DOI: 10.18721/JCSTCS.11205
UDC 004

SEQUENCE-TO-SEQUENCE BASED ENGLISH-CHINESE TRANSLATION MODEL

Tian Zhaolin, Zhang Weiwei

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

In recent years, with the continuous improvement of theory in artificial intelligence, artificial neural networks has become novel tools for machine translation. Compared with traditional Statistical Machine Translation (SMT), neural network based Neural Machine Translation (NMT) transcends SMT in many aspects such as translation accuracy, long distance reordering, syntax, tolerance to noisy data et al. In 2014, with the emergence of sequence-to-sequence (seq2seq) models and attentional mechanisms introduced into the model, NMT was further refined and its performance was getting better and better. This article uses the current popular sequence-to-sequence model to construct a neural machine translation model from English to Chinese. In addition, this paper uses Long-Short Term Memory (LSTM) to replace the traditional RNN in order to solve the problem of gradient disappearance and gradient explosion that it faces in long-distance dependence. The attention mechanism has also been introduced into this article. It allows neural networks to pay more attention to the relevant parts of the input sequences and less to the unrelated parts when performing prediction tasks. In the experimental part, this article uses TensorFlow to build the NMT model described in the article.

Keywords: NMT, seq2seq, LSTM, attention mechanism, encoder-decoder, tensorflow.

Citation: Tian Zhaolin, Zhang Weiwei. Sequence-to-Sequence based english-chinese translation model. St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems, 2018, Vol. 11, No. 2, Pp. 55–63. DOI: 10.18721/JCSTCS.11205

МОДЕЛЬ SEQUENCE-TO-SEQUENCE В АНГЛО-КИТАЙСКОМ ПЕРЕВОДЕ

Тянь Чжаолинь, Чжан Вэйвэй

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

В последние годы, в связи с постоянным совершенствованием теории искусственного интеллекта, новыми инструментами машинного перевода стали искусственные нейронные сети. Нейронный машинный перевод (NMT) имеет значительные преимущества по сравнению с традиционно используемым методом статистического машинного перевода (SMT) в таких аспектах, как точность перевода, изменение порядка слов в длинных предложениях, синтаксис, помехоустойчивость и т. д. После того как в 2014 году появились модели перевода по схеме «последовательность-в-последовательность» (seq2seq) и механизмы внимания, введенные в модель, методы NMT продолжали совершенствоваться, улучшалась их производительность. В данной статье для построения модели нейронного машинного перевода с английского на китайский использована популярная в настоящее время схема перевода seq2seq. Кроме того, вместо традиционно применяемой рекуррентной нейронной сети в статье для решения возникающей проблемы взрыва и исчезновения градиента для длинных строк использован метод долгой краткосрочной памяти (LSTM). Рассмотрен механизм, позволяющий нейронным сетям уделять больше внимания соответствующим частям входных последовательностей и меньше – несвязанным частям при выполнении задач прогнозирования. В экспериментальной части статьи для построения описанной модели NMT использован TensorFlow.

Ключевые слова: NMT, seq2seq, LSTM, механизм внимания, кодировщик-декодер, TensorFlow.

Ссылка при цитировании: Тянь Чжаолинь, Чжан Вэйвэй. Модель Sequence-to-Sequence в англо-китайском переводе // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. 2018. Т. 11. № 2. С. 55–63. DOI: 10.18721/JCSTCS.11205

1. Introduction

Introducing more reforms and implementing the One Belt One Road strategy, China is increasingly participating in international affairs. However, due to the peculiarity of the Chinese language, it is difficult for non-native speakers to master Chinese in a short period of time. In addition, due to the differences in grammatical logic between Chinese and western languages such as English and Russian, traditional statistical machine translation often fails to achieve satisfactory results. As more and more Chinese people travel around the world, and people in other countries are increasingly interested in China, Chinese, the world's most spoken language, and English, the world's most widely used language, produce an inevitable intersection.

