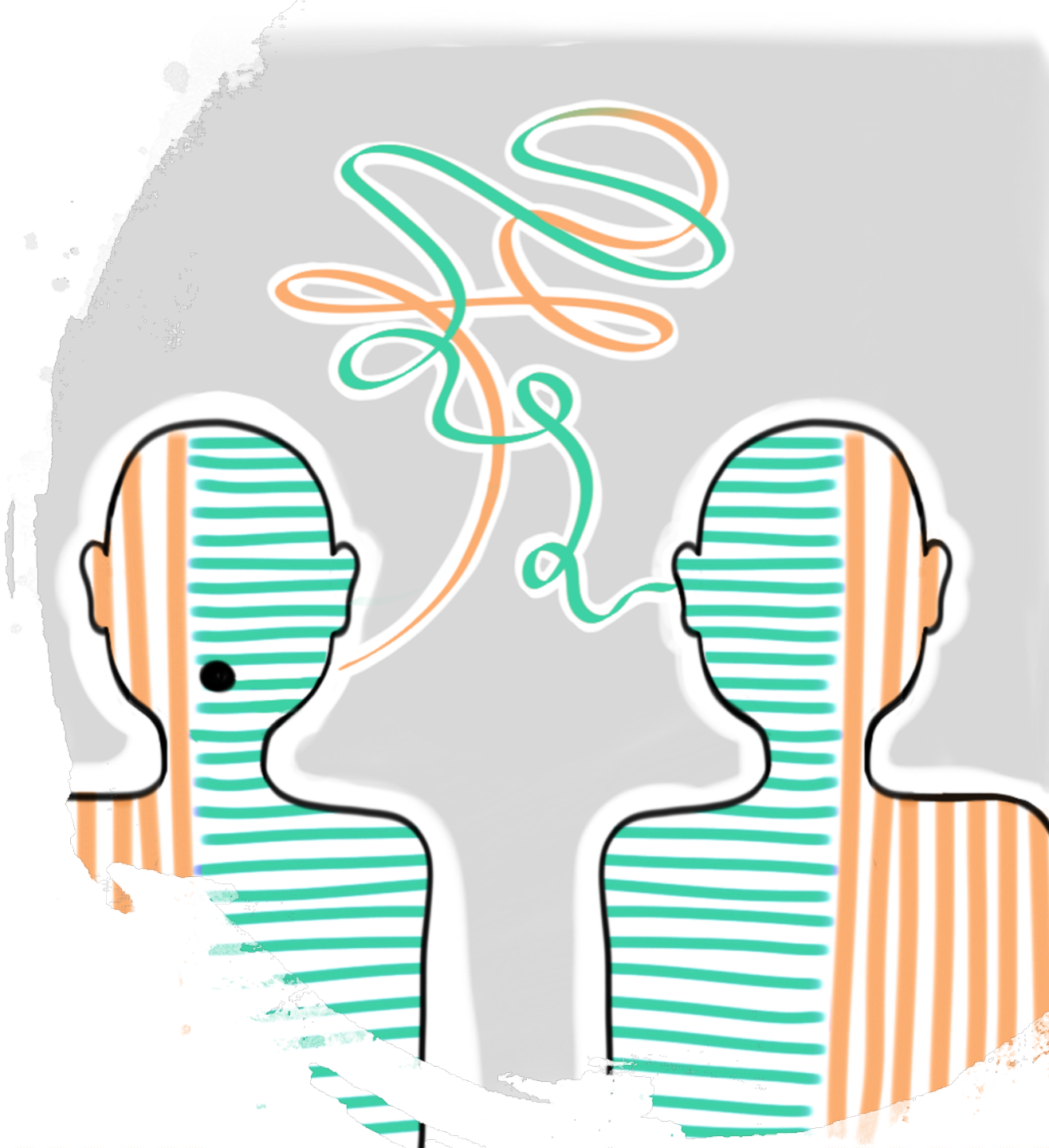
Code-Switching Pre-Training for Machine Translation (CSP)

