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

2006－SLP－61 (5)
2006／5／11

単語の共起関係と構文情報を利用した単語階層関係の統計的自動識別

大石 康智† 伊藤 克亘†† 武田 一哉† 藤井 敦†††

† 名古屋大学大学院情報科学研究科
〒464-8603 愛知県名古屋市千種区不老町
†† 法政大学情報科学部
〒184-0002 東京都小金井市梶野町 3-7-2
††† 筑波大学 図書館情報メディア研究科
〒305-8550 茨城県つくば市春日 1-2
E-mail: †ohishi@sp.m.is.nagoya-u.ac.jp, takeda@is.nagoya-u.ac.jp,
††itou@k.hosei.ac.jp, †††fujii@slis.tskuba.ac.jp

あらまし 本研究では，Webから抽出し，構築した大規模かつ情報量の多いテキストコーパスを用いて，統計的に単語間の意味的な階層関係を自動判定する手法を提案する．本手法は，説明文の方向性を考慮した出現頻度モデルと構文情報に基づく統計モデルからなる．説明文の方向性を考慮するということは，下位語の説明文は，その上位語を含む傾向があるが，一方，上位語の説明文は，下位語のすべてを含んでいるというわけではないという統計的な性質である．また，構文情報に基づく統計モデルとは，上位語，下位語と連接して出現する形態素の統計量の違いに基づいたモデルである．実験結果より，本手法が既存のシソーラスに記述された単語間の階層関係のうち74.1%を検出することが可能であった．

キーワード 事典コーパス，シソーラス，単語階層関係，構文情報，線形判別分析

Statistical Discrimination Method for the Hierarchical Relation of Word Pairs Using Co-occurrence and Syntactic Information

Yasunori OHISHI†, Katunobu ITOU††, Kazuya TAKEDA†, and Atsushi FUJII†††

† Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8603, Japan
†† Faculty of Computer and Information Sciences, Hosei University
3-7-2, Kajino-cho, Koganei, Tokyo, 184-0002, Japan
††† Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
E-mail: †ohishi@sp.m.is.nagoya-u.ac.jp, takeda@is.nagoya-u.ac.jp,
††itou@k.hosei.ac.jp, †††fujii@slis.tskuba.ac.jp

Abstract This paper proposes a discrimination method for hierarchical relations between word pairs. The method is a statistical one using an "encyclopedic corpus" extracted and organized from Web pages. In the proposed method, two types of the statistical model are used; the directional occurring model and the syntactic statistics model. The directional occurring model reflects that hyonyms' descriptions tend to include hypernyms whereas hypernyms' descriptions do not include all of the hyponyms. The syntactics statics model bases on the difference of the syntactic patterns observed around hypernyms and hyponyms. Experimental results show that the method detected 74.1% of the relations in an actual thesaurus.

Key words Encyclopedic Corpus, Thesaurus, Hierarchical relations, Syntactics information, Linear discriminant analysis

1. はじめに

"事典的なコーパス" を利用してシソーラスを構築するために単語間の意味的な関係を自動判定する手法を提案する．

シソーラスは，情報検索における検索質問の拡張や機械翻訳に利用される．しかし，既存のシソーラスは人手によって作成されるのが主流であり，単語間の意味関係を手作業で調べるために時間と手間を要する．また，多くの技術的な専門用語について，十分に整備されていない．

そこで，シソーラスの自動構築の試みとして，特定の構文パターン (例えば，"〜の一つ"，"〜は〜である"，"〜のような") をもつ文章や，辞書における見出し語とそれに対する説明文，特別な文書構造 (Web ページを項目別に分類したもの) などの表現形式から下位語，同義語，上位語を抽出するための手法が提案されている [1]〜[4]．しかし，構文パターンを網羅的に収集しなければならないこと，また見出し語に対する説明が少量であるため，大語彙を対象として意味的な関係を自動抽出することは難しい．

また，既存のシソーラスに未知語を配置することによってシソーラスを拡張する研究がある [5]．しかし，既存のシソーラス自体が人手で作られたものであるため，完全な自動化へは至っていない．

