The NiuTrans Machine Translation System for WMT18

Qiang Wang[12], Bei Li[1], Jiqiang Liu[1], Bojian Jiang[1], Zheyang Zhang[1], Yinqiao Li[12], Ye Lin[1], Tong Xiao[12], Jingbo Zhu[12]
[1]Natural Language Processing Lab., Northeastern University
[2]NiuTrans Co., Ltd., Shenyang, China
wangqiangneu@gmail.com, {xiaotong, zhujingbo}@mail.neu.edu.cn
{libeinlp, liujiqiang, jiangbojian}@stumail.neu.edu.cn
{zhangzheyang, liyinqiao, linyeneu}@stumail.neu.edu.cn

Abstract

This paper describes the submission of the *NiuTrans* neural machine translation system for the WMT 2018 Chinese ↔ English news translation tasks. Our baseline systems are based on the Transformer architecture. We further improve the translation performance 2.4-2.6 BLEU points from four aspects, including architectural improvements, diverse ensemble decoding, reranking, and post-processing. Among constrained submissions, we rank 2nd out of 16 submitted systems on Chinese → English task and 3rd out of 16 on English → Chinese task, respectively.

1 Introduction

Neural machine translation (NMT) exploits an encoder-decoder framework to model the whole translation process in an end-to-end fashion, and has achieved state-of-the-art performance in many language pairs (Wu et al., 2016; Sennrich et al., 2016c). This paper describes the submission of the *NiuTrans* neural machine translation system for the WMT 2018 Chinese ↔ English news translation tasks.

Our baseline systems are based on the Transformer model due to the excellent translation performance and fast training thanks to the self-attention mechanism. Then we enhance it with checkpoint ensemble (Sennrich et al., 2016c) that averages the last N checkpoints of a single training run. To enable open-vocabulary translation, all the words are segmented via byte pair encoding (BPE) (Sennrich et al., 2016b) for both Chinese and English. Also, we use back-translation technique (Sennrich et al., 2016a) to leverage the rich monolingual resource.

Beyond the baseline, we achieve further improvement from four aspects, including architectural improvements, diverse ensemble decoding, reranking and post-processing. For architectural improvements, we add relu dropout and attention dropout to improve the generalization ability and increase the inner dimension of feed-forward neural network to enlarge the model capacity (Hassan et al., 2018). We also use the novel Swish activation function (Ramachandran et al., 2018) and self-attention with relative positional representations (Shaw et al., 2018). Next, we explore more diverse ensemble decoding via increasing the number of models and using the models generated by different ways. Furthermore, at most 17 features tuned by MIRA (Chiang et al., 2008) are used to rerank the N-best hypotheses. At last, a post-processing algorithmic is proposed to correct the inconsistent English literals between the source and target sentence.

Through these techniques, we can achieve 2.4-2.6 BLEU points improvement over the baselines. As a result, our systems rank the second out of 16 submitted systems on Chinese → English task and the third out of 16 on English → Chinese task among constrained submissions, respectively.

2 Baseline System

Our systems are based on Transformer (Vaswani et al., 2017) implemented on the Tensor2Tensor [1]. We use base Transformer model as described in (Vaswani et al., 2017): 6 blocks in the encoder and decoder networks respectively (word representations of size 512,

[1]https://github.com/tensorflow/tensor2tensor/tree/v1.0.14. We choose this version because we found that this implementation is more similar to the original model described in (Vaswani et al., 2017) than newer versions.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 528–534

https://doi.org/10.18653/v1/W18-64057

feed-forward layers with inner dimension 2048, 8 attention heads, residual dropout is set to 0.1). We use negative Maximum Likelihood Estimation (MLE) as loss function, and train all the models using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The learning rate is scheduled as described in (Vaswani et al., 2017): $lr = d^{-0.5} \cdot min(t^{-0.5}, t \cdot 4000^{-1.5})$, where d is the dimension of word embedding, t is the training step number. To enable the open-vocabulary translation, we use byte pair encoding (BPE) (Sennrich et al., 2016b) for both Chinese and English. All the models are trained for 15 epochs on one machine with 8 NVIDIA 1080 Ti GPUs. We limit source and target tokens per batch to 4096 per GPU, resulting in approximate 25,000 source and 25,000 target tokens in one training batch. We also use checkpoint ensemble by averaging the last 15 checkpoints, which are saved at 10-minute intervals.

