LF-001

事典コーパスを用いた単語階層関係の統計的解析
Statistical Analysis for the Word Hierarchy Using an Encyclopedic Corpus

大石 康智* 伊藤 克亘* 武田 一哉* 藤井 敦† 板倉 文忠‡
Yasunori OHISHI Katsunobu ITOU Kazuya TAKEDA Atsushi FUJII Fumitada ITAKURA

1. はじめに

高度な情報検索や自然言語処理には，多様な辞書が必要である．その中でも，有用な辞書に，単語間の上位下位関係や同義関係を表現したシソーラスがある．シソーラスは情報検索におけるユーザが入力した検索式の拡張や単語間の意味的な距離を計算することで多義性の解消，機械翻訳といった多くのシステムにおいて利用されている．

シソーラスは，人手によって作成されるのが主流である．しかし単語と単語の関係を手作業で調べるため，時間と手間を要する．また単語間の関係を理解するためには，情報量の多い意味解析のできる辞書が必要となる．つまり単一の国語辞典や百科事典のみでは，語の説明が少量であり，また新しい事柄や専門技術，新しい定義などが頻繁に改定されるわけではないため，単語と単語の深い意味的な関係を見出すことができないという問題点がある．

一方，シソーラスの自動構築への試みとして，文書中の語の並列関係の表現形式のパターン化をすることによって，テキストコーパスから単語の同義関係を自動的に抽出する研究がある [1][2]．しかし，文章中には様々な表現があるためその表現のパターンを網羅的に特定することは困難である．

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 [3]．既存のシソーラス自体が人手で作られたものなので完全な自動化へは至っていない．

本研究では，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な手法で単語間の意味的な関係を自動的に推定する．学習用には，Web から構築した事典コーパス [4] を用いる．このコーパスを用いて統計的に2単語間の階層関係を判別する手法を提案し，評価をおこなう．

2. 事典コーパス

本研究で用いる事典コーパスは，Web を事典的に利用することを目的として、約 20 万語のデータを整備して作成したものである．
当事典コーパスの構築は，以下のように行った．

1. Web 検索エンジンを用い，ある見出し語を含む 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ページのレイアウトを解析し，見出し語を含む領域 (段落) を抽出し，これを見出し語に対する説明文とする

*名古屋大学大学院情報科学研究科
†筑波大学大学院図書館情報メディア研究科
‡名城大学理工学部

つまり1つの見出し語につき，Web から集めた説明文が多数存在する．その結果，20 万語の見出し語に対し，平均 10 以上の説明文を集めることができている．現時点では，情報通信技術 (IT) 分野における見出し語の各説明文に対して，以下のような判定を行った．

- A (見出し語を説明している)
- 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)

この3段階で判定し，さらに用語の語義や分野に応じて分類をする．
本研究ではこのコーパスが1つの見出し語につき，多数の説明文をもつという構成により，出現頻度に基づく単語の共起が明確になり，単語と単語の上位下位関係，同義関係を推定できると考えられるため使用することとした．

3. 単語間の上位下位関係の推定

3.1 単語間の上位下位関係の指標

ある単語を説明するとき,「～の種類」「～のひとつ」というような表現をするのが一般的である．ここで「～」は見出し語の上位語にあたる．例えば,「ライオン」の説明文では「ネコ科の哺乳類」と表現する．つまり説明文中において「哺乳類」という単語の出現頻度が高い．しかし，見出し語「哺乳類」を説明するとき「ライオン」という単語を説明に用いることは少ない．つまり説明文において「ライオン」の出現頻度が低い．このことから説明文に出現する単語は，見出し語の上位語の出現頻度が高い。この場合,「哺乳類」が「ライオン」の上位語であると推定できる．つまり，見出し語 w_i の説明文中における見出し語 w_j の出現頻度と見出し語 w_j の説明文中における見出し語 w_i の出現頻度を比較することが単語間の上位下位関係を推定する1つの指標であると考えられる．

