Goku's Participation in WAT 2020

Dongzhe Wang, Ohnmar Htun
Rakuten Institute of Technology
Rakuten, Inc.
{first.last}@rakuten.com

Abstract

This paper introduces our neural machine translation systems' participation in the WAT 2020 (team ID: *goku20*). We participated in the (i) Patent, (ii) Business Scene Dialogue (BSD) document-level translation, (iii) Mixed-domain tasks. Regardless of simplicity, standard Transformer models have been proven to be very effective in many machine translation systems. Recently, some advanced pre-training generative models have been proposed on the basis of encoder-decoder framework. Our main focus of this work is to explore how robust Transformer models perform in translation from sentence-level to document-level, from resource-rich to low-resource languages. Additionally, we also investigated the improvement that fine-tuning on the top of pre-trained transformer-based models can achieve on various tasks.

1 Introduction

This paper introduces our neural machine translation (NMT) systems' participation in the 7th Workshop on Asian Translation (WAT-2020) shared translation task (Nakazawa et al., 2020). We participated in the (i) JPO Patent, (ii) Document-level Business Scene Dialogue (BSD) translation, and (iii) Mixed-domain tasks. In particular, the document-level translation tasks are newly introduced for WAT 2020 as traditional translation tasks such as ASPEC usually focus on sentence-level translation, whose quality tends to saturation.

We built our NMT systems based on the standard Transformer (Vaswani et al., 2017) for the JPO Patent and Mixed-domain tasks. In addition to standard Transformer, a pre-training auto-encoder model mBART (Liu et al., 2020) has been explored in the JPO patent task. In terms of the document-level translation task, we evaluated on the BSD corpus using the hierarchical Transformer models (Miculicich et al., 2018) and compared the results with our fine-tuned mBART models, which were initially built to deal with the document-level translation as a downstream task.

The NMT systems for the JPO patent task have been trained in a constrained manner, which means no other resources were used except training corpus provided by the shared task organizers, and achieved remarkable performance. On the other hand, we leveraged other data resources when only limited number of data provided for model training. For instance, we included the Japanese-English Subtitle Corpus (JESC) (Pryzant et al., 2018) and Myth Corpus (Susanto et al., 2019) as auxiliary training data for the document-level and mixed-domain translation tasks, respectively. Our main findings for each task are summarized in the following:

- **Patent task:** We built several Transformer-based systems with and without pre-training approach and compared the performance for the sentence-level translation tasks.
- **Document-level translation task:** We applied two document-level NMT systems and found that the mBART model pre-trained on the large-scale corpora greatly outperformed.
- **Mixed-domain task:** We designed constrastive experiments with different data combinations for Myanmar↔English translation, and validated the effectiveness of data augmentation for low-resource translation tasks.

2 JPO Patent Task

2.1 Task Description

In the patent translation task, we conducted the experiments on the JPO Patent Corpus (JPC) version 4.3 that is constructed by the Japan Patent

Proceedings of the 7th Workshop on Asian Translation, pages 135–141

Office (JPO). Same as the previous tasks in WAT 2019 (Nakazawa et al., 2019), it consists of patent description translation sub-tasks for Chinese-Japanese, Korean-Japanese, and English-Japanese. Each language pair's training set contains 1M parallel sentences individually, which cover four patent sections: Chemistry, Electricity, Mechanical engineering, and Physics, based on International Patent Classification (IPC). Using the official training, develop, and test split provided by the organizer without other resources, we trained individual unidirectional Transformer models for each language pair. In addition, pre-training approach for sentence-level translation has been explored in this task.

2.2 Data Processing

As the baseline NMT systems data preparation suggested[1], we pre-tokenized the data with the following tools: Juman version 7.01[2] for Japanese; Stanford Word Segmenter version 4.0.0[3] for Chinese; Mecab-ko[4] for Korean, and Moses tokenizer for English.

