Master Thesis

Linguistically-driven Multi-task Pre-training for Low-resource Neural Machine Translation

Supervisor: Prof. Sadao Kurohashi

Department of Intelligence Science and Technology
Graduate School of Informatics
Kyoto University

Zhuoyuan Mao

February 1st, 2021

Linguistically-driven Multi-task Pre-training for Low-resource Neural Machine Translation

Zhuoyuan Mao

Abstract

Although language-agnostic sequence-to-sequence pre-training methods using large monolingual corpora lead to nontrivial improvements for low-resource neural machine translation (NMT), such methods still can not generate adequate and fluent translations. Previous work showed that specific linguistic knowledge can improve low-resource NMT, thus we suppose that injecting linguistic knowledge in the pre-training phase will result in further significant improvements for low-resource NMT.

In this work, we employ one of the state-of-the-art sequence-to-sequence pre-training methods for low-resource NMT, MASS, as the main baseline, and propose novel pre-training alternatives to MASS: JASS (JApanese-specific Sequence to Sequence) for language pairs involving Japanese as the source or target language, and ENSS (ENglish-specific Sequence to Sequence) for language pairs involving English. JASS focuses on masking and reordering Japanese linguistic units called bunsetsu, while ENSS is proposed based on phrase structure masking and reordering tasks.

Experiments on ASPEC Japanese–English & Japanese–Chinese, Wikipedia Japanese–Chinese, News Commentary Japanese–Russian translation show that JASS and ENSS outperform MASS by up to +2.9 BLEU point for the Japanese–English tasks, +7.0 BLEU point for the Japanese–Chinese tasks, and +1.4 BLEU for the Japanese–Russian tasks. This shows the effectiveness of our newly proposed linguistically-driven methods, ENSS and JASS.

Empirical analysis, focusing on the relationship between individual parts within JASS and ENSS, reveals the complementary nature between their subtasks. Adequacy evaluation by using LASER, human evaluation, and case study focusing on JASS show that our proposed methods have a larger positive impact on the adequacy as compared to the fluency.

低資源ニューラル機械翻訳のための言語知識に基づくマルチタスク事前学習

内容梗概

特定の言語によらない sequence-to-sequence 事前学習は、大規模な単言語コーパスを活用することにより、ニューラル機械翻訳（NMT）の低資源言語対での翻訳精度を向上させることができるが、適切で流暢な翻訳文を既に生成できるわけではない。言語知識は低資源 NMT モデルを改善できることが先行研究で明らかにされたため、事前学習の段階で言語知識を注入することで低資源 NMT を大幅に改善できると考えられる。

本研究は、機械翻訳に対する state-of-the-art 事前学習手法の一つ、MASS をベースラインとし、新たな事前学習タスクとして、日本語をソース言語またはターゲット言語とする NMT モデルにおいて JASS (JApanese-specific Sequence to Sequence)、英語に対して ENSS (ENglish-specific Sequence to Sequence) を提案する。JASS と ENSS は、各々日本語文節、英語句構造を基に Masked Language Model と並べ替え (Reordering) タスクを同時に行う事前学習手法である。

ASPEC 日本語-英語と日本語-中国語、Wikipedia 日本語-中国語、News Commentary 日本語-ロシア語に行った翻訳実験によると、JASS と ENSS は日本語-英語、日本語-中国語、日本語-ロシア語の翻訳対にて MASS をそれぞれ最大+2.9 BLEU、+7.0 BLEU、+1.4 BLEU スコアで上回っている。これは、提案した言語知識に基づく事前学習手法の有効性を示している。

実証的分析により、JASS と ENSS 内部の各サブタスク間の補完的な性質が明らかになった。LASER を使用した適切性評価、人手評価、およびケーススタディは、JASS が流暢さに比べて適切性に相当大きなプラスの影響を与えることを示している。

Linguistically-driven Multi-task Pre-training for Low-resource Neural Machine Translation

Contents

1 Introduction **1**

2 Related Work **5**

2.1 Low-resource Neural Machine Translation 5

2.2 Pre-training Tasks for Neural Machine Translation 6

3 Proposed Methods **8**

3.1 Preliminary Background 8

3.1.1 MASS 8

3.1.2 Bunsetsu 9

3.1.3 Head-driven Phrase Structure Grammar 9

3.1.4 Head Finalization 10

3.2 Proposed Methods for Japanese 11

3.2.1 BMASS 11

3.2.2 BRSS 12

3.3 Proposed Methods for English 13

3.3.1 PMASS 13

3.3.2 HFSS 16

3.4 Multi-task Pre-training 18

4 Experimental Settings **20**

4.1 Pre-training and Fine-tuning for NMT 20

4.2 Datasets 20

4.3 Pre-processing 22

4.4 Training and Evaluation Details 23

4.5 Baselines 23

4.6 Pre-trained Models 24

4.7 Fine-tuned NMT Models 26

5 Results and Analyses **27**

5.1 NMT Results 27
5.2 Adequacy Evaluation 33
5.3 Human Evaluation 34
5.4 Case Study 35
5.5 Pre-training Accuracy 36

6 Conclusion **39**

Acknowledgments **40**

Appendix **A-1**
A.1 Results in Middle/High-resource Scenarios A-1

1 Introduction

Neural machine translation (NMT) [2, 48] has led to large improvements in machine translation quality when large parallel corpora are available for training. However, this need for parallel corpora strongly limits its usefulness for many language pairs (Russian–Japanese, Marathi–English) and domains (tourism, medical, social media) for which such corpora do not exist. Often, these “poor" language pairs consist of languages that have “rich" monolingual corpora. Therefore, it is possible to compensate the lack of bilingual training corpora by leveraging large monolingual corpora. One popular approach for this is data augmentation, e.g. by back-translation [42]. Another approach is pre-training the NMT model on tasks that only require monolingual corpora [37, 46].

Pre-training has seen a surge in popularity in NLP ever since models such as BERT [7] have led to new state-of-the-art results in text understanding. However, BERT-like models were not designed to be used for NMT in the sense that they are essentially language models and not sequence to sequence models. To address this, Song et al. [46] recently proposed masked sequence to sequence pre-training (MASS), a pre-training task for NMT and obtained new state-of-the-art results in low-resource settings.

Languages that are “rich" enough to have large monolingual corpora often have available tools for linguistic analysis. In addition, works such as Sennrich and Haddow [41] and Murthy et al. [28] have shown that “linguistic knowledge" can improve NMT without using additional corpora. It seems, therefore, natural to use both monolingual corpora and linguistic tools in bilingual low-resource scenarios. However, because NMT models are end-to-end, the manner in which linguistic hints should be provided is not always clear.

It is practical to extract linguistic features on the monolingual side. Therefore, pre-training provides an ideal framework both for leveraging monolingual corpora and improving NMT models with linguistic information. First, we focus on language pairs involving Japanese. Japanese is a language for

which very high quality syntactic analyzers have been developed [20, 27]. Moreover, large parallel corpora involving Japanese exist only for a few number of language pairs and domain. As such it is critical to leverage both monolingual corpora and the syntactic analysis of Japanese for optimal translation quality. On the other hand, pre-training for low-resource NMT is required not only for Japanese, but also for any other language because the number of the middle- or high-resource parallel corpora are definitely limited until today.[1)] Thus, we further extend our proposal to more general scenarios which can be transferred onto most of the languages.

First, we propose a linguistically motivated pre-training approach called JASS (Japanese-specific Sequence to Sequence). JASS is inspired by MASS, but focuses on syntactic analysis obtained by using a parser. In particular, we add syntactic constraints to the sentence-masking process of MASS to obtain our BMASS (Bunsetsu MASS) task.[2)] We also propose, BRSS (Bunsetsu Reordering based Sequence to Sequence), a linguistically motivated reordering task. Several previous works [21, 39] provided the evidence that "multi-task" pre-training combining various styles of self-supervised training tasks significantly results in superior results for NMT. Thus, our JASS is proposed upon a combination of the above-mentioned two tasks and is tailored for NMT involving Japanese.

Second, we also propose methods for English to leverage syntax-specific information in the pre-training phase. They are respectively named as PMASS (Phrase structure based MASS) & HFSS (Head Finalization based Sequence to Sequence), and their combination is denoted as ENSS (ENglish-specific Sequence to Sequence).[3)] Moreover, unlike proposed methods for Japanese, our proposed methods for English can be transplanted onto any SVO language.

[1)] Although language pairs involving English are usually middle- or high-resource scenarios (parallel corpora size over 100k), we deem that it is worth proposing methods for English because there still exist a large number of low-resource language pairs involving English.

[2)] For BMASS, bunsetsus are the elementary syntactic component of Japanese. It can be extracted by using KNP. [20, 27]

[3)] Head finalization [14] is the technique to reorder sentences in SVO language to be SOV-like sentences.

We experiment on ASPEC Japanese–English & Japanese–Chinese, Wikipedia Japanese–Chinese, News Commentary Japanese–Russian in a variety of pre-training settings. Our results show that BMASS, BRSS, and HFSS significantly outperform the state-of-the-art MASS pre-training while PMASS yields marginal improvements. Furthermore, we show that linguistically-driven multi-task pre-training methods (JASS & ENSS) lead to further improvements of up to +2.9 BLEU for Japanese to English, +2.7 BLEU for English to Japanese, +4.3 BLEU for Japanese to Chinese , +7.0 BLEU for Chinese to Japanese, +0.5 BLEU for Japanese to Russian, and +1.4 BLEU for Russian to Japanese respectively in low-resource scenarios.

Our analysis focuses on the relationship between different pre-training tasks, and the adequacy & fluency of corresponding translations. Specifically, we validate the superior translation adequacy improvement of the linguistically-driven methods by implementing automatic adequacy evaluation by using LASER, human evaluation, and case study. Furthermore, we confirm the complementary nature between masked language model and reordering pre-training task by pre-training accuracy evaluation.

