Dual Supervised Learning

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Motivations

- NLG는 auto-regressive task로, teacher forcing을 통해 학습
  - 학습과 추론 방법 사이에 괴리가 생겨, 성능이 저하될 수 있음

- MRT(RL)은 괴리가 없으나, 샘플링 기반 방식이므로 매우 비효율적임

- MLE 방식 위에서, regularization을 통해 문제를 풀 수는 없을까?

Bayes Theorem

- From Bayes theorem:

$$P(\mathrm{y}|\mathrm{x}) = \frac{P(\mathrm{x}|\mathrm{y})P(\mathrm{y})}{P(\mathrm{x})}$$

- Given dataset, it would be always true,

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

$$P(y^n|x^n)P(x^n) = P(x^n|y^n)P(y^n)$$
$$\log P(y^n|x^n) + \log P(x^n) = \log P(x^n|y^n) + \log P(y^n)$$

Equations

- Objectives:

$$\hat{\theta}_{x\rightarrow y} = \underset{\theta_{x\rightarrow y}\in\Theta}{\text{argmin}}\sum_{i=1}^{N}\ell\big(f(x^i;\theta_{x\rightarrow y}),y^i\big)$$

$$\hat{\theta}_{y\rightarrow x} = \underset{\theta_{y\rightarrow x}\in\Theta}{\text{argmin}}\sum_{i=1}^{N}\ell\big(f(y^i;\theta_{y\rightarrow x}),x^i\big)$$

$$\text{s.t. } P(y^i|x^i)P(x^i) = P(x^i|y^i)P(y^i).$$

Equations

- New Objectives:

$$\mathcal{L}(\theta_{x\to y}) = \sum_{i=1}^{N} \Big(\ell\big(f(x^i;\theta_{x\to y}), y^i\big) + \lambda \mathcal{L}_{\text{dual}}(x^i, y^i; \theta_{x\to y}, \theta_{y\to x}) \Big)$$

$$\mathcal{L}(\theta_{y\to x}) = \sum_{i=1}^{N} \Big(\ell\big(f(y^i;\theta_{y\to x}), x^i\big) + \lambda \mathcal{L}_{\text{dual}}(x^i, y^i; \theta_{x\to y}, \theta_{y\to x}) \Big)$$

$$\text{where } \mathcal{L}_{\text{dual}}(x^i, y^i; \theta_{x\to y}, \theta_{y\to x}) = \Big\| \big(\log P(y^i|x^i;\theta_{x\to y}) + \log \hat{P}(x^i) \big) - \big(\log P(x^i|y^i;\theta_{y\to x}) + \log \hat{P}(y^i) \big) \Big\|_2^2.$$

Equations

- Note that we need to get derivative of each model:

$$\mathcal{L}(\theta_{x\rightarrow y}) = \sum_{i=1}^{N} \Big(\ell\big(f(x^i;\theta_{x\rightarrow y}), y^i\big) + \lambda \mathcal{L}_{\text{dual}}(x^i, y^i; \theta_{x\rightarrow y}, \theta_{y\rightarrow x}) \Big)$$

$$\mathcal{L}(\theta_{y\rightarrow x}) = \sum_{i=1}^{N} \Big(\ell\big(f(y^i;\theta_{y\rightarrow x}), x^i\big) + \lambda \mathcal{L}_{\text{dual}}(x^i, y^i; \theta_{x\rightarrow y}, \theta_{y\rightarrow x}) \Big)$$

$$\text{where } \mathcal{L}_{\text{dual}}(x^i, y^i; \theta_{x\rightarrow y}, \theta_{y\rightarrow x}) = \Big\| \big(\log P(y^i|x^i;\theta_{x\rightarrow y}) + \log \hat{P}(x^i) \big) - \big(\log P(x^i|y^i;\theta_{y\rightarrow x}) + \log \hat{P}(y^i) \big) \Big\|_2^2.$$

$$\nabla_{\theta_{x\rightarrow y}} \mathcal{L}_{\text{dual}}(x^i, y^i; \theta_{x\rightarrow y}, \theta_{y\rightarrow x}) = \nabla_{\theta_{x\rightarrow y}} \Big\| \big(\log P(y^i|x^i;\theta_{x\rightarrow y}) + \log \hat{P}(x^i) \big) - \big(\log P(x^i|y^i;\theta_{y\rightarrow x}) + \log \hat{P}(y^i) \big) \Big\|_2^2.$$

Evaluation

- MLE 방식이지만, 기존의 RL 방식을 뛰어넘음

Table 2. Summary of some existing En→Fr translations

Model	Brief description	BLEU
NMT[1]	*standard NMT*	33.08
MRT[2]	*Direct optimizing BLEU*	34.23
DSL	*Refer to Algorithm 1*	**34.84**
[1] (Jean et al., 2015); [2] (Shen et al., 2016)		

← MLE with Cross Entropy

← Reinforcement Learning

Summary

- MLE 방식 위에서, teacher forcing으로 인한 괴리를 해결하려 함
  - 비효율적인 RL을 쓰지 않고도, 더 뛰어난 성능 확보 가능
- Bayes Theorem을 통해 유도된 매우 간단하고 직관적인 regularization term
- 하나의 병렬 코퍼스로부터, 두 개의 모델을 동시에 학습하며 시너지 효과
  - 언어 모델이 추가로 필요한 것은 단점으로 작용