認知症高齢者の発話意図推定に基づく注意発話検出システムの開発 —帰宅願望や不安などを特定するコーパス構築—

有國 開成[1] 神崎 享子[2] 井佐原 均[1]
[1]追手門学院大学 心理学部 心理学科 人工知能・認知科学専攻 [2]群馬県立女子大学文学部文化情報学科

概要

本研究では、認知症高齢者の「帰宅願望」や「不安」といった注意すべき発話を特定するシステムの構築を目的とした。介護施設で収集した対話データを基に、実データの収集が困難な環境下でも活用可能なデータ拡張手法を活用し、日本語 BERT を基盤とした発話意図推定モデルを構築した。特に、ChatGPT(GPT-4o)を用いた疑似データ生成により、少数データクラスのデータ不足を補完し、分類精度の向上を実現した。

評価の結果、疑似データ生成がデータ不足を補う有効な手法である一方、施設特有の文脈や固有表現への完全な対応には最低限の実データが不可欠であることが判明した。本研究は、疑似データ生成を活用したコーパス構築の有効性を示すとともに、施設ごとに適応可能なモデル設計の方向性を示す一助となるものである。

1 はじめに

1.1 研究の背景

高齢化が進む現代社会において、認知症は医療および社会的課題として注目されている。認知症は、記憶障害や判断力の低下といった認知機能の障害に加え、不安や混乱、攻撃性など情緒的・行動的な変化を伴うことが多い。特に、介護施設における日常的な対話では、「帰宅願望」や「不安」といった特定の発話が頻繁に見られる。これらの発話に適切に対応しない場合、認知症高齢者は心理的ストレスを抱え、パニックや攻撃的行動を引き起こす場合がある【1】。

介護者は、認知症高齢者の発話意図を正確に理解し、迅速かつ適切に対応することが求められる。しかし、過重労働や専門知識の不足がその妨げとなる場合が少なくない。このような状況において、AI を活用した発話意図推定システムは、介護現場の負担を軽減し、認知症高齢者の生活の質を向上させる可能性を秘めている。

本研究では、これらの課題に取り組み、注意すべき発話を特定するシステムの開発を目指す。

1.2 研究の目的

本研究の目的は、認知症高齢者の「帰宅願望」や「不安」といった注意を要する発話を特定するためのシステムを構築することである。特に、実データの収集が困難な環境においても、データ拡張手法を活用することで発話意図を推定可能なモデルを開発することを目指す。

具体的には、以下の課題に取り組む。

1. 注意すべき発話意図を分類するためのコーパスを構築する。

2. GPT-4o を用いて少数データのクラスを補完し、モデルの分類精度を向上させる。

3. データ不足環境下でのモデル適用性を検証し、施設特有の文脈や固有表現を考慮したモデル設計の必要性を明らかにする。

本研究は、疑似データ生成がデータ不足を補う有効な手法であることを示すとともに、施設ごとに適応可能なデータセットの構築が重要であることを明確にするものである。また、これにより認知症高齢者ケアにおける AI 技術の実用化に向けた具体的な指針を提示することを目指す。

2 関連研究

2.1 認知症高齢者の発話解析

認知症高齢者と介護者の対話における発話解析は、心理的ケアや行動管理の向上に寄与することが知られている【2】。特に「帰宅願望」や「不安」といった発話は、適切な対応がなされない場合、高齢者の心理的負担を増加させることが指摘されている。これらの発話を自動的に検出する技術の開発は、介護現場の負担軽減と認知症ケアの質の向上において重要である。

2.2 データ拡張手法

少数データクラスの課題を克服するため、データ拡張手法が広く研究されている。日本語 WordNet【3】を用いた同義語置換は、データの多様性を向上させる方法として広く活用されている。一方で、大規模言語モデル(Large Language Model:LLM)による疑似データ生成は、特定のタスクにおける実データ不足を補完し、モデル性能を向上させる手法として注目されている【4】。

