CSP: Code-Switching Pre-training for Neural Machine Translation

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang and Qi Ju*
Tencent Minority-Mandarin Translation
{zieenyang, bojiehu, ambyera, springhuang, damonju}@tencent.com

Abstract

This paper proposes a new pre-training method, called Code-Switching Pre-training (CSP for short) for Neural Machine Translation (NMT). Unlike traditional pre-training method which randomly masks some fragments of the input sentence, the proposed CSP randomly replaces some words in the source sentence with their translation words in the target language. Specifically, we firstly perform lexicon induction with unsupervised word embedding mapping between the source and target languages, and then randomly replace some words in the input sentence with their translation words according to the extracted translation lexicons. CSP adopts the encoder-decoder framework: its encoder takes the code-mixed sentence as input, and its decoder predicts the replaced fragment of the input sentence. In this way, CSP is able to pre-train the NMT model by explicitly making the most of the cross-lingual alignment information extracted from the source and target monolingual corpus. Additionally, we relieve the pretrain-finetune discrepancy caused by the artificial symbols like [mask]. To verify the effectiveness of the proposed method, we conduct extensive experiments on unsupervised and supervised NMT. Experimental results show that CSP achieves significant improvements over baselines without pre-training or with other pre-training methods.

1 Introduction

Neural machine translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015) which typically follows the encoder-decoder framework, directly applies a single neural network to transform the source sentence into the target sentence. With tens of millions of trainable parameters in the NMT model, translation tasks are usually data-hungry, and many of them are low-resource or even zero-resource in terms of training data. Following the idea of unsupervised and self-supervised pre-training methods in the NLP area (Peters et al., 2018; Radford et al., 2018, 2019; Devlin et al., 2019; Yang et al., 2019), some works are proposed to improve the NMT model with pre-training, by making full use of the widely available monolingual corpora (Lample and Conneau, 2019; Song et al., 2019b; Edunov et al., 2019; Huang et al., 2019; Wang et al., 2019; Rothe et al., 2019; Clinchant et al., 2019). Typically, two different branches of pre-training approaches are proposed for NMT: ***model-fusion*** and ***parameter-initialization***.

The ***model-fusion*** approaches seek to incorporate the sentence representation provided by the pre-trained model, such as BERT, into the NMT model (Yang et al., 2019b; Clinchant et al., 2019; Weng et al., 2019; Zhu et al., 2020; Lewis et al., 2019; Liu et al., 2020). These approaches are able to leverage the publicly available pre-trained checkpoints in the website but they need to change the NMT model to fuse the sentence embedding calculated by the pre-trained model. Large-scale parameters of the pre-trained model significantly increase the storage cost and inference time, which makes it hard for this branch of approaches to be directly used in production. As opposed to ***model-fusion*** approaches, the ***parameter-initialization*** approaches aim to directly pre-train the whole or part of the NMT model with tailored objectives, and then initialize the NMT model with pre-trained parameters (Lample and Conneau, 2019; Song et al., 2019b). These approaches are more production-ready since they keep the size and structure of the model same as standard NMT systems.

While achieving substantial improvements, these

* indicates corresponding author.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2624–2636,
November 16–20, 2020.

pre-training approaches have two main cons. Firstly, as pointed out by Yang et al. (2019), the artificial symbols like [mask] used by these approaches during pre-training are absent from real data at fine-tuning time, resulting in a pretrain-finetune discrepancy. Secondly, while each pre-training step only involves sentences from the same language, these approaches are unable to make use of the cross-lingual alignment information contained in the source and target monolingual corpus. We argue that, as a cross-lingual sequence generation task, NMT requires a tailored pre-training objective which is capable of making use of cross-lingual alignment signals explicitly, e.g., word-pair information extracted from the source and target monolingual corpus, to improve the performance.

To address the limitations mentioned above, we propose Code-Switching Pre-training (CSP) for NMT. We extract the word-pair alignment information from the source and target monolingual corpus automatically, and then apply the extracted alignment information to enhance the pre-training performance. The detailed training process of CSP can be presented in two steps: 1) perform lexicon induction to get translation lexicons by unsupervised word embedding mapping (Artetxe et al., 2018a; Conneau et al., 2018); 2) randomly replace some words in the input sentence with their translation words in the extracted translation lexicons and train the NMT model to predict the replaced words. CSP adopts the encoder-decoder framework: its encoder takes the code-mixed sentence as input, and its decoder predicts the replaced fragments based on the context calculated by the encoder. By predicting the sentence fragment which is replaced on the encoder side, CSP is able to either attend to the remaining words in the source language or to the translation words of the replaced fragment in the target language. Therefore, CSP trains the NMT model to: 1) learn how to build the sentence representation for the input sentence as the traditional pre-training methods do; 2) learn how to perform cross-lingual translation with extracted word-pair alignment information. In summary, we mainly make the following contributions:

- We propose the code-switching pre-training for NMT, which makes full use of the cross-lingual alignment information contained in source and target monolingual corpus to improve the pre-training for NMT.
- We conduct extensive experiments on supervised and unsupervised translation tasks. Experimental results show that the proposed approach consistently achieves substantial improvements.
- Last but not least, we find that CSP can successfully handle the code-switching inputs.

