事前学習済み言語モデルによるエンティティの概念化

坂田将樹 [1,2]　横井祥 [1,2]　Benjamin Heinzerling[2,1]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
sakata.masaki.s5@dc.tohoku.ac.jp
benjamin.heinzerling@riken.jp,　{yokoi, kentaro.inui}@tohoku.ac.jp

概要

事前学習済みマスク言語モデルは，事実知識に関する穴埋め問題に正答できるなど，エンティティを含むテキストをある程度うまく処理することができる．果たしてこの挙動は言語モデルがエンティティを人間同様に「知っている」証拠となりうるのだろうか？本研究では，言語モデルがエンティティを概念化できているか，すなわち異なる文脈や表層をもって言及される同一のエンティティを同一の事物として認識できているかどうかについて，その内部表現が十分に密で他と分離されたクラスタを成しているかという視点で検証を行った．BERT を対象とした実験の結果，約 7 割のエンティティについて，その埋込表現が他の概念と完全に分離したクラスタを形成していること（すなわち内部表現の意味での概念化に成功していること）が確認できた．[1)]

1　はじめに

近年登場した事前学習済みマスク言語モデル [1, 2] は，実世界の事物である**エンティティ**を含むテキストをうまく処理できているように見える．例えば BERT [1] は “*Hillary Clinton was born in [MASK].*” という入力に対して，“*Chicago*” という実世界の事実と整合した出力を返すことができる [3, 4].

しかし，このように言語モデルがエンティティを含む個別のテキストをある程度の精度で「処理できる」という事実は，モデルがエンティティを人間同様に「知っている」証拠となりうるのだろうか．本稿では，**マスク言語モデルがどの程度エンティティについて「知って」いるか**についてある程度明らかにすることを試みる．もしモデルのエンティティに対する理解を正確に把握する手段が手に入れば，モデルの解釈性が上がり，さらに高品質な言語モデルの実現に繋がり得るだろう．

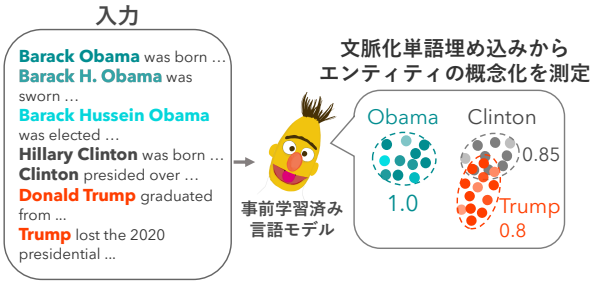

図 1　エンティティの概念化の測定の概要

我々は検証の第一歩として**概念化**という考え方を採用する．エンティティを「知っている」人間は，同一のエンティティが異なる周辺文脈で言及されたときでも，あるいは同一のエンティティを指し示すために異なるメンション（文字列）が用いられたときでも，それらを同一のものだと正しく認識できる．例えば “*Barack Obama was born in Hawaii.*” と “*Barack Hussein Obama was elected...*” に登場する人名を同じ人物（同じ概念）であると認識できる．すなわちエンティティを概念化できている．はたしてマスク言語モデルは人間同様の概念化をおこなえているのだろうか？この問いに答えるため，本研究では，**同一のエンティティに対する内部表現が十分に密な（他と分離された）クラスタを成しているか**という視点での検証をおこなう．言い換えれば，マスク言語モデルがエンティティをどの程度概念化しているかという問いに，埋め込み空間の配置を通じて答えることを試みる．本論文の貢献をまとめると以下の通り：

- 事前学習済みマスク言語モデルがエンティティを概念化できているかについて，同一の事物に対応するエンティティが埋め込み空間で他の埋め込みと混ざり合っていないかという観点で定量的に評価する方法を提案した．
- BERT を対象とした検証の結果，エンティティ

1) ソースコードは https://github.com/cl-tohoku/Geo-Ent-in-LMs にて公開している．

の周辺文脈やメンションが多様である場合でも，約 7 割のエンティティは他の概念と区別できていたことがわかった．

2 関連研究

言語モデルの内部表現の分析手法 言語モデルの内部表現に言語的特徴が反映されているかを分析している先行研究として，[5] と [6] が挙げられる．[5] は，普通名詞や動詞の語義によって文脈化単語埋め込みの位置が異なることを可視化によって定性的に示した．[6] は，言語の持つ特徴が単語埋め込み空間に反映されているかを明らかにするために，教師なしクラスタリングを用いた分析手法を提案した．

我々は [5] と [6] とは異なり，新たに**エンティティ**の情報がモデルの内部表現に反映されているかを探る．また，[6] の手法は生成されたクラスタが分離可能か不可能かの 2 値しかわからないため，各クラスタがどの程度混ざっているのかは測定できない．したがって，本研究ではエンティティが埋め込み空間で他の埋め込みと混ざっているか，もしくは分離しているかについて連続評価を行う．

