Neurocomputing 420 (2021) 15–26

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Improving neural machine translation with sentence alignment learning

Xuewen Shi [a,b], Heyan Huang [a,b], Ping Jian [a,b,*], Yi-Kun Tang [a,b]

[a] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[b] Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing, China

ARTICLE INFO

Article history:
Received 31 December 2019
Revised 23 April 2020
Accepted 6 May 2020
Available online 17 September 2020

Keywords:
Neural machine translation
Sentence alignment
Adversarial training

ABSTRACT

Neural machine translation (NMT) optimized by maximum likelihood estimation (MLE) usually lacks the guarantee of translation adequacy. To alleviate this problem, we propose an NMT approach that heightens the adequacy in machine translation by transferring the semantic knowledge from bilingual sentence alignment learning. Specifically, we first design a discriminator that learns to estimate sentence aligning score over translation candidates. The discriminator is constructed by gated self-attention based sentence encoders and trained with an *N*-pair loss for better capturing lexical evidences from bilingual sentence pairs. Then we propose an adversarial training framework as well as a sentence alignment-aware decoding method for NMT to transfer the discriminator's learned semantic knowledge to NMT models. We conduct our experiments on Chinese → English, Uyghur → Chinese and English → German translation tasks. Experimental results show that our proposed methods outperform baseline NMT models on all these three translation tasks. Further analysis also indicates the characteristics of our approaches and details the semantic knowledge that transfered from the discriminator to the NMT model.

1. Introduction

Recently, with the renaissance of deep learning, end-to-end neural machine translation (NMT) [1–4] has gained remarkable performances [5–7]. Conventional NMT approaches are typically optimized by maximizing the likelihood estimation (MLE) of each word in the ground truth translations during the training procedure. However, such an objective cannot guarantee the sufficiency of the generated translations, due to the lack of mechanism for quantitatively measuring the information transformational completeness from the source side to the target.

Some existing work alleviates this problem by directly incorporating coverage or fertility mechanism to an NMT model [8–10]. However, the problem is that attention weights based coverage calculation for NMT [8–10] is insensitive to translation errors and sometimes makes mistakes. Furthermore, it is also unreasonable to consider all kinds of source words equally, since disparate words contribute differently to sentences in semantics and syntax. For example, as illustrated in Fig. 1, translation errors are recorded as positive examples for calculating the coverage, and the alignments between function words also dilute the contribution of key words.

In this paper, we address the inadequate translation problem by introducing novel sentence alignment constrains to NMT. Specifically, we first propose a sentence alignment oriented discriminator *D* that learns to estimate an alignment score between source and target sentences. We employ a gated self-attention based encoder for bilingual sentences encoding in *D* in order to capture the semantic alignment evidence of the input data. An *N*-pair loss [11] is introduced to the training procedure of *D* for preventing the correct but not human generated translations from being overly penalized.

Then, we apply an adversarial training framework as well as an alignment-aware decoding strategy to incorporate the sentence alignment constrain into NMT. Under the adversarial training framework, a standard NMT model *G* is trained to produce an appropriate translation that gains higher score assigned by *D*. We leverage Gumbel-Softmax (GS) [12,13] approximation for *G* to solve the problem of discrete samples, making the response from *D* to *G* differentiable. As for alignment-aware decoding, we conduct *D* to guide the NMT generated translation by combining the alignment score with the decoding log-likelihood. A sentence alignment based value-network [14] is also employed to study the effectiveness of our approaches.

To sum up, the proposed approach has the following advantages:

* Corresponding author at: School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China.
E-mail addresses: xwshi@bit.edu.cn (X. Shi), hhy63@bit.edu.cn (H. Huang), pjian@bit.edu.cn (P. Jian), tangyk@bit.edu.cn (Y.-K. Tang).

https://doi.org/10.1016/j.neucom.2020.05.104
0925-2312/

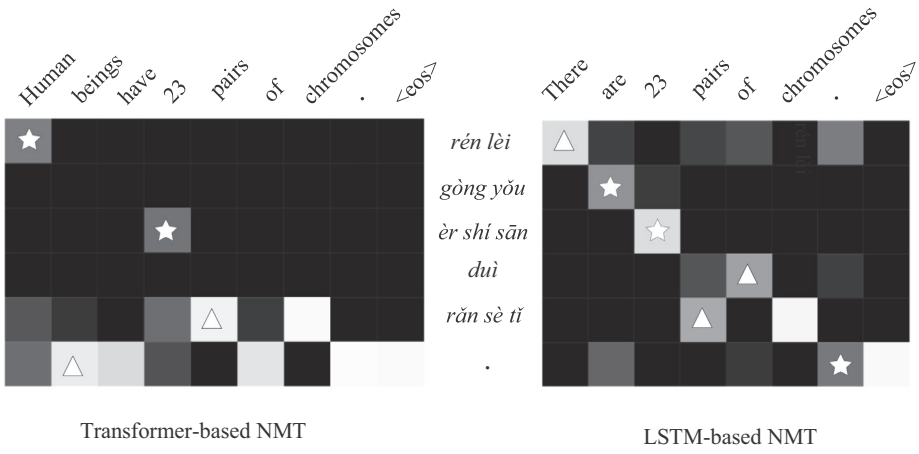


Fig. 1. Model generated attention weights matrix of two Chinese → English translation examples of Transformer-based and LSTM-based NMT systems. The source language (Chinese) is in the vertical direction, and the target language (English) is in the horizontal direction. The lighter color in the attention matrix represents the higher attention weight. The token "⟨eos⟩" is the end-of-sequence symbol. "☆" and "△" in the matrix represent the "good" and "bad" word alignments (attention), respectively. Each alignment will be counted as a coverage of the source word (each source word is covered at most once). According to the attention-based coverage, LSTM-based NMT covers more source words than Transformer-based NMT in this example, which is opposite to human judgments.

- We propose a gated self-attention based encoder for bilingual sentence embedding. The proposed encoder learns to focus on important lexical evidence for sentence aligning and enhance the contribution of the key words. This lexical and semantic knowledge can be transfered to *G* through the proposed training and decoding framework.
- We introduce a novel end-to-end NMT adversarial training framework that heightens adequacy in translation. Under the framework, an NMT model is encouraged to generate translations that match the semantic knowledge learned by a discriminator for sentence aligning. This can be viewed as an instance of "knowledge transfer".
- We also propose an alignment-aware decoding method for NMT. It incorporates the sentence alignment score into the NMT decoding step, which allows the NMT decoder to take into account both the adequacy and fluency of translations.
- The *N*-pair loss [11,15] encourages samples closed to the gold-standard to get higher score. Unlike the binary classification used in previous work [16–18], translations that are correct but different from the ground-truth ones will not be overly penalized.

We apply two popular NMT models, LSTM-based NMT [5] and Transformer [7], as the baseline model architectures and conduct experiments on Chinese → English, Uyghur → Chinese and German → English translation tasks. Experimental results show that our proposed approach achieves significant improvements on all the three language pairs. We also evaluate the performance of the discriminator on both sentence alignment and translation candidate re-ranking tasks, which proves its independence and transferability. Further analyses also show the detailed alignment-oriented knowledge that the discriminator transfers to the NMT model.

2. Related work

2.1. Neural attention model

A neural attention mechanism enables a neural network to focus more on relevant elements of the input than on irrelevant parts. At present, attention mechanism is a vital component for various natural language processing approaches, including, but not limited to language modeling [19], speech recognition [20], machine translation [4], text summarization [21] and question answering [22]. With its initial success in the field of natural language processing, attention modeling also rapidly finds its applications in various computer vision and vision-language tasks, such as object segmentation [23,24], image captioning [25], video segmentation [26], and human-object relation reasoning [27].

For NMT, one problem of early NMT solutions is that they often produce poor translations for long sentences [3,28]. Cho et al. [2] suggest that this weakness is due to the fixed-length of source encoding in conventional encoder-decoder [29]. Bahdanau et al. [4] introduce the concept of attention to NMT to avoid keeping a fixed source side representation. Luong et al. [30] follow the framework of [4] and utilize a dot-product scoring function for computing attention weights. Vaswani et al. [7] propose an NMT architecture based solely on attention mechanisms without any recurrence operations. A multi-head attention and a scaled dot-product are also employed in [7]. Choi et al. [31] propose a fine-grained attention to learn different attention weights for each dimension in the values for NMT.

