Improving Neural Machine Translation by Achieving Knowledge Transfer with Sentence Alignment Learning

Xuewen Shi[1,2], **Heyan Huang**[1,2], **Wenguan Wang**[1,3], **Ping Jian**[1,2]*, **Yi-Kun Tang**[1,2]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
[3]Inception Institute of Artificial Intelligence, UAE
{xwshi, hhy63, wenguanwang, pjian, tangyk}@bit.edu.cn

Abstract

Neural Machine Translation (NMT) optimized by Maximum Likelihood Estimation (MLE) 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 learned from bilingual sentence alignment. Specifically, we first design a discriminator that learns to estimate sentence aligning score over translation candidates, and then the learned semantic knowledge is transfered to the NMT model under an adversarial learning framework. We also propose a gated self-attention based encoder for sentence embedding. Furthermore, an N-pair training loss is introduced in our framework to aid the discriminator in better capturing lexical evidence in translation candidates. Experimental results show that our proposed method outperforms baseline NMT models on Chinese-to-English and English-to-German translation tasks. Further analysis also indicates the detailed semantic knowledge transfered from the discriminator to the NMT model.

1 Introduction

Recently, with the renaissance of deep learning, end-to-end Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014a; Sutskever et al., 2014; Bahdanau et al., 2014) has gained remarkable performance (Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017). Early NMT solutions are typically optimized to maximize 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 in the NMT model, due to the lack of quantitative measurement for the information transformational completeness from the source side to the target side.

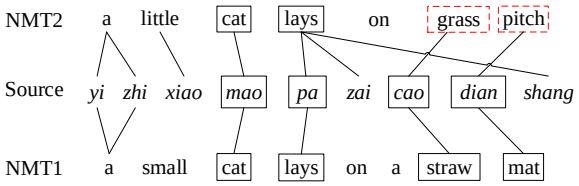

Figure 1: **Comparison between two Chinese-to-English translation examples of two independent NMT systems.** Lines between Source and NMTs represent model generated alignments (each source word cannot be covered more than once). Words in boxes are key words and red dotted dashed boxes indicate incorrect translations. Based on the model generated attention weights, NMT2 covers more source words than NMT1, which is opposite to human judgments.

Some existing work alleviates this problem by directly incorporating coverage or fertility mechanisms to an NMT model (Tu et al., 2016; Feng et al., 2016; Kong et al., 2019). However, the problem is that attention weights based coverage calculation for NMT is insensitive and sometimes even inaccurate to translation errors. Furthermore, it is unreasonable to consider the coverage of all kinds of source words equally, since various words contribute differently to sentences in semantics and syntax. For example, as illustrated in Fig. 1, translation errors are recorded as positive examples, and the alignments between function words also dilute the impact of key words alignments.

In this paper, we address the problem of inadequate translation by introducing a novel sentence alignment constrain under an adversarial training framework (Goodfellow et al., 2014; Lu et al., 2017; Yang et al., 2018). Specifically, our approach contains two sub-models: i) a sentence alignment oriented discriminator D learns to estimate the alignment score and sort the translation candidates by mainly considering the weighted

*Corresponding author: Ping Jian

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 260–270
Hong Kong, China, November 3-4, 2019.

alignment pairs (Ma, 2006) at a sentence representation level; and ii) a standard NMT model G aims to produce an appropriate translation with the highest ranking score (assigned by D) in the candidate list. To better capture the semantic alignment evidence of the input data, we also propose a novel gated self-attention based encoder for bilingual sentences encoding in discriminator D. Then, an N-pair training loss (Sohn, 2016) is introduced to select appropriate translation results from the candidates. We also leverage Gumbel-Softmax (GS) (Jang et al., 2017; Kusner and Hernández-Lobato, 2016) approximation for G to solve the problem of discrete samples, making the response from D to G differentiable.

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

- We apply 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 semantic knowledge learned by a discriminator for sentence aligning, which can be viewed as an instance of "knowledge transfer".

- Benefited from the gated self-attention mechanism, the proposed encoder learns to focus on important lexical evidence for sentence aligning and enhance the contribution of the key words. This knowledge will be transfered to G through the proposed framework.

- The N-pair loss (Sohn, 2016; Lu et al., 2017) encourages samples closed to the gold-standard one to get higher score. Unlike a binary classification used in previous work (Yu et al., 2017; Yang et al., 2018; Wu et al., 2018), translations that are correct but different from the ground-truth ones will not be over penalized.

