Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6704–6710
Marseille, 20-25 June 2022

JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan
{makoto.morishita.gr, katsuki.chousa.bg, masaaki.nagata.et}@hco.ntt.co.jp
jun.suzuki@tohoku.ac.jp

Abstract

Most current machine translation models are mainly trained with parallel corpora, and their translation accuracy largely depends on the quality and quantity of the corpora. Although there are billions of parallel sentences for a few language pairs, effectively dealing with most language pairs is difficult due to a lack of publicly available parallel corpora. This paper creates a large parallel corpus for English-Japanese, a language pair for which only limited resources are available, compared to such resource-rich languages as English-German. It introduces a new web-based English-Japanese parallel corpus named JParaCrawl v3.0. Our new corpus contains more than 21 million unique parallel sentence pairs, which is more than twice as many as the previous JParaCrawl v2.0 corpus. Through experiments, we empirically show how our new corpus boosts the accuracy of machine translation models on various domains. The JParaCrawl v3.0 corpus will eventually be publicly available online for research purposes.

Keywords: parallel corpus, machine translation, English, Japanese

1. Introduction

The current neural machine translation models are generally trained by supervised approaches (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), denoting reliance on parallel corpora. However, since publicly available parallel corpora remain limited, training a model for many language pairs is difficult. Thus, constructing a parallel corpus is crucial for expanding the applicability of machine translation.

This paper introduces a new large-scale web-based parallel corpus for English-Japanese for which only limited parallel corpora are available. One of the current largest parallel corpora for this language pair is JParaCrawl (Morishita et al., 2020), which is constructed by crawling the web and automatically aligning parallel sentences. However, this corpus contains around 10 million sentence pairs, which is still limited compared to the other resource-rich language pairs, and it is somewhat outdated because it was created two years ago. We entirely re-crawled the web to update the corpus and applied a different approach to extract parallel sentences. We collected more than 21 million unique sentence pairs, which is more than twice as many as the previous JParaCrawl corpus. We experimentally show how the new crawled corpus increases the accuracy of machine translation for English-Japanese and Japanese-English. Our new corpus, named JParaCrawl v3.0, will be publicly available through our website[1] for further researches.

Our contributions can be summarized:

- We constructed a large-scale English-Japanese parallel corpus, which contains more than 21 million sentence pairs, on top of the previous JParaCrawl corpus.
- We empirically confirmed that our corpus boosted the accuracy of English-Japanese and Japanese-English machine translation in broad domains.
- We plan to release our new parallel corpus for further researches.

2. Related Work

There are several sources for creating parallel corpora. One typical source is the parallel documents written by international organizations. An example is Europarl (Koehn, 2005), which was created from the proceedings of the European Parliament. Ziemski et al. (2016) complied the United Nations parallel corpus from the translated documents of the UN. Professional translators usually translate these texts, which sometimes contain such meta-data as document IDs that allow easy alignment of them. Unfortunately, since these parallel documents are not commonly available, these corpora are limited to a few language pairs and narrow domains.

Another critical source is the web. Many websites are written in several languages, and parallel sentences can be extracted from them. Thus the web is a fruitful source for creating a large parallel corpus in many languages and broader domains. In an earlier work, Uszkoreit et al. (2010) created a large-scale distributed system to mine parallel sentences from the web and books. Smith et al. (2013) proposed a method to mine parallel sentences from Common Crawl[2], a free web crawl archive. Recently, some works created large parallel corpora from Wikipedia or Common Crawl (Schwenk et al., 2019a; Schwenk et al., 2019b) with the latest parallel sentence

[1] http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

[2] https://commoncrawl.org/

Version	# sentences	# words
v1.0	$4,817,172$	$125,216,523$
v2.0	$8,809,771$	$234,393,978$
v3.0	$21,891,738$	$516,218,177$

Table 1: Number of unique sentence pairs and words on English side in JParaCrawl corpus. In this work, we created version 3.0.

alignment method, which uses multilingual sentence embeddings.

ParaCrawl is an important work that creates a large-scale parallel corpus for 24 European languages from the web (Bañón et al., 2020). Since it continuously updates the corpus, it continues to grow. Inspired by that work, Morishita et al. (2020) created a web-based large-scale parallel corpus for English-Japanese, where no large parallel corpus was available. Their corpus, called JParaCrawl, amassed more than 10 million sentences and is the largest publicly available parallel corpus for that language pair. However, the current JParaCrawl corpus remains tiny compared to the other resource-rich language pairs, and thus its translation accuracy is inferior to other resource-rich languages. Thus a larger parallel corpus must be created for English-Japanese. In this work, we extend the JParaCrawl corpus by re-crawling the web and applying a new parallel sentence extraction method, as described in the following section.

