PheMT: A **Phe**nomenon-wise Dataset for **M**achine **T**ranslation Robustness on User-Generated Contents

Ryo Fujii[1], Masato Mita[2,1], Kaori Abe[1], Kazuaki Hanawa[2,1], Makoto Morishita[3,1], Jun Suzuki[1,2], Kentaro Inui[1,2]

1. Tohoku University 2. RIKEN 3. NTT Communication Science Laboratories

TOHOKU UNIVERSITY

Summary

- A new dataset for evaluating the robustness of Japanese-to-English MT systems on UGC
- Provide focused evaluation on four linguistic phenomena with the idea of contrastive datasets
- Evaluated the effect of the phenomena with both in-house and widely used off-the-shelf systems
- Discovered a unique preprocessing method towards improving the performance on *Variant*

Background

- UGC are prevailing in our real-life communication
  - e.g., social media, blog posts, user reviews
- A shared task on machine translation robustness [1]

More attention towards handling UGC to promote cross-cultural communication

The performance of current MT systems on UGC is still far behind

Q. Why is it difficult to translate UGC ?

Still not clear…

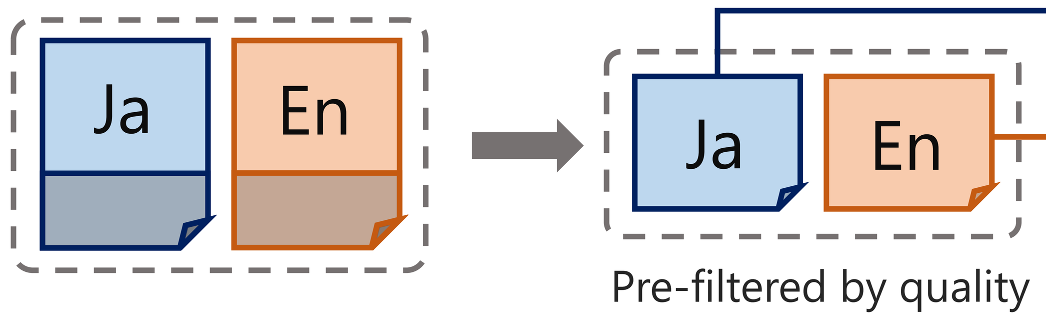
We need a solid basis for more detailed analysis !

[1] Li et al. (2019), Findings of the first shared task on machine translation robustness.

[2] Michel and Neubig (2018), MTNT: A Testbed for Machine Translation of Noisy Text.

Creating phenomenon-wise dataset

MTNT corpus [2]



Label *	Examples { Orig. / Norm. }
Proper Noun	安倍首相 (*abeshushō*, meaning PM Abe)
Abbreviated Noun	{ アプデ / アップデート } (*apude*, meaning update)
Colloquial Expression	{ かなちい / かなしい } (*kanachii* / *kanashii*, meaning sad)
Variant	{ アリガトウ / ありがとう } (*arigatou*, meaning thank you)

* Please refer to the paper for the definition

Step1: Annotating phenomena labels

Proper / *Abbrev.* / *Colloq.* / *Variant* / *None*

Step2: Extracting targeted expressions / alignments

Ja: 地味な**アプデ** (*apude*, meaning update, abbreviated) だが

En: That's a plain **update** though

Step3: Normalizing the expressions (except *Proper Noun*)

Orig.: 地味なアプデ (*apude*) だが

Norm.: 地味な**アップデート** (*update*, canonical) だが

Translation models

- The five **in-house models** :

Q1. Effect of training data size ?

1. SMALL 3.9 M pairs vs. **2. LARGE** 14.0 M pairs

Q2. Effect of tokenization ?

2. LARGE BPE-based vs. **3. CHAR** Char-based

Q3. Susceptible to local improvement ?

Trained on a fully-pronunciation based corpus to absorb symbolic differences in *Variant*

4. PRON Phonetic and **5. CAT** Concatenated

- **Off-the-shelf systems** : Google, DeepL

Phenomenon-wise evaluation

Our robustness measure :

The difference of arbitrary metrics for (Orig. / Norm.) input

Translation **accuracy** with extracted alignment (raw acc. only for *Proper*)

	SMALL	LARGE	CHAR	PRON	CAT	Google	DeepL
Proper	(34.3)	(49.7)	(47.1)	(43.2)	(48.0)	(55.2)	(50.5)
Abbrev.	+6.4 (24.1 / 30.5)	-0.6 (33.6 / 33.0)	+0.6 (34.2 / 34.8)	+1.1 (30.2 / 31.3)	-1.2 (34.2 / 33.0)	-4.3 (41.1 / 36.8)	-1.2 (39.1 / 37.9)
Colloq.	+5.8 (18.0 / 23.8)	+9.9 (14.5 / 24.4)	+4.1 (17.4 / 21.5)	+21.5 (8.7 / 30.2)	+16.9 (15.7 / 32.6)	+7.0 (19.2 / 26.2)	+5.8 (22.7 / 28.5)
Variant	+19.5 (15.5 / 35.0)	+25.2 (13.6 / 38.8)	+20.4 (13.6 / 34.0)	+10.7 (25.2 / 35.9)	+8.8 (26.2 / 35.0)	+14.6 (23.3 / 37.9)	+16.6 (18.4 / 35.0)

A1. High coverage with larger training data was effective for nouns, while not for UGC-specific phenomena

A2. Char-based tokenization worked well with *Colloq.*, which share most of the characters with their canonical forms

A3. Our dataset could detect the improvement against *Variant*, which was proven to be more problematic to current systems