To the best of our knowledge, this is the first time syntactic information is used in a sequence-to-sequence pre-training setting for NMT. The contributions of this paper can be summarized as follows:

1. **BMASS and BRSS:** Linguistically-driven novel pre-training methods for NMT involving Japanese.
2. **PMASS and HFSS:** Linguistically-driven novel pre-training methods for NMT involving English.
3. **Multi-task pre-training:** Showing that the multi-task training by combining masked language model and reordering task leads to better performance. Particularly, BMASS and BRSS can complement each other more provided that they are performed based on analogous syntactic units.
4. **Empirical evaluation:** A comparison of MASS, JASS, ENSS and other baseline methods for 6 translation directions and 3 different domains in several data size settings to identify situations where each technique can

be the most effective.

5. **Analysis:** Linguistic and statistical analysis of pre-training methods, their inter-relationships, and corresponding translations.

The JASS part of this work has been presented in LREC 2020 [24] and NLP 2020 [25].

2 Related Work

2.1 Low-resource Neural Machine Translation

There are mainly three lines of work related to improving NMT in low-resource situations: cross-lingual transfer, data augmentation and monolingual pre-training. These approaches are potentially complementary. Our work belongs to the monolingual pre-training category.

Cross-lingual transfer solves the low-resource issue by using data from different language pairs. One can use a richer language pair [59], or several language pairs at once [6, 8]. Murthy et al. [28] also proposed to reorder the assisting languages to be similar to the low-resource language.

Data augmentation involves the creation of synthetic bilingual data from monolingual data. In the popular back-translation approach [9, 11, 42], the source side of the data is synthesized using a MT system to back-translate the target side data. Recently, Zhou et al. [57] proposed to create this source side by rule-based reordering followed by word-by-word translation.

In monolingual pre-training approaches, all or part of a model is first trained on tasks that require monolingual data.[1)] Pre-training has enjoyed great success in other NLP tasks with the development of GPT [38], BERT [7] and many others [33, 47, 53].

Pre-training schemes like BERT were designed for natural language understanding (NLU) tasks and are not directly suitable for NMT. Conneau and Lample [5] and Ren et al. [40] proposed variants that can be trained in a multilingual way. However, they train the encoder and decoder independently. To address this, Song et al. [46] 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 adds more diverse and linguistically motivated training objectives.

Linguistic information is known to be useful for NMT [41], especially in low-resource scenarios. Outside of pre-training, the works [28, 55, 57] have

[1)] This is an instance of “transfer learning" just like Cross-lingual transfer. “Pre-training" often implies that the training task differs from the target task.

successfully used a linguistically-motivated reordering like the one in our BRSS task. Sun et al. [47] used some linguistically motivated pre-training tasks for Text Understanding. We are not aware of previous work on linguistically motivated pre-training tasks for NMT.

2.2 Pre-training Tasks for Neural Machine Translation

After the appearance of BERT [7], several pre-training methods were proposed for enhancing NMT [5, 21, 22, 23, 39, 40, 44, 45, 46, 50, 51, 52]. Specifically, Song et al. [46] proposed a random span reconstruction task to pre-train a sequence-to-sequence framework for NMT; Wang et al. [50] first proposed using shuffling, deleting, and replacing operations to implement the denoising pre-training for NMT system; following them, Lewis et al. [21] combined the denoising methods with masked language model pre-training of Song et al. [46] and provided detailed empirical results for a large number of language pairs; mBART [23] is a multilingual sequence-to-sequence denoising pre-training pre-trained by denoising tasks on 25 languages including Japanese, English, Chinese, Russian and others, which can be deemed as an extension of Lewis et al. [21]; other works focus on leveraging the cross-lingual supervision between languages by word alignment [22], phrase alignment [40], sentence-level alignment [5], code switching technique [52], or assisting languages (shared scripts) [45].

Among the above mentioned tons of pre-training techniques for NMT, we observe that no work has focused on leveraging specific linguistic features for NMT. Syntactic span masking [58] and semantic-aware BERT [56] have been proposed by using linguistically-driven supervisions for language understanding tasks. However, linguistically-driven methods for sequence-to-sequence pre-training should be considered and explored.

There also exist works focusing on improving MASS. Siddhant et al. [44] adapted MASS to multilingual scenarios; Qi et al. [36] proposed using n-stream self-attention mechanism to enhance MASS for language generation task. No previous work attempted to enhance MASS from the linguistic point of view, which will be first explored in our work.

Moreover, Wang et al. [51] pointed out that multi-task learning can significantly benefit multilingual NMT. Besides the MT task, the essential jointly learned tasks should be masked language model task and denoising (reconstruction) tasks, which are two basic pre-training styles based on which we propose our linguistically-driven methods.

3 Proposed Methods

We first give some background information and then describe JASS and ENSS, our pre-training techniques.

3.1 Preliminary Background

We describe the linguistic unit bunsetsu and MASS, both of which form the base of JASS pre-training. Then we provide the background information of head-driven phrase structure grammar and head finalization by which we propose ENSS pre-training.

3.1.1 MASS

MASS is a pre-training method for NMT proposed by Song et al. [46]. As shown in Figure 1, in MASS pre-training, the input is a sequence of tokens where a part of the sequence is masked and the output is a sequence where the masking is inverted.

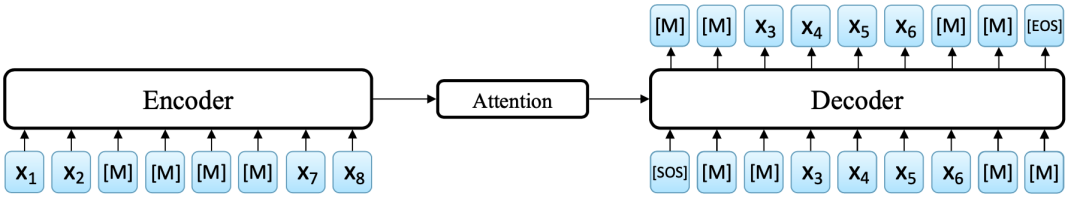

Figure 1: **Sequence to Sequence structure for MASS.** x_i represents a token and x_3 to x_6 are consecutive tokens to be masked/predicted.

Formally, consider $x \in \mathcal{X}$ which is a sequence of tokens where $\mathcal{X}$ is a monolingual corpus. Consider $C = [p_i, p_j]$ where $0 < p_i \leq p_j \leq len(x)$ and $len(x)$ is the number of tokens in sentence x. We denote by x^C the masked sequence where tokens in positions from p_i to p_j in x are replaced by a mask token $[M]$. $x^{!C}$ is the sequence with inverted mask, i.e. where tokens in positions other than the aforementioned fragments are replaced by the mask token $[M]$. In MASS, the pre-training objective is to predict the masked fragments in x using an encoder-decoder model where x^C is the input to

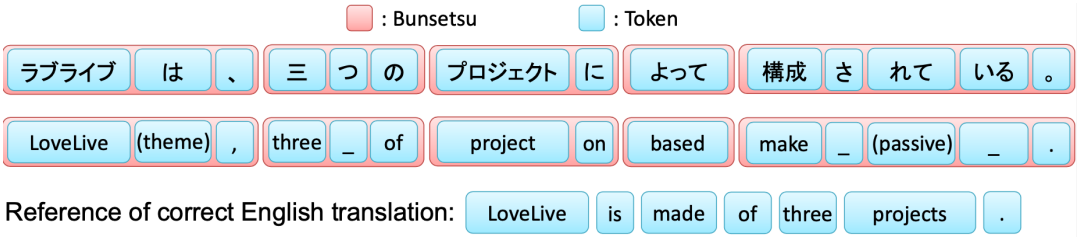

Figure 2: Word and bunsetsu segmentations for a Japanese sentence with meaning 'LoveLive is made of three projects." In the word-by-word English translations, "_" represents words with no meaningful translations.

the encoder and $x^{!C}$ is the target output for the decoder. The log likelihood objective function is:

$$\mathcal{L}_{mass}(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log P\left(x^{!C} | x^{C}; \theta\right) \tag{1}$$

where θ is set of model parameters. The number of tokens to be masked is a hyper-parameter of MASS. The NMT model is pre-trained with the MASS task jointly for both source and target languages.

3.1.2 Bunsetsu

Bunsetsu is the syntactic component of Japanese sentences [20, 27]. It is roughly equivalent to the concepts of noun phrases or verb phrases in English syntax and constitutes a minimal unit of meaning. The concept of "word" is ambiguous for writing systems like Japanese where word-separators are not applicable, and Japanese segmenters [20, 27] can segment a Japanese sentence either in words or in bunsetsus. As such, bunsetsu is also more likely to correspond to a well-defined entity or concept than words. The difference between word-level and bunsetsu-level segmentation is illustrated in Figure 2. Note that each bunsetsu contains self-contained information and case markers, which indicates its relation with other bunsetsus.

3.1.3 Head-driven Phrase Structure Grammar

As opposed to a dependency based grammar, Head-driven Phrase Structure Grammar (HPSG) [34, 35] is a lexicalism based grammar that focuses on generalizing phrase structures. HPSG primarily handles word and phrase signs within a sentence in terms of syntactic and semantic role of themselves.

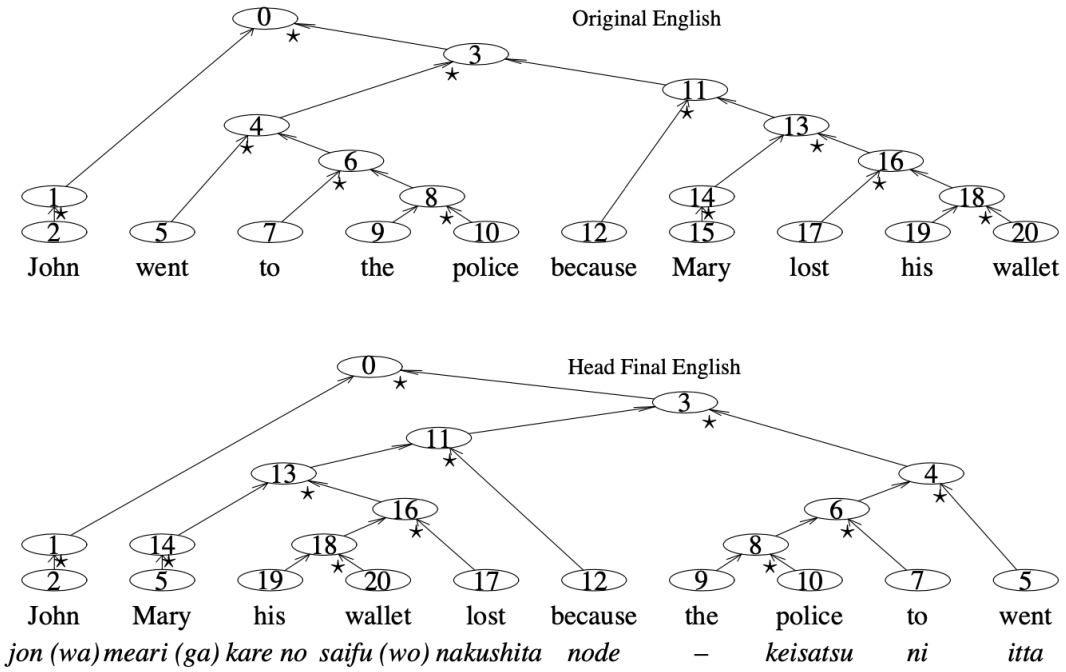

Figure 3: An example of HPSG parsing result and head finalization. Head finalization reorders an English sentence to Japanese-like sentence. [14]

Thus, HPSG should be an appropriate parsing rule for extracting phrase structures in sentences and applying the following proposed pre-training techniques. Figure 3 above shows an instance of parsing an English sentence by HPSG grammar.

