Meta Ensemble for Japanese-Chinese Neural Machine Translation: Kyoto-U+ECNU Participation to WAT 2020

Zhuoyuan Mao[♡] **Yibin Shen**[♢]
Chenhui Chu[♡] **Sadao Kurohashi**[♡] **Cheqing Jin**[♢]
[♡]Kyoto University, Japan
[♢]East China Normal University, China
{zhuoyuanmao, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp
ybshen@stu.ecnu.edu.cn, cqjin@dase.ecnu.edu.cn

Abstract

This paper describes the Japanese-Chinese Neural Machine Translation (NMT) system submitted by the joint team of Kyoto University and East China Normal University (Kyoto-U+ECNU) to WAT 2020 (Nakazawa et al., 2020). We participate in APSEC Japanese-Chinese translation task. We revisit several techniques for NMT including various architectures, different data selection and augmentation methods, denoising pre-training, and also some specific tricks for Japanese-Chinese translation. We eventually perform a meta ensemble to combine all of the models into a single model. BLEU results of this meta ensembled model rank the first both on 2 directions of ASPEC Japanese-Chinese translation.

1 Introduction

Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) has led to large improvements in machine translation quality when large parallel corpora are available for training. In this work, we revisit several existing NMT based techniques on ASPEC Japanese-Chinese translation task. Furthermore, we conduct a meta ensemble to fuse various NMT based systems. The aspects this work feature can be summarized as:

- We revisit various NMT based systems including LSTM (Sutskever et al., 2014), Transformer (Vaswani et al., 2017), ConvS2S (Gehring et al., 2017) and Lightconv (Wu et al., 2019) with different data augmentation and filtering techniques.
- We implement a meta ensemble on various NMT systems and obtain the state-of-the-art results on ASPEC Japanese-Chinese translation task.
- We empirically compare fully (semi-) supervised training with the recent popular language model pre-training based methods (Conneau and Lample, 2019; Song et al., 2019; Lewis et al., 2020; Liu et al., 2020).
- We revisit and explore the trick of character mapping (Song et al., 2020; Chen et al., 2020) between Chinese and Japanese on ASPEC translation task.

Although only our team participated in the ASPEC Japanese-Chinese translation task this year, BLEU results we report on the WAT official leader-board rank 1st both on ja→zh and zh→ja compared with all the previous submitted systems.

2 ASPEC Japanese-Chinese Translation Task

	Number of parallel sentences
train	672,315
dev	2,090
devtest	2,148
test	2,107

Table 1: **ASPEC-JC overview.** We merge "dev" and "devtest" as development set in our experiments.

ASPEC (Asian Scientific Paper Excerpt Corpus) (Nakazawa et al., 2016) was constructed in the Japanese-Chinese machine translation project from 2006 to 2010 by Japan Science and Technology Agency. ASPEC Japanese-Chinese (ASPEC-JC, shown in Table 1) and ASPEC Japanese-English (ASPEC-JE) respectively consists approximately 0.68M and 3M parallel sentences for training. In this work, we focus on NMT system training for ASPEC-JC while parts of ASPEC-JE is leveraged for data augmentation.

Proceedings of the 7th Workshop on Asian Translation, pages 64–71

3 Our System

3.1 Sequence-to-Sequence Framework

Sequence-to-sequence framework (S2S) is the basic technique being used to learn a mapping from a source sentence to a target sentence in an end-to-end manner. In this section, we revisit four different architectures for sequence-to-sequence learning including LSTM, ConvS2S, Transformer, and Lightconv. In our system, most experimental settings are based on Transformer while we also implement other 3 architectures to compare the performance on ASPEC task.

LSTM (Sutskever et al., 2014). Long-Short-Term Memory (LSTM) is a special recurrent neural network (RNN), which solves the problem of gradient vanishing/exploding on the long sequence by integrating three gates (one input gate, one forget gate, and one output gate) and memory cells. Therefore, LSTM is capable of storing and forgetting information for longer periods of time on the sequence.

ConvS2S (Gehring et al., 2017). Compared with RNNs that maintain a hidden state of the entire past, convolution operations can be fully parallelized during training. ConvS2S integrates the convolution operations into the sequence-to-sequence framework, which not only improves computation efficiency but also captures the long-range dependencies over the input sequence through multi-layer hierarchical structure.

Transformer (Vaswani et al., 2017). Transformer eschew recurrence and convolutions entirely, relying on the attention mechanism to capture global dependencies between input and output sequence. Due to its high parallelism and the high quality in the translation task, it has become the most popular model among researchers.