In this paper, we adopt two brands of attention-based NMT approaches, Bahdanau et al. [4] and Vaswani et al. [7], as the baseline methods. Our methods are built upon the above typical baselines, which helps us to verify the effectiveness and universality of our approaches. For the proposed sentence encoder, it is built upon a gated self-attention mechanism [32] in which all of the keys, values and queries come from the same place.

2.2. Translation adequacy

Most of the state-of-the-art NMT models are optimized by MLE-based objectives [4–7]. However, likelihood fails to measure whether the source information is completely transformed to the target side. Thus, it can hardly handle the translation adequacy problem [33].

One way to alleviate these problems is to apply coverage and fertility to an NMT model to record the translated and untranslated source words. Some efforts employ coverage vector or coverage ratio into NMT in order to represent whether a source word is translated or not [8,9]. Explicitly tracking PAST and FUTURE [34,35] is another method to help NMT to recognize the dynamic translated and untranslated contents. He et al. [14] use a prediction

network to estimate the future cost of translating the uncovered source words.

On the other hand, some recent efforts introduce an additional source side constraint for NMT to improve translation adequacy. Tu et al. [33] add a re-constructor to traditional NMT model, which introduces an auxiliary score to measure the adequacy of translation. Dual learning [36,37] and dual inference [38] are also proposed to exploit the probabilistic correlation between dual tasks to regularize the training process. These previous approaches apply a reconstruction reward by comparing the source input and the reconstructed sentence, while we use alignment score directly to model the discrepancy between the source and the translation.

2.3. Adversarial learning

GAN [39] is another promising framework to leverage sentence-level objectives in NMT. Recently, there is some remarkable work in NMT [17,18]. The framework comprises of two sub-models: i) an NMT model aims to produce sentences which are hard to be discriminated from the gold-standard sentences; and ii) a discriminator makes efforts to differentiate the model generated translations from the ground-truth ones. However, these approaches rarely take account of translation adequacy. Furthermore, the discriminators of these work refer the target sentence in the corpus as the single gold-standard regardless the quality of model generated translations, which usually punish too much for the good translations generated by the model. Kong et al. [10] propose an adequacy-oriented discriminator which is trained to estimate the Coverage Difference Ratio (CDR) given the source and the generated translation. However, CDR is unable to distinguish translation errors and it also neglects the importance of diversity between different words (as the examples shown in Fig. 1).

Unlike the discriminators in [17,18,10], our alignment-oriented discriminator learns a specific function to measure alignment score between source and target sentences, which is trained totally independently by the NMT generator. The proposed discriminator assigns different weights to words and is sensitive to translation errors. We also apply *N*-pair loss to the training process of *D* for ensuring that *D* will not punish the translations closed to the gold-standard overly.

3. Background: attention-based neural machine translation

NMT models are usually built upon an encoder-decoder framework [29]. In the encoder-decoder framework, the encoder reads an input sequence $\mathrm{X} = \{x_1, \ldots, x_{T_x}\}$ into a hidden state sequence $\mathrm{H} = \{h_1, \ldots, h_{T_x}\}$, and the decoder is designed to define a probability over the translation $\mathrm{Y} = \left\{y_1, \ldots, y_{T_y}\right\}$ by:

$$p(\mathrm{Y}|\mathrm{X}) = \prod_{t=1}^{T_y} p(y_t|y_{<t}, \mathrm{H}), \tag{1}$$

where T_x and T_y are sequence length of X and Y, $\mathrm{y}_{<\mathrm{t}}$ represents previous generated translations, and

$$p(y_t|y_{<t}, \mathrm{H}) = g(Ey_{t-1}, s_t, c_t), \tag{2}$$

where $g(\cdot)$ is an NMT decoder, Ey_{t-1} is the word vector of the word y_{t-1}, s_t is the decoder hidden state, and c_t is an attention based context representation.

To verify the effectiveness of the proposed method, we take two different NMT models, the recurrent neural network (RNN) based NMT and Transformer, as the implementations of our proposed approaches.

3.1. RNN-based approach

RNN-based NMT is a traditional NMT architecture which has been widely explored [4,30,5]. The encoder encodes the input sequence X into a hidden state sequence H by a bidirectional RNN. The decoder is another RNN that predicts a target sequence $\mathrm{Y} = \left\{y_1, \ldots, y_{T_y}\right\}$. Each word y_i is predicted via a recurrent hidden state s_t, the previously predicted word y_{t-1} and a context vector c_t as shown in Eq. 2. The s_t in Eq. 2 is computed as:

$$s_t = \mathrm{RNN}(Ey_{t-1}, s_{t-1}, c_t), \tag{3}$$

where $s_t, c_t \in \mathbb{R}^{d_m}, Ey_{t-1} \in \mathbb{R}^{d_e}$, and in this paper, d_m and d_e are set to 512. $\mathrm{RNN}(\cdot)$ is an RNN-based model architecture which can be implemented as long short term memory network (LSTM) [40] or gated recurrent unit (GRU) [29]. The c_i is computed as a weighted sum of the encoded annotations H:

$$c_t = \sum_{j=1}^{T_x} a_{t,j} h_j, \tag{4}$$

where $h_j \in \mathbb{R}^{d_m}$, and the weight $a_{t,j}$ of each annotation h_j is computed by the attention mechanism:

$$a_{t,j} = \frac{\exp(e_{t,j})}{\sum_{k=1}^{T_x} \exp(e_{t,k})}, \tag{5}$$

where

$$e_{t,j} = \alpha(s_{t-1}, h_j), \tag{6}$$

is an attention model which scores how well the inputs around position j match up with the output at position t.

3.2. Transformer

The Transformer [7] is one of the most popular state-of-the-art NMT models. In the Transformer, the encoder contains a stack of 6 identical layers. Each layer is consist of two sub-layers: i) a multi-head self-attention mechanism, and ii) a position-wise fully connected feed-forward network. Suppose a set of packed queries matrix Q, keys matrix K and values matrix V, the attention is computed as:

$$\alpha(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^\top}{\sqrt{d_k}}\right)\mathrm{V}, \tag{7}$$

where $1/\sqrt{d_k}$ is the scaling factor for the dot-product $\mathrm{QK}^\top$, and d_k is the dimension of a key vector. In a self-attention layer, all of the keys, values and queries are from the same place.

Transformer employs multi-head attention for jointly attending to information from disparate representation subspaces at different positions:

$$\mathrm{MultiHead}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_{N_h})W^O, \tag{8}$$

and

$$\mathrm{head}_n = \alpha\left(\mathrm{Q}W_n^Q, \mathrm{K}W_n^K, \mathrm{V}W_n^V\right), \tag{9}$$

where $\mathrm{Concat}(\cdot)$ is a concentrate operator, $W_n^Q \in \mathbb{R}^{d_m \times d_k}$, $W_n^K \in \mathbb{R}^{d_m \times d_k}$, $W_n^V \in \mathbb{R}^{d_m \times d_k}$ and $W^O \in \mathbb{R}^{n_h d_v \times d_m}$.

In this work, we employ $n_h = 8$ parallel attention heads. For each of these, we use $d_k = d_v = d_m/n_h = 64$. A residual connection [5] is applied around each of the two sub-layers, followed by layer normalization [41]. The decoder is also composed of a 6 identical layers stack. Besides the two sub-layers stated above, a third

sub-layer is inserted in each layer that performs multi-head attention over the output of the encoder.

4. Approach

In this section, we describe our approach that enhances the performance of NMT by incorporating sentence alignment knowledge. We first propose a sentence alignment-based discriminator D that learns to estimate the alignment score and sort the translation candidates. Then, we employ two frameworks to apply D to NMT: i) transferring sentence alignment knowledge to NMT model by applying adversarial training (see section 4.2). and ii) feeding sentence alignment knowledge back to NMT dynamically by decoding with D (see section 4.3).

4.1. Self-attention based discriminator

For the discriminator D, we propose a gated self-attention based sentence encoder to perform source and target sentence encoding, and then calculate the alignment score using the encoding pairs. The model architecture is shown in Fig. 2.

Self-Attention based sentence encoder. As depicted in Fig. 2, we conduct a shallow network architecture: one gated hidden layer and one self-attention layer as the sentence encoder. The self-attention mechanism and shallow architecture will help the encoder select more important lexical evidences to estimate the alignment score between two sentences.

