自然言語処理 80－2
（1990．11．22）

表層より得られる単語共起関係とその評価

潤潟謙一† 荒木健治‡ 宮永喜一† 栃内香次†

† 北海道大学　‡ 北海学園大学

本報告では、表層より得られる共起出現単語の有効性について述べる。音声入力によるテキストプロセッシングでは、音声認識段階における曖昧さのために多数の候補文字が出現し、これらの組合わせからなる単語候補の個数は極めて多くなる。この問題に対して本稿では、単語間の共起関係に着目し、候補に重みを付加することで効果を上げる手法を提案し、この手法の有効性を確認するために行なった実験と、その結果について報告する。さらに本手法を実際にシステムに適用した結果、85%の音声認識データに対して、共起単語は97%程度の精度で当てはめられており、その有効性を示すことができた。

The Estimation of Word Co-occurrence Relationship from The Strings of Character Level

Ken-ichi MAGATA† Kenji ARAKI‡ Yoshikazu MIYANAGA† Koji TOCHINAI†

† Hokkaido University
N13-W8,Kita-ku,Sapporo 060,Japan

‡ Hokkai-Gakuen University
S6-W11,Chuo-ku,Sapporo 064,Japan

This paper describes the effect of co-occurrence word from the strings of character level. In Japanese text processing by speech input, there are much possibilities of word matching, because the result of speech recognition is ambiguous. In this paper, we focus our attention to the word co-occurrence relationship, and we report the method of matching the correct word using this co-occurrence relationship and the result of the experiment show us that the proposed method is effective. In the practical experiment, when the correct rate of speech recognition is 85%, it is shown that the correct rate of the co-occurrence word matching is about 97%.

1．はじめに

計算機に日本語を入力するのに音声を用いることは、人間にとって最も簡便で、また最も自然な方法であるといえる。ところが、音声認識がもつ曖昧さのために、完全に正しい結果を得るのには強い拘束力のある文法や、それを処理する高度のシステムが必要である[1),2)]。このような言語情報を用意することは困難であるため、扱うタスクは小規模なものとなってしまう。

先に我々は、大規模な文書を対象に、音声入力において発生する誤りを正しい文字列に復元するために、統計的手法を用いて段階的に自動訂正する多段階分割復元法を提案した[3)]。しかしながら、この手法においても、多数の単語候補の中から1つを決定するには相当な処理時間を要する。この復元システムは段階的に処理するため、精度や処理時間は、上位でいかに多数の正しい復元が出来るかが非常に大きな影響を及ぼすといえる。

この問題に対して、我々は表層より得られる単語共起関係に着目した。共起関係を用いた日本語文解析[4),5)]に関しては種々検討されているが、本稿では表層に出現する単語間の共起関係を用いて、これにより単語候補に重みを付加することで処理時間の短縮と精度の向上を行なう手法を提案する。

ここでは、この手法の有効性を確認するために行なった実験とその結果について述べ、続いて実際に共起情報を多段階分割復元法に適用した実験について報告する。

2．キーワード共起単語と意味的共起単語との関係

2．1　共起出現単語

人間が文章を読む場合、文脈などを考慮してその文における核となる単語を認識して、検索空間を絞ることにより次にくる単語の範囲をある程度予測しながら読み進めているものと考えられる。本研究における復元処理においても、文章中の核となる単語が当てはめられた後、同一文に出現している他の単語は核となる単語に意味的に関わっていて出現しているのではないかと予測して、共起しやすい単語群を当てはめ候補とすることが有効であると考えられる。本稿では、このようにある単語になんらかの関係があって出現したものを共起出現単語、あるいは共起単語と呼ぶことにする。

2．2　実験手順

前述のような文章においての核となる単語を、キーワードと呼ぶことにする。このキーワードと同一文中に出現した単語が、人間の考える意味的に関連のある単語とどの程度一致するかを実験により調べた。ここでは、前者をキーワード共起単語、後者を意味的共起単語と名付ける。

実験資料として、今回は表1に示す情報処理に関する学術論文10編を対象とした。これらを主に自立語、付属語という分割を中心に形態素解析を行ない、その資料をもとに両者の関係を調べた。

