Associative Memory based on the Co-occurrence Relations between Words

単語間の共起関係に基づく連想記憶

三 好 義 昭

Associative Memory based on the Co-occurrence Relations between Words

Yoshiaki MIYOSHI

1．まえがき

音声を利用したマン・マシーン・コミュニケーションにおける実用的な音声認識方法として限定単語認識がある[1], [2]。これは，認識システムが使用する単語（短い音韻列）についての音響的特徴量をあらかじめ記憶しておき，入力された音声情報に対して，記憶しておいた単語群を検索，照合し，一致する単語を選び出して認識する方式である。単語単位での認識によって，入力された音声中に部分的に間違った，あるいは認識困難な音韻が混入していても，用意された単語の中から最も共通点の多いものを選ぶことによって，良い認識率が得られるという利点があり，近年の分析技術や情報処理装置の進歩によって音声入力装置として実用化されている。しかし，より人間に近いマン・マシーン・コミュニケーション・システムが要求されている今日，限定単語の認識という制限から解放された，自然言語的な会話音声の自動認識・理解システムの実現が期待される。

ところで，人間の音声認識過程は，音韻レベルでの認識，単語レベルでの認識，そして構文的レベルでの認識と低次から高次への処理に区分できる[3]。また，単語レベルでの認識はさらに，符号として単語を認識する単語認識と，認識された単語を記憶にある様々な情報と結び付ける意味理解の二つに分けることができる。ここで言う様々な情報とは情緒や視覚，感覚などの情報で，単語の持つ本質的な意味である。人間の音声認識過程を考えた場合，この二つが密接に関連していることが分かる。例えば，

（a）日常の会話で，話し手からまったく知識のない専門的な話を聞いたとき，正確に発音された単語でも，意味を理解するどころか，聞き取ることもできない場合があること。

（b）会話の途中に話し手がそれまでの話の内容となんの脈絡もない単語を突然話すと，聞き手はその単語を理解するどころか聞き取れない場合があること。

（c）会話の途中に話し手が単語の音韻の一部分を誤って話した時，聞き手がその間違いを知覚し，または無意識に，正しい単語を認識する場合があること。

等々を日常的によく経験する。

第１の例(a)は言語情報の意味理解が音声の認識(聞き取り)に影響を与えていることを示している。すなわち，人間の音声認識モデルで認識部から意味理解部への流れとは別に意味理解部から認識部へのフィードバック・ループがあることになる。

第２及び第３の例(b)(c)から，直前までの文章を理解した上での，文章または単語の流れの予測が行われていると考えることができる。つまり，(b)では予測外の単語によって，聞き手の誤った認識や混乱が起こり，また反対に(c)では予測の範囲内なら多少，間違いのある単語でも修正して認識していると考えられる。

以上のように人間は，意味理解に基づいた予測によって単語の認識を制御しているといえる。本論文では，この単語の認識と意味理解を

平成 18 年 10 月 2 日受理

関連づける一方法として，単語間の共起関係に注目し，ある対象とする単語について連想される単語群を，対象単語の近くに出現する単語を集積することによって得る方法を検証する。以下，2．において同音異義語の区別を例に自然言語の理解と単語認識の関係を示し，3．において単語間の共起関係に基づく連想単語抽出の手順について述べ，4．では，本手法を実際に小学校1年および6年の国語の教科書に適用して，その有効性を示す。

2．自然言語の理解と単語認識

人間の音声認識における単語レベルでの意味理解と同等の機能をシステム的に実現することは，現時点での情報処理技術レベルでは，記憶される情報の性質上，実現は非常に困難といえる。しかし，単語と結び付く種々の情報はある程度の不正確さを許せば言語情報に置き換えることが可能である。この性質から言語情報のみによる意味理解の方法を考察する。

