日本語文書処理 21－2
（1985. 5. 8）

連語解析を用いたべた書きかな漢字変換

本間 茂　山階 正樹　小橋 史彦
（NTT　横須賀電気通信研究所）

1．まえがき

かな漢字変換方式は、ワード・プロセッサ等における日本語入力法の主流となっており、その入力方式は、図1に示すように、字単位入力方式、漢字指定入力方式、文節指定入力方式、べた書き入力方式へと発展してきている。当初は、図1に示すように漢字を一文字毎に変換する字単位入力方式であったため、入力に要するキータッチ数が非常に多く入力速度の低いものであった。その後、入力規則が緩和された漢字指定入力方式、文節指定入力方式が実用化され、キータッチ数から見た入力速度は大幅に向上した。

しかし、文節指定入力では、文節に区切るための煩雑なキー操作が必要であること、文節に区切るための文法知識が必要であること等の問題点がある。このため、これらの問題が解決でき、オペレータにとって自然で、しかも高速に入力が可能なべた書き入力方式の実現が要望されている。

べた書きされたかな文字列は、分かち書きによる曖昧さがあり、さらに、文節入力の場合と同様に同音語による曖昧さがある。このため、入力文に対して、膨大な数の解釈が得られる場合が頻繁に発生する。

こうした問題を解決するために、図2に示すような数々の方法が研究されてきた[1),~,13)]。これらを大別すると、以下の5つに分類することができる。

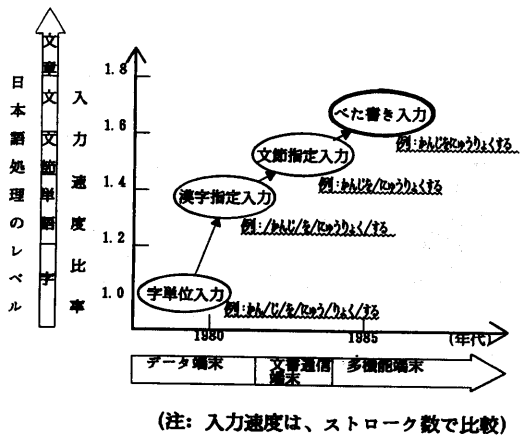
(1) 文字列中に存在する最長の文節形を1つの文節と見なし、辞書との照合、接続検定を繰り返すことによって解析を進め、接続に失敗した場合にバックトラックする方法。[1)]

(2) 2文節としての解釈が最長のものを採用する方法。[2)]

(3) 助詞や4音節単語等、文節区切りの手掛かりにより前処理を行ってから単語辞書を検索する方法。[3),~,6)]

(4) 総当り的に解釈を求め、文節数等の指標を基に候補を選択する方法。[7),8)]

(5) 構文解析（格文法）によって同音異義語を選択する方法。[9),~,13)]



（注：入力速度は、ストローク数で比較）

図1 かな漢字変換技術の発展動向

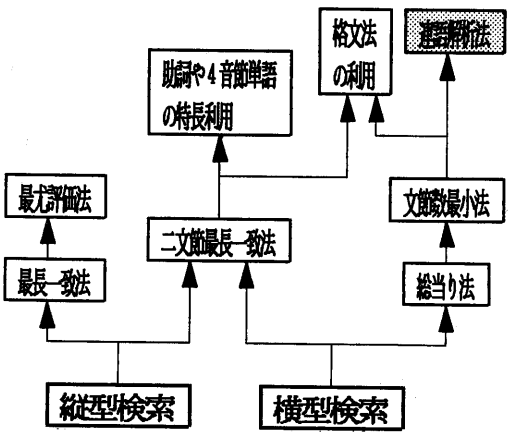


図2 べた書きかな漢字変換の研究

(1)、(2)、(3)の場合、文節区切り位置を確定しながら左から右に解析を進めるため、途中で誤った単語選択を行っても、文法上接続可能であれば、文末までの解釈が行われてしまう場合がある。(4)の場合は、最終的には無駄となる辞書検索が多く、また、最終候補を選択する指標の設定方法により正解を漏らす場合がある。(5)の場合、十分な性能を得るためには、格解析用の辞書を充実させる必要がある。

