社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－159　(11)
2004／1／13

係り受け関係からの格フレーム辞書自動生成システム

上之薗 和宏[†]　榎津 秀治[‡]　古宮 誠一[†]

あらまし　本稿では，係り受け解析に基づく，遺伝的プログラミングによる，格フレーム辞書生成システムを示す．格フレーム辞書を生成するに当たって，文の構成要素が他の構成要素に対して，どのような意味関係をもつか，その分類が目的，ドメインによって多様なものとなる．また，そのクラス分けを行う決定木（規則）を作成するには，膨大な労力が必要となる．そのため，効率よく決定木を学習し，それを利用して，自動的に格フレーム辞書を生成するシステムが有効である．そこで，フリーの解析ソフトを使って文節間の係り受け関係を獲得するとともに, 帰納学習と GP に基づく格決定ルールによって格を定め，複数の格フレーム（の要素）を暫定的に作成し，格ごとにクラスタリングを行う手法を述べる．

キーワード　格フレーム，意味解析，格解析，遺伝的プログラミング，帰納学習

The Case Frame Dictionary Automatic Generation System by the Dependency Relation between Clauses

Kazuhiro UENOSONO[†]　Hidezi ENOKIZU[‡]　Seiichi KOMIYA[†]

Abstract　This paper shows the case frame dictionary generation system by the genetic programming based on a dependency analysis between clauses. When generating a case frame dictionary, case classification will become various by the purpose and the domain with what semantic relation to the composition element. Moreover, a huge labor is needed in order to create the decision tree which performs the class division. Therefore, a decision tree is learned efficiently and generates a case frame dictionary automatically. Then, while acquiring the dependency relation between clauses using free analysis software, a case is defined by the case determination rule based on inductive study and GP, and the technique of clustering for every rank is described.

Keyword　case frame, semantic analysis, case analysis, genetic programming, inductive study

1. はじめに

近年，既存の文書の電子化や，インターネットの普及により，自然言語で書かれた電子文書の急増している．それに伴い，それらの効率的な有効利用を実現する研究が求められている．

そこでは，それらの文章を要約・分類し，そこに記載されている情報を抽出・要約する技術が必要である．その為，意味要素である語や句の意味関係を明らかにするとともに，文の意味構造を生成する意味解析が必要である．これにより，文章中の情報や意図を抽出することができる．

しかし，ある事柄を表現する方法が幾通りも存在したり，意味要素の相互関係が複雑に作用したりするため，意味解析を行うことは非常に困難である．そのため，コンピュータを用い，その意味構造を生成する規則，または辞書を自動的に学習・生成することが求められ，実際に半自動で生成する研究が行われている[1][2].

そこで，格文法によって，文を構成している要素が文中の述語に対して果たす機能を示す格構造を，文節間の係り受け関係と格決定ルールにより，格フレーム辞書として自動的に生成するシステムを提案する．

本稿では，格フレーム辞書自動生成システムの全体像と,そこ用いる格を決定するルールを，

†芝浦工業大学大学院
Graduate School of Engineering, Shibaura Institute of Technology, {m102184, skomiya}@sic.shibaura-it.ac.jp
‡芝浦工業大学
Shibaura Institute of Technology, enokizu@sic.shibaura-it.ac.jp

機械学習の一つである帰納学習と，進化的計算手法の一つである GP を用いて学習する，格決定ルール学習システムに付いて述べる．

2. 意味解析での格フレームの位置付け

本稿における意味とは，文の構成要素が持つ意味，すなわち意味要素と，その相互関係が示す意味,すなわち意味構造を併せたものである．意味要素には文節と形態素を用い，意味要素が持つ意味は，語義辞書や語彙体系等を用いる．意味関係は格文法に従って格フレームを用いて表現する．

2.1. 意図理解モデル

自然言語で書かれた文章の意図を人間が理解する過程をモデルとして図１に示す.これは，与えられた自然言語を表現する音声や文字画像から，音声・文字認識，形態素解析，構文解析，意味解析，文脈解析，意図解析を順に行うものである．本稿で扱う意味解析は，この意図理解モデルにおいて，構文構造を基に意味構造を同定する処理系である意味解析器によって行われる．

