名詞シソーラスを用いた派生語の処理

1E−4

市丸 夏樹, 中村 貞吾, 日高 達

九州大学

1 概要

派生語は語幹と接辞の連接によって作られる.

派生語の読みを漢字に変換する際, 語幹と接辞の読み仮名に各々複数の漢字が対応する場合が多いため, 派生語の漢字候補はそれらの組み合わせの数だけ現われ得る.

漢字の選択作業の手間を軽減するためには, 解候補を語幹と接辞の妥当な組み合わせのみに絞り込むことが必要となる. 派生語自体を一つの単語として辞書に登録する方法をとるにしても, 膨大な組み合せの全てを網羅することは難しい. むしろ, 語幹に成り得る単語と接辞との意味的な連接可能性を捉えて派生語か否かを判定することが望ましい.

そこで, 我々は名詞のシソーラスを語幹の意味的な分類として利用して, 2文字の語幹名詞と1文字の接尾語からなる3漢字語を取り扱う派生語文法を構成し, 大量データについての仮名漢字変換実験を行なった.

2 シソーラスを用いた派生語文法

2.1 名詞の上位下位関係

シソーラスは単語間の推移的な上位下位関係を表したものである. 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. よって以下の仮定をおく.

仮定 2.1 (単語と接尾語の連接性)
シソーラスに含まれる単語$\overline{w}$と接尾語$\overline{b}$は以下の条件の何れかを満すとき, 連接して派生語を成す.

1. $\overline{w}$と$\overline{b}$とが結び付いて派生語となることが予めわかっている.
2. $\overline{w}$のある上位語$\overline{w_1}$と$\overline{b}$が連接して派生語を成す.

2.2 派生語文法

定義 2.1 (派生語文法)
仮名漢字変換用の派生語文法Gは, $G=\langle \boldsymbol{N}, \boldsymbol{\Sigma}, \boldsymbol{P}, S\rangle$と表される. ただし, $\boldsymbol{N}$は非終端記号の有限集合, $\boldsymbol{\Sigma}$は終端記号の有限集合, Sは開始記号, $\boldsymbol{P}$は生成規則の有限集合であり, $\boldsymbol{N}\cap\boldsymbol{\Sigma}=\phi$ (空集合), $S\in\boldsymbol{N}$である.

$\boldsymbol{P}$は以下のような5種類の生成規則からなる.

$$N \longrightarrow H \mid \overline{w_0} \quad (1)$$
$$H \longrightarrow \overline{w}\ \ \overline{B} \quad (2)$$
$$\overline{w} \longrightarrow \overline{w'} \quad (3)$$
$$\overline{w} \longrightarrow w \quad (4)$$
$$\overline{B} \longrightarrow B \quad (5)$$

ただし, $N, H, \overline{w_0}, \overline{w}, \overline{w'}, \overline{B} \in \boldsymbol{N}$, $w, B \in \boldsymbol{\Sigma}$, $S=N$である.

ここで$N, H, \overline{B}$はそれぞれ名詞, 派生語, 接尾語のカテゴリーであり, また$\overline{w_0}$はシソーラスの頂点にある単語で, $\overline{w}$はシソーラス上のノードや葉となっている単語を表す. ただし, $\overline{w_0}, \overline{w}$にはどちらも語義番号が付与されている. そして, w, Bは各々$\overline{w}, \overline{B}$の読みである.

シソーラスには単語間の上位下位関係が記載されている. この関係は, シソーラス中の上位語が下位語を生成するという規則として(3)のように表現されている.

規則(1),(2)はそれぞれ, 名詞が派生語かシソーラス中の任意の単語を生成すること, 派生語が語幹単語と接尾語を生成することを表す. そして規則(4),(5)は, 単語, 接尾語が, 仮名漢字変換での終端記号としての読み仮名を生成することを表している.

2.3 確率文法の利用

確率文法とは, 文法G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. 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, 共通の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である.

