Tencent Neural Machine Translation Systems for WMT18

Mingxuan Wang, Li Gong, Wenhuan Zhu, Jun Xie, Chao Bian
Mobile Internet Group, Tencent Technology Co., Ltd
{xuanswang,ligong,wenhuanzhu,stiffxie,chaobian}@tencent.com

Abstract

We participated in the WMT 2018 shared news translation task on English↔Chinese language pair. Our systems are based on attentional sequence-to-sequence models with some form of recursion and self-attention. Some data augmentation methods are also introduced to improve the translation performance. The best translation result is obtained with ensemble and reranking techniques. Our Chinese→English system achieved the highest cased BLEU score among all 16 submitted systems, and our English→Chinese system ranked the third out of 18 submitted systems.

1 Introduction

In recent years, the emergence of seq2seq models has revolutionized the field of MT by replacing traditional phrase-based approaches with neural machine translation (NMT) systems based on the encoder-decoder paradigm. A successful extension of encoder-decoder models is the attention mechanism which conducts a soft search over source tokens and yields an attentive vector to represent the most relevant segments of the source sentence for the current decoding state (Luong et al., 2015; Bahdanau et al., 2014; Wu et al., 2016; Sutskever et al., 2014; Tu et al., 2016; Zhou et al., 2016). Most recently, the Transformer model, which is based solely on a self-attention mechanism and feed-forward connections, has further advanced the field of NMT, both in terms of translation quality and speed of convergence(Vaswani et al., 2017; Ahmed et al., 2018). In this paper, we describe the Tencent NMT (TNMT) systems submissions for the WMT 2018 Chinese→English and English→Chinese translation task.

We propose two different architectures as our end to end approaches namely RNMT and Transformer. For RNMT, we implemented a hybrid multi-layer attention-based encoder-decoder model. The decoder was implemented as Recurrent Neural Networks (RNNs) and the encoder was represented with self-attention layers. We also integrated with some recent promising techniques in RNMT including the methods which made significantly contribution to the success of Transformer. In doing so, we come up with an enhanced version of RNMT that achieves comparable performance with Transformer. For Transformer, we follow the latest version of the Transformer model in the public Tensor2Tensor[1] codebase. The Transformer model replaces the recurrent connections with self-attention which can be taken as a complement with the RNMT model.

For data augmentation, we used automatic back-translation of a sub-selected monolingual News corpus as additional training data(Sennrich et al., 2015). To achieve strong machine translation performance, we further leverage the joint training method described in (Hassan et al., 2018) to optimize both the target-to-source (T2S) and source-to-target (S2T) model by extending

[1] https://github.com/tensorflow/tensor2tensor

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 526–531

https://doi.org/10.18653/v1/W18-6429

the back-translation method. The joint training method uses both the monolingual and bilingual data and updates NMT models through several iterations. We also apply several knowledge distillation methods to leverage the information gain of different architectures. To alleviate the exposure bias problem of the *left-to-right* (L2R) model, Agreement Regularization was introduced as a teacher network (Hassan et al., 2018; Liu et al., 2016). Ensemble teacher networks and architecture teacher networks are also introduced to boost the performance of a single model.

In addition, we consider the system combination and improve the performance by reranking (Koehn et al., 2003) the n-best translation outputs of the ensemble models with some effective features, including the *target-to-source* (T2S) score, *left-to-right* (L2R) score, *right-to-left* (R2L) score, Transformer score and RNMT score. The ensemble models are trained with different architectures or parameter settings to increase the diversity of the system. As a result, our Chinese→English system achieved the highest cased BLEU score among all 16 submitted systems, and our English→Chinese system ranked the third out of 18 submitted systems.

2 NMT Baseline System

We apply two different NMT architectures for the shared news translation task as our baseline systems.

1. RNMT: A hybrid deep attentional encoder-decoder networks with a stack Long Short Term Memory (LSTM) recurrent neural network for decoder and a deep self-attention network for encoder. Inspired by Transformer, Multi-head additive attention is used instead of the single-head attention in the RNMT model. Layer normalization is also applied within the output of LSTM cells. In our setup, the dimension of word embeddings and the hidden layers are both set to 1024. The encoder has 6 self-attention layers and the decoder has 3 LSTM layers.
2. Transformer: Our reimplementation of tensor2tensor with minor changes. We also implement a C++ version of the system for speeding up the decoding process. The default parameters of Transformer Big model is adopted as our transformer baseline and we further change the hyper-parameters to find the best settings on the develop set.

We train the models with adadelta(Zeiler, 2012), reshuffling the training corpus between epochs. We batched sentence pairs by approximate length, and limited input and output tokens per batch to 8192 per GPU. Each resulting training batch contained approximately $60,000$ source and $60,000$ target tokens. To avoid gradient explosion, the gradients of the cost function which had ℓ_2 norm larger than a predefined threshold 25 were normalized to the threshold. During training, we employed label smoothing of value ranging from 0.05 to 0.2 and set dropout rate from 0.01 to 0.3(Hinton et al., 2012; Pereyra et al., 2017). We perform early stopping on the baseline system and validate the model every 1000 mini-batches against BLEU on the WMT 17 news translation test set.