Continuous improvement of relevant theories of artificial intelligence in the field of machine translation and the continuous popularization of high-performance hardware devices in the 21st century have paved the way for large-scale application of artificial neural networks in machine translation and created a rare opportunity for further development of neural machine translation. In 2013, Kalchbrenner and Blunsom proposed an end-to-end encoder-decoder model for machine translation. However, the traditional RNN used in the decoder has a problem of gradient disappearance and gradient explosion, making the model difficult to practically handle long-distance dependences. Also in 2013, Gravse et al. applied deep bi-directional LSTM to speech recognition, paving the way for deeper applications of bi-directional LSTM in Neural Machine Translation (NMT). In 2014, Cho et al. proposed a new sequence-to-sequence model and used LSTM (actually, LSTM is a variation of RNN) instead of traditional RNN as encoder and decoder. In the same year, Bengio et al. introduced the attention mechanism into NMT so that the neural network can pay more attention to the relevant part of the input sequences and pay less attention to the unrelated part when performing prediction tasks.

This paper uses a mature seq2seq model to construct a translation model from English to Chinese. The structure of this paper is roughly as follows. The second part introduces data sources and data preprocessing. The third part introduces the theoretical part of the encoder, attention mechanism and decoder in the seq2seq model in detail. The fourth part introduces the experimental results, the approximate implementation of the model, and the evaluation results of the model. The fifth section introduces recommendations for further research in the future.

2. Data Source and Data Preprocessing

The data used in this paper comes from the United Nations Parallel Corpus [1]. The English-Chinese parallel corpus contains almost fifteen million sentences. All these materials contain content which was produced and manually translated from 1990 to 2014, including sentence level alignments.

Before building our seq2seq model, we need to do some preprocessing with the parallel corpus.

- Handling training/testing dataset: We extract 100,000 sentences from the parallel corpus as the testing dataset and the remains are the training dataset.
- Handling source sentences: Add "BOS" in the beginning of sentences and "EOS" in the end of sentences.
- Handling dictionaries: Generate two dictionaries for Chinese and English based on the training dataset.
- Handling unknown words: If some vocabularies/words from the testing dataset do not exist in those two dictionaries, use "UNK" to replace them.
- Handling input sequences: Generate one-hot vectors based on the original sentence and two dictionaries. Then combine these one-hot vectors as input sequences.

3. The Model

Encoder. Assuming the input sequence $\boldsymbol{x} = (x_1, \cdots, x_T)$, the traditional recurrent neural network (RNN) calculates the hidden state vector $\boldsymbol{h} = (h_1, \cdots, h_T)$ and output $\boldsymbol{y} = (y_1, \cdots, y_T)$ by iterating the following equations from $t = 1$ to T:

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \tag{1}$$

$$y_t = W_{hy}h_t + b_y, \tag{2}$$

where W_{xh} denotes the input-hidden weight matrix, W_{hh} denotes the hidden-hidden weight matrix, W_{hy} denotes the hidden-output weight matrix, b_h denotes the hidden bias vector, b_y denotes the output bias vector and $\mathcal{H}$ is the hidden layer function.

However, we found out that the Long-Short Term Memory [2] has its advantages by using a gate mechanism in dealing with long distance dependences. So we use the LSTM cell proposed by Gers et al. in 2002 [3]. So here in our model, $\mathcal{H}$ is implemented by the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \tag{3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \tag{5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \tag{6}$$

$$h_t = o_t \tanh(c_t), \tag{7}$$

where σ denotes the logistic sigmoid function, and i, f, o, c denote the input gate, the forget gate, the output gate and the cell activation vectors respectively, all of which are the same size as the hidden vector h.