2 Related works

Several approaches have been proposed to improve NMT with pre-training. Edunov et al. (2019) proposed to feed the last layer of ELMo to the encoder of NMT and investigated several different ways to add pre-trained language model representations to the NMT model. Weng et al. (2019) proposed a bi-directional self-attention language model to get sentence representation and introduced two individual methods, namely weighted-fusion mechanism and knowledge transfer paradigm, to enhance the encoder and decoder. Yang et al. (2019b) proposed a concerted training framework to make the most use of BERT in NMT. Zhu et al. (2020) proposed to fuse the representations from BERT with each layer of the encoder and decoder of the NMT model through attention mechanisms. Large-scale parameters of the pre-trained model in these approaches discussed above significantly increase the storage cost and inference time, which makes these approaches a little far from production.[1] The other branch of approaches aims to keep the structure and size the same to the standard NMT system and designs some pre-training objectives tailored for NMT. Lample and Conneau (2019) proposed Cross-Lingual Language Model (XLM) objective and built a universal cross-lingual encoder. To improve the cross-lingual pre-training, they introduced supervised translation language modeling objective relying on the parallel data available. Song et al. (2019b) proposed the MASS objective to pre-train the whole NMT model instead of only pre-training the encoder by XLM. CSP builds on top of Lample and Conneau (2019) and Song et al. (2019b), and it explicitly makes full use of the alignment information extracted from the source and target monolingual corpus to enhance pre-training.

There have also been works on applying pre-specified translation lexicons to improve the performance of NMT. Hokamp and Liu (2017) and Post

[1] To be used in production easily, these models need to be distilled into a student model with the structure and size same as standard NMT systems.

and Vilar (2018) proposed an altered beam search algorithm, which took target-side pre-specified translations as lexical constraints during beam search. Song et al. (2019a) investigated a data augmentation method, making code-switched training data by replacing source phrases with their target translations according to the pre-specified translation lexicons. Recently, motivated by the success of unsupervised cross-lingual embeddings, Artetxe et al. (2018b), Lample et al. (2018a) and Yang et al. (2018) applied the pre-trained translation lexicons to initialize the word embeddings of the unsupervised NMT model. Sun et al. (2019) applied translation lexicons to unsupervised domain adaptation in NMT. In this paper, we apply the translation lexicons automatically extracted from the monolingual corpus to improve the pre-training of NMT.

3 CSP

In this section, we firstly describe how to build the shared vocabulary for the NMT model; then we present the way extracting the probabilistic translation lexicons; and we introduce the detailed training process of CSP finally.

3.1 Shared sub-word vocabulary

This paper processes the source and target languages with the same shared vocabulary created through the sub-word toolkits, such as SentencePiece (SP) and Byte-Pair Encoding (BPE) (Sennrich et al., 2016b). We learn the sub-word splits on the concatenation of the sentences equally sampled from the source and target corpus. The motivation behind is two-fold: Firstly, with processing the source and target languages by the shared vocabulary, the encoder of the NMT model is able to share the same vocabulary with the decoder. Sharing the vocabulary between the encoder and decoder makes it possible for CSP to replace the source words in the input sentence with their translation words in the target language. Secondly, as pointed out by Lample and Conneau (2019), the shared vocabulary greatly improves the alignment of embedding spaces.

3.2 Probabilistic translation lexicons

Recently, some works successfully learned translation equivalences between word pairs from two monolingual corpus and extracted translation lexicons (Artetxe et al., 2018a; Conneau et al., 2018). Following Artetxe et al. (2018a), we utilize unsupervised word embedding mapping to extract probabilistic translation lexicons with monolingual corpus only. The probabilistic translation lexicons in this paper are defined as one-to-many source-target word translations. Specifically, giving separate source and target word embeddings, i.e., X_e and Y_e trained on source and target monolingual corpus X and Y, unsupervised word embedding mapping utilizes self-learning or adversarial-training to learn a mapping function $f(X) = WX$, which transforms source and target word embeddings to a shared embedding space. With word embeddings in the same latent space, we measure the similarities between source and target words with the cosine distance of word embeddings. Then, we extract the probabilistic translation lexicons by selecting the top k nearest neighbors in the shared embedding space. Formally, considering the word x_i in the source language, its top k nearest neighbor words in the target language, denoted as $y'_{i1}, y'_{i2}, \ldots, y'_{ik}$ are extracted as its translation words, and the corresponding normalized similarities $s'_{i1}, s'_{i2}, \ldots, s'_{ik}$ are defined as the translation probabilities.

