ベクトル空間モデルによる人の四項類推の最適近似
Optimal Estimation of Analogy using Vector-Space Models

加藤 龍彦 †，日高 昇平 †
Tatsuhiko Kato, Shohei Hidaka
† 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
skylark@jaist.ac.jp

概要

人の記憶は物事の間の関係性，つまり意味を基盤として構造化されている．本研究ではこうした意味構造のモデルとしてベクトル空間モデル，特に Skip-gram に着目して分析，その四項類推が一部単語クラスについて最適化されていないことを示す．その上で，四項類推になりたつ関係を用いて，モデルの類推能力を人により近似するような演算を提案，それによって四項類推の性能を大きく向上可能であることを示す．このことは，四項類推関係を一般化することで単語空間の関係性をよりよく捉えることができることを示唆する．

キーワード：意味表現, 類推, ベクトル空間モデル

1. はじめに

象は動物の一種である，リスは木の実を食べる，といったように，環境中の物事は相互に関係して存在する．限られた資源の下で効率的な情報処理を行う必要がある脳やコンピュータにとって，これらの関係をどう利用して情報を処理するかは重要な問題である．このような意味表現の問題は，記憶や学習といった認知の重要な課題と密接に関連し，認知科学の根本的な問いとされる [3].

意味表現をモデル化する試みとして，これまでに意味ネットワークモデル [2]，ベクトル空間（または意味空間）モデル [4]，トピックモデル [3] といったモデルが提案されてきた．各モデルはそれぞれ，どのように単語や概念，またそれらの間の意味を表現するか，という点で異なる（表 1）．中でも近年，機械学習分野においては単語表現のモデルとしてベクトル空間モデルが注目されてきた [11, 7, 6, 8]．例えば word2vec (Skip-gram モデル) [7, 6] は，簡単な四項類推課題においてではあるが，人と近似した回答を高精度で与えることが示されている．認知科学においても，潜在意味解析 (LSA)[4] をはじめ，多くの言語モデルがベクトル空間モデルの発想に則って提案されてきた．本稿では，ベクトル空間モデルの中でも特に Skip-gram に焦点を当てて分析を行う．その理由としては，本稿執筆現在最高精度で四項類推課題に回答可能なモデルであること [5]，またモデルの学習方法が共起関係を直接圧縮するような単純で本質的ものであること，の二点がある．

	意味ネットワークモデル	ベクトル空間モデル	トピックモデル
概念	グラフのノード	空間上の点	単語上の確率分布
意味	グラフのエッジ	点間の距離	トピック上の確率分布

表 1　意味表現の代表的なモデルクラス．1 行目はモデルにおいて概念がどのように表現されているか，2 行目は同様に意味がどのように表現されているかを示す．

このようなモデルの基本的な仮定は分布仮説である [10]．それによれば，コーパス中の単語間のある文脈における共起関係を知ることで，その意味的関係についても多くを学ぶことができるという．例えば「王」,「女王」,「男」,「女」という 4 つの単語を考えてみよう.「王」や「女王」は「戴冠」などの王位に関連する語と多く共起し,「男」や「王」は「髭」などの男性に関する語と,「女」や「女王」は「化粧」など女性に関する語とそれぞれ多く共起する．そのため，共起関係は「王」と「女王」が同じ「王族」という単語クラスに属する，といった意味的な関係を捉えていると考えられる．

本研究では，ベクトル空間モデルが人の意味表現のモデルとしてどこまで妥当かを検討するため，モデルの代表例として Skip-gram を取り上げ分析，四項類推関係を一般化することでより人の類推を近似するような演算を定式化する．

2. 四項類推課題と Skip-gram モデル

認知科学において四項類推は,「関係性の間の関係性」という高次の理解を必要とするため，人の推論

能力の重要な特質とされてきた [12]．四項類推課題とは，(a : b) と (c : d) の対からなる 4 つの単語や物の内 a,b,c のみが与えられたとき，(a : b) の関係と (c : d) の関係から d を推定する，という課題である [12, p.147]．この課題に対して想定される回答を得るためには，(a : b)，(c : d) それぞれの関係を発見し，そこから D を推定する必要がある．例えば,「大阪府」と「大阪市」に対して「石川県」に対する d は何か，という問いには県-県庁所在地という関係を満たす「金沢市」が想定される回答となる．