実験の手順は以下の通りである。

（1）キーワードの決定
（2）キーワード共起単語と意味的共起単語の抽出
（3）結果評価

以下、各々の手順にそって述べる。

2．3　キーワードの決定

前述のように、キーワードはその文章において中心的な単語の一つである。ここでは、各々の実験資料において出現頻度が高く、その論文特有の語であるとみなすことができる自立語を、キーワードとする。キーワードと見なせる単語はその論文に関して何語か存在するが、その中から任意に1語を選択した。実験を行なう際に決定したそれぞれ論文のキーワード（KW）を、表1に示す。

2．4　キーワード共起単語と意味的共起単語の抽出

キーワード共起出現単語とは、キーワードと同一文中に出現する単語のことである。抽出する単語は自立語のみとした。ただしここで括弧などの記号は自立語扱いにしているが、共起出現単語としての抽出は行わない。また、「～いる」「～ある」「～する」「～ない」などの自立語は付属語扱いにしている。これは、これらの語が付属語と

同様に単語に付随して頻繁に出現するからである。

また、これと同時にキーワードと意味的に関係をもつ自立語、すなわち意味的共起出現単語の抽出を行なった。意味的に関係がある語として、ここではキーワードと次の関係にあるもののみを抽出した。

・直接係り受け関係にある自立語
・キーワードと複合名詞を構成している自立語

実験に使用した10編の文献のうち、表１における文献８の１文を例に説明する。
まず、キーワードを含む文を抽出する。文献８の場合、キーワード「故障」を含んでいる文を抽出することになる。

【抽出文章】

「ＶＬＳＩの故障検査を行う場合，回路を分割して試験したり，回路中の信号をモニタすることが困難であるために，大規模な回路では故障試験に膨大な入力系列（検査系列）を必要とする。」

この文章を形態素解析し、分割する。

【形態素解析】

「/ＶＬＳＩ/の/故障//検査/を/行/$う$/場合/，/回路/を/分割/して/試験/した/り/，/回路/<中>/の/信号/を/モニタ//する//こと/が/困難/で/ある//ため/に/，[大]/規模/な/回路/では/故障//試験/に/膨大/な/入力//系列//（//検査//系列//）/を/必要/と/する/。」

//：自立語　/：付属語　$$：活用語尾
[]：接頭語　<>：接尾語

キーワードを含んだこの文章に出現する自立語が共起出現単語として抽出される。

【キーワード共起単語】

「回路」(3)、「試験」「検査」「系列」(2)、「ＶＬＳＩ」「行」「場合」「分割」「信号」「モニタ」「こと」「困難」「ため」「規模」「膨大」「入力」「必要」(1)

() 内は頻度を示す。

この自立語のうち、意味的共起出現単語は上の形態素分割の下線部の自立語が相当し、以下のようになる。

【意味的共起単語】

「ＶＬＳＩ」「検査」「行」「試験」「必要」(1)

() 内は頻度を示す。

２．５　キーワード共起単語と意味的共起単語との関係

このようにして、キーワードを含む同一文中に出現する自立語と、その中でキーワードと意味的に関係があるとみなせるものを頻度により順位を付け、比較する。

以下表２に、文献８の場合を例にして両者をそれぞれ頻度順に並べた結果を示す。なお、() 内は出現頻度である。

No.	著者	題名	出典	文字数	KW
1	荒木他	音声認識における誤りの段階的復元手法	北学園大工研報16号	7,840	「辞書」
2	荒木他	日本語文の表層レベルでの浅い解析	北学園大工研報16号	6,232	「解析」
3	荒木他	付属語パターンを用いたルールによる表層レベルでの係り受け関係の解析	情処研報ＮＬ68-6	6,455	「係り受け」
4	荒木他	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	情処学論 Vol.30 No.2	10,036	「認識」
5	前島他	高集積マイクロコンピュータに適したマイクロプログラム制御方式	情処学論 Vol.23 No.1	10,594	「マイクロ」
6	任他	日中機械翻訳における慣用表現のエンコードおよびデコード	信学技報ＮＬＣ89-30	5,599	「表現」
7	任他	コード方式日中機械翻訳の実験システムＪＣＭＴの概要	情処研報ＮＬ72-7	6,079	「コード」
8	任他	グラフ故障シミュレーションアルゴリズムに関する研究	信学技報ＶＬＤ88-63	5,754	「故障」
9	佐藤他	意味記述を考慮したパーサについて	信学技報ＮＬＣ89-31	5,344	「構文」
10	荒木他	帰納的学習による文字列中の語の認識	信学技報ＮＬＣ89-32	10,326	「認識」
			文字数合計	74,259	

