Tencent AI Lab Machine Translation Systems for WMT20 Chat Translation Task

Longyue Wang[§] **Zhaopeng Tu**[§] **Xing Wang**[§] **Li Ding**[†*] **Liang Ding**[‡*] **Shuming Shi**[§]
[§]Tencent AI Lab {*vinnylywang,zptu,brightxwang,shumingshi*}*@tencent.com*
[†]Hong Kong Polytechnic University *dingli@oppo.com*
[‡]The University of Sydney *ldin3097@uni.sydney.edu.au*

Abstract

This paper describes the Tencent AI Lab's submission of the WMT 2020 shared task on chat translation in English⇔German. Our neural machine translation (NMT) systems are built on sentence-level, document-level, non-autoregressive (NAT) and pretrained models. We integrate a number of advanced techniques into our systems, including data selection, back/forward translation, larger batch learning, model ensemble, finetuning as well as system combination. Specifically, we proposed a hybrid data selection method to select high-quality and in-domain sentences from out-of-domain data. To better capture the source contexts, we exploit to augment NAT models with evolved cross-attention. Furthermore, we explore to transfer general knowledge from four different pre-training language models to the downstream translation task. In general, we present extensive experimental results for this new translation task. Among all the participants, our German⇒English primary system is ranked the second in terms of BLEU scores.

1 Introduction

Although neural machine translation (NMT, Bahdanau et al., 2015; Vaswani et al., 2017; Gehring et al., 2017) has achieved great progress in recent years, translating conversational text is still a challenging task due to its inherent characteristics such as discourse awareness (Maruf et al., 2018; Wang et al., 2019), informality (Wang et al., 2018; Yang et al., 2019) and personality (Mirkin et al., 2015; Wang et al., 2016). This is a task-oriented chat translation task (Wang et al., 2017a; Farajian et al., 2020), which aims to translating conversations between customers and agents. As a customer and an agent can respectively natively speak in German and English, the systems should translate the customer's utterances in German⇒English (De⇒En) while the agent's in German⇐English (De⇐En).

In this paper, we present our submission to the novel task in De⇔En. We explore a breadth of established techniques for building Chat NMT systems. Specifically, our systems are based on the self-attention networks including both sentence- and document-level Transformer (Vaswani et al., 2017; Wang et al., 2017b). Besides, we investigated non-autoregressive translation (NAT) models augmented with our recently proposed evolved cross-attention (Ding et al., 2020). Technically, we used the most recent effective strategies including back/forward translation, data selection, domain adaptation, batch learning, finetuning, model ensemble and system combination. Particularly, we proposed a multi-feature data selection on large general-domain data. We not only use three language models (i.e. n-gram, Transformer and BERT based LMs) to filter low-quality sentences, but also employ feature decay algorithms (FDA, Biçici and Yuret, 2011) to select domain-relevant data. In addition, we explore large batching (Ott et al., 2018) for this task and found that it can significantly outperform models with regular batching settings. To alleviate the low-resource problem, we employ large scale pre-training language models including monolingual BERT (Devlin et al., 2019a), bilingual XLM (Conneau and Lample, 2019) and multilingual mBART (Liu et al., 2020), of which knowledge can be transferred to chat translation models.[1] For better finetuning, we investigate homogenous and heterogeneous strategies (e.g. from sentence-level to document-level architectures). Simultaneously, we conduct fully-adapted data processing, model ensemble, back/forward translation and system combination.

[*] This work was conducted when Li Ding and Liang Ding were interning at Tencent AI Lab. Li Ding is now working at OPPO Research Institute.

[1] We experimented mBART after the official submission.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 483–491
Online, November 19–20, 2020.

