非タスク指向型対話における話題の深さ推定モデルの構築

三野星弥 [1]　伴碧 [1]　吉川雄一郎 [1]　石黒浩 [1]
[1] 大阪大学大学院 基礎工学研究科
{mitsuno.seiya, ban, yoshikawa, ishiguro}@irl.sys.es.osaka-u.ac.jp

概要

非タスク指向型対話システムが，話題の深さを考慮しながらユーザに応じて適切な話題を提示することは，より人間らしく自然な対話を実現するために重要であると考えられる．話題の深さを考慮しながら対話できるシステムの実現のための第一歩として，本研究では，様々な話題について，その深さを推定できるモデルを構築することを目指す．我々は，独自に作成した話題の深さデータセット[1)]を用いて，GPT-3.5 をファインチューニングし，話題の深さ推定モデルを構築した．そして，評価実験を通して，本モデルが，ベースラインと比較して，高い精度で話題の深さを推定できることを確認した．

1　はじめに

近年，娯楽 [1]，福祉 [2]，教育 [3] といった様々な領域で，人の代わりの雑談相手となる，非タスク指向型対話システムの活躍が期待されている．このような非タスク指向型対話システムがユーザに提示し得る話題の候補は，ユーザにとって浅く取るに足らないものから，自身の本質にかかわるような深いものまで多岐に渡る [4, 5]．この話題の深さという概念は，主に自己開示研究の分野で用いられており，対話における重要な要素の一つであると考えられている [6]．人は一般に，深い話題を他者に対してあまり話したがらない傾向にあると言われている [4, 5, 7]．また，対話相手との関係性の深化に伴って，人は徐々に深い話題を話すようになるとも言われている [6, 7]．以上より，非タスク指向型対話システムが，話題の深さを考慮しながらユーザに応じて適切な話題を提示することは，より人間らしく自然な対話を実現するために重要であると考えられる．

話題の深さを考慮しながら対話できるシステムの実現のための第一歩として，本研究では，様々な話題について，その深さを推定できるモデルを構築することを目指す．具体的には，まず，大規模言語モデルで拡張した話題リストを用意し，人間のアノテータに深さラベルを付与させ，話題の深さデータセットを作成する．次に，このデータセットを用いて，OpenAI[2)] の GPT-3.5-turbo をファインチューニングし，話題の深さを推定できるモデルを構築する．そして，構築したモデルが，ベースラインと比較して，高い精度で話題の深さを推定できることを評価実験を通して確認する．

2　データセット

我々は，大規模言語モデルで拡張した話題リストを用意し，各話題に対して人間のアノテータが話題の深さのラベルを付与したものを，ファインチューニングに使用するデータセットとした．

2.1　話題リスト

話題リストの作成フローを図 1 に示す．まず，初期データセットとして，話題の深さについての調査研究 [5, 4]，人–対話システム間で実際に行われた対話で登場した話題 [8]，不安感や不眠症等のメンタルヘルス関連項目（PHQ[9, 10]，GAD[11, 10]，ISI[12]）から 106 話題を用意した（e.g.,“あなたの一番好きな食べ物は何ですか？”）．

次に，大規模言語モデルを用いたデータセット拡張手法である *Evol-Instruct* [13, 14] を参考に，用意した話題リストを拡張した．具体的には，話題をより深くさせる指示（In-Depth Evolving）と話題の幅（種類）を増加させる指示（In-Breadth Evolving）を与え，GPT-4 を用いて，話題リストを拡張した（プロンプトは付録 A.1 を参照）．また，拡張した話題リストを OpenAI の text-embedding-ada-002 でベクトル化し，類似度の高い（コサイン類似度が 0.95 以上となる）話題を削除した．この処理を繰り返し，最終的に，1,638 話題を用意した．

1）本研究で作成したデータセットは GitHub（https://github.com/IshiguroLab/TopicDepthDataset/）にて公開している．

