Dual Unsupervised Learning

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Equations

- Given datasets:

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

- Marginal Distribution:

$$\begin{aligned} P(y) &= \mathbb{E}_{x \sim P(\mathrm{x})}[P(y|x)] \\ &= \sum_{x \in \mathcal{X}} P(y|x)P(x) \end{aligned}$$

New Objective

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

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

$$\mathcal{L}(\theta_{x\rightarrow y}) = -\sum_{n=1}^{N} \log P(y^n|x^n;\theta_{x\rightarrow y}) + \lambda \sum_{s=1}^{S} \Big\| \log \hat{P}(y^s) - \log \frac{1}{K}\sum_{k=1}^{K} P(y^s|x_k;\theta_{y\rightarrow x}) \Big\|_2^2,$$

$$\text{where } x_k \sim P(\mathrm{x}).$$

Thus, we need

- Importance Sampling:

$$\begin{aligned}\mathbb{E}_{x\sim p(\mathrm{x})}\left[f(x)\right] &= \int f(x)p(x)dx \\ &= \int \frac{f(x)p(x)}{q(x)}q(x)dx \\ &= \mathbb{E}_{x\sim q(\mathrm{x})}\left[f(x)\frac{p(x)}{q(x)}\right]\end{aligned}$$

Re-write Objective

- By importance sampling,

$$
\begin{aligned}
P(y) &= \mathbb{E}_{x\sim P(\mathrm{x})}[P(y|x)] \\
&= \sum_{x\in\mathcal{X}} P(y|x)P(x) \\
&= \sum_{x\in\mathcal{X}} \frac{P(y|x)P(x)}{P(x|y)} P(x|y) \\
&= \mathbb{E}_{x\sim P(\mathrm{x}|y)}\left[\frac{P(y|x)P(x)}{P(x|y)}\right] \\
&\approx \frac{1}{K}\sum_{k=1}^{K} \frac{P(y|x_k)P(x_k)}{P(x_k|y)}, \text{ where } x_k \sim P(\mathrm{x}|y)
\end{aligned}
$$

Re-write Objective

- Our new objective:

$$\mathcal{L}(\theta_{x\rightarrow y}) = -\sum_{n=1}^{N}\log P(y^n|x^n;\theta_{x\rightarrow y}) + \lambda\mathcal{L}_{\text{dul}}(\theta_{x\rightarrow y})$$

$$\mathcal{L}_{\text{dul}}(\theta_{x\rightarrow y}) = \sum_{s=1}^{S}\Bigg\|\log\hat{P}(y^s) - \log\frac{1}{K}\sum_{k=1}^{K}\frac{P(y^s|x_k^s;\theta_{x\rightarrow y})\hat{P}(x_k^s)}{P(x_k^s|y^s;\theta_{y\rightarrow x})}\Bigg\|_2^2$$

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

Evaluation

Table 1: BLEU scores on En→Fr and De→En translation tasks. Δ means the improvement over the basic NMT model, which only used bilingual data for training. The basic model for En→Fr is the RNNSearch model (Bahdanau, Cho, and Bengio 2015), and for De→En is a two-layer LSTM model. Note that all the methods for the same task share the same model structure.

System	En→Fr	Δ	De→En	Δ
Basic model	29.92		30.99	
Representative semi-supervised NMT systems				
Shallow fusion-NMT (Gulcehre et al. 2015)	30.03	+0.11	31.08	+0.09
Pseudo-NMT (Sennrich, Haddow, and Birch 2016)	30.40	+0.48	31.76	+0.77
Dual-NMT (He et al. 2016a)	32.06	+2.14	32.05	+1.06
Our dual transfer learning system				
This work	**32.85**	**+2.93**	**32.35**	**+1.36**

Summary

- Back Translation, Dual Learning for Machine Translation 과 달리, 수학적으로 매우 잘 정의된 깔끔한 objective function이 매력
  - Back Translation은 학습이 끝난 반대쪽 모델을 주로 활용하는 형태 (offline 학습)
  - Dual Learning for MT는 RL이 활용되므로, 비효율적인 학습이 될 수 있음
  - 하지만 언어 모델이 필요한 것이 단점