3.1.4 Head Finalization

By HPSG mentioned above, sentences in any language can be characterized by phrase structures. By the definition of phrase, the “head" of a phrase is then defined to be the syntactically determinant part within a phrase. In other words, “head" determines the syntactic category of the phrase and its “dependents." Specifically, English is called “head-initial" language because “head" appears before its “dependents" while Japanese is named by "head-final" language for that “head" usually follows “dependents" within a phrase.

The deliberate phrase structures given by HPSG parser are utilized in several scenes in NLP. Particularly, Isozaki et al. [14] proposed a simple re-ordering rule for SVO language (head-initial languages) by using the phrase

structure information given by HPSG parser. Figure 3 above shows an example of reordering an English sentence to be SOV-like sentence on the basis of the result of HPSG parsing. Via reordering sentences in SVO languages like English to be SOV-like sentences, performances of Statistical Machine Translation (SMT) have been improved. Specifically, Isozaki et al. [14] first proposed head finalization and applied it into English-to-Japanese SMT; Han et al. [10] applied it into Chinese-to-Japanese SMT and obtained significant improvements; more recently, Zhou et al. [57] utilized this reordering technique to generate synthetic parallel sentences in the back-translation phase on the translation between SOV and SVO languages. This time we utilize this reordering rule in the pre-training phase for NMT (see Section 3.3.2).

3.2 Proposed Methods for Japanese

Our methods are built on the ideas of the original MASS and are improved by jointly learning multiple linguistics-aware tasks. For Japanese, we propose BMASS (Bunsetsu-based MAsked Sequence to Sequence pre-training) and BRSS (Bunsetsu Reordering based Sequence to Sequence pre-training). Their combination, JASS (Japanese-specific Sequence to Sequence pre-training), is introduced in the following section.

3.2.1 BMASS

We propose BMASS, which leverages syntactic parses of Japanese monolingual data for sequence to sequence pre-training. While MASS pre-trains a NMT model by making it predict random parts of a sentence given their context, BMASS involves making the model predict a set of bunsetsus given the contextual bunsetsus. We expect this will let the model learn about bunsetsus and thereby focus on predicting meaningful subsequences instead of random albeit fluent ones.

To perform BMASS, we modify the definition of mask C in Equation 1: $C = [[p_{i_1}, p_{j_1}], [p_{i_2}, p_{j_2}], ...[p_{i_n}, p_{j_n}]]$, where $0 < p_{i_1} \leq p_{j_1} \leq p_{i_2} \leq p_{j_2} \leq ...p_{i_n} \leq p_{j_n} \leq len(x)$ and $len(x)$ is the number of tokens in sentence x. Subsequently, the $k-th$ position span p_{i_k} to p_{j_k} correspond to the start and end of the specific bunsetsu within a Japanese sentence. Consequently we denote the

Figure 4: An example of source and target for MASS, BMASS, BRSS with the meaning “LoveLive is made of three projects."

BMASS loss as $\mathcal{L}_{bmass}$. The main difference between MASS and BMASS is that in MASS we mask random token spans whereas in BMASS we only mask tokens spans that are complete bunsetsus. The number of bunsetsus to be masked constitutes a hyper-parameter for BMASS. Figures 4-b and 4-c give training pair examples for MASS and BMASS.

Note that our BMASS pre-training task differs from the entity masking task of ERNIE [47] and random span masking of SpanBERT [16]. ERNIE and SpanBERT are proposed without using syntactic units and are for language understanding downstream tasks.

3.2.2 BRSS

Japanese sentences are typically in a SOV word order which can be reordered to SVO in order to reduce the difficulty of translation to languages with SVO order. We first define here a simple process for reordering a (typically SOV) Japanese sentence into a “SVO Japanese" pseudo-sentence which will be used in BRSS. There exist several previous works about reordering a SOV-ordered sentence to a SVO-ordered sentence [12, 19]. In our case, in order to leverage bunsetsu units in Japanese consistently with BMASS, we propose Bunsetsu-based Reordering, which is able to generate a SVO-ordered Japanese sentence while retaining syntactic information at the bunsetsu-level. We first define

“chunking signal words" as any punctuation mark or the topic marker “は". Our re-odering process is as follows:

1. split the sentence into bunsetsus
2. select sequences of bunsetsus bounded by chunking signal words
3. simply reverse the order of the bunsetsus in these sequences without using rules

We can now propose BRSS which involves a Japanese sentence and its reordered version obtained using the aforementioned procedure. Refer to Figure 4-d for an example of a bunsetsu reordered sentence. The pre-training objective here is a reordering task. We expect that this will let the system learn the structure of Japanese language, as well as prepare it for the reordering operation it will have to perform when translating to a language with different grammar. We have two choices where we can make the NMT system predict the original sentence given the reordered sentence (BRSS.F) or vice-versa (BRSS.R). We will experiment with both options.

3.3 Proposed Methods for English

Similar to the proposed methods for Japanese, we propose two linguistically-driven methods for English which are respectively based on the masked sequence-to-sequence language model and reordering sequence-to-sequence language model. One is PMASS (Phrase-based MAsked Sequence-to-Sequence pre-training) and the other is HFSS (Head Finalization based Sequence-to-Sequence pre-training). The combination of PMASS and HFSS, ENSS (ENglish-specific Sequence to Sequence pre-training), is introduced in the next section.

3.3.1 PMASS

We propose PMASS by leveraging phrase span information within an English sentence. Generally speaking, we construct PMASS pre-training by limiting the masked tokens in MASS to be an entire phrase span or phrase spans. Thus, for masking plural phrase spans, we denote it as PMASS.P; for masking just a single phrase span, we name it PMASS.S. Specifically, the source and target for PMASS.P and PMASS.S pre-training can be generated by our proposed

Algorithm 1: The algorithm of determining masked phrase spans for PMASS.P.

Input: Length of the sentence L, tree of HPSG parsing result for the sentence T.

Output: M. (tokens to be masked)

```
Function Pmass(N, L, l, M):
    if tag of N is sentence then
        return Pmass(child of N, L, l, M)
    else if tag of N is tok then
        if int(L/2) − l > 0 then
            M.append(token on N)
        return M
    else if N only has one child and N.tag is cons then
        return Pmass(child of N, L, l, M)
    else
        ll ← number of tokens on the left child of N;
        lr ← number of tokens on the right child of N;
        if ll is 1 and lr is 1 then
            if int(L/2) − l > 1 then
                M.append(token on N)
            return M
        else if int(L/2) <= l then
            return M
        else if ll <= int(L/2) − l and lr > int(L/2) − l then
            if random p < 0.5 then
                M ← M+ tokens on the left child of N;
                l ← l + ll;
                return Pmass(right child of N, L, l, M)
            else
                return Pmass(right child of N, L, l, M)
            end
```

```
27
28
29      else if lr <= int(L/2) − l and ll > int(L/2) − l then
30          if random p < 0.5 then
31              M ← M+ tokens on the right child of N;
32              l ← l + lr;
33              return Pmass(left child of N, L, l, M)
34          else
35              return Pmass(left child of N, L, l, M)
36          end
37      else if ll > int(L/2) − l and lr > int(L/2) − l then
38          if random p < 0.5 then
39              return Pmass(left child of N, L, l, M)
40          else
41              return Pmass(right child of N, L, l, M)
42          end
43      else
44          M ← M+ tokens on the left child of N;
45          l ← l + ll;
46          return Pmass(right child of N, L, l, M)
47      end
48  Initialize Current Node N by ROOT of T, Empty Token List M;
49  l ← 0 ;
50  Pmass(N, L, l, M)
```

Algorithm 2: The algorithm of determining masked phrase spans for PMASS.S.

Input: Length of the sentence L, tree of HPSG parsing result for the sentence T.

Output: Token List M consisting of all the tokens on N. (to be masked)

```
Initialize Current Node N by ROOT of T;
while number of tokens on N > int(L/2) do
    if number of tokens on left child of N > number of tokens on right child
     of N then
        N ← left child of N;
    else
        N ← right child of N;
    end
end
```

phrase-masking algorithms described in Algorithm 1 and 2. Inspired by MASS, we force the number of the masked tokens to approximate the half of the length of the sentence to guarantee the effectiveness of sequence-to-sequence masked language model. Examples of PMASS.P and PMASS.S are given in Figure 5-c. We can observe that several phrase spans in PMASS.P and a single long phrase span in PMASS.S are masked. We expect such special masking patterns can force the NMT system to extract more phrase-level syntactic information in the pre-training phase.