3. JParaCrawl v3.0

This paper extends the current JParaCrawl v2.0 corpus by further crawling the web and extracting parallel sentences. Our methods are based on previous ParaCrawl and JParaCrawl projects (Bañón et al., 2020; Morishita et al., 2020). We described the detailed process in the following sections.

3.1. Find Websites Written in Parallel

Our method extracts parallel sentences from the web. Thus, the first step is finding a website that has parallel sentences. This method is based on the hypothesis that websites containing the same English and Japanese sentences might have parallel texts. To list such parallel websites, we analyzed all the Common Crawl text archive data released from March 2019 to August 2021[3]. We identified the language in the archive by `CLD2`[4] and listed 100,000 large websites that roughly have the same size of English and Japanese texts. For this step, we used `extractor`[5] that was provided by the ParaCrawl project.

We ignored the data released before March 2019 because they were already analyzed by the previous JParaCrawl project and focused more on the latest Common Crawl archive. We checked the website lists generated by this procedure and found that 70% were not listed in the previous JParaCrawl (Morishita et al., 2020).

3.2. Crawl the Found Websites

Next we crawled the websites listed in the previous step with `Heritrix`[6] at most 48 hours for each one. Although the previous JParaCrawl only focused on plain texts, in this work, we also crawled PDF and Microsoft Word documents in addition to plain text to extract more parallel sentences because the Japanese government and companies sometimes release their news on PDFs.

3.3. Extract Parallel Sentences

Next we extracted parallel sentences from the crawled archives with `Bitextor`[7] provided by the ParaCrawl project. We added Japanese support on Bitextor 8 and used it for this work. For parallel document and sentence alignment, we used machine a translation-based alignment toolkit `bleualign` (Sennrich and Volk, 2011). It first translates the Japanese sentences into English with a machine translation system and finds an English-Japanese sentence pair that maximizes the BLEU scores (Papineni et al., 2002). For the Japanese-English translations, we used a Transformer-based neural machine translation (NMT) model trained with JParaCrawl v2.0. Preliminary experiments found that `bleualign` outperformed a bilingual lexicon-based method[8].

3.4. Filter Out Noisy Sentences

As the last step, we filtered out the incorrectly aligned or poorly translated noisy sentence pairs with the `Bicleaner`[9] toolkit (Sánchez-Cartagena et al., 2018). Then we concatenated the clean parallel sentences and JParaCrawl v2.0 and deduplicated them. From these steps, we created a new large JParaCrawl v3.0 that contains more than 21 million sentences, which is more than twice as many as the previous JParaCrawl v2.0. Table 1 shows the number of unique sentence pairs in the previous and the new JParaCrawl v3.0 corpus. Note that this number is different from our previous paper (Morishita et al., 2020), because here we are reporting the number of *unique* sentence pairs.

4. Experiments

4.1. Experimental Settings

As an experiment, we trained an NMT model with JParaCrawl v3.0 and evaluated its accuracy on various

[3] During this period, the Common Crawl project released 25 archives, and their text size was about 212 TB.

[4] `https://github.com/CLD2Owners/cld2`

[5] `https://github.com/paracrawl/extractor`

[6] `https://github.com/internetarchive/heritrix3`

[7] `https://github.com/bitextor/bitextor`

[8] The previous JParaCrawl used `hunalign` (Varga et al., 2005), which relies on a bilingual lexicon.

[9] `https://github.com/bitextor/bicleaner`

Test set	Domain	# sentences	# words
ASPEC	Scientific Papers	1,812	39,573
JESC	Movie Subtitles	2,000	13,617
KFTT	Wikipedia Articles	1,160	22,063
TED (tst2015)	TED Talks	1,194	20,367
Business Scene Dialogue Corpus	Dialogues	2,120	19,619
WMT20 News En-Ja	News	1,000	22,141
WMT20 News Ja-En	News	993	24,423
WMT21 News En-Ja	News	1,000	23,305
WMT21 News Ja-En	News	1,005	24,771
WMT19 Robustness En-Ja (MTNT2019)	Reddit	1,392	19,988
WMT19 Robustness Ja-En (MTNT2019)	Reddit	1,111	13,390
WMT20 Robustness Set1 En-Ja	Wikipedia Comments	1,100	29,419
WMT20 Robustness Set2 En-Ja	Reddit	1,376	20,011
WMT20 Robustness Set2 Ja-En	Reddit	997	15,866
IWSLT21 Simultaneous Translation En-Ja Dev	TED Talks	1,442	20,677

