2004－NL－162　(27)
2004／7／16

シソーラスを用いた派生語の仮名漢字変換の特性

市丸 夏樹*,　中村 貞吾†,　日高 達‡

日本語の文章を仮名書きするとセグメンテーションの曖昧性や同音異義語が多発し読みにくい．そのため仮名漢字変換は通常の漢字仮名混じり文の形態素解析よりも難しい問題である．特に派生語には多くの同音異義語が存在するため，変換誤りの大きな原因の１つとなっている．しかし従来の１層の意味分類を用いる手法では十分な正解率を得ることはできなかった．本稿では PCFG においてシソーラス中の様々な階層まで汎化した大量の規則と品詞や単語のレベルの規則とを組み合わせて重み付けすることを試みた．コーパスを用いた実験の結果，学習サンプル数に応じて最適な分類数を選択することによって 95%の正解率が得られることがわかった．

Characteristics of Japanese Derivative Word Kana-Kanji Conversion with Thesaurus

Natsuki Ichimaru*,　Teigo Nakamura†,　Toru Hitaka‡

Japanese sentences written in kana characters are uneasy to read even for Japanese, because of their lexical ambiguity in segmentation and homonymy. Thus, their kana-kanji conversion is more complicated than usual kanji-kana mixture's morphologic analysis. Derivative words have a lot of homonyms, which often causes annoying problem in conversion. And the former method that uses single-layer semantic-markers was not so effective. In this paper, we combine three types of weighted rules to form a PCFG grammar of derivative words: generalized rules on several layers in a thesaurus, a part-of-speech level rule, and word level rules. In an experiment, our method achieves an accuracy of 95%.

1　はじめに

接尾語は 1 文字から数文字程度の漢字表記と短い読みをもつ．接尾語には同音異義語が多数存在し，助詞等の機能語とまぎらわしい場合がある．このことが，仮名漢字変換の文節区切りの誤りや変換誤りの大きな原因となっている．

派生語を成す語基と接尾語間の連接には，曖昧性解消のために有効な構文的な手がかりが存在しない．自然言語処理では，形態素解析の次に構文解析あるいは係り受け解析を行うのが通例となっているが，派生語は特殊なケースであって，形態素解析処理と同時に意味的な処理による曖昧性の絞り込みが必要とされる．本稿では,「名詞＋接尾語」型の派生語の仮名漢字変換に用例とシソーラスによる手法を表す PCFG を用いることによって，高い精度の派生語処理の実現を目指す．

関連研究としては，一般的な文を対象とし，WordNet を用いて汎化した用例を用いて smoothing を行う手法 [1] が提案されている．ただしその実験では，我々の初期の実験 [9] と同じく，用例の汎化は下から 1,2,3 段までに制限されていた．これは当時の PC のメモリ量などによるものであると考えられる．しかし現在では計算機の性能が大幅に向上したことにより，もはや汎化を制限する必要性は無くなっていると考えてよい．

本研究では，派生語の仮名漢字変換について，より詳細に様々な学習条件下での特性を調査した．その結果，学習サンプル数に応じて適切な学習条件を選択することによって 95%以上という高い正解率が得られることがわかった．

*九州大学システム情報科学研究院
Graduate School of Information Science and Electrical Engineering, Kyushu University
†九州工業大学情報工学部
Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
‡九州大学名誉教授
Emeritus Professor, Kyushu University

表 1: 各派生語処理手法における語基のレベル

	nb	ex	sem	A	B
シソーラス					
語基のレベル	品詞	単語	意味分類	3 層の組合せ	全段累積

2 シソーラスを用いた派生語処理

2.1 従来手法

一般名詞と接尾語の連接によって作られる派生語を処理する手法には例えば次のようなものがある．

任意連接の受理 (nb) 任意の名詞と任意の接尾語の連接を受理する．それぞれ独立に頻度による優先順位付けを行う．

単語登録 (ex) 派生語を単語として扱って辞書登録する．同音異義語は生起頻度順に優先付ける．

意味分類 (sem) 全ての単語を意味的に分類し，接尾語と連接可能な語基をその分類コードによって指定し，語基-接尾語間の選択制約により曖昧性を絞り込む．

以下本稿では簡単のため各手法を記号で表す．

接尾語は固有の意味を持つ名詞に連接する傾向がある．上記の手法はどれも，接尾語との連接が許される派生語基のカテゴリをシソーラス中のある 1 つの階層から選んで用いる手法であり，その抽象度のレベルが異ったものと捉えられる (表 1)．

