対義関係バイアス: 事前訓練済み言語モデルと人間の意味関係間の弁別能力に関する分析

Cao Zhihan
東京工業大学 情報理工学院
cao.z.ab@m.titech.ac.jp

山田寛章
東京工業大学 情報理工学院
yamada@c.titech.ac.jp

徳永健伸
東京工業大学 情報理工学院
take@c.titech.ac.jp

概要

意味関係の弁別は，人間にとっても，機械にとっても，容易なタスクではない．本研究では，多様な下流タスクで卓越した性能を示した事前訓練済み言語モデルが意味関係の弁別ができているか否かという問いに，混淆度という尺度を提案し，人間との比較の上でアプローチした．結果として，事前訓練済み言語モデルは，意味関係の弁別能力は，人間に比べて下回ることと同時に，非対義関係を対義関係として誤認識するバイアスが観察された．

1 イントロダクション

意味関係とは，二語が成す意味的な関係のことである．対義関係と類義関係は意味関係の代表例である．それ以外に，例えば，*bird* と *robin* のような上位・下位関係を結ぶ 2 語は，前者の意味が後者の意味を抽象している．この時，*bird* は *robin* に対する上位語であり，*robin* は *bird* に対する下位関係にある．同様に，*furniture* と *sofa* は全体と部分の関係にある．全体・部分関係を結ぶ二語は，前者の意味が後者の意味を一部として包含する．この場合，*furniture* は *sofa* に対する全体語であり，一方で *sofa* は *furniture* に対する部分語である．

異なる意味関係の間の境界は，常に明確だとは限らない．類義関係と上位・下位関係が人間にとって類似していることは，心理言語学の実験で観察されている [1, 2]．上位・下位関係と全体・部分関係の境界は，英語の集合名詞において不明瞭だと理論的に論じられている [3]．例えば，*furniture* は家具の集合を指すが，その部分語の *sofa* は具体的な家具を指すため，下位語としても考えられる．また，計算言語学では，類義関係をなす二語と対義関係を成す二語は，統合的に（syntagmatically）類似しているため，共起特徴量を用いて両関係を区別することが困難だと知られている [4]．

つまり，意味関係の弁別は人間にとっても，計算機にとっても簡単なタスクではないことが示唆されている．そこで，多様な下流タスクで優秀な性能を示しているとされる事前訓練済み言語モデル（Pretrained Language Model，PLM）は，意味関係の弁別ができているか．本研究は，混淆度という新たな尺度を提案し，この問題にアプローチしていく．

2 提案尺度

本研究では，語義関係の知識の評価に多用されているプロンプトに基づく探査（Prompt-based Probing）[5, 6, 7] という手法を用いる．この手法では，プロンプトによる探査子（Probe）を構築して，PLM が特定の知識を獲得しているかどうかを検証する．

本研究では，“*a W is a kind of V*” のように，W と V の単語が入るプレースホルダーを 2 個持つ文字列のことをプロンプトと定義する．本研究のプロンプトは W と V の間に必ず 1 つのみの意味関係が成立するよう設計した．W をあるターゲット単語 w_i で埋め，V は空欄のままのプロンプトのことを探査子と呼ぶ．探査子をモデルに入力すると，モデルは，V を埋めて文を完成させる．この時，文を完成させるタスクは，w_i とプロンプトが表す意味関係 r を持つ単語を予測させるタスクと同義である．以下，w_i と意味関係 r を持つ単語のことを略して w_i の r-関係語と呼ぶ．w_i が持つ r-関係語の集合を y_i^r とする．

i 番目のターゲット単語 w_i と，意味関係 r を意味する j 番目のプロンプト p_j^r によって構成される探査子を x_{ij}^r とする．モデル m は x_{ij}^r を入力として受け取った際に，探査子の V を補完できる単語の確率分布を出力する．$l_k(x_{ij}^r; m)$ を，予測確率が上位 k 個によって構成され，かつその予測確率によって降順にソート済みのリストとする．

