シソーラスと確率文法による派生語解析

市丸夏樹[†] 中村貞吾[††]
宮本義昭[†††] 日高達[††††]

派生語は語幹を成す名詞と接尾語から作られる．接尾語には同音語が多く存在し，派生語の解析時に多くの解候補が得られる．また，派生語はあまりに膨大で，その全てを収集し辞書に登録することは困難である．そのため派生語は仮名漢字変換の失敗の原因として大きな割合を占めていた．我々は用例に基づく手法と確率文法による統計的な手法を組み合わせることで，この問題に対処することを試みた．この方法では，大規模コーパスから収集した大量の派生語用例を直接，生成規則の形で記述する．更に単語間の上位下位関係を記述したシソーラスも同様に，各ノードを構文的カテゴリーとして取り扱うことにより，生成規則として記述される．生成規則の適用確率は用例から統計的に学習される．用例としては機械可読辞書と新聞記事コーパスから抽出したものを用いている．用例の頻度にはシソーラス中の用例の分布に従った重み付けを行った．確率文法の学習には，それらの用例に加え，語幹を上位語で置換した用例も使用し，文法の生成能力を拡張した．このことによって，解析候補に対して用例との類似性や頻度を反映した優先付けを行うことができ，未登録語の9割以上を上位10位以内の解候補として受理することができるようになった．最後に，大量データに対する仮名漢字変換実験を行い，用例の一般化によって正解率が大幅に向上することを確認した．

A Morphologic Analysis for Japanese Derivatives Based on Stochastic Grammar and a Thesaurus

NATSUKI ICHIMARU,[†] TEIGO NAKAMURA,[††] YOSHIAKI MIYAMOTO[†††] and TORU HITAKA [††††]

In Japanese, some nouns and some suffixes are concatenated into derivatives. We found difficulties in typing derivatives using Japanese input method. Because so many suffixes have the same pronunciation that the number of candidates for the morpheme tends to become large. In addition, derivatives are too huge in number to collect all of them and store them into a dictionary. Our solution to this problem is a combination of Example-Based Approach and Stochastic Context Free Grammar. About 20,000 sample derivatives are assembled from corpus texts into production rules. A large scale noun thesaurus is also represented as production rules, by treating each node in the thesaurus as a syntactic category. We extracted these samples from dictionary entries and newspaper articles. Besides raw samples, our grammar learns generalized samples to expand the size of its language. Therefore, our morpheme analyzer can rate the candidates by evaluating similarity to samples and these samples' frequency, and can recognize more than 90% of unregistered derivatives as top 10 answers. In an experiment on *kana-kanji* conversion, we confirmed that generalization of samples brought great improvement in the success rate of conversion.

† 九州大学大学院総合理工学研究科情報システム学専攻
Information Systems, Interdisciplinary Graduate School of Engineering Science, Kyushu University
†† 九州工業大学情報工学部知能情報工学科
Department of Artificial Intelligence, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
††† 日本ユニシス株式会社東京ベイ開発センター
Tokyo Bay Development Center, Nihon Unisys
†††† 九州大学工学部情報工学科
Department of Computer Science and Communication Engineering, Faculty of Engineering, Kyushu University

1. 序　論

単語間に区切りのない日本語を解析対象とする形態素解析は，仮名漢字変換や機械翻訳等の自然言語処理システムの性能に多大な影響をもたらす．派生語は接尾語の読みが助詞と紛らわしく，同音異義語が多いため，従来仮名漢字変換の失敗の原因として大きな割合を占めていた．本稿では，名詞と漢字一文字の接尾語（または造語単位）が連接して生成される派生語の，語幹の意味を考慮した形態素解析について述べる．

従来の形態素解析では派生語の読みが入力される際に次のような方法がとられてきた．その1つは辞書中の語と一致するもののみを受理する方法である．これは多くの市販の仮名漢字変換システムで採用されている．また2番目は接尾語と連接可能な名詞を数十種の意味分類を用いて判別する方法[3]である．

前者の方法では，語幹の解釈候補数を a，接尾の候補数を b とすると，派生語としての全部の組み合わせの数は ab 個になる．そこでユーザが語幹と接尾を別々に変換・確定する場合を想定すると，最悪の場合 $a+b$ 回の変換が必要となる．また派生語を収集して辞書に登録しようとしても，派生語の総数は非常に膨大なためすべてを網羅することは難しい．

一方後者の方法では，意味分類によって妥当だと判定された組み合わせに限定される．しかし同じ分類中に語幹が $a'(\leq a)$ 通り，接尾が $b'(\leq b)$ 通り存在する場合には，派生語としての解析候補の数が a' と b' の積となる．このため意味分類が粗い場合には解候補数がかなり大きくなってしまうことがあった．

従って，細かな意味分類を採用することによって語幹と接尾語の組み合わせを意味的に妥当なものだけに絞り込み，連接の妥当性を評価して信頼性の高い解候補を優先する方法が望まれていた．

近年，計算機の高速化と，機械可読辞書や大規模コーパスが利用できるようになったことから，用例に基づく言語処理 (Example-Based Approach[7]) についての研究が盛んになってきている．用例に基づく手法は，大量の用例をコーパスから収集しておいて，シソーラス等の知識を利用しながら，用例との類似性を基準として解候補に優先付けを行うものであり，言語現象の多様性や慣用性に有効に対処することができるものと期待される．さらに我々は，確率文法に基づく統計的言語処理の手法と組み合わせ，テキストコーパスから大量の派生語用例を抽出して統計的な学習を行う方法を採用した．ここで，確率文法とは，文法の生成規則にその規則が適用される確率を付与したものである．生成規則の適用確率は，大量の用例から自動的に学習される．導出に使用する生成規則の適用確率の積を求めることによって，導出木の生起確率が求められ，これを解析候補に対する優先度として利用することができる．

