単語共起行列の内部構造解明のための構成論的アプローチ
Constructive approach to study the internal structure of word co-occurrence matrix

前田　晃弘 †，鳥居　拓馬 †, 日髙　昇平 †
Akihiro Maeda, Takuma Torii, Shohei Hidaka
† 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
akihiro.maeda@jaist.ac.jp

概要

単語の意味を分散的に表現する単語ベクトルが四項類推課題を解くことはよく知られているが，そのメカニズムについては必ずしも明らかではない．本研究は，人工的なトイコーパスを用いて，文型や意味関係による言語構造上の制約に加え，文の出現頻度が四項類推課題を解くための平行四辺形の出現条件に関わっていることを示す．また，実コーパスを用いてその共起頻度行列の部分空間に平行関係が埋め込まれることを検証し，単語共起自体に内在する自然言語の構造を捉えることを試みる．

キーワード：単語ベクトル、四項類推課題、単語共起行列、人工コーパス、アフィン空間

1.　単語ベクトルモデルによる四項類推

単語の意味を数百次元のベクトル空間上に分散的に表現する単語埋め込み（以下、単語ベクトルという）は、近年の自然言語処理の要となっている [1]．BERT など最近の大規模言語モデル [2] においても、そのベースにあるのはあらかじめ別に学習された単語ベクトルであり，それらが大規模言語モデルへのインプットとして用いられている．また，認知科学分野においても脳の言語認知における分散表現の存在可能性を示唆するものとして単語ベクトルと脳波の対応関係が指摘されている [3].

代表的な単語ベクトルである word2vec[4] は，単語ベクトル間に四項類推演算（例えば $queen = king - man + woman$ や $works = speaks - speak + work$）が成り立ち，gender や royalty に関する意味的対称関係や動詞の活用のような統語的関係にある単語群がベクトル空間上で平行的な配置関係（線形構造）を持つことを示した．実際に，word2vec の単語ベクトル群を用いて，平行四辺形をなす 4 単語のうち 3 つの単語を与えると，残りの 1 単語を高い精度で予測することはよく知られている．

2.　類推課題を解く単語ベクトルの構造

なぜ word2vec は四項類推課題を解けるのであろうか．単語間の意味的・統語的関係が単語ベクトル空間内の平行関係として出現する現象に動機づけられて，いくつかの研究が単語ベクトルモデルの解明を試みている．例えば Levy & Goldberg [5] は、word2vec の学習が，単語の共起頻度行列の各要素に対して自己相互情報量 PMI[6] を求める計算に相当することを主張している．PMI は自然言語処理でその有効性がよく知られた前処理であるが，より正確には word2vec を生成するために用いられる Skip-gram negative sampling model[4] における最適化の結果，学習されるターゲット単語 w のベクトル $\mathbf{w}$ とコンテクスト単語 c のベクトル $\mathbf{c}$ の内積が，その PMI から定数 k の対数を減じたものに等しくなるとしている（次式）.

$$\mathbf{w} \cdot \mathbf{c} = PMI(w, c) - \log k = \log \frac{P(w, c)}{P(w)P(c)} - \log k$$

k は negative sampling の個数を表すパラメータである．また，$P(w, c)$ はターゲット単語とコンテクスト単語の同時確率を表し，$P(w)$ と $P(c)$ は，それぞれターゲット単語，コンテクスト単語の unigram 確率を表す．式の形から明らかなように，PMI は，ターゲット単語とコンテクスト単語の同時確率が，相互に独立である場合の同時確率（すなわち unigram 確率の積）と異なるのか否かを表す尺度となっている．これより PMI が共起頻度行列に内在する確率分布構造を抽出して，コーパスにあるなんらかの規則性を捉えているのは明らかである．しかしながら，共起分布がどのように四項類推演算を可能としているのか，単語ベクトル間の平行関係をもたらす単語間の共起構造が何に由来するのかは必ずしも明らかであるとはいえない.

