SJTU-NICT's Supervised and Unsupervised Neural Machine Translation Systems for the WMT20 News Translation Task

Zuchao Li[1,2,3], **Hai Zhao**[1,2,3,*],
Rui Wang[4,*], **Kehai Chen**[4], **Masao Utiyama**[4], and **Eiichiro Sumita**[4]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU)
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institude, Shanghai Jiao Tong University, China
[4]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
charlee@sjtu.edu.cn, zhaohai@cs,sjtu.edu.cn, wangrui@nict.go.jp

Abstract

In this paper, we introduced our joint team SJTU-NICT 's participation in the WMT 2020 machine translation shared task. In this shared task, we participated in four translation directions of three language pairs: English-Chinese, English-Polish on supervised machine translation track, German-Upper Sorbian on low-resource and unsupervised machine translation tracks. Based on different conditions of language pairs, we have experimented with diverse neural machine translation (NMT) techniques: document-enhanced NMT, XLM pre-trained language model enhanced NMT, bidirectional translation as a pre-training, reference language based UNMT, data-dependent gaussian prior objective, and BT-BLEU collaborative filtering self-training. We also used the TF-IDF algorithm to filter the training set to obtain a domain more similar set with the test set for finetuning. In our submissions, the primary systems won the first place on English to Chinese, Polish to English, and German to Upper Sorbian translation directions.

1 Introduction

Our SJTU-NICT team participated in the WMT20 shared task, including supervised track, unsupervised, and low-resource track. During the participation, we placed our attention on Polish (PL) → English (EN) and English (EN) → Chinese (ZH) on the supervised track, while on the unsupervised and low-resource track, the German (DE) ↔ Upper Sorbian (HSB) both directions are focused.

Our baseline system in supervised track is based on the Transformer big architecture proposed by Vaswani et al. (2017), in which its open-source implementation version `Fairseq` (Ott et al., 2019) is adopted. In the unsupervised and low-resource track, we draw on the successful experience of the `XLM` framework (Conneau et al., 2019), and used the two-stage training mode of masked language modeling (MLM) pre-training + back-translation (BT) finetune to obtain a very strong baseline performance. `Marian` (Junczys-Dowmunt et al., 2018) toolkit is utilized for training the decoder in reranking using machine translation targets instead of common GPT-style language modeling targets.

In order to better play the role of WMT evaluation in polishing the methods proposed or improved by our team (He et al., 2018; Li et al., 2018; Zhang et al., 2018; Zhang and Zhao, 2018; Xiao et al., 2019; Zhou and Zhao, 2019; Li et al., 2019b; Luo and Zhao, 2020), we divided the three language pairs we participated in into three categories:

1. Traditional language pair with rich parallel corpus: EN-PL,
2. Language pair with document-level information: EN-ZH,
3. Language pair with no or low parallel resources: DE-HSB.

In the supervised PL→EN translation direction, we based on the XLM framework to pre-train a Polish language model using common crawl and news crawl monolingual data, and proposed the XLM enhanced NMT model inspired from the idea of incorporating BERT into NMT (Zhu et al., 2020). Besides, we trained a bidirectional translation model of EN-PL based on the parallel corpus and further finetuned it to the PL→EN

* Corresponding authors. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011), Huawei-SJTU Long Term AI Project, Cutting-edge Machine Reading Comprehension and Language Model. Rui Wang was partially supported by JSPS grant-in-aid for early-career scientists (19K20354): "Unsupervised Neural Machine Translation in Universal Scenarios" and NICT tenure-track researcher startup fund "Toward Intelligent Machine Translation".

Proceedings of the 5th Conference on Machine Translation (WMT), pages 218–229
Online, November 19–20, 2020.

direction.

In the supervised EN→ZH translation with document information, we propose a document enhanced NMT model based on Longformer (Beltagy et al., 2020). The training of our proposed document enhanced NMT model is split into three stages. In the first stage, we pre-train the Longformer document encoder with MLM target on the document text in Wikipedia dumps, UN News, and News Commentary monolingual corpus. A conventional Transformer-big NMT model is trained in the second stage. In the final stage, the Longformer encoder and conventional Transformer big NMT model are used to initialize the full document-enhanced NMT model parameters, in which the Longformer encoder is adopted to extract representations for the document of an input sequence, and then the document representations are fused with each layer of the encoder and decoder of the NMT model through attention mechanisms.

In the unsupervised machine translation track on DE-HSB, we experimented with the reference language based UNMT (RUNMT) (Li et al., 2020b) framework we proposed recently. Under this framework, we choose English as the reference language, and use the Europarl parallel corpus of EN-DE to enhance the unsupervised machine translation between DE and HSB. Specifically, we adopted reference language translation (RAT), reference language back-translation (RABT), and cross-lingual back-translation (XBT) three training targets with the help of the cross-lingual agreement provided by the EN-DE parallel corpus to enhance the unsupervised translation performance.

Due to the introduction of more explicit supervision signals brought by parallel corpus in the low-resource machine translation track on DE-HSB, we discarded the use of the weaker agreement provided by the reference language, conducted joint training on the unsupervised back-translation and the supervised (forward-)translation directly, and introduced BT-BLEU based collaborative filtering technology for further self-training. In addition, inspired by our previous work (Sun et al., 2020b), we also use MLM and translation language modeling (TLM) to continue pre-training the model while machine translation training.