本手法の特徴を挙げる．まず第1に，シソーラスと用例は確率文脈自由文法で表現される．選択制約を生成規則に直接記述する，いわゆる Semantic Grammar の考え方に従い，各シソーラスノードは階層的な構文カテゴリーとして扱われ，派生語用例はシソーラスノードと接尾語の連接として捉えられる．第2に，用例の語幹部分を上位語で置換（一般化）し，その上位語を推移的にシソーラス中の下位語で置換する．これにより，受理可能な単語数が増加し，用例辞書に直接含まれない派生語の多くを網羅できるようになる．第3に，確率文法に基づいて解候補の生起確率を計算し，優先付けを行う．このことによって，より派生語らしく，より使用頻度が高いと推定される導出木を優先解とすることができる．本稿では更に，用例の多義性の絞り込みや，大量の派生語データに対する仮名漢字変換実験について述べる．この実験によって，一般化した用例で学習した場合には，未登録語に対する仮名漢字変換の上位3位までの正解率が72.0%程度まで向上し，未登録語の90%以上を10位までの解として受理できることが示された．

2. 確率文法概説

用例を用いて自然言語を解析するためには，用例の生起頻度の大きさや，シソーラスを辿って派生した解候補と用例との類似性を評価して優先付けを行うことが必要である．ここでは，確率文法に基づいて解候補の導出木の生起確率を計算し，評価値として使用することを考える．学習サンプルの頻度が大きく導出ステップ数が少ないほど，導出木の生起確率は大きくなる傾向があるため，生起確率の大きな解析結果を優先解として判定できる．また，生成規則の適用確率は用例からの学習によって統計的に計算されうるという利点がある．本章では，確率文法について概説する．

2.1 確率文脈自由文法

確率文法 (Stochastic Grammar[1]) は，通常の文脈自由文法の生成規則 $\alpha\rightarrow\beta$ に，α からの書き換えが生起するという条件の下で α から β への書き換えが生起する確率 $\Pr(\alpha\rightarrow\beta)$ を付与したものである．

ここで，集合 $\boldsymbol{X}$ の要素の任意の長さの連接によって作られる記号列の全体集合 (Kleene 閉包) を $\boldsymbol{X}^*$ と書くものとする．

定義 2.1（確率文脈自由文法） 確率文脈自由文法は，5項組 $G_s=(\boldsymbol{V}, \boldsymbol{\Sigma}, \boldsymbol{P}, S, p)$ で表される．ただし，$\boldsymbol{V}, \boldsymbol{\Sigma}$ はそれぞれ非終端記号，終端記号の有限集合で $\boldsymbol{V}\cap\boldsymbol{\Sigma}=\phi$（空集合）であり，$\boldsymbol{P}$ は生成規則の有限集合で $\boldsymbol{P}\subseteq\{A\rightarrow\alpha | A\in\boldsymbol{V}, \alpha\in(\boldsymbol{V}\cup\boldsymbol{\Sigma})^*\}$，を満たし，$S$ は開始記号 $S\in\boldsymbol{V}$，p は生成規則から確率値への写像 $\boldsymbol{P}\mapsto(0, 1]$ である．ここで，任意の生成規則 $A\rightarrow\alpha\in\boldsymbol{P}$ に対し，$p(A\rightarrow\alpha)$ は規則 $A\rightarrow\alpha$ の適用確率，すなわち A が α へと書き換わる確率を表し，$0\leq p(A\rightarrow\alpha)\leq 1$，$\sum_{\gamma\in\{x | A\rightarrow x\in\boldsymbol{P}\}} p(A\rightarrow\gamma)=1$ を満たす．

なお，確率文脈自由文法 G_s から写像 p を取り除くことにより，通常の文脈自由文法 $G=(\boldsymbol{V}, \boldsymbol{\Sigma}, \boldsymbol{P}, S)$ を構成できる．

確率文脈自由文法により形成される導出木の生起確率は，導出に使用する確率生成規則の適用確率の積で与えられる．確率文脈自由文法 G_s における開始記号 S から終端記号列 $x\in\boldsymbol{\Sigma}$ への導出 $S\overset{*}{\underset{G_s}{\Rightarrow}}x$ によって導出木 t が形成されるとき，生成規則 $r\in\boldsymbol{P}$ を使う回数を $c(r, t)$ と書く．すると，導出木 t の生起確率 $\Pr(t)$ は，$\Pr(t)=\prod_{r\in\boldsymbol{P}}p(r)^{c(r,t)}$ となる．

確率文法を用いた解析は，入力記号列 $\sigma\in\boldsymbol{\Sigma}^*$ に対応する導出 $S\overset{*}{\underset{G_s}{\Rightarrow}}\sigma$ における導出木 $t_1, t_2, \cdots, t_n$ を求め，生起確率 $\Pr(t_i)$ の降順に出力する問題となる．また，確率文法を仮名漢字変換に応用する場合は，変換後の文字列の生起確率を求めその降順に出力することになる．各導出木 t_i から得る漢字または正書法の出力用の文字列を $k(t_i)$ とおくと出力文字列 $k'\in\{k(t_i)|i=1, 2, \cdots, n\}$ の生起確率 $\Pr(k')$ は，$\Pr(k')=\sum_{t\in\{t'|k(t')=k'\}}\Pr(t)$ となる．

2.2 文法規則の適用確率の学習

確率生成規則の適用確率を設定する方法としては，Inside-Outside アルゴリズム等，様々な方法が用いられてきた[9]．ここでは非終端記号数が7万個を超えるため，複雑な繰り返し計算は用いず，テキストコーパスから収集した大量のサンプルデータから直接，CKY パージングによる方法[9]に準じた初期確率を与えるものとする．

ただし本研究では学習用のサンプルデータとして，センテンス（終端記号列）や括弧付きテキスト[6]ではなく，導出木を用いる．その理由は，1つの終端記号列に複数の導出木が対応する場合にシソーラスを利用して望ましい木構造を選び出す，非統計的な手法を併用することによって，我々の期待する文法をより正確に構成するためである．この手法については 4.2.2 項で述べる．

確率生成規則の適用確率は，以下のように計算される．生起頻度情報付きのサンプルデータを $\boldsymbol{T}=\{(t_i, f_i)|1\leq i\leq M\}$ とおく．ただし t_i は与えられた文脈自由文法 G における学習サンプルの導出木で，f_i はその頻度情報である．また M は導出木の異なり個数である．確率文脈自由文法 G_s の生成規則 $A\to\alpha\in\boldsymbol{P}$ の適用確率 $p(A\to\alpha)$ は，

$$p(A\to\alpha)=\frac{\sum_{i=1}^{M}f_i\cdot c(A\to\alpha, t_i)}{\sum_{\gamma\in\{x|A\to x\in\boldsymbol{P}\}}\sum_{j=1}^{M}f_j\cdot c(A\to\gamma, t_j)}, \quad (2.1)$$