For evaluation, we use beam search with length normalization (Wu et al., 2016). By default, we use beam size of 12, while the coefficient of length normalization is tuned on development set. We use the home-made C++ decoder as a more efficient alternative to the tensorflow implementation, which is also necessary for our diverse ensemble decoding (Section 3.2). The hypotheses that own too many consecutive repeated tokens (e.g. beyond the count of the most frequent token in the source sentence) are removed. We report all experimental results on *newsdev2018* by the official evaluation tool *mteval-v13a.pl*.

3 Improvements

We improve the baseline system from four aspects, including architectural improvements, ensemble decoding, reranking and post-processing.

3.1 Architectural Improvements

Dropout The original Transformer only uses residual dropout when the information flow is added between two adjacent layers/sublayers, while the dropouts in feed-forward neural network (e.g. relu dropout) and self attention weights (e.g. attention dropout) are not in use. In practice, we observed the consistent improvements than baseline when we set relu dropout to 0.1 and attention dropout to 0.1, thanks to the regularization effect to overcome the overfitting.

Larger Feed-Forward Network Limited by the size of GPU memory, we can not directly train a big Transformer model with the batch size as large as the base model. To solve this, we resort to increase the inner dimension (refer to d_{ff}) of feed-forward network while other settings stay the same. It is consistent with the finding of (Hassan et al., 2018) that the transformer model can benefit from larger d_{ff}.

Swish Activation Function The standard Transformer model has a non-linear expression capability due to the use of Rectified Linear Unit (ReLU) activation function. Recently, Ramachandran et al. (2018) propose a new activation function called Swish by the network automatic search techniques based on reinforcement-learning. They claim that Swish tends to work better than ReLU on deeper models and can transfer well to a number of challenging tasks. Formally, Swish is computed as:

$$Swish(x) = x \cdot sigmoid(\beta x),$$

where β is either a constant or a learnable parameter. In practice, we replace ReLU with Swish ($\beta = 1$) and do not change any other settings.

Relative Positional Representation Transformer uses the absolute position encodings based on sinusoids of varying frequency, while Shaw et al. (2018) point out that the representations of relative position can yield consistent improvement over the absolute counterpart. They equip the representations of both key and value with some trainable parameters (e.g. a_{ij}^K, a_{ij}^V in (Shaw et al., 2018)) when calculating the self attention. We re-implement this model, and use clipping distance $k = 16$ with the unique edge representations per layer and head. We use both the absolute and relative positional representations simultaneously.

3.2 Diverse Ensemble Decoding

Ensemble decoding is a widely used technique to boost the performance by integrating the predictions of several models, and has been

Source:	于是就有了这个去年 9 月发布的P@@ ass@@ p@@ ort 。
Translation:	so there is the Pas@@ port , which was released last September .
Post-Processing:	so there is the Passport , which was released last September .
Source:	Furious residents have savaged Sol@@ i@@ hull Council saying it was “ useless at dealing with the problem ”.
Translation:	愤怒的居民猛烈抨击了 S@@ ol@@ i@@ h@@ ou@@ s@@ 委员会, 称它 “ 在处理这个问题上是无用的” 。
Post-Processing:	愤怒的居民猛烈抨击了 Solihull 委员会, 称它 “ 在处理这个问题上是无用的” 。

Table 1: Samples of the inconsistent translation of the constant literal between source and target sentence. The subword is split by “@@”. The two samples are picked up from *newstest2018*.

proved effective in the WMT competitions (Sennrich and Haddow, 2016; Sennrich et al., 2017; Wang et al., 2017). Existing experimental results about ensemble decoding mainly concentrate upon a small number of models (e.g. 4 models (Wang et al., 2017; Sennrich et al., 2016c, 2017)). Besides, the ensembled models generally lack of sufficient diversity, for example, Sennrich et al. (2016c) use the last N checkpoints of a single training run, while Wang et al. (2017) use the same network architecture with different random initializations.

