J-027

大規模言語モデルによる発話文テキストを用いた認知症スクリーニング手法の提案

A dementia screening method from transcribed text of utterances using a large language model

朽名 彩音[1)] 伊藤 有生[2)] 加藤 昇平[2)] 佐久間 拓人[2)] 大嶽 れい子[3)]
Ayane Kutsuna Yuki Ito Shohei Kato Takuto Sakuma Reiko Ohdake
桝田 道人[4)] 伊藤 信二[3)] 渡辺 宏久[3)]
Michihito Masuda Shinji Ito Hirohisa Watanabe

1 はじめに

日本の高齢者人口は増加の一途を辿っている．2024 年現在，65 歳以上の人口は 3623 万人となり，総人口の 29.1 %を占めている [1]．さらに，国立社会保障・人口問題研究所の推計によると，高齢者の割合は 2040 年には 34.8 %，2045 年には 36.3 %になると見込まれている．また，高齢者の増加に伴って認知症患者も増加することが見込まれており，日本全国の 65 歳以上の人口における認知症患者の数は 2020 年に 964 万人，2070 年に 2828 万人に達すると予測されている．

認知症が発生する要因は複数存在し，外科的な処置から症状が改善する例もあれば，ホルモンの異常を治す等内科的な処置から症状が改善する例もある．そのため，発症の原因に応じた早期の治療が必要とされており，効果的な認知症スクリーニング手法が求められている [2]．現在は認知症スクリーニング検査として，専門医による問診や身体検査に加え，改訂長谷川式簡易知能評価スケール（HDS-R）や Mini-Mental State Examination（MMSE），Clinical Dementia Rating などの神経心理学的検査が広く用いられており [3][4]，MMSE の感度と特異度はそれぞれ 0.81 と 0.89 と報告されている [5]．しかし，現在用いられている検査は専門医による診断が不可欠であり，さらに，診断にかかる時間が長いため患者の身体的・精神的負担に繋がる．

認知症には様々な基礎疾患が存在する．本研究では，認知症の原因疾患の中から前頭側頭葉変性症（Frontotemporal Lobar Degeneration：FTLD），アルツハイマー病（Alzheimer's Disease：AD）の二つに着目する．前頭側頭葉変性症は日本の指定難病に認定されており，症状が進行するにつれ言語障害が見られやすくなる [6]．同疾患は他症例と比較して発症数が少なく，専門医以外による診断は困難である．一方，認知症の原因となる基礎疾患の中でアルツハイマー病が占める割合は 50 %以上であり，最も高くなっている．さらに，アルツハイマー病の患者の言語的変化は病気の経過の早い段階で起こることが知られている [7]．そのため，本研究では二疾患の引き起こす言語障害に着目し，音声課題を自然言語モデルの入力に用いたスクリーニング手法を提案する．この手法により，スクリーニングが簡便化され，患者と専門医の負担を減らすことが期待される．

1） 名古屋工業大学 大学院工学研究科 工学専攻 創造工学プログラム
Creative Engineering Program, Dept. of Engineering, Graduate School of Engineering, Nagoya Institute of Technology
2） 名古屋工業大学 大学院工学研究科 工学専攻 情報工学系プログラム
Computer Science Program, Dept. of Engineering, Graduate School of Engineering, Nagoya Institute of Technology
3） 藤田医科大学 医学部 脳神経内科学
Department of Neurology, Fujita Medical University School of Medicine
4） 名古屋大学 大学院医学系研究科 神経内科
Department of Neurology, Nagoya University Graduate School of Medicine

