統計情報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, 佐野 智久, 佐藤 健吾, 大森 久美子, 中西 正和
慶應義塾大学大学院理工学研究科計算機科学専攻

1 はじめに

中国語の入力に関する研究開発は現状ではまだ十分ではない。発音による入力法の中で連単語組一括変換法は実現されているが、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、本研究では近年考案された d-bigram(距離付き bigram) を用いて、文法的な情報を一切使わずに、発音による入力法の中で良く使われる拼音 (Pinyin) 入力法 (四声の入力は必要としない) に基づいた文単位一括変換法を提案する。更にこれについての実験結果を検討し、本手法が有効であることを示す。

2 中国語入力変換方法の概説

2.1 中国語入力法

現在、中国語入力方法に関しては多数の方法があるが、通常大きく二つに分類される。

2.1.1 字音 (発音) による入力方法

発音による入力方法は各漢字の発音によって入力する方法である。中国語系の人々に対して共通の音声記号が拼音 (Pinyin) と注音の二種類があり、拼音が今現在中国とシンガポールで良く使われている。注音が拼音よりも前に使われた音声記号であり、今台湾と香港で利用されている。

2.1.2 字形 (部首) による入力方法

部首による入力方法は各漢字の構成要素によって入力する方法である。これも多数代表的な例がある (例えば、五筆字形 (Wubi) や倉頡 (Cangjie) 等)。

A Sentence-Unit Pinyin-Hanzi Conversion Using Statistical Information.
Sun Da Jiang, Junya Tsutsumi, Shiho Nobesawa, Tomohisa Sano, Kengo Sato, Kumiko Oomori and Masakazu Nakanishi.
Department of Computer Science, Keio University.

2.2 中国語入力法の問題点

中国語入力法の問題点には以下のようなものが挙げられる。

- 発音による入力の問題点

1. 地域ごとに異なる発音を区別できないこと
   拼音の例で挙げると、ある地域の人が 'zh-', 'ch-', 'sh-' と 'z-', 'c-', 's-' 各々の異なる発音を区別できないことが良くある。
2. 声調 (即ち、四声) の問題
   だれにも、いつでも正確的に四声を把握するのは難しい。
3. 拼音で発音を表現する際にも曖昧性がある
   ここでの曖昧性というのは、拼音で文書を一括で表現する際、拼音間に区切りがないと幾つかの分け方が生じることである。例えば、拼音列 "angang" と書かれた時に、"an gang" と "ang ang" の二通りの分け方があるので、曖昧性が生じる。
   しかし、本論文では一つの入力 (区切ってある拼音列、四声を入力しない) に対して、複数の正しい文 (文法的及び意味的に正しい文) が生じる場合、こういう現象を曖昧性と定義する。
   具体的な例を挙げると表 1のようになる。拼音入力列は "zhe shi shi yan" であり、二通りの変換可能性がある。

表 1: Pinyin 入力によって生じる曖昧性

入力拼音列	中国語の表現
zhe shi shi yan	这是实验 (This is a test.)
zhe shi shi yan	这是食盐 (This is salt.)

- 部首による入力の問題点

今まで、数百種類の部首入力法が開発されていたが、漢字字形をどのように分解すれば良いかまだ統一されていない。又、分解したものをどのようにグループ化するかという問題がある。

中国語入力方法の現状をまとめると、表2のようになる[7]。中国本土とシンガポールでは、拼音入力法が全ての入力方法の7割を占め、五筆字形法が2割を占め、その他が1割を占める。それに対して、台湾と香港では注音法が全ての入力方法に3割を占め、倉頡入力法が5割を占め、その他が2割を占める。

表 2: 中国語入力方法の現状

代表的な地域	発音入力	部首入力	
中国	拼音 70%	五筆 20%	その他 10%
台湾	注音 30%	倉頡 50%	その他 20%

2.3 中国語変換法

今のところ、一般の拼音漢字変換システムはほぼ文字レベルの変換方法であり、毎回に一文字のみを入力できるシステムである。最近 cWnn のような連単語組一括拼音漢字変換法も方法として実現されているが、特定の文法に基づいて動作する。発音による入力法に関して、文単位一括拼音漢字変換法をサポートしていない現状である。