Lightconv (Wu et al., 2019). Lightconv is a lightweight convolution model that utilizes lightweight convolutions and dynamic convolutions instead of the self-attention mechanism. The attention weights of self-attention are computed by comparing the current time-step to all elements on the sequence, which brings a great challenge for longer sequences due to the quadratic complexity. Lightconv builds dynamic convolutions to predict a different kernel at each time-step rather than the entire sequence, which drastically reduces the number of parameters.

Corpus	Number of parallel sentences
Ubuntu	92,250
Open Subtitles	914,355
TED	376,441
Global Voices	16,848
Wikipedia	228,565
Wiktionary	62,557
News Commentary	570
Tatoeba	4,243
WikiMatrix	267,409
Total	1,963,238

Table 2: Out-of-domain Chinese-Japanese parallel datasets collected by IWSLT 2020.

3.2 Data Augmentation by Filtering Out-of-domain Parallel Data

NMT quality depends highly on the size of the training data. Thus, high quality augmented training dataset help ameliorate NMT. There exist just around 0.68M parallel sentences in ASPEC-JC, so we expect that extra parallel data can significantly improve the translation quality. We use Japanese-Chinese parallel datasets on other domains collected by IWSLT 2020 (Ansari et al., 2020).[1] All the out-of-domain datasets used are shown in Table 2 including Ubuntu corpora from OPUS (Tiedemann, 2012), Global Voices, and News Commentary; OpenSubtitles (Lison and Tiedemann, 2016); TED talks (Dabre and Kurohashi, 2017); Wikipedia (Chu et al., 2014, 2016); Wikitionary.org;[2] Tatoeba.org under CC-BY License;[3][4] and WikiMatrix (Schwenk et al., 2019). In total, over 1.9M out-of-domain parallel data are publicly available which can be leveraged to enhance the performance on the ASPEC task. However, we observe some sentence alignments are of low accuracy. Thus, we also conduct a filtering by using LASER (Artetxe and Schwenk, 2019). Specifically, we remove sentence pairs with the cos-similarity score less than a fixed threshold where similarity scores are calculated by LASER embeddings of two sentences. We observe that alignment quality tends to be high if the similarity score is over 0.6, so we use this value as the filtering thresh-

[1] https://github.com/didi/iwslt2020_open_domain_translation

[2] https://dumps.wikimedia.org/

[3] https://tatoeba.org/eng/

[4] https://creativecommons.org/licenses/by/2.0/fr/

old for most experiments. This results in 1.5M filtered out-of-domain parallel sentences which we leverage to train the NMT system jointly with ASPEC-JC dataset. We also revisit the domain adaption method (Chu et al., 2017) by adding tags of $\langle 2indomain \rangle$ and $\langle 2outofdomain \rangle$ during the training phase.

3.3 Data Augmentation by Back Translation

Besides using out-of-domain parallel data, we also implement back translation (Sennrich et al., 2016a; Edunov et al., 2018), another effective data augmentation technique for NMT. For monolingual corpora, we use the Japanese sentences in ASPEC-JE as in-domain Japanese monolingual data, where 3M Japanese sentences are used to augment the ASPEC-JC corpus. We do not perform the back translation by using Chinese monolingual data because no in-domain Chinese sentences are available. We neither do not consider using other out-of-domain monolingual data. For the translation direction of ja→zh, we name it *forward translation* because of the absence of the target-side monolingual data, which means we use the pre-trained system of ja→zh to forward translate monolingual Japanese sentences into Chinese to augment the original dataset.

3.4 Character Mapping

Character mapping is another essential trick frequently being used in Japanese-Chinese translation tasks (Song et al., 2020; Chen et al., 2020) because there are a large number of shared Chinese characters in Chinese and Japanese. Usually they not only share the character but also share the semantic function within a sentence, so pre-mapping Chinese characters to the target-side can help amplify the cross-lingual supervision. More precisely, for Kana characters in Japanese, we remain them with the same tokenization granularity whereas for Chinese characters, we first tokenize them into by characters, then use the character mapping table (Chu et al., 2012) to pre-map them to target-side.

3.5 mBART: Multilingual Denoising Pre-training

After the apperance of BERT (Devlin et al., 2019), several pre-training methods are proposed for enhancing NMT (Conneau and Lample, 2019; Song et al., 2019; Ren et al., 2019; Mao et al., 2020; Lewis et al., 2020). Recently, mBART (Liu et al., 2020) is a multilingual sequence-to-sequence language model pre-trained by denoising tasks on 25 languages including Japanese and Chinese. Specifically, we fine-tune mBART25[5] by Japanese-Chinese parallel sentences and compare this kind of multilingual pre-training with other fully (semi-) supervised baselines.

