連語登録による形態素解析システムJUMANの精度向上

山地治　黒橋禎夫　長尾眞

京都大学工学部電気工学第二学科

1　はじめに

英語などとは違って日本語は単語単位に分かち書きされていない。したがって日本語の形態素解析では品詞を明らかにすることを、文を形態素単位に分割することと同時に行わなければならない。これを機械で行う場合、次のような点が問題となる。

1. 構文的・意味的情報を深く考慮しなければ形態素を決定できないという本質的な多義性の問題。
2. 人間には当たり前だが機械には解析が難しい形態素並びの問題。特に付属語や補助動詞など平仮名が連続している部分が問題となる。

しかし、これら2つの問題ははっきりと区別できるものではない。特に、形態素解析で一般的に用いられるコスト計算という方法では、2の問題を改善しようとしても副作用として1の問題について余計な誤りを生じてしまうという問題が生じる。

そこで本論文では、形態素解析システムJUMAN[1]を対象として、複数の形態素からなる平仮名の連なりを連語として登録することにより、副作用を生じることなく上記2の問題だけを解決する方法を提案する。さらにEDRコーパスのテキストを用いて実験を行い、この改良による解析精度の向上を評価する。

2　従来の形態素解析における問題

2.1　コスト計算による形態素解析

従来の形態素解析の基本的方針はコストの最小化であった。つまり、文をさまざまな箇所で形態素へと切り分けた結果に対して適当なコスト計算を行い、コストが最小になるような切り分け方を解析結果として出力する。

JUMANにおいてはコストとして、個々の形態素に与える形態素コストと、隣接する2つの形態素の間に与える連接コストを用いる。

2.2　コスト調整の限界

コスト計算による形態素解析では長い形態素を含む解が優先させる。そのため、次のような解析誤りが生じる。

例1　　討論｜は｜しない｜ことに(副詞)｜なった｜。
(正解)　討論｜は｜し｜ない｜こと｜に｜なった｜。
例2　　問題｜だ｜と｜はいって(動詞)｜も｜、
(正解)　問題｜だ｜と｜は｜いって｜も｜、

従来は形態素コスト・連接コストを調整することによってこういった解析誤りを改善してきた。ところが、副詞「ことに」や動詞「はいる」は出現頻度の低い形態素ではないのでこれらのコストを上げると副作用を生じる恐れが強い。また連接コストは隣接する2形態素間の関係にしか反映できないので、たとえば例2について考えると「と｜はいる」、あるいは「はいって｜も」の連接に大きなコストを与えるか禁止するかしかできない。しかし「子供とはいっていく」、「はいっても何もない」などの例もあるので、このような処理も副作用を生じてしまう。したがってこれらの問題は形態素コストや連接コストの調整では解決できない。

3　連語を用いた形態素解析

3.1　連語登録の必要性と有効性

前節で述べたような解析誤りの多くは付属語や補助動詞などの連続部分である。このような連続部分は大抵ひとまとまりで意味をなすので、文法的には複数の

```
(連語    ;「といってよい」
 ((助詞 (引用助詞 ((読み と)(見出し語 と))))
  (動詞 ((読み いう)(見出し語 いう)
     (活用型 子音動詞ワ行)(活用形 タ系連用テ形)))
  (形容詞 ((読み よい)(見出し語 よい)
     (活用型 イ形容詞アウオ段)(活用形 *)))
 )
 0.5     ;連語コスト
)
```

図 1: 連語辞書の記述例

形態素だが1つの形態素であるかのように扱っても問題ないと考えられる。そこで、前節のような解析誤りを解決する方法として意味のまとまりを成す形態素並びを連語として一括登録して扱うという方法が考えられる。このようにすれば連語内の形態素の切り方を誤ることはなくなる。また、一つの意味まとまりを成すもの、すなわちその前後の形態素とはある程度独立に判断できるものだけを連語として登録することにすれば、このような連語の登録が他の解析に副作用を与えることもほとんどないと考えられる。

例えば前節で挙げた例文の場合、「こと｜に｜なる」、「と｜は｜いって｜も」などを連語として登録しておき、これらが解析の時に優先されるようにすれば、解析に失敗することはなくなる。

3.2 連語を用いた解析の方法

ここでは実際に連語を用いたJUMANの形態素解析の手法について説明する。登録すべき連語は連語辞書に登録する。連語辞書に記述すべき情報は、連語を構成する各形態素の情報と連語コストである(図1)。各形態素の情報は形態素辞書と同じ記述でよいが、中に活用する形態素が含まれている場合にはその活用形も指定しておく。ただし、連語を構成する形態素のうち最後のものには活用形を記述しなくてもよい。記述しなかった場合はその連語が活用できることを表す。