言語モデルにエンコードされている情報の分析手法 言語モデルにエンコードされている情報を分析する手法には，内部表現の分析の他にプロンプトを用いる手法がある [3, 4]．例えば，“*Paris is the capital of [MASK].*” というプロンプトに対して “*France*” が出力されるかによって，言語モデルが持つ事実知識を測定している．ただし，プロンプトから得られる評価は信頼性に欠ける可能性が指摘されている．具体的には，プロンプトが意味的に同じであっても，予測結果が異なる点 [7] や，各モデルによって，正解しやすいプロンプトの選好が存在する点 [8] が報告されている．よって，本研究ではプロンプトを用いずにモデルの内部表現を直接分析する．

また，内部表現を分析する手法は言語モデルが「知って」いることは測れていても，「使っている」ことを正しく測れているかはわからない点が指摘されている [9]．そこで，[9] は「知って」いることと「使って」いることを切り分けて分析する立場をとっている．我々も [9] と同じ立場をとる．つまり，言語モデルがエンティティに関する情報を「知って」いることと，それを認識した上で情報抽出や単語生成時に「使って」いるかを切り分けて分析する．本研究では，まず「知って」いることを調査する．

3 エンティティの概念化の測定方法

3.1 クラスタ間の分離度合いを測定

本研究の目的は，**言語モデルがエンティティを概念化できているか**を定量的に確かめることである．例えば，“*Barack Obama was born in Hawaii.*” と “*Barack Hussein Obama was elected...* ” の 2 つの文章が与えられた場合，登場する人名を同じ人物 (同じ概念) であると認識できれば，その人物について概念化しているといえる．

概念化ができているかについては，言語モデルの文脈化単語埋め込みから形成されるクラスタの分離度合いを測定することで判断する．多義語の文脈化単語埋め込みはその語義毎に埋め込み空間上で偏在することが知られている [5]．言い換えれば，同じ意味で用いられているトークン集合の文脈化単語埋め込みは空間上で凝集する．もし，言語モデルがエンティティを他の概念と混同しているならば，エンティティの文脈化単語埋め込みは他の概念と混ざりあっていることが予想される．逆に，エンティティと他の概念を区別できているのであれば，概念の違いによって分離していることが予想される．このことを踏まえて，文脈化単語埋め込みを用いて以下を測定する．

- エンティティの各クラスタと，他の概念のクラスタとの分離度合い
- エンティティ以外の各クラスタと，他の概念のクラスタとの分離度合い

ここでのクラスタとは，センテンス中に現れる “*Barack Obama*” や “*Barack Hussein Obama* ” の文脈化単語埋め込みを 1 つにまとめた集合を指す．

もし，エンティティの大多数のクラスタがエンティティ以外のクラスタと比べてうまく分離しているのであれば，言語モデルはエンティティの概念をよりよく区別できているといえる．

凝縮率 クラスタ間の分離度合いを測定するために，各文脈化単語埋め込みの最近傍のクラスタの中心が自クラスタのクラスタ中心である割合 (以下**凝縮率**と呼ぶ) を算出する．

エンティティの集合を $\mathscr{E}$，エンティティ $e \in \mathscr{E}$ に対応するベクトル集合を $\boldsymbol{X}_e := \{\boldsymbol{x}_e^1, \boldsymbol{x}_e^2, \cdots\}$，その重心を $\boldsymbol{b}_e := \frac{1}{|\boldsymbol{X}_e|} \sum_{\boldsymbol{x}_e \in \boldsymbol{X}_e} \boldsymbol{x}_e$, さまざまなエンティティの重心を集めた集合を $\boldsymbol{B} := \{\boldsymbol{b}_e \mid e \in \mathscr{E}\}$ とおく．エ

表 1　検証用データセットの詳細

		クラスタ数	センテンス数
固有表現	人名	280	8,661
	地名	87	8,160
固有表現以外		7,791	946,586

表 2　クラスタが凝縮率=100%である割合 (%)

	異なる部分	
	周辺文脈	周辺文脈と表層
エンティティ	**82.83**	**69.48**
エンティティ以外	28.85	28.85

ンティティ e の凝縮率，すなわち $\boldsymbol{X}_e$ に含まれるベクトルたちが一箇所に固まっており他のエンティティのベクトルたちと十分に分離している度合い $P(e)$ を以下で定義する．

$$P(e) \coloneqq \frac{1}{|\boldsymbol{X}_e|} \sum_{\boldsymbol{x}_e \in \boldsymbol{X}_e} \mathbb{1}\left[\arg\min_{\boldsymbol{b} \in \boldsymbol{B}} d(\boldsymbol{x}_e, \boldsymbol{b}) = \boldsymbol{b}_e\right] \quad (1)$$