もし，モデルが x_{ij}^r が表す意味関係 r を正確に認識していれば，y_i^r に属する単語は $l_k(x_{ij}^r; m)$ 中で高順位に位置する．一方で，もしモデルが x_{ij}^r が表す意味関係 r を別の関係 s に間違えていれば，$l_k(x_{ij}^r; m)$ 中で，y_i^s に属する単語が高順位に位置する．

以上を踏まえて，モデルが 2 つの関係を混淆する度合いを評価する混淆度を定義する．モデル m が探査子 x_{ij}^r を受け取った際の出力 $l_k(x_{ij}^r; m)$ における，ターゲット単語 w_i の s-関係語の平均順位スコア $\alpha(s, x_{ij}^r; m)$ を，以下のように定義する．

$$\begin{aligned} \alpha(s, x_{ij}^r; m) &= \frac{1}{|y_i^s|} \sum_{w \in y_i^s} \mathrm{score}(w, l_k(x_{ij}^r; m)) \\ \mathrm{score}(w, l) &= \frac{|l| - w \text{ の } l \text{ における順位} + 1}{|l| + 1} \end{aligned} \tag{1}$$

ただし，$|l|$ はリスト l の長さとする．したがって，$l_k(x_{ij}^r; m)$ を考える場合，それは k である．k の設定については 3 節で説明する．

$\alpha(s, x_{ij}^r; m)$ は関係 r を表す各探査子に対して定義されている．それを関係 r について集計した $\alpha(s, r; m)$ を以下のように定義する．

$$\alpha(s, r; m) = \frac{1}{\#x_{ij}^r} \sum_{i,j} \alpha(s, x_{ij}^r; m) \tag{2}$$

$\alpha(s, r; m)$ は直観的には，関係 r を表す探査子が入力された時にモデルが関係 s として認識している度合いを表している．したがって，$\alpha(s, r; m)$ は，$s = r$ の時に高く，$s \neq r$ の時に低いことが望ましい．

$\alpha(s, r; m)$ を $\alpha(r, r; m)$ を用いて 0 から 1 までに以下のように正規化し，それを r から s への混淆度（Confusability）として定義する．

$$\mathrm{Confusability}(s, r; m) = \min\left(\frac{\alpha(s, r; m)}{\alpha(r, r; m)}, 1\right) \tag{3}$$

$\frac{\alpha(s,r;m)}{\alpha(r,r;m)}$ は 1 より大きくなる場合があるため，閾値処理を行う．

$\mathrm{Confusability}(s, r; m)$ は 1 に近いほど，モデル m が r を s として混淆する度合いが高いことを意味している．一方で 0 に近いほど，r を s として混淆する度合いが低いことを意味している．したがって，各モデルは，$[0, 1]^{|R| \times |R|}$ の混淆度行列で特徴付けられる．ここで，R は意味関係の集合を表す．混淆度行列の対角線は定義上 1 であるために本研究では考慮しない．なお，混淆度は必ずしも対称ではなく，$\mathrm{Confusability}(s, r; m) = \mathrm{Confusability}(r, s; m)$ は必ずしも満たされない．そのため，混淆度行列も必ずしも対称行列ではない．

3 実験

3.1 対象モデル

本研究では，PLM と人間との比較を行う．特に，PLM のうち，Masked Language Model (MLM) と Causal Language Model (CLM) を考える．具体的には，前者は BERT [8]，ALBERT [9] と RoBERTa [10] の全てのサイズを，後者は OPT [11] [1] の全てのサイズを評価に含める．

3.2 対象意味関係と評価データ

本研究では 6 つの意味関係を考える．それぞれは上位関係（hypernymy, HYP），下位関係（hyponymy, HPO），全体関係（holonymy, HOL），部分関係（meronymy, MER），対義関係（antonymy, ANT），類義関係（synonymy, SYN）である．

