言語モデルにおける共参照解決を担う注意ヘッドの分析

澤田紘希 [1]　佐々木睦史 [1]　坂口慶祐 [1,2]　Benjamin Heinzerling [2,1]
[1] 東北大学　[2] 理化学研究所
{sawada.hiroki.s4,sasaki.mutsumi.q5}@dc.tohoku.ac.jp
keisuke.sakaguchi@tohoku.ac.jp　benjamin.heinzerling@riken.jp

概要

共参照関係を捉えることは言語理解に必要不可欠だが，言語モデルの内部でどのように実現されているかは十分に解明されていない．既存研究では初期の Transformer モデル（BERT）において，文字列一致や簡単な先行詞選択を行う注意ヘッドが特定されているものの，分析は短い入力に限定されている．本研究では，現代の言語モデルにおける共参照解決メカニズムの解明を目指し，長文データセットを用いて注意ヘッドを分析した．実験の結果，ルールベースの性能を上回る複数のヘッドを特定し，それらが協調的に先行詞選択に寄与していることが示唆された．

1　はじめに

共参照解決は表層的なパターンマッチを超えた言語理解をどの程度獲得しているかを評価する上で重要であるが [1–3]，言語モデルが共参照関係をどのように解決するのかは明らかにされていない．

既存研究では，言語モデルの中で共参照解決に主に貢献しているのはマルチヘッド注意機構であり，共参照解決に寄与するヘッドは先行詞トークンへ強い注意を向けるという仮説のもと，注意ヘッドの分析が行われている．例えば，先行研究 [4] では初期の Transformer モデル（BERT）を分析し，文字列一致や簡単な先行詞選択を行う注意ヘッドを特定した．ただし，初期のモデルを対象としている点や，分析に用いるデータの入力長が短く近年の言語モデルの長文脈処理を反映していない点などの制約がある．また，注意ヘッドの先行詞選択性能はパターンマッチのルールベースによる性能を上回っていないため，そのヘッドが単なる文字列一致等を扱うヘッドであるか，表層的な理解を超えた共参照解決も行っているヘッドであるかは不明である．

本研究では，現代の言語モデルにおける共参照解決メカニズムの解明を目指す．その初歩として，共参照関係の一部である先行詞選択に着目し，平均約 1,592 トークンと長文である BookCoref データセット [5] を用いて，近年のデコーダ言語モデルがどのように先行詞選択を行うかをヘッドレベルで分析した．実験では，メンション間の注意スコアに着目し，各ヘッドのスコアと共参照関係の有無の相関を分析するとともに，各ヘッドを注意スコアによる先行詞選択モデル [6] として評価した．さらに，ルールベースの先行詞選択手法 [7] との比較を行うことで，ヘッドが表層的なパターンマッチを超えて共参照関係を真に捉えているかを検証した．

その結果，各ヘッドの注意スコアと共参照関係の有無に相関があることを確認した．また，中間層のヘッドにおいて，ルールベースによるベースラインの性能を大きく上回るヘッドが観察され，表層レベルを超えたより深い言語理解を実現している可能性が示された．定性分析では，高スコアの注意ヘッドが距離的に遠く，ルールベースでは解決できない共参照を解決している結果が得られた．さらに，特定の 1 つのヘッドではなく，中間層の複数のヘッドが協調的に先行詞選択に寄与していることを示唆する結果が得られた．

2　実験設定

この節では，本研究で使用するモデル・データセットについて示す．モデルはデコーダモデルの Qwen2.5-1.5B-Instruct [8] を用いた．これは，28 層・各層 12 注意ヘッドで，最大系列長が 32,768 トークンと長文の入力に対応している．また，データセットは，英語小説に共参照関係のアノテーションを加えた BookCoref [5] を用いた．テスト用データセットを分割した 152 サンプルのうち，エンティティ分布の偏りが大きいサンプルを除外し，130 サンプルを使用した．偏りが大きいサンプルでは，単純なパターンマッチで高スコアが得られやすく，ヘッドの

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

図 1　共参照・非共参照の例と文字列一致で解決できる共参照の例．本研究は注意スコアと共参照関係の有無に着目し，表層的な文字列一致等による先行詞選択性能を越える注意ヘッドが存在するかを分析する．

真の共参照解決能力を評価しにくいためである．具体的には，以下のサンプルを除外した．

- 最頻出エンティティのメンション割合が 60%を超える
- エンティティ分布のエントロピーが 0.6 未満

3　注意スコアと共参照の相関

3.1　手法

注意ヘッドが共参照関係を捉えているかを確認するために，各ヘッドの注意スコアと共参照関係の有無の相関を調べる．具体的には，最大注意スコアによってメンションペアを作成し，それらが共参照関係であるか否かを注意スコアと対応づけて相関係数を計算する．

