Winograd Schema Challenge への ニューラル言語モデルの適用における人名の影響

米川和仁　　松崎拓也
東京理科大学大学院 理学研究科 応用数学専攻
1419523@ed.tus.ac.jp　　matuzaki@rs.tus.ac.jp

1 はじめに

機械学習モデルが獲得した常識的知識を評価するタスクの一つとして，Winograd Schema Challenge (以降，WSC) [7] というものが構想された．WSC における問題文の例を以下に示す．

When Debbie splashed **Tina**, she got wet.

この例に対して，回答者は先行詞候補 Debbie と Tina から代名詞 she に対応する先行詞を特定することを求められる．適切に解くためには「水をかけられたら濡れる」といった常識的な知識が必要となる．しかし，このような知識は明文化されにくい．そのため，単語の共起などの単純な統計的手法では WSC に対して十分な精度を得ることが難しいとされている．

一方で，近年の研究では BERT [3] などのニューラル言語モデルによる WSC に対する大幅な精度向上が報告されている [5, 10, 12]．このことから，ニューラル言語モデルが常識的な知識を獲得しつつあるのではないかと考えられる．BERT 等に基づく手法の訓練は，大量の教師無しデータによる pre-training と比較的少量の教師付きデータによる fine-tuning の 2 つの段階から成る．この学習方法は従来の手法と大きく異なる特徴の一つであり，大幅な精度向上の理由の一つと考えられている．しかしながら，各段階においてデータセットからどの様な知識が抽出されているかは明らかでない．

本研究では，WSC へのニューラル言語モデルの適用において，先行詞候補である人名がモデルの予測に与える影響を分析した．結果として，人名の組み合わせの変化がモデルの予測に大きく影響する問題の存在が確認できた．また，確認された人名の影響が fine-tuning に使用したデータセットにおいて先行詞候補として出現する人名の偏りによるものではないことがわかった．

2 関連研究

機械学習手法を用いて WSC を解くために，多くの常識的知識をどの様に素性として表現するかが議論されてきた．Rahman ら [9] は素性として，「A さんが B さんに水をかけたら，B さんは濡れる」といった事態間知識を機械的に収集したものを用いている．他にも，彼らは接続詞や単語の極性，FrameNet などを素性として利用している．

人手による素性抽出の取り組みに対して，ニューラル言語モデルによる WSC へのアプローチが盛んになりつつある．Kocijan ら [5] は BERT を用いた手法を提案している．彼らの手法は，[MASK] トークンに置き換えられた代名詞の部分に先行詞候補のどちらが当てはまるのかをモデルに学習させるというものである．この手法によって，WSC に対する機械学習手法の精度はニューラル言語モデルを用いない手法に比べ 8.8 ポイントも向上した．

この様な精度向上への取り組みの一方で，モデルが本当に常識的知識を基に WSC を解いているのかという議論が注目されている．Abdou ら [1] は，先行詞の性別の変更や同義語への置換などの言語的摂動に対してモデルによる予測が大きく変化することを示している．また，同様の摂動に対して人間の回答が安定していることを確認している．以上の結果から，彼らはモデルが論理的根拠を背景とした回答を常にしているわけではないと主張している．

3 準備

本節では，本研究で用いたデータセットや手法の詳細について述べる．

3.1 Winogrande

近年の研究では，オリジナルの WSC データセットの一部は統計的手法でも解くことが可能であるこ

とが報告されている [13]．また，データセットには 273 件の問題しか含まれておらず，訓練及び評価データをどのように分割すべきかという問題も議論されてきた．

本研究では，WSC データセットの一つである Winogrande を訓練及び評価データとして用いた．Winogrande は，上で述べたオリジナルの WSC データセットの弱点を改良したデータセットとされている [11]．

本論文でのニューラル言語モデルに対する Winogrande の入力例を表 1 に示す．モデルへの入力文は代名詞（_）を先行詞候補（太字）で穴埋めしたものであり，各問題に対して 2 つ用意される．この 2 つの文を同一のモデルへと入力し，[CLS] ベクトルと分類のためのパラメータベクトルとの内積が大きい方の文を回答として選択する．Winogrande に対するニューラル言語モデルの精度を表 2 に示す．表中の RoBERTa-large（人名）は後述する先行詞候補がいずれも人名である問題に対する精度である．