3.3.2 HFSS

We propose HFSS by head finalization technique [14] for pre-training English. As shown by Figure 5-d, the pre-training task is also a reordering task, which simulates the translation from SOV languages to English. More precisely, the source sentence for sequence-to-sequence pre-training is the reordered (SOV-like or head-finalized) English sentence and the target sentence is the original English monolingual sentence. We expect HFSS can help the system learn the word reordering pattern of the translation between SVO (head-initial) and

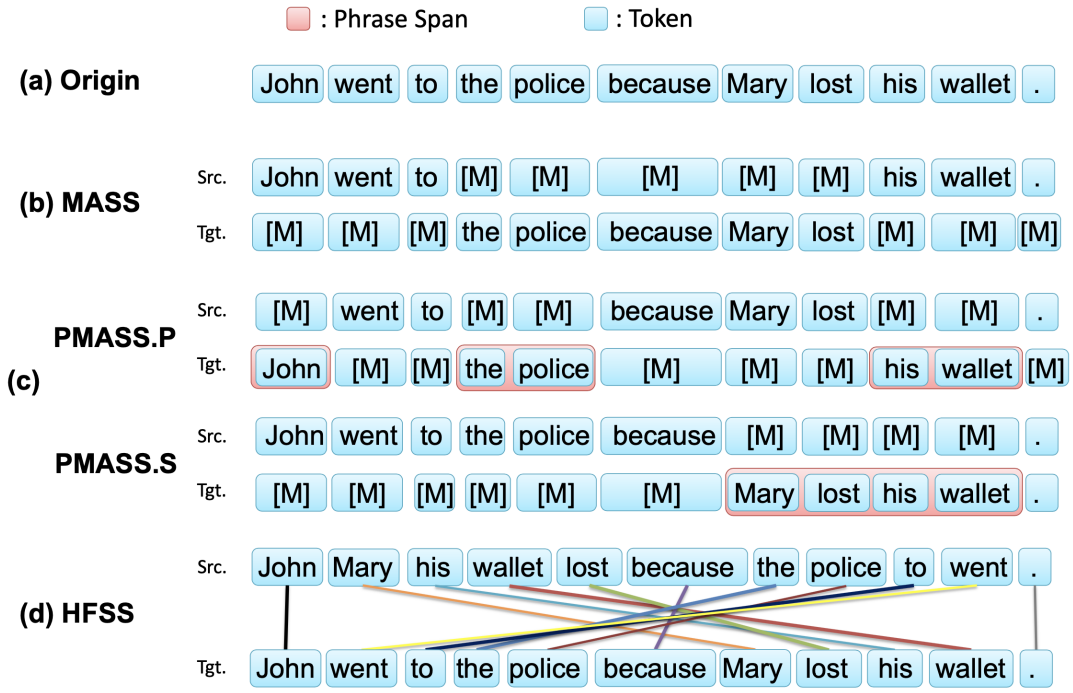

Figure 5: An example of source and target for MASS, PMASS, HFSS of a sentence in English."

SOV (head-final) languages in advance.

Based on the prior experiments for Japanese (see 5.1), we here do not distinguish HFSS with HFSS.F and HFSS.R (HFSS.F performs pre-training with SOV-SVO pattern while HFSS.R performs the resverse pattern). Instead, we directly define HFSS by the pre-training pattern of HFSS.F. Moreover, here HFSS is built on the base of head finalization, which utilizes the results from HPSG parsers. It is consistent with PMASS in which we extract phrases by HPSG parsing results.

We build up our proposal on English by head finalization while for SOV languages like Japanese, it is unmanageable to reorder SOV sentences to SVO-like ones [14]. Furthermore, HFSS can be used for all the head-initial languages besides English as well-developed reordering rules have been proposed and demonstrated effective for NMT. However, BRSS can only be implemented for Japanese-involved translation pairs because bunsetsu information is required to establish the source and target sentence for the sequence-to-

sequence pre-training.

3.4 Multi-task Pre-training

Multi-task pre-training objectives lead to robuster initial state for NMT system [21, 39]. Because our proposed methods can also be categorized into two groups of the pre-training tasks, here we give the proposal of the multi-task pre-training tasks for both Japanese and English.

We define JASS (JApanese-specific Sequence to Sequence) pre-training which is a combination of the previous two procedures: BMASS and BRSS. Our actual pre-training will consist of joint execution of these two pre-training. The pre-training objective for JASS is therefore:

$$\mathcal{L}_{jass}(\mathcal{X}_{ja}) = \mathcal{L}_{bmass}(\mathcal{X}_{ja}) + \mathcal{L}_{brss}(\mathcal{X}_{ja}) \tag{2}$$

where $\mathcal{X}_{ja}$ represents the monolingual corpus of Japanese and $\mathcal{L}_{brss}$ is the reordering loss using the forward or the reverse variant mentioned in Section 3.2.2. We expect BMASS & BRSS to learn syntactic knowledge jointly and BRSS to learn word ordering knowledge.

For English, we similarly define ENSS (ENglish-specific Sequence to Sequence) pre-training, which combines PMASS and HFSS. More precisely, the training objective is:

$$\mathcal{L}_{enss}(\mathcal{X}_{en}) = \mathcal{L}_{pmass}(\mathcal{X}_{en}) + \mathcal{L}_{hfss}(\mathcal{X}_{en}) \tag{3}$$

where $\mathcal{X}_{en}$ denotes the monolingual corpus of English, $\mathcal{L}_{pmass}$ indicates the PMASS.P or PMASS.S loss, and $\mathcal{L}_{hfss}$ indicates the reordering loss of HFSS.

JASS is specifically designed for Japanese while theoretically ENSS can be transplanted onto any SVO language as long as we can extract the phrase structure information of the corresponding language from a HPSG parser.

We also mix JASS pre-training for Japanese with MASS pre-training for the other language involved in the translation. In practice, we therefore designate by JASS the pre-training of the NMT system that uses Japanese monolingual data with BMASS and BRSS objectives, and “other language" monolingual data with MASS objective. Likewise, for English, ENSS pre-training consists

of PMASS & HFSS for English and MASS for “other language" involved in the fine-tuning translation pair.

We also consider attempting the combination of our proposed linguistically-driven methods with strong baseline pre-training objective, MASS, which we call MASS + JASS (or ENSS) in the following sections. To let the pre-training model know about which language and sub-task (MASS, BMASS, BRSS, PMASS, HFSS) it should perform, we prepend tags to inputs similar to the ones used in Johnson et al. [15] (see section 4.3 for details).

4 Experimental Settings

In this section, we evaluate our pre-training methods on simulated low-resource scenarios for ASPEC Japanese–English [31], Japanese–Chinese translation [29] and on realistic low-resource scenarios for Wikipedia Japanese–Chinese [3, 4], News Commentary Japanese–Russian translation [13].[1)]

4.1 Pre-training and Fine-tuning for NMT

We first introduce the pre-training and fine-tuning pipeline for NMT. As shown in Figure 6 below, we first utilize monolingual corpora to pre-train the initialized sequence-to-sequence model. Subsequently, we use the parallel corpus of interested languages to fine-tune the pre-trained models. The fine-tuned model will be the final NMT model. All the experiments in this thesis will be conducted on the basis of this pre-training and fine-tuning pipeline for NMT.

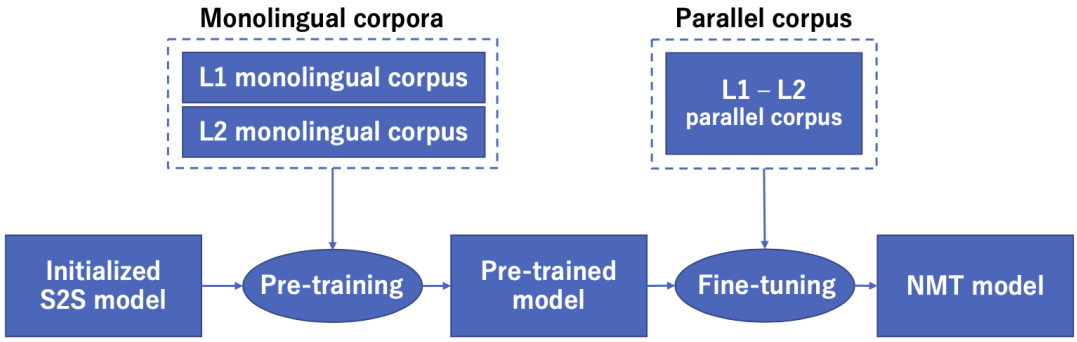

Figure 6: Pre-training and fine-tuning for NMT. “L1" and “L2" means two languages involved in the translation task. “S2S" denotes sequence-to-sequence.

4.2 Datasets

We use the monolingual data for pre-training and the parallel data for fine-tuning. Refer to Table 1 for an overview.

[1)] Japanese and Russian are individually resource-rich languages but Ja-Ru can be regarded as a low-resource language pair because of the limited amount of the parallel data.

	Language	Dataset	Size
Mono	Ja	Common Crawl	22M
	Zh	Common Crawl	22M
	En	Common Crawl	22M
	En	News Crawl	22M
	Ru	News Crawl	22M
Parallel	Ja-En	ASPEC-JE	1M
	Ja-Zh	ASPEC-JC	670k
	Ja-Zh	Wikipedia	258k
	Ja-Ru	JaRuNC	12k
	Ja-En	JaRuNC	42k
	Ru-En	JaRuNC	84k

Table 1: Overview of data. “Size" denotes the number of the monolingual sentences or parallel sentences.

Monolingual data: For pre-training, we use monolingual data of 22M lines each for Japanese, English, Russian, and Chinese, randomly sub-sampled from Common Crawl and News crawl[1)] mentioned in the official WMT monolingual training data.[2)] For pre-training on Japanese–English and Japanese–Russian, given that these three languages have different scripts and thus have few common words, the pre-training objectives for each language will relatively work separately even though they are performed jointly for two languages. However, for the pre-training on Japanese and Chinese, they share more characters, which indicates that the monolingual pre-training tasks will be run in a pseudo cross-lingual manner. Thus, we also expect to see whether such pre-training will benefit the fine-tuning more.