Table 2: Number of sentences and words on English side in test sets

Source	院内に「濃厚接触者」はいませんが、接触者全員にPCR検査を実施し、 女性が関係した病棟などを閉鎖して徹底的に消毒するということです。
Reference	There are no known "close contacts" in the hospital, but all contacts will be subjected to PCR tests, and the wards and other areas where the women had been will be closed and thoroughly disinfected.
JParaCrawl v1.0	There is no "strong contact person" in the hospital, but a PCR test will be conducted for all the contacts, and women will close the wards and thoroughly disinfect them.
JParaCrawl v2.0	Although there is no "strong contact person" in the hospital, PCR tests will be performed on all contact persons, and the wards related to women will be closed and thoroughly disinfected.
JParaCrawl v3.0	There are no "close contacts" in the hospital, but PCR tests will be conducted for all contacts, and the wards related to women will be closed and thoroughly disinfected.

Figure 1: Example translations of trained models. Example is from WMT21 News Ja-En test set.

Data	# sentences	# words
ASPEC	3,008,500	68,929,413
JESC	2,797,388	19,339,040
KFTT	440,288	9,737,715
TED	223,108	3,877,868

Table 3: Number of sentences and words on English side in training sets. Original version of ASPEC contains 3.0 million sentences, but we used only first 2.0 million for training based on previous work (Neubig, 2014).

test sets to confirm the effect of our new collected corpus.

4.1.1. Test Sets

To evaluate the NMT models on various domains, we tested our models on the 15 test sets listed in Table 2. We used all the test sets in our previous work (Morishita et al., 2020), which included the Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016), the Japanese-English Subtitle Corpus (JESC) (Pryzant et al., 2017), the Kyoto Free Translation Task (KFTT) (Neubig, 2011), and TED talks (tst2015) (Cettolo et al., 2012). We also evaluated our models on the Business Scene Dialogue Corpus (Rikters et al., 2019) to check whether they worked on conversations. We also added test sets from shared tasks: WMT 2020, 2021 news translation shared tasks (Barrault et al., 2020; Akhbardeh et al., 2021), WMT 2019, 2020 robustness shared tasks (Li et al., 2019; Specia et al., 2020), and the IWSLT 2021 simultaneous translation task (Anastasopoulos et al., 2021). Although some of the test sets are intended for specific translation directions (e.g., En→Ja), we used them for both En→Ja and Ja→En directions for reference.

Some corpora have an in-domain training set, as shown in Table 3. For comparison, we trained our model with these training sets and reported the BLEU scores.

4.1.2. Training Settings

First, we tokenized the corpus into sub-words with the `sentencepiece` toolkit with a vocabulary size of 32,000 (Kudo and Richardson, 2018). Then we trained the NMT models with the `fairseq` toolkit (Ott et al., 2019). Our models are based on Transformer (Vaswani et al., 2017) and trained with three settings: small, base, and large. Table 4 shows the detailed training settings. We used the small model for TED (tst2015), the base model for KFTT, and the big model for the others. Note

Common Settings	
Architecture	Transformer (Vaswani et al., 2017)
Enc-Dec layers	6
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$) (Kingma and Ba, 2015)
Learning rate schedule	Inverse square root decay
Warmup steps	4,000
Max learning rate	0.001
Dropout	0.3 (Srivastava et al., 2014)
Gradient clipping	1.0 (Pascanu et al., 2013)
Weight Decay	0.0
Label smoothing	$\epsilon_{ls} = 0.1$ (Szegedy et al., 2016)
Mini-batch size	512,000 tokens (Ott et al., 2018)
Number of updates	36,000 steps (v3.0), 24,000 steps (v1,0, v2.0)
Averaging	Save checkpoint for every 100 steps and take an average of last 8 checkpoints
Beam size	6 with length normalization (Wu et al., 2016)
Small Settings	
Attention heads	4
Word-embedding dimension	512
Feed-forward dimension	1,024
Base Settings	
Attention heads	8
Word-embedding dimension	512
Feed-forward dimension	2,048
Big Settings	
Attention heads	16
Word-embedding dimension	1,024
Feed-forward dimension	4,096