For the byte-pair encoding (BPE)-based SentencePiece model (Kudo and Richardson, 2018) training, we set the vocabulary size to 100,000 and threshold of occurrence to 10 times for subword units (Sennrich et al., 2016) removal from the vocabulary, following same data preparation by BPE for the baseline NMT system released by the organizer[5]. Moreover, we merged the source and target sentences and trained a joint vocabulary for the NMT systems. For the text input to mBART fine-tuning, we used the same 250,000 vocabulary as in the pre-trained mBART model across the 25 languages, which was also tokenized with a SentencePiece model based on BPE method. Note that the aforementioned pre-tokenization was not applicable to the fine-tuning approach.

2.3 Model

Firstly, we built models based on the standard Transformer (Vaswani et al., 2017) with the implementation in the Fairseq toolkit (Ott et al., 2019).

Models	Transformer	mBART
Vocab size	100k	250k
Embed. dim.	1024	1024
Tied embed.	Yes	Yes
FFN dim.	4096	4096
Attention heads	8	16
En/Decoder layers	6	12
Label smoothing	0.1	0.2
Dropout	0.3	0.3
Attention dropout	0.1	0.1
FFN dropout	0.1	0.1
Learning rate	$1e^{-3}$	$3e^{-5}$

Table 1: JPO models settings comparison.

Intuitively, we tied the input embedding layers of encoder and decoder together with the decoder output embedding layers (Press and Wolf, 2017) for the tokenized input as well as the detokenized output. As a result, a large amount of parameters were automatically saved without depressing the performance. The model was optimized with Adam (Kingma and Ba, 2015) using $\beta 1 = 0.9$, $\beta 2 = 0.98$, and $\epsilon = 1e^{-8}$. Same as (Susanto et al., 2019), we used the learning rate schedule of 0.001 and maximum 4000 tokens in a batch, where the parameters were updated after every 2 epochs.

Secondly, we fine-tuned on the JPO patent corpus using the mBART auto-encoder model (Liu et al., 2020), which has been pre-trained on large-scale monolingual CommonCrawl (CC) corpus in 25 languages using the BART objective (Lewis et al., 2020). Specifically, we used the mBART models in a teacher-forcing manner, where the pre-trained mBART weights[6] ($\sim$ 680M parameters) were loaded. Then, our student models were utterly built upon the bi-text data, which fed the source language and target language into the pre-trained encoder and decoder for fine-tuning. We experimented our mBART and standard Transformer with the hyper-parameters summarized in Table 1 on 4 Nvidia V100 GPUs.

Finally, the best performing models on the validation sets was selected and applied for decoding the test sets. Furthermore, we trained three independent models with different random seeds in order to perform ensemble decoding.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/baseline/dataPreparationJEp.html

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[3] https://nlp.stanford.edu/software/segmenter.shtml

[4] https://bitbucket.org/eunjeon/mecab-ko/

[5] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/baseline/dataPreparationBPE.html

[6] https://dl.fbaipublicfiles.com/fairseq/models/mbart/mbart.CC25.tar.gz

Task	Model	BLEU	Human
N zh-ja	XFMR, sing.	48.17	-
N zh-ja	XFMR, ens.	**48.44**	-
N zh-ja	mBART sing.	48.17	-
N zh-ja	mBART ens.	48.09	4.51
N ja-zh	XFMR, sing.	39.24	-
N ja-zh	XFMR, ens.	**41.65**	-
N ja-zh	mBART sing.	40.53	-
N ja-zh	mBART ens.	41.52	4.64
N ko-ja	XFMR, sing.	71.47	-
N ko-ja	XFMR, ens.	**72.20**	-
N ko-ja	mBART sing.	68.32	-
N ko-ja	mBART ens.	69.37	4.64
N ja-ko	XFMR, sing.	69.45	-
N ja-ko	XFMR, ens.	**71.30**	-
N ja-ko	mBART sing.	70.77	-
N ja-ko	mBART ens.	70.48	4.73
N en-ja	XFMR, sing.	44.02	-
N en-ja	XFMR, ens.	**45.43**	-
N en-ja	mBART sing.	44.21	-
N en-ja	mBART ens.	44.52	4.42
N ja-en	XFMR, sing.	41.89	-
N ja-en	XFMR, ens.	**43.57**	-
N ja-en	mBART sing.	43.01	-
N ja-en	mBART ens.	43.51	4.59
EP zh-ja	XFMR, sing.	39.41	-
EP zh-ja	XFMR, ens.	**40.60**	-
EP zh-ja	mBART sing.	38.56	-
EP zh-ja	mBART ens.	38.54	-