3.2 影響関数

あるテストサンプルの予測に特定の学習サンプルが与える影響を定量的に解析する場合，学習サンプルを 1 つずつ除いたデータでモデルを再学習させるという方法が考えられる．この手法は単純である一方で，解析にかかる時間とコストが膨大になるという問題がある．影響関数 [6] は，この問題を解決し効率的に各訓練データの影響を測ることを可能にする手法である．まず，以下の式で訓練データ $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ から学習サンプル (x_i, y_i) を取り除いた場合のモデルのパラメータ θ の変化 θ_{change} を求める．

$$\theta_{\text{change}} = -\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_\theta^2 L(x_j, y_j, \hat{\theta})\right)^{-1}\nabla_\theta L(x_i, y_i, \hat{\theta}) \quad (1)$$

ここで，L は学習に用いた損失関数，$\hat{\theta}$ は学習済みのモデルのパラメータを示す．次に，連鎖律を適用しテストサンプル $(x_{\text{test}}, y_{\text{test}})$ に対する損失の変化 L_{change} を以下の式で求める．

$$L_{\text{change}} = \nabla_\theta L(x_{\text{test}}, y_{\text{test}}, \hat{\theta})^\top \theta_{\text{change}} \quad (2)$$

この L_{change} を基に，あるテストサンプルの予測に学習サンプルが与える影響を定量的に評価できる．Han ら [4] は，この手法が BERT の fine-tuning で用いられるデータセットの分析方法として有効であると示している．

4 手法

本研究は，先行詞候補がいずれも人名である問題を対象として，人名に対する摂動によるモデルの挙動の変化を調べる．具体的な摂動の手法は，表 1 に示すように先行詞候補を異なる人名で置き換えるというものである．摂動に用いる人名及び対象となる問題は CoreNLP [2] の Stanford NER (3 class) を用いて抽出した．Winogrande の訓練データ全体に対し検出された 63 個の人名を組み合わせて，問題当たりに 3,905 件の摂動例を生成した．また，抽出された問題に対する RoBERTa [8] の精度を表 2 に示す．

5 実験

本節では，前節で述べた摂動がニューラル言語モデルに与える影響について分析を行う．まず，摂動による予測値の変化の有無について確認する．次に，予測値の変化と問題文中の文法誤りなどノイズ的要素の関係について述べる．最後に，fine-tuning で用いた訓練データが摂動による予測値の変化の原因となっているのかについて分析を行う．

5.1 実験設定

Winogrande に対して最も精度が良い RoBERTa を摂動実験の対象とする．摂動は評価データのみに与え，38,000 件の訓練データで fine-tuning されたモデルの，評価データに対する予測スコアの変化を観察した．本研究における予測スコアとは，代名詞を正しい先行詞候補で置き換えた問題文に対する出力値のことである．

5.2 摂動の影響

まず，各問題における摂動の影響の度合いを測る．影響の度合いを示す $\text{score}_{\text{diff}}$ は，原文における予測スコア score と各摂動例における予測スコア score'_i の差分の平均として計算する:

$$\text{score}_{\text{diff}} = \frac{1}{N}\sum_{i=1}^{N}(\text{score'}_i - \text{score}) \quad (3)$$

ここで N は摂動例の件数である．

摂動による予測スコアの変化の分布を図 1 に示す．図 1 の 0 付近のビンから，1684 件中 1400 件ほどのテストデータで $\text{score}_{\text{diff}}$ の絶対値が 0.1 未満であることがわかる．このことから，RoBERTa は人名に関する単純な摂動の影響に対して頑健であると考えられる．一方で， 200 件ほどの問題で摂動の影響

表 1　ニューラル言語モデルに対する入力文の例

	例文	選択される回答
原文	**Rebecca** wanted to hire **Kayla** to build the website _ needed.	-
入力文 1	[CLS] **Rebecca** wanted to hire **Kayla** to build the website **Rebecca** [SEP] needed.	Rebecca
入力文 2	[CLS] **Rebecca** wanted to hire **Kayla** to build the website **Kayla** [SEP] needed.	Kayla
摂動例	**Steven** wanted to hire **Craig** to build the website _ needed.	-

表 2　Winogrande に対する各モデルの正解率

モデル	正解率 (%)	テストデータ数
random	50.0	2863
BERT-large	68.6	2863
RoBERTa-large	82.8	2863
RoBERTa-large（人名）	83.0	1684