We use one of the state-of-the-art NMT models, Transformer (Vaswani et al., 2017), as the baseline model architecture and conduct experiments on Chinese-to-English and German-to-English translation tasks. Experimental results show that our proposed approach achieves significant improvements on both language pairs. We also evaluate the performance of the discriminator on both sentence alignment and translation candidate reranking tasks, which proves its independence and transferability. Further analyses show the specific alignment-oriented knowledge that the discriminator transfers to the NMT model.

2 Related Work

Most of the state-of-the-art NMT models are optimized by MLE-based objectives (Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017), but likelihood fails to measure whether the source information is completely transformed to the target side. Thus, it cannot handle translation adequacy problem (Tu et al., 2017).

One way to alleviate these problems is to apply coverage and fertility to NMT model. Feng et al. (2016) aim at controlling the fertilities of source words by appending additional additive terms to train objectives. Tu et al. (2016) employ coverage vector or coverage ratio as a lexical-level indicator to represent whether a source word is translated or not.

On the other hand, some recent efforts introduce additional source side constraints and explore duality properties of NMT (He et al., 2016; Cheng et al., 2016; Xia et al., 2017; Tu et al., 2017). Cheng et al. (2016) present a semi-supervised approach to train bidirectional NMT models and reconstruct the monolingual corpora using an auto-encoder (Socher et al., 2011). Tu et al. (2017) add a re-constructor to traditional NMT model, which introduces an auxiliary score to measure the adequacy of translation. Dual learning (He et al., 2016) and dual supervised learning (Xia et al., 2017) 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.

GAN (Goodfellow et al., 2014) is another promising framework to leverage sentence-level objectives in NMT. Recently, there is some remarkable work in NMT (Wu et al., 2018; Yang et al., 2018). 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. A policy gradient method is leveraged to co-train the NMT model and the discriminator. However,

those approaches rarely take account of translation adequacy. Furthermore, the discriminators of those work refer the target sentence in the corpus as the single gold-standard regardless the quality of model generated translations, which will punish too much to the good model generated translations. Kong et al. (2019) 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 (Wu et al., 2018; Yang et al., 2018; Kong et al., 2019), 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 for training D to ensure that D will not punish the translations closed to the gold-standard overly.

3 Approach

In this section, we describe our approach that can transfer knowledge from a sentence alignment oriented discriminator D to an NMT model G. Our approach mainly consists of two sub-models: i) 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 Fig. 2: for each sentence pair (X, Y) sampled from the training corpus, the NMT model G generates a translation $\hat{\mathrm{Y}}$ given X, and queries the discriminator D with $\hat{\mathrm{Y}}$ to get feedback and update itself. In order to obtain more stable training, we also leverage a teacher-forcing (Li et al., 2017) step to our approach.

3.1 NMT Generator

In this paper, we take the Transformer (Vaswani et al., 2017), one of the popular state-of-the-art NMT models, as the specific implementation of the NMT model G. This helps to better illustrate the effectiveness of the proposed method. The Transformer in this paper follows the conventional encoder-decoder framework (Cho et al., 2014b). Specifically, the encoder contains a stack of six 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. A residual connection is applied around each of the two sub-layers, followed by layer normalization (Ba et al., 2016). The decoder is also composed of a stack of six identical layers. 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.

Following the base model setups of the Transformer (Vaswani et al., 2017), we use 8 attention heads, 512-dimensional output vectors for each layer, and 2048-dimensional inner-layer of the feed-forward network.

1: Pre-train a generator G (see section 3.1) and a discriminator D (see section 3.2), individually.
2: **for** number of training iterations **do**
3: Sample $(\mathrm{X}, \mathrm{Y}^{+})$ from training corpus
4: Sample $\hat{\mathrm{Y}} \sim G(\mathrm{X})$ with a Gumbel-Softmax sampler (see section 3.4)
5: Compute loss $\mathcal{L}_G$ for $(\mathrm{X}, \hat{\mathrm{Y}})$ using D (see section 3.3)
6: Update G with the learning rate η:
$$\theta_G \leftarrow \theta_G - \eta \bigtriangledown_{\theta_G} \mathcal{L}_G \quad (1)$$
7: Teacher-Forcing: update G on $(\mathrm{X}, \mathrm{Y}^{+})$ (see section 3.5)
8: **end for**

Figure 2: A brief overview of the proposed training framework. See section 3 for more details.