In addition, in all basic NMT models, we empower the training process with our proposed data-dependent gaussian prior objective (D2GPo) (Li et al., 2020a), so that the model can maintain the diversity of the output. When the main model training is finished, the TF-IDF algorithm is employed to filter the training set according to the input of the test set, a training subset whose domain is more similar to the test set is obtained, and then used to finetune the model for reducing the performance degradation caused by domain inconsistency. For the final submission, an ensemble of several different trained models outputs the n-best predictions, and used the decoder trained with `Marian` toolkit to performs reranking to get the final system output.

2 Methodology

2.1 XLM-enhanced NMT

Pre-trained language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLM (Conneau et al., 2019), XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019) etc. have recently demonstrated a very dominant effect on natural language processing tasks. Several works (Clinchant et al., 2019; Imamura and Sumita, 2019; Zhu et al., 2020) leveraged a pre-trained BERT model for improving NMT and found that BERT can bring significantly better results over the baseline.

Since BERT and other pre-trained language models are trained on large scale corpus beyond the data provided by the WMT20 organizers, the direct use of BERT will make the system submitted unconstrained. Using an XLM model, a variant of BERT, pre-trained from scratch on the monolingual data provided by the official to enhance our NMT model, is a good choice to keep the system constrained. Moreover, the XLM model has the advantages of simple training preprocessing, low requirement for training environment that no specialized hardware such as TPU is needed. Inspired by the *BERT-fused model* proposed by Zhu et al. (2020), we built a *XLM-enhanced model*, in which we utilize XLM context-aware representations to adaptively interact with all layers in the NMT model with attention mechanism, instead of serving it as input embeddings only.

In the *XLM-enhanced model*, XLM as an additional encoder and the original encoder of NMT constitute a dual-encoder structure, which is very similar to our previous work (Li et al., 2019a). The XLM-encoder attention and XLM-decoder attention are essentially the same with the

Representation Learning Frameworks (RLFs) we proposed: Source-side fusion RLF (SRLF), Target-side fusion RLF (TRLF), and both-side fusion RLF (BRLF, which is a combination of SRLF and TRLF). Specifically, in the SRLF, given a source language input x, a Pre-trained Language Modeling (PLM) encoder (like BERT, XLM) first encodes it into a context-aware representation:

$$H_P = \texttt{PLM}^k(x), \tag{1}$$

where H_P is the output of the k-th layer of the PLM encoder. As PLM and NMT models adopt different sub-word segmentation rules or algorithms and the addition of special tokens are different, the input sequence length of PLM and NMT encoders is inconsistent or cannot correspond in every position. Assuming that i represents the position of the input sequence of NMT encoder, the hidden state H_E^l after fusion with H_P in SRLF of the l-th layer is:

$$\begin{aligned} H_E^l = \frac{1}{2}(&\texttt{attn}_\texttt{S}(H_E^{l-1}, H_E^{l-1}, H_E^{l-1}) \\ &+\texttt{attn}_\texttt{P}(H_E^{l-1}, H_P, H_P)), \end{aligned} \tag{2}$$

where $\texttt{attn}_S$ is a multi-head self-attention layer and $\texttt{attn}_P$ is the multi-head attention layer. H_E will eventually be output from the last layer as the final representation.

In the TRLF framework, the dual-encoder provides two encoded outputs; the decoder will use both contexts at the same time. In the case of layer l in the decoder, we have

$$\begin{aligned} H_{DS}^l &= \texttt{attn}_\texttt{MS}(H_D^{l-1}, H_D^{l-1}, H_D^{l-1}), \\ H_D^l &= \frac{1}{2}(\texttt{attn}_\texttt{EC}(H_{DS}^l, H_E, H_E) \\ &\quad +\texttt{attn}_\texttt{PC}(H_{DS}^l, H_P, H_P)), \end{aligned} \tag{3}$$

where $\texttt{attn}_\texttt{MS}$ is the multi-head future-masked self-attention layer, $\texttt{attn}_\texttt{EC}$ and $\texttt{attn}_\texttt{PC}$ are independent multi-head attention layer for context query.

In the condition that SRLF framework is only used, the representation of PLM is only fused into the final representation H_E in the encoder side; then the decoder side continues to use the original decoding ways: $H_D^l = \texttt{attn}_\texttt{PC}(\texttt{H}_\texttt{DS}^\texttt{l}, \texttt{H}_\texttt{E}, \texttt{H}_\texttt{E})$. While the the TRLF framework is only adopted, the output of NMT encoder is $H_E = \texttt{attn}_\texttt{S}(\texttt{H}_\texttt{E}^{\texttt{l}-1}, \texttt{H}_\texttt{E}^{\texttt{l}-1}, \texttt{H}_\texttt{E}^{\texttt{l}-1})$. A BRLF framework is a combination of these two frameworks.

