CASICT-DCU Neural Machine Translation Systems for WMT17

Jinchao Zhang[1] **Peerachet Porkaew**[1] **Jiawei Hu**[1] **Qiuye Zhao**[1] **Qun Liu**[2,1]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences
[2]ADAPT Centre, School of Computing, Dublin City University
{zhangjinchao,dingchunfa,hujiawei,zhaoqiuye,liuqun}@ict.ac.cn

Abstract

We participated in the WMT 2016 shared news translation task on English ↔ Chinese language pair. Our systems are based on the encoder-decoder neural machine translation model with the attention mechanism. We employ the Gated Recurrent Unit (GRU) with the linear associative connection to build deep encoder and address the unknown words with the dictionary replace approach. The dictionaries are extracted from the parallel training data with unsupervised word alignment method. In the decoding procedure, the translation probabilities of the target word from different models are averagely combined as the ensemble strategy. In this paper, we introduce our systems from data preprocessing to post-editing in details.

1 Introduction

We build the Neural Machine Translation systems CASICT-DCU for WMT17 English ↔ Chinese news translation task. Our systems are based on the encoder-decoder model with the attention mechanism, which is also known as the RNNSearch model (Bahdanau et al., 2015). To construct the deep RNN network, we employ the Gated Recurrent Unit (Cho et al., 2014b) with the linear associative connection (Wang et al., 2017) to ensure the fluent gradient propagation. Adadelta (Zeiler, 2012) algorithm is used to optimize the parameters and stochastic gradient descent algorithm with small learning rate is used in the fine-tuning stage. We extract dictionaries from parallel training data with the unsupervised method to address the unknown words in target translation according to the word alignment vector. During the decoding, the ensemble strategy is used to combine the translation probabilities of the target word from different models.

2 System Description

The neural machine translation model (Kalchbrenner and Blunsom, 2013; Cho et al., 2014b; Sutskever et al., 2014) aims to capture the translation knowledge through training a neural network in the end-to-end style. Our systems are built on the RNNSearch neural machine translation model. Formally, given a source sentence $\mathbf{x} = \mathbf{x}_1, ..., \mathbf{x}_m$ and a target sentence $\mathbf{y} = \mathbf{y}_1, ..., \mathbf{y}_n$, NMT models the translation probability as

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{n} P(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}), \tag{1}$$

where $\mathbf{y}_{<t} = \mathbf{y}_1, ..., \mathbf{y}_{t-1}$. The generation probability of $\mathbf{y}_t$ is

$$P(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}) = g(\mathbf{y}_{t-1}, \mathbf{c}_t, \mathbf{s}_t), \tag{2}$$

where $g(\cdot)$ is a softmax regression function, $\mathbf{y}_{t-1}$ is the newly translated target word and $\mathbf{s}_t$ is the hidden states of decoder which represents the translation status. The attention $\mathbf{c}_t$ denotes the related source words for generating $\mathbf{y}_t$ and is computed as the weighted-sum of source representation $\mathbf{h}$ upon an alignment vector α_t shown in Eq.(3) where the $align(\cdot)$ function is a feedforward network with $softmax$ normalization.

$$\begin{aligned} \mathbf{c}_t &= \sum_{j=1}^{m} \alpha_{t,j} \mathbf{h}_j \\ \alpha_{t,j} &= align(\mathbf{s}_t, \mathbf{h}_j) \end{aligned} \tag{3}$$

The hidden states $\mathbf{s}_t$ are updated as

$$\mathbf{s}_t = f(\mathbf{s}_{t-1}, \mathbf{y}_{t-1}, \mathbf{c}_t), \tag{4}$$

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 428–431
Copenhagen, Denmark, September 711, 2017.

