Utilizing Phonetic Information for Textual NLP Tasks

Kenan Tang

June 2nd

Outline

- Motivating examples*
- Definition of phonetic units and features
- Conversion systems and challenges
- Applications in NLP

* Consider other languages?

Outline

- Motivating examples*
- Definition of phonetic units and features
- Conversion systems and challenges
- Applications in NLP

Motivating Examples*

- User-generated content (UGC) normalization
  - This is 2 difficult 4 me.
- Spelling correction
  - discreet math
- Historical text normalization
  - canuaise, likelyhode, resberyes
- Named entity recognition (NER)
  - shikago シカゴ

Phonotactic segmentation. Humor detection. Transliteration.

Motivating Examples

- Distance
  - Meaning (semantic): low
  - Spelling (orthographic): high
  - Pronunciation (phonetic): low

- Pipeline?
  - Spelling → pronunciation

Other distance combinations. Other pipelines.

Outline

- Motivating examples*
- Definition of phonetic units and features
- Conversion systems and challenges
- Applications in NLP

Phonemes and Phones*

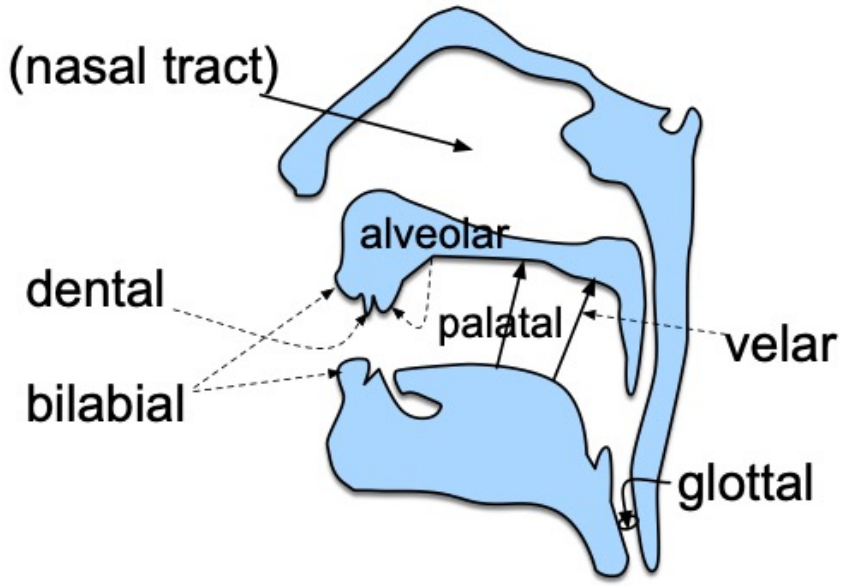
- One example:
  - peak /pik/ [pʰik]
  - speak /spik/ [spik]

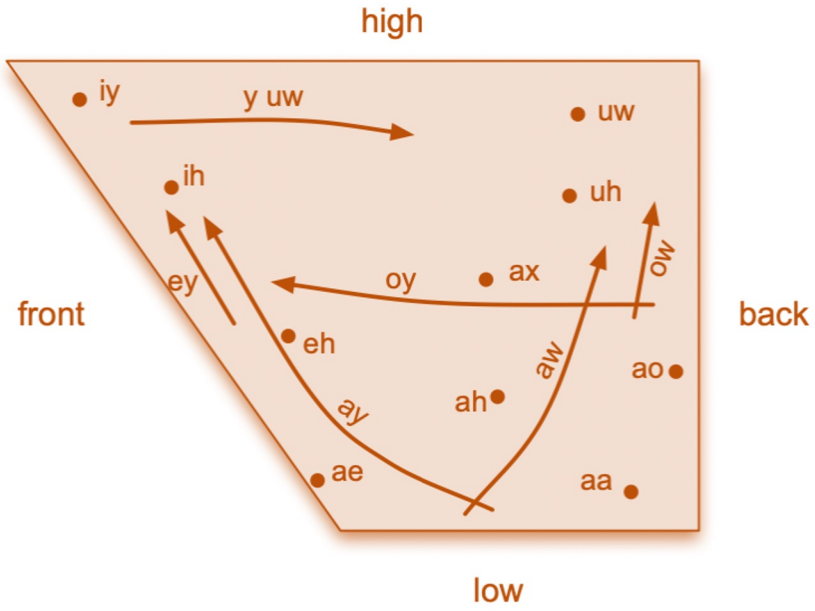
- Another example:
  - fight /faɪt/ [faɪt]
  - write /ɹaɪt/ [ɹaɪt]

Alternative phone inventories.