Moreover, in the training of the RLFs, a same drop-net trick proposed by Zhu et al. (2020) is adopted to ensure that the features output by PLM and the conventional encoder are fully utilized. In this method, the interval of 0-1 is divided into three parts according to the pre-set drop-net ratio p_{net}, where $[0, \frac{p_{net}}{2})$ is the probability of attending to the final sum for the first `attn` in H_E^L and H_D^L, $[\frac{p_{net}}{2}, 1 - \frac{p_{net}}{2})$ is the probability for the whole H_E^L and H_D^L equation, $[1 - \frac{p_{net}}{2}, 1]$ is the probability for the second `attn` in in H_E^L and H_D^L.

2.2 Bidirectional NMT

Machine translation, in general, is unidirectional, that is, from the source language to the target language. The encoder-decoder framework for NMT has been shown effective in large data scenarios, and the more high-quality bilingual training data, the better performance the model tends to achieve. Recent works (Zoph et al., 2016; Kim et al., 2019) on translation transfer learning (Torrey and Shavlik, 2010; Pan and Yang, 2009) from rich-resource language pairs to low-resource language pairs demonstrate that translation has some universal nature in essence between different language pairs. As the source-to-target (S2T) forward translation and target-to-source (T2S) backward translation can be seen as two special language pairs in bilingual translation, it can make use of the translation universal nature to improve each other, i.e., dual learning (He et al., 2016). Based on this motivation, we developed a bidirectional NMT model, in which the S2T and T2S translation were trained and optimized jointly. Therefore, the training data was doubled to make better and full use of the costly bilingual corpus.

Given parallel corpus $\mathcal{C} = \{(x^{(n)}, y^{(n)})\}_{n=1}^N$, the bidirectional NMT model is trained in two phase. In the first *bidirectional translation as pre-training* phase, a joint training objective is used to jointly maximize the likelihood of both translation direction on the bilingual data:

$$\mathcal{L}(\theta_{parent}) = \sum_{n=1}^{N}(\log p(y^{(n)}|x^{(n)}) + \log p(x^{(n)}|y^{(n)})), \tag{4}$$

where θ_{parent} is the parameters of the model, namely *parent model*, obtained in this phase.

The second phase is *unidirectional translation fine-tuning*. Although there are commonalities in different translation directions, the differences are also very obvious. To further expose the model to the direction difference and improve the effect of unidirectional translation, we further finetune the

bidirectional pre-trained model on the bilingual data. Take S2T translation as an example; the model is optimized as follows:

$$\mathcal{L}(\theta_{S\to T}) = \sum_{n=1}^{N} \log p(y^{(n)}|x^{(n)}), \qquad (5)$$

where $\theta_{S\to T}$ is the parameters of *child model* which is initialized with θ_{parent}. Similarly, the T2S *child model* can also be obtained.

Due to the introduction of bidirectional translation in one model, follow the practice of Conneau and Lample (2019), shared subword vocabulary and shared encoder-decoder (source and target) embedding were employed to improves the alignment of embedding spaces across languages. In addition, since the encoder and decoder need to be able to handle two languages simultaneously, a language embedding was used to indicate the language being processed, so as to reduce confusion of the model.

2.3 Document-enhanced NMT

In spite of its success (Vaswani et al., 2017), sentence-level NMT has been based on strong independence and locality assumptions generally, in which the interrelations among these discourse (Jurafsky, 2000) elements were ignored. This results in that the translations may be perfect at the sentence-level but lack crucial properties of the text, hindering understanding (Maruf et al., 2019). To help to resolve ambiguities and inconsistencies in translations, some MT pioneers (Bar-Hillel, 1960; Xiong et al., 2013; Sennrich, 2018) exploit the underlying discourse structure information of a text to address this issue, while others (Bawden et al., 2018; Voita et al., 2018; Jean and Cho, 2019; Wang et al., 2019; Scherrer et al., 2019) extend the translation units with the context or use an additional context encoder and attention. It is worth noting that the essence of the document-level NMT claimed with additional context and attention is still sentence-level MT, whose translation is still output sentence by sentence. We named it as document-enhanced NMT more precisely.

Due to computational efficiency and tractability concerns, the document-enhanced NMT models mostly used document embedding, document topic information, and limited past or future context sentences, etc., rather than the truly whole document information. Recently, with the increase in computational power available to us and the well-designed neural network structures (Dai et al., 2019; Kitaev et al., 2019; Beltagy et al., 2020) for long sequence encoding, we are finally in a position to employ the whole document information for enhancing sentence-level NMT. In addition, we argue that since long sequences encoding is easier than decoding, truly whole document-level translation is still a long way off, since the bidirectional context is available in the encoder, but only the past is visible by the decoder.

Longformer To make the long documents processed with Transformer (Vaswani et al., 2017) architecture feasible or easier, a modified Transformer architecture named Longformer was proposed by Beltagy et al. (2020), in which the limitation for memory and computational requirements is addressed with a novel self-attention operation scales linearly with the sequence length.

In Longformer, the original full self-attention ($O(n^2)$ time and memory complexity) is sparsified to makes it efficient for longer sequences. There are three "attention patterns" for specifying pairs of input locations attending to one another.

- **Sliding Window** Self-attention is performed in a fixed-size window w and multiple stacked layers of such sliding windowed attention results in a large receptive field as analogs to CNNs.
- **Dilated Sliding Window** Inspired by the dilated CNNs (Oord et al., 2016), dilation gaps of size d is introduced to the window to further increase the receptive field without increasing computation.
- **Global Attention** Though the receptive field is enlarged by stacking multiple layers and dilation in sliding window and dilated sliding window attention patterns, some part of the long sequence has the requirement for keeping the full and global receptive field due to the downstream tasks, so global attention is introduced to make up this need.

