数理解析研究所講究録
229巻 1975年 134-144

構文情報を用いた一つの音声認識法について
——未知語の取り扱い——

九州大学理学部　武谷 峻一
九州大学工学部　河口 英二
九州大学工学部　田町 常夫

1. まえがき

連続音声の認識における言語的な情報の重要性は古くから指摘されており，近年いろいろなレベルの言語情報を利用する認識システムが研究されている[(1),(2),(3)]．これらの多くは，入力文に対して文法的な制限を設け（対象を特定の文法構造をもつ言語分野に限定するなど），その文法規則を利用することにより認識の質を高め，処理効率の向上を図っている．こうしたシステムでは，処理にあたっていずれも"単語"という概念を媒介としなければならず，単語辞書が不可欠である．

辞書を用いる場合，認識の対象を拡大しようとするとそれに伴って内蔵すべき語彙の増大は避けられない．しかし，実

際に入力に現われる単語の頻度にはいちじるしい偏りが見られる(4). 天気予報文448文章を対象としてそれに用いられている単語151語の使用頻度を調べた結果，最も頻度の高い13語で全体の50%，150語では97%以上が尽されている。そして残りの2～3%のために頻度1または2の単語約100語が必要となっている．こうした傾向は自然語の一般的法則としてよく知られている．したがって，辞書の利用効率から見れば，すべての単語を辞書に登録しておくよりも，頻度の低い単語（たとえば固有名詞など）を“未知語”として検出できる方式であることが望ましい．

こうした観点から，本稿では文の構文的な構造を利用して未知語を検出する一つの方法を提案する．

2. 未知語取り扱い上の問題点

未知語を扱う場合，入力系列を単語に区切ってゆく操作が必要である．単語への分割操作ということは通常の言語処理ではあまり問題にならない．すなわち，入力があいまいさのない文字列であれば，見出し語と完全に一致する部分を単語として区切ればよい．この場合，未知語の検出は比較的容易である．とくに英語では空白によって単語が区切られているのでほとんど問題がない．しかし音声認識では，一般に入力系列は物理的な特徴によって識別されたあとの不完全な（あ

いまいな記号列であり，見出し語との照合において完全一致となるものを探し出せることは期待できない．そこで，相対的な一致の尺度を設定して判定せざるを得ないが，そのために分割の多様性が生じることになる．したがって，未知語を取り扱うには，何らかの形で構文的な情報を利用する分割法が必要となる．その際，一般によく行なわれる左から右への分割法は不適当であろう．

3. 未知語を考慮した単語分割

2.で述べた点を考慮して，構文を利用し未知語の検出およびその構文上の機能の推定が可能な分割法について述べる．

3.1 入力に対する仮定

まず，入力に対してつぎのような仮定を設ける．すなわち，対象とする音声は以下に定める生成文法Gにより生成された文を区切り記号の部分で区切りながら連続音声として発話したものであると仮定する．

文法Gは

H：開始記号，

$S = \{a_1, a_2, \cdots, /\}$：音韻記号と区切り記号の有限集合，

$W = \{w_1, w_2, \cdots, w_N, w_{N+1}, \cdots\}$：単語の集合，Nは辞書に登録されている単語数，

$C = \{c_1, c_2, \cdots\}$：構文的機能（品詞，区切りなど）を表わ

す記号の有限集合，

$V = \{\alpha, \beta, \cdots\}$：書き換えのための変数の有限集合

とする．Sは終端記号の集合，$W \cup C \cup V$は非終端記号の集合であり，生成規則Pはつぎの4つの形式からなる．

(i) $H \rightarrow /\alpha/$

(ii) $C_i \beta C_j \rightarrow C_i \gamma C_k \delta C_j$

(iii) $C_i \rightarrow w_n$

(iv) $w_n \rightarrow s_1 s_2 \cdots s_l$

ただし，区切り記号は単語でもあり構文的機能を表わす記号でもある（$/ \in W$, $/ \in C$）．また，$\varepsilon \in V$（$|\varepsilon| = 0$）とする．Gの文は一般に