表1　実験に使用した資料

キーワードを同一文中に出現する自立語	キーワードと意味的に関係のある自立語
「回路」（34）	「節点」（15）
「節点」（33）	「ｼﾐｭﾚｰｼｮﾝ」（15）
「ｼﾐｭﾚｰｼｮﾝ」（26）	「ﾘｽﾄ」（15）
「ﾘｽﾄ」（15）	「可検」（14）
「ｸﾞﾗﾌ」（15）	「記憶」（ 8）
「可検」（14）	「ｸﾞﾗﾌ」（ 8）
「処理」（12）	「伝播」（ 6）
「記憶」（11）	「処理」（ 4）
「こと」（10）	「方法」（ 3）
「方法」（ 9）	「部分」（ 3）
：	：
：	：

表２．ＫＷ共起単語と意味的共起単語

このようにして得られた結果を上位より、10位、20位、30位、40位と抽出し、それぞれにおける両者の一致度を比較した。表２の例では、頻度上位10単語のうち一致しているものは８単語で、その一致度は80%ということになる。

文献10編での平均は表３の通りであった。

	上位10位	20位	30位	40位
一致度	60%	66%	65%	65%

表３．ＫＷ共起単語と意味的共起単語の一致度（句点単位）

これから、同一文中でキーワードと意味的に共起している単語と、文中に高頻度で出現する単語とには、ある程度の一致があるといえる。

２．６　意味的共起単語の抽出方法

意味的に共起していると思われる単語を、計算機で正確に自動抽出することは困難である。一方、意味的共起単語としている直接の係り受け関係にあるものは、キーワードの前後数語の範囲内にあることが多い。そこで、表層で得ることのできる確実な情報である句読点を利用し、その間の文字列を抽出範囲として実際に人間が抽出した意味的共起単語との一致度を再び調べた。結果は次の通りであった。

	上位10位	20位	30位	40位
一致度	73%	78%	77%	73%

表４．ＫＷ共起単語と意味的共起単語の一致度（句読点範囲）

例に示した資料（文献８）の場合、上位30位において抽出範囲を同一文中とすると、両者の一致度は50%であった。ところが範囲を句読点単位にすると、一致度は77%に上昇している。しかし同じ句読点範囲であっても40位までとると率が73%に下がることから、上位30位くらいまでに大部分の意味的共起単語は含まれていると考えられる。他の資料についても傾向は同じで、これにより計算機で抽出した表層上より得られる共起単語を意味的共起単語とみなすことが可能であると考えられる。

３．多段階分割復元法

３．１　概要

人間は誤りを含む文字列をみたとき、最初に手掛かりとなる部分を見出して誤りの訂正を試み、前後のつながりに矛盾を生じないようであれば同様の操作をすすめていき、原文を正しく読み取っていくと考えることができる。多段階分割復元法は、誤りを含む文字列の複数の単語候補の中から、前後関係を考慮しながら、統計頻度に基づいた確実度の高いものより順次段階的に単語を決定し、原文を復元していくというものである。本手法は、上位の段階において単語を決定することにより、その後における段階での単語候補が急激に減少するため、誤りを含む文字列を対象とした音声入力テキストプロセッシングにおける単語分割に有効である。

３．２　システム構成

システムは、音声認識部と復元処理部とからな

る。音声認識部では、市販の単音節認識装置によって音声認識データを得る。復元処理を行なうプログラムは、北海道大学大型計算機センターのHITAC-M682H上に作成されている。使用言語はＰＬ/Ｉである。なお、音声認識部と復元処理部は現在オフラインとなっている。

