Vol.2022-ICS-206 No.10
2022/3/10

空間的分布に基づく学術用語の階層化

甘川 由理[1]　浅谷 公威[2]　磯沼 大[2]　坂田 一郎[1,2]

概要：学術領域を理解する上で学術用語同士の関係性を理解することは重要であるが，専門分野の細分化等による学術用語の増加により，用語の関係性を人手によって整理することは困難である．本論文ではその中でも学術用語の上位下位関係に注目する．これまでも単語の上位下位関係を推定する研究は単語の共起に基づく方法をはじめとして多く行われているが，本論文では，ある学術用語が使用される論文の分布の広がりや分布の包含関係を指標化することによって，2 つの学術用語間の上位語下位語関係の有無およびその方向性の判別手法を提案する．この手法では WordNet のような大規模な上下関係のコーパスを用いずに，小規模な学習データにより推定が行える．Scopus の論文情報と JST 科学技術用語シソーラスを用いた検証では，既存の指標に比べ，学術用語の上位下位関係を高精度で判別することが可能であるとの結果が得られた．

1. 序論

1.1 研究背景

学術論文の出版数は年々指数関数的に増加しており [1]，それに伴って学術用語の種類も線形に増加している [2]．学術用語に表される概念の増加に伴って学術分野の専門性は深まり [3]，研究者は自分自身の専門領域ですら最新の動向を追うことが困難になっている．一方で，引用数の多い論文はそうでない論文に比べて高い学際性を持っている傾向にあり，科学的に大きなインパクトのある研究には幅広い分野の知識が必要である [4]．したがって，研究者は自身の専門分野に関しての最新の動向を追いつつも，関連する幅広い分野の知識もある程度理解しておく必要がある．そのためには学術概念を整理する必要がある．

単語の上位下位関係の推定タスクは自然言語処理の分野で研究が行われており，2 単語間の上位下位関係の有無を判別する Detection タスクと 2 単語が上位下位関係にあるときの方向性を判別する Directionality タスク [5] として一般化されており，近年では高精度の手法も多く提案されている [6][7]．これらの論文のほとんどは，一般的な単語を対象に上位下位関係を定義した WordNet[8] 等をもとにして学習をおこなっている．しかし，学術用語の場合そういったシソーラスが少なく専門性の高い用語が多いため適用できる既存手法は限られてくる．また，新しい用語が次々と生まれてくることから学術用語の上位下位関係の推定は一般的な単語の上位下位関係の推定に比べより難しいタスクとなっている．

このような学術用語の上位下位関係の推定においてはシソーラス等の情報を必須としない教師なし学習指標が適用可能であるが，既存手法 [9][10][11] では精度が低く精度向上の余地がある．既存手法は基本的には単語間の文章内での共起関係とその各単語の使用範囲の比較により包含関係を判別してきた．しかし，論文のアブストラクトといった現実に入手可能なデータを考えた場合，2 単語間に上下関係がある場合においても，必ずしも共通する共起があるとは限らない．さらに，同じ回数出現する単語でも，狭い範囲で使用されるものと広い範囲で使用されるものなどの違いは大きい．したがって，2 単語の包含関係を考える場合，このような空間的分布の比較を行う必要があるが，上位下位関係の推定において空間的分布を考慮し単語間の関係性を比較する手法は提案されてこなかった．

本研究では学術用語の空間的分布をその単語が出現する論文の空間的分布として求め，その分布を地理情報学で使われる手法 [12][13] を応用して定量化し，上位下位関係の推定を行う．論文の空間的分布は，論文間の引用ネットワークの表現学習 [14] によって得ることができる．このことにより，単語の使用分野の広さ・狭さを定量化するとともに，共通の共起語がない場合でも 2 単語間の分布の近さや包含関係を求めることが可能となる．本研究では JST 科学技術シソーラスによって上位下位関係が定義された単語ペアを用いて精度評価実験を行い，既存手法と比較して高

[1] 東京大学工学部システム創成学科
Systems Innovation,Faculty of Engineering,The University of Tokyo
[2] 東京大学大学院工学系研究科 技術経営戦略学専攻
Department of Technology Management for Innovation, The University of Tokyo

い精度精度を達成した.

2. 関連研究

2.1 単語の上位下位関係推定の既存のアプローチ

単語間の上位下位関係の推定は古くから研究されているテーマであり，大きく，パターンベースのアプローチと分布ベースのアプローチが存在している [7].

2.1.1 パターンベースのアプローチ

パターンベースのアプローチは 1992 年に Hearst によって最初に提案された手法であり，文法的な構造から単語の上下関係を推定する方法である [15]．例えば，X such as X1, X2, ..., Xn というような文章において X1, X2, ..., Xn は X の下位語であると定義することができる．このように文章の特定のパターンから単語間の上位下位関係を推定するのがパターンベースアプローチである．このような文法上のパターンはあらかじめ定義したものを使用することも可能だが，自動的な学習によって有効なパターンを見つけることも可能である [16][17]．この手法の問題点は，同じ文章内にパターン化された文章表現によって 2 つの単語が記されている必要があり，検出できないケースが多く存在してしまうことである．近年では，大規模なコーパスを使用することが可能になったことから，このような問題を解決することができる分布ベースのアプローチが主流となってきている.

2.1.2 分布ベースのアプローチ

分布ベースのアプローチは単語が使用される文脈の分布の特徴を捉えることで単語の上位下位関係を推定する方法である．この方法では，基本的に分布包含仮説 [18] と分布一般性 [9] という 2 つの分布意味論的直観が前提になっている.

