3Q－3

d-bigramを用いた中国語における文単位一括変換法

孫 大江, 堤 純也, 延澤 志保, 佐藤 健吾, 佐野 智久, 中西 正和
慶應義塾大学理工学研究科計算機科学専攻

1. 背景

中国語の入力に関する研究開発はまだ十分ではない。特に連単語組一括変換法は方法として実現されてはいるが、特定の文法に基づいて動作する。良い変換率を達成するためには、良い中国語文法（中国語言語学の文法ではなく、拼音漢字変換のための文法）の実現が必要である。しかしながら、より良い拼音漢字変換法の研究開発にはまだまだ課題がある[1]。

2. 中国語入力、変換方法の現状

2.1 中国語入力法

中国語入力方法は大きく二つに分類される。

- 字音（発音）による入力方法
- 字形（部首）による入力方法

表 1: 中国語入力現状 [1]

地域	発音入力	部首入力	その他
大陸	拼音入力 70%	五筆字形 20%	10%
台湾	注音入力 30%	倉頡入力 50%	20%

2.2 中国語入力法の問題点

- 発音による入力の問題点
  1. 地域によって異なる発音を区別できない
  2. 声調（即ち、四声）の問題
  3. 発音を表現する際にも曖昧性がある
- 部首による入力の問題点
  1. 漢字字形をどのように分解するか
  2. 分解したものをどのようにグループ化するか

2.3 中国語変換法

中国語入力においては入力をいかに効率的に行うかが大問題であり、発音による入力よりもキータッチが少なくて済み、同音字が少ない筆画入力、或は発音と字形の組み合わせが主流である。

変換法は漢字一字ごとの変換から単語、熟語単位の変換に進化しており、更に文単位の変換も実用段階に入っている。

3. d-birgamを用いた文単位一括変換法

3.1 方法

1. d-bigram

   d-bigramとは、2つの単語 w_1, w_2が単語間距離 d で現れる確率を統計情報として持つものである [2][3][4]。

2. 相互情報量（Mutual Information）

   2単語間の d-bigram の相互情報量：

$$MI(x_i, x_{i+d}, d) = \log_2 \frac{P(x_i, x_{i+d}, d)}{P(x_i)P(x_{i+d})} \quad (1)$$

   x_i : 入力列中の i 番目の要素
   d : 2要素間の距離
   $P(x)$: 要素 x が現れる確率
   $P(x, y, d)$: 要素 x, y が距離 d で現れる確率

   ここで、距離の概念を導入したことによって、隣接していない単語同士の関係も情報として持っているため、単語間の意味的な性質も得ることが可能である [2][3][5]。

3. 評価値の計算式の定義

   文に対する評価値は、文の中の全ての単語の組に対してのそれぞれの相互情報量の和を値とし、文に対する評価式は次のようになる。

$$I_d(W) = \sum_{d=1}^{m} \sum_{i=0}^{n-d-1} \frac{MI_w(x_i, x_{i+d}, d)}{g(d)} \quad (2)$$

   MI に対する重み付け

$$g(d) = d^2$$

A Sentence-Unit Pinyin-Hanzi Transcription Using D-bigram
Da Jiang SUN, Junya TSUTSUMI, Shiho NOBESAWA, Kengo SATO, Tomohisa SANO, Masakazu NAKANISHI
Department of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

W : 入力列
n : 入力列の要素数
m : d-bigram の最大距離

4. 実験及び検討

4.1 実験前提及び対象

- 拼音入力変換法のみを対象とする
- 変換の速さを考えない
- 変換正当率の評価を目的とする

4.2 PH コーパスを用いた実験

PH コーパスは中国新華ニュース機関 (XinHua News Agency of China) からの出版物 (1990 年 1 月から 1991 年 3 月まで) が集められたもので、約 400 万の中国の標準文字コードである GB (国家標準信息交換用漢字編碼字符集) コードから構成されるものである [6]。[1]

実際に本手法の実現は枝刈り探索と全探索二つ方法で実験してきたが、表 2 と表 3 のような結果を得られた。但し、対照辞書中で拼音で表さない lü と nü

表 2: 実験結果 (枝刈り探索)

コーパス : PH コーパス
辞書 : 6763 字 (自作した拼音
: と常用漢字の対照辞書)
入力 : 拼音のみ (声調を入力しない)
実験数 : 約 100 文
平均入力単語数/文 : 約 9 語

	1 位	～ 5 位まで	曖昧性
α	59.3 %	67.7 %	22.2 %
β	50.0 %	72.2 %	16.7 %

α : コーパス中に存在する文
β : コーパス中に存在しない文

表 3: 実験結果 (全探索)

	1 位	～ 5 位まで
α	81.5 %	88.9 %
β	67.7 %	88.9 %

は lu :と nu :と定義されている。

4.3 検討

正解文が 5 位までに含まれない理由は以下のように考えられ、その結果表 4 で示す。

(a) 表 2 示したように枝刈り探索で実験した方が、入力したい漢字がコーパス中に出現確率が低い場合、枝刈り対象となってしまった。
(b) 中国語は英語のように S+V+O という語順が強いので、使役或は命令形などに対して弱い。これはコーパスにそのような文が少ないことも関係ある。
(c) その他 (部分的に正しい或は原因が明確でない)。

表 4: 検討結果

	(a)	(b)	(c)
α	60.0 %	20.0 %	20.0 %
β	0.0 %	16.7 %	83.3 %

5. 結論

従来の中国語における拼音漢字変換法の課題に対して、本論文では d-bigram を用いて、文法的な情報を一切使わずに、新しい中国語入力変換方法を提案した。更に実験で文単位一括変換法について結果を検討し、本手法が有効であることを示した。

参考文献

[1] Zhong, X., and Kuribayashi, H. A Chinese Input Environment on UNIX—The Inplementation of cWnn (written in Japanese). *Proceedings of the 16th JUS UNIX Symposium*, November, 1990.

[2] 堤 純也, 新田 朋晃, 小野 孝太郎, and 延澤 志保. 統計情報を用いた多言語間機械翻訳システム. 人工知能学会研究会, pages 7–12, 1993.

[3] Nobesawa, S., Tsutsumi, J., Nitta, T., Ono, K., Sun, D. J. and Nakanishi, M. Segmeting a Japanese Sentence into Morphemes Using Statistical Information between Words. *Coling*, 1994.

[4] Sun, D. J., Tsutsumi, J., Nitta, T., Ono, K., Nobesawa, S. and Nakanishi, M. An intelligent Chinese input system using statistical information between words. *Qualico*, pages 102–107, 1994.

[5] Tsutsumi, J., Nitta, T., Ono, K., Nobesawa, S. and Nakanishi, M. Multi-lingual Machine Translation Based on Statistical Information. *Qualico*, pages 147–152, 1994.

[6] Guo, J., and Liu, H. C. PH—a Chinese corpus for pinyin-hanzi transcription. *ISS Technical Report TR93-112-0*, 1992.

[1] Prof. GuoJin at the Institute of Systems Science, National University of Singapore.