J-029

Doc2Vec による発話文の分散表現を用いた認知症の病型判別

Discrimination of Dementia Types Using Distributed Representation of Speech Sentences by Doc2Vec

須嵜　愛梨[1)]　伊藤　有生[1)]　加藤　昇平[1)]　佐久間　拓人[1)]
Airi Susaki　Yuki Ito　Shohei Kato　Takuto Sakuma

大嶽　れい子[2)]　桝田　道人[3)]　伊藤　信二[2)]　渡辺　宏久[2)]
Reiko Ohdake　Michihito Masuda　Shinji Ito　Hirohisa Watanabe

1　はじめに

日本における超高齢社会の進展は著しく，認知症患者の急速な増加が問題視されている．認知症は脳の神経細胞が壊れるために起こる症状で，認識力，記憶力，判断力などに支障をきたす．認知症の根本的な治療方法はいまだ見つかっておらず，症状の進行抑制のため早期発見が非常に重要である．また，認知症は症状の異なる様々な病型が存在し，適切な治療を施すために病型を正しく識別する必要がある．アルツハイマー型認知症 (AD)，血管性認知症 (VaD)，レビー小体型認知症 (DLB)，前頭側頭葉変性症 (FTLD) は 4 大認知症と呼ばれ、認知症患者全体の約 80 %を占めている [1]. 中でも FTLD は日本の指定難病とされており，特有の言語障害や行動障害が見られ，認知症の中でも特に識別が困難であるとされている [2].

認知症の臨床診断には MRI による画像診断や血液・髄液検査が用いられることがあるが，これらの検査は費用が高額であり侵襲性が高いという問題点が指摘されている．現在，それらに代わる認知症スクリーニング検査として改訂長谷川式簡易知能評価スケール [3] や Mini-Mental State Examination（MMSE）[4]，Clinical Dementia Rating[5] などが広く用いられている．しかし，これらの検査は経験を積んだ医師や言語聴覚士，臨床心理士などの専門家によって行われる必要があり、質問や採点などの負担がかかっている．このような現状から，専門知識がない人でも検査を可能にする認知症の簡易的なスクリーニング技術が必要である．

近年では，発話音声に着目した認知症スクリーニングシステムに関する研究が提案されている．Horigome ら [6] は，自由会話の発話文から分散表現を抽出し，認知症と非認知症の判別を試みた．その結果，0.90 という高い正診率を得ているが，詳細な病型判別はしていない．花井ら [7] は，認知症の中でも判別が難しいとされる FTLD，AD および健常者 (HC) を分類した．計 385 種の発話特徴量を用いた 3 群分類の分類性能は 0.73 であった．言語特徴量として品詞の種類や総語数など計 17 種を抽出しているが，文の意味に関する特徴は考慮していない．そこで，本研究では発話文の意味に注目し，FTLD，AD および HC の分類を試みる．発話文から得られる分散表現を特徴量として判別器を構築し，花井らのモデルとの判別性能を比較する．

1）　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
2）　藤田医科大学　医学部　脳神経内科学
Department of Neurology, Fujita Medical University School of Medicine
3）　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

表 1: 実験協力者

	FTLD	AD	HC	p 値
Male/Female	12/18	11/20	36/59	p=0.93
Age	68.5±8.4	70.8±8.5	67.9±8.7	p=0.21

2　音声データ収集

2.1　実験協力者

表 1 に実験協力者の内訳を示す．実験には，FTLD 30 名，AD 31 名および HC 95 名の計 156 名が参加した．HC は認知機能を評価する検査により定め，Mini-Mental State Examination（MMSE）が 26 点以上，かつ Addenbrooke's Cognitive Examination Revised（ACE-R）が 89 点以上を満たすものとする．なお発話は年齢や性別によって影響するため，HC の年齢および性別は疾患群に合うように統制されている．FTLD，AD，HC の性別に対してカイ二乗検定，年齢に対してクラスカル・ウォリス検定をしたところ，共に有意差は認められなかった．