単語レベルでの意味理解の課題に日本語の同音異義語の区別が挙げられる。例えば，

（a）“はし（橋）”の下を川は流れる。
（b）“はし（箸）”を使って米を食べる。
（c）机の“はし（端）”から鉛筆が落ちる。

この例の区別が人間には難しくないのは，それぞれの“はし”についての情報が言語情報（単語または文）として記憶されているからだと考えられる。例えば，橋（川，渡る，道路），箸（食べる，持つ，竹），端（角，机，落ちる）などは人間の連想記憶にあたり，これらの単語を発見することで，“はし”は区別することができる。これはシステム的にも可能なので，ある単語について連想される単語群を記憶しておくことによって，この問題は解決できるといえる。従って，音声認識システムの単語レベルでの意味理解は，符号としての単語と，連想される単語群とを対応させることで実現することができる。また，このように単語を理解することによって，単語認識のための次出現単語の予測が可能になる。ただし，ここで言う予測は次出現単語をある範囲に限定することに留まる。しかし，無数に単語のある自然言語を処理する場合，検索，照合する単語数をある範囲に限定できることは大きな意味を持つといえる[4]。そして，単語を認識，理解すると同時に次出現単語の予測をするという過程を繰り返すことで連続音声の認識の可能性も出て来る。例えば，図1で“かわ”という音韻列を単語として認識した場合，“かわ”を川と仮定して連想される単語をある範囲まで探すことによって，“流れる”，“魚”の2単語を発見することができれば，最初の単語を“川”と理解する。次に“流れる”から“穏やか”を，“魚”から“集まる”を連想し発見するというように単語認識，理解と連想，を繰り返すことで文章全体の単語列が明らかにできると考えられる。また可能性のある単語の認識をある程度試みて意味理解のできない場合は，その音韻列の認識を保留して時間的に後方の単語の意味理解を待つことも考えられる。

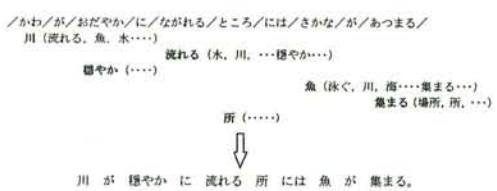


図1　連続音声の単語列の認識と意味理解

3．単語間の共起関係に基づく連想記憶

3．1　重み付き連想単語の抽出

人間の言語学習の発展段階で単語間の関連性が初めて言語情報として加えられるのは単語列を理解するようになってからだと考えられる。文章（単語列）の最小単位は2単語からなる（主語・述語），（修飾語・被修飾語）で，これらは様々な概念に関する言語情報の基本形になる

といえる。このレベルでは構文的に順序関係を考慮する必要もなく単語間の関連性の強さは理解できる。また，これらはかなり複雑な文章になっても保存されていて，同じ文中の近い位置にある単語間の関連性は強いという原則になっていると考えることもできる。前述の連想記憶による単語の意味理解，予測という観点からしても関連性の強い単語は近い位置になければ意味がないといえる。

具体的には，対象とする単語との文中の距離や前後関係に応じてそれぞれの位置に荷重設定し，荷重範囲内に出現する単語を記憶し，出現した位置に応じた荷重を加算していく。その様子を図２に示す。

ここで対象とする単語の位置を原点にとり，文章を読んで行く方向を横軸のプラス方向にとる。音声を処理する場合を考えると，横軸のプラス方向は時間的に後方であり対象単語の出現以後の単語列になり，マイナス方向は時間的に対象単語の出現以前である。本論文では言語情報を，文字による文章中の単語列に限定して，プラス方向からマイナス方向へ文章を分解した単語列を移動することで単語列中のすべての単語について荷重による順位を持った単語群を集める。ただし，単語列は文章中の句読点で分断し，句読点を超えた位置の単語は対象としないことにする。図２は“流れる”を対象として，前２語～後３語迄にそれぞれ/0.3/0.7/対象語/1.0/0.6/0.2/ の荷重を設定した場合の例であり，その処理結果を表１に示す。表１は“流れる”を対象語として得られる単語群とその関連度を示したもので，“流れる”に関連する単語群として“所”，“ゆっくり”，“魚”，“川”，“集まる”が得られ，それらの関連度をそれぞれ1.0，0.7，0.6，0.3，0.2 と得点付ける。この方法で多くの文章を処理し，ある対象とする単語について得られた関連単語群の上位の単語が文章を読んだ人の持った印象や連想と一致するなら，この方法は人間が連想記憶を得る方法に近いといえる。