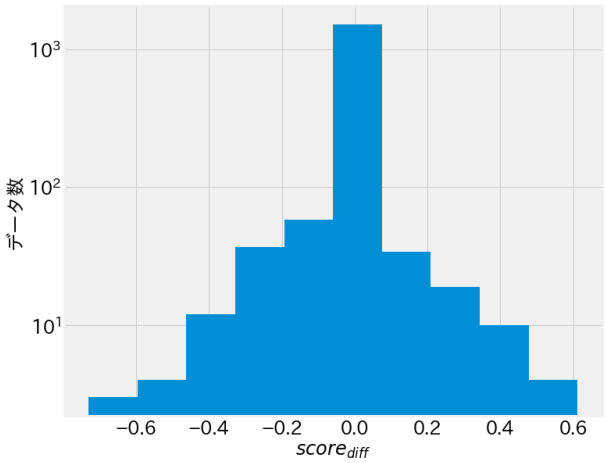

図 1　摂動による予測スコアの変化の分布

が観測された．

5.3　摂動とノイズの関係

評価データにおける問題文と摂動の影響の例を表 3 に示す．摂動の影響を大きく受けた問題について見てみると文法的もしくは意味的な誤りなどのノイズが含まれた文が見られた．例えば，番号 1 の問題では文中の they という代名詞が指し示す対象が不明瞭である．また，番号 6 の問題のように代名詞の直前に冠詞がついているものも確認された．その一方で，ノイズがあるにも関わらず摂動の影響を受けていない問題も多く見られた．例えば，番号 3 の問題では trilingual な人物を選択する必要があるが，先行詞候補はどちらも trilingual では無い．このようなノイズがあるにも関わらず，番号 3 の問題の $score_{diff}$ は 1.31×10^{-4} と非常に小さい．したがって，ノイズと摂動の影響の間に強い関係があると考えるのは難しい．

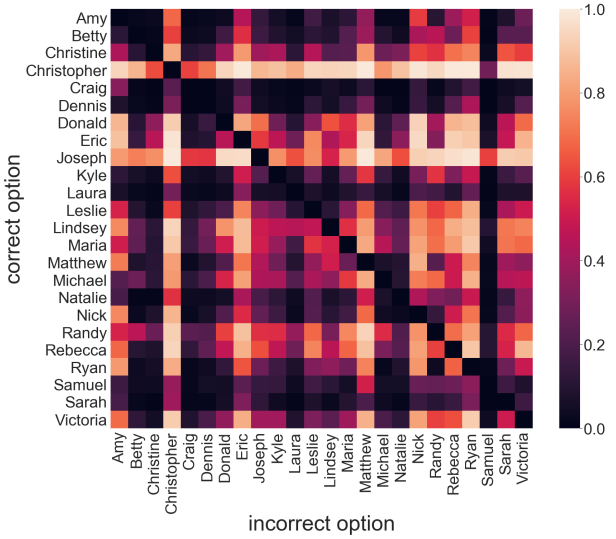

図 2　人名の組み合わせによる予測スコアの分布．縦軸が正解，横軸が誤りの先行詞候補を表す．

5.4　摂動と人名の関係

特定の人名に対して予測スコアの偏りがあるのか，すなわち特定の人名に対して大きな/小さなスコアが付与される傾向の有無を検証する．図 2 に，表 3 の番号 2 の問題に対する各摂動についての予測スコアの分布（一部）を示す．先行詞候補の組み合わせによって，予測スコアが大幅に変化していることがわかる．また，この図で見られるような縦及び横方向にのびる縞模様が多くの問題で確認された．例えば，**Joseph** や **Christopher** という人名が正解となる摂動例はもう一方の人名とのどの組み合わせでもほぼ一貫して高い予測スコアを示している．このことから，モデルの予測と人名の間に強い関係が存在する場合があると考えられる．

5.5　訓練データの影響

摂動例において，特定の人名に対する予測スコアが高くなる例が見られた．このことから，訓練データにおいて正解になりやすい人名がテストデータで選択されやすくなっていることが予想される．本項ではこれを検証する．各摂動例に対して，影響が

表 3　摂動の影響の例

	問題文	$\text{score}_{\text{diff}}$
1	**Rachel** was standing uphill from **Maria** so _ could see them at the bottom of the hill.	−0.730
2	**Eric** decided to eat a lot more yucca than **Nick** did. _ was no longer hungry.	−0.518
3	**Maria** can speak English and German but **Monica** can only speak English because _ is trilingual.	1.31×10^{-4}
4	**Victoria** read the contract carefully before signing it but **Betty** didn't. _ got ripped off by the shady merchant.	1.81×10^{-9}
5	**Eric** was eating more than **Joel** was eating for dinner because _ had skipped lunch.	0.552
6	**Steven** hit his leg on the leg that **Adam** stretched on the floor and fell down because the _ is sitting.	0.612

