Pre-training via Leveraging Assisting Languages for Neural Machine Translation

Haiyue Song[1] **Raj Dabre**[2] **Zhuoyuan Mao**[1]
Fei Cheng[1] **Sadao Kurohashi**[1] **Eiichiro Sumita**[2]
[1]Kyoto University, Kyoto, Japan
[2]National Institute of Information and Communications Technology, Kyoto, Japan
{song, feicheng, zhuoyuanmao, kuro}@nlp.ist.i.kyoto-u.ac.jp,
{raj.dabre, eiichiro.sumita}@nict.go.jp

Abstract

Sequence-to-sequence (S2S) pre-training using large monolingual data is known to improve performance for various S2S NLP tasks. However, large monolingual corpora might not always be available for the languages of interest (LOI). Thus, we propose to exploit monolingual corpora of other languages to complement the scarcity of monolingual corpora for the LOI. We utilize script mapping (Chinese to Japanese) to increase the similarity (number of cognates) between the monolingual corpora of helping languages and LOI. An empirical case study of low-resource Japanese–English neural machine translation (NMT) reveals that leveraging large Chinese and French monolingual corpora can help overcome the shortage of Japanese and English monolingual corpora, respectively, for S2S pre-training. Using only Chinese and French monolingual corpora, we were able to improve Japanese–English translation quality by up to 8.5 BLEU in low-resource scenarios.

1 Introduction

Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) is known to give state-of-the-art (SOTA) translations for language pairs with an abundance of parallel corpora. However, most language pairs are resource poor (Russian–Japanese, Marathi–English) as they lack large parallel corpora and the lack of bilingual training data can be compensated by by monolingual corpora. Although it is possible to utilise the popular back-translation method (Sennrich et al., 2016a), it is time-consuming to backtranslate a large amount of monolingual data. Furthermore, poor quality backtranslated data tends to be of little help. Recently, another approach has gained popularity where the NMT model is pre-trained through tasks that only require monolingual data (Song et al., 2019; Qi et al., 2018).

Pre-training using models like BERT (Devlin et al., 2018) have led to new state-of-the-art results in text understanding. However, BERT-like sequence models were not designed to be used for NMT which is sequence to sequence (S2S). Song et al. (2019) recently proposed MASS, a S2S specific pre-training task for NMT and obtained new state-of-the-art results in low-resource settings. MASS assumes that a large amount of monolingual data is available for the languages involved but some language pairs may lack both parallel and monolingual corpora and are "truly low-resource" and challenging.

Fortunately, languages are not isolated and often belong to "language families" where they have similar orthography (written script; shared cognates) or similar grammar or both. Motivated by this, in this paper we hypothesize that we should be able to leverage large monolingual corpora of other assisting languages to help the monolingual pre-training of NMT models for the languages of interest (LOI) that may lack monolingual corpora. Wherever possible, we subject the pre-training corpora to script mapping which should help minimize the vocabulary and distribution differences, respectively, between the pre-training, main training (fine-tuning) and testing time datasets. This should help the already consistent pre-training and fine-tuning objectives leverage the data much better and thereby, possibly, boost translation quality.

To this end, we experiment with ASPEC Japanese–English translation in a variety of low-resource settings for the Japanese–English parallel corpora. Our experiments reveal that while it's possible to leverage *unrelated languages* for pre-training, using *related languages* is extremely important. We utilized Chinese to Japanese script mapping to maximize the similarities between the assisting languages (Chinese and French) and the languages of interest (Japanese and English).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 279–285
July 5 - July 10, 2020.

We show that only using monolingual corpora of Chinese and French for pre-training can improve Japanese–English translation quality by up to 8.5 BLEU.

The contributions of our work are as follows:

1. Leveraging assisting languages: We give a novel study of leveraging monolingual corpora of related and unrelated languages for NMT pre-training.

2. Empirical evaluation: We make a comparison of existing and proposed techniques in a variety of corpora settings to verify our hypotheses.

2 Related work

Our research is at the intersection of works on monolingual pre-training for NMT and leveraging multilingualism for low-resource language translation.

Pre-training has enjoyed great success in other NLP tasks with the development of methods like BERT (Devlin et al., 2018). Song et al. (2019) recently proposed MASS, a new state-of-the-art NMT pre-training task that jointly trains the encoder and the decoder. Our approach builds on the initial idea of MASS, but focuses on complementing the potential scarcity of monolingual corpora for the languages of interest using relatively larger monolingual corpora of other (assisting) languages.

