人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第13回)
SIG-AM-13-03

語の分散表現と上位下位関係ー研究動向と今後への試案ー

Recent Research Trends and Proposal of Distributional Word Representations for Hypernymy Detection

鷲尾光樹 [1*]
Koki Washio[1]

[1] 東京大学大学院総合文化研究科
[1] Graduate School of Arts and Sciences, The University of Tokyo

Abstract: Distributional representation for words is an important model of word senses which reflects several types of semantic relations, not only similarity relations. In this paper, we provide an overview of unsupervised/supervised approaches with word distributional representations to detect hypernym-hyponym relations and a recently reported problem associated with these approaches. Moreover, we propose a future research direction to solve this problem and prove the validity of this direction with small experiments.

1 はじめに

語の分散表現では、分布意味論に基づいて、語の分布の情報からベクトルを作成し、単語ベクトルとして語の意味を表現する [22][7]。単語ベクトル間の距離などを用いて、語の意味の類義性などが判断できることから、語の分散表現は質問応答システムや情報抽出などの様々なタスクに貢献している。また、語の分散表現はコーパスから自動的に獲得できるため、人手で作成したリソースに存在しない語に対応できることも大きな利点となっている。近年では、語の類義関係に留まらず、アナロジーなど、単語間の様々な意味関係を分散表現を用いて推測するという意味タスクの研究がなされている。そのような研究の一つとして、分散表現を用いた二語の上位下位関係の推測・学習が注目されている。これは、「動物」が「犬」の上位語であり、「犬」が「動物」の下位語であるというような意味関係を、それぞれの語の分散表現から推定するタスクである。

本稿では、分散表現からある二語の上位下位関係を推測する研究と問題点について概観し、今後の研究の方向性を提案するとともに、その提案の妥当性を裏付ける簡単な実験の報告を行う。

2 分散表現

分布意味論においては、語の出現文脈によって、単語の意味を捉えるため、分散表現を獲得する際は、まず単語の分布を見る際の文脈を規定する。文脈に何を用いるかは様々だが、代表的なものとしては、近傍共起や依存構造に基づく関係を利用するものが挙げられる。近傍共起では、対象の語の前後数語を文脈窓として共起を測る。一方、依存構造においては、主語と動詞の関係など、文中において対象の単語が他の語とどのように関わっているかを文脈として用いる。分散表現の性質はこのような文脈の選択により決定される。たとえば、文脈窓を採用した場合は話題的・領域的な類似性がベクトル空間上で表現され、依存構造などの統語的な文脈を用いた場合は、語の品詞などを考慮した機能的な類似性が捉えられるという報告がある [9]。しかし、様々な意味タスクにおける性能の良さや、コーパスを処理するコストの低さから、自然言語処理における分布意味論の分野では文脈窓が用いられることが多い。

以上から、本稿では文脈窓の利用を前提としつつ、本節においては二つの分散表現の獲得法について説明する。一つは共起頻度に基づく古典的な共起頻度ベクトル（カウントベースの分散表現）であり、もう一つは近年台頭してきたニューラルネットワークによる学習から獲得する単語埋め込みベクトル（ニューラルベースの分散表現）である。

なお、本稿では以下の記法を用いる。単語 w の集合を V_W、単語の文脈（文脈窓中に出現する語）c の集合を V_C とし、w の単語ベクトルを $\vec{w}$、文脈 c の単語ベクトルを $\vec{c}$ とする。また、コーパスで観察された w と c の共起 (w, c) の集合を D とし、コーパスで観察された要素の頻度を返す関数を f とする。

*連絡先： 東京大学総合文化研究科言語情報科学専攻
〒 153-8902 東京都目黒区駒場 3-8-1
E-mail: kkwashio3333@gmail.com

2.1 カウントベースの分散表現

カウントベースの分散表現獲得では、文脈に基づいて共起頻度を集計し、$|V_W| \times |V_C|$ の共起頻度行列 M を作成する。単語 w に該当する行を行列 M から切り出すことで、単語ベクトル $\vec{w}$ が得られる。カウントベースの分散表現は、高次元でスパースな表現であり、M のほとんどの要素は 0 である。また、後に述べるニューラルベースの分散表現と異なり、カウントベースの分散表現においては $\vec{w}$ の各次元が単語 w と文脈 c の結びつきの強さを表しており、各次元の持つ意味が明確であることが特徴である。