分布包含仮説とは，ある 2 単語が上位下位関係にあるならば，上位語は下位語の構文的特徴を含んでいるという仮説である．これを共起する単語の観点で考えると，例えば，「猫」という単語は「走る」や「鳴く」といった単語と共起しやすいという特徴を，その上位語である「動物」も持っているということである．このような単語の包含関係を見ることで，上位下位関係を判断できると考えられる.

一方で，分布一般性とは上位語の方が下位語に比べてより広く分布するという性質である．例えば，「動物」という単語は「走る」，「泳ぐ」といった単語と共起しやすいと考えられるが，「猫」は「走る」と共起しやすいが，「泳ぐ」とは共起しにくいということである．このことからより広い文脈で出現する単語の方がより上位語らしいと判断することができる.

2.2 分布ベースのアプローチの詳細

前項の分布意味論的仮説に基づき，主に以下の 3 つのアプローチによって上位下位関係の推定が行われている.

- 教師なし学習：単語の出現文脈の分布特徴を数値化
- 教師あり学習：単語の分散表現を使った二値分類
- 半教師あり学習：単語の分散表現を上位語下位語を判定しやすいように加工して使用

2.2.1 教師なし学習

教師なし学習は基本的に分布意味論的観点に基づいた仮説を数値化することで 2 単語の上位下位関係を判断する方法である．以下ではある 2 つの単語 w_1，w_2 の上位下位関係を判定するものとする．分布意味論的観点から最初に提案された手法として Weeds[9] がある．この指標は共起している単語の相互情報量をもとに分布包含仮説における単語の含意関係を表現した指標である．単語 w_1 のベクトルは $\vec{w}_i = (w_{i1}, \ldots, w_{in})$ と表現され，このベクトルは共起頻度行列に基づく正の相互情報量 (PPMI) 行列の w_1 に該当する行のベクトルである．PPMI 行列は単語 w と w との共起語を c として以下の数式で表される.

$$PMI(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)} \tag{1}$$

$$PPMI(w,c) = \begin{cases} 0 & (PMI(w,c) \leq 0) \\ PMI(w,c) & (PMI(w,c) > 0) \end{cases} \tag{2}$$

$P(w,c)$: 単語 w と c が共起する確率
$P(w)$: 単語 w の出現確率
$P(c)$: 単語 c の出現確率

この PPMI 行列のベクトルの要素 $w_{i1}, \ldots, w_{in}$ を用いて WeedsP と WeedsR は式 3 で算出される.

$$\begin{aligned} WeedsP\,(w_1,w_2) &= \frac{\sum_{i \in F(w_1) \cap F(w_2)} w_{1i}}{\sum_{i \in F(w_1)} w_{1i}} \\ WeedsR\,(w_1,w_2) &= \frac{\sum_{i \in F(w_1) \cap F(w_2)} w_{2i}}{\sum_{i \in F(w_2)} w_{2i}} \end{aligned} \tag{3}$$

$F(w)$ は $\vec{w}$ の 0 ではない要素の列番号の集合を示している．したがって，例えば，
$\sum_{i \in F(w_1) \cap F(w_2)} w_{1i}$ は w_1 と w_2 の共通の共起語と単語 w_1 の PPMI の和を示している．ここで w_1 が下位語，w_2 が上位語であった場合，WeedsP は 1 に近い値を取り，WeedsR は 0 から 1 の間に収まることになる．そこで，上位下位関係の有無を判定する際には WeedsP の値もしくは $WeedsP - WeedsR$ が閾値を超えているかを基準とするのが指標 Weeds である.

Weeds にはいくつかの派生指標が提案されているがその中でも代表的な Clarke[10] と invCL[11] を紹介する．Clarke はほとんど Weeds 指標と同様の指標であり，invCL はこの Clarke を用いて分布包含仮説に加えて下位語の出現文脈を除いた文脈における上位語の分布の広さも考慮した指標である．Clarke の定義式を式 4 に invCL の定義式を式 5 に示す.

$$\begin{aligned} ClarkeP(w_1,w_2) &= \frac{\sum_{i \in F(w_1) \cap F(w_2)} \min(w_{1i}, w_{2i})}{\sum_{i \in F(w_1)} w_{1i}} \\ ClarkeR(w_1,w_2) &= \frac{\sum_{i \in F(w_1) \cap F(w_2)} \min(w_{1i}, w_{2i})}{\sum_{i \in F(w_2)} w_{2i}} \end{aligned} \tag{4}$$

$$invCL(w_1, w_2) = \sqrt{ClarkeP(w_1, w_2)(1 - ClarkeR(w_1, w_2))} \quad (5)$$

Clarke も Weeds と同じように閾値を設定して上下関係の有無を判定でき, ClarkeP の値もしくは$ClarkeP-ClarkeR$が閾値を超えているかを基準とする．invCL は，値が 1 に近いほど w_1 が下位語，w_2 が上位語という関係があると判断できる．

2.2.2 教師あり学習

この手法では，2 つの単語のベクトル表現に対して差を取る，2 つのベクトルを結合するなどの処理を行って生成した特徴量から，2 単語が上位下位関係を持つかどうかの二値分類を行う．学習アルゴリズムには，SVM やロジスティック回帰などが用いられている [19]．この手法は教師なし学習の場合に比べて精度が高いが，一方で 2 単語の関係性を学習しているのではなく，上位語になりやすい単語を学習しているだけであるという問題点が指摘されている [20]．この問題への対応として，教師なし学習で使われる分布の意味的特徴をとらえた特徴量（Weeds や invCL など）を用いて学習する方法が提案されており，これにより 2 単語の関係性の学習を促進できることが示されている [21]．

