Minimizing the Bag-of-Ngrams Difference for Non-Autoregressive Neural Machine Translation

Chenze Shao[1,2], **Jinchao Zhang**[3], **Yang Feng**[1,2*], **Fandong Meng**[3] and **Jie Zhou**[3]

[1] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)

[2] University of Chinese Academy of Sciences

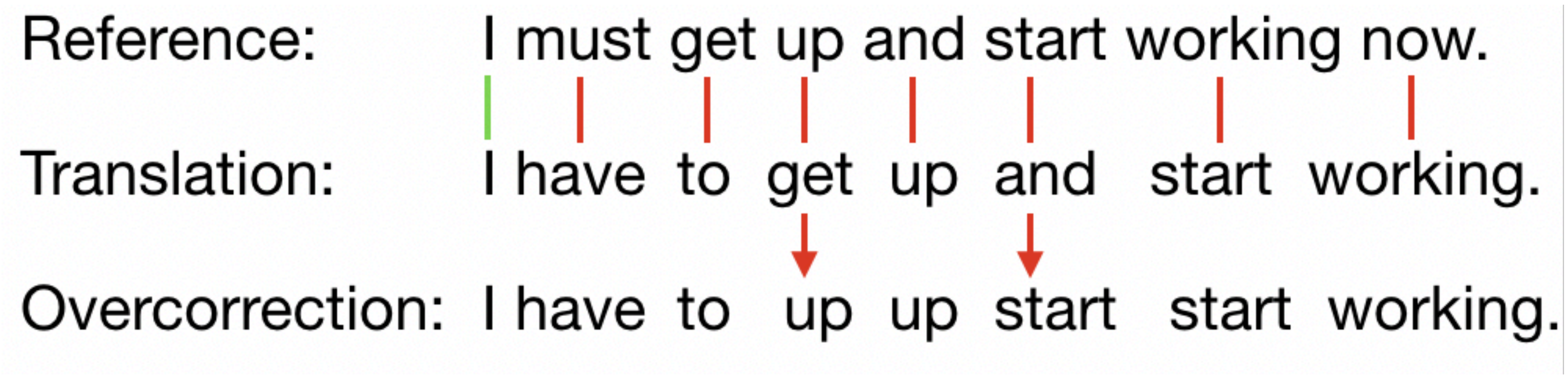
[3] Pattern Recognition Center, WeChat AI, Tencent Inc, China

邵晨泽

2019.12.22

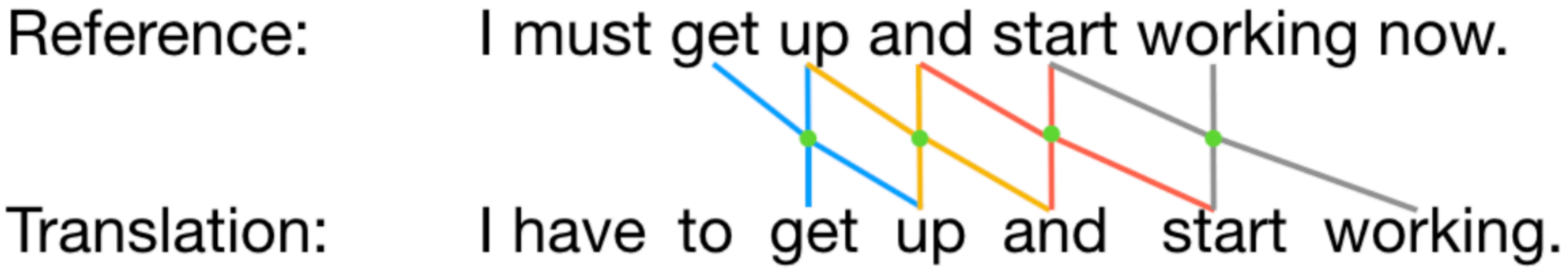
Motivation

- NAT: generate target words independently
- The cross-entropy loss is not suitable for NAT
  - cannot model target-side sequential dependency properly
  - Requires strict alignment, hard for NAT
  - Overcorrection, repeated tokens