3.6 Meta Ensemble

Systems mentioned above can be classified into 3 groups. First, *different training data*. This group means we train the NMT systems with various kinds of the training data. We introduce changes on the original dataset by different data augmentation methods and settings (filtering threshold etc.). Second, *different S2S frameworks* by which we train the NMT system with different model architectures. Third, *different model capacities*. Different hyperparameter settings of transformer model are changed in this class by which we can obtain various styles of translated sentences. This benefits the quality of the ensembled model. Specifically, we focus on exploring deeper architectures and the deep encoder, shallow decoder setting (Kasai et al., 2020). Finally, we conduct an self-ensemble to combine all the above introduced systems. We take the checkpoint average for each system, thus the final ensemble is an ensemble of self-ensemble. Consequently, we call the final ensemble phase by *meta ensemble*.

4 Preprocessing and Training Details

We conduct tokenization by using *Juman*[6] (Kurohashi et al., 1994; Morita et al., 2015) for Japanese and *stanford parser pku*[7] for Chinese. Sentences over 175 tokens are removed for training. We build a joint vocabulary with 30k merge operations by Byte-Pair Encoding (Sennrich et al., 2016b). This results in a joint vocabulary with approximately 40k tokens. Note that we build a single vocabulary for all the settings except fine-tuning mBART25. We oversample ASPEC-JC for BPE codes learning and NMT model training to balance the in-domain and out-of-domain tokens during the model training. We use the provided vocabulary constructed by SentencePiece (Kudo and Richardson, 2018)

[5] https://github.com/pytorch/fairseq/blob/master/examples/mbart/README.md
[6] https://github.com/ku-nlp/jumanpp
[7] https://nlp.stanford.edu/software/lex-parser.shtml

#	Augmentation	Threshold	Vocabulary	Architecture	ja-zh		zh-ja	
					Best	Avg.10	Best	Avg.10
Vanilla transformer								
1	-	-	BPE-30k	big	35.69	35.82	48.53	49.03
Different training data								
2	out-of-domain w/o tag	0.6	BPE-30k	big	35.32	35.42	48.54	49.02
3	out-of-domain w tag	0.6	BPE-30k	big	35.51	35.66	48.40	49.45
4	BT / FT (1)	0.6	BPE-30k	big	**36.16**	**36.52**	**49.66**	**50.17**
5	BT / FT (1)	0.6	BPE-30k	big + ls(0.2)	**36.23**	**36.52**	**49.75**	**50.35**
6	BT / FT (2)	0.6	BPE-30k	big	**36.29**	36.22	**49.67**	49.97
7	out-of-domain + BT / FT (1)	0.9	BPE-30k	big + ls(0.2)	**36.21**	36.25	48.94	49.64
Different S2S frameworks								
8	BT / FT (1)	0.6	BPE-30k	LSTM + ls(0.2)	35.64	35.94	47.57	47.56
9	BT / FT (1)	0.6	BPE-30k	ConvS2S + ls(0.2)	35.34	35.97	46.71	47.65
10	BT / FT (1)	0.6	BPE-30k	Lightconv + ls(0.2)	**36.33**	**36.63**	**49.86**	**50.32**
Different model capacities								
11	BT / FT (1)	0.6	BPE-30k	big(12/1) + ls(0.2) + dp(0.1)	35.69	**36.55**	49.07	**50.10**
12	BT / FT (1)	0.6	BPE-30k	big + ls(0.2) + ffhd(8192)	36.08	**36.69**	49.23	49.83
13	BT / FT (1)	0.6	BPE-30k	big(9/9) + ls(0.2)	35.95	36.42	**49.71**	**50.34**
Character mapping								
14	-	-	BPE-30k	big	35.44	36.05	47.88	48.74
Fine-tune pre-trained models								
15	mBART25	-	SPM-200k	big(12/12)	34.36	35.35	49.35	49.71
16	mBART25 + BT / FT (1)	0.6	SPM-200k	big(12/12)	34.87	35.20	49.27	49.89

Table 3: **BLEU results on test sets of ASPEC ja-zh and zh-ja translation).** Bold denotes top-5 BLEU scores of each column. "Best" and "Avg.10" respectively means the best checkpoint and the average of last 10 checkpoints. "out-of-domain" means training with additional out-of-domain parallel data. "BT / FT (i)" denotes i-th turns of the back translation or forward translation. "big" means transformer-big setting and "big(12/1)" means transformer-big with 12 encoder layers and 1 decoder layer. "ls", "dp" and "ffhd" represents label smoothing, dropout and feed-forward hidden dimension, respectively. Model settings without declarations of "ls", "dp" and "ffhd" are set to "ls(0.1)", "dp(0.3)" and "ffhd(4,096)" for transformer-big and default for other S2S frameworks. "Threshold" means the filtering threshold value for LASER embedding.

