Prior Knowledge Integration for Neural Machine Translation using Posterior Regularization

Andrew Drozdov

November 28, 2017

Background

- Neural Machine Translation
- Posterior Regularization

Background: Neural Machine Translation

- Standard Seq2seq.

$$\mathcal{L}(\theta) = \Sigma_n \log P(y^n | x^n; \theta)$$

- It's difficult to incorporate linguistic prior knowledge in discrete symbolic forms (like phrase tables).
  - There are other ways to incorporate linguistic priors. [1]
- Example of prior knowledge: coverage constraint.
- Existing approaches are not scalable.
  - Only works for simple constraints.
  - Not flexible.

[1] Eriguchi et al. 2017 Learning to Parse and Translate Improves Neural Machine Translation

Background: Posterior Regularization

- Posterior regularized likelihood:

$$F(\theta, q) = \lambda_1 \mathcal{L}(\theta) - \lambda_2 \Sigma_n \min_{q \in \mathcal{Q}} KL\big(q(y) || P(y|x^n; \theta)\big)$$

- $\mathcal{Q}$ is a constrained posterior set.
  - For instance, bijectivity and symmetry in machine translation.

$$\mathcal{Q} = \{q(y) : \mathbb{E}_q[\phi(x, y)] \leq b\}$$

- Can be optimized using a simple EM scheme.

$$\mathrm{E} : q^{(t+1)} = \operatorname*{argmin}_{q} \mathrm{KL}\Big(q(\mathbf{y}) \Big|\Big| P(\mathbf{y}|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{(t)})\Big)$$

$$\mathrm{M} : \boldsymbol{\theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q^{(t+1)}}\Big[\log P(\mathbf{y}|\mathbf{x}^{(n)}; \boldsymbol{\theta})\Big]$$

Model

- Use a log-linear model rather than a constrained posterior set.

$$\mathcal{J}(\theta, \gamma) = \lambda_1 \mathcal{L}(\theta) - \lambda_2 \Sigma_n KL\big(Q(y|x^n; \gamma) || P(y|x^n; \theta)\big)$$

$$Q(y|x^n; \gamma) = \frac{\exp(\gamma \cdot \phi(x, y))}{\Sigma_{y'} \exp(\gamma \cdot \phi(x, y'))}$$

Feature Design

- Bilingual Dictionary $\phi \in \{0, 1\}$
- Phrase Table $\phi \in \{0, 1\}$
- Coverage Constraint $\phi_{CP}(x, y) = \Sigma_{i \in |x|} \log\big(\min(1.0, \Sigma_{j \in |y|} a_{i,j})\big)$
  - Dependent on sentence lengths.
- Length Ratio $\phi_{LR}(x, y) = \frac{\beta|x|}{|y|}$ if $\beta|x| < |y|$ otherwise $\frac{|y|}{\beta|x|}$

Training and Inference

- During training, subsample predicted target sentences.

$$\mathrm{KL}\Big(Q(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\gamma})\Big|\Big|P(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\theta})\Big)$$
$$\approx \sum_{\mathbf{y}\in\mathcal{S}(\mathbf{x}^{(n)})} \tilde{Q}(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\gamma}) \log \frac{\tilde{Q}(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\gamma})}{\tilde{P}(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\theta})}$$

- Normalize the sampled subspace.

$$\tilde{Q}(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\gamma}) = \frac{\exp(\boldsymbol{\gamma}\cdot\boldsymbol{\phi}(\mathbf{x}^{(n)},\mathbf{y}))}{\sum_{\mathbf{y}'\in\mathcal{S}(\mathbf{x}^{(n)})}\exp(\boldsymbol{\gamma}\cdot\boldsymbol{\phi}(\mathbf{x}^{(n)},\mathbf{y}'))}.$$

$$\tilde{P}(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\theta}) = \frac{P(\mathbf{y}|\mathbf{x}^{(n)};\boldsymbol{\theta})^{\alpha}}{\sum_{\mathbf{y}'\in\mathcal{S}(\mathbf{x}^{(n)})}P(\mathbf{y}'|\mathbf{x}^{(n)};\boldsymbol{\theta})^{\alpha}}. \quad (18)$$

- During inference, generate a candidate list using maximum likelihood only, then rerank by incorporating prior knowledge.

$$\hat{y} = arg \max_{y\in\mathcal{C}(x)} \{\log P(y|x;\theta) + \gamma\cdot\phi(x,y)\}$$

Experimental Setup

- English-to-Chinese Translation.
- Train RNNSearch for 300k iterations (4 days). Train model of this work (3 days).
- RNNSearch
  - Word Embedding: 620
  - Hidden Layer: 1000
  - Batch Size: 80
  - AdaDelta
  - Beam Size of 10 during decoding.
- RNNSearch+ThisWork
  - Batch Size: 1
  - Candidates: 80 and $\alpha = 0.2$
  - $\lambda_1 = 8 \times 10^{-5}$, $\lambda_2 = 2.5 \times 10^{-4}$