と求められる．ただし，$c(A\to\alpha, t_i)$ は導出木 t_i の導出に生成規則 $A\to\alpha\in\boldsymbol{P}$ を使用する回数である．こうして求めた適用確率は，サンプル導出木集合の生起確率の総積 $\prod_{i=1}^{M}\Pr(t_i)^{f_i}$ を最大にすることが知られている[4]．

3. シソーラスによる派生語文法

3.1 語幹と接尾の連接性

派生語は名詞と接尾語との連接によって作られる．各接尾ごとに連接できる語幹はまちまちであり，構文的な手がかりで規則性を把握することは難しい．例えば，接尾語「展」は「作品」と連接し派生語「作品展」を成す．また，「絵画」や「油絵」等，「作品」の多くの下位（具象的）概念とも連接可能である．このように，「展」は物のなかでも特に芸術作品や希少価値のある自然物という意味を持つ名詞と結び付きやすい．そこで我々は各接尾と結び付く語幹の意味的な傾向に着目した．このような語幹と接尾の意味的な結び付きを捉えるためには，語幹の名詞を意味的に細かく分類した知識が必要とされる．

シソーラスには単語間の様々な関係が記述されている．そのうち上位下位関係は，抽象的な概念を表わす単語を上位に，そして対応する具体的な単語を下位に階層的に記述したものであり，**図1**のような有向非周回グラフ（DAG）で表される．したがって，互いに類似した単語はシソーラス中で，兄弟ノードやその下位語付近に位置することになる．シソーラスの各ノードはその下位語を包括する概念を成し，単語を意味的に細分類したカテゴリー名として捉えられる．これは，丁度名詞という構文的なカテゴリーを更に小さなカテゴリーに分割していったことに相当する．そこでこの

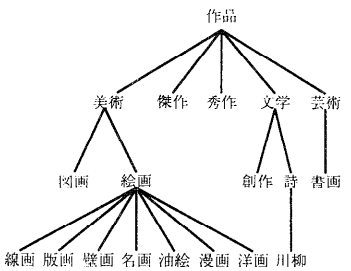


図1 名詞のシソーラス（「現代日本語名詞シソーラス」より抜粋）．
Fig. 1 A fragment of a thesaurus.

シソーラスノードを構文的カテゴリーと同一視することによって，シソーラスを確率文脈自由文法の枠組で記述することを試みる．

なお，各々のシソーラスノードは，表記文字列とその読み仮名の組に語義番号を付加することによって区別されている．従って，同音同表記で異なった意味を持つ単語は，シソーラス中では別々のノードとして取り扱われる．

3.2 確率派生語文法

派生語用例とシソーラスを結びつけて単一の文法で記述するためには，用例の語幹部をシソーラスノードと結び付ける必要がある．そしてシソーラス中の全部の単語を導出し，更に，用例そのものに加え，用例から派生した語，すなわち語幹を意味カテゴリー内の別の単語で置き換えた語を導出できなければならない．

定義 3.1（確率派生語文法） 確率派生語文法とは，以下のような5種類のテンプレートで規定される生成規則の集合 $\boldsymbol{P}$ を持つ確率文脈自由文法 $G_s=(\boldsymbol{V}, \boldsymbol{\Sigma}, \boldsymbol{P}, S, p)$ である．

$$N \rightarrow H|W_0, \quad (1)$$
$$H \rightarrow W\,B, \quad (2)$$
$$W \rightarrow W', \quad (3)$$
$$W \rightarrow w, \quad (4)$$
$$B \rightarrow b, \quad (5)$$

ただし，$N, H, W_0, W, W', B \in \boldsymbol{V}$ はそれぞれ名詞のカテゴリー，派生語のカテゴリー，シソーラスの根，シソーラス中のノードの上位語と下位語，そして接尾語のカテゴリーを表す．$w, b \in \boldsymbol{\Sigma}$ はそれぞれ，W, B の表記であり，開始記号 $S=N$ である．

各生成規則について説明する．

（1） 名詞が派生語を生成する．また，シソーラスの頂点から規則（3）を辿ってシソーラス中の任意の単語を生成する．

（2） 派生語が語幹を成す単語と接尾語を生成する．さらにこの語幹部から規則（3）を辿って下位語を生成可能である．この型の規則は大量の派生語用例から構成されるもので，学習した用例数だけ存在する．「H→作品展」等．

（3） シソーラス中の上位語が下位語を生成する．この規則はシソーラス中の上位下位関係と1対1に対応する．「作品→絵画」等．

（4） シソーラス中の単語がその綴りを生成する．この型の規則はシソーラス中の全単語について作成される．「作品→'さくひん'」等．

（5） 接尾語がその綴りを生成する．この型の規則は接尾語辞書から作成される．「展→'てん'」等．

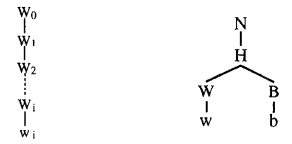
このように確率文法でシソーラスを利用することは，純粋に統計的な手法（マルコフモデル等）からは逸脱したものである．しかし，学習した用例を忠実に反映する文法では，逆に未知語を取り扱えない．我々の文法は，解析結果の導出木の生起確率の計算に厳密さを追い求めることよりも，用例そのものに加え，シソーラスを辿って用例に似て非なる派生語を導出することを目指したものである．

3.3 確率派生語文法における生成規則の適用確率の学習

確率生成規則の適用確率の学習サンプルとしては，シソーラスの頂点からシソーラス中の全単語への導出木と，派生語用例の導出木を用いる（**図2**参照）．各生成規則の適用確率は，これらの木構造とその頻度情報が与えられれば（2.1）式により計算される．

シソーラスの頂点から各ノードを導出する導出木についてすべて同じ生起頻度を設定して学習すると，それらの導出木の生起確率はすべて等しくなる（**図3**）．それに続いて図2（b）の形をした派生語の用例の導出木を学習していくと，（4）型の生成規則 $A \rightarrow a$ をすべての学習導出木の導出に使用する回数の総和が大きくなり，それにより（2.1）式の分母の値が大きくなる．このような確率値の正規化の作用により，（3）型の生成規則 $A \rightarrow B$ の適用確率の値が相対的に小さくなる．以上のような学習によって，導出の際にシソーラスを辿る段数が多いほど生起確率が小さくなっていくのである．

ところで，この方法には，シソーラス中で用例語幹



(a) シソーラスノード (b) 派生語サンプル

(a) A thesaurus node. (b) A sample derivative.