2 対象疾患

本研究は非専門医による FTLD+AD と HC の診断を支援するスクリーニングツールの開発を目的とする．以下に各疾患について詳述する．

2.1 前頭側頭葉変性症（FTLD）

FTLD は大脳の前頭葉や側頭葉を中心に神経細胞の変性・脱落によって発生する認知症である．言葉の意味の理解や物の名前などの知識が失われる「語義失語」や発語量の減少などの症状が見られ，行動障害，認知機能障害などが徐々に進行する神経変性疾患であり，難病指定を受けている [8]．FTLD は脳の病変部位により分けられており，意味性認知症（Semantic Dementia: SD），進行性非流暢性失語症（Progressive Non-Fluent Aphasia: PNFA），行動障害型前頭側頭型認知症（behavioral variant Frontotemporal Dementia: bvFTD）の 3 つの臨床疾患が存在する．主として初老期に発症し，人格の変化や社会行動の乱れなど，他の認知症には見られにくい症状が現れるため，診断が遅れる症例や社会的に問題になる症例が報告されている．礼節や社会常識が欠けることで，他の人からどう思われるかを気にしなくなり，自己本位的な行動を繰り返す，暴力的な言動を起こす，信号無視や万引きをするといった犯罪行為に及ぶこともある．

2.2 アルツハイマー病（AD）

AD は溜まったタンパク質が神経細胞を破壊し，脳が萎縮することによって発症する．認知症の中で最も多く見られる疾患であり，発症者の大半が 65 歳以上であるという特徴がある．経過の早い段階で言語的変化が起こり，進行すると徘徊や性格の変化等が現れ，日常生活を送るにもサポートが必要となる．早期段階で投薬治療をすることで，進行を遅延させることも可能である．

3 関連研究

Toshiro ら [9] は，構造化されていない自由会話データから特徴を抽出し，認知症の有無を識別する機械学習モデルを提案した．認知症罹患者 127 名，健常者 197 名のデータセットを使用して 2 群分類を実行した結果，精度 0.90，感度 0.88，特異度 0.92 を得た．Toshiro らは，録音

表 1: 実験協力者（N=156）

	FTLD+AD		HC	p 値
	FTLD	AD		
Male/Female	12/18	11/20	36/59	p=1.00
Age	68.5±8.4	70.8±8.5	67.9±8.7	p=0.19
WAB（/100）	69.3±15.8	90.0±10.1	97.1±2.3	p<0.01
MMSE（/30）	19.3±6.6	19.3±4.2	28.8±1.1	p<0.01

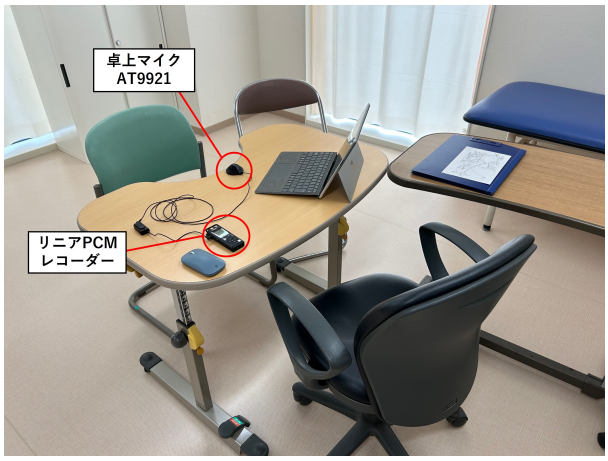

図 1: 録音環境 (共同研究者撮影)

データから被験者の発話のみをフィラーを含めて書き起こし，BERT と TF-IDF を用いて 200 次元の発話特徴量を抽出した．Toshiro らの手法は MMSE と比較して高い精度を算出しているが，入力データの前処理の負担が大きい．また，実際に認知症スクリーニング手法として使用する際には，録音データから発話を書き起こす作業は必要な労力が大きいことに加え，同じ録音データを使用しても作業者によって書き起こした内容に差が生じる可能性がある．そのため，本研究では音声認識ソフトを用いてテキストを書き起こし、そのテキストのみ入力した．

