長距離相互作用する文脈依存言語における相転移現象 — 言語モデルの創発現象を統計力学の視点で理解する —

都地悠馬 [1] 高橋惇 [2] 横井祥 [3,4] 栗林樹生 [5] 上田亮 [6] 宮原英之 [7]
[1] 北海道大学 工学部 [2] CQuIC, University of New Mexico [3] 東北大学 大学院情報科学研究科
[4] 理化学研究所 革新知能統合研究センター [5] MBZUAI NLP Department
[6] 東京大学 情報理工学系研究科 [7] 北海道大学 大学院情報科学研究院
toji.yuma.c1@elms.hokudai.ac.jp, juntakahashi@unm.edu, yokoi@tohoku.ac.jp,
tatsuki.kuribayashi@mbzuai.ac.ae, ryoryoueda@is.s.u-tokyo.ac.jp,
miyahara@ist.hokudai.ac.jp

概要

最近，大規模言語モデル (LLM) のスケーリング則や創発的な能力が報告され，LLM のメカニズムを理解する上で重要な手掛かりになることが期待されている．実は統計力学においてもこれらの概念に相当する「相転移」と呼ばれる概念が存在する．本研究では，言語モデルの性質を相転移の観点で再検討する．具体的には，長距離相互作用を持つ 1 次元イジング模型を念頭にした単純な言語モデルを構成し，統計力学的な意味での相転移現象が起きることを示す．さらに，シンボル数が増えるという統計力学モデルにはない言語モデル特有の過程に注目することで，統計力学モデルにはない現象を発見したことを報告する．

1 はじめに

ChatGPT [1] や Google Bard [2] などの大規模言語モデル (LLM）がどのような性質を持つのか，またなぜ優れた性能を示すのかといった問いに対して，LLM の機序と原理の解明に向けた研究が盛んに行われている．面白い知見として，[3] では LLM におけるスケーリング則が報告され，[4] では LLM の創発的な性能の向上が示されている．これらの現象はシステムサイズ (学習データ，学習時間，パラメータ数) を大きくしたときに現れる現象であり，統計力学における相転移現象との類似性が見られる．本研究では，言語モデルを相転移の観点から理解することを試みる．

相転移とは，系を指定するなんらかのパラメータ (通常は温度や圧力など) を変化させたときに，ある値を境に急激に系の巨視的な性質 (通常は磁化，密度，結晶構造など) が定性的に変化する現象で，統計力学においてよく研究されている．例として，水を冷やすと 0 °C で突然性質の異なる氷になる, あるいは 磁石を温めるとある温度で突然磁力が 0 になる，といった現象が知られている．相転移現象は，物質の状態が大きく変わる現象としても興味深いが，ある種の相転移が起こる点 (臨界点) 直上では興味深いスケーリング則が現れることも重要な特徴である．また，相転移はシステムサイズを大きくしたとき [1] に現れる現象で，この点でも LLM との対応がある．

なお，自然言語と相転移の関連を論じること自体は本研究がはじめてではなく，特に文脈自由文法（CFG）の相転移を論じる既存研究が複数ある．しかしこれらの研究は依然共通見解に至っておらず，とくに，どういうクラスの言語が相転移を持つのかについては不明な状況が続いている [5, 6, 7, 8][2]．

本研究では，CFG よりも強力な文法である文脈依存文法 (CSG) の枠組み内で明確な相転移現象が起こることを，長距離イジング模型から着想を得て構成した CSG の数値計算に基づいて主張する．自然言語を統計力学的にモデル化する際には，素朴には「文の初めから終わり」に対応する 1 次元方向のみを持つと考えるのが自然であろう．統計力学分野において相転移を示す最も単純な模型の 1 つであるイジング模型は，空間次元を 1 次元としたときに短距離相互作用のみでは相転移を示さず，長距離相互作用を持つ場合 (長距離 1 次元

1） 系の構成要素の数 N が無限に大きくなる極限（熱力学極限と呼ばれる）を数学的には考えることになる．

2） [5, 6] は CFG 相転移があるという主張をし，一方で [7] は [5, 6] の解析が間違いであると指摘し，さらに [8] は [7] に対して反論を行っている．

