大規模言語モデルによるポライトネス理論の検証

高橋 哲朗 [1]　宇佐美 まゆみ [2]
[1] 鹿児島大学　[2] 東京外国語大学
takahashi@ibe.kagoshima-u.ac.jp　usamima@tufs.ac.jp

概要

社会科学の分野で研究の蓄積があるポライトネス理論の検証を目的とし，対話コーパス中の発話における発話の丁寧さや，話者間の距離，力関係，発話内容の相手への負荷を大規模言語モデルにより推定した．推定結果を用いてフェイス侵害度と丁寧さを比較したところ実験に用いた 120 対話のうち 100 対話で有意な相関が見られ，これらの対話においてはポライトネス理論が示す通りの結果を確認できた．

1　はじめに

対話システムの研究においては，大規模言語モデル (LLM) の活用により人間との自然な対話を行なうシステムが実現されてきている [1, 2, 3]．これらの研究においてシステムを評価する際には，タスクの達成度に加えて対話の自然さが用いられることが多いが，その判断基準はタスクの理解や発話における矛盾の無さ，首尾一貫性などの発話の命題の範囲に留まっている．一方で，言語学や社会科学の分野では対話や談話に関する研究 (以降，対話研究[1)] と表記する) の蓄積があり，これらの研究では，対話相手による言葉遣いの違い，円滑なコミュニケーションのためのストラテジーとしてのポライトネス理論，会話における人間の相互作用，話題展開，発話行為（謝罪，勧誘，断り，依頼）などに関する様々な研究がなされてきた [4, 5, 6]．

これらの研究が対象としている「言語使用場面における対人関係」は，上述した対話システムの研究においてはまだ十分には考慮されていない．そのため現行の対話システムは，タスクが明確なやりとりにおいてその対象とするタスクはある程度行なえ，また加えて雑談も継続することができるようになったが，まだ人間同士の対話とは違い，対話システムに人が歩み寄って使用している状況にあると言える．従来は，対話研究の成果を対話システムに実装することは容易ではなく，理論研究と対話システムの間に隔りがあったが，LLM を活用することによりこの隔りを克服し，対話研究で蓄積されてきた理論や成果を実装できる可能性が出てきている．

本研究では，そういった取り組みの一つとして対話研究の中でポライトネス理論 [7] に焦点を当て，LLM がポライトネス理論を扱えるかの検証を目的とする．具体的には以下に取り組む．

- ポライトネス理論で定義されている D, P, R_x の値や発話の丁寧度 (丁寧さの度合い) を LLM を用いて対話コーパス中の発話から推定する
- 対話コーパスを用いたフェイス侵害度と丁寧さの比較を通し，ポライトネス理論を定量的に検証する

2　ポライトネス理論

ポライトネス理論とは円滑な人間関係の確立・維持のための言語的ストラテジーを体系化したものでありブラウンとレビンソン (以降 B&L)[8] の定義がその後の研究に強い影響を与えている．

発話におけるポライトネスに関し B&L はフェイスという概念を導入した．

ポジティブ・フェイス
: 他者に理解されたい，好かれたい，賞賛されたいというプラス方向の欲求

ネガティブ・フェイス
: 賞賛されないまでも，少なくとも，他者に邪魔されたり，立ち入られたくないというマイナス方向の欲求

そして人がコミュニケーションを行なう際には相手のフェイスを侵害する行為を行なわざるを得ないと捉え，このフェイス侵害度を埋め合わせるために行なう言語的スタラテジーとしてポライトネスを捉えた．そして，このフェイス侵害の度合い (W_x) を見

1)　対話システムの研究とは区別する

積もる公式として式 (1) を提案した．

$$W_x = D(S,H) + P(H,S) + R_x \quad (1)$$

W_x　フェイス侵害度．行為 (x) がフェイスを脅かす度合い

D　S (Speaker) と H (Hearer) の社会的距離 (Social Distance)

P　H (Hearer) の S (Speaker) に対する力 (Power)

R_x　特定の文化で，ある行為 (x) が相手にかける負荷度の絶対的順序に基づく重み

この公式が示していることの一つは「フェイス侵害度 W_x が高いときに丁寧な表現になる」ということである．そこで本研究では対話コーパスを用いてこの現象が表われていることを定量的に確認する．