Parallel Data: We use scientific abstracts domain ASPEC parallel corpus

[1)] The pre-training will be much more effective if the domains of the pre-training and fine-tuning dataset overlap more. [39]

[2)] `http://www.statmt.org/wmt19/translation-task.html`

for training Japanese–English and Japanese–Chinese models and the news commentary domain JaRuNC parallel corpus for training Japanese–Russian models. For Japanese–Chinese fine-tuning, we also try the Wikipedia dataset which is a real low-resource scenario. For ASPEC, We use the official train, development and test splits provided by WAT 2019.[1)2)] For Wikipedia, we use the dataset released by Kyoto University.[3)]

4.3 Pre-processing

We tokenize the monolingual data by using Moses tokenizer for English and Russian[4)], Jumanpp for Japanese[5)], and jieba for Chinese[6)]. We get the bunsetsu information by using KNP[7)] and obtain the HPSG parsing results by using enju[8)]. Sentences over 175 tokens are removed. For each language pair, we built a joint vocabulary with 60,000 sub-word units via byte-pair encoding (BPE) [43] on the concatenated monolingual corpora involved during pre-training.[9)] As we do multi-task pre-training, each sentence is prepended with a task token $[MASS]$, $[BMASS]$, $[BRSS]$, $[PMASS]$, or $[HFSS]$ and a language token $[Ja]$, $[En]$, $[Ru]$, or $[Zh]$. This ensures that the model learns to distinguish between different pre-training objectives and languages. This token can be used when monolingual pre-training is conducted jointly by multiple languages and multiple tasks.

---

1) http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/index.html\#task.html

2) For ASPEC Japanese–English, we use the first 1M parallel sentences. Parallel sentences for different fine-tuning size settings are randomly sampled from selected 1M dataset.

3) http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?Wikipedia\%20Chinese-Japanese\%20Parallel\%20Corpus

4) https://github.com/moses-smt/mosesdecoder

5) https://github.com/ku-nlp/jumanpp

6) https://github.com/fxsjy/jieba

7) https://github.com/ku-nlp/pyknp

8) https://mynlp.is.s.u-tokyo.ac.jp/enju/

9) Specifically, 30,000 BPE merge operations will lead to a joint vocabulary with the size around 60,000 for Japanese–Chinese, while 40,000 BPE merge operations is set for other language pairs.

4.4 Training and Evaluation Details

In our experiments we use the open source OpenNMT [17] implementation of the Transformer [49] NMT model.[1)] The hyperparameters are set to the Transformer-big setting in OpenNMT. In particular our model has a 6-layer encoder and decoder, a hidden size of 1024, a feed-forward hidden layer size of 4096, batch-size of 2048, dropout rate of 0.3 and 16 attention heads. An ADAM optimizer with a learning-rate of 10^{-4} is used both for pre-training and fine-tuning. All the pre-training tasks are run till convergence on 4 TITAN V100 GPU cards and fine-tuning uses only 1 GPU. It takes around 2 days for each pre-training run. Mixed precision training [26] is used for both pre-training and fine-tuning. For multi-task pre-training, data is randomly shuffled so that even in each mini-batch, different pre-training objectives will appear, corresponding to a real joint pre-training. Pre-training tasks are evaluated using perplexity and the checkpoint with the lowest pre-training perplexity is selected for fine-tuning. We use BLEU [32] for automatic evaluation, and adequacy and fluency for human evaluation. We do early stopping using 1-gram accuracy and perplexity on development-set. We evaluate the statistical significance of our BLEU scores by bootstrap resampling [18].

4.5 Baselines

Besides MASS, we also define two pre-training baselines here to compare with our proposed methods. They are named as MultiMASS (Multi-span based MAsked Sequence to Sequence) and Deshuffling. Moreover, the joint training with MASS and Deshuffling is set as the multi-task pre-training baseline. All the baselines are as following:

MASS. Using the same settings as in Song et al. [46].

MultiMASS. MultiMASS is a baseline method added to help demonstrate the effectiveness of masking specific syntactic units like bunsetsu or phrase spans within a sentence which we propose as BMASS and PMASS.

As shown in Figure 7, MultiMASS predicts several randomly masked

[1)] `https://github.com/OpenNMT/OpenNMT-py`

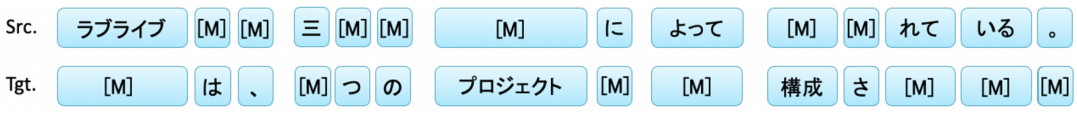

Figure 7: An example of source and target for MultiMASS with the meaning "LoveLive is made of three projects."

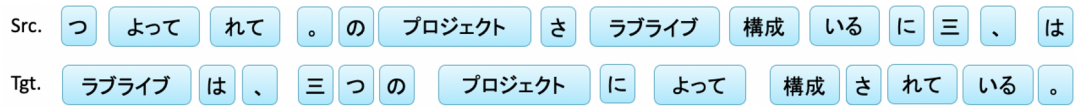

Figure 8: An example of source and target for Deshuffling with the meaning "LoveLive is made of three projects."

token spans within a sentence, which differs from the single masked span in MASS, masked busetsu spans in BMASS, several phrase spans in PMASS.P and the single phrase span in PMASS.S.

Deshuffling. Deshuffling denotes the pre-training task of random shuffling based sentence reconstruction, which is also one of the crucial pre-training tasks in BART [21]. We perform this pre-training task as another baseline to confirm the effectiveness of reordering syntactic units in BRSS and the reordering driven by head finalization of HFSS. A pre-training example is shown in Figure 8.

Multi-task Baseline. Multi-task baseline is the combination of respective best baseline methods from masked language model and reordering pre-training. Thus, here multi-task baseline consists of MASS[1)] and Deshuffling. It is formulated by:

$$\mathcal{L}(\mathcal{X}) = \mathcal{L}_{mass}(\mathcal{X}) + \mathcal{L}_{deshuffling}(\mathcal{X}) \tag{4}$$

where $\mathcal{X}$ represents the monolingual corpora.

4.6 Pre-trained Models

We pre-train our NMT models by leveraging the monolingual data of the source and target languages. For Japanese we can use MASS, BMASS, or

[1)] MASS outperforms MultiMASS, so we use MASS rather than MultiMASS. (See 5.1)

#	Pre-trained Model	Details
Main baseline		
1	MASS	Using the same settings as in Song et al. (2019).
Proposed methods for Japanese		
2	BMASS	Similar to MASS, we mask half of the number of the bunsetsus during pre-training.
3	BRSS	We separately pre-train on SVO–SOV (BRSS.F) as well as SOV–SVO (BRSS.R) models.
4	JASS	Multi-task training of BMASS and BRSS.
Combinations of proposed methods with MASS		
5	MASS+BMASS	Multi-task training of MASS and BMASS.
6	MASS+BRSS	Multi-task training of MASS and BRSS.
7	MASS+BMASS+BRSS	Multi-task training of BMASS, BRSS and MASS.
Other baselines for Japanese		
8	MultiMASS (Ja)	Based on MASS, several random token spans are masked rather than one consecutive span.
9	Deshuffling (Ja)	Random shuffling based original sentence reconstruction.
10	MASS+Deshuffling (Ja)	Multi-task pre-training baseline for Japanese.
Proposed methods for English		
11	PMASS	Similar to MASS, we mask an entire phrase span based on head-driven phrase structure grammar. We respectively perform the experiments for PMASS.P and PMASS.S.
12	HFSS	We train SOV(head finalized)–SVO(original) models for English.
13	ENSS	Multi-task training of MASS and HFSS.
Other baselines for English		
14	MultiMASS (En)	Based on MASS, several random token spans are masked rather than one consecutive span.
15	Deshuffling (En)	Random shuffling based original sentence reconstruction.
16	MASS+Deshuffling (En)	Multi-task pre-training baseline for English.
Combination of proposed methods for English and Japanese		
17	JASS+ENSS	Multi-task training of JASS and ENSS
Baseline for #17		
18	MASS+Deshuffling	Multi-task pre-training baseline for JASS+ENSS

Table 2: Settings of pre-trained models.

BRSS, while for English we can use MASS, PMASS, or HFSS. For Russian and Chinese, we only use MASS. Following Imankulova et al. [13], we experimented with multilingual models for Japanese–Russian translation. For this purpose, we pre-train for all three languages with MASS and/or JASS as applicable to each language.[1)] In particular we pre-train different types of models in Table 2. Note that we use MASS for ENSS because PMASS underperforms MASS by a significant margin (see 5.1).

4.7 Fine-tuned NMT Models

We fine-tune to improve Japanese-English, English-Japanese, Japanese-Chinese, Chinese-Japanese, Japanese-Russian, Russian-Japanese translation. We train the following NMT models:

1. **Ja–En and En–Ja:** Japanese to English and English to Japanese models using from 3k to 50k parallel sentences randomly sampled from ASPEC for fine-tuning.
2. **Ja–Zh and Zh–Ja:** Japanese to Chinese and Chinese to Japanese models using from 3k to 50k parallel sentences randomly sampled from ASPEC and Wikipedia respectively for fine-tuning.
3. **Ja–Ru and Ru–Ja:** Japanese to Russian and Russian to Japanese models. We fine-tune on the Japanese–Russian data for unidirectional models (labelled as “UNI"). Following Imankulova et al. [13], we also trained multilingual models (labelled as “M2M") by fine-tuning on the combination of all news commentary data in Table 1.

We compare these models with baselines which are supervised NMT models on the same data settings but without pre-training. In addition, fine-tuning results under the high-resource scenarios (with over 50k parallel sentences) are given and discussed in Appendix A.1.

[1)] ENSS will be combined in this multilingual pre-training in future work.

5 Results and Analyses

Tables 3, 4, 5, 6, and 7 contain the NMT BLEU results of our proposed methods for Japanese–English, Japanese–Chinese, and Japanese–Russian translation on a variety of translation domains respectively. Afterwards, we will provide in-deep analysis for translation quality in terms of adequacy by using LASER [1], human evaluation scores, specific cases for the real low-resource scenario of Wikipedia Ja-Zh. Last but not the least, we conduct an investigation on the pre-training accuracy to analyze how different the pre-trained models are and how they complement with each other.