これに対して，加藤ら [7] は，PMI のような特定の前処理によらずとも共起頻度行列自体が四項類推課題に高精度で正答するための必要な情報をすでに持つこ

とを主張して，共起頻度行列に対して PMI を算出するのではなく，対数化した共起頻度行列を用いるのみで，PMI に匹敵する類推精度を得られることを実証的に示した．さらに，特異値分解 (SVD) を行った場合には，対数共起頻度行列は PMI 行列よりも高い類推精度を上げたことを示している．

3. 単語共起行列の線形構造を再現するための構成論的アプローチ

これらの研究が示唆するのは，word2vec，PMI，対数化，さらには SVD の手法のいずれもが類推課題を一定程度解くことができ，精度の優劣はあるにせよ，単語共起頻度行列に内在するなんらかの分布構造を抽出していることである．一方，これらの研究で答えられていない問いとは，四項類推を可能とする分布構造とはどのようなものか，どのような分布構造が単語ベクトル間の平行関係，すなわち，単語間の意味的・統語的関係を表現しているのかである．さらに言えば，類推関係に現れる平行関係以外にも単語間の意味的・統語的関係が存在することは十分に考えられる．例えば，上位語と下位語の単語ベクトルはどのような幾何的関係として表されるのだろうか．単語ベクトルが四項類推課題に正答を与える理論的根拠については必ずしも明らかではない．

そこで本研究では，単語共起行列に内在する線形構造の由来と性質を明らかにすることを目的として，人工的なコーパスを用いて単語ベクトル間の平行関係を再現する構成論的アプローチをとる．特に，単語間の平行関係のような規則性が，言語の生成プロセスに由来する統語構造や単語間の意味的制約関係から生じているのではないかとの初期的予想のもとで，コーパス中の文構造と共起行列ならびに単語ベクトル空間の関係性を調べる．また，このアプローチで得られた仮説を実証するために，実コーパスから得られた共起行列（対数頻度による共起行列）を人工コーパスによる結果と比較する．

4. 人工コーパスの作成方法

具体的な単語共起のモデルとして，3 組の平行関係を持つ六面体（平行六面体）を再現することを想定して，24 文から構成される小規模の人工コーパス（以下，トイコーパスという．）を設計する．Mikolov et al.[4] をはじめとする従来研究では，2 組の単語ペア間における類推関係が単語ベクトル空間における 4 つの単語ベクトルがなす平行四辺形関係（2 組の平行関係）として表現されるとの事実が示されてきているが，本研究では文構造との関連性を調べるために，3 組の平行関係（平行六面体関係）へ拡張したモデルを検討する．

トイコーパスは，24 の英文，18 語の語彙から構成されている (図 1)．各文は，主語-動詞-目的語-副詞の

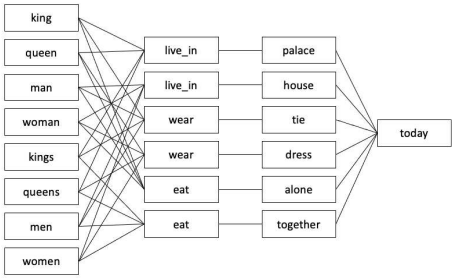

図 1　トイコーパスの 24 文の文構造

4 語からなる同一文型を有しており，品詞ごとに分類された語彙中の単語を組み合わせて作られている (例:“king live_in palace today”)．冠詞や動詞活用を省いた人工的な文法に従う文である．具体的な語彙としては，主語となる名詞が 8 単語（king，queen，man，woman，kings，queens，men，women)，述語となる動詞（句動詞）が 3 単語 (live_in，wear, eat)，動詞句を構成する名詞または副詞が 6 単語 (palace, house, tie, dress, alone, together)，文全体を修飾する副詞が 1 単語 (today) の計 18 単語である．その際，動詞と目的語，主語と目的語・副詞の間には意味的な対応関係があると想定する．例えば，king は palace と組み合わせることができるが，house とは組み合わないなどである．文型中のそれぞれの位置にある品詞に対応した単語を自由に組み合わせれば，$8 \times 3 \times 6 \times 1 = 144$ 文となるが，個々の単語間に意味的制約を課すことにより，その部分集合 (24 文）のみがトイコーパスに含まれることになる．これらの意味的対応関係が自由な文の生成を制約する結果として，単語間の共起頻度に一定の規則性が出現することになる．意味的制約の設計にあたっては，主語となる名詞 8 単語間に 3 軸の意味的対称性が生じるよう 3 つの動詞を用い，それぞれの動詞が目的語・副詞としてとれる単語 2 語のうち，主語の属性に応じてどちらかひとつの単語しか対応しないよう制約をした（図 1)．すなわち，3 つの動詞に対して 2 つの選択肢があるので，主語となる各単語は 3 ビットで表現されていることになる．このトイコーパスに対し，同一文内に共起する単語ペア (各文は 4 単語から構成されるので ${}_4C_2 = 6$ 組のペア）をカウントすることにより共起頻度行列（18 行 18 列）を作成し，各単語に対応する行ベクトルをそれぞれの単語ベクトルとした．