Given a one-hot encoded input sequence X and a word embedding lookup table $\mathrm{E} \in \mathbb{R}^{|V| \times d_e}$, where $|V|$ is the size of vocabulary and d_e is the embedding dimension. The input X will be represented as a corresponding word embedding matrix $\mathrm{E}x \in \mathbb{R}^{T_x \times d_e}$. We apply a gating mechanism [32] to compute the hidden layer H:

$$\mathrm{H} = (U_h \mathrm{E}x + b_h) \otimes \sigma(U_g \mathrm{E}x + b_g), \tag{10}$$

where U_h and $U_g \in \mathbb{R}^{d_m \times d_e}$, $\sigma(\cdot)$ is a logistic sigmoid function, and $\otimes$ is element-wise product between matrices. Then the self-attention weights W is computed as:

$$\mathrm{W} = \mathrm{softmax}(\tanh(U_a \mathrm{H})), \tag{11}$$

where $U_a \in \mathbb{R}^{d_m \times d_m}$. The output of the gated self-attention encoder is formulated as:

$$e = U_o(\mathrm{W} \times \mathrm{H}) + b_o, \tag{12}$$

where $e \in \mathbb{R}^{d_m}$, and $U_o \in \mathbb{R}^{d_m \times d_m}$. We add layer normalization [41] to the output layer, and in this paper, the model dimension d_m and the embedding dimension d_e are set to 512 and 256, respectively.

Alignment score and loss function. With the source and target sentence encoding vectors e_x and e_y, the alignment score $s_{(\mathrm{X,Y})}$ can be computed as:

$$s_{(\mathrm{X,Y})} = D(\mathrm{X}, \mathrm{Y}) = e_x^\top e_y. \tag{13}$$

Given a candidate target sentence list $\mathscr{Y}$, D produces a distribution over $\mathscr{Y}$ and aims to maximize the log-likelihood of the gold-standard alignment sentence Y^+. Since sentence-level alignments in automatic extracted corpora are usually not very precise, we expect the loss function for training D not to be too strict with candidates closed to the gold-standard one. Therefore, we apply a metric-learning multi-class N-pair loss [11,15] to our model, which can be defined as:

$$\begin{aligned}\mathscr{L}_D &= \mathscr{L}_{N-\mathrm{pair}}\left(\left\{\mathrm{X}, \mathrm{Y}^+, \{\mathrm{Y}_n^-\}_{n=1}^{N-1}\right\}; \theta_D\right) \\ &= \log\left(1 + \sum_{n=1}^{N-1} \exp(D(\mathrm{X}, \mathrm{Y}_n^-) - D(\mathrm{X}, \mathrm{Y}^+))\right),\end{aligned} \tag{14}$$

where Y^+ is the alignment target sentence to the source sequence X, and Y_n^- is one of the $N-1$ unaligned samples. $D(\cdot)$ is the sentence-alignment based discriminator parameterized with θ_D.

Compared to the cross entropy loss used in some previous work [16–18], the N-pair objective encourages the translation candidates which are similar to the given golden-standard one to be scored higher than the dissimilar ones. In this way, translations that are correct but different from the ground truth will not be overly penalized.

4.2. Adversarial training for NMT with sentence alignment model

In this section, we describe our approach that transfers knowledge from D to an NMT model G by adversarial training. This approach mainly consists of two sub-models: ii) a discriminator D learns to estimate the alignment score and sort the translation candidates, and ii) an NMT model G aims to generate translations with higher score assigned by D. A sketch of the proposed training framework is shown in Algorithm 1: for each sentence pair (X, Y) sampled from the training corpus, the NMT model G generates a translation $\widehat{\mathrm{Y}}$ given X, and queries D with $\widehat{\mathrm{Y}}$ to get feedback and update itself. In order to obtain more stable training, we also leverage a teacher-forcing [42] step to our approach.

Algorithm 1: The proposed training framework. See section 4.2 for more details.

Input: Training corpus, a pre-trained generator G and a pre-trained discriminator D.

1: **for**number of training iterations **do**

2: Sample $(\mathrm{X}, \mathrm{Y}^+)$ from training corpus

3: Sample $\widehat{\mathrm{Y}} \sim G(\mathrm{X})$ with a Gumbel-Softmax sampler

4: Compute 1-pair loss $\mathscr{L}_G$ for $\left(\mathrm{X}, \widehat{\mathrm{Y}}\right)$ with D by:

$$\mathscr{L}_G = \log\left(1 + \exp\left(D(\mathrm{X}, \mathrm{Y}^+) - D\left(\mathrm{X}, \widehat{\mathrm{Y}}\right)\right)\right)$$

5: Update G with the learning rate η_G:

$$\theta_G \leftarrow \theta_G - \eta_G \nabla_{\theta_G} \mathscr{L}_G$$

6: (Optional) Update D with the learning rate η_D, $\mathscr{L}_D = -\mathscr{L}_G$:

$$\theta_D \leftarrow \theta_D - \eta_D \nabla_{\theta_D} \mathscr{L}_D$$

7: Teacher-forcing: update G on $(\mathrm{X}, \mathrm{Y}^+)$ by MLE

8: **end for**

Out put: ENSURE The generator G with parameter θ_G

4.2.1. The discriminative loss for generative training

In our framework, G aims to generate a translation scoring higher than the golden-standard under the judgment of D. Specifically, for each sentence pair $(\mathrm{X}, \mathrm{Y}^+)$ in training sets, first, G samples translation $\widehat{\mathrm{Y}}$ given X with greedy searching. Second, D takes $\widehat{\mathrm{Y}}$ as well as $(\mathrm{X}, \mathrm{Y}^+)$ as inputs to compute alignment scores, and then G gets the feedback from D. Eq. 15 gives the perceptual loss that G aims to optimize.

$$\begin{aligned}\mathscr{L}_G &= \mathscr{L}_{1-pair}\left(\left\{\mathrm{X}, \mathrm{Y}^+, \widehat{\mathrm{Y}}\right\}\right) \\ &= \log\left(1 + \exp\left(D(\mathrm{X}, \mathrm{Y}^+) - D\left(\mathrm{X}, \widehat{\mathrm{Y}}\right)\right)\right).\end{aligned} \tag{15}$$

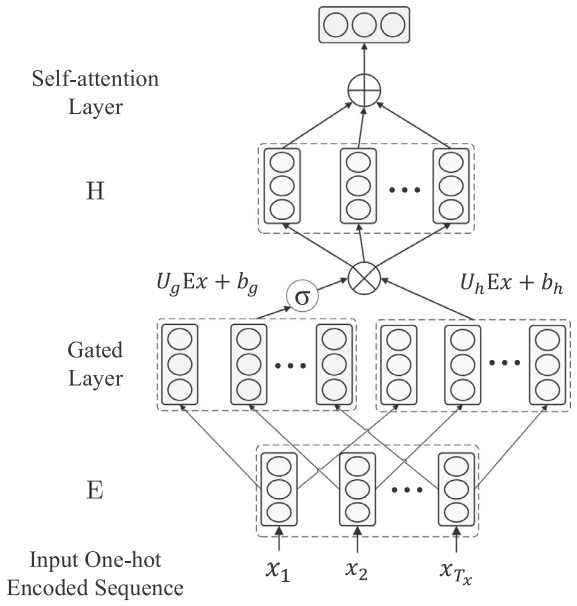


Fig. 2. Model architecture of the gated self-attention sentence encoder. See section 4.1 for more details.

Intuitively, updating generator parameters to minimize $\mathscr{L}_G$ can be interpreted as learning to produce a translation $\widehat{\mathrm{Y}}$ that "fools" the discriminator into believing that this translation should score higher than the human translation Y^+ under the D's scoring function.

4.2.2. Gumbel-softmax sampler

The process of sampling a translation $\widehat{\mathrm{Y}}$ with G is not differentiable, since it includes argmax$(\cdot)$ operator for performing one-hot encoding. To alleviate this problem, we leverage the Gumbel-softmax [12] sampler to the NMT generator. Formally, at the decoding step j, suppose that $p_j \in \mathbb{R}^{|V_y|}$ contains the model output log-probabilities over target vocabulary V_y, and $g_j \in \mathbb{R}^{|V_y|}$includes i.i.d samples drawn from the standard distribution Gumbel$(0,1)$. A sample y_j is transformed as:

$$\hat{y}_j = \mathrm{softmax}\big((p_j + g_j)/\tau\big), \tag{16}$$

where τ is a temperature parameter and it is set to 0.5 in our experiments.