本研究では、文解釈候補の中に正解が含まれる可能性が最も高いという観点から(4)の総当り的な解釈方法を採用し、この手法で問題となる多数の候補が出力される点を解決するため、連語解析法を検討した。本報告では、連語解析を用いたべた書きかな文字列の変換アルゴリズムについて述べるとともに、その評価実験の結果を述べる。

2. 連語解析

従来、日本語処理では、品詞、使用頻度等、単語単体での属性を記述した辞書を使用する場合がほとんどであった。しかし、処理の高精度化を図るためには、単語単体での属性ばかりでなく、単語間の関係を用いることが必要であると考えられる。この情報の一つに連語情報（2つ以上の単語が結合して、まとまった意味をなすもの）があげられる。例えば、図3左側に示すように、「花」という単語は、「美しい」、「咲く」等の単語と強い結び付をもって文章中に出現する。「人」という単語についても「孤独な」、「会う」等の単語との間で同様のことがいえる。このような連語情報をあらかじめ用意しておくことによって、文に複数の解釈が生じた場合に、連語情報を用いることによって正しい候補の選択が行える。図3右側に示したように、「はながさきます」、「ひとにあう」には、それぞれ2通りの解釈がえられる。そこで、連語情報を参照すれば、「花が咲きます」、「人に会う」の方を正しく選択することができる。

図4はべた書き入力の処理の流れを示しており、連語解析は、文法接続検定が終了し、文解釈候補が得られた時点で行う。連語解析を行っても、同音語が原因で文解釈候補が絞りきれないものについては、単語の使用頻度等をもとに最終決定を行う。

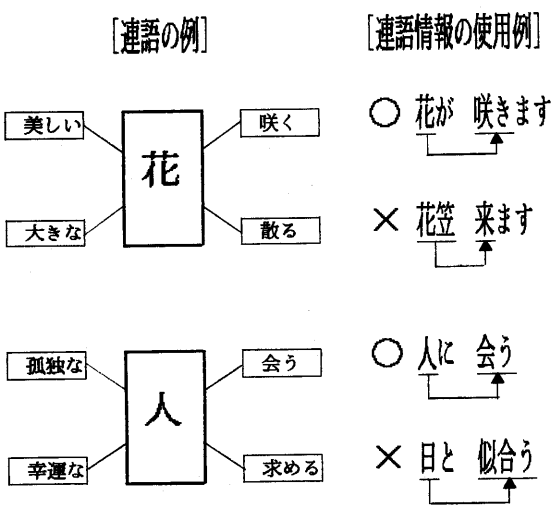


図3　連語を用いた解析

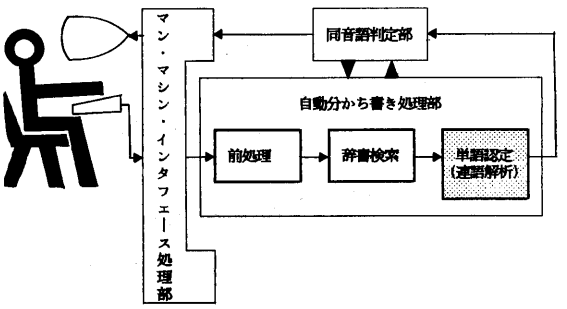


図4　べた書き入力の処理の流れ

3．連語辞書

連語情報を人手で収集しようとすると、膨大な工数が必要であるばかりでなく、網羅的な収集は困難であると考えられる。そのため、ここでは実用的な連語辞書の構築を目指し、連語情報を自動抽出した。以下に自動抽出の手法を述べる。