On the other hand, leveraging multilingualism involves cross-lingual transfer (Zoph et al., 2016) which solves the low-resource issue by using data from different language pairs. Dabre et al. (2017) showed the importance of transfer learning between languages belonging to the same language family but corpora might not always be available in a related language. A mapping between Chinese and Japanese characters (Chu et al., 2012) was shown to be useful for Chinese–Japanese dictionary construction (Dabre et al., 2015). Mappings between scripts or unification of scripts (Hermjakob et al., 2018) can artificially increase the similarity between languages which motivates most of our work.

3 Proposed Method: Using Assisting Languages

We propose a novel monolingual pre-training method for NMT which leverages monolingual corpora of assisting languages to overcome the scarcity of monolingual and parallel corpora of the languages of interest (LOI). The framework of our approach is shown in Figure 1 which consists of script mapping, data selection, pre-training and fine-tuning.

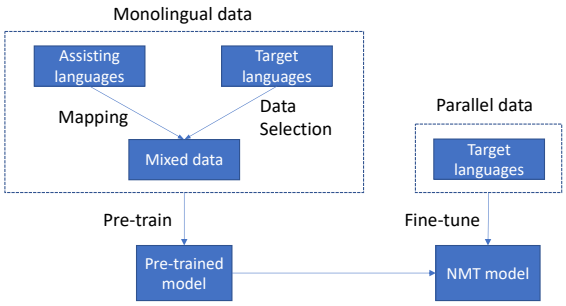

Figure 1: An overview of our proposed method consisting of script mapping, data selection, pre-training and fine-tuning

3.1 Data Pre-processing

Blindly pre-training a NMT model on vast amounts of monolingual data belonging to the assisting languages and LOI might improve translation quality slightly. However, divergences between the languages, especially their scripts (Hermjakob et al., 2018) and also the distributions of data between different training phases is known to impact the final result. Motivated by past works on using related languages (Dabre et al., 2017), orthography mapping/unification (Hermjakob et al., 2018; Chu et al., 2012) and data selection for MT (Axelrod et al., 2011), we propose to improve the efficacy of pre-training by reducing data and language divergence.

3.1.1 Script Mapping

Previous research has shown that enforcing shared orthography (Sennrich et al., 2016b; Dabre et al., 2015) has a strong positive impact on translation. Following this, we propose to leverage existing script mapping rules[1] or script unification mechanisms to, at the very least, maximize the possibility of cognate sharing and thereby bringing the assisting language closer to the LOI. This should strongly impact languages such as Hindi, Punjabi and Bengali belonging to the same family but written using different scripts.

For languages such as Korean, Chinese and Japanese there may exist a many to many mapping between their scripts. Thus, incorrect mapping of

[1]Transliteration is another option but transliteration systems are relatively unreliable compared to handcrafted rule tables.

characters (basic unit of a script) might produce wrong words and reduce cognate sharing. We propose two solutions to address this.

1. One-to-one mapping: Here we do not care about word level information and map each character in one language to its corresponding character in another language. Here, we just select the first mapping in the mapping list.

2. Many-to-many mapping with LM scoring: A more sophisticated solution is where for each tokenized word-level segment in one language we enumerate all possible combinations of mapped characters and use a language model in the other language to select the character combination with the highest score as the result.

3.1.2 Note on Chinese–Japanese Scripts

Japanese is written in Kanji which was borrowed from China. Over time the written scripts have diverged and the pronunciations are naturally different but there are a significant number of cognates written in both languages. As such pre-training on Chinese should benefit translation involving Japanese. Chu et al. (2012) created a mapping table between them which can be leveraged to further increase the number of cognates.

3.1.3 Data Selection

Often, the pre-training monolingual data and the fine-tuning parallel data belong to different domains. (Axelrod et al., 2011; Wang and Neubig, 2019) have shown that proper data selection can reduce the differences between the natures of data between different training domains and phases. In this paper we experiment with **(a)** Scoring monolingual sentences using a language model (LM) and selecting the highest scoring ones and **(b)** Selecting monolingual sentences to match the sentence length distribution of the development set sentences in the parallel corpus.

1. LM based data selection: We use a language model trained on corpora belonging to the domain that the fine-tuning data belongs to. We use this sort monolingual sentences according to LM score and use the top N sentences that are expected to be the most similar to the domain of the fine-tuning data.

2. Length based data selection: Algorithm 1 describes how to use the in-domain dataset ($TargetFile$; typically the sentences from the fine-tuning parallel corpus) to select $SelectNum$ lines from the out-of-domain dataset ($InputFile$; typically the monolingual corpus). When selecting monolingual data of languages of interest, we can first calculate the length distribution of parallel data as target distribution (the ratio of all lengths in $TargetFile$) and we fill the length distribution by selecting sentences from monolingual data of same language. As a result, the monolingual data and parallel data have similar length distribution.