for fine-tuning mBART25. For parts of the out-of-domain Chinese sentences that are in traditional Chinese, we transfer Traditional Chinese characters to Simplified Chinese ones.[8]

We conduct all the experiments by using Fairseq[9] (Ott et al., 2019), an open source sequence-to-sequence framework implementation. Most systems are set to the Transformer-big setting except those built up by other architectures or different model capacities. In particular, our model has a 6-layer encoder and decoder, a hidden size of 1,024, a feed-forward hidden layer size of 4,096, batch-size of 2,048, dropout rate of 0.3 and 16 attention heads. For LSTM, we also use a 6-layer encoder and decoder architecture. For ConvS2S and Lightconv, we use the default settings in Fairseq. All the systems are early stopped if BLEU does not improve for continuous 50,000 steps. Experiments are run on 8 TITAN X (Pascal) GPUs except that single V100 GPU is used for mBART25 fine-tuning. We use BLEU (Papineni et al., 2002) for automatic evaluation.

[8] https://github.com/berniey/hanziconv

[9] https://github.com/pytorch/fairseq

5 Results

5.1 Single Model Results

Results of 16 single models with different settings are shown in Table 3.

First, we report the results of vanilla transformer-big model (#1). By averaging last 10 checkpoints, BLEU results of 35.82 and 49.03 on ja-zh and zh-ja are obtained, which is slightly higher than results on the best checkpoints.

Second, we train NMT systems with out-of-domain parallel corpora (#2 & #3). We observe almost no BLEU improvements from additional training data, which can be attributed to extremely high alignment quality of the ASPEC-JC training data. BLEUs of averaged models increase by adding the domain tag (#3) that has once been demonstrated effective in LSTM architecture on the same task

#	Models	ja-zh		zh-ja	
		dev	test	dev	test
With other resources					
17	1	36.44	35.82	49.95	49.03
18	1+2	37.64	37.27	51.71	50.63
19	1+2+4	38.24	37.79	52.56	51.73
20	1+2+4+7	38.51	37.92	52.84	52.12
21	1+2+4+7+8	38.53	38.10	53.03	52.37
22	1+2+4+7+8+11	38.84	38.32	53.17	52.59
23	1+2+4+7+8+11+13	38.85	38.44	53.42	52.69
24	1+2+4+7+8+9+11+13	39.00	38.50	53.51	52.71
25	1+2+4+7+8+9+10+11+13	38.98	38.49	**53.53**	**52.75**
26	2+4+7+8+9+10+11+13	**39.07**	**38.63**	52.30	52.62
27	2+4+7+8+9+11+13	38.97	38.48	-	-
Without other resources					
28	1+4+8+9+10+11+13	-	-	-	52.59
29	4+8+9+10+11+13	-	38.50	-	-

Table 4: **Ensembled BLEU results of ASPEC ja-zh and zh-ja translation.** For each model, we use Avg.10 for the ensemble. Model numbers are referred from Table 3. Results without using other resources are reported in the last 2 rows.

by Chu et al. (2017).

Third, by conducting back translation and forward translation (#4 $\sim$ #7) with Japanese in-domain monolingual data from ASPEC-JE, we obtain significant improved BLEUs. However, iterative back (or forward) translation have marginal contributions. Furthermore, we also mix out-of-domain parallel data with back (or forward) translated synthetic parallel data to train the NMT system with the threshold of 0.9. Although no BLEU improvements observed, different generations will contribute to model ensembling.

Fourth, we report the results of systems trained by different architectures. For the implementation of LSTM, ConvS2S, and Lightconv, we utilize the same training set as that of back (or forward) translation with 0.6 filtering threshold. As shown by #8, 9, 10 in Table 3, Lightconv yields better results compared with transformer architecture (#5). LSTM and ConvS2S underperform the other 2 architectures but are capable to yield generations with discrepancy (see 5.2).

Fifth, we report results of deep encoder shallow decoder system (#11), larger transformer (#12) and deeper transformer (#13). These 3 systems yield comparable results as the standard transformer-big system (#5). Whether they give contributions to model ensembling will be demonstrated in the next section.

Sixth, we revisit the trick of character mapping on Japanese-Chinese translation task. The results of #14 show that only averaged result on ja-zh marginally outperforms the vanilla transformer (#1). The failure of this trick on APPEC task can be ascribed to the high quality of ASPEC-JC, which indicates that this trick harms the Chinese character embedding learning on a well-constructed parallel corpus.