According to the official evaluation results, our systems in De⇒En and De⇐En are respectively ranked 2nd and 4th.[2] Furthermore, a number of advanced technologies reported in this paper are also adapted to our systems for biomedical translation (Wang et al., 2020) and news translation (Wu et al., 2020) tasks, which respectively achieve up to 1st and 2nd ranks in terms of BLEU scores. Though our empirical experiments, we gain some interesting findings on the chat translation task:

1. The presented data selection method improves the baseline model by up to +18.5 BLEU points. It helps a lot for small-scale data.
2. The large batch learning works well, which makes sentence-level NMT models perform the best among different NMT models.
3. Our proposed method can improve the NAT model by +0.6 BLEU point, which is still hard to beat its autoregressive teachers.
4. Document-level contexts are not useful on the chat translation task due to the limitation of contextual data.
5. It is difficult to transfer general knowledge from pretrained LMs to the downstream translation task.

The rest of this paper is organized as follows. Section 2 introduces data statistics and our processing methods. In Section 3, we present our system with four different models: sentence-level NMT, document-level NMT, non-autoregressive NMT and NMT with pre-training LMs. Section 4 describes advanced technique integrated into our systems such as data selection and system combination. In Section 5, we reports ablation study and experimental results, which is followed by our conclusion in Section 6.

2 Data and Processing

2.1 Data

The parallel data we use to train NMT systems consist of two parts: in-domain and out-of-domain corpora. The monolingual data used for back/forward translation are all out-of-domain. Table 1 shows the statistics of data in En-De.

[2]The primary systems are ranked according to BLEU. And the official results are listed in http://www.statmt.org/wmt20/chat-task_results_DA.html.

Data	# Sents	# Ave. Len.
Parallel		
In-domain	13,845	10.3/10.1
Valid	1,902	10.3/10.2
Test	2,100	10.1/10.0
Out-of-domain	46,074,573	23.4/22.4
+filter	33,293,382	24.3/23.6
+select	1,000,000	21.4/20.9
Monolingual		
Out-of-domain De	58,044,806	28.0
+filter	56,508,715	27.1
+select	1,000,000	24.2
Out-of-domain En	34,209,709	17.2
+filter	32,823,301	16.6
+select	1,000,000	14.5

Table 1: Data statistics after pre-processing. Note that in-domain/valid/test set is speaker-ignored combined and their average lengths are counted based on En/De.

In-domain Parallel Data The small-scale in-domain corpus is constructed by the task organizer.[3] The training, validation and test sets contain utterances in task-based dialogues with contextual information. We use both w/ and w/o context formats for training corresponding models. Although there exists duplicated/noisy sentences, we do not further filter such limited data.

Out-of-domain Parallel Data The participants are allowed to use all the training data in the News shared task.[4] Thus, we combine six corpora including Europarl, ParaCrawl, CommonCrawl, TildeRapid, NewsCommentary and WikiMatrix. We first filter noisy sentence pairs (as detailed in Section 2.2) and simultaneously select parts of them as pseudo-in-domain data (as detailed in Section 4.1).

Out-of-domain Monolingual Data Due to the high degree of sentence similarity within the TaskMaster monolingual corpus,[5] participants are not allowed to use the in-domain monolingual data to train their systems. Thus, we collect part of monolingual data in news domain, which consists of CommonCrawl and NewsCommentary. We conduct data selection (in Section 4.1) to select similar amount of sentences for back/forward translation.

[3]https://github.com/Unbabel/BConTrasT.

[4]http://www.statmt.org/wmt20/translation-task.html.

[5]https://github.com/google-research-datasets/Taskmaster.

We do not use larger monolingual corpora (e.g. CommonCrawl) and leave this for future work.

2.2 Processing

Pre-processing To pre-process the raw data, we employ a series of open-source/in-house scripts, including full-/half-width conversion, Unicode conversation, punctuation normalization, tokenization and true-casing. After filtering steps, we generate subwords via Joint BPE (Sennrich et al., 2016b) with 32K merge operations.

Filtering To improve the quality of data, we filter noisy sentence pairs according to their characteristics in terms of language identification, duplication, length, invalid string and edit distance. According to our observations, the filtering method can significantly reduce noise issues including misalignment, translation error, illegal characters, over-translation and under-translation.