これらの問題を解決するために，本研究では，Cyclone コーパス [6] と呼ばれる事典的なコーパスを利用する．このコーパスは，75 万語という大語彙に及ぶものであり，各見出し語は，複数の説明文によって様々な観点から説明されている．

Cyclone コーパスのような事典的なコーパスを利用することは，新聞のような一般的な文書を利用するよりも，単語間の階層関係を抽出することに対してより効果的であることが期待される．なぜなら，事典的なコーパスは，その構造上，見出し語とその説明文中の単語となんらかの意味的な関係をもっているということが期待されるためである．また，随時更新されていく Web ページを整備して構築されたコーパスであるので，新しい専門用語や稀な用語を含んでいる．Web における文書のいくつかは，信頼性の低いものであるため，Cyclone コーパスでは，それらの表現については排除するように構築されている．

事典的なコーパスを利用して，本研究では，ある 2 つの単語が与えられたとき，それらが階層関係であるのか，無関係であるのかを出力する自動識別器を設計する．今回は，同義語と関連語は考慮しない．この識別器は，見出し語とその説明文に出現する単語との共起関係を考慮した出現頻度に基づくモデルと，単語に連接する形態素に着目した構文情報の統計量に基づくモデルからなる．これらを利用して識別を行ったところ，既存のシソーラスに記述された単語間の階層関係のうち 74.1%を検出することができた．

2. Cyclone コーパス

Cyclone コーパスは，Web を事典的に利用することを目的として，量質ともに優れたコーパスを構築することを目指している．そのために，Web ページに含まれる用語説明を抽出して語義 (専門分野) に基づいて分類し，さらに良質な説明を選択的に取得してコーパスを自動構築する．現在，約 75 万語の用語 (見出し語) を整備して作成されている．この構築手順は以下のようになる．

(1) Web 検索エンジンを用い，見出し語を含む Web ページを網羅的に取得する

(2) 取得したページにおける HTML のタグ構造を利用して，見出し語を含む領域 (段落) を抽出し，これを見出し語の説明文とする

3. 単語間の階層関係の推定手法

3.1 説明文の方向性を考慮した出現頻度モデル

百科事典における見出し語の説明文は，"方向性" をもつ．例えば，"ライオン" の説明文には，"ネコ科の哺乳類" というように，"ネコ" や "哺乳類" という上位語を含んでいる．しかし，"哺乳類" の説明文には，"犬や猫のような動物" というように，必ずしも "ライオン" という下位語を利用して説明するとはかぎらない．一般に，見出し語に関する説明文を複数集めてきた場合，その上位語は，どの説明文にも共通して含まれていると考えられるが，必ずしも，その下位語が，どの説明文にも共通して含まれているとはかぎらない．なぜなら，説明文における下位語の使用は，見出し語を説明する観点に依存するためである．

一方，"ライオン" の説明文に，"爬虫類" という語が使用される頻度は，限りなく 0 に近い．また，"爬虫類" の説明文に，"ライオン" という語が使用される頻度も，限りなく 0 に近いと考えられる．すなわち，見出し語に対して無関係な語が，その説明文に使用される頻度は明らかに 0 に等しい．

そこで，見出し語とその説明文の方向性を考慮した確率的な手法を提案する．まず，$H(X|Y)$ を以下のように定義する．

$$H(X|Y) = C(X|Y) - C(Y|X) \qquad (1)$$

ここで $C(X|Y)$ は，見出し語 Y の説明文における単語 X の出現確率であり，$C(Y|X)$ はその逆である．このとき $|H(X|Y)|$ の値によって，X と Y には階層関係があるのか，無関係であるのかを識別するという問題設定に以上の例を置き換えることができる．この $C(X|Y)$ と $C(Y|X)$ を計算するために，Cyclone コーパスは有効である．なぜなら見出し語に対して，多くの観点に基づく様々な説明文が収集されているためである．