Algorithm 1: Length Distribution Data Selection

Input : $TargetFile$, $InputFile$, $SelectNum$
Output : $SelectedLines$

$TargetDistribution \leftarrow \{\}$;
$CurrentDistribution \leftarrow \{\}$;
$SelectedLines \leftarrow \{\}$;
$TargetNum = \#\ of\ Lines\ in\ TargetFile$;
foreach $Line \in TargetFile$ **do**
  $TargetD[len(Line)] += 1$;
foreach $Line \in InputFile$ **do**
  if $CurrentD[len(Line)]/SelectNum < TargetD[len(Line)]/TargetNum$ **then**
    $CurrentD[len(Line)] += 1$;
    $SelectedLines \leftarrow SelectedLines \cup \{Line\}$;

3.2 NMT Modeling

In order to train a NMT model we first use the pre-processed monolingual data for pre-training and then resume training this model on parallel data to fine-tune for the languages of interest.

We use MASS, which is a pre-training method for NMT proposed by Song et al. (2019). In MASS, the input is a sequence of tokens where a part of the sequence is masked and the pre-training objective is to predict the masked fragments using a denoised auto-encoder model. The NMT model is pre-trained with the MASS task, until convergence, jointly for both the source and target languages. Thereafter training is resumed on the parallel corpus, a step known as fine-tuning (Zoph et al., 2016).

4 Experimental Settings

We conducted experiments on Japanese–English (Ja–En) translation in a variety of simulated low-resource settings using the "similar" assisting lan-

guage pairs Chinese (Zh) and French (Fr) and the "distant" assisting language pairs Russian (Ru) and Arabic (Ar).

4.1 Datasets

We used the official ASPEC Ja–En parallel corpus (Nakazawa et al., 2016) provided by WAT 2019[2]. The official split consists of 3M, 1790 and 1872 train, dev and test sentences respectively. We sampled parallel corpora from the top 1M sentences for fine-tuning. Out of the remaining 2M sentences, we used the En side of the first 1M and the Ja side of the next 1M sentences as monolingual data for language modeling for data selection. We used Common Crawl[3] monolingual corpora for pre-training. To train LMs for data-selection of the assisting languages corpora, we used news commentary datasets [4]. While this data selection step for the assisting languages won't minimize the domain difference from the parallel corpus, it can help in filtering noisy sentences. In this paper we consider the ASPEC and news commentary data as in-domain and the rest of the pre-training data as out-of-domain.

4.2 Data Pre-processing

1. Normalization and Initial Filtering: We applied NFKC normalization to data of all languages. Juman++ (Tolmachev et al., 2018) for Ja tokenization, jieba[5] for Zh tokenization and NLTK[6] tokenization for other languages. We filtered out all sentences from the pre-training data that contain fewer than 3 and equal or more than 80 tokens. For Chinese data, we filtered out sentences containing fewer than 30 percent Chinese words or more than 30 percent English words.

2. Script Mapping: Chinese is the only assisting language that can be mapped to Japanese reliably. We converted Chinese to Japanese script to make them more similar by using the mapping table from (Chu et al., 2012) and the mapping approaches mentioned in the previous section. French and English are written using the Roman alphabet and do not need any script mapping. We did not perform script mapping for Arabic and Russian to show the impact of using distant languages (script-wise as well as linguistically).

3. Data selection: We used KenLM (Heafield, 2011) to train 5-gram LMs on in-domain data for LM scoring based data selection and use ASPEC dev set for length distribution based data selection.

5 Results and Analysis

5.1 Training and Evaluation Settings

We used the tensor2tensor framework (Vaswani et al., 2018) [7], version 1.14.0., with its default "*transformer_big*" setting.

We created a shared sub-word vocabulary using Japanese and English data from ASPEC mixing with Japanese, English, Chinese and French data from Common Crawl. We used SentencePiece (Kudo and Richardson, 2018) and obtained a vocabulary with the size of roughly 64k . We used this vocabulary in all experiments except unrelated language experiment where Arabic and Russian were used instead of Chinese and French data.

We combined monolingual data of assisting languages and languages of interest (LOI; Japanese and English) for pre-training. When mixing datasets of different sizes, we always oversampled the smaller datasets to match the size of the largest.

For all pre-training models, we saved checkpoints every 1000 steps and for all fine-tuning models, we saved checkpoints every 200 steps. We used early-stopping using approximate-BLEU as target and stops when no gain after 10,000 steps for pre-training and 2,000 steps for fine-tuning. We fine-tuned different fine-tune settings from the last checkpoint of each pre-trained model.