連語に対しての連接規則としては、左連接規則(その連語と1つ前に現われる形態素との間の連接規則)には連語を構成する先頭の形態素の連接規則を、右連接規則(その連語と1つ後に現われる形態素との間の連接規則)には、末尾の形態素の連接規則をそのまま利用する。これによって、従来の形態素に対して記述した連接規則を連語にもそのまま適用できるようになる。

ところが、場合によっては形態素が連語として現われたときにはその中の先頭、または末尾の形態素とは異なる、特別な連接規則を記述したいことがある。連語登録の副作用をなるべく防ぐために、解析誤りを生じているような最低限の連接に対してのみ連語を適用したい、といった場合である。そのような場合には、連接規則辞書に連語の連接規則を個別に記述しておき、解析のときには連語の連接規則を優先する。

例えば「〜してあまりある」という表現について、「あまり(名詞)｜ある」を連語登録する場合を考える。すると、「そんなことがあまりあるとは思えない」のように「あまり」が副詞として現われたときにも名詞と解析してしまう。これを防ぐために「あまり｜ある」の左連接規則として活用形がタ系連用テ形(〜して)のものだけを許すようにすれば、「〜してあまりある」という文以外では連語が適用されなくなるわけである。

連語部分のコストは、従来のJUMANと同様に形態素コストと連接コストの総和として計算する。ただし、連語として登録されたものが優先されて解析結果に出てくるように、連語内の形態素コストと連語コストはα倍($0 < \alpha < 1$)する(図2)。このαの値(連語コスト)は各形態素の情報とともに連語辞書で指定する。今のところ標準的な連語コストは0.5にしているので絶対的に連語が優先されるようになっているが、連語コストを微調整することによって連語とは別の解析結果と競合させることも可能である。

4 実験と考察

連語登録によるJUMANの解析精度の変化を調べるために評価実験を行った。

この実験では、連語抽出用テキストおよび評価用テキストとしてEDR電子化辞書[2]のEDR日本語コーパス(以下EDRコーパスと略す)を用いた。EDRコーパスの文書データには形態素情報が付加されているので、これを正しい形態素解析結果として利用できるからである。

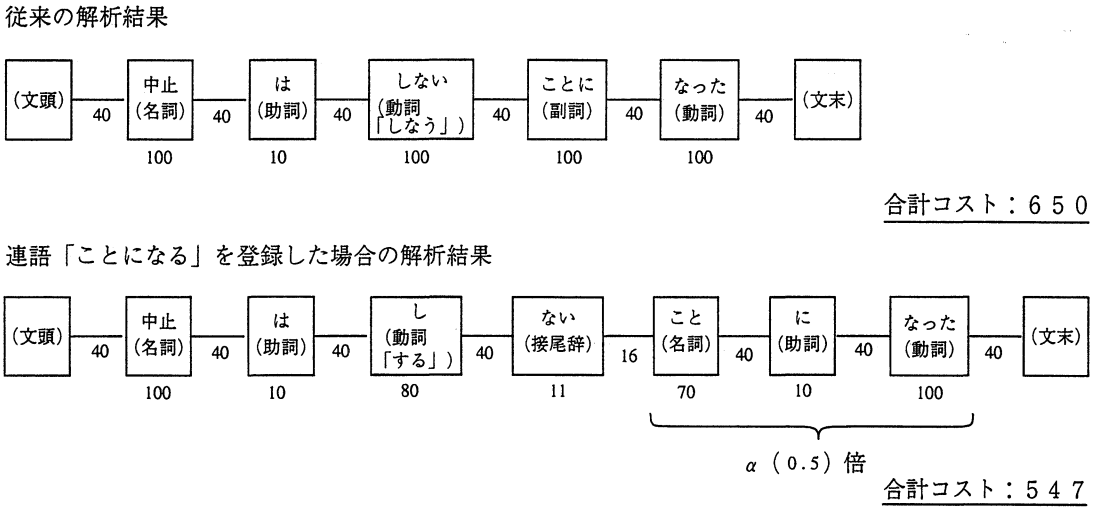

図 2: 連語のコスト計算

表 1: 登録した連語の例

こと (名詞) ｜に (助詞) ｜なる (動詞)
し (動詞) ｜かね (接尾辞) ｜ない (接尾辞)
と (助詞) ｜し (動詞) ｜ない (接尾辞)
な (形容詞) ｜さ (接尾辞) ｜そうだ (接尾辞)
と (助詞) ｜いって (動詞) ｜いい (形容詞)

4.1 連語の抽出

まず登録すべき連語を抽出するために EDR コーパス中の 25,000 文を JUMAN で解析し、その結果をコーパスの形態素情報と比較して従来の JUMAN の誤り箇所を洗い出した。