また，$C(X|Y)$ と $C(Y|X)$ を見出し語の説明文から直接的に計算するだけでなく，単語の間接的な関係を考慮した，説明文を再帰的に展開する手法を利用する．例えば，図 1 のように，"ペルシャ猫" の説明文には，"哺乳類" が含まれていなくとも，もし，"ネコ" の説明文に，"食肉目ネコ科の哺乳類" という記述があれば，間接的に "哺乳類" は "ペルシャ猫" の上位語であるという階層関係を見つけることができる．このように，説明文を見出し語の集合であると考えると，見出し語に対する説明文は繰り返して展開できる．この手法は拡張説明文 [3] と呼ばれ，見出し語と意味的な関係はあるが，説明文には出現しない単語の出現確率を，説明文を再帰的に展開することによって間

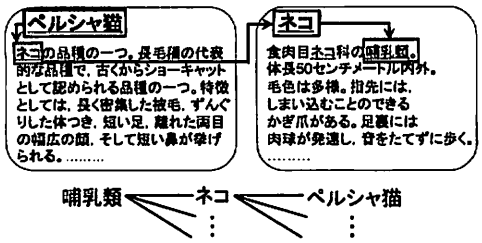

図 1　単語の間接的な関係を考慮した階層関係の推定

接的に推定することが可能となる．以下，説明文中の語を n 回展開した説明文を "n 次説明文" と呼ぶ．まず，見出し語 w_j の 1 次説明文 (見出し語を直接的に説明した文書) 中の単語 w_i の出現確率は，以下の正方行列の要素として定義される．

$$\mathbf{A}_{i,j} = P(w_i^{(1)}|w_j), \qquad i,j = 1,\ldots,K \tag{2}$$

ここで K は見出し語数であり，$P(w_i^{(1)}|w_j)$ は，最尤推定により，以下に定義される．

$$P(w_i^{(1)}|w_j) = \frac{F(w_i|w_j)}{\sum_{k=1}^{K} F(w_k|w_j)} \tag{3}$$

ここで，$F(w_i|w_j)$ は見出し語 w_j の説明文中の単語 w_i の頻度である．見出し語 w_j の 2 次説明文における単語 w_i の出現確率は，図 2 の左図に示されるように，見出し語 w_j の説明文に出現する各単語の出現確率を遷移確率と考え，それらの単語を見出し語とした説明文における単語 w_i の出現確率とともに，

$$P(w_i^{(2)}|w_j) = \sum_k P(w_i|w_k)P(w_k|w_j) \tag{4}$$

を計算することになる．1 次説明文の全体を表す正方行列 $\mathbf{A}$ を用いれば，2 次説明文の全体は $\mathbf{A}^2$ と表され，$P(w_i^{(2)}|w_j)$ は，i 行 j 列目に位置する．同様に，n 次説明文についても，図 2 の右図のように考えることにより，出現確率の全体は $\mathbf{A}^n$ として表される．最終的に，各次数の説明文の出現頻度行列の線形結合を拡張説明文として定義する．

$$\mathbf{C} = \sum_{n=1}^{N} [\alpha_n \mathbf{A}^n] \tag{5}$$

ここで N は説明文を展開する次数であり，α_n は，展開された n 次説明文の重みと考える．この拡張説明文 $\mathbf{C}$ の要素 $\mathbf{C}_{i,j} = P(w_i|w_j)$ は，見出し語 w_j との間接的な関係も考慮に入れた説明文における単語 w_i の出現確率である．このように単語間の間接的な関係を考慮することは，シソーラスを構築する上で重要なことである．なぜなら，直接的な関係だけを利用して，意味的に大規模な階層を構築することは不十分であるためである．

以上の拡張説明文 $\mathbf{C}$ を利用して式 (1) の $H(X|Y)$(ここでは $H(w_i|w_j)$) を計算する．学習では，$|H(w_i|w_j)|$ の値によって，単語 w_i と w_j が階層関係か無関係か識別できるように，α_n を線形判別分析 (LDA) により推定する．評価では，図 3 に示す