花井ら [10] は，FTLD，AD の患者と健常者（Healthy Control：HC）を判別した．自発話課題の音声を形態素解析することで，音響言語と時間特徴量からなる計 404 種の発話特徴量を抽出し，判別に用いた．FTLD11 名，AD15 名，HC63 名に対して FTLD または AD，HC を判別する 2 群判別を実行した結果，感度 0.62，特異度 0.78 を得た．この結果により，音声課題による FTLD と AD 検出の有効性が示されているが，花井らの手法では録音した課題に対して形態素解析や時間特徴量を抽出する必要があり，モデルの入力データ作成にかかるコストが大きい．そのため，本研究ではテキストのみを入力に使用し学習の利便性の向上を図る．本稿では，音声課題の類似性から花井らの研究をもとに比較検討した．

4 方法

4.1 実験協力者

表 1 に実験協力者の内訳を示す．実験には FTLD の患者が 30 名（68.5±8.4 歳），AD の患者が 31 名（70.8±8.5 歳）と HC が 95 名（67.9±8.7 歳），すなわち計 156 名が参加した．HC の対象者は認知機能検査によって定め，MMSE が 26 点以上かつ修正版 Addenbrooke の認知検査（ACE-R）が 89 点以上である者とした．ここで，FTLD の被験者には行動異常型前頭側頭型認知症（bvFTD）が

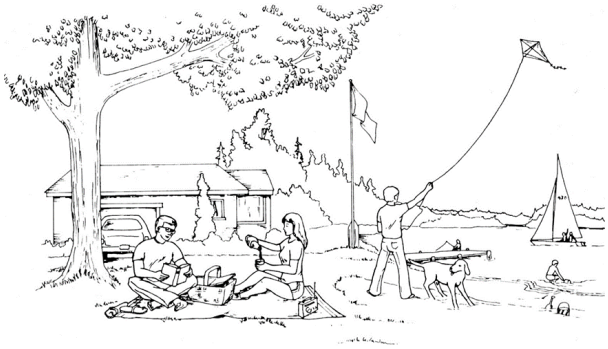

図 2: 自発話課題で示した絵 [11]

8 名（62.5±9.0 歳），進行性非流暢性失語（PNFA）が 7 名（73.1±6.7 歳），意味性認知症（SD）が 15 名（69.6±7.3 歳）含まれている．年齢，WAB 失語症検査の点数，MMSE の点数の三項目について，FTLD+AD 群と HC 間の平均値に差がないことを帰無仮説とする t 検定を実行した結果，WAB 失語症検査と MMSE の点数において有意差が認められた．また，FTLD+AD 群と HC の性別に関してフィッシャーの正確確率検定を実行したが，有意差は認められなかった．

4.2 音声データ収集及びデータセット構築

音声のデータは藤田医科大学，名古屋大学医学部附属病院，名古屋大学医学部保健学科大幸キャンパス，大阪大学医学部附属病院にて専門の言語聴覚士により収集された．録音機器はオーディオテクニカ製の単一指向性の卓上マイク AT9921 を用いた．図 1 に音声収録時の実験環境を示す．

実施手順を以下に示す．まず，WAB 失語症検査と ACE-R を実施し，発話音声を録音する．本研究では自発話部分の録音音声のみを使用した．自発話課題とは，絵を提示してその様子を自由に述べてもらう課題である．図 2 に提示した絵を示す．次に，実験参加者 156 名の発話音声を Microsoft 社製の音声認識システムである Speech to Text によりテキスト化し，判別モデルの入力に用いて fine-tuning を実行する．このとき，音響特徴および言語特徴，時間特徴等は抽出しないため，先行研究と比べてデータセットの前処理にかかる負担を軽減することができる．

4.3 判別モデルの構築

判別には大規模自然言語モデルの一種である Bidirectional Encoder Representations from Transformers（BERT）[12] と Distilled version of the RoBERTa-base model（DistilRoBERTa）[13]，Decoding-enhanced BERT with disentangled tention（DeBERTa）[14] を使用し，モデルごとの判別性能の差や事前学習タスクの影響について考察した．