$Q = /s_1 s_2 \cdots s_l / s_{l+1} \cdots / \cdots s_m /$

となる．

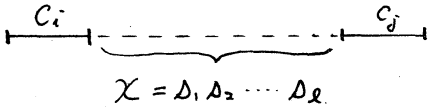
3.2 単語分割のアルゴリズム

Qを連続発声した音声は，その物理的な特徴によって識別され，一たん音韻記号列Q'（$\in S^*$）に変換される．この記号列は一般に誤って識別された音韻記号などを含む不完全な記号列である．本稿ではこのQ'を入力系列として取り扱う．Q'中の音韻記号列と辞書$D = \{w_1, w_2, \cdots, w_N\}$（$\subset W$）中の単語との相対的な一致の尺度はつぎのように定める．いま，$S_0 = S - \{/\}$，$x \in S_0^*$とするとき，xと$w \in D$との一致の度合を

$\omega(w, x)$とする．このとき，wに固有のしきい値$\theta(w)$を考え，$\omega(w, x) \geq \theta(w)$であれば$w$と$x$が一致したとする．

Gの文Qを認識する立場から見ると，未知語$(w_{N+1}, w_{N+2}, \cdots)$に関しては$G$の生成規則(iii)，(iv)が欠落していたと見なすことができる．そこで単語分割にあたっては，生成規則(i)，(ii)の形式を構文的な情報として利用する．すなわち，関数亜：$C \times C \to 2^C$を考え，左および右側の構文的機能がそれぞれC_L，C_Rであるとき，$C_L - C_R$間の構文的機能を亜(C_L, C_R)で規定する．この亜と辞書Dを用いる分割のアルゴリズムをつぎのように定める（例参照）．

分割は段階的に行なうこととし，いま注目している音韻記号列$x = a_1 a_2 \cdots a_\ell$（$x \in S_0^*$，$Q'$の部分記号列）に関して，$C_L = C_i$，$C_R = C_j$とするとき，



(I) 亜$(C_i, C_j) = \phi$の場合．

$C_i - C_j$間に許される構文的機能がない場合で，前段でのC_iあるいはC_jの決定が誤りである．前段へ戻り新たに別のC_i'あるいはC_j'を選ぶ．

(II) 亜$(C_i, C_j) \neq \phi$で，かつxの部分記号列y（$y \prec x$と表わす）に対して

$$T_y^x = \{w \in D \mid \omega(w, y) \geq \theta(w),\ w \in \psi(c),\ c \in 亜(C_i, C_j)\}$$

（ただし，ψは$C \to 2^D$であり$\psi(c)$で構文的機能cをもつ単語の集合を表わす）としたとき，すべての$y \prec x$に対して$T_y^x = \phi$の場合．

すなわち，$c_i - c_j$間に許される構文的機能はあるが，文中に$c \in 至(c_i, c_j)$をもつ既知語$w \in D$とも一致する部分がない場合は，xを未知語と決定する．

(III) $至(c_i, c_j) \neq \phi$で，かつ$T_y^x \neq \phi$となる$y \prec x$が存在する場合．

すなわち，$c_i - c_j$間にある$c \in 至(c_i, c_j)$をもつ既知語$w \in D$と一致する部分がある場合は，

$$Y_x = \{ y \prec x \mid T_y^x \neq \phi \},$$

$$\omega(w_0, y_0) = \max_{y \in Y_x} \max_{w \in T_y^x} \omega(w, y)$$

とおき，このy_0の部分をw_0とする．そしてQ'のy_0に相当する部分をw_0の構文的機能で置き換えて次段へ進む．なお，次段が(I)の場合で逆戻りしたときは，順次求め，n番目に大きい$\omega(w, y)$を選ぶことにする．

以上の分割法により，未知語部分は最後に音韻記号列のまま残る．いま，この両隣の構文的機能が$c_L = c_i$，$c_R = c_j$であるとすれば，その未知語の構文的機能は$至(c_i, c_j)$のいずれかであると推定できる．

実際の分割操作では，$C \in 表(C_L, C_R)$ に対して，C_L に隣接，C_R に隣接，両方に隣接，中間位置などの情報が利用できる．したがって，$w(w, x)$ の計算が単語の語尾からもできる方式であれば，これらの情報により照合の回数を減らすことが可能である．

3.3 表の構成

上に述べたアルゴリズムを用いて Q' を単語に分割してゆくとき，あらかじめ表をどのように定めておくかが問題となる．

表としては，まず

$$表_1(C_i, C_j) = \{C \in \mathcal{C} \mid C_i \alpha C_j \to C_i \beta C \gamma C_j \in P\}$$

