HOTATE：本音と建前の応答対からなる対話コーパスの構築

戸田 裕子 [1] 前川 大輔 [2] 眞鍋 光汰 [2] 米山 瑛人 [1] 野々村 奏 [1] 藤原 有希 [1] 梶原 智之 [2,3]
[1] 愛媛大学工学部工学科 [2] 愛媛大学大学院理工学研究科 [3] 大阪大学 D3 センター
{toda,maekawa,manabe,yoneyama,nonomura,fujiwara}@ai.cs.ehime-u.ac.jp
kajiwara@cs.ehime-u.ac.jp

概要

本研究では，大規模言語モデルが日本語対話における本音と建前をどれだけ正確に扱えるのかを明らかにする．日本語対話コーパスの発話に本音と建前の各応答を付与し，分類および変換の性能を評価した結果，既存の大規模言語モデルは本音と建前を充分に理解していない一方で，我々のコーパスによる学習でそれらの性能が大幅に向上することを確認した．また，変換の実験では建前よりも本音の生成が難しいことを自動評価と人手評価の両方で示した．

1 はじめに

我々は日々の対話において，人間関係を円滑に保つために，状況や相手との関係性に応じて「本音」と「建前」を使い分けている [1]．本音は話者の意図と発話内容が一致している一方で，建前は社会的配慮のために意図と発話内容が乖離している．本音と建前の認識は，コミュニケーションを円滑に進めるために重要だが，対話においてある発言が本音か建前かを判断することは必ずしも容易ではない．

本音／建前の関連研究として，直接／間接的表現 [2] や皮肉 [3–5] が挙げられる．直接／間接は，発話者の意図をどの程度明示するかという表現方法に関する概念であり，意味内容に関する概念である本音／建前とは独立している．建前とは異なり，直接的表現と間接的表現は，いずれも意図と発話内容が一貫している．皮肉は，発話者の意図と発話の字義的な意味が異なるという点で建前と類似した概念である．ただし，社会的配慮を目的とする建前とは機能が異なり，皮肉は攻撃的であり相手を批判する目的で使用される．これらの関連研究とは異なり，本音と建前を明示的に扱うコーパスは存在しないため，大規模言語モデル (Large Language Models; LLMs) を含む自然言語処理モデルによる本音と建前の認識および生成の能力は明らかになっていない．

そこで本研究では，既存の日本語対話コーパスを拡張し，7,964 件の対話に対して本音の応答および建前の応答を人手で付与した HOTATE[1)] コーパスを図 1 のように構築し，公開する．本コーパスを使用し，対話における最後の発話が本音か建前かを分類するタスクおよび本音と建前を変換するタスクを通じて，LLM が本音と建前をどれだけ理解可能かを検証した．実験の結果，既存の日本語 LLM は本音と建前を充分に理解できないことが明らかになった．

2 HOTATE コーパス

本節では，HOTATE コーパスの設計や統計情報について述べ，本音と建前の違いについて分析する．

2.1 データソース

日常やビジネスなどの多様なドメインを対象とするために，複数の日本語対話コーパスを利用した．まず，日常分野のデータソースとして，日常生活・学校・旅行・健康・娯楽の 5 つのトピックで構成される Japanese Daily Dialogue (JDD)[2)] [6] を採用した．また，ビジネス分野のデータソースとして，会議や雑談などの多様なビジネス対話を収録した The Business Scene Dialogue corpus (BSD)[3)] [7] を用いた．これらの 2 つのコーパスをもとに，次節で説明するアノテーションを追加した．

2.2 アノテーション

上記の対話コーパスにおいて，対話ごとに 1 つの発話を選択して本音および建前の応答を付与した．アノテーションの例を図 1 に示す．アノテータは，クラウドソーシングサービスのランサーズ[4)]で募集した．高品質なアノテーションのために，ランサー

1) A dialogue corpus consisting of pairs of **HO**nne and **TATE**mae https://github.com/EhimeNLP/HOTATE
2) https://github.com/jqk09a/japanese-daily-dialogue
3) https://github.com/tsuruoka-lab/BSD
4) https://www.lancers.jp