イジング模型) は相転移を示すことが知られている [9, 10, 11, 12, 13, 14, 15, 16, 17]．このイジング模型の性質に従い，長距離イジング模型に対応するような CSG を構成し，相転移の有無を調べる．

また，通常の統計力学モデルは持たず言語モデルは持つ特有の性質として，時事刻々とシンボルが増えるという点がある．このシンボルの増え方に対して，相転移点の挙動が変わることも議論する．この現象は，元々の統計力学のモデルにはなく自然言語をモデル化した際に生じる新しい現象であるという意味でも興味深い

2 問題設定

本研究で扱うモデルと計算する物理量を説明する．モデルに関しては，シンボル間の相互作用の入りかたを重点的に説明する．計算する物理量は，相転移を議論する上でよく用いられる量を導入し，それらが自然言語の何に対応し得るかを説明する．

2.1 提案モデル

本研究では，自然言語を大幅に簡略化した模型として，非終端記号と終端記号をそれぞれ 2 つ持ち，以下の 3 種の生成規則を持つ言語モデルを考える:

$$X \to x, \tag{1a}$$

$$X \to YZ, \tag{1b}$$

$$\alpha X \beta \to \alpha Y \beta. \tag{1c}$$

大文字の X，Y，Z は非終端記号であり，A，B をとる．また，小文字の x，y は終端記号であり，a，b をとる．式 (1a) は非終端記号がそれぞれ対応する終端記号になる規則群である．式 (1b) は非終端記号が 2 つの非終端記号に置き換わる規則群である．式 (1c) は非終端記号に文脈依存性を導入しており，前後の文脈 α と β をもとに，非終端記号を書き換える（後述）．自然言語の文脈依存性の程度については議論の余地があるが，文脈アクセスが強力な昨今のニューラル言語モデルを念頭におき CSG を採用する．なお式 (1c) を除くと CFG となり [5, 6, 7, 8] で扱われたモデルと一致する．式 (1a)，式 (1b)，式 (1c) の規則が採択される確率はそれぞれ qt，$q(1-t)$，$1-q$ とする．各規則群内で適用可能な規則が複数存在する場合は，それぞれを等確率で用いる．$t \to 0$ は非終端記号が現れない極限であり，相転移を議論する上で最も単純な熱力学極限（シンボル数無限大の極限 $N \to \infty$）を議論できるパラメータである．

文脈依存性: 式 (1c) の詳細を述べる．N を文字列全体の長さとし，$\alpha \coloneqq X_{-[Nr_\alpha]} X_{-([Nr_\alpha]-1)} \dots X_{-1}$，$\beta \coloneqq X_1 \dots X_{[Nr_\beta - 1]} X_{[Nr_\beta]}$ とする．X は任意の文字（非終端記号含む）を表し，下付きの値は，式 (1c) における A (B) を基準とした相対位置を表す．$[\cdot]$ はガウス記号であり，r_α，r_β は長距離イジング模型の相互作用範囲に対応するパラメータである．

以下の確率でシンボルを変更する:

$$p = \min[1, e^{-\Delta E/T}]. \tag{2}$$

ただし，ΔE の定義は以下である:

$$\Delta E \coloneqq 2J\left(\sum_{\ell=1}^{[Nr_\alpha]} \frac{\sigma_0 \sigma_{-\ell}}{\ell^{1+s}} + \sum_{\ell=1}^{[Nr_\beta]} \frac{\sigma_0 \sigma_\ell}{\ell^{1+s}} \right). \tag{3}$$

さらに，σ_j の定義は以下で与えられる:

$$\sigma_j \coloneqq \begin{cases} 1 & (X_j = A, a), \\ -1 & (X_j = B, b). \end{cases} \tag{4}$$