For decoding we averaged 10 checkpoints of the fine-tuning stage with $\alpha = 0.6$ and $beamsize = 4$. We used sacreBLEU[8] to evaluate BLEU score for all translation evaluation.

5.2 Models Trained and Evaluated

5.2.1 Pre-trained Models

We separated pre-training settings into different blocks as shown in Table 1. Baseline model without fine-tuning is shown as A1. Zero (0M), low (1M) and rich (20M) monolingual-corpus scenarios are shown in parts B, C and D, respectively. Part E explores the impact of the two script mapping techniques on pre-training. Part F shows the impact of using related versus unrelated assisting languages.

[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/index.html#task.html
[3] http://data.statmt.org/ngrams/
[4] http://data.statmt.org/news-commentary/v14/
[5] https://github.com/fxsjy/jieba
[6] https://www.nltk.org
[7] https://github.com/tensorflow/tensor2tensor
[8] https://github.com/mjpost/sacreBLEU

#	Pre-training					Fine-tuning							
	Data pre-processing	Zh	Ja	En	Fr	En→Ja				Ja→En			
						3K	10K	20K	50K	3K	10K	20K	50K
A1	-	-	-	-	-	2.5	6.0	14.4	22.9	1.8	4.6	10.9	19.4
B1	1-to-1 Zh→Ja mapping + LM	20M	-	-	-	**5.3**	**14.5**	**20.0**	**26.1**	**3.7**	**11.2**	**15.6**	**20.5**
B2	LM	-	-	-	20M	3.4	9.1	14.9	23.4	2.1	6.3	11.3	17.7
B3	1-to-1 Zh→Ja mapping + LM	20M	-	-	20M	2.1	6.7	12.6	21.9	2.2	6.3	10.7	16.8
C1	LD	-	1M	1M	-	7.7	15.8	**20.7**	26.3	7.2	**12.7**	15.7	19.6
C2	1-to-1 Zh→Ja mapping + LD	20M	1M	1M	-	**8.3**	**16.4**	20.2	**26.9**	**7.5**	12.5	**16.3**	**20.7**
C3	LD	-	1M	1M	20M	**8.3**	15.3	19.3	26.7	6.8	12.3	15.4	20.4
C4	1-to-1 Zh→Ja mapping + LD	20M	1M	1M	20M	7.1	15.2	19.4	26.5	6.6	12.0	15.4	19.9
D1	LD	-	15M	15M	-	9.6	**17.2**	21.5	**28.0**	**8.6**	**13.5**	**16.8**	**20.9**
D2	1-to-1 Zh→Ja mapping + LD	20M	15M	15M	-	**9.7**	17.1	**21.6**	27.2	8.3	13.3	16.7	20.6
D3	LD	-	15M	15M	20M	7.7	15.0	19.8	26.3	6.3	11.7	15.1	20.2
D4	1-to-1 Zh→Ja mapping + LD	20M	15M	15M	20M	7.7	14.9	19.7	26.1	6.5	11.4	15.4	19.8
E1	1-to-1 Zh→Ja mapping	20M	20M	20M	20M	**7.0**	**13.4**	**19.3**	**25.7**	**5.9**	**11.1**	**15.0**	**19.8**
E2	LM-scoring Zh→Ja mapping	20M	20M	20M	20M	6.3	12.7	18.1	24.7	5.7	10.3	13.5	18.9
F1	LM-scoring	-	20M	20M	-	4.7	11.7	16.6	23.9	4.5	9.1	12.9	18.3
F2	1-to-1 Zh→Ja mapping + LM-scoring	20M	20M	20M	20M	**7.0**	**13.4**	**19.3**	**25.7**	**5.9**	**11.1**	**15.0**	**19.8**
F3	LM-scoring + Ar20M + Ru20M	-	20M	20M	-	4.8	12.1	18.1	25.1	4.4	10.2	13.5	18.9

Table 1: Low-resource pre-training experiments. Part A shows the baseline results. Part B, C, and D show results on monolingual zero, low and rich-resource scenarios. Part E shows results of two different mapping methods. And part F shows results of using related and unrelated languages. LD is with the meaning of "length distribution". Best results of each part are in bold.

5.2.2 Fine-tuned Models

We evaluated both Ja→En and En→Ja models with four parallel dataset size settings, 3K, 10K, 20K and 50K, selected from the previously selected 1M ASPEC parallel sentences.

In Table 1, we show results of several experimental settings to analyse the effect of: pre-training data size, Zh→Ja mapping methods and choices of unrelated languages versus related languages.