2.2.3 半教師あり学習

半教師あり学習では word2vec などで単語の分散表現を獲得する際にシソーラスで定義された単語間の関係性に関する教師データを導入することでより上位下位関係の判定をしやすい分散表現を獲得する方法が提案されている [22]．また，分散表現を獲得する際にシソーラス情報を入れて学習するのではなく，あらかじめ用意された word2vec や Fasttext のベクトル表現をシソーラスの情報を使って加工することで上位下位関係を反映した分散表現を獲得する方法も提案されている [6]．

2.3 ネットワーク表現学習

本節では，上位下位関係を推定したい単語を含む論文を多次元ベクトル空間に射影するために用いたネットワーク表現学習についての関連研究をまとめる．引用ネットワークやソーシャルネットワークなどのネットワークを効率的に分析するためには，どのようにしてネットワーク特徴を捉えるかが重要となる．多くの場合，ネットワークのエッジに着目した隣接行列が用いられるが，これはノード間の隣接関係のみを捉えている．このような単純な表現では，ネットワーク上のパスや頻出部分構造など，より複雑で高次のネットワーク構造を捉えることはできない．そこで提案されたのが，ネットワーク表現学習であり，これによりネットワークのトポロジー構造やノードの特徴などの情報を保持したままネットワークのノードをベクトル空間に埋め込みネットワーク構造を捉えることが可能になった．ネットワーク表現学習として様々な手法が提案されており，単語の埋め込み手法をグラフに適用した DeepWalk[23]，間接的なノードの近さも考慮した LINE[24]，学習時のサンプリング方法が工夫された node2vec[25] などが提案されている．これらを用いて得られた分散表現はノードの分類やリンク予測等に用いられることで有用性が示されており，ネットワーク表現学習により得られた空間的な情報（特に概念間の距離）は多くの情報を含むことが示唆される．そこで，本研究では，ネットワーク表現学習によって得られた分散表現の空間的分布に着目することで，従来の共起語による推定に比べて高精度で学術用語の上位下位関係を推定できるのではないかと考えた．

2.4 分布の広がりの定量化手法

空間的分布に着目するにあたって，本研究では人文地理学や都市工学の分野で主に用いられる 2 次元の点分布分析手法を応用したため，本節では地理的な空間における分析方法についてまとめる．人口の分布のような 2 次元分布における分布の広がりを定量化する初期の試みは点集合座標の標準偏差に注目した標準偏差楕円である [26]．標準偏差楕円は x 軸方向と y 軸方向の標準偏差を用いて．分布のばらつきを反映した楕円を求め，広がりを定量化する方法である．また，x 軸方向と y 軸方向の分散を足し合わせた値の平方根である標準距離によっても分布のばらつきが定量化されることが示されている [27]．しかし，これらの手法は点の数が同数でない場合の比較において問題が生じることから点の数に左右されない散布疎度が提案された [12]．本研究ではこの散布疎度をベースとして空間上の広がりを定量化する．

$$\text{散布疎度} = \frac{\text{N 個の平均相互距離}}{\text{N 個が最も均一・密集して分布する場合の平均相互距離}} \quad (6)$$

また，2 つの分布が存在した時，2 つの分布が互いに接近しているのか，互いに避けあっているのか，そのどちらでもないのかを判別する方法として式 7 で算出される相互最近隣距離法 [13] が使われている．例えば駅の分布と商業施設の分布は類似しているといったことを分析できる．本研究では，分布の包含関係を評価する際にこの手法に着想を得た．

$$D = \frac{1}{n_a + n_b}\left(\sum_{i=1}^{n_a} d_{ai} + \sum_{i=1}^{n_b} d_{bi}\right) \quad (7)$$

d_{ai}: 分布 A の点 i から, 分布 B の最も近い点までの距離
d_{bi}: 分布 B の点 i から, 分布 A の最も近い点までの距離

これらの地理学における指標は小売店の出店計画モデル [28] や伝染病の発生地点の分析から伝染病の原因や特徴の分析 [29]，犯罪の発生しやすい地点を予測する分析 [30] などに適用されている．

3. 提案手法

3.1 使用データ

本論文では Elsevier が管理する Scopus から抽出した書誌情報を用いた．Scopus は査読済み文献等の世界最大級の抄録・引用文献データベースである．論文の Embedding においては，1970 年 1 月から 2020 年 12 月までに出版された全 73,103,156 件のデータを利用し，各指標の計算においては計算時間の問題から全体のうちランダムにサンプリングした 100 万件の論文を用いて実験を行った．

単語の上位下位関係推定の精度評価においては JST 科学技術用語シソーラスを用いた．JST 科学技術用語シソーラスは，国立研究開発法人科学技術振興機構（JST）によって作成された科学技術用語辞書である [31]．JST 科学技術用語シソーラスは，科学技術全分野の専門用語を収録しており，同義関係や意味上の類似関係，階層関係といった情報が整理されている．本研究は学術用語の階層構造を把握するという目的があるため，特に科学技術用語に特化したデータベースである JST 科学技術用語シソーラスを活用した．

