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

Binbin Shen, Jie Wang, Meng Meng, Yujun Wang

Xiaomi Inc., Beijing, China

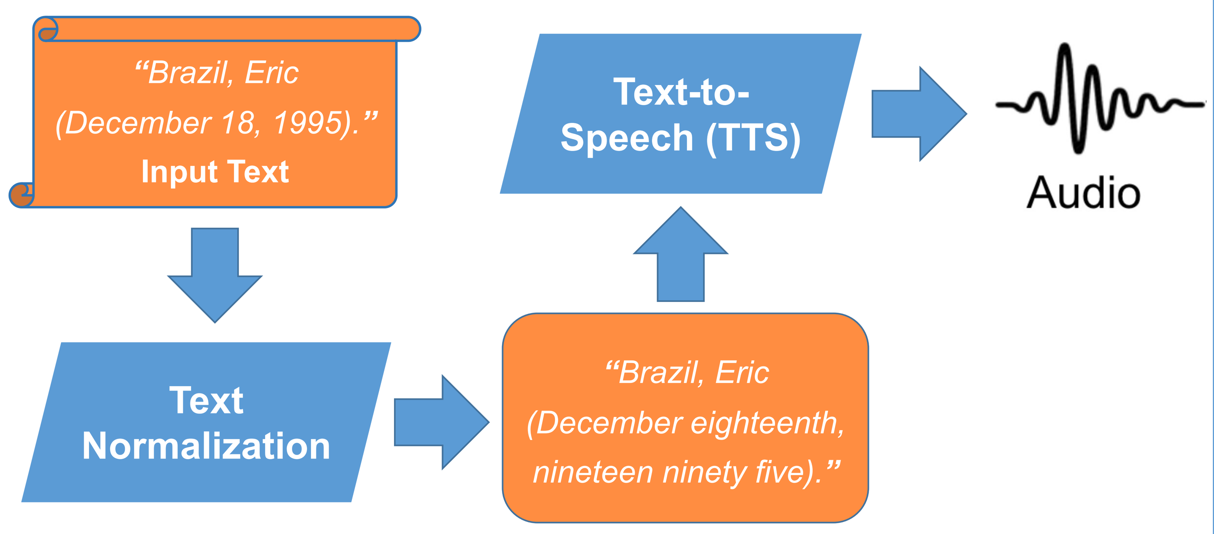
1. MOTIVATION

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

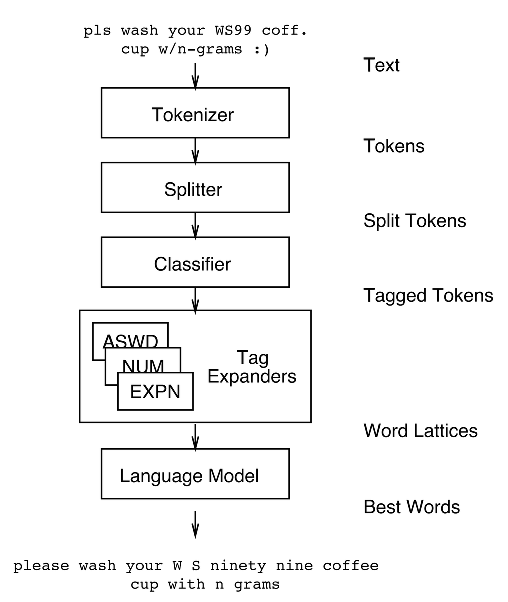
- Rule-based systems and WFSTs often struggle with context-dependent inputs and are not easily scalable.

