a method to automatically build a bilingual corpus from movie subtitle files, and also created a translation table by the method [7].

Baroni et al. introduced ukWaC, deWaC, and itWaC, three very large corpora of English, German, and Italian built by Web crawling, and described the methodology and tools used in their construction. The corpora contained more than a billion words each, and were thus among the largest resources for the respective languages [8].

Smith et al. used their open-source extension of the STRAND algorithm-mined 32 terabytes of the Common Crawl, a public Web crawl hosted on Amazon's Elastic Cloud. Even with minimal cleaning and filtering, the resulting data boosted translation performance across the board for five different language pairs in the news domain [9].

Chu et al. proposed a bilingual sentence extraction system to construct a Japanese-Chinese bilingual corpus from Wikipedia [10]. Using the system, they constructed a Japanese-Chinese bilingual corpus containing more than 126k highly accurate bilingual sentences from Wikipedia.

Benko reported on the first phase of an ongoing project to create a Web corpus and summarized the problems encountered in the process [11].

The United Nations Parallel Corpus v1.0 was composed of official records and other parliamentary documents of the United Nations that were in the public domain. These documents were mostly available in the six official languages of the United Nations. The current version of the corpus contained content that was produced and manually translated between 1990 and 2014, including sentence-level alignments [12].

Pryzant et al. constructed a Japanese-English correspondence database of movie and TV program subtitles crawled from the Internet. The JESC was the largest freely available Japanese-English bilingual corpus (about 2.8 million sentences) [13].

OpenSubtitles2018 was a multilingual parallel corpus of movie subtitle data [14]. The Japanese-English bilingual corpus was a parallel corpus of two million sentences consisting of approximately 2000 movies, and will be considered for use in the field of machine translation and other tasks that take advantage of the characteristics of movie subtitles.

Park et al. proposed a simple, linguistically motivated solution to improve the performance of Korean–Chinese neural machine translation models by using a common vocabulary [15].

Morishita et al. constructed JParaCrawl, a large-scale Web-based Japanese-English bilingual corpus, by crawling the Web on a large scale, automatically collecting Japanese-English bilingual sentences, and filtering out noisy bilingual pairs [16].

Guokun et al. automatically built a corpus by crawling language resources from the Internet, but the data were not well filtered [17].

Hasan et al. built a sentence segmentation method for Bengali, a sparse language, and constructed a non-English bilingual corpus [18]. Thus, the construction of an open Japanese-Chinese bilingual corpus for NMT has significant implications on the resource scarcity problem.

Václav et al. compared two corpora of Czech. One was a traditional corpus and the other was a Web-crawled corpus, which had been extensively compared and analyzed for quality [19].

The EuroparlTV Multimedia Parallel Corpus (EMPAC) was a collection of subtitles in English and Spanish for videos from the European Parliament's Multimedia Centre. The corpus covered a time span from 2009 to 2017 and it was made up of 4000 texts amounting to two and half million tokens for every language, corresponding to approximately 280 h of video [20].

Nakazawa et al. introduced the BSD corpus and the results of the BSD translation task in WAT2020. Additionally, they discussed the challenges of dialogue translation based on the analysis of the translation results [21]. The BSD corpus was constructed using "dialogue" in "business" as the domain of the bilingual corpus.

Liu et al. conducted a parallel corpus in the field of biomedicine for English-Chinese translation [22]. They compared the effectiveness of different algorithms/tools for sentence boundary detection and sentence alignment, and used the constructed corpus, fine-tuning the NMT models.

Dou et al. used pretrained language models, but by fine-tuning them on parallel texts with the aim of improving alignment quality, they proposed a method for efficiently extracting alignments from these fine-tuned models and demonstrated that their models can consistently outperform all previous state-of-the-art models of the species [23].

Schwenk et al. presented an approach based on multilingual sentence embeddings to automatically extract parallel sentences from the content of Wikipedia articles in 96 languages, extracting 135M parallel sentences for 16,720 different language pairs, and achieving strong BLEU scores for many language pairs [24].

For Japanese-Chinese translation, Zhang et al. proposed the following three data augmentation methods to improve the quality of Japanese-Chinese NMT: (1) radicals as an additional input feature [25]; (2) the created Chinese character decomposition table [26]; (3) a corpus augmentation approach [27], considering the lack of resources in bilingual corpora.

The related works above are sorted by year. We also show the classification of these related works on five aspects: (1) Corpus linguistics; (2) Japanese-Chinese bilingual corpora; (3) Web-crawled corpora; (4) other corpora; (5) corpus augmentation through a summarized Table 1.