In this paper, we study the effects of more diverse ensemble decoding from two perspectives: the number of models and the diversity of integrated models. We explore at most 15 models for jointly decoding by allocating two models per GPU device in our C++ decoder. In addition to using different random seeds, the ensembled models are generated from more diverse ways, such as different training steps, model sizes and network architectures (see Section 3.1).

Every ensembled model is also assigned a weight to indicate the confidence of prediction. In practice, we simply assign the same weight 1.0 for each model. We also study the greedy tuning strategy (randomly initialize all weights firstly, then fix other weights and only tune one weight each time), while there is no significant improvement observed. [2]

3.3 Reranking

We apply the reranking module to pick up a potentially better hypothesis from the n-best generated by ensemble decoding. The used features for reranking include:

- TFs: Translation features. We totally use eight types of translation features, and each type can be represented as a tuple with four elements: (L_s, D_s, L_t, D_t), where $L_s, L_t \in \{ZH, EN\}$ denotes the language of source and target respectively, and $D_s, D_t \in \{L2R, R2L\}$ denotes the direction of source and target sequence respectively. For example, (ZH, L2R, EN, R2L) denotes a system trained on ordinal Chinese $\rightarrow$ reversed English.
- LM: 5-gram language model of target side [3].
- SM: Sentence similarity. The best hypothesis from the target R2L system is compared to each n-best hypothesis and used to generate a sentence similarity score based on the cosine of the two sentence vectors. The sentence vector is represented by the mean of all word embeddings.

Given the above features, we calculate the ranking score by a simple linear model. All weights are tuned on the development set via MIRA. The hypothesis with the highest ranking score is chosen as the refined translation.

3.4 Post-Processing

Current NMT system generates the translation word by word [4], which is difficult to guarantee the consistency of some constant literals between source sentence and its translation.

In this section, we focus on the English literals in a Chinese sentence. For example, as

[2] We do not use some more sophisticated tuning methods, such as MERT, MIRA, due to the expensive cost for ensemble decoding, especially with a large beam size.

[3] All language models are trained by KenLM (Heafield, 2011).

[4] Actually it is subword by subword in this paper.

Algorithm 1 Post-processing algorithmic for inconsistent English literals translation.

Input: S: source sentence; T: NMT translation;

Output: T': translation after post-processing

Initialize: $T' = T$, create $\mathbb{S}(x, y)$ saves the similarity between x and y
Get the set of English literals $\mathbb{EL}$ from Chinese sentence (either S or T)
for each English literal el in $\mathbb{EL}$ **do**
    if el not in T **then**
        for each y in the set of n-gram of T $(1 \leq n \leq 3)$ **do**
            $\mathbb{S}(el, y) = sim(el, y)$
        end for
    end if
    $y^* = argmax_y \mathbb{S}(el, y)$
    replace el with y^* in T'
end for

shown in Table 3.2, the literal "Passport" in Chinese sentence is translated into "Pasport" wrongly, and a similar error happens between "Solihull" and its translation "Solihous".

To solve this issue, we propose a post-processing method to correct the unmatched translations for the constant literals, as shown in Algorithm 1. The basic idea is that the English literals appearing in Chinese sentence must be contained in English sentence. The challenge is that how to align the correct literal with its wrong one. In practice, we compute the normalized edit distance as the similarity:

$$sim(x, y) = \frac{D(x, y)}{L_x}, \quad (1)$$

where $D(x, y)$ denotes the edit-distance between x and y, L_x is the length of x. Then, the most similar translated literal is recovered by the original one.

Since the number of Chinese sentences containing the English literals is relatively small, our approach can not significantly improve the BLEU, but we find that it is very effective for human evaluation.