A disadvantage of the traditional recurrent neural networks is that they are not able to take advantage of subsequent context. In this paper we use a bidirectional recurrent neural network [4] to process input sequences in both directions by using two separate hidden layers, and then feedforward to a same output layer. The calculation of the forward hidden state $\vec{h}$, the backward hidden state $\overleftarrow{h}$ and the output sequence $\boldsymbol{y}$ is shown below:

$$\vec{h}_t = \mathcal{H}(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}), \tag{8}$$

$$\overleftarrow{h}_t = \mathcal{H}(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}), \tag{9}$$

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y. \tag{10}$$

Here we can also use LSTM, which is mentioned above, to replace the traditional recurrent neural network cell [5, 6]. As a result, bidirectional long-short term memory is the basic structure of the model in this paper.

Furthermore, it has been proved that the performance of a deep neural network is always better than of that with a single layer. In our case, it is totally possible to stack several layers of bidirectional RNN to generate a deep bidirectional RNN [9]. Assuming all the hidden layers are sharing the same function, the calculation of the hidden state in n^th^ layer is shown below:

$$\vec{h}_t^n = \mathcal{H}(W_{\vec{h}^{n-1}\vec{h}^n} + W_{\vec{h}^n\vec{h}^n}\vec{h}_{t-1}^n + b_{\vec{h}}^n), \tag{11}$$

$$\overleftarrow{h}_t^n = \mathcal{H}(W_{\overleftarrow{h}^{n-1}\overleftarrow{h}^n} + W_{\overleftarrow{h}^n\overleftarrow{h}^n}\overleftarrow{h}_{t-1}^n + b_{\overleftarrow{h}}^n). \quad (12)$$

If we define $\vec{h}^0 = \overleftarrow{h}^0 = \boldsymbol{x}$, then the output of network y_t is:

$$y_t = W_{\vec{h}^n y}\vec{h}_t^n + W_{\overleftarrow{h}^n y}\overleftarrow{h}_t^n + b_y. \quad (13)$$

Attention Mechanism. In this paper, we implement a global attention mechanism [7, 8] in our model. Firstly, we take the hidden state h_t at the top layer of deep LSTM and generate a probability distribution based on the context vector c_t to help predict the current target word y_t. The equations are shown below:

$$\tilde{h}_t = \tanh(W_c[c_t; h_t]), \quad (14)$$

$$p(y_t|y < t, x) = softmax(W_s\tilde{h}_t), \quad (15)$$

where $\tilde{h}_t$ denotes attentional vector.

The idea of the global attention mechanism is to consider all hidden states of the encoder when calculating the context vector c_t. In this mechanism, by comparing the current target hidden state h_t with every source hidden state $\bar{h}_s$, we may get a variable-length alignment vector a_t, whose size equals the number of time steps on the source side:

$$a_t(s) = \frac{\exp(score(h_t, \bar{h}_s))}{\sum_{s'}\exp(score(h_t, \bar{h}_{s'}))}. \quad (16)$$

The score can be calculated in three ways, all of which are shown below:

$$score(h_t, \bar{h}_s) = \begin{cases} h_t^T\bar{h}_s & dot \\ h_t^T W_a\bar{h}_s & general \\ W_a[h_t; \bar{h}_s] & concat \end{cases} \quad (17)$$

In our model, we use the general score (the second one in Eq. (17)), which has been proved to be the best one [7], to compute the alignment vector a_t. Consider the alignment vector is weights, we use the weighted average over all the source hidden states to generate the context vector c_t. An example of the attentional vector is shown in Fig. 1 [8].

Decoder. The decoder is trained to predict the next word y_t based on the given context vector c and all the previously predicted words $\{y_1, \cdots, y_{t-1}\}$ [10, 11]. The calculating function is shown below:

$$p(\boldsymbol{y}) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c), \quad (18)$$

where $y = (y_1, \cdots, y_{T_y})$. Here, the conditional probability in the recurrent neural network can be also defined as:

$$p(y_t|\{y_1, \ldots, y_{t-1}\}, c) = g(y_{t-1}, h_t, c), \quad (19)$$

where g denotes a nonlinear function that

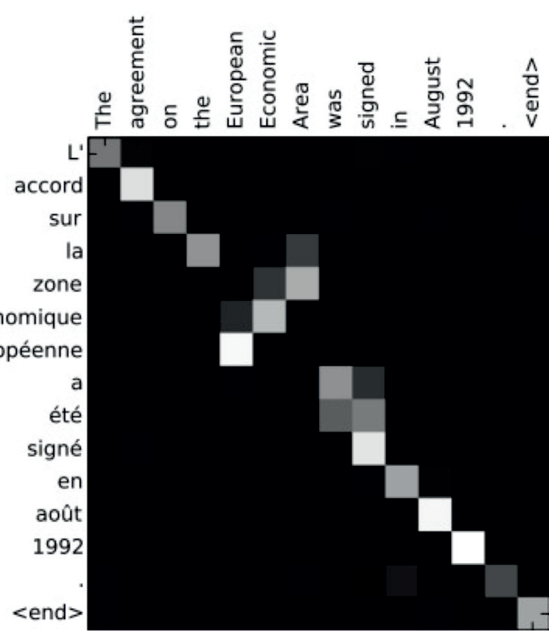

Fig. 1. English-French sample alignments found by RNN search-50. (Bahdanau et al., 2014)

outputs the probability of y_t, h_t denotes the hidden state of recurrent neural network [12].

In our model, every conditional probability described in Eq. (18) is defined as

$$p(y_t|y_1, ..., y_{t-1}, x) = g(y_{t-1}, h_t, c_t), \quad (20)$$

where h_t is computed by the following equation:

$$h_t = f(h_{t-1}, y_{t-1}, c_t). \quad (21)$$

Be advised, the probability here is conditioned on a distinct context vector c_t for each target word y_t. The context vector c_t can be obtained by the method described in the previous section.

4. Experiments

Dependency. First our model is running on Linux. And we need the following tools to be ready.

- Python >= 3.5
- TensorFlow >= 1.2
- Numpy >= 1.12

It is preferable to have a GPU to help speed up the training process [25].

Model Structure. Fig. 2. shows a brief structure of our model. Here in our model, we have a double-layer bidirectional LSTM as an encoder and two-layer LSTM as a decoder. The detailed initial configuration of the encoder and the decoder is shown below.

Encoder

- Hidden state size: 1024
- Number of layers: 2
- Input keep probability: 1.0
- Output keep probability: 1.0

Decoder

- Hidden state size: 1024
- Number of layers: 2
- Input keep probability: 1.0
- Output keep probability: 1.0

Other configurations

- Learning rate: 0.0005
- Batch size: 128
- Beam size: 5
- Size of attentional vector: 512

Model Evaluation and Result Analysis. First, we use cross entropy as the loss of our model. Fig. 3 shows the variation of cross-entropy during the iteration: as the training progresses, the