JST 科学技術用語シソーラスには，全 97,261 ペアの単語の上位下位関係が定義されており，どちらが上位語かを予測する方向性判別タスクのデータセットとして，上位下位関係のペアのデータの方向性をランダムに変更したデータを利用した．これらのデータのうち両方の単語が 100 万件サンプリングした論文データに 2 回以上含まれているペアを取り出すと全 30,786 件のペアが抽出された．このうちランダムに分割した 588 件を学習データもしくは教師なし学習の際の閾値調整用の検証用データとし，残りの 30,271 件のデータを精度を検証するためのテストデータとした．また，関係性判別タスクのデータセットとして，上位語-下位語のペア 97,261 件，下位語-上位語のペア 97,261 件，上位下位関係を持たないランダムに選択された単語のペア 97,250 件からなる全 291,772 件のペアのデータセットを作成した．このデータセットにおいて正例となるのは上位語-下位語のペア 97,261 件である．方向性判別の場合と同様に，ペアとなる 2 つの単語が対象データ中に 2 回以上出現する 175,220 件のデータを使用する．全 175,220 件のデータのうちランダムに分割した 3492 件を学習データもしくは教師なし学習の際の閾値調整用の検証用データとし，残りの 171,728 件のデータを精度を検証するためのテストデータとした．

3.2 提案指標の算出

本研究では，2 章で述べた分布意味論的観点を 2 つの指標を用いて捉える．一つは分布の広がりである分布一般性を捉える広範性指標である，2 つの分布が存在した時，それぞれの分布の広がりの大きさを定量化することで，上位語は下位語に比べて広く分布するという特徴を捉えることを目的とするのが広範性指標である．もう一つは分布包含仮説を捉える相互距離指標である．相互距離指標は 2 つの分布が存在した時，片方の分布から見たもう一方の分布までの距離を定量化することで，2 つの分布が包含関係にあるかを捉えることを目的に作成した指標である．

まず初めに Scopus の 1970 年から 2020 年までの全論文の引用ネットワークから最大連結成分に含まれる論文と引用関係を抽出し，重み無しの無効グラフとして LINE によって各論文の 128 次元のベクトル表現を獲得した．実装においては高速化のため Graphvite[32] というライブラリを用いた．

得られた各論文の分散表現を用いて提案指標の算出を行う．ここでは，ある 2 単語のペア w_1，w_2 の判定をする場合を考える．まず，100 万件サンプリングデータから w_1，w_2 をタイトルもしくはアブストラクトに含む論文群を取得する．次に，その論文群の LINE による分散表現を抽出する．この時，論文数が多いと計算に時間がかかることから，論文数が 100 本以上の場合は，ランダムに 100 論文を抽出し，この 100 本の論文のベクトルを使って指標の計算を行った．

1 つ目の指標である広範性指標は，w_1 に対する広範性指標を E_1，w_2 に対する広範性指標を E_2 として式 8，式 9 で算出する..

$$E_1 = \frac{\sum_{i \in P_1} \sum_{j \in P_1} ||\vec{e_i} - \vec{e_j}||}{\sum_{i \in R} \sum_{j \in R} ||\vec{e_i} - \vec{e_j}||} \tag{8}$$

$$E_2 = \frac{\sum_{i \in P_2} \sum_{j \in P_2} ||\vec{e_i} - \vec{e_j}||}{\sum_{i \in R} \sum_{j \in R} ||\vec{e_i} - \vec{e_j}||} \tag{9}$$

P_1: w_1 を含む論文群
P_2: w_2 を含む論文群
R: 全論文から P と同数だけランダムに抽出した論文群
$\vec{e_i}$: 点 i のベクトル

図 fig:E は広範性指標のイメージを示した図であり，w_1 が下位語で w_2 が上位語の場合，上位語の分布の方が大きくなるはずのため，$E1 < E2$ となる．

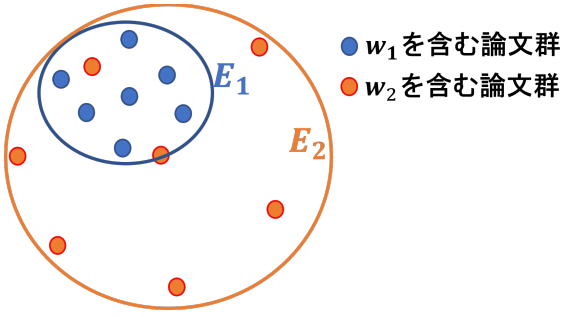

図 1 広範性指標のイメージ

2 つ目の指標である相互距離指標は，w_1 を含む論文群の

各点に対して w_2 を含む論文群の最近隣点を探し，その距離を平均したものと，w_1，w_2 を入れ替えて，w_2 を含む論文群の各点に対して w_1 を含む論文群の最近隣点を探し，その距離を平均したものをそれぞれ計算する．

$$MD_{12} = \sum_{i \in P_1} d_{2i} \tag{10}$$

$$MD_{21} = \sum_{i \in P_2} d_{1i} \tag{11}$$

P_1: w_1 を含む論文群
P_2: w_2 を含む論文群
$\vec{e_{1i}}$: P_1 の論文の分散表現
$\vec{e_{2i}}$: P_2 の論文の分散表現
d_{1i}: P_2 の論文のベクトルのうち $\vec{e_{1i}}$ 最も近いベクトルまでの距離
d_{2i}: P_1 の論文のベクトルのうち $\vec{e_{2i}}$ 最も近いベクトルまでの距離