3.3 Training process of CSP

CSP only requires monolingual data to pre-train the NMT model. Given an unpaired source sentence $x \in X$, where $x = (x_1, x_2, \ldots, x_m)$ is the source sentence with m tokens, we denote $x_{[u:v]}$ as the sentence fragment of x from u to v where $0 < u < v < m$, and denote $x^{\backslash u:v}$ as modified version of x where its fragment from position u to v are replaced with their translation words according to the probabilistic translation lexicons. Formally, $x^{\backslash u:v}$ is represented as:

$$x^{\backslash u:v} = (x_1, \ldots, x_{u-1}, y'_u, \ldots, y'_v, x_{v+1} \ldots, x_m) \quad (1)$$

where $x^{\backslash u:v}_{[u:v]} = (y'_u, \ldots, y'_v)$ is sampled based on the extracted probabilistic translation lexicons presented on Section 3.2. Here, we take the replacing process from x_u to y'_u as an example. Considering the source word x_u, its top k translation words $y'_{u1}, y'_{u2}, \ldots, y'_{uk}$ and the translation probabilities $s_{u1}, s_{u2}, \ldots, s_{uk}$, y'_u is calculated as:

$$y'_u = y'_{uj} (1 \leq j \leq k) \quad (2)$$

where y'_{uj} is decided by performing multinomial sampling on the distribution defined by translation probabilities $s'_{u1}, s'_{u2}, \ldots, s'_{uk}$. With higher translation probability s'_{uj}, the translation word y'_{uj} is more likely to be selected.

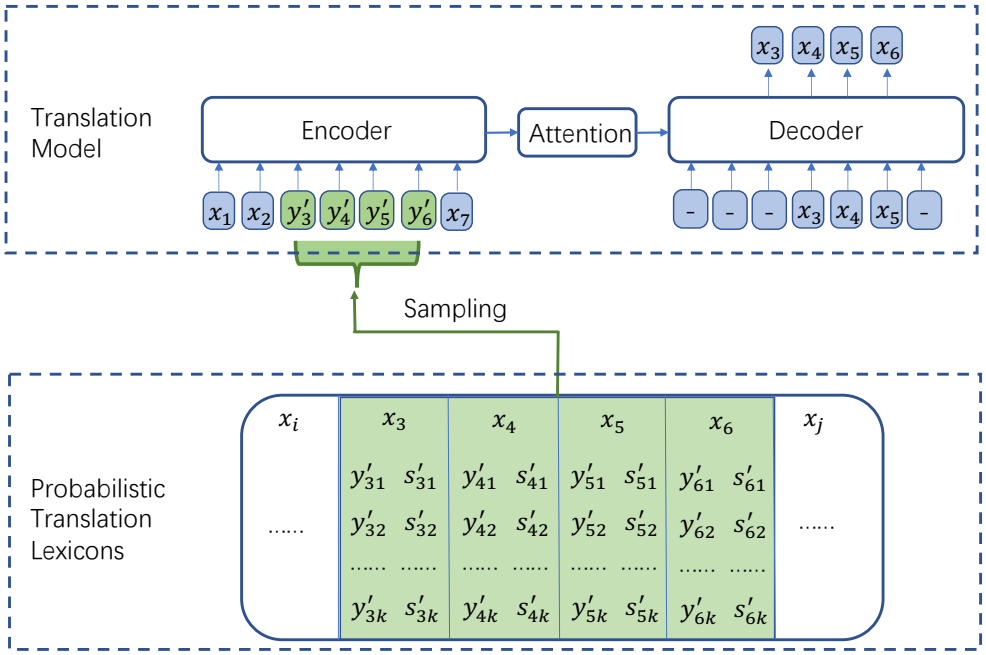

Figure 1: The training example of our proposed CSP which randomly replaces some words in the source input with their translation words based on the probabilistic translation lexicons. Identical to MAS, the token '-' represents the padding in the decoder. The attention module represents the attention between the encoder and decoder

Similar to Song et al. (2019b), CSP pre-trains a sequence to sequence model by predicting the sentence fragment $x_{[u:v]}$ with the modified sequence $x^{\backslash u:v}$ as input. With the log likelihood as the objective function, CSP trains the NMT model on the monolingual corpora X as:

$$L(\theta; X) = \frac{1}{|X|} \sum_{x \in X} logP(x_{[u:v]}|x^{\backslash u:v}; \theta)$$
$$= \frac{1}{|X|} \sum_{x \in X} log \prod_{t=u}^{v} P(x_t|x_{<t}, x^{\backslash u:v}; \theta) \quad (3)$$

Figure 1 shows an example for CSP training, where the original source sentence $(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ with the fragment (x_3, x_4, x_5, x_6) being replaced with their translation words, i.e., $(y_3^{'}, y_4^{'}, y_5^{'}, y_6^{'})$ sampled from the extracted probabilistic translation lexicons. The encoder takes the code-mixed source sentence as input, and the decoder only predicts the replaced fragment (x_3, x_4, x_5, x_6).

4 Experiments and Results

This section describes the experimental details about CSP pre-training and fine-tuning on the supervised and unsupervised NMT tasks. To test the effectiveness and generality of CSP, we conduct extensive experiments on English-German, English-French and Chinese-to-English translation tasks.

4.1 CSP pre-training

Model configuration We choose Transformer as the basic model structure. Following the base model in Vaswani et al. (2017), we set the dimension of word embedding as 512, dropout rate as 0.1 and the head number as 8. To be comparable with previous works, we set the model as 4-layer encoder and 4-layer decoder for unsupervised NMT, and 6-layer encoder and 6-layer decoder for supervised NMT. The encoder and decoder share the same word embeddings.