3　LLM を用いたフェイス侵害度および丁寧度の推定

3.1　対象データ

本研究は，宇佐美により構築された BTSJ1000 人日本語自然会話コーパス [9] を用いた．このコーパスには初対面，友人，教師と学生など，多様な話者間の対話のデータが合計 514 対話収録されており，様々な言語運用の研究のために有用なものである．BTSJ コーパスは 32 種類のフォルダに分けられているが，今回はその中から 1,2,3,4,5,17,18 番のフォルダに収録されている合計 120 の対話のデータを用いた．これらの対話における話者間の関係や対話の内容を 表 2 の「対話種別」に示す．

3.2　LLM による推定

本研究では対話コーパス中の発話の内容から，その発話において話者が相手にかけるフェイス侵害度 (W_x) と丁寧度 (Po) を LLM により推定する．モデルには "gpt-4o-mini"を用い，それぞれ付録 A の図 1，図 2 に示すプロンプトを用い推定した．120 対話に含まれる全 25,844 発話において，フェイス侵害度 (W_x) を求めるために D,P,R_x をそれぞれ LLM により推定し，それらの合計により W_x を算出した．

推定結果の例を 表 1 に示す．W_x の式において，話者間の社会的距離 D および，聞き手の話し手に対する力 P は対話の開始から終了までを通して変化する事は考えにくい．LLM が出力した結果を見ても，D と P の値はほとんど変化しなかった．表 1 の対話は，話者 1 が 話者 2 に依頼をする対話である．R_x については，「うん」や「えー」のような相槌においては 1 と低い値が推定されている一方で，依頼について話している発話では高い値が推定されている．相槌は相手に対して負荷をかけることが少ない一方で，依頼は相手に対して負荷をかける．そのためこの例においては正しい推定ができていると考えられる．

表 1　LLM よる負荷度と丁寧度の推定

話者	D	P	R_x	W_x	Po	発話
1	2	3	2	7	3	お願いがあって，電話したんだけど
2	2	3	1	6	1	うん．
1	2	3	1	6	3	今度の月曜日に
2	2	3	1	6	1	うん．
1	2	3	1	6	3	朝の 9 時にね
2	2	3	1	6	1	うん．
1	2	3	3	8	3	国立国語研究所に行ってね
2	2	3	1	6	1	うん．
1	2	3	4	9	3	私の代わりに
2	2	3	1	6	1	うん．
1	2	3	4	9	5	言語調査に関する実験に参加してはいただけないでしょうか?.
2	2	3	1	6	1	えー, へー, へ, えー.
1	2	3	3	8	3	しかも，韓国人と一緒に．

次に丁寧度を表す Po の推定結果を見てみると，「今度の月曜日に」や「朝の 9 時にね」のような，丁寧度において中立的な表現は 1〜5 の中間の 3 の値を得ており，一方で，「うん」や「えー」のような相槌は非常にくだけた表現であるため，低い値となっている．そして，「言語調査に関する実験に参加してはいただけないでしょうか?」という発話は敬体で丁寧な表現となっているが，この発話には 5 という高いスコアが推定されている．これらの結果から丁寧度 P_o の値も正しく推定できていると考えられる．

上記の結果から，フェイス侵害度 (W_x) と丁寧度 (Po) に対する LLM の推定はある程度正しくできていると判断できる．

W_x と Po それぞれに対して人手で正解データを作り，推定の精度を評価することも考えられるが，個別の推定の評価は今後の課題とし，今回はこれらの推定結果をもとにポライトネス理論を検証することに主眼を置いた．

4　推定結果の分析

今回対象としたコーパスに含まれる 120 の対話に含まれているすべての発話に対して，フェイス侵害度 (W_x) および丁寧度 (Po) の推定を行ない，カテゴリ毎に統計値を示した．結果を 表 2 に示す．

4.1　対話種別間の推定値の比較

本節では発話の丁寧度 Po とフェイス侵害度 W_x のそれぞれにおいて，対話種別間で値を比較する．

Po に着目すると，カテゴリ 1,2 に対して 3 は有意に高い数値になっていた．友人同士 (1,2) と比べると初対面 (3) の対話においては，対話の内容は雑談であっても丁寧な表現になることから納得できる結果となっている．同様に 4 の論文指導における Po の値は 1 や 2 と比較すると有意に高い値となっていた．5,6,7 は話者に学年の差がある場合 (5 と 7) と同級生の場合 (6) の間で，丁寧度が異なることが推測される．Po の推定値を見ると，5,7 は 6 よりも高い値となっているが，これらの間に有意差は無かった．また，同じ条件で性別の異なる 1 と 2，および 8 と 9 の Po を比較しても，大きな差は認められず，丁寧さにおいて性別による違いは，今回の実験結果からは認められなかった．Po における対話種別間の有意差のテスト結果 (p 値) を付録 B の図 3 に示す．