2）https://openai.com/

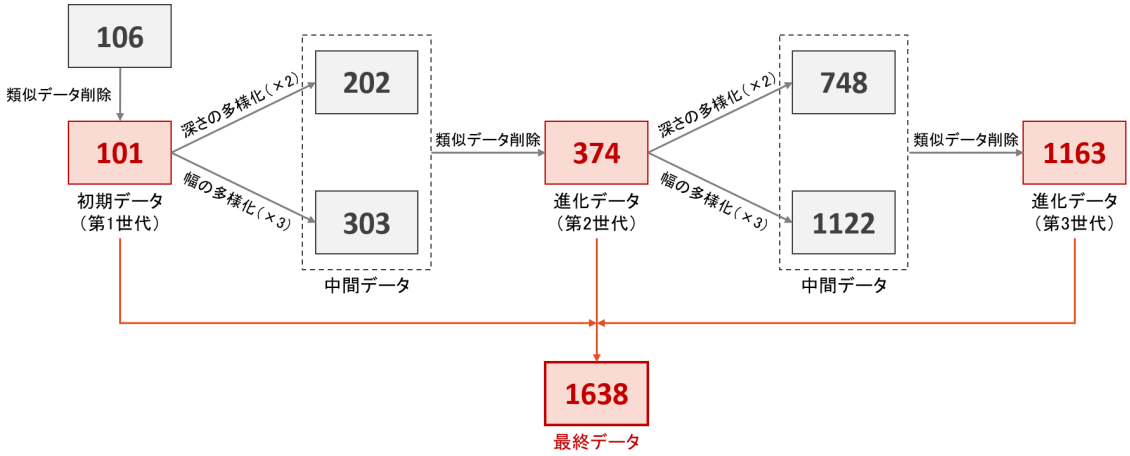

図 1　話題リストの作成フロー

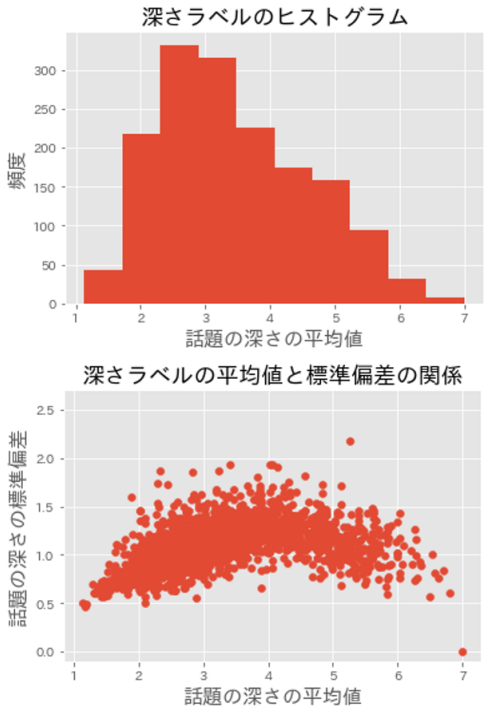

図 2　話題の深さデータセットの分布

2.2　深さラベル

用意した 1,638 話題に対し，人手で深さラベルを付与した．具体的には，クラウドソーシングサービスを用いて，各話題につきアノテータ 30 名に深さラベルを付与させ，その代表値（平均値）を最終的な深さラベルとした．深さラベルは，先行研究 [5, 6, 7, 15] を参考に，“あなたが「初対面の人とメッセージアプリ上で雑談をしている」という状況を想定してください．この状況において，以下の質問文は，質問しやすいもの（浅く，取るに足らないようなもの）でしょうか，あるいは，質問しにくいもの（深く，人の本質に関わるようなもの）でしょうか．”という教示のもと，1：とても質問しやすい（浅く，取るに足らない）～7：とても質問しにくい（深く，人の本質に関わる）の 7 件法で付与させた．この際，“質問文の意味が読み取れない”，という項目を同時に用意し，全体の 10%以上（3/30 名以上）のアノテータがこの項目を選択した場合，その話題は除外した（38 話題が該当）．また，Maniaci & Rogge の Directed Questions Scale (DQS) [16] を参考に，質問中にチェック項目（e.g.,“この項目は 1 を選択してください”）を複数設け，これらの項目に一つでも違反したアノテータのデータは平均値の算出時に除外した（728 データ=4 人 ×182 話題/人が該当）．