行列 M の要素として、生の共起頻度をそのまま用いた場合、w や c の頻度のばらつきが単語ベクトルに悪影響を及ぼす可能性がある。このような問題を回避するために、共起頻度行列 M の各要素を以下で定義される PMI(pointwise mutual information, 相互情報量) に変換することがよく行われる。

$$\begin{aligned} PMI(w,c) &= \log_2 \frac{P(w,c)}{P(w)P(c)} \\ &= \log_2 \frac{f(w,c)|D|}{f(w)f(c)} \end{aligned}$$

PMI は w と c が共起する確率からそれぞれの出現確率を差し引いており、正確に二語の結びつきの強さを測ることができる。しかし、w と c の共起が観察されなかった場合、$PMI(w,c) = -\infty$ になるため、実際は以下の PPMI(positive pointwise mutual information, 正の相互情報量) が用いられることが多い。

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

PPMI は生の共起頻度を用いるより、様々な意味タスクにおいて良い成績を残す一方で、低頻度な w や c を含むペアに高い値を返してしまうバイアスがあることが知られている [12]。

また、行列 M のスパースネスを解消するために、特異値分解によって $|V_W| \times |V_C|$ の行列 M を $|V_W| \times d$ (d は数百次元) に圧縮して用いることもある。この場合、文脈 c を表していた単語ベクトルの各次元の意味は失われる。

2.2 ニューラルベースの分散表現

ニューラルベースの分散表現獲得では基本的に、ニューラルネットワークの入力層と出力層に単語や文脈を配置して学習を行い、d 次元の隠れ層を低次元で密な単語の表現として取り出す。カウントベースの分散表現とは対照的に、各次元の意味は不明瞭であり、直観を働かせることはできない。本稿では、文献 [14] と文献 [15] で提案された、ネガティブサンプリングを用いた SkipGram モデル（SGNS）を説明する。SGNS は意味タスクにおいて汎用的に用いられている分散表現獲得法である。

SGNS は、コーパスで観察された共起 (w,c) に対し、d 個のユニットからなる隠れ層を持つニューラルネットワークの入力層に、w に該当する次元のみ 1（他の次元は 0) の $|V_W|$ 次元 one-hot ベクトルを配置し、出力層には c に該当する部分のみ 1 の $|V_C|$ 次元 one-hot ベクトルを配置して学習を行う。その結果において、入力層側にある $|V_W| \times d$ のパラメータ行列の w に該当する行を単語ベクトル $\vec{w}$、出力層側にある $d \times |V_C|$ のパラメータ行列の c に該当する列を文脈ベクトル $\vec{c}$ とみなす[1]。

以下では SGNS の目的関数について説明する。いま、(w,c) が D に含まれる確率を、

$$P(D=1|w,c) = \frac{1}{1+\exp(-\vec{w}\cdot\vec{c})}$$

のようにモデル化すると、目的関数は、

$$\begin{aligned} &\arg\max_{\vec{w},\vec{c}} \prod_{(w,c)\in D} P(D=1|w,c) \\ &\qquad = \arg\max_{\vec{w},\vec{c}} \sum_{(w,c)\in D} \log \frac{1}{1+\exp(-\vec{w}\cdot\vec{c})} \end{aligned}$$

となる。このままだと、すべての (w,c) に対し $P(D=1|w,c)=1$ になってしまうパラメータが存在するため、各 (w,c) に対し、D に存在しない k 個のペア（$(w,c_1),...,(w,c_k)$）を負例として考慮して学習する。c_j は $P(c)$ の 3/4 乗に従ってサンプリングを行う。負例の集合を D' とすると、ネガティブサンプリングを用いた場合の目的関数は、

$$\begin{aligned} &\arg\max_{\vec{w},\vec{c}} \prod_{(w,c)\in D} P(D=1|w,c) \prod_{(w,c)\in D'} P(D=0|w,c) \\ &\quad = \arg\max_{\vec{w},\vec{c}} \sum_{(w,c)\in D} \log \frac{1}{1+\exp(-\vec{w}\cdot\vec{c})} \cdot \\ &\qquad \sum_{(w,c)\sim P(c)^{\frac{3}{4}}\in D'}^{k} \log \frac{1}{1+\exp(\vec{w}\cdot\vec{c})} \end{aligned}$$

となる。この目的関数を確率的勾配降下法などを用いて最適化すると、d 次元の単語埋め込みベクトルが得られることになる。

ニューラルベースで獲得した低次元で密な分散表現はカウントベースの分散表現と同様に、様々な意味タスクに用いることができる。

[1] w の one-hot ベクトルに入力層側のパラメータ行列をかけると、d 次元の単語埋め込みベクトル $\vec{w}$ が得られる。

2.3 カウントベース vs. ニューラルベース

文献 [2] では、様々な類義語の認識やアナロジーの計算などの様々な意味タスクにおいて、カウントベースの分散表現とニューラルベースの分散表現を用いた際の性能が比較されており、基本的にはニューラルベースの分散表現の性能が良いと報告されている。