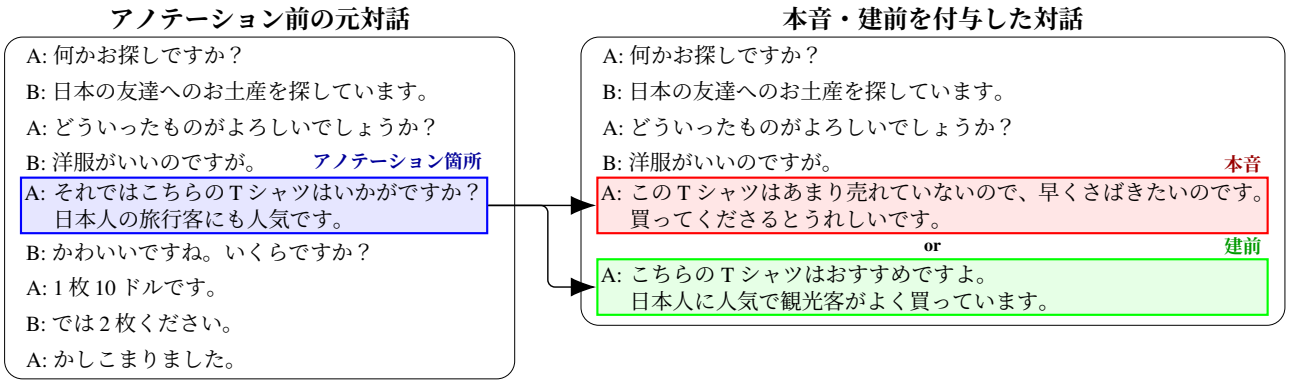

図 1: 対話への本音と建前の応答のアノテーション

表 1: HOTATE コーパスの統計情報

元コーパス	トピック	対話数	1 対話あたりの統計		1 発話あたりの統計	語彙サイズ
			平均発話数	平均単語数	平均単語数	
JDD	日常生活	1,648	5.930	90.350	15.300	4,727
	学校	1,692	5.822	99.236	17.119	4,798
	旅行	1,400	5.504	91.660	16.827	4,382
	健康	1,160	6.083	97.590	16.340	3,804
	娯楽	1,140	5.532	96.819	17.531	3,964
BSD	ビジネス	924	18.565	238.823	12.879	6,102
	合計	7,964	7.263	111.675	15.461	14,808

ズにて継続的な高評価の実績を持つ「認定ワーカ」に限定してアノテータを募集し 33 名を採用した.[5)]

アノテーション作業では，まずアノテータが対話を読み，文脈や発話スタイルを理解したうえで，本音の応答と建前の応答に分かれそうな箇所（図 1 左のアノテーション箇所）を 1 つだけ選択し，図 1 右のように本音と建前の応答を付与した．ここで，アノテーションした発話よりも後の対話履歴は HOTATE コーパスには含まれないことに注意されたい．その他，アノテータには，以下の指示を与えた．

1. 話者の意図を明確に設定し，意図と発話内容が，本音では一致し，建前では乖離するよう，対照的な意味の本音-建前の対を作成すること．
2. 対話履歴の文脈を踏まえた応答を作成すること．つまり，当該話者の発話スタイルとして自然で，発話内容が意味的に妥当であること．
3. 対話履歴から自然に導かれない新たな背景情報や状況設定を付加しないこと．

5) アノテーション作業には時給 1,400 円を支払った．これは，代表的なクラウドソーシングサービスのひとつである Prolific における最低報酬水準（時給 8 ドル）を超える充分な金額である． https://www.prolific.com/pricing

コーパスの品質を担保するため，2 段階の品質管理を実施した．まず，アノテーション結果を第一著者が確認し，指示に従っていない場合にはアノテータに修正を依頼した．その後，著者らが再度コーパス全体を見直し，必要に応じて修正を加えた．

2.3 統計情報

本研究では，3,982 件の対話に対して，それぞれ本音と建前の 2 種類の応答を付与し，7,964 対話を得た．本コーパスの統計情報を表 1 に示す．ビジネスドメインは，対話あたりの発話数が多く，発話あたりの単語数が少ない傾向が見られた．ここで，平均単語数や語彙サイズの計算には，Sudachi[6)] [8] による単語分割 (sudachidict_full) を採用した．

