Flexible Proper Name Search Using a Sound Approximation Approach

Ananlada Chotimongkol, Sumonmas Thatphithakkul, Santipong Thaiprayoon

National Electronics and Computer Technology Center, Thailand

{ananlada.chotimongkol, sumonmas.thatphithakkul, santipong.thaiprayoon}@nectec.or.th

Abstract

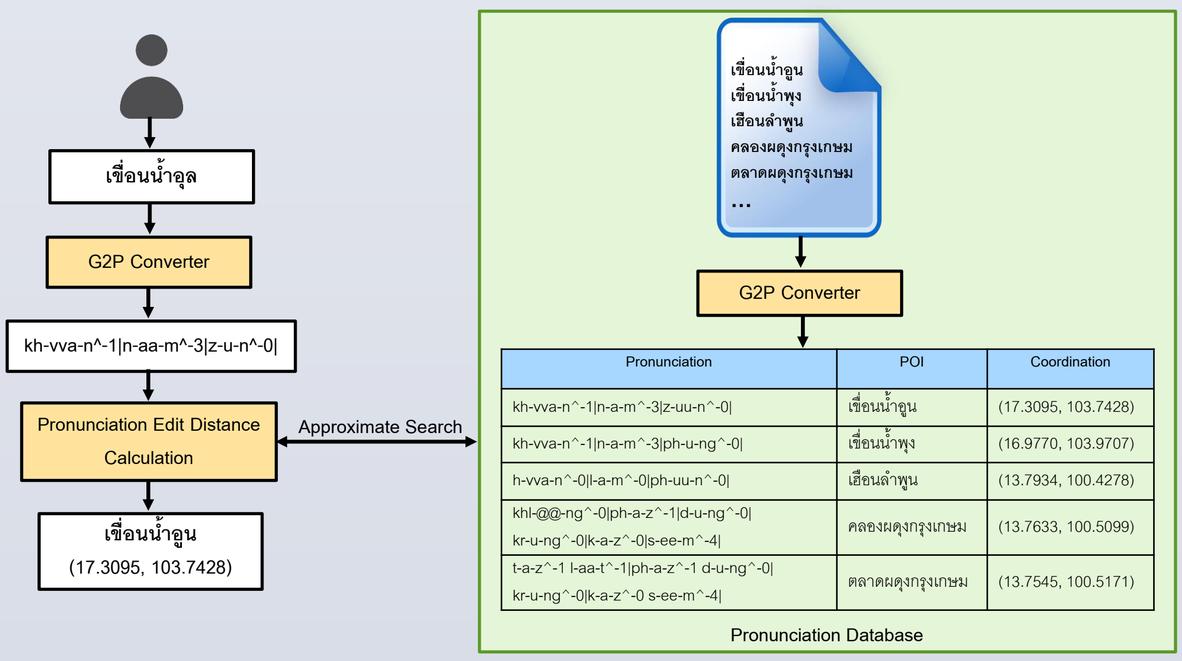
When people do not know the correct spelling of a proper name, they normally use the spelling that sounds similar instead. This phonetic error is problematic to name search applications e.g. directory search. To correct a phonetic error, namely to find a name in a database that sounds similar to a search term, we apply a grapheme-to-phoneme converter to both the search term and database and then perform approximate string matching on phoneme sequences where similarity is defined in terms of pronunciation edit distance. More improvement can be gained with modification of edit distance cost function to reflect sound similarity in Thai. We evaluated our proposed sound approximation approach on both place and person name search and achieved up to 86.5% search accuracy. The proposed approach can be easily combined with character edit distance to correct typological errors in addition to phonetic errors which achieved up to 94.2% accuracy.

Problems in Name Search

- Proper names are sometimes difficult to spell due to their uniqueness
- A misspelled search term is known to be problematic for a search engine
- There are 2 types of common errors
  1. **Typographical Error** (inaccurate typing) e.g. Anchorage is mistyped as Ancjorage
  2. **Phonetic** or **Cognitive Error** (similarly pronounced spelling is used when the correct one is not known) e.g. Anchorage is misspelled as Ankorage
- This work focuses on phonetic errors

Proposed Solution

- Correct a phonetic error by finding a name in a database that pronounced the closest to a search term
- Index names in a database by their pronunciations in terms of phoneme sequences using a grapheme-to-phoneme converter (G2P)
- Use approximate string matching to find words with similar pronunciations
  - Similarity is defined in terms of pronunciation edit distance between an indexed word and a search term
  - Cost of edit operations (insert, delete, substitute) can be modified to reflect common cognitive errors
- Pronunciation edit distance can be easily combined with character edit distance to also correct typographical errors