式 (2) より，ΔE が小さいほどシンボルは反転しやすくなる．また式 (3) より，ΔE は $\sigma_0 \sigma_{-\ell}$ と $\sigma_0 \sigma_\ell$ が負の値をとるときに小さくなる．式 (4) より，これは周辺記号が自身と異なる記号のときに記号が反転することを意味し，T は同じ記号を揃えたがる傾向の強さを定めるコントロールパラメータである[3)]．式 (3) が長距離相互作用と言われる理由は，シンボル間の相互作用がべき的（$1/\ell^{1+s}$）に減衰しているためである．またここで s はべき減衰の強度を決定するパラメータであり，本論文では $s = 0.9$ と固定した．まとめると，式 (1c) の規則は，確率 $1-q$ で適用され，さらに式 (2) によって反転するか否かが判定される．

式 (1c) の規則は統計力学的には，長距離 1 次元イジング模型の 1 シンボルのフリップに関する Metropolis-Hastings 法を用いた Monte Carlo シミュレーションの操作に対応している．式 (3) はその際に参照される，長距離 1 次元イジング模型において 1 スピン (1 シンボル) のみを変更した場合のエネルギー変化である．

2.2 相転移を議論するための物理量

相転移はオーダーパラメータの特異性で議論する．オーダーパラメータとは何かしらの確率変数の期待値であり，イジング模型の例では磁化 (アップスピンとダウンスピンの割合) を採用することが通例である．

3) 統計力学では絶対温度に対応する．

本研究では磁性体を扱うわけではないが，2 シンボルを考えるとイジング模型と同様にオーダーパラメータを定義することができる．そのオーダーパラメータを磁化と呼ぶことにし，以下で定義する:

$$M \coloneqq \frac{N_A - N_B}{N}. \tag{5}$$

ただし，N_A と N_B をそれぞれシンボル A と B の数とし，$N \coloneqq N_A + N_B$ とする[4]．M は「単語 A と単語 B の頻度がどの程度偏るか」を表す量である．以下では，独立変数である温度 T を動かした時に，従属変数である M がどう変化するかを観察することで，1 次元イジング模型（最も単純な統計力学模型）を元に構成される言語の性質が質的に急激な変化を起こす点（相転移点）が存在するかを考える．相転移点直上で，熱力学極限 ($N \to \infty$) において磁化 (式 (5)) が微分不可能になることが相転移の特徴づけになる．

続いて，帯磁率を以下で定義する:

$$\chi \coloneqq N(\langle |M|^2 \rangle - \langle |M| \rangle^2). \tag{6}$$

帯磁率 (式 (6)) は相転移点直上で発散することが知られており，実験的な測定のしやすさからよく調べられている．

磁化 (式 (5)) の微分不可能性，帯磁率 (式 (6)) の発散は，最も基本的な相転移の特徴付けであるが，数値実験においては熱力学極限 ($N \to \infty$) を取れないというシステムサイズの困難がある．そこで，Binder パラメータを用いた解析がよく用いられる．Binder パラメータを以下で定義する:

$$U \coloneqq \frac{1}{2}\left(3 - \frac{\langle M^4 \rangle}{\langle M^2 \rangle^2}\right). \tag{7}$$

スケーリング解析より Binder パラメータ (式 (7)) の温度依存性をシステムサイズを変えながらプロットすると，臨界点では交差することが知られている他，臨界的でない相転移点においては負に発散することなどが知られている．よって，数値計算における相転移点の同定には Binder パラメータ (式 (7)) のシステムサイズ依存性を調べるという手法が標準的になっている．

3 数値計算

提案モデル (式 (1)) の Binder パラメータ (式 (7)) の結果を示す[5]．式 (1b) の規則として，$X \to YZ$ (等確率でシンボルが増える場合) と $X \to XX$ (シンボルが複製される場合) を考える．

4) 2.1 節で定義した文字列の長さ N と等価である．

5) 磁化 (式 (5))，帯磁率 (式 (6)) の計算は，参考情報 (A) に示す．

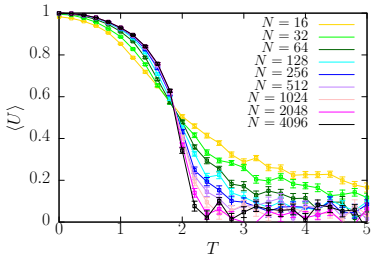

図 1 $X \to YZ$ の場合の Binder パラメータの温度依存性．$q = 10^{-2}$，$t = 0$，$s = 0.9$，$r_\alpha = r_\beta = 0.25$ とした．シンボル数 N を 16，32，64，128，256，512，1024，2048，4096 と変えた．

