大規模言語モデルの日本語理解能力検証のための「本音と建前」データセットの構築

子安隆人　　綱川隆司　　西田昌史
静岡大学情報学部情報科学科
koyasu.takato.16@shizuoka.ac.jp　　{tuna,nishida}@inf.shizuoka.ac.jp

概要

本研究では，大規模言語モデルの言語処理能力の検証を目的とし，大規模言語モデルが日本語の独特の文化的背景や言語構造を適切に理解し処理することができるかを明らかにするため，言語モデルが文中の非明示的な情報や隠された意味を理解し，それを基に推論する能力を有しているか検証するためのデータセットを構築する.本データセットを用いて，言語モデルがどの程度日本語文のニュアンスや文脈を適切に捉えることができるのかを検証した.

1 はじめに

近年，人工知能技術のブレイクスルーにより自然言語処理技術は目覚ましい発展を遂げている．この分野での飛躍的な発展はコンピュータが人間の言語をより深くより正確に理解し，それを基に適切な応答を生成する能力の向上を示している．この動向の中心には, OpenAI 社による大規模言語モデルである ChatGPT が存在し，その卓越した性能は注目を集めている.

しかしながら，大規模言語モデルの言語処理能力の検証においては，英語文における検証や評価の報告は数多く行われている一方，日本語における評価は英語と比較して少ないという現状がある[1]．日本語は独特の文化的背景や言語構造を持ち，これらの要素が言語処理において重要な役割を果たしている．英語とは異なる日本語のニュアンスや文脈を正確に捉えるための検証や研究は極めて重要であると言え，とりわけ文中の非明示的な情報や隠された意味を理解し，それを基に推論する能力に関する研究は限定的である．大規模言語モデル ChatGPT が，文中の非明示的な情報や意図をどの程度正確に理解し，適切に処理できるのかという点について，「本音と建前」に焦点を当てたデータセットを用いて検証する.

2 関連研究

言語モデルの一般的な日本語理解能力を測る研究としては， 栗原[1]らの研究が挙げられる． この研究では， 文章分類タスク， 文ペア分類タスク， QA タスクから構成される言語理解ベンチマーク JGLUE を構築しており，QA タスク以外においては人間のスコアよりも高いスコアを記録する言語モデルが存在することが示唆された.

一方，言語モデルの日本語における常識推論能力を検証する研究としては， 竹下[2]らの研究が挙げられる．文化相対性を考慮した英語圏以外の常識道徳を反映した，常識道徳の理解度評価用の日本語データセット JCommonsenseMorality を構築した． この研究によって，他の類似したベンチマークと比較して，より困難な常識理解を必要とすることが示唆された.

3 データセット構築手法

本研究では, 日本において「本音と建前」として扱われる習慣を焦点とし，大規模言語モデルが日本語文中の非明示的な情報や背後の意図をどれだけ的確に捉えられるのか, またその推論能力の限界や強みを明らかにするための手法を提案し，検証を行うことを目的とする.

本実験では，「本音と建前」を軸としたシナリオを複数用意し, それを基に「本音と建前」に関する日本語の読解問題 50 問からなるデータセット[i]を構築する．そして，作成した問題を大規模言語モデルに入力し，回答を記録して分析を行う.

本実験で作成した読解問題は 4 択の選択式問題お

[i] https://github.com/mrbte/honne_tatemae

よび自由記述問題であり，問題文は以下の 4 つの部分で構成される．

1. 以下に続く文章が日本の「本音と建前」文化を表すものであることに関する説明文
2. 各エピソードの前提となる説明文
3. 建前の形で表現される言葉に関する文
4. 隠された本音によってもたらされた結果を示す文