DistilRoBERTa と DeBERTa は BERT の提案後に発表された後継モデルである．Victor らは，BERT アーキテクチャをベースにして，パラメータが 40 %少なく，60 %高速に動作できるモデルとして Distilled version of BERT（DistilBERT）を提案した [13]．DistilRoBERTa は DistilBERT における改良方法を RoBERTa に使用したモデルである．Liu ら [15] は，BERT のハイパーパラメータ，事前学習手法，埋め込み手法を変えることで，どの要素がモデルの性能を上げるのかについて分析し，RoBERTa を提案した．その結果，RoBERTa は英語用の

表 2: FTLD+AD/HC（2 群分類）

事前学習タスク	判別モデル	正診率	感度	特異度	F 値
chABSA	BERT	0.63	0.63	0.63	0.62
日本語 Wikipedia	BERT	0.74	**0.74**	**0.86**	0.73
日本語 Wikipedia	DistilRoBERTa	0.73	0.54	0.84	0.71
日本語 Wikipedia	DeBERTa	**0.75**	0.61	0.84	**0.75**
なし（花井ら [10]）	SVM	0.72	0.62	0.78	0.71

Text	Target	Label
商品Aの売り上げが上がった。	商品A#売り上げ	Positive
商品Bについては、コストが上がった。	商品B#コスト	Negative
⋮	⋮	⋮

図 3: chABSA-dataset

一般言語理解評価ベンチマークである GLUE タスク [16] において複数の State-of-the-Art を達成した．

DeBERTa は RoBERTa の提案後に発表されたモデルである．Pengcheng らは，単語ペアの Attention の重みが相対的な位置に依存するという仮定の下，BERT や RoBERT のモデルの構造を変更したモデルとして DeBERTa を提案した．トークン i から見たときのトークン j に対する Attention の重み $A_{i,j}$ を以下の式で表す．

$$\begin{aligned} A_{i,j} &= \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^T \\ &= H_i H_j^T + H_i P_{j|i}^T + P_{i|j} H_j^T + P_{i|j} P_{j|i}^T \end{aligned}$$

ここで，$\{H_i\}$ は位置 i の単語を表すベクトルであり，$\{P_{i|j}\}$ は位置 i と j の相対位置を表すベクトルである．BERT や RoBERTa は $A_{i,j} = H_i H_j^T$ で Attention の重みを計算していたため，DeBERTa の方がより多くの情報を表現することができる．

その後，モデルにラベル付きテキストを入力して fine-tuning を実行する．ただし，これらのモデルは英語の Wikipedia や BooksCorpus を用いて学習されているため，日本語のデータセットで事前学習を実行する．事前学習に用いたデータセットについては後の章で詳述する．

本研究では，録音された被験者の発話音声をテキスト化してラベルを付与し，モデルの入力とする．このとき，形態素解析による言語特徴量の抽出や時間特徴量の抽出はせず，得られた自然言語文をそのままモデルに入力して分類する．また，HC の被験者数に対して AD や FTLD の被験者数が少なかったため，HC のデータ数を AD と FTLD の数に合わせて削除するアンダーサンプリングを実行してデータセットを作成した．データセット作成後は 5 分割交差検証によりデータを分割し，学習時のハイパーパラメータの値を調整して検証する．

4.4 事前学習のデータセット

英語テキストで学習済みのモデルを使用するとき，日本語データセットの判別性能を向上させるためには日本語での事前学習が必要である．そのため，異なるデータセットを使用して，事前学習データセットがモデルの判別性能に与える影響について考察した．

第一に使用したデータセットは chABSA-dataset である．図 3 に chABSA-dataset 内で用いられているデータのイメージ図を示す．chABSA-datset は上場企業の有価証券報告書をベースに作成されたデータセットで，各文に対して positive あるいは negative の感情分類ラベルだけでなく，「何が」positive，negative なのかという観点を表す情報が含まれている．事前学習のタスクはこのデータセットを positive あるいは negative に分類する 2 値分類である．

第二に使用したのは，東北大学が用意した日本語 Wikipedia からなるデータセットである．2019 年 9 月 1 日時点の日本語版 Wikipedia のデータを約 1700 万文使用しており，語彙数は 32000 である．Hugging Face に学習済みのモデルが公開されていたため，本研究では同データを使用した．