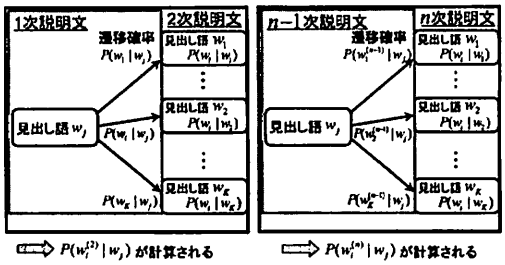

図 2　見出し語の説明文の展開方法; 2 次説明文と n 次説明文における単語の出現確率の計算方法を示す

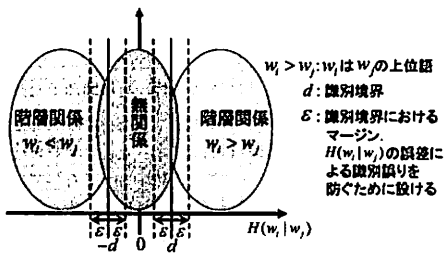

図 3　単語間の階層関係・無関係の識別結果の模式図

3 つのクラスの識別結果が，以下の式 (6) の $H(w_i|w_j)$ の値によって出力される．

$$H(w_i|w_j) = \mathbf{C}_{i,j} - \mathbf{C}_{j,i}$$

$$\begin{cases} H(w_i|w_j) > d + \epsilon & w_i \text{ は } w_j \text{ の上位語} \\ H(w_i|w_j) < -(d+\epsilon) & w_i \text{ は } w_j \text{ の下位語} \\ -d + \epsilon \leqq H(w_i|w_j) \leqq d - \epsilon & w_i \text{ と } w_j \text{ は無関係} \end{cases} \tag{6}$$

今回は式 (6) における ϵ を 0 に設定し，識別境界におけるマージンを考慮しないものとした．今後は，このマージンを考慮することにより，識別性能の改善が期待される．

3.2　局所的な構文情報に基づく統計モデル

前節のモデルは，単純に説明文の方向性を考慮したときの出現頻度の違いに着目したに過ぎないため，単語間に階層関係があることを判定できても，どちらが上位語でどちらが下位語であるかを正確に判定するためには，更なる情報を付加する必要がある．

そこで，"ライオン" の上位語となる "ネコ" に後接する "科"(ネコ科) や "哺乳類" の下位語となる "ネコ" に後接する "など"(ネコなど) というように，単語の周辺に頻出する構文情報を利用して，見出し語の説明文に出現する単語が，見出し語に対して，上位語であるのか下位語であるのかを推定する手法を提案する．本手法は，局所的な構文情報に基づいた統計量を利用するため先行研究のように階層関係を決定付ける特定の構文パターンを網羅的に見つける必要はない．

局所的な構文情報に基づいた統計量を計算するために，見出

し語の上位語，もしくは下位語の周辺に頻出する形態素からなる構文ベクトル $\mathbf{x}_{w_i|w_j}$ を導入する．この構文ベクトル $\mathbf{x}_{w_i|w_j}$ は，見出し語 w_j の説明文に出現する単語 w_i に後接する頻度の高い形態素の相対頻度ベクトルである．まず，見出し語 w_j の説明文中に出現する単語 w_i に後接する形態素をすべて抽出する．その頻度の高かった上位 W 種類の形態素を考えた場合，$\mathbf{x}_{w_i|w_j}$ は以下のように定義される．

$$\mathbf{x}_{w_i|w_j} = (x(s_0), x(s_1), \ldots, x(s_W)) \tag{7}$$

ここで s_k は k 番目に頻出する形態素であり，s_0 は上位の W 種類以外の残りの形態素を表す．確率 $x(s_k)$ は以下のように定義する．

$$x(s_k) = \frac{F(s_k)}{\sum_{l=0}^{W} F(s_l)} \tag{8}$$

ここで，$F(s_k)$ は形態素 s_k の頻度である．次に前節で述べた拡張説明文の考え方を導入する．これは，単純に見出し語 w_j の説明文から w_i に後接する形態素に基づく $\mathbf{x}_{w_i|w_j}$ を計算しただけでは，ベクトルが疎になり，w_i が見出し語 w_j に対して，上位語か下位語かを正しく識別できないためである．そこで，見出し語 w_j の n 次説明文における w_i の出現確率は，$P(w_i^{(n)}|w_j)$ なので，以下のように n 次説明文における構文ベクトル $\mathbf{x}^{(n)}_{w_i|w_j}$ を定義する．

$$\begin{aligned}\mathbf{x}^{(n)}_{w_i|w_j} &= P(w_i^{(n)}|w_j) \cdot \mathbf{x}_{w_i|w_j} \\ &= (A^n)_{i,j} \cdot \mathbf{x}_{w_i|w_j}\end{aligned} \tag{9}$$