近年自然言語処理分野においては，四項類推が単語埋め込みモデルの意味の表現能力の評価課題として採用されている．Mikolov ら [7, 6] は，単語ベクトルを用いて四項類推を行うための演算として以下を提示した:

$$f(w_a, w_b, w_c) = \arg\max_{w_d}(sim(w_c - w_a + w_b, w_d)) \quad (1)$$

ここで w_i はある語 i の単語ベクトル表現で，w_d が推定結果の単語である．また，関数 $sim(v, w)$ は特定の類似性の指標で，典型的にはベクトル v, w のコサイン類似度 $\sim (v, w) = \frac{v^\top w}{\|v\|\|w\|}$ を用いる. 例えば，(man, woman), (king, ?) が与えられた場合には，$w_{king} - w_{man} + w_{woman} \approx w_{queen}$ という結果が正答となる．

3. Skip-gram と人の類推

Skip-gram は現行のモデル中で最高精度の四項類推課題正答率を持つと述べたが，モデルの正答率の評価に使用されるスタンダードな課題セット (以下 Google テストセット) は，人の基準からすると単純なものである．このテストセットは 14 のクラスに分けられた 19544 組の類推課題からなり，例えば，よく知られた (man : woman :: king : queen) や (Germany : Berlin :: France : Paris) といった意味的課題と (bad : worse :: big : bigger) のような文法的課題を含む（表 2）．そのため人であれば，必要な単語さえ記憶していれば関係性を見抜くことは容易で，100 %に近い精度で回答できると思われる．しかし (1) による演算では，いくつかのクラスについてはモデルの正答率は 30-40%程度と低い．

一方，共起関係を用いた単語埋め込みと人の類推能力とが強く関連する，ということも調べられている．演算 (1) の結果が「正しい」回答になるためには，ベクトル空間上の単語表現の対の差が近似的に平行四辺形になっている必要があるが，Rumelhart ら [9] は 1973 年の時点で既にベクトル空間上でアナロジーに必要な演算（演算 (1) と同じもの) と，この演算が人の類推と合致する必要条件として空間上での平行四辺形関係を定式化し，人を用いた実験を行って理論の妥当性を検証している．また，人による単語の自由想起課題の評価スコアとコーパスの共起頻度によって作成した単語表現の間には強い正の相関関係が存在することも確かめられており [5]，ベクトル空間モデルによる人の類推能力の表現には一定の妥当性があると考えられる．

テストセット中でのクラス名	例題
capital-common-countries	Berlin : Germany :: Paris : France
capital-world	London : England :: Rome : Italy
currency	Japan : yen :: USA : dollar
city-in-state	Boston : Massachusetts :: Honolulu : Hawaii
family	man : woman :: king : queen
gram1-adjective-to-adverb	amazing : amazingly :: calm : calmly
gram2-opposite	acceptable : unacceptable :: aware : unaware
gram3-comparative	bad : worse :: big : bigger
gram4-superlative	bad : worst :: big : biggest
gram5-present-participle	code : coding :: dance : dancing
gram6-nationality-adjective	France : French :: Germany : German
gram7-past-tense	dancing : danced :: decreasing : decreased
gram8-plural	banana : bananas :: bird : birds
gram9-plural-verbs	decrease : decreases :: describe : describes

表 2　モデルの類推性能のテストで標準的に使用される Google テストセットの各課題クラスとその例題．:: で区切られた左右二組が関係性を持つ．

それにも関わらず先行研究のモデルが人の類推をうまく近似できない理由の 1 つとして，モデルが単語の共起関係の情報圧縮に主眼を置いており，類推課題への高い正答率はその副産物として表れる，ということが挙げられる．モデル上の類推である演算 (1) は必ずしも人の類推の近似を目的としていない．本研究では，演算 (1) を四項類推関係が成立するために適切な部分空間の選択も含めて一般化したベクトル演算を定式化し，この演算がよりよく人の類推を近似できるかを検討する．

3.1 Skip-gram のベクトル演算は最適か

我々は定式化の前段階として，Skip-gram [7, 6] に関して探索的な分析を行った [13]．その結果，以下の示唆を得た．（1）Skip-gram の単語表現では，語の関係を上記のベクトル演算で捉えられない単語クラスがある（関連研究 [1]）.（2）四項類推に本質的な特徴を捉えるには，単語の共起確率で十分である．（1）を端的に示す結果が，すでに参照した Google テストセット