Last, we explore recent popular pre-training methods on this task. We observe improvements by fine-tuning mBART25 on zh-ja whereas BLEU decreases on ja-zh (#15). Moreover, by fine-tuning mBART25 with the back (or forward) translated training set, no significant BLEU improvements are observed, which indicates the conflict nature between denoising pre-training and back translation in the semi-supervised scenario.

5.2 Ensembled Model Results

In this section, we report ensembled model results which are shown in Table 4. We have trained 16 NMT systems and 12 of them (#1, #2, #4 $\sim$ #13) can be directly ensembled because they share an identical vocabulary and the same format of the source sentence (without adding tags or pre-mapping). However, it is difficult to discover the

References	
ja	大規模テキストコレクションであるＮＴＣＩＲ３特許データコレクションのデータを用いて提案手法の評価を行い，間接評価と直接評価により本手法の有効性を示した．
zh	利用大规模文本集ＮＴＣＩＲ３专利数据集的数据对提案方法进行评价，并通过间接评价和直接评价证明了本方法的有效性。
Generated by our systems	
ja	大規模テキストコレクションＮＴＣＩＲ３特許データコレクションのデータを用いて提案手法の評価を行い，間接評価と直接評価により本手法の有効性を示した．
zh	利用大规模文本集ＮＴＣＩＲ３专利数据集的数据，对提案方法进行评价，并通过间接评价和直接评价证明了本方法的有效性。

Table 5: Translation examples of #25 on zh-ja and #26 on ja-zh.

Systems	ja-zh	zh-ja	Ranking
With other resources			
25	-	52.80	1^{st}
26	38.66	-	1^{st}
Without other resources			
28	-	52.65	2^{nd}
29	38.52	-	3^{rd}

Table 6: BLEU results of our submitted systems evaluated by organizers.

best ensemble combinations. Thus, we conduct model ensembling in a greedy manner by the model order (start from #1). In Table 4, we only list out positive ensemble combinations. We observe that BLEU improves significantly for the first 2 ensemble (#17 $\sim$ #19) while BLEU improvements slow down after #19. We also observe that most ensemble combinations give BLEU improvements (9 models give positive contribution on zh-ja while 8 on zh-ja). This demonstrates that ensembled systems can yield better results by changing the training set, utilizing different architectures, and modifying model capacities even though their single models do not provide performance improvements. As shown in Table 5, we observe generated results are of extremely high quality. Last, systems we submitted for official evaluation are shown in Table 6. It is worth mentioning that the result of #25 on zh-ja and the result of #26 on ja-zh rank first on ASPEC-JC leaderboard. Ensembled results without using other resources (#28 & #29) respectively rank second on zh-ja and third on ja-zh.

6 Conclusion

In this work, we participated in ASPEC-JC translation task. We revisited several strong NMT baselines, tricks for handling training set, and also pre-training + fine-tuning methods for Japanese-Chinese NMT. Furthermore, we conducted a deliberate search of better ensembled models and obtained state-of-the-art translation results on this task. Because of the high BLEU results on this task, in the future, adapting systems trained on ASPEC to other domains should be explored and unsupervised machine translation on this task can be focused on.

References

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Pinzhen Chen, Nikolay Bogoychev, and Ulrich Germann. 2020. Character mapping and ad-hoc adaptation: Edinburgh's IWSLT 2020 open domain translation system. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 122–129, Online. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2012. Chinese characters mapping table of Japanese, traditional Chinese and simplified Chinese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2149–2152, Istanbul, Turkey. European Languages Resources Association (ELRA).

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a Chinese—Japanese parallel corpus from Wikipedia. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 642–647, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2016. Integrated parallel sentence and fragment extraction from comparable corpora: A case study on chinese-japanese wikipedia. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 15(2):10:1–10:22.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Raj Dabre and Sadao Kurohashi. 2017. MMCR4NLP: multilingual multiway corpora repository for natural language processing. *CoRR*, abs/1710.01025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1243–1252.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation. *CoRR*, abs/2006.10369.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 22–28.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Zhuoyuan Mao, Fabien Cromieres, Raj Dabre, Haiyue Song, and Sadao Kurohashi. 2020. JASS: Japanese-specific sequence to sequence pre-training for neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3683–3691, Marseille, France. European Language Resources Association.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian*

Translation, Suzhou, China. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 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 (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Explicit cross-lingual pre-training for unsupervised machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 770–779, Hong Kong, China. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Eiichiro Sumita. 2020. Pre-training via leveraging assisting languages for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 279–285, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 5926–5936.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA).

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

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.