2.2 従来手法の問題点

これらの手法にはそれぞれ次のような問題点が存在する．nb では非派生語を派生語として受理する誤認識が多く，無意味な変換候補を大量に生成する．ex では未登録語を受理できないが，膨大な派生語全てを収集することは困難である．また，もし仮に全てを収集できたとしても，その辞書は巨大なものになると予想される．sem では分類が粗いとあまり良い正解率が得られない．また，分類を細分化しようとすると，用例が疎らになることを補うために大量の学習サンプルを増加させなければならないが，必要とされる量は指数関数的に増大する．そこで，これらの従来手法を表す文法を混合し，シソーラス上の様々なレベルの中間ノードで代表される語基と接尾語との連接規則を組み合わせて使用することを考える．

2.3 提案手法

PCFG では，複数のレベルの連接規則を組み合わせて使用できる．様々な組み合わせを試みた結果，本稿では特に次の 2 つの手法を提案する．

3 層の組み合わせ (提案手法 A) nb, ex, sem の 3 つのレベルの連接規則を組み合わせて使用し，$ex > sem > nb$ の順に優先付けする．

全段累積 (提案手法 B) 段階的に汎化した用例を全段にわたって累積的に使用し，シソーラス中の全ての階層の中間ノードを使用する．

nb の文法を包含していることから，A,B の文法から生成される言語は nb の言語と同一である．しかし手法 A では，ex の被覆率の低さがまず sem で補われ，次に全ての未登録語が nb によって救済される．手法 B も同様であるが，より多くの階層が用いられる．これらにより解候補の生起確率には，用例の頻度と語基の分布密度を反映した優先付けが施される．

3 確率派生語手法

3.1 派生語処理手法の PCFG による表現

派生語の品詞を表す非終端記号を **H**，名詞と接尾語についてそれぞれ，品詞を表す非終端記号を **N**，**B**，単語を表す非終端記号を W，B，仮名表記文字列を表す終端記号列を w, b とし，シソーラスの頂点ノードを N_0，語義番号 i をもつシソーラス中の中間ノードを頂点とする名詞のサブセットを表す非終端記号を N_i とするとき，前述の手法を統合した確率派生語文法の生成規則は次のように表される．

$$\mathbf{H} \xrightarrow{p_1} W\ B \quad (1) \qquad N_i \xrightarrow{p_6} N_j \quad (6)$$

$$W \xrightarrow{p_2} w \quad (2) \qquad N_i \xrightarrow{p_7} W \quad (7)$$

$$B \xrightarrow{p_3} b \quad (3) \qquad \mathbf{H} \xrightarrow{p_8} \mathbf{N}\ \mathbf{B} \quad (8)$$

$$\mathbf{H} \xrightarrow{p_4} N_0 \quad (4) \qquad \mathbf{N} \xrightarrow{p_9} W \quad (9)$$

$$\mathbf{H} \xrightarrow{p_5} N_i\ B \quad (5) \qquad \mathbf{B} \xrightarrow{p_{10}} B \quad (10)$$