5.1 NMT Results

In Tables 3, 4, 7, where we simulate several low resource settings for Japanese–English and Japanese–Chinese translation on ASPEC with different pre-training datasets, and in Table 5 and 6, where we use a realistic low-resource setting for Wikipedia Japanese–Chinese and JaRuNC Japanese–Russian translation, we can observe that all settings using pre-training outperform those without pre-training (#0), indicating the importance of pre-training. The results also show that JASS (#4) and ENSS (#13) are generally better than MASS (#1).

Specifically, for Japanese–English translation, BMASS (#2) is comparable to MASS; BRSS (#3 & #3(R)) and their combination, JASS (#5), are significantly better than MASS. However, as shown in Table 4 and 5, results on two parallel corpora on different domains for Japanese–Chinese show much more significantly better results by using our proposed BMASS and BRSS. We see that only few settings on Japanese-to-Chinese of BRSS yield lower BLEU results than MASS, while other settings by using proposed methods give better results than MASS by significant margins. Although, MASS is better than BMASS for Japanese–English translation, the reverse can be observed for Japanese–Chinese and also on Japanese–Russian translation, especially when multilingual data (M2M) is used for fine-tuning. This indicates that the effects of the proposed linguistically-driven techniques might correlate to

#	Model	Ja-En				En-Ja			
		3k	10k	20k	50k	3k	10k	20k	50k
Main baselines									
0	w/o pre-training	0.8	2.1	3.5	16.1	1.1	2.7	5.1	19.4
1	MASS	8.8	13.8	17.2	21.2	9.1	16.0	20.6	25.0
Proposed methods for Japanese									
2	BMASS	8.9	13.9	17.4	21.8	8.7	15.9	20.1	25.4
3	BRSS	8.8	14.9†	18.1†	22.0†	10.0†	17.3†	21.0	26.0†
3 (R)	BRSS.R	8.2	14.3†	17.7†	21.7†	10.0†	17.2†	20.5	25.7†
4	JASS	10.6†	15.7†	**18.9**†	**22.3**†	**11.5**†	17.7†	21.6†	26.5†
Combinations of proposed methods with MASS									
5	1 + 2	9.2	14.8†	17.7†	21.7†	9.7†	16.6†	20.9	25.9†
6	1 + 3	**10.9**†	15.9†	18.3†	22.2†	11.0†	17.7†	**21.7**†	**26.8**†
7	1 + 4	10.5†	15.5†	18.5†	22.0†	**11.5**†	**17.9**†	**21.7**†	26.4†
Other Baselines for Japanese									
8	MultiMASS (Ja)	7.1	12.1	15.1	20.5	6.9	13.0	17.7	24.1
9	Deshuffling (Ja)	6.8	12.7	16.6	21.0	7.8	14.7	19.3	24.9
10	1 + 9	8.2	13.3	17.0	21.4	8.3	15.5	19.5	25.4
Proposed methods for English									
11	PMASS.P	6.8	12.1	15.9	20.7	5.5	13.5	17.8	24.5
11*	PMASS.S	6.5	12.3	16.2	21.2	6.2	13.5	18.2	24.6
12	HFSS	10.5†	**16.3**†	**18.9**†	**22.6**†	9.8†	17.8†	**21.7**†	**26.8**†
13	ENSS	**11.2**†	**16.7**†	**19.0**†	22.1†	**11.7**†	**18.7**†	**22.5**†	**27.0**†
Other baselines for English									
14	MultiMASS (En)	6.9	12.0	15.2	20.1	7.0	12.8	17.5	23.8
15	Deshuffling (En)	6.6	12.5	15.9	20.9	6.8	14.1	19.2	24.7
16	1 + 15	7.7	13.2	16.7	21.0	8.6	15.7	20.4	25.6
Combination of methods for Japanese and English									
17	4 + 13	**10.9**†	**16.4**†	18.7†	**22.3**†	**11.9**†	**18.4**†	**22.0**†	26.5†
18	10 + 16 (baseline)	7.2	12.6	16.4	20.9	8.4	14.8	19.1	25.5

Table 3: BLEU scores for simulated low/high-resource settings for Japanese–English ASPEC translation using 3k to 50k parallel sentences for fine-tuning. Pre-trained models used for fine-tuning are numbered as per their description in section 4.6. Results better than MASS with statistical significance $p < 0.05$ are marked with †. Bold denotes the top-three scores.

#	Model	Ja-Zh				Zh-Ja			
		3k	10k	20k	50k	3k	10k	20k	50k
Main baselines									
0	w/o pre-training	0.7	3.4	11.5	21.0	1.9	4.5	16.0	28.2
1	MASS	15.7	20.3	22.4	24.7	19.4	25.9	29.4	32.9
Proposed methods									
2	BMASS	16.7†	21.1†	23.0†	25.3†	20.9†	27.2†	30.2†	33.7†
3	BRSS	15.6	21.1†	22.6	24.9	20.7†	26.8†	30.0†	33.3†
4	JASS	**17.1**†	**22.2**†	**23.2**†	25.2†	21.6†	27.5†	**30.4**†	**33.6**†
Combinations of proposed methods with MASS									
7	1 + 4	17.0†	21.7†	23.1†	**25.4**†	**21.8**†	**27.6**†	30.2†	33.4†
Other baselines									
8	MultiMASS	14.5	20.5	22.3	24.7	19.6	25.7	29.8	33.2
9	Deshuffling	14.1	19.5	21.6	24.3	18.4	25.0	28.7	32.8
10	1 + 9	15.0	20.2	22.1	25.0	18.9	25.9	29.3	33.1

Table 4: BLEU scores for simulated low-resource settings for Japanese–Chinese ASPEC translation using 3k to 50k parallel sentences for fine-tuning. Results better than MASS with statistical significance $p < 0.05$ are marked with †.

specific translation directions and domains.

As shown in Table 3, our proposed methods of leveraging linguistic knowledge for English obtain significant higher BLEU results when we perform the reordering pre-training task, HFSS (#12). However, the proposed linguistically-driven masked language model PMASS.P (#11) and PMASS.S (#11*) just yield comparable results to several other baseline methods like MultiMASS (#14) and Deshuffling (#15). This demonstrates that syntactical span based masked language model may merely work on head-final languages like Japanese. Considering the weak performance of PMASS, we combine HFSS with MASS for ENSS. The multi-task pre-trained ENSS gives highest results

#	Model	Ja-Zh				Zh-Ja			
		3k	10k	20k	50k	3k	10k	20k	50k
Main baselines									
0	w/o pre-training	0.9	2.9	2.9	6.0	1.6	2.9	3.9	6.5
1	MASS	7.7	15.4	18.3	23.4	9.6	17.6	23.3	27.1
Proposed methods									
2	BMASS	10.8†	15.7	**20.1**†	24.5†	16.2†	19.4†	25.4†	**30.0**†
3	BRSS	11.6†	16.2†	20.0†	24.6†	15.7†	21.6†	25.0†	28.3†
4	JASS	**12.0**†	**17.0**†	**20.1**†	**25.0**†	**16.6**†	21.2†	**26.5**†	29.2†
Combinations of proposed methods with MASS									
7	1 + 4	11.8†	16.8†	**20.1**†	24.6†	**16.6**†	**22.3**†	25.5†	29.6†
Other baselines									
8	MultiMASS	8.2	13.8	18.6	21.5	10.7	17.3	22.0	26.4
9	Deshuffling	9.3	14.2	18.7	22.7	12.4	18.4	23.2	27.4
10	1 + 9	8.7	13.8	19.4	23.2	14.3	18.8	24.8	27.8

Table 5: BLEU scores for simulated low-resource settings for Japanese–Chinese Wikipedia translation using 3k to 50k parallel sentences for fine-tuning. Results better than MASS with statistical significance $p < 0.05$ are marked with †.

almost on all the low-resource settings.

However, in Table 3, when performing a universal linguistically-driven pre-training simultaneously for Japanese and English (#17), we did not obtain further significant BLEU improvements. This can be attributed to that NMT may depend more on specific linguistic information on single language side, and the joint pre-training does not allow the linguistic knowledge transfer across languages between dissimilar languages.

Besides the main baseline MASS, we also conduct several other sequence-to-sequence pre-training baselines: MultiMASS (#8 & #14) and Deshuffling (#9 & #15) and their multi-task combinations (#10, #16 and #18) respectively

#	Model	UNI		M2M	
		Ja-Ru	Ru-Ja	Ja-Ru	Ru-Ja
0	w/o pre-training	0.50	0.72	1.70	4.49
1	MASS	0.96	2.84	3.08	6.54
2	BMASS	0.97	2.77	**3.57**†	7.12†
3	BRSS	0.85	2.36	3.11	6.98†
4	JASS	**1.20**	3.08	3.36	**7.92**†
7	1 + 4	1.07	**3.45**†	3.02	7.06†

Table 6: BLEU scores for news commentary Japanese–Russian translation. We evaluate uni-directonal ("UNI") as well as multilingual models ("M2M"). Results better than MASS with statistical significance $p < 0.05$ are marked with †.

for Japanese and English. As shown in Table 3, 4, 5, we observe that the proposed masked style pre-training task – BMASS and reordering pre-training tasks – BRSS & HFSS outperform these baselines by significant margins, which indicates that linguistically-driven methods should be superior to self-supervised pre-training without leveraging linguistic features.

As shown by Table 3, BRSS-F (English/Russian ordering to Japanese ordering) gave slightly better results than BRSS-R (vice-versa) and thus we only experimented with BRSS-F for remaining experiments. We suppose the reason is that training the decoder with the original sentence is more important than training the encoder with it. In other words, forcing the decoder to generate a natural sentence leads to better initialized decoder for NMT. Meanwhile, HFSS is performed by the analogous manner for the same reason.

As mentioned above, JASS gives the best results when we only consider the linguistically-driven methods for Japanese. After combining the proposed methods for Japanese with MASS respectively (#5∼#7 in Table 3, 7), no significant improvements can be observed. This demonstrates that linguistic-aware methods can substitute the linguistic-agnostic ones.