図 2 は相互距離指標を w_1，w_2 に適用した際のイメージである．分布包含仮説によると w_1 の分布は w_2 の分布に含まれるため，MD_{21} が MD_{12} に比べ大きくなる．

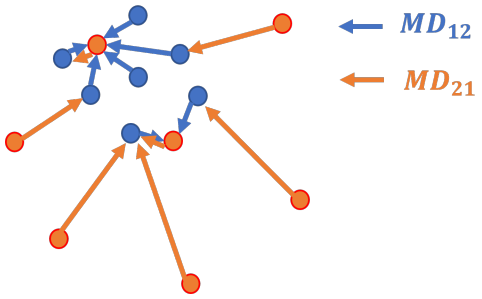

図 2 相互距離指標のイメージ

以上の処理により，方向性あるいは関係性を判別したい 2 単語のペアに対して E_1，E_2，MD_{12}，MD_{21} の 4 つの値が計算される．

3.3 比較指標

2 章で説明した定義に従って WeedsP, WeedsR, ClarkeP, ClarkeR，invCL をそれぞれ算出した．具体的な手順は以下の通りである．100 万件の論文サンプリングデータのタイトルとアブストラクトを結合し，小文字に変換したのち，統計的自然言語処理ライブラリである NLTK にストップワードとして登録されている単語を除く処理を行う．同一論文に出現する単語と 1 回ずつ共起しているとカウントして，各単語の PPMI 行列を作成した．次に，この PPMI 行列をもとに，計算対象となる単語を含む論文を抽出し，その論文に含まれる単語と共起しているとして，3 式，4 式，5 式で算出される各指標の数値を求めた．

既存指標に加えて，今回のタスクにおいて寄与すると考えられる，単語をタイトルもしくはアブストラクトに含む論文数も比較指標として用いる．

3.4 学習と精度評価方法

Weeds や invCL といった既存指標は当初は教師なし学習に使用される指標として提案された指標である．しかし，教師あり学習の特徴量として使用した場合，教師あり学習における関係性学習に寄与することが明らかになっているため [21]，本研究では教師なし学習として閾値の決定を行った場合と各指標を特徴量としてロジスティック回帰によって学習した教師あり学習の場合の 2 通りで検証を行った．

3.4.1 教師なし学習

既存指標と提案指標はどちらも値の大きさが意味を持っており，値に応じて判別が可能である．仮に，w_1 を下位語，w_2 を上位語とした場合，それぞれの指標の値がどのようになるのかまとめる．提案指標の広範性指標は w_1 を含む論文の広範性は小さく，w_2 を含む論文の広範性が大きくなることから $\frac{E_2}{E_1}$ が大きい方が w_1 が下位語で w_2 が上位語であると判定できることになる．提案指標の相互距離指標に関しては w_1 を含む論文群から w_2 の論文群に対する最近隣距離の方が w_2 を含む論文群から w_1 の論文群に対する最近隣距離に比べて小さくなると考えられることから $\frac{MD_{21}}{MD_{12}}$ が大きいほど w_1 が下位語であり，w_2 が上位語であると判別できる．また，広範性指標と相互距離指標を合わせた指標として，広範性指標と相互距離指標を掛け合わせたものを使用する．

Weeds は WeedsP が 1 に近い値となり，WeedsR が 0 から 1 の間の値を取る．したがって，WeedsP の値が大きいほど，あるいは，$WeedP - WeedR$ の値が大きいほど，w_1 が下位語で w_2 が上位語である可能性が高い判定することができる．Clake は Weeds の場合と同じであり，ClarkeP の値もしくは $ClakeP - ClakeR$ の値の大きさによって判定できる．invCL は 0 から 1 の値を取り，w_1 が下位語の場合には ClakeP が大きくなり，(1-ClakeR) が大きくなることから invCL の値が大きいほど w_1 が上位語で w_2 が下位語であると判断できる．

論文数は，上位語であるほど登場頻度が多く，下位語であるほど登場頻度が低いことが考えられるため，w_2 を含む論文数を w_1 を含む論文数で割った論文数の比が大きいほど，w_1 が上位語だと捉えることができる．

教師なし学習においては，検証用データにより F 値が最大となる閾値を求め，その閾値を使ってテストデータでの F 値を求め精度の比較を行う．評価する際に使用する指標は，$\frac{E_2}{E_1}$，$\frac{MD_{21}}{MD_{12}}$，$\frac{E_2}{E_1} \cdot \frac{MD_{21}}{MD_{12}}$，WeedsP，ClarkeP，invCL，論文数の比 の 7 つである．閾値の範囲は $\frac{E_2}{E_1}$，$\frac{MD_{21}}{MD_{12}}$，$\frac{E_2}{E_1} \cdot \frac{MD_{21}}{MD_{12}}$ は 0 から 1.5 の範囲で 0.015 刻みで最適なパラメータを求め，WeedsP，ClarkeP，invCL，論文数の比は 0 から 1 の範囲で 0.01 刻みで最適なパラメータを求めた．

3.4.2 教師あり学習

教師あり学習では，教師なし学習においてパラメータ調整用のデータとして使っていた検証用データを学習用データとしてロジスティック回帰を行い，各指標の F 値をテストデータを使って評価する．この時，特徴量として使用する組み合わせは以下の 7 通りである．