規則 (1), (5), (8) はそれぞれ単語，意味分類，品詞の各レベルの派生語基と，接尾語との連接規則を表す．

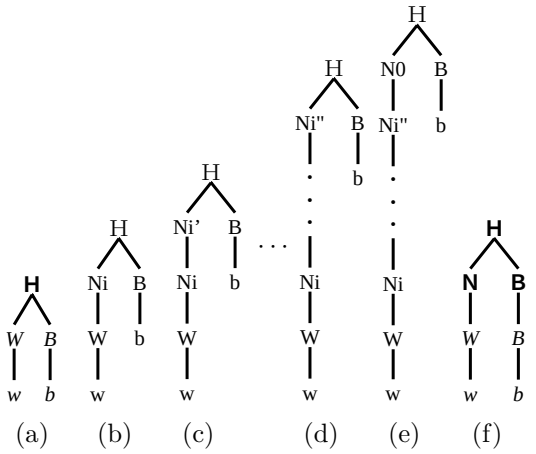


図 1: 学習用サンプル構文木

表 2: 各手法で使用する生成規則と構文木

手法	生成規則	構文木
ex	(1)〜(3)	(a)
nb	(8)〜(10)	(f)
sem	(4)〜(7)	(b)
提案手法 A	(1)〜(10)	(a), (b), (f)
提案手法 B	〃	(a)〜(f)

3.2 確率派生語文法の学習

これらの連接規則をテキストコーパスから収集した派生語用例から自動的に獲得し，最尤推定法 [6, 5] により適用確率値 $p_1, \cdots, p_{10}$ を求める．各手法の生成規則と学習に用いる主な構文木の形状を図 1，対応表を表 2 に示す．

sem の意味分類としては，シソーラスを動的に縮小して生成した意味分類を与える．用例から作成する連接規則には縮小された意味分類の中で最も深い位置にあるノードを採用する．

提案手法 B においては，規則 (5) 型の連接規則を用例から獲得する際，上位にあるノードほど複数の用例から重複して学習されやすく，各中間ノードの学習のスピードはノードの子孫数に比例する傾向がある．そこで連接規則に対して 1/子孫数 の比を用いた重み付けを行い，汎化段数の少ない連接規則が優先されるようにする．

さらにゼロ頻度問題に対処するため，図 1 に示したものの他に規則 (4) から始まる全単語の導出パスを予め微少頻度で学習させておくことにより，学習サンプルに現れない導出パスを補完する．

以上により，高頻度でかつ派生語として尤もらしいものを優先解とするような PCFG を構築することができる．

表 3: 学習データ

データ名	用途
EDR CP	縮小して意味分類を作成.
EDR JWD	1 単語を N+B に分割.
EDR JCO	N+B を抽出.
毎日新聞'91〜95,98	N+B を抽出，分割. ('94 の 3000 記事を除く)

表 4: 獲得された連接規則数：(m=ノード数)

サンプル番号	sem m=2046	sem m=33193	B(nb,ex を除く) m=167967
S-1	10,421	11,513	50,727
S-2	16,493	18,635	77,167
S-3	25,830	29,850	115,150
S-4	38,649	45,995	165,863
S-5	55,742	68,527	232,302
S-6	78,433	99,958	319,504
S-7	108,244	144,182	435,244
S-8	145,179	203,226	581,043
S-9	188,835	278,659	761,704
S-10	235,955	368,670	967,957

4 派生語処理の特性

各手法を用いて派生語の仮名漢字変換実験を行い，その特性を求めた．

4.1 学習データ

PCFG の学習データとしては，表 3 に示す単語辞書，コーパスから抽出した次の 2 通りの派生語サンプルを使用した．

(1) 「名詞 接尾語」の品詞パターンに合致する部分形態素列.

(2) 新聞コーパスあるいは単語辞書中に出現する単語のうち「名詞 接尾語」に分割可能なもの.

いずれもコーパス等と EDR JWD の両方に記載された語からなるもので，コーパス等における品詞が「名詞」と「接尾語」であり，かつ，EDR JWD での品詞が「一般名詞 (JN1)」と「接尾語 (JB1, JUN, JN6)」であるものに限られる．

こうして獲得された派生語サンプル数を表 5 に示す．ここでは全量を $S-10$ とし，$S-i$ の 1/2 をランダムに取り出したものを $S-(i-1)$ とおく．獲得された連接規則の数を表 4 に示す.