３．３　処理手順

復元処理の過程を図１に示す。

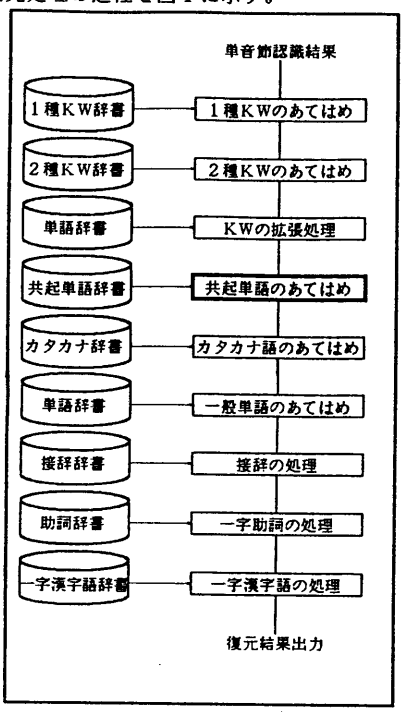

図１．復元処理の過程

まず単音節認識装置により得られたローマ字列の結果を復元処理部におくる。そして復元処理部においてそれを段階的に単語列に分割復元し、その結果を漢字かな混じり文で出力する。

最初の段階で使用される確実度の高い単語を、キーワードと呼ぶ。キーワードはその中で、さらに条件によって、１種キーワードと２種キーワードとに階層化されている。つぎに、キーワードを前後に拡張した文字列について検討し、以下、カタカナ語、一般単語、接辞、助詞、一字漢字語のあてはめ、復元処理が行なわれる。

今回ここでは、キーワードに関する分割復元が行われた後、キーワード共起単語による分割復元を試み、その結果について考察する。以下、この共起単語による復元処理段階を、共起処理と呼ぶことにする。

４．性能評価実験

４．１　実験手順

性能評価実験では、音声認識部で得られた音声認識データを復元処理部によって分割復元し、その結果を評価する。本実験では、共起処理を行なう場合とそうでない場合についてを同じ条件で行ない、それぞれの結果について比較検討する。

ここで，実験に用いた復元用辞書と音声認識データについて説明する。

復元処理部で用いる各辞書は表５に示される情報に関する14編の論文により作成されている。共起辞書は、キーワード辞書（１種、２種）にある

No.	著者	論文名	情処学論文誌	文字数
1	前島他	高集積マイクロコンピュータに適したマイクロプログラム制御方式	Vol.23 No.1	10,594
2	山本他	ＣＯＢＯＬマシンとその設計思想－ハードウェア構成について	Vol.23 No.1	8,451
3	松山他	フーリエ変換を用いたテクスチュアの構造解析	Vol.23 No.2	7,723
4	木村	日本語文入力用カタカナ語検出規則とオンライン国語辞典の一分析	Vol.23 No.2	8,468
5	有田	インテリジェント・コンソール－ＯＳの機能拡張の一方法	Vol.23 No.3	9,343
6	田村他	ポータブル画像処理ソフトウェア・パッケージＳＰＩＤＥＲの開発	Vol.23 No.3	9,624
7	高藤他	グラフィック・ディスプレイ・ターミナルのための端末作画システム	Vol.23 No.4	6,423
8	長岡他	オペレーティング・システムのファームウェア化対象選定法	Vol.23 No.4	7,594
9	酒井他	プログラム階層構造の生成、処理、文書化能力を有するテキスト・エディタ	Vol.23 No.4	7,916
10	中所他	パステストに本質的な分岐に着目した網ら率尺度の提案	Vol.23 No.5	10,970
11	吉住他	計算機システムにおける性能管理の一方法とそれを用いた実験	Vol.23 No.6	9,123
12	寺田他	高速パケット電送路用前置処理装置の一構成法	Vol.23 No.6	8,401
13	有澤他	ソフトウェア生産過程の評価実験に関する考察	Vol.23 No.3	6,453
14	吉村他	文節数最小法を用いたべた書き日本語文の形態素解析	Vol.24 No.1	9,658
			文字数合計	120,741

表5　辞書作成に使用した資料

キーワード単語のうち自立語であるものを見出し語として、１語につき30語の共起単語を登録している。共起単語は２．の実験結果を受けて句読点範囲から抽出しており、頻度順に上位より並べられている。同頻度のものは、文字数の多いものが上位になっている。