一方で、ニューラルベースの分散表現獲得法がカウントベースの獲得法よりアルゴリズムとして優れているという主張に疑問を投げかける、Levy らによる一連の報告がある [11][10][13]。

たとえば、文献 [11] では、SGNS などの word2vec 系の分散表現獲得法は、カウントベースの PMI 行列を行列分解していることと等価であると述べられている。この報告によると、SGNS の目的関数を展開し、$\vec{w} \cdot \vec{c}$ で微分することで、

$$\begin{aligned} \vec{w} \cdot \vec{c} &= \log \left(\frac{f(w,c) \cdot |D|}{f(w) \cdot f(c)} \right) - \log k \\ &= PMI(w,c) - \log k \end{aligned}$$

が得られる。このとき、SGNS における入力層側のパラメータ行列を W、出力層側のパラメータ行列を C とすると、

$$W_i \cdot C_j = PMI(w_i, c_j) - \log k$$

となる。この式の右辺はカウントベースの共起頻度行列 M の各要素を PMI 行列に変換し、各要素からネガティブサンプリングに用いた負例ペアの数の対数をとった値を差し引いた行列である。結局、SGNS で行われている最適化は、単語ベクトルと文脈ベクトルの内積が PMI 行列の各要素から $\log k$ を引いた値になるように学習しているということになる。以上のことから、SGNS は特異値分解とは異なるが、同じ行列分解の一種とみなすことができる[2]。

また、ニューラルベースの分散表現が注目された理由のひとつに、アナロジーの計算ができることが挙げられる。アナロジーの計算とは、たとえば、「女性」のベクトルから「男性」のベクトルを引き「王様」のベクトルを足すと、「女王」のベクトルに非常に近いベクトルになるというものであり、ニューラルベースの分散表現は意味関係を簡単な足し引きで表現できると言われていた [14][15][16]。しかし、アナロジーを計算するための式を改良することで、ニューラルベースの分散表現とカウントベースの分散表現の両表現でアナロジータスクの性能が向上し、またそれぞれの表現を用いた際の性能が拮抗するという報告がある [10]。このことから、アナロジーの計算に必要な情報は、ニューラルベースの分散表現のベクトル空間のみにおいて捉えられているわけではなく、カウントベースの分散表現獲得法でも捉えられていることがわかる。

さらに、文献 [13] では、SGNS などのニューラルベースの分散表現獲得法に実装されているネガティブサンプリングやサブサンプリング（高頻度語の文脈からの確率的な排除）などを、分散表現獲得の際のハイパーパラメータとみなし、同等の処理ををカウントベースの分散表現獲得法にも適用することで、様々な意味タスクにおける両分散表現の性能が拮抗するという実験結果が報告されている。この結果を踏まえて Levy らは、ニューラルベースの分散表現の優位性は、アルゴリズムそのものが優れているのではなく、獲得手法にデフォルトで設定されているハイパーパラメータが性能の向上に寄与した結果であり、カウントベースの分散表現獲得法とニューラルベースの分散表現獲得法に本質的な差はないと結論づけている。一方で Levy らは、SGNS は分散表現の獲得が早く、かつ、非常にロバストな表現であり、意味タスクにおいて際立った性能を出すことは少ないが、性能が急激に落ちることもなく、実験を行う際のベースラインとして優れていると報告している。

以上の一連の報告から、カウントベースの分散表現とニューラルベースの分散表現のどちらが優れているかは一概に論じることが難しく、意味タスクごとに最適な分散表現とハイパーパラメータを探索することが重要であることがわかる。

3 上位下位関係の学習

上位下位関係を分散表現から学習する場合、類義関係などの対称的な意味関係と異なり、二つの分散表現に対して非対称な指標や関数を求める必要がある。本節では、2 節の手法を用いて獲得した分散表現を用いて語の上位下位関係を学習するための二つのアプローチ（教師なし学習と教師あり学習）と、文献 [12] で報告された教師あり学習の問題点について概観する。

3.1 教師なし学習

分散表現を用いた上位下位関係推測研究の初期においては、二語の分布の包含性を測る指標が研究されていた。これらの一連の研究は、内省的な分析に基づいた指標の提案を行うものであり、訓練データを利用した機械学習によらないという意味で、上位下位関係の教師なし学習と呼ばれている。このアプローチにおいては、基本的に以下の二つの分布意味論的直観が前提とされている。