2.4 本音と建前の感情分布

本コーパスを観察すると，本音の応答にはネガティブな表現が多く，建前の応答にはポジティブな表現が多い傾向が見られた．本節では，感情の分布について，語レベルと文レベルの両方で分析する．

6) https://github.com/WorksApplications/SudachiPy

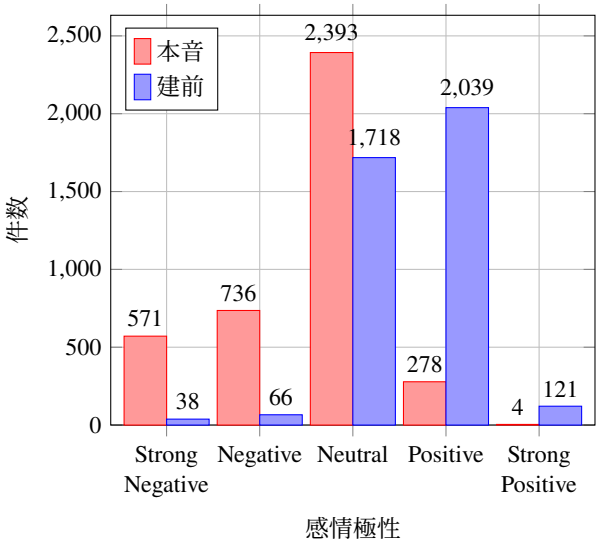

図 2: 本音と建前の文レベルの感情極性の分布

語レベルの感情分布　本音と建前の応答間の語彙的な特徴を分析するために，前節と同様に単語分割したうえで，それぞれに頻出する単語を数えた．ただし，両者に 30 回以上ずつ出現する高頻度語は除外し，特徴を観察した．分析の結果，本音の応答においては，面倒・苦手・無理など，ネガティブな語が頻度の上位を占めた．一方で，建前の応答においては，楽しみ・楽しい・嬉しいなど，ポジティブな語が頻度の上位を占め，対照的な傾向が見られた．

文レベルの感情分布　文レベルでも感情分布を調査した．日本語の感情分析コーパス WRIME[7] [9, 10] を用いて ModernBERT[8] [11] をファインチューニングし，本音と建前の応答に適用した結果を図 2 に示す．語レベルの分析と同様に，本音ではネガティブな応答が多く，建前ではポジティブな応答が多い傾向が見られた．特に，建前の応答はポジティブが多く，ネガティブな応答は 3% 未満と少数であった．

3　評価実験

LLM が本音と建前をどれだけ正確に扱えるかを，分類タスクおよび変換タスクにおいて検証する．

3.1　実験設定

本音と建前の分類および変換の両タスクに共通の実験設定として，2 節で構築した HOTATE コーパスを，対話単位で 8 : 1 : 1 に分割し，それぞれ訓練用・検証用・評価用に使用した．実験に使用した LLM は，

7）https://github.com/ids-cv/wrime
8）https://huggingface.co/sbintuitions/modernbert-ja-310m

表 2: 本音と建前の分類タスクにおける正解率 (%)

Model	0-shot	10-shot	SFT
gpt-oss-20b	**55.51**	55.39	89.72
llm-jp-13b	45.61	65.66	91.85
swallow-8b	50.38	**78.82**	**92.98**

日本語に特化していない `gpt-oss-20b`[9] [12] に加え，日本語に特化した `llm-jp-3.1-13b-instruct4`[10] [13] および `Llama-3.1-Swallow-8B-Instruct-v0.5`[11] [14] の 3 種類である．LLM の指示チューニング（SFT）には SFT Trainer[12] を使用し，学習率は $2e-4$，最適化手法は AdamW [15] とし，検証用データセットにおけるクロスエントロピー損失の 3 エポックの early-stopping によって訓練を終了した．

3.2　本音と建前の分類タスク

本タスクは，対話を入力とし，最後の発話が本音か建前かを推定する 2 値分類タスクである．

実験設定　3 種類の LLM を，0-shot，10-shot，SFT の各設定で使用し，分類の正解率を自動評価した．10-shot 推論には，訓練用データセットから本音と建前の対話を 5 件ずつ無作為に抽出して使用した．

