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

Presented by Zhiwei He (hezw.tkcw@gmail.com)

Shanghai Jiao Tong University Tencent AI Lab

Outline

Outline

Unsupervised Neural Machine Translation (UNMT)

- Goal: train a neural machine translation (NMT) system using only monolingual corpora

Unsupervised Neural Machine Translation (UNMT)

- Goal: train a neural machine translation (NMT) system using only monolingual corpora
- A critical component of UNMT: online back-translation (BT)

Unsupervised Neural Machine Translation (UNMT)

- Goal: train a neural machine translation (NMT) system using only monolingual corpora
- A critical component of UNMT: online back-translation (BT)

Steps of online back-translation

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. train the model using (x^*, y)

---

* denotes translated text.

Data Gap between Training and Inference of UNMT

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

Table 1: Types of training and inference data. $*$ stands for translated sentences.

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

* denotes translated sentences.

Data Gap between Training and Inference of UNMT

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

Table 1: Types of training and inference data. $*$ stands for translated sentences.

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

The source discrepancy between training and inference hinders the translation performance of UNMT models.

* denotes translated sentences.

Outline

The Overestimated UNMT

Supervised NMT (SNMT) v.s. Unsupervised NMT (UNMT)

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

The Overestimated UNMT

Supervised NMT (SNMT) v.s. Unsupervised NMT (UNMT)

Model	**En-Fr** $\Rightarrow$	**En-Fr** $\Leftarrow$	**En-De** $\Rightarrow$	**En-De** $\Leftarrow$	**En-Ro** $\Rightarrow$	**En-Ro** $\Leftarrow$	**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

- Full set: UNMT $\approx$ SNMT (previous works)

The Overestimated UNMT

Supervised NMT (SNMT) v.s. Unsupervised NMT (UNMT)

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

- Two parts of the test set
  - target-original: sentence pairs originally written in target language
  - source-original: sentence pairs originally written in source language

- Full set: UNMT $\approx$ SNMT (previous works)

The Overestimated UNMT

Supervised NMT (SNMT) v.s. Unsupervised NMT (UNMT)

Model	En-Fr $\Rightarrow$	En-Fr $\Leftarrow$	En-De $\Rightarrow$	En-De $\Leftarrow$	En-Ro $\Rightarrow$	En-Ro $\Leftarrow$	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

- Two parts of the test set
  - target-original: sentence pairs originally written in target language
  - source-original: sentence pairs originally written in source language

- Full set: UNMT $\approx$ SNMT (previous works)
- Tgt-Ori: UNMT $>$ SNMT

The Overestimated UNMT

Supervised NMT (SNMT) v.s. Unsupervised NMT (UNMT)

Model	**En-Fr**		**En-De**		**En-Ro**		**Avg.**
	$\Rightarrow$	$\Leftarrow$	$\Rightarrow$	$\Leftarrow$	$\Rightarrow$	$\Leftarrow$	
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

- Two parts of the test set
  - target-original: sentence pairs originally written in target language
  - source-original: sentence pairs originally written in source language

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

The Overestimated UNMT

Supervised NMT (SNMT) v.s. Unsupervised NMT (UNMT)

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

- Two parts of the test set
  - target-original: sentence pairs originally written in target language
  - source-original: sentence pairs originally written in source language

UNMT is overestimated on the previous benchmark.

Outline

Style Gap

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

Style Gap

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

Inference Input	PPL
Natural	242
Translated	219

Table 2: UNMT has a lower perplexity on the translated input than on natural input.

Style Gap

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

Inference Input	PPL
Natural	242
Translated	219

Table 2: UNMT has a lower perplexity on the translated input than on natural input.

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

Table 3: The performance of UNMT is significantly improved after the input is switched from natural to translated style.

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.

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.

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

10 most frequent entities in the source sentences of De-En translation

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.

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

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 rather than the expected source language German.

Content Gap - Hallucinated Translation

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

Table 4: Example translation that the UNMT model outputs the hallucinated translation "U.S.", which is biased towards target language English.

Outline

Online Self-training

Recap: steps of online back-translation

Given translation task $X \rightarrow Y$, for each batch:

1. $x^* = \arg\max_x P_{Y \rightarrow X}(x \mid y; \tilde{\theta})$
2. construct sample (x^*, y)
3. train the model using (x^*, y)

Ours: steps of online self-training

Given translation task $X \rightarrow Y$, for each batch:

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

[1] UNMT models are typically bi-directional.