プロンプト　プロンプトに基づく探査とパターンに基づく関係語ペアの自動抽出に関する先行研究を参照し，人手でプロンプトを設計した [13, 14, 15, 7]．その結果，上位関係，下位関係，全体関係と部分関係について 7 つ，部分関係については 6 つ，対義関係については 9 つのプロンプトを得た[2]．MLM と CLM 間でプロンプトの互換性を保つために，全てのプロンプトで補完対象のトークンが末尾に位置するに設計した．

ターゲット単語　それぞれの関係について，ターゲット単語は Hyperlex [2] と Overschelde らが構築したカテゴリーノルムのコーパス [16] を活用して決定した．Hyperlex は意味関係が付与されている単語ペアのコーパスであり，カテゴリーノルムのコーパスはある名詞（ノルム）とそのインスタンスによるコーパスである．Hyperlex とカテゴリーノルムの

1） BERT, ALBERT, RoBERTa と OPT の語彙の積集合のサイズとそれらの和集合のサイズとの比は 0.19 であるのに対し，BERT, ALBERT, RoBERTa と Llama 2 [12] のその比は 0.08 であったため，比較可能性を考慮して OPT を選定した．

2） 全てのプロンプトは https://github.com/hancules/Responses-to-probes にて公開している．

表 1　各意味関係の統計．r-関係語集合サイズの表記は平均 ± 標準偏差．

意味関係	# ターゲット単語	r-関係語集合サイズ	# プロンプト	# 探査子
上位関係 (HYP)	718	9.0 ± 7.1	7	5,026
下位関係 (HPO)	319	20.7 ± 43.5	7	2,233
全体関係 (HOL)	195	2.6 ± 2.1	7	1,365
部分関係 (MER)	146	3.0 ± 4.4	6	876
対義関係 (ANT)	105	1.1 ± 0.4	9	945
類義関係 (SYN)	218	3.0 ± 2.7	7	1,526

コーパスにある単語で，1) 名詞であり，2) 対象モデルの共通語彙に存在し, 3) いずれの対象モデルのトークナイザーにもサブワードに分割されない単語を，ターゲット単語とする．

関係語集合　各ターゲット単語とある意味関係 r について，WordNet [17]，[2] と [16] におけるターゲット単語の r-関係語を検索し，その結果を該当ターゲット単語の r-関係語集合とする．WordNet で上位語と下位語を検索する際は，ターゲット単語との距離が 2 以内の単語のみ含める．

表 1 に得られた評価データの統計を示す．例えば，上位関係については，718 個のターゲットの単語がある．それらが持つ上位語の集合の平均サイズは 9 語である．上位関係を意味するプロンプトが 7 つあるため，1 つのターゲット単語について，7 つの探査子が得られる．その結果，トータルで $7 \times 718 = 5026$ 個の探査子が得られる．

3.3　人間による回答の収集

人間の各探査子に対する回答は Amazon Mechanical Turk（Mturk）で収集した．参加者は，Mturk Master の資格を持ち，その回答が 500 回以上承認され，承認率が 95 %を超えており，さらに現在米国，英国，オーストラリア，カナダのいずれかに居住している者に限定した．1 つの探査子に対して，4 人分の回答を募集した．その結果，48 人の参加者から、1 人当たり平均 22 件の回答を得た．ある探査子に対する回答に出現した単語の頻度を集計し，その頻度に応じて降順に単語を並べ替えたリストをその探査子の人間による回答とする．各探査子に対する人間の回答の長さを式 (1) の k とする．このようにすることで，ランクに基づく提案尺度が計算可能になり，人間の混淆度行列が得られるようになる．

4　結果

図 1 に，MLM，CLM と人間（Human）の混淆度行列を示す．MLM の行列は，BERT，ALBERT と RoBERTa の全てのサイズの平均混淆度行列であり，CLM の行列は OPT の全てのサイズの平均混淆度行列である．例えば，MLM の行列の HYP の行と HPO の列で示されたのは，全ての MLM の Confusability(HPO, HYP, m) の平均値，つまり，上位関係を意味する探査子が入力された時に，MLM が下位関係に混淆してしまう平均的な度合いである．

