Bidirectional Generative Adversarial Networks for Neural Machine Translation

Zhirui Zhang[1], Shujie Liu[2], Mu Li[2], Ming Zhou[2] and Enhong Chen[1]

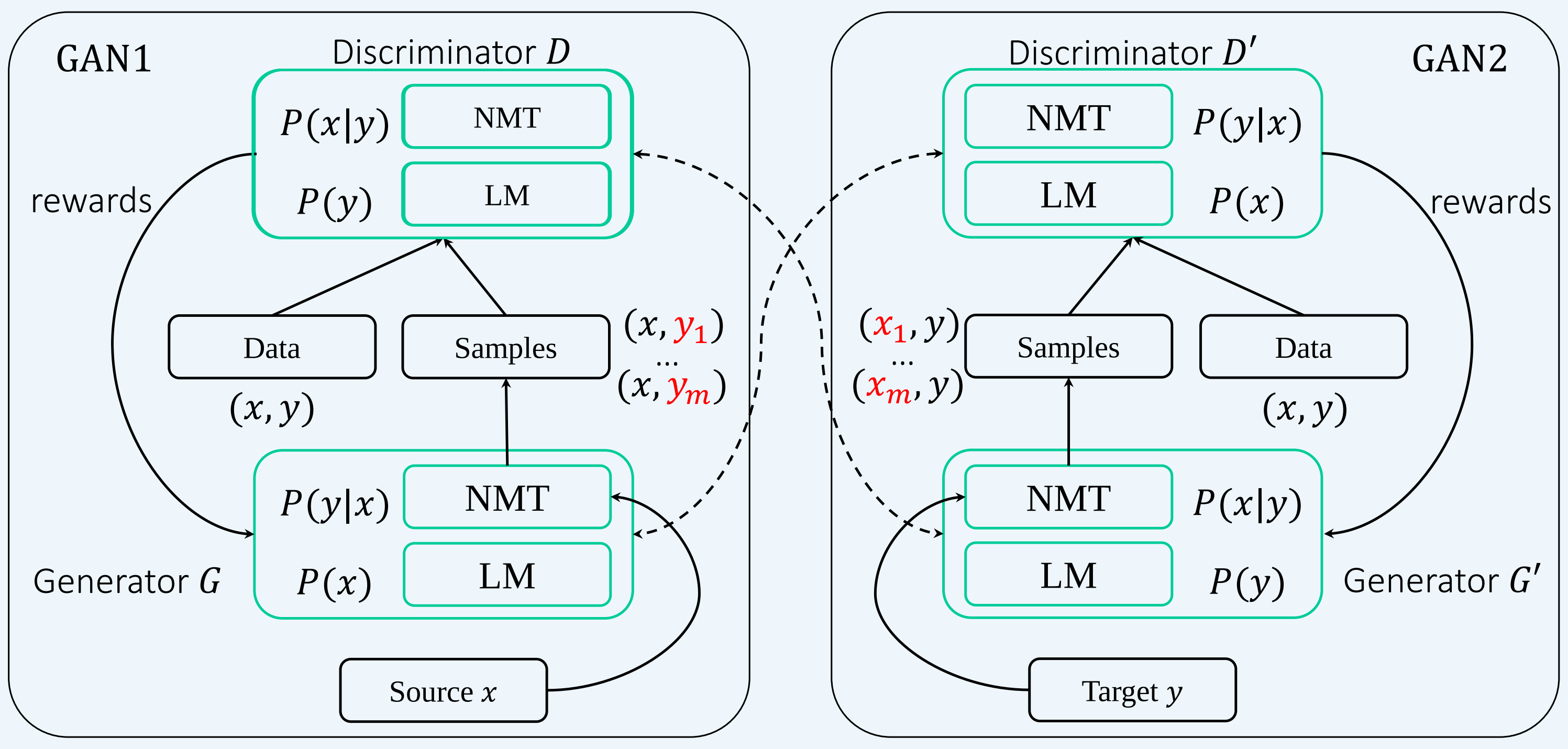
[1]University of Science and Technology of China [2]Microsoft Research Asia

