Bilingual Correlation Optimization Model for Chinese-Uyghur Machine Translation (Post-print)

Authors: Yirong Pan, Li Xiao, Yang Yating, Dong Rui

Date: 2019-01-03T00:00:00+00:00

Abstract

To address the semantic irrelevance issue in statistical machine translation systems for Chinese-Uyghur language pairs, we propose a bilingual association optimization model based on neural machine translation methods. This model leverages attention mechanisms to capture word alignment information, incorporates semantic correlation and internal lexical matching degree between bilingual phrases, and predicts the generation probability of bilingual phrases as the bilingual association degree to optimize phrase translation scores in statistical translation models. Experimental results on the Chinese-Uyghur open machine translation dataset from the 11th China Workshop on Machine Translation (CWMT 2015) demonstrate that, compared with the baseline system and under conditions of using smaller-scale training data and vocabulary, the proposed method can effectively improve the performance of both phrase-level and sentence-level machine translation tasks simultaneously, achieving maximum BLEU score improvements of 2.49 and 0.59, respectively.

Full Text

Preamble

Vol. 37 No. 3
Application Research of Computers
ChinaXiv Cooperative Journal

Bilingual Relatedness Optimization Model for Chinese-Uyghur Machine Translation

Pan Yirong[1,2,3], Li Xiao[1,3], Yang Yating[1,3], Dong Rui[1,3]
(1. Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. Xinjiang Laboratory of Minority Speech & Language Information Processing, Urumqi 830011, China)

Abstract: Addressing the issue of semantic independence in Chinese-Uyghur statistical machine translation (SMT) systems, this paper proposes a bilingual relatedness optimization model based on neural machine translation methods. The model employs an attention mechanism to capture word alignment information while incorporating semantic relevance between bilingual phrases and internal lexical matching degree to predict the generation probability of bilingual phrases, which is then used as bilingual relatedness to optimize phrase translation scores in the statistical translation model. Experimental results on the 11th China Workshop on Machine Translation (CWMT 2015) Chinese-Uyghur public machine translation datasets demonstrate that, compared with the baseline system and under conditions of smaller-scale training data and vocabulary, the proposed method can effectively improve performance in both phrase-level and sentence-level machine translation tasks, achieving maximum BLEU score improvements of 2.49 and 0.59 points, respectively.

Key words: Uyghur; neural machine translation; attention mechanism; word alignment; generation probability

0 Introduction

In phrase-based statistical machine translation (SMT) systems [**?**], translation models model bilingual phrases extracted from parallel corpora, primarily including parameters such as phrase translation probability and lexical weight. These parameters serve as feature functions combined with log-linear methods to train machine translation systems, thereby obtaining optimal weight distributions that enable the decoder to search for the most probable translation options during bilingual conversion. Although various machine translation methods have made tremendous progress in recent years and translation quality continues to improve, issues such as lexical translation errors and semantically irrelevant content in translation results remain to be addressed.

SMT constructs phrase translation scores based on statistical methods that only consider co-occurrence frequencies of bilingual phrases, largely ignoring semantic relevance. Additionally, since word alignment results originate from statistical alignment models [**?**] and are learned using maximum likelihood estimation (MLE), they suffer from missing, redundant, and erroneous information. This leads to data sparsity issues when evaluating lexical weights in statistical translation models, reducing accuracy and consequently affecting machine translation quality. To address these problems, this paper applies neural machine translation (NMT) methods based on attention mechanisms to capture word alignment information, introduces semantic relevance and internal lexical matching degree of bilingual phrases, and predicts bilingual relatedness scores to optimize phrase

translation probabilities in statistical translation models, demonstrating their effectiveness through experiments.