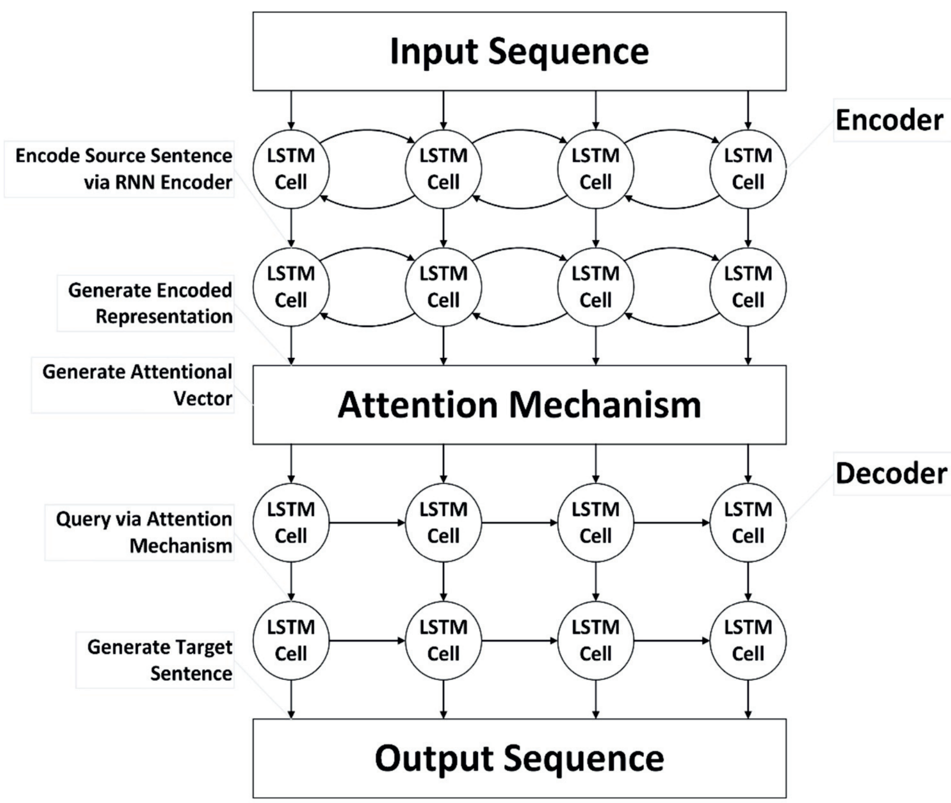

Fig. 2. The structure of the EN-CH model

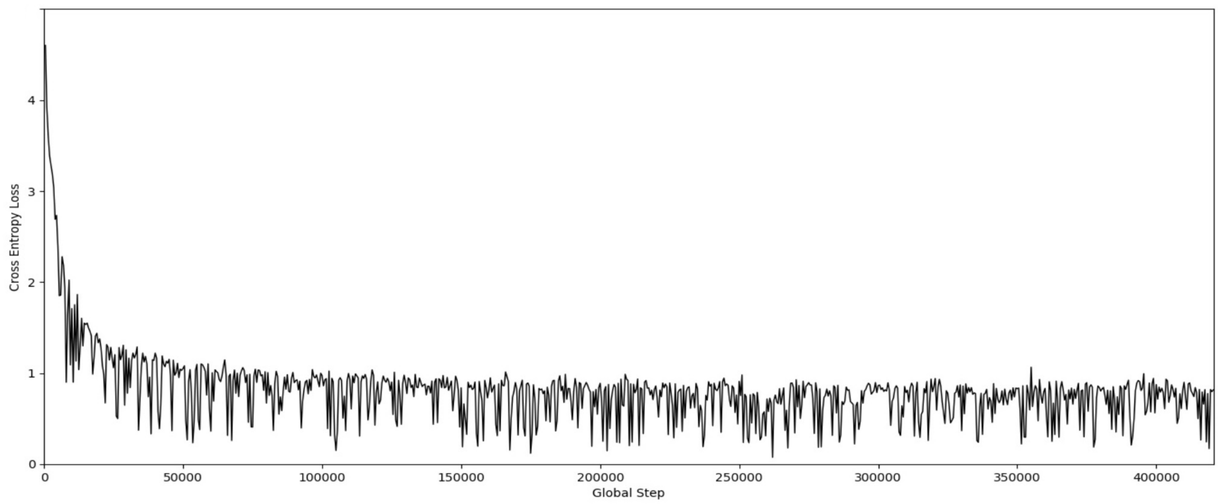

Fig. 3. Cross Entropy Loss

cross entropy decreases gradually and settles at around 0.5 and between 0.1 and 0.9, showing that the model is still quite effective.

We also notice that the losses vary in a relatively large range after the model was trained over 50,000 steps. That's because the model performs better with shorter sentences then with longer ones. When input sentences are short, the translations by the model are exactly the same with the standard translation most of the time, and, therefore, the losses in this moment may be close to zero. When input sentences are long, the translations by the model may not be accurate, but still acceptable, or, compared to the standard translations, express the idea in another way. All of this is the reason why the losses under such circumstances are relatively larger than those for short sentences.

Table 1 shows some examples of short sentences.

Table 2 shows some examples of long sentences.