3.1 $X \to YZ$ (等確率でシンボルが増える場合)

$X \to YZ$ (等確率でシンボルが増える場合) を考える．非終端シンボルの数は 2 つなので，8 つの過程があることに注意する:

$$\begin{aligned} &A \to AA,\ A \to AB,\ A \to BA,\ A \to BB,\\ &B \to AA,\ B \to AB,\ B \to BA,\ B \to BB. \end{aligned} \tag{8}$$

また，これらの過程が A，B に対してそれぞれ等確率で起こるとする．

図 1 に数値計算の結果を示す．$q = 10^{-2}$，$t = 0$，$s = 0.\,9$，$r_\alpha = r_\beta = 0.25$ とした．図 1 において，綺麗な Binder パラメータの交差が見え，臨界点が $T \sim 2$ に存在することがわかる．

3.2 $X \to XX$ (シンボルが複製される場合)

$X \to XX$ (シンボルが複製される場合) を考える．非終端シンボルの数は 2 つなので，2 つの過程があることに注意する:

$$A \to AA,\ B \to BB. \tag{9}$$

図 2 に数値計算の結果を示す．$q = 10^{-2}$，$t = 0$，$s = 0.9$，$r_\alpha = r_\beta = 0.25$ とした．$X \to YZ$ の場合 (図 1) と同様に，図 2 において，綺麗な Binder パラメータの交差が見え，臨界点が $T \sim 2$ に存在することがわかる．

3.3 q 依存性

式 (1a)，あるいは式 (1b) の規則が採択される確率 q を固定し，$X \to YZ$ の場合 (図 1) と $X \to XX$ の場合 (図 2) を比較すると，ほとんど見分けのつかない結果が得られた．続いて，これらの結果が q を変えたときにどのように変化するかをみる．図 3 に数

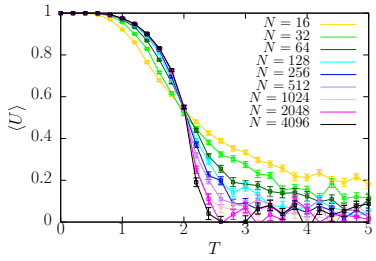

図 2　$X \to XX$ の場合の Binder パラメータの温度依存性．$q = 10^{-2}$，$t = 0$，$s = 0.9$，$r_\alpha = r_\beta = 0.25$ とした．シンボル数 N を 16，32，64，128，256，512，1024，2048，4096 と変えた．

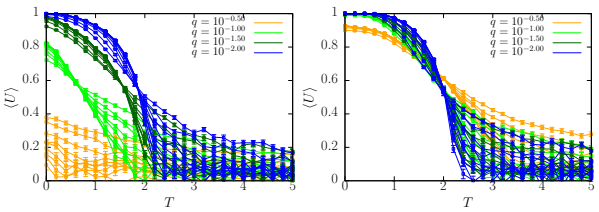

図 3　(左) $X \to YZ$，(右) $X \to XX$ の場合の Binder パラメータの q 依存性．$t = 0$，$s = 0.9$，$r_\alpha = r_\beta = 0.25$ とした．図 3 では，同じ q に対して異なるシステムサイズ (異なる N) を同じ色でプロットしている．$q = 10^{-0.5}$，$10^{-1.0}$，$10^{-1.5}$，$10^{-2.0}$ とした．