3．1　ソース・データ

各単語の典型的な使われ方が比較的短い文で記載されている国語辞典等の用例に着目し、これらの用例約30万件をファイル化して自動抽出用のソースデータとした。

3．2　自動抽出の流れ

図5は連語抽出の流れを示している。

①ソース・データを形態素解析し、各単語の品詞、活用形を認定する。

②文節の形態的な性質に着目して係り受け解析を行い、係り受け関係を持ち得る文節を認定する。

③解析の結果、係り受け関係に曖昧さがある場合、文節間の距離関係等を用いて、係り受け関係を持つ確率が高い文節を選択する。

④名詞連続、あるいは、名詞と接辞の連続を複合語と認定する。

⑤係り受け関係を持つと認定された文節から、連語情報を抽出する。なお、係り文節が自立語と接辞の組み合せ以外の複合語の場合、末尾の単語が係りの機能を持つと考える。受け文節が複合語の場合［情報の／処理／技術：新薬の／治療／効果］には、複合語を構成するどの単語が受け機能を持つかを認定することは困難なため、受け側では、複合語形で連語情報を抽出する。

3．3　意味分類番号

以上の処理によって得られた連語情報を図6に示す。この連語情報をそのまま辞書化して連語処理に用いることも可能であるが、見出しが一致しなければ連語として抽出できない。このため、連語解析の効果を高めるには、膨大な件数を辞書に登録する必要がある。もし、そのような辞書を用いると、容量が大きくなるばかりでなく、検索に多くの時間を要する問題がある。

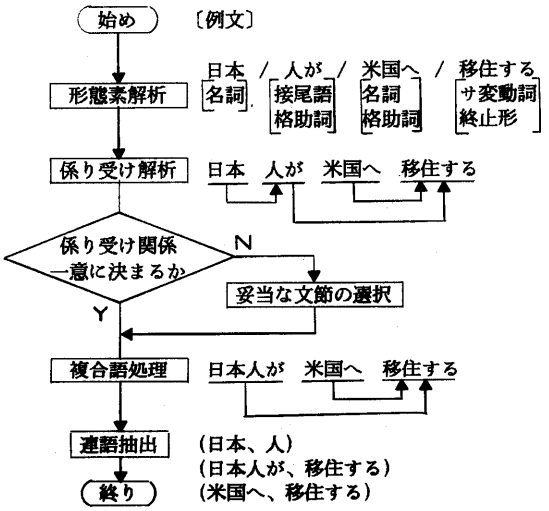


図5　連語抽出の手順

係り単語			助詞	受け単語		
見出し	意味分類	意味分類番号		見出し	意味分類	意味分類番号
藍	染料	917	で	染める	色付	256
哀愁	悲喜	493	を	帯びる	包含	268
愛情	愛憎	481	を	持つ	存廃	285
合図	合図	334	の	口笛	音楽	870
生地	実質	130	を	染める	色付	256
花	花	056	が	咲く	花	056
人	自他	505	に	合う	出会	781
…	…	…	…	…	…	…

図6　連語情報

一方、「藍で染める」と「顔料で着色する」のように、似た概念を表現しているものは、係り単語あるいは受け単語を、意味の類似した単語と置換えても、連語関係が保たれる場合が多い。そこで、本研究では、意味の類似した単語を1カテゴリとして扱うこととし、単語を意味分類番号[14]で代表させて表現した。意味分類、意味分類番号の具体例を図6の連語情報に示す。

意味分類番号は3桁の数字で上位の桁から、大分類、中分類、小分類に相当する。従って、カテゴリ数は1000である。これにより、前述の連語情報は、1000×1000のマトリクス上の、'1'、'0'で表現できる。図7に連語辞書の概念を示す。

3．4　連語の拡張解釈

上述のように意味分類番号で単語を代表させて表現すると、以下のような利点がある。図8中央下段に波線を付して示した、「藍で染める」がソースデータにあれば、染料(917)と、色付(256)の間の連語情報が登録される。すると、「顔料で着色する」、「晒し粉で漂白する」等は、連語として登録されていなくても、染料と色付の関係であるため連語として抽出可能となる。このように、ソース・データを拡張解釈することにより、実効上は多数の単語が登録されているのと同等の効果が得られる。

