

階層的ファイルを用いた中国語単語の検索法

1 H-1

丁 光躍* 手塚 祐一: 田中 栄一: 池田 満: 金 明哲:**
*西南交通大学 : 宇都宮大学工学部 **: 長春郵電学院

1. はじめに

拼音 (Pin Yin) は中国漢字の標準発音であり, 中国語の入力方式としてよく使われている。しかし, 拼音は日常的な表記法ではないため, 誤入力することが多い。この原因は幾つかある。a) 中国語には方言が多く, 地域によって発音の差が大きい。b) 拼音の音素数に地域差がある。c) 類似した音素が多い。d) 拼音の記憶間違い, 等である。

このため, 拼音表記の中国語単語を辞書上で検索する際には, 誤りを訂正する能力が望まれる。筆者らは既に大規模辞書において, 誤りのある入力英文字列から正しい英単語を高速に検索する手法[1]を提案している。この方法では階層的辞書ファイルの構成が検索効率と誤り訂正能力を考える上で重要である。

中国語の単語は1漢字単語の出現頻度が高く, 1単語中の音素数が少ないため, 英単語のような高い訂正率を持つ誤り訂正検索は期待できない。そこで, 限られた範囲の拼音入力誤りに対して対処できる方法を考える。

[2]の階層的ファイルでは, 各段で誤り訂正操作を行っているが, 本稿では, 第2段で誤り訂正操作を行い, 第1段は検索速度を上げる働きをする。また, 入力誤りは発音の似た音素の置換誤りだけを考え, 全ての候補単語を出力するようにする。

2. 中国語単語の性質

2.1 単語の構成と発音

中国語の単語は幾つかの漢字から構成される。一般に, 全ての漢字の発音は単音節 [声母+韻母] で発音する。声母と韻母は音素と言う。声母だけ, あるいは韻母だけの発音の漢字もある。例えば:

単語	拼音	音素構成
“北”	bei	声母+韻母
“蝦”	e	韻母

音節の種類は400 余りあり, それらを構成する声母の種類は21種, 韻母は36種である。

声母類: {b p m f d t l n g k h j q x z c s r
zh ch sh};

韻母類: {a o e i u v ai ao an ang ou ong ei
en eng er ia iao ie iou ian in iang
ing iong ua uai uei uo uan uen uang
ueng ve van vn}。

中国語には韻母を発音するとき, 音の高低の変化があり, これを声調という。声調は方言差が大きいため拼音で入力するとき, 通常声調は考慮しない。

2.2 拼音入力の特徴

漢字の発音は拼音入力誤りの構成的特徴を決めている。仮に, 音素を分割できない符号とすると, 拼音入力誤りは以下のような特徴を持っている[2]。

- a) 声母の脱落と挿入, 及び韻母の脱落と挿入誤りは少ない。
- b) 声母と韻母は誤りに関して独立である。即ち, 声母は声母に誤り, 韻母は韻母に誤る。
- c) 音素の誤りには, 一定の傾向が認められる。例えば:
“eng” は {en, uen, ueng, un} に間違い易いが, 他の韻母には間違いにくい。

3. 音素の分類と類名表記

誤りの傾向に基づいて, 音素を分類すると以下のようになる:

声母の分類:

A={zh, z}; B={ch, c}; C={sh, s}; D={n, l, r}
E={b, p}; F={d, t}; G={q, j}; H={g, k}; I={m};
J={f}; K={h}; L={x}; M={無声母類}。

韻母の分類:

a={ao, iao}; b={ong, iong}; c={a, ia, ua};
d={i, u, v}; e={in, ing, vn};
f={o, ou, uo, iou, iu};
g={en, eng, uen, ueng, un};
h={an, ang, ian, iang, uan, uang, van};
i={e, ei, ai, uai, ui, uei, ue, ve, ie, er};
j={無韻母類}。

ここで A, ..., M; a, ..., j を類名と言う。次に, 単語の類名表記を定義する。“bei jing” (北京)の音素を類名で書くとEiGeとなる。これを類名表記という。

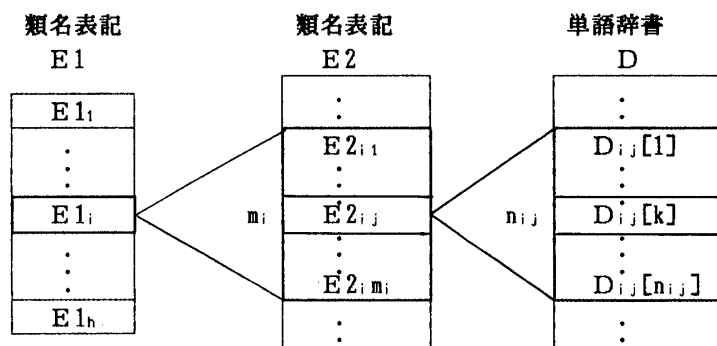


図1 辞書ファイルの構成

上記の分類を

$C1 = \{C1_1, C1_2, \dots, C1_n\}$ ($n=23$) (小分類)
とする。更に高速化[3]のために小分類をまとめた類 (大分類) を定義する。即ち、