2.3　データの分布

話題の深さデータセットのヒストグラムを図 2 上部に示す．最終的に採用した 1,600 話題の深さラベルは 1 点から 7 点まで幅広く分布しており，特に 3 点付近に多く分布していることが確認された．また，深さラベルの平均値と標準偏差を確認すると，浅い話題や深い話題と比較して中程度の深さの話題が標準偏差が高い傾向にあり，アノテータによって評価する深さにばらつきが出やすいことが示唆された（図 2 下部）．

また，表 1 に話題と深さラベルの例を示す．浅い話題としては，好きな食べ物や趣味といったポジティブな内容のものや，今日の天気といった客観的情報に関するものが多くみられた．その一方で，深い話題としては，精神的に疲れを感じているかや自己嫌悪を感じた経験などのネガティブな内容のものや，金銭や性に関するものが多くみられた．

表 1　話題と深さラベルの例

話題	深さ
今日の天気はどうですか？	1.20
あなたの一番好きな食べ物は何ですか？	1.53
あなたの趣味は何ですか？	2.07
最近どこかに旅行の予定はありますか？	2.41
今日は何か面白いことがありましたか？	2.93
あなたの人生のモットーは何ですか？	3.63
最近起きた困った出来事は何ですか？	3.93
友人に求める理想の価値観と性格は何ですか？	4.53
最近、精神的に疲れを感じていますか？	5.13
一番自己嫌悪を感じた経験は何ですか？	5.76
預貯金の目標額はいくらですか？	6.21
あなたは株でいくらくらい儲けていますか？	6.63
最近、パートナーとのセックスライフに何か変化はありましたか？	7.00

3　話題の深さ推定モデル

作成した話題の深さデータセットを用いて，OpenAI の gpt-3.5-turbo-0613 をファインチューニングし，話題の深さ推定モデルを構築した．具体的には，話題の深さデータセットを訓練用，テスト用に 9:1 に分割し，OpenAI の API を用いて，ファインチューニングを実施した[3)]．ファインチューニング時のプロンプトは，深さラベル付与の際に用いた教示と類似したものとなるようにした（付録 A.2 参照）．また，ファインチューニング時のハイパーパラメータは，OpenAI が事前に用意した，訓練データに合わせて自動で設定される値のセットを用いた（batch_size: 2, learning_rate_multiplier: 2, n_epochs: 3）．

4　評価実験

4.1　ベースライン

gpt-3.5-turbo-0613 をファインチューニングして作成した話題の深さ推定モデル（以下，Finetuned GPT-3.5）を評価するための実験を実施した．ベースラインとして，Zero-shot の gpt-3.5-turbo-0613（以下，GPT-3.5），Zero-shot の gpt-4-1106-preview（以下，GPT-4），7-shot の gpt-4-1106-preview（以下，GPT-4 (7-shot)）を用意した．GPT-4 (7-shot) は，Finetuned GPT-3.5 の訓練データを浅い順に並べ替え，等間隔にピックアップした 7 データをプロンプトに含め，In-context Learning を行った．

3）　ファインチューニングの所要時間は約 75 分で，費用は約 1,500 円であった．

表 2　実験結果

	MAE	$RMSE$	R^2	r
GPT-3.5	1.76	2.07	-2.13	0.63
GPT-4	0.82	1.08	0.15	0.80
GPT-4 (7-shot)	0.64	0.87	0.45	0.84
Finetuned GPT-3.5	**0.25**	**0.33**	**0.92**	**0.96**

