形態素タイプを用いた日本語構文解析前処理

1S−6

長瀬　友樹

富士通株式会社

1．はじめに

近年，機械翻訳の利用形態はたいへん多様になってきている．例えば，電子メールやテレックスの翻訳，外国人との会話翻訳システムなどが考えられているが，これらに共通する傾向として「即時性」がある．

「即時性」を要求されることは，翻訳前に固有名詞を始めとする未登録語を登録することができないことを意味　する．したがって，翻訳システムはこれまで以上の正確さで固有名詞などを認識し，翻訳精度を上げる必要がある．

そこで我々は，形態素解析と構文解析の間に，固有名詞等を認識するための新しいフェーズをつくり，実験を進めている．これは，字種，長さ，品詞などをもとに，形態素リストの各ノードにタイプ付けを行い，このタイプを終端記号としてＣＦＧ規則を適用するものである．記号列のまとめ，ひらがなの未登録語認識なども同時に処理できるが，ここでは固有名詞の認識に絞って説明する．

2．辞書にない固有名詞などに伴う問題点

人名や地名などは膨大な数であり，これらのすべてを辞書に登録することは不可能である．原文中に辞書にない固有名詞が含まれると，機械翻訳などでは次のような問題生じる．

(1)一般に，辞書にない語が存在すると構文解析が失敗する

「列車は倶利伽羅駅を通過した」

⇩

／列車／は／倶／利／伽／羅／駅／を／通過／し／た／

(?)　(?)

⇩

（解析失敗）

(2)たとえ固有名詞全体で辞書に存在しなくても，それらを構成する語（漢字）によって単語分割が成功してしまう場合がある．

「私は昨日愛鷹へ登った」

⇩

／私／は／昨日／愛／鷹／へ／登／っ／た／

⇩

「 I climbed the love hawk yesterday. (???) 」

3．処理の概要

(1) 従来の処理と同様に形態素解析する．

注）形態素解析の結果は辞書引き結果がリスト形式につながれたもので，１つの単語に相当するノードを「形態素ノード」と呼ぶ．形態素ノードには表記，品詞，概念記号，文法属性などが含まれている．(a)

(2) 形態素ノードの情報を基にして，それぞれのノードにタイプをつける．タイプの種類を 表１ に示す．

(3) タイプを終端記号とみなして，形態素まとめルール（ＣＦＧ規則（図２））を適用する．

注）ここで求めるべきものはゴールが成り立つ範囲（部分文字列）であって，通常の構文解析のような文全体のPARSE TREEではない．本処理にとって文全体で解析が成功する必要性は全くないことに注意して欲しい．句構造解析はボトムアップの並列型アルゴリズム（チャート法）を用いる．

(4) (3)の結果をみて形態素ノードをまとめた新たなノードを生成し，(1)のリストを更新する．(b)

(5) 従来と同様に構文解析以降の処理をする．

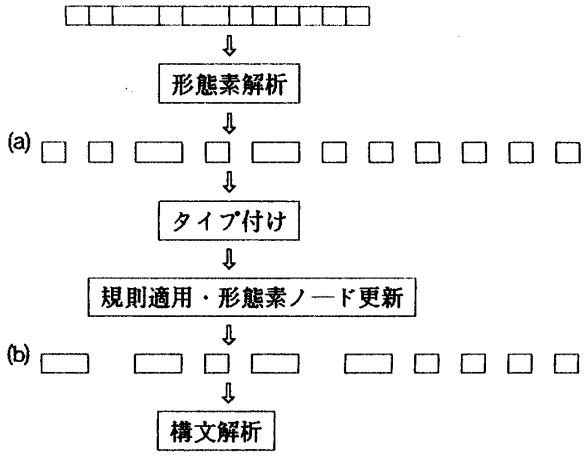


図１　処理の概要

Pre-processor for Japanese Parsing using Morpheme Type

Tomoki NAGASE

FUJITSU, Ltd.