Table 2: JPO task results. "XFMR" is short for Transformer and **HUMAN** refers to the final results provided by the task organizers. Readers may refer to the task overview for the detailed breakdown for each test set.

2.4 Results

As shown in Table 2, our model performance for the patent task has been split into four parts for standard Transformer and mBART approaches, with respect to the single and ensemble models. Note that only the results of the `test-N`[7] set and the Expression Pattern task (JPCEP) for were reported in the table for brevity. Here, we present the results based on the automatic metrics scores, as well as the human evaluation results[8].

In general, the Transformers' single model decoding results lagged behind that of the ensemble decoding in all directions. Without using any other resources, our best submissions of Transformer models obtain the first place on the WAT leaderboard[9] for ja-zh, and ja-en.

In terms of the fine-tuning results, we observed that the mBART single models outperformed the Transformer single models in 5 out of 7 language pairs, where the maximum margins can reach as much as 1.3 BLEU points (i.e., ja-zh and ja-ko). However, the ensemble model decoding of the mBART models could hardly boost the gains as we expected, which indicates that the advantages of Transformer-based pre-training approach can not be reflected in the JPO patent tasks when the training data size is sufficient (e.g., 1M).

3 Document-Level Translation Task

3.1 Task Description

In this year, WAT workshop introduced a new document-level translation task with sub-tasks from the perspective of two different domains: scientific paper and business conversation. In particular, we participate in the business conversation sub-task in WAT 2020. We followed the instruction of the shared-task organizer, using the Business Scene Dialogue (BSD) corpus for the dataset including training, development and test data. The BSD corpus consist of 20,000 training, 2,051 development and 2,120 test sentences from 670, 69, 69 documents, respectively.

Considering the limited document-level parallel data (<1k) in BSD training and development sets, we supposed that auxiliary document-level resources would be necessarily important. Therefore, we performed constrastive experiments with and without additional resources for this task. In particular, we appended the Japanese-English Subtitle Corpus (JESC) training set to the original BSD corpus, which brings in about 2.8M ja↔en sentences. We trained a context-aware hierarchical attention network (HAN) from scratch and fine-tuned on the BSD corpus using the mBART models.

3.2 Data Processing

For the document-level NMT tasks, we utilized the contextual information of 3 sentences instead of the entire documents in the dataset for both the HAN and mBART models. Similar to the data preprocessing illustrated in Section 2.2, we ran the

[7] is a union of JPCN{1,2,3} subsets

[8] Human evaluation results of the JCPEP tasks are not yet visible as the time of this writing.

[9] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Models	HAN$_{\texttt{joint}}$	mBART
Vocab size	32k	250k
Embed. dim.	512	1024
Tied embed.	Yes	Yes
FFN dim.	2048	4096
Attention heads	8	16
En/Decoder layers	6	12
Label smoothing	0.1	0.2
Dropout	0.1	0.3
Attention dropout	0.1	0.1
FFN dropout	0.1	0.1
Learning rate	$1e^{-2}$	$3e^{-5}$
Context size	3	3

Table 3: Comparison of models settings on the BSD tasks.

Juman analyzer to segment the Japanese characters but did nothing on the English documents for the HAN models. After pre-tokenization, we fed the Japanese and English documents into separate SentencePiece models (SPM) to train BPE subword units. The subword vocabulary size is 32,000 with 100% character coverage. On the other hand, we tokenized for the fine-tuning model with the pre-trained mBART multilingual vocabulary with 250,000 subword tokens. None of additional preprocessing was required in this implementation. For both two experimental settings, all empty lines and sentences exceeding 512 subword tokens have been removed from the training set.