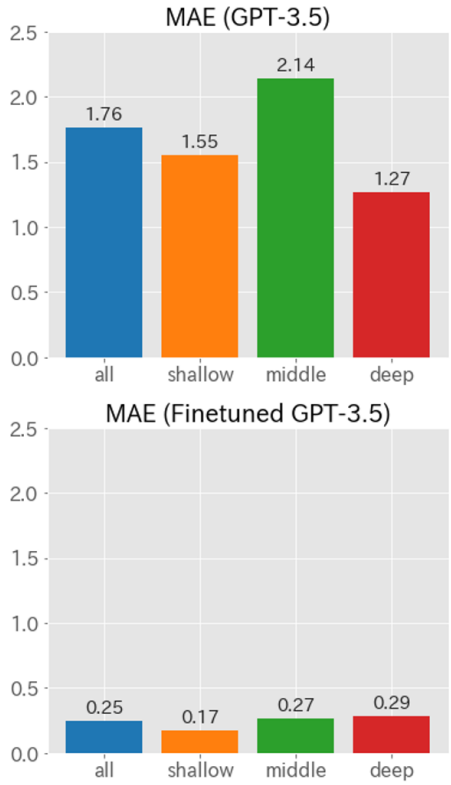

図 3　話題の深さ群毎の推定精度

4.2　結果

実験の結果を表 2 に示す．テストデータセットにおいて，MAE, RMSE, R^2, Peason's r の全ての指標で，Finetuned GPT-3.5 が最も高い性能を示したことが確認された．

また，話題の深さ毎の推定精度を評価するため，テストデータセットを，深さラベルの四分位数をもとに，浅い話題群（shallow; 0~25%, 1.00~2.50 点），中程度の深さの話題群（middle; 25~75%, 2.50~4.23 点），深い話題群（deep; 75~100%, 4.23~7.00 点）に分割した．そして，GPT-3.5 と Finetuned GPT-3.5 それぞれについて，各話題群における推定精度（MAE）を評価した（図 3）．その結果，GPT-3.5 は，中程度の深さの話題群の推定精度が，浅い話題群，深い話題群と比較して低いことが確認された．一方で，Finetuned GPT-3.5 は，全ての話題群において高い精

度で話題の深さを推定できており，特に浅い話題群での推定精度が高いことが確認された．

5 おわりに

本研究では，独自に作成した話題の深さデータセットを用いて，GPT-3.5 をファインチューニングし，話題の深さ推定モデルを構築した．そして，評価実験を通して，本モデルが，ベースラインと比較して，高い精度で話題の深さを推定できることを確認した．本モデルの課題として，話題の深さに影響し得る様々な要因を十分に考慮できていないことが挙げられる．具体的には，対話相手の性別や関係性（家族，友人，仕事の同僚）などが話題の深さに影響を与えると考えられているため [7, 17]，これらの要素を考慮できるようにする必要がある．今後は，本モデルの拡張を行うとともに，本モデルを用いて，話題の深さを考慮しながら対話できる非タスク指向型対話システムの実現を目指す予定である．

謝辞

本研究は JSPS 科研費 JP20H00101, JP19H05691, JP23KJ1462, Society 5.0 実現化研究拠点支援事業（グラント番号: JPMXP0518071489）の助成を受けた．

参考文献

[1] 杉山弘晃, 古賀光, 西島敏文. 移動体から見える風景を話題とする雑談対話システム. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 2N5OS7a04–2N5OS7a04, 2022.

[2] Joost Broekens, Marcel Heerink, Henk Rosendal, and Others. Assistive social robots in elderly care: a review. **Gerontechnology**, Vol. 8, No. 2, pp. 94–103, 2009.

[3] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. **Human–Computer Interaction**, Vol. 19, No. 1-2, pp. 61–84, June 2004.

[4] 三野星弥, 伴碧, 吉川雄一郎, 石黒浩. 初対面ロボットにユーザは何を話すのか—対話場面における話題の深さの検討—. 2024.

[5] 三野星弥, 伴碧, 吉川雄一郎, 石黒浩. ロボットとの対話意欲と話題の深さの関係のモデル化. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 2O5OS2a01–2O5OS2a01, 2023.

[6] Irwin Altman and Dalmas A Taylor. Social penetration: The development of interpersonal relationships. Vol. 212, , 1973.

[7] 丹羽空, 丸野俊一. 自己開示の深さを測定する尺度の開発. パーソナリティ研究, Vol. 18, No. 3, pp. 196–209, 2010.