3.2 拡張説明文

説明文における単語の出現頻度を比較することで見出し語の上位語を推定することは可能である．ただ，本研究で用いる事典コーパスは Web から集めてきた説明文であるため，各説明文における信頼性の水準に差が生じている．
そこで，ある見出し語の説明文中に出現する単語も説明文をもつ，と考えることで説明文を再帰的に展開する手法を利用する．例えば,「ROM」の説明文中に見出し語「RAM」が出現していたとする．また「RAM」の説明文では「記憶装置」が出現していたとする．このとき，説明文を展開することで「ROM」の上位語として「記憶装置」を推定することが可能となる．

つまり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無限に展開できる．この手法は拡張説明文 [5] 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 w_i の n 次説明文中に単語 w_j が現れる確率を $P(w_j^{(n)}|w_i)$ とすると，1 次説明文での単語間の関係は

$$A = \begin{bmatrix} P(w_1^{(1)}|w_1) & P(w_1^{(1)}|w_2) & \dots & P(w_1^{(1)}|w_m) \\ P(w_2^{(1)}|w_1) & \ddots & & P(w_2^{(1)}|w_m) \\ \vdots & & \ddots & \vdots \\ P(w_m^{(1)}|w_1) & P(w_m^{(1)}|w_2) & \dots & P(w_m^{(1)}|w_m) \end{bmatrix}$$

と表される．$N(w_j^{(1)}|w_i)$ を説明文中の単語の出現頻度とすると，各要素は

$$P(w_j^{(1)}|w_i) = \frac{N(w_j^{(1)}|w_i)}{\sum_{k=1}^{m} N(w_k^{(1)}|w_i)} \quad (1)$$

と表される．2 次説明文に関しては，

$$P(w_j^{(2)}|w_i) = \sum_{all\ k} P(w_j^{(2)}|w_k^{(1)}) P(w_k^{(1)}|w_i) \quad (2)$$

が成立ち，行列 A を用いれば全体は A^2 と表せる．同様に，n 次説明文に関しては，全体を表す式は A^n となる．
ここで 1 次説明文から ∞ 次説明文までの全てをまとめた，拡張説明文

$$C = \alpha_1 A + \alpha_2 A^2 + \ldots + \alpha_n A^n + \ldots \quad (3)$$

を定義する．α_n は n 次説明文の全体に対する重みである．拡張説明文 C の要素にあたる $C(w_j, w_i)$ の値は，見出し語 w_i の拡張した説明文中における見出し語 w_j の頻度に基づく出現確率を表している．

3.3 類似研究との比較

拡張説明文 [5] は本来，単語間の類似度を算出する手法として提案された．見出し語 w_i から見出し語 w_j を想起する確率は，

$$P(w_j|w_i) = \sum_{all\ k} P(w_j|w_k) P(w_k|w_i) \quad (4)$$

により与えられる．そこで国語辞典を用いて拡張説明文 C を算出し，その要素を用いて計算される $P(w_j|w_i)$ を単語間の類似度として考え，同義語を抽出する．ここで拡張説明文 C における各 n 次説明文の重みは 3.4.2 節で述べる指数重みを用いている．
また文献 [9] では，ある観点に対する単語間の類似性判別を行なっている．この文献においても辞書の語義文中における単語を再帰的に展開することの有効性が示されている．しかし，単語の類似性判別にとどまり，単語の上位下位構造の推定は行なわれていない．

本研究では，拡張説明文 C にこそ，単語と単語の意味的なベクトルが含まれていると考えることで，単語間の上位下位関係を推定するために利用する．
先に述べた単語間の意味的な上位下位関係を導く指標を用いると見出し語 w_i と w_j における $C(w_j, w_i)$ の値とその対称の成分 $C(w_i, w_j)$ の値を比較することになる．本研究では，

$$d = C(w_j, w_i) - C(w_i, w_j) \quad (5)$$

を計算し，d の値が正であれば w_j は w_i の上位の語，負であれば w_j は w_i の下位の語であると推定することにする．

3.4 拡張説明文における重み

式 (3) の拡張説明文 C を算出するときにおける各 n 次説明文の重みのつけ方について二種類の方法を提案する．

3.4.1 最適な重みの推定

式 (3) において ∞ 次説明文までを考慮するのではなく，低次の説明文のみを用いて，単語と単語の上位下位関係を推定する．各低次の説明文に対する最適な重みを線形判別分析により学習し，拡張説明文 C を計算する．これは見出し語数を増やし大規模にモデルを試すときに生じる計算量の問題に対処するためである．この学習，評価法については 4 章で説明する．