が考えられる．これは単語分割を G の生成過程と全く同じに進ませることを表わしている．この場合，たとえば Q の両端の区切り記号 / に関して P に $/\alpha/ \to /\beta C_0 \gamma/$ がただ1個しかないときには，$表(/,/) = \{C_0\}$ となる．この C_0 に相当する部分が未知語であれば，分割は第1段目で未知語を検出して止まり，Q' 全体が未知語と決定される．

つぎに，表として $L(G)$ のすべての文に対して任意の C_L — C_R 間の構文的機能を調べ，可能な構文的機能をすべて $表_2(C_L, C_R)$ とする方法がある．一般に，任意の C_L, C_R に対して $表_1(C_L, C_R) \subseteq 表_2(C_L, C_R)$ となる（例参照）．この $表_2$ では，Q の両端の / に対して $表(/,/) = \mathcal{C}$（すべての構文的機能が

含まれる）となり，分割の第1段目でD内のすべての単語とQのすべての部分系列との照合が必要となる．

以上の点を考えれば，T_1は未知語が存在するQに対しては不適当であり，またT_2は単語との照合回数が多く処理効率が悪い．したがって，現実のTとしてはT_2を採用し，$T(C_L, C_R)$内のCに対して優先度を設けて処理を行なうか，あるいは未知語を許す構文的機能を限定しておく（たとえば，助詞，助動詞などは未知語を許さず，固有名詞は許すなど）などの制限を設ける必要があろう．また，いずれの方法を用いても，3.2のアルゴリズムでは$Q \notin L(G)$を受け入れる可能性があり，単語分割のあとに構文的なチェック（未知語に対しては推定された構文的機能をあてはめてチェックを行なう）が必要である．

4. むすび

本稿で示した分割法の特徴は，構文的な情報を形式的にTで取り扱い，上から下へ分割してゆくことによって，未知語の検出およびその構文上の機能の推定が可能な点である．しかし，本方式で未知語として処理される音韻記号列には，(1)真の未知語である場合と，(2)既知語でありながら音声の物理的な特徴による識別段階で誤りが大きく，分割を誤った場合とがある．この両者を区別する情報はなく，いずれも

同一の処理を行なわざるを得ない．また，連続した未知語は取り扱えないなどの問題点が残されている．

現在，本方式による認識実験を行なっており，近くその成果を報告する予定である．

参考文献

(1) A. Newell et al. : " Speech Understanding Systems, Final Report of a Study Group ", Carnegie-Mellon Univ., Pittsburgh (1971-05)

(2) 斎藤他：" 音声認識における言語情報の利用 "，昭和49年電気四学会連合大会論文集，No. 193～199（昭49-10）

(3) 武谷，河口："言語構造情報を利用した連続音声認識システムのシミュレーション "，電子通信学会論文誌(A)，Vol. 56-A, No. 9（昭48-09）

(4) 河口，武谷，田町：" 音声認識における文法規則の利用 "，昭和49年電気四学会連合大会論文集，No. 195（昭49-10）

構文的機能と単語

T：明日は，明後日は，…　C：曇り，晴れ，雨，…

F：のち，一時，時々，…　X：です，でしょう，…

生成規則（(i)，(ii)の形式のみ）

$H \to /\alpha/$　　$/\beta C \to /T\delta C$

$/\alpha/ \to /\beta C/$　　$T\delta C \to TC\varepsilon C$

$/\alpha/ \to /\beta C\gamma/$　　$C\varepsilon C \to CFC$

$/\beta C \to /TC$　　$C\gamma/ \to CX/$

文

$/T[CF]C[X]/$　　[ ]は省略可

{ 明日は晴れ．　明後日は雨でしょう．
{ 明日は曇りのち晴れです．　…

$\mathfrak{T}_1$

C_L \ C_R	/	T	C	F	X
/	{C}	φ	{T}	φ	φ
T	φ	φ	{C}	φ	φ
C	{X}	φ	{F}	φ	φ
F	φ	φ	φ	φ	φ
X	φ	φ	φ	φ	φ