4.2.3. Teacher-forcing step

$\mathscr{L}_G$ in Eq. 15 mainly considers the discrepancy of alignment and integrity between the model output and the ground-truth, though it rarely inspects grammar correctness and language fluency. To alleviate this problem, following [42,15], we adopt a teacher-forcing step to our training process. In this paper, we perform a MLE training ($\mathrm{O}_{\mathrm{MLE}}$) for the teacher-forcing step. Given a training sample $(\mathrm{X}, \mathrm{Y}^+)$, $\mathrm{O}_{\mathrm{MLE}}$ maximize the log-likelihood of the training data: $\hat{\theta}_G = \mathrm{argmax}_{\theta_G}(LL)$, where

$$LL = \sum_{t=1}^{T_y} \log\ p(y_t | y_{<t}, \mathrm{X}; \theta_G), \tag{17}$$

where T_y is the length of Y^+.

4.3. Sentence alignment-aware decoding for NMT

In this section, we describe our approach that utilizes sentence alignment knowledge to guide NMT decoding. Specifically, we redefine the score function $\Theta(\mathrm{X}, \mathrm{Y})$ of the beam search algorithm for NMT by applying a discriminator score $D(\mathrm{X}, \mathrm{Y})$. The decoding process with $\Theta(\cdot)$ is presented in Algorithm 2. We propose two brands of $\Theta(\mathrm{X}, \mathrm{Y})$: i) $\Theta_D(\cdot)$, combining the NMT decoding score $p(\mathrm{Y}|\mathrm{X})$ and the sentence alignment score $D(\mathrm{Y}|\mathrm{X})$ linearly and ii) $\Theta_v(\cdot)$, introducing a value network (VNN) [14] that estimates the sentence alignment score for NMT decoding.

Algorithm 2: Beam search with sentence alignment model for NMT.

Input: Testing example X, an NMT model $p(\mathrm{Y}|\mathrm{X})$, target vocabulary V_y, a discriminator $D(\mathrm{X}, \mathrm{Y})$, beam search size K, maximum sentence length L.

1: S and U are candidate sets, $S = \varnothing = U$
2: **for** $t = 1$; $t \leftarrow t + 1$; $t < L$ and $|S| < K$ **do**
3: $U_{expand} \leftarrow \left\{\widehat{\mathrm{Y}}_i + \{w\} | \widehat{\mathrm{Y}}_i \in U, w \in V_y\right\}$
4: $U \leftarrow \{\text{top } (K - |S|) \text{ candidates that maximize } \Theta\left(\mathrm{X}, \widehat{\mathrm{Y}}_{1:t}\right) | \widehat{\mathrm{Y}}_{1:t} \in U_{expand}\}$
5: $S \leftarrow S \cup \left\{\widehat{\mathrm{Y}}_{1:t} | \widehat{\mathrm{Y}}_{1:t} \in U, y_t = \text{"}\langle\text{eos}\rangle\text{"}\right\}$
6: $U \leftarrow U \setminus \left\{\widehat{\mathrm{Y}}_{1:t} | \widehat{\mathrm{Y}}_{1:t} \in U, y_t = \text{"}\langle\text{eos}\rangle\text{"}\right\}$
7: **end for**

Out put: $\widehat{\mathrm{Y}} = \mathrm{argmax}_{\widehat{\mathrm{Y}} \in S \cup U} \Theta\left(\mathrm{X}, \widehat{\mathrm{Y}}\right)$

4.3.1. Decoding with discriminator

For $\Theta_D(\cdot)$, we simply combine the outputs of both the NMT model generative likelihood and the discriminator score linearly. Formally, given a translation model score $p(\mathrm{Y}|\mathrm{X})$, a discriminator score $D(\mathrm{X}, \mathrm{Y})$ and a hyperparameter $\beta_D \in (0, 1)$, the score $\Theta_D(\mathrm{X}, \mathrm{Y}_{1:t})$ of partial sequence $\mathrm{Y}_{1:t}$ at decoding step t for X is computed by:

$$\Theta_D(\mathrm{X}, \mathrm{Y}_{1:t}) = \beta_D \times \frac{1}{t} \log\ p(\mathrm{Y}_{1:t}|\mathrm{X}) + (1 - \beta_D) \times \log\ D(\mathrm{X}, \mathrm{Y}_{1:t}). \tag{18}$$

4.3.2. Decoding with sentence alignment based value network

Although $\Theta_D(\cdot)$ introduce the additional alignment score, it is still faced with myopic bias without estimating long-term reward. Following the framework of VNN-NMT [14], we apply a prediction model VNN_D to estimate the expected long-term alignment score according to the source input and the partial generated target sequence. In the decoding step, the decoder selects the best candidates not only based on the generative probability, but also based on the estimated alignment score. The score function $\Theta_v(\cdot)$ is computed by:

$$\Theta_v(\mathrm{X}, \mathrm{Y}_{1:t}) = \beta_v \times \frac{1}{t} \log\ p(\mathrm{Y}_{1:t}|\mathrm{X}) + (1 - \beta_v) \times \log\ \mathrm{VNN}_D(\mathrm{X}, \mathrm{Y}_{1:t}), \tag{19}$$

where $\mathrm{VNN}_D(\mathrm{X}, \mathrm{Y}_{1:t})$ is the score of the sentence alignment based value network and $\beta_v \in (0, 1)$ is a hyperparameter. Since the model estimate the decoding value based on the RNN hidden states of NMT model, we only perform VNN_D on the LSTM-based NMT model in this paper. The detailed training process of VNN_D is shown in section 5.3.

5. Experiment

5.1. Datasets

We evaluate the proposed approach on three translation tasks: Chinese → English (Zh → En), Uyghur → Chinese (Ug → Zh), and English → German (En → De). For all translation tasks, we first tokenize all corpora with the Moses [43] *tokenizer.perl*.[1] Sentences

[1] Moses scripts used in this work: https://github.com/moses-smt/mosesdecoder/blob/master/scripts/.

longer than 100 tokens are discarded, and all the sentences are encoded with byte-pair encoding (BPE) [44]. For Zh → En and Ug → Zh translation tasks, the Chinese parts of both training and test sets are segmented by the LTP Chinese word segmentor [45] before applying BPE [44] to the corpus. In the following parts of the paper, the Chinese examples are written in Chinese Pinyin and presented by segmented italic form.

Chinese → English. For Chinese → English translation, the training data is extracted from four LDC corpora.[2] The training set finally contains 1.3 M parallel sentence pairs in total. After preprocessing, we get a Chinese vocabulary of about 39K tokens, and an English vocabulary of about 30K tokens. We use NIST2005 dataset for validation and NIST2002, NIST2003, and NIST2004 datasets for testing.

5.1.1. Uyghur → Chinese

For Uyghur → Chinese translation, our training corpus is from Uyghur to Chinese News Translation Task in CCMT2019 Machine Translation Evaluation[3]. The training set of Ug → Zh contains 0.17M parallel sentence pairs. Apart from the Moses [43] tokenizer, we do not use any other tools to segment Uyghur. We get vocabularies of 30K tokens for both Uyghur and Chinese corpus. We use CWMT2018-uc-news-test and CCMT2019_UC_test as the validation set and the test set respectively.

5.1.2. English → German

For English → German translation, we conduct experiments on the publicly available corpora WMT'14[4] En → De. The training set of En → De task totally contains 4.5M sentence pairs, and we use a shared source-target vocabulary of about 39K tokens. We use newstest2013 as the validation set and report the results on newstest2014.

Discriminative corpus construction.[5] Different from parallel corpora for training NMT, the training data for D provides a candidate translation list for each input source sentence. We manually construct the training corpus for D using the original parallel corpus. For each source sentence, we set the size of candidate list to 100. The list contains one golden standard translation and the other 99 candidates are interference. The translation candidates are obtained from the context of the golden standard translation in the comparable paragraph. If the number of context sentences N_c is less than 99, then we sample another $99 - N_c$ sentences randomly from the rest whole corpus. The test sets for discriminator is similar to the training set, where each input corresponds to one hundred candidates extracted from the adjacent document context. As for the format of data construction, we follow most of VisDial[6] [47].