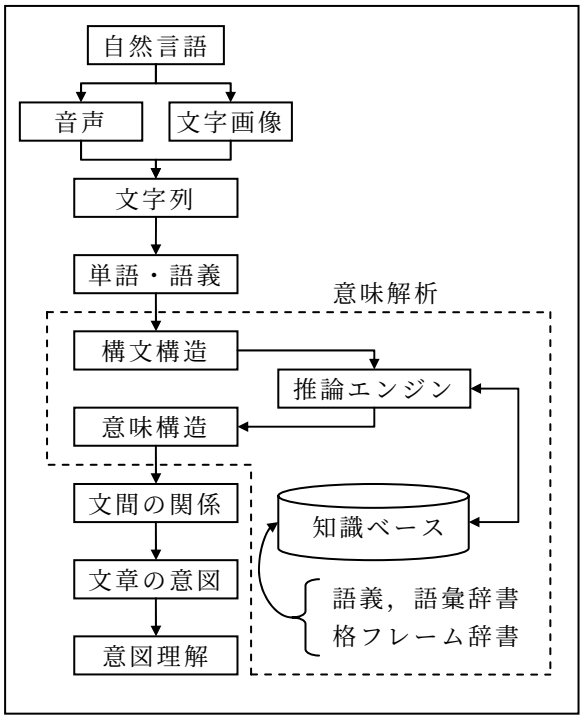

図１　意図理解モデル

2.2. 格フレームの位置付け

人間が文章の意図を理解するのは，記号列である自然言語表現から，意図を抽出する能力を有すると捉え，この言語能力の，構文構造から文を構成している要素，すなわち意味要素の間の意味構造を同定する能力の獲得を目指す．

意味構造を同定するには，意味解析器が構文解析の結果を受け取り，汎用的な推論エンジンに渡す．推論エンジンは意味構造を同定するために必要な知識を知識ベースから参照して用いる．格フレーム辞書はこの知識ベースに含まれており，目的，用途に応じて交換可能である．また，格フレーム辞書は，文中の意味構造を作成するだけでなく，文の意味を説明するのにどのような情報が不十分であるかを知るためにも利用される．

3. 格決定ルール

一般に日本語において，述語に対してある語が割り当てられる格は，格助詞・係助詞によって決まるとされている．しかし，同じ格助詞であっても異なる格を持つ場合がある．また，格助詞がなくても，語の意味から格をある程度特定できる．このことから，格を決定するルールは，助詞等の自立語に付随するもの，意味要素間の依存関係，その他得られる情報を基に生成する必要があると考える．

そこで，既存の助詞に基づく格解析システムによって格を割り当て，係り受け結果にある情報を出来るだけ用いて，それらを訓練事象として新たな有限の格を決定するルールを獲得する．格を決定するルールを，決定木を用いて表現する．これを格決定ルールと呼ぶ．決定木とは，中間ノード（リーフ以外のノード）がテストされるべき属性を，枝がその属性値を，またリーフノードがクラスを表している木構造である．また，得られた格決定ルールの精度を高めるために，遺伝的プログラミングの手法を用い，精度の向上を目指す．

格決定ルールは，係り受け結果を参照し格を割り当てる．格決定ルールは，戻り値として

１)一つの決定木で全ての格の割り当てを行い，参照した係り受けに割り当てる格を返す．

２）それぞれの格ごとに決定木を持ち，その格を割り当てるかどうかを true/false で返す．

の２つの場合が考えられる．

格決定ルールは大きく分岐条件部である Condition Cell と，実行部である Action Cell の 2

つに分類される要素から構成される．Condition Cell は実際の比較演算による Condition Statement Cell と，論理演算の And，Or，Not の Operator Cell を含む．Action Cell は入力された係り受け情報に実際に各を割り当てる Action Statement Cell と，再帰的に if-then ルールを扱えるように，if-then-else Operator Cell を含む．

