単体複体を用いた次単語予測分布の幾何的解釈

石峯 拓海 [1,2]　日野 英逸 [3]　横井 祥 [2,4,5]
[1] 東京大学　[2] 国立国語研究所　[3] 統計数理研究所　[4] 東北大学　[5] 理化学研究所
ishimine-t@g.ecc.u-tokyo.ac.jp　hino@ism.ac.jp　yokoi@ninjal.ac.jp

概要

言語モデルの次トークン予測は，形式上は数万項の単語集合の上の確率分布であるが，実際には文脈に応じた小規模な語群の上だけに確率値が集中する．この，事前学習から自然に生じる単語予測分布の低ランクな構造を，単体複体という概念を用いて表し，「次単語予測分布の幾何構造」を考えたい．提案アルゴリズムは，未知文脈の候補集合を包含するように単体を構成することで，コーパスサイズによらない（Heaps 則に適した）推定を可能にする．Olmo 7B と Dolma を用いた実験で，名詞の文脈と動詞をとる文脈が自然に分離されるなど，次単語予測が言語的に自然な幾何構造を内包していることを確かめた．さらに，単体複体を構成する単体（＝次トークン候補の集合）の頻度が，単語と同様に種々のべき乗則を満たすことを確認した．

1　はじめに

大規模言語モデル（LLM）の興隆に伴い，各種の概念が線形に埋め込まれているとする線形表現仮説 [1, 2] や，その概念に相当するベクトルを特定しようとする線形プロービング [3] など，内部状態の幾何学的な解釈が多数試みられている．

また LLM の主機能である次単語予測に着目すると，形式上は数万項に及ぶ単語の上の分布が出力されるが，経験的には文脈に応じた小規模な語群の上の低ランクな分布をなす事実が知られている [4, 5, 6]．例えば，I never . . . という文脈では，thought, knew, などの動詞の確率が高くなり，文法的にとりづらい単語の確率は 0 とみなせる．

本稿では，この次単語予測分布の低次元性を，位相幾何学で用いられる単体複体という幾何的概念（図 1 右下）を介して理解する方法を提案する．文脈 c の次単語予測の確率値が語群 s_c 上に集中するとしよう．s_c を（三角形や四面体の一般的概念である）単体とみなすと，次単語予測分布の全体は，小さな単体の集まりである単体複体とみなせる．

言語モデルに対応する単体複体を推定するため，まず LLM から「文脈に応じた次トークン候補集合」を集め，ここから単体複体を自然に誘導する枠組みを定式化する．ただしここで言語データの Heaps 性，すなわちコーパスの拡大に伴って新しい単体が観測されつづける特性 [7] が障害となる．観測された単体から素朴に作成した単体複体は利用する有限のコーパスに依存してしまうため，言語全体を表現する単体複体を工夫して構成する必要がある．そこで我々は，未知の単体をどれだけ予測可能かを評価する閉性の指標を導入し，与えられたコーパスには現れない次トークン候補集合を包含するような単体を推定する手続きを提案する．

実験では，推定した単体複体が言語的に自然なクラスタ構造を捉えていること，未観測コーパスに対して一定の頑健性を持った推定が可能であること，また得られる単体（単語集合）とその頻度がべき乗的な挙動を示すことを確認した．この研究により，これまで一つの巨大な空間として扱われてきた [8, 1, 9] 言語モデルの内部機序に，たくさんの小規模な空間の集積という新しい視点 [10] を提供したい．

2　準備

2.1　言語モデルに関する記号と仮定

V をトークンの集合とする．**言語モデル** LM は，文脈 $c \in C \subset V^*$（V^* は V の有限列全体の集合，C はコーパスとトークナイザで決まる文脈の集合）を受け取り，確率分布 $\boldsymbol{p} \in \Delta^V$ を返す関数

$$\mathrm{LM} := V^* \to \Delta^V \tag{1}$$

である．$\Delta^V := \{\boldsymbol{p} \mid p_i \geq 0, \sum_i p_i = 1\}$ は次に定義する単体の一つである．