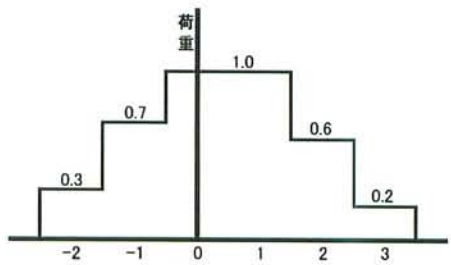


図２　単語の重み付け

表１　抽出される単語群

対象単語:流れる	
0.3/0.7/対象単語/1.0/0.6/0.2	
単語群	関連度
所	1.0
穏やか	0.7
魚	0.6
川	0.3
集まる	0.2

３．２　単語列への変換手順

単語の予測なしでは連続音声を認識するのは困難である。従って，その単語予測と単語の意味理解に必要な連想単語記憶を得るには，構文的に簡単な文章を単語列に変換した上で，できるだけ大量にデータを処理して，統計的に単語間の関連性を明らかにしていかなければならない。前節の重み付きの単語記憶方法によって単語列の統計的な解析を行なって，記憶される一つ一つの単語に対して関連の強い単語群を集め，連想単語を得る。以下の処理を行なって，文章を単純化し単語列とする。

① 文章を可能な限り漢字仮名混じり文で分かち書きする。

② 漢字三字以上の複合語は分解する。一般に一漢字が一概念を持つので，複数の漢字を含む単語は分解できる。ただし，漢字二字の単語はそれ以上分解しない。

（例）　水中翼船　→　／水中翼／船
　　→　水中／翼／船
　　発電所　→　発電／所

③ 助詞や助動詞等の付属語や接続詞，疑問詞，等は消去する。

すなわち，品詞の格情報は使用しない（構文レベルでの処理を行なうことなく，単語間の関連性がどの程度まで明らかになるかを見出すのが本論文の目的である）。

④ 固有名詞，感動詞，記号，等は消去する。

⑤ 活用のある品詞は，基本形にもどす。また，自動詞と他動詞の区別のある動詞は自動詞の形に統一する。

（例）　冬眠／して／います　→　冬眠／する
　　水／止める（他動詞）
　　→　水／止まる（自動詞）
　　電気／起こす（他動詞）
　　→　電気／起きる（自動詞）

以上の処理によって例えば，

（例）原　文：水を塞き止めて発電所で電気を起こしています。
　　処理文：水　塞き　止まる　発電　所　電気　起きる　。

となる。

4．処理結果

4．1　処理資料

個人によって日常的に読む，あるいは過去に読んだ書籍が異なり，さらには生活環境が異なる事から，同一の単語から連想される単語は個人によって当然異なる。しかし，義務教育では年代が同じであれば，内容的に全員ほぼ同等の教科書を使用しており，単語間の共起関係に少なからず影響を与えていると思われる。そこで，重み付き連想単語の抽出対象として，構文的にも簡潔な小学校教育において使用されている教科書を使用した。特に，国語は日本語を学ぶうえで重要であり，また低学年と高学年の違いにも注目すべく，小学校1年生と6年生の国語の教科書[5]-[8]を処理した結果を示す。

4．2　単語数／文章の分布

図3に1文章が何単語で構成されているかの分布を示す。国語1年の総文章数は277文章で1文章の平均単語数は5.0単語であった。一方，国語6年の総文章数は773文章で1文章の平均単語数は6.9単語と，当然の事ながら，学年をおう毎に文章が複雑になり1文章の構成単語数が多くなっていることが分かる。ところで，両

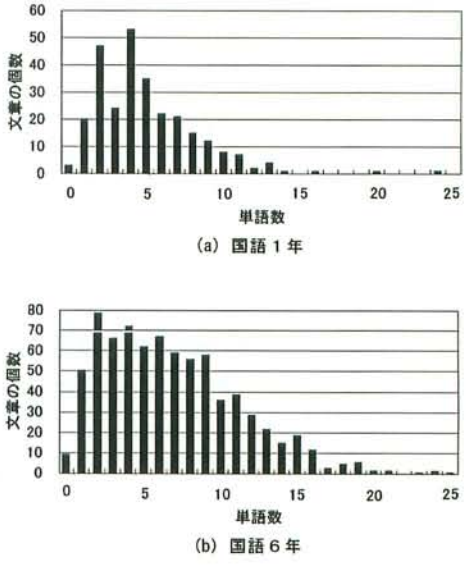


図3　単語数／文章の分布

学年とも1単語から成る文章，さらには単語数0の文章がある一方，国語1年の文章中に，1文章が24単語で構成された文章，国語6年の文章の中にも，1文章が25単語で構成された文章が存在する。それぞれの例を表2に示す。表2(a)(b)の例のように主に会話文中の短文の場合に3.2節の単語列変換則により1単語で構成ないし0単語（文中に該当単語無し）で構成となる文章が生じる。一方，20単語以上からなる文章となるのは，表2(c)の例のように，国語1年の文章では単純な繰り返しの連鎖で長くなって