Table 1. Summary of related works.

Classification	Related Works
Corpus linguistics	[3]
Japanese-Chinese bilingual corpora	[1], TED talks, [4,5]
Web-crawled corpora	[7–11,13,14,16,17,19,24]
Other corpora	[6,12,18,20–22]
Corpus augmentation	[15,23,25–27]

The above related research showed that corpora play an important role in improving translation accuracy and in other directions of language processing. Thus, the construction of a Japanese-Chinese bilingual corpus for NMT has significant implications for the resource

scarcity problem. In this research, we aim to develop Japanese-Chinese machine translation and construct a Japanese-Chinese bilingual corpus of a certain scale.

3. Construction of Japanese-Chinese Bilingual Corpus

The corpus to be constructed, WCC-JC, is a collection of Japanese-Chinese bilingual sentences from the Web. This method discovers websites that may contain Japanese-Chinese bilingual sentences, and attempts to extract bilingual sentences from the Web data.

3.1. Web Crawling

Considering a website that contains many Japanese-Chinese bilingual texts, we use Scrapy (https://scrapy.org/ (accessed on 10 October 2021)) to retrieve subtitle files from the website (http://assrt.net/ (accessed on 10 June 2020)) that contains subtitle files of movies, dramas, and TV series. In these subtitle files, there are bilingual translations of slang, spoken language, explanatory text, and story commentary. These are areas that have not been dealt with much in the existing corpora.

3.2. Extraction of Bilingual Sentences

Most of the acquired subtitle files are advanced SubStation Alpha (ASS) files. Figure 1 shows an example of the contents of an ASS file (dummy contents). As shown in the figure, in the dialogue line of the ASS file, the information of display start time, end time, style, subtitle display content, and so on are described. The language information often appears in the layer field (the first "0" in the figure) and the style field ("DefaultJp" in the figure).

```
Dialogue: 0,1:44:02.47,1:44:05.56,DefaultJp,,0,0,0,,{\blur4}すみません、タクシー一台急いでお願いします。
Dialogue: 0,1:44:08.61,1:44:13.53,DefaultJp,,0,0,0,,{\blur4}すぐ行きますが、何か目印を教えてください。
Dialogue: 0,1:44:14.19,1:44:16.66,DefaultJp,,0,0,0,,{\blur4}山手通りを渋谷方面へ来てください。
Dialogue: 0,1:44:48.27,1:44:50.40,DefaultJp,,0,0,0,,{\blur4}そうすると、东急线の踏切があるから、それを渡ると、大きな交差点に出ます。
Dialogue: 0,1:44:50.98,1:44:54.99,DefaultJp,,0,0,0,,{\blur4}路线の下を通るんですか。
Dialogue: 0,1:44:56.28,1:45:01.58,DefaultJp,,0,0,0,,{\blur4}じゃなくて、踏切。
Dialogue: 0,1:45:02.45,1:45:03.16,DefaultJp,,0,0,0,,{\blur4}で、初めの大きな交差点を右ね。
Dialogue: 0,1:45:09.18,1:45:10.06,DefaultJp,,0,0,0,,{\blur4}しばらく来ると、左側にお寺があります。
Dialogue: 0,1:45:17.69,1:45:20.57,DefaultJp,,0,0,0,,{\blur4}その手の前の细い道を左。
Dialogue: 0,1:45:23.53,1:45:23.78,DefaultJp,,0,0,0,,{\blur4}突きあたったら、右に曲がってください。
Dialogue: 0,1:45:24.82,1:45:25.45,DefaultJp,,0,0,0,,{\blur4}三轩目の左側の家です。
Format: Layer, Start,    End,      Style, Name, MarginL, MarginR, MarginV, Effect, Text
```

Figure 1. An example of the contents of an ASS file (dummy content, the main story is about a conversation asking for the directions).

If there is a dialogue line containing one of "ja", "jp", "日 (Japan)" in the style name, the style is judged to be Japanese.

If there is a dialogue line containing one of "cn", "ch", "zh", "中 (China)", or "default" in the style name, the style is considered to be Chinese. If neither Japanese style nor Chinese style is found, it is judged that the subtitle file does not contain Japanese-Chinese bilingual subtitles.