3.3 Model

Firstly, we explored the context-aware based HAN models on the BSD corpus with the OpenNMT toolkit (Klein et al., 2017), where the document context of 3 previous sentences were integrated for global context encoding and decoding of the source and target languages, respectively. Intuitively, we trained the HAN$_{\texttt{base+}}$ models as baselines, which were essentially sentence-level Transformer-based models. Then, a multi-encoder and multi-decoder Transformer were learned based on sentence-level models. Finally, we built HAN$_{\texttt{joint+}}$ models upon the multi-encoder and multi-decoder models.

Besides the HAN models, we fine-tuned on the BSD corpus using the mBART auto-encoder pre-trained model via the Fairseq toolkit, as mentioned in Section 2.3. Since the pre-traind mBART model initially can handle more than one sentences, it owns very good compatibility of the document-level machine translation tasks. In this case, we considered the tri-sentence segments[10] as documents of the training sets, and fed them into the pre-trained model to learn dependencies between sentences. We trained the HAN$_{\texttt{joint+}}$ and mBART models on 4 V100 GPUs, whose model parameters have been shown in Table 3.

Task	Model	BLEU	Human
en-ja	HAN$_{\texttt{joint+}}$ sing.	13.58	-
en-ja	mBART$_{\texttt{doc+}}$ sing.	19.28	-
en-ja	mBART$_{\texttt{doc+}}$ ens.	19.43	4.20
ja-en	HAN$_{\texttt{joint+}}$ sing.	17.77	-
ja-en	mBART$_{\texttt{doc+}}$ sing.	22.10	-
ja-en	mBART$_{\texttt{doc+}}$ ens.	23.15	4.19

Table 4: Comparisons of HAN and mBART best models results in the BSD task. The results shown with + used JESC auxiliary corpus during training.

3.4 Results

We show the best BLEU scores that the HAN and mBART models can achieve in Table 4. Under single model decoding, we observed that the mBART$_{\texttt{doc+}}$ models could lead far ahead the HAN$_{\texttt{joint+}}$ models by 5.7 and 4.3 BLEU scores in the BSD en-ja and ja-en tasks, respectively. It indicates that the advantages of pre-trainining are substantial in the BSD translation tasks. Moreover, our best submissions of the mBART$_{\texttt{doc+}}$ models with ensemble model decoding achieved the first place on the WAT leaderboard in human evaluation scores for both directions.

To investigate how important the document-level translation is and how much gains can be achieved by using other resources, we performed the ablative studies upon several mBART settings, where the results are shown in Table 5. On one hand, HAN$_{\texttt{base+}}$ sentence-level models performed worst among all the listed models. However, mBART$_{\texttt{sen}}$ models incredibly outperformed the baselines due to the pre-training manner, even without additional resources. On the other hand, we observed that the mBART$_{\texttt{doc}}$ could hardly overwhelm the mBART$_{\texttt{sen}}$ until additional JESC corpus was leveraged, where over 1 BLEU gains were obtained for both directions. Furthermore, we found that the mBART$_{\texttt{sen+}}$ and mBART$_{\texttt{doc+}}$ models have achieved remarkable improvements by adding the

[10]The BSD training and JESC corpus have been expanded into 6,927 and 959,399 tri-sentence segments, respectively.

Task	Model	BLEU	Human
en-ja	HAN$_{\texttt{base+}}$ sing.	13.05	-
en-ja	mBART$_{\texttt{sen}}$ sing.	14.74	3.55
en-ja	mBART$_{\texttt{doc}}$ sing.	14.49	-
en-ja	mBART$_{\texttt{sen+}}$ sing.	18.30	-
en-ja	mBART$_{\texttt{doc+}}$ sing.	19.28	-
en-ja	mBART$_{\texttt{doc+}}$ ens.	19.43	4.20
ja-en	HAN$_{\texttt{base+}}$ sing.	16.88	-
ja-en	mBART$_{\texttt{sen}}$ sing.	17.02	3.57
ja-en	mBART$_{\texttt{doc}}$ sing.	15.62	-
ja-en	mBART$_{\texttt{sen+}}$ sing.	20.68	-
ja-en	mBART$_{\texttt{doc+}}$ sing.	22.10	-
ja-en	mBART$_{\texttt{doc+}}$ ens.	23.15	4.19

