言語データベース作成のための形態素解析における未知語検出の検討

4P－5

石川永和　伊藤彰則　牧野正三
東北大学応用情報学研究センター

1　はじめに

現在我々は大規模な言語データベースの構築を行なおうとしている。データベース作成にあたっては大量のテキストを解析することが必要であり、これらのテキストに対して十分な語彙を持つ辞書を用意する必要がある。しかしながらあらゆるテキストに対処できる辞書を構成することは日本語の造語能力などの点から、困難である。またデータベース作成の趣旨からはテキスト中に辞書に記載されていない語が存在した場合この語の文法的性質や意味推定を行ない、最終的には新語として辞書に単語登録する段階に達することが望まれる。

本稿ではデータベース作成の第一段階として行なわれる形態素解析において辞書未登録語を検出することを目的とする。従来さまざまな形態素解析法が提案されているがこれらは解析対象となるテキストに辞書未登録語が現れないことを前提としているものが多く、未登録語が存在する場合の動作は保証されていない。ここでは一旦形態素候補を作成した後、新たな形態素候補を加えることにより、未知語が存在しても形態素解析が行なえるアルゴリズムを開発することをねらう。

2　検出の方針

通常の形態素解析においては最初に文節内文法により文節候補リストを得るわけであるが、未知語が解析文に存在する場合、本来ならばあるはずの文節候補が、このリスト内に存在しないことになる。そこでこのリストに、元々存在する文節候補の情報を元に疑似文節を加え、この疑似文節候補にコストを与えて、文生成の段階で文全体のコストが最小となるような文節系列を選択することにする。このとき選択された文節の中に疑似文節が存在すれば、文節内文法を適用して、未知語区間を最終的に切り出すことになる。この未知語検出までの過程を図1に示す。

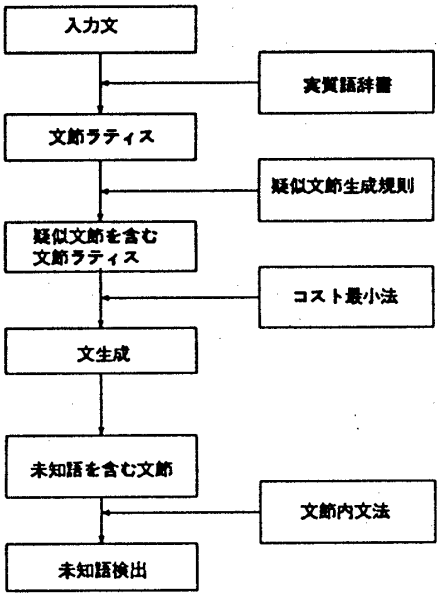

図 1: 未知語検出の過程

前述した方法で未知語を検出する場合にまず問題となるのは、疑似文節の加え方と、そのコストの付加の仕方である。解析対象に未知語が含まれていた場合、未知語の種類に応じて、リストの内容にさまざまな状態が生じると考えられる。そのためこれらの状態を知るために未知語の性質を知る必要がある。[1] によれば、未知語は外来語異表記、人名・地名などの固有名詞、複合名詞等の名詞が大部分である。これらのことから、未知語が名詞であると仮定して疑似文節を加える方法を考える。

3　疑似文節生成

まず以下で用いる記号等の定義を行なう。
入力文を $s(1)s(2)\cdots s(i)\cdots s(N-1)s(N)$ で表すことにする。ここで $s(i)$ は入力の i 番目の文字を表している。また $s(i+1)\cdots s(j)$ が文法情報 a、コスト c である文節であることを $b(i,j,a,c)$ で示し、j 文字目で終

Detection of Unknown Words in the Morphemic Analysis for Corpus
Hisakazu ISHIKAWA,Akinori ITO,Shozo MAKINO
Reseach Center for Applied Information Science,Tohoku University