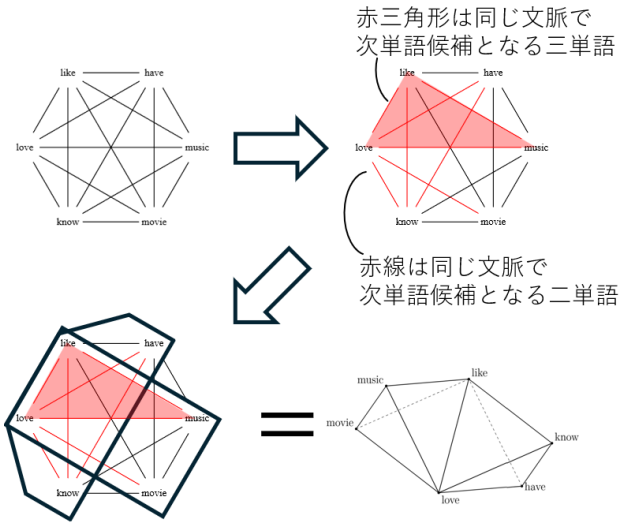

図 1　確率分布から単体複体を導く様子．言語モデルの次トークン予測は全語彙中の確率分布=Δ^V 中の一点を与える（左上）が，その確率は文脈に応じて少数語彙に集中している（右上）．これは四項上の確率分布が二つ重なったものと考えられ（左下），その確率分布の幾何構造は正四面体二つが張り合わさった単体複体をなす（右下）．

2.2　単体複体について

単体とは，$|s|$ 個の頂点で構成される $|s|-1$ 次元の幾何学的対象である．点，線分，正三角形，正四面体，正 $|s|$ 胞体は単体である．単体は，低次の単体を辺や面にもつ．例えば，正四面体は線分や正三角形を構成要素に持つ．単体複体とは，複数の単体が張り合わされたものである [10]．例えば図 1（右下）は，二つの四面体を共通する線分ではり合わせたものである．この構造を，単体を集合 s，単体複体をその集まりによってもあらわす [11]．

定義 1 (単体複体とその実現)**.** 語彙集合 V の部分集合 s や，s **のみが正の確率を持つような確率分布全体**を単体と呼ぶ．単体が集まったものを単体複体と呼ぶ．単体同士は，その共通部分のなす部分単体同士で張り合わされている．

3　理論

3.1　確率分布を単体複体中の点とみなす

前述のとおり，コーパス C の中の文脈 c での LM の確率分布 $\mathbf{p}^c$ は，c に応じた一部の語群 s_c が多くの確率を担っているので，自然に $\mathbf{p}^c$ は単体 s_c に属しているとみなせる．コーパス C の各元 c について，次トークン候補集合 s_c を集めると単体複体となる．これを $\Sigma(C)$ とかく．

3.2　Heaps 則によるコーパス依存性

定義した $\Sigma(C)$ は，図 1 の右上にあたるが，言語データの経験的事実から，C を十分大きくしたとき，$\Sigma(C)$ はある一定の単体複体に収束していくとは考えづらい．一般に，言語データの語彙数 $\#V$ とコーパスサイズ $\#C$ には，$\beta \in [0,1]$ によるべき乗則

$$\#V \sim \#C^{\beta}$$

が成立することが知られている [7]．この法則は $\Sigma(C)$ に属する単体を語彙とみなしても成立し，単体複体はコーパスに依存して拡大していく．コーパスサイズに依存しない単体複体を構成することを以降の目標とする．図 1 の例では，右上の赤線/赤三角が与えられたとき，未知のサンプルを包含するように黒枠を設定することに相当する．これは，以下のような学習問題であると言い換えられる．

- 教師データ C_{train}：与えられたコーパス/図 1 赤線/赤三角
- 学習モデル $\Sigma^+(C_{\text{train}})$：単体複体＝語彙の組の集合/図 1 黒枠
- 評価データ C_{test}：未知の文脈（または別のコーパス）
- 目的関数：$P(\mathrm{LM}(c_{test}) \in \Sigma^+)$/未知の候補語彙集合を黒枠内に収めること

3.3　未知文脈に対応した単体複体の構成

3.3.1　指標の定義

未知の予測語彙集合を含むような組=単体を作成するためには，似た文脈で発生しやすい単語の組を探せばよい．そのような語群で高くなりやすいと考えられる指標を定義し，求める単体複体を探索的に構築することを目指す．