4 Experiments and Results

4.1 Chinese → English Results

For Chinese → English task, we use all the CWMT corpus and partial of UN and News-Commentary combined corpus [5]. We also augment the training data by back-translation of the *NewsCraw2017* corpus using the baseline system based on the parallel data only. All texts are segmented by home-made word segmentation toolkit [6]. We remove the parallel sentence pairs which is duplicated, exceptional length ratio, or bad alignment score obtained by fast-align [7]. As a result, we use 7.2M CWMT corpus, 4.2M UN and News-Commentary combined corpus, and 5M pseudo parallel data. Detailed statistical information of training data is shown in Table 2. Then we learn BPE codes with 32k merge operations from independent Chinese and English text, resulting in the size of source and target vocabulary is 47K and 33K respectively. We also study the effect of merge operations, however no significant gain is found when we shrink or expand the number of merge operations.

Table 3 presents the BLEU scores on *newsdev2018* for Chinese → English task. Firstly, we can see that using checkpoint ensemble brings +0.82 BLEU than the baseline of single model. When we equip the Transformer base model with larger d_{ff} and relu & attention dropout, +0.56 BLEU are improved further. However, to our disappointment, we do not observe consistent improvement via Swish or relative positional representations.

Based on the strong single model baseline, we firstly study the conventional ensemble decoding: 4 models with different random seeds, resulting in a significant gain of 0.72 BLEU point. Then we use 4 models with different architectures: *baseline*, $d_{ff} = 4096$, *dropout* and d_{ff}*=4096 + dropout*, then an interesting result is that the diverse ensemble decoding is superior than the ensemble of $d_{ff} + dropout$, which provides an evidence that diverse models may be more important than homogeneous strong models. The beam size of 100 is a bit better than 12. This result is inconsistent with previous work claiming that larger beam size can badly drop down the performance (Tu

[5] We randomly sample 30% data, and found that it can achieve comparable performance with the full data. In this way, we can train more models for our diverse ensemble decoding and reranking.

[6] For Chinese, the word segmentation is done based on unigram language model with Viterbi algorithm.

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

Direction	Lang.	Sentences	Tokens	Ave. sentence length
ZH → EN	ZH	16.5M	391M	23.7
	EN	16.5M	415M	25.2
EN → ZH	EN	16.9M	505M	29.9
	ZH	16.9M	465M	27.5

Table 2: Statistics of the training data

System		beam size	Valid.
Baselines	Transformer-Base	12	25.09
	+checkpoint ensemble	12	25.91
Architectural Improvements	+d_{ff} =4096	12	26.17
	+dropout	12	26.45
Diverse Decoding	4 same models with different random seeds	12	27.21
	4 diverse models	12	27.67
	4 diverse models with large beam	100	27.69
	8 diverse models	100	28.06
	15 diverse models	80	28.18
Re-ranking	14 features	-	28.46
Post-processing	**English literal revised***	-	28.46

Table 3: BLEU scores [%] on *newsdev2018* Chinese-English translation. * denotes the submitted system.

et al., 2017), which needs to be invested further. Additionally, we expand the number of models from 4 to 8 and 15 [8], the overall performances are further improved +0.35 and +0.52 respectively. For 15 models ensemble decoding, we arrange every two models on one GPU via our C++ decoder except the big model which requires one GPU.

Then we rerank the n-best from diverse ensemble decoding (at most 80 candidates) with 14 features [9], we achieve +0.28 BLEU improvement thanks to the complementary information brought by the features. At last, we do post-processing for the reranking output, but almost no effect on BLEU due to limited English literals are found in Chinese sentences.

4.2 English → Chinese Results

For English → Chinese translation, the training data also consists of three parts: CWMT corpus, part of UN and News-Commentary combined data and pseudo parallel data from back-translation. The differences from Chinese → English translation are that the UN and News-Commentary combined data is selected by XenC (Rousseau, 2013) [10] according to the *xmu* Chinese monolingual corpus from CWMT, and *xin_cmn* monolingual corpus is used for back-translation. Data preprocessing is same as Section 4.1, resulting in 7.2M CWMT corpus, 3.5M UN and News-Commentary combined corpus, and 6.2M pseudo parallel data. Then 32k merge operations are used for BPE.