Introduction

Generative Adversarial Network (GAN) has been proposed to tackle the exposure bias problem of Neural Machine Translation (NMT). However, the discriminator typically results in the instability of the GAN training due to the inadequate training problem: the search space is so huge that sampled translations are not sufficient for discriminator training. To address this issue and stabilize the GAN training, in this paper, we propose a novel Bidirectional Generative Adversarial Network for Neural Machine Translation (BGAN-NMT), which aims to introduce a generator model to act as the discriminator, whereby the discriminator naturally considers the entire translation space so that the inadequate training problem can be alleviated. To satisfy this property, generator and discriminator are both designed to model the joint probability of sentence pairs, with the difference that, the generator decomposes the joint probability with a source language model and a source-to-target translation model, while the discriminator is formulated as a target language model and a target-to-source translation model. To further leverage the symmetry of them, an auxiliary GAN is introduced and adopts generator and discriminator models of original one as its own discriminator and generator respectively. Two GANs are alternately trained to update the parameters. Experiment results on German-English and Chinese-English translation tasks demonstrate that our method not only stabilizes GAN training but also achieves significant improvements over baseline systems.

Bidirectional Generative Adversarial Network

The overall architecture of BGAN-NMT consists of an original GAN (GAN1) and an auxiliary GAN (GAN2):



- Both generator and discriminator of original GAN are defined to model the joint probability of sentence pairs $P(x,y)$. Since $P(x,y)$ can be decomposed into two ways: $P(x,y) = P(x)P(y|x)$ and $P(x,y) = P(y)P(x|y)$, they are used as generator G and discriminator D for GAN1 respectively.
- Auxiliary GAN (GAN2) employs G and D of GAN1 as its own discriminator D' and generator G' to better exploit the symmetry between G and D

Training Objective

During adversarial training, G and D of GAN1 play a two-player minmax game with the following value function $V(D,G)$:

$$\min_G \max_D V(D,G) = \mathrm{E}_{(x,y)\sim P_d(x,y)}[\log P(x|y)P(y)] + \mathrm{E}_{x\sim P_d(x), y'\sim P(y|x)}[\log(1 - P(x|y')P(y')]$$

- **The objective function of D is to maximize $V(D,G)$:**

$$L_D = \mathrm{E}_{(x,y)\sim P_d(x,y)}[\log P(x|y)P(y)] + \mathrm{E}_{x\sim P_d(x), y'\sim P(y|x)}[\log(1 - P(x|y')P(y')]$$

$$\Rightarrow \frac{\partial L_D}{\partial \theta_D} = \mathrm{E}_{(x,y)\sim P_d(x,y)}\left[\frac{\partial \log P(x|y)}{\partial \theta_D}\right] + \mathrm{E}_{x\sim P_d(x), y'\sim P(y|x)}\left[\left(1 - \frac{1}{1-P(x|y')P(y')}\right)\frac{\partial \log P(x|y')}{\partial \theta_D}\right] \quad (1)$$

- **The objective function of G is to maximize the expected rewards (the probability of D) instead of directly maximizing $V(D,G)$:**

$$\mathrm{L}_G = \mathrm{E}_{x\sim P_d(x), y'\sim P(y|x)}[P(x|y')P(y')] \Rightarrow \frac{\partial L_D}{\partial \theta_D} = \mathrm{E}_{x\sim P_d(x), y'\sim P(y|x)}\left[P(x|y')P(y')\frac{\partial \log P(y'|x)}{\partial \theta_G}\right] \quad (2)$$

Besides, GAN2 is designed by exchanging G and D of GAN1, so the update equations of D' and G' as following:

$$\frac{\partial L_{D\prime}}{\partial \theta_G} = \mathrm{E}_{(x,y)\sim P_d(x,y)}\left[\frac{\partial \log P(y|x)}{\partial \theta_G}\right] + \mathrm{E}_{y\sim P_d(y), x'\sim P(x|y)}\left[\left(1 - \frac{1}{1-P(y|x')P(x')}\right)\frac{\partial \log P(y|x')}{\partial \theta_G}\right] \quad (3)$$

$$\frac{\partial L_D}{\partial \theta_D} = \mathrm{E}_{y\sim P_d(y), x'\sim P(x|y)}\left[P(y|x')P(x')\frac{\partial \log P(x'|y)}{\partial \theta_D}\right] \quad (4)$$