In our document-enhanced NMT model, some heads in multi-head attention are set to use the sliding window pattern to focus on the local context which was revealed very important (Kovaleva et al., 2019), while others with dilation focus on longer context. Besides, as Longformer is incorporated into the NMT model, we perform

global attention on the position of `[CLS]` token in which the representation of the whole sequence (i.e., the document embedding) is generated. This makes the previous document-enhanced model with document embedding as a special case of ours. It is worth noting that since the sentence being translated is part of the document, setting its positions in the document to use global attention pattern will improve the performance; but to reduce the document computation and use cache for acceleration (not recalculate the document for each sentence), we only attend the `[CLS]` position globally.

In our document-enhanced model, the Longformer is first pre-trained with the masked language modeling objective on the monolingual document corpus. It is fixed throughout the NMT training to reduce the model parameters optimized in the training stage. Thus, Longformer can also be thought of as a pre-trained language model, as it provides a document context representation H_P for the NMT model, the integration of Longformer in *Document-enhanced NMT* is consistent with the XLM model in *XLM-enhanced NMT*.

2.4 Reference Language based UNMT

The rise of UNMT almost completely relieves the parallel corpus curse, though UNMT is still subject to unsatisfactory performance due to the vagueness of the clues available for its core back-translation training. Further enriching the idea of pivot translation by extending the use of parallel corpora beyond the source-target paradigm, we propose a new reference language-based framework for UNMT, RUNMT, in which the reference language only shares a parallel corpus with the source, but this corpus still indicates a signal clear enough to help the reconstruction training of UNMT through a proposed reference agreement mechanism.

Specifically, we proposed three kinds of reference agreement utilization approaches in (Li et al., 2020b): reference agreement translation (RAT), reference agreement back-translation (RABT), and cross-lingual back-translation (XBT).

RAT RAT utilizes the principle for translating paired sentences into the target language $\mathcal{T}$ of the source $\mathcal{S}$ and reference $\mathcal{R}$ language. Since the input the parallel, the both translation outputs should be the same. Given a parallel sentence pair $\langle s, r\rangle$ between language $\mathcal{S}$ and $\mathcal{R}$, we would ideally have $\mathbb{P}(\cdot|s;\theta_{\mathcal{S}\to\mathcal{T}}) = \mathbb{P}(\cdot|r;\theta_{\mathcal{R}\to\mathcal{T}})$, where $\theta_{\mathcal{S}\to\mathcal{T}}$ and $\theta_{\mathcal{R}\to\mathcal{T}}$ represent $\mathcal{S}\to\mathcal{T}$ and $\mathcal{R}\to\mathcal{T}$ translation models respectively. However, as the two models are trained on different data, the agreement may be corrupted. Therefore, we combine the two models to obtain the agreed-upon translation output $\tilde{t}_a$:

$$\tilde{t}_a \sim \mathbb{P}(\cdot|s,r;\theta_{\mathcal{S}\to\mathcal{T}},\theta_{\mathcal{R}\to\mathcal{T}}), \tag{6}$$

where $\mathbb{P}(\cdot|s,r;\theta_{\mathcal{S}\to\mathcal{T}},\theta_{\mathcal{R}\to\mathcal{T}})$ is

$$\prod_{i=1}^{J}[\frac{1}{2}(\mathbb{P}(\cdot|s,\tilde{t}_{<i};\theta_{\mathcal{S}\to\mathcal{T}}) + \mathbb{P}(\cdot|r,\tilde{t}_{<i};\theta_{\mathcal{R}\to\mathcal{T}}))], \tag{7}$$

$\tilde{t}_{<i}$ indicates the decoded tokens before the i-the generation step.

Finally, two synthetic sentence pairs $\langle s,\tilde{t}_a\rangle$ and $\langle r,\tilde{t}_a\rangle$ are used to train the models $\mathcal{S}\to\mathcal{T}$ and $\mathcal{R}\to\mathcal{T}$. Since the silver learning target is optimized, the smoothed cross-entropy loss $\mathcal{L}_\epsilon$ is used instead of the ordinary cross-entropy loss $\mathcal{L}$. The learning objective for RAT can be written as:

$$\mathcal{L}_{\mathbf{RAT}}(\mathcal{S},\mathcal{T},\mathcal{R}) = \mathcal{L}_\epsilon(\theta_{\mathcal{S}\to\mathcal{T}}) + \mathcal{L}_\epsilon(\theta_{\mathcal{R}\to\mathcal{T}}), \tag{8}$$

RABT With the regularized pseudo-parallel sentences in RAT, we not only train the $\mathcal{S}\to\mathcal{T}$ and $\mathcal{R}\to\mathcal{T}$ forward-translation models (as the generation direction is the same as the training direction), but also train the BT models, i.e., $\mathcal{T}\to\mathcal{S}$ and $\mathcal{T}\to\mathcal{R}$. The learning objective of RABT can be described as:

$$\mathcal{L}_{\mathbf{RABT}}(\mathcal{S},\mathcal{T},\mathcal{R}) = \mathcal{L}(\theta_{\mathcal{T}\to\mathcal{S}}) + \mathcal{L}(\theta_{\mathcal{T}\to\mathcal{R}}). \tag{9}$$

