Sentence Embedding for NMT Domain Adaptation

Rui Wang, Andrew Finch, Masao Utiyama and Eiichro Sumita
National Institute of Information and Communications Technology, Kyoto, Japan
`https://wangruinlp.github.io/`

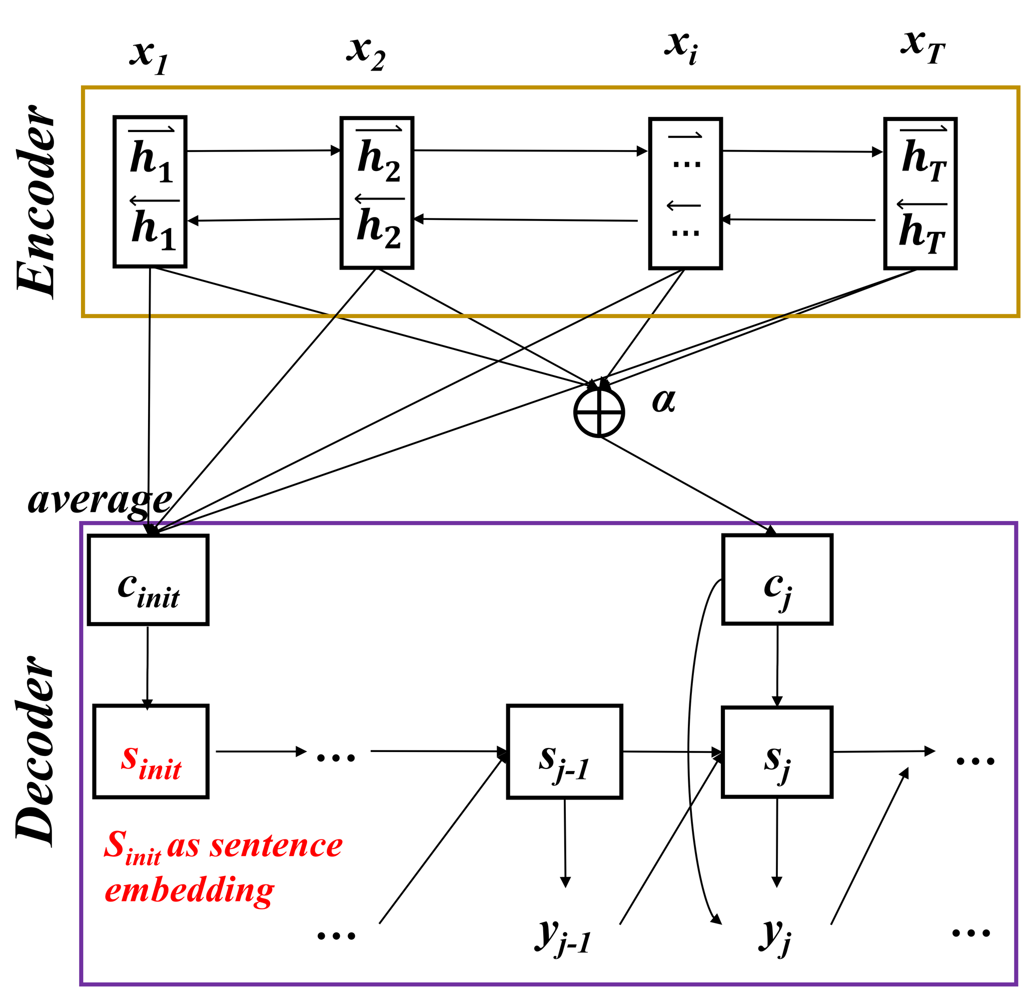
Hypotheses

For some specific translation task, such as IWSLT:

- In-domain corpus is not enough to train a robust NMT system.
- Adding out-of-domain corpora directly, cannot benefit NMT.
- There are some pseudo "in-domain" data in out-of-domain corpora.
- How to select the pseudo "in-domain" data and make use of them?
- How about using NMT internal embedding?

Methods	SMT	NMT
Sentence Selection	Many	This work
Model Combination	Many	Ensemble
Instance Weighting	Many	To appear [5]

NMT Background



In the decoder, an RNN hidden state s_j for time j is computed by [3]:

$$s_j = f(s_{j-1}, y_{j-1}, c_j). \tag{1}$$

The context vector c_j is then, computed as a weighted sum of these annotations $\mathbf{H} = \{h_1, ..., h_{T_x}\}$, by using alignment weight α_{ji}:

$$c_j = \sum_{j=1}^{T_x} \alpha_{ji} h_i. \tag{2}$$

Sentence Embedding

The initial hidden layer state s_{init} for the decoder as this vector:

$$s_{init}(\mathbf{X}) = tanh(\mathbf{W} \frac{\sum_{i=1}^{T_x} h_i}{T_x} + \mathbf{b}), h_i \in \mathbf{H}. \tag{3}$$

Sentence Selection

- 1) We train a French-to-English NMT system $\mathbf{N_{FE}}$ using the in-domain and out-of-domain data together as training data.
- 2) Each sentence f in the training data F (both in-domain F_{in} and out-of-domain F_{out}) is embedded as a vector $v_f = s_{init}(f)$ by using $\mathbf{N_{FE}}$.
- 3) The sentence pairs (f, e) in the out-of-domain corpus F_{out} are classified into two sets: the sentences close to in-domain sentences, and those that are distant.