$d(\boldsymbol{x}, \boldsymbol{y})$ は距離関数であり，検証ではユークリッド距離を使用した．

この凝縮率は，あるクラスタに所属している割合を算出するという意味で，クラスタリングの良さを測る尺度である純度 （Purity） と非常によく似ている．

もし，あるエンティティの凝縮率が 100%であれば，そのエンティティは他のクラスタと分離しており，他の概念と区別ができているといえる (図 1, “*Obama*”)．逆に凝縮率が低い場合，他の概念と混ざり合っているといえる (図 1, “*Clinton*” と “*Trump*”).

4　検証

言語モデルが各エンティティの概念を区別できているかについて，文脈化単語埋め込みの分離度合いを測定することで確かめる．具体的に以下の 2 点について検証する．

1. 言語モデルは**周辺文脈が異なるエンティティ同士**を “同じもの” であるとわかるか？
2. 言語モデルは**周辺文脈とメンションが異なるエンティティ同士**を “同じもの” であるとわかるか？

1 点目では，例えば “*Obama*” は，大統領に就任する文やゴルフをしている文など，多様な文脈に登場する．この “*Obama*” を言語モデルは同じ人物だと認識しているかを検証する．2 点目では，例えば，エンティティのメンションが “*Obama*” である文と，“*Barack H. Obama*” である文に対して，言語モデルは同じ人物だと認識しているかについて検証する．

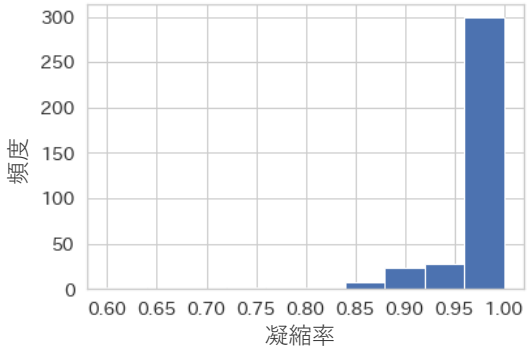

図 2　エンティティの周辺文脈とメンションが異なる場合のエンティティのクラスタの凝縮率の頻度

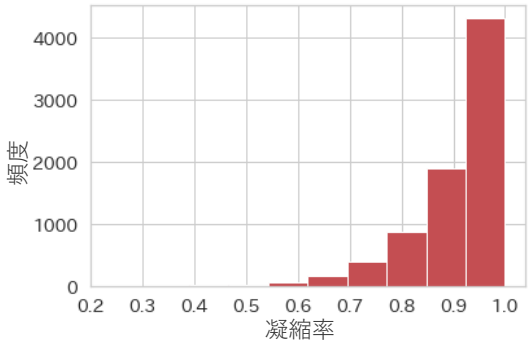

図 3　エンティティの周辺文脈とメンションが異なる場合のエンティティ以外のクラスタの凝縮率の頻度

言語モデルが各エンティティに対して概念の区別をしているかどうかは，3 節で述べたクラスタ間の分離度合いを測定し，エンティティ以外のクラスタよりも凝縮率が高いかどうかで判断する．

4.1　設定

モデル　事前学習済み BERT-base (uncased) を使用した．取得する文脈化単語埋め込みは BERT 最終層の隠れ状態ベクトルを使用した．

使用データ　wikilinks[10] データセットを使用した．このデータセットには，Wikipedia に存在する 300 万エンティティと，その周辺文脈が収録されている．検証では，上記データセットから約 97 万センテンスをランダムに抽出した．抽出条件として，1 クラスタ内の点群は必ず 10 個以上とした．また，検証に使用するエンティティは人名と地名とした．「1. 言語モデルは**周辺文脈が異なるエンティティ同士**を “同じもの” であるとわかるか？」を検証する際には，エンティティのメンションの揺れを Wikipedia 上のタイトルに統一した．入力時には，センテンスの文頭に [CLS] トークン，文末に [SEP] トークンを挿入した．検証用データの詳細は表 1 に示した．

表 3　メンションが他のエンティティクラスタと混ざる事例．括弧内は単語のカテゴリを表す

メンション	混ざっているクラスタ
The Breeders (musician)	The Irish Rovers (musician), The Jackson Five (musician)
San Clemente (location)	San Diego (location)
Roger McGuinn (musician)	The Irish Rovers (musician), John McGahern (author)