5.2. Evaluation

For all NMT models, we apply beam search during decoding with the beam size of 6. For the Transformer baseline, following [7], we report the result of a single model obtained by averaging the 5 checkpoints around the best model selected on the development set. The translation results are measured in case-insensitive BLEU [48], which is widely used in machine translation community [43,4,5,7]. Formally, a BLEU score is calculated as:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4}\log p_n\right), \tag{20}$$

where p_n is the case-insensitive n-gram precision. The brevity penalty BP is computed as:

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \operatorname{len}\left(\hat{\mathrm{Y}}\right) > \operatorname{len}\left(\mathrm{Y}^{+}\right) \\ \exp\left(1 - \frac{\operatorname{len}\left(\mathrm{Y}^{+}\right)}{\operatorname{len}\left(\hat{\mathrm{Y}}\right)}\right)) & \text{if } \operatorname{len}\left(\hat{\mathrm{Y}}\right) \leqslant \operatorname{len}\left(\mathrm{Y}^{+}\right) \end{cases}, \tag{21}$$

where the function $\operatorname{len}(\cdot)$ returns the length of the inputs. For Zh → En and En → De, the BLEU score is reported at word level with the Moses tokenizer[1], and for Ug → Zh, the BLEU score is evaluated at Chinese character level. In this paper, we apply the *multi-bleu.perl* script[1] to calculate BLEU scores.

For the discriminator, the performance is evaluated on mean reciprocal rank score (MRR), recall@k(R@k) and mean rank score (Mean). Those metrics are usually used to measure the model performance for ranking answers with multiple candidates [47,49]. The detailed calculation methods of the above metrics are described below.

5.2.1. MRR and mean

For a single query, the reciprocal rank is 1/rank, where rank is the order of the correct answer. In this paper, the maximum value of rank is 100. For multiple queries, MRR is the mean of the N_Q reciprocal ranks:

$$\mathrm{MRR} = \frac{1}{N_Q}\sum_{i=1}^{N_Q}\frac{1}{\mathrm{rank}_i}, \tag{22}$$

where N_Q is the number of queries and Mean is computed as:

$$\mathrm{Mean} = \frac{1}{N_Q}\sum_{i=1}^{N_Q}\mathrm{rank}_i. \tag{23}$$

5.2.2. Recall@k

Recall@k is computed as $N_{\mathrm{rank}\leqslant k}/N_Q$, where $N_{\mathrm{rank}\leqslant k}$ is the total number of the correct answers with rank $\leqslant$ k.

5.3. Training details

5.3.1. Model setups

In this paper, we take two specific implements of NMT: a stacked-LSTM based NMT [5] and the Transformer [7]. For the LSTM-based NMT, the model consists of an encoder with a bi-directional LSTM and a decoder with two stacked LSTM layer. The model dimension d_m of each LSTM layers and the attention model is set to 512. For the Transformer, following the setups of the base model [7], we use 8 attention heads, 512-dimensional output vectors for each layer, and 2048-dimensional inner-layer of the feed-forward network. For decoding with D and v_D, we set β_D and β_v to be 0.95 and 0.9 respectively.

5.3.2. Pre-train discriminator and NMT model

We first pre-train the discriminator and NMT models separately until the performance on development set does not improve for a warm start. For training discriminator, we apply an Adam [50] optimizer with $\beta_1 = 0.8, \beta_2 = 0.99$ and a base learning rate of 4×10^{-4}. The mini-batch size is 100 and the dropout rate is set to 0.1. As for the NMT model, we apply the dropout rate of 0.1 to

[2] LDC2005T10, LDC2003E14, LDC2004T08 and LDC2002E18. Since LDC2003E14 is a document-level alignment comparable corpus, we use Champollion Tool Kit [46] to extract parallel sentence pairs from it.

[3] https://ccmt2019.jxnu.edu.cn/

[4] http://www.statmt.org/wmt14/.

[5] The source codes and demo data will be released at: https://github.com/PolarLion/Sentence-Alignment-Learning.

[6] https://visualdialog.org/

Zh → En and En → De, and 0.3 to Ug → Zh, respectively. We follow the base model of Transformer [7] except employing label smoothing [51].

5.3.3. Frozen discriminator v.s. adversarial discriminator

We also study the effects of two training setups of the discriminator: updating D (adversarial D) or not (frozen D) with the NMT model. When we perform frozen D, the NMT model is trained with a combination of the discriminative perceptual loss [15] and the teacher-forcing loss [42,15]. Each mini-batch contains 32 sentence pairs due to the limitation of memory size of a single GPU. For the adversarial discriminator, we alternately update G and D under the adversarial learning framework [17,18,10]. An adversarial D is to maximize the score of the human translation Y^{+} and minimize the score of the generated translation $\widehat{\mathrm{Y}}$. Then the training loss for adversarial D can be represented as: $\mathscr{L}_D = -\mathscr{L}_G$.

5.3.4. The training of sentence alignment based value network

We perform VNN_D on LSTM-based NMT only and follow most setups of [14]. For VNN_D, we replace the averaged BLEU score used in [14] by averaged sentence alignment score as the reward. The average sentence alignment score is computed by a pre-trained discriminator:

$$Averaged\_Alignment\left(\mathrm{X}, \widehat{\mathrm{Y}}_{\mathrm{p}}\right) = \frac{1}{K} \sum_{\widehat{\mathrm{Y}} \in S\left(\widehat{\mathrm{Y}}_{\mathrm{p}}\right)} D\left(\mathrm{X}, \widehat{\mathrm{Y}}\right), \tag{24}$$

where $\widehat{\mathrm{Y}}_{\mathrm{p}}$ is a partial target sentence generated by G with random early stop. We use G to finish the translation starting from $\widehat{\mathrm{Y}}_{\mathrm{p}}$, and obtain a set $S\left(\widehat{\mathrm{Y}}_{\mathrm{p}}\right)$ of $K = 6$ to complete target sentences using beam search.

5.4. Machine translation results

We report the experimental results on machine translation in this section. We build upon our approaches with two popular NMT model architectures: LSTM-based NMT and Transformer for comparison and perform translation on three translation tasks: Zh → En, Ug → Zh and En → De. The BLEU scores are shown in Table 1 and Table 2. The overall results show that: i) our proposed methods achieve better performances on BLEU for both model architectures; ii) the proposed methods are also effective on low resource translation task like Ug → Zh. These results indicate that the proposed methods make up for the shortcomings of the MLE training.

For the three translation tasks, the corpus size of En → De are relatively rich while the datasets of Ug → Zh are scarce. However, the proposed methods are effective for all the three translation tasks. Both of the proposed training framework and the decoding method can be seen as knowledge transfer from the discriminator to NMT model, through N-pair loss and scoring function, respectively. We suppose that if D can capture lexical knowledge from the corpus, D will successfully enhance NMT using the learned information. In this paper, we employ a light and flat attention-based discriminator architecture, which is less sensitive to the data size. Further experiments also show that the performance of D is stable for all language pairs (the evaluation of discriminator is shown in section 5.5).

Table 2
BLEU scores on Ug → Zh and En → De translation task. For CWMT2019 Ug-Zh testset, the BLEU scores are reported at Chinese character level (not word level). See section 5.4 for more details.

#	Model	CWMT2019	newstest2014
	LSTM-based NMT	31.51	21.71
	+decoding w/ D	31.86	21.63
	+decoding w/ VNN_D	31.62	21.95
	+frozen D + $\mathrm{O}_{\mathrm{MLE}}$	32.13	23.64
	+adversarial D + $\mathrm{O}_{\mathrm{MLE}}$	**32.40**	**23.67**
	Transformer	32.56	26.52
	+decoding w/ D	32.93	26.57
	+frozen D + $\mathrm{O}_{\mathrm{MLE}}$	33.24	27.10
	+adversarial D + $\mathrm{O}_{\mathrm{MLE}}$	**33.78**	**27.16**

For the proposed adversarial training framework (see section 4.2), we compare two setups of frozen D and adversarial D for the discriminator. Experimental results show that continuing to update D along with G gains better BLEU scores for all the three translation tasks. Continuing to update D can be seen as fine-tuning D with G's dynamically generated data, and the above results illustrate that fine-tuning D can further improve the training efficiency. Further analyses about the comparison between "froze D" and "adversarial D" are presented in section 6.4.

For decoding with D and VNN_D (see section 4.3), the improvements seem to be narrow since G gets limited guidance from D's learned knowledge on sentence alignment. Introducing D's score enriches the searching strategy with considering alignment, although it does not alleviate the short sighted actions of beam search framework. For VNN_D, it is training with D's estimated score which increases the uncertainty comparing to model-agnostic BLEU score. Therefore, the improvement of the effect may be affected.

5.5. Discriminator performance

In our approach, D aims at ranking appropriate translations as high as possible. We present the performance of D on Zh → En, Ug → Zh and En → De test sets in Table 3. It shows that for all test sets of all language pairs, our proposed discriminator performs steadily at high recall rate of more than 96% on recall@1, and nearly 100% on recall@5 and recall@10, which demonstrates that the ground-truth translations are always assigned to high alignment score.

