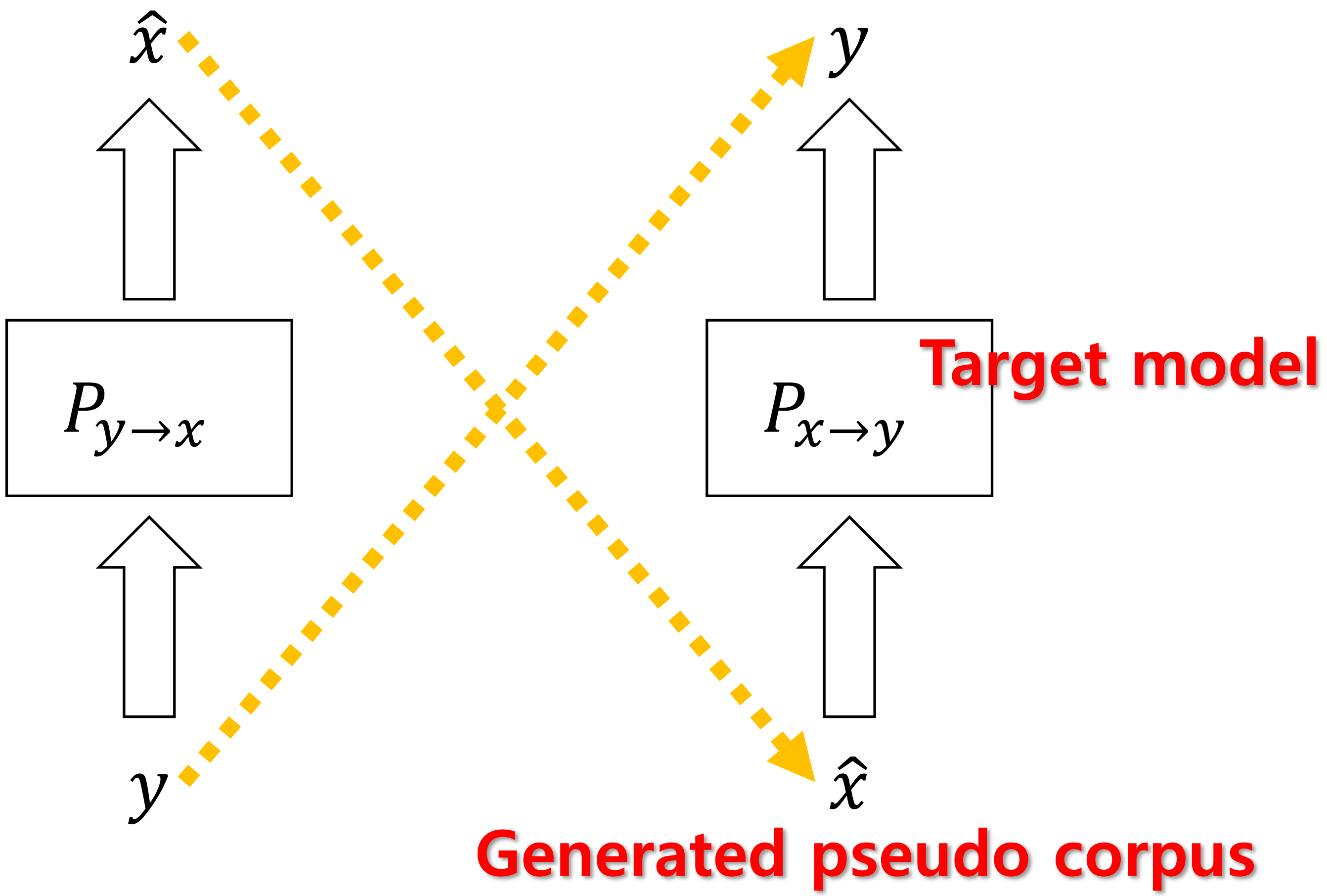
Appendix:
Back Translation Review

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Back Translation

- 보통 번역은 두 개의 모델이 동시에 나오기 마련
  - 반대쪽 모델을 활용하여 synthetic corpus를 만들 수 있음



Equations

- Given datasets:

$$\mathcal{B} = \{x^n, y^n\}_{n=1}^{N}$$
$$\mathcal{M} = \{y^s\}_{s=1}^{S}$$

- We need to minimize:

$$\mathcal{L}(\theta_{x\rightarrow y}) = -\sum_{n=1}^{N} \log P(y^n | x^n; \theta_{x\rightarrow y}) - \sum_{s=1}^{S} \log P(y^s)$$

Equations

- By Marginal Distribution:

$$
\begin{aligned}
\log P(y) &= \log \sum_{x \in \mathcal{X}} P(x, y) \\
&= \log \sum_{x \in \mathcal{X}} P(y|x)P(x) \\
&= \log \sum_{x \in \mathcal{X}} \frac{P(y|x)P(x)}{P(x|y)} P(x|y)
\end{aligned}
$$