Datasets and pre-processing Following the work of Song et al. (2019b), we use the monolingual data sampled from WMT News Crawl datasets for English, German and French, with 50M sentences for each language.[2] For Chinese, we choose 10M sentences from the combination of LDC and WMT2018 corpora. For each translation task, the source and target languages are jointly tokenized into sub-word units with BPE (Sennrich et al., 2016b). The vocabulary is extracted from the tokenized corpora and shared by the source and target languages. For English-German and English-French translation tasks, we set the vocabulary size as 32k. For Chinese-English, the vocabulary size is set as 60k since few tokens are shared by Chinese

[2]In this paper, we lower-cased all of the case-sensitive languages by default, such as English, German and French.

System	en-de	de-en	en-fr	fr-en	zh-en
Yang et al. (2018)	10.86	14.62	16.97	15.58	14.52
Lample et al. (2018b)	17.16	21.0	25.14	24.18	-
Lample and Conneau (2019)	27.0	34.3	33.4	33.3	-
Song et al. (2019b)	28.1	35.0	37.5	**34.6**	-
Lample and Conneau (2019) (our reproduction)	27.3	33.8	32.9	33.5	22.1
Song et al. (2019b) (our reproduction)	27.9	34.7	37.3	34.1	22.8
CSP and fine-tuning (ours)	**28.7**	**35.7**	**37.9**	34.5	**23.9**

Table 1: The translation performance of the fine-tuned unsupervised NMT models. To reproduce the results of Lample and Conneau (2019) and Song et al. (2019b), we directly run their released codes on the website.[3]

and English. To extract the probabilistic translation lexicons, we utilize the monolingual corpora described above to train the embeddings for each language independently by using word2vec (Mikolov et al., 2013) . We then apply the public implementation of the method proposed by Artetxe et al. (2017) to map the source and target word embeddings to a shared-latent space.[4]

Training details We replace the consecutive tokens in the source input with their translation words sampled from the probabilistic translation lexicons, with random start position u. Following Song et al. (2019b), the length of the replaced fragment is empirically set as roughly 50% of the total number of tokens in the sentence, and the replaced tokens in the encoder will be the translation tokens 80% of the time, a random token 10% of the time and an unchanged token 10% of the time. [5] In the extracted probabilistic translation lexicons, we only keep top three translation words for each source word and also investigate how the number of translation words produces an effect on the training process. All of the models are implemented on Py-Torch and trained on 8 P40 GPU cards.[6] We use Adam optimizer with a learning rate of 0.0005 for pre-training.

4.2 Fine-tuning on unsupervised NMT

In this section, we describe the experiments on the unsupervised NMT, where we only utilize monolingual data to fine-tune the NMT model based on the pre-trained model.

Experimental settings For the unsupervised English-German and English-French translation tasks, we take the similar experimental settings to Lample and Conneau (2019); Song et al. (2019b). Specifically, we randomly sample 5M monolingual sentences from the monolingual data used during pre-training and report BLEU scores on WMT14 English-French and WMT16 English-German. For fine-tuning on the unsupervised Chinese-to-English translation task, we also randomly sample 1.6M monolingual sentences for Chinese and English respectively similar to Yang et al. (2018). We take $NIST02$ as the development set and report the BLEU score averaged on the test sets $NIST03$, $NIST04$ and $NIST05$. To be consistent with the baseline systems, we apply the script *multi-bleu.pl* to evaluate the translation performance for all of the translation tasks.

Baseline systems We take the following four strong baseline systems. Lample et al. (2018b) achieved state-of-the-art (SOTA) translation performance on unsupervised English-German and English-French translation tasks, by utilizing cross-lingual vocabulary, denoising auto-encoding and back-translation. Yang et al. (2018) proposed the weight-sharing architecture for unsupervised NMT and achieved SOTA results on unsupervised Chinese-to-English translation task. Lample and Conneau (2019) and Song et al. (2019b) are among the first endeavors to apply pre-training to unsupervised NMT, and both of them achieved substantial improvements compared to the methods without utilizing pre-training.

Results Table 1 shows the experimental results on the unsupervised NMT. From Table 1, we can find that the proposed CSP outperforms all of the previous works on English-to-German, German-to-

[3] https://github.com/facebookresearch/XLM
https://github.com/microsoft/MASS

[4] The configuration we used to run these open-source tool kits can be found in appendix A.

[5] We test different length of the replaced segment and report the results in the appendix B. We find similar results to Song et al. (2019b).

[6] The code we used can be found in the attached file.

System	en-de	en-fr	zh-en
Vaswani et al. (2017)	27.3	38.1	-
Vaswani et al. (2017) (our reproduction) / + BT	27.0 / 28.6	37.9 / 39.3	42.1 / 43.7
Lample and Conneau (2019) (our reproduction) / + BT	28.1 / 29.4	38.3 / 39.6	42.0 / 43.7
Song et al. (2019b) (our reproduction) / + BT	28.4 / 29.6	38.4 / 39.6	42.5 / 44.1
CSP and fine-tuning (ours) / + BT	**28.9 / 30.0**	**38.8 / 39.9**	**43.2 / 44.6**

Table 2: The translation performance of supervised NMT on English-German, English-French and Chinese-to-English test sets. (+ BT: trains the model with back-translation method.)