さらに，展開次数ごとにの加重和により，拡張説明文の考え方を導入した構文ベクトル $\mathbf{X}_{w_i|w_j}$ を以下に定義する．

$$\mathbf{X}_{w_i|w_j} = \sum_{n=1}^{N} q_n \mathbf{x}^{(n)}_{w_i|w_j} \tag{10}$$

q_n は，今回はすべて 1 に固定する．

ロジスティック回帰分析を利用して，学習データから計算される構文ベクトル $\mathbf{X}_{w_i|w_j}$ から単語間の階層関係を判定する．そのために，重みベクトル $\boldsymbol{\beta}$(構文ベクトル $\mathbf{X}_{w_i|w_j}$ の各次元に対応する形態素の出現確率の重み) を推定し [7]，単語 w_i と w_j 間の階層関係を判定するために，以下の 4 つの確率を計算する．

$$\begin{aligned}
\Pr(w_i > w_j|\mathbf{X}_{w_i|w_j}) &= \frac{\exp(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_{w_i|w_j})}{1 + \exp(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_{w_i|w_j})} \\
\Pr(w_i < w_j|\mathbf{X}_{w_i|w_j}) &= \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_{w_i|w_j})} \\
\Pr(w_i > w_j|\mathbf{X}_{w_j|w_i}) &= \frac{\exp(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_{w_j|w_i})}{1 + \exp(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_{w_j|w_i})} \\
\Pr(w_i < w_j|\mathbf{X}_{w_j|w_i}) &= \frac{1}{1 + \exp(\beta_0 + \boldsymbol{\beta}\boldsymbol{X}_{w_j|w_i})}
\end{aligned} \tag{11}$$

ここで $w_i > w_j$ は，単語 w_i が w_j の上位語であることを意味する．以上の式 (11) の中で最も確率の大きいものを階層関係の判定結果とする．

表 1 Cyclone コーパスにおけるコンピュータ関連の専門用語 2,074 語の説明文の精度

判定	見出し語	平均説明文数	正解データ	無関係データ
A	1624	6.62	136 組	301 組
A+B	1803	10.4	168 組	366 組
ALL	2074	80.7	206 組	497 組

4. 評価実験

4.1 実験条件

ある 2 つの単語を入力したとき，その意味関係を判定するための 3 節で提案した自動識別器の性能を評価するために，Cyclone コーパスのコンピュータ関連の見出し語 2,074 語を利用して実験を行う．本実験では，まず，2 つの見出し語間に階層関係があるのか，無関係であるのか判定を行う．さらに階層関係であると判定されたものについては，その上下関係の自動識別を行う．

まず，各見出し語に対する複数の説明文の精度を手動で以下のように判定した．

- A(見出し語を正しく説明している)
- B(見出し語を部分的に説明している)
- C(見出し語を説明していない)

以上の基準に従って，判定された説明文をもつ見出し語数と，見出し語あたりの平均説明文数を表 1 に示す．ALL とは判定 A，B，C をすべてまとめたものとなる．

次に判定された見出し語間の関係が正しいものであるかどうかを確認するために，手動で構築されたシソーラス (JICST 科学技術シソーラス 1992 年度版) から抽出される 2 つの単語の関係を正解データとして利用した．この JICST シソーラスは様々な文献を探索することによって構築された約 43,000 語にのぼる科学技術分野の検索用キーワード集である．そのうち，本実験で利用する Cyclone コーパスのコンピュータ関連の見出し語の 172 語が取り上げられていた．それらを説明文の精度に基づいて分類した結果を表 1 に示す．例えば，判定 A に関して正解データが 136 組とは，JICST シソーラスから抽出された見出し語 172 語の中で階層関係である組み合わせのうち，判定 A の説明文をもつ見出し語の組数を表す．

