最長一致法に基づく3種のアルゴリズムを融合した形態素解析

池戸 俊之　兵藤 安昭　奥村 将司　栄留 孝行　池田 尚志
岐阜大学 工学部

1　はじめに

形態素解析における曖昧さは、最終的には構文や意味を考慮しなければ解消できない。そのため、形態素解析のみで解を一つに絞ろうとするとすると、正解の可能性を逃してしまうことがある。

本論文で述べる形態素解析システムは、唯一の解を求めることには主眼を置かず、複数の単純なパス選択アルゴリズムを用いて、正解を含んだ少数の解集合を得ることを目的としている。複数のパス選択アルゴリズムとは、単純な前方・後方最長一致法と、後方最長一致法に補正を加えた方法の3種である。新聞記事500文を対象とした解析実験を行った結果、高い精度(98.6%)で正解を含んだ少数の解集合(平均1.7個)を得られることが確認できた。

2　システムの概要

本システムでは、CYK表と同様の解析三角表を用いた。以下に、本システムの解析の流れを示す。

1. 字種切りを伴った辞書検索
2. 形態素間の接続判定
3. 3種のアルゴリズムによるパスの選択
4. 3種の解の合成

未登録語については、字種切り法を用いて対処した。すなわち、ひらがなを含む語は全て辞書に登録してあるものとし、その他の字種が連続する語については未登録語があると仮定して処理を行った。具体的には、辞書検索時にひらがな以外の同一字種(漢字、片仮名、ローマ字)の連続部分に対して、名詞またはサ変名詞のラベル付けを行い、その後の処理で自立語として扱う。

解析用の自立語辞書は、EDR日本電子化辞書研究所の「日本語単語辞書 Ver. 1.0」をベースに、ひらがな以外の同一字種からなる名詞、サ変名詞を削除し、見出し語数を25万語から17万語に縮小したものを用いた。これは、このような単語は字種切り法により切り出されるため、辞書に登録しなくても解析が行えるからである。また機能語については、単純な機能語に加えて複合的な機能語を構築し、約1800語を登録してある。

3　3種のパス選択アルゴリズム

パスの選択においては、前方・後方最長一致法、および後方最長一致法に補正を加えた方法の3種類のアルゴリズムを用いて解を求める。

前方・後方の最長一致法は、それぞれ文頭・文末から見て最も長い文節を探索するだけの単純な手法であり、以下に示すような文で解析に失敗する。

- **字種連続部分に含まれる単語**
  「今度旅行しよう」という文で、漢字列の連続部分である「今度旅行」(サ変名詞)を切り出してしまうため、「今度」が得られない。
- **前方・後方最長一致の間にある解**
  「いつもより」という文に対して、

  (いつ) (もより)　　... 後方最長一致
  (いつも　よ) (*り*)　... 前方最長一致

  という2解[1]が得られるが、(いつも) (より)という解が得られない。

後方最長一致の補正は、このような文に対しても正解が得られるよう、ヒューリスティックな条件を加えた処理である。

[1]「もより」は「最寄り」の異表記。「*り*」は平仮名の未登録語。

例文：現在一部門だけで行われている調査を、全部門で行うことになった

- **後方最長一致による解**
  ((現在一部門 だけ で) (行われ ている) (調査 を) (、) (全部門 で) (行 う) (こと) (にな った))
- **前方最長一致による解**
  ((現在一部門 だけ で) (行われ ている) (調査 を) (、) (全部門 で) (行 う) (こと に) (な った))
- **後方最長一致+補正による解**
  ((現在) (一部門 だけ で) (行われ ている) (調査 を) (、) (全部) (門 で) (行 う) (こと に) (な った))

3種の解を組み合わせる

⇓

({ (現在一部門 だけ で) / (現在) (一部門 だけ で) } (行われ ている) (調査 を) (、) { (全部門 で) / (全部) (門 で) } (行 う) { (こと) (になった) / (こと に) (なった) })