端する文節の集合を $B(j)$ で示す。文法情報は機能語の付属した名詞節の場合を nf、名詞のみからなる文節の場合を nn、疑似文節の場合を pn で表すことにする。手順を次のように定める。

1. 実質語辞書と文節構造を示す文節オートマトンにより、既知語による文節候補ラティスを作成する。今注目する文字を $s(n)$ とし、$n = N$ として3へ;
2. $n \leftarrow n-1$。$n = 0$ ならば終了。
3. $B(n) = \{b(i,n,a,c)\}$ のうち、$a = nf, nn, pn$ である要素に対し4を行なう。全ての要素について終えたならば2へ。
4. $B(i)$ の各要素に対する条件判断を行なう。全ての要素に対し判断し終えたなら、3へ。
   (a) $B(i) \neq \phi$ の場合
   対象としている文節が $b(i',i,nn,c')$ の形であれば、$b(i',n,nn,c'')$ を $B(n)$ に加える。コストの決定方法は後ほど検討する。4へ。
   文法情報が上記以外のときは4へ。
   (b) $B(i) = \phi$ の場合
   解析文の始端方向に存在する最近接の文節が $s(i')$ で終端しているとすると $b(i',i,pn,c')$ を疑似文節として加える。4へ。

この方法は辞書引きにより得られた文節候補の名詞節に注目して検出されなかった名詞節を救うものである。次に各文節候補に対するコストの与え方を次のように定める。

- 既知語による文節 $b(i,j,a,c)$ にはコスト値として $c = w1$ を与える。
- 疑似文節に対しては
  - 2つの文節の連接により生成された場合これらの文節のコストを $c1, c2$、生成された文節のコストを $c3$ とすると $c3 = c1 + c2 + w2$ で与える。ここで $w2$ は連接によって生じると考えられるコストである。
  - 文節未検出の区間を文節 $b(i,j,a,c)$ として生成した場合、$c = j - i$ で与えることにする。

ここで $w1, w2$ は適当に与えられる値である。

4 未知文節の検出実験

検出はコスト最小法によって行なう。パラメータ $w1$, $w2$ の値を変化させ、最適となる値を求める。

まずあらかじめ辞書を用いて形態素解析を行ない名詞のリストを作成する。次にこの内容の一部を辞書から削除する。そしてこの辞書を用いて、解析を行なう。削除する語数はリストの全体の個数に比べ、十分小さいものとする。ここではパラメータの値を $w1 = 1, w2 = 1$ とした場合の動作を例示しておく。スペースで区切られた文字列は正しく解析が行なわれた場合の文節の区切りであり、ゴチック体で示した語が未知語を意味している。また、下線部で示された部分はその部分が未知語を含む文節であるとして選択されたことを意味する。

- 解析に成功した例
  帯域フィルタや 周波数分析器という
  <u>アナログ技術は</u> 使えなかった

- 解析に失敗した例
  カードに <u>パンチ機で穴を</u> 空けて バッチ処理を依頼していた

  このケースにおいてはこのような「未知語を含む文節＋既知語を含む文節」が1つの「未知語を含む文節」として検出される誤りがもっとも多く観察された。

5 まとめ

形態素解析の段階において未知語を検出することを考え、コスト最小法を用いて検出実験を行なった。今後はさらに名詞の詳細な構造 [5] などを考慮して検出を行なうことを検討していきたい。

参考文献

[1] 亀田, 森田, 倉島, 藤崎: 未知語の分類とその処理規則 情報処理学会第36回全国大会, (1988)

[2] 大沢, 藤崎: 未知語を含む文の形態素解析システム 情報処理学会第42回全国大会, (1991)

[3] 荒木, 栃内: 帰納的学習による形態素解析手法における適応能力の評価 信学技報, Vol.90, No.375, pp.1-8, (1990)

[4] 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, 情報処理学会論文誌, Vol.30, No.3, (1989)

[5] 植田, 小松, 横尾, 宮崎: 部分複合語による複合名詞構造解析 情報処理学会第43回全国大会, (1991)