3Q−3　コアワードとその共起単語の自動抽出による誤りを含む文の復元

佐藤　智榮†　荒木　健治††　宮永　喜一†　栃内　香次†
†北海道大学工学部　††北海学園大学工学部

1　はじめに

音声認識や文字認識など自然言語を対象とした認識では，信号情報のボトムアップ処理の他に，自然言語の特性を生かしたトップダウン処理過程が存在する．さきに発表した多段階分割復元法[1]でトップダウン的手法による復元処理の有効性が確認されたが，認識候補が多く，処理量が大きいという欠点があった．

このような欠点を補うため，本報告では，このトップダウン処理に利用する情報として，単語共起関係に着目し，手がかり語とそれに共起する語を表層情報から自動的に抽出し，利用する復元手法を提案する．また，本手法に基づく実験システムを作成し，その有効性について述べる．

2　概要

音声認識装置の認識結果など，誤りを多く含む文字列を人間が復元する場合，はじめに比較的少数の手がかり語を見つけ，その語と何らかの意味的関係を持つ語を優先的に復元していくと考えられる．したがって，計算機で同様な処理を行うには，この意味的関係をあらかじめ辞書化しておく必要がある．

しかし，このような辞書は作成に膨大な手間がかかり，また，単語の意味が使われ方に依存するという理由により，実際の文書から自動的に抽出することが必要となる．

本手法では，上述の意味的情報に相当するものとして単語共起関係を想定し，手がかり語（コアワード）とその共起単語を自動的に抽出，利用することにより復元処理を行なう．

以下ではこの方法を音声認識結果からの原文の復元に応用したシステムについて報告する．

3　処理過程

本手法の応用として音声認識装置の認識結果を復元するシステムについて述べる．本システムは音素表記された誤りを含む文字列を受け取り，漢字かな混じりの復元結果を出力するシステムである．

本システムの構成を図1に示す．

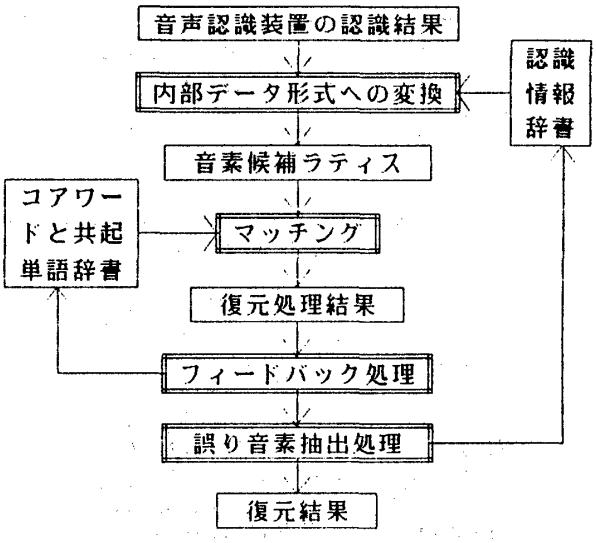

図1　復元システムの構成

3．1　内部データ形式への変換

誤りを含んだ文字列データは認識情報辞書を参照し，マッチング処理に向いた内部形式の音素候補ラティスに変換される．図2に認識情報辞書の例を，図3に音素候補ラティスの例を示す．また，図2中の「確からしさ」とは過去に「対象の音素」が「候補の音素」であった割合のことである．

対象の音素	候補の音素	確からしさ
o	o	0.9932
a	a	0.9848
d	d	0.9804
……	……	……
zu	yo	0.8421

図2　認識情報辞書の例

m	a	i	k	u	r	o	k	o	n	p	y	u	h	t	a
m	a	i	s	u	u	o	s	o	m		zy	u	u	t	a
u		s	k	o	rp		k		n	p	w	o	h	p	
n			t		n		t		no				s		
			p				p								

図3　音素候補ラティスの例

Recovery of Errors in Sentences Using Automatic Extraction of Core Words and Co-occurrence Relation Words