なお、数字、記号、アルファベットなどは辞書には登録していない。

表６に作成された各辞書の登録語数を示す。

辞書名	語数
１種ＫＷ辞書	74
２種ＫＷ辞書	350
共起単語辞書	348
カタカナ語辞書	200
単語辞書	3,262
接辞辞書	37
助詞辞書	15
一字漢字語辞書	116

表６　辞書登録語数

音声入力用の資料は、表５中の文献１の第１章、第２章（総音節数 2,953）である。これを、各章それぞれ前半後半に分け音声認識部で認識させ、その結果を音声認識データとした。

今回は音声認識率と復元結果を比較するため、得られた音声認識結果（正音声認識率　67%）の認識誤りをランダム訂正し、正音声認識率70%、75%、80%、85%、90%の５種類のデータを用意した。以上の音声データと復元辞書を用いて性能評価実験を行なった。

４．２　復元例

音声認識率70%のデータの１文を例に復元結果を得るまでの過程を示す。

【資料】

これによって解決することが論理設計の単純化の見地からも理想である。

【音声認識結果】

こ？によずて？いへ？することがのんりけぜけいのたんぎゆんがのけ？じがらもびそうである。

？：未認識音節

【分割復元結果】

!ここ!(に)@よって@!解決!@する@@こと@(が)[論理]@設計@(の)[単純]!かの!!見地!@から@(も)!理想!@である@。

@@：１種ＫＷ　[]：２種ＫＷ　!!：共起単語

!!：一般単語　()：助詞

４．３　実験結果

こうして得られた復元結果の復元率、および、処理時間を項目をわけて示す。

（１）復元率

復元率とは以下の式により算出されるものである。

$$復元率=\frac{復元結果における正音節数}{総音節数}\times 100$$

復元処理によって得られた結果の復元率を、表６、表７に示す。表は、復元処理における共起処理の有無によって分けている。表中の総合復元率とは、音声認識データ（ 2,953 音節 ）を復元したもののうち、正しい音節が占める割合を表わす。また、ＫＷ復元率とはその中で、キーワードによる分割復元によって、当てはめられた単語の復元率を表わす。共起復元率も同様である。

音声認識率	総合復元率	ＫＷ復元率	共起復元率
70%	86.1%	86.1%	—
75%	87.4%	87.4%	—
80%	89.9%	89.6%	—
85%	92.5%	92.3%	—
90%	95.1%	94.6%	—

表７．共起処理を行なわない場合の復元率

音声認識率	総合復元率	ＫＷ復元率	共起復元率
70%	85.4%	86.1%	82.7%
75%	87.0%	87.4%	87.1%
80%	89.6%	89.6%	89.0%
85%	92.6%	92.3%	96.9%
90%	95.0%	94.6%	97.1%

表８．共起処理を行なう場合の復元率

（２）処理時間

共起処理の有無による処理時間の変化についてその結果を表９に示す。ここでは、入力資料の第１章前半に相当する音声認識データに対する復元処理時間を示す。短縮率とは、共起処理の行なわない場合の処理時間を 100%としたときの、それを行なった場合の処理時間の割合を表わす。

定義式を次式に示す。

$$\text{短縮率}(\%) = \frac{\text{共起復元有の時の処理時間}}{\text{共起復元無の時の処理時間}} \times 100$$

音声認識率	処理時間 共起　無	処理時間 共起　有	短縮率
70%	4'58"36	3'39"16	73.5%
75%	4'11"66	3'01"84	72.3%
80%	4'19"34	3'33"05	82.2%
85%	3'20"92	2'49"47	84.3%
90%	2'31"28	1'57"84	77.9%