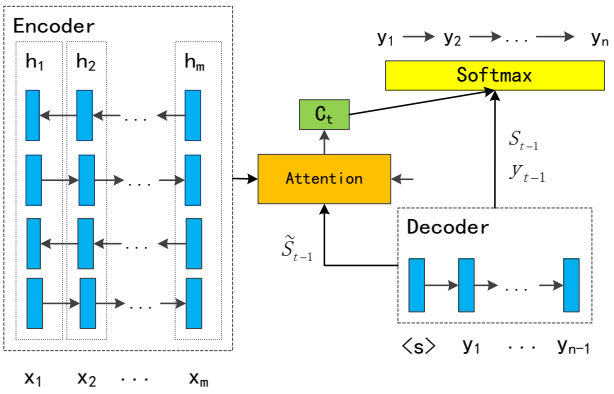

Figure 1: The general architecture of our systems.

where $f(\cdot)$ is a recurrent function.

We adopt a varietal attention mechanism[1] in our system which is implemented as

$$\begin{aligned} \widetilde{\mathbf{s}}_t &= f_1(\mathbf{s}_{t-1}, \mathbf{y}_{t-1}), \\ \alpha_{t,j} &= align(\widetilde{\mathbf{s}}_t, \mathbf{h}_j), \\ \mathbf{s}_t &= f_2(\widetilde{\mathbf{s}}_t, \mathbf{c}_t), \end{aligned} \tag{5}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are recurrent functions.

To construct deep network, we use the linear associative unit (LAU) to ensure fluent gradient propagation. The LAU is computed as

$$\begin{aligned} r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1}), \\ z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1}), \\ g_t &= \sigma(W_{sg}x_t + W_{hg}h_{t-1}), \\ \widetilde{h}_t &= tanh((1 - r_t) \odot W_{xh}x_t + W_{hh}(r_t \odot h_{t-1})), \\ h_t &= ((1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t) \odot (1 - g_t) + g_t \\ &\odot (W_x x_t) \end{aligned} \tag{6}$$

where W_* is the weight matrices, x_t is the input at time t and h_{t-1} is the hidden states at time $t-1$. The LAU allows the input linearly forward propagates in a certain scale to acquire fluent gradient back propagation. It works like residual connections (He et al., 2016) and fast-forward connections (Zhou et al., 2016) and makes build deep network possible. Our encoder is a 4 layers LAU network where forward LAU and backward LAU are alternately stacked. The general architecture of our systems is shown in Figure 1.

3 Pipeline Description

We introduce the pipeline of building the translation systems from data preprocessing to post editing in this section.

3.1 Data Preprocessing

For English ↔ Chinese news translation task, WMT 2017 provides tree parts of data: News Commentary v12, UN Parallel Corpus V1.0 and CWMT Corpus. We used all corpora to train our translation systems. For English sentences, the Moses tokenization script[2] is employed to execute the tokenization processing. For Chinese sentences, we used our in-house word segmentor called "PBCLAS" to do the word segmentation. The word segmentation criterion follows the Chinese People's Daily format. We filter the duplicated sentences and the sentences that are too long (more than 120 words) or too short (less than 5 words). The training corpus is case-sensitive.

3.2 Vocabulary

Our systems are based on the words rather than sub-words (Sennrich et al., 2016; Wu et al., 2016). For our system is serially trained on the single GPU with restricted memory space, the source vocabulary size is set to 100,000 and the target vocabulary size is set to 50,000. The words that out of the vocabulary are represented by the "UNK" symbol.

3.3 Training Details

The sentence length for training systems is up to 120. The word embedding dimension is set to 512 and the hidden layer size is 512. Square matrices are initialized in a random orthogonal way. Non-square matrices are initialized by sampling each element from the Gaussian distribution with mean 0 and variance 0.01^2. All biases are initialized to 0. Parameters are updated by Mini-batch Gradient Descent and the learning rate is controlled by the AdaDelta algorithm with the decay constant $\rho = 0.95$ and the denominator constant $\epsilon = 1e{-}6$. The batch size is 80. We use stochastic gradient descent with small learning rates as 0.0001 to fine-tune the models. Dropout strategy (Srivastava et al., 2014) is applied to the output layer with the dropout rate 0.5 to avoid over-fitting. The gradients of the cost function which have $L2$ norm larger than a predefined threshold 1.0 is normalized to the threshold to avoid gradients explosion (Pascanu et al., 2013). We exploit length normal-

[1] https://github.com/nyu-dl/dl4mt-tutorial/tree/master/session2

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

ization (Cho et al., 2014a) on candidate translations and the beam size for decoding is 12.