表1　形態素タイプ（一部）とその例

タイプ名	長さ	例	記号
人名接尾語	1	君，殿，氏，様	c(n t k)
〃	>2	社長，首相，教諭，さん	s(n t)
地名接尾語	1	村，町，峠，竹，駅，前	c(p t k)
〃	>2	工場，営業所，支店，神社	s(p t)
地名接頭語	1	市，区，町，村	c(p h k)
〃	>2	備後，加賀，筑前，信濃	s(p h)
一般接尾語	1	性，的	c(t k)
一般接頭語	1	反，非，無	c(h k)
苗字	>2	山田，鈴木，山本	s(m)
名前	>2	太郎，次郎，花子	s(f)
その他漢字	1		s(k)
〃	>2		s(k)

注）記号の n p t h kはそれぞれ，人名，地名，接頭語，接尾語，漢字の属性を表す．

```
goal-n  →  l(n)   cm(k)   s(*)          ;                    (1)
       →  s(*)   cm(k)   r(n)          ;                    (2)
       →  s(*)    s(m)   c(k)   s(*)   ;                    (3)
       →  s(*)    c(k)   s(f)   s(*)   ;                    (4)

r(n)   →  s(n t)                       ;  ～首相            (5)
       →  c(h k)        s(n t)         ;  ～元首相          (6)
       →  c(n t k)      s(*)           ;  ～君…             (7)
       →  s(f)                         ;  ～太郎            (8)
l(n)   →  s(m)                         ;  山本～            (9)
       →  s(*)     c(h k)              ;  …反～             (10)

cm(k)  →  c(k)     c(k)                ;                    (11)
cm(k)  →  c(k)     cm(k)               ;                    (12)

s(k X)→  s(k)     s(k X)               ;                    (13)
```

図2　形態素まとめ規則（一部）

4．固有名詞の認識

ここでは，処理の概要に従って，固有名詞が認識されるようすを例をあげながら説明する．

〔入力文〕「奥田敬和元郵政大臣」

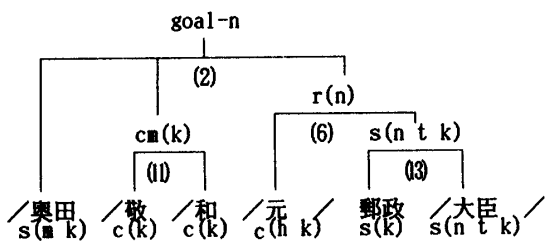
〔形態素解析結果〕

／奥田／敬／和／元／郵政／大臣／

〔タイプ付け〕

／奥田	／敬	／和	／元	／	郵政	／大臣	／
s(m k)	c(k)	c(k)	c(h k)		s(k)	s(n t k)	

〔ルール適用〕



〔固有名詞の取出し〕

実際の固有名詞は，ゴール(goal-n,goal-p,goal)からみて，左右1つずカテゴリを除いた残りである．この場合，s(m k)と r(n) を除いた残りの cm(k)の部分，すなわち，

「　敬　和　」

が，人名属性を持った固有名詞ということになる．

〔形態素リストの更新〕

／奥田／敬和／元／郵政／大臣／

5．結果

本処理を機械翻訳システム　(ATLASⅡ）に組み込んだところ，解析失敗文を減らすだけでなく，翻訳精度も向上した．例えば，「石原慎太郎がアメリカへ行った．」という文は，次のように訳が変わった．（ATLAS では，未登録語を日本語のままで訳文中に表示している．）

旧　：Ishihara modesty Taro went to United States.

新　：慎太郎 Ishihara went to United States.

固有名詞の認識精度は，人名（地名）接尾（頭）語を含むものについては90% 以上正確に認識された．まちがって認識されるのは，「後藤田正晴」の未登録固有名詞が「田正晴」になることなどであるが，実用上問題にならない．

6．おわりに

形態素解析と構文解析の間で，語種などをもとに形態素列をまとめるためのヒューリスティックを実現するフェーズについて述べた．この処理は，固有名詞の多い即時翻訳などに有効であり，未登録語のために解析が失敗する文を減らすことができた．実際には，固有名詞認識の他に，記号処理，ひらがな未登録語処理[1]も同じフェーズで実現する．

本報告と同じようなヒューリスティックを形態素解析内で実現しようという試みもなされている．[2] 逆に，構文解析内で，本来の句構造解析と同時進行的に実現することも考えられので，今後は，これらの可能性も含めて評価をしていく必要がある．

参考文献

(1) 長瀬　「ＡＴＬＡＳⅡにおける未登録語の抽出とその扱い」，情報処理学会大第36回全国大会講演論文集，1988.

(2) 西野　「未登録語テンプレートを用いた日本語形態素解析」，情報処理学会大第39回全国大会講演論文集，1989.