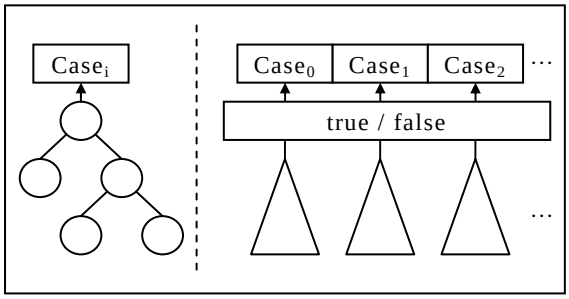

図２．決定木

3.1. 帰納学習法 C4.5

帰納学習法（帰納推論）とは，「大量のデータ，即ち訓練事例から，そのデータを説明する一般的な概念を小さな有限の表現として抽出・獲得するプロセス」のことであり，概念学習（concept learning）とも称される．対象の属性値とそれが属するクラス（集合）の対である，訓練データ集合から，そのデータをクラスに分類する決定木を生成するのが，決定木の帰納学習である．この学習法は Quinlan によって開発された汎用の学習手法である[3][4][5]．決定木の帰納学習法（C4.5,ID3）では，訓練データの集合から有限個の属性の連言形で表現可能な既知のクラスへの分類規則を決定木の形で帰納的に学習可能である．しかし，帰納学習だけでは，訓練データがノイズを含んでいる場合，属性地が不明な場合，連続値の場合などに対処しきれない．

以下に帰納学習 C4.5 により決定木の帰納学習アルゴリズムを示す．

3.2. 係り受け解析への適応

係り受け解析の結果は，既存の係り受け解析ツールである cabocha[6]を同格，並列の情報を出力できるモデルを用いて得る．得られた係り受け解析の結果は係り先文節を親とし，係り元文節を子とする木構造に変換する．

係り受け解析の結果と，係り元文節の助詞に着目した，既存の格決定ルールによって割り当てられた格との組を訓練データ集合とし，係り受け解析の結果を参照して格を割り当てる分類規則を格決定ルールとし，決定木で表す．

係り受け情報は係り受け文節情報，文節情報，形態素情報からなり，形態素情報以外の階層は可変長である．これを決定木の帰納学習法を適応するために，固定長にする．これを訓練データの属性値とし帰納学習を行う．図４に係り受け情報の変化を示す．

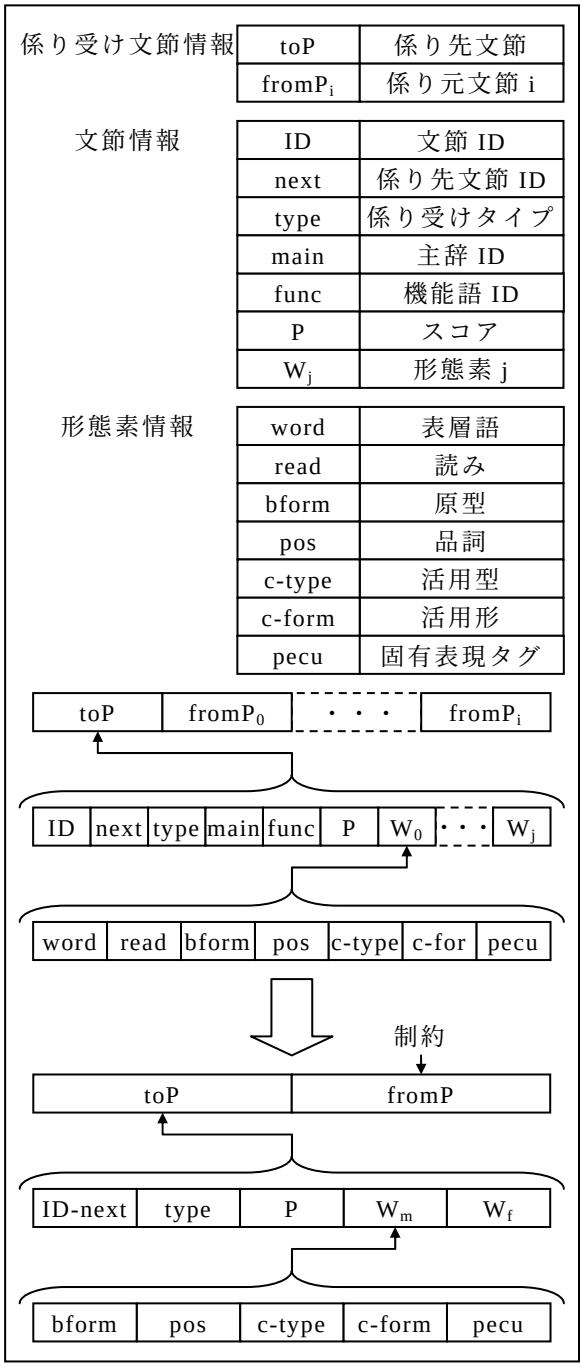

図４．係り受け情報の変化

3.2.1. 形態素情報

cabocha の出力に現れる形態素解析の結果を元に，形態素の情報を扱う．形態素情報は，表層語，読み，基本形，品詞，活用型，活用形，固有表現タグを持つ．ただし，形態素情報の表層語，読みは，基本形と活用型，活用形から導き出されるため，これらは格を決定する情報としては用いない．品詞，活用型，活用形，固有表現タグは予め定めたリストを参照する index 値で表す．

