Appendix: Back Translation Review

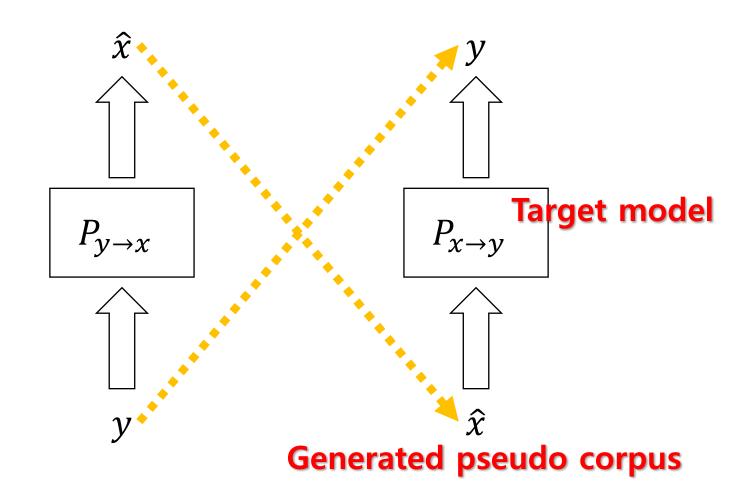
Ki Hyun Kim

nlp.with.deep.learning@gmail.com



Back Translation

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



Given datasets:

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

• We need to minimize:

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

• By Marginal Distribution:

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

By Jenson's Inequality,

$$egin{aligned} \log P(y) &= \log \sum_{x \in \mathcal{X}} P(x,y) \ &= \log \sum_{x \in \mathcal{X}} P(y|x) P(x) \ &= \log \sum_{x \in \mathcal{X}} rac{P(y|x) P(x)}{P(x|y)} P(x|y) \ &\geq \sum_{x \in \mathcal{X}} P(x|y) \log rac{P(y|x) P(x)}{P(x|y)} \ &= \mathbb{E}_{x \sim P(\mathbf{x}|y)} \Big[\log rac{P(y|x) P(x)}{P(x|y)} \Big] \ &= \mathbb{E}_{x \sim P(\mathbf{x}|y)} \Big[\log P(y|x) \Big] - \mathrm{KL} \Big(P(\mathbf{x}|y) | P(\mathbf{x}) \Big) \end{aligned}$$

• Re-write the objective:

$$egin{aligned} \mathcal{L}(heta_{x o y}) &\leq -\sum_{n=1}^N \log P(y^n|x^n; heta_{x o y}) - \sum_{s=1}^S \left(\mathbb{E}_{x\sim P(\mathbf{x}|y; heta_{y o x})}ig[\log P(y^s|x; heta_{x o y})ig] - \mathrm{KL}ig(P(\mathbf{x}|y^s; heta_{y o x})|P(\mathbf{x})ig)
ight) \ &pprox - \sum_{n=1}^N \log P(y^n|x^n; heta_{x o y}) - \sum_{s=1}^S \left(rac{1}{K}\sum_{k=1}^K \log P(y^s|x_k^s; heta_{x o y}) - \mathrm{KL}ig(P(\mathbf{x}|y^s; heta_{y o x})|P(\mathbf{x})ig)
ight), ext{ where } x_k^s \sim P(\mathbf{x}|y^s; heta_{y o x}) \ &= ilde{\mathcal{L}}(heta_{x o y}) \end{aligned}$$

If we get derivative of the loss:

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

$$egin{aligned}
abla_{ heta_{x o y}} ilde{\mathcal{L}}(heta_{x o y}) &= -
abla_{ heta_{x o y}} \sum_{n=1}^N \log P(y^n|x^n; heta_{x o y}) -
abla_{ heta_{x o y}} rac{1}{K} \sum_{s=1}^S \sum_{k=1}^K \log P(y^s|x_k^s; heta_{x o y}) \ &pprox -
abla_{ heta_{x o y}} \sum_{n=1}^N \log P(y^n|x^n; heta_{x o y}) -
abla_{ heta_{x o y}} \sum_{s=1}^S \log P(y^s|\hat{x}^s; heta_{x o y}), ext{ where } \hat{x}^s \sim P(ext{x}|y^s; heta_{y o x}) ext{ and } K = 1. \end{aligned}$$

$$heta_{x
ightarrow y} \leftarrow heta_{x
ightarrow y} - \eta
abla_{ heta_{x
ightarrow y}} ilde{\mathcal{L}}(heta_{x
ightarrow y})$$

Summary

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