4F−4

文節オートマトンを用いた未知語検出法

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 ‡東北大学情報処理教育センター

1 はじめに

最近の連続音声認識システムや自然言語処理システムでは、文法・意味などの言語情報の大部分を単語に付属させることが多いが、この情報付与は手作業であるため、自動化が望まれている。本研究は、未知語を検出しその語の文法的・意味的属性を自動的に付与することを目的としている。

われわれは形態素解析において辞書未登録語を検出するため、未知語に対応できる形態素解析法を提案した[1]。しかし、この方法は経験則に大きく依存し、十分な効果を上げることはできなかった。本稿では確率的手法を用いることにより、未知語検出を高精度化する方法を提案する。

2 未知語検出法

文節数最小法に代表される形態素解析法は、全体を通してコストのもっとも少ない経路を選択する、コスト最小法の一種であると言える。文節数最小法は、各文節候補に対し一律のコストを与えることに相当する。本手法ではこれを拡張し、各文節候補に対し、それがどれだけもっともらしいか、つまり「文節らしさ」を反映するようにコストを設定した。

ここで提案する方法では、文節内文法を表現した文節オートマトンを用いて文節を検出するだけでなく、任意区間を実質語と仮定して文節オートマトンを駆動し、未知語を含む文節候補を生成する。このようにして得られた候補を「疑似文節」と呼ぶ。

まず、記号の定義を行なう。

$S = S_0 S_1 S_2 \cdots S_N S_{N+1}$：文節数 N からなる文

$S_i (i = 0, 1, \ldots, N, N+1)$：文節
（ただし、S_0, S_{N+1} は始端、終端を表す、仮想的な文節）

$L = \{X_j\}$：解析によって得られた文節ラティス

$X = X_0 X_1 X_2 \cdots X_M X_{M+1}$：文節ラティスより得られる文として可能な系列
（ただし、X_0, X_{M+1} は始端、終端を表す、仮想的な情報）

$X_j = (x_i, x_2, \cdots, x_m)_j$：解析によって得られた文節に関する情報

今、ある文 S から X を得る過程を確率現象と考える。文 S が生起する確率を $P(S)$、X を観測する確率を $P(X)$ とすると、解析によって複数の X が得られたとき、形態素解析を S と X の同時確率 $P(S,X)$ を最大にする S と X の組を求める問題であるととらえる。すなわち、

$$(S,X) = \underset{S,X}{\operatorname{argmax}}[P(S,X)] = \underset{S,X}{\operatorname{argmax}}[P(S)P(X|S)] \quad (1)$$

$S \rightarrow X$ の生成過程において、文節の脱落、付加、および、分割誤りが生じないものと仮定し、各文節に関する情報は、対応する文節のみに依存すると考えると、$P(X|S)$ は次のように近似的に表される。

$$P(X|S) \simeq \prod_{i=1}^{M} P(X_i|S_i) \quad (2)$$

また $P(S)$ を一次のマルコフ過程で近似すると、

$$P(S) \simeq \prod_{i=0}^{M+1} P(S_i|S_{i-1}) \quad (3)$$

となる。(1)式の対数をとって符号を反転すると、形態素解析は

$$(S,X) = \underset{S,X}{\operatorname{argmin}}[-\log\{P(X|S)P(S)\}] = \underset{S,X}{\operatorname{argmin}} \left[\sum_{i=1}^{M} I(X_i|S_i) + \sum_{i=0}^{M+1} I(S_i|S_{i-1})\right] \quad (4)$$

となり、この式を計算し最小値を与える S, X を選択すれば良いことになる。ここでは $I(X_i|S_i)$、$I(S_i|S_{i-1})$ がそれぞれ文節に対するコスト、文節の接続に対するコストに対応づけられる。

Detection of Unknown Words using a Bunsetsu Automaton

Hisakazu ISHIKAWA†, Akinori ITO‡, Shozo MAKINO†
†Research Center for Applied Information Science, Tohoku University
‡Education Center for Information Processing, Tohoku University

3 検出実験

アルゴリズムの有効性を検証するため、実験を行なった。解析対象は音声関係の科学技術論文より抽出した136文である。これらはあらかじめ記号等を取り除き、整形してある。