定義 2 (cloneness)**.** $u \subseteq V$ と $i \le \#u$ について，閉性 (closeness) を

$$\mathrm{closeness}_i(u) := \frac{P(\mathrm{LM}(c) \in \Delta^u)}{P\left(\mathrm{LM}(c) \in \bigcup_{\#(u \cap s) \ge i} \Delta^s\right)} \tag{2}$$

と定義する．任意の u, i について，$\#(u \cap u) \ge i$ より，$\cup\Delta^s \supseteq \Delta^u$ に注意すると，$\mathrm{closeness}_i(u) \in [0,1]$ である．

この指標の解釈を与えておく．異なる文脈が，共通部分を多く持つ次トークン候補集合 s_1, s_2 を持つとき，これらの文脈は似ているといえる．未知の似た文脈の候補集合 s_x も，これらの候補集合と多く

の共通部分を持つだろう．s_x を含む蓋然性の高い単体 u としては，$s_1 \cup s_2$ が考えられる．同様に $s_3, s_4 \ldots$ があれば，これらの和集合は s_x をより確実に含むといえる．closeness ~ 1 は，このように構成できる極大の集合であることを表す．

3.3.2 アルゴリズム

初期状態を空集合 $\Sigma' = \varnothing$ とする．単体 $u \subset V$ を以下のように逐次的に n_{simplex} 件構成し，付け加えていくことで $\Sigma^+(C_{\text{train}})$ を構成する．

1. 事前に，各トークンに対する重み $\{w_v\}_{v \in V}$ を，v が候補集合に現れる確率の，上に凸な関数 f による変換と正規化によって計算しておく
2. u をランダムに選んだ $s \in \Sigma(C\text{train})$ で初期化
3. 各 v について，以下のように score 付けをする
   - v を含む候補集合 s について，$k_s := |s \cap u|$ を計算する
   - $k_s > i$ か s が u に含まれるなら正，それ以外なら負の重みをつけ，和を score とする
4. score と w_v の積の上位 $n_{\text{add\_vertex}}$ 件を u に追加
5. 3，4 を n_{cycle} 回繰り返す
6. $|u| > d_{\min}$ を満たすうち closeness_i が最大のものを $\Sigma^+(C_{\text{train}})$ に追加

4 実験

4.1 前処理と Σ_C の構成

公開 LLM である Olmo 7B[12] とその学習データの一部 [13]36M トークンを用いて次トークン予測を計算した．確率分布上位 150 件中，top-p サンプリング [6]（$p = 0.99$）によって候補集合を得た． 図 2

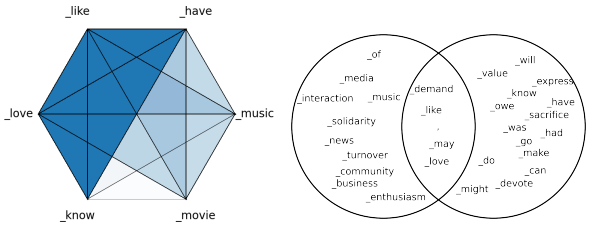

図 2 図 1 の実際の例（左）と二つの候補集合が重なる様子（右）．図 1 で図示した二つの四頂点の内部だけでなく，"music"-"have"も同時に候補集合に現れうることがわかる．ベン図では，実際に名詞をとる文脈-動詞をとる文脈で"like", "love"といった単語が共通部分になっていることがうかがえる．

は図 1 で例示した六単語の実験での振る舞い，図 3

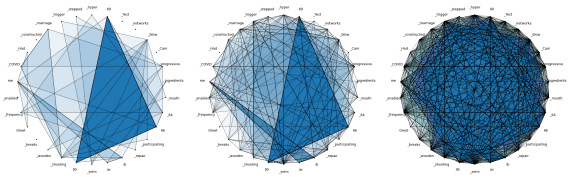

図 3 一部のトークンでの，コーパス C 中の次トークン候補集合の集まりを単体複体 $\Sigma(C)$ とみなしたもの．色付きの多角形が単体＝同時に次トークン候補となった組み合わせを，色の濃さが頻度を表す．それぞれ $|C| = 10^4$（左），10^5（中），10^6（右）．