Post-processing After decoding, we process detokenizer and de-truecaser on system outputs. We found that the toolkit can not precisely deal with all cases. Thus, we automatically fix these bugs according to bilingual agreement.

3 Models

We adopt four different model architectures namely: SENT, DOC, NAT and PRETRAIN.

3.1 Sentence-level NMT (SENT)

We use standard TRANSFORMER models (Vaswani et al., 2017) with two customized settings. Due to data limitation, we use the small settings (SENT-S)[6] with regular batch size (4096 tokens $\times$ 8 GPUs). Based on the base settings (SENT-B),[7] we also empirically adopt big batch learning (Ott et al., 2018) (16348 tokens $\times$ 4 GPUs) with larger dropout (0.3).

3.2 Document-level NMT (DOC)

To improve discourse properties for chat translation, we re-implement our document-level model (Wang et al., 2017b) on top of TRANSFORMER. Its addition encoder reads $N = 3$ previous source sentences as history context and the representations are integrated into the standard NMT

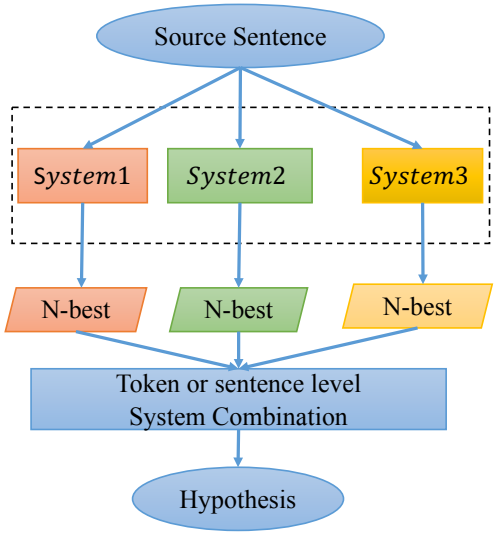

Figure 1: The simplified system combination process, into which we feed each system/model with the source sentence, in turn obtain corresponding n-best result. After pooling all system results, we can perform the token-level or sentence-level system combination decoding and obtain the final hypothesis.

for translating the current sentence. The other configures are same as SENT with small settings.

3.3 Non-autoregressive NMT (NAT)

Different from autoregressive NMT models that generate each target word conditioned on previously generated ones, NAT models break the autoregressive factorization and produce target words in parallel (Gu et al., 2018). Although NAT is proposed to speed up the inference, we exploit it to alleviate sequential error accumulation and improve the diversity in conversational translation. To adequately capture the source contexts, we proposed evolved cross-attention for NAT decoder by modeling the local and global attention simultaneously (Ding et al., 2020). Accordingly, we implement our method based on the advanced MaskPredict model (Ghazvininejad et al., 2019)[8], which uses the conditional mask LM (Devlin et al., 2019a) to iteratively generate the target sequence from the masked input.

3.4 Pretraining NMT (PRETRAIN)

To transfer the general knowledge to chat translation models, we explore to initialize (part of) model parameters with different pretrained language/generation models. Li et al. (2019) showed

[6] https://github.com/pytorch/fairseq/blob/master/fairseq/models/transformer.py#L947.

[7] https://github.com/pytorch/fairseq/blob/master/fairseq/models/transformer.py#L902.

[8] https://github.com/facebookresearch/Mask-Predict.

#CP	En-De	De-En
1	60.32	59.51
5	**60.33**	**59.53**
10	60.26	59.42
15	60.19	59.34
20	60.23	59.22
ENS	**60.49**	**60.08**

(a) Model average and ensemble.

#BM	En-De	De-En
4	60.33	59.23
8	60.33	**59.53**
12	60.33	59.24
14	60.34	59.27
16	**60.37**	59.28
20	60.28	59.19

(b) Beam size.

#LP	En-De	De-En
0.8	57.78	57.27
0.9	57.82	57.31
1.0	57.83	57.46
1.1	**57.90**	**57.50**
1.2	57.84	57.49
1.3	57.82	57.49

(c) Length penalty.