Empirical and principled studies indicate that high initial accuracy of binary classification based discriminator usually leads to

Table 1
BLEU scores on Zh → En translation task. Transformer is the baseline model. "Average" is the averaged BLEU scores on test sets. See section 5.4 for more details.

#	Model	NIST2002	NIST2003	NIST2004	Average
	LSTM-based NMT	37.20	34.87	38.07	36.71
	+decoding w/ D	37.34	34.89	38.10	36.78
	+decoding w/ VNN_D	37.89	35.04	38.23	37.05
	+frozen D + $\mathrm{O}_{\mathrm{MLE}}$	38.42	36.20	39.75	38.12
	+adversarial D + $\mathrm{O}_{\mathrm{MLE}}$	**38.61**	**36.23**	**39.87**	**38.24**
	Transformer	41.56	39.95	42.05	41.19
	+decoding w/ D	41.63	40.10	42.01	41.25
	+frozen D + $\mathrm{O}_{\mathrm{MLE}}$	42.51	**40.74**	42.42	41.89
	+adversarial D + $\mathrm{O}_{\mathrm{MLE}}$	**42.67**	40.67	**42.51**	**41.95**

Table 3
Discriminator performance on Zh → En, Ug → Zh and En → De test sets. See section 5.5 for more details.

Testset	MRR	R@1	R@5	R@10	Mean
NIST2002	98.28	97.04	99.77	99.89	1.06
NIST2003	98.36	97.50	99.34	99.67	1.10
NIST2004	98.43	97.25	99.94	99.94	1.05
CCMT2019	97.76	96.30	99.50	99.70	1.20
newstest2014	98.07	96.90	99.43	99.80	1.10

worse model performance for GANs [52,17], due to the adoption of the Jensen-Shannon divergence [52] between two data distributions. In this paper, G is trained with 1-pair loss defined on sentence alignment score instead of classification based cross-entropy, which could avoid the vanishing gradient in conventional GANs. Therefore, the high efficacy of the discriminator would not make negative impact on the proposed adversarial training procedure. Further discussions about the impact of pre-training of D are presented in section 6.4.

6. Analysis

In this section, we will study the characteristics of our proposed approaches and report some detailed experimental results. We also give a specific translation example to illustrate how knowledge transferring improves NMT performance.

6.1. Knowledge learned by self-attention mechanism

In this section, we attempt to answer "what kind of knowledge does D transfer to NMT?". In order to illustrate the lexical-level knowledge learned by D, we give a visual example in Fig. 3. It shows self-attention weights of the encoders for the given source and the target sentences from a human reference and an NMT's generation. In the example, the source sentence is "*rénlèi gòngyǒu èrshís=an duì rǎnsètǐ*.", the reference sentence and the NMT output are "Humans have a total of 23 pairs of chromosomes." and "Human beings have 23 pairs of chromosomes.". We notice that the source language words "rénlèi", "èrshís=an" and "rǎnsètǐ", and their corresponding target language words "Humans (Human beings)", "23" and "chromosomes" are assigned higher self-attention weights than other tokens, which demonstrates that the encoder regard those tokens as important lexical evidences to estimate alignment scores. Those self-learned attention weights share the same spirit with the weighted translation pairs in Champollion [46]. During the adversarial training process, the NMT model leans to treat these important words carefully and avoid missing them in order to get higher score with the judgment of D. This process can be considered as transferring lexical knowledge from D to NMT model.

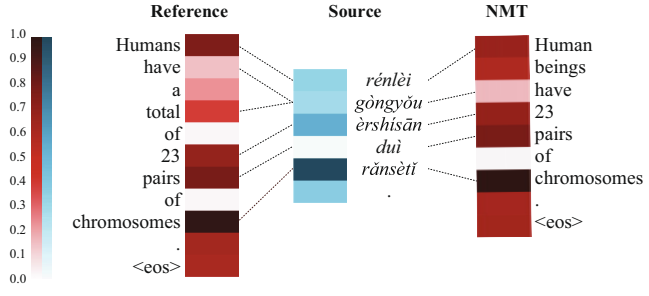


Fig. 3. Example of the self-attention weights for the source (blue) and the target (red) language encoders. Sentences in the example are selected from NIST2003 Zh → En test set. "Reference" and "NMT" represent ground-truth and Transformer generated target sentences. All weights in this example are scaled by Min-Max scaling method for better visualization. The darker color represents higher attention weights. Aligned words are manually connected by dashed lines. See section 6.1 for more details. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We also give averaged sentence alignment scores between translations and source inputs estimated by D on different model setups in Table 4. The results show that the output alignment scores of the proposed methods are higher than the baseline models, which illustrates that G learns the knowledge on measuring alignment from D successfully under the proposed training and decoding frameworks.

6.2. Translations re-ranking by D

Since D is trained independently in our framework, there is a doubt that whether the discriminator can distinguish good and bad translation candidates generated by the NMT model correctly. Therefore, to verify whether an individual discriminator is suitable for the model generated data, we conduct further experiments on translation re-ranking task. The experiments are based on the baseline Transformer [7] model, and we use Zh → En corpus as the dataset. Moreover, in order to obtain more translation candidates, we expand the beam size to 24 and then re-order the N-best translation candidates by D. Experimental results are shown in Table 5. As prior observations in previous work [53] we can observe that the larger beam search size leads to the worse performance, since the likelihood score for decoding tends to score short translations higher than long sentences. Larger searching space also brings more good translation candidates, and D re-orders them using alignment score and gains better BLEU scores than most baseline setups as shown in Table 5. The above observations indicate that D can profitably handle the unseen data generated by NMT models. Previous work [5,53] introduces length normalization to solve the above beam search decoding problem, whose results are also presented in Table 5 for a fair comparison.

Table 4
Averaged sentence alignment scores on Zh → En, Ug → Zh and En → De test sets. The higher score represents the better sentence alignment quality under D's view. See section 6.1 for more details.

Model setups	Zh → En	Ug → Zh	En → De
Transformer	11.15	10.04	13.24
+decoding w/ D	11.17	10.05	13.28
+frozen D + O_{MLE}	11.31	**10.43**	13.35
+adversarial D +O_{MLE}	**11.34**	10.29	**13.36**

Table 5
BLEU scores on Zh → En translation re-ranking task. The "beam N" represents the decoding beam search size. The "+length penalty" means using length normalization [5] when performing beam search. The "+D re-ranking" represents that the translation candidates are re-ranked by D. See section 6.2 for more details.

Setups	NIST2002	NIST2003	NIST2004
Transformer (beam 6)	41.56	39.95	42.05
Transformer (beam 24)	40.72	38.64	41.13
+length penalty	41.93	**40.26**	42.49
+D re-ranking	**42.22**	40.20	**42.56**

6.3. Translation performance over source lengths

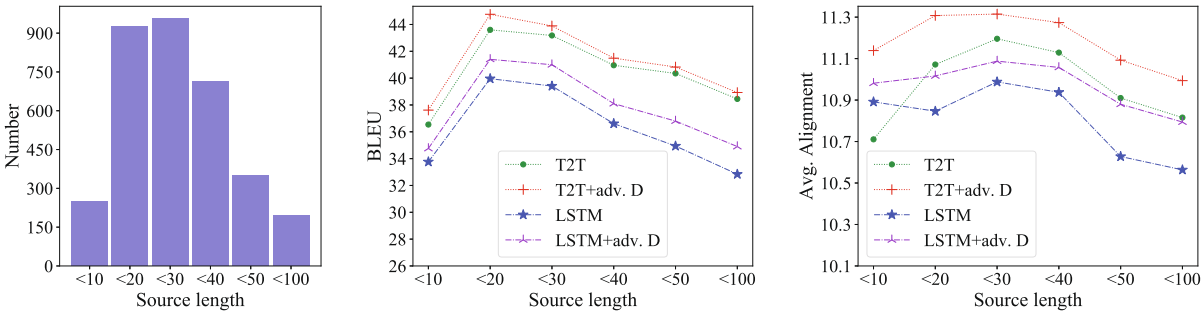
Translating long sentences is a well-known challenge for NMT [53]. We take the translation results on NIST2002∼2004 testset as examples to show translation performance over different source sentence lengths. We group source sentences in similar lengths and compute both the BLEU score and average alignment score for each group, as shown in Fig. 4. Fig. 4a shows the source sequence length distribution. We observe that the length of source sentences concentrate between 10-40. Specifically, length spans contains 251, 926, 956, 714, 350, and 196 sentences, respectively.