[2] word2vec 系でも CBoW モデルや、別のニューラルネットワークモデルでは事情が異なる。一方、同じくニューラルベースの分散表現獲得法である GloVe は、行列分解とみなせるとの報告がある [20]。

- 分布一般性 (Distributional generality)[22]
- 分布包含仮説 (Distributional inclusion hypotheses)[5]

分布一般性とは、意味の広い語はコーパス上に広く分布するという直観である。たとえば、「動物」という単語は「犬」という単語に比べて、「犬」とは共起しにくそうな「泳ぐ」や「飛ぶ」といった単語とも共起するはずである。この直観をもとに、広く分布する語はより上位語らしく、分布が狭い語はより下位語らしいと判断することができる。一方、分布包含仮説とは、ある二語が上位下位関係にあるならば、下位語の出現文脈がある程度上位語の出現文脈に含まれているという直観である。たとえば、「犬」は「走る」や「吠える」などと共起するはずだが、「動物」も同様の文脈に出現するはずである。ここから、分布の包含関係を見ることで、上位下位関係を判断できると期待できる。この二つの前提のもとで、二語の分散表現の各次元の値を見比べることで、上位下位関係性を判断するための指標が提案されてきた。なお、教師なし学習においては各次元の意味が明確である必要があるため、基本的に分散表現はカウントベースのものを用いる。以下では、代表的な指標を紹介する。なお、以降では w_1 と w_2 の二語を扱うとする。$\vec{w_i} = (w_{i1}, ..., w_{in})$ であり、関数 F をベクトルの 0 ではない素性の集合を返す関数とする。

分布意味論的観点から提案された最初の指標は、$Weeds$ という指標である [22]。

$$WeedsP(w_1, w_2) = \frac{\sum_{i \in F(w_1) \cap F(w_2)} w_{1i}}{\sum_{i \in F(w_1)} w_{1i}}$$

$$WeedsR(w_1, w_2) = \frac{\sum_{i \in F(w_1) \cap F(w_2)} w_{2i}}{\sum_{i \in F(w_2)} w_{2i}}$$

これは、情報検索などの評価指標として用いられる precision と recall を二語の関係性を捉える際に適用し、分布の包含性を捉えようとしたものである。いま w_1 が下位語、w_2 が上位語であるとすると、下位語の分布は上位語の分布にある程度包含されるため、$WeedsP$ の値は 1 に近づき、$WeedsR$ の値は 0 と 1 の間に収まるという指標である。判別においては、$WeedsP$ のみ、あるいは $WeedsP - WeedsR$ を用いて、閾値を設定して上位下位関係の有無を判断する。

Weeds から派生した指標として、$Clarke$[3] と $invCL$[8] という指標がある。

$$ClarkeP(w_1, w_2) = \frac{\sum_{i \in F(w_1) \cap F(w_2)} \min(w_{1i}, w_{2i})}{\sum_{i \in F(w_1)} w_{1i}}$$

$$ClarkeR(w_1, w_2) = \frac{\sum_{i \in F(w_1) \cap F(w_2)} \min(w_{1i}, w_{2i})}{\sum_{i \in F(w_2)} w_{2i}}$$

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

$Clarke$ は $Weeds$ に近い指標であり、$ClarkeP$ のみ、あるいは $ClarkeP - WeedsR$ を用いて、閾値を設定する。$invCL$ は $Weeds$ や $Clarke$ と異なり、下位語が上位語にどれだけ包含されているかのみならず、下位語の出現文脈の以外の上位語の分布の広さを考慮した指標になっている。$invCL$ も閾値を設定して、二語の上位下位関係の有無を判断させる。

また、$invCL$ のような指標の派生として、以下の様な指標も考えることができる [21]。

$$simdiff(w_1, w_2) = \frac{\sum_{i \in F(w_1) \cup F(w_2)} \min(w_{1i}, w_{2i})}{\sum_{i \in F(w_1) \cup F(w_2)} \max(w_{1i}, w_{2i})} \cdot \frac{\sum_{i \in F(w_2) - F(w_1)} w_{2i} - \sum_{i \in F(w_1) - F(w_2)} w_{1i}}{\sum_{i \in F(w_1) \cup F(w_2)} \max(w_{1i}, w_{2i})}$$

これは Jaccard 尺度（第 1 項）と分布が重なっていない部分の文脈の広さの差（第 2 項）の積である。つまり、二語がどれくらい似ているかと上位語と下位語にどれくらい分布の広さに差があるかを考慮している。