いるだけで，文章自体は簡単な文章である。また，国語６年の例文では複数の文が読点で連結されて長文となっているだけで，構文的には簡単な文章といえる。いずれにしても，２０単語以上となる文章は両学年とも数例(国語１年：277 文中２例，国語６年：773 文中８例)であることから特段の考慮は不要と考える。

表２　処理文の例

(a) 0単語となる例

	国語1年	国語6年
原　文	「うんとこしょ、どっこいしょ。」	えっ。
処理文	。　(0単語)	。　(0単語)

(b) 1単語となる例

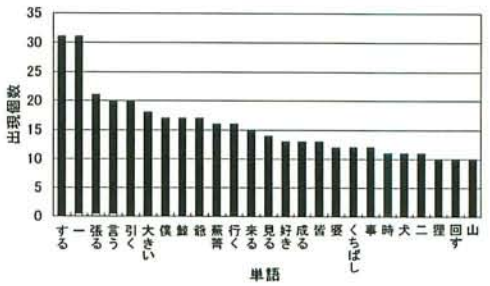
	国語1年	国語6年
原　文	おや、もう　おひるだ。	だめだよ、こんなのじゃ。
処理文	昼。　(1単語)	駄目。　(1単語)

(c) 20単語以上となる例

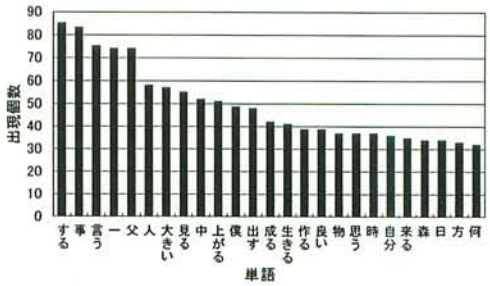
	国語1年	国語6年
原　文	かぶをおじいさんが　ひっぱって、おじいさんをおばあさんが　ひっぱって、おばあさんをまごが　ひっぱって、まごを犬が　ひっぱって、犬をねこが　ひっぱって、ねこをねずみがひっぱって、「うんとこしょ、どっこいしょ。」	多くの人が共に考え、工夫し合うことで、ユニバーサルデザインでの物作りがあたりまえになり、いろいろな人が、いっしょに、安心して暮らせる社会が実現すると思います。
処理文	蕪菁　爺　引く　張る　爺　婆　引く　張る　婆　孫　引く　張る　孫　犬　引き　張る　犬　猫　引く　張る　猫　鼠　引く　張る。　(24単語)	多い　人　共　考える　工夫　する　合う　事　ユニバーサル　デザイン　物　作る　当たる　前　成る　色々　人　一緒　安心　する　暮らし　社会　実現　する　思う。　(25単語)

４．３　出現単語の分布

両学年の出現単語数はそれぞれ国語１年 445 単語，国語６年 1,145 単語であった。このうち出現頻度の高い順に上位 25 単語までをそれぞれ図４(a)(b)に示す。両学年とも“する”と数詞の“一”が可成りの頻度で現れていることが分かる。また，上位 25 位以内に入るには国語１年では 10 回以上出現していれば良いが，国語６年では 30 回以上出現している必要がある一方，１回しか出現しない単語も，両学年それぞれ国語１年 210 単語（率にして，46.2%），国語６年 469 単語（率にして，41.0%）もあった。



(a) 国語１年



(b) 国語６年

図４　各単語の出現個数

４．４　連想単語の抽出

国語１年の「蕪菁(かぶ)」を対象単語，「抜く」，「大きい」，「成る」を連想単語として，荷重範囲を前後２語及び前後３語として得られた結果を表３(a)に示す。なお，荷重は全て 1.0（すなわち，対象単語からの位置関係に関わらず，重みは全て 1.0）にしたときが連想単語の関連度の総和がいずれも最大値 2.0 となった。しかしながら，関連度の総和は同じながら，荷重範囲を前後３語とすると「張る」が８位から３位に上昇し，「大きい」，「成る」の順位が１ランク下がる結果となる。国語１年の文章は主格・目的格が近接した簡潔な文章であることから，荷重範囲は余り広げずに前後２語程度までとすれば良いといえる。

