

係り受け解析のための辞書の構成とその学習機能†

白井克彦^{††} 林良彦^{††*}
平田裕一^{††**} 久保田淳市^{††***}

日本語文の構造は、文の構文的・意味的な構造に大きく関わっている。よって、日本語文の処理を行う場合、その係り受け構造を明らかにすることが重要である。計算機により日本語文の係り受け解析を行う研究は広く行われているが、そのために必要な知識の獲得・構造化を半自動的にかつ適応的に行うための研究は少ないように思われる。本論文では、テキストデータに対する分析から直接的に以後の成長の核となる初期辞書データベースを構成する方法、および成長のための学習機能について検討した。この辞書データベース中では、単語はその係り受け特性に基づいてクラスタリングされ、分析対象としたテキスト中の単語間の係り受け関係は、クラスタ間の係り受け可能関係として抽象化されて記述される。本辞書データベースは、実験文解析システム ESSAY による文解析に適用され、その評価を受ける。さらに、解析が不成功である文において、そのネガティブな状況より獲得される情報を用い、学習構造化の処理を受ける。このように、言語要素(単語等)の使われ方に基づいて知識の獲得を行うため、対象世界における拘束を緩やかに含んだ形の知識を得ることができる。また辞書データベースという記述的な形で構造化を行うため、それ自身インクリメンタルに成長することが可能となった。

1. ま え が き

自然言語の複雑でかつ多様な現象に対処可能な処理システムを実現しようとする際、それに必要な種々の知識をどのようにして獲得し、利用するかという問題は、知識表現の問題とも関連し、人工知能等の分野で中心的な課題の一つとなっている^{1),2)}。

本研究は、日本語文の係り受け構造を解析するのに必要な知識のなかで、とくに言語要素(自立語、付属語等)の用法に関するものに焦点を定めて、これを半自動的に手法により、辞書データベースとして記述的な形で構造化することを試みたものである^{3),4)}。

本研究のアプローチの概要を図1に示す。このアプローチの特徴的な点は次の2点にまとめられる。

(1) 辞書の構成手法は、テキスト文における係り受け関係のクラスタ分析に基づいている。作成された辞書はテキスト文には陽に含まれていない単語間の関係を分析できる可能性をもつ。

(2) 文解析実験を通して辞書データベースの評価を行い、その結果に基づいて辞書の内容を改良するフィードバック径路をもち、実験の不成功の状況から得

られる情報を辞書記述に反映するための学習機能をもつ。

文解析については、ESSAY という実験システムを用意している⁴⁾。

二文節間の係り受け関係を主にした解析は、九州大学で試みられている⁵⁾が、これは、品詞水準で行われている。しかしわれわれの手法では言語要素の使われ方に注目して、知識の獲得を行うため、対象世界における拘束等もゆるやかに含んだ形で知識を得ることができる。また、辞書データベースは学習機能を用いて、インクリメンタルに成長しうるので、高い適応性をもっている。

本論文で対象としたテキストは小学校算数の教科書であり、次の2点を考慮して選定した。

- (1) 数詞等を除けば、語彙数が限られている。
- (2) 曖昧さを含む文、複雑な文が少ない。

以下、2章では係り受け解析のための辞書の構成法と初期辞書データベースについて述べる。また、3章では辞書の評価、4章では学習機構および学習実験について述べる。

2. 辞書の構成と初期辞書データベース

本論文のアプローチの基本は、実文章に含まれる単語間の係り受け関係を、クラスタ技法により分析することにある。その結果は辞書中に登録され、つぎの文解析に用いられる。本章では、その辞書の構成法について述べる。辞書データベースは文解析に使用される

† Database Formulation and Learning Procedure for Kakari-Uke Dependency Analysis by KATSUHIKO SHIRAI, YOSHIHIKO HAYASHI, YUICHI HIRATA and JUNICHI KUBOTA (Department of Electrical Engineering, Waseda University).