図2 学習用導出木

Fig. 2 Derivation trees for learning.

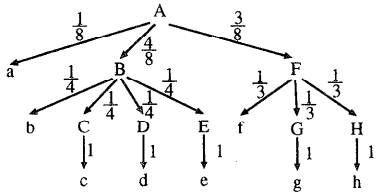


図3 派生語用例語幹の存在しない部分のシソーラスノードへの導出の道筋

Fig. 3 Production paths to each thesaurus node which is connectable to no suffix.

が付近に現れない部分ではノードの表記を表す終端記号の生起確率がどれも等しいままとなってしまうという欠点がある．従って，生起確率の学習には十分な量の用例を確保することが必要である．

4. 用例からの文法規則の構成

前章までに，学習用の派生語の導出木が与えられれば，確率派生語文法を構成できることが示された．次に，テキストコーパスから収集された派生語の用例とシソーラスから，生成規則とサンプルデータの導出木を組み立てる方法について述べる．

シソーラスノードの集合を $\boldsymbol{W}$，接尾語の集合を $\boldsymbol{B}$，そのそれぞれの表記（仮名漢字変換の場合は読み）の集合を w, b と記す．すると，派生語文法 $G=(\boldsymbol{V}, \boldsymbol{\Sigma}, \boldsymbol{P}, N)$ と，確率派生語文法 $G_s=(\boldsymbol{V}, \boldsymbol{\Sigma}, \boldsymbol{P}, N, p)$ における非終端記号の集合 $\boldsymbol{V}$ と終端記号の集合 $\boldsymbol{\Sigma}$ は，$\boldsymbol{V}=\boldsymbol{W}\cup\boldsymbol{B}$, $\boldsymbol{\Sigma}=w\cup b$ となる．規則（3）型の生成規則は，シソーラスノード $W, W'\in\boldsymbol{W}$ について，W が W' の上位語であるならば $W\rightarrow W'\in\boldsymbol{P}$，とすることで構成できる．

ここで，集合 $\boldsymbol{X}$ の要素数を $\#\boldsymbol{X}$ と書く．また，語義番号付きのシソーラスノード $W_1, \cdots, W_n\in\boldsymbol{W}$ が共通した表記文字列とその読み仮名列の組 s を持つとき，$\boldsymbol{nodes}(s)=\{W_1, \cdots, W_n\}$ と書くものとする．

4.1 用例の収集

本稿では，あらかじめ形態素解析済みのテキストコーパスと辞書の見出し語から，派生語の表記文字列とその読み情報を抽出しているものとする．収集した派生語文字列のうち，語幹の表記文字列とその読み仮名がシソーラスノードと一致し，かつ接尾部分が接尾語辞書に含まれるものを選び，次のような形式の派生語原用例データを構成する．

$$\boldsymbol{E}\subset\{(s, B, f)|\boldsymbol{nodes}(s)\cap\boldsymbol{W}, B\in\boldsymbol{B}, f>1\}. \tag{4.1}$$

ここで s は収集した派生語文字列の語幹部の表記文字列とその読み仮名の組，B は接尾語の1つを表す非終端記号，f は派生語データ "sB" の生起頻度である．

4.2 派生語用例の語幹部とシソーラスノードの対応付け

派生語原用例 $(s, B, f)\in\boldsymbol{E}$ の語幹部分 s と同じ表記と読みを持つシソーラスノードは，$\boldsymbol{nodes}(s)$ により得られる．しかし，テキストコーパスから収集したばかりの派生語原用例の語幹部分の単語には，同音同表記の異義語が多数存在するかもしれない（図4）．そのうち本来派生語として妥当な語義は幾つかに限られる．そこでシソーラスと用例自体を利用して，意味的に信頼性が高いと推定される語義ノードを優先することを考える．その手順は以下のようになる．

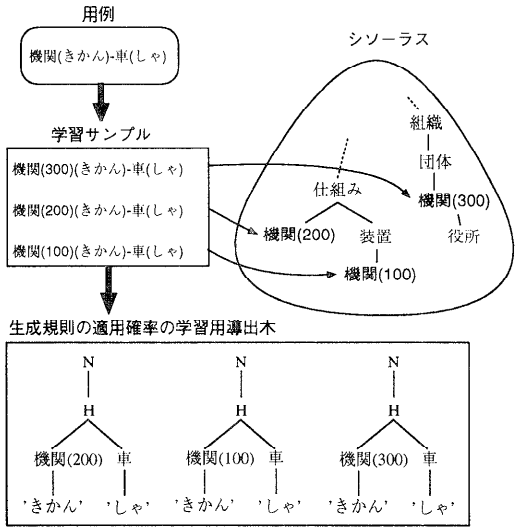


図4 派生語用例への語義番号の設定と学習用導出木の例
Fig. 4 Identification of sample derivatives and forming derivation trees.

1. 原用例 "sB" の頻度を $\boldsymbol{nodes}(s)$ の各要素を語幹として持つ派生語データ（学習サンプル）に等分する．
2. そうして得られた学習サンプルのシソーラス中の分布を調べ，お互いに近くに寄り集まっているものに優先的に大きな頻度を与える．

ここで学習サンプルとは，2.2節で説明した，図2の形の導出木 t とその頻度情報 f の組 (t, f) の集合である．以下では簡単のために，語幹 W と接尾 B から一意的に決まる図2(b)の形の導出木を $W-B$ と略記する．

4.2.1 用例の頻度を等分したサンプル

原派生語用例集合 $\boldsymbol{E}$ に含まれる各派生語データ "sB" の頻度 f' を，語幹 s に対応する語義集合 $\boldsymbol{nodes}(s)$ の要素ごとに等分したサンプル集合 $\boldsymbol{eq}(\boldsymbol{E})$ は，

$$\boldsymbol{eq}(\boldsymbol{E})=\left\{(W-B, f)|\exists sf'.\ W\in\boldsymbol{nodes}(s),\ (s, B, f')\in\boldsymbol{E}, f=\frac{f'}{\#\boldsymbol{nodes}(s)}\right\},$$

と表わされる．

ここで，任意のサンプル集合 $\boldsymbol{S}$ における派生語サンプル $W-B$ の頻度を

$$fs(W\quad B)=\begin{cases}f & \text{if}(W-B, f)\in\boldsymbol{S},\\ 0 & \text{otherwise},\end{cases}$$

と書くものとする．

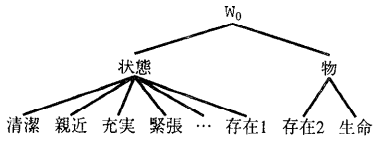


図5 接尾語「感」と連接するシソーラス中の単語の例
Fig. 5 A thesaurus that consists of words connectable to a suffix.