3.4.2 指数重み

1 次説明文ほどその見出し語を直接的に表現しているという考えから 1 次説明文に最も高い重みを与え，n の値に応じて指数的に減少するような重みを考える．a を定数とするとき，

$$C = b(aA + a^2 A^2 + \ldots + a^n A^n + \ldots) \quad (6)$$

$$b = 1/\sum_{k=1}^{\infty} a^k, \quad 0 < a < 1 \quad (7)$$

となり，これらの式から特に $\det(I - aA) \neq 0$ ならば，

$$C = abA(I - aA)^{-1} \quad (8)$$

となり，拡張説明文 C を計算する．

4. 評価実験

語彙中の語と語の上位下位関係を語に対応する説明文を用いて推定をおこなう．このとき説明文の質による推定精度を確認するために，事典コーパスにおいて説明文が人手によって A，B，C と判定されている IT 用語に限定し，その中の語の上位下位関係を推定する．今回はその見出し語に対応する複数の説明文すべてをまとめて，1 つの説明文と考える．表 1 に使用した IT 用語の説明文の判定別データを示す．
評価用としては JICST 科学技術シソーラス 1999 年度版 (約 43000 語を記述)[6] を用いる．その中で表 2 のよう

表 1: 事典コーパスにおける IT 関係の見出し語の説明文

判定*	見出し語	平均説明文数**	判定データ***
A	1625	6.62	376 組
B	1594	7.84	324 組
A,B	1801	10.4	376 組
C	2065	74.3	375 組
A,B,C	2077	80.7	376 組

* 説明文の判定
** 見出し語あたりの平均説明文数
*** 見出し語のうち JICST シソーラスにおいて上位下位の判定がされていたもの

に IT 用語中の二つの見出し語の上位下位関係の判定がされているものを抽出した．JICST シソーラスに記述されていない IT 用語の上位下位関係については，今回は評価をおこなっていない．

JICST シソーラスから抽出したデータを 4 等分し，そのうち 1 つを評価用に，残り 3 つを学習用データとする．この 4 等分とは，判定の対象となる二つの見出し語のもつ説明文数の和の多い順に並び替え，均等にデータセットを 4 つ作成することである．

4.1 重みの学習

まず式 (5) を n 次説明文まで用いて，以下のように展開する．

$$\begin{aligned} d &= C(w_j, w_i) - C(w_i, w_j) \\ &= \alpha_1\{P(w_j^{(1)}|w_i) - P(w_i^{(1)}|w_j)\} \\ &+ \alpha_2\{P(w_j^{(2)}|w_i) - P(w_i^{(2)}|w_j)\} \\ &+ \ldots \\ &+ \alpha_n\{P(w_j^{(n)}|w_i) - P(w_i^{(n)}|w_j)\} \end{aligned} \tag{9}$$

式 (9) を線形判別関数と考え，JICST シソーラスからの見出し語間の正しい上位下位関係を示す学習用データを用いて d が正の値と，負の値の 2 クラスに判別できるように係数 α を求める．

ここでフィッシャーの線形判別法 [7] を利用する．これは p 個の特徴量に対して

$$z = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p \tag{10}$$

という線形結合 z の値によって判別ができるように係数 α を決定する手法である．2 クラスの判別を行なうときの z の変動を表す平方和は，

$$\begin{aligned} \sum_{k=1}^{2}\sum_{i=1}^{n_k}(z_i^{(k)} - \bar{z})^2 &= \sum_{k=1}^{2} n_k(\bar{z}^{(k)} - \bar{z})^2 \\ &+ \sum_{k=1}^{2}\sum_{i=1}^{n_k}(z_i^{(k)} - \bar{z}^{(k)})^2 \end{aligned} \tag{11}$$

