**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.


---

**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**

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

<table>
<thead>
<tr>
<th>Translation accuracy with extracted alignment (raw acc. only for Proper)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALL</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Proper</td>
</tr>
<tr>
<td>Abbrev.</td>
</tr>
<tr>
<td>Abbrev.</td>
</tr>
<tr>
<td>Colloq.</td>
</tr>
<tr>
<td>Colloq.</td>
</tr>
<tr>
<td>Variant</td>
</tr>
<tr>
<td>Variant</td>
</tr>
</tbody>
</table>

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