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang and Qi Ju*

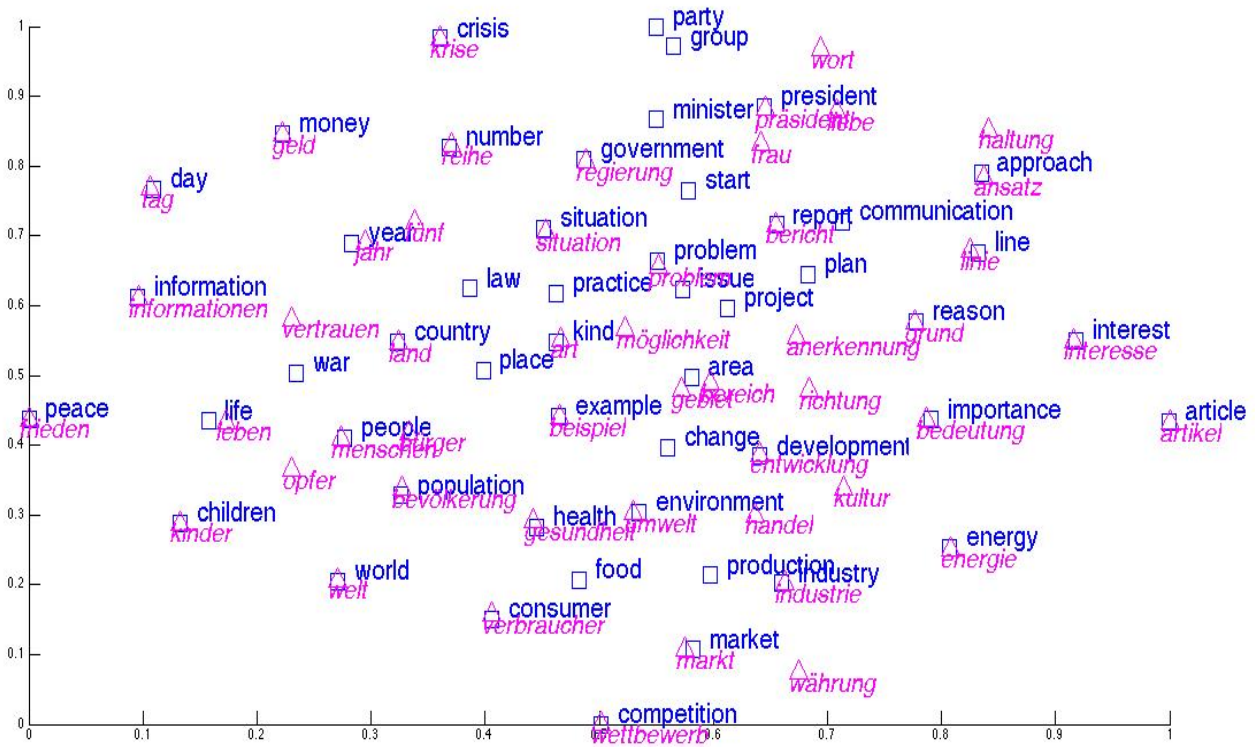
Why CSP?

- Artificial symbols like [mask] used during pre-training are absent from real data at finetuning time, resulting in a pretrain-finetune discrepancy.
- Pre-training step only involves sentences from the same language and are unable to make use of the cross-lingual alignment information contained in the source and target monolingual corpus.
- 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.



How does CSP Work?

- Cross-lingual word embeddings
  - Perform lexicon induction to ge translation lexicons by unsupervised word embedding mapping (Artetxe et al., 2018a; Conneau et al., 2018);
- Randomly replace some words in th input sentence with their translatio words in the extracted translation lexicons and train the NMT model t predict the replaced words.



How does CSP Work?

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.

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.

CSP Architecture

Shared sub-word vocabulary

- learn the sub-word splits on the concatenation of the sentences equally sampled from the source and target corpus.

Probabilistic translation lexicons

- one-to-many source-target word translations.
- learn a mapping function f(X) = W X, which transforms source and target word embeddings to a shared embedding space
- measure the similarities between source and target words with the cosine distance of word embeddings.
- extract the probabilistic translation lexicons by selecting the top k nearest neighbors in the shared embedding space.
  - For 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**.

Training process of CSP

- Given an unpaired source sentence x ∈ X where $x=(x_1,x_2..x_m)$ with m tokens
  - $x_{[u:v]}$= sentence fragment $u<v<m$
  - $x^{\backslash u:v}$ = modified sentence where fragments from u to v are replaced with their translation words
- The translation words are selected by multinomial sampling with words with highest probability
- 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
- 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)
- Log likelihood objective function