Processing derivative nouns with a thesaurus
Natsuki ICHIMARU, Teigo NAKAMURA, Toru HITAKA
Kyushu University

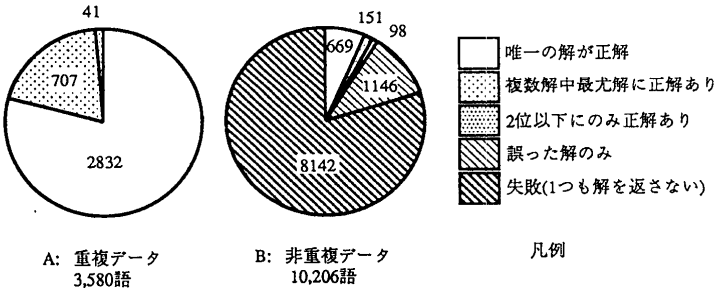

図 1: 仮名漢字変換実験結果

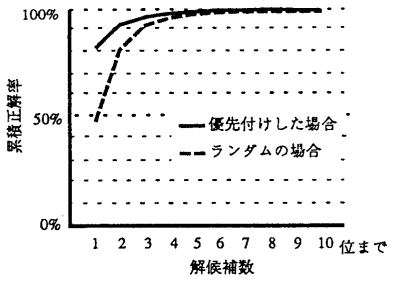

図 2: 優先付けの効果

派生語文法の生成規則 (2) は，生起頻度情報付きのサンプルデータ $S = \{(t_i, m_i)|i = 1, 2, \cdots, M\}$ から構成される．ただし t_i は G における派生語の導出木で，m_i はその頻度情報である．また M は派生語の異なり語数である．

G を用いて t_i を生成する際に生成規則 $\alpha \to \beta$ を使う回数を $C(\alpha \to \beta, t_i)$ とすると，適用確率 $p(\alpha \to \beta)$ は，

$$p(\alpha \to \beta) = \frac{\sum_{i=1}^{M} m_i \cdot C(\alpha \to \beta, t_i)}{\sum_{\gamma} \sum_{i=1}^{M} m_i \cdot C(\alpha \to \gamma, t_i)} \qquad (6)$$

と求められる．

実験においては，まずサンプルデータに対する構造木を一意的に与え，その中で各生成規則が使用されている回数を集計し，生成規則の適用確率を上式により設定した．

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，「現代用語の基礎知識 1989 年版 CD-ROM」より抽出された (○○) + ○型の 3 漢字語データ 9939 語 (出現頻度付き) を使用して派生語文法を構成した．次に，公用データベース日本語単語辞書の原データの 3 漢字語 (21245 語) のうち語幹がシソーラス，接尾がサンプルデータの接尾語にそれぞれ含まれる 13786 語の，読み仮名と正解となる漢字の組みを実験入力データとして用いた仮名漢字変換実験を行なった．

3.1 実験結果

実験結果を，(A) 入力派生語がサンプルデータに元々含まれていたもの，(B) それ以外の派生語，に分類して集計したグラフを図 1に示す．

次に確率による優先付けの効果を見るために，誤りを含む複数解が得られる入力派生語 997 語に対して，確率値による優先付けを行なった場合と，ランダムに解を取り出す場合について，解候補の中から n 位までを取り出した時に正解が含まれている割合を比較した．これを図 2に示した．

3.2 問題点

実験結果 (図 1B) において，一つも解を返さない場合が約 80%にも達している．これは，本来接続すべき語幹と接尾語の組合せを，規則 (2) として網羅できていないためであると思われる．この問題に対処して，今後サンプルデータを増加させることが望まれる．

4 結論

シソーラス (類語辞典) は既に英文ワープロ等に実装され，作文や推敲の支援ツールとして実用に供されている．もし日本語ワープロでもシソーラスが利用できるようになれば，そのような機能に加えて，仮名漢字変換自体もより使いやすいものになることが期待される．

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，3 漢字派生語データを提供して下さった日本ユニシスの方々に深く感謝致します．

参考文献

[1] 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