2.2　収録機器および実験環境

音声データは，専門の言語聴覚士によって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，並びに藤田医科大学の協力により収集された．図 1 に藤田医科大学における検査実施環境を示す．オーディオテクニカ製の単一指向性の卓上マイク AT9921 を実験協力者との距離が 30~40 cm になるように設置し，騒音レベルが 60 dB を超えない環境で実験した．また，録音時には実験協力者の発話と院内放送や検査者の声が重ならないように留意した．

2.3　実施手順

実験参加者に，認知課題として WAB 失語症検査と ACE-R を実施した．WAB 失語症検査については「自発話」「復唱」「物品呼称」の項目に加え，復唱課題において使用された文章の「音読」，ACE-R については「見当識」「3 単語記銘」「計算」「語の逆唱」「3 単語再生」を実施し，その回答音声を録音した．

FTLD 患者は，物の名前や使い方がわからなくなる「意味記憶障害」や，文法に則った正しい文を話せなくなる「失文法」といった言語障害をきたす．そこで FTLD の特有症状を考慮し，WAB 失語症検査で実施される「自発話課題」への回答音声を解析対象とした．自発話課題とは，絵を提示しその様子を自由に述べてもら

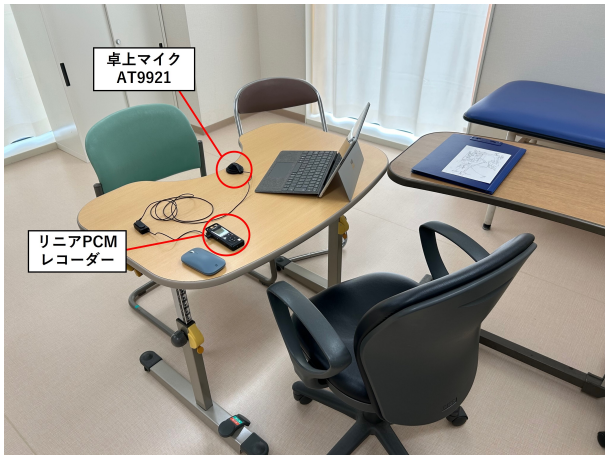

図 1: 録音環境 (共同研究者撮影)

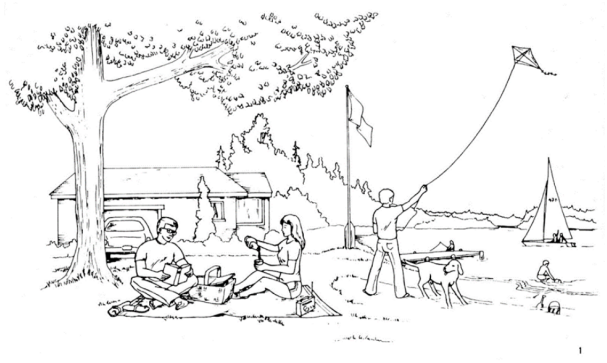

図 2: 自発話課題で示した絵 [8]

う検査である．図 2 に自発話課題に使用した絵を示す．実験協力者が単語の羅列のみで回答した場合，可能な限り文章のかたちで話すよう指示し，絵の一部のみについて回答を行った場合は「他にはありますか？」と回答を促した．

3 手法

3.1 特徴量抽出

収集した計 156 名の音声データから特徴量を抽出する．本研究では，文章を分散表現に変換する Doc2Vec[9] の手法である Distributed Bag of Words version of Paragraph Vector (PV-DBOW) と Distributed Memory Model of Paragraph Vectors(PV-DM) による特徴量，先行研究で花井ら [7] が用いた 17 種の言語特徴量の 3 種類を抽出し，それぞれの特徴量を用いて判別した結果を比較する．データの前処理として，収集した音声データを Microsoft 社製の音声認識システム Speech to Text を用いてテキスト化し，MeCab[10] により形態素解析した．