Table 5: Ablative study on the mBART in the BSD task. "sen" means using the mBART pre-training for the sentence-level translation evaluation, and the BLEU score of it calculated on the concatenation of all translated sentences.

JESC corpus for training, which explicitly reflects that data hungry effect of the BSD corpus remains a challenge. Some examples whose translation quality was improved by considering context in BSD tasks have been illustrated in Table 6.

4 Mixed-domain Task

4.1 Task Description

Despite the Myanmar-English mixed-domain tasks were excluded in the final evaluation this year, our experimental task is described in this section. We trained the models on both the University of Computer Studies, Yangon (UCSY) corpus only (Ding et al., 2018) and evaluated the model with a portion of the Asian Language Treebank (ALT) corpora (Ding et al., 2019, 2020). The UCSY corpus consists of approximately 200,000 sentences, while the ALT validation and test sets include 1,000 sentences respectively. Due to the low resource nature of the Myanmar-English language pair and the added difficulty of domain adaptation, we trained additional models that compiled with Myth Corpus[11] as other resources for the task participation, and compared them with the models using training data provided by the shared task only.

4.2 Data Processing

For the mix-domain task, some noisy double quotes from training data were cleaned first. Then we tokenized it using Pyidaungsu Myanmar Tokenizer[12] in syllable and word level tokenization for Myanmar sentences, and English sentences were fed directly to the SentencePiece model to produce subword units. Accordingly, we augmented the Myanmar data by three types (i) original, (ii) syllable, and (iii) word, where the training datasets could be built upon different combinations of these three types of Myanmar data, e.g., my (original+word)-en, my (original+syllable+word)-en, etc. In practice, we simply replicated the English sentences accordingly to match the number of sentences for the augmented Myanmar data during training.

4.3 Model

We experimented with several Transformer models using Marian[13] toolkit (Junczys-Dowmunt et al., 2018) for my-en and en-my, respectively. We separately trained four models for both direction with the hyper-parameter setting shown in Table 7, each of which corresponds to one combination of training data as mentioned in Section 4.3. Therefore, we had eight models to be trained in total, which can be denoted as: (i) my (original)↔en (BASE), (ii) my (original+word)↔en (WORD), (iii) my (original+syllable+word)↔en (ALL), and (iv) my (original+word)↔en with Myth corpus (WORD+). All experimental models in this task were trained on 3 GP104 machines with 4 GeForce GTX 1080 GPUs in each, and the experimental results will be shown and analyzed in the following section.

4.4 Results

Table 8 presents the results of our experiments on the given ALT test dataset evaluation for two directions. The baseline model BASE performed the poorest in the en↔my translation models solely trained on the original dataset. By using data augmentation, however, we observed significant improvements in the BLEU scores in en-my and my-en models that trained together with Myanmar word and syllable data. Interestingly, we also found that the BLEU score dropped down by 4.7 when syllable data was added during en-my model training (ALL vs. WORD), yet the similar performance decay did not appear in the my-en models. On the other hand, the models trained with additional Myth corpus (WORD+) outperformed the other three models for both directions because it could help on

