3G－7

正規表現を辞書項目とする形態素解析辞書の構成と利用

建石由佳、伊東伸泰
日本アイ・ビー・エム株式会社 東京基礎研究所

1　はじめに

形態素解析において、一定のパターンに従った文字列集合が文中で同一の役割を持つとみなせることがある。文字列パターンは未知語テンプレート [1] として、未知語（OCR後処理の場合の文字選択誤りを含む）の影響が周辺に及ぶことの抑止にも利用できる。このようなパターンを辞書上にコンパクトに記述できれば、辞書サイズの縮小、管理の手間の軽減に役立つ。このために、日本語形態素解析用トライ辞書 [2] を拡張し、正規表現を辞書項目とできるようにしたのでその構成と利用について報告する。

2　正規表現

パターンを記述するために正規表現を利用することは、検索コマンドなどでなじみが深い。ここでは、OCR後処理に利用することを考えて、つぎのような文法を持つ正規表現を辞書項目とすることにした。

リテラル 全角文字一個は、そのものにマッチする。

字種クラス 半角の A,N,K,H,J,C,O はそれぞれ一個の英字、数字、かたかな、ひらがな、日本文字 (漢字とかたかなとひらがな)、漢字、記号にマッチする。

範囲 [と] にかこまれた全角文字の並びはその中の文字一個にマッチする。また、[^ と] にかこまれた全角文字の並びはその中にない文字一個にマッチする。-の両側に文字を書くことによって、シフトJISコードでの範囲を指定することができる。

また、R,Sを正規表現、Rにマッチする相手を $l(R)$ として

連 RSは $l(R)$ と $l(S)$ の concatenation にマッチする。

選択 $R|S$は $l(R)$ または $l(S)$ にマッチする。

繰返し $R*$は $l(R)$ の 0 個以上の繰返し、$R+$は $l(R)$ の 1 個以上の繰返し、$R?$は $l(R)$ または空ストリングにマッチする。

グループ (R) は $l(R)$ にマッチする。

「任意の1文字」を直接示す表現は持たない。これは、OCR後処理において、任意の1文字という記述のしかたでは、候補の優先順位付けに対する寄与が少いからである。また、OCRの場合、読まれた文字に全半角の区別はないので、リテラルは全角文字のみで、半角文字はメタ記号専用とすることによって、メタ記号と一致する文字がリテラルにあらわれるのを避けている。

3　トライの構造

R、Sを正規表現、Q、Tを正規表現または空ストリングとして、$Q(R|S)T$を登録することは、QRTと QST を登録することと同等であるので、これは自然にトライに組込むことができる。繰返しは、繰返される正規表現と、繰返しを示すマーカーの組を1つの単位と扱う。従って、Rと $R+$とは、トライ上で別のノードとみなす。また、字種クラスはそれぞれ異る（リテラルとも異る）文字のようにみなしてトライ上に登録する。範囲も、同じ範囲を示すものごとに異る文字のようにみなしてトライ上に登録する。例えば、

```
大宮市
大和市
C+町
C+町([1-9]|[一二三四五六七八九])丁目
[1-9]N*号
[1-9]N+―N+
A(A|N)+@A(A|N)+.(JP|jp)
```

を辞書項目とするトライは図1のようになる。

ここで、繰返し、字種クラス、範囲をあらわす正規表現が存在すると、探索の際に、それらのそれぞれの可能性を並列に調べる必要が生じる。例えば、図1において入力の1文字目が漢字であったとき、C（漢字を示す字種クラス）とマッチしても、普通のトライの場合と異り、入力を次に進めることはできず、リテラルの「大」にマッチする可能性を調べなければならない。これは、トライの本質ともいえる決定性をこわすことである。しかし、定型パターンを登録する場合でも、未知語テンプレートとして正規表現を使用する場合でも、少ない種類で記述できることが正規表現を利用する本質である。

A Dictionary for Japanese Morphological Analysis with Regular Expressions
Yuka Tateisi, Nobuyasu Itoh
IBM Research, Tokyo Research Laboratory