After determining whether the dialogue lines are Japanese or Chinese, the dialogue lines with Japanese style and Chinese style are sorted in ascending order by start time (value of the start field in dialogue lines), respectively. The correspondence is stored by extracting the corresponding bilingual sentences in the timeline of start (value of the start field in dialogue lines) and end (value of the end field in dialogue lines). If the timelines do not correspond exactly, we also consider contextual one-to-many and many-to-many relationships to check whether the timelines between multiple sentence pairs correspond correctly. If the timelines of multiple sentence pairs correspond accurately, we treat the multiple sentence pairs as one bilingual pair. This allows corresponding to as many pairs of bilingual sentences as possible. Finally, we obtained the raw parallel corpus.

3.3. Text Processing

Before corpus segmentation, the text in the raw parallel corpus needs to be normalized. We perform text normalization by changing traditional Chinese characters to simplified Chinese characters and using zenhan (https://pypi.org/project/zenhan/ (accessed on 10 October 2021)) to normalize Japanese katakana to full-width format, respectively. Finally, the text is sorted and duplicates are removed to obtain the filtered parallel corpus.

3.4. Corpus Segmentation

Since the generated corpus WCC-JC consists of subtitle data of many works, it is necessary to extract validation data (development data) and test data for NMT. Therefore, using other corpora as a reference, we decided that 2000 sentence pairs of ten characters or more were randomly extracted twice as the development data and the test data from the collected sentence pairs, and the remaining sentence pairs were used as training data.

Figure 2 shows the whole workflow of the corpus construction. It is mainly divided into four steps: (1) Web crawling; (2) extraction; (3) text processing; (4) corpus segmentation. The above are also the contents of Sections 3.1–3.4.

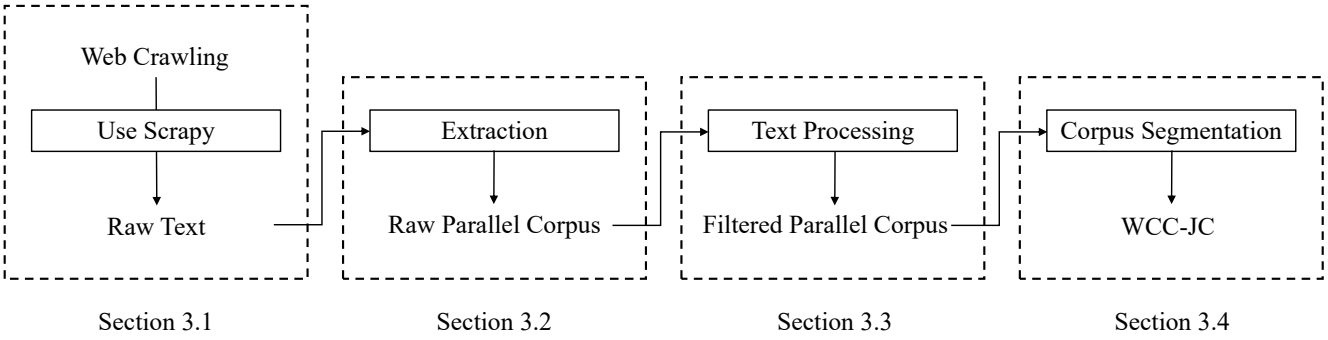

Figure 2. Workflow for the creation of the corpus from raw subtitle files.

Table 2 shows the number of sentences in the ASPEC-JC, OpenSubtitles, and the constructed corpus. It can also be seen that, in terms of capacity, ASPEC-JC > OpenSubtitles > WCC-JC. This is mainly due to the length of the sentences in the corpus. Figure 3 shows a right-skewed sentence length distribution. The average length of Chinese sentences and Japanese sentences are 10.6 and 13.7, respectively. The content of the WCC-JC is composed of spoken subtitles, and the sentences are generally shorter. Even though the number of sentences is higher than ASPEC-JC, it does not contain as much information as ASPEC-JC.

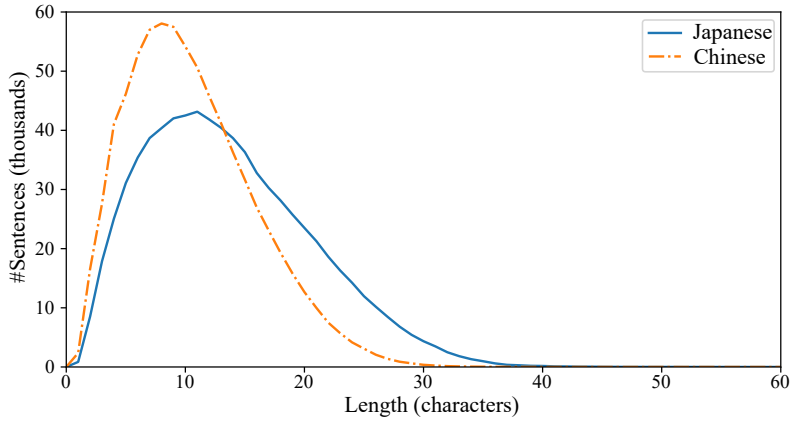

Figure 3. WCC-JC exhibits a right-skewed sentence length distribution: 5 Chinese and 14 Japanese sentences have length >60.