5. 共起行列における線形構造の分析結果

5.1 一様なトイコーパスの場合

まず，トイコーパス中の 24 文が一様確率（各文の出現確率がすべて $\frac{1}{24}$）で出現すると仮定して共起頻度行列より単語ベクトルを生成した．この時，主語となる名詞 8 単語の単語ベクトルが平行六面体を構成した (図 2)．すなわち 8 つの単語ベクトルを頂点とする立体において，6 つの面それぞれが平行四辺形となった．平行六面体の各面を構成する 4 点に対応する 4 つの単語は類推関係にあり，これらの単語ベクトルを用いた四項類推課題を解くとその正答率は 100%となる．例えば，$king - queen + woman$ の単語ベクトルを計算すると，これとのコサイン類似度が最も高いベクトルに対応する単語が $queen$ となる．6 つの面ごとに 2 組の平行関係があるので，四項類推課題は合計 12 組あるが，そのすべてにおいて正答することができる．

生成された共起頻度行列は，2 つのブロック行列 $C_0 \in \mathbb{R}^{8\times10}, C_1 \in \mathbb{R}^{10\times10}$ から構成され，次のように表現できる．

$$C = \begin{bmatrix} \mathbf{0}_{8,8} & C_0 \\ C_0^T & C_1 \end{bmatrix}$$

平行六面体の頂点を構成する 8 単語の単語ベクトルは，C の先頭 8 行の行ベクトルとなっており，そのうち非ゼロの部分ベクトルに対応するブロック行列が C_0 であるが，C_0 は階数 4 であり，3 次元アフィン空間に存在していることが明らかとなった．このとき，ブロック行列 C_0 は，3 つの線形独立なベクトル基底 $b_1, b_2, b_3 \in \mathbb{R}^8$ と平行移動 $b_0 \in \mathbb{R}^8$ を使って次のように表現することができる．

$$C_0 = (b_1, b_2, b_3)\mathbf{A} + b_0\mathbf{1}_{1,10}$$

$\mathbf{A} \in \mathbb{R}^{3\times10}$ は，アフィン基底 $\mathbf{B} = (b_0, b_1, b_2, b_3)$ に対応して一意に定まる行列である．

次に，これらの単語ベクトルを 3 次元空間上の座標へ写像することを考えるため，ブロック行列 C_0 の転置行列 C_0^T を考え，さらに $C_0^T \in \mathbb{R}^{10\times8}$ を各列毎に正規化した行列を $\bar{C_0}$ とする．

$$\bar{C_0} := C_0^T - \frac{1}{10}\mathbf{1}_{10,10}C_0^T$$

$\bar{C_0}$ は階数 3 の行列となるので，その先頭 3 列の列ベクトルを座標軸 $\bar{b_1}, \bar{b_2}, \bar{b_3} \in \mathbb{R}^{10}$ として $\bar{B} = [\bar{b_1}\ \bar{b_2}\ \bar{b_3}]$ とおくと $\bar{B}X = \bar{C_0}$ を満たす $X \in \mathbb{R}^{3\times8}$ の各列ベクトルが, 8 つの単語ベクトル（転置ベクトル C^T の列ベクトル）の座標を与える．具体的には次の式より算出される．

$$X = (\bar{B}^T\bar{B})^{-1}B^T\bar{C_0}$$

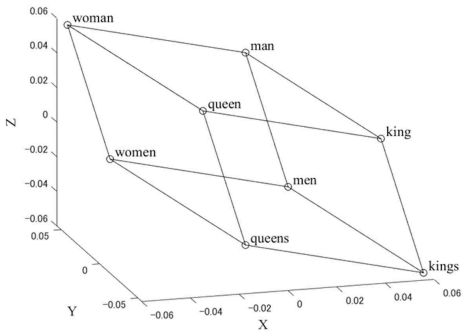

図 2 単語共起行列空間における主要 8 単語ベクトルの配置（一様なトイコーパスの場合）

これらの座標を XYZ 空間に図示すると, 図 2 のとおり平行六面体が現れ, 共起頻度行列の部分空間に 8 単語の平行関係が埋め込まれていることが確認できる．

5.2 非一様なトイコーパスの場合

次に，共起頻度行列において 8 つの単語群が平行六面体を構成しない可能性を調べるために，24 文の出現頻度が異なるトイコーパスを新たに作成した．24 文ごとにランダムに確率を付与して生成されたトイコーパスを用いて，あらためて共起頻度行列から単語ベクトルを生成すると，今度は 8 つの主語名詞に対応する単語ベクトルの配置は，平行六面体を構成していなかった（図 3)．