†† 早稲田大学理工学部電気工学科

* 現在 日本電信電話(株)横須賀電気通信研究所

** 現在 九州電力(株)

*** 現在 松下電器(株)中央研究所

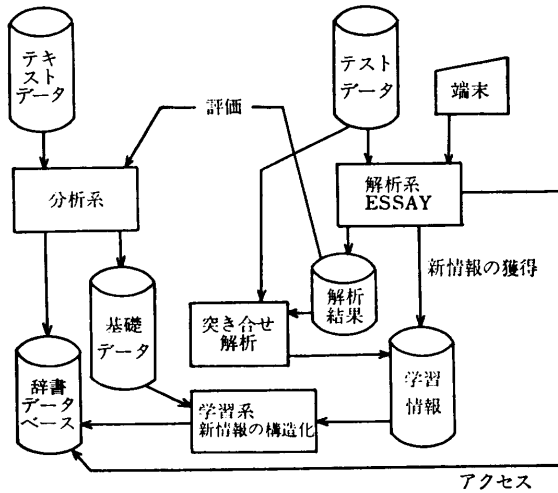
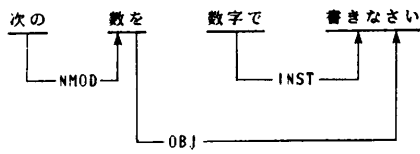
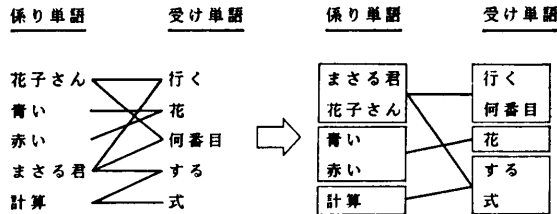


図1 アプローチの概要

Fig. 1 Overview of research approach.



(a) 係り受け構造



(b) 係り受け対応関係とそのクラスタリング

図2 単語のクラスタリング

Fig. 2 Clustering of words.

表1 テキストデータの統計量

Table 1 Fundamental statistical amount of accumulated text data.

文数	2,503
自立語数	2,526
付属語数	248
平均文節数	5.2

ことにより成長していくので、ここで述べる辞書は実際には初期辞書となる。

2.1 単語のクラスタリング

先に述べたテキスト文に対し、係り受け構造を図2(a)に示すような文節間二項関係として与える。ま

た、各文節は、自立語とそれに続く付属語列(ない場合は何も付かないという付属語として扱う)により構成されるとすると、テキスト文より図2(b)に示すような係り語対受け語の対応関係を得ることができる。

そこで、まず係り側の単語に関して、同じような係り特性をもつものをクラスタリングによってまとめる。

次に、受け側の単語に関しては、クラスタリングによってできた係り側単語クラスと対応において同じような受け特性をもつものをまとめる。

係り側クラスタリングを先に行うのは、一般に受け単語より係り単語のほうが、係り受けの特徴が現れやすいということが、いくつかの予備的な実験によってわかったためである。

2.1.1 単語間距離の定義

単語クラスタリングのための距離尺度として、単語間の距離(非類似度) $D(a, b)$ を係り側、受け側についてそれぞれ以下のように定義する。

係り側

$$D(a, b) = 1 - \frac{\sum_{c \in U} \sum_{f \in F} \Phi(M(a, f, c), M(b, f, c))}{\sum_{c \in U} \sum_{f \in F} (M(a, f, c) + M(b, f, c))} \quad (a, b \in K) \quad (1.1)$$

受け側

$$D(a, b) = 1 - \frac{\sum_{A \in C} \sum_{f \in F} \Phi(M_c(A, f, a), M_c(A, f, b))}{\sum_{A \in C} \sum_{f \in F} (M_c(A, f, a) + M_c(A, f, b))} \quad (a, b \in U) \quad (1.2)$$

ただし、

a, b, c : 自立語, A : クラスタ

f : 付属語

K : 係り側に現れた自立語の集合

U : 受け側に現れた自立語の集合

C : 係り側クラスタの集合

F : すべての付属語の集合

$M(a, f, b)$: 単語 a が付属語 f を介して単語 b に係る頻度

$M_c(A, f, b)$: クラスタ A が付属語 f を介して単語 b に係る頻度

$$\Phi(x, y) = \begin{cases} x + y & (x > 0 \text{ かつ } y > 0) \\ 0 & (x = 0 \text{ または } y = 0) \end{cases}$$

単語間距離の定義としては、このほかに頻度まで考慮しないものや、付属語を考慮しないものなどで実験

を行ったが、単語間の係り受け関係が最も明確に表され、評価結果もよいということで、上に示した定義を用いることにした。

2.1.2 クラスタリングアルゴリズム

以上に述べたクラスタリングに用いたアルゴリズムはセントロイド法⁵⁾である。これは、セントロイド法によれば、特定のクラスタに多数の単語が集中すること(チェイニング効果)を避けることができるためである。