実験結果　表 2 に実験結果を示す．まず 0-shot 設定では，どのモデルも 50% 前後の正解率にとどまり，ランダム推定と同等の性能であった．次に 10-shot 設定では，日本語モデルの `llm-jp-13b` および `swallow-8b` においては性能改善を確認できた．そして，指示チューニングによって，全てのモデルにおいて大きな性能向上が見られ，90% 前後の正解率に到達した．これらの実験結果は，既存の LLM が日本語の対話における本音と建前を充分に理解できない一方で，我々のコーパスによる学習で著しい性能改善が得られることを示している．本実験から，HOTATE コーパスの有用性を確認できた．

訓練データ量と性能の関係　LLM が本音と建前を充分に学習するために必要なデータ量を明らかにするために，訓練データ量と分類性能の関係を分析した．訓練用データセットのうち，先頭から $N \in \{300, 500, 1000, 2000, 4000, 6000\}$ 件を用いて

9）https://huggingface.co/openai/gpt-oss-20b
ただし，`reasoning_effort` は `medium` の設定で使用した．
10）https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4 （以降 `llm-jp-13b` と記載）
11）https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5 （以降 `swallow-8b` と記載）
12）https://huggingface.co/docs/trl/sft_trainer

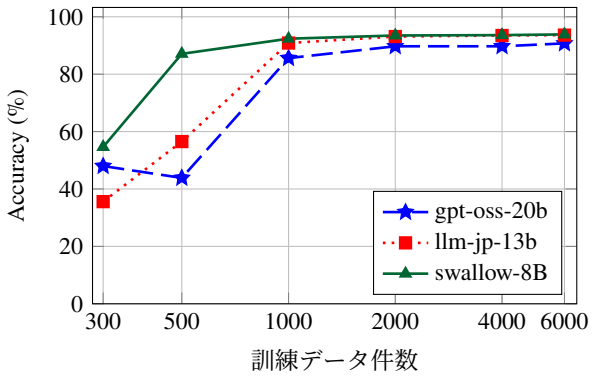

図 3: 訓練データ量と分類正解率の関係

学習した際の LLM の分類性能を図 3 に示す．まず，全てのモデルにおいて，1,000 件の訓練で 90% の正解率を達成し，2,000 件の訓練で最高性能に到達できることがわかる．次に，表 2 と比較すると，`gpt-oss-20b` では 10-shot 設定で 0.55，`llm-jp-13b` では 10-shot 設定で 0.66 の正解率であったため，500 件未満の訓練データ量であれば，文脈内学習 [16] を採用すべきであることが示唆される．最後に，`swallow-8b` は，訓練データ量の多少に関わらず，一貫して他のモデルを上回る性能を達成しており，本音と建前の学習に適した LLM であると言える．

3.3 本音と建前の変換タスク

本タスクは，対話を入力とし，最後の本音（建前）の発話を建前（本音）に変換する生成タスクである．

実験設定 分類タスクと同じ 3 種類の LLM を用いて SFT の設定で実験した．自動評価は，評価用データセットの全体に対して生成文と参照文を比較し，BLEU[13] [17, 18] による表層的類似度および SBERT[14] [19, 20] による意味的類似度を評価した．また，評価用データセットから無作為抽出した 50 対話について，人手評価も実施した．評価者は日本語母語話者の大学生 3 名である．評価項目は以下の 3 項目であり，それぞれ 1〜3 の 3 段階で評価した．

発話 文法性や流暢性など，発話自身の良さ．

対話 履歴を考慮した上での発話内容の妥当性や発話スタイルの一貫性など，対話としての良さ．

変換 本音では発話意図と発話内容が一致しており，建前ではそれらが乖離しているという，本音と建前の対としての良さ．

13) https://github.com/mjpost/sacrebleu
14) https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