From the perspective of the BLEU score (as shown in Fig. 4b), the proposed models ("+adv. D") outperform both Transformer (i.e. T2T) and LSTM baselines in all length of sets. Fig. 4b also presents that T2T-based methods gain higher BLEU scores than LSTM-based methods over all lengths. While translating sentences longer than 30 tokens, the decrease in the translation performance of the LSTM-based approach is more obvious than that of the Transformer-based models, and our proposed method alleviates the drawbacks to a certain extent.

As for alignment score, in Fig. 4c, it is less sensitive for variations in sentence length than BLEU, and the proposed approaches also gain better performance than baselines models. Though the alignment score degrades significantly when the length of source sentence is longer than 40 for all model setups, the proposed methods can alleviate the downward trend of scores effectively. There is an interesting finding that the Transformer baseline gains the lowest score when source sentence is less than 10 tokens, which is different from the trend of BLEU (as shown in Fig. 4b). We speculate that the reason for the above phenomenon is that comparing to the abstract attention of multi-layer Transformer, shallow and specific attention of LSTM-based model works better on short sentences while translating some key words.

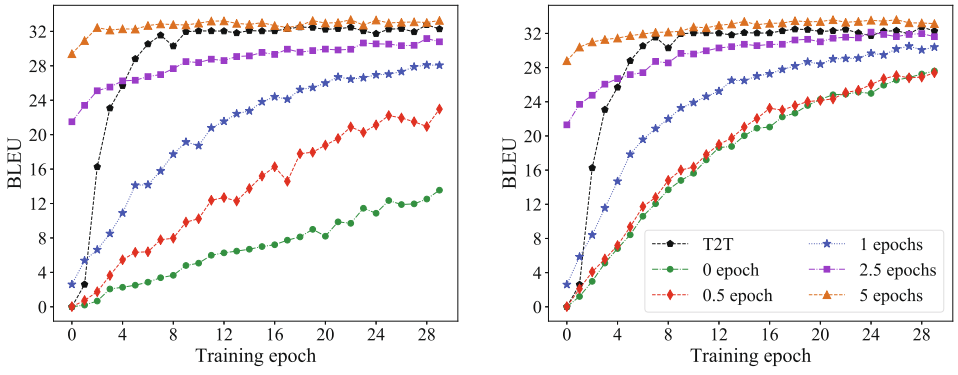
6.4. Initial pre-training steps

As mentioned in section 5.3, we pre-train the NMT model in our approach until the translation performance on the validate set no longer improves. A natural question is that whether the approach is end-to-end trainable. The number of the initial pre-training steps of the NMT model can be viewed as a hyper-parameter. We study the impacts of translation performance on our approaches ("+frozen *D*" and "+adversarial *D*") on Ug → Zh validate set. For "+froze *D*", we apply a well pre-trained discriminator with fixed parameters to the adversarial training, while for "+adversarial *D*", we adopt the discriminator that shares the same pre-training steps with the NMT model. We compare our approaches on five different initial pre-training epochs (0, 0.5, 1, 2.5, 5), and the experimental results are shown in Fig. 5.



(a) Length distribution (b) Length-BLEU (c) Length-Alignment

Fig. 4. BLEU and Alignment scores of the translations with respect to the input lengths on Zh → En NIST2002∼2004 testset. (a) is a histogram of the length distribution of the source sequence. For both (b) and (c), the x-axis represents the sequence length of source inputs, and the y-axis represents the BLEU and average alignment score respectively. "+adv. D" is an abbreviation of "+adversarial *D*". See section 6.3 for more details.



(a) +frozen D (b) +adversarial D

Fig. 5. BLEU score on Ug → Zh testset for "+frozen *D*" and "+adversarial *D*" with different initial pre-training steps. "T2T" represents the Transformer baseline. The number in the legend represents the pre-training epochs for the NMT model to perform warm start (e.g. "0" means no warm start). We use epochs instead of training steps as the x-axis since the baseline and our approaches use different training batch sizes. See section 6.4 for more details.

Benefit from the teacher-forcing step, our proposed approaches can be trained without a warm start, though the training is inefficient before the NMT model is well pre-trained. As Fig. 5 shows, a well pre-trained initial NMT model helps the proposed approach get the best performance on the validate set with less epochs. Furthermore, for each training epoch, the time costs of the proposed methods are 15 times more than that of the Transformer. The reason for inefficient end-to-end training is that for the adversarial training step, the training batch size is small and it is more computationally intensive than baseline setups.

On the other hand, for the discriminator, pre-training using *N*-pair loss makes *D* learn to distinguish positive samples from various candidates. Under end-to-end training, *D* will be trained on 1-pair loss function, which has negative effects on *D*'s training efficiency. We suppose that it is another reason why the model performance improves slower with less initial pre-training steps.

Comparing Fig. 5a) and (b), we find that the training performance of "+adversarial *D*" is more stable and rises more rapidly than that of "+frozen *D*", especially for the models with few pre-training steps. We suspect the reasons for the above phenomenons are:

- Models with few pre-training steps usually generate inferior translations, which are certainly different with the training data of *D*. So that the discriminator in "+frozen *D*" cannot provide a rational loss signal for *G*. On the contrast, "+adversarial *D*" uses a trainable discriminator, which makes *D* adapt to the outputs of the model quickly during the training process. As a result, the slopes of the most curves in Fig. 5b are higher than the curves in Fig. 5a.
- In "+adversarial *D*", we continue updating *D* with the NMT outputs and *D* will fit *G*'s outputs during the whole training process. Therefore, the curves in Fig. 5b are smoother than those in Fig. 5a with smaller epoch numbers. For methods with more pre-training steps ("epoch 5" and "epoch 10"), the curves is smooth for both of "+frozen *D*" and "+adversarial *D*", which illustrates that the well pre-trained *D* can handle those translations better.

6.5. Example translations

We provide example translations on Zh → En translation task in Fig. 6. Although the translation result of the Transformer in Fig. 6 is correct in syntax, its logic is wrong on account of missing important source information of "*x=ıngq=ılìù fù xi=anggǎng qǔdé qi=anzhèng* (went to Hong Kong on Saturday for obtaining a visa)". All the translations generated by our proposed methods do not make the baseline's mistake, since they learn the knowledge from the discriminator that the verbs "*fù* (go to)" and "*qǔdé* (obtain)", and the nouns '*x=ıngq=ılìù* (Saturday)' and "*qi=anzhèng* (visa)" are important lexical evidences. We also show a translation re-ranking example, which gains a similar result to other proposed methods. An alignment score evaluated by the discriminator and a sentence-level BLEU score measured by *sentence-bleu*[1] are also shown under the corresponding translations. Both the golden reference and the model generated translations gain higher alignment score from *D*, which illustrates the rationality of the discriminator design.

7. Conclusion

In this paper, we propose an adversarial training framework and an alignment-aware decoding method to address the inadequacy translation problem in NMT. The proposed method can achieve sentence alignment oriented knowledge transfer and improve the

Source	*chénjīndé **xīngqīliù fù** xiānggǎng **qǔdé qiānzhèng** , zuótiān dǐ jīng fǎngwèn 10 tiān .*
Reference	Chen Chin-teh **went to** Hong Kong **on Saturday for his visa** and arrived in Beijing yesterday for his 10-day visit . *(align: 11.15)*
Transformer	Chen Jinde arrived in Beijing yesterday for a 10-day visit to Hong Kong . *(align: 9.91, BLEU: 28.33)*
+decoding w/ *D*	Chen Jinde **obtained a visa** in Hong Kong **on Saturday** and yesterday arrived in Beijing for a 10-day visit . *(align: 10.69, BLEU: 29.23)*
+frozen *D* re-ranking	after **receiving a visa** in Hong Kong **on Saturday** , Chen Jinde arrived in Beijing yesterday for a 10-day visit . *(align: 10.75, BLEU: 39.33)*
+frozen *D* $+O_{MLE}$	Chen Jinde **went to** Hong Kong **to obtain a visa on Saturday** and yesterday arrived in Beijing for a 10-day visit . *(align: 10.97, BLEU: 29.55)*
+adversarial *D* $+O_{MLE}$	Chen Jinde **went to** Hong Kong **for a visa on Saturday** and yesterday arrived in Beijing for a 10-day visit . *(align: 10.92, BLEU: 37.49)*

Fig. 6. Example translations on the Zh → En translation task. The example is selected from the NIST2002 test set. "Source" and "Reference" are the source input and one of the four given references. **Words in red bold fonts** represent the missing part of the translation generated by the baseline model. An alignment score (*al1ign*) and a sentence-level BLEU are given below the target sentence. See section 6.5 for more details.

translation adequacy in NMT. We design a discriminator to measure sentence alignment by mainly considering lexical evidence via a gated self-attention mechanism. A discriminative loss as well as a teacher-forcing objective are used to constrain NMT model to generate sufficient and fluent translations during the training procedure. The sentence alignment-aware decoding method also guides the NMT to achieve adequate translation. Experimental results on three different language pairs show that our proposed approach outperforms the conventional NMT models. Further analysis indicates the proposed discriminator captures the weighted lexical relationships among sentences well and transfers the knowledge to the NMT model successfully.