2.1.3 閾値の定め方

単語間の距離行列を入力としてクラスタリングを行うが、その際問題となるのは閾値の定め方である。

一般に、閾値を高くすると抽象化された記述が得られるが、文解析に用いたとき曖昧さが生じやすく、形成される単語クラスタはサイズが大きくなり、その性質が不明確となる、といった問題が生じる。そこで、閾値を決定する指標として、次の品詞分散と距離歪という2種類の量を定義し用いることとした。

品詞分散 E_c は、クラスタ内の単語の品詞的なまとまりを表す値であり、以下のように定義する。

$$E_c = \sum_{c=1}^Q E_{cc} \quad (2)$$

Q : クラスタ数

$$E_{cc} = \sum_{T=1}^{N_{cl}} |H_{cl} - H_T|^2 \quad (3)$$

N_{cl} : あるクラスタ内に含まれる単語数

$$H_{cl} = \frac{\sum_{T=1}^{N_{cl}} H_T}{N_{cl}} \quad (4)$$

$H_T = (x_1, x_2, \dots, x_7)$

H_T : 単語の品詞を表すベクトルで x_i は七つの品詞カテゴリ(名詞, 動詞, 形容詞, 連体詞, 副詞, 代名詞, 接続詞)に相当し、単語の品詞に該当する要素が1となる。

H_{cl} : あるクラスタ内の平均品詞ベクトル

クラスタリングを行った後では、同一のクラスタに属する単語間の距離は0とみなされるため、クラスタリング前の距離との間に差異が生じる。この差異は、元の情報の保存の指標として用いることができる。そこで、距離 E_d を以下のように定義する。

$$E_d = \sum_{c=1}^Q E_{dc} \quad (5)$$

$$E_{dc} = \sum_{i=1}^{N_{cl}-1} d_i \quad (6)$$

d_i : 単語がクラスタに含まれたときの単語とクラスタの中心間の距離

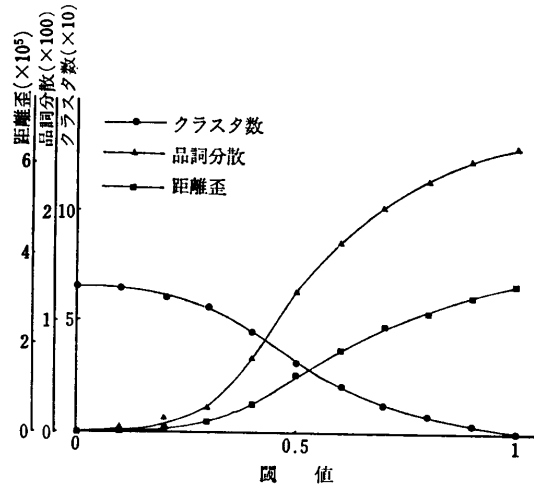


図3 閾値と歪とクラスタ数の関係

Fig. 3 Relation between threshold value, distortion and number of clusters.

つまり、品詞分散とは、クラスタリングによって、一般に用いられている文法的情報が、どの程度保たれているかを示す値で、距離歪とは、クラスタリングによって、初期のデータ量がどの程度保たれているかを示す値である。

単語間距離行列に対し、クラスタリングを行った際の、閾値に対するクラスタ数、品詞分散、距離歪の変化を図3に示す。閾値を高くとるほど、クラスタ数は減少し、品詞分散、距離歪は増加する傾向が読みとれるが、この例では各量が急激な変化を示すのは閾値が0.35の付近であった。よって、閾値は0.35程度が適当であると考えられる。

2.1.4 単語の出現頻度の考慮

現在のテキストデータ中において、総異なり単語1,529語中、係り単語数約1,500語、受け単語数約1,000語であったが、全テキストを通して出現頻度が1のものが、係り単語で約40%、受け単語で約44%存在する。よって、出現頻度が高い単語と低い単語を同一の基準で取り扱おうと、頻度の低い単語(一般に重要度が低いと考えられる)が雑音的な存在となってしまう頻度の高い単語(重要度が高いと考えられる)の特性が明確にならないという問題が生じる。そこで、今回は頻度が2以下の単語と3以上の単語はそれぞれ別に処理し、結果を合体した。