Table 2: Effects of different hyper-parameters on translation quality of SENT-B model. The BLEU score is calculated based on *combined* and *tokenized* validation set by *muti-bleu.perl*, which is different from official evaluation.

that large scale generative pretraining could be used to initialize the the document-level translation model by concatenating the current sentence and its context. We follow their work to build the BERT→DOC model. Furthermore, Conneau and Lample (2019) proposed to directly train a novel cross-lingual pretraining language model (XLM) to facilitate translation task. Accordingly, we adopt XLM pretrained model[9] to sentence-level NMT (XLM→SENT). More recently, Liu et al. (2020) proposed a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective. As they showed promising results on document translation, we additionally conducted the experiment on Chat data after submitting our systems.[10]

4 Approaches

We integrated advanced techniques into our systems, including data selection, model ensemble, back/forward translation, larger batch learning, finetuning, and system combination.

4.1 Data Selection

Inspired by Ding and Tao (2019), multi-feature language modelling can select high-quality data from a large monolingual or bilingual corpus. We present a four-feature selection criterion, which scoring each sentence by BERT LM (Devlin et al., 2019b), Transformer LM (Bei et al., 2018), N-gram LM (Stolcke, 2002) and FDA (Biçici and Yuret, 2011). Three LMs are complement each other on measuring qualities of sentences while FDA can measure its domain relevance given a in-domain dataset. Sentence pairs in the out-of-domain corpus are ranked by a sum of the above feature scores, and we selected top-M instances as pseudo-in-domain data. According to our observations, the selected data can maintain both high-quality and in-domain properties. For BERT LMs, we exploit two models built by Google[11] and our Tencent AI Lab, which are trained on massive multilingual data. The Transformer LM is trained on all in-domain and out-of-domain data via Marian.[12] Besides, we used FDA toolkit[13] to score domain relevance between in-domain and out-of-domain data.

4.2 Checkpoint Average and Model Ensemble

For each model, we stored the top-L checkpoints according to their BLEU scores (instead of PPL or training time) on validation set and generated a final checkpoint with averaged weights to avoid stochasticity. To combine different models (maybe different architectures), we further ensembled the averaged checkpoints in each model. In our preliminary experiments, we find that this hybrid combination method outperforms solely combining checkpoints or models in terms of robustness and effectiveness.

4.3 Finetuning

We employ various finetuning strategies at different phases of training. For Sent-Out→Sent-In finetune (same architecture but different data), we first train a sentence-level model on large pseudo-in-domain data and then continuously train it on small in-domain data. We apply similar strategy for Doc-Out→Doc-In finetuning, and the only difference is to use document-level data. However, pseudo-in-domain data have no document-level contexts

[9] https://github.com/facebookresearch/XLM.

[10] https://github.com/pytorch/fairseq/tree/master/examples/mbart.

[11] https://github.com/google-research/bert.

[12] https://github.com/marian-nmt/marian.

[13] https://github.com/bicici/FDA.

Method	# Sent.	BLEU
SENT-B	10K	41.87
+Bi-FDA	300K	59.36
	500K	59.81
	1M	**59.96**
+Bi-FDA-XL	500K	59.86
	800K	**59.95**
	1M	59.68
+Mono-FDA-XL	800K	**60.36**
	1M	59.80

Table 3: BLEU scores of SENT-BASE model on En⇒De task with different FDA variants (three LMs scoring are consistent).

and we use "$\langle/s\rangle$" symbols as their pseudo contexts (Kim et al., 2019; Li et al., 2020). Besides, we conduct Sent-Out→Doc-In finetuning (different architectures and data). Specifically, we first train a sentence-level model on pseudo-in-domain data and then use parts of corresponding parameters to warm-up a document-level model, which will be continuously trained on in-domain data.