また，正解データに加えて，2 つの単語間には何も関係がないという無関係データを作成するために，見出し語 2,074 語からランダムに 500 組を抽出し，手動で単語間の関係を調査したところ，497 組が無関係であると判定された．説明文の精度に基づいて分類した結果を表 1 に示す．

評価実験では，以上に述べた正解データ，無関係データに対する識別率だけを報告する．すなわち，JICST シソーラスは見出し語 2,074 語のすべての階層関係を含んでいるわけではないので，テストセット (正解データと無関係データ) に含まれていないその他の語の関係については，今回調査は行わない．

4.2 単語間の階層関係・無関係の 2 群識別実験

説明文の方向性を考慮した出現頻度モデルを利用して，2 つの単語間に階層関係があるのか，無関係であるのかの識別実

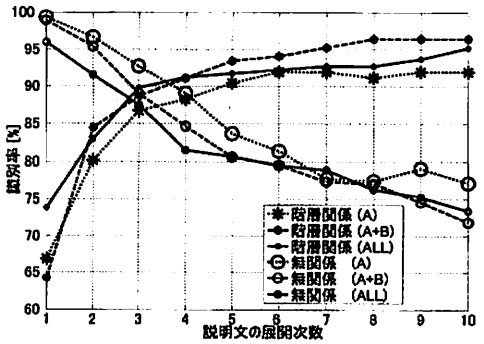

図 4　説明文の方向性を考慮した出現頻度モデルを利用したときの説明文の展開次数に対する単語間の階層関係・無関係の識別性能

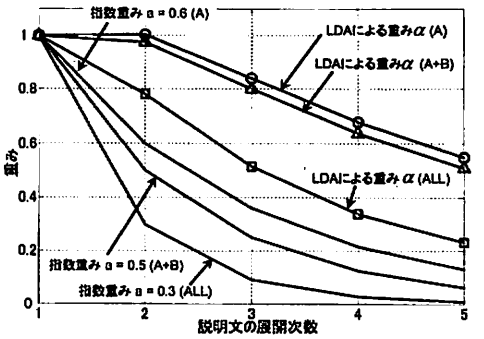

図 5　提案手法と指数重み手法の展開された説明文の重み

表 2　提案手法と指数重み手法の識別性能の比較

テストセット	A	A+B	ALL
提案手法 (%)	88.2	88.7	90.8
(展開次数)	4	3	3
指数重み手法 (%)	88.5	88.0	86.7
(階層関係)(%)	80.4	80.9	76.9
(無関係)(%)	96.5	95.1	96.4
(重み)(a)	0.6	0.5	0.3

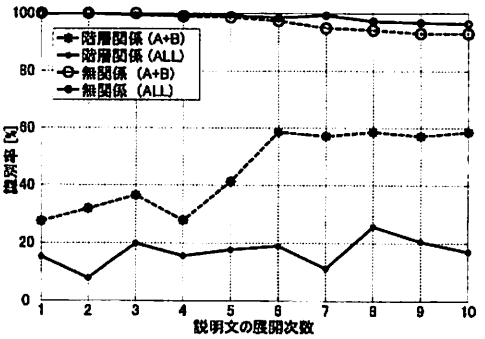

図 6　単一の説明文を利用したときの展開次数に対する単語間の階層関係・無関係の識別性能

験を行う．具体的には，式 (6) で計算される $|H(w_i|w_j)|$ の値を利用して，2 群識別を行う．3 つのテストセット A，A+B，ALL に対して，4-fold クロスバリデーションで評価を行う．学習データに対して LDA を行い，式 (5) の重み α_n を推定する．比較として，鈴木が提案した手法 [3](以後，「指数重み手法」とよぶ) も評価したこれは以下の式で拡張説明文 $\mathbf{C}$ を計算する．

$$\mathbf{C} = \lim_{k\to\infty} b(a\mathbf{A} + a^2\mathbf{A}^2 + \ldots + a^k\mathbf{A}^k) \quad 0 < a < 1 \quad (12)$$

ここで b は正規化係数であり，無限等比級数を考えているので，$b = (1-a)/a$ となる．こちらも 4-fold クロスバリデーションで評価を行う．つまり，テストセットの 3/4 を利用して最も識別率が高い a の値を推定し，その a を利用して残りの 1/4 のデータの識別率を計算する．