表９．復元処理時間

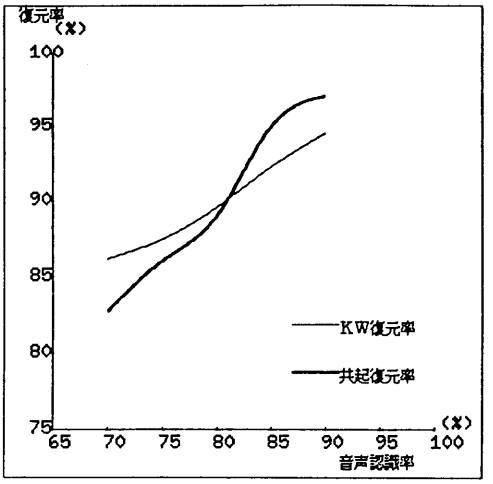

図２．ＫＷ復元率と共起復元率の関係

４．４　考察

（１）復元率

共起単語による復元率は、音声認識率が80%以上あれば90%を超えることが分かる。また、音声認識率が85%になると共起単語は約97%の精度で当てはめられており、その有効性を確認できた。

共起単語による分割復元はその性質上、キーワードの当てはめの結果に大きな影響を受けるものと思われる。そのことを表わすキーワード復元率と共起復元率の関係を図２に示す。

キーワード復元率は、音声認識率に対してほぼリニアに上昇しているのに比べて、共起復元率はキーワード復元率が90%に近づくと、急激に上昇することが分かる。キーワード復元率が92%をこえると、共起復元率は97%となり復元精度は非常に高くなっている。したがって、キーワードの復元率が90%程度あれば、共起段階は高い精度で復元処理を行なうことができると考えられる。

逆に正しいキーワードが用意されないと、誤った共起単語が多数用意されることになる。総合復元率が共起処理では大幅に改善されず、１%未満のゆれしか見せていないのは、このようなことが一因にあると考えられる。

今回の実験では、共起復元処理段階を加えるだけで、他の処理段階には変更を加えていない。このようなことから、多段階分割復元法で共起処理を考える場合、当てはまったキーワードの結果だけを受けるのではなく、たとえば、候補として上がったけれども条件により当てはめられなかったキーワードの検討など、キーワードの処理アルゴリズムも含めて考える必要があると思われる。

（２）処理時間

処理時間に関しては、共起処理を加えることにより約15～30%の短縮がなされている。これは処理語数が多い、そして最も処理時間のかかる一般単語による分割復元の前に、共起処理により、一般単語による分割復元の対象範囲を減少させることができるからである。出力結果をみても、従来一般単語として当てはめられていたものが共起段階で当てはめられている例が最も多かった。復元アルゴリズムの性格上、上位段階で当てはめられれば下位での候補は激減し、そのことにより単語検索時間が短縮されたためと考えられる。

5．おわりに

以上本稿では、ある文章において中心的な単語（キーワード）の含まれている文に高頻度で出現する語（共起単語）と当該キーワードとの意味的関係について述べ、次いで意味的共起単語を表層よりある範囲で推定し自動抽出した結果と人間が考えた結果との関係について述べた。そして、そのようにして得られた共起単語のデータを用いて、実際に復元システム上にこの処理段階を組み込んで実験を行ない、その結果について報告した。

今後は、表層上からのより有効な共起情報抽出方法や、共起復元処理を生かすためのキーワードの利用方法、及びそれを用いた復元方法の検討などを行なう予定である。

参考文献

1）堀正洋、辻野克彦、溝口理一郎、角所収：「音声理解システムSPURT-1 －動的クラスタリング方式と文節発声による性能評価－」電子情報通信学会論文誌 J72-DII No.8 pp.1291-1298（1989）

2）劉学敏、西田豊明、堂下修司：「統語パーサによる統合的自然言語解析」情報処理学会論文誌 Vol.31 No.2 pp.1293-1301（1990）

3）荒木健治、宮永喜一、栃内香次：「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 No.2 pp.169-178（1989）

4）高橋直人、板橋秀一：「単語共起頻度を利用した形態素解析」情報処理学会自然言語処理研究報告69-5（1988）

5）松川智義、長尾真、中村順一：「共起関係に注目したＤＭ分解と確率的推定による単語のクラスタリング」情報処理学会自然言語処理研究報告72-8（1989）

6）潤潟謙一、荒木健治、宮永喜一、栃内香次：「表層からの単語共起関係の推定」1990年電子情報通信学会春季全国大会講演論文集D-88 pp.6-88（1990）