次に W_x に着目すると，カテゴリ 3 と比べ 11 は有意に高い値となっていた．初対面の女性同士という点は同じ設定であるが，対話の内容が雑談と討論という違いがある．カテゴリ 8,9 の依頼の対話においては，発話の内容に依頼が含まれることから W_x が高くなることが予想されたが，その予想に反し低い値となった．対話の内容を詳しく調べてみると，確かに依頼をしている発話においては W_x の値が高かったが，対話全体を通してみればそれ以外の発話においては補足説明や弁解をする発話が多くあり，対話全体としての W_x は高い値にはならなかった．また依頼をしている発話においても直接的な表現を用いる事は少なく，湾曲的な表現が多く使われており，そのために W_x の値は顕著には高くはならなかったと考えられる．カテゴリ 1 の雑談と比較したときにカテゴリ 4 の論文指導は高い値となっていた．これは，教師による論文の内容の確認や指示などによりフェイス侵害度の高い内容が話されていたことが要因として考えられる．ただ，平均値の間に差はあったものの，この差は有意ではなかった．W_x における対話種別間の有意差のテスト結果 (p 値) を付録 B の図 4 に示す．

4.2　W_x と Po の間の関係

3.2 節で述べたように，式 (1) 中の D と P は対話を通してほとんど変化はしない．そのため，W_x の変化は主に R_x の変化によって引き起こされていた．ポライトネス理論によるとこの W_x が大きくなるほど丁寧度は高くなるはずである．そのため，今回の実験結果においてもフェイス侵害度 W_x と丁寧度 Po は正に相関していることが予想される．そこで W_x と Po の間の相関係数を調べた．ノンパラメトリックの検定方法であるスピアマンの相係数を用い W_x と Po の相関係数を計算したところ，120 対話のうち表 2 中の「対話数*」で示す数の対話 (全 100 対話) において，W_x と Po は有意に相関していた．有意に相関していた対話における両者の相関係数の平均値を 表 2 中の「相関係数」に示す．

カテゴリ 4 の論文指導の 10 対話において，W_x と Po の間に有意な相関のあった 9 対話の中で 2 対話においては，相関係数がマイナスの値 (–0.647 と –0.295) であった．これらの対話内での発話を詳細に調べたところ，発話内容に指摘や指示が無い発話 (すなわち W_x が低い発話) においても教師が全体を通して敬体で丁寧な発話をしていた．そのためこの対話では，W_x と Po の間にフェイス侵害の見積もりの公式に当てはまらなかったことが考えられる．このような対話を正しく解析するためには，宇佐美 [7] がディスコース・ポライトネス理論として提案する時間軸上での基本状態の把握やそこからの相対的な変化も考慮する必要がある．

5　考察

今回は発話の解析の一つとして，LLM を用いてフェイス侵害度 W_x と丁寧度 Po を推定した．120 対話のうち 100 対話 (83.3%) においては，W_x と Po が有意に相関していたことから，実際の対話の多くにおいてフェイス侵害の度合いを見積もる公式が成り立っていることを確認できた．

フェイス侵害度や丁寧度の推定誤りもまだ多くあると考えられるが，一方で，推定誤りが理由ではなくフェイス侵害の度合いの見積もりと丁寧度が一致しない場合も考えられる．すなわち，推測されるよりも丁寧な (もしくは丁寧でない) 表現が使われていることもある．そのような不一致は，皮肉などの

表 2 BTSJ の中で使った対話カテゴリにおける統計値
(「対話数*」は W_x と Po が有意に相関している対話の数を表す)

カテゴリ	対話種別	対話数	対話数*	W_x 平均値	Po 平均値	相関係数
1	友人同士雑談 (男男)	10	9	6.05	2.31	0.559
2	友人同士雑談 (女女)	21	19	5.50	2.29	0.699
3	初対面雑談 (女女)	11	11	5.23	3.38	0.421
4	論文指導 (教師と学生)	10	9	6.79	3.32	0.447
5	断りの電話会話 (対先輩; 女女)	13	12	4.62	2.84	0.615
6	断りの電話会話 (対同級生; 女女)	13	9	4.15	2.73	0.607
7	断りの電話会話 (対後輩; 女女)	13	6	4.12	3.19	0.170
8	友人同士依頼の電話会話 (男男)	10	7	4.11	2.66	0.594
9	友人同士依頼の電話会話 (女女)	10	9	3.69	2.61	0.535
10	友人同士討論 (男女)	5	5	7.18	2.66	0.753
11	初対面討論 (女女)	4	4	7.76	3.49	0.123