3.2.2. 文節情報

文節情報は，cabocha の出力に現れる「*」で始まる文節の開始位置を意味する行を元に，文節の情報を扱う．文節 ID，係り受け先文節 ID，係り受け type，主辞 ID，機能語 ID，スコア，形態素情報集合を持つ．

文節情報は，その文節に複数の形態素を持つため可変長である．これを，帰納学習を行う際に文節がもつ主辞と機能語がそれぞれ一つである事から，これらの情報を主にし，残りの形態素情報を付加させる形で固定する．

しかし，主辞と機能語の情報だけでは不十分ではないかと考えられるが，帰納学習の際の，文節情報の固定長化と，同じ意味をもつ属性値の比較に有効と考えられる．但し，品詞がサ変動詞の動詞が主辞の場合，直前の名詞を主辞と置き換える．

3.2.3. 係り受け文節情報

係り受け文節情報は cabocha の出力から，文節を係り先文節と，その文節に係っている係り元文節の組を取り出して扱う．係り先文節は述語であり，係り受け文節情報に唯一存在する．係り受け文節は同一文中の述語に係る全ての文節である．すなわち，述語に複数の文節が係るため可変長である．これを，帰納学習を行う際に，係り元文節と係り先文節を一対一の組に分割する．

ただし，同じ述語にかかった語は，異なる格を持つので，同じ係り先文節に係る文節は，cabocha で係り受けのタイプが D または O と判定された場合，異なる格を持つという制約を加える．同格または並列と判定された場合，その係り先文節がかかる，文末以外の文節に係るとし，同じ格を持つという制約を加える．

3.3. GP オペレータ

3.1 で述べたとおり，帰納学習のみでは，高い精度の決定木を生成できるとは限らない．そこで GP（Genetic Programming）[7]を用いる．GP は GA（Genetic Algorithms）の表現である遺伝子型を,構造的な表現を扱えるように拡張し，プログラム生成や学習，推論，概念形成などに応用することを目指したものである．3.2 で得られた格決定ルールである決定木の精度向上のため，図５に示すような GP オペレータを用いる．GP のオペレータには G-inversion（逆位），G-crossover （交叉），G-mutation（突然変異）を用いる．これは GA オペレータを決定木に適応するための自然な拡張である．各オペレータの適用率は，その世代の個体数に対して適応する個体数の割合である．以下に各オペレータについて述べる．また，決定木の評価値である適合度と，自然淘汰について述べる．

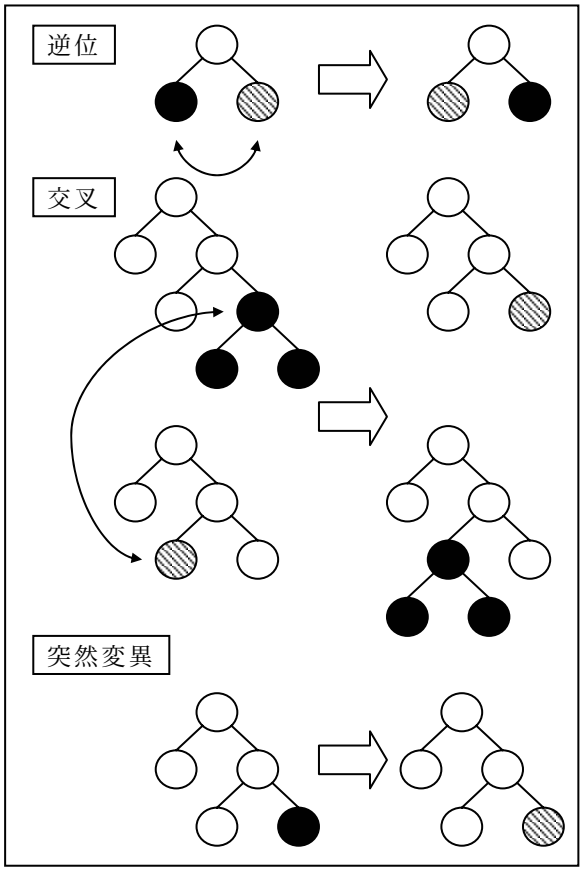

図５　GP オペレータ

3.3.1. 逆位

決定木同じノードが持つ子ノードの位置を変える．格決定ルールでは，if-then-else Operator

Cell の実行部を交換する．Condition Cell では，And と Or が複数の子を持つが,逆位オペレータを適応しても,役割が変わらないため用いない.