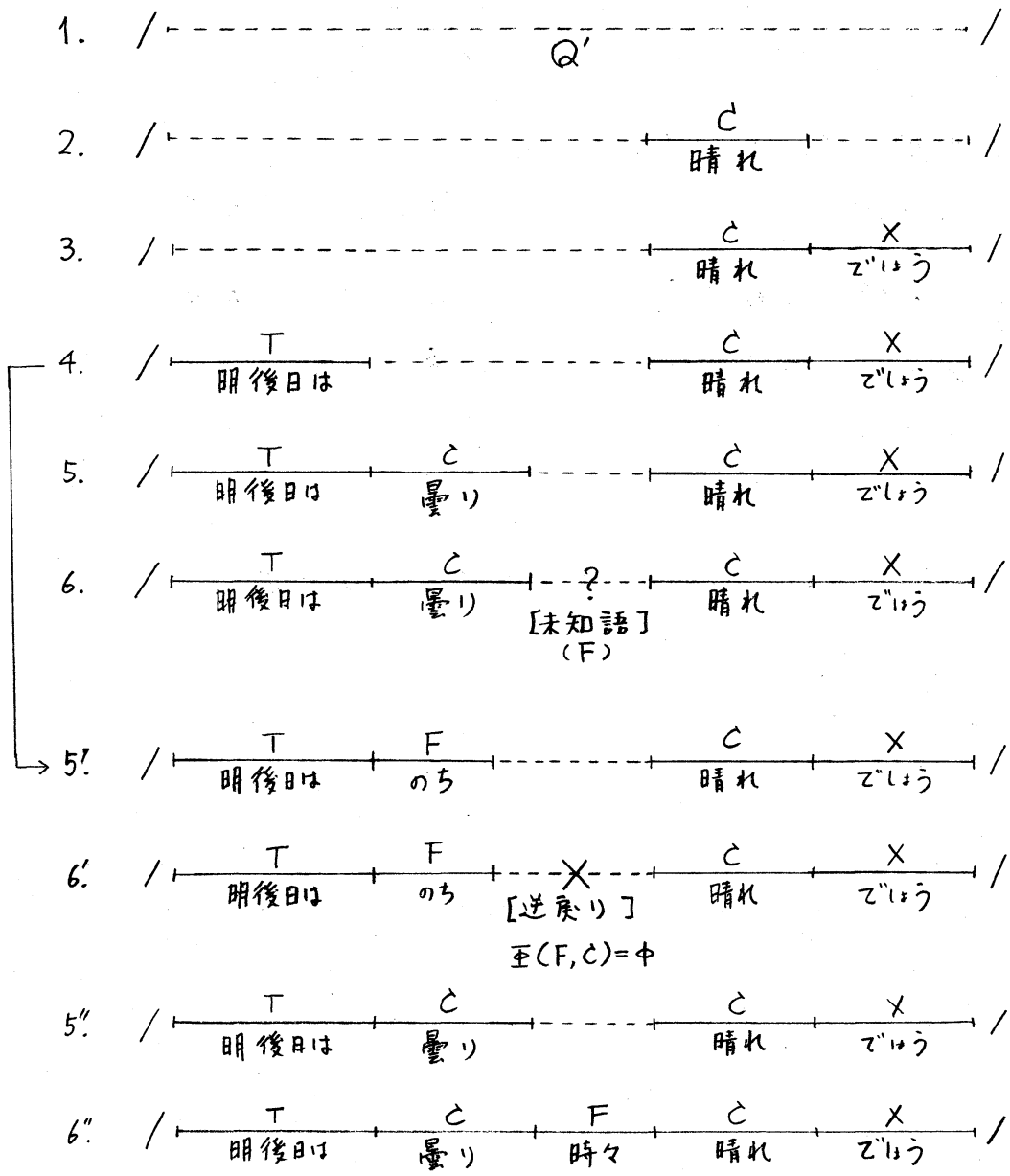
$\mathfrak{T}_2$

C_L \ C_R	/	T	C	F	X
/	{T,C,F,X}	φ	{T,C,F}	{T,C}	{T,C,F}
T	{C,F,X}	φ	{C,F}	{C}	{C,F}
C	{C,F,X}	φ	{F}	φ	{C,F}
F	{C,X}	φ	φ	φ	{C}
X	φ	φ	φ	φ	φ

例

144

単語分割の進み方



例（続き）