Norihide SATO[1], Kenji ARAKI[2], Yoshikazu MIYANAGA[1], Koji TOCHINAI[1]
1 Hokkaido University, 2 Hokkai-Gakuen University

3．2　マッチング処理

本手法では，先に述べたように，比較的少数の手がかり語（コアワード）とその共起単語を使って復元していく．例えば，図4（a）に示す音素列を復元する場合，コアワードと共起単語辞書が（b）であったとすると，まずラティス構造に変換し，コアワードの音素列と一致するパスをラティス構造から探す．この場合，コアワードである「プログラム」が当てはまったとすると，その共起単語，「マイクロ」と「方式」がマッチングして復元され，（c）の復元結果が得られる．

（a）入力データ
haikugopusogumamihousikigasaiyoupareru

（b）コアワードと共起単語辞書

コアワード	共起単語
プログラム (puroguramu)	マイクロ(maikuro) 命令(meirei) 制御(seigyo) 方式(housiki)
制御 (seigyo)	マイクロ(maikuro) 方式(housiki)
フィールド (fihrudo)	アドレス(adoresu) マクロ(makuro) デコード(dekohdo)
論理 (ronri)	方式(housiki) 柔軟(zyuunan) 制御(seigyo)

（c）復元結果
マイクロプログラム方式gasaiyoupareru

図4　マッチング処理の動作例

3．3　フィードバック処理

フィードバック処理はマッチング結果をもとに，コアワードとその共起単語を抽出する部分である．

このコアワードとその共起単語の辞書は，多段階分割復元法の欠点であった候補数の多さを改善するため，復元処理の最初の手がかりとしてふさわしくない一般的単語を取り除く．そして，どのような単語からも共起される語は一般的単語であると仮定し，辞書を自動的に作成する．次に，出現率と一般化率の定義を示す．

出現率　＝　単語の出現回数／自立語の延べ数
一般化率＝　共出された回数／自立語の数

各々の語について出現率，一般化率を求め，出現率が高く，一般化率の低い単語を取り出す．これをコアワードとする．次に，個々のコアワードの共起単語を求めるため，一般化率の高い語を取り除いた後で共起率を求め，共起率の高い語をコアワードに付随する共起単語とする．共起率の定義を次に示す．

共起率　＝コアワードとの共出回数／コアワードが共出させた語の延べ数

3．4　誤り音素抽出処理

この段階では，正しい音素列がマッチング処理等の出力から得られるので，最初に入力された誤りを含む文字列と比較することにより，最初に入力されたデータのどの部分が誤っていたのかがわかる．このデータを積み重ねていくことにより，認識情報辞書が作成される．

本手法では，連続音声の復元を対象としているため，音素の付加・脱落に対応できるようになっている必要がある．一般に，付加・脱落を考慮すると，その誤りの対応関係は一意に定まらない．そこで，誤りの部分が最小になるような対応関係を抽出する方法を用いた．この方法は，連続して一致する部分を長さと位置で評価し，評価の高いもので木を構成する．この木の中から誤りの最も少ないパスを見つけていく．図5にその結果を示す．

正しいデータ　：　ronrinozyuunanseinikaketeiru
誤を含むデータ：　sonrinozyuumanseinikaketeigi

正しいデータ	誤りを含むデータ
r	s
onrinozyuu	onrinozyuu
n	m
anseinikaketei	anseinikaketei
ru	gi

図5　得られるデータの対応関係

4　おわりに

本稿では，単語共起関係を利用した認識誤り復元手法を提案し，音声認識装置の認識結果を復元するシステム作成した．今後の課題としては，大量のデータを用いた実験により本システムを評価することを考えている．

謝辞

本研究の一部は平成3年度特定研究経費によって行われた．

参考文献

1）荒木，宮永，栃内：多段階分割復元法による誤りの多い文字列からの原文の復元．情報処理学会論文誌　Vol.30,No2,pp169-178(1989)