In our preliminary experiments we found out that 1-to-1 script mapping was not only faster but better than LM-scoring based script mapping. Furthermore, using length distribution was better than LM based data selection for the languages of interest (Japanese and English). Due to lack of space we only report core results using 1-to-1 script mapping (for assisting languages) and length distribution based data selection (for languages of interest).

5.3 Monolingual Zero and Low-resource scenario

The results of zero-resource and low-resource scenario are shown in parts B and C of Table 1. In these settings we used either no monolingual data or very little (1M) monolingual data for Japanese and English.

In part B, for a zero-monolingual data scenario, we observed large improvements, a maximum of 8.5 BLEU score over the baseline setting (A1), on all fine-tuning settings over model without fine-tuning when using only Chinese monolingual data (B1). Using only French data also gives better results on almost all fine-tuning settings, but not as large as that of using only Chinese data. Combining Chinese and French data, led to reduction in scores indicating some incompatibility between them.

In part C of the table, when there are 1M Japanese and English monolingual sentences, combining them with 20M Chinese data also gives improvements up to 1.1 BLEU points over A1. Combining with French data only gives occasional improvements. In this setting too, combining Chinese and French data led to reduction in performance.

Although French and English share cognates and have similar grammar, we have not performed explicit script mapping like we did for Chinese to make it more similar to Japanese. In the future we will investigate whether using a simple dictionary to map French to English can alleviate this issue.

We can draw the following conclusions,

1. Utilizing monolingual corpora of other languages **IS** beneficial.
2. Using similar languages (French and English) will **sometimes** give better results.
3. There may be **conflicts** between data of different assisting languages.

5.4 Monolingual resource-rich scenario

In part D, we found that there is less need to combine related language data when we use a large monolingual data of target languages. Only combining with Chinese data (D2) is comparable with pure Japanese-English monolingual pre-training (D1). Using French data degrades the translation quality in most settings. Thus, assisting languages become interfering languages in scenarios where large amounts of monolingual data are available for languages to be translated.

5.5 Chinese to Japanese mapping

In part E, we compared our two proposed script mapping methods. Results showed that the one-to-one mapping (character-level mapping) gives better BLEU score than word-level mapping consistently on most fine-tuning settings, about 0.7 to 1.0 in most cases. The word-level mapping gives lower score than baseline in Ja→En 50K case. One possible reason is that the Chinese and Japanese tokenizers cut the words in different granularity. So that applying Japanese LM to Chinese data may not work well. Therefore, we focus on 1-to-1 mapping experiments.

5.6 Unrelated language VS related language

In part F of the table, we compare pre-training on related languages versus unrelated languages. We saw that using Arabic and Russian as unrelated assisting languages in addition to Japanese and English, gives about 0.1 to 1.5 BLEU improvement over the baseline (G1) which uses only Japanese and English monolingual data. This is surprising and it shows that leveraging any additional language is better than not leveraging them. However, using (mapped) Chinese and French instead of Arabic and Russian yields about 2 to 2.7 BLEU score improvements. This clearly indicates that language relatedness is definitely important. In the future, we will consider more rigorous ways of increasing relatedness between pre-training corpora by using existing dictionaries and advanced script unification/mapping techniques instead of simple script mapping techniques.

6 Conclusion

In this paper we showed that it is possible to leverage monolingual corpora of other languages to pre-train NMT models for language pairs that lack parallel as well as monolingual data. Even if monolingual corpora for the languages of interest are unavailable, we can successfully improve translation quality by up to 8.5 BLEU, in low-resource settings, using monolingual corpora of assisting languages. We showed that the similarity between the other (assisting) languages and the languages to be translated is crucial and leveraged script mapping wherever possible. In the future, we plan to experiment with even more challenging language pairs such as Japanese–Russian and attempt to leverage monolingual corpora belonging to diverse language families.We might be able to identify subtle relationships among languages and approaches to better leverage assisting languages for several NLP tasks.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA.

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'12)*, pages 2149–2152, Istanbul, Turkey. European Language Resources Association (ELRA).

Raj Dabre, Chenhui Chu, Fabien Cromieres, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. Large-scale Dictionary Construction via Pivot-based Statistical Machine Translation with Significance Pruning and Neural Network Features. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 289–297, Shanghai, China.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box Universal Romanization Tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

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.

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)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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 *Proceedings of the 27th Neural Information Processing Systems Conference (NIPS)*, pages 3104–3112, Montréal, Canada.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, USA. Association for Machine Translation in the Americas.

Xinyi Wang and Graham Neubig. 2019. Target Conditioned Sampling: Optimizing Data Selection for Multilingual Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1575, Austin, USA.