4.2.2 サンプル頻度の重み付け

派生語を成す接尾語は特定の意味を持つ語幹を好む．よって，特定の接尾語と連接可能な語幹のシソーラス中の分布に着目すると，孤立した語義よりも，用例の語幹が付近に密集している語義の方が，用例の解釈として信頼性が高いものと考えられる．そこで用例の語幹を意味カテゴリーごとにまとめて考え，より多くの用例を含んだカテゴリーに属する語義を用例の解釈として優先することができる．

用例 $\boldsymbol{E}$ に含まれる派生語 "sB" の語幹部 s に n 個のシソーラスノード $\boldsymbol{nodes}(s)=\{W_1,\cdots,W_n\}$ が対応している場合を考える．シソーラスノード W_i と同じ意味カテゴリー中に位置する単語の集合を $\boldsymbol{c}(W_i)$ と書き，

$$\boldsymbol{c}(W_i)=\{W' \mid \exists W''.\ W''\rightarrow W_i\in\boldsymbol{P},\ W''\overset{*}{\underset{G}{\Rightarrow}}W',\ W'\in\boldsymbol{V}\},$$

とおく．まず，$\boldsymbol{c}(W_i)$ 中に語幹が含まれるすべての派生語サンプルについて，サンプル $\boldsymbol{eq}(\boldsymbol{E})$ の下での頻度の総和を求める．するとその各々が和集合 $\bigcup_{i=1}^{n}\boldsymbol{c}(W_i)$ 中の頻度の総和に占める割合を計算できる．これを，元々の用例 $\boldsymbol{E}$ の頻度に掛け，派生語サンプルの頻度の目安とすることができる．こうしてカテゴリー $\boldsymbol{c}(W_i)$ 中にサンプル頻度の多い語義 W_i を優先して重み付けできる．すなわち，この派生語サンプルを $\boldsymbol{cat}(\boldsymbol{E})$ とすると，

$$\boldsymbol{cat}(\boldsymbol{E})=\left\{(W-B,f) \mid \exists sf'.\ W\in\boldsymbol{nodes}(s),\ (s,B,f')\in\boldsymbol{E},\ f=\frac{\sum_{W'\in\boldsymbol{c}(W)} f_{\boldsymbol{eq}(\boldsymbol{E})}(W'-B)}{\sum_{W''\in\bigcup_{W'\in nodes_{(s)}}\boldsymbol{c}(W')} f_{\boldsymbol{eq}(\boldsymbol{E})}(W''-B)}f'\right\},$$

と求められる．

ここで頻度計算の例を挙げる．シソーラスが図5のようになっており，(存在−感, 1)$\in\boldsymbol{E}$ という用例が得られ，接尾「感」と連接する語幹単語が，それを含めて存在$_1$ の下位に 80 語，存在$_2$ の下位に 2 語ある場合，$f_{\boldsymbol{cat}(\boldsymbol{E})}$(存在$_1$−感)$=80\cdot 1/82\simeq 0.98$, $f_{\boldsymbol{cat}(\boldsymbol{E})}$(存在$_2$−感)$=2\cdot 1/82\simeq 0.02$ となる．よって派生語「存在感」の語幹としては，「物」の下位語の「存在$_2$」よりも「状態」を表す「存在$_1$」の方が優先されることになる．解析時にはこのようにして派生語を成す語幹部分の語義が確定される．この手法は自然言語理解や機械翻訳等の応用においても，派生語語幹の解釈の曖昧さの絞り込みに役立つものと思われる．

4.3 派生語サンプルの一般化

5章で述べるように大量データに対する仮名漢字変換実験を行ったところ，上記のサンプルをそのまま使用して学習した際には，未知語をほとんど受理できないことが判明した．そこで我々はまずサンプルデータを増加させようとした．しかし，新たに用例を収集してもそのほとんどは元の用例と重複した派生語になってしまい，新規の派生語を獲得するには莫大な量のテキストが必要となることがわかった．そのため現在保有している用例を利用して生成能力を拡張することを考えた．

シソーラスの中で類義語は，DAG 構造中の親ノードを共有する兄弟ノードに相当する．したがって我々は，用例の語幹部を上位語で置き換えた派生語を，仮想的な学習サンプルとして使用することにした．こうして作成した一般化サンプルそのものは必ずしも派生語として妥当でないものの，語幹と接尾の妥当な組み合わせの多くを生成するためには不可欠である．

頻度等分サンプル集合 $\boldsymbol{eq}(\boldsymbol{E})$ またはカテゴリー内頻度によって優先付けされたサンプル集合 $\boldsymbol{cat}(\boldsymbol{E})$ を $\boldsymbol{S}$ とおくと，$\boldsymbol{S}$ の一般化サンプル $\boldsymbol{g}(\boldsymbol{S})$ は，

$$\boldsymbol{g}(\boldsymbol{S})=\{(W-B,f) \mid \exists W'f'.\ W\rightarrow W'\in\boldsymbol{P},\ (W'-B,f')\in\boldsymbol{S},(\forall f''.(W-B,f'')\notin\boldsymbol{S}),\ f=C\sum_{W'\in\boldsymbol{k}(W)} fs(W'-B)\},$$

となる．ここで，$\boldsymbol{k}(W)$ はシソーラスノード W の下位語の集合，$\boldsymbol{k}(W)=\{W' \mid W\rightarrow W'\in\boldsymbol{P}\}$，であり，定数 C は下位語を語幹として持つ用例の頻度の総和と上位語を語幹として持つ用例の頻度の比を表し，

$$C=\frac{1}{\#\boldsymbol{t}(\boldsymbol{S})}\sum_{W-B\in\boldsymbol{t}(\boldsymbol{S})}\frac{fs(W-B)}{\sum_{W'\in\boldsymbol{k}(W)} fs(W'-B)},$$

