7B-2

日本語文書リーダ後処理における未登録語処理

西野文人　　高尾哲康

㈱富士通研究所

1. はじめに

日本語文書リーダの認識率を向上させるためには言語解析を行う後処理が不可欠である．この後処理として，単語照合検査と単語間接続検査により，言語的制約を検査する方法[1]を採用している（図1）．

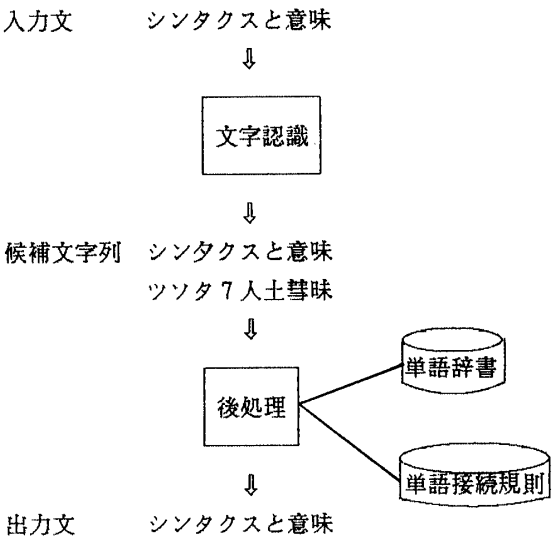

図1　日本語文書リーダ処理の流れ

単語照合検査を行うには単語辞書が必要であるが，単語辞書の語数をどんなに多くしても未登録語はなくならない．実用的な文書リーダシステムを考えたとき，未登録語の処理は避けて通れない問題である．本稿では日本語文書リーダ後処理における未登録語処理の方法について述べる．

2. 日本語文書リーダ後処理における未登録語処理

日本語文書リーダ後処理を機械翻訳などで使用されている一般の形態素解析と比べると以下のような点で異なる．

(1) 一般の形態素解析では，入力文は正しいものと仮定できる．これに対して，文書リーダ後処理では，第1位候補として与えられた文字が常に正しいとは限らない．したがって，第1位候補の組み合わせの綴りが辞書になかったからといって未登録語であるとは限らない．正解の文字が第2位以降に存在するかもしれないからである．

(2) 機械翻訳や自然言語による問い合わせシステムなどでは，ユーザがある目的のために自然言語を入力する．そこで入力される文体などの種類は限定されたもの，あるいはシステムが制限をつけたものである．これに対して文書リーダでは，既存の文書を読むわけであるから，入力文書の種類・文体などについては，あらゆるものに対処する必要がある．

(3) 正解文字を推定することに主眼があるので，未登録語として認定した後のその語の意味属性についてはあまり重要でない．したがって，複合語についてはそんなに考慮しなくてもよい．

2. 未登録語処理の方法

日本語文書リーダの未登録語処理は2つのステップに分かれる．(1)どの部分が未登録語であるかを認定する単語範囲認定部と，(2)その単語のつづりを認定する文字認定部である．

2.1 単語範囲の認定

未登録語の範囲を決定する方法としては，形態素解析が失敗したら，その前後を解析し直して未登録語の範囲を決定する方法が提案されている[2]．しかし，この方法では未登録語処理用の特殊な処理が必要である．また，各文字に対して候補文字が複数ある文書リーダ後処理では，そのすべての組み合わせを調べて失敗したら未登録語であると認定する方法では以下の点で問題である．

① 時間がかかること．たとえ全く意味のなさないような文字列（あるいは非文字列）を読み込ませたとしても，何らかの処理を行って許容できる時間内に応答が戻ってこなければならない．

② 本来正解でない文字の組み合わせでたまたま辞書に存在する過った単語を認識してしまうかもしれないということ．

そこで，我々は未登録語テンプレートを利用した方法を提案する．この方法では，候補文字集合列から単語辞書を検索した時に，文字候補の組み合わせから可能な文字列としての単語に加えて，未登録語テンプレートが検索リストにつけ加えられる（この前後の詳しいアルゴリズムに関しては文献(1)を参照のこと）．テンプレートにはマッチする

Post-Processing Unregistered Words in a Japanese Document Reader

Fumihito NISHINO and Tetsuyasu TAKAO

Fujitsu Laboratories Ltd.