Secondly, we use a BLEU score, a kind of self-evaluation method for a machine translation model, to evaluate our model. The BLEU score is computed based on the test dataset. BLEU score is computed by the following equations [13]:

Table 1

Example of short sentences

Step/Loss	Translations by model	Standard Translations
229500/ 0.360127	纽约办事处	纽约办事处
	(议程项目5)	(议程项目5)
	联合国发展援助框架	联合国发展援助框架
277000/ 0.335871	秘书长办公厅	秘书长办公厅
	占总数的百分比	总量的百分比
	秘书处的说明	秘书处的说明
283500/0.315215	议程项目40	议程项目40
	给秘书长的信	给秘书长的信
	其他亚洲国家	亚洲其他国家

Table 2

Examples of long sentences

Step/Loss	Translations by model	Standard Translations
276500/0.715091	委员会关切地注意到 ， 现有住房方案没有充分解决穷人的住房需要 。	委员会关切地注意到， 现行住房方案没有充分满足贫困者的住房要求 。
	有人提醒说 ， 对某些发展中国家来说， 缓解行动可能集中在这个部门	缔约方提请注意， 对某些发展中国家来说， 缓解行动很可能会集中于这一部门 。
	在同次会议上 ， 理事会决定将这个项目交由理事会主席协商 。	在同次会议上， 理事会决定将这一项目提交理事会主席举行的磋商会议 。
277500/0.894594	在生物多样性高的地区建立基于市场的可持续发展 ， 通过可持续利用自然资源解决农村创收发展问题 ；	在高度生物多样化区域， 以基于市场的方法进行可持续的社区发展， 并通过可持续地使用自然资源解决农村创收问题 ；
	由于外地办事处报告的数据列于财务报表附注中 ， 儿童基金会总部必须合理保证非消耗性财产数字准确 。	由于外地办事处报告的这些数字都要在财务报表附注中披露， 儿童基金会总部必须合理地保证非消耗性财产数字的准确性 。
	委员会面前的一个关键问题是发展筹资 ， 2001年在召开一次关于这一问题的高级别政府间会议上已经开展了筹备工作 。	第二委员会所要审议的一个基本问题是为发展筹措资金 ； 将于2001年专门讨论这个问题的政府间高层会议正在准备之中 。

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-r/c} & if\ c \le r \end{cases} \tag{22}$$

$$Bleu = BP \cdot \exp\left(\sum_{n=1}^{N} \omega_n \log p_n\right), \tag{23}$$

$$\log Bleu = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} \omega_n \log p_n, \tag{24}$$

where BP denotes the Brevity Penalty, c denotes the length of the output sentence, r denotes the length of the standard translation sentence, N equals 4 and ω_n equals 0.25.

In our model, we got a 26.6 bleu score for the English-Chinese task.

5. Further Development of NMT

In the rapidly developing and highly competitive environment, the NMT technology is making significant progress. NMT will also be continuously improved in many aspects, including:

- Rare word problem [14, 15]
- Use of single-language data [16, 17]
- Multilingual Translation / Multilingual NMT [18]
- Memory mechanism [19]
- Language fusion [20]
- Coverage issues [21]
- Training process [22]
- A priori knowledge fusion [23]
- Multi-modal translation [24]

Acknowledgement

We would like to thank Professor E.A. Rodionova for her great help. We would also like to express our gratitude to all the people who have helped and supported us to finish this article.

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

1. **Ziemski M., Junczys-Dowmunt M., Pouliquen B.** Parallel Corpus United Nations. *Language Resources and Evaluation (LREC'16)*, Portorozh, Sloveniya, 2016.

2. **Hochreiter S., Schmidhuber J.** Long Short Term Memory. *Neural Computation*, 1997, Vol. 9, No. 8, Pp. 1735–1780.

3. **Gers F.A., Schraudolph N.N, Schmidhuber J.** Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, 2002, Vol. 3, Pp. 115–143.