ただし，$\boldsymbol{t}(\boldsymbol{S})=\{W-B \mid \exists W'.\ W\rightarrow W'\in\boldsymbol{P}, fs(W'-B)>0\}$，である．

この頻度の設定の方法としては，他にも様々な方法を試みた．しかしその結果にあまり差はなく，頻度が登録語に比較して十分小さければ実用上特に問題はないものと思われる．

さらに，この $\boldsymbol{g}(\boldsymbol{S})$ 自体を $\boldsymbol{g}()$ に再帰的に適用して累積していくことにより，n 段階一般化したサンプル

を得ることができる．これを $\boldsymbol{g}^n(\boldsymbol{S})$ と書くと，$\boldsymbol{g}^0(\boldsymbol{S})=\boldsymbol{S}, \boldsymbol{g}^i(\boldsymbol{S})=\boldsymbol{g}(\boldsymbol{g}^{i-1}(\boldsymbol{S}))\cup\boldsymbol{g}^{i-1}(\boldsymbol{S})$ $(i\geq 1)$ と定義される．

このように一般化したサンプルを用いる利点として，以下のようなことが挙げられる．まず第1に，一般化後の仮想的な派生語サンプルは，確率文法における通常の学習サンプルとして使用できる．そして第2に，一般化したサンプルを用いれば，類義語を検索する際に語幹部の親ノードから下位語を探索する必要がないため，計算量を増大させることなく生成能力を拡張することが可能である．

4.4 生成規則の構成

最後に，派生語サンプルから確率文脈自由文法 G_s を構成する．原用例の頻度を等分または優先付けした分配を行った派生語サンプル S を，i 段一般化したサンプル $\boldsymbol{g}^i(\boldsymbol{S})$ が与えられたとき，$(W-B, f)\in\boldsymbol{g}^i(\boldsymbol{S})$ の各々について生成規則 $H\rightarrow WB$ を $\boldsymbol{P}$ に加え，文脈自由文法 G における導出木 $W-B$ を図2 (b) の形に構成し，$(W-B, f)$ をサンプルデータ $\boldsymbol{T}$ に加える．こうして構成した $\boldsymbol{T}$ から，生成規則の適用確率 p を (2.1) 式により設定する．以上により，確率派生語文法が完成する．

5. 派生語の仮名漢字変換実験

確率派生語文法がどの程度有効かを確認するために，大量データと大規模なシソーラスを使用して実験を行った．派生語は読み仮名で表現された時に特に曖昧さが増大するので，応用形態として仮名漢字変換を想定した．

5.1 使用データ

シソーラスとしては筑波大学で作成された「現代日本語名詞シソーラス[8]」を使用した．ノード数は71,915，上位下位関係数は75,028である．確率の学習の際，図2 (a) 型の木構造を構成し，その頻度としては簡単のため一律に約2.6を設定した．この値は日経新聞コーパスから抽出した派生語と，同じコーパスに含まれるシソーラス中の単語の頻度の総和の比に，派生語学習サンプルの頻度の和を掛けたものを，全シソーラスノードに等分したものである．

派生語用例としては，(A) 九大公用データベース日本語単語辞書の見出し語から抽出した異なり語数で24,509語，および，(B) 日本経済新聞記事1982年1〜3月の形態素解析済みデータからの10,315語を (4.1) 式の形式に従って収集した．データ (B) の頻度の総和は49,543であった．これをランダムに10個に分割したものを，$\boldsymbol{R}_1, \cdots, \boldsymbol{R}_{10}$ とおく．データ (A) の方は頻度情報が付属していないため各派生語の頻度を1と設定した．これは $\boldsymbol{D}$ とおく．

用例の増加による正解率の変動を調査するため，$\boldsymbol{E}_0=\boldsymbol{D}, \boldsymbol{E}_i=\boldsymbol{E}_{i-1}\cup\boldsymbol{R}_i$ $(i=1..5)$ の6種類を学習サンプルとして準備した．これらは辞書のみから始めて，新聞記事に含まれる派生語用例をほぼ9日分ずつ学習サンプルに加えていったものとなっている．

試験データは，$\boldsymbol{I}=\bigcup_{i=6}^{10}\boldsymbol{R}_i$，とした．これは学習サンプルと似た頻度傾向を持つ別の派生語データで，新聞記事にして約1か月半の分量に相当する．

5.2 実験手順

用例の添字 $j=0, \cdots, 5$，一般化段数 $i=0, \cdots, 3$ のすべての組み合わせについて以下の手順にしたがって実験を行った．

1. 用例 $\boldsymbol{E}_j$ を収集し，4章にしたがって $\boldsymbol{eq}(\boldsymbol{E}_j)$, $\boldsymbol{cat}(\boldsymbol{E}_j)$，そして i 段一般化したサンプル $\boldsymbol{g}^i(\boldsymbol{eq}(\boldsymbol{E}_j))$, $\boldsymbol{g}^i(\boldsymbol{cat}(\boldsymbol{E}_j))$ を作成した．これから3.2節のような確率派生語文法を構成し，図2の形の導出木を用いて生成規則の適用確率を (2.1) 式により計算した．

2. 派生語の表記と読み仮名と頻度からなる3種類のデータ，(a) ランダムに選択した試験データ $\boldsymbol{I}$，(b) $\boldsymbol{I}$ 中の未登録語 $\boldsymbol{I}-\boldsymbol{E}_j$，(c) 学習サンプル $\boldsymbol{E}_j$ 自体，について，読み仮名の部分を変換の入力，表記の部分を正解データ，頻度情報の部分を入力の回数として仮名漢字変換を行った．ここで，入力された1つの読み仮名に対して幾つかの変換候補がその生起確率とともに得られる．それらの解候補を生起確率の大きな順にソートして正解の順位 n を求め，第 n 位までには必ず正解が得られる入力の生起回数として集計した．

