5G－1

日本文音声出力のための数詞標準形変換のルール化

高橋博之　宮崎正弘

新潟大学大学院工学研究科

1　はじめに

文字音韻変換処理は一般の単語については辞書を用いて行われる。しかし、数詞は表記が無数に生成できるため、全表記を辞書に収録することはできない。そのため、数詞は別処理で音韻付加する必要がある。

日本語における数詞の表記は、算用数字表記、漢数字表記に加え両者の混ぜ書きも行われるなど多様である。本稿ではこれら多様な表記の数詞を処理するための可読性、拡張性に富むルールを提案し、その有効性を示す。

2　型分類と標準表記

数詞はその読み方の特徴から七つの型に分類できる。整数型、小数型、分数型、「十五、六」などおよその数を表す概数型「十二～十五」など数の範囲を表す範囲型、「十三、十四」など複数の数の並びを表す並記型、および電話番号などを表す棒読み型の七つである。これを表1に示す[1]。

日本語の数表記では漢数字表記と算用数字の二種類の表記が行われ、それらの混ぜ書きも行われるなど多様である。そのため、例えば「13000」「13,000」「一万三〇〇〇」「一万三千」のように、同じ数についていくつもの表記法がある。そこでこれら表記のゆれを吸収するため、標準的な表記法に変換することとした。標準表記は音韻付加の便宜上その読み方を反映したものが望ましい。そこで実際の読み方に近い漢数字表記を基本にした。

表1: 数表記の分類

No.	型名	表記例	標準表記
1	整数型	一万二千三百四十 一万二三四〇 12340 12,340	一万二千三百四十
2	小数型	13.57 十三・五七	十三・五七
3	分数型	$\frac{5}{13}$ 5/13	五／十三
4	概数型	二、三〇万 二、三十万 二三十万	二、三十万
		12,3 十二、三 十二三	十二、三
		一億二、三千万 一億二三千万	一億二、三千万
5	範囲型	52-55 52～55 五十二～五十五	五十二～五十五
6	並記型	12,13,14 十二、十三、十四	十二、十三、十四
7	棒読み型	233-2501	二三三－二五〇一

3　分類・変換ルール

数詞を前述の七つの型に分類する処理は、手続きで書くと繁雑なものとなってしまい、可読性・拡張性に乏しい。そこで、宣言型のルールを用いて記述することにした。

各型の数表現のパターンは正規文法のクラスで記述可能である。そこで、正規文法と等価な有限オートマトンを用いて、数詞のパターンを記述することにした。オートマトンは、記述しやすさを考慮して、ε規則ありの、非決定性オートマトンとした。

標準表記への変換はそのパターンに依存した処

A Standardization Rule of Japanese Numeral for Japanese Text-to-Speech System
Hiroyuki Takahashi, Masahiro Miyazaki
Niigata University

理となる。そこで、この変換操作をオートマトンに埋めこむことにした。具体的にはオートマトンの枝（状態遷移）に手続きを付加し、状態遷移時に実行させることにした。手続きには状態遷移時にその遷移で読み込まれた文字が入力として与えられる。手続きは入力を変換して出力し、その出力の並びが標準表記となる。

手続きは以下の５種類の命令の組み合わせで記述される。

ＮＯＰ 何もしない。

ＯＵＴ 入力文字をそのまま出力する。

［文字］ 文字を出力する。

ＰＵＳＨ／ＰＯＰ この二つの命令は組みで用いられ、算用数字表記をを漢数字表記に変換するために用いられる。PUSH は入力文字をスタックに積む。POP はスタックに積まれた文字列 (算用数字表記) を漢数字表記に変換して出力する。

オートマトンの例を図 1に示す。これは漢数字表記 (「100 万」「一万二千」等) を受理する。

4 評価実験

この規則化の有効性を実証するため評価実験を行った。評価には、新聞記事などから収集した数詞、約 1000 語を用いた。その結果、約 99%の数詞について正しい型分類と標準表記への変換が行われた。うまくいかなかった例としては、住所や記号の特殊な使い方 (「モデル 5/10/15」) 等が挙げられる。

5 おわりに

多彩な表記形態を持つ数詞の型分類、標準表記への変換処理をオートマトンを用いたルールで記述した。宣言型のルールを使用することにより、記述性、可読性が良く、文書の性格にあわせてルールセットを入れ替えるようなことも可能となった。

参考文献

[1] 宮崎 正弘: 日本文音声変換のための数詞読み規則, 情報処理学会論文誌, Vol.25, No.6, pp.1035-1043 (1984)

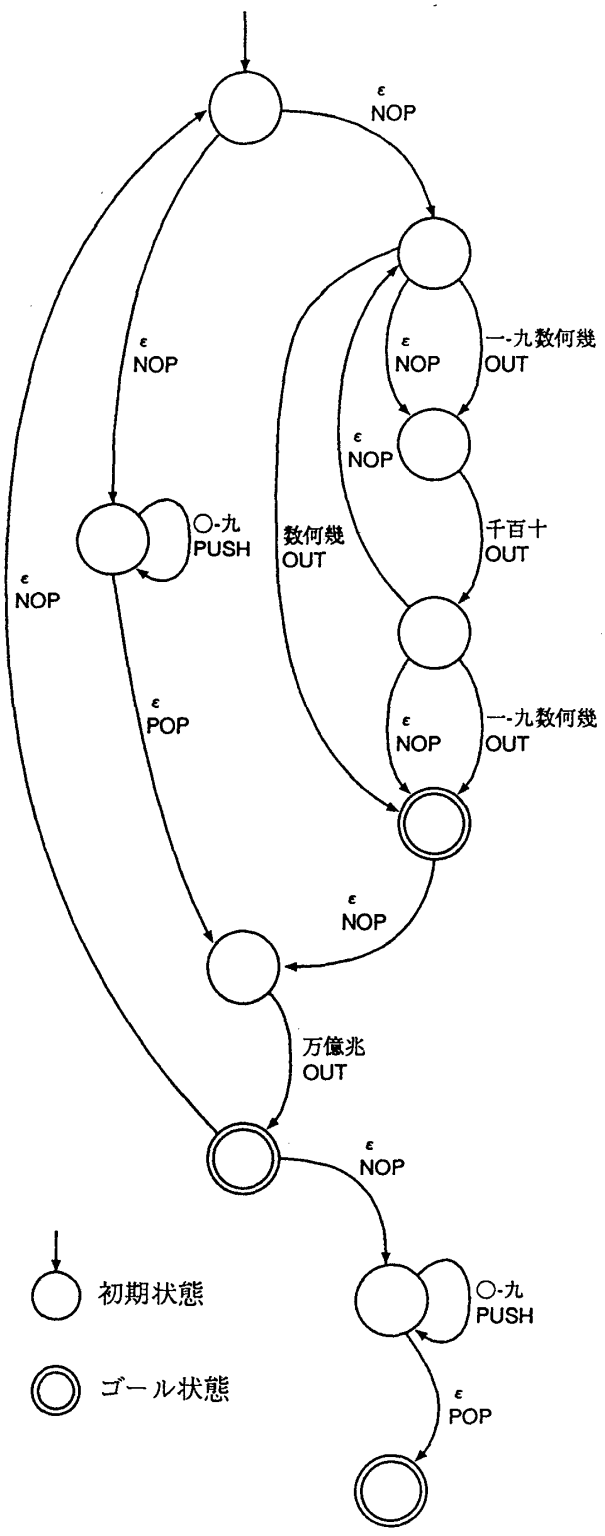

図 1: オートマトン