English, English-to-French and Chinese-to-English unsupervised translation tasks, with as high as +0.7 BLEU points improvement in German-to-English translation task. In French-to-English translation direction, CSP also achieves comparable results with the SOTA baseline of Song et al. (2019b). In Chinese-to-English translation task, CSP even achieves +1.1 BLEU points improvement compared to the reproduced result of Song et al. (2019b). These results indicate that fine-tuning unsupervised NMT on the model pre-trained by CSP consistently outperforms the previous unsupervised NMT systems with or without pre-training.

4.3 Fine-tuning on supervised NMT

This section describes our experiments on supervised NMT where we fine-tune the pre-trained model with bilingual data.

Experimental settings For supervised NMT, we conduct experiments on the publicly available data sets, i.e., WMT14 English-French, WMT14 English-German and LDC Chinese-to-English corpora, which are used extensively as benchmarks for NMT systems. We use the full WMT14 English-German and WMT14 English-French corpus as our training sets, which contain 4.5M and 36M sentence pairs respectively. For Chinese-to-English translation task, our training data consists of 1.6M sentence pairs randomly extracted from LDC corpora.[7] All of the sentences are encoded with the same BPE codes utilized in pre-training.

Baseline systems For supervised NMT, we consider the following three baseline systems. [8] The first one is the work of Vaswani et al. (2017), which achieves SOTA results on WMT14 English-German and English-French translation tasks. The other two baseline systems are proposed by Lample and Conneau (2019) and Song et al. (2019b), both of which fine-tune the supervised NMT tasks on the pre-trained models. Furthermore, we compare with the back-translation method which has shown its great effectiveness on improving NMT model with monolingual data (Sennrich et al., 2016a). Specifically, for each baseline system, we translate the target monolingual data used during pre-training back to the source language by a reversely-trained model, and get the pseudo-parallel corpus by combining the translation with its original data. [9] At last, the training data which includes pseudo and parallel sentence pairs is shuffled and used to train the NMT system.

Results The experimental results on supervised NMT are presented at Table 2. We report the BLEU scores on English-to-German, English-to-French and Chinese-to-English translation directions. For each translation task, we report the BLEU scores for the standard NMT model and the model trained with back-translation respectively. As shown in Table 2, compared to the baseline system without pre-training (Vaswani et al., 2017), the proposed model achieves +1.6 and +0.7 BLEU points improvements on English-to-German and English-to-French translation directions respectively. Even compared to stronger baseline system with pre-training (Song et al., 2019b), we also achieve +0.5 and +0.4 BLEU points improvements respectively on these two translation directions. On Chinese-to-English translation task, the proposed model achieves +0.7 BLEU points improvement compared to the baseline system of Song et al. (2019b). With back-translation, the proposed model still outperforms all of the baseline systems. Experimental results above show that fine-tuning the supervised

[7] LDC2002L27,LDC2002T01,LDC2002E18,LDC2003E07, LDC2004T08,LDC2004E12,LDC2005T10

[8] Since ***model-fusion*** approaches incorporate too much extra parameters, it is not fair to take them as baselines here. We leave the comparison between CSP and ***mode-fusion*** approaches in the appendix C.

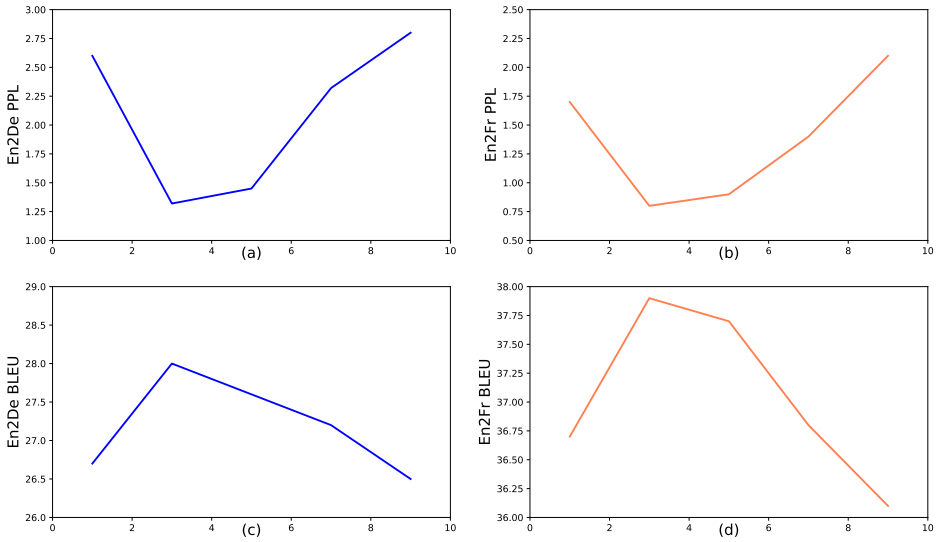

Figure 2: The performance of CSP with the probabilistic translation lexicons keeping different translation words for each source word, which includes: (a) the PPL score of the pre-trained English-to-German model; (b) the PPL score of the pre-trained English-to-French model; (c) the BLEU score of the fine-tuned unsupervised English-to-German NMT model; (d)the BLEU score of the fine-tuned unsupervised English-to-French NMT model.