3 Experiment Techniques

3.1 Back Translation

In statistical machine translation, large monolingual corpora in the output language have traditionally been used for training language models to make the system output more fluent. However, it is difficult to integrate language models in current NMT architectures. Instead of ignoring such large monolingual corpora, Sennrich et al. (2015) exploited large corpora in the output language by translating a subset of them into the input language and then using the resulting synthetic sentence pairs as additional training data. We translated monolingual English text into Chinese using our English→Chinese system and translated monolingual Chinese text into English using our Chinese→English system described in Section 2. To improve the quality of the synthetic corpus we propose to use the ensemble models to translate the target sentence.

To select sentences for back-translation, we used semi-supervised convolutional neural network classifiers (Chen et al., 2017) and LSTM language models respectively. We selected 80M sentences from the target monolingual corpus based on both their classifier and language model scores, which reflect their similarity to the in-domain corpus. The selected sentences are then translated and divided into 8 portions with each contains 10M synthetic sentence pairs. Each portion is used to enhance an individual baseline model.

3.2 Joint Training of Source-to-Target and Target-to-Source Models

Back translation augments parallel data with plentiful monolingual data, allowing us to train source-to-target (S2T) models with the help of target-to-source (T2S) models. In order to leverage both source and target language monolingual data, and also let S2T and T2S models help each other, we leverage the joint training method to optimize them by extending the back-translation method(Zhang et al., 2018).

The joint training method uses both the source and the target monolingual data and updates NMT models through several iterations. In iteration 1, the process can be viewed as traditional back translation methods. The T2S model translated the target monolingual data to help the S2T model. Similarly, we can optimize the T2S translation model with the help of S2T translation model. In iteration 2, the above process is repeated, and the synthetic training data are re-translated with the updated T2S and S2T model. It is worth noting that ensemble models are used to generate the synthetic corpus so that the negative impact of noisy translations can be minimized. In order to increase the robustness of the system, we also re-translated the target of the bilingual corpus as the synthetic data. The joint training process continues until the performance on a development data set is no longer improved. We repeated three iterations for all our systems.

3.3 Knowledge Distillation

Knowledge distillation describes a method for training a student network to perform better by learning from a stronger teacher network. In our experiments, it is surprising to find that the teacher network is not necessarily stronger than the student network. The student network is capable of learning complementary information from even a worser heterogeneous teacher. We therefore investigated three different kinds of teacher networks to enhance the translation performance of a student NMT network.

R2L Teacher The approach is also referred as *Agreement Regularization of Left-to-Right and Right-to-Left Models* to integrate the information of R2L models to L2R ones (Hassan et al., 2018) . Following this work, we translate the source sentences of the parallel data with R2L model and use the translated pseudo corpus to improve the L2R model. It is worth noting that we filter the pseudo corpus with BLEU score lower than 30.

Ensemble Teacher We also apply knowledge distillation on ensemble teacher models (Freitag et al., 2017). Similar with R2L teacher model, we use ensemble models to translate the source side sentence of the parallel corpus and then apply the pseudo corpus to the training corpus.

Architecture Teacher The RNMT and Transformer models achieve similar performances but use very different ways to encode and decode context which leverage the advantages by combine the information of both architectures. We therefore use a teacher network to boost a student network with different arctectures.

3.4 System Combination and Re-ranking

For single models, we average the last 60 checkpoints to avoid overfitting. The checkpoint are saved every 600 seconds. For ensemble models, we trained 8 systems with different parameters and the different portion of monolingual corpus selected in Section 3.1. Since both the source

and target sentences can be generated from left to right and from right to left, we can have a total of eight ensemble systems, which including RNMT-S2T-R2L, RNMT-S2T-L2R, Transformer-S2T-L2R, Transformer-S2T-R2L, RNMT-T2S-R2L, RNMT-T2S-L2R, Transformer-T2S-L2R and Transformer-T2S-R2L.

For both S2T and T2S direction, we rescored 200-best lists output from four ensemble systems (S2T or T2S) using a rescoring model consisting of eight features: four S2T ensemble model scores and four T2S ensemble model scores.

4 Experiments Settings and Results

4.1 Pre-processing and Post-processing

We first segmented the Chinese sentences with our Chinese word segmentation tool and tokenized English sentences with the scripts provided in Mosess[2]. To enable open-vocabulary, we use BPE (Sennrich et al., 2016) with 50K operations. In our preliminary experiments, we found that BPE works better than UNK replacement techniques. We also filter bad sentences according to the alignment score obtained by fast-align toolkit [3] and remove duplications in the training data. The preprocessed training data consists of 19M bilingual pairs.

For Chinese→English translation, the final output was true-cased and de-tokenized with the scripts provided in Moses. For English→Chinese translation, we normalized the punctuations of the outputs with our in-house script and remove the space between the Chinese characters.