全体的な傾向として，人間の混淆度はモデルより低い．したがって，CLM・MLM 問わず，モデルの意味関係間の弁別能力は人間と比較して相対的に低いと言える．

人間は，類義関係の上位・下位関係への混淆度が高い．同時に，上位・下位関係の類義関係への混淆度も比較的高い．したがって，人間は類義関係と上位・下位関係を混淆してしまう傾向がある．この傾向は，[1] と [2] の結果と一致する．

類義関係と上位・下位関係の混淆は，MLM と CLM でも観察される．MLM と CLM の類義関係の上位・下位関係への混淆度と，上位・下位関係の類義関係への混淆度は，人間よりも高い．つまり，モデルは類義関係と上位・下位関係について人間以上に混淆して学習してしまっている．

人間と比較してモデルの混淆度行列で際立っているのは，対義関係の列のスコアが全体的に高いことと同時に，対義関係の行のスコアが全体的に低いことである．このことは，モデルは非対義関係を対義関係と誤認識する傾向があるが，対義関係を非対義関係から区別できていることを意味する．つまり，入力探査子が意図する意味関係に関わらず，常にモデルはターゲット単語の対義語に対して大きな確率を割り当ててしまう．このように，PLM には対義語が選好されている対義関係バイアスが観察された．

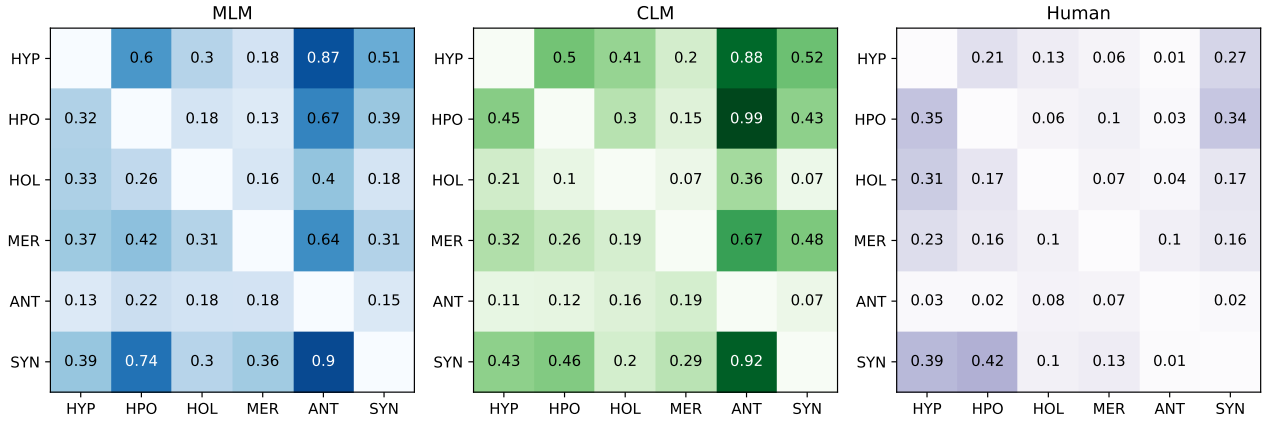

図 1　MLM，CLM および人間の混淆度行列

実験対象としたモデルは，考慮できる文脈（MLM vs CLM）が異なるにも関わらず，対義関係バイアスは共通して観察された．したがって，対義関係バイアスは PLM の特徴だと考えられる．

表 2　ターゲット単語が *niece* の上位関係を意味する探査子 *"the word 'niece' has a more specific sense than the word V"* に対するモデルと人間の回答（Top 2 まで）．

Model	Top 1	Top 2
ALBERT-base	evalle	jovah
ALBERT-large	nephew	niece
ALBERT-xlarge	niece	aunt
ALBERT-xxlarge	niece	cousin
BERT-base	mother	love
BERT-large	sister	niece
RoBERTa-base	sister	daughter
RoBERTa-large	aunt	cousin
OPT-125M	"	n
OPT-350M	"	niece
OPT-1.3B	nephew	cousin
OPT-2.7B	nephew	cousin
OPT-6.7B	nephew	cousin
OPT-13B	nephew	sister
OPT-30B	nephew	cousin
OPT-66B	niece	nephew
Human	relative	kin