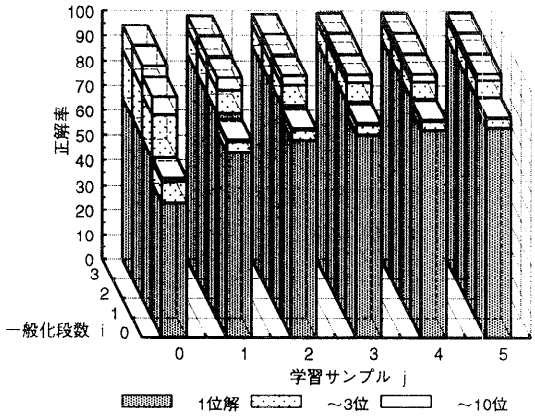
ただし本手法で未登録語を取り扱うためには，語幹と接尾の両方が辞書に登録されている必要がある．そのため (b) の未登録語としては，語幹部の名詞がシソーラスに含まれていて，かつ接尾語の部分が接尾語辞書に登録されており，派生語として妥当であるが，学習サンプルに使用した用例には含まれないような派生語データを使用している．

5.3 結果と考察

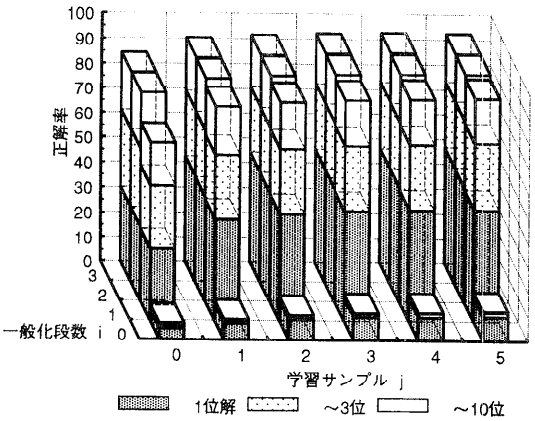
以上のような実験を行い，第 n 位までの正解率（正解の順位が n 以下となった入力の頻度の和が，入力の頻度の総和に占める割合）を求めた．このグラフを**図6**(a)〜(c)に示す．それぞれの実験の入力の総数は**表1**のようなものである．

5.3.1 用例増加の効果

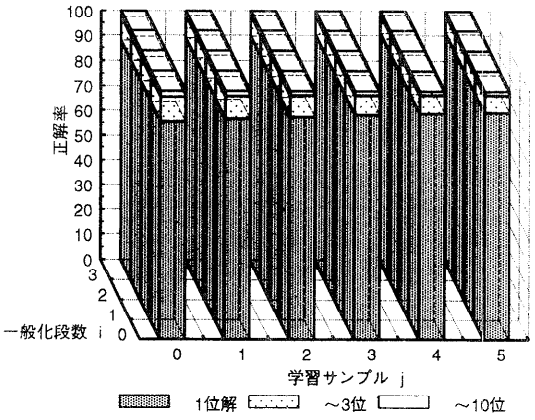
用例を $\boldsymbol{E}_0$ から $\boldsymbol{E}_5$ へと増加するにつれて，ランダム試験データ $\boldsymbol{I}$ を変換入力とした実験（図6(a)）の正解



(a) 入力：I



(b) 入力：$I-E_j$



(c) 入力：E_j

図 6 サンプル $g^i(cat(E_j))$ で学習した際の仮名漢字変換の正解率

Fig. 6 The success rates of *kana-kanji* conversion, where the sample for learning is $g^i(cat(E_j))$.

表 1 実験の入力として使用された派生語用例

Table 1 The input derivative data for experiment.

用例	異なり語数(頻度)	用例	異なり語数(頻度)
I	11771 (24728)		
E_0	24509 (24509)	$I-E_0$	6582 (9352)
E_1	25859 (29454)	$I-E_1$	4770 (5355)
E_2	26842 (34399)	$I-E_2$	3973 (4267)
E_3	27689 (39344)	$I-E_3$	3501 (3701)
E_4	28384 (44289)	$I-E_4$	3165 (3308)
E_5	29052 (49234)	$I-E_5$	2916 (3030)

率は次第に上昇している．一方 I 中の未登録語についての場合（図 6(b)）は変化が緩やかである．従って，用例を増加することによって生じる正解率の変動は，入力 I のうち登録語 $I \cap E_j$ が占める割合の増加を反映したものと思われる．

用例 E_i は E_{i-1} に一定量の用例を加えたものであるけれども，異なり語数の増加分は表 1 のように i が大きくなるに従って小さくなっていく．これは用例が増すごとに重複分が大きくなるためである．そのためこれ以上用例を増加させても正解率の向上幅は次第に小さくなり，図 6(c) 程度の正解率に漸近的に近づいていくものと考えられる．

5.3.2 一般化の効果

一般化を行ったサンプルで学習した場合，図 6(b) に示すとおり，認識可能な派生語数と仮名漢字変換の正解率に大きな改善が見られる．一般化の段数による正解率の伸びは 1 段目が最も大きく，2 段以降は小さくなり 2 段と 3 段とではほとんど違いがなくなる．特に上位 3 位までの正解率は一般化を 1 回行った後はほぼ一定である．よって，一般化の段数は 1 段程度で十分であると考えられる．

また登録語については，一般化した仮想的なサンプルを導入することによって，不正解が増加する懸念があった．しかし，図 6(c) によると，正解率の低下は見られない．よって，一般化サンプルの頻度が十分に小さく設定されており，解候補に対する優先付けが有効に機能しているものと思われる．

5.3.3 用例頻度の重み付けの評価

次に，我々の学習サンプル頻度への重み付け手法が用例の語義を正しく選択できているのかどうかを調査した．まず，288 種類の接尾語からなる派生語用例の一部約 15,000 語（約 30,000 語義）について人手で語義番号を設定した．こうして作成した語義設定の正解と，自動的に重み付けられた用例 $cat(E_5)$ とを比較した．その結果の一部を**表 2** に示す．複数の語義候補の得られる用例のうち優先順位の上位半数に正解が得られる

表 2 語義候補数別に集計した，設定頻度の順位で等 i 位に正解が現れる用例の数
Table 2 The number of sample derivatives whose correct meaning is the ith answer in n candidates.

語義候補数 n	分配頻度の順位 i									
	1	2	3	4	5	6	7	8	9	10
1	4407									
2	1744	371								
3	633	187	100							
4	304	101	67	53						
5	133	61	31	19	7					
6	41	23	13	8	17	5				
7	37	12	8	8	8	5	5			
8	15	4	6	3	3	5	1	0		
9	2	1	2	0	3	1	0	1	2	
10	13	3	1	0	0	2	0	0	2	1

表 3 学習用の用例の頻度を，語義ごとに等分した場合とカテゴリー内の頻度和の比で重み付けして分配した場合の正解率の比較
Table 3 The success rates of *kana-kanji* conversion in which sample for learning is $g^1(eq(E_5))$ or $g^1(cat(E_5))$.

