Bridging the Data Gap between Training and Inference for Unsupervised Neural Machine Translation

Zhiwei He[1], Xing Wang[2], Rui Wang[1], Shuming Shi[2], Zhaopeng Tu[2]
[1]Shanghai Jiao Tong University, [2]Tencent AI Lab
hezw.tkcw@gmail.com

1. Summary

- Findings
  - Unsupervised neural machine translation (UNMT) suffers from the data gap between training and inference. This data discrepancy results in the overestimation of UNMT on the previous benchmark, which is reflected in the performance gap between UNMT and supervised neural machine translation (SNMT) on the source-original test sets.
  - We identify two representative characteristics of the data gap: style gap and content gap.
- Solution
  - We propose an online self-training approach, which simultaneously uses the pseudo parallel data {natural source, translated target} to mimic the inference scenario.
- Results
  - Our method achieves significant improvement on the source-original test sets.
  - Better natural-to-natural and named entities translation (more details in the paper).

2. Data Gap

Existing methods adopt online back-translation causing data gap between training and inference.

- The model is trained with translated source ($\mathcal{X}^*$).
- But it translates natural source ($\mathcal{X}$) sentences in inference.

Types of training and inference data. * stands for translated sentences

	Source	Target
Train	$\mathcal{X}^*$	$\mathcal{Y}$
Inference	$\mathcal{X}$	$\mathcal{Y}^*$

3. The Overestimated UNMT

- Full set: SNMT≈UNMT (previous works)
- Tgt-Ori: SNMT<UNMT
- Src-Ori: SNMT>UNMT (what we need)

Model	En-Fr ⇒	En-Fr ⇐	En-De ⇒	En-De ⇐	En-Ro ⇒	En-Ro ⇐	Avg.
Full Test Set							
SNMT	38.4	33.6	29.5	33.9	33.7	32.5	33.6
UNMT	37.8	34.9	27.1	35.2	35.1	33.4	33.9
Target-Original Test Set / Translated Input							
SNMT	37.4	32.4	25.6	37.1	38.2	28.2	33.2
UNMT	**39.2**	**37.6**	**27.0**	**42.9**	**43.1**	**35.6**	**37.6**
Source-Original Test Set / Natural Input							
SNMT	**38.2**	**34.1**	**32.3**	**28.8**	**29.4**	**35.9**	**33.1**
UNMT	35.2	30.2	26.1	23.6	27.4	30.8	28.9

4. Two Factors of Data Gap

1 Style Gap

When training, the input is in translated style; while in inference, it's in the natural style.

UNMT has a lower perplexity on the translated input than on natural input

Inference Input	Perplexity
Natural	242
Translated	219

UNMT improves significantly when the input style switches from natural to translated

Model	Natural In. BLEU	Natural In. Δ	Translated In. BLEU	Translated In. Δ
SNMT	28.8	–	44.9	–
UNMT	22.5	-6.3	42.1	-2.8

2 Content Gap

The content of input in training is biased towards the target language. While the input during inference is more biased towards the source language.

10 most frequent entities in the source sentences of De-En translation. The training data of UNMT has more entities biased towards the target language English

Data	Most Frequent Name Entities
Src-Ori Test	Deutschland, Stadt, CDU, deutschen, Zeit SPD, USA, deutsche, China, Mittwoch
Tgt-Ori Test	Großbritannien, London, Trump, USA, Russland, Vereinigten Staaten, Europa Mexiko, Amerikaner, Obama
UNMT Train	Deutschland, dpa, USA, China, Obama, Stadt Hause, Europa, Großbritannien, Russland

UNMT model outputs the hallucinated translation "U.S." which is biased towards target language English

Input	Die deutschen Kohlekraftwerke ... in Deutschland emittierten ...
Ref	German coal plants , ..., total amount emitted in Germany.
SNMT	..., German coal-fired power stations ..., emissions in Germany.
UNMT	U.S. coal-fired power plants ... amount emitted in the U.S. ... .

5. Our Approach

We incorporate the self-training method into UNMT framework to remedy the data gap between the training and inference.
Given translation task $X \to Y$, for each batch:

1. $x^* = \arg\max_x P_{Y\to X}(x \mid y; \tilde{\theta})$
2. construct sample (x^*, y)
3. reverse the sample and get (y, x^*)
4. train the model using (x^*, y) and (y, x^*) [a]

[a]UNMT models are typically bi-directional.

6. Experiments

Our method achieves significant improvement on the source-original test sets

Testset	Model	Approach	En-Fr ⇒	En-Fr ⇐	En-De ⇒	En-De ⇐	En-Ro ⇒	En-Ro ⇐	Avg.	Δ
Our Implementation										
Full set	XLM	UNMT	37.4	34.5	27.2	34.3	34.6	32.7	33.5	–
		+Self-training	**37.8**	**35.1**	**28.1**	**34.8**	**36.2**	**33.9**	**34.3**	+0.8
	MASS	UNMT	37.8	34.9	27.1	35.2	35.1	33.4	33.9	–
		+Self-training	**38.0**	**35.2**	**28.9**	**35.6**	**36.5**	**34.0**	**34.7**	+0.8
Src-Ori	XLM	UNMT	34.7	**30.4**	26.6	22.5	27.4	30.6	28.7	–
		+Self-training	**35.4**⇑	30.2	**28.0**⇑	**23.1**↑	**29.6**⇑	**32.7**⇑	**29.8**	+1.1
	MASS	UNMT	35.2	30.2	26.1	23.6	27.4	30.8	28.9	–
		+Self-training	**35.9**⇑	**30.9**↑	**28.7**⇑	**24.9**⇑	**30.1**⇑	**31.9**⇑	**30.4**	+1.5

Natural-to-natural and named entities translation

Note: HQ(R) and HQ(all 4) are natural-to-natural test sets provided by Google (detailed in the paper).

Model	HQ(R)	HQ(all 4)
XLM+UNMT	24.5	19.6
+Self-training	**25.9**	**20.7**
MASS+UNMT	24.3	19.6
+Self-training	**26.0**	**20.8**

Model	Approach	NE Acc.
XLM	UNMT	0.46
	+Self-training	**0.53**
MASS	UNMT	0.44
	+Self-training	**0.52**