ケース・スタディとして，ターゲット単語が *niece* の上位関係を意味する探査子 *"the word "niece" has a more specific sense than the word V"* に対するモデルと人間の回答を表 2 で示した．この探査子に対する正解は，*niece* の上位語の *relation* と *relative* である．人間による回答には，正解である *relative* が含まれているのに対して，いずれのモデルも *relation* または *relative* が回答に含まれていない．*niece* の対義語である *nephew* が，8 個中 1 個の MLM，8 個中 6 個の CLM で回答に含まれている．この事例からも，モデルがターゲット単語の対義語を優先して出力する傾向があることが観察できる。

5　結論

本研究では，PLM が意味関係をどれほど弁別できるかに着眼して，分析を進めてきた．主な貢献は，1) PLM の意味関係の弁別性能評価の新しい指標である混淆度の提案，2) PLM の意味関係弁別性能を人間と比較して分析する初の評価実験の実施の二点である．実験の結果，MLM・CLM 共に人間を下回る弁別性能であることが示された．PLM は人間よりも特に類義関係と上位・下位関係を混淆している傾向にあることが観察された．さらに，PLM が他の関係よりも対義関係を選好し、非対義関係を対義関係として誤認識してしまう対義関係バイアスが存在することを明らかにした．

謝辞

本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2106 の支援を受けたものです．

参考文献

[1] Roger Chaffin and H H Clark. The similarity and diversity of semantic relations. **Memory & Cognition**, Vol. 12, pp. 134–141, 1984.

[2] Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. Hyperlex: A large-scale evaluation of graded lexical entailment. **Computational Linguistics**, Vol. 43, pp. 781–835, 12 2017.

[3] Frank Joosten. Collective nouns, aggregate nouns, and superordinates. **Lingvisticae Investigationes**, Vol. 33, pp. 25–49, 7 2010.

[4] Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. Computing lexical contrast. **Computational Linguistics**, Vol. 39, No. 3, pp. 555–590, September 2013.

[5] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 34–48, 1 2020.

[6] Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. On the systematicity of probing contextualized word representations: The case of hypernymy in bert. **Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics**, pp. 88–102, 12 2020.

[7] Michael Hanna and David Mareček. Analyzing bert's knowledge of hypernymy via prompting. In **Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP**, pp. 275–282. Association for Computational Linguistics, 2021.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **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)**, Vol. 1, pp. 4171–4186, 10 2018.

[9] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[10] 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. **CoRR**, Vol. abs/1907.11692, , 2019.

[11] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. **CoRR**, Vol. abs/2205.01068, , 2022.

[12] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. **CoRR**, Vol. abs/2307.09288, , 2023.

[13] Tim Vor Der Brück. Learning semantic network patterns for hypernymy extraction. **Proceedings of the 6th Workshop on Ontologies and Lexical Resources**, pp. 38–47, 2010.

[14] Andrey Simanovsky and Alexander Ulanov. Mining text patterns for synonyms extraction. In **2011 22nd International Workshop on Database and Expert Systems Applications**, pp. 473–477. IEEE, 8 2011.

[15] Tuğba Yıldız, Banu Diri, and Savaş Yıldırım. Acquisition of turkish meronym based on classification of patterns. **Pattern Analysis and Applications**, Vol. 19, pp. 495–507, 5 2016.

[16] James P Van Overschelde, Katherine A Rawson, and John Dunlosky. Category norms: An updated and expanded version of the battig and montague (1969) norms. **Journal of Memory and Language**, Vol. 50, pp. 289–335, 4 2004.

[17] George A. Miller. Wordnet. **Communications of the ACM**, Vol. 38, pp. 39–41, 11 1995.