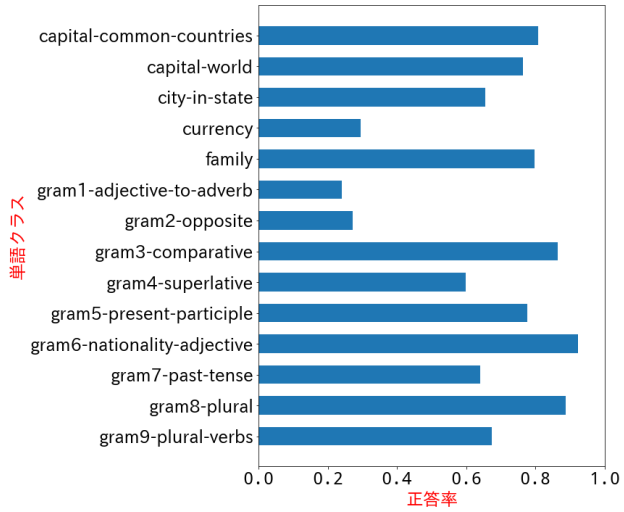

図 1　Google テストセットのクラス毎正答率．縦軸はデータセット中でのクラス名，横軸は各クラスの全データ中の正解の割合を示す．

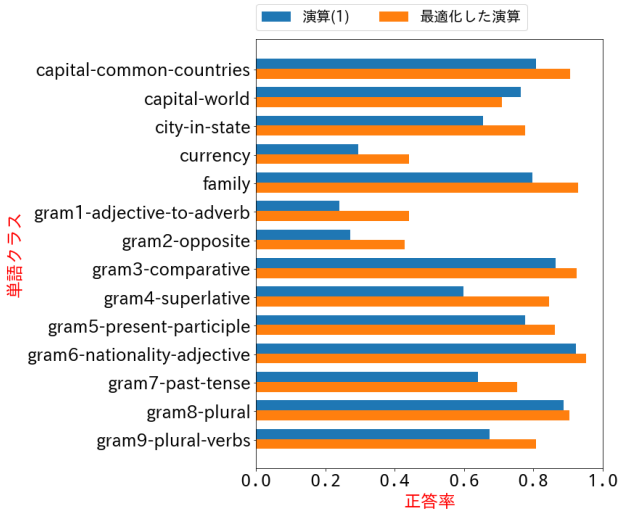

図 2　Google テストセットのクラス毎正答率に関して，演算（1）と提案手法を比較した．縦軸はデータセット中でのクラス名，横軸は各クラスの全データ中の正解の割合を示す．

[6][1]に対する単語クラス毎の正答率である（図 1）．この図では，対義語（gram2-opposite），形容詞－副詞（gram1-adjective-adverb），通貨（currency）の 3 クラスだけ正答率が 30%程度と低くなっていた．この結果は，Skip-gram のベクトル演算が，ある単語クラスの四項関係を適切に表現していないことを示唆する．

4.　定式化

前節の Skip-gram の検討において示された，類推精度の低い種類の単語群にも有効なベクトル演算を調べるため，本研究では以下のように定式化を行う．4 単語 w_i, w_j, w_k, w_l の四項類推に相当するベクトル演算を以下のように定義する：

$$g_{M_1,M_2}(w_a, w_b, w_c) = \arg\max_{w_d}(sim(M_1 w_c - M_1 w_a + M_2 w_b, M_2 w_d))$$

ここで M_1, M_2 は回転や平行移動などのある線形変換である．この変換の対 (M_1, M_2) で定義されるベクトル演算 g_{M_1,M_2} により先行研究のモデルを拡張する．この定式化で，$M_1 = M_2 = I$ （ここで I は単位行列）とすると，先行研究 (式 (1)) と一致する $(f = g_{I,I})$．このような演算の一般化により，人の四項類推を表現するのに平行四辺形関係より優れた演算を模索する．

4.1　最適ベクトル演算の推定

演算の一般化は，Skip-gram の単語の持つ次元の分布の性質を考慮し，類推によって発見したい単語の分布とそれ以外の単語の分布を切り分けるような次元を重み付けるよう行う．

Mikolov らが Skip-gram を用いて訓練した単語表現[2]は各 300 次元を持つ約 300 万の単語によって構成される．これら約 300 万単語の各次元は，近似的に指数分布

$$P(V|\Lambda) \propto \lambda e^{-\lambda|V|}$$

に従う．ここで V は単語表現中の全単語からなる行列である．Mikolov らの演算による四項類推においては，四単語 w_a, w_b, w_c, w_d が与えられた時これら約 300 万単語の中から $w_a - w_b + w_c$ にコサイン類似度で最も近い単語として w_d が選択される必要がある．そのため，$w_1 = w_c - v_a + v_b$ とすると，このような類推において誤った回答 (w_d 以外の単語を選択する確率) を得る確率は以下のように定式化できる．

$$P(error) = P(\|V_i - w_d\| < \|w_1 - w_d\|)$$

式中の $\|V_i - w_1\| < \|w_1 - w_d\|$ は，全単語の内 w_1 との距離が w_d よりも近い単語が存在することを表す．このエラー確率を最小化するような重みを用いて単語の次元を選択することで，演算 (1) を次元の選択も含めて最適化することができる．積分に関する計算から