Articulatory Features

- Illustration (from SLP3)
  - Left: consonants; Right: vowel, schematic





Articulatory Features Quantification

- Vector form
  - ± nasal
  - ± labial
  - ± high
  - ± low
  - etc.

- Distance calculation

Missing ablation studies. Alternative definitions of distance.

Outline

- Motivating examples*
- Definition of phonetic units and features
- Conversion systems and challenges
- Applications in NLP

G2P Conversion Systems

- Pipeline: Grapheme → Phoneme → Articulatory Features

- Two challenges
  - Context dependency
  - OOV words

Context Dependency

- Example*
  - I see an *object* in the distance.
  - Sustainability groups and some New York state lawmakers *object* to the practice because of the environmental impact.

- Solution
  - Sentence-level datasets
  - Ignore stress and tones

OOV words

- Example*
  - path
  - pothole

- Solution
  - Letter-level transduction models
  - Ignore ambiguity

Outline

- Motivating examples*
- Definition of phonetic units and features
- Conversion systems and challenges
- Applications in NLP

Application 1: UGC Normalization

- Jahjah et al. (2016)

- Rule-based signature generation
  - Simplified IPA symbol
  - Examples: ks → X, as in the alternative spelling *axent* of *accent*
  - About 100 rules and 200 exceptions for 78,000 word-IPA pairs
- (Heuristic) similarity matching by signature

Application 2: Spelling Correction

- Zhang et al. (2021)

- MLM-phonetics
- Linear combination

MLM-phonetics

- Intuition*: misspellings with close pronunciation

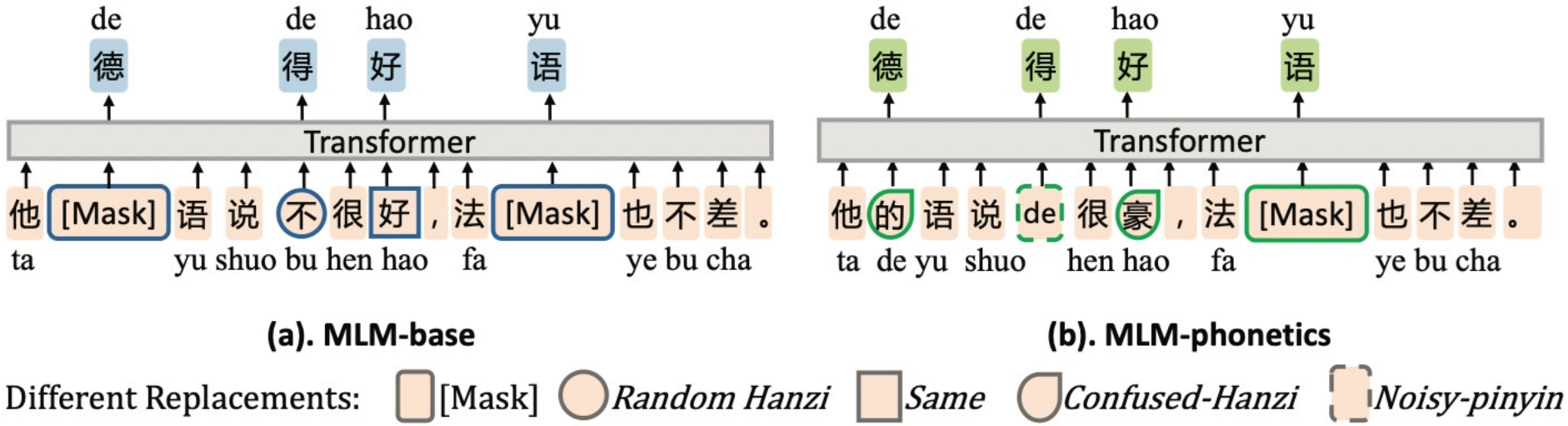
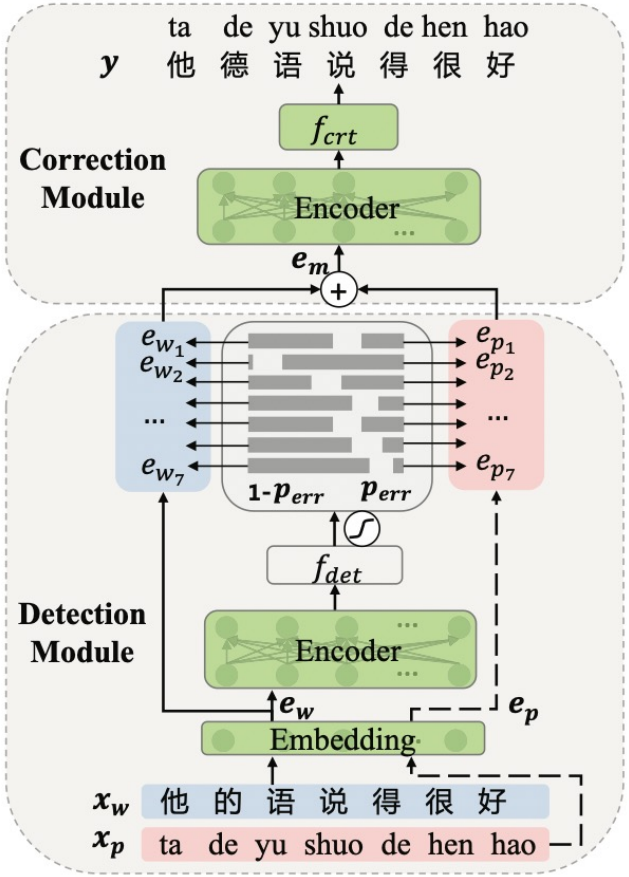