4.2 Chinese→English Systems

Table 1 shows the Chinese→English translation results on validation set (WMT2017). We reported cased BLEU scores calculated with Moses mteval-v13a.pl script[4]. The Transformer and RNMT model achieved similar results in terms of the mean BLEU scores which is consistent with the observations of Chen et al., (2018). In order to obtain more diverse models and better ensemble results, we trained eight models independently with different random initializations and dropout rate ranging from 0.01 to 0.3.

SYSTEM	BLEU
RNMT	
Baseline	24.2
+ Back Translation	25.4
+ Joint Training	26.1
+ R2L Teacher	27.1
+ Transformer Teacher	27.3
+ Ensemble Teacher	27.7
Transformer	
Baseline	24.3
+ALL features	27.6
System Combination	
Ensemble Baseline + Rerank	26.1
Ensemble BT + Rerank	27.2
Ensemble Best	27.9
Ensemble Best + Rerank	28.5

Table 1: Chinese→English Systems BLEU results on development set (WMT17). Submitted system is the last system.

The synthetic data plays an import role in the success of our system. As for the single model, back translation improved the strong baseline by 1.2 BLEU score. Even for system combination, the synthetic data still achieved a stable improvements from 26.1 to 27.2 in terms of BLEU. As an extension of the back translation method, the joint training approach interactively makes data augmentation by boosting source-to-target and target-to-source NMT systems. The method again obtained a substantial improvements up to 0.7 BLEU score.

Among knowledge distillation methods, the R2L teacher significantly enhanced our single system by 1.0 BLEU score. The Transformer teacher and ensemble teacher further get an improvements by 0.2 and 0.4 in terms of BLEU.

Applying different combinations of the techniques described in Section 3.4, we build eight single systems with all the optimization techniques described in Section 3. We then obtained

[2]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[3]https://github.com/clab/fast_align

[4]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl

four ensemble models including Transformer-L2R, Transformer-R2L, RNMT-L2R and RNMT-R2L. We then rescored 800 best lists output from the our ensemble NMT systems using a rescoring mode consisting of eight features. As can be seen in the Table 1. After ensemble a little improvement over the best single model by 0.2 BLEU is achieved. One possible explanation is that the information gain of the ensemble model has been obtained by the distillation method. For rerank model, we finally achieved an improvements of 0.6 BLEU score with fine-tuned feature weights.

4.3 English→Chinese Systems

SYSTEM	BLEU
RNMT	
Baseline	35.9
+ Joint Training	38.5
+ ALL features	40.1
Transformer	
Baseline	35.0
+ALL features	39.8
System Combination	
Ensemble Best	40.4
Ensemble Best + Rerank	41.1

Table 2: English→Chinese Systems BLEU results on development set (WMT17). Submitted system is the last system.

Table 2 shows the English→Chinese translation results on development set. All results are evaluated by character-level BLEU. We followed exactly the same settings with the Chinese→English translation system. In this case, the Joint Training method brought a substantial improvement over 2.6 BLEU scores showing the advantages of using the monolingual data and integrating the S2T model and T2S model. For knowledge distillation, We observed an improvement of 1.6 BLEU score. Finally, we applied ensemble and reranking methods, which provided 1.3 BLEU improvements over the best single model.

5 Conclusion

We present the *Tencent* NMT systems for WMT 2018 Chinese↔English news translation tasks. For both translation directions, our final systems achieved substantial improvements up by $4 \sim 5$ BLEU score over baseline systems by integrating the following technique:

1. Back translation the target monolingual data set
2. Joint training of the S2T and T2S systems
3. Knowledge distillation with R2L teacher networks, architecture teacher networks and ensemble teacher networks
4. System combination and reranking.

As a result, our submitted Chinese→English system achieved the highest cased BLEU score among all 16 submitted systems and our English→Chinese system ranked the third out of 18 submitted systems.

References

[Ahmed et al.2018] Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. 2018. Weighted transformer network for machine translation. *arXiv: Artificial Intelligence*.

[Bahdanau et al.2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Chen et al.2017] Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46.

[Chen et al.2018] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George F Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. *meeting of the association for computational linguistics*.

[Freitag et al.2017] Markus Freitag, Yaser Alonaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv: Computation and Language*.

[Hassan et al.2018] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

[Hinton et al.2012] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors.

[Koehn et al.2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

[Liu et al.2016] Lemao Liu, Masao Utiyama, Andrew M Finch, and Eiichiro Sumita. 2016. Agreement on target-bidirectional neural machine translation. pages 411–416.

[Luong et al.2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

[Pereyra et al.2017] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv: Neural and Evolutionary Computing*.

[Sennrich et al.2015] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

[Sennrich et al.2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *meeting of the association for computational linguistics*, 1:1715–1725.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Tu et al.2016] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. *ArXiv eprints, January*.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Llion Jones, Jakob Uszkoreit, Aidan N Gomez, and ukasz Kaiser. 2017. Attention is all you need. *neural information processing systems*, pages 5998–6008.

[Wu et al.2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

[Zeiler2012] Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

[Zhang et al.2018] Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. *national conference on artificial intelligence*.

[Zhou et al.2016] Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *arXiv preprint arXiv:1606.04199*.