1 Related Work

As a low-resource language with complex morphological structures, Uyghur has seen research on Chinese-Uyghur statistical machine translation focusing on two main aspects. The first involves parallel corpus construction. For instance, Peng et al. [**?**] utilized spatial vector representations of Chinese-Uyghur bilingual sentences to extract parallel corpora based on similarity between source and target sentences, ensuring semantic relevance of parallel sentence pairs. The second aspect concerns Uyghur grammar and morphological analysis. For example, Miliwan et al. [**?**] treated Uyghur word stems and affixes as basic translation units, proposing a stem-affix language model based on directed graphs that leverages Uyghur' s agglutinative characteristics for machine translation experiments. Furthermore, Pan et al. [**?**] employed deep learning techniques to analyze semantic features of Chinese-Uyghur phrases, using recurrent neural networks (RNN) to learn reordering information and reconstruct reordering models, thereby assigning more reasonable reordering directions and probability distributions to reordering rules.

These methods primarily model monolingual linguistic properties (Uyghur stems, affixes, morphology, etc.) and external bilingual correspondences (Chinese-Uyghur parallel sentence pairs, reordering directions, etc.), lacking research and analysis on semantic relevance and internal lexical matching degree of bilingual aligned phrases. Consequently, semantic independence issues persist in Chinese-Uyghur statistical translation models. Since phrase translation probability distributions are not entirely reasonable, they fail to correctly assess bilingual phrase relatedness at both semantic and lexical levels, leaving room for improvement in statistical machine translation research.

With the application and development of deep learning techniques in SMT, numerous studies have utilized neural network methods to improve statistical translation models with promising results. Schwenk [**?**] proposed a continuous space translation model for phrases that employs vector representations to predict phrase translation probabilities. Son et al. [**?**] introduced a hierarchical neural network translation model that jointly evaluates continuous space vector representations and related parameters of translation units. Zou et al. [**?**] proposed a bilingual word vector representation model to compute semantic similarity between words, incorporating these scores as additional features during system training. Cho et al. [**?**] presented an encoder-decoder based phrase representation model that uses RNNs to maximize the conditional probability of aligned phrases and evaluate generation probability scores for bilingual phrases in translation models.

Building upon Cho et al. (2014), this paper re-evaluates phrase translation probabilities in statistical translation models and proposes a neural-based bilingual

relatedness optimization model (NBROM). Unlike previous approaches, our model first utilizes the Bahdanau et al. (2014) framework [**?**] with attention mechanisms to capture word alignment information between bilingual phrases, incorporating phrase semantic relevance and internal lexical matching degree to optimize phrase translation probabilities. Second, when training this model, we address the out-of-vocabulary (OOV) problem in Uyghur using three models for OOV generation probability prediction: the Unk model, MultiClass model, and byte pair encoding (BPE) model [**?**], assigning appropriate weights to different OOV words. Experimental results on the CWMT 2015 Chinese-Uyghur dataset demonstrate that under conditions of relatively small-scale training data and vocabulary, our proposed method simultaneously improves performance in both phrase-level and sentence-level machine translation tasks, achieving maximum BLEU improvements of 2.49 and 0.59 points respectively, thereby validating the effectiveness of our approach.

2 Statistical Machine Translation System

Given a source language sentence f, SMT aims to find the optimal target language translation e that maximizes the conditional probability, as shown in Equation (1):

$$e^* = \arg\max_e p(e|f)$$

where $p(e|f)$ represents the translation model and $p(e)$ denotes the language model. Generally, SMT incorporates multiple feature functions with their corresponding weights into a log-linear framework for modeling, as shown in Equation (2):

$$p(e|f) \propto \exp\left(\sum_{n=1}^{N} w_n \log \phi_n(e,f)\right) = \frac{1}{Z_f} \exp\left(\sum_{n=1}^{N} w_n \phi_n(e,f)\right)$$

where ϕ_n and w_n represent the n-th feature function and its corresponding weight, respectively, and Z_f is the normalization constant. Based on this framework, the translation model factorizes into a weighted sum of feature functions. SMT optimizes weight parameters w_n on a development set to maximize translation performance metrics (BLEU score [**?**]) and uses these parameters during decoding to search for optimal candidate phrase translations.