Table 2. Japanese-Chinese bilingual corpora.

Contents	Number of Bilingual Sentences		
	ASPEC-JC (184.8 MB)	OpenSubtitles (72.4 MB)	WCC-JC (52.2 MB)
Training data	672,315	1,087,295	749,017
Development data	2090	2000	2000
Test data	2107	2000	2000

4. Experiment and Evaluation

In order to confirm the effectiveness of this corpus, we conducted experiments. In Section 4.1, we configured the NMT system of the following experiments. In Section 4.2, we calculated the BLEU scores of ASPEC-JC, OpenSubtitles, and the constructed corpus with their own generated NMT models, and used JPO scores to manually evaluate the test data W's translation results.

We showed the ability of the constructed corpus to generalize to different domains. For Japanese-Chinese translation, the character level and LSTM models were found to be the most effective [28]. In our experiments, we compared character level, subword level, LSTM model, and transformer model, respectively.

4.1. Configuring the NMT System

In this experiment, we trained the NMT model using fairseq [29]. We used two predefined architectures of fairseq, lstm-wiseman-iwslt-de-en [30], and transformer-iwslt-de-en [31], as the LSTM model and the transformer model. The LSTM model had the embedding size of 512, 1 encoder layer, and 1 decoder layer. The transformer model had the embedding size of 512, 6 encoder layers with 8 encoder attention heads, and 6 decoder layers with 8 decoder attention heads. The two models' remaining parameters were the same, the dropout rate of 0.1; the Adam optimizer with betas of 0.9 and 0.98; the learning rate of 1×10^{-7}; the max tokens of 4096; the max update steps of 150,000; the batch size of 128; and for the translation process, the beam size of 5. Subword-level realization used subword-nmt (https://github.com/rsennrich/subword-nmt (accessed on 23 May 2017)), with the vocabulary size of 32,000.

Because sentences in Japanese and Chinese are written without spaces, we tokenized them with MeCab (http://taku910.github.io/mecab (accessed on 4 August 2017)) for Japanese and Jieba (http://github.com/fxsjy/jieba (accessed on 4 August 2017)) for Chinese.

BiLingual Evaluation Understudy (BLEU) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another [32]. The BLEU scores were calculated for each experiment, using the "fairseq-score" command after word segmentation. In other words, we took the word-level evaluation.

4.2. Evaluation

In this section, the human evaluators are native Chinese and Japanese. The Chinese evaluators had studied abroad in Japan and had passed the N2 or higher level of the Japanese Language Proficiency Test (JLPT).

4.2.1. Alignment Evaluation

We checked the validity of bilingual sentence alignments based on the procedure of [33]. Human evaluators randomly sampled 1000 sentence pairs. On average, 88% of these pairs were perfectly aligned, 7% partially aligned, and 5% misaligned. Thus, we may conclude that WCC-JC is noisy but has a significant signal that there is a lot of room for improvement.

4.2.2. Translation Evaluation

In addition to alignment, we evaluated the quality of the translations of WCC-JC. Our evaluators used the JPO adequacy criterion with the level of content transfer. This is a subjective evaluation of how accurately the machine translation results convey the substantive content of the source text in light of the content of the reference translation, based on a 5-point (5 being the best and 1 being the worst) system for scoring the quality of a translation pair (https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html (accessed on 5 September 2021)). Table 3 shows the JPO adequacy criterion from 5 to 1. This criterion is also used in the next Section 4.2.3.

Table 3. The JPO adequacy criterion.

Scores	Scoring Criteria
5	All important information is transmitted correctly. (100%)
4	Almost all important information is transmitted correctly. (80%∼)
3	More than half of important information is transmitted correctly. (50%∼)
2	Some of important information is transmitted correctly. (20%∼)
1	Almost all important information is NOT transmitted correctly. (∼20%)

We sampled and evaluated 1000 sentence pairs from the pool of non-misaligned sentences, observing an average JPO adequacy score of 4.67, implying that the amateur and crowd-sourced translations are high quality.