NMT on the model pre-trained by CSP achieves substantial improvements over previous supervised NMT systems with or without pre-training. Additionally, it has been verified that CSP is able to work together with back-translation.

5 Analysis

5.1 Study the number of translation words

In CSP, the probabilistic translation lexicons only keep the top k translation words for each source word. For each word in the translation lexicons, the number of translation words k is viewed as an important hyper-parameter and can be set carefully during the process of pre-training. A natural question is that how much of translation words do we need to keep for each source word? Intuitively, if k is set as a small number, the model may lose its generality since each source word can be replaced with only a few translation words, which severely limits the diversity of the context. And if otherwise, the accuracy of the extracted probabilistic translation lexicons may get significantly diminished, which shall introduce too much noise for pre-training. Therefore, there is a trade-off between the generality and accuracy. We investigate this problem by studying the translation performance of unsupervised NMT with different k, where we vary k from 1 to 10 with the interval 2. We observe both the performance of CSP after pre-training and the translation performance after fine-tuning on the unsupervised NMT tasks, including the English-to-German and English-to-French translation directions. For each translation direction, we firstly present the perplexity (PPL) score of the pre-trained model averaged on the monolingual validation sets of the source and target languages.[10] And then we show the BLEU score of the fine-tuned model on the bilingual validation set. Figure 2 (a) and (c) illustrate the PPL score of the pre-trained model and BLEU score of the fine-tuned unsupervised NMT model respectively on English-to-German translation. Figure 2 (b) and (d) present the PPL and BLEU score respectively for English-to-French translation. From Figure 2, it can be seen that, when k is set around 3, the pre-trained model achieves the best validation PPL scores on both of the English-to-German and English-to-French translation directions. Similarly, CSP also achieves the best BLEU scores on the unsupervised translation tasks when k is set around 3.

5.2 Ablation study

To understand the importance of different components of the model pre-trained by CSP, we perform an ablation study by training multiple versions of

[9]We randomly select the target monolingual data with the same size to the bilingual data.

[10]For English-German translation, the monolingual validation set for English is built by including all English sentences in the bilingual English-German validation set, and the monolingual validation set for German is built in the same way.

the supervised NMT model with some components initialized randomly: the word embeddings, the encoder, the attention module between the encoder and decoder, and the decoder. Experiments are conducted on English-to-German and English-to-French translation tasks. All models are trained without back-translation and results are reported in Table 3. We can find that the two most critical components are the pre-trained encoder and attention module. It shows that CSP enhances NMT not only on the ability of building sentence representation for the input sentence, but also on the ability of aligning the source and target languages with the help of word-pair alignment information. Additionally, the experimental results indicate that the pre-trained decoder shows little effect on the translation performance. This is mainly because the decoder only predicts the source-side words during pre-training but predicts the target-side words during fine-tuning. This pretrain-finetune mismatch makes the pre-trained decoder less helpful for performance improvement.

System	en-de	en-fr
No pre-trained embeddings	28.4	38.5
No pre-trained encoder	27.9	38.2
No pre-trained attention module	28.1	38.3
No pre-trained decoder	28.8	38.8
Full model pre-trained by CSP	28.9	38.8

Table 3: Ablation study on English-German and English-French translation tasks. The embeddings include the source-side and target-side word embeddings.

5.3 Code-switching translation

Code-switching, which contains words from different languages in single input, has aroused more and more attention in NMT (Johnson et al., 2017; Menacer et al., 2019). In this section, we show that the proposed CSP is able to enhance the ability of the fine-tuned NMT model on handling the code-switching input. To present quantitative results, we build two test sets for the supervised Chinese-to-English translation task to evaluate the performance of the translation model on handling code-switching inputs. We randomly select 200 Chinese-English sentence pairs from $NIST02$, based on which we build two code-switching test sets. The first test set, referred to as test A, is built by randomly replacing some phrases in each Chinese sentence with their counterpart English phrases, where the English phrase is the translation result by feeding the corresponding Chinese phrase to the Google Chinese-to-English translator; The second test set, referred to as test B, is constructed by randomly replacing parts of the words in each Chinese sentence with their nearest target words in the shared latent embedding space (the same way used by CSP in Section 3.2). Table 4 shows the translation performance of NMT systems on the two code-switching test sets.[11] Besides the baseline systems mentioned in section 4.3, we also train a Chinese-English multi-lingual system (Johnson et al., 2017) based on Transformer, which has shown the ability of handling code-switching inputs. From Table 4, We can find that the proposed approach achieves significant improvements over previous works. Compared to multi-lingual system, we achieve +2.3 and +3.0 BLEU points improvements respectively on test A and test B. The case study can be found in appendix D.

System	test A	test B
Vaswani et al. (2017)	28.17	32.51
Lample and Conneau (2019)	28.82	32.90
Song et al. (2019b)	28.70	33.21
Multi-lingual system	30.51	35.10
CSP and fine-tuning	**32.84**	**38.17**

Table 4: The performance of Chinese-to-English translation on in-house code-switching test sets.