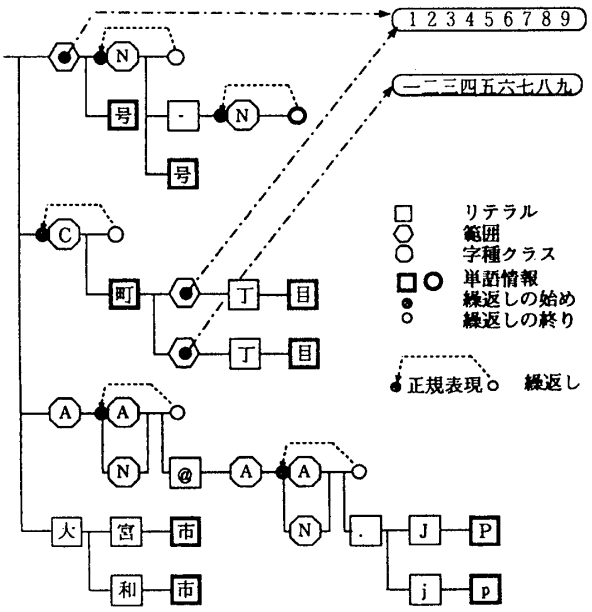

図 1: 辞書の構成

すなわち、正規表現は、日本語の辞書中の特別な場合を記述するために用いるので、辞書エントリーのほとんどはリテラル・ストリングであり、メタ記号の入ったものは少ないと考えられる。従って、ここで部分的に非決定性が発生しても、探索効率に対する影響は少いと判断した。

4 応用例

OCR 後処理に正規表現が有用な例として FORM 類の例をあげる。また、OCR 処理とは直接関係しないが、正規表現が有用な例として文書論理構造ラベルの解釈への応用をあげる。

4.1 FORM 類

FORM 類には数字と特定の記号からなる不定長のパターンで表現できるフィールド（金額、電話番号など）が多い。例えば、

```
¥([1-9]|[1-9]N|[1-9]NN)(,NNN)*(.N*)?
```

は、(英米式記数法による) 金額をしめすパターンである。

また、住所など、特定の字を含むことはわかっている文字列を解析する場合、一部未知語があっても全体の解析への影響を小さくすることができる。例えば、

```
横浜市CC?区C+[町台]NN?—N+
```

は横浜市内の住所（の一部）を示すパターンであるが、OCR において、「区」「町」などの文字さえ確実に読む（候補に入る）ことができれば、他の文字の読み誤りによって全体の解析に失敗する、という事態は避けることができる。

4.2 文書論理構造ラベルの解釈

章、節、リストの始めを示すキーワードを基に、文書に論理構造タグを付与することができる（[3] など）が、キーワードを正規表現で記述し辞書化することで、多様な書式の文書に柔軟に対応できる [4]。例えば、

```
第N+章
(N+. )+
```

などを辞書項目とし、品詞体系をヘッダー・リスト・段落の始めなどとすることにより、文書論理構造解釈用の辞書ができる。

5 まとめ

正規表現を辞書項目として登録できるトライ辞書を作成した。これは、定型文字列パターンや未知語テンプレートを容易に記述し、これらを一般の単語と同様に扱った形態素解析を可能にする。

参考文献

[1] 高尾、西野: 「日本語文書リーダ後処理の実現と評価」、情報処理学会論文誌、Vol. 30, No. 11, pp. 1394-1402 (1988).

[2] 伊東、丸山: 「OCR 入力された日本語文の誤り検出と自動訂正」、情報処理学会論文誌、Vol. 33, No. 5, pp. 664-670 (1992).

[3] 土井、福井、山口、竹林、岩井: 「文書構造抽出技法の開発」、電子通信学会論文誌、Vol. J76-D-II, No. 9, pp.2042-2052 (1993).

[4] Tateisi, Y., Itoh, N.: "Using Stochastic Syntax Analysis for Extracting a Logical Structure from a Document Image," *To appear in Proc. 12th Int'l Conf. on Pattern Recognition*, IEEE Computer Society Press, (1994).