3.3.2. 交叉

二つの決定木のノードをそれぞれで選び，そのノード以下の部分木を交換する．図５ではそれぞれの決定木から，一つのノードを選択する一点交叉である．決定木の交叉には，複数箇所のノードを交換する複数点交叉や，二つの決定木の共通部分をさがし，効率よく交叉を行う一様交叉（uniform crossover），二つの決定木の同じ位置に対応するノードで行う狭義の交叉，同じ決定木の二つノード間の交差がある.

3.3.3. 突然変異

決定木のノードのラベルをランダムに変化させる. Condition Cell では参照する係り受け情報の遺伝子座の index 値と，その遺伝子座の属性値の範囲を乱数によって変化させたり，And と Or のオペレータを入れ替えたりする. Action Cell では，実行文のノードの返す値を乱数によって変化，または true，false を入れ替える．If-then-else Operator Cell は逆位と同様になる.

3.3.4. 適合度

決定木の精度を，適合度として t 世代の集団 P(t)の固体，すなわち格決定ルールの仮説 h∈P(t)が，格の割り当てが終了している評価用データに対し，正確にただし区画を割り当てられたデータの割合を決定木の適合度として用いる.，ある種の MDL（Minimum Description Length）基準として，決定木のサイズが小さいほど,適合度が高くなるように設定すべきだが,決定木がどの程度の大きさになるか検証してから採用する.

3.3.5. 淘汰

GP では通常,初期集団をランダムに上限個体生成し，世代を重ねるが，格決定ルールの決定木は，初期集団が 1 個である．これは，帰納学習によって，決定木を一意に獲得するためである. そこで, GP オペレータを同一の親に対しても行えるようにし，また，親自身も次世代に残るようにする．こうすることで，少ない母集団から，より多くの次世代を生成することが出来る．そして，ある世代で設定した個体の上限を超える世代で,淘汰処理を行う.淘汰処理では，適合度が高いものほど次世代に残りやすく，逆に低いものほど残りにくくする．そこで，個体集合をキューであるとし，世代 t に対して GP オペレータを適用した中間集合から，個体を一つ pop し，その適合度をそのまま，次の世代へ残る確率として判定する．もし，残ることになれば，次の世代の集合に push し，残らないことになれば，元の中間集合の最後に push する．次の世代の個体数が設定数に達したときに，淘汰を終了する．図６に淘汰処理の流れを示す.

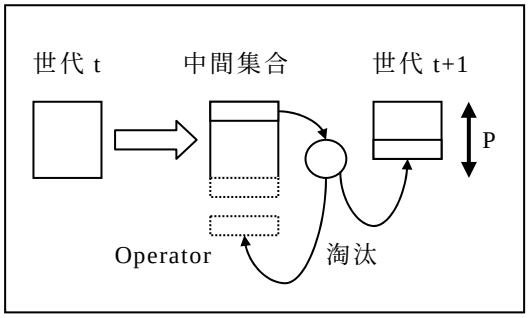

図６．淘汰処理

4. 格フレーム

格フレーム（case frame）は Fillmore によって唱えられた格文法（case grammar）に従い，意味要素が述語に対して果たす役割・機能に格（case）を，意味構造として格構造（case structure）を表現する．ただし，本稿では，表記上現れる構文的な表層格（surface case）ではなく，意味上の役割である深層格（deep case）を指す．従って格フレームは，述語に着目し，その述語がどのような格を持つか，又どのような意味要素が格となるかを示すものである．格の定義については，基本となる動作主格，経験者格，道具格，対象格，源泉格，目標格，場所格，時間格の八種の格を用い，必要に応じて修正する．この格は名詞節が述語に対して持つ役割を示すものだが，同様の手法によって，他の品詞間についても獲得可能である.

また，格フレームには以下で説明する格フレームオペレータによって生成する過程を保存するために，その格フレームが生成された元の格フレームの情報を残すことが出来るようになっている．これをサブ格フレームと呼ぶ.