- E_1， E_2， MD_{12}， MD_{21} (提案指標)
- E_1， E_2 (提案指標)
- MD_{12}， MD_{21} (提案指標)
- WeedsP， WeedsR
- ClarkeP， ClakeR
- invCL
- w_1 を含む論文数， w_2 を含む論文数

ロジスティック回帰を使用する場合においては，w_1，w_2 でそれぞれ算出した値を割ったり，広範性指標と相互距離指標を掛け合わせて一つの値にしたりするよりもそれぞれの値を直接特徴量とした方が最適な組み合わせ方を学習可能であると考えたため，算出した値を特徴量として使用した．

4. 結果と考察

4.1 精度評価の結果

まず，教師なし学習における各指標のそれぞれのタスクにおけるテストデータの F 値を表 1 に示す．

表 1 教師なし学習による精度

	方向性判別	関係性判別
$\frac{E_2}{E_1}$ ・ $\frac{MD_{21}}{MD_{12}}$	0.847	0.686
$\frac{E_2}{E_1}$	0.836	0.672
$\frac{MD_{21}}{MD_{12}}$	0.813	0.683
WeedsP	0.747	0.679
ClarkeP	0.770	0.681
invCL	0.784	0.667
論文数の比	0.787	0.621

方向性判別タスク，関係性判別タスクとも，広範性指標と相互距離指標の 2 つを用いた指標が最も高い精度となった．しかし，関係性判別タスクにおいては他の指標と比べて精度にほとんど差が生じなかった．

次に，ロジスティック回帰を用いた場合の F 値を表 2 に示す．

表 2 ロジスティック回帰による精度

	方向性判別	関係性判別
E_1,E_2,MD_{12},MD_{21}	0.849	0.754
E_1,E_2	0.835	0.593
MD_{12},MD_{21}	0.805	0.696
WeedsP,WeedsR	0.774	0.678
ClarkeP,ClarkeR	0.773	0.678
invCL	0.776	0.673
論文数	0.703	0.482

教師なし学習の場合と同様に，広範性指標と相互距離指標の 2 つを組み合わせた場合が両方のタスクにおいて最も精度が高くなった．関係性判別タスクに注目すると，相互距離指標単体で用いた場合でも既存指標よりもわずかに精度は上回っている．一方で，広範性指標のみを用いた場合には，論文数を特徴量とした場合よりは上回っているものの，Weeds，Clarke，invCL に比べると精度が低い．これは関係性判別においてはそれぞれの単語が使用されている分布の広さだけでは，その単語同士が用いられる分布の近さを考慮できないためであると考えられる．一方で，相互距離指標はそれぞれの分布の距離に加えてある程度広さも考慮できるため単体でもそれほど精度が低下しなかったと考えられる．

広範性指標と相互距離指標の両方を用いる場合に関して，方向性判別タスクにおいては，教師なし学習を用いた場合と教師あり学習を用いた場合の精度の差は 0.002 であるのに対して，関係性判別タスクにおいてはその差が 0.068 と比較的大きくなっている．関係性判別タスクは方向性判別タスクと比較するとより複雑なタスクであるため，単純な両方の指標の積を用いるのではなく，ロジスティック回帰を使うことでより精度が高くなったのではないかと考えられる．

4.2 提案指標と比較指標の相関

関係性判別タスクにおいて算出した提案指標と比較に用いた各指標の相関関係を示したマトリクスを図 3 に示す．

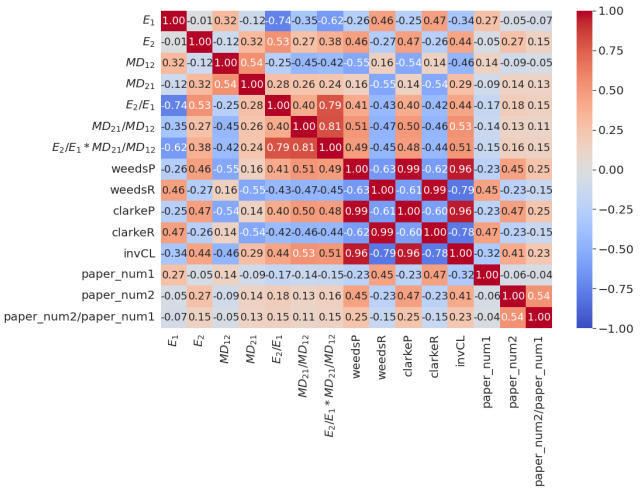

図 3 各指標の相関

まず，比較指標である Weeds，Clarke，invCL について確認すると，それぞれの指標間に強い相関が確認される．WeedsP と WeedsR の相関は-0.63 と負の相関を持っており，ClarkeP と ClarkeR も同様に互いに負の相関を持っている．また，WeedsP，ClarkeP，invCL の相関は 0.9 を超えておりお互いに非常に強い相関を持っている．それぞれ

