Dynamic Sentence Sampling for Efficient Training of NMT

Rui Wang, Masao Utiyama, and Eiichro Sumita

National Institute of Information and Communications Technology, Kyoto, Japan

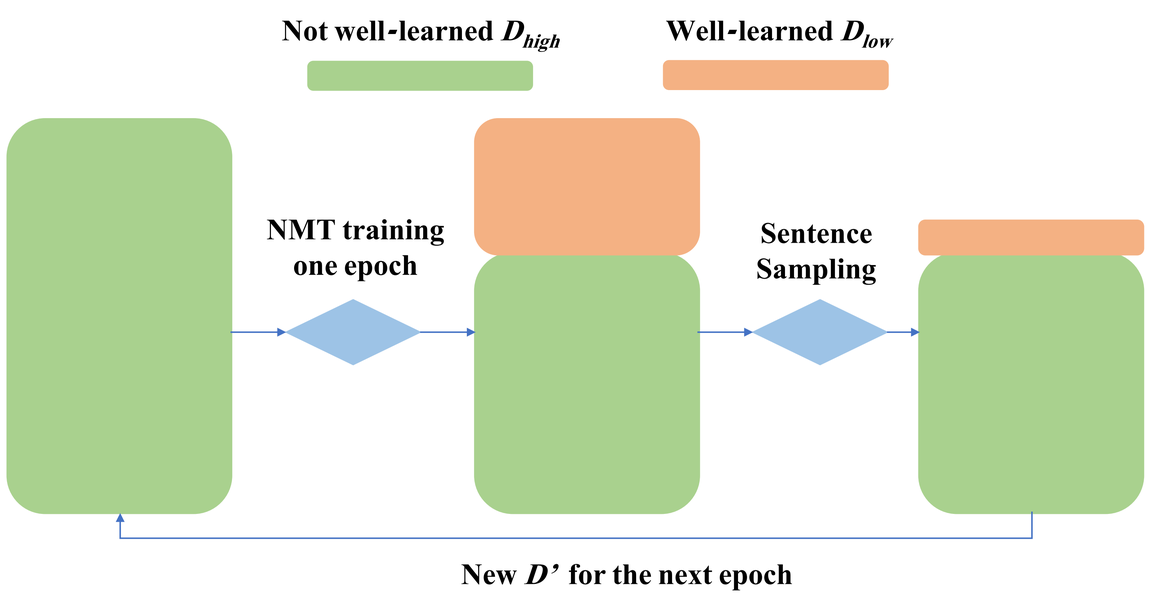
Hypotheses

NMT involves a fixed training procedure where each sentence is sampled once during each epoch:

- Some sentences are well-learned during the initial few epochs.
- Some sentences were not well learned until 10-30 epochs.
- Training these two type sentences together results in a wastage of time.

We proposes a Dynamic Sentence Sampling (DSS) method:

- We use the training cost difference as the criterion to measure which sentence has been well-learned.
- We propose two sentence sampling mechanisms: Weighted Sampling (WS) and Review Mechanism (RM).



Criterion

The training cost of a sentence pair $\langle x, y\rangle$ from corpus D during the ith iteration can be calculated as:

$$cost^{i}_{\langle x,y\rangle} = -\log P(y|x,\boldsymbol{\theta}). \quad (1)$$

We adopt the ratio of differences (dif) between training costs of two training iterations to be the criterion,

$$dif^{i}_{\langle x,y\rangle} = \frac{cost^{i-1}_{\langle x,y\rangle} - cost^{i}_{\langle x,y\rangle}}{cost^{i-1}_{\langle x,y\rangle}}. \quad (2)$$

Dynamic Sentence Sampling (DSS)

(1) Weighted Sampling (WS)

Weighted sampling without any replacement was used to select a small subset, such as 80% of the entire corpus, as the corpus D^{i+1}_{ws} to perform the subsequent iteration.

$$J_{ws} = \sum_{\langle x,y\rangle \in D_{ws}} -\log P(y|x,\boldsymbol{\theta}). \quad (3)$$

(2) Review Mechanism (RM)