4.4 Back/Forward Translation

Following Section 2, we obtain processed monolingual data. For back translation (BT), we use the best backward translation model to translate from target to source side and produce the synthetic corpus, which is used to enhance the autoregressive NMT models (Sennrich et al., 2016a). About forward translation (FT), we employ forward translation model to perform sequence distillation for NAT models (Kim and Rush, 2016) .

4.5 System Combination

As shown in Figure 1, in order to take full advantages of different systems ($\mathcal{M}odel_1$, $\mathcal{M}odel_2$ and $\mathcal{M}odel_3$), we explore both token- and sentence-level combination strategies.

Token-level We perform token-level combination with confusion network. Concretely, our method follows Consensus Network Minimum Bayes Risk (ConMBR) (Sim et al., 2007), which can be modeled as $E_{ConMBR} = \text{argmin}_{E'}\mathcal{L}(E', E_{con})$, where E_{con} was obtained as backbone through performing consensus network decoding.

Sentence-level We employ the reranking strategy to combine sentence-level systems. Particularly, the sentence reranker contains the best left-to-right (L2R) translation model, R2L (right-to-left) translation model and T2S (target-to-source) translation model. They are integrated by K-best batch MIRA training (Cherry and Foster, 2012) on valid set.

Systems	Integration	BLEU
Models		
SENT-B	IN	42.56
	IN+OUT	**59.81**
SENT-S	IN	41.87
	IN+OUT	58.62
DOC	IN	45.65
	IN+OUT	51.12
	IN→IN	51.93
NAT	IN+OUT	54.01
	*IN+OUT	54.59
Pretrain		
SENT→DOC	OUT→IN	49.77
	OUT→IN+OUT	51.58
XLM→SENT	IN+OUT	**59.61**
BERT→DOC	IN+OUT	56.01
MBART→SENT	IN+OUT	57.48

Table 4: BLEU scores of SENT, DOC, NAT and PRE-TRAIN with different finetuning strategies on En⇒De.

5 Experimental Results

Unless otherwise specified, reported BLEU scores are calculated based on *combined* and *tokenized* validation set by *muti-bleu.perl*, which is different from the official evaluation method.

5.1 Ablation Study

Table 2 investigates effects of different settings on translation quality. We then apply the best hyper-parameters to the models in Section 4 if applicable.

Effects of Model Average and Ensemble Following Section 4.2, we averaged top-L checkpoints in SENT-B model and found that it performs best when $L = 5$. We followed the same operation for SENT-S model and then combined two best averaged models (one from SENT-B and the other one from SENT-S) via ensemble method. As shown in Table 2(a), the ENS model (i.e. "average + ensemble") performs the best.

Effects of Beam Size and Length Penalty Table 2(b) and 2(c) report BLEU scores of SENT-B model using different beam size and length penalty,

# Methods	En⇒De	De⇒En
SENT-S	**59.12**	**59.61**
+BT	59.05	59.22
SENT-B	**60.33**	**59.53**
+BT	59.34	58.94
+FT	59.80	58.94
NAT	54.01	56.58
+FT	**56.56**	**56.69**
XLM	**59.61**	**60.96**
+BT	59.43	58.84

Table 5: BLEU scores of back-translation and forward-translation strategies for different models.

Models	En⇒De		De⇒En	
	-Dom.	+Dom.	-Dom.	+Dom.
Valid Set (combined)				
SENT-S	**60.47**	60.31	**62.66**	61.19
SENT-B	**62.28**	62.08	**64.99**	63.00
XLM	**61.12**	60.85	64.19	**61.30**
Valid Set (split)				
SENT-S	**60.69**	60.48	60.05	**62.09**
SENT-B	61.65	**61.93**	59.64	**63.31**
XLM	**60.90**	60.74	61.12	**62.04**
AVE.	**61.08**	61.05	62.27	**62.48**

Table 6: BLEU scores of domain adaptation strategy for different models.

respectively. As seen, it obtains the best performance when using larger beam size (e.g. 8 or 16). The length penalty prefers around 1.0 because En and De belong to similar language family.