[8] 三野星弥, 吉川雄一郎, 伴碧, 石黒浩. 友人グループ内での長期間利用による他者情報のやり取りを行う日常対話チャットボットの評価. 人工知能学会論文誌, Vol. 37, No. 3, pp. IDS–I_1–14, 2022.

[9] Kurt Kroenke, Robert L. Spitzer, and Janet B. Williams. The PHQ-9: validity of a brief depression severity measure. **J. Gen. Intern. Med.**, Vol. 16, No. 9, pp. 606–613, September 2001.

[10] 村松公美子, Others. Patient health questionnaire (PHQ-9, PHQ-15) 日本語版および generalized anxiety disorder-7 日本語版-up to date. 新潟青陵大学大学院臨床心理学研究, Vol. 7, pp. 35–39, 2014.

[11] Robert L Spitzer, Kurt Kroenke, Janet B W Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the GAD-7. **Arch. Intern. Med.**, Vol. 166, No. 10, pp. 1092–1097, May 2006.

[12] Cèlyne H. Bastien, Annie Vallières, and Chales M. Morin. Validation of the insomnia severity index as an outcome measure for insomnia research. **Sleep Med.**, Vol. 2, No. 4, pp. 297–307, July 2001.

[13] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. WizardLM: Empowering large language models to follow complex instructions. April 2023.

[14] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. WizardMath: Empowering mathematical reasoning for large language models via reinforced Evol-Instruct. August 2023.

[15] 飯長喜一郎. グループ合宿における自己開放性. 東京大学教育学部紀要, Vol. 17, pp. 77–84, February 1978.

[16] Michael R Maniaci and Ronald D Rogge. Caring about carelessness: Participant inattention and its effects on research. **J. Res. Pers.**, Vol. 48, pp. 61–83, February 2014.

[17] 榎本博明. 自己開示の心理学的研究 北大路書房. 1997.

A プロンプト

A.1 Evol-Instruct によるデータ拡張

A.1.1 In-Depth Evoloving

話題をより深くさせる指示（In-Depth Evolving）は，以下の通りである．なお，{given_question} には，質問文が挿入される．

指示
・以下の指示に従って，対話における質問文の作成をしてください．
・具体的には，元となる質問文から，より深くプライベートな質問文（新しい質問文）を作成してください．
・元となる質問文に対する回答と新しい質問文に対する回答は，類似したものにならないようにし，異なる回答が得られるものにしてください．
・新しい質問文は自然な日本語になるようにしてください．
・質問は 30 文字以内にしてください．
・質問する内容は一つにし，クエスチョンマークは 1 度しか使用しないでください．

元となる質問文
{given_question}

新しい質問文

A.1.2 In-Breadth Evoloving

話題の幅（種類）を増加させる指示（In-Breadth Evolving）は，以下の通りである．なお，{given_question} には，質問文が挿入される．

指示
・以下の指示に従って，対話における質問文の作成をしてください．
・具体的には，元となる質問文から，異なる話題で類似した質問文（新しい質問文）を作成してください．
・元となる質問文に対する回答と新しい質問文に対する回答は，類似したものにならないようにし，異なる回答が得られるものにしてください．
・新しい質問文は自然な日本語になるようにしてください．
・質問は 30 文字以内にしてください．
・質問する内容は一つにし，クエスチョンマークは 1 度しか使用しないでください．

元となる質問文
{given_question}

新しい質問文

A.2 話題の深さの推定

ファインチューニングおよび深さの推定に使用したプロンプトは以下の通りである．なお，ファインチューニング時には，出力結果として，質問文の深さを表す数値もデータに含めている．また，{question} には，質問文が挿入される．

指示
・「初対面の人とメッセージアプリ上で雑談をしている」という状況において，以下の質問文は，質問しやすいもの（浅く，取るに足らないようなもの）か質問しにくいもの（深く，人の本質に関わるようなもの）かを評価してください．
・出力は，1.00 点が「とても質問しやすい（浅く，取るに足らない）質問文」， 7.00 点が「とても質問しにくい（深く，人の本質に関わる）質問文」となるようにしてください．
・出力は，1〜7 の間の小数点以下第 2 位までの実数で評価してください．

質問文
{question}

出力