3.1.1 PV-DBOW

図 3 に PV-DBOW のモデル図を示す．PV-DBOW は，語順を無視した上で，文書に含まれる単語を当てるように文書ベクトルを学習する．段落 ID を入力とし，文書に含まれる n 個の単語 ID を出力としている．同図において X は，単語 ID を表す one-hot ベクトルである．文書中からサンプリングした n 個の単語を予測するように学習を進め，中間層から出力層の重み D を最適化する．学習後の重み D が特徴量として用いる文書ベクトルである．本研究では，全被験者をまとめたテキストデータの

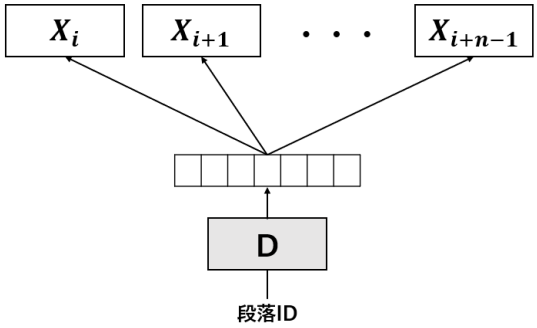

図 3: PV-DBOW

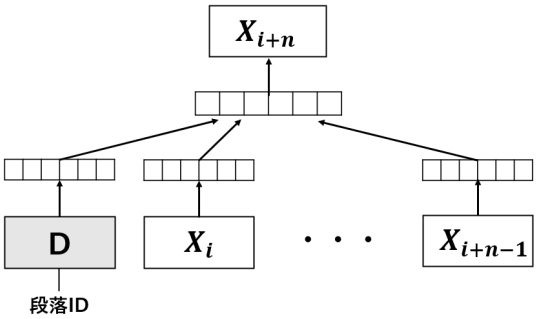

図 4: PV-DM

中で各被験者の発話を 1 つの段落とし，150 次元の文書ベクトルを抽出した．PV-DBOW の実装には python の自然言語用ライブラリ gensim[11] の doc2vec を使用した．

3.1.2 PV-DM

図 4 に PV-DM のモデル図を示す．PV-DM は，段落 ID と連続する単語 ID から，直後の単語を予測するように文書ベクトルを学習する．段落 ID と連続する n 個の単語 ID を入力とし，その直後の単語 ID を出力としている．同図において X は，単語 ID を表す one-hot ベクトルである．文書中からサンプリングした連続する n 個の単語を結合し，サンプリングした単語に続く次の単語を予測するように学習を進め，文書ベクトル D および中間層から出力層の重みを更新する．学習後の文書ベクトル D を特徴量として用いる．本研究では，全被験者をまとめたテキストデータの中で各被験者の発話を 1 つの段落とし，150 次元の文書ベクトルを抽出した．PV-DM の実装にも python の自然言語用ライブラリ gensim[11] の doc2vec を使用した．

3.1.3 先行研究 [7] で用いた言語特徴量

全体に占める各品詞 11 種の割合（名詞，動詞，形容詞，副詞，助詞，接続詞，助動詞，連体詞，感動詞，接頭詞，フィラー），総語数 N，異なり語数 $V(N)$，TTR，Simpson の D 値 [12] を算出した．TTR，D はそれぞれ式（1）と（2）で表される．

$$TTR = \frac{V(N)}{N} \tag{1}$$

$$D = \Sigma_{m=1}^{V(N)} V(m,N) \frac{m(m-1)}{N(N-1)} \tag{2}$$

$V(N)$ は異なり語数，すなわち，発話に含まれている語彙数を示し，Vm,N は，m 回出現する異なり語数の数を示す．TTR，D は共に標本に含まれる語彙の豊富さの指

標として用いられている．これらの特徴量に，音声認識システムの出力から得られる，フィラーの数，音声認識の確信度の平均を加えた，計 17 種類の言語特徴量を抽出した．