また、ここでは深く触れないが、w_1 と w_2 の分布の包含を測る際、AP(Average Precision) を応用することで、w_1 が w_2 とより結びつきが強い文脈語と共起する場合により重みを割り振った、$balAPinc$ という指標もある [7]。

教師なし学習の性能を測るために、Erk が行った実験 [4] を参考にして、以下の実験を行った。対象の語を動詞にしぼり、BNC(British National Corpus) コーパス前半の 5000 万語から、動詞の主語と目的語を依存構造文脈として集計し、PPMI 行列を作成した。データセットしては、WordNet3.0 から同義語集合を一つしか持たない単義の動詞と、その直接の上位語のペアを正例とし、負例は下位語と上位語の正例以外の組み合わせにより作成した。条件ごとに、負例から正例と同じ数のペアをランダムにサンプリングし、それぞれの指標を用いて最適な閾値を設定した場合の正解率は以下のようになった。なお、$Weeds$ と $Clarke$ に関しては precision から recall を引いたものをを指標として用いている。

表 1: 教師なし学習の性能

	全負例	負例 1	負例 2	負例 3
$Weeds$	0.74	0.59	0.83	0.72
$Clarke$	0.74	0.61	0.83	0.72
$invCL$	0.71	0.64	0.83	0.65
$simdiff$	0.74	0.64	0.83	0.72

全負例はすべての負例ペア、負例 1 はちぐはぐな上

位下位関係のペア、負例 2 は下位語同士のペア、負例 3 は上位語同士のペアからそれぞれサンプリングした場合の正解率である。ちぐはぐな上位下位関係のペアとは、下位語の集合と上位語の集合から、それぞれ一語ずつ抜き出し、正例にないペアを作ることによって作り出した負例である。用いるデータセットにもよるが、頻度の少ない下位語同士のペアを負例と判断するような簡単な場合（負例 2）を除くと、分布の包含関係に基づく指標は、だいたい 6 割から 7 割 5 分くらいの正解率であることがわかる。

また、分布包含仮説に依拠せず、文脈語の分布に着目した指標として、文献 [19] で提案された $SLQS$ という指標がある。この指標は、上位語はより広い意味を持つ文脈語と共起し、下位語は狭い意味を持つ文脈と共起するという直観に基づいている。たとえば、「動物」という語は「走る」や「飛ぶ」といった一般的な語と共起するが、「犬」は「吠える」などの具体的な語と共起しやすい。このような直観に基づけば、二語がよく共起する文脈語のベクトルを見て、それらのエントロピーを比べることで上位下位関係性が特定できるはずである。$SLQS$ では、対象の語と結びつきが強い上位 N 個の文脈語のエントロピーを計算し、その中央値をとって比較する。いま、文脈語ベクトル $\vec{c} = (c_1, ..., c_n)$ のエントロピーを以下のように定義する。

$$H(c) = -\sum_{i=1}^{n} p(c_i|c) \cdot \log_2(p(c_i|c))$$

$p(c_i|c)$ は文脈語 c の頻度と各共起頻度の割合である。これを MinMax スケーリングにより 0 から 1 の値を取るようにスケーリングし、その値を $H_n(c)$ と定義すると、単語 w_i の文脈語のエントロピーの中央値は、

$$E_{w_i} = Me_{j=1}^{N}(H_n(c_j))$$

と定義される。ただし、Me は中央値を返す関数である。このとき $SLQS$ は、

$$SLQS(w_1, w_2) = 1 - \frac{E_{w_1}}{E_{w_2}}$$

となる。$SLQS(w_1, w_2) > 0$ のとき、w_1 は w_2 の下位語と判断される。

3.2 教師あり学習

分散表現を用いた上位下位関係の教師あり学習では、二つの単語ベクトルに何らかの演算を施して特徴ベクトルとして扱い、二語が上位下位関係を持つか否かの二値分類を学習する。学習アルゴリズムには、SVM を用いるもの [18][23]、ロジスティック回帰を用いるもの [18] などがある。二つのベクトルに施す演算としては、差をとったり、二つのベクトルを結合して特徴量とする方法の性能がいいことが知られている。

教師あり学習の性能を確認するために、以下の上位下位関係性の二値分類実験を行った。

分散表現には、BNC 約 1 億語から、前後二語の文脈窓を採用した SGNS を用いて獲得したものを用いた[3]。データセットには BLESS[1] を用いた。このデータセットは曖昧性のない 200 語の名詞について、上位下位関係や類義関係、部分全体関係、ランダムな関係にある語などを収集して作成されたものである。