Motivation

- Minimize the bag-of-ngrams difference for NAT
  - Model sequential dependency: evaluate NAT outputs on n-gram level
  - Outputs may not be aligned with the reference: optimize bag-of-ngrams
- Bag-of-ngrams correlates well with the translation quality



Method

- Minimizing the Bag-of-Ngrams Difference
  - Define Bag-of-Ngrams (BoN) of NAT
  - Give a method to calculate BoN
  - Choose a distance metric to measure the BoN Difference
  - Give a method to calculate the metric
  - Bag-of-Ngrams difference as training objective

Method

- Bag-of-Ngrams (BoN): a vector of size V^n, where V is the vocabulary size. Each entry represents the number of occurrences of an n-gram **g** in sentence **Y**:

$$\mathrm{BoN}_{\boldsymbol{Y}}(\boldsymbol{g}) = \sum_{t=0}^{T-n} 1\{y_{t+1:t+n} = \boldsymbol{g}\}$$

- How to define the BoN of NMT?

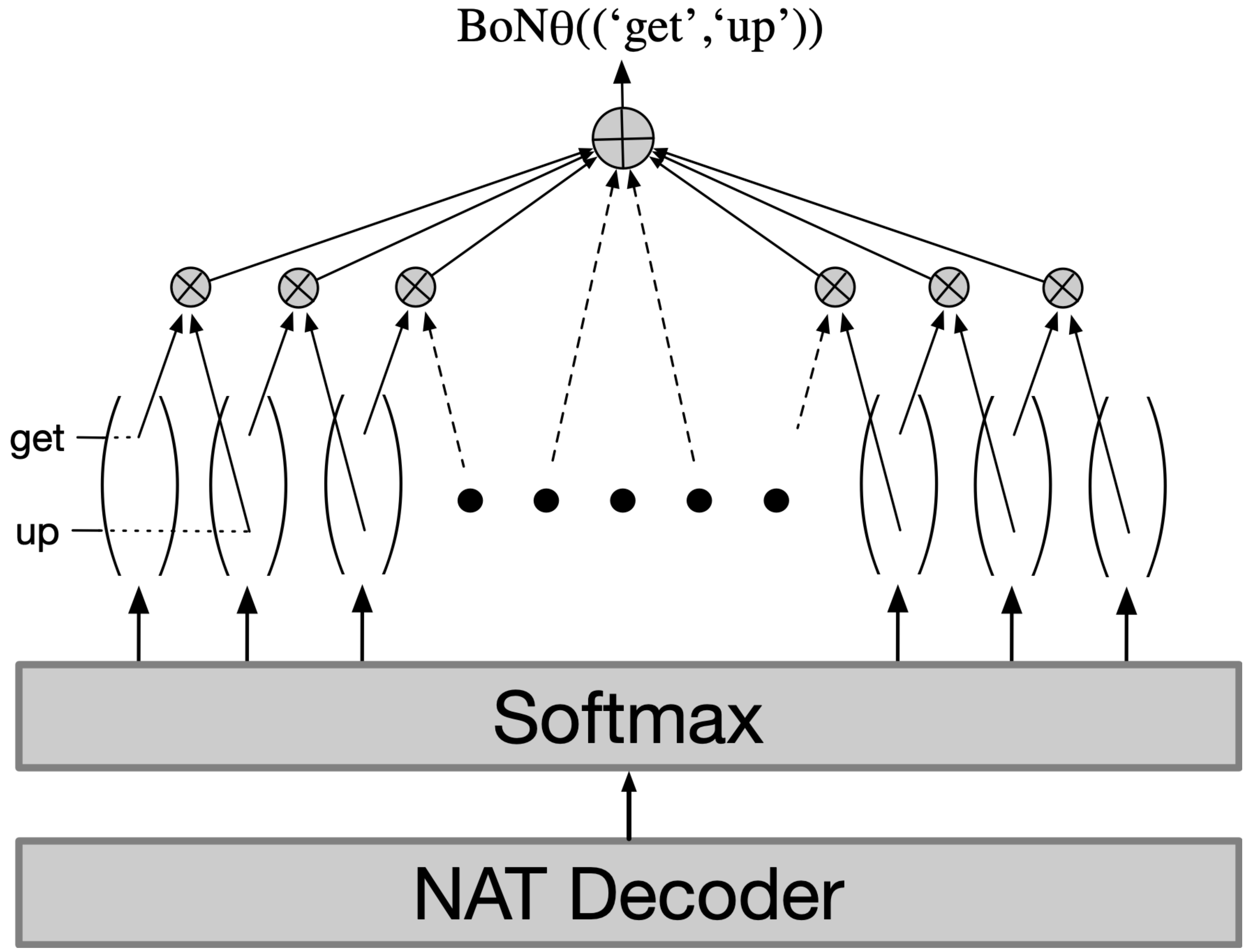
- Consider all possible translations, the expected BoN