国語６年の「森」を対象単語，「見る」，「木」，「熊」を連想単語として，得られた結果を表３(b)に示す。表３(b)より，荷重範囲を前後３語とし，荷重は全て 1.0 とした場合(表３(b)左)，連想単語の関連度の総和が 2.0 となり，何れも

高順位に抽出できることが分かる。そして，同一荷重範囲であっても，対象単語の前１語～前３語の荷重を1.0，後ろ１語の荷重0.4，後ろ２語及び３語の荷重をそれぞれ 0.2 とすれば(表(b)右），連想３単語の関連度の総和が2.29に増大し，「木」の順位は下がるものの，関連度自体は大きくなり，「熊」の関連度ならびに順位も良くなることから，対象単語の位置を基準位置として，単語の位置に応じた荷重を掛けることが有効であるといえる。

表３　連想単語抽出

(a) 国語１年

対象単語:蕪菁(かぶ)			
1.0/1.0/対象単語/1.0/1.0		1.0/1.0/1.0/対象単語/1.0/1.0/1.0	
単語群	関連度	単語群	関連度
抜く	1.00	抜く	1.00
畝	1.00	畝	1.00
引く	0.86	引く	0.86
大きい	0.57	張る	0.86
成る	0.43	大きい	0.57
甘い	0.29	成る	0.43
未だ	0.29	甘い	0.43
種	0.14	未だ	0.29
蒔く	0.14	種	0.14
張る	0.14	蒔く	0.14
する	0.14	婆	0.14
ー	ー	する	0.14

(b) 国語６年

対象単語:森			
1.0/1.0/1.0/対象単語/1.0/1.0/1.0		1.0/1.0/1.0/対象単語/0.4/0.2/0.2	
単語群	関連度	単語群	関連度
見る	1.00	見る	1.00
中	1.00	目	1.00
僕	0.71	僕	1.00
木	0.57	成る	0.94
入る	0.57	中	0.82
目	0.57	木	0.65
成る	0.57	熊	0.65
熊	0.43	流す	0.65
流す	0.43	間	0.65
間	0.43	道	0.59
上がる	0.43	上がる	0.41

５．むすび

単語間の共起関係に注目し，対象とする単語の位置を基準位置として，その前後に出現する単語を記憶し，かつ出現した位置に応じた荷重を与え，それを累積する事により単語間の関連度を抽出する手法を小学校の国語の教科書に適用することにより，その有効性を検証した。

国語１年の教科書に適用した結果，出現した単語の位置に応じた重み付けより，荷重範囲の方がより重要で，対象単語の前後２単語の範囲内の単語を蓄積すれば良いことが明らかとなった。一方，国語６年の教科書では，適切な荷重範囲に加えて，出現した単語の位置に応じた重み付けを行うことにより，関連度が増大することが明らかとなった。

これらの結果は，単語意味理解のための連想記憶は文章中の近い位置に在る単語の並びから得られることを示しており，連続音声中の単語認識に不可欠な出現単語の予測がシステム的に行なえることが明らかとなった。

ところで，国語１年の結果は，小学校１年生の段階で取り扱う文章は，主格・目的格が近接した簡潔な文章であることから当然の結果であるといえるが，このことは，小学校低学年を対象とした書籍を大量に処理すれば単語間の共起関係のみから，一般的な連想単語のデータベース化が可能となることを示唆しており，現在，他の教科の教科書に適用して，その有効性を検討中である。

文　献

[1]木村晋太：“音響セグメントネットワークを用いた大語彙音声認識”，電子情報通信学会論文誌，J71D-II，3，pp. 475-482(1994).
[2]古山純子，小林哲則：“部分隠れマルコフモデルによる単語音声認識”，電子情報通信学会論文誌，J83-D-II，11，pp. 2379-2387(2000).
[3]中川聖一：“確率モデルによる音声認識”，電子情報通信学会(1988).
[4]伊藤彰則，牧野正三：“拡張 RHA 法による連続音声認識のための単語予備選択”，電子情報通信学会論文誌，J78-D-2，3，pp. 400-408(1995).
[5]宮地裕他編：“こくご一(上)”，光村図書(2005).
[6]宮地裕他編：“こくご一(下)”，光村図書(2005).
[7]宮地裕他編：“国語六(上)”，光村図書(2005).
[8]宮地裕他編：“国語六(下)”，光村図書(2005).