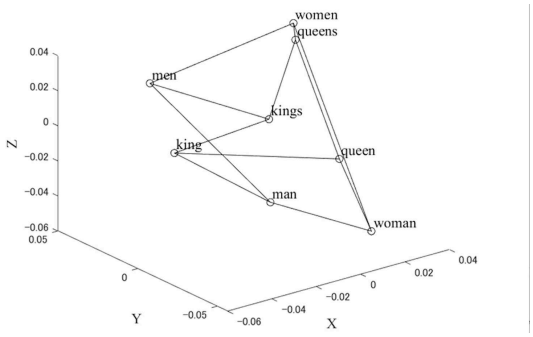

図 3 単語共起行列空間における主要 8 単語ベクトルの配置（非一様なトイコーパスの場合）

一様なトイコーパスの場合と比較すると，コーパス中の文の統語構造と各単語間の意味的制約は共通であるので生成される文自体は同じである．従って，非一様なトイコーパスにおいても語彙中の任意の単語の対が共起するかしないか，すなわち共起頻度行列中の要素がゼロであるか非ゼロであるかは，一様なコーパスのそれと同一である．異なるのは各文の出現頻度のみ

であることから，平行六面体が現れるためには，各文が出現する確率分布上になんらかの対称性が備わっていることが必要であると考えられる．

5.3 実コーパスによる検証

トイコーパスにより生成された二つの共起頻度行列に関する前節までの分析は，平行六面体関係として現れた単語ベクトル間の配置が，統語的・意味的構造のもとでのなんらかの統計的な規則性・対称性を反映していることを示唆していた．この予想を検証するため，実際にこのような規則性や対称性を持つ分布構造が自然言語の中に存在するのか，加藤ら [7] で用いた実コーパスから共起頻度行列を作成して検証を行った．実コーパスは 2017/10/1 の英語 Wikipedia ダンプ (Python の Gensim ライブラリ [8] 中のデータから取得）であり，コーパス全体のトークン数は約 79 億，語彙数は約 260 万である．共起頻度をカウントする際の窓枠は左右ともに 5 としている．本研究では，このうち類推に使用される Google テストセット [9] 中の単語と最も多く共起する単語 1000 語に，それらに含まれないが類推テストに含まれる単語を加えた 1487 語を対象として共起頻度行列 (1487×1487) を作成した．

その上で，各要素の対数をとった対数共起頻度行列に対して特異値分解によるスペクトル分解を行い，平行関係を埋め込んだ部分空間が対数共起頻度行列内に存在するかどうか探索を行った．探索は Family カテゴリーの単語群を対象に行い，その結果，図 4 に示すように，king, queen, man, woman, he, she, father, mother の 8 語間について，2 つの特異値に対応した行列の部分空間に 3 つの対称性が埋め込まれており，平行六面体ではないが，平行四辺形と台形に近い面から構成される六面体を形成していることが確認された (図 4)．つまり，トイコーパスによって再現された線形構造は，実コーパスにおいても共起頻度行列の部分空間に存在することが確認された．

6. 考察

本研究では，まず人工的な文構造と単語間の意味制約のもと一様分布による確率で生成されたトイコーパスにおいて，その共起頻度行列の部分空間に平行関係が出現することを示した．平行六面体を構成する 8 単語の単語ベクトルがなす行列は階数 4 で 3 次元のアフィン空間に存在していた．一方，同じ文構造と意味制約のもと各文の出現確率を非一様とした場合には，平行六面体の配置が崩れることとなった．また，実

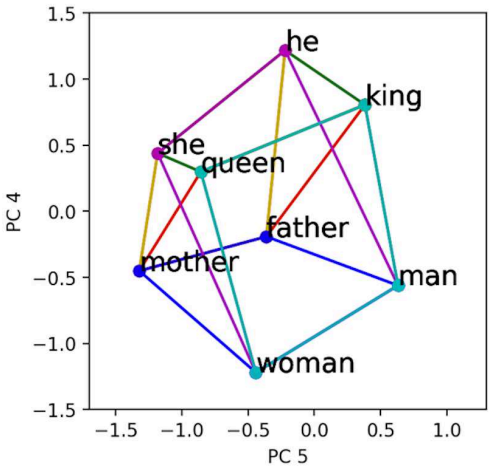

図 4 実コーパスから生成された単語共起行列空間における Family カテゴリーの 8 単語ベクトルの配置

コーパスから作成された対数共起頻度行列を特異値分解した部分空間において，意味的な対称性を持つと思われる単語群が平行六面体に近い幾何的関係を構成していることが確認できた．