Statistical translation models primarily model phrase translation probabilities and lexical weights. Phrase translation probability is calculated statistically, as shown in Equation (3):

$$\phi_{\text{phrase}}(f,e) = \frac{\text{count}(f,e)}{\sum_{e'} \text{count}(f,e')}$$

where f and e denote source and target language aligned phrases, respectively, and count(f, e) represents their co-occurrence frequency in large-scale parallel sentence pairs. The translation model uses this value as the translation probability for the aligned phrase.

Lexical weight score uses word alignment results as a baseline, dividing source and target language phrases into lexical units to evaluate matching degree between bilingual words, as shown in Equation (4):

$$\phi_{\text{lex}}(f, e) = \prod_{i=1}^{|e|} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall j \in \{j|(i,j) \in a\}} p(e_i|f_j)$$

where a represents word alignment, e_i denotes the i-th word in the target phrase, and $p(e_i|f_j)$ represents the co-occurrence frequency of word pair (e_i, f_j) in large-scale parallel sentence pairs. For each word in the target phrase e, the translation model traverses all words in sequence and multiplies their probability values, using this product as the lexical weight score.

Word alignment is obtained using the GIZA++ [**?**] tool, which is based on the EM algorithm and evaluates statistical alignment probabilities of word pairs, considering bilingual words with maximum probability as aligned in parallel corpora. Since both phrase translation probability and lexical weight are obtained through statistical methods, translation models suffer from semantic independence issues.

Figure 1 [Figure 1: see original paper] illustrates an example of Chinese-Uyghur bilingual aligned phrases (Uyghur is written right-to-left). For the source phrase "为人民群众服务" (serve the masses), the statistical translation model retains the aligned target phrase " " extracted from training corpora and assigns corresponding word alignment information. As shown, the source word "服务" (serve) aligns with two target words " " and " ", yet only " " meets semantic alignment requirements while " " contains erroneous alignment information. Meanwhile, words "为" and "人民" lack aligned target words. Since lexical weight score is the product of translation probabilities for aligned words, missing, redundant, or erroneous alignment information prevents lexical weights from correctly evaluating matching degree between words in bilingual phrases, thereby reducing statistical translation model accuracy and SMT system performance.

3 Bilingual Relatedness Optimization Model

3.1 Model Overview

Based on deep learning techniques and NMT methods, our model first uses an encoder to convert source language phrases into fixed-dimensional feature vectors, then employs a decoder to transform these vectors into variable-length target phrases while introducing an attention mechanism to capture semantic

information and aligned words in bilingual phrases, compensating for semantic independence and word alignment errors in statistical translation models.

The encoder is a bidirectional recurrent neural network (BiRNN) [**?**] that processes input sequences in both directions, updating forward hidden states $\vec{h}_t$ and backward hidden states $\overleftarrow{h}_t$ as shown in Equations (5) and (6):

$$\vec{h}_t = \tanh(W_{xh}x_t + W_{hh}\vec{h}_{t-1} + b_h) \quad \text{(forward traversal from } x_1 \text{ to } x_t\text{)}$$

$$\overleftarrow{h}_t = \tanh(W_{xh}x_t + W_{hh}\overleftarrow{h}_{t+1} + b_h) \quad \text{(backward traversal from } x_t \text{ to } x_1\text{)}$$

where tanh is the nonlinear activation function. For each word x_k in the source phrase $[x_1, x_2, ..., x_T]$, we annotate it with surrounding context using $h_k = [\vec{h}_k; \overleftarrow{h}_k]$.