3.4 UNK Replace

As the vocabulary sizes are restricted, target sentences may contain "UNK" symbols, which leads to sense ambiguity. We attempt to extract a dictionary to replace the "UNK" symbol in target sentence. We use the "fast_align" [3] word alignment tool to generate the word alignment and extract the dictionary through keeping the highest translation probability. We extract English → Chinese and Chinese → English dictionaries in this way.

At the decoding stage of NMT, we regard the source word that possesses highest alignment probability as the one that generates the target word. Once a "UNK" symbol is generated, we locate the corresponding source word and translate it with the dictionary. If the source word is not in the dictionary, it will be presented in the target sentence.

3.5 Model Ensemble

To add the diversity of systems, we train several models and combine them with the ensemble strategy. These models are initialized with different weight parameters. Each model produces the probability distribution on the target vocabulary at each step of decoding procedure. These probability distributions are averagely combined as the ultimate distribution for beam searching. For our UNK replace strategy, the word alignment vectors that produced by models are also averagely combined to determine the corresponding source word.

4 Experimental Results

4.1 English to Chinese

We ensemble 5 models for English to Chinese translation. The performance of the system on the validation set is presented in Table 1. We figure that the ensemble strategy brings +0.86 BLEU points improvement and the UNK replace approach provide further +1.57 BLEU points.

4.2 Chinese to English

We ensemble 6 models for Chinese to English translation. Table 2 presents the performance of system on the validation set. Same as the English to Chinese translation, the ensemble and UNK replace approaches can enhance the system performance over a single model. The ensemble strategy improves the system by +0.74 BLEU points and the UNK replace approach achieves further +0.51 BLEU point gain. Table 3 shows the performance of our systems on the test set.

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

Model	BLEU
Single Model	25.22
Ensemble 6	$26.08^{+0.86}$
+UNK Replace	$27.65^{+1.57}$

Table 1: The model performances on the validation set in English to Chinese direction.

Model	BLEU-cased
Single Model	18.13
Ensemble 5	$18.87^{+0.74}$
+ UNK Replace	$19.38^{+0.51}$

Table 2: The model performances on the validation set in Chinese to English direction.

Direction	BLEU	BLEU-cased
English → Chinese	30.5	30.5
Chinese → English	23.4	22.3

Table 3: The performance of our systems on the test set.

5 Conclusion

We present CASICT-DCU neural machine translation systems for the WMT17 shared news translation task on English ↔ Chinese language pair. The Gated Recurrent Unit (GRU) with the linear associative connection are employed to build the deep encoder. We extract dictionaries from the parallel training data with unsupervised word alignment approach. We locate the source word that generates the "UNK" symbol in target sentence according to the word alignment vector and translate it with the dictionary. In the decoding procedure, the translation probabilities of the target word from different models are averagely combined as the ensemble strategy to further improve the performance.

Acknowledgments

Qun Liu's work is partially supported by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR2015*.

Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoderdecoder approaches. In *Eighth Workshop on Syntax,Semantics and Structure in Statistical Translation*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014*. Doha, Qatar, pages 1724–1734.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 770–778.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of EMNLP2013*. Seattle, Washington, USA, pages 1700–1709.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)* 28:1310–1318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL2016*. pages 1715–1725.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR* abs/1409.3215.

Mingxuan Wang, Zhengdong Lu, Jie Zhou, and Qun Liu. 2017. Deep neural machine translation with linear associative unit. *CoRR* abs/1705.00861.

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* .

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

Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. In *Proceedings of EMNLP2016*.