3 統計情報を用いた中国語における文単位一括変換法

本論文では統計情報を用いて、文法的な情報を一切使わずに、発音による入力法の中で良く使われる拼音入力法 (四声の入力は必要としない) に基づいた文単位一括変換法を提案する。この方法は近年考案された d-bigram(距離付き bigram)[3][4][5][6] を用いて、任意の長さの拼音入力列に対して、四声を入力しなくても、文単位で一括変換できる新しい方法である。

3.1 システムについて

本システムでは、入力する際に拼音の音声記号列のみを、拼音間に区切りを入れて、文単位で入力する。そして、対照辞書によって拼音に与えられた中国語を探し出す。最後に、各拼音に相応する中国語の全ての組合せを作成する。音節列は対照辞書を利用して簡単に相応する漢字に書き換えられる。従って、これらの全ての組合せから多数の中国語文が構成される。ここで、問題点となるのは膨大な中国文の中でどれが正しい文章であるか判断することが極めて難しいことである。

このような問題点を解決するために、単語間 (中国語文字間) の統計情報 (d-bigram) を用いて各文の評価値を計算し、評価値の高い順に序列化し、正しい変換を求める。以下はこれらの簡単な流れを示している。

拼音のみを、拼音間を区切って文単位で入力

⇓ 対照辞書チェック

拼音に与えられた中国語を探し出す

⇓

各拼音に相応する中国語の全ての組合せを作成

⇓ d-bigram を用いて計算式 (2) による

単語間統計情報を用いて評価値を計算

⇓

評価値の高い順に序列化
正しい変換を求める

3.2 d-bigram を用いた中国語変換法

自然言語処理 (NLP) の分野では、bigram 情報と trigram 情報という概念が良く応用される。bigram 情報、trigram 情報とは各々2単語、3単語がコーパスの中に、連続して出現する確率である。

本論文では、通常の情報を用いずに、距離の概念を入れた統計モデル (d-bigram) を導入する [3] [4] [5][6]。d-bigram モデルとは、2単語が距離 d だけ離れて出現する確率のモデルである。

3.2.1 2単語間の d-bigram 相互情報量 (MI_d)

相互情報量 (MI:Mutual information) とは事情間の関係の情報量のことである [1]。これが様々な分野では広く応用され、自然言語処理の分野でも単語同士の関係を得るために用いられる。しかし、本論文では計算時に通常の相互情報量の代わりに、2単語間の d-bigram 相互情報量 (MI_d) を使う。以下の計算式 (1) は2単語間の d-bigram 相互情報量を計算する方法である [3][4][5][6]。

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

ここで、$P(x_i)$ はコーパスの中に i 番目に単語が現れる確率であり、dが2単語間の距離である。$P(x_i, x_{i+d}, d)$ はコーパスの中に単語 x_iと x_{i+d}が距離 dで現れる確率を意味する。

又、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 [3][5]。

3.2.2 文評価値の計算式の定義

ある文 Wに対する評価値 $I(W)$ は、先に与えられた距離 dまで文の中の全ての漢字の組について、その間の d-bigram を計算し、それらに距離に応じた重み付けを施したものを足し合わせることで得ている。文の評価値に対する計算式 (2) は次のようになる [3][5]。

$$I(W) = \sum_{d=1}^{d_{max}} \sum_{i=0}^{n-d-1} \frac{MI_d(x_i, x_{i+d}, d)}{g(d)} \quad (2)$$

ここで、nが入力文の中の単語の数であり、d_{max}が d-bigram の最大距離 (事前に決められたある固定値) である。$g(d)$ が MI に対する重み付けで、ここに $g(d) = d^2$とする。

4 実験及び検討

4.1 PH コーパスを用いた実験

PH コーパス[1]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 [2]。

PH コーパスを用いる目的は、節 (3) で説明した方法で中国語文字列から構成された膨大な文章の集合の中で、どれが'中国語らしい'か決定することである。即ち、拼音入力変換法のみを対象とし、変換の速さを考えず、変換正当率の評価を目的とする。表3に現在の実験環境を示す。

本システムで使われた辞書は我々が自作した、拼音と6763個常用漢字の対照辞書である。但し、普通の方法で 'ü' を入力できないため、対照辞書の中では記号 'lu:' と 'nu:' を各々の拼音 'lü' と 'nü' の音節と同じ表現と定義する。

[1]これはシンガポール (The Institute of Systems Science, National University of Singapore.) から FTP で自由に利用できるものである。