識別結果を図 4 に示す．階層関係の識別性能は，展開次数が増えるにつれて上昇する．一方，無関係の識別性能は展開次数の増加とともに低下する．この結果より，単語間の関係が階層関係であると識別するためには，説明文を展開することが有効であると考えられる．すなわち，単語間の間接的な関係を考慮することにより，その識別性能が向上したと考えられる．一方，無関係であると識別するためには，式 (6) の $H(w_i|w_j)$ が 0 に近いことが期待される．したがって，1 次説明文だけを利用した場合に，性能が最も高いのは明らかであり，説明文を展開することにより，逆に様々な単語間の間接的な関係が抽出され，識別性能が低下したと考えられる．

表 2 は，展開次数ごとに階層関係と無関係の識別率の平均値を算出し，その中で最も高い識別率とその展開次数を各テストセットごとに示す．同時に，鈴木の提案した指数重み手法による識別性能も示す．各テストセットに関しては，見出し語数が違うが，説明文の判定を厳しくすると説明が得られる語の数が変わるのは不可避的なことであるので，今回はこれらが別の基準で選んだ語彙の中での自然なばらつきと考えて，比較を行う．ALL に関して，提案手法は，もっとも識別性能の高い 90.8%に到達した．このことから，階層関係と無関係の識別には，見出し語に対して大量の説明文を収集することが有効であることが明らかとなった．一方，ALL のテストセットに指数重み手法を適用したとき，86.7%という最も低い性能であった．指数重み手法は無関係に関しては，高い識別率が得られた．しかし，階層関係に関しては低い識別率であった．なぜなら，指数重みは，図 5 に示されるように 1 次説明文を表す頻度行列に大きい重みを設定し，展開された説明文に対しては非常に小さい重みを設定しているためである．ここで，重みは各 1 次の重みによって正規化した．よって，提案手法は，説明文を展開することにより階層関係が明らかになるという手法であるため，指数重みの場合，高い識別率を得ることが難しいことがわかる．

図 6 は，各テストセットの中で，もっとも見出し語を適切に説明していると判定された説明文を利用したときの提案手法の結果である．ここでテストセット A に関しては，正方行列 A のスパース性のため，LDA により正しく重み α を推定できず，結果を得ることができなかった．全体的に階層関係の識別率は，

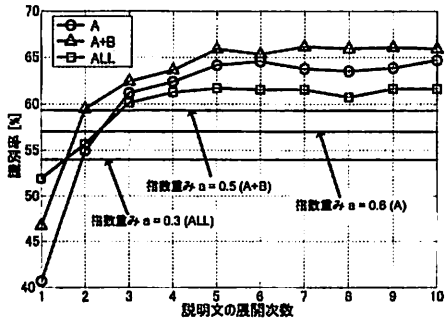

図 7　説明文の方向性を考慮した出現頻度モデルを利用した説明文の展開次数に対する単語間の階層関係の識別性能 (上下判定)

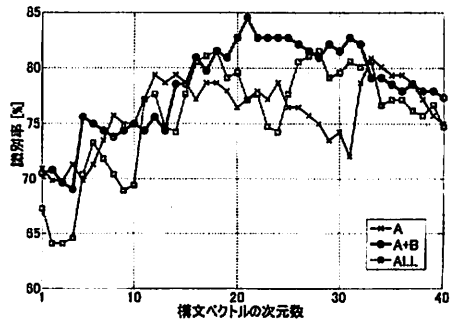

図 8　構文ベクトルを利用したモデルによる単語の上位下位の識別

表 3　構文ベクトルを利用したときの最も高い識別率とその展開次数，また全体の識別性能

テストセット	A	A+B	ALL
構文ベクトルによる識別性能 (%)	80.9	84.5	81.6
(構文ベクトルの大きさ)	34	22	29
階層関係 or 無関係 (%)	88.2	88.7	90.8
全体 (%)	71.4	75.0	74.1
出現頻度モデルによる識別性能 (%)	64.6	66.1	61.7