6 Conclusions and Future work

This work proposes a simple yet effective pre-training approach, i.e., CSP for NMT, which randomly replaces some words in the source sentence with their translation words in the probabilistic translation lexicons extracted from monolingual corpus only. To verify the effectiveness of CSP, we investigate two downstream tasks, supervised and unsupervised NMT, on English-German, English-French and Chinese-to-English translation tasks. Experimental results show that the proposed approach achieves substantial improvements over strong baselines consistently. Additionally, we show that CSP is able to enhance the ability of NMT on handling code-switching inputs. There are two promising directions for the future work. Firstly, we are interested in applying CSP to other

[11] The two in-house code-switching test sets can be found in the attached files.

related NLP areas for code-switching problems. Secondly, we plan to investigate the pre-training objectives which are more effective in utilizing the cross-lingual alignment information for NMT.

Acknowledgement

We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Stéphane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of bert for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019)*, pages 108–117.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

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

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. 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 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.

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

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *neural information processing systems (2019)*, pages 7057–7067.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Mohamed Amine Menacer, David Langlois, Denis Jouvet, Dominique Fohr, Odile Mella, and Kamel Smaïli. 2019. Machine translation on a parallel code-switched corpus. In *Canadian Conference on Artificial Intelligence*, pages 426–432. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation 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 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *arXiv preprint arXiv:1907.12461*.

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.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019a. Code-switching for enhancing NMT with pre-specified translation. 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 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019b. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1235–1245, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, pages 3104–3112.

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

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.

Rongxiang Weng, Heng Yu, Shujian Huang, Weihua Luo, and Jiajun Chen. 2019. Improving neural machine translation with pre-trained representation. *arXiv preprint arXiv:1908.07688*.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. 2019b. Towards making the most of bert in neural machine translation. *arXiv preprint arXiv:1908.05672*.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

A The configurations for the open-source toolkit

A.1 Get word embeddings with word2vec

We train the word embedding use the following script:

```
word2vec -train text -output
    embedding.txt -cbow 0 -size
    512 -window 10 -negative 10 -
    hs 0 -sample 1e- -threads 50 -
    binary 0 -min-count 5 -iter 10
```

A.2 Word embedding mapping with Vecmap

After we get the embeddings for both the source and target languages, namely s_embedding.txt and t_embedding.txt, we use the open-source Vecmap to map these embeddings to a shared-latent space with the following scripts:[12]

```
python normalize_embeddings.py
    unit center -i s_embedding.txt
     -o s_embedding.normalized.txt
```

```
python normalize_embeddings.py
    unit center -i t_embedding.txt
     -o t_embedding.normalized.txt
```

```
python map_embeddings.py --
    orthogonal s_embedding.
    normalized.txt t_embedding.
    normalized.txt
s_embedding.mapped.txt
    t_embedding.mapped.txt --
    numerals --self_learning -v
```

B Study of different length of the replaced segment

The length of the replaced fragment is an hyper-parameter which can be set by the user beforehand. We are curious to know how the length of the replaced fragment shows effect on CSP. In this section, we study the performance of CSP with different length of the replaced fragment, where we set the length of the replaced fragment from 10% to 90% percentage of the sentence length with a step size of 10%. Similar to section 5.1, we report both the performance of CSP after pre-training and the translation performance after fine-tuning on the unsupervised NMT tasks, including the English-to-German and English-to-French translation directions. For each translation direction, we firstly present the perplexity (PPL) score of the pre-trained model averaged on the monolingual validation sets of the source and target languages. And then we show the BLEU score of the fine-tuned model on the bilingual validation set. Figure 3 (a) and (c) illustrate the PPL score of the pre-trained model and BLEU score of fine-tuned unsupervised NMT model respectively on English-to-German translation direction. Figure 3 (b) and (d) present the PPL and BLEU score respectively for English-to-French translation direction. We can find that when the length of the replaced fragment is set nearly 50% of the sentence length, CSP achieves best performance not only on the pre-training task but also on the downstream unsupervised NMT task. Therefore, we set the length of the replaced fragment as 50% of the sentence length in our experiments.