と展開される．右辺第 1 項はクラス間平方和 S_B，第 2 項はクラス内平方和 S_W である．n_k は各クラスにおける要素数である．

表 2: JICST シソーラスから抽出した事典コーパスの IT 用語

抽出した IT 用語の例
$w_1 > w_2$
高水準言語 > C 言語
数理計画法 > 線形計画法
統計手法 > 数量化理論

* $w_1 > w_2$ は w_1 が w_2 の上位語であること示す

すなわち z をよく判別できるためにはクラス内平方和 S_W を小さく，クラス間平方和 S_B を大きくするように係数 α を決定することに帰着する．そこでフィッシャーは評価基準として式 (12) を定義した．

$$J_S = \frac{S_B}{S_W} \tag{12}$$

J_S の値が大きくなるように係数 α を決定すればよい．式 (12) をさらに展開すると固有値問題が得られる．この最大固有値に対応する固有ベクトルが係数 α となる．

4.2 重みの評価

学習した重み α を用いて，JICST シソーラスから得られた評価用データの正解率を算出する．表 2 のような評価用データにおける見出し語 w_1 と w_2 の組に対して，式 (9) より

$$d = C(w_1, w_2) - C(w_2, w_1) \tag{13}$$

を計算し，値が正であったものを評価用データにおける全ての見出し語の組で割ったものを正解率，式 (14) とした．

$$\text{正解率} = \frac{d > 0 \text{ の見出し語の組}}{\text{評価用セットにおける見出し語の組の総数}} \tag{14}$$

4.3 線形判別法により推定した重みの検証

1. 1 次説明文 (A) と 2 次説明文 (A^2) を計算し，JICST シソーラスにおける学習用データを用いて，その重み α をフィッシャーの線形判別法により学習する
2. 学習した重み α を用いて，JICST シソーラスから得られた評価用セットにおける正解率を計算する
3. 学習用，評価用データの組み合わせにより，4 回のクロスバリデーション (交叉検定) をおこない，正解率の平均値をもとめる
4. 展開した説明文を特徴量として増やし，同様の学習，評価を繰り返す

説明文の判定別に重みの推定をおこなったときの正解率の推移を図 1 に示す．

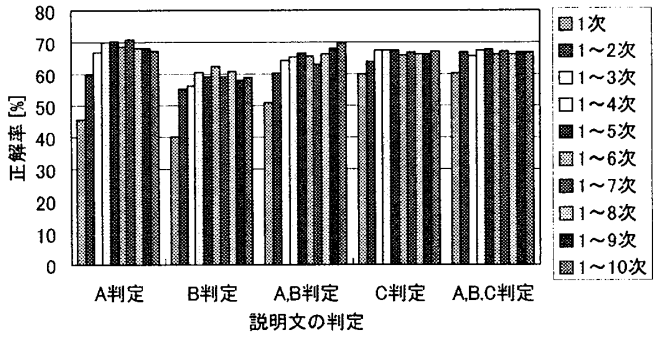

図 1: 特徴量 (次元) の増加に伴う正解率の推移

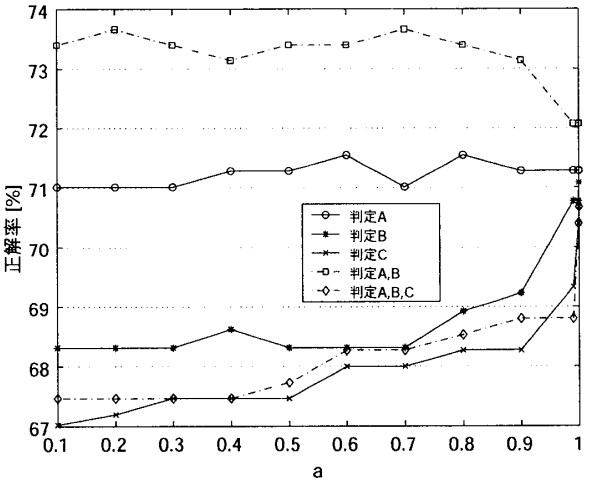

図 2: 指数的な重みの変化による正解率の推移

4.4 指数重みの検証

式 (8) の定数 a の値を変化させながら拡張説明文を算出する．そして JICST シソーラスからの見出し語間の正しい上位下位関係を示すデータに対して，式 (13) による d の値と式 (14) から正解率を算出する．そのときの正解率の推移を図 2 に示す．