BLESS において、上位下位関係にある 1337 ペアから、分散表現が獲得できた 1155 ペアを正例として扱い、同じく分散表現が獲得できた上位下位関係以外の名詞ペアを負例とした。正例ペアと同じ数だけ負例ペアをサンプリングし、SGNS によって獲得した 500 次元の単語ベクトルの差を特徴量として、10 分割交差検定のロジスティック回帰を行う。これを 50 回繰り返したとき、平均スコアとして 0.93 の分類正解率、0.93 の F 値、また、ペアに割り当てた確率をもとにデータを並べ替え AP(Average Precision) を計算したところ、0.97 をマークした。文脈や分散表現の性質、用いているデータセットが異なるものの、教師あり学習は 3.1 節で紹介した教師なし学習の手法よりも良い成績を残している。

3.3 教師あり学習の問題点

教師あり学習は高い分類精度を誇る一方で、上位語に位置しやすい単語を覚えているだけであるという問題点が指摘されている。文献 [12] では、このような教師あり学習の振る舞いを調べるために、様々なデータセットを用いて二つの実験が行われている。

ひとつは、訓練データとテストデータの単語ペアの語彙の重なりをなくして分類を行う実験である。これによって、教師あり学習の分類性能が大きくに下がり、データセットによっては教師なし学習を下回るの性能になってしまうことが観測されている。また、二つのベクトルの差や結合を用いた場合と、上位語のみを用いて学習した場合の性能の比較を行うと、性能差がごく小さいことがわかった。この実験によって、教師あり学習は下位語の情報をほぼ無視していることが明らかになった。

もうひとつの実験は、様々な条件で学習を行った分類器に、ちぐはぐな上位下位関係のペアを判断させる実験である。ちぐはぐな上位下位関係を誤って正例として分類してしまう割合を match error として算出し、様々なモデルで recall との相関を調べたところ非常に強い相関があり、$matcherror = 0.935 \cdot recall$ と線形回帰できてしまうことが明らかになった。つまり、正

[3] 分散表現の獲得には、Omer Levy が公開している hyperwords を用いた。https://bitbucket.org/omerlevy/hyperwords

しく上位下位関係を持つペアを正例として多く分類できる分類器は、その分、ちぐはぐな上位下位関係を持つペアも誤って正例として分類してしまうことになる。

これらの実験から文献 [12] は、教師あり学習では典型的な上位語を覚えているだけであり、二語の関係性は学習できていないと結論づけている。

この結論を追証するために、3.2 節で行った実験と、コーパスや分散表現は同じ条件で、訓練データとテストデータの語彙の重なりをなくした場合の結果を以下に示す。

表 2: 教師あり学習における語彙の重なりの影響

	重なりあり	重なりなし	Δ
分類正解率	0.93	0.68	0.25
F 値	0.93	0.61	0.32
AP	0.97	0.77	0.20

（Δ は「重なりあり」と「重なりなし」の差を表す）

表 5 を見ると重なりがある場合と重なりがない場合に性能差があることがわかる。教師あり学習が二語の関係性を学習できていないとすると、そのようなモデルは訓練データにない語彙に対応できない。これは、コーパスからの自動的学習のそもそものメリットである、「人手によるリソースに存在しない語彙に対応できる」という点においては致命的である。

4 考察と今後の方向性

3.3 節で述べたような教師あり学習の典型的な上位語の記憶という問題への対処としては、二つの方向性が考えられる。ひとつは分散表現獲得法の最適化であり、もうひとつは学習法の見直しである。

4.1 分散表現の最適化

分散表現は、語の文脈によって語の意味を表現するという分布意味論に基づいて獲得されるが、そのモデル自体が特定の意味タスクを志向しているわけではない。よって、ある意味タスクの性能の向上のためには、それぞれの意味タスクが着目している語の意味の側面を正しく反映するようなハイパーパラメータの調整や、モデルそのものの変更が必要になる。たとえば、上位下位関係の認識においては話題的・領域的な類似性だけでなく、語の機能的な類似性も見る必要があるはずである。現在は他の様々な意味タスクでの性能の良さや処理コストの低さから文脈窓を用いられることが多いが、上位下位関係認識においては、文脈に依存構造を採用することによって、ちぐはぐな上位下位関係を正例と判断してしまう割合が減少することが考えられる[4]