3.2 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 encodings pair.

Gated Self-Attention Sentence Encoder. As depicted in Fig. 3, we opt a shallow network architecture: one gated hidden layer and one self-attention layer as the sentence encoder. This lightweight encoder mainly captures the lexical meanings of the sentence. The self-attention mechanism helps the encoder select more impor-

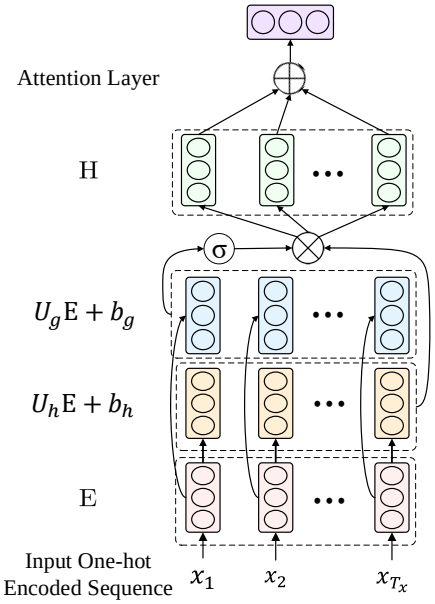

Figure 3: **Model architecture of the gated self-attention sentence encoder.** See section 3.2 for more details.

tant lexical evidences to estimate the alignment score between two sentences.

Given a one-hot encoded input sequence X and a word embedding lookup table $\mathrm{D} \in \mathbb{R}^{|V|\times d}$, where d is the model dimension. The input X will be represented as a corresponding word embedding matrix $\mathrm{E} \in \mathbb{R}^{\mathrm{T_x}\times \mathrm{d}}$. We apply a gating mechanism (Dauphin et al., 2017) to compute the hidden layer H:

$$\mathrm{H} = (U_h\mathrm{E} + b_h) \otimes \sigma(U_g\mathrm{E} + b_g), \tag{2}$$

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

$$\mathrm{W} = \mathrm{softmax}(\tanh(U_a\mathrm{H})), \tag{3}$$

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

$$e = U_o(\mathrm{W} \times \mathrm{H}) + b_o, \tag{4}$$

where $e \in \mathbb{R}^d$, and $U_o \in \mathbb{R}^{d\times d}$. We add layer normalization (Ba et al., 2016) to the output layer. The model dimension d is set to 512.

Alignment Score and Discriminator Loss. With the source and target sentence encodings e_x and e_y, the alignment score $s_{(\mathrm{X},\mathrm{Y})}$ can be computed as:

$$s_{(\mathrm{X},\mathrm{Y})} = e_x^\top e_y. \tag{5}$$

Given the candidate target sentences list $\mathcal{Y}$, the discriminator produces a distribution over $\mathcal{Y}$ 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 that are closed to the gold-standard one. Therefore, following Lu et al. (2017), we apply a metric-learning multi-class N-pair loss (Sohn, 2016) to our model, which can be defined as:

$$\begin{aligned}\mathcal{L}_D &= \mathcal{L}_{N-\mathrm{pair}}(\{\mathrm{X}, \mathrm{Y}^+, \{\mathrm{Y}_n^-\}_{n=1}^{N-1}\}) \\ &= \log(1 + \sum_{n=1}^{N} \exp(s_{(\mathrm{X},\mathrm{Y}_\mathrm{n}^-)} - s_{(\mathrm{X},\mathrm{Y}^+)})),\end{aligned} \tag{6}$$

where Y^+ is alignment target sentence to the source language X, and Y_n^- is one of the $N-1$ unaligned samples.

Compared to cross entropy loss used in previous work (Yu et al., 2017; Yang et al., 2018; Wu et al., 2018), the N-pair objective encourages the score of target sentences which are similar to the given golden-standard one to be higher than the dissimilar ones. In this way, translations that are correct but different from the ground truth will not be over penalized, and thus this can be useful to provide a reliable signal for the generator.

In later sections, we will analyze the semantic information learned by the model through some visualization examples, shown in section 5.1, and the experimental results show that it achieves sufficient accuracy for scoring the alignment between source and target sentences.