図 2: 3種の解の合成

以下に、**図1**の2文を例にした補正の処理を示す。

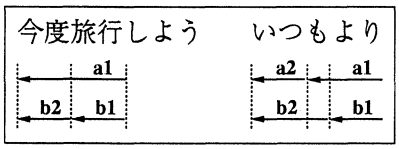


図 1: 後方最長一致の補正

始めに準備として、後方最長一致解 (a1,a2) と、それを後ろにずらした解 (b1,b2) を探しておき、以下の2つの処理をこの順序で行う。

[字種連続の途中に区切りを入れる処理]

・「今度旅行しよう」

「ずらした解の前部 (b2) が辞書に存在する1文字でない単語」で、「後部 (b1) が漢字文節である」時、ずらした解を正解とする。

[文節区切りの場所を変更する処理]

・「いつもより」

「ずらした解の前部 (b2) が辞書に存在する1文字でない単語」であるか、「前部 (b2) が1個以上の機能語を含む漢字文節で、後方最長一致解の前部 (a2) が漢字文節である」時、ずらした解を正解とする。

この例では、ずらした解の前部 (b2) が辞書に存在するため、そちらが正解となる。

最後に、このようにして得られた3種の解を、図 **2** に示すように合成して、解析結果とする。

4 解析実験

4.1 3種のアルゴリズムによる解析結果

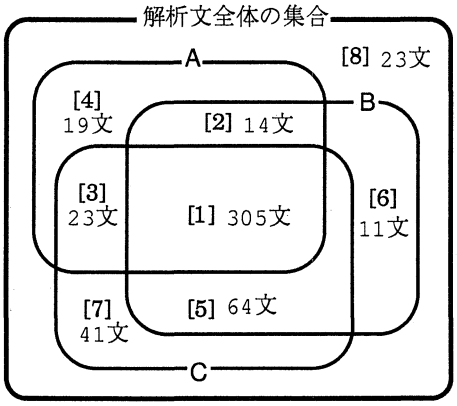
本システムは、最終的には3種類のアルゴリズムによる解を合成して出力するが、比較のため、それぞれの方法のみによる解析実験を行った。

実験は、朝日新聞記事 500 文 (平均 52.4 文字) を対象にして、3種類のそれぞれのアルゴリズムによる正解/不正解の分布を調べた。この結果を図**3**に示す。ただし、正解の判定は「1文全体を通した文節の分割が正しいかどうか」で行っている。

[8] の 23 文は、全ての方法において正解が得られなかった文を表す。

この図から、それぞれの方法による正解率は以下のようになる。

- 後方最長一致 (**A**) = [1]+[2]+[3]+[4]
  = 305+14+23+19 = 361 (72.2%)
- 前方最長一致 (**B**) = [1]+[2]+[5]+[6]
  = 305+14+64+11 = 394 (78.8%)
- 後方最長一致+補正 (**C**) = [1]+[3]+[5]+[7]
  = 305+23+64+41 = 433 (86.6%)



A : 後方最長一致法による正解
B : 前方最長一致法による正解
C : 後方最長一致法 + 補正による正解

図 3: 3 種のアルゴリズムによる解の分布

後方最長一致法による解の中には、「１４日午後に」という文を「(１４)(日午後 に)」と解析してしまうなど、「数字 + 単位 + 漢字の単語」のパターンの文で失敗するものが多く (28 例) 存在し、解析率を低下させる原因の一つとなった。

次に、それぞれの方法のみでしか正解が得られなかった例を挙げる。

- 後方最長一致のみで正解した例 [4]

  ○ **A** : (屈折) (はしご) (車)
  B,C : (屈折 は) (しご) (車)

  ○ **A** : (受け た) (とき から)
  B,C : (受け たと) (き から)

- 前方最長一致のみで正解した例 [6]

  ○ **B** : (計 ２４万 台) (生産 し ている)
  A : (計 ２４) (万台生産 し ている)
  C : (計 ２４万) (台生産 し ている)