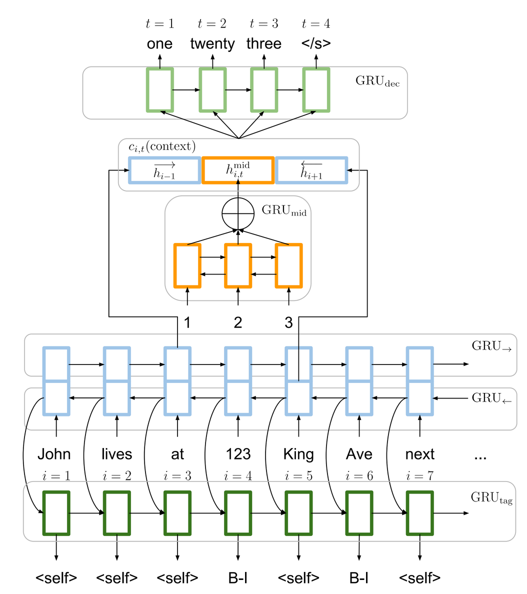
Neural Models

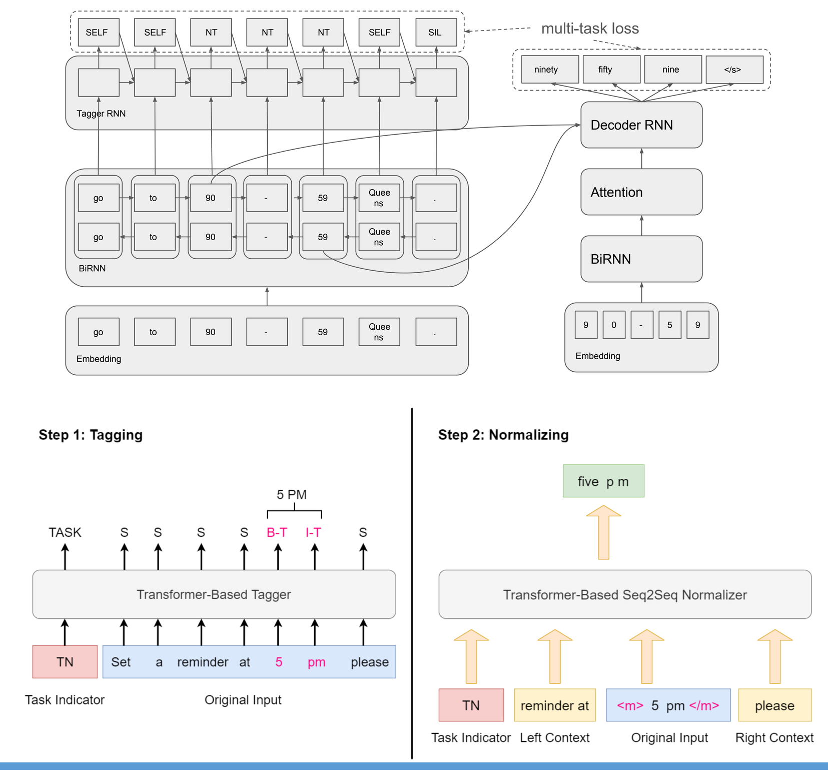
- Enhanced accuracy with neural models, often in two steps: locating non-standard words and contextual normalization.

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

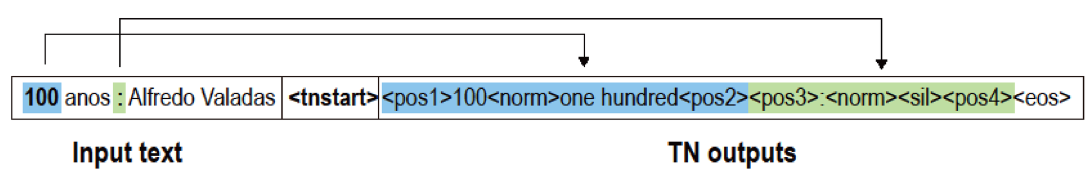
- Leverages pre-trained GPT-2 models fine-tuned for English and Chinese languages.
- Employs position markers and a <tnstart> token to facilitate the normalization process.
- Outputs normalized text alongside position information.

Source Text Validation

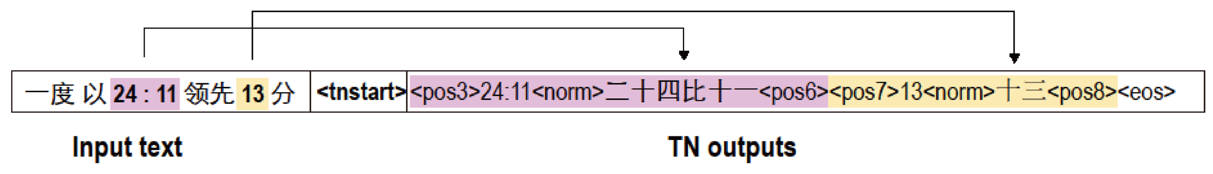
- Accuracy Verification: Ensures the transcribed source text matches the predicted start and end positions.
- Error Correction: In cases of discrepancies, the source text is re-transcribed to align with the correct positions.
- Handling Omissions: Detects and reintroduces any missing elements in the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