これらの結果より，単語ベクトルが平行六面体をなすための条件として，言語の生成プロセスと意味的・統語的関係（言語構造的要件）と文の出現頻度に関する要件の両方が関わっていることが示唆される．まず，後者に関しては，一様分布により生成されるトイコーパスが高い対称性を有していたが，どこまでその対称性を緩めることができるのか，平行六面体が成立するための必要十分条件を数学的に導出する必要がある．また，四項類推課題のテストセットに含められるような単語群では高い精度で平行四辺形が構成していることになるが，その場合，出現頻度に関する必要十分条件が充足されているのか否か実コーパスでの検証が必要である．

前者の言語構造的要件に関しては，一様分布のトイコーパスと非一様分布のトイコーパスのそれぞれから生成された共起頻度行列において，ゼロと非ゼロの要素が同じ位置にあることから，文の出現頻度ではなく言語構造が共起行列の構造を制約している側面もあると考えられる．その場合，仮に文の出現頻度が非一様でなかったとしても共起頻度行列の構造を抽出することで，言語構造に由来する平行関係を抽出することができる可能性も考えられる．その場合，どのようなバイアス補正を行えば，非一様な共起分布から言語構造の規則性を抽出することができるのか，また word2vec や，PMI を用いた既存の単語ベクトル生成手法がそのような言語構造的規則性を抽出しているのか否か今後

の研究課題となる．

7. 結論

本研究は，これまで研究対象とされてこなかった単語共起行列に対して，人工的なトイコーパスを用いた構成論的アプローチをとることで、単語共起頻度行列に部分空間として内在する線形構造の存在を明らかにした．単語ベクトル間に四項類推課題を解くための平行四辺形関係が成立するための要件として，文型と意味的・統語的制約による言語構造，並びに，出現頻度が関与している可能性があることが示唆される．先行研究で指摘されてきたように，共起行列にすでに自然言語の構造が内在しており，word2vec などの学習モデルや PMI などの前処理が一定の行列変換を行うことにより自然言語の構造を抽出しているという予想に対して理論的根拠を与える試みの端緒となるものである．今後の研究課題として，平行四辺形をなすための言語構造的要件と文の出現頻度に関する必要十分条件の導出と検証を行うとともに，平行関係以外の幾何的関係の存否についても検討していきたい．自然言語に内在する構造を文の生成プロセスまで遡って定式化しようとする試みはこれまでなされてこなかったことであり，今後の共起行列の内部構造解明に対する有効な手法となりえると考えられる．

謝辞

本研究は JST さきがけ JPMJPR20C9 の助成を受けて行われた．

文献

[1] Lenci, A. (2018) “Distributional models of word meaning” The Annual Review of Linguistics, Vol. 4, pp. 151-171.
[2] Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019) “BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Vol. 1(Long and Short Papers), pp. 4171-4186.
[3] Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., & Hasson, U., (2022) “Shared computational principles for language processing in humans and deep language models” Nature Neuroscience, Vol. 3, pp. 369–380.
[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J., (2013) “Distributed representations of words and phrases and their compositionality”, Proceedings of the 26th International Conference on Neural Information Processing Systems, Vol. 2, pp.3111-3119.
[5] Levy, O.,Goldberg, Y. & Dagan, I., (2015) “Improving distributional similarity with lessons learned from word embeddings” Transactions of the Association for Computational Linguistics, Vol. 3, pp. 211-225.
[6] Church, K. W., & Hanks, P. (1989) “Word Association Norms, Mutual Information and Lexicography” ACL
[7] 加藤龍彦，日高昇平，鳥居拓馬，(2020) “対数共起頻度を用いた四項類推：word2vec と PMI の比較” 2020 年度第 34 回人工知能学会全国大会
[8] Rehurek, R. & Sojka, P., (2010) “Software framework for topic modelling with large corpora”, Proceedings of the LRECWorkshop on New Challanges for NLPFrameworks, pp.45-50.
[9] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Google Analogy Test Set