実験方法は以下の通りである。まず、あらかじめこれらの文章を解析し、実質語のリストを作成しておく。次に、このリストのうち任意の語を未知語と仮定して、辞書項目から削除し、この辞書を用いて解析を行なう。

実験にあたっては、未知語の性質、すなわちどのような語が未知語となり易いかということを知る必要がある。しかし未知語はほとんど名詞であるという報告がある[3]。そこで、未知語は名詞であると仮定する。1文中に現れる未知語は1語であるとしたが、解析の際には未知語の個数の情報は利用していない。

4 実験結果

観測情報には文節長、実質語長、字種変化数（文節内・実質語）、始終端字種、文節構成品詞、実質語の品詞の7種類、接続情報としては始終端字種、文節構成品詞、実質語の品詞の3種類を考えた。確率の推定値は解析対象から求めた値を用いた。表1に結果を示す。「候補数」は正解を持つ文に対する平均文候補数を表している。なお未知語区間が正しく未知語として切り出されているどうかを検出の基準とした。

文節については実質語の字種変化数を、接続については品詞情報を用いるのが有効であることが解析結果よりわかる。逆に語長や、字種の接続をコスト決定に用いるのはあまり効果がないことがわかった。

表1の結果は、文節候補が辞書登録語によって生成されたものか、疑似文節であるかに関わらず、一律のコストを設定している。しかし、辞書登録語によって得られた文節の方がより「文節らしい」と考えられる。そこで辞書未登録語に対し、付与するコストを0とした実験を行なった。結果を表2に示す。前実験に比べ全体的な検出率の向上とともに、候補数が減少する傾向がみられた。

5 おわりに

本稿で提案した方法により未知語検出の可能性を示すことができた。また検出に有効な情報を知ることができた。今後の課題としてはさらに多くの文にあたり、アルゴリズムを検証する必要がある。また検出誤りの内容を検討し、アルゴリズムの改善を図りたい。

参考文献

[1] 石川, 伊藤, 牧野: 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法の検討 情報処理学会第44回全国大会, (1991)

[2] 伊藤: タスクに依存しない日本語文音声の認識に関する研究 東北大学審査博士学位論文, (1991)

[3] 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)

表 1: 一律のコストを用いた場合

条件	事象	接続	検出率	候補数
なし	文節長	品詞（文節）	24(%)	2.33
		品詞（実質語）	6.5	15.16
		字種	14	66.6
	実質語長	品詞（文節）	32	2.23
		品詞（実質語）	25	1.98
		字種	15	6.77
	字種変化数（文節）	品詞（文節）	13	5
		品詞（実質語）	15	5.7
		字種	0.5	6.5
	字種変化数（実質語）	品詞（文節）	89	5.05
		品詞（実質語）	86	15.76
		字種	74	55.37
文節の構成品詞	文節長	品詞（文節）	28	2.39
		品詞（実質語）	30	2.37
		字種	28	1.93
	実質語長	品詞（文節）	38	2.61
		品詞（実質語）	43	2.52
		字種	38	2.03
	字種変化数（文節）	品詞（文節）	77	5.66
		品詞（実質語）	80	8.91
		字種	82	24.7
	字種変化数（実質語）	品詞（文節）	91	5.07
		品詞（実質語）	92	5.23
		字種	82	2.26

表 2: 辞書による候補にコスト = 0 を用いた場合

条件	事象	接続	検出率	候補数
なし	文節長	品詞（文節）	50(%)	1.17
		品詞（実質語）	37	2.70
		字種	18	7.93
	実質語長	品詞（文節）	63	1.02
		品詞（実質語）	49	1.19
		字種	27	2.32
	字種変化数（文節）	品詞（文節）	80	1.20
		品詞（実質語）	73	4.75
		字種	79	6.51
	字種変化数（実質語）	品詞（文節）	93	1.06
		品詞（実質語）	84	2.41
		字種	77	2.84
文節の構成品詞	文節長	品詞（文節）	53	1.18
		品詞（実質語）	45	1.30
		字種	32	1.36
	実質語長	品詞（文節）	61	1.15
		品詞（実質語）	54	1.28
		字種	46	1.91
	字種変化数（文節）	品詞（文節）	81	1.37
		品詞（実質語）	81	2.31
		字種	81	96.84
	字種変化数（実質語）	品詞（文節）	92	1.21
		品詞（実質語）	90	1.63
		字種	83	7.63