新聞コーパスから抽出された派生語用例の大半は漢字表記のみのデータであるため，その多義性が問題となる．そこで本稿の実験では，人手で設定された正解語義を持つ EDR JCO 中の用例をシソーラス

表 5: 派生語学習サンプル数と試験データの内訳

サンプル番号	学習データ数		試験データ数					
			正例					
			登録済		未登録		負例	
	異なり語数	頻度総和 n	異なり語数	頻度総和	異なり語数	頻度総和	異なり語数	頻度総和
S-1	3,738	7,074	842	2,569	1,832	2,408	965	1,163
S-2	6,223	14,148	1,124	3,006	1,550	1,971	965	1,163
S-3	10,011	28,296	1,425	3,434	1,249	1,543	965	1,163
S-4	15,554	56,592	1,661	3,727	1,013	1,250	965	1,163
S-5	23,624	113,185	1,863	3,954	811	1,023	965	1,163
S-6	35,510	226,370	2,035	4,159	639	818	965	1,163
S-7	52,520	452,741	2,173	4,345	501	632	965	1,163
S-8	76,772	905,482	2,265	4,468	409	509	965	1,163
S-9	110,554	1,810,965	2,340	4,555	334	422	965	1,163
S-10	155,399	3,621,930	2,412	4,653	262	324	965	1,163

表 6: 試験データ

データ	用途等
RWC-DB-TEXT-95-2 毎日新聞'94 の 3000 記事.	
うち 2/3	品詞の bigram モデルの学習に使用.
残り 1/3	N+B の正例を抽出.
〃	形態素解析結果より負例を抽出.

上にマッピングし，正解が近傍に存在する割合を用いることにより，多義の用例の語義候補に対し頻度の重み付けを行っている．

m 分類の意味分類には，EDR CP(シソーラス) の末端のノードを省いて縮小したものを用いた．その際汎化段数を用いると粒度が揃わなくなるため，ここでは子孫数の降順で上位 m 個のノードを用いるものとし，ほぼ $m = 2^i$ となるようにノード数を選んだ．

4.2 試験データ

本実験で用いた試験データを表 6 に示す．RWC-DB-TEXT-95-2 (ver.1) は毎日新聞'94 の 3000 記事に対する読み付きの人手修正済み形態素データである．

実験では，派生語正例および負例の仮名表記を試験入力とし，漢字表記を出力とする．正解は正例の漢字表記である．負例としては，品詞レベルの bi-gram モデルによる仮名表記文の形態素解析の最尤解に含まれる派生語候補のうち，正解に含まれないものを用いた．ただし,「接頭語 名詞」,「活用語 活用語尾」のパターンに合致するものや，コーパスと単語辞書の品詞名の不一致によって派生語として受理された単語や複合語は，必ずしも負例とは言えないため除外した．

4.3 適合率と再現率

試験データ中の正例数を t，正解を含んだ出力が得られる数を c，不正解のみが出力される数を e とするとき，適合率 $\overset{Precision}{p}$, 再現率 $\overset{Recall}{r}$ (または正解率) を次のように定義する．

$$p = \frac{c}{c+e}, \qquad r = \frac{c}{t} \tag{11}$$

以下，本稿では最尤解の値を示す．

4.4 各手法の特性と考察

各手法について，様々な分類数 m の意味分類と 10 通りのサンプル S-i を与えた派生語の仮名漢字変換実験を行った．

まず，手法 nb, ex, sem(m) の適合率 p – 再現率 r の特性曲線を図 2 に示す．各手法の適合率 p_{nb}, p_{ex}, $p_{sem(i)}$ には次のような傾向が見られる．

$$p_{nb} < p_{sem(i)} < p_{ex} \tag{12}$$

$$i < j \Rightarrow p_{sem(i)} < p_{sem(j)} \tag{13}$$

適合率は学習サンプル数によって変動するが，この相対的な順序関係は保存されている．

nb の場合の適合率 p，再現率 r の特性は，試験データ中の正例数を t，負例数を f とおくと，次式で表される直線上を動く．

$$r = (1 + \frac{f}{t}) \cdot p \tag{14}$$

sem の特性はノード数に応じて nb と ex の間を連続的に変化する．学習サンプルを固定してノード数を変化させた場合，ピークを持ったへの字型の特性変化が観察される．そのため，学習サンプルが少ないうちはより汎化された規則を用いた方が再現率 r が高くなるが，ある程度学習が進むと順番が入れ替