Like Chinese → English, using checkpoint ensemble can bring a gain of +0.62 BLEU solidly. Besides, increasing the dimension of d_{ff} and activate more dropout are proved effective again. The biggest difference from Chinese → English is that diverse ensemble decoding improves the performance at most +1.33 BLEU when we integrate 10 models. However, increasing either the number of models or the diversity is helpful for ensemble decoding. As for reranking, although we only use four (EN, ZH, L2R, R2L) models as features due to time constraint. there is still +0.35 BLEU improvement obtained. At last, post-processing makes an more obvious effect for English → Chinese translation than Chinese → English, because the BLEU4 is computed on characters rather than tokens.

[8]The types of used models include *baseline*, d_{ff}, *dropout*, d_{ff} + *dropout*, *Swish*, *RPR* (relative position representation), *big* (Transformer big model with small batch size) and *baseline-epoch20* (training 20 epochs rather than 15).

[9]Four (ZH, EN, L2R, L2R) models, four (ZH, EN, L2R, R2L) models, one (ZH, EN, R2L, L2R) feature, one (ZH, EN, R2L, R2L) feature, one (EN, ZH, R2L, L2R) feature, one (EN,ZH,R2L,R2L) feature, one LM feature and one SM feature.

[10]https://github.com/antho-rousseau/XenC

System		beam size	Valid.
Baselines	Transformer-Base	12	38.41
	+checkpoint ensemble	12	39.03
Model Variance	+d_{ff}=4096	12	39.48
	+dropout	12	39.61
Diverse Decoding	4 same models with different random seeds	12	40.19
	4 diverse models	12	40.46
	4 diverse models + big beam	50	40.54
	10 diverse models	50	40.94
Re-ranking	4 features	-	41.29
Post-processing	**English literal revised***	-	41.41

Table 4: BLEU scores [%] on *newsdev2018* English → Chinese translation. * denotes the submitted system.

5 Conclusion

This paper presents the *NiuTrans* system to the WMT 2018 Chinese ↔ English news translation tasks. Our single model baseline use the Transformer architecture, and has achieve comparable performance than the last year's best ensembled results. We further improve the baseline's performance from four aspects, including architectural improvements, diverse ensemble decoding, reranking and post-processing. We find that increasing the number of models and the diversity of models is crucial for ensemble decoding. In addition, as the improvement of ensemble decoding, the gain from reranking gradually decreases. Among all the constrained submissions to the Chinese ↔ English news task, our submission is ranked 2nd out of 16 submitted systems on Chinese → English task and the 3rd out of 16 on English → Chinese task, respectively.

Acknowledgments

This work was supported in part by the National Science Foundation of China (No. 61672138 and 61432013), the Fundamental Research Funds for the Central Universities.

References

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In Proceedings of the conference on empirical methods in natural language processing, pages 224–233. Association for Computational Linguistics.

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.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, pages 187–197, Edinburgh, Scotland, United Kingdom.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2018. Searching for activation functions.

Anthony Rousseau. 2013. Xenc: An open-source tool for data selection in natural language processing. The Prague Bulletin of Mathematical Linguistics, (100):73–82.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh' s neural mt systems for wmt17. WMT 2017, page 389.

Rico Sennrich and Barry Haddow. 2016. Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers, chapter Linguistic Input Features Improve Neural Machine Translation. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August

7-12, 2016, Berlin, Germany, Volume 1: Long Papers.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, chapter Edinburgh Neural Machine Translation Systems for WMT 16. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), volume 2, pages 464–468.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In AAAI, pages 3097–3103.

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 6000–6010.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for wmt17. In Proceedings of the Second Conference on Machine Translation, pages 410–415.

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.