3 データ収集とラベリング

3.1 データ収集

本研究では、認知症高齢者の発話意図を特定するため、先行研究【2】で収集された介護施設での対話データを基礎データとして利用した。このデータは、

社会福祉法人が運営する認知症に優しいグループホームで、軽度から中等度の認知症高齢者の日常的な対話を記録したものである。

録音は、介護者と高齢者に IC レコーダーを装着し、施設内の会話を収集する形式で実施され、約 9,600 発話が記録された。データは聴き取れない部分や不要な発話を除去し、高齢者の発話部分を抽出して最終的に 3,973 発話に絞り込まれた。なお、データ収集は大学および社会福祉法人の倫理委員会の承認を得た上で実施され、対象者の同意を得た録音データは研究目的にのみ使用された。

3.2 ラベリング手法

収録した発話データには、以下の 5 つの発話意図ラベルを付与した。

1. **帰宅願望**: 自宅や過去の生活への帰属意識を示す発話。

2. **不安**: 混乱や将来への不安を表す発話。

3. **自己否定**: 自己評価の低下や否定的な感情を示す発話。

4. **活動拒否**: 日常生活や活動への拒否的な態度を表す発話。

5. **不満**: 環境や対応に対する不満を表す発話。

これらに該当しない発話は「その他」と分類し、特別な対応を要しないものとして扱った。ラベル付けは、研究者自身が行い、基準の一貫性を保つために、曖昧なケースについては複数の研究者で協議を行い、参考文献【1】の基準を参考に決定した。

3.3 データ分布

ラベル付けの結果、各クラスの発話数は以下の通りである。

- **帰宅願望**: 7 発話
- **不安**　　: 8 発話
- **自己否定**: 29 発話
- **活動拒否**: 11 発話
- **不満**　　: 41 発話
- **その他**　: 3,875 発話

これらのデータ分布から、「帰宅願望」や「不安」といった少数クラスのデータ不足が課題であることが明らかとなった。

4. モデル構築とデータ拡張手法

本章では、日本語 BERT を基盤とした発話意図推定モデルの構築手法および、少数クラスの課題を克服するためのデータ拡張手法について述べる。本研究では、GPT-4o を用いた疑似データ生成を主軸とし、日本語 WordNet を用いた同義語置換手法についても補足的に検討した。

4.1 日本語 BERT を基盤としたモデル構築

日本語 BERT（cl-tohoku/bert-base-japanese-v3）を基盤とする発話意図推定モデルを構築した。このモデルは、認知症高齢者の対話データに特化し、特に「帰宅願望」や「不安」など少数データクラスの分類精度向上を目的としている。

4.1.1 データ構成

本研究で使用したデータセットは以下の通りである。

1. **訓練データ**:

・GPT による疑似データ生成: 少数クラス(例:「帰宅願望」「不安」)を補完するため、GPT-4o を用いて生成した発話データ。

・**元データ（その他ラベル）**：元データ中の「その他」ラベルをそのまま使用。

2.テストデータ:

テストデータには、元データに含まれる重要ラベル（「帰宅願望」「不安」「自己否定」「活動拒否」「不満」）を持つ発話と、GPT-4o による疑似データ生成で作成したテスト用データを含む計 183 発話を使用した。**表 1** に各クラスごとの件数を示す。

表 1: テストデータにおけるクラス別件数

クラス	件数
帰宅願望	32
不安	30
自己否定	39
活動拒否	31
不満	51
合計	183

なお、テストデータと訓練データの重複を防ぐため、元データの『帰宅願望』や『不安』などの重要ラベルに該当する発話は、訓練データから完全に除外した。これにより、モデルの学習と評価の際にデータリークが発生しないよう配慮した。

4.1.2 モデル設計

日本語 BERT を基盤モデルとして採用し、発話意図分類タスクに適応させた。モデル設計の要点は以下の通りである。

- **基盤モデル**：日本語 BERT（cl-tohoku/bert-base-japanese-v3）
- **出力層**：全結合層を追加し、6 つの発話意図クラス（「帰宅願望」「不安」「自己否定」「活動拒否」「不満」「その他」）を分類。
- **データのトークン化**：AutoTokenizer を使用し、最大 128 トークンに制限してトークン化。長すぎる発話は切り捨て、短い発話にはゼロパディングを適用。