5 実験結果と考察

本研究では，提案手法を用いて FTLD+AD/HC の 2 群分類について実験した．

5.1 実験結果

表 2 に結果を示す．chABSA-dataset を使用して事前学習したとき，正診率が 0.63，感度は 0.63，特異度は 0.63 だった．日本語 Wikipedia を使用して事前学習したとき，正診率が 0.74，感度は 0.74，特異度は 0.86 だった．また，DistilRoBERTa を用いて判別したとき，正診率が 0.73，感度は 0.54，特異度は 0.84 だった．DeBERTa を使用したときは，正診率が 0.75，感度は 0.61，特異度は 0.84 だった．

BERT について，森廣ら [17] の論文を参考にハイパーパラメータを調整した結果，学習率が $5e^{-6}$，batch size が 16 のとき最も性能が高かった．また，DistilRoBERTa，DeBERTa についてハイパーパラメータを調整した結果，学習率が $5e^{-6}$，batch size が 16 のとき最も性能が高かった．

5.2 考察

同じ判別モデルを使用したとき，事前学習において，有価証券報告書を基にした chABSA-dataset より，日本語 Wikipedia を基にしたデータセットを使用した方が精度が向上した．有価証券報告書は日本語 Wikipedia より偏った表現が多く見られたため，判別精度が低下したと考えられる．また，花井らが用いた特徴量が計 404 種であったのに対し，提案手法ではテキストのみで正診率，感度，特異度，F1 スコア共に先行研究より優れており，特に感度が高かった．Toshiro らの実行した自由会話データを用いた分類結果には及ばないものの，音声認識システムと自発話のテキストのみを用いた 2 群分類についても有用性が示されたと考えられる．

また，DistilRoBERTa を用いた実験では，BERT を用いたときより精度が低下する結果となった．モデル自体の性能は向上しているため，日本語で十分に事前学習ができていなかった可能性がある．今後は事前学習の方法について見直し，さらなる精度向上を達成したい．DeBERTa を用いた実験では，BERT を用いたときより正診率と F1 スコアが向上し，感度と特異度が悪化した．

FTLD+AD 群と比べ，HC をより正確に判別するモデルになっていると考えられる．

5.3 今後の展望

今後は判別性能を向上させるため，以下の三つの課題に取り組む．第一にデータセットおよび特徴量の変更である．自発話ではなく導入発話を用いた判別を検討する．導入発話は「最近何か困ったことはありますか」等と質問し，自由に答えてもらう課題である．自然な発話を促すことができるため，自発話を用いた判別より性能が向上する可能性がある．

第二にモデルの改良である．BERT-base のモデルにおいて，学習時に最終層のベクトルのみを使用するより、最終 4 層のベクトルを結合して使用する方が精度が向上するという結果が報告されている [12][18]．さらに，今回の実験ではトレーニング時の学習率を一律に設定していたため，今後はトークンを結合したり，事前学習済みの層と全結合層で学習率を変更したりして精度の向上を図りたい．

第三にサンプリングの改良である．本研究では HC の被験者数に対して FTLD や AD の被験者数が少なかったため，HC のデータ数を FTLD+AD の数に合わせて削除するアンダーサンプリングを実行した．しかし，元々のサンプル数が多くないため，さらにデータを収集してサンプル数を増やすか，オーバーサンプリングすることにより判別性能が向上する可能性がある．入力が自然言語文のため単純なオーバーサンプリング手法は使用できなかったが，生成 AI を使用してデータを生成したり，同ラベル群の文章を継ぎ接ぎして新しいデータを生成する手法について調査，実験し，有用性を検討したい．