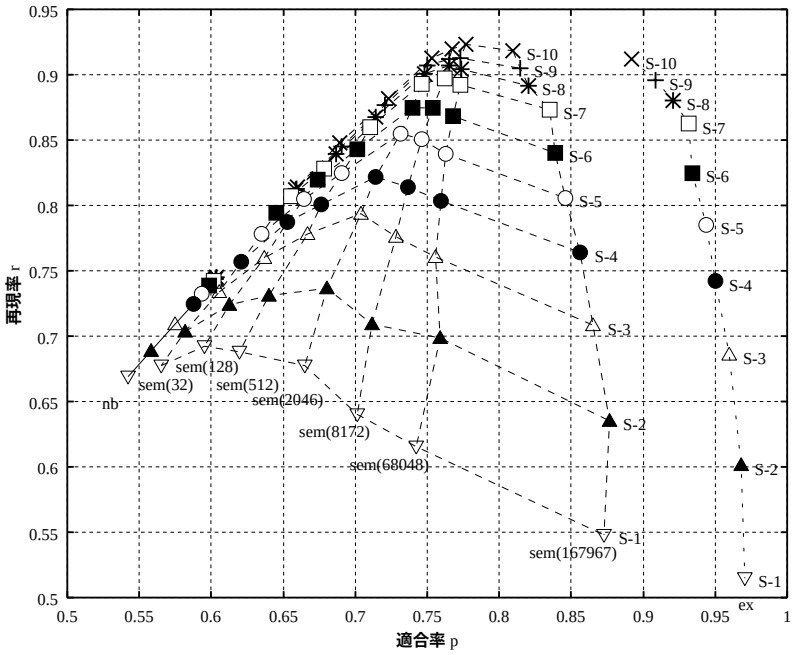


図 2: 1 層の連接規則を用いる手法の特性

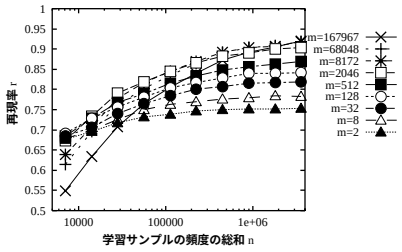


図 3: sem の正解率 r の変動

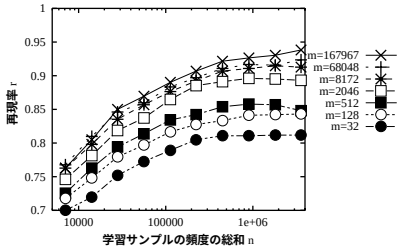


図 4: 手法 B の正解率 r の変動 (nb,ex を除いた場合)

表 7: 各手法の正解率 r の最大値

サンプル番号	sem 正解率 r	sem ノード数 m	A 正解率 r	A ノード数 m	B 正解率 r	B ノード数 m
S-1	0.6923	(128)	0.7825	(128)	0.7900	(128)
S-2	0.7365	(2,046)	0.8259	(4,067)	0.8295	(128)
S-3	0.7935	(2,046)	0.8653	(16)	0.8704	(4,067)
S-4	0.8256	(4,067)	0.8927	(4,067)	0.8905	(4,067)
S-5	0.8571	(4,067)	0.9123	(4,067)	0.9128	(1,024)
S-6	0.8790	(4,067)	0.9287	(4,067)	0.9269	(2,046)
S-7	0.8984	(4,067)	0.9365	(4,067)	0.9378	(2,046)
S-8	0.9108	(4,067)	0.9427	(2,046)	0.9399	(4,067)
S-9	0.9151	(33,193)	0.9467	(2,046)	0.9464	(8,172)
S-10	0.9242	(33,193)	0.9513	(4,067)	0.9514	(167,967)