4.1. 格フレーム辞書

格フレーム辞書は図 7 のような述語の集合と，その述語が持ちうる格と，その格を割り当てられる意味要素の集合をひとまとめにしたものが登録されている．意味解析器は，文中の述語に対する格フレームを検索し，その述語に係っている語にどの格が割り当てられるかを調べる．また，格フレーム辞書の格に要素が含まれているが，文中の語に格を割り当てられない場合，そのスロットを示し，格が不足していることを表す．

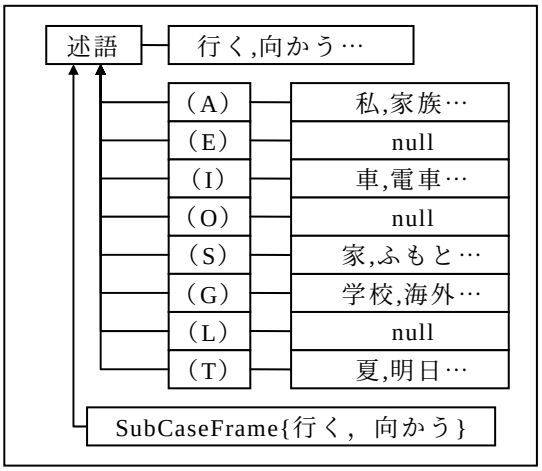

図 7．格フレームの例

4.2. 格フレームオペレータ

格フレーム辞書は係り受け情報に格を割り当てたデータから格フレームオペレータを使用して生成する．オペレータは CF-generate（作成），CF-delete（削除），CF-join（統合），CF- divide（分類）の四種類を用意する．

4.2.1. 作成

格を割り当てられた係り受け情報を入力とし，格フレームのインスタンスを出力する．このインスタンスは最小の格フレームとなり，実際に文に現れた具体例となる．

4.2.2. 削除

格フレーム辞書に登録されている格フレームを入力とし，削除する．ただし，その格フレームが持つサブ格フレームを格フレーム辞書に登録する．

4.2.3. 統合

統合オペレータは，二つの機能を持つ．一つは，同じ述語を持つ格フレームを一つにまとめる．それぞれの格において類似度を求め，類似度が高ければ和集合をとり新しい格フレームとする．類似度はシソーラスを基に，集合の大きさを踏まえて決定する．図 8 に，統合オペレータの簡単な例を示す．

もう一つは，同様の格を持つ述語を一つにまとめる．ある程度の大きさを持つ格フレーム間で同様にそれぞれの格においての類似度を求め，類似度が高ければ述語を一つにまとめ，新しい格フレームとする．格フレーム辞書は，主にこの統合オペレータを適用する事で生成される．

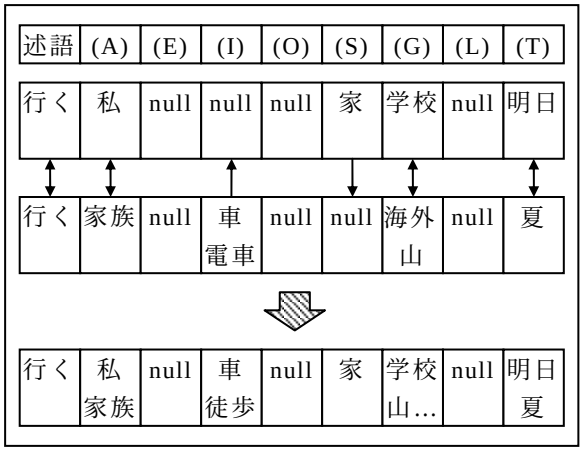

図 8．統合オペレータ

4.2.4. 分類

一つの格フレームの，それぞれのスロットの要素について，集合を二つに分割する．分割した集合間で類似度を求め，類似度が低い分割が出来た場合，格フレームからその集合を取り除き，新しい格フレームとする．この際，サブ格フレームを参照して，実際の用例に近い組み合わせになるよう，他の格から同時に要素を移動する．

5. 格フレーム辞書自動生成システム

本システムは，生コーパスに対して，cabocha を用い，形態素解析・係り受け解析を行った結果を入力とし，格フレーム辞書を出力とする．また，格フレームに用いる深層格は，前記した八種類の深層格を用い，出力結果を検討した結果，修正が必要と判断した時に修正を行う．

