用例に基づく派生語の確率的解析

3M－7

市丸 夏樹*,　中村 貞吾†,　宮本 義昭‡,　日高 達*

*九州大学,　†九州工業大学,　‡日本ユニシス

1　はじめに

日本語における派生語とは，語幹の単語に接辞や造語成分が連接して元の意味から離れた意味や異なった品詞を持つようになった転成語を指す．

仮名漢字変換等における派生語の解析には，従来，派生語自体を一つの単語として辞書に登録する方法や，接尾の連接可能な語幹を数十程度の意味分類によって判別する方法がとられてきた．しかし，派生語の数は非常に膨大であるため前者の方法では十分な量の派生語を取り扱うことが難しい．また後者を用いると，分類が荒いために正解率の向上があまり望めなかった．

そのため，我々は語幹と接尾の意味的な連接性を派生語用例と名詞のシソーラスを用いて捉えることを考えた [1]．ところが，用例をそのまま学習に用いた場合，語幹と接尾の連接関係を網羅できないため入力派生語の多くが解析不能となってしまうことが判明した．

そこで本稿では，用例の語幹部をシソーラス中の上位語で置き換えて作り出した派生語データ (一般化サンプル) を学習に使用する方法を提案する．その結果，学習サンプルに出現していなかった派生語についての正解率を大幅に向上させることに成功した．以下，この実験の概要と結果について報告する．

2　語幹と接尾の連接性

シソーラスには単語間の上位下位関係が有向非周回グラフ構造として記述されている．シソーラスのノードに当たる単語は，その単語の下位語の集合を包括する概念となっている．接尾語は特定の意味を持つ語幹に選択的に連接するものと考えられるため，シソーラス中の単語を，特定の接尾語との連接性を表わす意味分類として捉えることができよう．これは，名詞「作品」と接尾語「展」が連接して「作品展」という派生語を成し,「作品」の下位語の「美術」「油絵」も「展」と連接してそれぞれ派生語「美術展」「油絵展」を成すという事実により例証される．

2.1　確率派生語文法

言語の解析に用例を用いる場合，用例から派生した解候補や用例そのものを如何に優先付けするかということが重要な問題である．我々は派生語解析に確率文法を導入することにより，用例の頻度情報に基づいた優先付けを自然な形で実現することができた．

確率文法とは，通常の文脈自由文法の生成規則 $\alpha \rightarrow \beta \in \boldsymbol{P}$に，適用確率 $p(\alpha \rightarrow \beta)$ を付与したものである．確率文法により生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．確率文法を用いた仮名漢字変換は，読みに対応する導出木を求め，生起確率の降順に出力する問題となる．

文法規則の適用確率は通常大量のサンプルデータから学習され，その計算方法としては，サンプルデータの生起確率の総和や積を最大にする方法が知られている．前者の方法では生成，学習の繰り返し計算が必要である．しかし我々の文法では莫大な量の派生語を生成してしまうため，この収束計算は実質的に不可能である．従って，サンプルデータの生起確率の総積を最大にする後者の簡便な方法を用いている．

以下に，学習に用いたサンプルの木構造の例と，我々の文法から生成される導出木の例を示す．

Example-Based Stochastic Analysis of Derivatives
Natsuki ICHIMARU*, Teigo NAKAMURA†,
Yoshiaki MIYAMOTO‡, Toru HITAKA*
*Kyushu University, †Kyushu Institute of Technology, ‡Nihon Unisys

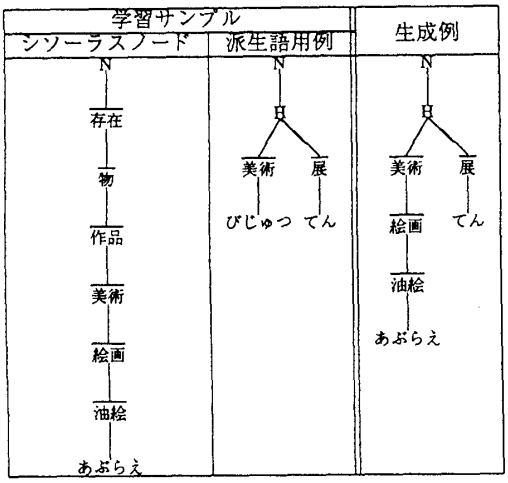


表 1: 学習サンプルと導出木の例

3 仮名漢字変換実験

名詞のシソーラスとして「現代日本語名詞シソーラス」，サンプル派生語データとして，表2のような(○○)＋○型の3漢字語データを使用して確率派生語文法を構成し，仮名漢字変換実験を行なった．

	派生語データ源	単語数	頻度和
A	九大公用データベース	13786	—
B	広辞苑	948	—
C	現代用語の基礎知識 1989 年版	9939	42800
D	日本経済新聞 (1982 年 1〜3 月)	4497	16299
学習入力	A+B(D による頻度付き)	14733	25188
	C(語幹がシソーラスノード)	9190	40379

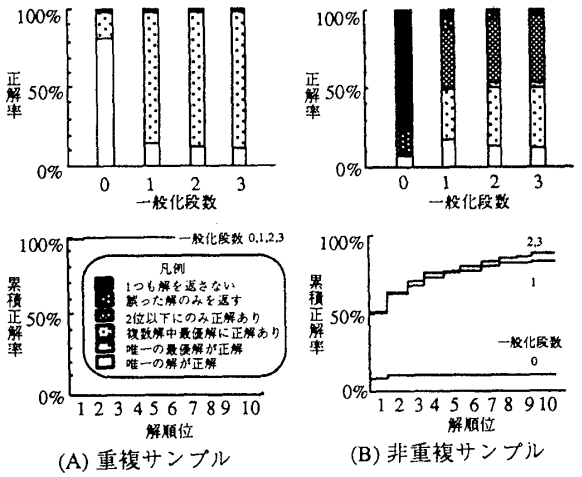
表 2: 実験に使用した用例データの一覧

実験の際には3回まで一般化を行い，(0) 上の用例データ，(1) それに一般化したデータを加えたもの，(2)1 に一般化したデータを加えたもの，(3)2 に一般化したデータを加えたもの，の4種類それぞれのデータについて学習を行って正解率の変動を観察した．

3.1 実験結果

仮名漢字変換の正解率と最尤解からの累積正解率を，(A) 入力派生語がサンプルデータに元々含まれていたのべ 28046 語，(B) それ以外ののべ 12333 語，に分類して集計したグラフを図1に示す．

この結果から，用例を一般化した場合には連接可能な語幹と接尾語の組合せをほぼ網羅できることがわかった．



(A) 重複サンプル　(B) 非重複サンプル

図 1: 仮名漢字変換実験結果

4 おわりに

派生語用例と名詞シソーラスに基づく確率派生語文法と，その仮名漢字変換への応用について述べた．用例を一般化することは，接尾の連接性をシソーラス中の下位語のみならず兄弟にあたる類義語にまで拡張して考えることに相当する．現在は，一般化によって生成派生語数が過大になることを防ぐため，用例への語義番号の設定や，一般化する用例の限定，また，接尾の連接関係の反例の利用等，生成能力を制限する方法を検討中である．

謝辞

「現代日本語名詞シソーラス」を作成された，筑波大学の荻野綱男先生,「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」の原データを作成された，九州芸術工科大学の稲永紘之先生に深く感謝致します．

参考文献

[1] 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，情報処理学会第 45 回全国大会講演論文集 (3)，pp71-72，平成 4 年 10 月

[2] 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]，NLC92-17，pp39-46，平成 4 年 10 月

[3] 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，平成 4 年 3 月