Two Examples:

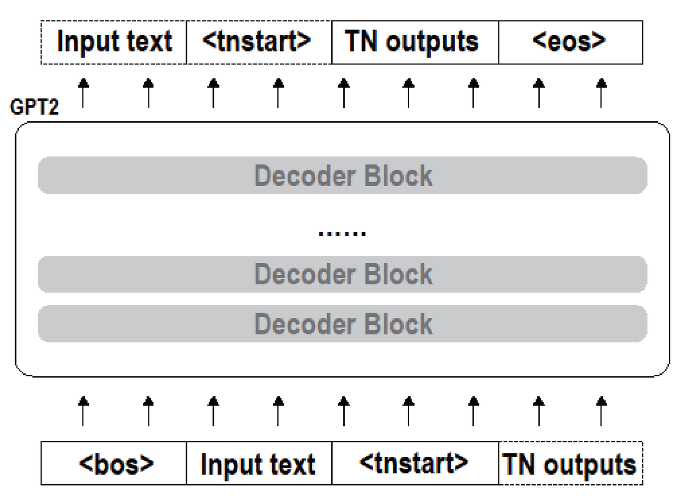
English example:



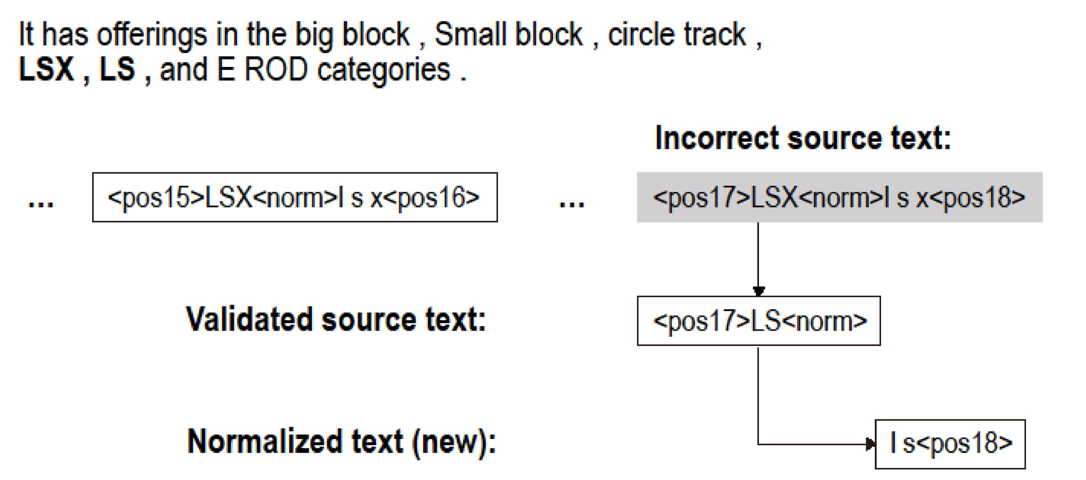
Chinese example:



TNFormer Model:



Source Text Validation:



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.

Performance on GoogleTN test set

Model	Sentence Accuracy (%)
Duplex [15]	98.41
WFST+LM [19]	97.79
tokenized single-pass edits [16]	97.96
two-pass BERT fine-tuned [16]	**98.58**
TNFormer-En	98.27

Performance on FlatTN test set (F1-score)

Category	FlatTN [18]	TNFormer-Zh
PUNC	**0.9965**	0.9943
MINUTE_CARDINAL	0.9851	**0.9907**
POINT	0.9689	**0.9823**
CARDINAL	0.9641	**0.9898**
DIGIT	0.9527	**0.9807**
SLASH_PER	**0.9412**	0.9375
HYPHEN_RATIO	0.9375	**0.9565**
VERBATIM	0.9057	**0.9183**
HYPHEN_RANGE	0.8599	**0.9226**
HYPHEN_IGNORE	0.8428	**0.9548**

Ablation study

Configuration	GoogleTN	FlatTN	Internal
normal	**98.27**	**93.26**	**97.33**
w/o src text val	97.92	91.79	95.22

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