選択肢式問題における選択肢は以下の 4 つの構成とする．

1. 建前の文を肯定した文
2. 建前の文を否定した文
3. 本音を的確に表した文
4. 本音にやや近いが最も的確に表しているとはいえない文

上述の問題構成および選択肢の作成ルールに基づいて作成した選択肢式の問題例を図 1 に示す．また，自由記述の問題例を図 2 に示す．

以下の文章は日本の本音と建前の文化を表す文章です。A さんは若い頃、自分のアパートで何回かパーティを開きました。すると近所の人たちは「お友達といつも楽しそうにしてて、羨ましい」といったような言葉を A さんにかけました。そして、A さんが次のパーティに近所の人たちを誘ったところ、警察が呼ばれました。A さんは警察から「騒がしすぎる」と注意を受けました。

近所の人の「お友達といつも楽しそうにしていて、羨ましい」という言葉はどのような意味・意図であったか、最も適当なものを以下の 1 から 4 の選択肢の中から一つ選んでください。その選択肢を選んだ理由も簡潔に述べてください。

1.友達とパーティを楽しんでいることが羨ましいと伝えたい
2.友達とパーティを楽しむことは羨ましくはないと伝えたい
3.パーティの音がうるさく迷惑であることを伝えたい
4.パーティを楽しんでいることは羨ましいが音が大きすぎるということを伝えたい

図 1　選択肢式の問題例

以下の文章は日本の本音と建前の文化を表す文章です。
A さんは若い頃、自分のアパートで何回かパーティを開きました。すると近所の人たちは「お友達といつも楽しそうにしていて、羨ましい」といったような言葉を A さんにかけました。
そして、A さんが次のパーティに近所の人たちを誘ったところ、警察が呼ばれました。A さんは警察から「騒がしすぎる」と注意を受けました。
近所の人の「お友達といつも楽しそうにしていて、羨ましい」という言葉はどのような意味・意図であったと考えられるか、簡潔に答えてください。その理由も述べてください。

図 2　自由記述の問題例

4 実験設定

本実験のテスト対象となる言語モデルは OpenAI 社の大規模言語モデルである GPT-4 および GPT-3.5 Turbo である．選択肢式問題における選択肢の 1 から 4 の並び順は全ての問題で昇順であるが，一つの問題につき新しいセッションを生成して問題の入力を行っているため，ある問題への回答が前の問題の情報によって影響を受けることはない．評価実験は，2023 年 11 月時点で利用可能な最新のモデルを API 経由で用いて実施した．

4.1 問題の入力条件

大規模言語モデルのバージョンによる回答の違い，そして問題入力時の指示，すなわちプロンプトエンジニアリング手法[3][4]による回答の違いを検証するため，言語モデルのバージョンおよび入力方法によって回答を分けて記録する．問題の入力方法を表 1 に示す．テストは一つのモデルにつき入力方法ごとに 5 回ずつ実行し，回答を記録する．

表 1　問題の入力方法

入力方法	説明
Zero-Shot(normal)	作成した選択式問題をそのまま入力する
Few-Shot	例題とその解答を複数提示した後に問題を入力する
Zero-Shot CoT	「ステップバイステップで考えてみましょう」という指示を与え段階的に推論するように促す
Self-Consistency	例題と解答、およびその解説を複数提示した後に問題を入力する
本音と建前の説明文なし	問題文が日本の「本音と建前」文化を表すものであるという説明文を省く
自由記述	問題に選択肢を含めず回答を自由に記述させる

4.2 評価項目

言語モデルの回答における評価項目は，選択肢式問題では問題の正解率に加え，表 4 中の(P), (Q)の 2 つの項目を評価する．選択肢式問題における評価対象は表 1 中の Zero-Shot とする．

自由記述問題における回答の評価は，表 5 中の(R), (S), (T)の 3 つの項目を評価する．

4.3 結果

言語モデルの選択肢式問題における正解率を表 3 に示す．

表 3 選択肢式問題における正解率

	正解率(%)	
入力方法/モデル	GPT-3.5 Turbo	GPT-4
Zero-Shot	44.8	59.2
Few-Shot	32.4	57.2
Zero-Shot CoT	41.2	58.8
Self-Consistency	44.4	72.4
本音と建前の情報なし	40.0	58.4