3.3 Discriminative Losses for Generative Training

In our framework, G aims to generate a translation score higher than the golden-standard, under the premise of encoders and the scoring function learned by D. Specifically, for each sentence pairs $(\mathrm{X}, \mathrm{Y}^+)$ in training sets, first, G samples translation $\hat{\mathrm{Y}}$ given X with greedy searching. Second, D takes $\hat{\mathrm{Y}}$ as well as $(\mathrm{X}, \mathrm{Y}^+)$ as inputs to compute alignment scores, and then G gets the feedback from D. Eq. (7) gives the perceptual loss that G aims to optimize.

$$\begin{aligned}\mathcal{L}_G &= \mathcal{L}_{1-pair}(\{\mathrm{X}, \mathrm{Y}^+, \hat{\mathrm{Y}}\}) \\ &= \log(1 + \exp(s_{(\mathrm{X},\mathrm{Y}^+)} - s_{(\mathrm{X},\hat{\mathrm{Y}})})).\end{aligned} \tag{7}$$

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

3.4 Gumbel-Softmax Sampler

The process of sampling a translation $\hat{\mathrm{Y}}$ with G is not differentiable, since it includes $\mathrm{argmax}(\cdot)$ operator to perform one-hot encoding. We leverage the Gumbel-softmax (Jang et al., 2017) sampler to solve this problem. Formally, at the decoding step j, suppose that $p_j \in \mathbb{R}^{V_y}$ contains the model output log-probabilities over target vocabulary, and $g_j \in \mathbb{R}^{V_y}$ includes i.i.d samples drawn from the standard distribution $\mathrm{Gumbel}(0, 1)$. A sample y_j is transformed as:

$$\hat{y}_j = \mathrm{softmax}((p_j + g_j)/\tau), \tag{8}$$

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

3.5 Teacher-forcing Step

Since $\mathcal{L}_G$ in Eq. (7) mainly considers the discrepancy of alignment and integrity between the model output and the ground-truth, it rarely inspects grammar correctness and language fluency. To alleviate this problem, following (Li et al., 2017; Lu et al., 2017), we adopt the similar teacher-forcing step to our training process.

We perform two different teacher-forcing objectives for comparison: i) a likelihood objective $\mathrm{O_{LM}}$ and ii) a BLEU score reward ($\mathrm{R_{BLEU}}$), under the training strategies of MLE and MRT (Shen et al., 2016), respectively.

4 Experiments

4.1 Datasets and Setups[1]

We evaluate the proposed approach on Chinese-to-English (Zh-En) and English-to-German (En-De) translation tasks. For both of the two translation tasks, we tokenize all corpora with the Moses tokenizer[2]. Sentences longer than 100 words are discarded, and all the sentences are encoded with byte-pair encoding (bpe) (Sennrich et al., 2016).

Chinese-to-English. For Chinese-to-English translation, our training data are extracted from four LDC corpora[3]. The training set contains totally 1.3M parallel sentence pairs. For preprocessing, the Chinese part for both training and testing sets is segmented by the LTP Chinese word segmentor (Che et al., 2010) before applying bpe (Sennrich et al., 2016) to the corpus. 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. In the following parts of the paper, the Chinese examples are presented by segmented italic romanized form, and different Chinese characters are delimited by single quotation marks.

English-to-German. For English-to-German translation, we conduct experiments on the publicly available corpora WMT'14 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. Different from parallel sentence pairs for training generative models, the corpus for training D needs to provide a candidate translations list for each source sentence. Therefore, we need to manually construct the corpus for training D using the original parallel corpus. For each source sentence, we set the size of candidate list to 100. The translation candidates are preferentially obtained from the context of the golden standard translation in the comparable paragraph. If the context sentence number N_c is less than 99, we will randomly sample another $99 - N_c$ sentences from the whole rest target corpus. As for the data format, we follow most of Das et al. (2019).

Evaluation. As for generative models, following Vaswani et al. (2017), we report the result of a single model obtained by averaging the 5 checkpoints around the best model selected on the development set. We apply beam search during decoding with the beam size of 6. The translation results in this paper are measured in case-insensitive BLEU (Papineni et al., 2002) by the

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

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

[3]LDC2005T10, LDC2003E14, LDC2004T08 and LDC2002E18. Since LDC2003E14 is a document-level alignment comparable corpus, we use Champollion Tool Kit (Ma, 2006) to extract parallel sentence pairs from it.

multi-bleu.perl script[4]. For the discriminator, the performance is evaluated on recall@k and mean rank score.