XBT The parallel corpus between languages $\mathcal{S}$ and $\mathcal{R}$ can not only bring agreement in the translations of the same target language $\mathcal{T}$, but also cross-lingual agreement, that is, using the target language as the bridge to form pivot translation (Wu and Wang, 2007; Utiyama and Isahara, 2007; Paul et al., 2009) patterns: $\mathcal{S}\to\mathcal{T}\to\mathcal{R}$ and $\mathcal{R}\to\mathcal{T}\to\mathcal{S}$. In XBT, paired sentences s and r are translated to language $\mathcal{T}$: $\tilde{t}_s$ and $\tilde{t}_r$, and forms two new pseudo-parallel pairs: $\langle\tilde{t}_s,r\rangle$ and $\langle\tilde{t}_r,s\rangle$, which promote the training of translation $\mathcal{T}\to\mathcal{R}$ and $\mathcal{T}\to\mathcal{S}$. The objective function of XBT is:

$$\mathcal{L}_{\mathbf{XBT}}(\mathcal{S},\mathcal{T},\mathcal{R}) = \mathcal{L}(\theta_{\mathcal{T}\to\mathcal{R}}) + \mathcal{L}(\theta_{\mathcal{T}\to\mathcal{S}}), \tag{10}$$

2.5 CFST: Collaborative Filter for Self-Training with BT-BLEU

Self-training, proposed by Scudder (1965), is a semi-supervised approach that utilizes unannotated

Algorithm 1 Classic Self-training

1: Train a base NMT/UNMT model $f_{\theta_{S\to T}}$ on $\mathcal{C}$
2: **repeat**
3: Apply $f_{\theta_{S\to T}}$ to the unlabeled instances $\mathcal{U}$
4: Select a subset $\mathcal{Q} \subset \{(x, f_{\theta_{S\to T}}(x))|x \in \mathcal{U}\}$
5: Update model $f_{\theta_{S\to T}}$ on $\mathcal{Q}$ with self-training objective and $\mathcal{C}$ with original objective
6: **until** convergence or maximum iterations are reached

data to create better models. Recently, self-training has been successfully applied to both NMT and UNMT fields (He et al., 2019; Sun et al., 2020a), especially for the unbalanced low-resource training data scenarios.

Formally, in self-training strategy for machine translation, a parallel dataset $\mathcal{C} = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$ in NMT and a unpaired monolingual dataset $\mathcal{D} = \{x^{(m)}\}_{m=1}^{M} \cup \{y^{(n)}\}_{n=1}^{N}$ in UNMT is used to train the initial model. Then, a subset of pseudo parallel data is incorporated to update the model with a pseudo-supervised NMT (PNMT) objective (including forward translation and backward translation) for both NMT and UNMT as shown in Algorithm 1. In NMT, a large unlabeled dataset $\mathcal{U} = \{x^{(j)}\}_{j=1}^{L}$ is used for the synthesis of pseudo-parallel corpora. While in UNMT, since the model is trained with back-translation on unpaired monolingual data, the pseudo-parallel corpora is synthesized by the monolingual data, i.e., $\mathcal{U} = \{x^{(m)}\}_{m=1}^{M}$.

Considering the translation quality can't effectively be evaluated across languages in machine translation with only the monolingual data, therefore the selection of the subset $\mathcal{Q}$, is one of the key factors for self-training. It is usually selected based on some confidence scores (e.g. log probability or perplexity, PPL) (Yarowsky, 1995), but it is also possible for S to be the whole pseudo parallel data (Zhu and Goldberg, 2009). In the backward translation based on the pseudo-parallel data, the DAE method widely used in UNMT can alleviate the impact of the noise resulted from the synthesized sentences on model training, since the synthesized sentences are only used as input. However, in the forward translation training, the quality of noisy targets will directly affect the success of the model training. Therefore, the selection of synthetic parallel corpus becomes particularly critical.

Algorithm 2 BT-BLEU based Collaborative Filter

1: Split $\mathcal{U}$ equally into two subsets $\mathcal{U}_1 = \{x^{(j)}\}_{j=1}^{L/2}$ and $\mathcal{U}_2 = \{x^{(j)}\}_{j=L/2+1}^{L}$
2: Apply $f_{\theta_{S\to T}}$ to the unlabeled instances $\mathcal{U}_1$ and $\mathcal{U}_2$
3: Train two backward translation models $f^{(1)}_{\theta_{T\to S}}$ with $\{(f_{\theta_{S\to T}}(x), x)|x \in \mathcal{U}_1\}$ and $f^{(2)}_{\theta_{T\to S}}$ with $\{(f_{\theta_{S\to T}}(x), x)|x \in \mathcal{U}_2\}$ respectively
4: Translate $\{f_{\theta_{S\to T}}(x)|x \in \mathcal{U}_2\}$ with model $f^{(1)}_{\theta_{T\to S}}$, while $\{f_{\theta_{S\to T}}(x)|x \in \mathcal{U}_1\}$ with model $f^{(2)}_{\theta_{T\to S}}$
5: Calculate BT-BLEU $\mathcal{B}$ for two subsets: $\textbf{BLEU}(f^{(2)}_{\theta_{T\to S}}(f_{\theta_{S\to T}}(x)), x), \forall x \in \mathcal{U}_1$ and $\textbf{BLEU}(f^{(1)}_{\theta_{T\to S}}(f_{\theta_{S\to T}}(x)), x), \forall x \in \mathcal{U}_2$
6: $\mathcal{Q} = \{(x, f_{\theta_{S\to T}}(x))|x \in \mathcal{U}_1, \mathcal{B} > \gamma\} \cup \{(x, f_{\theta_{S\to T}}(x))|x \in \mathcal{U}_2, \mathcal{B} > \gamma\}$