In the future, we would like to make the discriminator learn more semantic related knowledge including the word relative and absolute position, dependency, and semantic role. Employing richer semantic knowledge can help the discriminator judge the translation quality more equitably comparing to only considering the key word alignments. We will also integrate our proposed method with existing adversarial training techniques [16,17,10] and study the interaction between different discriminators.

CRediT authorship contribution statement

Xuewen Shi: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft. **Heyan Huang:** Conceptualization, Resources, Supervision, Funding acquisition, Writing - review & editing. **Ping Jian:** Supervision, Funding acquisition, Writing - review & editing. **Yi-Kun Tang:** Software, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all reviewers for their valuable comments. This work is supported by the National Key Research and Development Program of China (Grant No. 2017YFB1002103) and the National Natural Science Foundation of China (No. 61732005).

References

[1] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 1700–1709.
[2] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder–decoder approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 103–111.
[3] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inf. Process. Syst. 27 (2014) 3104–3112.
[4] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: 3rd International Conference on Learning Representations, ICLR 2015, 2015.
[5] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, CoRR abs/1609.08144. arXiv:1609.08144.
[6] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y.N. Dauphin, Convolutional sequence to sequence learning, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, 2017, pp. 1243–1252.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 6000–6010.
[8] Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li, Modeling coverage for neural machine translation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 76–85.
[9] S. Feng, S. Liu, N. Yang, M. Li, M. Zhou, K.Q. Zhu, Improving attention modeling with implicit distortion and fertility for machine translation, in: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11–16, 2016, Osaka, Japan, 2016, pp. 3082–3092.
[10] X. Kong, Z. Tu, S. Shi, E.H. Hovy, T. Zhang, Neural machine translation with adequacy-oriented learning, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 2019, pp. 6618–6625.
[11] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, Adv. Neural Inf. Process. Syst. 29 (2016) 1849–1857.
[12] E. Jang, S. Gu, B. Poole, Categorical reparameterization with gumbel-softmax, in: 5th International Conference on Learning Representations, ICLR 2017, 2017.
[13] M.J. Kusner, J.M. Hernández-Lobato, GANS for sequences of discrete elements with the gumbel-softmax distribution, CoRR abs/1611.04051. arXiv:1611.04051.
[14] D. He, H. Lu, Y. Xia, T. Qin, L. Wang, T. Liu, Decoding with value networks for neural machine translation, Adv. Neural Inf. Process. Syst. 30 (2017) 178–187.
[15] J. Lu, A. Kannan, J. Yang, D. Parikh, D. Batra, Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model, Adv. Neural Inf. Process. Syst. 30 (2017) 313–323.
[16] L. Yu, W. Zhang, J. Wang, Y. Yu, Seqgan Sequence generative adversarial nets with policy gradient, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 2852–2858.
[17] Z. Yang, W. Chen, F. Wang, B. Xu, Generative adversarial training for neural machine translation, Neurocomputing 321 (2018) 146–155.
[18] L. Wu, Y. Xia, F. Tian, L. Zhao, T. Qin, J. Lai, T. Liu, Adversarial neural machine translation, in: Proceedings of The 10th Asian Conference on Machine Learning, 2018, pp. 534–549.
[19] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, 2017, pp. 933–941.
[20] J. Chorowski, D. Bahdanau, K. Cho, Y. Bengio, End-to-end continuous speech recognition using attention-based recurrent NN: first results, CoRR abs/1412.1602. arXiv:1412.1602.
[21] A.M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, 2015, pp. 379–389.
[22] K.M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, Adv. Neural Inf. Process. Syst. 28 (2015) 1693–1701.
[23] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1–15.
[24] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1448–1457.
[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on Machine Learning, Vol. 37 of Proceedings of Machine Learning Research, PMLR, Lille, France, 2015, pp. 2048–2057.
[26] W. Wang, X. Lu, J. Shen, D.J. Crandall, L. Shao, Zero-shot video object segmentation via attentive graph neural networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 9236–9245.
[27] T. Zhou, W. Wang, S. Qi, H. Ling, J. Shen, Cascaded human-object interaction recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[28] P. Sountsov, S. Sarawagi, Length bias in encoder decoder models and a case for global conditioning, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1516–1525.
[29] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1724–1734.
[30] T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1412–1421.
[31] H. Choi, K. Cho, Y. Bengio, Fine-grained attention mechanism for neural machine translation, Neurocomputing 284 (2018) 171–176.
[32] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, 2017, pp. 933–941.
[33] Z. Tu, Y. Liu, L. Shang, X. Liu, H. Li, Neural machine translation with reconstruction, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3097–3103.
[34] Z. Zheng, H. Zhou, S. Huang, L. Mou, X. Dai, J. Chen, Z. Tu, Modeling past and future for neural machine translation, TACL 6 (2018) 145–157.

[35] Z. Zheng, S. Huang, Z. Tu, X.-Y. Dai, J. Chen, Dynamic past and future for neural machine translation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 931–941.
[36] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, W. Ma, Dual learning for machine translation, Adv. Neural Inf. Process. Syst. 29 (2016) 820–828.
[37] Y. Xia, T. Qin, W. Chen, J. Bian, N. Yu, T. Liu, Dual supervised learning, in: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, 2017, pp. 3789–3798.
[38] Y. Xia, J. Bian, T. Qin, N. Yu, T. Liu, Dual inference for machine learning, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 3112–3118.
[39] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial networks, CoRR abs/1406.2661. arXiv:1406.2661.
[40] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[41] L.J. Ba, R. Kiros, G.E. Hinton, Layer normalization, CoRR abs/1607.06450. arXiv:1607.06450.
[42] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, D. Jurafsky, Adversarial learning for neural dialogue generation, in, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2157–2169.
[43] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses open source toolkit for statistical machine translation, in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 177–180.
[44] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725.
[45] W. Che, Z. Li, T. Liu, LTP: A chinese language technology platform, in: COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, 2010, pp. 13–16.
[46] X. Ma, Champollion A robust parallel text sentence aligner, in, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, 2006, pp. 489–492.
[47] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J.M.F. Moura, D. Parikh, D. Batra, Visual dialog, IEEE Trans. Pattern Anal. Mach. Intell. 41 (5) (2019) 1242–1256.
[48] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[49] Z. Zheng, W. Wang, S. Qi, S. Zhu, Reasoning visual dialogs with structural and partial observations, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 6669–6678.
[50] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015.
[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2818–2826.
[52] M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, in: 5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings, 2017.
[53] P. Koehn, R. Knowles, Six challenges for neural machine translation, in: Proceedings of the First Workshop on Neural Machine Translation, 2017, pp. 28–39.

Xuewen Shi received the B.S. degree from Hunan University in 2013, and now he is a Ph.D. candidate of computer science at Beijing Institute of Technology.

Heyan Huang is currently professor and dean of School of Computer Science and Technology in Beijing Institute of Technology of China. She received her Ph.D. degree in Computer Science and Technology in 1989 from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her current research interests include machine learning, information retrieval, and natural language processing. She has published over 100 research papers in reputed journals and conferences, such as TKDE, IJCAI, AAAI, ACL, WWW, COLING. She serves on the editorial boards of International Journal of Advanced Intelligence and Journal of Computer Research and Development. She has undertaken 20 more research projects including National 863 Project of China, National 973 Project of China, National Natural Science Foundation of China, etc.

Ping Jian is a full-time lecturer in the Department of Computer Science and Technology, Beijing Institute of Technology, China. She received her Ph.D. degree in pattern recognition and intelligent systems from Graduate University of Chinese Academy of Science in 2010. Her area of research includes natural language syntactic parsing and discourse analysis, which connect to the fields of natural language understanding, machine learning and pattern recognition.

Yi-Kun Tang received the B.S. degree from Beijing Institute of Technology in 2016, and now she a Ph.D. candidate of computer science at Beijing Institute of Technology.