表 3: 実験環境

コーパス	PH corpus
対照辞書 (自作した拼音と漢字)	6763 常用漢字
入力文	拼音のみ 声調を入力しない
実験数	100
平均入力字数/文	10

4.2 実験結果

図 (1) と図 (2) で実験の例を挙げ、表4に我々が実験で得た結果を示す。

```
Corpus : PH Corpus (Chinese)
======= initializing =======
Distance_Max = 10
MI_Weight : square
Window = 0
Window_Weight : square
Border = -10.000000
Input Pinyin Sequence: zhe shi shi yan
============================
( 1)  @@这是实验。    : Point = 17.9095272
( 2)  @@这是食盐。    : Point = 14.0365443
( 3)  @@这实实验。    : Point = 13.4332374
( 4)  @@这事实验。    : Point = 13.4070197
( 5)  @@这是视眼。    : Point = 12.9505929
( 6)  @@这是事演。    : Point = 12.5668581
( 7)  @@这是世言。    : Point = 12.4253549
( 8)  @@这时事演。    : Point = 11.5987054
( 9)  @@这时实验。    : Point = 11.0857440
(10)  @@这使实验。    : Point = 10.7470051
```

図 1: 実験の例

表 4: 実験結果

	1位	～ 5位まで
α	82%	89%
β	68%	89%

αは入力文の全てが PH コーパス中に存在する文

```
Corpus : PH Corpus (Chinese)
====== initializing ======
Distance_Max = 10
MI_Weight : square
Window = 0
Window_Weight : square
Border = 0.000000
Input Pinyin Sequence: han zi shi zhong guo li shi de zai ti
=======================
( 1)  @@汉字是中国历史的载体。  : Point = 28.9007359
( 2)  @@汉子是中国历史的载体。  : Point = 26.2998249
( 3)  @@汉字是中国历史地再提。  : Point = 25.0766495
( 4)  @@汉字是中国历史的在题。  : Point = 24.5871161
( 5)  @@汉字是中国历史地在题。  : Point = 24.4303050
( 6)  @@汉字是中国历史的再提。  : Point = 24.3359287
( 7)  @@汉子世中国历史的载体。  : Point = 23.3773029
( 8)  @@汉字市中国历史的载体。  : Point = 23.2189987
( 9)  @@汉字是中国历史地在梯。  : Point = 22.8136240
(10)  @@汉子是中国历史地再提。  : Point = 22.2276424
```

図 2: 実験の例

の時の変換正当率であり、βは入力文が PH コーパス中に存在しない文の時の変換正当率である。明らかに、入力文が PH コーパス中に存在しない時にはベスト結果の変換正当率を下げる。しかし、α、βとも 5 位までの変換正当率が約 9 割であるとの良い結果を得られた。

又、αの場合には入力文の 22 パーセントは節 (2.2) で説明した曖昧性があり、βの場合には 17 パーセントが曖昧である。

実験の結果によって、表 4で示されたような 89 パーセントの変換正当率で、入力文がコーパスの中に存在するかどうか全く関係なく、正しい中国語に変換でき、統計情報を用いた文単位一括変換の実験システムが相当に良い結果を得られたと言える。

5 結論

現在、中国語における入力変換法については、一般の拼音漢字変換システムはほぼ文字レベルの変換方法であり、最近連単語組一括拼音漢字変換法も実現されているが、特定の文法に基づいて動作する。

以上のような課題に対して、本論文では統計情報 (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。この方法は d-bigram(距離付き bigram) を用いて、任意の長さの拼音入力列に対して、四声を入力しなくても、文単位で一括変換できる新しい方法である。

更に文単位一括変換法についての実験結果を検討し、本手法が有効であることを示した。

参考文献

[1] P.Brown, J.Cocke, S.Della Pietra, V.Della Pietra, F.Jelinek, R.Mercer, and P.Roossin. A Statistical Approach to Language Translation. *Proc. of COLING-88*, pages 71–76,1989

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

[3] Tsutsumi, J., Nitta, T., Ono, K., Nobesawa, S. A Multi-Lingual Translation System Based on A Statistical Model (written in Japanese). *JSAI Technical report, SIG-PPAI-9302-2*, pages 7–12,1993

[4] Nobesawa, S., Tsutsumi, J., Nitta, T., Ono, K., Sun, D. J., Nakanishi, M. Segmenting a Japanese Sentence into Morphemes Using Statistical Information between Words. *Coling-94*, pages 227–233,1994

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

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

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