We propose a collaborative filtering algorithm based on BT-BLEU to select high quality pseudo-parallel pairs, as shown in Algorithm 2. The BT-BLEU, as defined in (Li et al., 2020b), is a BLEU of $\boldsymbol{x} \in \mathcal{S}$ and $\tilde{\boldsymbol{x}}$ generated in the $\mathcal{S} \to \mathcal{T} \to \mathcal{S}$ back-translation process. As long as the model of $\mathcal{T} \to \mathcal{S}$ is fixed and the preference for translation of certain sentences is reduced as much as possible, BT-BLEU can reflect the translation quality of $\mathcal{S} \to \mathcal{T}$ to some extent, because of the necessary but insufficient condition that only the better the translation of $\mathcal{S} \to \mathcal{T}$ is, the better the translation of $\mathcal{T} \to \mathcal{S}$ can be.

To achieve the goal of reducing translation preferences, we split the pseudo parallel set into two subsets, ensure no overlap between two subsets. The model trained on subset 1 is used for back-translation on the subset 2, while the model on subset 2 back-translate the subset 1. This collaborative translation process enables the two models not to see the sentences to be translated, which guarantees the translation not relies on tricks. Additionally, we found that the sentences in different lengths have different difficulties for back-translation; we further divide the sentences into different bags according to their lengths and use different BT-BLEU threshold γ for filtering.

2.6 TF-IDF Finetune

NMT has been prominent in many machine translation tasks. However, in some domain-specific tasks, only the corpora from similar

Systems	Dev	Test	
	BLEU	BLEU	chrF
Base Data:			
Transformer big	25.8	-	-
XLM-enhanced	26.8	-	-
Base Data + ParaCrawl:			
Transformer big	30.0	32.2	0.596
+D2GPo	30.9	-	-
XLM-enhanced	31.4	-	-
Bidirectional NMT	29.5	-	-
+Finetune	31.2	-	-
Ensemble	32.0	34.0	0.606
++TF-IDF finetune	32.3	34.2	0.609
++Re-ranking	32.5	34.6	0.610

Table 1: PL→EN performance (sacreBLEU and chrF score) for different models.

domains can improve translation performance. If a trained NMT model is evaluated on a domain mismatch corpus, the translation performance may even degrade. Therefore, domain adaptation techniques are essential to solve the NMT domain problem. It is a very common domain adaptation approach to further finetune the translation model trained on the domain-mixed corpus by using data that is the same or similar to the test set in domain. Therefore, we need to select sentences that are as close to the input domain as possible in the domain-mixed training set.

We argue that low-frequency words contain more domain information than high-frequency words, since low-frequency words are mostly domain-specific nouns, etc., which may indicate the topic directly. Therefore, we adopt the TF-IDF algorithm to search and filter on the whole training set. In fact, the improved version of TF-IDF algorithm, BM25 (Robertson and Zaragoza, 2009), is employed to calculate the sentence similarity. BM25 is based on probabilistic information retrieval theory, whose score for a term q to a sequence Q is:

$$s(Q,q) = \frac{\text{IDF} * ((k+1) * \text{TF})}{(k * (1.0 - b + b * \frac{L_Q}{L_{\text{avg}}}) + \text{TF})}, \quad (11)$$

where IDF is the Inverse Document Frequency for term q appears in the whole corpus, TF is the Term Frequency for q in D, L_Q represents the sequence length, L_{avg} is the average length of corpus D, k and b is the adjustable parameters.

With this scorer, every sequence will obtain a BM25 vector on the terms of the corpus:

$$V = [s(Q,t), \quad \forall t \in D_{\text{terms}}], \quad (12)$$

where D_{terms} indicates the all terms set in corpus D. We calculate the cosine similarity as final scores between the query and every source sentence in corpus, and ranked on the scores to get the top-K pairs (K=1000 in our experiments) as the sub-training set for finetuning.

3 Data Preprocessing and Model Setup

Before model training, we preprocessed the data uniformly and customized the processing according to the requirements of each model. We normalized punctuation, remove non-printing characters, and tokenize all data with the Moses tokenizer (Koehn et al., 2007) except for the Chinese. For Chinese, we removed the segmentation space in some training data and then use PKUSeg (Luo et al., 2019) toolkit to cut all Chinese sentences, so as to obtain unified word segmentation annotations. We use joint byte pair encodings (BPE) with 40K split operations for subword segmentation (Sennrich et al., 2016).