#	Model	Ja-En				En-Ja			
		3k	10k	20k	50k	3k	10k	20k	50k
Main baselines									
0	w/o pre-training	1.3	2.6	9.5	17.6	1.5	3.7	11.5	21.0
1	MASS	9.9	15.2	18.7	22.3	11.0	18.1	21.6	28.0
Proposed methods									
2	BMASS	8.5	14.3	18.6	22.3	9.8	17.3	21.1	27.1
3.1	BRSS.F	8.6	14.8	18.4	22.1	10.3	17.7	21.6	27.5
3.2	BRSS.R	7.4	13.7	17.3	21.8	9.8	16.9	20.8	26.3
4	JASS	10.1	15.6†	19.1†	22.9†	11.4	18.5†	22.2†	27.3
Combinations of proposed methods with MASS									
5	1 + 2	8.7	14.2	18.2	22.2	10.0	17.4	21.5	27.2
6	1 + 3.1	**10.8**†	16.0†	19.0	22.7	12.0†	19.0†	22.3†	27.3
7	1 + 4	10.7†	**16.1**†	**19.3**†	**23.2**†	**12.6**†	**19.2**†	**23.0**†	**28.1**
Other baselines									
8	MultiMASS	5.2	10.8	14.7	20.6	6.1	12.9	17.8	25.0
9	Deshuffling	5.6	10.9	15.2	21.1	6.9	14.0	18.9	25.1

Table 7: BLEU scores for simulated low-resource settings for Japanese–English ASPEC translation using 3k to 50k parallel sentences for fine-tuning (News Crawl for English used for pre-training). Results better than MASS with statistical significance $p < 0.05$ are marked with †.

Moreover, as shown by Table 4 and 5, we observe that on ASPEC domain, JASS improves up to 2.2 BLEU scores while on Wikipedia domain, JASS gives up to 7.0 BLEU improvements. On one hand, this demonstrates the promising performance of our proposed methods. On the other hand, this indicates that the more pre-training domain overlaps with fine-tuning domain, the more improvements linguitically-driven pre-training methods will produce. Furthermore, by seeing Table 3 and 7, we can find that BLEU scores with New Crawl (Table 7) are better than those on another setting, which shows that pre-

training with high-quality monolingual dataset leads to superior fine-tuning results.

In Table 6, the improvement contributed by bilingual pre-training is limited on JaRuNC. When we use multilingual (M2M) fine-tuning, the translation quality is significantly better than unidirectional fine-tuning but the BLEU scores are all less than 10. As such, it is difficult to make serious claims about which pre-training method is the best. Our scores are smaller than the ones in Imankulova et al. [13] because unlike us, they performed bilingual pre-training as well as leveraged in-domain monolingual data for back-translation. Note that Japanese–Russian is a difficult language pair, the fine-tuning data is small and the news commentary domain is much harder than the ASPEC and Wikipedia domain. Our future efforts will be directed towards effective pre-training methods on more difficult fine-tuning domains in addition to using back-translation.

5.2 Adequacy Evaluation

Reference-free MT evaluation is evaluating the translation system without using the target reference. Such kinds of evaluation can help circumvent the noise existing in the references of translation target. After the emergence of multilingual sentence encoder [1], Yankovskaya et al. [54] proposed using multilingual sentence embeddings encoded by LASER to implement the reference-free MT evaluation. More precisely, we first apply LASER to respectively encode the source sentence and the translated sentence, then the cosine value of those two embeddings is used to evaluate the similarity between the source and the translation. This cosine value is thus the metric to evaluate the translation adequacy. There exist two advantages here. One is that target references are not required as above-mentioned. The other is that every two translation directions can be compared with each other because of using the language-agnostic embedding for evaluation.

We report the adequacies in Table 8. First, we observe that methods with pre-training can yield more semantically correct translations than those without pre-training. Second, our proposed methods can significantly obtain

#	Model	ASPEC		ASPEC		Wikipedia	
		Ja-En	En-Ja	Ja-Zh	Zh-Ja	Ja-Zh	Zh-Ja
*	Reference	80.78		86.10		87.26	
0	w/o pre-training	52.59	45.89	69.54	67.08	57.55	56.48
1	MASS	75.63	76.09	85.52	86.32	81.08	78.52
2	BMASS	75.75	76.68	85.42	86.49	80.91	81.36
3	BRSS	78.34	76.66	85.87	86.54	81.71	**84.29**
4	JASS	**80.00**	**77.63**	**85.96**	**86.58**	**85.39**	83.08

Table 8: Adequacy evaluated by LASER embedding based cosine similarity for ASPEC Japanese–English, Japanese–Chinese, and Wikipedia Japanese–Chinese translations respectively using 10k sentences for fine-tuning. Reference is the cosine similarity between the test-sets in two languages.

higher LASER similarity scores than MASS baseline, especially the results on ASPEC Japanese to English translation and both two directions of Chinese–Japanese translation on Wikipedia.

5.3 Human Evaluation

Following Nakazawa et al. [30], we performed adequacy and fluency evaluation for the Japanese–Chinese translation when 10k Wikipedia parallel sentences were used for fine-tuning the pre-trained models. We randomly sampled 100 test set English sentences and blindly evaluated their translations across various models. We scored each sentence on a scale of 1 to 5, with 1 being the worst score. The higher the score, the more adequate (meaningful) or fluent (well-formed), the sentence is. The final score we report is the average of the scores of 100 sentences. We did not look at the references but only looked at the sources for our evaluation.

In Table 9, we can observe that NMT models even if without pre-training are capable to generate rather fluent sentences, and the lack of parallel sen-

#	Model	BLEU		Adequacy		Fluency	
		Ja-Zh	Zh-Ja	Ja-Zh	Zh-Ja	Ja-Zh	Zh-Ja
0	w/o pre-training	2.9	2.9	1.22	1.05	3.90	3.99
1	MASS	15.4	17.6	2.72	2.33	4.11	4.09
2	BMASS	15.7	19.4	3.12	2.88	4.34	4.32
3	BRSS	16.2	**21.6**	3.30	3.35	4.30	**4.40**
4	JASS	**17.0**	21.2	**3.79**	**3.44**	**4.47**	4.36

Table 9: Adequacy and fluency of Wikipedia Japanese–Chinese translations using 10k sentences for fine-tuning.

tences (low-resource scenario) will mainly influence the translation adequacy (refer to the extremely low adequacy of models without pre-training). On the other hand, we can see that our proposed BMASS, BRSS, and JASS result in large improvements in adequacy and moderate improvements in fluency, for both two translation directions. The better performance of adequacy than MASS demonstrates the effectiveness of the linguistically-driven pre-training methods. Moreover, we can see that the results of human evaluation are almost consistent to those of BLEU.

5.4 Case Study

We conduct case studies on English-to-Japanese translation fine-tuned by 10k ASPEC parallel sentences and Chinese-to-Japanese translation fine-tuned by 10k Wikipedia parallel sentences to make improvements shown by BLEU score evaluations visible. As shown in Table 10 and 11, we find that the vanilla NMT system trained by 10k parallel sentences without pre-training can hardly implement the translation. With regard to models with pre-training, we observe that MASS and other baseline models generated several incorrect tokens in terms of semantics whereas the whole sentence is somehow fluent. However, our proposed methods can generate sentences with superior adequacy and fluency, where much less missing keywords are observed.

	Reference–Ja	水の性質の多様性について, まず, 水分子同士の間に働く力である水素結合と, そのネットワーク構造について解説した。
	Reference–En	To begin with, various properties of water were explained on hydrogen bond in which the force works among the water molecules and the network structure.
English to Japanese translation fine-tuned by 10k ASPEC-JE parallel sentences		
0	w/o pre-training	ここでは, これらのデータを述べ, これらの研究に, これらの研究について解説した。
1	MASS	まず, 水分子の力を中心とした水素結合について様々な特性を説明した。
2	BMASS	まず, 水分子とネットワーク構造をもつ水素結合について, 様々な性質を説明した。
3	BRSS	まず, 水分子とネットワーク構造としての力がある水素の性質を説明した。
4	JASS	まず, 水分子とネットワーク構造を融合させる力で水素結合に関する種々の性質を説明した。
8	MultiMASS(Ja)	まず, 水分子との間の水素結合について, 酸素分子とネットワーク構造を説明した。
9	Deshuffling(Ja)	まず, 水分子とネットワーク構造について, 様々な特性を説明した。
10	Multi-task baseline (Ja)	まず, 水分子の力とネットワーク構造について水素を中心に, 様々な特性を説明した。
11	PMASS	まず, 水分子とネットワーク構造の相互作用を持つ水素結合について解説した。
12	HFSS	まず, 水分子とネットワーク構造を持つ力を持つ水素結合について, 様々な性質を説明した。
13	ENSS	まず, 水中の様々な性質について, 水分子とネットワーク構造の中で力を働く水素結合について説明した。
14	MultiMASS(En)	まず, 水分子を中心に水素結合の種々の性質を説明した。
15	Deshuffling(En)	まず, 水分子とネットワーク構造に関係する水素の様々な特性を説明し, 水分子とネットワーク構造について解説した。
16	Multi-task baseline (En)	まず, 水分子における力とネットワーク構造に関係する水素結合について種々の性質を説明した。

Table 10: Japanese–English translation examples fine-tuned by 10k ASPEC parallel sentences.

	Reference–Ja	翌嘉永２年（１８４９年）１１月２２日に婚姻し、御廉中様（将軍世子の正室）と称された。
	Reference–Zh	翌嘉永２年（１８４９年）１１月２２日成婚，被称作御廉中 样（将 军世子的正室）。
Chinese to Japanese translation fine-tuned by 10k Wikipedia parallel sentences		
0	w/o pre-training	慶長２年（１５８３年)、劉宮将軍となり、幕府の命を率いた。
1	MASS	翌嘉永２年（１８４９年）５月２７日に成婚。
2	BMASS	翌嘉永２年（１８４９年）１１月２３日に結婚。
3	BRSS	翌嘉永２年（１８４９年）１１月２３日に成婚し、御廉の中（将軍世子の正室）と呼ばれる。
4	JASS	翌嘉永２年（１８４９年）１１月２２日成婚し、御廉中様（将軍世子の正室）と称した。
8	MultiMASS	翌嘉永２年（１８４９年）１月２９日に成婚した。
9	Deshuffling	翌嘉永２年（１８４９年）１１月２３日、御廉中の正室（将軍世子の正室）と称された。
10	Multi-task baseline	翌嘉永２年（１８４９年）１１月２２日成婚し、御廉の代わりに義子と呼ばれる。

Table 11: Japanese–Chinese translation examples fine-tuned by 10k Wikipedia parallel sentences.