Joint Training Algorithm

1. Pre-train $P(y|x)$ and $P(x|y)$ on bilingual data with MLE principle;
2. Get golden and generated samples from bilingual data and generator respectively, and then update discriminator models with Equation (1) and (3) ;
3. Get generated samples and their rewards from generator and discriminator respectively, and then update generator models with Equation (2) and (4);
4. Repeat step 2 and 3;

Experimental Results

- **Results on German-English Translation**

Methods	Baseline	Model
MIXER (Ranzato et al., 2015)	20.10	21.81
MRT (Shen et al., 2016)	-	25.84
BSO (Wiseman and Rush, 2016)	24.03	26.36
Adversarial-NMT (Wu et al., 2017)	-	27.94
A-C (Bahdanau et al., 2016)	27.56	28.53
Softmax-Q (Ma et al., 2017)	27.66	28.77
Adversarial-NMT*	27.63	28.03
BGAN-NMT	27.63	**29.17**

Table 1: Comparison with previous work on IWSLT2014 German-English translation task. The "Baseline" means the performance of pre-trained model used to warmly start training.

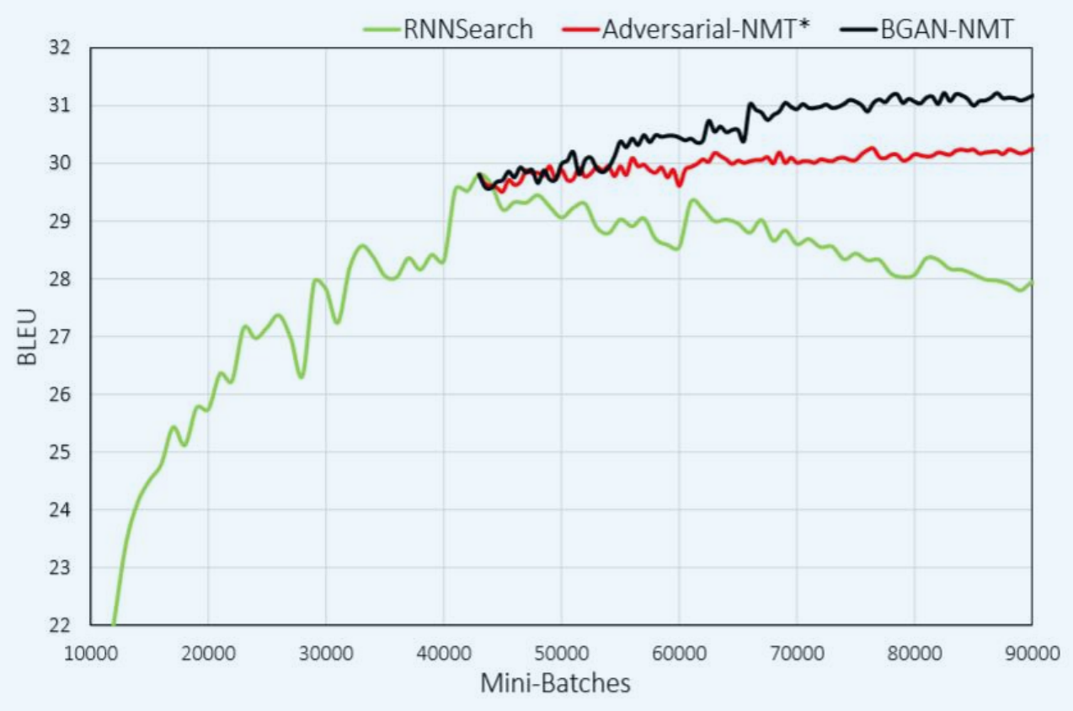


Figure 2: The BLEU score changes on IWSLT2014 German-English validation set for RNNSearch, Adversarial-NMT* and BGAN-NMT as training progresses.