5.2 Main Results

This section mainly reports translation qualities across different models and approaches (in Section 3 and 4). Finally we combine all of them via techniques integration and system combination.

Data Selection Table 3 demonstrates the translation performances of SENT-BASE on different FDA variants. "+Bi-FDA" means using bilingual in-domain data as seed to select N sentences from out-of-domain data while "+Bi-FDA-XL" indicates using larger seed (iteratively add selected pseudo-in-domain data to seed). "Mono" means that we only use source-side data for data selection. As seen, selected data from News domain can help to significantly improve translation quality. However, monolingual selection ("+Mono-FDA-XL") performs better than bilingual one ("+Bi-FDA-XL") and obtain the best BLEU score when $N = 800K$.

Models and Pretraining Table 4 illustrates the translation performances of various NMT models (i.e. SENT, DOC, NAT) with different training strategies. As seen, all models are hungry for larger in-domain data due to the data limitation problem (IN+OUT vs. IN). About sentence-level models, the "base + big batch" setting performs better than the "small" one (SENT-B vs. SENT-S). However, it is difficult for document-level models to outperform sentence-level ones (DOC vs. SENT). The interesting finding is that the document-level model trained on pseudo contexts ("IN+OUT") can improve the baseline that is trained on only real context ("IN") by +5.47 BLEU points. We think there are two main reasons: 1) it lacks of large-scale training data with contextual information; 2) it is still unclear how the context help document translation, which is similar to the conclusion in previous work(Kim et al., 2019; Li et al., 2020). About NAT models, our proposed approach can improve the vanilla NAT by +0.6 BLEU point, which are lower than those of autoregressive NMT models.

About pre-training, we first explore SENT→DOC, which train a sentence-level model and then use part of their parameters to warm-up a document-level model. However, it is still lower than sentence-level models. The performance of BERT→DOC is much better than pure document-level models (56.01 vs. 51.93), which confirms our hypothesis that contextual data is limited in this task. Furthermore, the XLM→SENT can obtain 59.61 BLEU points which are closed to that of SENT-B. The MBART→SENT with CC25 pretrained model can achieve 57.48 BLEU points. We find that performances of most pretraining models can not beat that of the best sentence-level model. There are two possible reasons: 1) needing a number of tricks on finetuning; 2) it is not easy to transfer general knowledge to downstream specific tasks.

Back/Forward Translation Table 5 empirically shows the translation performances of BT and FT for different models, including SENT-S, SENT-B, NAT and PRE-TRAIN. In particular, we performed BT for all systems except NAT, while deploying FT on NAT and SENT-B. As seen, augmenting with monolingual data via BT/FT can not achieve better performances than pure models. The reason

Combination type	En⇒De	De⇒En
Token-level	58.91	59.53
Sentence-level	60.41	62.41

Table 7: Model performance after system combination.

may be that we only use a small part of large-scale monolingual data in news domain. In future work, we will exploit to select in-domain data from the larger monolingual corpus.

Sub-domain Adaptation Modeling of all the speakers and language directions involved in the conversation, where each can be regarded as a different sub-domain. We conduct domain adaptation for different models to avoid performance corruption caused by domain shifting in Table 6. Specifically, we finetune the well-trained models w/ and w/o domain adaptation, denoted as "-Dom." and "+Dom.", and evaluated them on domain *combined* and *split* valid sets. As seen, domain adaptation helps De⇒En more on valid set ("AVE." 61.27 vs. 61.48), while has no much benefits on En⇒De tasks. While evaluating on combined valid sets has a bias towards models without domain adaptation. We attribute this interesting phenomenon to personality and will explore it in the future.

System Combination In order to make full use of the optimal models obtained by the above strategies, we perform token- and sentence-level system combination simultaneously. For each strategy, we generate the n-best candidates to perform the combination. As shown in Table 7, although token-level combination preserves more lexical diversity and avoids the stochasticity, its translation performance is significantly weaker (averagely -2.19 BLEU points) than sentence-level combination. Encouragingly, the sentence-level combination outperforms token-level one on valid set, which is thus used in our final system (in Table 8).