The decoder is a single-layer RNN that, given the previous predicted word y_{i-1}, current RNN hidden state s_i, and context vector c_i, predicts the current output word y_i, where s_i and c_i depend on s_{i-1} and y_{i-1}, as shown in Equation (7):

$$p(y_i|y_{i-1}, ..., y_1, x) = g(y_{i-1}, s_i, c_i)$$

where g is the nonlinear softmax activation function, c_i is the weighted sum of h_j, and α_{ij} represents alignment weights that evaluate matching degree between word e_j in the source phrase and word y_i, as shown in Equation (8):

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(a(s_{i-1}, h_j))}{\sum_{k=1}^{T} \exp(a(s_{i-1}, h_k))}$$

where a is the alignment model and W_a and U_a are neural network parameters:

$$a(s_{i-1}, h_j) = v_a^T \tanh(W_a s_{i-1} + U_a h_j)$$

By training the encoder and decoder to maximize the log conditional probability score, the model predicts the most likely target phrase, as shown in Equation (9):

$$\theta^* = \arg\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y^{(n)}|x^{(n)})$$

where θ represents the model parameter set and $(x^{(n)}, y^{(n)})$ are aligned phrase pairs in the training dataset. Figure 2 [Figure 2: see original paper] illustrates our model framework using Chinese-Uyghur bilingual phrase generation as an example. The model maps source language phrases into continuous space vector representations to capture semantic information while using attention mechanisms to assign different weights to each word in the source phrase, representing its importance. This model can predict the most probable target language phrase given a source phrase and estimate bilingual relatedness scores to re-evaluate lexical weights in statistical translation models, thereby improving both phrase-level and sentence-level machine translation quality.

3.2 Target Phrase Prediction

Using the above model, we can predict target phrase e given source language phrase f to maximize the score in Equation (9), as shown in Equation (10):

$$e^* = \arg\max_e \log p(e|f) = \arg\max_e \sum_{t=1}^{I} \log p(e_t|e_{<t}, f)$$

where e_t is the t-th word in the predicted phrase and I is the number of words in e.

3.3 Bilingual Relatedness Score

The model can also re-evaluate lexical weight values in statistical translation models by considering semantic relevance and internal lexical matching degree of bilingual phrases to predict reasonable bilingual relatedness scores. Given source phrase f, the relatedness score for predicting target phrase e is calculated as shown in Equation (11):

$$\text{Rescore}(f, e) = \log p(e|f) = \sum_{t=1}^{I} \log p(e_t|e_{<t}, f)$$

where e_t is the t-th word in target phrase e and I is the number of words in e.

3.4 Out-of-Vocabulary Handling Strategies

When training NMT systems, only the top K most frequent target words are retained due to time and space complexity considerations, with K typically ranging between 30k (Bahdanau et al., 2015) and 80k (Sutskever et al., 2014) [**?**]. This prevents effective prediction of low-frequency rare words, thereby reducing NMT system translation quality. To address this issue, we employ three models for OOV generation probability prediction, validating their respective effectiveness and limitations through experiments. For all OOV words in source phrases, we uniformly annotate them with the [UNK] symbol and set source vocabulary size to 50k and target vocabulary size to 30k-50k.

3.4.1 Unk Model Following Sutskever et al. (2014), Cho et al. (2014), and Bahdanau et al. (2015), this model uniformly annotates all OOV words with the [UNK] symbol, a very common OOV handling strategy in NMT system training where all OOVs are assigned identical weights.

3.4.2 MultiClass Model Following Jean et al. (2015) [**?**], this model classifies all OOV words into lexical categories, where OOV words in the same category share identical weights. For each OOV word, the model decomposes its generation probability prediction into the product of category probability score and internal lexical probability score, thereby reducing training complexity. We use four lexical categories in our experiments: numbers [NUM], symbols [SYM], named entities [NOUN], and other words [UNK]. Vocabulary category identification is modeled as a classification problem, training a classifier to categorize OOV words, with detailed implementation described in Section 4.2.