無関係の識別率に対して低い性能である．しかし，A+B の方が ALL よりも高い識別率であった．以上より，見出し語に対する説明文が 1 つであったとしても，見出し語を適切に説明している方 (A+B) が識別には有効であることが明らかとなった．

以上の実験より，階層関係か無関係かを識別するためには，見出し語に対する説明文を大量に集めること，また説明文の説明の精度が高いことを要求されることが明らかとなった．同時に，説明文を多く集めることによって，説明文の正確さを補償することになるということも明らかとなった．

4.3　階層関係をもつ単語間の上下判定のための識別実験

階層関係にある単語間のうち，どちらが上位語で，下位語であるかを判定するための識別実験を行う．図 7 は式 (6) から計算される階層関係の上下判定の性能を示す．A+B を利用した場合，展開次数 7 次で最も高い 66.1%になった．それゆえに，階層関係の上下判定には，見出し語に対して正確な説明文が有効であることがわかる．一方で，指数重みを使った場合は，どのテストセットに対しても性能は 60%以下である．これは高次の展開説明文に対する重みが小さいためであると考えられる．

また，局所的な構文情報に基づいた統計モデルを利用して識別性能を評価する．ここでは $N = 5$ として 5 次までの展開説明文を利用して式 (10) を計算した．これは，図 7 において，展開次数が 5 次のときに，すべてのテストセットに対して，平均的に高い識別性能が得られたためである．図 8 は構文ベクトルの次元を変化させたときの識別結果である．テストセットごとの最も高い識別性能と，全体の識別性能 (表 2 の提案手法の性能は，単語間の階層関係の有無を識別した結果であり，これに本節で求められる構文ベクトルを利用した単語間の上下判定の識別性能を掛け合わせたもの) を表 3 に示す．表 3 より，どのテストセットに対しても構文ベクトルによる単語間の上下判定の識別率は 80%以上得られた．特に ALL に対しては構文ベクトルを 29 次元まで利用した場合 81.6%であった．これは，説明文の方向性を考慮した出現頻度モデルによる階層関係の識別性能 61.7%と比較して，51.9%の誤りが改善された．最終的に，ALL に関する全体の識別性能では，74.1%($0.908 \times 0.816 \times 100$) となった．

5.　まとめと今後の課題

Cyclone コーパスを利用して，単語間の階層関係を判定するための統計的推定手法を提案した．説明文の方向性を考慮した出現頻度モデルと局所的な構文情報に基づく統計モデルにより，JICST シソーラスに記述された単語間の階層関係のうち，74.1%を検出することができた．実験結果より，単語間の階層関係を識別するためには，見出し語を説明する精度の高い説明文を数多く集めることが重要であった．したがって，この条件を満たしている Cyclone コーパスは，語彙における階層関係を推定するために効果的なコーパスであることがわかった．今後の課題は，単語間の意味的な関係として同義語，関連語の概念を導入することである．これは，単語ペアが，階層関係であるのか，同義・関連関係であるのか，または全く意味的な関係をもたないかの 3 カテゴリの識別問題に拡張することである．

文　献

[1] Marti, A. H.: Automatic Acquisition of Hyponyms from Large Text Corpora, *COLING-1992* (1992).

[2] 鶴丸弘昭，竹下克典，伊丹克企，柳川俊英，　吉田将：国語辞典情報を用いたシソーラスの作成について，情処研報自然言語処理，Vol. 1991, No. 37 (1991).

[3] Suzuki, S.: Probabilistic Word Vector and Similarity based on dictionaries, *CICLing-2003*, pp. 564–574 (2003).

[4] Shinzato, K. and Torisawa, K.: Extracting Hyponyms of prespecified hypernyms from itemizations and headings in web documents, *COLING-2004*, pp. 938–944 (2004).

[5] Uramoto, N.: Automatic Placing Unknown Words to Thesaurus Using Information from Corpora, *COLING-1996* (1996).

[6] Fujii, A., Itou, K. and Ishikawa, T.: Cyclone: An Encyclopedic Web Search Site, *WWW-2005*, pp. 1184–1185 (2005).

[7] Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning -Data Mining, Inference and Prediction-*, Springer (2001).