言外の意図[2)]が含まれていることを知る手掛かりとなる．

今回の実験で推定した D,P,R_x および Po のうち，R_x と Po は発話の内容や文体からある程度推測できるが，D と P を発話のみから推測することは難しい．4.2 節で述べた教師と学生との関係などがその一例である．実際の対話が行なわれている場面においては話者同士のそれぞれの背景知識に基き D や P が決まることになるため，発話の内容だけでなく外部から情報を与えるべきであるとも考えられる．

今回用いた対話コーパスにおいては，話者間の関係が「教師-学生」や「依頼主-依頼相手」のように対等ではない対話も含まれている．1 そのため，P や R_x にも偏りがあることが推測されるが，4.1 節の分析では話者の区別をせずに対話全体での平均値を用いたため，話者毎の特徴が見えにくくなっていた．話者を分離した上で各推定値の比較をすることによって，それぞれの対話の特徴がよりはっきり見える可能性があるため今後の課題としたい．

6 おわりに

本研究ではポライトネス理論や DP 理論を活用することによって人間とのより自然な対話を実現する対話システムの構築を目指し，その第一歩として，発話の丁寧さや発話内で示す行為が相手にかける負荷の度合いを LLM を用いて発話内容から推定した．そして，推定した値からポライトネス理論で提唱されているフェイス侵害の見積もりの公式と丁寧度の関係を確認した結果，120 対話中 100 対話においてフェイス侵害度 R_x と丁寧度 Po が有意に相関していることを確認できた．

本稿で述べたような定量的なアプローチが可能となれば，単一の発話におけるポライトネスだけではなく，談話の中でのポライトネスの解析も可能となり，宇佐美 [7] がディスコース・ポライトネス理論として提案する時間軸上での基本状態の把握やそこからの相対的な変化，及び，話し手と聞き手の丁寧度の捉え方のギャップとそれが生み出すイン／ポライトネス効果も定量的に捉えられるようになるため，この方向で今後の研究を発展させていきたい．

本研究の応用方法としては，発話の生成および解釈において，それぞれ以下がある．

- 対話システムがユーザの発話を理解する際に，ユーザの発話から丁寧度 Po と フェイス侵害度 W_x を推定することによりユーザの言語的ストラテジーを推測する．たとえばもしフェイス侵害度から予想される丁寧度と推定した丁寧度との間に差分があれば，ユーザ発話に何か言外の意図が含まれているの可能性を把握し，それに基づく対話戦略を取ることができる．
- 対話システムが発話をする際に，命題+フェイス侵害度の計算を基に言語的配慮を決定し，常体と敬体などの丁寧さを意図的に操作する．

今回の研究は 1 つ目の推定に主眼を置いたが，2 つ目の発話の生成においても活用していきたい．

2) 例として，普段常体で話している夫婦間において「今日も飲み会なんですか」と敬体で話した際に含まれているような言外の意図

参考文献

[1] Vojtěch Hudeček and Ondřej Dušek. Are LLMs all you need for task-oriented dialogue? **arXiv preprint arXiv:2304.06556**, 2023.

[2] Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. PlatoLM: Teaching LLMs in Multi-Round Dialogue via a User Simulator. In **Proc. ACL (Volume 1: Long Papers)**, pp. 7841–7863, 2024.

[3] 東中竜一郎, 高橋哲朗, 稲葉通将, 斉志揚, 佐々木裕多, 船越孝太郎, 守屋彰二, 佐藤志貴, 港隆史, 境くりま, 船山智, 小室允人, 西川寛之, 牧野遼作, 菊池浩史, 宇佐美まゆみ. 対話システムライブコンペティション 6. 第 99 回人工知能学会 言語・音声理解と対話処理研究会 (第 14 回対話システムシンポジウム), 2023.

[4] 智子堀田, 薫堀江. 日本語学習者の「断り」行動におけるヘッジの考察 : 中間言語語用論分析を通じて. 語用論研究 = Studies in pragmatics, No. 14, pp. 1–19, 2012.