$C2 = \{C2_1, C2_2, \dots, C2_m\}$ (大分類)

操作の高速化を図るため、音素の分類を $T = \{C_1, C_2, \dots, C_u\}$ とするとき、 C_i を数値 $(i-1)$ で表わすと、ある単語の類名表記は1つの u 進数で表わすことができる。辞書中の単語を類名表記で表わすと同一の類名表記を持つ単語が多数あるので、辞書中の単語は類名表記 $E_1, E_2, \dots, E_n$ で分類できる。分類された部分辞書に対応する類名表記を部分辞書の見出しという。

音素の誤りが類内で起こった場合、これを類内置換と言う。別の類の音素に誤った場合、類外置換と言う。

4. 階層的な中国語単語辞書ファイルの構成

3節の分類に基づいて、類名表記で作った辞書ファイルの構成を、図1に示す。

ここに

$E1_i$: 一つの大分類の類名表記 ($1 \leq i \leq h$)

h : 大分類の類名表記数

$E2_{ij}$: $E1_i$ を見出しとする一つの小分類の類名表記 ($1 \leq i \leq h, 1 \leq j \leq m_i$)

m_i : $E1_i$ を見出しとする小分類の類名表記数

$D(E1_i)$: $E1_i$ を見出しとする部分辞書

$D(E2_{ij})$: $E2_{ij}$ を見出しとする小辞書

n_{ij} : $E2_{ij}$ を見出しとする単語数

$D_{ij}[k]$: $E2_{ij}$ を見出しとする部分辞書 D_{ij} の

k 番目の単語 ($1 \leq i \leq h, 1 \leq j \leq m_i, 1 \leq k \leq n_{ij}$)

図1のように、大分類と小分類による2種類の類名表記を用いて階層ファイルを作成すると、各部分辞書の間で以下の式が成り立つ:

$$D = \bigcup_{i=1}^h D(E1_i), \quad D(E1_p) \cap D(E1_q) = \phi$$

ここで $p \neq q, 1 \leq p, q \leq h$

$$D = \bigcup_{i=1}^h \bigcup_{j=1}^{m_i} D(E2_{ij})$$

$$D(E1_i) = \bigcup_{j=1}^{m_i} D(E2_{ij}) \quad (1 \leq i \leq h)$$

$$D(E2_{ts}) \cap D(E2_{qp}) = \phi, \quad (t \neq q, s \neq p)$$

5. 検索法

ここで拼音の類内置換誤りだけを考え、 X を入力単語の拼音と仮定する。

- 1) X の大分類と小分類の類名表記 $E1(X), E2(X)$ を作る。
- 2) 部分辞書 $D(E2_{ij})$ を、 $E1(X), E2(X)$ を見出しとして捜し出す。
- 3) X を $D(E2_{ij})$ の全ての単語と比べ、 $D(E2_{ij})$ になければ4)へ、あれば $\{D_{ij}[k] \mid D_{ij}[k] = X, 1 \leq k \leq n_{ij}\}$ を X の正しい候補単語として全部出力する。
- 4) X と $D(E2_{ij})$ の全ての単語とのレーベンシュタイン距離を計算し、集合 $\{D_{ij}[k] \mid HD(D_{ij}[k], X) \leq 1, 1 \leq k \leq n_{ij}\}$ を X の訂正候補単語として全部出力する。

6. 結び

本稿では拼音入力の類内置換誤りの訂正方法を提案した。実際この訂正法を少し変更すると韻母の脱落、挿入、類外置換、及び音素外などの誤りも訂正できる。現在は単語の出現頻度を利用した辞書構成の最適化を検討している。

謝辞: 本研究にご協力頂いた細島美智子技官に感謝致します。

参考文献

- [1] 田中, 小橋口, 島村, “綴りの置換誤りの訂正法”, 情報論, Vol. 27, No. 2, pp. 177-182 (1986).
- [2] E. Tanaka, Y. Kojima, “A High Speed String Correction Method Using a Hierarchical File” IEEE Trans. PAMI, Vol. 9, No. 6, pp. 806-815 (1987).
- [3] 王, 鈴木, “中国語音声認識の研究”, 信学会研資, SP86-83, pp. 25-32 (1987).