# Dual Learning for Machine Translations

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

# Cycle GAN [Zhu et al., 2017]

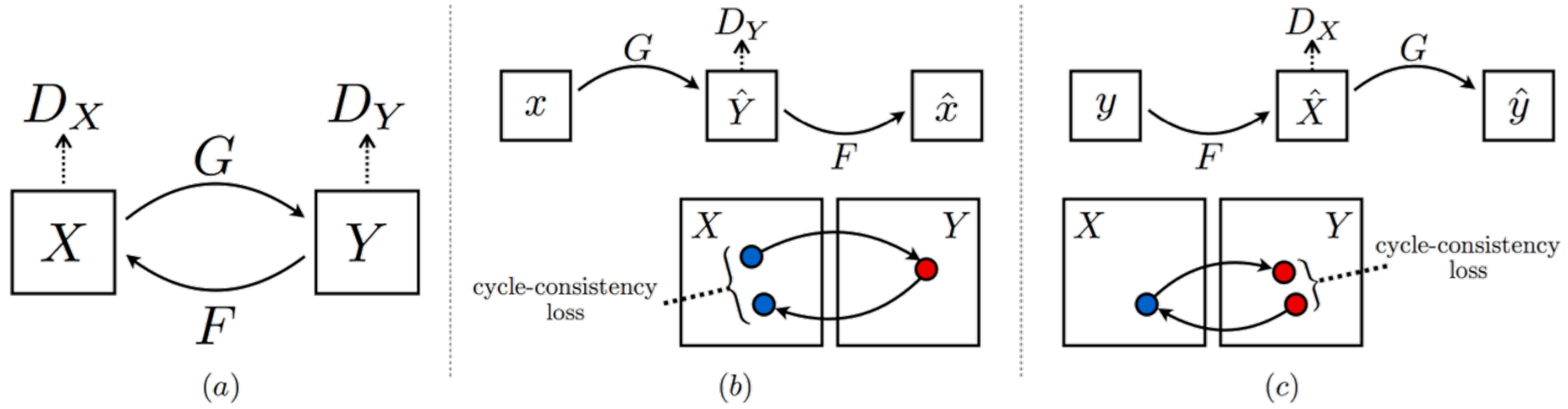
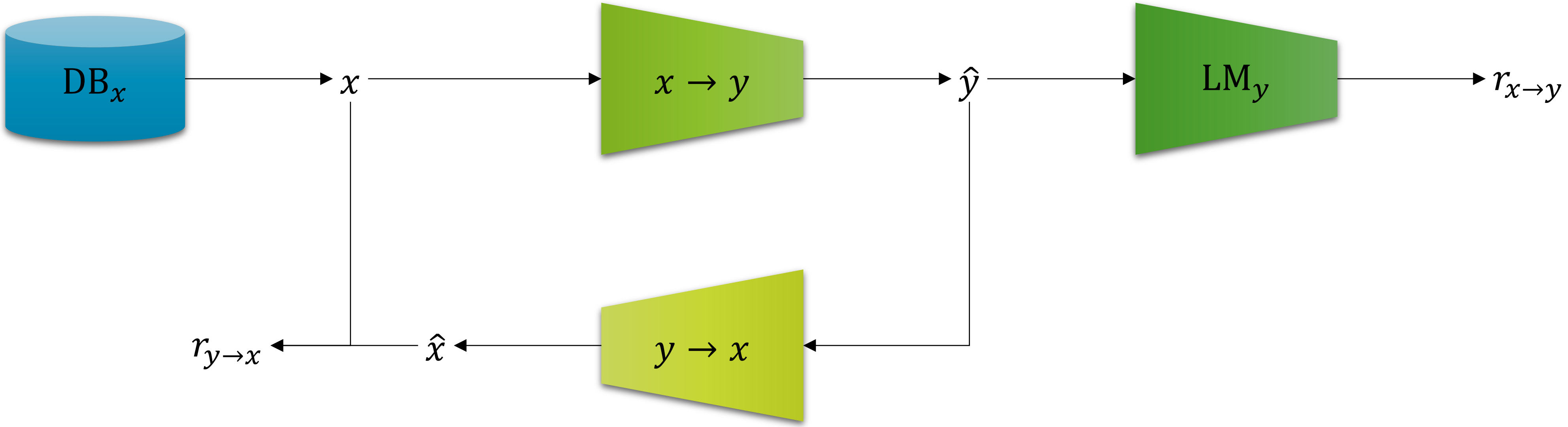



Figure 3: (a) Our model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators $D_Y$ and $D_X$. $D_Y$ encourages $G$ to translate $X$ into outputs indistinguishable from domain $Y$, and vice versa for $D_X$ and $F$. To further regularize the mappings, we introduce two *cycle consistency losses* that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

# Dual Learning for Machine Translations [Xia et al., 2016]

- Using monolingual corpus, fine-tune both pretrained models.



# Equations

- Using Policy Gradients:

$$r = \alpha \times r_{x \rightarrow y} + (1 - \alpha) \times r_{y \rightarrow x}$$

$$r_{x \rightarrow y} = P(\hat{y}), \text{ where } \hat{y} \sim P(\mathrm{y}|x; \theta_{x \rightarrow y})$$ ← **Reinforcement Learning**

$$r_{y \rightarrow x} = \log P(x|\hat{y}; \theta_{y \rightarrow x})$$ ← **MLE using Back Translation?**

$$\theta_{x \rightarrow y} \leftarrow \theta_{x \rightarrow y} - \eta \frac{1}{K} \sum_{k=1}^{K} \left[ r_k \nabla_{\theta_{x \rightarrow y}} \log P(\hat{y}_k | x; \theta_{x \rightarrow y}) \right]$$

$$\theta_{y \rightarrow x} \leftarrow \theta_{y \rightarrow x} - \eta \frac{1}{K} \sum_{k=1}^{K} \left[ (1 - \alpha) \nabla_{\theta_{y \rightarrow x}} \log P(x | \hat{y}_k; \theta_{y \rightarrow x}) \right]$$

# Summary

- 소량의 parallel corpus와 다량의 monolingual corpus가 있을 때, Dual learning을 통해 성능을 큰 폭으로 개선할 수 있음
  - RL에 기반하고 있는 점은 아쉬움

Table 1: Translation results of En↔Fr task. The results of the experiments using all the parallel data for training are provided in the first two columns (marked by "Large"), and the results using 10% parallel data for training are in the last two columns (marked by "Small").

| | En→Fr (Large) | Fr→En (Large) | En→Fr (Small) | Fr→En (Small) |
|---|---|---|---|---|
| NMT | 29.92 | 27.49 | 25.32 | 22.27 |
| pseudo-NMT | 30.40 | 27.66 | 25.63 | 23.24 |
| dual-NMT | **32.06** | **29.78** | **28.73** | **27.50** |

[Xia et al., 2016]