4. **Schuster M., Paliwal K.K.** Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997, Vol. 45, Pp. 2673–2681.

5. **Graves A., Mohamed Abdel-Rahman, Hinton G.** Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, Pp. 6645–6649, preprint arXiv: 1303.5778 03.2013 (https://arxiv.org/pdf/1303.5778.pdf).

6. **Graves A., Schmidhuber J.** Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 2005, Vol. 18, Issue 5-6, Pp. 602–610.

7. **Minh-Thang Luong, Hieu Pham, Christopher D.** Manning Bilingual word representations with monolingual quality in mind. *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, Pp. 151–159, preprint arXiv: 1508.04025 08.2015 (http://www.aclweb.org/anthology/W15-1521).

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

9. **Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Yoshua Bengio**. *How to construct deep recurrent neural networks*, preprint arXiv: 1312.6026, 12.2013.

10. **Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio.** *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, preprint arXiv: 1406.1078, 2014.

11. **Sutskever I., Vinyals O., Le. Quoc V.** Sequence to sequence learning with neural networks. *Proceedings of the Advances in Neural Information Processing Systems*, 2014, Pp. 3104–3112, preprint arXiv: 1409.3215S.

12. **Kalchbrenner N., Blunsom Ph.** Recurrent continuous translation models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, Pp. 1700–1709.

13. **Papineni K., Roukos S., Ward T., Wei-Jing Zhu.** BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 2002, Pp. 311–318.

14. **Jean S., Kyunghyun Cho, Memisevic R., Bengio Y.** On using very large target vocabulary for neural machine translation. *Conference ACL-2015*, preprint arXiv: 1412.2007, 2014.

15. **Minh-Thang Luong, Sutskever I., Quoc V. Le, Vinyals O., Wojciech Zaremba**. *Addressing the rare word problem in neural machine translation*, preprint arXiv: 1410.8206, 2014.

16. **Sennrich R., Haddow B., Birch A.** Improving neural machine translation models with Monolingual Data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, preprint arXiv: 1511.06709, 2015

17. **Cheng Y., Xu W., He Z., He W., Wu H., Sun M., Liu Y.** *Semi-supervised learning for neural machine translation*, preprint arXiv: 1606.04596, 2016.

18. **Dong D., Wu H., He W., Yu D., Wang H.** Multi-task learning for multiple language translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, Vol. 1, Pp. 1723–1732.

19. **Wang M., Lu Z., Li H., Liu Q.** *Memory-enhanced decoder for neural machine translation*, preprint arXiv: 1606.02003, 2016.

20. **Sennrich R., Haddow B.** Linguistic input features improve neural machine translation. *Proceedings of the First Conference on Machine Translation*, 2016, Pp.83–91, preprint arXiv: 1606.02892, 2016.

21. **Tu Z., Lu Z., Liu Y., Liu X., Li H.** Modeling coverage for neural machine translation. *ACL Conference*, preprint arXiv: 1601.04811, 2016.

22. **Shen S., Cheng Y., He Z., He W., Wu H., Sun M., Liu Y.** *Minimum risk training for neural machine translation*, preprint arXiv: 1512.02433, 2015.

23. **Cohn T., Duy C., Hoang V., Vymolova E., Yao K., Dyer C., Haffari G.** *Incorporating structural alignment biases into an attentional neural translation model*, preprint arXiv: 1601.01085, 2016.

24. **Hitschler J., Schamoni S., Riezler S.** Multimodal Pivots for Image Caption Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, Pp. 2399–2409.

25. **Britz D., Goldie A., Minh-Thang Luong, Quoc Le.** *Massive exploration of neural machine translation architectures*, preprint arXiv:1703.03906, 2017.

Received 08.05.2018. / Статья поступила в редакцию 08.05.2018.

THE AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

TIAN Zhaolin
ТЯНЬ Чжаолинь
E-mail: peter0431peter@gmail.com

ZHANG Weiwei
ЧЖАН Вэйвэй
E-mail: soszhang@outlook.com