の指標の定義が類似していることからも，既存の指標のWeedsP，WeedsR，invCL は互いに似たような傾向を捉えていることがわかる．

次に，広範性指標と WeedsP，ClarkeP，invCL の相関に注目する．WeedsP，ClarkeP，invCL は単語 w_1 の広範性指標とは-0.3 程度の非常に弱い負の相関を持っており，単語 w_2 の広範性指標に対しては 0.5 弱の正の相関を持っている．WeedsP，ClarkeP，invCL 同士の相関に比べて弱い相関であり，既存の指標とは異なる部分も捉えていることがわかる．相互距離指標と WeedsP，ClarkeP，invCL の間にも同様の傾向がみられるが，w_1 に対する相互距離指標との相関は-0.5 程度であるのに対して，w_2 に対する相関は 0.2 弱でありほとんど相関を持っていない．このような差が生じるのは，w_1 が上位語，w_2 が下位語であるときには，WeedsP は小さくなるのに対して MD_{12} は大きく，MD_{21} は小さくなり，w_1 が下位語で w_2 が上位語の場合は WeedsP は大きくなり，MD_{12} は小さく，MD_{21} は大きくなり，w_1 と w_2 がランダムな組み合わせの場合，WeedP は小さくなり，MD_{12} は大きく，MD_{21} は大きくなるためだと考えられる．MD_{12} は基本的にどのパターンも WeedsP とは逆の方向に動くのに対して，MD_{21} は w_1 と w_2 が上位語下位語関係もしくは下位語上位語関係にある場合には同じ方向に動くのに対して，ランダムなペアの時は逆方向に動くためである．

教師なし学習において用いた $\frac{E_2}{E_1}$，$\frac{MD_{21}}{MD_{12}}$，$\frac{E_2}{E_1} \cdot \frac{MD_{21}}{MD_{12}}$ との相関は 0.5 程度であり正の相関が確認されるが強いものではない．広範性指標と相互距離指標を組み合わせた $\frac{E_2}{E_1} \cdot \frac{MD_{21}}{MD_{12}}$ との相関は $\frac{E_2}{E_1}$ や $\frac{MD_{21}}{MD_{12}}$ と比べてほとんど差がないことから広範性指標と相互距離指標を組み合わせた場合でも既存指標とは異なる特徴を捉えていることが考えられる．

続いて，広範性指標と相互距離指標間の相関を確認すると 0.32 であり弱い正の相関が確認された．相互距離指標は互いの単語の類似性を反映すると同時に分布の広さもある程度反映する指標であるため正の相関を持つと考えられる．例えば，w_1 に対する相互距離指標が比較的大きい場合に考えられるパターンは，w_1 が上位語，w_2 が下位語であるために大きいパターンと，w_1 と w_2 は関連性の低い単語であり，それぞれの単語が使用される単語の分布が離れているパターンが考えられる．前者の場合，w_1 に対する広範性指標は比較的大きいと考えられるが，後者の場合は大きい場合も小さい場合もどちらも同程度存在すると考えられる．

最後に，論文数との相関に注目する．論文数と強い相関を持っている指標は存在していないが提案指標に比べて既存指標の方がやや相関が強い．

5. 結論

本研究では単語の上位下位関係を対象の単語を含む論文の広がりの広さと分布間の距離に基づいた包含関係から単語間の上位下位関係の有無と方向性を予測し，既存指標に比べて高精度であることを示した．本手法は教師なし学習においても一定の精度を出せていることから学術用語のように次々と新しい単語が登場し，単語の専門性が高く，一般的な単語に基づいたシソーラスではうまくカバーできないという性格を持つ用語一般に対しても適用可能な手法である．

今後の展望として，特定の分野の学術用語に対して上位下位関係の有無と方向性を推定し，その結果を階層構造として可視化することで，その分野の学術概念の関係性の理解を促進することができると考えられる．このような可視化は，新しく該当の分野に参入を試みる企業や研究者等にとって有益である．

また，本研究においては学術論文の引用関係によって得られた分散表現を活用したが，本手法は引用関係を持たないような文章に対しても Sentence BERT 等を活用することで適用可能であり，応用範囲を拡大することも可能である．

参考文献

[1] Bornmann, L. and Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, *Journal of the Association for Information Science and Technology*, Vol. 66, No. 11, pp. 2215–2222 (online), DOI: https://doi.org/10.1002/asi.23329 (2015).

[2] Milojević, S.: Quantifying the cognitive extent of science, *Journal of Informetrics*, Vol. 9, No. 4, pp. 962–973 (online), DOI: https://doi.org/10.1016/j.joi.2015.10.005 (2015).

[3] Wray, K. B.: Rethinking Scientific Specialization, *Social Studies of Science*, Vol. 35, No. 1, pp. 151–164 (online), DOI: 10.1177/0306312705045811 (2005). PMID: 15991447.

[4] Chen, S., Arsenault, C. and Larivière, V.: Are top-cited papers more interdisciplinary?, *Journal of Informetrics*, Vol. 9, No. 4, pp. 1034–1046 (online), DOI: https://doi.org/10.1016/j.joi.2015.09.003 (2015).

[5] Kiela, D., Rimell, L., Vulić, I. and Clark, S.: Exploiting Image Generality for Lexical Entailment Detection, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, Association for Computational Linguistics, pp. 119–124 (online), DOI: 10.3115/v1/P15-2020 (2015).

[6] Vulić, I. and Mrkšić, N.: Specialising Word Vectors for Lexical Entailment, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, Association for Computational Lin-

guistics, pp. 1134–1145 (online), DOI: 10.18653/v1/N18-1103 (2018).

[7] Roller, S., Kiela, D. and Nickel, M.: Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, pp. 358–363 (online), DOI: 10.18653/v1/P18-2057 (2018).

[8] Miller, G. A.: WordNet: A Lexical Database for English, *Proceedings of the Workshop on Human Language Technology*, HLT '93, USA, Association for Computational Linguistics, p. 409 (online), DOI: 10.3115/1075671.1075788 (1993).