The vector centers of in-domain $C_{F_{in}}$ and out-of-domain $C_{F_{out}}$ corpora, respectively.

$$C_{F_{in}} = \frac{\sum_{f \in F_{in}} v_f}{|F_{in}|}, \text{ and } C_{F_{out}} = \frac{\sum_{f \in F_{out}} v_f}{|F_{out}|}. \tag{4}$$

We use the difference δ_f of these two distances d to classify each sentence, where d is Euclidean distance:

$$\delta_f = d(v_f, C_{F_{in}}) - d(v_f, C_{F_{out}}). \tag{5}$$

By using an English-to-French NMT system $\mathbf{N_{EF}}$, corresponding distance difference δ_e is,

$$\delta_e = d(v_e, C_{E_{in}}) - d(v_e, C_{E_{out}}). \tag{6}$$

δ_f, δ_e and $\delta_{fe} = \delta_f + \delta_e$ can be used to select sentences. That is, the sentence pairs (f, e) with δ_f (or δ_e, δ_{fe}) less than a threshold are the new selected in-domain corpus. This threshold is tuned by using the development data.

Data sets

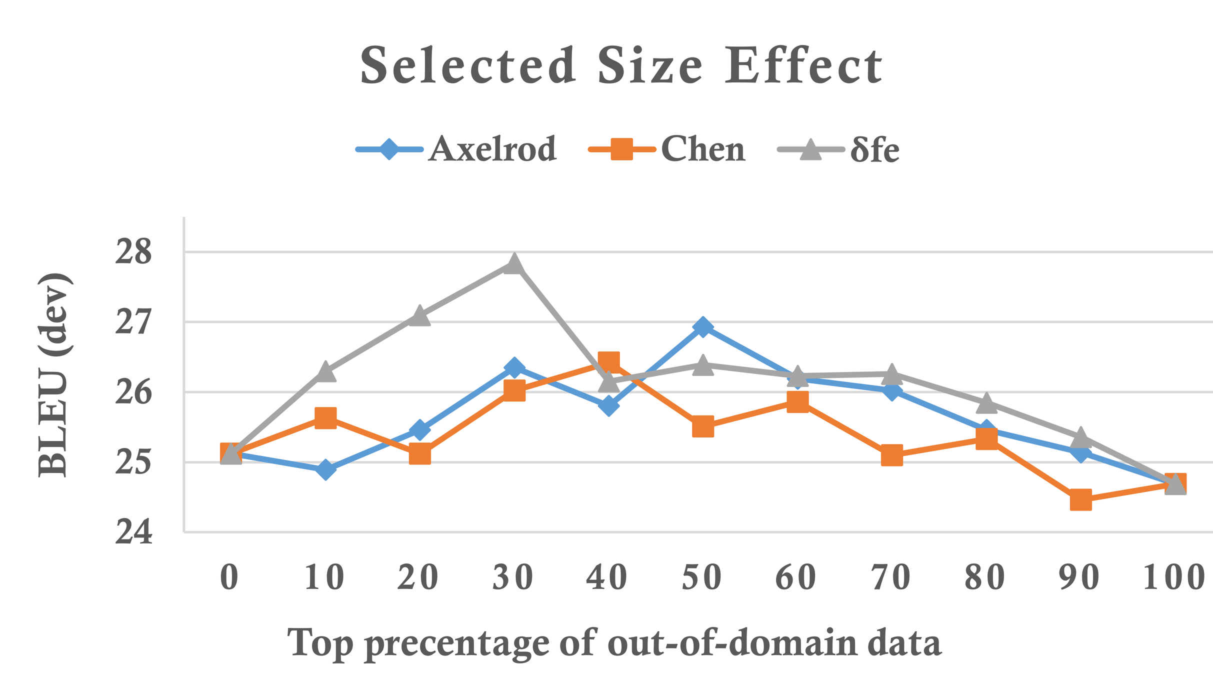
IWSLT EN-FR	Sentences	Tokens
TED training (in-domain)	178.1K	3.5M
WMT training (out-of-domain)	17.8M	450.0M
TED dev2010	0.9K	20.1K
TED test2010	1.6K	31.9K
TED test2011	0.8K	15.6K
NIST ZH-EN	Sentences	Tokens
NIST training (in-domain)	430.8K	12.6M
UN & NTCIR training (out-of-domain)	8.8M	249.4M
dev (MT02-04)	3.4K	106.4K
test (MT05)	1.0K	34.7K
test (MT06)	1.6K	46.7K