また、意味分類を中分類以上のレベルで見ることにより、連語の拡張解釈が可能である。図8中央中段の、「アルコールで溶かす」は、染料(917)の上位概念である薬品(91)と色付(256)の上位概念である変質(25)の間の関係であるため、連語として抽出できる。

しかし、拡張解釈を行うと、一つの意味分類が表現する範囲が広くなるため、連語としてふさわしくないものまで、連語として抽出する場合が発生する。特に、大分類の場合はほとんどのカテゴリ間に連語が成立しているので連語解析を行うための情報源とならない。実験では、意味分類番号を3桁（小分類）用いた場合、上位2桁（中分類）用いた場合についてその効果を検討した。

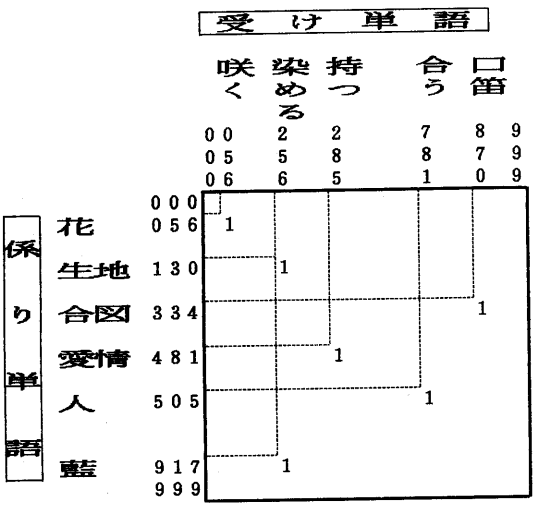


図7　連語辞書の概念図

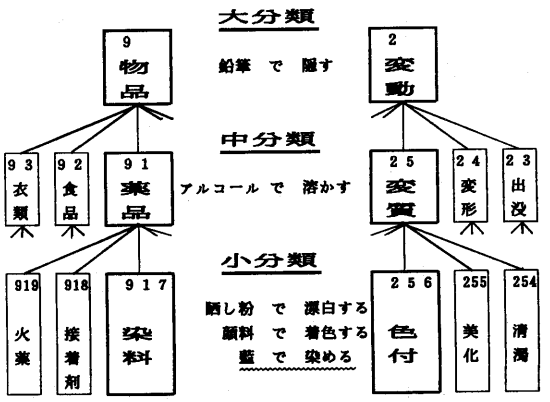


図8　連語の拡張解釈

4．実験方法

4．1　辞書

実験には以下の辞書を用いた。

①自立語辞書

一般語（複合語を含む）約6万5千語を収録している。収録情報は、かな見出し、漢字表記、品詞の他に、使用頻度、意味分類番号である。

②付属語辞書

助詞、助動詞等、合計 422語を収録している。

③接辞辞書

接頭語、接尾語、数詞を合計1085語収録している。

④文法辞書

自立語－付属語、付属語－付属語の接続規則を表形式で表現したものである。

⑤連語辞書

3章で述べた辞書内容を持ち、自立語辞書の意味分類番号によって参照する。

4．2　自動分かち書き処理

本方式の処理の流れを図9に示す。

①入力文字列に対して左から右へ順次、自立語、付属語、接辞の辞書検索を総当りで行う。

②部分解釈の組み合せの中から文法的に成立するものを選択する。

③文節数最小の候補を選択する。ここでは接辞も文節数1とする。

④単語数最小の候補を選択する。ただし、「自立語－付属語－自立語」を、「自立語－自立語」という複合語に解釈してしまうのを防ぐため、複合語を構成する単語パターンとして使用頻度の低いものが得られた場合は採用しない。また、「自立語－付属語」、「自立語－活用語尾」のいずれにも解釈できるものは、活用語尾も単語数の計数に加える。