5. 考察

図 1 において，低次の展開した説明文のみを用いて学習による最適な重みを推定したところ，A 判定の説明文では 1～7 次説明文を用いたときに 70.8% の正解率を得た．また C 判定の説明文では 1～5 次説明文を用いたときに 67.5% と高い正解率を得ることができた．A 判定の説明文は，1 次説明文のみを用いただけでは正解率が 45.5% であったものの，説明文を展開することによって正解率は向上し，見出し語間の上位下位関係を推定することが可能であるといえる．また，判定 C の説明文に関しては 1 見出し語あたりの平均説明文数が 74.3 文と非常に多いため，上位語の推定性能が高いと考えられる．つまり，Web のような信頼性の水準に差があり，説明の仕方における視点が異なる説明文であっても，大規模に説明文を集めることで上位語の推定は可能であると考えられる．

図 2 から，パラメータ a が 1 に近づくにつれて，判定 B，C の説明文を用いた拡張説明文では，正解率が向上していくのがわかる．つまり 1 次説明文ほど高い重みを与え，指数的になだらかに重みを減少させていくことで正解率が上がっている．これは，低次の説明文ほど見出し語の上位の語が含まれている可能性が高いことを示している．最も高い正解率は判定 A,B の説明文を使用し $a = 0.7$ のときであり，73.7% 得られた．

重みのつけ方について比較すると図 2 では一見，図 1 よりも高い正解率が得られているが，見出し語数を増やすにつれて，式 (8) における逆行列の計算量が増えることにつながる．しかし図 1 における学習による最適な重みの推定を行えば，さらに展開した説明文 (次元) を増やしても次元数の固有値問題を解くことに帰着するので少ない計算量で高い正解率が得られると予想される．

また上位下位の判定において，閾値を 0 にするのではなく，d の値が 0 付近のものを 1 つのクラスとして抽出することができれば，同義語として単語をまとめることが可能であると考えられる．この閾値のとり方について今後考えていく必要がある．

6. まとめと今後の展開

大規模かつ事典的な性質を持つコーパスを用いて、単語の共起から意味的な上位下位関係を推定した．その結果，ある見出し語の説明文を再帰的に展開すること，説明文の量を増やすことにより，70%に近い正解率で見出し語間の上位と下位を推定することができた．今後は，重みの学習における判別の閾値について検討する必要がある．

参考文献

[1] Marti A.Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora" *Proceedings of the Fourteenth International Conference on Computational Linguistics*,July.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" 情報処理学会研究報告, 1991-NL-83

[3] 浦本直彦, "コーパスに基づくシソーラス－統計情報を用いた既存のシソーラスへの未知語の配置" 情報処理学会論文誌,Vol.37,No.12,pp.2182-2189,Dec.1996.

[4] 藤井敦, 伊藤克亘, 石川徹也, "WWW は百科事典として使えるか？－大規模コーパスの構築－" 情報処理学会研究報告, 2002-NL-149

[5] 鈴木敏, "辞書に基づく単語の確率ベクトル" 技術情報レターズ (FIT2002),vol.1,pp.79-80,2002.

[6] JST(JICST) 科学技術シソーラス 1999 年版, http://jois.jst.go.jp/JOIS/html/thesaurus_index.htm.

[7] 石井健一郎, 上田修功, 前田英作, 村瀬洋, "パターン認識" オーム社 (2002)

[8] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明, "日本語形態素解析システム「茶筌」version 2.3."

[9] 笠原要, 松澤和光, 石川勉, "国語辞典を利用した日常語の類似性判別" 情報処理学会論文誌, Vol.38,No.7,pp.1272-1284,1997