[5] ボイクマン総子, 森一将. 日本語におけるスピーチ・レベルの分析ー親しい間柄の依頼，勧誘，謝罪の力関係と負担度ー. 社会言語科学, Vol. 21, No. 1, pp. 225–238, 2018.

[6] 丹楠趙, タンナンチョウ, Dannan ZHAO. 依頼発話行為のポライトネス選択における負担度の影響. 語学教育研究論叢, Vol. 39, pp. 137–150, 03 2022.

[7] 宇佐美まゆみ. ポライトネス理論ー発話行為から談話へ. 大修館書店, 2024.

[8] Penelope Brown and Stephen C Levinson. **Politeness: Some universals in language usage**. No. 4. Cambridge university press, 1987.

[9] 宇佐美まゆみ監修. 『BTSJ1000 人日本語自然会話コーパス』、科研基盤研究 (A)「語用論的分析のための日本語 1000 人自然会話コーパスの構築とその多角的研究」（研究代表者：宇佐美まゆみ）及び、国立国語研究所、機関拠点型基幹研究プロジェクト「日本語学習者のコミュニケーションの多角的解明」（2016〜2021）, 2023.

A　プロンプト

ポライトネス理論においてフェイス侵害度 (Face Threatening Act) Wx は以下の式により定式化される．
Wx=D(S,H)+P(H,S)+Rx
ここで，D,P,Rx はそれぞれ以下のように定義される．
D (社会的距離): 話し手と聞き手の親密さや関係性に基づいて，1 (近い関係) から 5 (遠い関係) まで．
P (力): 聞き手が話し手に対して持つ影響力や地位．1 (話し手が強い) から，3 (対等)，5 (聞き手が強い)．
Rx (負荷度): 発話内容が聞き手に与える負担や影響の度合い．1 (ほぼ負荷なし) から 5 (高い負荷)．
以下のそれぞれ発話において，D,P,Rx の値を推定してください．
解説は不要です．D,P,Rx の値のみを出力してください．

図 1　フェイス侵害度を推定するプロンプト

以下の日本語のそれぞれの発話において丁寧さのスコアを 1(丁寧ではない) から 5(非常に丁寧) までの 5 段階で判断し，ID と丁寧さのスコアをカンマ区切りで出力してください．
—
ID \t 話者 \t 発話内容

図 2　丁寧度を推定するプロンプト

B　対話種別間の差異の検定結果

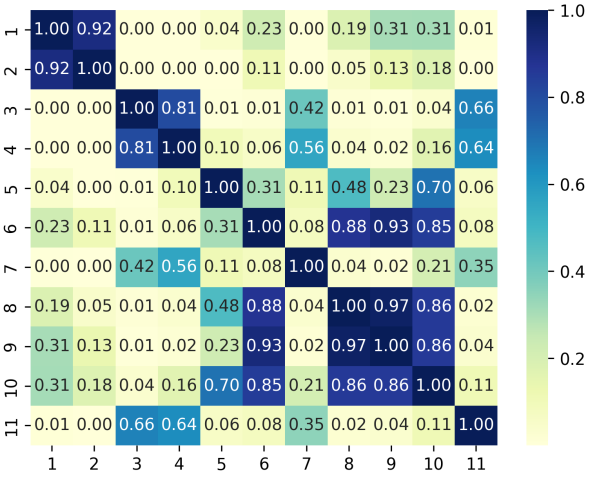

図 3　対話種別間の Po の差 (U 検定結果の p 値)

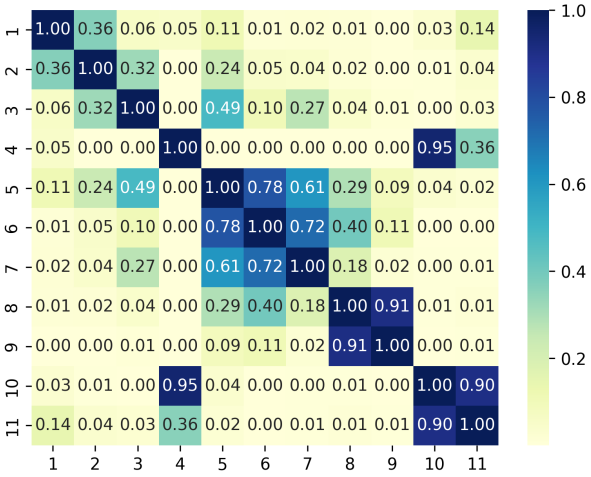

図 4　対話種別間の W_x の差 (U 検定結果の p 値)