[1] https://github.com/tmikolov/word2vec/blob/master/questions-words.txt

[2] "GoogleNews-vectors-negative300.bin.gz", https://code.google.com/archive/p/word2vec/

この定式化の厳密解を得ることは難しいため，本研究ではその 1 つの近似として，$\epsilon = \|V_{a,i} - V_{b,i}\|$ として以下の重みを用いる．

$$w_i = e^{\max(V_{a,i}) - \max(V_{b,i}) + \epsilon} \quad (2)$$

以下の実験は約 300 万の単語ベクトル表現の各 300 次元をこの w_i によって重み付けした演算 (1) によって行った．これは，四項類推に最適な部分空間を用いることで演算 (1) を拡張することを意味する．

5. 結果

提案手法で Google テストセットの四項類推の正答率を計算した結果を図 2 に示した．多くのクラスにおいて 10-20 %程度正答率が向上していることがわかる．また以下の表は，2 つの演算のテストセット全体での正答率である．全体としても，正答率が 7 %程度向上していることがわかる．

演算（1）	提案手法
69%	76%

今回の結果とは訓練コーパスなどの実験環境が異なるため厳密な比較はできないが，現在自然言語処理分野において最高精度を持つとされる演算 (3CosMul) は，(1) と比較して 2〜3 %程度の向上しか見せておらず，その性能も 71%に留まる [5]．そのため，提案手法はモデルの類推性能を大きく向上させたと言える．これは，四項単語の対応付けを最大化しつつエラーとなる確率を最小化するような次元によって重み付けを行う，という本研究のアプローチの有効性を示している．

6. 議論・考察

上の結果から，我々が提案した部分空間の選択についても考慮したベクトル演算によって，四項類推の性能が他の関連する先行研究の手法と比べても一部大きく向上することが分かった．本研究の限界として，上述のように本研究の実験で用いた重みが近似的な計算に基づくものであることが挙げられる．そのため，より厳密な計算に基づく重みを使用することで，四項類推関係を捉えるような演算を推定可能であると考えられる．

本研究の結果は，類推関係を捉えるための演算・規則が，従来想定されてきたよりも多様であり得ることを意味する．言い換えれば，こうした多様な四項関係を積極的に利用する事で，単語の共起関係に潜在する意味構造を探索できる可能性を示唆する．本研究では Google テストセットを用い，人が事前に選択した四項類推課題を最適に解く方法を模索したが，今後の研究としては四項類推関係を用いて，言語中にどのような単語間の関係を発見可能かを検討予定である．

文献

[1] Dawn Chen, Joshua C Peterson, and Thomas L Griffiths. Evaluating vector-space models of analogy. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, pp. 1–6, 2017.

[2] Allan M Colins and Elizabeth F Loftus. A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, Vol. 82, No. 6, pp. 407–428, February 1975.

[3] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological Review*, Vol. 114, No. 2, pp. 211–244, 2007.

[4] Thomas K Landauer and Susan T Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, Vol. 104, No. 2, pp. 211–240, 1997.

[5] Omer Levy and Yoav Goldberg. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 211–225, March 2015.

[6] T M Mikolov, K Chen, G Corrado, and J Dean. Efficient Estimation of Word Representations in Vector Space . *ICLR Workshop Papers*, pp. 1–12, 2013.

[7] T M Mikolov, I Sutskever, K Chen, G Corrado, and J Dean. Distributed Representations of Words and Phrases and their Compositionality. *NIPS*, pp. 1–9, 2013.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[9] David E Rumelhart and A Abrahamsen. A Model for Analogical Reasoning. *Cognitive Psychology*, No. 5, pp. 1–28, 1973.

[10] Magnus Sahlgren. The distributional hypothesis. *Rivista di Linguistica*, Vol. 20, No. 1, pp. 33–35, 2008.

[11] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, Vol. 37, pp. 141–188, 2010.

[12] ホリオーク, キース J, サガード, ポール著, 鈴木宏明, 河原哲雄訳. アナロジーの力－認知科学の新しい探求. 新曜社, 1998.

[13] 加藤龍彦, 日高昇平. 意味表現の解明に向けて: ベクトル空間モデルのアナロジー推論の分析. 第 8 回知識共創フォーラム, S4, 2018.