未登録語テンプレートを用いた日本語形態素解析

2F-2

西野文人
(nisino@flab.fujitsu.junet)
富士通研究所

1. はじめに

日本語の形態素解析ツールは，日本語の解析，未登録語の抽出，誤り文の指摘，文字・音声認識後処理など様々な応用に使われ，それぞれ重点となる項目は異なるが，1)精度が良いこと，2)様々な文書（様々な文体，未登録語）に適応できること，3)高速であること，4)部品として利用できること，5)辞書の拡張が容易であること，が望まれる．今回，隣接可能性(親和力)，単語頻度に基づく最良優先探索を行い，単語の文字構成パターンのテンプレートによる未登録語処理を行うことにより，精度の向上，高速化，柔軟性の向上を図った日本語の形態素解析を作成した．

2. 形態素解析手法

本形態素解析は与えられた入力文字列に対して，辞書と接続表で認められる形態素列を，単語表記の長さ情報，単語の頻度，単語間の接続親和力から計算される評価値に基づいた最良優先探索によって見つけ出すものである．辞書には各形態素について表記，接続情報，頻度が設定されている．接続情報としては前接情報（先行語との接続関係を示すもの）と後接情報（後続語との接続関係を示すもの）を有する．接続表には各形態素クラス間の接続のし易さ(親和力)を示す整数値が登録されている．

形態素クラスは，単語の連接親和力の違いを明確にするための分類であり，品詞，意味属性，字種などを考慮して設定したものである．例えば『投手』と『ピッチャー』では，品詞も意味属性も同じであるが直前の姓名の連接しやすさが異なるので形態素クラスは異なる．

3. 未登録語解析手法

未登録語の存在を決定する方法としては，接続可能な形態素候補がなくなった時点で未登録語を認定する方法[1,2]が利用されることもあるが，たまたま構文的に接続可能な形態素列が存在すると，その形態素列の各形態素の頻度が非常に小さいものであってもそれを採用したり，本来の未登録語と別の位置の文字列を未登録語と認識してしまうことがある．例えば，『ほうりべ』という未登録語を認識せずに，『ほう（ラ行五段）り（連用形）べ（名詞）』のように誤った形態素解析をしてしまう．別の未登録語解析手法として，1文字の漢字・平仮名や片仮名・アルファベット列を単語と同等に扱う方法[3,4]が報告されているが，実行速度の低下が心配されている．また，これらの未登録語解析ではどういうものを未登録語とするかの判断がプログラムに組み込まれているなど，適用分野による柔軟性に問題がある．

本形態素解析システムでは，形態素候補が存在するかどうかにかかわらず常に未登録語の可能性を考慮しながらも，実行速度の低下があまりなく，また柔軟性のある手法として，未登録語テンプレートによる未登録語処理手法を開発した．本手法では，単語辞書中に，マッチングする文字条件，接続条件，および評価値を持つ未登録語テンプレートを登録しておき，単語検索時に一般単語と同様に条件を満たす未登録語テンプレートを検索する．未登録語テンプレートは，未登録語の文字パターンにしたがって登録することができる．例えば，2文字以上からなる未登録語は，

＜語頭文字＞＜語中文字＞*＜語末文字＞

というパターン（*は0回以上の繰り返しを示す）をしているので，それぞれに対する3つの未登録語テンプレートT_h, T_m, T_tを登録し，単語としての接続情報はT_hの前接情報と T_tの後接情報とに反映すればよい．例えば，『ほうりべ』という文では，『ほう(ラ行五段動詞)－り(連用形)－べ(名詞)』という形態素解析結果も存在するが，『ほ』がT_h, 『う』と『り』がともにT_m, 『べ』がT_tに対応してできる『ほうりべ(名詞未登録語)』の評価値の方が高くなれば，未登録語と認定される．

未登録語処理には柔軟性も要求される．例えば，未登録語抽出ルーチンではなるべく多くの未登録語の可能性を抽出したい．これに対して一度未登録語抽出したあとの日本語解析（例えば機械翻訳）では，高速性が要求される．また，マニュアル文では人名はほとんどみられないが，新聞記事などでは人名は多く見られる．本方式では，テンプレートの追加・変更や文字制約の変更，評価値の変更によって柔軟に未登録語処理をすることができる．

4. 実験

登録するテンプレートを変えることによって，精度および実行速度がどう変化するかを99,000語の辞書を使って実験した．なお，辞書は前方一致圧縮(0.7Mbytes)し，単語検索は先頭1文字のインデックスによるシーケンシャルサーチである．

Japanese Morphological Analysis using Templates for Unregistered Words
Fumihito NISHINO
FUJITSU LABORATORIES LIMITED

実験には次の4つのシステムを用いた.