4.2 Training Details

Pre-train discriminator and NMT model. During the training procedure, we first pre-train the discriminator and the generator separately for a warm start. We pre-train the model until the performance of D and G on development set dose not improve. For training discriminator on all language pairs, we use the Adam 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 follow the base model of Transformer (Vaswani et al., 2017) for most of training setups, except the label smoothing (Szegedy et al., 2016).

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 learned with a combination discriminative perceptual loss (Lu et al., 2017) and teacher-forcing loss (Li et al., 2017; Lu et al., 2017). Each mini-batch contains 30 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 (Yang et al., 2018; Wu et al., 2018; Kong et al., 2019). An adversarial D is to maximize the score of the human translation Y^+ and minimize the score of the generated translation $\hat{\mathrm{Y}}$. Then the training loss for adversarial D can be represented as: $\mathcal{L}_D = -\mathcal{L}_G$.

4.3 Machine Translation Results

We report the experimental results on machine translation in this section. Table 1 shows the BLEU scores of Zh-En and En-De translation tasks. Our approach achieves an improvement up to +0.76 BLEU points averagely on Zh-En testing sets and +0.64 BLEU points on En-De testing set. It should be noted that we do not apply label smoothing (Szegedy et al., 2016) due to using Gumbel-Softmax approximation, which results in a decline in En-De translation performance compared to the reported result of BLEU 27.3 in Vaswani et al. (2017).

We compare two setups of frozen D (Row 3-4) and adversarial D (Row 5-6) for the discriminator. Experimental results show that continuing to update D along with G gains best BLEU score for the both translation tasks. It means that fine-tuning D with the model generated data can further improve training quality. We also evaluate two different training objectives for the teacher-forcing step (Row 3,5 vs Row 4,6) and Row 2 is another baseline for training only on $\mathrm{R}_{\mathrm{BLEU}}$ (similar to MRT (Shen et al., 2016)). We can see from the results that applying MLE or BLEU reward does not make much difference. These results indicate that the proposed method makes up for the shortcomings of MLE training.

4.4 Discriminative Results

The format of the test datasets for discriminator is similar to the training set described in section 4.1, where each input corresponds to one hundred candidate translations extracted from the document context. The goal of the discriminator is to rank the correct translation as high as possible. We present recall@k and mean rank of the discriminator on Zh-En and En-De test sets in Table 2. It shows that for all test sets of both 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. Both of the high recall@k and ranking mean closed to 1 indicate 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 may lead to worse model performance for GANs (Salimans et al., 2016; Arjovsky and Bottou, 2017; Yang et al., 2018). In this paper, G is trained with a specific 1-pair loss defined on sentence alignment score, instead of Jensen-Shannon divergence (Arjovsky and Bottou, 2017; Arjovsky et al., 2017) between two data distributions, which could avoid the vanishing gradient problem in GANs. Therefore, the high accuracy of the proposed discriminator would not make negative impact on G.

5 Analysis

In this section, we will study characteristics of the proposed approach and report some detailed ex-

[4] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

#	Model	Zh-En				En-De
		NIST2002	NIST2003	NIST2004	Average	newstest2014
1	Transformer	41.56	39.95	42.05	41.19	26.52
2	+$\mathrm{R}_{\mathrm{BLEU}}$	42.30	40.47	42.46	41.74	26.73
3	+ frozen D + O_{LM}	42.51	**40.74**	42.42	41.89	27.10
4	+ frozen D + $\mathrm{R}_{\mathrm{BLEU}}$	41.63	40.06	42.16	41.28	26.77
5	+ adversarial D +O_{LM}	42.67	40.67	42.51	**41.95**	**27.16**
6	+ adversarial D +$\mathrm{R}_{\mathrm{BLEU}}$	**42.75**	40.28	**42.67**	41.90	27.05

Table 1: **BLEU scores on Zh-En and En-De translation task.** Transformer is the baseline model. "Average" is the averaged BLEU scores on testing sets. Following Vaswani et al. (2017), we report the result of a single model obtained by averaging the 5 checkpoints around the best model selected on the development set. See section 4.3 for more details.