2.2 階層多重クラスタの構成

以上のようなクラスタ化だけでは、抽象化の能力が不十分であるので、単語レベルのクラスタの上にさらにクラスタを構成することとし、それを階層多重クラスタと呼ぶ。階層多重クラスタは係り単語クラスタの

上位集合であり、一つの係り単語クラスタは、複数の階層多重クラスタに属しうることになっている。したがって、単語-単語クラスタ-階層多重クラスタの階層構造は、完全な木構造とはならない。しかし、このように多重レベルの情報を用いることにより、複雑な日本語文の係り受け構造をより大きく捉えることができる。

本クラスタの構成手法としては、単語クラスタ間の距離を求め、再度クラスタリング技法を適用するという方法はとらなかった。その理由は、少数のサイズの大きな階層クラスタと、単語クラスタと同一の多数の階層クラスタが形成されることが予備実験により判明しているためである。

そこで、以下のような構成手法をとっている。すなわち、同一の受け単語クラスタに、同一の付属語を介して係っている係り単語クラスタを要素とする一つの階層多重クラスタを形成する。これは、同じ係り受けを示すことのある係り単語クラスタ間には、何らかの共通の要素があるはずであるという仮定に基づくものである。

2.3 辞書データベースの構成

以上のように、分析対象のテキスト文中に存在する係り受け関係を成立させている自立語間の依存・修飾関係のクラスタ分析によって、初期辞書が作られる。

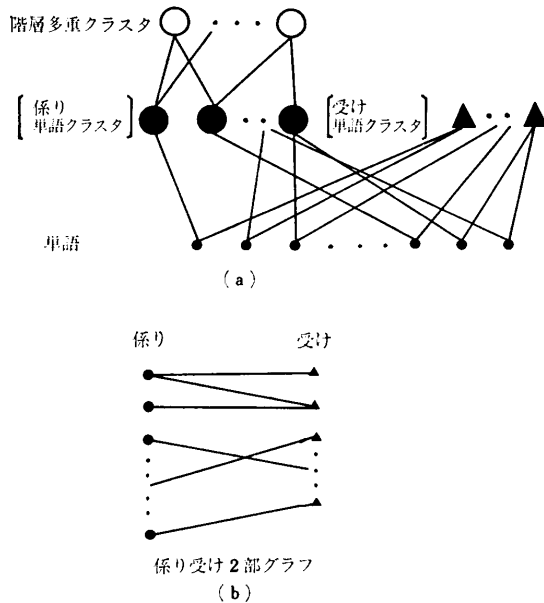


図 4 辞書記述における階層構造
Fig. 4 Hierarchical knowledge structure of dictionary description.

その結果、自立語の集合である単語クラスタ間の係り受け可能関係が図 4 (a) のように構造化されている。おのおのの自立語は一つの係り単語クラスタ、一つの受け単語クラスタに属し、一つ以上の階層多重クラスタに属することになる。

階層多重クラスタには、階層多重クラスタ-受け単語クラスタ間の 2 部グラフは用意されておらず、図 4 (b) の 2 部グラフが用いられる。そのため、各係り単語クラスタには、それが属する階層多重クラスタに共通に属する他の係り単語クラスタとの密接度の情報を

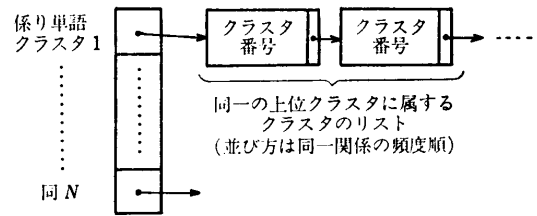


図 5 階層多重クラスタの利用
Fig. 5 Utilization of hierarchical multi clusters.