4.2.3. Machine Translation Performance and Manual Evaluation

In Tables 4–7, A was the test data of ASPEC-JC; O was the test data of OpenSubtitles; W was the test data of WCC-JC and N was the test data of 185 sentences extracted from the text of NHK Radio's "まいにち中国語 (Everyday Chinese)" (https://www2.nhk.or.jp/gogaku/onayami/chinese/ (accessed on 15 November 2021)) to verify the generalization capability of the translation models. The two predefined architectures of fairseq, lstm-wiseman-iwslt-de-en and transformer-iwslt-de-en were abbreviated as LSTM and Transformer, respectively.

Table 4. Japanese→Chinese translation experiment results in character level (BLEU values).

Data\Method	Character Level							
	LSTM				Transformer			
Test data→	A	O	W	N	A	O	W	N
ASPEC-JC	**32.0**	1.0	2.6	1.9	**34.5**	1.2	3.2	4.9
OpenSubtitles	0.3	**5.0**	2.4	7.1	0.0	2.1	0.0	3.0
WCC-JC	2.6	2.7	**12.1**	**9.5**	3.2	**3.9**	**15.9**	**13.7**

Table 5. Japanese→Chinese translation experiment results in subword level (BLEU values).

Data\Method	Subword Level							
	LSTM				Transformer			
Test data→	A	O	W	N	A	O	W	N
ASPEC-JC	**31.3**	1.1	2.3	3.0	**34.0**	1.2	3.3	6.6
OpenSubtitles	0.1	**4.3**	2.1	5.7	0.0	**3.5**	0.9	4.8
WCC-JC	1.9	2.4	**11.3**	**7.1**	2.0	2.9	**14.3**	**10.3**

Table 6. Chinese→Japanese translation experiment results in character level (BLEU values).

Data\Method	Character Level							
	LSTM				Transformer			
Test data→	A	O	W	N	A	O	W	N
ASPEC-JC	**38.8**	1.2	2.8	2.8	**44.8**	1.7	4.2	5.1
OpenSubtitles	0.2	**5.5**	4.0	5.6	0.1	**4.0**	2.3	4.1
WCC-JC	3.2	3.4	**13.8**	**8.5**	3.5	3.9	**17.1**	**9.0**

Table 7. Chinese→Japanese translation experiment results in subword level (BLEU values).

Data\Method	Subword Level							
	LSTM				Transformer			
Test data→	A	O	W	N	A	O	W	N
ASPEC-JC	**39.8**	1.1	2.9	3.3	**44.3**	1.4	4.0	5.3
OpenSubtitles	0.2	**4.2**	2.6	3.6	0.2	**4.6**	2.6	4.0
WCC-JC	2.3	2.7	**13.0**	**6.4**	2.8	2.3	**15.3**	**7.7**

From Tables 4–7, we could conclude that different corpora had the highest BLEU values on their own test data, except for the test data O under Japanese→Chinese (J→C)'s character level with transformer, where our WCC-JC was more effective than OpenSubtitles' own translation. This also showed the generalizability of WCC-JC. Additionally, WCC-JC achieved better results than ASPEC-JC and OpenSubtitles on both test data W and N in all cases. Moreover, by comparing the results of different tables, we could basically conclude that the transformer model of character level was more effective in both J→C and Chinese→Japanese (C→J) directions.

There are many possible reasons that affect low BLEU values. The most likely cause is that one Japanese sentence had been translated into many different Chinese sentences, also known as a one-to-many situation. Regarding the problem that our BLEU values in Tables 4–7 are not very high, we investigated the duplicate Japanese sentences, which are one-to-many Japanese-Chinese sentences. The details are shown in Table 8, which shows the top 10 duplicate Japanese sentences. As we can see, these sentences were all very common daily-use utterances in spoken language, and they were translated differently when used in different scenarios, that is, one-to-many.

Table 9 shows the results of one-to-many translations (only ten kinds are shown here), and the corresponding English translations show that these were colloquial short sentences. Although the characters were different, the meanings were actually very similar, that is, only because the spoken language in different scenarios will be a little different, even if when translated into English it has basically a similar meaning. However, when evaluating the translation results, we cannot include all the translation references. This was also a problem that was difficult to avoid with the evaluation metric of machine translation.

Table 8. Top 10 Japanese duplicate sentences in the corpus (one-to-many).