5.5 Pre-training Accuracy

Pre-training accuracy is the accuracy of the monolingual pre-training tasks and it can be an indicator of task complexity and pre-training objective's performance. Table 12 shows the component-wise and overall pre-training

#	Model	Overall	MASS	BMASS	BRSS
1	MASS	69.66	69.66	-	-
2	BMASS	77.88	-	77.88	-
3	BRSS	87.35	-	-	87.35
4	JASS	85.61	-	77.49	93.51
5	1 + 2	74.04	70.02	78.09	-
6	1 + 3	77.19	69.89	-	84.34
7	1 + 4	80.61	70.12	77.83	93.58

Table 12: Component-wise and overall pre-training accuracies on ASPEC Japanese development sentences. Note in particular how BRSS accuracy is boosted in multi-task settings, while the opposite could have been expected.

accuracies for various models on the ASPEC Japanese development set sentences. Regarding individual component methods, it can be seen that MASS is the hardest task, given its low accuracy whereas BRSS is the easiest one. Moreover, the accuracy of MASS and BRSS improves when coupled with BMASS. Cross-referencing these accuracies with the BLEU scores in Table 3, we can see that an increase in BLEU scores almost has no relationship with the pre-training accuracy here. However, BMASS seems to act as an accuracy improving catalyst for BRSS and MASS which in turn has a positive impact on the translation quality.

One possible reason for this is that multi-task training of different pre-training methods helps boost the performance of individual methods. This is in accordance with several past works on multi-task training for NMT [8, 21, 23, 39]. As such, we recommend that such an analysis of multi-objective pre-training methods can help isolate the importance of individual pre-training objectives. Nevertheless, our analysis shows that the components of JASS, BMASS and BRSS, are certainly responsible for improving translation quality for Japanese-involved language pairs. The analysis of this section is not applicable to English-side because PMASS did not yield significant improvements

against MASS which can be attributed to the absence of specific syntactic unit like bunsetsu in English.

6 Conclusion

In this thesis, we proposed the JASS and ENSS, pre-training methods that leverage information from syntactic structures of sentences and novel alternatives to language-agnostic pre-training schemes such as MASS for NMT. Our work leveraged abundant monolingual data and syntactic analysis so that the pre-training phase becomes aware of specific language structures. Our experiments on ASPEC Japanese–English, Japanese–Chinese, Wikipedia Japanese–Chinese and News Commentary Japanese–Russian translation showed that JASS and ENSS outperform MASS in most low-resource settings. Furthermore, we showed that JASS and ENSS can completely substitute the corresponding language-agnostic pre-training tasks and enhance the performance for low-resource NMT. This demonstrates the importance of injecting language-specific information into the pre-training objective as well as the benefit of multi-task pre-training with diverse objectives. Our adequacy evaluation through LASER, human evaluation, and case study also showed that our methods result in a significant improvement in terms of adequacy and fluency of translations. The analysis of pre-training accuracy reveals the complementary nature of individual tasks within JASS.

Our future work will focus on implementing linguistic-aware multilingual pre-training massively by more languages for robuster pre-trained models. We will also work on determining the impact of multi-task pre-training using a combination of a wide variety of pre-training approaches that focus on different aspects of language structure. We also note that Raffel et al. [39] has recently shown that many NLP tasks such as text understanding could be reformulated as text-to-text tasks. This broadens a lot the domain of usefulness of sequence-to-sequence pre-training tasks such as ours, and we will be interested in evaluating our approach on a wider range of NLP tasks.

Acknowledgments

First, I would like to sincerely express my gratitude to my supervisors, Prof. Sadao Kurohashi, Associate Prof. Chenhui Chu, and Associate Prof. Fabien Cromières, who have been leading me into the great world of NLP and tutoring me forward on my academic path during these precious two years.

I would like to thank all the members in Kurohashi & Chu & Murawaki Lab. for your zealous support and prompt advice for my research. Without your help, this thesis could not have been forwarded.

I would like to say special thanks to other coauthors during these two years. Dr. Raj Dabre from NICT helped mentor part of this work and lead me to the forefront of machine translation at the very beginning. Yibin Shen from East China Normal University helped submit a state-of-the-art Japanese–Chinese NMT system for WAT. I am also grateful to my collaborators, Prakhar Gupta and Associate Prof. Martin Jaggi from EPFL, for their great support to my another research topic. Also thanks to my other coauthors, Haiyue Song, Assistant Prof. Fei Cheng, Prof. Cheqing Jin, and Dr. Eiichiro Sumita.

Thanks to Kyoto University – for granting me a maximum of freedom.

Thanks to Kyoto University LoveLive! Kenkyūkai – for together participating in NF, giving ōen to Aqours, and improving my spoken Japanese.

Thanks to Hyakumeizan Dōkōkai – for giving me stimulating and enjoyable trekking experiences.

Thanks to friends in Kyoto who play badminton and basketball together – for enriching my extracurricular life.

Thanks to Qianying Liu and Yujie Qian – for giving very helpful comments for my research.

Thanks to Wanzhou Zhang – for creating wonderful memories together.

Lastly, I would like to thank my parents, Mr. Jianping Mao and Ms. Xuefei Yuan, who are supporting my studying abroad and providing guidance for my lifetime.

References

[1] 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.

[2] 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.

[3] 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'14)*, pages 642–647, Reykjavik, Iceland. European Language Resources Association (ELRA).

[4] 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.

[5] 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, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

[6] Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural 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 1410–1416, Hong Kong, China. Association for Computational Linguistics.

[7] 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.

[8] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

[9] 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.

[10] Dan Han, Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2012. Head finalization reordering for Chinese-to-Japanese machine translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 57–66, Jeju, Republic of Korea. Association for Computational Linguistics.

[11] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

[12] Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. 2013. Two-stage pre-ordering for Japanese-to-English statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1062–1066, Nagoya, Japan. Asian Federation of Natural Language Processing.

[13] Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139,

Dublin, Ireland. European Association for Machine Translation.

[14] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010. Head finalization: A simple reordering rule for SOV languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 244–251, Uppsala, Sweden. Association for Computational Linguistics.

[15] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

[16] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

[17] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

[18] Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

[19] Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. 2006. Phrase reordering for statistical machine translation based on predicate-argument structure. In *2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006*, pages 77–82.

[20] 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.

[21] 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.

[22] Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

[23] 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. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

[24] 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.

[25] Zhuoyuan Mao, Raj Dabre, Fabien Cromieres, Haiyue Song, Ryota Nakao, and Sadao Kurohashi. 2020. ニューラル機械翻訳のための言語知識に基づくマルチタスク事前学習. In 言語処理学会 第 *26* 回年次大会, 茨城, pages 1061–1064.

[26] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

[27] 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.

[28] Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. 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 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

[29] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd workshop on Asian translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 1–28, Kyoto, Japan. Workshop on Asian Translation.

[30] Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

[31] 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'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

[32] 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.

[33] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[34] Carl Pollard and Ivan A. Sag. 1988. *Information-Based Syntax and Semantics: Vol. 1: Fundamentals*. Center for the Study of Language and Information, USA.

[35] Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

[36] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

[37] 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.

[38] Alec Radford. 2018. Improving language understanding by generative pre-training.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

[40] 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.

[41] Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

[42] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

[43] 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.

[44] Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

[45] 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.

[46] 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.

[47] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8968–8975.

[48] 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.

[49] 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 *Proceedings of the 30th Neural Information Processing Systems Conference (NIPS)*, pages 5998–6008, Long Beach, USA.

[50] Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. 2019. Denoising based sequence-to-sequence pre-training for text generation. 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 4003–4015, Hong Kong, China. Association for Computational Linguistics.

[51] Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

[52] Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.

[53] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

[54] Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.

[55] Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

[56] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9628–9635.

[57] Chunting Zhou, Xuezhe Ma, Junjie Hu, and Graham Neubig. 2019. Handling syntactic divergence in low-resource 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 1388–1394, Hong Kong, China. Association for Computational Linguistics.

[58] Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.

[59] 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*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Appendix

A.1 Results in Middle/High-resource Scenarios

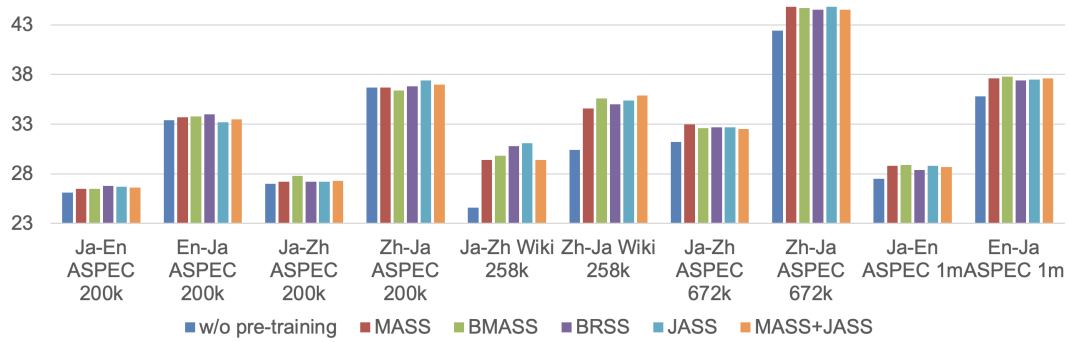

Figure A.1: BLEU results overview of middle/high-resource scenarios on ASPEC Japanese–English, Japanese–Chinese and Wikipedia Japanese–Chinese translations.

As shown in Figure A.1, we here report a BLEU result overview in middle/high-resource scenarios. The fine-tuning is performed by over 200k parallel sentences on respective language pair and domain. By comparing with models without pre-training, we find that pre-training can still contribute some improvements whereas much less than those in low-resource scenarios. Second, we observe that most pre-training methods obtained comparable BLEU results regardless of whether they are linguistically-driven methods or not. This indicates that in middle/high-resource scenarios, our proposed methods might be limited, which also means that linguistically-driven supervisions can be utilized to compensate the lack of the parallel sentences.