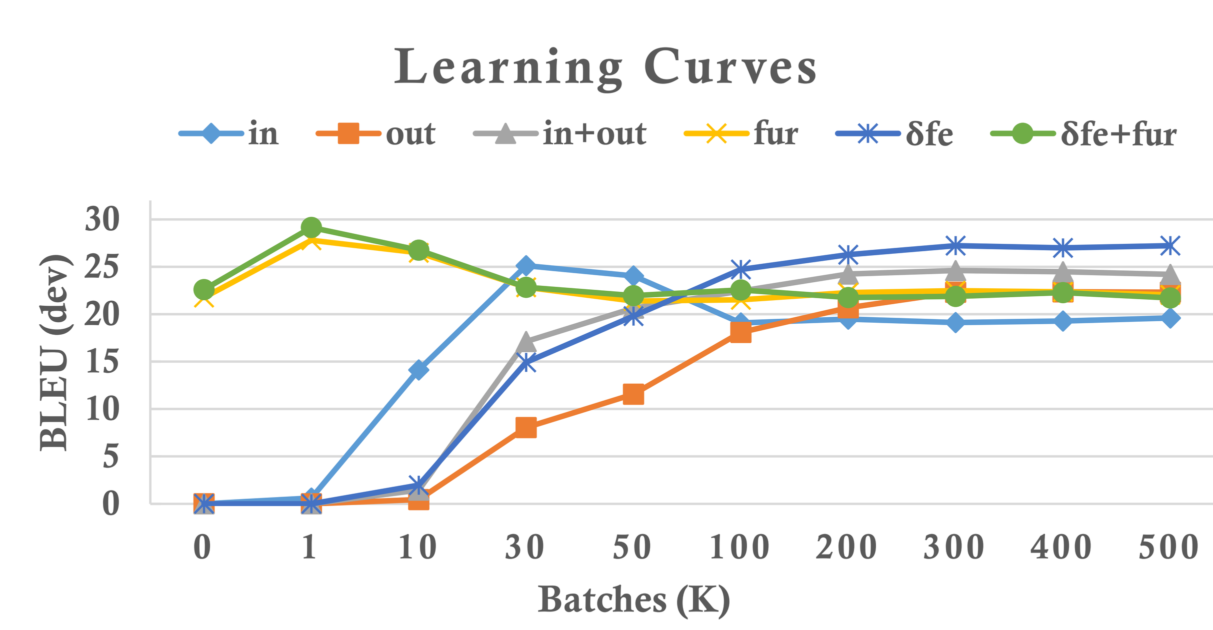
Results and Analyses

We implemented the proposed method in Groundhog [3]. the in, out and $in+out$ indicate that the in-domain, out-of-domain and their mixture were used as the NMT training corpora. δ_f, δ_e and δ_{fe} indicate that corresponding proposed criterion was used to select sentences. $+fur$ indicates that the selected sentences were used to train an initial NMT system, and then this initial system was further trained by in-domain data [4].

IWSLT EN-FR	Sent. No.	SMT tst10	SMT tst11	NMT tst10	NMT tst11
in	178.1K	31.06	32.50	29.23	30.00
out	17.7M	30.04	29.29	27.30	28.48
$in+out$	17.9 M	30.00	30.26	28.89	28.55
Random	5.5M	31.22	33.85	30.53	32.37
Luong [4]	17.9 M	N/A	N/A	32.21	35.03
Axelrod [1]	9.0M	32.06	34.81	32.26	35.54
Chen [2]	7.3M	31.42	33.78	30.32	33.81
δ_f	7.3M	31.46	33.13	32.13	34.81
δ_e	3.7M	**32.08**	35.94	32.84	36.56
δ_{fe}	5.5M	31.79	**35.66**	32.67	36.64
δ_f+fur	7.3M	N/A	N/A	34.04	37.18
δ_e+fur	3.7M	N/A	N/A	33.88	38.04
$\delta_{fe}+fur$	5.5M	N/A	N/A	**34.52**	**39.02**

NIST ZH-EN	Sent. No.	SMT MT05	SMT MT06	NMT MT05	NMT MT06
in	430.8K	29.66	30.73	27.28	26.82
out	8.8M	29.91	30.13	28.67	27.79
$in+out$	9.3M	30.23	30.11	28.91	28.22
Random	5.7M	29.90	30.18	28.02	27.49
Luong	9.3M	N/A	N/A	29.91	29.61
Axelrod	2.2M	30.52	30.96	28.41	28.75
Chen	4.8M	30.64	31.05	28.39	28.06
δ_f	4.8M	30.90	**31.96**	29.21	30.14
δ_e	2.2M	**30.94**	31.33	30.00	30.63
δ_{fe}	5.7M	30.72	31.43	30.13	31.07
δ_f+fur	4.8M	N/A	N/A	30.80	31.54
δ_e+fur	2.2M	N/A	N/A	30.49	31.13
$\delta_{fe}+fur$	5.7M	N/A	N/A	**31.35**	**31.80**





References

[1] Amittai Axelrod et al. Domain adaptation via pseudo in-domain data selection. In *EMNLP*, 2011.

[2] Boxing Chen et al. Bilingual methods for adaptive training data selection for machine translation. In *AMTA*, 2016.

[3] Dzmitry Bahdanau et al. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[4] Minh-Thang Luong et al. Stanford neural machine translation systems for spoken language domains. In *IWSLT*, 2015.

[5] Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichro Sumita. Instance weighting for neural machine translation domain adaptation. In *EMNLP*, 2017.