T0　未登録語テンプレートを用いず，形態素解析が失敗したら今までで一番深く解析した位置に未登録語があると判定する.

T1　接続条件は緩いが評価値の低い単語を構成する一連の3つの未登録語テンプレートを持つ.

T2　T1に加えて，前接条件は厳しいが評価値の高いテンプレート群を登録したもの（姓名の後には名前，括弧内には読みがくるというヒューリスティックを実現したもの）

T3　T2に加えて，後接条件は厳しいが前接続条件は緩い評価値の高いテンプレート群を登録したもの（川，町，社長などの単語の前は固有名詞が来るというヒューリスティックスを実現したもの）

4.1. 精度

文中にある未登録語を正しく認定できたかどうかは次のとおりである.

	正しく認定
T0	63%
T1	78%
T2	86%
T3	90%

T0で認識できなかったが，T1では認識できたものとしては，次のようなものがある.

(1) 接続可能な単語列が存在するが，その親和力は弱い
・たい　まつ，　お　はら　い

T1で認識できなかったが，T2では認識できたものとしては，次のようなものがある.

(2) 前接語に強い手掛かりがある
・山本　卓　真，（もの　の　ふ）

T2で認識できなかったが，T3では認識できたものとしては，次のようなものがある.

(3) 後接語に強い手掛かりがある
・部　子　川

T3でも正しく認識できなかったものには次のようなものがある.

(4) 名詞連続の方が未登録語の評価値より高かった
・池　北　線，　祝　部

(5) 助詞・助動詞と一致する平仮名が使用されたために文節と認識されたもの
・いけ　だ，　錦　のみ　旗

4.2. 実行速度

各システムの1000文あたりの速度(sun4/260)は以下のとおりである．なお未登録語を含む文は1文あたり32語で平均1.3語の未登録語を含み，未登録語を含まない文は1文あたり23語である.

	未登録語を含む	未登録語を含まない
T0	23.0秒	11.5秒
T1	30.2秒	11.5秒
T2	30.2秒	11.5秒
T3	58.3秒	24.8秒

あらゆる部分で未登録語の可能性を考慮する方式では，これまでは速度低下が心配されてきた．しかし，未登録語のテンプレートの左側の接続条件を厳しくしておくことで，多くの未登録語の可能性は，接続検査で排除されたり，あるいは評価値の低さから活性化されない．したがって，ほとんど速度低下をおこさない．その結果，約30%の文に未登録語を含む1万文の日本語形態素解析を行っても2.8分(T2)と高速化が図れた.

5. 応用

単に単語が分割されていれば良い場合もあるが，意味的にも正しい形態素を抽出して欲しい場合もある．その場合には『日米関係』を『日(day)』と『米 (rice)』とに分割されたのでは困る．もちろん，ここまで形態素解析で行うべきかどうかは議論のあるところだが，もし形態素解析で簡単に出来るならば構文解析の負荷の減少になる．本形態素解析では，未登録語のテンプレートと同様に，『平成＃年』とか『第＃位』，『※$_1$※$_2$関係(※は国を表す1字)』というようなパターンとしての単語を登録することにより，このようなパターンの形態素解析ができる．つまり，未登録語処理とパターン語処理とは同じなのである.

6. 今後の課題

本形態素解析ルーチンでは高速化と精度向上を図った．そして，未登録語テンプレートと親和力の調整によって柔軟に対応できた．しかし，これは逆に言えば，親和力の値や未登録語テンプレートの評価値をどう決めるかが難しいことを示している．現在，統計値に基づいて値を設定するが，まだ完全に自動化はできていない．形態素クラスをどう分割するか，どのような未登録語テンプレートを登録するかとともに今後の課題である.

参考文献

[1] 芦沢，平井，梶：日英機械翻訳用前編集支援システム(2)－形態素の曖昧性の検出方式－，情報処理学会全国大会第36回全国大会2U-3(1988)

[2] 長瀬：ATLAS IIにおける未登録語の抽出とその扱い，情報処理学会全国大会第36回全国大会4U-7(1988)

[3] 吉村，武内，津田，首藤：未登録語を含む日本語文の形態素解析，情報処理，Vol.30，No.3，pp294-301 (1989)

[4] 大場，元吉，井佐原，横山，石崎，板橋：未定義語を含む文の多段階構文解析手法，自然言語処理70-4(1989)