**TNFormer: Single-Pass Multilingual Text Normalization with a Transformer Decoder Model**

Binbin Shen, Jie Wang, Meng Meng, Yujun Wang

Xiaomi Inc., Beijing, China

## 1. MOTIVATION

### Challenges in Text-to-Speech (TTS) Systems
- TTS systems must convert **varied text forms** into a **canonical format** for accurate synthesis.
- Contextual ambiguities in text pose significant challenges in normalization.

### Innovations of TNFormer
- **Single-Pass TN**: Efficiently identifies and normalizes Non-Standard Words (NSWs) in one go.
- **Multilingual Support**: Effectively handles normalization for both English and Chinese datasets.
- **Context-Driven**: Capable of understanding the surrounding context to improve accuracy.

## 2. RELATED WORK

### Traditional Methods
- Rule-based systems and WFSTs often struggle with context-dependent inputs and are not easily scalable.

### Neural Models
- Enhanced accuracy with neural models, often in two steps: locating non-standard words and contextual normalization.

### Advancements
- Hybrid approaches combining rule-based and neural systems have emerged for better context handling.

## 3. PROPOSED APPROACH

### TNFormer Model
- A **decoder-only** Transformer architecture designed for single-pass text normalization.

### Key Features
- Leverages pre-trained GPT-2 models fine-tuned for English and Chinese languages.
- Employs position markers and a `<instart>` token to facilitate the normalization process.
- Outputs normalized text alongside position information.

### Source Text Validation
- **Accuracy Verification**: Ensures the transcribed source text matches the predicted start and end positions.
- **Error Correction**: In cases of discrepancies, the source text is re-transcribed to align with the correct positions.
- **Handling Omissions**: Detects and reintroduces any missing elements in the input text sequence.

## 4. EXPERIMENTS

### Datasets
- **English**: Google Text Normalization dataset (GoogleTN)
- **Chinese**: FlatTN and an in-house developed Internal Chinese TN Dataset

### Methodology
- Position markers assigned to facilitate normalization based on space-delimited words; Chinese text pre-tokenized.
- Models trained on respective datasets using TensorFlow and the Transformers library.

### Results
- TNFormer demonstrates superior performance compared to several existing models.

<table>
<thead>
<tr>
<th>Category</th>
<th>TNFormer Model</th>
<th>GoogleTN</th>
<th>FlatTN</th>
<th>Internal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance (%)</td>
<td>F1-score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text Normalization</td>
<td>99.27</td>
<td>99.38</td>
<td>99.33</td>
<td></td>
</tr>
</tbody>
</table>

## 5. CONCLUSIONS

### Model Efficacy
- Effectively transforms text normalization into a next-token prediction problem, enhancing efficiency.
- Exhibits strong performance across different languages without being explicitly designed for multilingual support.

### Future work
- Handling more complex text and multilingual mixtures.
- Integrating with covering grammars to handle unrecoverable errors.

### More questions?
- For inquiries or further information about TNFormer, please contact (shenbinbin, wangjie50)@xiaomi.com