4.2 学習法の見直し

上位下位関係認識の教師あり学習においては、特徴量として二つのベクトルの差や結合を用いるだけでは、二語の関係性を学習できていないことがわかっている [12]。これに対しては、上位下位関係に関して分布的な意味付けがより明確な特徴を用いることによって、二語の関係性の学習を促進できる可能性がある。そのような特徴として、今まで研究されてきた教師なし学習の指標を挙げることができる。これらの指標を特徴として採用することで、二語の分布の包含関係や形状の違いなどの関係性が学習されるはずである。

4.3 実験

これらのアプローチの妥当性を検討するために、3.1 節で述べたような、教師なし学習として提案された指標群を、教師あり学習における特徴量として採用した場合の性能評価を行った。

分散表現として BNC 約 1 億語から前後二語の文脈窓を共起として獲得した PPMI 行列を採用し、データセットは 3.2 節と同じく BLESS における名詞ペアを用いた。

特徴量としては以下のものを採用した。

- 二語の類似度： cos 類似度
- 分布の包含関係: $WeedsP/R, ClarkeP/R, invCL$
- 分布の形状： 二語の分布のエントロピーの差と比、それぞれのベクトルにおいて値が高い上位 50 次元の値の平均の差
- 文脈の分布: $SLQS$

まず、二語の類似度を見るために cos 類似度を特徴として考慮し、分布の包含関係を見るために、$WeedsP/R$、$ClarkeP/R$、$invCL$ を採用した。また、分布の形状の違いを見るために二語の分布のエントロピーの差と比、それぞれの単語ベクトルにおいて値が高い上位 50 次元の値の平均の差を採用した。これは上位語は分布がなだらかで広く、下位語は狭い文脈にしか出現しないという直観に基づく。さらに結びつきが強い文脈語の分布を見るために $SLQS$ を特徴として考慮する。これによって、二語の分散表現から 10 次元の特徴ベクトルを算出することができる。

[4]ただし、大規模コーパスに対し依存構造解析を行うコストや、精度の良いパーサーが存在する言語の少なさを考えると、文脈窓の採用には妥当性がある。

表 3: 提案手法における語彙の重なりの影響

	重なりあり	重なりなし	Δ
分類正解率	0.61	0.55	0.06
F 値	0.59	0.50	0.09
AP	0.64	0.59	0.05

表 4: cos 類似度のみ (baseline)

	重なりあり	重なりなし	Δ
分類正解率	0.54	0.49	0.05
F 値	0.51	0.41	0.10
AP	0.56	0.56	0.0

この特徴ベクトルをもとに、上位下位関係のペアを正例として、正例と負例の数を揃えてロジスティック回帰を行い、10 分割交差検定を行った場合の性能と、3.3 節で行った実験と同様に、訓練データとテストデータの語彙の重なりをなくした場合の性能を比較すると、表 3, 表 4 のようになった。

なお、baseline は cos 類似度のみを用いた場合とした。提案手法はいずれの場合でもベースラインを上回りつつ、SGNS よりも語彙の重なりの影響が少ないことがわかる。この結果から、二語の関係性としての意味付けが明確なものを特徴量に用いると、性能は低いものの、語彙の重なりの影響が少ないことがわかる。

さらに、SGNS で獲得した二語の単語ベクトルの差に、今回算出した特徴ベクトルを結合して学習した場合の性能（SGNS+提案手法）と、単語ベクトルの差のみを特徴ベクトルとして学習した場合の性能（SGNS のみ）を比較した。訓練データとテストデータの語彙の重なりをなくした際の結果は以下のようになった。

表 5: 提案手法を付加した場合の性能

	SGNS のみ	SGNS+提案手法
分類正解率	0.68	**0.73**
F 値	0.61	**0.68**
AP	0.77	**0.83**

提案手法の特徴ベクトルを従来の教師あり学習の特徴ベクトルに結合した場合、正解率、F 値、AP において性能が向上している。この結果から、従来の特徴ベクトルに教師なし学習の指標を付加することで、二語の関係性の学習が促進されていることがわかる。これによって、教師あり学習の際に、上位下位関係に関して意味付けが明確な分布的特徴を用いることで、二語の関係性の学習が可能であることが示された。

5 おわりに

本稿では、分散表現を用いてある語のペアが上位下位関係を持つか否かを判別する研究とその問題点を概観した。今後の研究の方針として、二語の分散表現の差をや結合を用いて教師あり学習を行った場合、二語の関係性を学習できないが、分散表現から分布的特徴を適切に抽出すれば、二語の関係性の学習を促進できることを示した。これからは、上位下位関係の認識に最適な分散表現の獲得法と、影響力のある分布的特徴を模索していきたい。