一般に共参照関係にある 2 つのメンションのうち，文中で先に現れるものを先行詞，後に現れるものを照応詞と呼ぶ．また，照応詞と先行詞のつながりをリンクと呼ぶ．各照応詞から注意スコアが最大であるメンションを特定し，照応詞・先行詞候補のメンションペアを作成する．各メンションペアが共参照関係である場合を 1，ない場合を 0 とし，対応する注意スコアとの Spearman の順位相関係数を算出する．本研究ではヘッドの先行詞選択の性能を正確に測るために，リンク作成時に以下のルールを適用する．

- メンション以外への注意スコアを除外
- 二つのメンション間のスコアは，全トークンペアの中の注意スコアのうち，最大なものとする
- 入れ子構造を持つメンションが照応詞の場合は，その最後のトークンのスコアを使用する

なお，先行研究 [4] では全トークンを先行詞候補としているが，本研究ではアノテーションされたメンションのみを対象とする．これにより，ヘッドがメンション間の共参照関係をどの程度捉えているかを直接評価できる．

また，共参照関係の有無とは異なる，トークン距離による注意スコアへの影響を排除するため，スコアに $\log(d+1)$（d はトークン距離）を重みとして適用し，同様に相関係数を算出する．

3.2　結果

相関係数は最も高い (L18, H3) で 0.41 であり，トークン距離の重み付けを適用した場合も 0.42 と同等であった．これより，特定のヘッドの注意スコアと共参照関係の有無には弱い相関があり，共参照関係をある程度捉えていると考えられる．

4　注意ヘッドの性能評価

4.1　手法

先行詞選択を主に行う注意ヘッドを特定するために，各ヘッドで注意スコアによる先行詞選択を行い，ルールベースのベースラインと比較する．リンクの作成には，§ 3.1のルールに加えて，先行詞が存在するメンションのみを照応詞とするという条件を適用する．これは，先行詞が存在しない照応詞を含めると，ヘッドの先行詞選択性能を適切に評価できないためである．作成されたメンションペアが共参

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

図 2　各ヘッドとベースラインの CoNLL F1 スコアにおける差 (ベースラインの CoNLL F1 スコア:0.55)．CoNLL F1 スコアは共参照解決タスクに一般的に使用される指標であり，赤色が濃いほど性能の高さを表している．最もスコアが高い (L18, H3) について，以下で定性的な分析を行う．

照関係か否かを，F1 スコアで評価する．また，作成した全てのリンクに推移的閉包を適用することでクラスタを作成し，CoNLL F1 スコア [9] で評価する．CoNLL F1 スコアは MUC [10]，B^3 [11]，$CEAF_e$ [12] という 3 つの異なる観点から共参照を評価する指標の平均値であり，共参照解決研究において標準的に用いられている [2] [3]．

注意ヘッドが共参照関係を捉えているように見えても，単なるパターンマッチを行っているだけという可能性がある．そこで，ヘッドが共参照関係を捉えていることを確認するために，ルールベースによる先行詞選択手法をベースラインとし，各ヘッドの性能と比較する．ベースラインを上回る性能のヘッドは，ルールベース以上の共参照解決を行っていることになり，真に共参照解決を捉えていると考えられる．具体的には，先行研究 [7] のルールを簡略化した手法を用いて，照応詞候補に対して，

1. 文字列の完全一致
2. メンション末尾単語の一致
3. 代名詞に限り，辞書による数・性の一致

を判定し，いずれかが成立する最近傍のメンションを先行詞とする．また，注意ヘッドによる先行詞選択と同様に，全てのリンクに推移的閉包を適用してクラスタを作成する．性能評価も同様に，共参照関係のメンションペアを F1 スコアで，作成されたクラスタを CoNLL F1 スコアで評価する．各層・各ヘッドにおける，CoNLL F1 スコアのベースラインとの差を図 2に示す．

4.2　結果と考察

図 2より，ヘッドごとに CoNLL F1 スコアの差があり，ベースラインを上回るヘッドが複数存在することが確認できる．最もスコアが高いヘッドは (L18, H3) で，ベースラインの CoNLL F1 スコアを 0.22 上回っている．また，高性能ヘッドは L11〜L18 の中間層に集中していた．メンションペア F1 スコアでも図 2と同様の傾向が見られ，複数のヘッドがベースラインを上回った（§ A）．高性能ヘッドは両指標で一致しており，表層的なパターンマッチを超えた共参照解決能力を持つことを示している．複数のヘッドがベースラインを上回ることから，単一のヘッドではなく，複数のヘッドが協調的に先行詞選択を実現していることが示唆される．

5　定性的な分析

5.1　手法

性能が高いヘッドがどのように注意を向けているのかを確認するために，メンション間の注意スコアを可視化する．§ 4.1と同様に，各ヘッドの注意スコアをもとにメンションペアを作成する．各メンションペアのリンクが共参照関係である場合は青色，そうでない場合は赤色の矢印で表す．また，注意スコアの大きさを矢印の太さで表し，メンションを各エンティティで色分けしている．図 2を参考に，リンクの F1 スコアが最大である層 18 のヘッド 3 を可視