Number of Duplicates	Sentence Content (in English)
183	はい (Yes)
75	いや (No)
71	ありがとうございます (Thank you)
70	ああ (Yeah)
69	そうだな (Right)
56	そうですね (Sure)
55	行くぞ (Let's go)
54	そうだね (Okay)
52	それは (That is)
52	ほら (Look)

Table 9. Examples of one-to-many translations.

Japanese Sentence	Chinese Sentence (in English)
ああ そうだな	嗯 我赞成 (Yes, I agree)
ああ そうだな	对 就是这样 (Yeah, that's it)
ああ そうだな	嗯 是的 (Yes, it is)
ああ そうだな	嗯 是啊 (Yes, right)
ああ そうだな	嗯 没错 (Yes, that's right)
ああ そうだな	对 你说的对 (Yeah, you're right)
ああ そうだな	嗯 你说得对 (Yes, you are right)
ああ そうだな	嗯 确实啊 (Yes, indeed)
ああ そうだな	说的也对 (That's right)
ああ そうだな	说得也是 (That's true)
.....	.....

For the test data W of WCC-JC, the highest BLEU values were observed for these two experiments, character level with transformer on J→C and C→J, and to make a better evaluation, we also used JPO scores to manually evaluate the test data W's translation results for these two experiments. Figures 4 and 5 show the results of the manual evaluation on J→C and C→J.

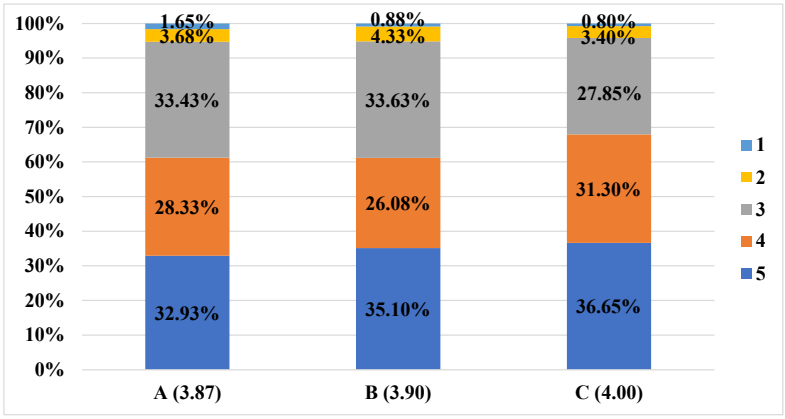

Figure 4. Results of the manual evaluation (Japanese→Chinese).

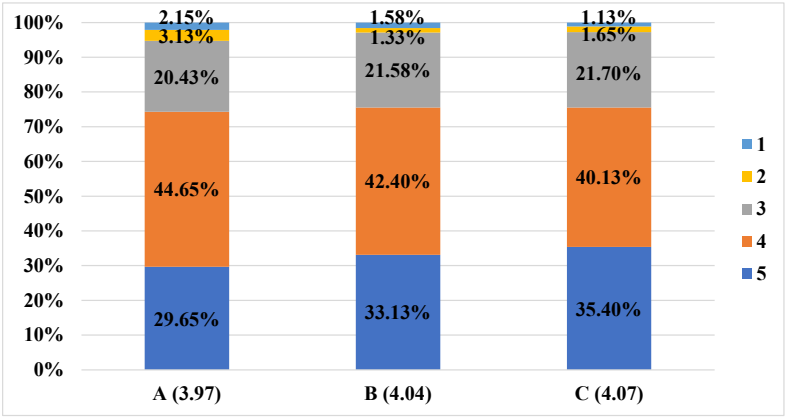

Figure 5. Results of the manual evaluation (Chinese→Japanese).

We set up three groups of A, B, and C. Two people in each group were evaluated independently with the criteria from the previous Section 4.2.2. The specific information is represented in Table 10.

Table 10. The specific information of the human evaluators.

	Group A		**Group B**		**Group C**	
Age:	32	25	33	29	28	33
Gender:	Male	Male	Female	Male	Female	Male
Occupation:	Lab Researcher	International Student	University Lecturer	International Student	Publisher's Staff	University Lecturer
Chinese and Japanese language proficiency:	Excellent	Proficient	Excellent	Proficient	Proficient	Excellent

Then, the average scores were calculated. The numbers in the bar chart indicate the percentage of each group's evaluation value. The numbers in parentheses after the group name represent the average of the JPO scores. We can see that our average results of JPO scores were 3.87, 3.90, and 4.00; 3.97, 4.04, and 4.07 on J→C and C→J, respectively. We obtained relatively high-level results. However, due to the nature of the WCC-JC, lots of the sentences are very short, so a deeper analysis is needed to determine how satisfactory the results are. We should also perform some evaluations of the other experiments to find out the differences in the results and the related reasons.

