NTT's Machine Translation Systems for WMT19 Robustness Task

Soichiro Murakami[†*] **Makoto Morishita**[§*] Tsutomu Hirao[§] Masaaki Nagata[§]

[†]Service Innovation Department, NTT DOCOMO, INC., Japan

[§]NTT Communication Science Laboratories, NTT Corporation, Japan

[*]Equal contribution

1. Abstract

- We participated in En-Ja and Ja-En tasks.
- Our system combined techniques including
  ① **utilization of a synthetic corpus**,
  ② **domain adaptation**,
  ③ **placeholder mechanism**.
- The placeholder mechanism improves translation accuracy even with noisy texts.

Translating text on social media is a challenging task due to **various noise.**

(a)	I'll let you know bro, thx "abbreviations"
(b)	She had a ton of rings. "grammatical errors"
(c)	oh my god it's beatiful "misspellings"
(d)	Thank you so much for all your advice!!😭💕 "emojis"
(e)	(\＊´∀`＊) so cute "emoticons"

Table1: Example of comments from Reddit.

2. System Details

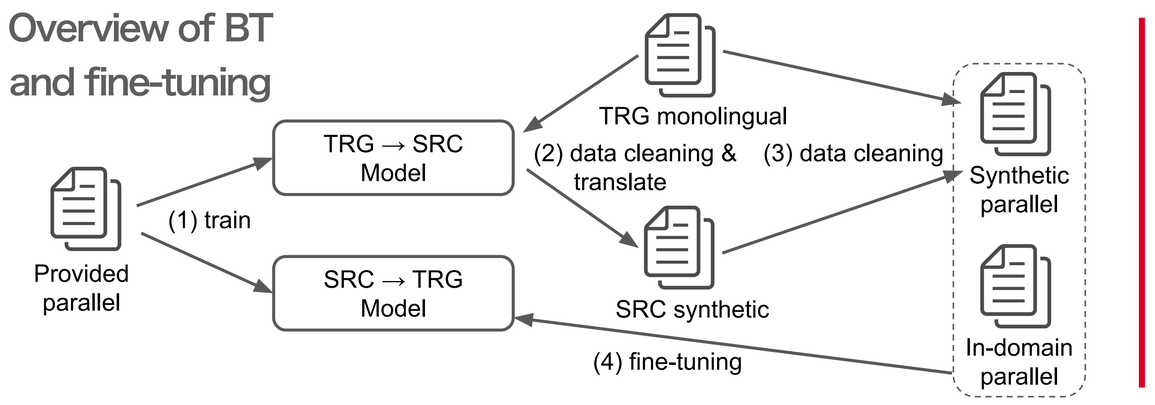
① Utilization of a Synthetic Corpus
② Domain Adaptation

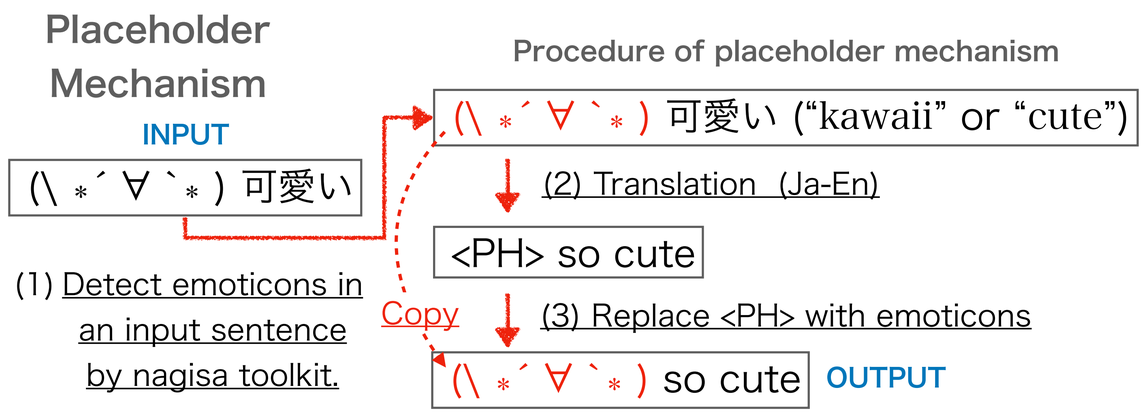
Problem

- The lack of an in-domain parallel corpus.

Method

(1) Construct SRC→TRG and TRG→SRC models.
(2) Create a synthetic corpus with TRG→SRC.
(3) Apply filtering techniques to synthetic corpus.
(4) Fine-tune the SRC→TRG model on both synthetic corpus and in-domain parallel.





③ Placeholder Mechanism

Problem

- Noisy text on social media often contains tokens that do not require translation such as emojis "❤️, 😎" and emoticons, "＼(^o^)／, (*^^*)".

Method

- We replace the emojis and emoticons with the placeholders "<PH>" in the training data.
- In the prediction phase, we copy the emojis and emoticons from SRC, and replace "<PH>" with them.

3. Experimental Results

Effect of Fine-tuning

- The placeholder mechanism achieved improvements of **+1.4 points** for Ja-En and **+0.7 points** for En-Ja.

Input	Woah woah, hang on a minute, let's hear this guy out. Amazing title 😂
Baseline	うわぁ ちょっと 待 っ て こいつ の 話 を 聞 い て み ま しょ う 驚 く よう な 名前 だっ た わ ね (Well wait a minute let's listen to this story It was an amazing name)
+ Fine-tuning	うわー 、 うわー 、 ちょっと 待 っ て 、 この 男 の 話 を 聞 こ う ぜ 。 すご い タイトル だ 😂 (Wow, wow, wait a minute and hear this guy talk. It's an amazing title 😂 .)

Effect of Placeholders

- The Fine-tuning with in-domain and synthetic corpora led to substantial gain: **+3.2 points** for Ja-En and **+1.9 points** for En-Ja.

	Case 1	Case 2
Input	(つ・ω・)つ許す！	かわいい♪(*・ω・人)
Reference	(つ・ω・)つ I Approve!	Kawaii♪(*・ω・人)
Our system	(つ ・ω・) I forgive you!	Cute (*・ω・人)
Another system	I forgive you!	It's cute.

	Ja-En		En-Ja	
Baseline	10.8		14.3	
+placeholders	12.2	(+1.4)	15.0	(+0.7)
+fine-tuning	11.9	(+1.1)	16.2	(+1.9)
+synthetic	14.0	(+3.2)	—	
+ 4-model ensemble	14.9	(+4.1)	17.0	(+2.7)
Submission	14.8		17.0	

Table2: Case-sensitive BLEU scores

	Improved	Degraded	Unchanged
Ja-En	9 (53%)	0 (0%)	8 (47%)
En-Ja	14 (82%)	1 (1%)	2 (12%)

Table3: The number of improved/degraded sentences by the placeholder mechanism compared with the baseline.