$$\mathrm{BoN}_{\theta}(\boldsymbol{g}) = \sum_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{X}, \theta) \cdot \mathrm{BoN}_{\boldsymbol{Y}}(\boldsymbol{g})$$

Method

- BoN definition $\mathrm{BoN}_\theta(\boldsymbol{g}) = \sum_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{X},\theta) \cdot \mathrm{BoN}_{\boldsymbol{Y}}(\boldsymbol{g})$
- The first difficulty: exponentially large search space
- Use the property of NAT: probabilities at different positions are independent

$$\begin{aligned}\mathrm{BoN}_\theta(\boldsymbol{g}) &= \sum_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{X},\theta) \cdot \sum_{t=0}^{T-n} 1\{y_{t+1:t+n} = \boldsymbol{g}\} \\ &= \sum_{t=0}^{T-n} \prod_{i=1}^{n} p(y_{t+i} = g_i|\boldsymbol{X},\theta).\end{aligned}$$



Method

- Calculate the BoN difference between NAT and reference (L_1, L_2, cosine, etc.)
- The second difficulty: V^n n-grams, costly to calculate and store all of them
- How to simplify?
  - The BoN of NAT is dense.
  - The BoN of reference is very sparse, only a few entries are non-zero.
  - Use the sparsity to simplify the calculation of L_1 distantce

Method

- Intuition: a sentence of length T has T−n+1 n-grams:

$$\sum_{\boldsymbol{g}} \mathrm{BoN}_{\boldsymbol{Y}}(\boldsymbol{g}) = \sum_{t=0}^{T-n} \sum_{\boldsymbol{g}} 1\{y_{t+1:t+n} = \boldsymbol{g}\} = T - n + 1. \tag{9}$$

$$\begin{aligned} \sum_{\boldsymbol{g}} \mathrm{BoN}_{\theta}(\boldsymbol{g}) &= \sum_{\boldsymbol{g}} \sum_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{X}, \theta) \cdot \mathrm{BoN}_{\boldsymbol{Y}}(\boldsymbol{g}) \\ &= \sum_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{X}, \theta) \cdot \sum_{\boldsymbol{g}} \mathrm{BoN}_{\boldsymbol{Y}}(\boldsymbol{g}) = T - n + 1. \end{aligned} \tag{10}$$