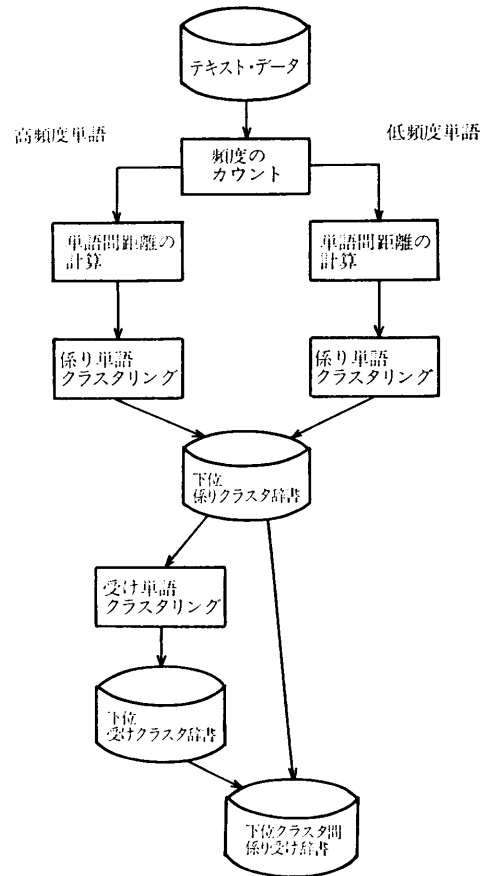


図 6 辞書構成処理の流れ
Fig. 6 Flow of dictionary construction process.

表 2 初期辞書データベースの評価
Table 2 Evaluation of initial dictionary data base.

	単語レベル	クラスタレベル	階層使用
係りクラスタ数	1,635	839	—
受けクラスタ数	1,635	1,215	—
係り受けリンク数	6,162	4,202	—
カバーリンク数	6,162	106,819	—
品詞分散	0	35.4	—
文解析率(%)	32.5	46.7	54.2

ものから順に、文の解析結果として出力する（出力例を図7に示す）。

実験は辞書構成に用いた文とは別の500文について行い、一文について出力される複数の解析結果の中に正しい解析結果が含まれていた文の割合を文解析率として評価指標に用いた。また、品詞分散、距離歪、カバーリンク数を用いて性質を評価した。カバーリンク数 NL とは、処理可能な係り受け単語のペア数であり、次式で定義する。

$$NL = \sum_{i=1}^{N_K} \sum_{j=1}^{N_U} R(i, j) |K_i| \cdot |U_j| \quad (8)$$

N_K : 係り単語クラスタ数

N_U : 受け単語クラスタ数

$|K_i|$: 係り単語クラスタ K_i の要素数

$|U_j|$: 受け単語クラスタ U_j の要素数

$$R(i, j) = \begin{cases} 1: K_i \text{ から } U_j \text{ へ係り受け可能} \\ 0: \text{それ以外} \end{cases}$$

表2に結果を示す。クラスタリングによる抽象化の効果は、カバーリンク数、文解析率に現れているが、得られている解析率は十分とはいえない。文解析の失敗には、係り先が不明の文節を含む場合と、得られた係り受け構造が違っている場合の2タイプがあるが、失敗のほとんどは前者のタイプであった。また、辞書検索によって係り先が求められない文節数は、失敗した文の総文節数のうち約30%であり、これを補うような学習処理を行えば、大幅に解析率を向上させることが可能であることが判明した。

4. 辞書に対する学習機能と評価

初期辞書作成のためのデータには実際上限りがあるので、辞書の性能向上のためには学習機能が不可欠である。辞書に対する学習機能は、次の2段階のフェーズにより実現される。

(1) 文解析における何らかの意味の不成功の状況から学習情報を抽出・獲得する。

(2) (1)により獲得された学習情報を既存の構造

中へ構造化する。

(1)の処理に対する支援機能は、文解析実験システム ESSAY に用意されているので文献4)に譲り、以下では(2)について述べる。ここで問題とする学習情報とは、既存の辞書中に記述されていない係り受けに関する情報であるが、これをたんに係り受け2部グラフ上の新しいリンクとして付加するよりも、構造化したほうが合理的であり、獲得した情報を抽象化して保持することも可能となる。

したがって、学習処理は単語クラスタ形状を再構成する処理であるといえる。しかしながら、2章で述べた初期辞書の構成手法はかなり複雑な処理によっているので、文解析された文をすべて原テキストデータに付加し、すべてのクラスタ化処理を再度行うことは実際的ではない。そこで、次のような再構成法を試みた。