4.1.3 学習と評価

学習設定や実装環境の詳細は付録に記載するが、以下に概要を示す。

- **学習設定**：学習率、エポック数、バッチサイズなどのハイパーパラメータ設定。
- **クロスバリデーション**：Stratified K-Fold（5 分割）を採用し、ラベル分布を保ちながらモデルの汎化性能を評価。
- **評価指標**：Accuracy、Precision、Recall、F1 スコアを使用。

4.2 GPT による疑似データ生成

本研究では、少数クラス（例：「帰宅願望」「不安」）のデータ不足を補うため、GPT-4o を用いて疑似データを生成した。参考文献【1】を基にプロンプトを設計し、元データを参照しない生成手法と元データを活用した追加生成手法を併用した。

4.2.1 元データを参照しない生成手法

参考文献【1】で示された「注意すべき発話」の基準を基に、各クラスの抽象的な特徴を反映したプロンプトを設計し、GPT-4o に入力してデータを生成した。この手法では、各クラスごとに 200 件の発話を生成し、元データにはない表現や文脈を含む多様性の高い訓練データを構築した。

4.2.2 元データを参照した追加生成手法

初期生成データで明らかになった課題を解決するため、元データを活用して生成データの品質向上を図った。この手法では、以下の改善を行った。

1.課題の特定:

・**固有表現の不足**：例：「フクシ村」などの施設名や固有表現が生成データに含まれていない。

・**短い発話や曖昧な表現の対応不足**：「ダメだ」「今日入らん」などの短い発話が不十分。

・**クラス間の曖昧さ**：「不安」や「自己否定」など、類似クラス間での誤分類が発生。

2.プロンプトの改良:

元データの具体例を GPT-4o に読み込ませ、文法構造や文脈、固有表現を反映したプロンプトを再設計。特に短い発話や施設名を含む文脈を重視し、自然かつ文脈特化型の発話を生成可能な形に調整した。

3.追加生成:

各クラスごとにさらに 200 件の発話を生成し、初期生成データと統合して最終的な訓練データを**表 2** のように拡張した。

表 2: 初期生成データと追加生成データの構成

クラス	元データ	初期生成データ（元データ参照なし）	追加生成データ（元データ参照）	合計
その他	3,877	0	0	3,877
帰宅願望	0	200	200	400
不安	0	200	200	400
自己否定	0	200	200	400
活動拒否	0	200	200	400
不満	0	200	200	400
合計	**3,877**	**1,000**	**1,000**	**5,877**

4.3 実験結果と考察

4.3.1 モデル性能評価

本生成データを活用したモデルのテストデータに対する評価結果を**表 3** に示す。

表 3: テストデータに対するモデル評価結果

クラス	Precision	Recall	F1	件数
その他	0.00	0.00	0.00	0
不安	1.00	0.97	0.98	30
不満	0.96	1.00	0.98	51
帰宅願望	1.00	0.94	0.97	32
活動拒否	1.00	0.97	0.98	31
自己否定	1.00	0.97	0.99	39
Weighted Avg	0.99	0.97	0.98	183

結果の概要:

- **高い性能**: Precision、Recall、F1 スコアがほとんどのクラスで 0.97 以上を達成し、全体として Weighted Avg の F1 スコアは 0.98 に到達した。
- **Accuracy**: モデルの全体的な Accuracy は **0.97** で、テストデータに対して高い正解率を示した。
- **例外**: 「その他」クラスはデータが含まれていないため、評価値が 0.00 となった。

4.3.2 誤分類の傾向

生成データを活用したモデルは高い性能を示したが、「自己否定」や「不満」などに含まれる、文脈に応じて意味が変わる発話は、冗談や世間話と真剣な発話を区別できず、誤分類されやすいことが確認された。

4.3.3 考察

実験結果を基に、以下の知見が得られた。

1. **生成データの有効性**
   GPT-4o を用いた疑似データ生成は、少数クラスのデータ不足を補い、モデルの分類性能を大幅に向上させる有効な手法であることが確認された。
2. **改良の必要性**
   固有表現や短い発話への対応が必要であり、施設ごとの文脈や特徴を反映したデータ生成の重要性が示唆された。
3. **汎化性能の限界**
   クラス間の曖昧な発話への対応や、新しい文法的・意味的表現の生成には、プロンプト設計やコーパスの改良のさらなる工夫が求められる。