- L_1 distantce: only consider n-grams in the reference

$$\begin{aligned} \text{BoN-}L_1 &= \sum_{\boldsymbol{g}} |\mathrm{BoN}_{\theta}(\boldsymbol{g}) - \mathrm{BoN}_{\hat{\boldsymbol{Y}}}(\boldsymbol{g})| \\ &= \sum_{\boldsymbol{g}} (\mathrm{BoN}_{\theta}(\boldsymbol{g}) + \mathrm{BoN}_{\hat{\boldsymbol{Y}}}(\boldsymbol{g}) - 2\min(\mathrm{BoN}_{\theta}(\boldsymbol{g}), \mathrm{BoN}_{\hat{\boldsymbol{Y}}}(\boldsymbol{g})) \\ &= 2(T - n + 1 - \sum_{\boldsymbol{g}} \min(\mathrm{BoN}_{\theta}(\boldsymbol{g}), \mathrm{BoN}_{\hat{\boldsymbol{Y}}}(\boldsymbol{g}))). \end{aligned}$$

Method

- Normalize the L_1 distance to range [0,1]

$$\mathcal{L}_{BoN}(\theta) = \frac{\text{BoN-}L_1}{2(T-n+1)}$$

- BoN-FT: fine-tune the pre-trained NAT model

- BoN-Joint: combine the BoN loss and cross-entropy loss

$$\mathcal{L}_{joint}(\theta) = \alpha \cdot \mathcal{L}_{MLE}(\theta) + (1-\alpha) \cdot \mathcal{L}_{BoN}(\theta)$$

- BoN-Joint+FT: joint training and fine-tuning

Experiments

		IWSLT'16 En-De			WMT'16 En-Ro			WMT'14 En-De		
		En→	speedup	secs/b	En→	Ro→	speedup	En→	De→	speedup
AR	b=1	28.64	1.09×	0.20	31.93	31.55	1.23×	23.77	28.15	1.13×
	b=4	28.98	1.00×	0.20	32.40	32.06	1.00×	24.57	28.47	1.00×
NAT Models	NAT-FT (Gu et al. 2017)	26.52	15.6×	–	27.29	29.06	–	17.69	21.47	–
	IRNAT(iter=2) (2018)	24.82	6.64×	–	27.10	28.15	7.68×	16.95	20.39	8.77×
	IRNAT(adaptive) (2018)	27.01	1.97×	–	29.66	30.30	2.73×	21.54	25.43	2.38×
	NAT-REG (Wang et al. 2019)	–	–	–	–	–	–	20.65	24.77	27.6×
	Reinforce-NAT (Shao et al. 2019)	25.18	8.43×	13.40	27.09	27.93	9.44×	19.15	22.52	10.73×
Our Models	NAT-Base	24.13	8.42×	0.62	25.96	26.49	9.41×	16.05	19.46	10.76×
	BoN-FT (n=2)	25.03	8.44×	1.41	27.21	27.95	9.50×	19.27	23.20	10.72×
	BoN-Joint (n=2, α=0.1)	25.63	8.39×	1.49	28.12	29.03	9.44×	20.75	24.47	10.79×
	BoN-Joint+FT (n=2, α=0.1)	25.72	8.40×	1.41	28.31	29.29	9.51×	20.90	24.61	10.77×

- 1. BoN-FT outperforms Reinforce-NAT: faster and better
- 2. BoN-Joint achieves considerable improvements over BoN-FT
- 3. BoN-Joint+FT achieves about 5 BLEU improvements on WMT14 EN<->DE

Experiments

- The correlation between BoN difference and translation quality?
  - WMT14 En->De dev set, randomly divide the 3000 sentences to 30 groups of size 100
  - Calculate the cross-entropy loss, BoN-L_1 loss and BLEU score on each group
  - The pearson correlation between BLEU score and losses:

Loss function	CE	$n = 1$	$n = 2$	$n = 3$	$n = 4$
Correlation	0.37	0.56	0.70	0.67	0.61