Testset	R@1	R@5	R@10	Mean
NIST2002	97.04	99.77	99.89	1.06
NIST2003	97.50	99.34	99.67	1.10
NIST2004	97.25	99.94	99.94	1.05
newstest2014	96.60	99.43	99.73	1.12

Table 2: **Discriminator performance on Zh-En and En-De test sets.** R@k and Mean are abbreviations for recall@k score and mean rank score, respectively. See section 4.4 for more details.

Model setups	Align
Transformer	11.15
+$\mathrm{R}_{\mathrm{BLEU}}$	11.18
+ frozen D + O_{LM}	11.31
+ frozen D + $\mathrm{R}_{\mathrm{BLEU}}$	**11.38**
+ adversarial D +O_{LM}	11.34
+ adversarial D +$\mathrm{R}_{\mathrm{BLEU}}$	11.36

Table 3: **Averaged sentence alignment scores on Zh-En NIST2002∼2004 test sets.** "Align" means the averaged sentence alignment score estimated by D. The higher score represents the better alignment quality in D's view. See section 5.1 for more details.

perimental results. We also give a specific translation example to illustrate how knowledge transferring improves NMT performance.

5.1 What kind of Knowledge Does D Transfer to G?

In this paper, we propose a discriminator that directly learns to measure alignment and then transfers the learned knowledge to an NMT model. D is designed to capture lexical evidence for sentence alignment by learning a self-attention encoder. We give averaged sentence alignment scores between translations and source inputs on different model setups in Table 3. Those alignment scores are estimated by a pre-trained D. Table 3 shows that the output alignment scores of the proposed approaches are all higher than the baseline methods, which illustrates that G can learn the knowledge on measuring alignment from D under the proposed training frameworks.

In order to illustrate the lexical-level knowledge learned by D, we give a visual example in Fig. 4. It shows self-attention weights of the encoders for the given source and the target sentence. In the example, the source sentence is "*bao'wei'er 12'ri yu sha'long ju'xing le hui'tan*" and the target sentence is "Powell hold a talk with Sharon on the 12th ." We notice that the source language words "*bao'wei'er*", "*12'ri*" and "*sha'long*", and their corresponding target language words "Powell", "12th" and "Sharon" are assigned higher attention weights than others. This means that the encoders regard those words as important lexical evidences for estimating the alignment score. Those self-learned attention weights share the same spirit with the weighted translation pairs in Champollion (Ma, 2006). During the training process, G leans to treat those important words carefully and avoid missing them to get higher score with the judgment of D. This process can be considered as transferring the semantic knowledge leaned by D to G.

5.2 Can discriminator distinguish good and bad translation results?

Since D is trained independently in our framework, it is difficult to estimate whether the discriminator can correctly distinguish the good and bad translations generated by the NMT model. Therefore, to verify whether an individual discriminator is suitable for the model generated data,

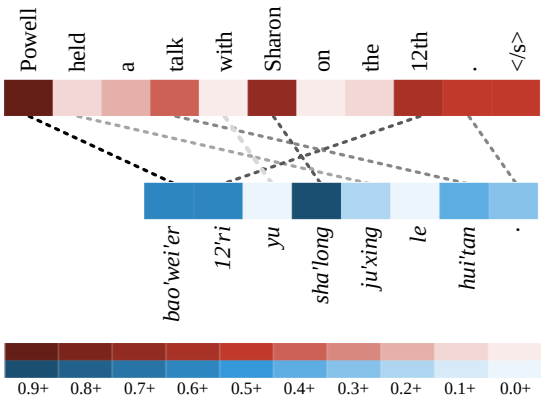

Figure 4: **Example of the self-attention weights for the source (lower) and the target (upper) language encoders.** Sentences in the example are selected from NIST2002 Zh-En test set. All weights in this example are scaled by Min-Max scaling method for better visualization and darker colors represent higher attention weights. Aligned words are manually connected by dashed lines. See section 5.1 for more details.

we conduct further experiments on translation re-ranking task on the baseline Transformer (Vaswani et al., 2017) model on Zh-En translation. 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 4. 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 by alignment score and gains better BLEU scores than most baseline setups as shown in Table 4. The above observation indicates that D can successfully handle the unseen data generated by NMT models. Previous work (Wu et al., 2016; Koehn and Knowles, 2017) introduces length normalization to solve the above beam search decoding problem, whose results are also presented in Table 4 for a fair comparison.