わり，より具体的な規則を用いた方がかえって高い値が得られるようになる (図 3)．なお，手法 B から nb,ex を除いたものでは最初からノード数 m が多い方が再現率 r が高く，入れ替わりは起らない (図 4)．

手法 sem, A，B の最尤解の正解率 r のピークの値とその時のノード数 m を表 7 に示す．

sem の正解率は，意味分類の適度な細分化によって $r = 92.42\%$ まで向上している．しかし，最適なノード数 m が変動するため，用例の追加の度に最適値を求め直し，意味分類および PCFG の連接規則を再構築する必要がある．また，未登録語の再現率 r は単純な nb よりも劣っている (表 8)．

A,B の正解率はサンプル数の全域において従来手法の sem を上回り，手法 B(m=167,967) でサンプル S-10 を使用した場合に最も高い正解率 $r = 95.14\%$ が得られている．A,B ではノード数 m による正解率 r の差が小さく，1% 未満程度であるので，もし m を固定して使用しても比較的安定した性能が得られるものと思われる．

提案手法 A,B を比較すると，総合的な正解率はほぼ同等である．ただし A の方がメモリ使用量は少なく，未登録語に対する正解率は B より 4 ポイントほど高い．(表 8)．

同音異義の接尾語に対する正解率の例を表 9 に示す．接尾語は生起頻度の降順に並んでいる．nb と ex では，同音異義語の中で頻度が 1 位のものは良いが，2 位以下の正解率は大きく悪化している．それに対し提案手法 A, B では低頻度の接尾語の正解率があまり低下しない．これは提案手法が，連接する語基の分布の微妙な差異により適切な接尾語を選択できることを示している．従って，これらの提案手法を仮名漢字変換に利用すれば，多くの場合 1 回の変換で適切な語基と接尾語の組を選択できるようになるものと考えられる．

5 おわりに

EDR の単語辞書，コーパス，及び新聞記事 6 年分のコーパスから派生語用例を収集し，縮小したシソーラスを用いて 3 層または全階層のレベルまで汎化したその構文木を，より具体的な連接規則を優先するよう重み付けしながら，累積的に PCFG に学習させることによって派生語文法を構築した．これにより，派生語単体での仮名漢字変換の正解率の目標

表 8: 正例/負例に対する実験結果の内訳：(サンプル S-10 使用時)