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

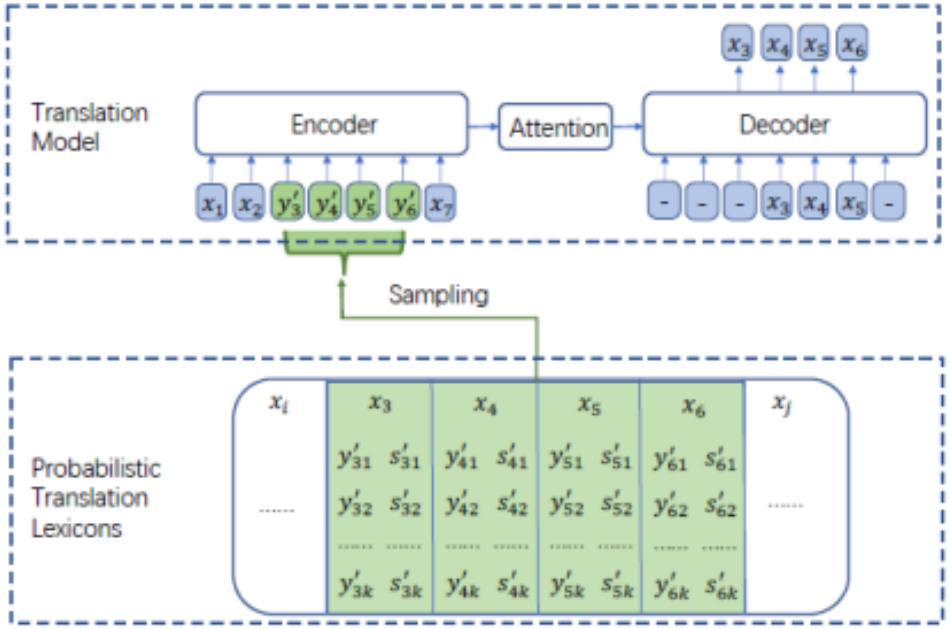


Figure 1: The training example of our proposed CSP which randomly replaces some words in the source input with their translation words based on the probabilistic translation lexicons. Identical to MAS, the token '-' represents the padding in the decoder. The attention module represents the attention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

Model configuration

- Transformer;
- 512 embedding dimension;
- dropout rate = 0.1;
- 8 heads
- 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.

Data and Pre-Processing

- Dataset
  - WMT News Crawl datasets for English, German & French, with 50M sentences for each language
  - For Chinese, 10M sentences from the combination of LDC and WMT2018 corpora
- Preprocessing
  - For each translation task, the source and target languages are jointly tokenized into sub-word units with BPE
  - The vocabulary is extracted from the tokenized corpora and shared by the source and target languages.
  - English-German and English-French = 32k BPE.
  - Chinese-English, the vocabulary size = 60k BPE
  - Use the monolingual corpora to train the embeddings for each language independently by using word2vec.
  - Use VecMap to map the source and target word embeddings to a shared-latent space.[4]

Training Details

- The number of replaced tokens is about 50% of number of tokens in a sentence
- 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.
- In the extracted probabilistic translation lexicons, we only keep top three translation words for each source word
  - investigate how the number of translation words produces an effect on the training process.
- All of the models are implemented on PyTorch and trained on 8 P40 GPU cards. We use Adam optimizer with a learning rate of 0.0005 for pre-training.

Fine Tuning on Unsupervised NMT

- 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.
- Randomly sample 1.6M monolingual sentences for Chinese and English respectively
- We take NIST02 as the development set and report the BLEU score averaged on the test sets NIST03, NIST04 and NIST05.
- Apply the script multi-bleu.pl to evaluate the translation performance for all of the translation tasks

Fine Tuning on Unsupervised NMT

- WMT14 English-French 4.5 M sentences
- WMT14 English-German 36M sentences
- LDC Chinese-to-English corpora 1.6M sentences
- All of the sentences are encoded with the same BPE codes utilized in pre-training.

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]

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.)

Results

Analysis

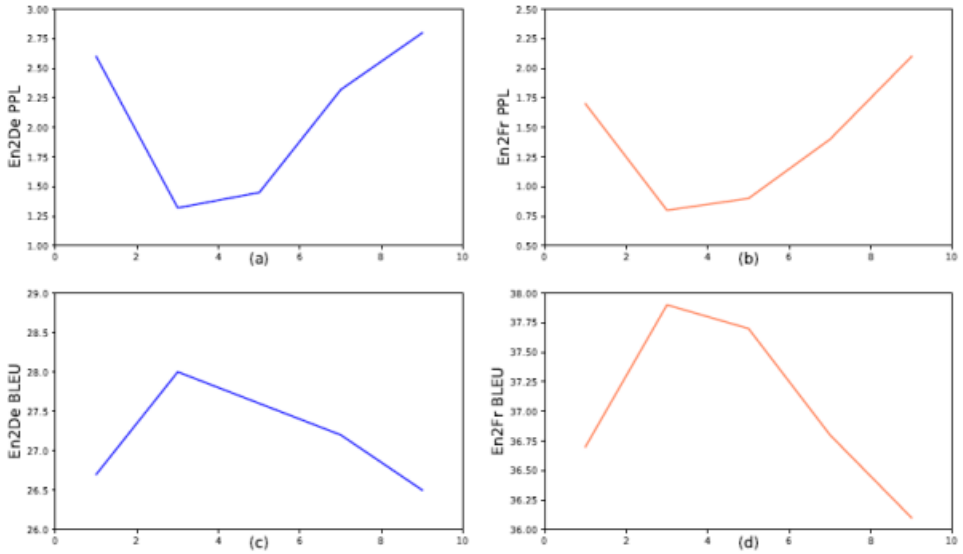


Figure 2: The performance of CSP with the probabilistic translation lexicons keeping different translation words for each source word, which includes: (a) the PPL score of the pre-trained English-to-German model; (b) the PPL score of the pre-trained English-to-French model; (c) the BLEU score of the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

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.

Code Switching Translation

- Test A – replacement with English phrases of the corresponding Chinese phrase in the Google Chinese-to-English translator
- Test B - replacement with English phrases of the corresponding Chinese phrase with translation phrases

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.

Conclusion

- CSP is able to enhance the ability of NMT on handling code-switching inputs.