4.1 単語クラスタ形状の再構成

新しい係り受け情報の付加に対して、単語クラスタ

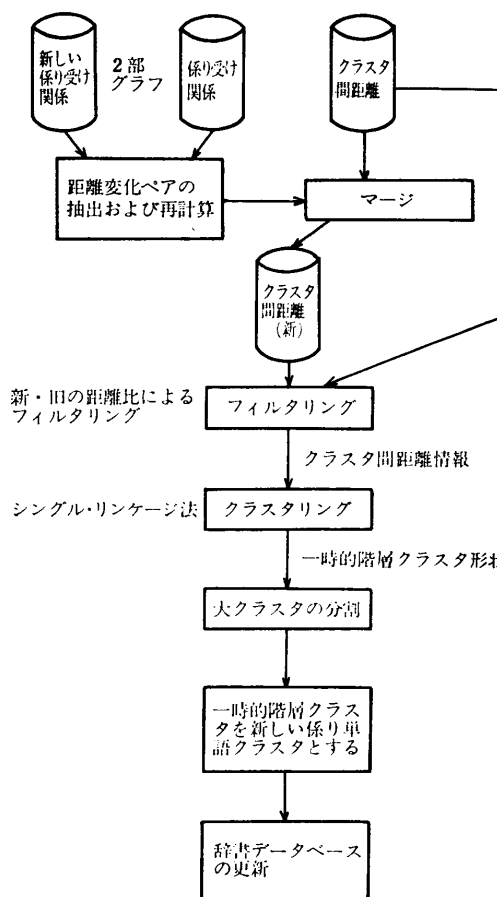


図8 係りクラスタ形状の再構成処理の流れ
Fig. 8 Flow of reconstruction process for kakari clusters configuration.

間の距離の変化に注目しながら、一時的階層クラスタ(2.2節の階層多重クラスタとは異なる)にその情報を集約し、図8のような処理に従って、係りクラスタ形状の再構成を行う。

一時的階層クラスタを用いることにより、単語レベルにおける複雑な処理(頻度による処理の振り分け等)を避けることができる。以下、図8の流れに従い述べる。

4.1.1 距離変化ペアの検出と再計算

係り単語クラスタ i, j 間の距離 D_{ij} は、クラスタレベルでの係り受け2部グラフ上でのリンクの共通性にに基づき次式で定義している。

$$D_{ij} = \frac{\sum_r \{w(i, r) + w(j, r)\}}{\sum_p w(i, p) + \sum_q w(i, q)} \quad (9)$$

$w(i, r)$: 係りクラスタ i から受けクラスタ r の係り受けリンクの頻度

新しい係り受けが影響するクラスタは距離の変化が生じるクラスタ・ペアとして容易に検出できる。よって、これらのペアについてのみ距離の修正計算を行う。

4.1.2 クラスタ間距離行列によるフィルタリング

クラスタリングに対して入力となるのは、新しい状況下でのクラスタ間の距離行列であるが、クラスタ間距離を、ある程度以上変化させるような情報が重要であるから、それを抽出するため、本行列に対してフィルタリング処理を施した。すなわち、新旧の行列をそれぞれ $D=(D_{ij}), D'=(D'_{ij})$ とするとき、次の変化率行列 R を定義し、

$$R=(R_{ij}) = \left(\frac{|D_{ij} - D'_{ij}|}{D_{ij}} \right) \quad (10)$$

R_{ij} が閾値 $T(T \leq 1)$ に対し

$$R_{ij} > T$$

となるデータだけを用いて一時的階層クラスタを構成した。

結局、この処理は、新しいデータによって距離が近くならなかったペアは一時的階層クラスタにおいては同一のクラスタに含まれないようにするという意味をもつ前処理である。

4.1.3 クラスタリング・アルゴリズム

一時的階層クラスタを構成するのに既存の単語クラスタを距離 D_{ij} によってクラスタリングする。この場合のクラスタリングには、単語クラスタの構成の場合に比べ精密さが要求されないため、アルゴリズムが単純な Single-Linkage 法⁵⁾を用いた。

Single-Linkage 法においては、要素数の多いクラスタが少数構成され、他の多くのクラスタは一つの要素しか含まないという chaining-effect が問題となるが、これは4.1.2項のフィルタリングにより避けることができた。すなわち、フィルタリングを行わない予備実験においては、複数の要素を含む一時的階層クラスタは一つであり、その要素数は150程度(要素全体の約1/6)であったのに対し、フィルタリングによれば複数の要素を含む一時的階層クラスタの数は67と妥当な結果が得られた。

なお、得られる一時的階層クラスタには、既存の単語クラスタと比較して、サイズの大きすぎるものも含まれるので、細分割処理を行う。細分割は、多くの単語を含む単語クラスタを核クラスタとして選出し、核クラスタ以外の単語クラスタは、それに最も距離の近い核クラスタに結合するという処理により行った。