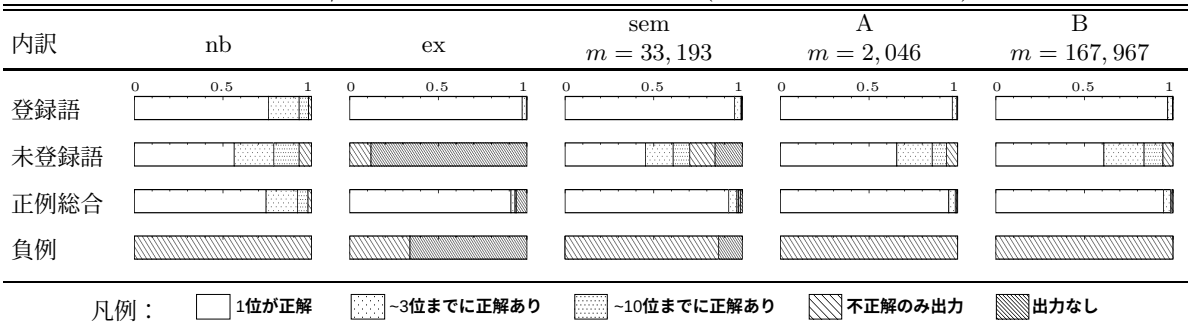
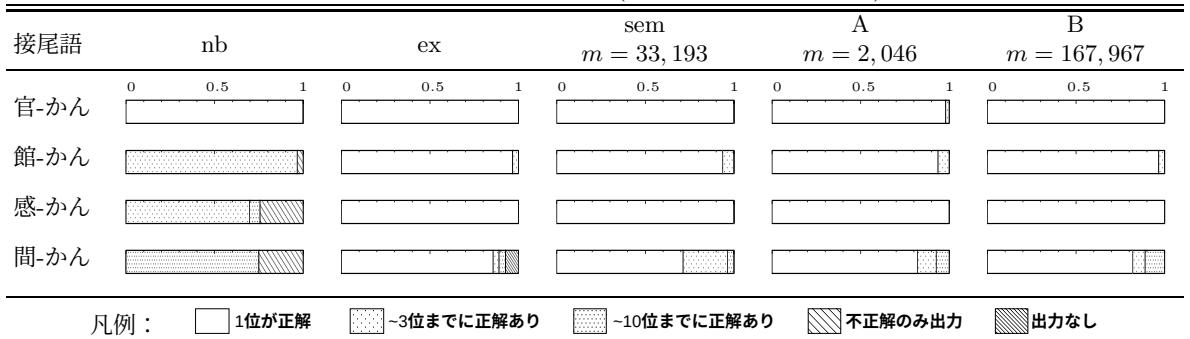


表 9: 接尾語別の正解率 r：(サンプル S-10 使用時)



値であった 95%を越える正解率を得ることができた．今後学習サンプルを増加させれば，正解率は nb の直線的な特性 (式 (14)) に沿ってさらに上昇していくものと考えられる．

参考文献

[1] Daniel M. Bikel. A statistical model for parsing and word-sense disambiguation. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000.

[2] Eugene Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*. AAAI Press/MIT Press, 1997.

[3] 松延栄治, 日高達, 吉田将. 確率文節文法による構文解析. 情報処理学会 自然言語処理研究会 研究報告 86-NL-56-3, Vol. 86, No. 52, pp. 17–24, 1986.

[4] 池原悟, 村上仁一, 木本泰博. 単語意味属性を使用したベクトル空間法. 自然言語処理, Vol. 10, No. 2, pp. 111–128, 2003.

[5] 日高達. 確率文脈自由文法におけるパラメタの最尤推定法. 電子情報通信学会 信学技報 NLC-93-57, Vol. 93, No. 366, pp. 23–28, 1993.

[6] 日高達. 確率文法. 情報処理, Vol. 36, No. 2, pp. 169–176, 1995.

[7] 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. 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. 電子情報通信学会, 1992.

[8] Natsuki Ichimaru, Teigo Nakamura, Yoshiaki Miyamoto, and Toru Hitaka. Example-based stochastic analysis of japanese derivative words. *Natural Language Processing Pacific Rim Symposium '93*, pp. 368–371, 1993.

[9] 市丸夏樹, 中村貞吾, 宮本義昭, 日高達. シソーラスと確率文法による派生語解析. 情報処理学会論文誌, Vol. 36, No. 4, pp. 849–858, 1995.

[10] 市丸夏樹, 中村貞吾, 日高達. PCFG による派生語処理手法の比較と検討. 九州大学システム情報科学研究科 研究科報告, Vol. 4, No. 1, 1999.

[11] 日本電子化辞書研究所. EDR 電子化辞書 version 2. CDROM, 1999.

[12] 田中英輝. シソーラスを利用した言語データ最適一般化アルゴリズム. 信学技報 [言語理解とコミュニケーション] NLC 95-19〜29, Vol. 95, No. 169, pp. 9–14, 1995.

[13] 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, 1992.

[14] 毎日新聞社. CD-毎日新聞 '91–'95, '98. CDROM, 1991–1995, 1998.

[15] 松本祐治. 形態素解析システム「茶筌」. 情報処理学会誌, Vol. 41, No. 11, pp. 1208–1214, 2000.