Results

Method	Feature	MT02	MT03	MT04	MT05	MT06	MT08	All
RNNSEARCH	N/A	33.45	30.93	32.57	29.86	29.03	21.85	29.11
CPR	N/A	33.84	31.18	33.26	30.67	29.63	22.38	29.72
POSTREG	BD	34.65	31.53	33.82	30.66	29.81	22.55	29.97
	PT	34.56	31.32	33.89	30.70	29.84	22.62	29.99
	LR	34.39	31.41	34.19	30.80	29.82	22.85	30.14
	BD+PT	34.66	32.05	34.54	31.22	30.70	22.84	30.60
	BD+PT+LR	34.37	31.42	34.18	30.99	29.90	22.87	30.20
this work	BD	**36.61**	33.47	36.04	32.96	32.46	**24.78**	32.27
	PT	35.07	32.11	34.73	31.84	30.82	23.23	30.86
	CP	34.68	31.99	34.67	31.37	30.80	23.34	30.76
	LR	34.57	31.89	34.95	31.80	31.43	23.75	31.12
	BD+PT	36.30	**33.83**	36.02	32.98	32.53	24.54	32.29
	BD+PT+CP	36.11	33.64	36.36	**33.11**	32.53	24.57	32.39
	BD+PT+CP+LR	36.10	33.64	**36.48**	33.08	**32.90**	24.63	**32.51**

Table 1: Comparison of BLEU scores on the Chinese-English datasets. RNNSEARCH is an attention-based neural machine translation model (Bahdanau et al., 2015) that does not incorporate prior knowledge. CPR extends RNNSEARCH by introducing coverage penalty refinement (Eq. (11)) in decoding. POSTREG extends RNNSEARCH with posterior regularization (Ganchev et al., 2010), which uses constraint features to represent prior knowledge and a constrained posterior set to denote the desired distribution. Note that POSTREG cannot use the CP feature (Section 3.2.3) because it is hard to bound the feature value appropriately. On top of RNNSEARCH, our approach also exploits posterior regularization to incorporate prior knowledge but uses a log-linear model to denote the desired distribution. All results of *this work* are significantly better than RNNSEARCH ($p < 0.01$).

Ablation

Feature	Rerank	MT02	MT03	MT04	MT05	MT06	MT08	All
BD	w/o	36.06	32.99	35.62	32.59	32.13	24.36	31.87
	w/	**36.61**	33.47	36.04	32.96	32.46	**24.78**	32.27
PT	w/o	34.98	32.01	34.71	31.77	30.77	23.20	30.81
	w/	35.07	32.11	34.73	31.84	30.82	23.23	30.86
CP	w/o	34.68	31.99	34.67	31.37	30.80	23.34	30.76
	w/	34.68	31.99	34.67	31.37	30.80	23.34	30.76
LR	w/o	34.60	31.89	34.79	31.72	31.39	23.63	31.03
	w/	34.57	31.89	34.95	31.80	31.43	23.75	31.12
BD+PT	w/o	35.76	33.27	35.64	32.47	32.03	24.17	31.83
	w/	36.30	**33.83**	36.02	32.98	32.53	24.54	32.29
BD+PT+CP	w/o	35.71	33.15	35.81	32.52	32.16	24.11	31.89
	w/	36.11	33.64	36.36	**33.11**	32.53	24.57	32.39
BD+PT+CP+LR	w/o	36.06	33.01	35.86	32.70	32.24	24.27	31.96
	w/	36.10	33.64	**36.48**	33.08	**32.90**	24.63	**32.51**

Table 2: Effect of reranking on translation quality.

Examples

Source	*lijing liang tian yu bingxue de* ***fenzhan*** *, 31ri shenye 23 shi 50 fen , shanghai jichang jituan* ***yuangong yinglai*** *le 2004nian de zuihou yige hangban .*
Reference	after **fighting** with ice and snow for two days , **staff** members of shanghai airport group **welcomed** the last flight of 2004 at 23 : 50pm on the 31st .
RNNSEARCH	after a two - day and two - day journey , the team of shanghai 's airport in shanghai has ushered in the last flight in 2004 .
+ BD	after two days and nights **fighting** with ice and snow , the shanghai airport group 's **staff welcomed** the last flight in 2004 .
Source	*suiran tonghuopengzhang* ***weilai ji ge yue*** *reng jiang weizhi zai* ***baifenzhier yishang*** *, buguo niandi zhiqian keneng jiangdi .*
Reference	although inflation will remain **above 2 % for the coming few months** , it may decline by the end of the year .
RNNSEARCH	although inflation has been maintained for more than two months from the year before the end of the year , it may be lower .
+ PT	although inflation will remain at **more than 2 percent in the next few months** , it may be lowered before the end of the year .
Source	*qian ji tian ta ganggang* ***chuyuan*** *, jintian* ***jianchi*** *lai yu lao pengyou* ***daobie*** *.*
Reference	just **discharged from the hospital** a few days ago , he **insisted on** coming to **say farewell** to his old friend today .
RNNSEARCH	during the previous few days , he had just been given treatment to the old friends .
+ CP	during the previous few days , he had just been **discharged from the hospital** , and he **insisted on goodbye** to his old friend today .
Source	*(guoji) yiselie fuzongli fouren jihua kuojian* ***gelan gaodi*** *dingjudian*
Reference	(international) israeli deputy prime minister denied plans to expand **golan heights** settlements
RNNSEARCH	(world) israeli deputy prime minister denies the plan to expand **the golan heights** in **the golan heights**
+ LR	(international) israeli deputy prime minister denies planning to expand **golan heights**

Table 3: Example translations that demonstrate the effect of adding features.