順位	学習サンプル	
	$g^1(eq(E_5))$	$g^1(cat(E_5))$
1	89.44%	89.35%
～3	96.04%	96.09%
～10	98.75%	98.79%

ものが約 80% を占めている．従って，このカテゴリー内の用例頻度を使った手法は，必ずしも最優先解が正解するとは言えないけれども，大量のデータのかなりの部分に対してどちらかというと優先順位の前半に正解を集めており，いわゆる weak method として働いていると言える．

5.3.4 用例頻度の重み付けの効果

4.2.2 項で述べた用例頻度の重み付け手法の仮名漢字変換の正解率に及ぼす効果について調べた．用例語幹の語義に用例の頻度を等分したサンプル $g^1(eq(E_5))$ で学習した場合と，それを重み付けした $g^1(cat(E_5))$ で学習した場合の正解率を**表 3** に示す．これによると，最尤解を除いては重み付けを行った方が若干正解率が向上するものの，両者にほとんど差は現れないことがわかった．すなわち，仮名漢字変換に関する限り，用例の語義の優先付けを行わなくても，行った場合とほぼ同等の正解率を得ることができると言える．

6. 結　論

コーパスから抽出した「語幹名詞 + 接尾語」型の派生語用例と名詞のシソーラスを確率文脈自由文法で記述する方法を提案した．用例そのものに加えて一般化したサンプルを用いることで派生語文法の生成能力を大幅に拡張することができた．

用例に適切な語義番号を設定できれば実行時に必要な用例数の減少が見込まれるため，更に用例の語義の曖昧さを絞り込む手法を開発することが必要となるだろう．今後は本手法を拡張し，複合語も取り扱う予定である．

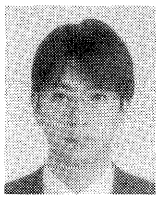
謝辞　本研究では，筑波大学の荻野綱男先生の作成された「現代日本語名詞シソーラス」，九州芸術工科大学の稲永紘之先生が原データを作成された「九州大学大型計算機センター公用データベース日本語単語辞書」，NTT 情報通信網研究所から戴いた日経新聞コーパスを使用しました．また九州大学工学部情報工学科の有馬信宏君，松永　晃君，西村武司君には，用例の語義チェックの作業をして戴きました．皆様に深く感謝いたします．

参考文献

1) Lee, H. C. and Fu, K.-S.: A Stochastic Syntax Analysis Procedure and Its Application to Pattern Classification, *IEEE Trans. Comput.*, Vol. C-21, No. 7, pp. 660-666 (1972).
2) 吉田　将，日高　達，稲永紘之，田中武美，吉村賢治：公用データベース日本語単語辞書の使用について，九州大学大型計算機センター広報，Vol. 16, No. 4, pp. 335-361 (1983).
3) 稲永紘之，新谷隆之：意味的なつながりを考慮した接尾語辞書の作成について，情報処理学会自然言語処理研究会研究報告，Vol. 86, No. 60, NL 57-3, pp. 1-7 (1986).
4) 杉本　洋：接辞の意味的結合性に基づく派生語文法，九州大学大学院総合理工学研究科修士論文，p. 43 (1992).
5) 市丸夏樹，中村貞吾，日高　達：名詞シソーラスを用いた派生語の処理，電子情報通信学会技術研究報告［言語理解とコミュニケーション］，NLC 92-17, pp. 39-46 (1992).
6) Pereira, F. and Schabes, Y.: Inside-Outside Reestimation From Partially Bracketed Corpora, *90th Annual Meeting of ACL*, pp. 128-135 (1992).
7) 隅田英一郎：用例を使った言語処理，ATR Technical Report, TR-I-0374, pp. 33-40 (1993).（人工知能学会研究会資料 SIG-SLUD-9204-4 (2/5)）

8) 荻野綱男：現代日本語名詞シソーラスから見た語彙の意味分類，平成4年度科学研究費助成金一般研究（C）研究成果報告書，p. 69 (1993).
9) 中川聖一：確率・統計的手法による連続音声認識アルゴリズム，電子情報通信学会「コーパスに基づく自然言語処理」講習会，pp. 1-28 (1993).
10) 市丸夏樹，中村貞吾，宮本義昭，日高　達：用例に基づく派生語の確率的解析，情報処理学会自然言語処理研究会研究報告 93-NL-97, pp. 21-28 (1993).
11) Ichimaru, N., Nakamura, T., Miyamoto, Y. and Hitaka, T.: Example-Based Stochastic Analysis of Japanese Derivative Words, *Natural Language Processing Pacific Rim Symposium'93*, pp. 368-371 (1993).

（平成6年10月31日受付）
（平成7年 2 月10日採録）



市丸　夏樹（正会員）

昭和42年生．平成2年九州大学工学部電子工学科卒業．平成4年同大学院総合理工学研究科修士課程情報システム学専攻修了．同博士後期課程情報システム学専攻在学中．修士（工学）．自然言語処理，コンピュータプログラミングに興味を持つ．

中村　貞吾（正会員）

昭和34年生．昭和57年九州大学工学部電子工学科卒業．昭和59年同大学院工学研究科修士課程電子工学専攻修了．昭和62年同博士後期課程電子工学専攻単位取得退学．九州工業大学情報工学部知能情報工学科講師．工学博士．人工知能，自然言語処理，計算言語学に関する研究に従事．電子情報通信学会，人工知能学会各会員．

宮本　義昭（正会員）

昭和30年生．昭和53年東京理科大学理工学部経営工学科卒業．昭和55年同理工学研究科経営工学専攻修士課程修了．同年日本ユニシス株式会社（当時日本ユニバック）に入社．入社時より現在まで，日本語処理・文書処理に関連するソフトウェアの開発・保守に従事．認知科学学会会員．

日高　　達（正会員）

昭和14年生．昭和40年九州大学工学部電子工学科卒業．昭和42年同大学院工学研究科電子工学専攻修士課程修了．昭和44年同博士課程中退．昭和63年九州大学工学部情報工学科教授，現在に至る．工学博士．形式言語の方程式論，自然言語処理，手書き文字認識の研究に従事．電子情報通信学会，人工知能学会各会員．