学習対象の入力は標準入力とファイルシス

テムから，cabocha によって係り受け解析済みのテキストデータを受け取る．格フレーム辞書の出力は，Object Stream，text，DB，XML など選択可能にする．また，出力された格フレーム辞書を検討するために，途中経過をファイルで出力し，編集する機能を持たせる．このシステムは，係り受け解析情報と，格決定ルールがあれば，自動的に辞書を生成する事が出来る．図９にシステムの概要を示す．

格決定ルール学習システムは，格フレーム辞書自動生成システムに必要な格決定ルールを,3 で述べた手法で生成する内部システムである．係り受け解析の結果を入力とし,帰納学習と GP によって生成される決定木を出力する．

格決定システムは，係り受け情報を入力とし，格決定ルール学習システムによって得られた格決定ルールに従って，それぞれの要素に対して格を割り当てる内部システムである．

格フレームオペレートシステムは，格決定システムから格を割り当てられた係り受け情報を基に，4.2 で述べたオペレータを用いて格フレーム辞書を作成する内部システムである．

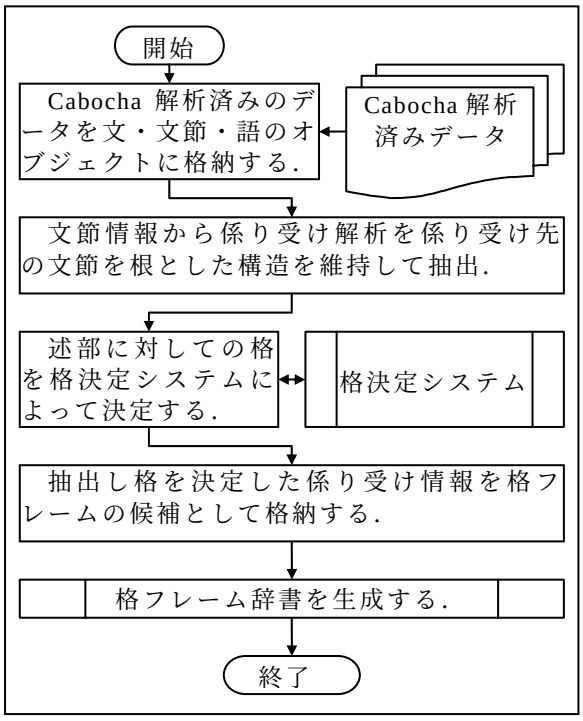

図９．格フレーム辞書自動生成システム

6. まとめ

本稿では，格フレーム辞書を生成するにあたり，もっとも大きな労力を要する格を割り当てる処理に付いて,帰納学習,GP の考えを導入し，より高精度のシステムを提案した．

今後，GP の効率化を図る．これは GP の遺伝子表現の簡略化と,GP のオペレータのより知的な操作を指す．例えば，交叉を行う際に，親の決定木の，精度の高い部分木同士を組み合わせる．

また，格決定ルールを学習させるには，訓練データが少ないというボトルネックに対して，GA オペレータを用いる事で，少ない訓練データから，より高い適合度の決定木を帰納学習する方法を検討している．

格フレームオペレータに付いては，類似度の計算アルゴリズム，メタフレームの設計，述語部分に名詞，形容詞，形容動詞が入るように拡張を目指す．

参考文献

[1] Baker, Collin F., Fillmore, Charles J., and Lowe, John B. The Berkeley FrameNet project, Proceedings of the COLING-ACL, Montreal, Canada, 1998

[2] 河原大輔，黒橋禎夫，用言と直前の格要素の組を単位とする格フレームの自動構築，自然言語処理，Vol.9，No.1，2002.1

[3] Quinlan,J.R.,C4.5 Programs For Machine Learning, Morgan Kaufman Publishers,Inc. (1993)(古川康一監訳,“ AI によるデータ解析 ”，トッパン，1995)

[4] 手塚勝，帰納的学習法 C4.5 による知識の獲得とその評価，平成 12 年度学位論文概要集 pp.124-125（2001）

[5] 吉田倫洋，遺伝的アルゴリズムによる規則の獲得, 2001 年度芝浦工業大学システム工学部電子情報システム学科総合研究論文

[6] 工藤拓，松本裕治，“チャンキングの段階適用による日本語係り受け解析，” 情報処理学会論文誌, vol43, no.6, pp.1834-1842, 2002

[7] 伊庭斉志.遺伝的プログラミング入門. 東京大学出版会（2001）