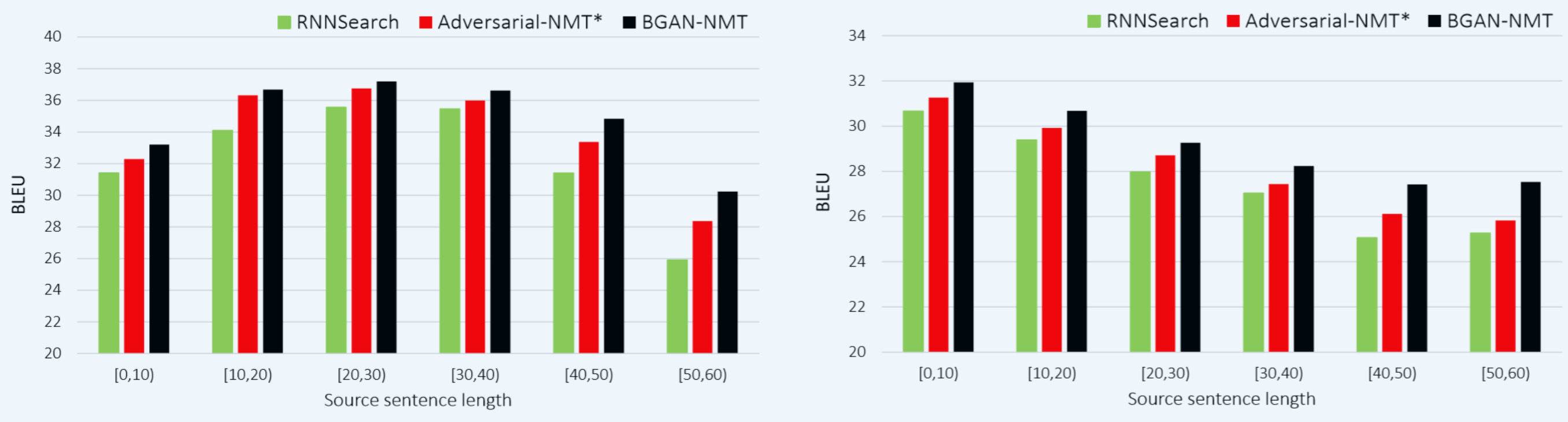
- **Results on Chinese-English Translation**

System	NIST2006	NIST2005	NIST2008	NIST2012	Average
HPSMT	32.46	32.42	25.23	26.20	29.08
RNNSearch	38.61	38.31	30.04	28.48	33.86
Adversarial-NMT*	39.79	38.81	31.86	30.19	35.16
BGAN-NMT	**40.74**	**39.20**	**33.55**	**31.30**	**36.19**

Table 2: Case-insensitive BLEU scores (%) on Chinese-English translation. The "Average" denotes the average results of all datasets.

Analysis

- **Effect on Long Sentences**



(a) German-English Translation

(b) Chinese-English Translation

Figure 3: Performance of the generated translations with respect to the length of source sentences on different datasets. For Chinese-English, we merge all NIST datasets in this experiment. For German-English, we only use test datasets.

- **Effect of Discriminative Loss**

Model	DE-EN	ZH-EN
BGAN-NMT	29.17	36.19
-Discriminative Loss	28.59	35.46

Table 3: Translation performance of BGAN-NMT without discriminative loss on German-English (DE-EN) and Chinese-English (ZH-EN) translations. The BLEU score for Chinese-English translation is the average results of all datasets we used in the experiment.

Conclusion

- In this paper, we have presented a Bidirectional Generative Adversarial Network for Neural Machine Translation. The entire framework consists of an original GAN and an auxiliary GAN, in which generator and discriminator are designed to model the joint probability of sentence pairs.
- In the future, we plan to extend this method to other sequence-to-sequence NLP tasks.