は，一部トークン 30 件（＝ 30 頂点）での単体複体の様子をコーパスサイズ別に示したものである．図 3 でもっとも顕著な三角形は"00"，"66"，"69"を結ぶものである．この 3 トークンが同時に候補集合となりやすいことは，直感的に明らかである．コーパスサイズが増えるにつれ，登場する単体の種類が増大していることも見て取れる．図 4 は，単体の次数の

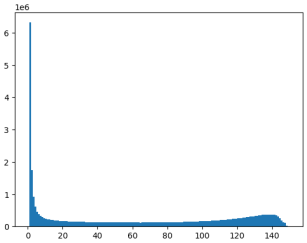

図 4 $\Sigma(C)$ 中の単体の次数の分布

分布を示している．低次元での分布が突出していることが見て取れる．

4.2 Σ^+ の作成

計算したトークン予測のうち 10%（3.6M トークン）を C_{train} として，$\Sigma^+(C_{\text{train}})$ を計算した．$\Sigma^+(C_{\text{train}})$ と，訓練データに含まれない $1M$ 件の文脈の次トークン候補集合 s に対して，共通部分の要素数の分布，すなわち図 1 での（候補集合のサイズ，候補集合のうち黒枠内にある要素の数）の分布を計算した（図 5）．理論の節で目標とした $P(\text{LM}(c_{test}) \in \Sigma^+(C_{\text{train}}))$（図 1 において未知文脈の候補集合が完全に黒枠内に入る確率），は，図 5 の分布中の対角線上の割合を示しており，$n_{\text{simplex}} = 180000$ で 39.3%, $n_{\text{simplex}} = 18000$ で 34.8%であった．

4.3 次単語候補集合によるべき則の確認

図 6 のとおり，単体＝各文脈での次トークン候補集合の頻度と種類による Zipf 則（左）と Heaps 則（右）が確かめられた．線形回帰の結果，単体の種類を S として，$S = k\#C^{\beta}$ の係数はそれぞれ

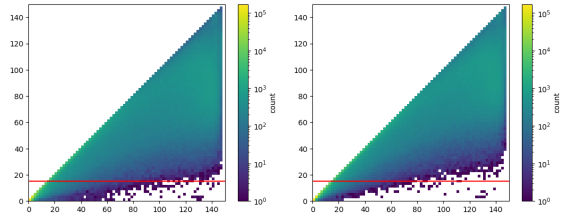

図 5　単体の要素数（横軸）と $\Sigma^+(C_{\text{train}})$ との最大共通部分の要素数（縦軸）の分布（1M 件中）．$n_{\text{simplex}} = 18000$（左），180000（右）