3.4.3 BPE Model Following Sennrich et al. (2015) [**?**], this model processes all OOV words using Uyghur subword units representation, segmenting low-frequency words to increase co-occurrence counts of subwords in sparse words and effectively solve data sparsity issues. We use the open-source segmentation tool subword-nmt (https://github.com/rsennrich/subword-nmt) to process OOV words, and following Halidanmu et al. [**?**], we do not perform morphological segmentation on Uyghur.

4 Experiments

4.1 Experimental Setup

We conduct experiments on a Chinese-Uyghur statistical machine translation system using the open-source translation platform Moses (http://www.statmt.org/moses/) as our baseline. Training data originates from the CWMT 2015 publicly available Chinese-Uyghur news domain corpus, containing 110k parallel sentence pairs with 64,851 Chinese words and 104,992 Uyghur words. Development and test sets use in-domain data containing 1,095 and 1,000 parallel sentence pairs, respectively. We train a 5-gram language model using SRILM [**?**] on the training data, perform Chinese word segmentation using the Stanford segmenter (https://nlp.stanford.edu/software/segmenter.html), apply the grow-diag-final-and alignment strategy, set maximum phrase extraction length to 8, and extract 4,934,572 bilingual aligned phrases. We tune the SMT system using MERT [**?**] and evaluate machine translation performance using case-insensitive BLEU scores.

We train our model using bilingual aligned phrases extracted from the SMT system. First, considering semantic matching degree of bilingual phrases, we retain the top 1 million phrase pairs with highest phrase translation probabilities. Then, we select bilingual phrases where each word in the target phrase has at least one aligned word in the source phrase, ensuring internal lexical matching

degree to some extent. Finally, for identical source phrases, we retain longer target phrases to improve model adaptability to long phrases. After these steps, approximately 400k aligned phrase pairs are selected as training data for our proposed model.

4.2 Model Training

For the vocabulary category classifier, we use an RNN network structure with hidden layer units of 256-128-64-16-4, training for 200 epochs with an initial learning rate of 0.1, using stochastic gradient descent (SGD) to update network parameters for loss function minimization, and adding a softmax activation function at the output layer to produce probabilities for the four word classes. Upon training completion, optimized parameters predict OOV word classes, replacing each OOV word with its most probable category.

The encoder in our model consists of forward and backward RNNs, each containing 100 hidden units. The decoder contains 100 hidden units using a neural network structure with maxout hidden layers [**?**]. We train the model using mini-batch SGD combined with the Adadelta learning rate update method [**?**], setting batch size to 100 and training for 500 epochs. Upon training completion, optimized network parameters evaluate bilingual relatedness scores and predict the most probable target phrases. Following Schwenk (2012), we replace all lexical weights in the statistical translation model with bilingual relatedness scores.

4.3 Experimental Results

We compare experimental performance between the statistical machine translation system and NBROM for sentence-level machine translation tasks, considering OOV handling strategies, training data scale, and target vocabulary size. Results are shown in Table 1 .

Table 1 Experimental performance comparison of machine translation tasks

Training Data Scale	Vocabulary Size	BLEU Score
Moses (baseline)	-	38.16
NBROM + Unk	50k / 30k	38.65 (+0.49)
NBROM + MultiClass	50k / 30k	38.68 (+0.52)
NBROM + BPE	50k / 30k	38.75 (+0.59)

Table 1 shows that re-evaluating lexical weights in statistical translation models using NBROM significantly improves machine translation performance. Under training data scale of 50k and vocabulary size of 30k, BLEU scores increase by 0.49-0.59 points, validating our model' s effectiveness. NBROM combined with BPE achieves the highest BLEU score of 38.75 in our experiments. This method segments all low-frequency OOV words into subword units, increasing

co-occurrence counts of subwords in sparse words and mitigating OOV impact, effectively predicting generation probabilities of unseen words and achieving a 0.59-point improvement over the baseline system. For NBROM combined with Unk and MultiClass models, the latter slightly outperforms the former because Unk model uniformly replaces all OOV words with [UNK] symbols, assigning identical weights, whereas MultiClass model incorporates word class information during prediction, further improving OOV prediction accuracy. Experimental results demonstrate that our bilingual relatedness optimization model, leveraging semantic relevance and internal lexical matching degree of bilingual phrases, can effectively improve Chinese-to-Uyghur machine translation performance using small-scale training data and vocabulary.