謝辞

本研究に関して、終始あたたかくご指導ご鞭撻をして頂き、また本稿の執筆の機会を与えてくださった、東京大学大学院総合文化研究科加藤恒昭教授に心より感謝いたします。

参考文献

[1] Baroni, M., Lenci, A.: How we BLESSed distributional semantic evaluation., In *Proc. of the ofthe GEMS 2011Workshop on GEometrical Models of Natural Language Semantics*, pp. 1–10 (2011)

[2] Baroni, M., Dinu, G., Kruszewski, G.: Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors., In *Proc. of the Annual Meeting of the Association for Computational Linguistics(ACL)*, Vol. 2, long paper, pp. 238–247 (2014)

[3] Clarke, D.: Context-theoretic semantics for natural language: an overview., In *Proc. of the EACL 2009Workshop on GEMS: GEometrical Models of Natural Language Semantics*, pp. 112–119 (2009)

[4] Supporting inferences in semantic space: representing words as regions., In *Proc. of the International Conference on Computational Semantics(ICCS)*, pp. 104–115 (2009)

[5] Geffet, M., Dagan, I.: The distributional inclusion hypotheses and lexical entailment., In *Proc. of the Annual Meeting of the Association for Computational Linguistics(ACL)*, pp. 107–114 (2005)

[6] Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al. ' s negativesampling., *arXiv preprint*, arXiv:1402.3722. (2014)

[7] Kotlerman, L., Dagan, I., Szpektor, I., Geffet, M., Directional Distributional Similarity for Lexical Inference., Natural Language Engineering, Vol. 16(4), pp. 359–389 (2010)

[8] Lenci, A., Benotto, G.: Identifying hypernyms in distributional semantic space., In *SEM 2012 The First Joint Conference on Lexical and Computational Semantics*, Vol. 2, pp. 75–79 (2012)

[9] Levy, O., Goldberg, Y.: Dependencybased word embeddings., In *Proc. of the Annual Meeting of the Association for Computational Linguistics(ACL)*, Vol. 2, Short Paper (2014)

[10] Levy, O., Goldberg, Y.: Linguistic regularities in sparse and explicit word representations., In *Proc. of the Conference on Computational Natural Language Learning.*, pp. 171-180 (2014)

[11] Levy, O., Goldberg, Y.: Neural word embeddings as implicit matrix factorization., In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pp. 2177-2185 (2014)

[12] Levy, O., Remus, S., Biemann, C., Dagan, I.: Do Supervised Distributional Methods Really Learn Lexical Inference Relations?, In *Proc. of the 2015 North American Chapter of the Association for Computational Linguistics(NAACL): Human Language Technologies*, pp. 970–976 (2015)

[13] Levy, O., Goldberg, Y., Dagan, I. Ramat-Gan, I.: Improving distributional similarity with lessons learned from word embeddings., *Transactions of the Association for Computational Linguistics*, 3 (2015)

[14] Mikolov, T., Chen, K., Corrado, G. S., Dean, J.: Efficient estimation of word representations in vector space., *CoRR*, abs/1301.3781. (2013)

[15] Mikolov, T., Sutskever, I., Chen, K. Corrado, G. S., Dean, J.: Distributed representations of words and phrases and their compositionality., In *Advances in neural Information Processing Systems*, pp. 3111–3119 (2013)

[16] Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations., In *Proc. the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751 (2013)

[17] Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation., In *Proc. of the Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2, Short Paper (2015)

[18] Roller, S., Erk, K., Boleda, G.: Inclusive yet selective: Supervised distriubtional hypernymy detection., In *Proc. of the International Conference on Computational Linguistics(COLING)*, pp. 1025–1036 (2014)

[19] Santus, E., Lenci, A., Lu, Q., Walde, S.: Chasing hypernyms in vector spaces with entropy., In *Proc. ofn the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 38–42 (2014)

[20] Suzuki, J., Nagata, M.: A Unified Learning Framework of Skip-Grams and Global Vectors., In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, Vol. 2, Short Paper (2015)

[21] 鷲尾光樹 , ベクトル空間表現における動詞の意味関係の現れ, 東京大学教養学部教養学科超域文化科学分科 卒業論文 (2015)

[22] Weeds, J., Weir, D., McCarthy, D.: Characterising measures of lexical distributional similarity., In *Proc. of the International Conference on Computational Linguistics(COLING)*, pp. 1015–1021(2004)

[23] Weeds, J., Clarke, D., Reffin, J., Weir, D., Keller, Bill.: Learning to distinguish hypernyms and co-hyponyms., In *Proc. of the International Conference on Computational Linguistics(COLING): Technical Papers*, pp. 2249–2259 (2014)