選択肢式問題における 2 つの評価項目の結果を表 4 に示す．

表 4 選択肢式問題における 2 つの評価項目

	評価(%)	
評価項目/モデル	GPT-3.5 Turbo	GPT-4
言語モデルの回答が本音の存在を見抜いている,またはその言及があるかどうか(P)	76.0	94.0
回答に整合性があるかどうか(Q)	99.6	100.0

自由記述問題における 3 つの評価項目の結果を表 5 に示す．

表 5 自由記述問題における 3 つの評価項目

	評価(%)	
評価項目/モデル	GPT-3.5 Turbo	GPT-4
回答に整合性があるかどうか(R)	100.0	100.0
回答が選択肢式問題における正解の選択肢の内容と一致しているかどうか(S)	42.0	74.0
回答が選択肢式問題における正解の選択肢の内容とは一致しないが，一定以上の水準で正解に近い，または現実的な推論ができているかどうか(T)	48.0	26.0

表 3 から，入力方法によらず GPT-3.5 Turbo より GPT-4 の方が高い平均正解率を記録している．一方，各モデルの入力方法ごとの平均正解率の変化に着目すると, GPT-3.5 Turbo では問題入力時に回答についての指示を与えた全ての場合で平均正解率が減少している. GPT-4 では入力方法として Self-Consistency を採用した場合に平均正解率が大きく上昇しており，それ以外の 3 つの入力方法では Zero-Shot の場合と平均正解率に大きな差はない．

表 4 から，評価項目(P)において GPT-4 は GPT-3.5 Turbo に比べ 124%程度の数値を記録している. (Q)では, GPT-3.5 Turbo で 0.996, GPT-4 で 1.000 とほぼ 1 に近い数値である．

表 5 から, (R)において GPT-3.5 Turbo と GPT-4 の両方で 1.000 を記録している．一方, (S)においては GPT-3.5 Turbo で 0.420, GPT-4 で 0.740 となり, GPT-4 は GPT-3.5 Turbo に比べ 76%程度高い数値となっている. (T)においては, GPT-3.5 Turbo で 0.480, GPT-

4 で 0.260 となっており，それぞれのモデルで(S)と(T)の数値を加算した合計値は GPT-3.5 Turbo で 0.900, GPT-4 で 1.000 である．

以上の結果から，選択肢式問題と自由記述問題の 2 つの出題方法にかかわらず，GPT-4 は GPT-3.5 Turbo に比べ，より正解に近い，すなわち非明示的な情報や意図を適切に推測し処理していることが示唆された．一方で，出題方法間のモデルごとの回答に着目すると，表 4 中の(P)と，表 5 中の(S)と(T)の合計値との比較から，それぞれのモデルにおいて選択肢式問題として出題した場合より自由記述問題として出題した場合の方がより適当な推測を行うことができる傾向があることが示唆された．

5 おわりに

本論文では，言語モデルが文中の非明示的な情報や隠された意味を理解し，それを基に推論する能力を有しているか検証するためのデータセットを構築した．今後は言語モデルの言語処理能力を評価するためのベンチマークとして，より複雑で高度な言語処理能力を必要とするデータセットの構築を進めていく予定である．

参考文献

[1] 栗原健太郎，河原大輔，柴田知秀:” JGLUE: 日本語言語理解ベンチマーク”，言語処理学会第 28 回年次大会, 2022.

[2] 竹下昌志，ジェプカ・ラファウ，荒木健治:” JCommonsenseMorality:常識道徳の理解度評価用日本語データセット”，言語処理学会第 29 回年次大会, 2023.

[3] Jie Huang, Kevin Chen-Chuan Chang: “Towards Reasoning in Large Language Models: A Survey”, In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065. 2023.

[4] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao: “ReAct: Synergizing Reasoning and Acting in Language Models”, ICLR2023, 2023.