6 おわりに

本研究では，BERT，DistilRoBERTa，DeBERTa を用いて FTLD，AD と HC を分類する手法を提案した．事前学習におけるデータセットを変更した結果，有価証券報告書を基にした chABSA-dataset より，日本語 Wikipedia を基にしたデータセットを使用した方が精度が向上した．また，日本語 Wikipedia を使用して事前学習したとき，正診率が 0.74，感度は 0.74，特異度は 0.86 だった．この結果から，400 種以上の特徴を抽出しなくても，自発話課題から FTLD あるいは AD を判別する手法の有効性が示されたと考えられる．

また，Local Interpretable Model-agnostic Explanations（LIME）等によって判断根拠を可視化することで，Attention の重さを視覚的に把握することが容易になると考えられる．分類結果だけでなく判断根拠も併せて出力することで，利用者が妥当性を感じる可能性が高くなる．そのため，より説明性の高い出力方法を検討する必要があると思われる．以上の方法により，より説明性の高い検出方法を実現し，利用者に認知症の兆候を知らせるようなシステムを目指す．

謝辞

本研究は，一部，文科省科研費（JP24H00741），ならびに，NICT 委託研究の助成により行われた．

参考文献

[1] 統計からみた我が国の高齢者, 統計局, 統計トピックス 138 (2023).
[2] 下濱俊：認知症の早期発見と予防, 学術の動向, Vol. 20, No. 6, pp. 76–80 (2015).
[3] 三木隆己：認知症診療における最近の検査, 老年期認知症研究会誌誌, Vol. 18, pp. 88–92 (2011).
[4] Valcour, V. G., Masaki, K. H., Curb, J. D. and Blanchette, P. L.: The detection of dementia in the primary care setting, *Archives of internal medicine*, Vol. 160, No. 19, pp. 2964–2968 (2000).
[5] Tsoi, K. K., Chan, J. Y., Hirai, H. W., Wong, S. Y. and Kwok, T. C.: Cognitive tests to detect dementia: a systematic review and meta-analysis, *JAMA internal medicine*, Vol. 175, No. 9, pp. 1450–1458 (2015).
[6] 大槻美佳：FTLD: 言語および関連症候の特徴とその診方, 臨床神経学, Vol. 52, No. 11, pp. 1224–1227 (2012).
[7] M., V. and J., H. R.: Semantic memory and language dysfunction in early Alzheimer's disease: a review, *International journal of geriatric psychiatry*, Vol. 27, No. 12, pp. 1209–1217 (2012).
[8] 池田学：難病指定からみた FTLD, 高次脳機能研究（旧 失語症研究）, Vol. 36, No. 3, pp. 376–381 (2016).
[9] Horigome, T., Hino, K., Toyoshiba, H., Shindo, N., Funaki, K., Eguchi, Y., Kitazawa, M., Fujita, T., Mimura, M. and Kishimoto, T.: Identifying neurocognitive disorder using vector representation of free conversation, *Scientific reports*, Vol. 12, No. Article number. 12461 (2022).
[10] 花井俊哉, 加藤昇平, 坂口巧一, 佐久間拓人, 大嶽れい子, 桝田道人, 渡辺宏久：認知課題遂行時の発話特徴を用いた認知症希少疾患の簡易検出, 電子情報通信学会論文誌, Vol. J104-D, No. 4, pp. 198–206 (2021).
[11] 杉下守弘訳: WAB 失語症検査, 医学書院 (1986).
[12] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *Proc. of NAACL*, Vol. 1, pp. 4171–4186 (2019).
[13] SANH, V., DEBUT, L., CHAUMOND, J. and WOLF, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910. 01108* (2019).
[14] He, P., Liu, X., Gao, J. and Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention, *arXiv preprint arXiv: 2006. 03654* (2020).
[15] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint arXiv: 1907. 11692* (2019).
[16] Wang, A. and al., et : GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING, *arXiv preprint arXiv: 1804. 07461* (2018).
[17] 森廣勇樹, 南條浩輝, 馬青：日本語レビューに対するレーティング予測の精度比較, 言語処理学会第 29 回年次大会 (2023).
[18] Let's Complicate Things [Bert]https://www.kaggle.com/c/google-quest-challenge/discussion/123770, 最終アクセス 2024-6-9.