表 4　メンションの変化によって他クラスタと混ざった事例．括弧内は単語のカテゴリを表す

メンションの変化前	メンションの変化後	混ざったクラスタ
Stone Phillips (actor)	Stone	stone (普通名詞)
Tom clancy (author)	Mr. Clancy	Dr. Oz (doctor)
Noam Chomsky (author)	Professor Noam Chomsky	The Irish Rovers (musician)
Millau Viaduct (location)	the Millau Viaduct	The Irish Rovers (musician) , the

4.2　検証結果と考察

各単語の分離度合いを検証した結果を表 2 に示す．

4.2.1　エンティティの周辺文脈のみが異なる場合

エンティティの周辺文脈が多様であっても，BERT はエンティティを区別できる：各単語の文脈化単語埋め込みの分離度合いを測定したところ，エンティティのクラスタの約 83%が，凝縮率 100%であり，他のクラスタと分離できていた．対して，エンティティ以外のクラスタで凝縮率 100%であるものは約 29%であった．

エンティティとエンティティ以外の凝縮率を比較すると，エンティティのほうが他の概念と分離できていることが確認できる．したがって，エンティティの周辺文脈が多様であっても，BERT はエンティティ以外の単語と比べて他の概念と区別できるといえる．

4.2.2　エンティティの周辺文脈とメンションが異なる場合

エンティティの周辺文脈やメンションが多様であっても，BERT はエンティティを区別できる（図 2, 3）：各単語の文脈化単語埋め込みの分離度合いを測定したところ，エンティティのクラスタの約 70%が，凝縮率 100%であった．対して，エンティティ以外のクラスタで凝縮率 100%であるものは約 29%であった．

エンティティとエンティティ以外の凝縮率を比較すると，エンティティのほうが他の概念と分離できていることが確認できた．したがって，エンティティの周辺文脈とメンションが多様であっても，BERT はエンティティ以外の単語と比べて他の概念と区別できるといえる．

しかし，「周辺文脈のみが異なる場合」と比較すると，エンティティが凝縮率 100%である事例は約 13%減少する．すなわち，メンションが異なることで，他のクラスタと混ざるようになり，各エンティティの区別が若干難しくなることがわかる．

実際に，エンティティが他のクラスタと混ざってしまう事例を表 3, 表 4 に示した．これらの事例を観察すると，同じカテゴリ（人名や地名）同士で混ざっているように見える．実際に集計すると，人名と地名の両者において同じカテゴリ同士で混ざる比率は，チャンスレベルの比率より高いことがわかった（Appendix A, 表 5）．よって，BERT は少なくともオントロジーレベルの概念化はできていることがわかる．しかしながら，“Stone (actor)” と “stone (noun)” が混ざり合っているように，BERT のエンティティの概念化は完璧とは言えないことも同時にわかる．

5　おわりに

本研究では，マスク言語モデルがエンティティをどのような形で「知って」いるかについて検証を行った．検証では，文脈化単語埋め込みから形成されるクラスタの分離度合いを分析することで，エンティティの概念化を測定した．BERT を対象とした実験の結果，約 7 割のエンティティについて，他の概念と分離したクラスタを形成していたことが確認できた．すなわち内部表現の意味での概念化には成功していることがわかった．

ただし，本研究の検証では言語モデルがエンティティの概念化を情報抽出や単語生成時に「使って」いるかはわからない．今後は言語モデルがタスクを解く上で，エンティティの概念化を実際に使っているかを検証する．これにより，言語モデルの解釈性を向上することが可能となる．

謝辞

本研究は JST CREST JPMJCR20D2，JST ACT-X JPMJAX200S，JSPS 科研費 22H05106，22H03654，21K17814 の助成を受けたものです．また，本研究の実装に関して東北大学の穀田一真氏に多くのご助言を頂きました．感謝致します．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training 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, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[3] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 423–438, 2020.

[5] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[6] Yichu Zhou and Vivek Srikumar. DirectProbe: Studying representations without classifiers. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5070–5083, Online, June 2021. Association for Computational Linguistics.

[7] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1012–1031, 2021.

[8] Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5796–5808, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. Probing for the usage of grammatical number. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8818–8831, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. 2012.

A Appendix

エンティティの各単語埋め込みが最近傍クラスタと混ざっている場合，同じカテゴリ同士で混ざるのか，もしくは違うカテゴリと混ざるのかについて検証した．得られた結果は表 5 に示した．チャンスレベルの比率と比較すると，同じカテゴリ同士で混ざる比率がより多いことがわかる．ここでのカテゴリは，人名，地名，エンティティ以外の 3 種類である．

表 5 最近傍クラスタが異なるクラスタである場合の各カテゴリの比率

カテゴリ	同じカテゴリ：異なるカテゴリ	チャンスレベルの比率
人名	1 : 0.991	1 : 28.136
地名	1： 7.12	1 : 92.77