$k = 1.03, \beta = 0.97$ となった．

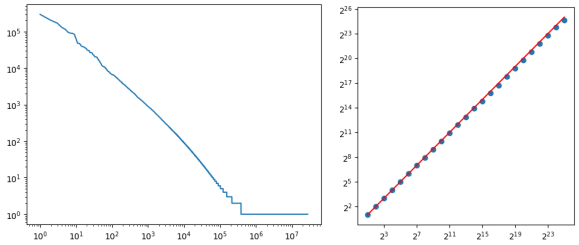

図 6　単体の頻度順の rank（横軸）と頻度（縦軸）の関係（左），コーパスのサイズ（横軸）と現れる単体の種類（縦軸）の関係と線形回帰の結果（右）

また，各トークンが候補集合中に現れる確率について，図 7 の**非** Zipf 性がみられた．

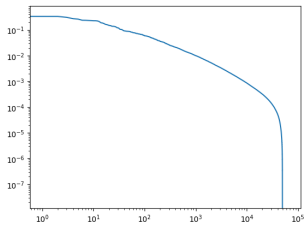

図 7　トークンの頻度順の rank（横軸）と頻度（縦軸）の関係

4.3.1　考察

図 5 から，単体複体 $\Sigma^+(C_{\text{train}})$ の元が増えると，未知の候補集合が完全に含まれる確率が上昇している．しかし，多くの未知文脈が $\Sigma^+(C_{\text{train}})$ に含まれないままであった．図 2 の例でいえば，図 1 の二つの黒い四角形内に含まれない (music, love,have) の三角形を同時に含む候補集合をとるような文脈が存在したことに相当する．一方で，実験において，ほとんどの未知文脈の次トークン候補集合で，$\Sigma^+(C_{\text{train}})$ との共通部分が空でないことが確認できた．例えば，$\#s_c \geq 15$ を満たす c のうち，共通部分の要素数 ≥ 15 である c は 98.4%であった．これは，先の例では，枠内に含まれない (music, love, have) の組でも，二単語は黒い四角に入っていることに対応する．これは幾何的には，全体の幾何構造をなす単体複体のほとんどの単体が，$\Sigma^+(C_{\text{train}})$ と 15 次以上の共通部分によって接していることを示している．C を大きくしたときの $\Sigma(C)$ の元は，$\Sigma^+(C_{\text{train}})$ に低次の単体が付け加えられた形により与えられることが推測される．比喩的には，$\Sigma^+(C_{\text{train}})$ が「幹」のように振る舞い，ここにユニークな「葉」が無数に加えられたような幾何構造になっているといえるだろう．

Heaps 則は β が非常に 1 に近い，減衰の少ない形で確認された．これは，人間の認知能力によって上限がある [14] 語彙集合に対し，その部分集合は爆発的な組み合わせ数を持つことが影響していると考えられる．非べき乗則は，高頻度語が，実際に観測されるより多くの文脈で，低いが 0 でない確率で候補集合に現れていることを示唆している．例えば，図 2（右）のベン図においても，高頻度語である","が両者において観測された．

5　結論と展望

本稿では，LLM の確率分布を，単体複体の構造を持つ幾何的対象とみなし，コーパスから幾何構造を推定する方法を提案した．実験では，提案手法により得られる $\Sigma^+(C_{\text{train}})$ が，すべての未知文脈の次トークン候補集合を含まないまでも，ほとんどの候補集合と一定以上の共通部分を持ち，$\Sigma(C)$ の幹といえる単体複体を構成することを確認した．これにより，言語の幾何構造に一つの示唆と定式化を与えたといえる．この定式化に基づき，より高度な数学の道具立ての適用や，より厳密で再現性の高い内部機序理解が期待されるほか．特に，内部機序理解の代表的な対象の一つである隠れ状態 [3, 1] のうち最終層は，softmax と線形変換のみによって確率分布と結びついている [15, 12]．本稿の議論を引き戻した最終隠れ状態や，一般の隠れ状態の幾何的考察が，今後の展望である．

提案手法の課題として，closeness の高い集合を作成するヒューリスティックが必要とされること，特に任意に与えられた文脈 c について LM(c) を含み高い閉性を達成する単体が存在するか/現実的な計算量で構成が可能かどうかという問題がある．

謝辞

本研究の一部は，JSPS 科研費 JP23K24909, JP25H01494, JP22H05106, および JST 創発 JPMJFR2331 の支援を受けて実施されました．

参考文献

[1] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. **arXiv [cs.CL]**, November 2023.

[2] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. **arXiv [cs.AI]**, October 2023.

[3] Wes Gurnee and Max Tegmark. Language models represent space and time. **arXiv [cs.LG]**, October 2023.

[4] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. **arXiv [cs.CL]**, May 2018.

[5] Alec Radford, Jeff Wu, R Child, D Luan, Dario Amodei, and I Sutskever. Language models are unsupervised multitask learners. 2019.

[6] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. **arXiv [cs.CL]**, April 2019.

[7] 田中久美子. 言語とフラクタル : 使用の集積の中にある偶然と必然. 東京大学出版会, 2021.

[8] Tomas Mikolov, Wen-Tau Yih, and G Zweig. Linguistic regularities in continuous space word representations. **North Am Chapter Assoc Comput Linguistics**, pp. 746–751, May 2013.

[9] Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. **arXiv [cs.LG]**, May 2024.

[10] Allen Hatcher. **Algebraic Topology**. Cambridge University Press, Cambridge, England, December 2001.

[11] 服部晶夫. 位相幾何学. 岩波基礎数学選書 / 小平邦彦監修 ; 岩堀長慶 [ほか] 編集. 岩波書店, 1991.

[12] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. **arXiv [cs.CL]**, February 2024.

[13] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An open corpus of three trillion tokens for language model pretraining research. **arXiv [cs.CL]**, January 2024.

[14] Alessandro Bellina and Vito D P Servedio. Cognitive limits shape language statistics. **arXiv [physics.soc-ph]**, March 2025.

[15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. **arXiv [cs.CL]**, February 2023.