The Overall Process of Sound Approximation Approach

Thai Phonology

- Thai syllable structure is $Ci(C)V^{T}Cf$ or $Ci(C)V^{T}V(Cf)$
  - 21 single initial consonants (Ci) and 12 double initial consonants (CiC)
  - 12 short vowels (V) and 12 long vowels (VV)
  - 9 single final consonants (Cf)
  - 5 tones (T)
- Initial consonants and final consonants are classified by place and manner of articulation e.g. stop, nasal and glide
  - Consonants that have the same place and manner of articulation could be confusedly used
    e.g. Liquid and Trill (/r/, /l/) confusion
    ลัดดารมย์ [látdāarōm] vs. ลัดดาลมย์ [látdāalōm]
- Vowels are classified by tongue height and tongue advancement
  - A pair of short-long vowel with the same tongue height and tongue advancement could be confusedly used
    e.g. /uu/, /u/ confusion
    เขื่อนน้ำอูน [kʰɯ̀ɯannáamʔūun] vs. เขื่อนน้ำอุล [kʰɯ̀ɯannáamʔūn]
- Tones
  - Tone variations e.g. a high tone / ˊ/ becomes a falling tone / ˆ/
    กว๊านพะเยา [kwáanpʰáʔjāw] vs. กว้านพะเยา [kwâanpʰáʔjāw]

Pronunciation Edit Distance

- Calculate from the difference between phoneme sequences
- Substitution, insertion, deletion → distance = 1
  เขื่อนน้ำอุล = kh-vva-n^-1|n-aa-m^-3|z-u-n^-0|
  เขื่อนน้ำอูน = kh-vva-n^-1|n-aa-m^-3|z-uu-n^-0|
  - Pronunciation distance = 1 (substitute /u/ with /uu/)
  - Character distance = 2 (substitute / ุ / with / ู / and 'ล' with 'น')

Phoneme Similarity Edit Distance

- Adjust the cost of edit operations according to Thai phoneme similarity
  - Substitution within a similarity group → distance = 0.5
  - Other operations → distance = 1
- Similarity group
  - Consonant: similar place and manner of articulation
    e.g. {kl, kr, kw, khl, khr, khw}, {m^, n^, ng^}
  - Vowel: a short and long pair e.g. {a, aa}, {u, uu}
  - Tone variations

พดุงกุลเกษม = ph-a-z^-3|d-u-ng^-0|k-u-n^-0|k-a-z^-1|s-ee-m^-4|
ผดุงกรุงเกษม = ph-a-z^-1|d-u-ng^-0|kr-u-ng^-0|k-a-z^-1|s-ee-m^-4|

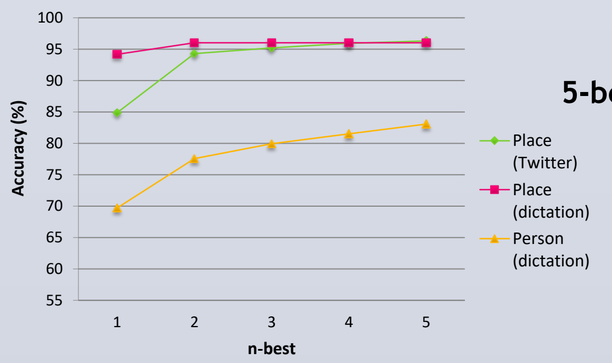
- Phoneme similarity = 1.5 (tone 3 → tone 1, /k/ → /kr/, and /n^/ → /ng^/)
- Pronunciation distance = 3
- Character distance = 3 ('พ' → 'ผ', insert 'ร', and 'ล' → 'ง')

Experiments

- We evaluated the proposed sound approximation approach by search accuracy
- 2 proper name databases
  - Point Of Interest (place name) 61,738 entries
  - Company directory (person name) 2,890 entries
- Test sets:
  - Place-Twitter: 542 misspelled place names extracted from Twitter
  - Place-Dictation: 326 misspelled place names from transcribed recorded speech
  - Person-Dictation: 254 misspelled person names from transcribed recorded speech
  - The Twitter set contains both typographical and phonetic errors while the dictation sets contain mostly phonetic errors

Approximate Search Method	Place		Person
	Twitter	Dictation	Dictation
Pronunciation distance	71.4%	83.7%	62.6%
Phoneme similarity distance	74.9%	**86.5%**	**65.8%**
Character distance	**85.1%**	84.4%	33.9%
Phoneme+Character	84.9%	94.2%	69.7%

- Phoneme similarity distance which uses cost functions that reflect sound similarity in Thai achieved higher accuracy than pronunciation distance
- Character distance is suitable for correcting typological errors
- Phoneme similarity distance is suitable for correcting phonetic errors
- The combined method which chooses a candidate with smaller distance between phoneme similarity distance and character distance achieves the highest accuracy



5-best accuracy of the combined method

- 96% accuracy on place name search
- 83% accuracy on person name search

Conclusions

- We proposed a flexible proper name search technique based on a sound approximation approach to alleviate the problem of misspelling
- Similarity is defined based on pronunciation edit distance
- We achieved up to 86.5% search accuracy on place name and 65.8% on person name with the modification of edit distance cost function to reflect sound similarity in Thai
- Pronunciation edit distance which good at correcting phonetic errors can be easily combined with character edit distance which good at correcting phonological errors, and can achieve up to 94.2% accuracy on place name search and 69.7% on person name search