[9] Weeds, J., Weir, D. and McCarthy, D.: Characterising Measures of Lexical Distributional Similarity, *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, COLING, pp. 1015–1021 (online), available from ⟨https://aclanthology.org/C04-1146⟩ (2004).

[10] Clarke, D.: Context-theoretic semantics for natural language: an overview, *Proceedings of the workshop on geometrical models of natural language semantics*, pp. 112–119 (2009).

[11] Lenci, A. and Benotto, G.: Identifying Hypernyms in Distributional Semantic Spaces, SemEval '12, USA, Association for Computational Linguistics, p. 75–79 (2012).

[12] 龔建新：事象の空間的分布に関する散布疎度統計量の構築，人文地理， Vol. 46, No. 5, pp. 455–473 (1994).

[13] Lee, Y.: A Nearest-Neighbor Spatial-Association Measure for the Analysis of Firm Interdependence, *Environment and Planning A: Economy and Space*, Vol. 11, No. 2, pp. 169–176 (online), DOI: 10.1068/a110169 (1979).

[14] Zhang, D., Yin, J., Zhu, X. and Zhang, C.: Network Representation Learning: A Survey, *IEEE Transactions on Big Data*, Vol. 6, No. 1, pp. 3–28 (online), DOI: 10.1109/TBDATA.2018.2850013 (2020).

[15] Hearst, M. A.: Automatic Acquisition of Hyponyms from Large Text Corpora, *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, (online), available from ⟨https://aclanthology.org/C92-2082⟩ (1992).

[16] Snow, R., Jurafsky, D. and Ng, A. Y.: Learning Syntactic Patterns for Automatic Hypernym Discovery, NIPS'04, Cambridge, MA, USA, MIT Press, p. 1297–1304 (2004).

[17] Shwartz, V., Goldberg, Y. and Dagan, I.: Improving Hypernymy Detection with an Integrated Path-based and Distributional Method, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2016).

[18] Geffet, M. and Dagan, I.: The Distributional Inclusion Hypotheses and Lexical Entailment, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, Association for Computational Linguistics, pp. 107–114 (online), DOI: 10.3115/1219840.1219854 (2005).

[19] Roller, S., Erk, K. and Boleda, G.: Inclusive yet selective: Supervised distributional hypernymy detection, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1025–1036 (2014).

[20] Shwartz, V., Santus, E. and Schlechtweg, D.: Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, Association for Computational Linguistics, pp. 65–75 (online), available from ⟨https://aclanthology.org/E17-1007⟩ (2017).

[21] 鷲尾光樹：語の分散表現と上位下位関係―研究動向と今後への試案―，*SIG-AM*, Vol. 13, No. 03, pp. 14–21 (2016).

[22] Nguyen, K. A., Köper, M., Schulte im Walde, S. and Vu, N. T.: Hierarchical Embeddings for Hypernymy Detection and Directionality, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 233–243 (online), DOI: 10.18653/v1/D17-1022 (2017).

[23] Perozzi, B., Al-Rfou, R. and Skiena, S.: DeepWalk: Online Learning of Social Representations, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, New York, NY, USA, Association for Computing Machinery, p. 701–710 (online), DOI: 10.1145/2623330.2623732 (2014).

[24] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. and Mei, Q.: LINE: Large-Scale Information Network Embedding, WWW '15, Republic and Canton of Geneva, CHE, International World Wide Web Conferences Steering Committee, p. 1067–1077 (online), DOI: 10.1145/2736277.2741093 (2015).

[25] Grover, A. and Leskovec, J.: Node2vec: Scalable Feature Learning for Networks, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, Association for Computing Machinery, p. 855–864 (online), DOI: 10.1145/2939672.2939754 (2016).

[26] Lefever, D. W.: Measuring geographic concentration by means of the standard deviational ellipse, *American journal of sociology*, Vol. 32, No. 1, pp. 88–94 (1926).

[27] Furfey, P. H.: A note on Lefever's" standard deviational ellipse", *American Journal of Sociology*, Vol. 33, No. 1, pp. 94–98 (1927).

[28] Vega, R. S., Acuña, J. L. G. and Díaz, M. R.: SPATIAL ANALYSIS OF CONSUMER BEHAVIOR IN A FOOD PRODUCTS MARKET, *Theoretical and Empirical Researches in Urban Management*, Vol. 10, No. 1, pp. 25–42 (online), available from ⟨https://www.proquest.com/scholarly-journals/spatial-analysis-consumer-behavior-food-products/docview/1658831491/se-2⟩ (2015).

[29] Selvin, S., Shaw, G., Schulman, J. and Merrill, D. W.: Spatial Distribution of Disease: Three Case Studies2, *JNCI: Journal of the National Cancer Institute*, Vol. 79, No. 3, pp. 417–423 (online), DOI: 10.1093/jnci/79.3.417 (1987).

[30] Wang, Z. and Zhang, H.: Understanding the spatial distribution of crime in hot crime areas, *Singapore Journal of Tropical Geography*, Vol. 40, No. 3, pp. 496–509 (online), DOI: https://doi.org/10.1111/sjtg.12293 (2019).

[31] 株式会社ジー・サーチ：JST 科学技術用語シソーラスって何？ https://jdream3.com/jd-room/start/20190628/（参照 2021-12-25）.

[32] Zhu, Z., Xu, S., Qu, M. and Tang, J.: GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding, *The World Wide Web Conference*, ACM, pp. 2494–2504 (2019).