Table 4: List of hyperparameters

that these settings are almost the same as our previous work (Morishita et al., 2020) for a fair comparison, except we changed the number of updates for the v3.0 models because the new corpus is too large to converge in 24,000 steps. We evaluated the model with the `sacreBLEU` toolkit (Post, 2018) and reported the BLEU scores (Papineni et al., 2002). For the evaluations, we NFKC-normalized all the test sets for consistency with our previous experiments (Morishita et al., 2020).

4.2. Experimental Result

Table 5 shows the BLEU scores on various test sets. Our corpus is not designed as a specific domain but as a general one. Thus, unsurprisingly, the JParaCrawl model did not reach the model's score trained with in-domain data. The model trained with JParaCrawl v3.0 achieved the best score on 14 of 15 test sets on both Japanese-English and English-Japanese. These results clearly show that our new parallel corpus increased the accuracy of the NMT models on various domains, including scientific papers, news, and dialogues.
Our v3.0 model worked especially well on the WMT21 news translation tasks. We believe that this is because the previous JParaCrawl v2.0 was based on the web in 2019, and so it might not have included terms frequently used in 2021. For example, news articles in 2021 cited words related to COVID-19, a term that was not obviously less frequently used in 2019. Perhaps we need to continue to update the parallel corpus to reflect the latest terms.

4.3. Translation Example

Figure 1 shows an example translation of JParaCrawl v1.0, v2.0, and v3.0. We chose this example from the WMT21 news translation test set because it is related to COVID-19. This input includes the Japanese phrase "濃厚接触者", which should have been translated to "close contacts." But the v1.0 and v2.0 models incorrectly translated the language to "strong contact person." In contrast, the model trained with v3.0 correctly translated the phrase to "close contacts." Similar to this example, we identified many improvements in the articles related to COVID-19. These results support our hypothesis that our model trained with the v3.0 corpus correctly translated the terms and language frequently used in recent years.

5. Conclusion

This paper introduced an updated version of the large English-Japanese parallel corpus called JParaCrawl. We re-crawled parallel websites by analyzing the latest CommonCrawl archive and extended the crawl target to PDF and Word documents. After filtering out noisy sentences, the new JParaCrawl v3.0 included more than 21 million unique sentence pairs. We empirically confirmed that the new corpus boosts the translation accuracy on various domains, especially on the trendiest news articles. Our future work will update the JParaCrawl corpus and propose better alignment/filtering techniques. The new JParaCrawl v3.0 will be available on our website for further research. We expect that JParaCrawl v3.0 will support future research and products.

6. Acknowledgements

We gratefully acknowledge the ParaCrawl project for releasing its software and its members with whom we engaged in valuable discussions. We thank the three anonymous reviewers for their comments.

7. Bibliographical References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz,

	English-to-Japanese				Japanese-to-English			
		JParaCrawl				JParaCrawl		
Test set	**In-domain**	**v1.0**	**v2.0**	**v3.0**	**In-domain**	**v1.0**	**v2.0**	**v3.0**
ASPEC	44.3	24.7	26.5	**27.0**	28.7	18.3	19.7	**21.0**
JESC	14.5	6.6	6.5	**6.8**	17.8	7.0	7.5	**8.3**
KFTT	31.8	17.1	**18.9**	17.9	23.4	13.7	16.2	**17.0**
TED (tst2015)	11.1	11.5	12.6	**13.0**	13.7	11.0	11.9	**12.0**
Business Scene Dialogue Corpus	—	12.4	13.5	**14.1**	—	17.4	19.6	**19.9**
WMT20 News En-Ja	—	20.7	21.9	**23.5**	—	21.3	23.3	**24.0**
WMT20 News Ja-En	—	20.1	22.8	**23.7**	—	19.2	21.0	**21.6**
WMT21 News En-Ja	—	21.1	21.8	**25.1**	—	21.9	23.1	**23.9**
WMT21 News Ja-En	—	19.6	21.5	**22.8**	—	18.1	20.7	**21.7**
WMT19 Robustness En-Ja (MTNT2019)	—	12.4	12.5	**14.4**	—	15.6	16.8	**17.2**
WMT19 Robustness Ja-En (MTNT2019)	—	11.5	12.3	**13.0**	—	16.0	17.2	**17.7**
WMT20 Robustness Set1 En-Ja	—	15.2	15.8	**18.7**	—	20.0	20.6	**21.4**
WMT20 Robustness Set2 En-Ja	—	12.7	13.0	**14.5**	—	16.4	**17.4**	17.2
WMT20 Robustness Set2 Ja-En	—	7.9	8.2	**8.9**	—	12.0	12.6	**13.8**
IWSLT21 Simultaneous Translation En-Ja Dev	—	12.5	13.3	**14.5**	—	12.9	14.3	**15.0**