5.3 Example Translations

We provide example translations on Zh-En translation task in Fig. 5. From Fig. 5, we can see that though the translation results of the baseline model is correct in syntax, its logic is wrong on account of missing an important source information of "went to Hong Kong on Saturday for a visa". All the translations generated by our proposed method do not make this mistake, since it is learned and transfered from the discriminator where the verb "went", nouns 'Saturday' and "visa" are important lexical evidence for estimating alignment score. 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[5] (Papineni et al., 2002) score 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 discriminator design.

Setups	NIST02	NIST03	NIST04
beam 6	41.56	39.95	42.05
beam 24	40.72	38.64	41.13
+length penalty	41.93	**40.26**	42.49
+re-ranking	**42.22**	40.20	**42.56**

Table 4: **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 (Wu et al., 2016) when performing beam search. The "+re-ranking" represents that the translation candidates are re-ranked by D. See section 5.2 for more details.

6 Conclusion

In this work, we propose a novel training framework which achieves sentence alignment oriented knowledge transfer to improve the NMT. We design a discriminator to measure sentence alignment by mainly considering lexical evidence via a gated self-attention mechanism. Then, a discriminative loss as well as a teacher-forcing objective is used to make NMT model generate sufficient and fluent translations during training procedure. Experimental results on different language pairs show that our proposed approach outperforms standard NMT models. Further analysis indicates the proposed discriminator well captures the weighted lexical relationships among sentences and successfully transfers the knowledge to the NMT model.

In the future, we would like to make discriminator learn more semantic related knowledge like dependency, and combine our approach with other advanced techniques in reinforcement learning and adversarial learning (Yu et al., 2017; Yang et al., 2018; Kong et al., 2019).

[5]Evaluated by Moses (Koehn et al., 2007) sentence-bleu script.

Source	*chen'jin'de* ***xing'qi'liu fu*** *xiang'gang* ***qu'de qian'zheng*** *, zuo'tian di jing fang'wen 10 tian .*
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)*
+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: 39.32)*
+frozen D +O_{LM}	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)*
+frozen D +R_{BLEU}	Chen Jinde **went to** Hong Kong **on Saturday to obtain a visa** , and yesterday arrived in Beijing for a 10-day visit . *(align: 10.99, BLEU: 37.49)*
+adversarial D +O_{LM}	Chen Jinde **went to** Hong Kong **on Saturday to obtain a visa** and yesterday arrived in Beijing for a 10-day visit . *(align: 10.92, BLEU: 40.19)*
+adversarial D +R_{BLEU}	Chen Jinde **went to** Hong Kong **on Saturday for a visa** , and yesterday arrived in Beijing for a 10-day visit . *(align: 11.01, BLEU: 44.53)*

Figure 5: **Example translations on the Zh-En translation task.** The example is selected from the NIST2002 testing 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. A alignment score (*align*) and a sentence-level BLEU are given below the target sentence. See section 5.3 for more details.

Acknowledgments

We thank all anonymous reviewers for their valuable comments. This work was 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

Martín Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR*, abs/1701.07875.

Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A chinese language technology platform. In *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China*, pages 13–16.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2019. Visual dialog. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(5):1242–1256.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. 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*, pages 933–941.

Shi Feng, Shujie Liu, Nan Yang, Mu Li, Ming Zhou, and Kenny Q. Zhu. 2016. 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*, pages 3082–3092.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on*

Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 1243–1252.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *CoRR*, abs/1406.2661.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1700–1709.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39.

Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard H. Hovy, and Tong Zhang. 2019. 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, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 6618–6625.

Matt J. Kusner and José Miguel Hernández-Lobato. 2016. GANS for sequences of discrete elements with the gumbel-softmax distribution. *CoRR*, abs/1611.04051.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2157–2169.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 313–323.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 489–492.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 151–161.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing*

Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 1849–1857.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3097–3103.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.

Lijun Wu, Yingce Xia, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Adversarial neural machine translation. In *Proceedings of The 10th Asian Conference on Machine Learning, ACML 2018, Beijing, China, November 14-16, 2018.*, pages 534–549.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual supervised learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3789–3798.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Improving neural machine translation with conditional sequence generative adversarial nets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1346–1355.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 2852–2858.