  ○ **B** : (曲げ たり) (し やすい)
  A : (曲げ) (たり し やすい)
  C : (曲げ た) (り し やすい)

- 後方最長一致+補正のみで正解した例 [7]

  ○ **C** : (十分) (予想 され る)
  A,B : (十分予想 され る)

  ○ **C** : (当面 は) (この) (方式 を)
  A : (当面) (はこ の) (方式 を)
  B : (当面 は) (この方) (式 を)

  ○ **C** : (今月) (初め に)
  A : (今) (月初め に)
  B : (今月初) (め に)

4.2 複数解の合成による解析結果

3 種のアルゴリズムによって得られた解を合成した時の解析結果を、表 1に示す。対象とした文章は、3 種のアルゴリズムによる解析で用いたものと同一の 500 文である。

正解が含まれる文	479 文	95.8%
未登録語による失敗	14 文	2.8%
その他の失敗	7 文	1.4%

表 1: 解析結果

ここで「正解が含まれる文」とは、合成された解の中に、文節の分割が正しく行われたものを含んでいる文を示す。

「未登録語による失敗」とは、「あさ子」「ぼっ発」といった固有名詞や表記のゆらぎなどの、ひらがなを含んだ未登録語による失敗のことである。これらの単語を登録するとすれば、正解が含まれる割合は 98.6% となる。

以下は、本手法により解析を失敗する例である。

- 複合的な機能語に含まれる単語が、本来の意味で切れなくなる。

  「～を始め」
  “今年度から調査を始め”
  (今年度 から) (調査 を始め)
  「～でない」
  “５０分ほどでなくなった”
  (５０ 分 ほど でな く な った)
  「～ものを」
  “危機的なものを感じる”
  (危機 的 な もの を) (感 じる)

上の文は、解析実験で実際に出現した例である。このような失敗は、ほぼ決まった機能

語に対して起こっている。そのため、そのような語が出現した場合、途中で分割する解も出力することで対処できる。

- 補正のアルゴリズムの中で、1文字漢字を切り出さないようにしているために失敗する。

> "今教育現場では"
> (今教育現場 では)

1文字漢字を切り出さないのは、「見直し」「来賓」「本社」「当選」など、最初の1文字が辞書に登録されているような単語[2]を細かく分割することを防ぐためである。
なお、今回の解析実験ではこのような例は出現しなかった。

次に、複数解の個数を調べたものを、**表2**に示す。複数解の個数とは「文全体を通した、異なる文節区切りの組合せの個数」を言う。つまり、異なる文節区切りが、文中のある場所で2個、別の場所で3個あった場合、解の数は6個と数える。

複数解の個数	文の数	文の長さの平均
1	305	46.9
2	124	56.4
3	23	57.3
4	35	73.8
6	6	81.0
8	6	86.9
12	1	68.0

平均の個数 ... 約 1.7 個

表 2: 複数解の個数

対象とした文の約61%は解の数が1個であり、3個以内であるものが約90% であった。

5 おわりに

前方・後方最長一致法、後方最長一致に補正を加えた方法の、3種のアルゴリズムによる解を合成して解を求める形態素解析システムについて述べた。それぞれのアルゴリズムは非常に単純なものであり、正解率もそれほど高くはないが、3種の解を組み合わせることによって、高い精度で正解を含む解集合が得られることを確認できた。

また、このような方法では複数解が出ることはやむを得ないが、今回の解析実験では、組み合わせによる解の数は予想していた程多くはならなかった。

今後の課題は、ひらがなを含む未登録語に対する処理と、複数解の個数を減らす処理である。

参考文献

[1] EDR 日本電子化辞書研究所 : EDR 電子化辞書 利用マニュアル 第2.1版 (1994).

[2] 「見」- 動詞 ,「来」- 動詞 ,「本」- 副詞 ,「当」- 副詞