5. 結論

本研究では、認知症高齢者の「帰宅願望」や「不安」など注意すべき発話を特定する発話意図推定モデルを構築するため、GPT-4o を用いた疑似データ生成手法を検討した。以下に本研究の主要な成果をまとめる。

1. **生成手法の有効性**
   GPT-4o を用いて少数クラスのデータ不足を補い、モデル性能を向上させた。
2. **コーパス設計の方向性**:
   施設特有の表現を反映したコーパス設計の必要性を示し、施設ごとに適応可能なデータセットの構築が重要であることを明らかにした。

参考文献

1. 鈴木 みずえ (2017). 『認知症の人の気持ちを理解するための聞き方・話し方』. 池田書店. ISBN:978-4-262-14591-4. https://www.ikedashoten.co.jp/book-details.php?isbn=978-4-262-14591-4.
2. Kanzaki, K., & Isahara, H. (2022). "Analysis of Dialogue Between Caregivers and Recipients for a Communication Robot." Proceedings of the 7th International Conference on Business and Industrial Research (ICBIR2022), pp. 446–450. DOI: https://doi.org/10.1109/ICBIR54589.2022.9786451.
3. Bond, F., Isahara, H., Fujita, S., Uchimoto, K., & Kuribayashi, T. (2009). "Enhancing the Japanese WordNet." ALR7 Workshop on Asian Language Resources, pp. 1–8. https://aclanthology.org/W09-3401/
4. 藤井巧朗, 勝又智 (2024). 『日本語タスクにおけるLLMを用いた疑似学習データ生成の検討』言語処理学会第 30 回年次大会論文集, P8-8. https://yuji96.github.io/NLP2024-miniconf/poster_P8-8.html.

A 付録

A.1 実装環境

本研究の実験は以下の環境で実施した。

1.使用ハードウェア: Google Colab, GPU (Tesla T4)

2.使用ソフトウェア・ライブラリ:

・Python 3.9

・PyTorch 1.10

・Hugging Face Transformers 4.20

・その他: fugashi, ipadic, Datasets (v2.14.4), evaluate

3.学習条件:

・学習率: 2e-5

・エポック数: 5

・バッチサイズ: 16

・損失関数: クロスエントロピー損失関数（クラスウェイト適用）

・最適化手法: AdamW

A.2 生成条件（GPT-4o へのプロンプト）

1.元データを参照しないプロンプト:
参考文献[4]に記載されている注意すべき発話の基準を基にプロンプトを設計した。。詳細は以下のページを参照。

- **『帰宅願望/不安』①（参考文献[4], P146）**
- **『自己否定/活動拒否』②（参考文献[4], P152）**
- **『不満/攻撃的な言動』③（参考文献[4], P176）**

2.元データを参照したプロンプト:
元データを参照しないプロンプトで生成したデータを学習させたモデルを、テストデータで評価した際に発生した誤分類を基に作成しました。以下の手順でプロンプトを設計した：

1. **誤分類例の抽出**:
テストデータにおける誤分類例を分析し、元データに特有の固有表現や短いフレーズが誤分類の原因となっていることを確認した。
2. **誤分類例の説明**:
誤分類例に含まれる文脈や表現を GPT-4o に説明し、これを反映したプロンプトを作成した。

A.3 評価指標

本研究では以下の指標を用いてモデル性能を評価した：

- **Accuracy**: 全推論結果のうち正解の割合。
- **Precision**: 各クラスの予測結果における正解率。
- **Recall**: 各クラスにおける実際の正解が検出された割合。
- **F1 スコア**: Precision と Recall の調和平均。
- **Weighted Avg**: クラスごとのサンプル数に基づいて重み付けした平均値。

A.4 詳細な結果と GitHub リポジトリ

本研究の再現性を確保し、詳細な結果を共有するために、以下のような内容を GitHub リポジトリに公開している。

1.日本語 WordNet によるデータ拡張のテスト結果

・同義語置換による拡張データの効果。

・テストデータでの評価分析。

2.元データを参照しないプロンプトで作成したテストデータの評価結果

・誤分類例の詳細とその傾向。

その他、詳細なプロンプト例、評価結果、データ生成の手法についても公開している。
GitHub リポジトリ:
https://github.com/kaiseiarikuni/dementia-speech-intent-estimation