4.2 階層多重クラスタ形状の再構成

本処理は、再構成された係り単語クラスタ形状との対応をとり、学習情報に従って新しいクラスタを生成するか、または、既存の該当クラスタへの追加を行うだけでよく簡便である。

4.3 学習実験と評価

学習処理により再構成された辞書データベースの性能と性質の学習情報の量に対する変化を調べた。

実験条件は、初期辞書構成に用いない文を100文×3セット=300文用意し、学習を行った。新しい単語を含む文は除外した。このとき得られるクラスタレベルの新しい係り受けリンク数は251であり、これは初期辞書中のリンク数の約6%に相当する。

カバーリンク数と解析率の評価結果を図9, 10に示す。図中、構造化Aとは、2章に述べたクラスタリング手法をすべてのデータに対し再度適用するものであり、構造化Bとは本章に述べた手法によるものである。また、非構造化とは新しい係り受けリンクを2部グラフにたんに付加するものである。

カバーリンク数については、構造化Bの場合が最も増加率が大きい。これは、一時的に上位レベルの情報を用いていることによるものである。

解析率をみると、構造化A、構造化B、非構造化の順により値を示しており、構造化による抽象化の効果が明らかである。また、構造化Bは構造化Aに比べカバーリンク数で勝るが、解析率では若干劣っており、実験に用いたテキストでは、カバーリンク数は解析率に必ずしも直結しないことがわかる。

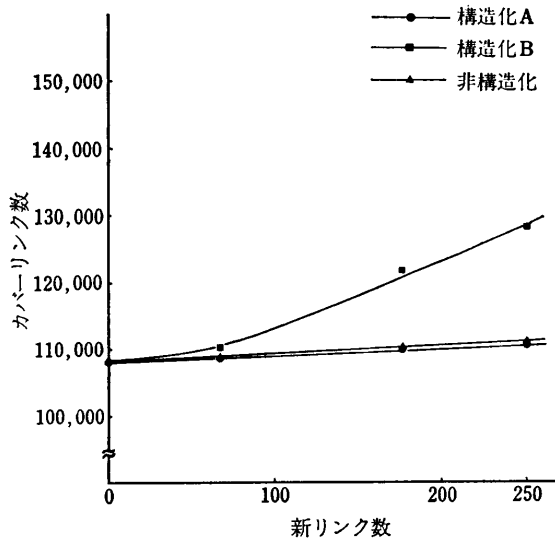


図 9 学習情報量とカバーリンク数の関係
Fig. 9 Relation between size of learning information and number of possible links.

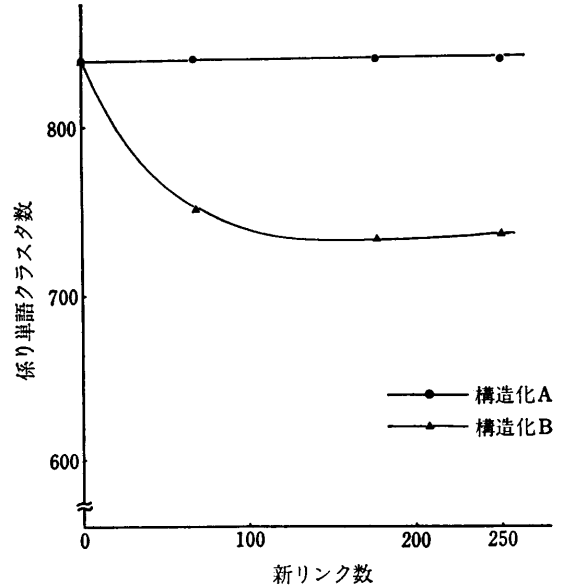


図 11 学習情報量と係り単語クラスタ数の関係
Fig. 11 Relation between size of learning information and number of kakari-words clusters.