値計算の結果を示す．$t = 0$，$s = 0.9$，$r_\alpha = r_\beta = 0.25$ とした．図 3 では，同じ q に対して異なるシステムサイズ (異なる N) を同じ色でプロットしている．図 3 では，$X \to YZ$ の場合 (図 1) と $X \to XX$ の場合 (図 2) で，Binder パラメータの交差点の挙動が大きく異なることがわかる．図 3(左) に示されているように，$X \to YZ$ (等確率でシンボルが増える場合) は q を大きくすると相転移温度 (Binder パラメータの交差点) が下がっていることがわかる．さらに，q が十分大きいときに相転移点がなくなっていることがわかる．一方で，$X \to XX$ (シンボルが複製される場合) は大きく挙動が異なる．図 3(右) を見ると，q を大きくしても相転移温度 (Binder パラメータの交差点) が変化していないことがわかる．

4　議論

本研究では長距離 1 次元イジング模型に類似する CSG を構成し，連続的な相転移 (臨界点) が存在することを数値的に示した．本研究で発見された相転移はイジング模型と同様，$\mathbb{Z}_2$ 対称性の自発的破れであり，[7] 等で議論されている CSG における相転移とは別種の相転移である．しかし，相転移の存在が現状論争になっている CFG と比較して，CSG では明確に相転移を発生させることが比較的容易であることを本研究は示している．これは，我々が構成した CSG が CFG となる $q = 1$ 付近にて相転移が消失するという事実からも裏付けられる．この結果は言語モデルにおいて長距離相互作用が相転移を安定化させることを示唆している．

また，今回発見された相転移が明確に臨界的であったことも重要である．相転移が発生するような系を適当に構成した際には，一般的には臨界的でない不連続転移になることも多々あるためである．不連続転移においては非自明なスケーリング則が存在しないため，自然言語にみられる Zipf 則などの説明には適さない．本研究では Bidner パラメータが負に発散する傾向無く綺麗に交差することを確認したことで，不連続転移よりも非自明な臨界転移であることも明確にできた．

なお，通常の統計力学モデルには系自体が時間と共に増大する式 (1b) のような規則は存在せず，これらの規則は言語モデル特有のものである．そのため，本研究の相転移現象は単に統計力学の長距離 1 次元イジング模型の計算を自然言語処理の文脈でやり直しただけではない．言語に限らず，時間と共に空間的に拡大する系の統計力学を一般に考える際にも，本研究の結果は新しい洞察を与えうる．

また，自然言語処理の深層学習モデルにおいて，Attention 機構が重要であると考えられている [18]．Attention 機構は長距離相互作用を実現しており，本研究の研究の結果と整合するように思われる．

5　おわりに

本研究では，シンプルな相転移を示す言語モデルの研究を行った．さらに，言語モデル特有の規則に対して，非自明な相転移現象を示すことが明らかになった．今後，[5, 7] で議論されているような更に非自明な相転移が，今回のような CSG やその確率的拡張模型で実現されるかを考察する．また，より現実な言語モデルを考えることで，LLM で観測されているスケーリング則，創発的な言語能力のメカニズムを明らかにすることを目指す．なお，本研究で長距離相互作用が言語モデルの相転移に重要な役割を果たしていることが明らかになったが，Attention 機構などとの関係は明らかになっていない．Attention 機構の長距離極限における挙動を今後明らかにし，長距離相互作用，Attention 機構，相転移，スケーリング則，創発的言語能力の関係を明らかにする．

謝辞

本研究を行うにあたり，宮原英之は吉岡真治教授に有益なコメントをいただいた．本研究は JSPS 科研費 JP23H04489，JP22H05106，及び NSF 科研費 No. 2116246 の助成を受けた．

参考文献

[1] OpenAI. Chatgpt, 2023. https://chat.openai.com/chat.

[2] Google. Google bard, 2023. https://bard.google.com.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361**, 2020.

[4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. **arXiv preprint arXiv:2206.07682**, 2022.

[5] E. DeGiuli. Random language model. **Phys. Rev. Lett.**, Vol. 122, p. 128301, Mar 2019.

[6] Eric De Giuli. Emergence of order in random languages. **Journal of Physics A: Mathematical and Theoretical**, Vol. 52, No. 50, p. 504001, nov 2019.

[7] Kai Nakaishi and Koji Hukushima. Absence of phase transition in random language model. **Phys. Rev. Research**, Vol. 4, p. 023156, May 2022.