5. Dataset Publication and Copyright Law

In the case of collecting data from the Web, and publishing the data, there is a problem of copyright infringement because the data contain the copyrighted works of others.

We have consulted with relevant professional lawyers, and WCC-JC does not violate the Copyright Law of the People's Republic of China.

According to the revised Copyright Law of Japan, Article 30-4, which came into effect in 2019, "It is permissible to exploit a work, in any way and to the extent considered necessary, in any of the following cases, or in any other case in which it is not a person's purpose to personally enjoy or cause another person to enjoy the thoughts or sentiments expressed in that work", and even works of others can be widely used for data analysis (translation model creation, image recognition model creation, etc.) (https://www.bunka.go.jp/seisaku/chosakuken/hokaisei/h30_hokaisei, https://www.japaneselawtranslation.go.jp/en/laws/view/3379 (both accessed on 15 November 2021)). Our data were obtained for Japanese-Chinese NMT research through an automated acquisition process without any human intervention, which does not violate the Copyright Law of Japan.

6. Conclusions

In this research, we introduced a Japanese-Chinese bilingual corpus: WCC-JC. This corpus was constructed by crawling the Web on a large scale and automatically collecting Japanese-Chinese bilingual sentences. In the end, about 753k sentence pairs of Japanese-Chinese bilingual data were obtained. The corpus is one of the largest Japanese-Chinese corpora available at present, and includes bilingual texts in spoken languages, which have not been widely treated in existing corpora.

In the experiments using conversational sentences extracted from language course textbooks, we confirmed that although the BLEU values were low, the translation accuracy was the highest among the compared Japanese-Chinese corpora. We also obtained relatively high-level results from the manual evaluations.

Future works include constructing a larger-scale Web-crawled corpus. Another important issue is to improve the accuracy of the alignment of bilingual sentences by the subtitle display time. We are also considering adding more language pairs in the future.

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

Funding: This research was funded by the General Young Talents Project for Scientific Research grant of the Educational Department of Liaoning Province (LJKZ0267), the Research Support Program for Inviting High-Level Talents grant of Shenyang Ligong University (1010147001004), and the 2021 Shenyang Ligong University Research and Innovation Team Development Program Support Project (SYLUTD202105).

Data Availability Statement: A demo (100k) of the WCC-JC presented in this research is openly available on the Github (https://github.com/zhang-jinyi/Web-Crawled-Corpus-for-Japanese-Chinese-NMT (accessed on 6 May 2022)). If you would like to obtain all the data, please contact the following email address on the Github's page while ensuring that it is for your own use and for research purposes only.

Acknowledgments: All authors wish to express their gratitude to the many people who helped them in the construction and evaluation of the WCC-JC.

Conflicts of Interest: The authors declare no conflict 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. Nakazawa, T.; Yaguchi, M.; Uchimoto, K.; Utiyama, M.; Sumita, E.; Kurohashi, S.; Isahara, H. ASPEC: Asian Scientific Paper Excerpt Corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016; European Language Resources Association (ELRA): Portorož, Slovenia, 2016; pp. 2204–2208.
2. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. ImageNet: A large-scale hierarchical image database. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
3. Stefanowitsch, A. *Corpus Linguistics*; Number 7 in Textbooks in Language Sciences; Language Science Press: Berlin, Germany, 2020. [CrossRef]
4. Xu, Y.; Cao, D. *A Study on the Development and Application of Chinese-Japanese Bilingual Corpus: A Collection of Papers*; Foreign Language Teaching And Research Press: Beijing , China, 2002. (In Chinese)
5. Zhang, Y.; Uchimoto, K.; Ma, Q.; Isahara, H. Building an Annotated Japanese-Chinese Parallel Corpus—A Part of NICT Multilingual Corpora. In Proceedings of the Machine Translation Summit X: Papers, Phuket, Thailand, 13–15 September 2005; pp. 71–78.
6. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Machine Translation Summit X: Papers, Phuket, Thailand, 13–15 September 2005; pp. 79–86.
7. Lavecchia, C.; Smaili, K.; Langlois, D. Building a bilingual dictionary from movie subtitles based on inter-lingual triggers. In Proceedings of the Translating and the Computer 29, London, UK, 29–30 November 2007; Aslib: London, UK, 2007.
8. Baroni, M.; Bernardini, S.; Ferraresi, A.; Zanchetta, E. The WaCky wide Web: A collection of very large linguistically processed Web-crawled corpora. *Lang. Resour. Eval.* **2009**, *43*, 209–226. [CrossRef]
9. Smith, J.R.; Saint-Amand, H.; Plamada, M.; Koehn, P.; Callison-Burch, C.; Lopez, A. Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 4–9 August 2013; Association for Computational Linguistics: Sofia, Bulgaria, 2013; pp. 1374–1383.
10. Chu, C.; Nakazawa, T.; Kurohashi, S. Constructing a Chinese—Japanese Parallel Corpus from Wikipedia. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; European Language Resources Association (ELRA): Reykjavik, Iceland, 2014; pp. 642–647.
11. Benko, V. Aranea: Yet Another Family of (Comparable) Web Corpora. In Proceedings of the Text, Speech and Dialogue, Brno, Czech Republic, 8–12 September 2014; Sojka, P., Horák, A., Kopeček, I., Pala, K., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 247–256.
12. 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'16), Portorož, Slovenia, 23–28 May 2016; European Language Resources Association (ELRA): Portorož, Slovenia, 2016; pp. 3530–3534.
13. Pryzant, R.; Chung, Y.; Jurafsky, D.; Britz, D. JESC: Japanese-English Subtitle Corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.