⑤隣接する自立語（付属語は無視する）について連語関係の有無を調べ、各候補毎に連語数を計数する。

⑥連語抽出件数が最大の候補を最尤候補として出力する。ただし、拡張解釈を行うことにより、出現頻度が極端に低い単語同志の組み合せや、連語関係の希薄な単語の組み合せを、出現頻度が高く、しかも、関係が緊密な単語の組み合せと同等に扱うことによって誤りを生じる場合があるので、この誤りを防ぐために以下の規則を設けた。連語数最大の候補文の中で、連語関係が成立しているが、単語の頻度が極端に低い場合には、連語関係が成立していなくても、高頻度の単語を含む文解釈を候補に含める。

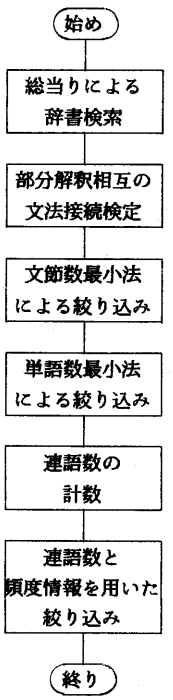


図9　処理の流れ

図10に連語解析の具体例を示す。3つの文解釈候補は、文節数最小で単語数最小の条件で選択されたものである。意味分類番号を上位2桁用いた場合は、3つの候補はいずれも連語が3件抽出され、全て最尤候補として残る。意味分類番号を3桁用いた場合は、上段の候補が連語抽出件数2件で、3つの候補の中で最大であるため、最尤候補となる。

5．実験結果

5．1　実験データ

入力データは、新聞社説10編から抽出した記号等を含まない句読点で区切られた単位の文字列、386件で文節数は、1411（文字数：約5500字）である。1文あたりの、文字列平均長は14文字で、平均 3.7文節からなっている。

5．2　実験結果

表1に本手法による文解釈候補の絞り込み効果と文節単位での誤り率（正解を漏らす率）を示す。

総当り的に出力した候補文の中から、文節数最小の候補のみを選択すると、平均文解釈候補数は、62.6であり、文節数により候補を絞り込む際の誤り率は、1.1%である。表2に示すように、接辞を文節数1としたことによる誤りが大半を占めた。

次に、単語数最小の候補を選択すると平均文解釈候補数は、16.6となる。単語数によって候補を絞り込む際の誤り率は、0.8%である。ここでの誤りは、表2に示したような「自立語－付属語」を「自立語」と解釈したこと等により生じている。

最後に、連語情報を用いた絞り込みでは、平均文解釈候補数は、

8.7　（意味分類番号上位2桁）
6.4　（意味分類番号　3桁）

となる。ここでの、誤り率は、

1.8%　（意味分類番号上位2桁）
1.6%　（意味分類番号　3桁）

であった。誤りは、表2に示したように、連語情報が辞書になかったために発生している場合が多い。

また、未登録語による誤りも1.84%生じている。

単語数	文解釈候補	連語抽出件数 意味分類番号 2桁	3桁
7	武器 禁輸 政策 を 堅持することは	3	2
7	武器 禁輸 製作 を 堅持することは	3	0
7	武器 禁輸 制作 を 堅持することは	3	1

：意味分類番号を3桁用いた場合の連語

：意味分類番号を上位2桁用いた場合の連語

図10　連語解析の具体例

表1　連語解析結果

処理段階		平均文解釈候補数	誤り率(%)（文節単位）
文節数最小		62.6	1.7
単語数最小		16.6	0.8
連語検出数最大（意味分類番号）	(上位2桁)	8.7	1.8
	(3桁)	6.4	1.6

入力文：新聞社説10編

表2　誤解析とその原因

原因	誤解析例	入力文例	件数
文節数最小	教科書の共通か	教科書の共通化	15
単語数最小	体制の元手	体制のもとで	11
未登録連語	弔旗の対応策	長期の対応策	23
未登録語	念書以来	年初以来	26