[8] Fatemeh Lalegani and Eric De Giuli. Robustness of the random language model. **arXiv preprint arXiv:2309.14913**, 2023.

[9] Freeman J Dyson. Existence of a phase-transition in a one-dimensional ising ferromagnet. 1969.

[10] Jürg Fröhlich and Thomas Spencer. The phase transition in the one-dimensional ising model with 1/r^2 interaction energy. 1982.

[11] Freeman J Dyson. Non-existence of spontaneous magnetization in a one-dimensional ising ferromagnet. 1969.

[12] Freeman J Dyson. An ising ferromagnet with discontinuous long-range order. **Communications in Mathematical Physics**, Vol. 21, pp. 269–283, 1971.

[13] D. J. Thouless. Long-range order in one-dimensional ising systems. **Phys. Rev.**, Vol. 187, pp. 732–733, Nov 1969.

[14] YunFeng Chang, Liang Sun, and Xu Cai. Phase transition of a one-dimensional ising model with distance-dependent connections. **Phys. Rev. E**, Vol. 76, p. 021101, Aug 2007.

[15] J.G. Martínez-Herrera, O.A. Rodríguez-López, and M.A. Solís. Critical temperature of one-dimensional ising model with long-range interaction revisited. **Physica A: Statistical Mechanics and its Applications**, Vol. 596, p. 127136, 2022.

[16] Yusuke Tomita. Monte carlo study of one-dimensional ising models with long-range interactions. **Journal of the Physical Society of Japan**, Vol. 78, No. 1, p. 014002, 2009.

[17] YunFeng Chang, Liang Sun, and Xu Cai. Phase transition of a one-dimensional ising model with distance-dependent connections. **Phys. Rev. E**, Vol. 76, p. 021101, Aug 2007.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

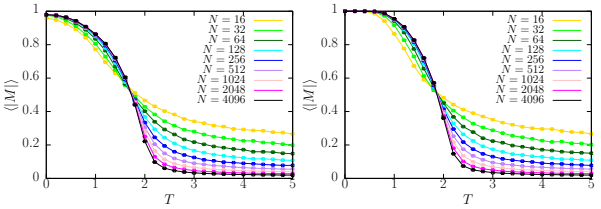

図 4　(左) $X \to YZ$ と (右) $X \to XX$ の場合の磁化の温度依存性．$q = 10^{-2}$，$t = 0$，$s = 0.9$，$r_\alpha = r_\beta = 0.25$ とした．シンボル数 N を 16，32，64，128，256，512，1024，2048，4096 と変えた．

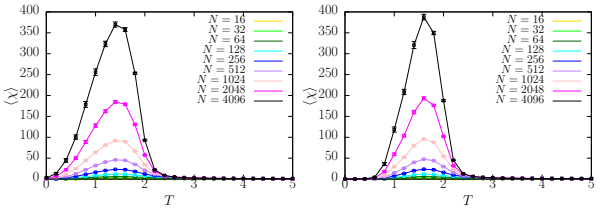

図 5　(左) $X \to YZ$ と (右) $X \to XX$ の場合の帯磁率の温度依存性．$q = 10^{-2}$，$t = 0$，$s = 0.9$，$r_\alpha = r_\beta = 0.25$ とした．シンボル数 N を 16，32，64，128，256，512，1024，2048，4096 と変えた．

A　参考情報

A.1　磁化の計算

章 3.1 と章 3.2 のセットアップの場合の帯磁率の計算を図 4 に示す．どちらの場合でも，転移点以下 $T \lesssim 2$ ではシステムサイズ N が大きくになるにつれて非ゼロの値に磁化が収束していき，転移点以上では 0 に収束していく様子が見られる．Binder パラメータの交差と併せ，臨界的な $\mathbb{Z}_2$ 対称性の破れと整合的である．

A.2　帯磁率の計算

章 3.1 と章 3.2 のセットアップの場合の帯磁率の計算を図 5 に示す．どちらの場合も相転移点近傍で発散的な振る舞いが見られ，相転移現象と整合する結果である．