Equations

- By Jenson's Inequality,

$$
\begin{aligned}
\log P(y) &= \log \sum_{x \in \mathcal{X}} P(x, y) \\
&= \log \sum_{x \in \mathcal{X}} P(y|x)P(x) \\
&= \log \sum_{x \in \mathcal{X}} \frac{P(y|x)P(x)}{P(x|y)} P(x|y) \\
&\geq \sum_{x \in \mathcal{X}} P(x|y) \log \frac{P(y|x)P(x)}{P(x|y)} \\
&= \mathbb{E}_{x \sim P(\mathrm{x}|y)} \left[\log \frac{P(y|x)P(x)}{P(x|y)} \right] \\
&= \mathbb{E}_{x \sim P(\mathrm{x}|y)} \big[\log P(y|x) \big] - \mathrm{KL}\big(P(\mathrm{x}|y) | P(\mathrm{x})\big)
\end{aligned}
$$

Equations

- Re-write the objective:

$$
\begin{aligned}
\mathcal{L}(\theta_{x\rightarrow y}) &\le -\sum_{n=1}^{N}\log P(y^n|x^n;\theta_{x\rightarrow y}) - \sum_{s=1}^{S}\Big(\mathbb{E}_{x\sim P(\mathrm{x}|y;\theta_{y\rightarrow x})}\big[\log P(y^s|x;\theta_{x\rightarrow y})\big] - \mathrm{KL}\big(P(\mathrm{x}|y^s;\theta_{y\rightarrow x})|P(\mathrm{x})\big)\Big) \\
&\approx -\sum_{n=1}^{N}\log P(y^n|x^n;\theta_{x\rightarrow y}) - \sum_{s=1}^{S}\left(\frac{1}{K}\sum_{k=1}^{K}\log P(y^s|x_k^s;\theta_{x\rightarrow y}) - \mathrm{KL}\big(P(\mathrm{x}|y^s;\theta_{y\rightarrow x})|P(\mathrm{x})\big)\right), \text{where } x_k^s \sim P(\mathrm{x}|y^s;\theta_{y\rightarrow x}) \\
&= \tilde{\mathcal{L}}(\theta_{x\rightarrow y})
\end{aligned}
$$

Equations

- If we get derivative of the loss:

$$
\begin{aligned}
\mathcal{L}(\theta_{x\rightarrow y}) &\le -\sum_{n=1}^{N}\log P(y^n|x^n;\theta_{x\rightarrow y}) - \sum_{s=1}^{S}\Big(\mathbb{E}_{x\sim P(\mathrm{x}|y;\theta_{y\rightarrow x})}\big[\log P(y^s|x;\theta_{x\rightarrow y})\big] - \text{KL}\big(P(\mathrm{x}|y^s;\theta_{y\rightarrow x})|P(\mathrm{x})\big)\Big) \\
&\approx -\sum_{n=1}^{N}\log P(y^n|x^n;\theta_{x\rightarrow y}) - \sum_{s=1}^{S}\left(\frac{1}{K}\sum_{k=1}^{K}\log P(y^s|x_k^s;\theta_{x\rightarrow y}) - \text{KL}\big(P(\mathrm{x}|y^s;\theta_{y\rightarrow x})|P(\mathrm{x})\big)\right), \text{where } x_k^s \sim P(\mathrm{x}|y^s;\theta_{y\rightarrow x}) \\
&= \tilde{\mathcal{L}}(\theta_{x\rightarrow y})
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\theta_{x\rightarrow y}}\tilde{\mathcal{L}}(\theta_{x\rightarrow y}) &= -\nabla_{\theta_{x\rightarrow y}}\sum_{n=1}^{N}\log P(y^n|x^n;\theta_{x\rightarrow y}) - \nabla_{\theta_{x\rightarrow y}}\frac{1}{K}\sum_{s=1}^{S}\sum_{k=1}^{K}\log P(y^s|x_k^s;\theta_{x\rightarrow y}) \\
&\approx -\nabla_{\theta_{x\rightarrow y}}\sum_{n=1}^{N}\log P(y^n|x^n;\theta_{x\rightarrow y}) - \nabla_{\theta_{x\rightarrow y}}\sum_{s=1}^{S}\log P(y^s|\hat{x}^s;\theta_{x\rightarrow y}), \text{where } \hat{x}^s \sim P(\mathrm{x}|y^s;\theta_{y\rightarrow x}) \text{ and } K=1.
\end{aligned}
$$

$$
\theta_{x\rightarrow y} \leftarrow \theta_{x\rightarrow y} - \eta\nabla_{\theta_{x\rightarrow y}}\tilde{\mathcal{L}}(\theta_{x\rightarrow y})
$$

Summary

- Back Translation을 수학적으로 다시 해석함
  - 기존 BT 방법이 정당성을 얻게 됨