In *XLM-enhanced NMT* and *Document-enhanced NMT*, we first train a basic NMT (Transformer big) model on the sentence-level data until convergence, then initialize the encoder and decoder of the *XLM-enhanced NMT* and *Document-enhanced NMT* full model with the obtained model. The PLM-encoder attention $\texttt{attn}_{\texttt{P}}$ and PLM-decoder attention $\texttt{attn}_{\texttt{PC}}$ are randomly initialized.

EN-PL On the language pair EN-PL, we explored performance in two training data settings. The first is *base data*, including Europarl v10, Tilde Rapid corpus, and WikiMatrix bitext data, whose raw data is on the sentence-level. In the second setting *base data* + *paracrawl*, we converted the paragraph-level alignment data in Paracrawl to sentence-level alignment and incorporated it with the *base data*. In the conversion process, we adopted the method and program proposed by (Gale and Church, 1993) for aligning sentences based on a simple statistical model of character lengths, which uses the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter

Systems	19test	Test	
	BLEU	BLEU	chrF
Transformer big	37.2	-	-
+D2GPo	37.7	-	-
XLM-enhanced	38.9	-	-
Document-enhanced	39.2	-	-
Ensemble	40.0	48.6	0.418
++TF-IDF finetune	40.2	48.8	0.422
++Re-ranking	40.5	49.1	0.427

Table 2: EN→ZH performance (charBLEU and chrF score) for different models.

sentences. A probabilistic score is assigned to each proposed correspondence of sentences, based on the scaled difference of lengths of the two sentences (in characters) and the variance of this difference. This probabilistic score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences.

For the Polish pre-trained XLM language model, we used all NewsCrawl monolingual data and some CommonCrawl monolingual data. Since the CommonCrawl data is very large and noisy and can potentially decrease the performance of LM if it is used in its raw form. We apply language identification filtering (`langid`; Lui and Baldwin (2012)), keeping sentences with correct languages. In order to filter out the sentences shorter than 5 words or longer than 150 words more precisely, we re-split sentences using Spacy (Honnibal and Montani, 2017) toolkit.

EN-ZH In EN-ZH, the pre-training of Longformer as a document encoder is unique. As described in (Beltagy et al., 2020), the Longformer needs a large number of gradient updates to learn the local context first; before learning to utilize longer context. In the first phase of the staged training procedure, an initial RoBERTa (Liu et al., 2019) model implemented in Fairseq (Ott et al., 2019) repository was trained on the sentence-level text available. In each subsequent phase, we trained the model on the paragraph text, doubled the window size and the sequence length, and halve the learning rate. For the paragraph text, the Wikidumps and NewsCommentary v15 have document intervals and can be used directly, while UN v1.0 has no document intervals but the sentence order is not interrupted. Therefore, we use the BERT Next Sentence Prediction (NSP) classification model provided by Google for document interval prediction to recover the documents.

DE-HSB In RUNMT on EN-DE-HSB, Europarl v10 EN-DE parallel corpus is used for EN-DE NMT and RAT/RABT/XBT training[1]. Additionally, the BPE size increases to 50K for three languages. In CFST, the filtering threshold of BT-BTBLEU is set to $\gamma = 50.0$.

4 Results and Analysis

Results and ablations for PL→EN[2] are shown in Table 1, EN→ZH in Table 2, unsupervised DE↔HSB in Table 3 and low-resource DE↔HSB in Table 4. We report case-sensitive SacreBLEU scores using SacreBLEU (Post, 2018) for EN-PL, DE-HSB, and BLEU based on characters for EN-ZH. In the results, "+" means addition based on baseline, and "++" means cumulative addition based on the previous one.

In PL→EN, the introduction of ParaCrawl data improves the baseline performance on the dev dataset by about 4.2 BLEU. +D2GPo, XLM-enhanced NMT, Bidirectional NMT, and ensembling outperforms our strong baseline by 2 BLEU point. Finally, finetuning and reranking further gives another 0.5 BLEU.

For EN→ZH, as with PL→EN, we see similar improvements with +D2GPo, XLM-enhanced NMT, ensembling and reranking. We also observe that the addition of Document-enhanced NMT is much more substantial, improving single model performance by over 1.5 BLEU.

In the unsupervised track, we compared CLM, MLM, and Explicit Sentence Compression (ESC) pre-training approaches joint trained with BT in the second stage of UNMT, respectively, and found that MLM and ESC had similar effects and were stronger than CLM. Moreover, the pre-training baseline of MLM was stronger than that of MASS. The combination of unsupervised training of DE-HSB and supervised training of EN-DE achieves the purpose of transfer learning, and the improvement is greater than 3 BLEU. Based on the conclusion of MLM and BT joint training on the UNMT Baseline, we also got a similar

[1]Our systems in unsupervised track are not a constrained unsupervised system due to the utilization of additional parallel corpora.

[2]The team name for PL→EN submission is "NICT-rui" in the OCELoT site to distinguish between different sub-teams.