[11] Available at https://github.com/alvations/myth

[12] https://github.com/kaunghtetsan275/pyidaungsu

[13] https://marian-nmt.github.io

Source	景気はどうです おかげさまで、**順調です**。 最近、新しい施設が稼働開始しまして、その管理で忙しくて。ああ、それ、御社のサイトで読みましたよ。**おめでとうございます**。
Reference	How's business lately? **It's been good**. **We** recently commissioned a new facility so I've been busy managing that. I read about that on your company website. **Congratulations**.
HAN`base+`	How's the economy? Thank you, **I'm good**. I've been busy with that management since the new facility started recently. Oh, I read that on your website. **Congratulations**.
HAN`joint+`	How's the economy? Thank you, **it's fine**. **There's been** a new facility running recently, and I've been busy managing it. Oh, I read it on your website. **Thank you**.
mBART`doc+`	How's the economy going? **It's going well** thanks to you. **We** recently opened a new facility and I've been busy managing it. Oh, I read that on your website. **Congratulations**.
Source	しかし、**どのような商品の取引であっても**、一般的に輸出入の手順**は同**じです。あなたの職務は主に、北米からアジアへの輸出品に関する輸出書類を用意することになります。 あとで、当**部署のエレ** インさんにやり方を説明してもらいます。
Reference	But **regardless of the product traded**, the procedures for exporting or importing are generally the same. Your task will mainly be preparing export documents for products from **North America going to Asia**. **Elaine in our department** will teach you how it's done later.
HAN`base+`	However, **even if it's a commodity exchange**, it's the same procedure as export procedures. You will mainly prepare export documents for exports from **North America**. I will explain **how Elaine will do it** later.
HAN`joint+`	But **any product transaction** is **commonly** the export process. You will mainly prepare export documents for exports **from North America to Asia**. I'm going to need you to explain **how you do it** later on in the department.
mBART`doc+`	But **regardless of the product deal**, the standard export procedure is the same. You will be required to prepare export documents on exports **from North America to Asia**, mainly. I will have **Elaine from our department** explain how to do it later.

Table 6: Translation examples: Comparison of the HAN and mBART models for BSD ja-en task. All the results shown here are obtained from single model decoding.

Vocabulary size	380k
Embedding dim.	1024
Tied embeddings	Yes
Transformer FFN dim.	4096
Attention heads	8
En/Decoder layers	4
Label smoothing	0.1
Dropout	0.1
Batch size	12
Attention weight dropout	0.1
Transformer FFN dropout	0.1
Learning rate	$1e^{-4}$

Table 7: Mixed-domain model parameter settings

Task	**Model**	**BLEU**
ALT2 my-en	`BASE`	6.9
ALT2 my-en	`WORD`	11.3
ALT2 my-en	`ALL`	12.9
ALT2 my-en	`WORD+`	14.2
ALT2 en-my	`BASE`	14.9
ALT2 en-my	`WORD`	22.1
ALT2 en-my	`ALL`	17.4
ALT2 en-my	`WORD+`	24.4

Table 8: Mixed-domain Task Results. "+" means the model was trained with additional Myth corpus.

the data hunger nature of low resource languages. Furthermore, our best BLEU results were achieved by the two `WORD+` models, which already or nearly surpassed the shared task organizer's baseline results on the WAT learderboard. Our approach in this way of amplifying training data size gave the improvement of BLEU score while using a single Marian NMT model. We need further discovery by turning model hyper-parameters and/or different modeling approaches.

5 Conclusion

We presented our submissions (team ID: *goku20*) to the WAT 2020 shared translation tasks in this paper. We trained Transformer-based NMT systems across different tasks. We found that additional training datasets from other resources could lead to substantial performance gains on smaller data sets. We also validated the capability of Transformers with pre-training in dealing with the sentence-level and document-level tasks, especially when the data hungry problem appeared. Finally, we attempted data augmentation approaches on the low-resource language translation tasks and achieved outperforming experimental results.

References

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):5.

Chenchen Ding, Sann Su Su Yee, Win Pa Pa, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. A Burmese (Myanmar) treebank: Guildline and analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3):40.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):17.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

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.

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.

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. *arXiv preprint arXiv:2001.08210*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. 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.

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.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. 2018. JESC: Japanese-English Subtitle Corpus. *Language Resources and Evaluation Conference (LREC)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Raymond Hendy Susanto, Ohnmar Htun, and Liling Tan. 2019. Sarah's participation in wat 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 152–158.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.