14. Lison, P.; Tiedemann, J.; Kouylekov, M. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.
15. Park, J.; Hai, Z. Korean-to-Chinese Machine Translation using Chinese Character as Pivot Clue. In Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation, Hakodate, Japan, 13–15 September 2019; Waseda Institute for the Study of Language and Information: Hakodate, Japan, 2019; pp. 522–530.
16. Morishita, M.; Suzuki, J.; Nagata, M. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 3603–3609.
17. Lai, G.; Dai, Z.; Yang, Y. Unsupervised Parallel Corpus Mining on Web Data. *arXiv* **2020**, arXiv:abs/2009.08595. [CrossRef]
18. Hasan, T.; Bhattacharjee, A.; Samin, K.; Hasan, M.; Basak, M.; Rahman, M.S.; Shahriyar, R. Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 2612–2623. [CrossRef]
19. Cvrček, V.; Komrsková, Z.; Lukeš, D.; Poukarová, P.; Řehořková, A.; Zasina, A.J.; Benko, V. Comparing Web-crawled and traditional corpora. *Lang. Resour. Eval.* **2020**, *54*, 713–745. [CrossRef]
20. Serrat Roozen, I.; Martínez Martínez, J.M. EMPAC: An English-Spanish Corpus of Institutional Subtitles. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 4044–4053.
21. Nakazawa, T.; Li, L.; Rikters, M. *Construction of a Corpus for Business Scene Dialogue and Issues in Dialogue Translation*; The Association for Natural Language Processing (Japan): Kitakyushu, Japan, 2021; pp. 1375–1380. (In Japanese)
22. Liu, B.; Huang, L. ParaMed: A parallel corpus for English-Chinese translation in the biomedical domain. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 258. [CrossRef] [PubMed]
23. Dou, Z.Y.; Neubig, G. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 2112–2128. [CrossRef]
24. Schwenk, H.; Chaudhary, V.; Sun, S.; Gong, H.; Guzmán, F. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 1351–1361. [CrossRef]
25. Zhang, J.; Matsumoto, T. Improving character-level Japanese-Chinese neural machine translation with radicals as an additional input feature. In Proceedings of the 2017 International Conference on Asian Language Processing (IALP), Singapore, 5–7 December 2017; pp. 172–175. [CrossRef]
26. Zhang, J.; Matsumoto, T. Character Decomposition for Japanese-Chinese Character-Level Neural Machine Translation. In Proceedings of the 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019; pp. 35–40. [CrossRef]
27. Zhang, J.; Matsumoto, T. Corpus Augmentation for Neural Machine Translation with Chinese-Japanese Parallel Corpora. *Appl. Sci.* **2019**, *9*, 2036. [CrossRef]
28. 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 Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 3242–3252. [CrossRef]
29. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MI, USA, 3–5 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 48–53. [CrossRef]
30. 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]
31. 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 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
32. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Virtual, 7–12 July 2002; Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318. [CrossRef]
33. Utiyama, M.; Isahara, H. A Japanese-English patent parallel corpus. In Proceedings of the Machine Translation Summit XI: Papers, Copenhagen, Denmark, 10–14 September 2007.