取り出された誤り箇所のうち、形態素コスト・連接コストの調整や辞書の登録語の追加で解決できるものは人手で個別に対処した。

以上のことを行っても残った誤りを調べ、そのなかから意味的まとまりをもつ形態素並び 139 個を連語として登録した。登録した連語の例を表 1に示す。

4.2 連語登録の評価実験

連語登録の有効性を調べるため、連語抽出に用いた 25,000 文とは別の 25,000 文を用いて実験を行った。そこでは 139 個の連語辞書を使用する場合と使用しない場合についてそれぞれ解析を行った。そして、解析結果が EDR コーパスの形態素情報と異なる部分を自動的に取り出し、それらが解析誤りであるかどうかを人手で判定した。

この結果を表 2に示す。表 2で「多品詞形態素」とは、複数の品詞解釈が可能な形態素をさす (例えば「極めて」には動詞と副詞の解釈がある)。このような問題に対して従来の JUMAN は、その前後の形態素を見る優先規則によって一意の解を出力していた (「助詞の後では動詞を優先する」などの規則)。しかし、本研究ではこのような優先規則は排除し、複数の品詞の解釈が可能である場合はそれらをすべて出力するようにした。「はじめに」で述べたように、連語登録によって本質的多義性の問題とそれ以外の (人間には当たり前の) 問題を区別できるようになったので、本質的な多義性は構文解析にゆだねるという立場で形態素解析での決定を保留したのである。

本質的多義性以外の問題に対する連語登録の効果は表 2に示した通りかなり大きい。すなわち、連語登録によって解析誤りの 3 分の 1 近くが改善された。また、改善形態素数 980 に対して副作用は 18 と非常に少ない。これらのことから連語登録の効果が十分に認められたといえる。

表 2: 連語登録による JUMAN の解析結果の変化

	連語登録前	連語登録後
総形態素数	526169	526169
多品詞形態素数	12219 (2.322%)	11695 (2.223%)
誤り形態素数	2621 (0.498%)	1659 (0.315%)
連語登録による改善		− 980 (0.186%)
連語登録による副作用		+ 18 (0.003%)
解析精度	99.50%	99.68%

$$\text{解析精度} = \frac{(\text{総形態素数}) - (\text{全誤り形態素数})}{(\text{総形態素数})}$$

表 3: 連語登録後の JUMAN の解析誤りの内訳

誤りの種類	誤り形態素数	割合
従来の方法で解決可能な誤り		
形態素コストによる誤り	244	14.71%
未登録語による誤り	561	33.82%
文法規則の不備による誤り	173	10.43%
話言葉・古語における誤り	113	6.81%
従来の方法では解決が難しい誤り		
連語登録すべきと思われる誤り	270	16.27%
連語登録による副作用	18	1.08%
その他の解決が難しい誤り	280	16.88%

4.3 解析誤りの分析と対処の方針

表3は、連語登録を行った後の解析誤りを分類したものである。

139 個の連語を登録した後でも、解析誤りの 16%(270 個) がさらに連語登録すべきと思われるものであった。そして、これらは異なりとしては 98 個に連語によってカバーされるものであった。これらのことから連語登録の継続的作業が有効であること、その作業はある程度収束していくと予想されることがわかる。

また、従来のような形態素コスト・連接コストの調整で解決できるような問題も依然として数多く残っている。特に全体の 15% 近くを占める「形態素コストによる誤り」の多くは、平仮名を含む見出し語の中で出現頻度が低いものが原因となっている。これらはコスト調整によってある程度解消できると思われる。

さらに、「未登録語による誤り」が全体の 35% 以上もある。これを解決するには新たに辞書登録をすればよいが、出現頻度の低い形態素の登録は副作用の原因ともなるので注意が必要である。他にも形態素辞書に本来とは異なる品詞で登録されているために生じる誤りもある。「アクセス｜で｜きる」という表現は、本来サ変名詞である「アクセス」が誤って普通名詞として登録されているための解析誤りとなっている。このような問題は表3では未登録語による誤りに含めている。

なお、「その他の解決が難しい誤り」の多くは副詞・連体詞に関係する問題である。これは、例えば複数の形態素から成る副詞「目に見えて」が登録されている場合に「失敗｜が｜目に見えて｜いる」と解析されるような問題のことである。この文では「目に見えて」は副詞的には用いられていないので、正しくは「目｜に｜見えて」と解析されるべきである。

5 おわりに

本研究では日本語形態素解析システム JUMAN の精度向上のために、従来からなされてきた形態素コスト・連接コストの調整といった方法に加えて新たに連語を登録するという手法を導入した。EDR コーパスをサンプルに実験を行った結果、JUMAN の解析精度を 99.50% から 99.68% に向上させることに成功した。誤り全体の 3 分の 1 近くを連語登録によって解消できた計算になる。

参考文献

[1] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾眞, 日本語形態素解析システム JUMAN 使用説明書 version2.0, (京都大学工学部 長尾研究室 , 奈良先端科学技術大学院大学松本研究室, 1994).

[2] EDR 電子化辞書仕様説明書, (日本電子化辞書研究所, 1993).