5.3 Official Results

The official automatic evaluation results of our submissions for WMT 2020 are presented in Table 8. For the primary submission, the SYS-1 combines SENT (ensembled SENT-B and SENT-S), DOC and NAT models. As contrastive submissions, the SYS-2 combines SENT and XLM models while the SYS-3 combines SENT, DOC, NAT and XLM ones. Among participated teams, our primary systems achieve the second and the forth BLEU scores on De⇒En and En⇒De, respectively.

Systems	En⇒De		De⇒En	
	Valid	Test	Valid	Test
SYS-1	60.41	58.6	62.41	62.3
SYS-2	58.91	53.6	59.53	54.0
SYS-3	60.42	58.6	62.40	61.9
BEST	–	60.4	–	62.4

Table 8: Official BLEU scores of our submissions for WMT20 Chat task. The BEST denotes the best BLEU scores of systems submitted by participants.

6 Conclusion

The paper is a system description for the Tencent AI Lab's entry into the WMT2020 Chat Translation Task. We explore a breadth of established techniques for building chat translation systems. The paper includes numerous models making use of sentence-level, document-level, non-autoregressive NMT. It also investigates a number of advanced techniques including data selection, model ensemble, finetuing, back/forward translation and initialization using a pretrained LMs. We present extensive experimental results and hope that this work could help both MT researchers and industries to boost the performance of discourse-aware MT systems (Hardmeier, 2014; Wang, 2019).

Acknowledgments

The authors wish to thank the organizers of WMT2020 Chat Translation for their prompt responses on our questions. The authors also specially thank Dr. Xuebo Liu (University of Macau) and Dr. Siyou Liu (Macao Polytechnic Institute), who kindly support us by their engineering and linguistic suggestions, respectively.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, and Conghu Yuan. 2018. An empirical study of machine translation for the shared task of WMT18. In *WMT*.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *WMT*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *NAACL*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Liang Ding and Dacheng Tao. 2019. The university of sydney's machine translation system for wmt19. In *WMT*.

Liang Ding, Longyue Wang, Di Wu, Dacheng Tao, and Tu Zhaopeng. 2020. Localness matters: The evolved cross-attention for non-autoregressive translation. In *COLING*.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the wmt 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *ICLR*.

Christian Hardmeier. 2014. *Discourse in statistical machine translation*. Ph.D. thesis, Acta Universitatis Upsaliensis.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *DiscoMT*.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. *arXiv preprint arXiv:2005.03393*.

Liangyou Li, Xin Jiang, and Qun Liu. 2019. Pretrained language models for document-level neural machine translation. *arXiv*.

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*.

Sameen Maruf, André FT Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *EMNLP*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *WMT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*.

Khe Chai Sim, William J Byrne, Mark JF Gales, Hichem Sahbi, and Philip C Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *ICASSP*.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *ICSLP*.

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 *NeurIPS*.

Longyue Wang. 2019. *Discourse-aware neural machine translation*. Ph.D. thesis, Dublin City University.

Longyue Wang, Jinhua Du, Liangyou Li, Zhaopeng Tu, Andy Way, and Qun Liu. 2017a. Semantics-enhanced task-oriented dialogue translation: A case study on hotel booking. In *IJCNLP*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *AAAI*.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One model to learn both: Zero pronoun prediction and translation. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017b. Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016. The automatic construction of discourse corpus for dialogue translation. In *LREC*.

Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2020. Tencent AI Lab machine translation systems for the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*.

Shuangzhi Wu, Xing Wang, Longyue Wang, Fangxu Liu, Jun Xie, Zhaopeng Tu, Shuming Shi, and Mu Li. 2020. Tencent neural machine translation systems for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*.

Jingxuan Yang, Jianzhuo Tong, Si Li, Sheng Gao, Jun Guo, and Nianwen Xue. 2019. Recovering dropped pronouns in Chinese conversations via modeling their referents. In *NAACL*.