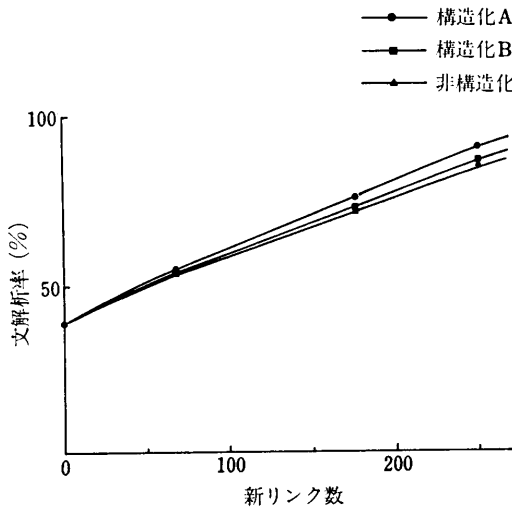


図 10 学習情報量と文解析率の関係
Fig. 10 Relation between size of learning information.

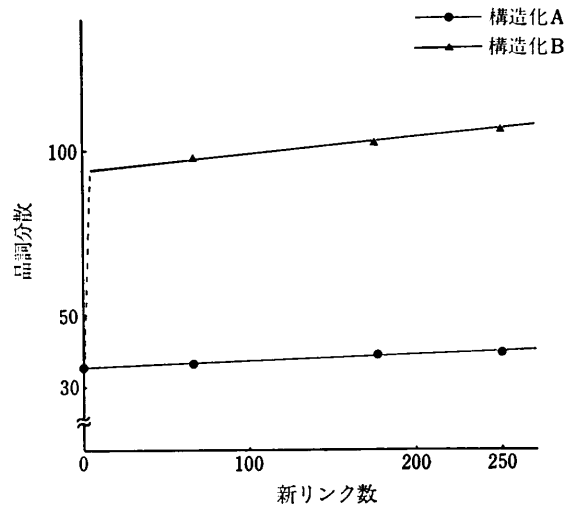


図 12 学習情報量と品詞分散の関係
Fig. 12 Relation between size of learning information and grammatical categorical distortion.

係り単語クラスタの数と、2.1.3 項に述べた品詞分散を指標として評価した結果を図 11, 12 に示す。本実験においては、新しい単語は対象文中には含まれないので、新しい係り受け関係を考慮しても、構造化Aの場合は、係り単語クラスタ数、品詞分散ともほぼ一定の値を示している。また、構造化Bの場合も学習情報の加え始めにおいて急激な変化がみられるが、その後では一定傾向を示している。

構造化Bの場合、クラスタ数は少なくなるにつれ品詞分散は大きくなっており、要素数の多いクラスタが

構成されていることがわかる。これは、一時的に上位レベルの情報を用いていることによるものである。

5. むすび

実文章における係り受け関係の分析に基づいて初期辞書データベースを構成することにより、言語の構成要素である単語の現実の機能を重視した辞書記述を得ることが可能となった。しかしながら、現在、辞書構成のための分析の基礎となっているのは、2 文節間の係

り受け関係を形づくる単語間の2項関係という最も単純な関係にしか着目していないので、今後は3項以上の関係なども考慮していくとともに、文解析システム ESSAY もそれに対応した改良を行っていく必要があると考えられる。

また、このような核というべき初期辞書は、5章に述べた学習処理により、インクリメンタルに成長可能であるが、これは辞書データベースという記述的な形式で、言語要素の用法に関する知識を表現することによる大きな利点の一つと考えられる。今後は、どの程度の量の学習情報を得たときに、構造化処理を行うべきかという点について検討を進める予定である。

いずれにせよ、図1のように大部分の処理を機械的に行うことにより、整合性のとれた記述を得ることが可能となるので、大規模なデータセットへの応用も可能であると考えられる。

参 考 文 献

- 1) 辻井潤一：表現と利用，信学技報，AL 80-27 (1980).
- 2) 大須賀節雄：知識の表現と利用—知識システムの満たすべき条件—，情報処理，Vol. 19, No. 10, pp. 944-951 (1978).
- 3) Shirai, K., Kubota J. and Hayashi, Y.: Japanese Sentence Analysis System ESSAY, *Bulletin of Science Engineering Research Laboratory, Waseda University*, No. 103, pp. 1-18 (1983).
- 4) 白井克彦, 林 良彦, 平田裕一：日本語文係り受けの分析・解析・学習，情報処理学会自然言語処理研究会資料 36-1(1983).
- 5) Hartigen, J. A.: *Clustering Algorithms*, John-Wiley & Sons, New York (1975).
- 6) 吉田 将：二文節間の係り受けを基礎とした日本語文の構文解析，信学論(D)，Vol. 55-D, No. 4, p. 238 (1972).

(昭和58年11月14日受付)

(昭和60年2月21日採録)