Main Results

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
Trg-Ori	XLM	UNMT	39.1	36.5	**26.6**	42.2	42.1	**34.4**	36.8	–
		+Self-training	**39.3**	**37.8**	26.5	**42.4**	**42.9**	34.1	**37.2**	+0.4
	MASS	UNMT	**39.2**	**37.6**	27.0	**42.9**	**43.1**	**35.6**	**37.6**	–
		+Self-training	39.0	37.3	**27.7**	42.7	42.9	35.3	37.5	-0.1
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

Main Results

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
Trg-Ori	XLM	UNMT	39.1	36.5	**26.6**	42.2	42.1	**34.4**	36.8	–
		+Self-training	**39.3**	**37.8**	26.5	**42.4**	**42.9**	34.1	**37.2**	+0.4
	MASS	UNMT	**39.2**	**37.6**	27.0	**42.9**	**43.1**	**35.6**	**37.6**	–
		+Self-training	39.0	37.3	**27.7**	42.7	42.9	35.3	37.5	-0.1
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

Main Results

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
Trg-Ori	XLM	UNMT	39.1	36.5	**26.6**	42.2	42.1	**34.4**	36.8	–
		+Self-training	**39.3**	**37.8**	26.5	**42.4**	**42.9**	34.1	**37.2**	+0.4
	MASS	UNMT	**39.2**	**37.6**	27.0	**42.9**	**43.1**	**35.6**	**37.6**	–
		+Self-training	39.0	37.3	**27.7**	42.7	42.9	35.3	37.5	-0.1
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

Output Fluency

Approach	En-Fr		En-De		En-Ro		Avg.
	$\Rightarrow$	$\Leftarrow$	$\Rightarrow$	$\Leftarrow$	$\Rightarrow$	$\Leftarrow$	
XLM							
UNMT	101	147	250	145	152	126	154
+ST	101	144	253	147	156	138	157
MASS							
UNMT	100	145	256	144	143	119	151
+ST	103	146	263	142	156	133	157

- We evaluate the output fluency in terms of perplexity (PPL) with trained language models.

Output Fluency

Approach	En-Fr $\Rightarrow$	En-Fr $\Leftarrow$	En-De $\Rightarrow$	En-De $\Leftarrow$	En-Ro $\Rightarrow$	En-Ro $\Leftarrow$	Avg.
XLM							
UNMT	101	147	250	145	152	126	154
+ST	101	144	253	147	156	138	157
MASS							
UNMT	100	145	256	144	143	119	151
+ST	103	146	263	142	156	133	157

- We evaluate the output fluency in terms of perplexity (PPL) with trained language models.
- Slight impact on the fluency of model outputs, with the average PPL of XLM and MASS models only increasing by $+3$ and $+6$, respectively.

Natural-to-Natural Translation

Model	HQ(R)	HQ(all 4)
Supervised Model	35.0	27.2
XLM+UNMT	24.5	19.6
+Self-training	**25.9**	**20.7**
MASS+UNMT	24.3	19.6
+Self-training	**26.0**	**20.8**

- Google provides natural-to-natural test sets based on WMT19 En⇒De, whose references have been paraphrased by experts[1].

[1] https://github.com/google/wmt19-paraphrased-references

Natural-to-Natural Translation

Model	HQ(R)	HQ(all 4)
Supervised Model	35.0	27.2
XLM+UNMT	24.5	19.6
+Self-training	**25.9**	**20.7**
MASS+UNMT	24.3	19.6
+Self-training	**26.0**	**20.8**

- Google provides natural-to-natural test sets based on WMT19 En⇒De, whose references have been paraphrased by experts[1].
- We adopt the HQ(R) and HQ(all 4), which have higher human adequacy rating scores.

[1] https://github.com/google/wmt19-paraphrased-references

Natural-to-Natural Translation

Model	HQ(R)	HQ(all 4)
Supervised Model	35.0	27.2
XLM+UNMT	24.5	19.6
+Self-training	**25.9**	**20.7**
MASS+UNMT	24.3	19.6
+Self-training	**26.0**	**20.8**

- Google provides natural-to-natural test sets based on WMT19 En⇒De, whose references have been paraphrased by experts[1].
- We adopt the HQ(R) and HQ(all 4), which have higher human adequacy rating scores.
- Our proposed method outperforms baselines on both kinds of test sets.

[1] https://github.com/google/wmt19-paraphrased-references

Named Entities Translation

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

- Our proposed method achieves a significant improvement in the translation accuracy of named entities compared to the baseline.

Outline

Summary

- We first point out the data gap between training and inference for UNMT.

Summary

- We first point out the data gap between training and inference for UNMT.
- Previous benchmark overestimates UNMT models $\rightarrow$ use source-original test set

Summary

- We first point out the data gap between training and inference for UNMT.
- Previous benchmark overestimates UNMT models → use source-original test set
- We identify two critical factors: style gap and content gap.

Summary

- We first point out the data gap between training and inference for UNMT.
- Previous benchmark overestimates UNMT models $\rightarrow$ use source-original test set
- We identify two critical factors: style gap and content gap.
- We propose a simple and effective approach for incorporating the self-training method into the UNMT framework to remedy the data gap.

Summary

- We first point out the data gap between training and inference for UNMT.
- Previous benchmark overestimates UNMT models → use source-original test set
- We identify two critical factors: style gap and content gap.
- We propose a simple and effective approach for incorporating the self-training method into the UNMT framework to remedy the data gap.
- Code, data, and trained models are available: `https://github.com/zwhe99/SelfTraining4UNMT`