Figure 3: An example of the different replacement strategy for **MLM-base** and **MLM-phonetics**.

Alternative pre-training from audio.

Linear Combination

- Detection module
  - p_{err} from only x_w
- Weighted combination
  - $e_m = (1 - p_{err}) \cdot e_w + p_{err} \cdot e_p$



Alternative weighting.

Application 3: Historical Text Search

- Jurish (2010)

- Equivalent when pronounced the same
  - likelyhode / lykelyhood → likelihood
- Much faster than purely textual rewriting rules
- Improves recall when used together with rewriting rules

Application 4: NER

- Bharadwaj et al. (2016)
- Character LSTM for OOV words
- One-hot identity vector and binary feature vector as input to character LSTM

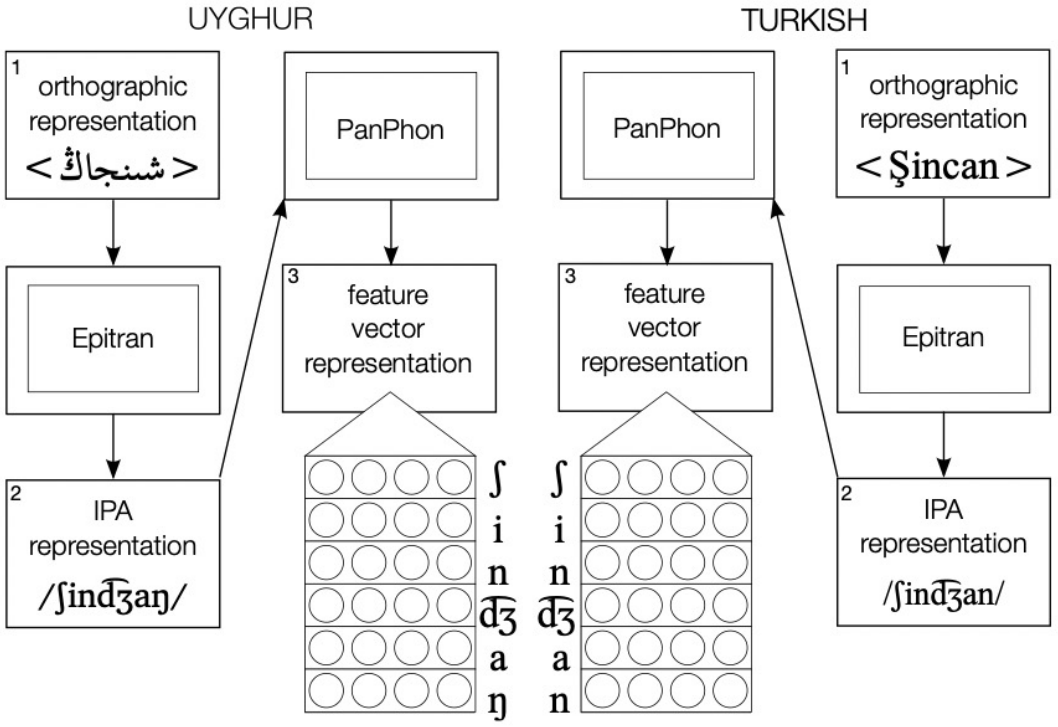


Figure 3: Use of Epitran and PanPhon to enable transfer across orthographies

Similarity (affinity) for transliteration.

Summary

- Possible improvements
  - Language-specific considerations*
  - Awareness of downstream tasks

- Full paper available at https://github.com/kt2k01/ttic-31210-survey/blob/main/phonetic.pdf

Q & A