Systems	DE→HSB			HSB→DE		
	Dev	Test	Official	Dev	Test	Official
UnsupSMT (Artetxe et al., 2018)	17.1	14.7	-	13.8	12.6	-
MASS baseline	29.8	26.0	-	31.4	27.3	-
UNMT baseline	31.1	27.2	-	31.3	27.2	-
+CLM finetune	29.2	25.6	-	28.6	24.5	-
+MLM finetune	32.4	28.3	-	32.4	27.3	-
+ESC finetune	32.1	28.3	-	32.2	27.8	-
EN-DE-HSB MUNMT baseline	29.3	25.6	-	30.0	26.2	-
++EN-DE NMT	33.6	29.3	-	33.6	29.6	-
++MLM finetune	35.1	30.5	28.6	34.9	30.7	28.6
++RAT + RABT + XBT	47.8	41.8	40.3	40.6	35.9	32.8

Table 3: DE↔HSB unsupervised performance (sacreBLEU score) for different models.

Systems	DE→HSB			HSB→DE		
	Dev	Test	Official	Dev	Test	Official
UNMT baseline	31.1	27.2	-	31.3	27.2	-
++MLM finetune	32.4	28.3	-	32.4	27.3	-
++DE-HSB NMT	59.9	53.0	52.5	61.6	53.1	54.6
++TLM finetune	60.2	53.2	-	61.4	52.7	-
++CFST	61.3	54.5	60.2	62.2	53.9	55.6
++D2GPo	61.4	54.6	60.4	62.9	54.5	56.6
Ensemble+Re-ranking	61.5	54.7	60.7	63.3	56.1	58.5
EN-DE-HSB MUNMT baseline	29.3	25.6	-	30.0	26.2	-
++EN-DE NMT + MLM finetune	35.1	30.5	28.6	34.9	30.7	28.6
++DE-HSB NMT	59.8	53.0	-	62.0	53.7	-

Table 4: DE↔HSB low-resource performance (sacreBLEU score) for different models.

trend on the MUNMT system. In the final system, the enhancement of RAT+RABT+XBT brought a BLEU increase of 11.7 and 4.2, respectively.

In the low-resource track, the model in the unsupervised track is used as the pre-trained model, and DE-HSB NMT and BT are jointly trained. Due to the DE-HSB parallel corpus, we can not only use MLM for monolingual pre-training, but also use TLM for cross-lingual pre-training. The addition of CFST and D2GPo further improves the effect of the model, indicating that these contributions are orthogonal. In addition, comparing UNMT with MUNMT given a parallel corpus, we found that although MUNMT used more data, it did not bring about a large enough effect improvement, so we will leave it for future research.

5 Conclusion

This paper describes SJTU-NICT's submission to the WMT20 news translation task. For three typical scenarios, we adopt different strategies. In this work, we not only study the pre-trained language model to enhance MT, but also consider the impact of document information on translation. We considered both the way of converting document alignment into sentence alignment and the use of BERT's NSP to recover the structure of documents. In addition, transfer learning from supervision is taken into account in unsupervised translation, and various means are used to enhance low-resource translation. Our systems performed strongly among all the submissions: we ranked 1st in PL→EN, EN→ZH, and DE→HSB respectively, and stayed Top-3 for the HSB→DE.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. In *Advances in computers*, volume 1, pages 91–163. Elsevier.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in 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 1304–1313.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Stéphane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of bert for neural machine translation. *EMNLP-IJCNLP 2019*, page 108.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

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.

William A Gale and Kenneth Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2061–2071, Melbourne, Australia. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained bert encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31.

Sébastien Jean and Kyunghyun Cho. 2019. Context-aware learning for neural machine translation. *arXiv preprint arXiv:1903.04715*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2019a. Explicit sentence compression for neural machine translation. *arXiv preprint arXiv:1912.11980*.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020a. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020b. Reference language based unsupervised neural machine translation. *arXiv preprint arXiv:2004.02127*.

Zuchao Li, Junru Zhou, Hai Zhao, and Rui Wang. 2019b. Cross-domain transfer learning for dependency parsing. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 835–844. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Ying Luo and Hai Zhao. 2020. Bipartite flat-graph network for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6408–6418, Online. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2019. A survey on document-level machine translation: Methods and evaluation. *arXiv preprint arXiv:1912.08494*.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 221–224, Boulder, Colorado. Association for Computational Linguistics.

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. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level nmt in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61.

H Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Rico Sennrich. 2018. Why the time is ripe for discourse in machine translation. In *Second Workshop on Neural Machine Translation and Generation*.

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.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020a. Self-training for unsupervised neural machine translation in unbalanced training data scenarios. *arXiv preprint arXiv:2004.04507*.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020b. Unsupervised neural machine translation with cross-lingual language representation agreement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1170–1182.

Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

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.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274.

Xinyi Wang, Jason Weston, Michael Auli, and Yacine Jernite. 2019. Improving conditioning in context-aware sequence to sequence models. *arXiv preprint arXiv:1911.09728*.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Fengshun Xiao, Jiangtong Li, Hai Zhao, Rui Wang, and Kehai Chen. 2019. Lattice-based transformer encoder for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3090–3097, Florence, Italy. Association for Computational Linguistics.

Deyi Xiong, Yang Ding, Min Zhang, and Chew Lim Tan. 2013. Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of the 2013 conference on empirical methods in Natural Language Processing*, pages 1563–1573.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Huan Zhang and Hai Zhao. 2018. Minimum divergence vs. maximum margin: an empirical comparison on seq2seq models. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junru Zhou and Hai Zhao. 2019. Head-Driven Phrase Structure Grammar parsing on Penn Treebank. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2396–2408, Florence, Italy. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

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.