Table 2 shows experimental performance comparison of BPE model in machine translation tasks.

Table 2 Experimental performance of BPE model in machine translation tasks

Training Data Scale	Vocabulary Size	BLEU Score
30k	20k	38.45
50k	30k	38.75
100k	50k	38.52

We also compare NBROM combined with BPE model performance, setting training data scale proportional to vocabulary size. Table 2 indicates that BPE model performs optimally with smaller training data and vocabulary scales, with performance decreasing as these scales expand. This may be because Uyghur' s complex morphological information and large vocabulary require BPE to segment low-frequency words into subword units added to the target vocabulary, which affects semantic information integrity and increases model training complexity. This introduces data sparsity issues when predicting OOV word generation probabilities, diminishing machine translation improvement effects.

4.4 Target Language Phrase Prediction

We compare experimental performance between statistical machine translation systems and NBROM for phrase-level machine translation tasks. Test data consists of 2,000 randomly extracted bilingual aligned phrases from our model' s training data, with statistics on Uyghur vocabulary size and average words per phrase shown in Table 3 .

Table 3 Experimental performance of phrase generation tasks

System	Vocabulary Size	Avg. Words/Phrase	BLEU Score
Moses	1,517	4.2	91.27

System	Vocabulary Size	Avg. Words/Phrase	BLEU Score
NBROM + Unk	1,517	4.2	93.56 (+2.29)
NBROM + MultiClass	1,517	4.2	93.76 (+2.49)
NBROM + BPE	1,517	4.2	92.13 (+0.86)

Table 3 demonstrates that NBROM significantly outperforms statistical machine translation systems in predicting target phrases corresponding to source phrases, validating the method' s effectiveness in phrase-level machine translation tasks. Among the three OOV handling strategies, MultiClass model achieves the best performance with a 2.49-point BLEU improvement over the baseline. NBROM combined with BPE shows less pronounced effects, possibly because BPE segmentation introduces excessive subword units during target word generation, reducing prediction accuracy and completeness. Additionally, MultiClass model uses vocabulary category information during training, further improving OOV prediction accuracy compared to Unk model.

As described above, NBROM combined with MultiClass model can predict the most probable aligned target phrase with highest matching degree to the source phrase and re-evaluate lexical weights to assign more reasonable relatedness scores. Table 4 presents examples of target language phrase prediction from statistical machine translation systems and NBROM, where Score represents lexical weight in statistical translation models and Rescore represents bilingual relatedness scores from NBROM.

Table 4 Example of Uyghur phrase prediction

Source Phrase	Moses Output	Score	NBROM + MultiClass Output	Rescore
金融等领域		2.0e-15		4.8e-11
教育、文化		-		-
活动今天启动		-		-

Statistical machine translation systems retain multiple target phrases for identical source phrases with corresponding lexical weight scores. Table 4 shows that under conditions of high semantic content and lexical matching degree, statistical machine translation systems produce small lexical weight scores that

fail to correctly assess bilingual phrase alignment probabilities, contradicting actual conditions and reducing translation model quality. In contrast, NBROM' s attention mechanism effectively captures aligned words in bilingual phrases, enabling reasonable prediction of target phrases with semantic relevance and lexical matching degree while assigning appropriate bilingual relatedness scores, thereby improving performance in phrase-level machine translation tasks.