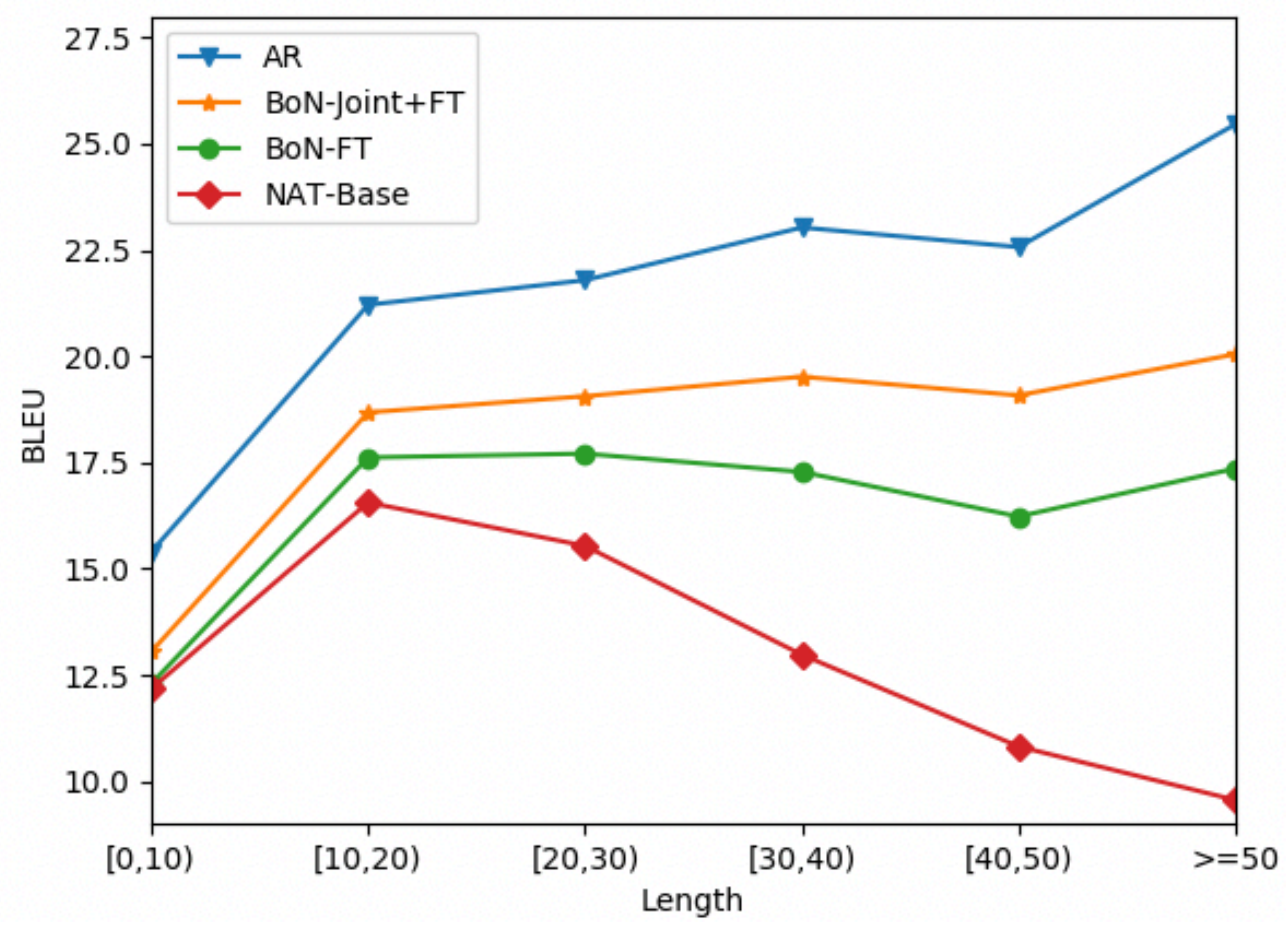
Experiments

- The effect of sentence length?
- Divide the devset evenly to short and long sentences
- Calculate the correlation on each part
- Performance of the cross-entropy loss drops sharply, where the BoN loss is robust to long sentences

	all	short	long
Cross-Entropy	0.37	0.52	0.21
BoN (n=2)	0.70	0.79	0.81

Experiments

- WMT14 En->De dev set
- BLEU performance on different length buckets



Thanks