6．むすび

本研究では、べた書きかな文字列の解析において、総当り的手法によって出力した文解釈候補を、文節数最小法、単語数最小法によって絞り込み、さらに、連語情報と使用頻度情報を用いた解析によって、最尤候補文を得る方法について検討した。

その結果、誤り率の増加を、

1.8%　（意味分類番号上位2桁）

1.6%　（意味分類番号　　3桁）

と、低く押さえた上で、平均候補数を、文節数最小法の

約1/7（意味分類番号上位2桁）

約1/10（意味分類番号　　3桁）

に絞り込むことができた。

今後は、以下の検討を行う予定である。

①現在、隣接した文節のみで行っている連語の検査を、離れた文節においても可能とするための構文解析法の検討。

②未登録語を抽出し、その影響を他文節に及ぼさない処理方式の検討。

謝辞

日頃御指導頂く、山崎宅内機器研究部長、小森入力装置研究室長に深謝する。

参考文献

1）内田、杉山、「自由入力形式のカナ漢字変換」、情報処理学会自然言語処理研究会資料、27-3、pp.1-8、1982年3月。

2）牧野、木澤、「べた書き文の分かち書きと仮名漢字変換」、情報処理学会論文誌、vol.20、no.4、pp.337-345、1979年7月。

3）赤川、牧野、「べた書きカナ文分かち書きについて」、情報処理学会第22回（昭和56年前期）全国大会講演論文集、論文番号1I-5、pp.857-858、1981年3月。

4）赤川、牧野、「仮名漢字変換システム「ＢＥＴＡ」」、情報処理学会第24回（昭和57年前期）全国大会講演論文集、論文番号5G-3、pp.995-996、1982年3月。

5）坂本、平塚、椎野、「前処理を導入したカナ漢字変換」、情報処理学会第26回（昭和58年前期）全国大会講演論文集、論文番号2H-2、pp.1167-1168、1983年3月。

6）館林、中馬、杉村、小林、向井、滝口、「自由文入力・仮名漢字変換方式」、情報処理学会第26回（昭和58年前期）全国大会講演論文集、論文番号2H-1、pp.1165-1166、1983年3月。

7）吉村、日高、吉田、「日本語文の形態素解析における最長一致法と文節数最小法について」、情報処理学会自然言語処理研究会資料、30-7、pp.1-6、1982年3月。

8）斎藤、河田、武田、矢内、山中、「かな漢字変換方式について」、情報処理学会第25回（昭和57年後期）全国大会講演論文集、論文番号6J-1、pp.1125-1126、1982年10月。

9）牧野、木澤、「べた書き文の仮名漢字変換システムとその同音語処理」、情報処理学会論文誌、vol.22、no.1、pp.59-67、1981年1月。

10）武市、阿部、大島、湯浦、「構文意味解析を適用したべた書き文仮名漢字変換システムの開発」、情報処理学会第28回（昭和59年前期）全国大会講演論文集、論文番号4M-5、pp.1325-1326、1984年3月。

11）湯浦、阿部、武市、「文節構成に関する経験的規則を用いた仮名べた書き文の形態素解析」、情報処理学会第28回（昭和59年前期）全

国大会講演論文集、論文番号4M-6
pp.1237-1238、1984年3月．
12) 大島、阿部、湯浦、武市、中島、「格文法による仮名漢字変換のための構文意味解析」、情報処理学会第28回（昭和59年前期）全国大会講演論文集、論文番号4M-7、
pp.1239-1240、1984年3月．
13) 阿部、湯浦、大島、武市、「べた書き文仮名漢字変換における最適文選択法」、情報処理学会第28回（昭和59年前期）全国大会講演論文集、論文番号4M-8、
pp.1241-1242、1984年3月．
14) 大野、浜西、「類語新辞典」、角川書店、1981年．