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

図 3 (L18, H3) の注意スコアによるリンクの可視化．青矢印は共参照関係，赤矢印は非共参照関係を示す．メンションはエンティティごとに色分けしている．

化したものを図 3に示す．

5.2 結果と考察

図 3より，正しい先行詞へのリンクが多い一方で，間違った先行詞へのリンクも存在していることがわかる．これは，(L18, H3) の精度の高さを表すと同時に，単体のヘッドでは完全な先行詞選択を達成できないことを示す．また，長距離でルールベース手法では解決困難な共参照関係を捉えていることも確認できる．さらに，常に直前の先行詞を選択する，あるいは必ずそのエンティティが最初に出現したメンションに注意を向けるといった，単純な傾向は見られなかった．したがって，先行詞選択に特化したヘッドは，文脈を理解することで複雑なルールを形成していると考えられる．

6 複数ヘッドの組み合わせ

6.1 手法

§ 4.2，§ 5.2より，モデルは単一のヘッドでではなく，複数のヘッドで先行詞選択を行なっていることが示唆された．そこで本節では，複数の注意ヘッドが協調的に先行詞選択を担っているという仮説を検証する．具体的には，上位ヘッドの組み合わせが正解の先行詞をどの程度カバーできるかを分析する．

具体的には図 2を参考に，スコアが上位の 5 つのヘッド (L18, H3)，(L17, H6)，(L13, H2)，(L13, H3)，(L11, H11) を用いて，いずれかのヘッドが正しい先行詞へ最大の注意を向けている場合，正解とする．複数のヘッドが正しい先行詞を指している場合は，最も注意スコアの高いリンクを採用する．いずれのヘッドも正しい先行詞を指していない場合は，5 つのヘッド全体で最大の注意スコアを示すリンクを予測結果とする．

6.2 結果と考察

複数のヘッドによる先行詞選択の結果，メンションペアの F1 スコアは 0.95，CoNLL F1 スコアは 0.88 となり，高精度な先行詞選択が可能であることが確認された．本分析はオラクル設定，すなわち各照応詞に対して正解の先行詞がどのメンションかを事前に知った上で行った評価である．したがって，実際の共参照解決タスクにおける性能を直接示すものではない．しかし，高いスコアが得られたことは，これら 5 つのヘッドが先行詞選択に必要な情報をほぼ網羅しており，モデルの共参照解決能力の中核を成していることを示唆する．

7 おわりに

本研究は，言語モデルが共参照解決をどう行うかを，注意ヘッドに注目して分析した．メンション間の注意スコアに着目した評価の結果，注意スコアと共参照関係の有無に相関があること，先行詞選択性能においてルールベースのベースラインを上回る複数のヘッドが存在することを確認した．また，定性分析により，高性能ヘッドが代名詞と先行詞の対応関係を捉えている様子を可視化し，長距離の共参照関係を捉えていることを示した．さらに，それらのヘッドが協調的に先行詞選択に寄与していることが示唆された．今後は，介入実験によりこれらのヘッドの因果的寄与を検証する予定である．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JSPS 科研費 JP24K03236，JP25K03175，及び JST BOOST JPMJBY24F9 の助成を受けたものです.

参考文献

[1] Yujian Gan, Massimo Poesio, and Juntao Yu. Assessing the capabilities of large language models in coreference: An evaluation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 1645–1665, May 2024.

[2] Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and analysis. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 5803–5808, November 2019.

[3] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. **arXiv preprint 1907.10529**, 2020.

[4] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286, August 2019.

[5] Giuliano Martinelli, Tommaso Bonomo, Pere-Lluís Huguet Cabot, and Roberto Navigli. BOOKCOREF: Coreference resolution at book scale. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 24526–24544, July 2025.

[6] Sebastian Martschat and Michael Strube. Latent structures for coreference resolution. **Transactions of the Association for Computational Linguistics**, Vol. 3, pp. 405–418, 2015.

[7] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**, pp. 492–501, October 2010.

[8] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. **arXiv preprint 2412.15115**, 2025.

[9] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In **Joint Conference on EMNLP and CoNLL - Shared Task**, pp. 1–40, July 2012.

[10] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In **Sixth Message Understanding Conference (MUC-6)**, 1995.

[11] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In **The first international conference on language resources and evaluation workshop on linguistics coreference**, Vol. 1, pp. 563–566, 1998.

[12] Xiaoqiang Luo. On coreference resolution performance metrics. In **Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing**, pp. 25–32, October 2005.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

図 4　各ヘッドとベースラインのメンションペア F1 スコアにおける差 (ベースラインのメンションペア F1 スコア:0.71)

A　各ヘッドのメンションペア F1 スコア

§ 4で分析した，各ヘッドとベースラインのメンションペア F1 スコアの差を図 4に示す．図 2の CoNLL F1 スコアと比較すると，両指標において高いスコアを示すヘッドが共通していることが確認できる．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).