表 3: 本音と建前の変換タスクにおける性能評価

	自動評価		人手評価		
Model	BLEU	SBERT	発話	対話	変換
本音 → 建前					
gpt-oss-20b	10.59	0.69	3.0	2.8	2.6
llm-jp-13b	12.14	0.70	2.9	2.7	2.6
swallow-8b	13.84	0.70	2.9	2.8	2.7
建前 → 本音					
gpt-oss-20b	8.01	0.67	2.6	2.2	2.0
llm-jp-13b	9.60	0.68	2.9	2.6	2.4
swallow-8b	8.73	0.68	2.9	2.5	2.5

実験結果 表 3 に実験結果を示す．まず，自動評価と人手評価で一貫して，建前 → 本音の方向よりも本音 → 建前の方向の変換の方が高い評価を得た．次に，自動評価に注目すると，両方向の変換において一貫して，`gpt-oss-20b` よりも日本語特化モデルである `llm-jp-13b` および `swallow-8b` の方が高い性能を達成した．最後に，人手評価に注目すると，本音 → 建前の方向の変換においてはモデル間に大きな性能差はなく，どのモデルも高品質な変換ができていると言える．一方で，建前 → 本音の方向の変換では，`gpt-oss-20b` が全ての評価項目において低評価であり，本音の生成が難しいことがわかる．

自動評価と人手評価の相関 自動評価と人手評価の間のピアソン相関は，BLEU で 0.04 から 0.18，SBERT で 0.00 から 0.23 であった．これらの参照文との一致に基づく自動評価は，いずれの人手評価項目とも充分な相関を持たないことが明らかになった．日本語における本音と建前の変換タスクのためのより良い自動評価の確立は，今後の課題である．

4 おわりに

本研究では，日本語の日常対話コーパス JDD およびビジネス対話コーパス BSD に対して本音と建前の応答対を付与し，LLM が日本語における本音と建前を理解および変換する能力について分析した．3 種類の LLM を対象とする評価実験の結果，現状の LLM にとって本音と建前の分類が難しいことを明らかにするとともに，我々のコーパスを用いた学習によって分類性能が大きく向上することを示した．また，生成タスクにおいては，建前よりも本音の発話を生成することが難しいことが明らかになった．

謝辞

本研究は，JSPS 科研費（若手研究，課題番号：JP24K20840）の助成を受けて実施した．

参考文献

[1] Raden Regine Melansyah and Nuria Haristiani. Analysis of Japanese Refusal Speech Acts to an Invitation as a Tatemae. In **Proceedings of the 3rd International Conference on Language, Literature, Culture, and Education**, pp. 112–115, 2020.

[2] Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. DIRECT: Direct and Indirect Responses in Conversational Text Corpus. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 1980–1989, 2021.

[3] Deirdre Wilson. The Pragmatics of Verbal Irony: Echo or Pretence? **Lingua**, Vol. 116, No. 10, pp. 1722–1743, 2006.

[4] Silviu Oprea and Walid Magdy. Exploring Author Context for Detecting Intended Vs Perceived Sarcasm. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2854–2859, 2019.

[5] Ibrahim Abu Farha, Steven Wilson, Silviu Oprea, and Walid Magdy. Sarcasm Detection Is Way Too Easy! An Empirical Comparison of Human and Machine Sarcasm Detection. In **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 5284–5295, 2022.

[6] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 日本語日常対話コーパスの構築. 言語処理学会 第 29 回年次大会発表論文集, pp. 108–113, 2023.

[7] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the Business Conversation Corpus. In **Proceedings of the 6th Workshop on Asian Translation**, pp. 54–61, 2019.

[8] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: A Japanese Tokenizer for Business. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**, pp. 2246–2249, 2018.

[9] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2095–2104, 2021.

[10] Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 7022–7028, 2022.

[11] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, pp. 2526–2547, 2025.

[12] OpenAI. Gpt-oss-120b & Gpt-oss-20b Model Card. **arXiv:2508.10925**, 2025.

[13] LLM-jp. Llm-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv:2407.03963**, 2024.

[14] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pretraining for Cross-lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In **Proceedings of the First Conference on Language Modeling**, 2024.

[15] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **Proceedings of the Seventh International Conference on Learning Representations**, 2019.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Proceedings of the Thirty-Fourth Annual Conference on Neural Information Processing Systems**, pp. 1877–1901, 2020.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[18] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, 2018.

[19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 3982–3992, 2019.

[20] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 4512–4525, 2020.