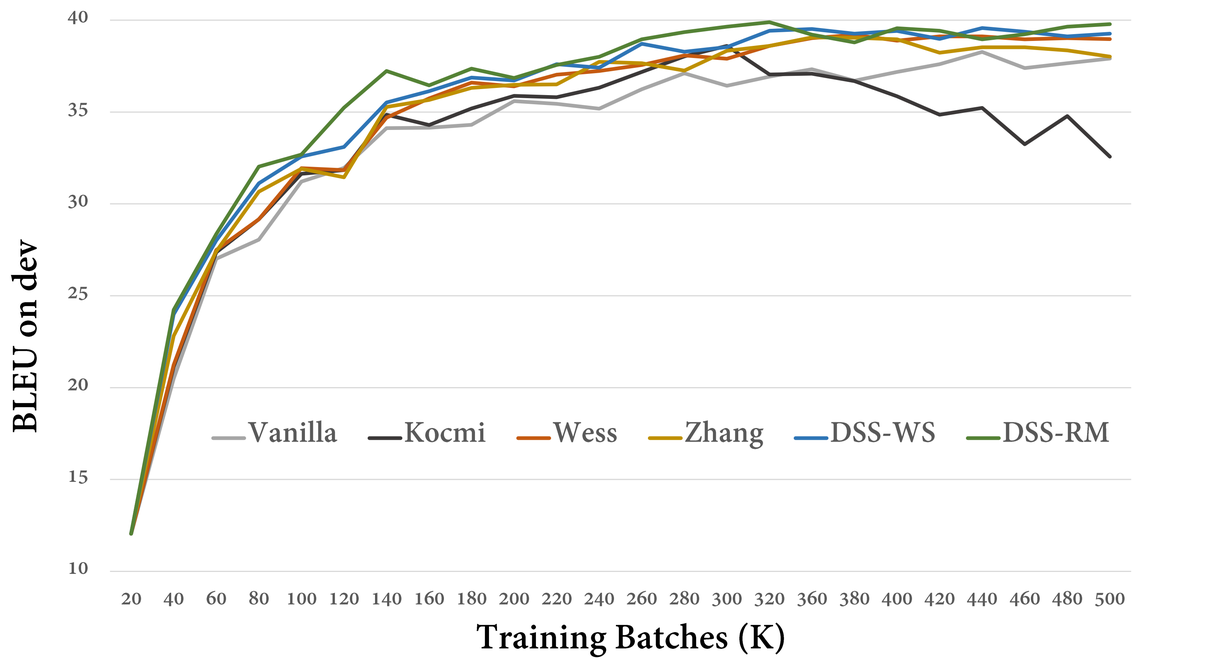
To prevent the loss of the knowledge that was obtained from the D_{low} group during NMT, a small percentage λ, such as 10%, of the D_{low} group is sampled as the knowledge to be reviewed.

$$J_{rm} = \sum_{\langle x,y\rangle \in D_{high}} -\log P(y|x,\boldsymbol{\theta}) + \sum_{\langle x,y\rangle \in \lambda D_{low}} -\log P(y|x,\boldsymbol{\theta}). \quad (4)$$

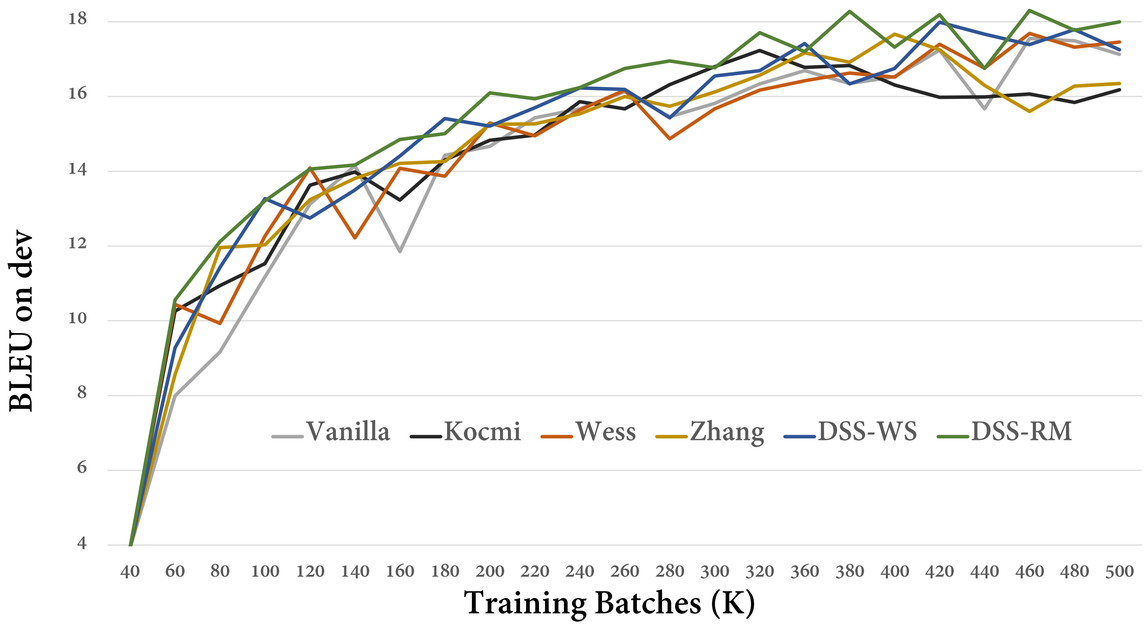
Results and Analyses

We adopted attentional RNN based NMT by Nematus.

ZH-EN (500K batches)	Dev (NIST02)	Test (NIST03-08)
PBSMT	33.15	29.66
Vanilla NMT	38.48	35.08
Random Sampling	38.35	34.62
Kocmi (Curriculum Learning)	38.51	35.19
Wees (Dynamic tuning)	**39.16**	**35.62**
Zhang (Boosting)	39.08	35.57
DSS-WS	39.54+	36.85++
DSS-RM	**39.89**++	**37.33**++



EN-DE (500K batches)	Dev (WMT12)	Test (WMT13-15)
PBSMT	14.89	16.35
Vanilla NMT	17.55	20.06
Random Sampling	17.39	19.61
Kocmi (Curriculum Learning)	17.63	20.18
Wees (Dynamic tuning)	**17.69**	20.19
Zhang (Boosting)	17.67	**20.30**
DSS-WS	17.99	20.96+
DSS-RM	**18.34**+	**21.22**++



Discussions

We would like to investigate what would happen if:

- Train on larger/extreme-large corpora.
- Keep training for longer time.
- There is noisy/low-quality data in the corpus.