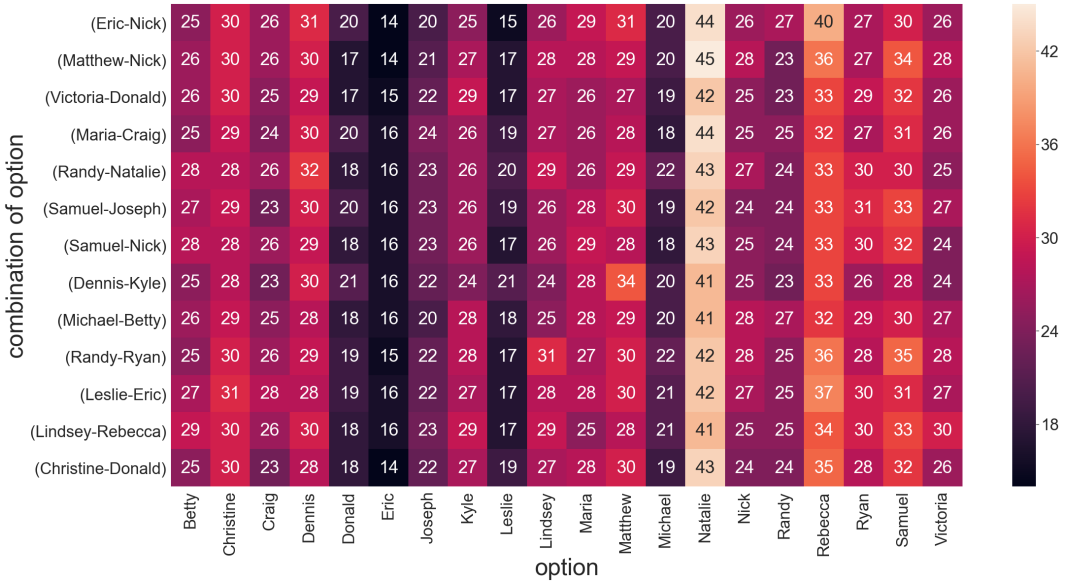

図 3　訓練データにおいて正解となる先行詞候補の出現頻度.
縦軸が摂動例（正解-誤り），横軸が訓練データにおける正解の先行詞候補を示す.

大きい訓練データの上位 3000 件を，影響関数の値に基づき抽出した．各摂動例に対し抽出された訓練データ中での，正解となる人名の出現頻度を図 3 に示す．図は表 3 における番号 2 の問題に対する結果の一部である．各行は一つの摂動例を表し，図に示した 13 の摂動例は全ての摂動例からランダムサンプリングしたものである．また，各列は訓練データにおける正解の先行詞候補を表し，訓練データに含まれる全ての先行詞候補から図に示した 13 個の摂動例に用いた人名のみ抽出している．図 3 から，Christine や Leslie の様にどの摂動例に対しても，影響が大きい訓練データに正解として現れる数がほとんど変化しない人名が多く見られる．一方で，Matthew や Samuel の様に出現頻度が大きく変化している列も確認できる．この様な列で確認された偏りの例として，Matthew という先行詞候補の出現頻度が (Dennis-Kyle) の組み合わせに対して高くなるというものがあげられる．このような例は他にもいくつか確認できたが，摂動に用いられた人名と同一のものが，摂動例に対し大きな影響を持つ訓練データの中で特に多い/少ないという例は確認できなかった．以上のことから，前節で確認された特定の人名に対する予測スコアの偏りは Winogrande の訓練データに起因するものではないと考えられる．

6　おわりに

本研究では，人名の変更という摂動によってモデルの挙動がどう変化するかを分析した．結果として，特定の人名に対して極端な予測を行う問題の存在を示した．しかし，確認された人名への偏りが fine-tuning に用いられたデータセットによるものではないことを確認した．よって，この偏りは pre-training に用いられたデータに起因すると予想される．今後の課題としては，今回確認された偏りが pre-trainig に用いられたデータセットのどの部分によるものかについての分析が挙げられる．

参考文献

[1] Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott and Anders Søgaard. The Sensitivity of Language Models and Humans to Winograd Schema Perturbations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[2] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.

[3] 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). 2019.

[4] Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[5] Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov and Thomas Lukasiewicz. A Surprisingly Robust Trick for the Winograd Schema Challenge. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[6] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. Proceedings of the 34th International Conference on Machine Learning. 2017.

[7] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning. 2012.

[8] 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. arXiv preprint arXiv:1907.11692. 2019.

[9] Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The winograd schema challenge. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012.

[10] Yu-Ping Ruan, Xiaodan Zhu, Zhen-Hua Ling, Zhan Shi, Quan Liu, and Si Wei. Exploring Unsupervised Pretraining and Sentence Structure Modelling for Winograd Schema Challenge. arXiv:1904.09705. 2019

[11] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. Proceedings of the AAAI Conference on Artificial Intelligence. 2019.

[12] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense Reasoning about Social Interactions. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[13] Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019