5 Conclusion

Addressing semantic independence issues in Chinese-Uyghur statistical machine translation systems, this paper proposes a bilingual relatedness optimization model based on neural machine translation methods. The model introduces attention mechanisms to capture word alignment information in bilingual phrases and re-evaluates bilingual relatedness scores based on semantic relevance and internal lexical matching degree to optimize lexical weights in statistical translation models. Additionally, given a source phrase, the model can predict the target phrase with highest matching degree. Experimental results demonstrate that under conditions of relatively small-scale training data and vocabulary, our proposed method can effectively improve performance in both phrase-level and sentence-level machine translation tasks.

Future research directions include: First, since word alignment results contain missing, redundant, and erroneous information that significantly affects model training, we will consider directly optimizing word alignment results. Second, as this paper only conducts data analysis and modeling for Chinese-Uyghur machine translation tasks, performance may vary for other language pairs, so we will conduct related experiments on other language pairs to improve model generalization. Third, we will segment Uyghur stems and affixes to learn more morphological information.

References

[1] Koehn P, Och F J, Marcu D. Statistical phrase-based translation [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technology-Volume 1. Association for Computational Linguistics. 2003: 48-54.

[2] Och F J, Ney H. A systematic comparison of various statistical alignment models [J]. Computational linguistics, 2003, 29 (1): 19-51.

[3] Peng Fei, Tuergen Yibulayin, Aishan Wumaier, et al. Construction of Chinese-Uyghur comparable corpus for alignment of bilingual technical terms [J]. Journal of Xinjiang University: Natural Science Edition, 2017, 34 (3): 316-321.

[4] Miliwan Xuehelaiti, Liu Kai, Turgun Ibrahim. Chinese-Uyghur machine translation model based on smallest translation units of stems and suffixes [J]. Journal of Chinese Information Processing, 2015, 29 (3): 201-206.

[5] Pan Yirong, Li Xiao, Yang Yating, et al. Reordering table reconstruction model for Chinese-Uyghur machine translation [J]. Journal of Computer Applications, 2018, 38 (5): 1283-1288.

[6] Schwenk H. Continuous space translation models for phrase-based statistical machine translation [J]. Proceedings of COLING 2012: Posters, 2012: 1071-1080.

[7] Son L H, Allauzen A, Yvon F. Continuous space translation models with neural networks [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2012: 39-48.

[8] Zou W Y, Socher R, Cer D, et al. Bilingual word embeddings for phrase-based machine translation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2013: 1393-1398.

[9] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. arXiv preprint arXiv: 1406.1078, 2014.

[10] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv: 1409.0473, 2014.

[11] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [J]. arXiv preprint arXiv: 1508.07909, 2015.

[12] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proc of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. 2002: 311-318.

[13] Och F J, Ney H. Giza++: training of statistical translation models [Z]. 2003.

[14] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Trans on Signal Processing, 1997, 45 (11): 2673-2681.

[15] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Advances in neural information processing systems. 2014: 3104-3112.

[16] Jean S, Cho K, Memisevic R, et al. On using very large target vocabulary for neural machine translation [J]. arXiv preprint arXiv: 1412.2007, 2014.

[17] Halidanmu Abudukelimu, Liu Yang, Sun Maosong. Performance comparison of neural machine translation systems in Uyghur-Chinese translation [J]. Journal of Tsinghua University: Science and Technology, 2017, 57 (1): 1-6.

[18] Stolcke A. SRILM: an extensible language modeling toolkit [C]// Proc of the 7th International Conference on Spoken Language Processing, volume 2. 2002: 901-904.

[19] Och F J. Minimum error rate training in statistical machine translation [C]// Proc of the 41st Annual Meeting on Association for Computational Linguistics, volume 1. 2003: 160-167.

[20] Goodfellow I J, Warde-Farley D, Mirza M, et al. Maxout networks [J]. arXiv preprint arXiv: 1302.4389, 2013.

[21] Zeiler M D. ADADELTA: an adaptive learning rate method [J]. arXiv preprint arXiv: 1212.5701, 2012.

Note: Figure translations are in progress. See original paper for figures.