Table 5: BLEU scores of models trained with in-domain training set, JParaCrawl v1.0, v2.0, and v3.0. Best scores among JParaCrawl models are highlighted in bold.

C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In Proceedings of the 6th Conference on Machine Translation (WMT), pages 1–88.

Anastasopoulos, A., Bojar, O., Bremerman, J., Cattoni, R., Elbayad, M., Federico, M., Ma, X., Nakamura, S., Negri, M., Niehues, J., Pino, J., Salesky, E., Stüker, S., Sudoh, K., Turchi, M., Waibel, A., Wang, C., and Wiesner, M. (2021). FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT), pages 1–29.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 4555–4567.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In Proceedings of the 5th Conference on Machine Translation (WMT), pages 1–55.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: web inventory of transcribed and translated talks. In Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), pages 261–268.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of the Machine Translation Summit X, pages 79–86.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 66–71.

Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N. K., Firat, O., Koehn, P., Neubig, G., Pino, J. M., and Sajjad, H. (2019). Findings of the first shared task on machine translation robustness. In Proceedings of the 4th Conference on Machine Translation (WMT).

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1412–1421.

Morishita, M., Suzuki, J., and Nagata, M. (2020). JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pages 3603–3609.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ASPEC: Asian scientific paper excerpt corpus. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC).

Neubig, G. (2011). The Kyoto free translation task. http://www.phontron.com/kftt.

Neubig, G. (2014). Forest-to-string SMT for asian language translation: NAIST at WAT2014. In Proceedings of the 1st Workshop on Asian Translation (WAT), pages 20–25.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. In Proceedings of the 3rd Conference on Machine Translation (WMT), pages 1–9.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). 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: Human Language Technologies (NAACL HLT), pages 48–53.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In Proceedings of the 30th International Conference on Machine Learning (ICML), volume 28, pages 1310–1318.

Post, M. (2018). A call for clarity in reporting BLEU scores. In Proceedings of the 3rd Conference on Machine Translation (WMT), pages 186–191.

Pryzant, R., Chung, Y., Jurafsky, D., and Britz, D. (2017). JESC: Japanese-English Subtitle Corpus. arXiv preprint arXiv:1710.10639.

Rikters, M., Ri, R., Li, T., and Nakazawa, T. (2019). Designing the business conversation corpus. In Proceedings of the 6th Workshop on Asian Translation (WAT), pages 54–61.

Sánchez-Cartagena, V. M., Bañón, M., Ortiz-Rojas, S., and Ramírez-Sánchez, G. (2018). Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In Proceedings of the 3rd Conference on Machine Translation (WMT), pages 955–962.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019a). WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. arXiv preprint arXiv:1907.05791.

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., and Joulin, A. (2019b). CCMatrix: Mining billions of high-quality parallel sentences on the web. arXiv preprint arXiv:1911.04944.

Sennrich, R. and Volk, M. (2011). Iterative, MT-based sentence alignment of parallel texts. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA), pages 175–182.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), pages 1374–1383.

Specia, L., Li, Z., Pino, J., Chaudhary, V., Guzmán, F., Neubig, G., Durrani, N., Belinkov, Y., Koehn, P., Sajjad, H., Michel, P., and Li, X. (2020). Findings of the WMT 2020 shared task on machine translation robustness. In Proceedings of the 5th Conference on Machine Translation (WMT), pages 76–91.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15:1929–1958.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS), pages 3104–3112.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), pages 2818–2826.

Uszkoreit, J., Ponte, J. M., Popat, A. C., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pages 1101–1109.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2005). Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing (RANLP), pages 590–596.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS), pages 6000–6010.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In Proceedings of the 10th International Conference on

Language Resources and Evaluation (LREC), pages 3530–3534.