文字パターンと前後に接続する文字列のパターン，及び評価値が記述されている．例えばカタカナ未登録語は$\alpha\beta^{+}\gamma$というパターンをしているので，，カタカナ未登録語テンプレートとしては次のようなα（語頭），β（語中），γ（語末）を用意する．

	左接続可能性	右接続可能性	文字条件
α	文頭，格助詞，…	β	語頭カタカナ
β	α	γ	カタカナ
γ	β	文末，格助詞，…	カタカナ

各テンプレートは以下の情報を持っている．

(1)文字条件

そのテンプレートが採用されるための候補文字集合中の文字の条件を示す．例えばαでは，語頭に来ることのできるカタカナ（カタカナの集合から『ン』，『ッ』，『ャ』，…を除いたもの）が候補文字集合中にあれば採用されることを意味する．

(2)接続情報

そのテンプレートの前後の単語との接続条件を示す．

(3)評価値

そのテンプレートが使用された時の評価値を示す．

例えば，文字認識の結果として次のような候補文字集合列があたえられたとする．

シソタクスと

ツンタ７人Ｋ

図１においては，『シン』という単語に加えて未登録語テンプレートα，β，γ，…の中で接続条件の一致するαが検索単語リストに加えられる．各候補単語の評価値が計算されるが，『シン』につながる候補がないため，結局『と』に接続する『$\alpha\beta\beta\beta\gamma$』と『人』につながる『$\alpha\beta\beta\gamma$』という未登録語が残ることになる．このどちらが選ばれるかは，文字認識の評価値と文字の連接確率によって計算される評価値によって決まる．

一般の未登録語処理では未登録語の範囲決定として文字種の変わり目を利用している．ここでの方法では未登録語のパターンとその接続条件を記述するのであって，必ずしも文字種の変わり目が未登録語の範囲でなくてもよい．（実際の多くの未登録語テンプレートでは字種の変わり目を利用するのであるが，固有名詞が前に接続することの多い語の情報[3]を利用するとか，外来語では『～ション』，『～メント』のような英語に多い語構造の知識を導入するなど，より柔軟に対処ができる）．実際に未登録語テンプレートとしてどのようなものを用意するかが問題である．未登録語の語構成パターンとしてどのようなものが多いかを調べることが必要であろう[4]．

2.2 単語の文字の決定

未登録語の範囲が決定された後，各位置に対する正解文字を推定する必要がある．この推定には各文字候補の文字認識結果による評価値と文字の連接確率とを用いて計算した評価値が使われる[5]．

3. 未登録語とユーザインタフェース

日本語文書リーダにおいて重要なことは，ユーザの満足度である．認識率の向上も一つの要因である．しかし，完全なる100%の認識率は望めないので，文書読み込み後，何らかの人間の介入が必ず必要である．たとえ認識率が99%であっても，それを修正するのに全文書を読み直さなければならないとしたら大変である．評価値の低かったところ（文字認識の評価値の低い文字，未登録語と判定された単語など）を表示するようなインタフェースが重要である．

また，未登録語の登録も重要である．すなわち，システムが未登録語と認識した部分は繰り返し現れることが普通であるので，認識率の向上のためにも，そのような単語は登録しておくことが望ましい．システムで未登録語と認識した部分の単語と推定した品詞を画面上に表示し，ユーザに確認して登録できるようにすることが必要である．さらに，単語が登録されたことによって，今まで未登録語として認定していた部分の再評価が必要である．

4. おわりに

本稿では日本語文書リーダにおける未登録語処理について述べた．ことばは常に増えつづけるものであり，人間は新しいことばを学習することができる．今までの自然言語処理というと，とかく小さな世界で閉じたシステムになりがちであった．日本語文書リーダ後処理のような開かれた世界に柔軟に対応できるシステムの研究にももっと目を向けるべきであろう．

参考文献

(1) 西野，高尾：『日本語文書リーダ後処理の実現』，情処学会NL研64-8（1987）

(2) 長瀬：『ＡＴＬＡＳⅡにおける未登録語の抽出とその扱い』，情報処理学会第36回全国大会(1988)

(3) 川崎：『知的情報検索システムIRISにおける固有名詞抽出用形態素解析』情報処理学会第37回全国大会(1988)

(4) 亀田他：『未知語の分類とその処理規則』，情報処理学会第36会全国大会(1988)

(5) 高尾，西野：『日本語文書リーダ後処理におけるヒューリスティック規則について』，情報処理学会第36回全国大会(1988)