[12] https://github.com/artetxem/vecmap

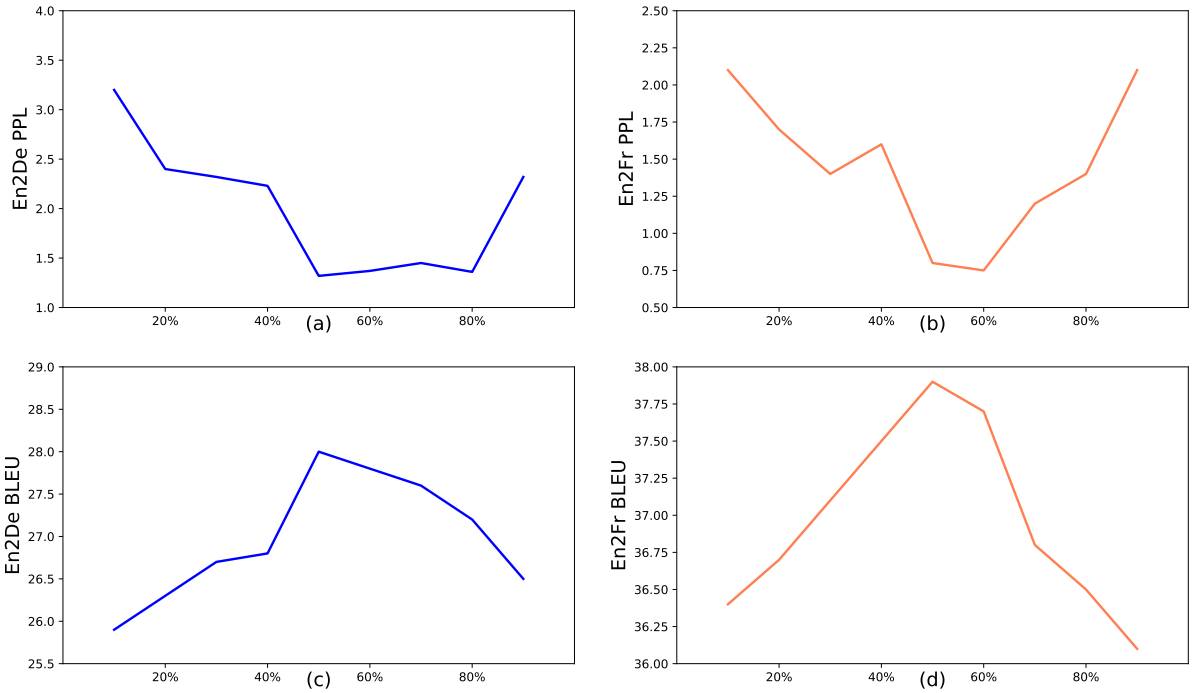

Figure 3: The performance of CSP with different length of the replaced fragment, which includes: (a) the PPL score of the pre-trained English-to-German model; (b) the PPL score of the pre-trained English-to-French model; (c) the BLEU score of the fine-tuned unsupervised English-to-German NMT model; (d)the BLEU score of the fine-tuned unsupervised English-to-French NMT model.

System	en-de	en-fr	zh-en
Zhu et al. (2020)	29.2	39.1	43.8
+Knowledge Distillation	28.7	38.5	**43.4**
CSP and fine-tuning (ours)	**28.9**	**38.8**	43.2

Table 5: The comparison between CSP and ***model-fusion*** approaches. We get the translation result of Zhu et al. (2020) by directly running their released codes on the website.[13]

C Compared to *model-fusion* approaches

In this section, we compare the proposed CSP with ***model-fusion*** approaches. We conduct experiments on supervised NMT where we fine-tune the pre-trained model with bilingual data. Experimental settings are identical to the settings in section 4.3. We report the performance of English-to-German, English-to-French and Chinese-to-English translation tasks. Since Zhu et al. (2020) released their code which makes their results reproducible, we take their system as the baseline. To make the comparison more fair, we distill the model of Zhu et al. (2020) to a student model which has the same size and structure to standard NMT model. For knowledge distillation, we utilized the sequence-level knowledge distillation proposed by Kim and Rush (2016). [14] Experimental results are presented in Table 5. We can find that, compared to the distilled student model of Zhu et al. (2020), CSP achieves better translation performance on two of three translation tasks.

D Case study for code-switching

In this section, we compare the performance of different NMT systems by case study. We randomly select some examples of the code-switching inputs and get the outputs by feeding the code-switching inputs into different NMT systems. The results are presented in Table 6. We can find that, for the two code-switching input sentences in Table 6, the standard Transformer and the multi-lingual system are both easily to give insufficient translations with some semantic contents untranslated. We assume that this is mainly because these systems are weak in encoding the full context of the code-switching input. Compared to the baseline systems, our system gives more sufficient and fluent translations. This shows that CSP enhances the model's ability

[13] https://github.com/bert-nmt/bert-nmt

[14] While variant distillation methods have been proposed recently, we only test the most simple and standard one.

Source sentence	切尼在访问所有海湾国家以后星期一到科威特，But even this most loyalUS ally is opposed to attacking Baghdad.
Output of Transformer	cheney arrived kuwait after visiting all Gulf, But even this most employed US
Output of Multi-lingual system	Cheney arrived in Kuwait after visiting, but even this most loyal is opposed to attacking Baghdad.
Output of our system	Cheney arrived in Kuwait on Monday after visiting all Gulf countries, but even the most loyal US ally is opposed to attacking Baghdad.
Reference	cheney arrived kuwait on monday after visiting all other gulf states. however , even this most loyal ally to u.s. opposes an attack on baghdad .
Source sentence	对于日本《朝日新闻》报道说 Megawati will send a personal letter from Kim Dae Jung to Kim Jong Il，韩国政府方面则予以否认。
Output of Transformer	as japan says, Megawati send a personal letter to Kim Jong, the south korea denied.
Output of Multi-lingual system	as for the news released in japan asahi that megawati will hand a letter from kim dae jung in his own handwriting to kim, the south korea denied this .
Output of our system	as for the news released in the japanese newspaper asahi that will hand a personal letter from kim dae jung in his own handwriting to kim jong , the south korean government denied .
Reference	as for the news released in the japanese newspaper asahi that megawati will hand a personal letter from kim dae jung in his own handwriting to kim jong - il , the south korean government denied this .

Table 6: Examples of the code-switching inputs and outputs of different NMT systems.

on encoding code-switching inputs.