3.2 分類モデル

自発話課題に対する回答音声から得られた特徴量を用いて，Support Vector Machine (SVM) により FTLD+AD および HC の 2 群，FTLD，AD および HC の 3 群を分類する．SVM のパラメータ C は $[10^{-5}, 10^{-4}, \cdots, 10^{4}]$ の範囲で grid search により決定した．今回使用するデータは，AD，FTLD に比べて HC の人数が多く不均衡である．不均衡データをそのまま扱うと，多数群を単に予想することで見かけの正答率が向上する．しかし，認知症が見逃される可能性が高まるため，スクリーニングシステムにおいてはデータの偏りを緩和する必要がある．不均衡データへの対処として，多数派データのデータ数を減少させるアンダーサンプリングと，多数決で出力を決定するアンサンブル学習を組み合わせたモデルを構築した．弱学習器の個数は 1000 個に設定した．

4 結果と考察

4.1 性能評価実験

自発話課題の回答音声から 3 種類の特徴量を抽出し，以下の 5 つのモデルで FTLD+AD/HC の 2 群，FTLD/AD/HC の 3 群を分類する．

1. PV-DBOW により抽出した文書ベクトルを特徴量とする
2. PV-DM により抽出した文書ベクトルを特徴量とする
3. 先行研究で花井ら [7] が用いた 17 種の言語特徴量を用いる
4. 花井ら [7] が用いた 17 種の言語特徴と，PV-DBOW による文書ベクトルを合わせたものを特徴量とする
5. 花井ら [7] が用いた 17 種の言語特徴と，PV-DM による文書ベクトルを合わせたものを特徴量とする

各モデルにおいて 5 分割交差検証を行い，全ての評価指標の平均を算出する．評価指標として，各クラスごとの適合率，再現率，全体の指標として正診率および F 値を算出した．

4.2 FTLD+AD/HC の 2 群判別

表 2 に，2 群判別の判別性能を示す．表中の疾患 F+A は FTLD と AD を合わせた疾患患者，H は HC を指す．PV-DBOW による 150 次元の文書ベクトルを特徴量としたモデル 1 と，PV-DM による 150 次元の文書ベクトルを特徴量としたモデル 2 では，どちらも先行研究で用いられた 17 種の言語特徴量を用いたモデル 3 と比べ正診率が低い結果となった．しかし，モデル 1 ではモデル 3 と比べ FTLD+AD の再現率が高いことから，PV-DBOW による文書ベクトルを特徴量とした場合，疾患患者と判別しやすいことが分かる．疾患患者を健常と判断してしまうと疾患患者を見逃してしまうため，健常者を疾患の疑いがあると判断するより疾患患者を健常と判断するほうがリスクが大きい．つまり，偽陰性より偽陽性が高いほうがスクリーニングシステムとしては望ましい．その点においては，モデル 3 よりもモデル 1 が優れていると言える．先行研究で用いられた 17 種の言語特徴量と PV-DBOW による文書ベクトルを合わせたものを特徴量としたモデル 4 が，5 つのモデルの中で最も高い正診率を示した．しかし，モデル 4 ではモデル 1 と比べ FTLD+AD の再現率が低下していることから，PV-DBOW による文書ベクトルに先行研究で用いた 17 種の言語特徴量を加えることで，偽陰性の割合が増加したことが分かる．以上から，PV-DBOW による文書ベクトルを用いることで疾患患者の見逃しを減少させることができるという点において，PV-DBOW による文書ベクトルを特徴量として用いることの有用性が示唆された．

表 2: 2 群判別における性能

モデル	疾患	正診率	適合率	再現率	F 値
1	F+A	0.717	0.601	**0.763**	0.712
	H		**0.821**	0.688	
2	F+A	0.579	0.460	0.492	0.558
	H		0.670	0.634	
3	F+A	0.724	0.649	0.627	0.707
	H		0.768	**0.785**	
4	F+A	**0.763**	**0.683**	0.729	**0.754**
	H		0.820	**0.785**	
5	F+A	0.689	0.596	0.586	0.661
	H		0.745	0.753	

表 3: 3 群分類における性能

モデル	疾患	正診率	適合率	再現率	F 値
1	F	0.500	0.292	0.452	0.461
	A		0.356	**0.571**	
	H		0.780	0.495	
2	F	0.447	0.292	0.452	0.414
	A		0.293	0.429	
	H		0.677	0.452	
3	F	0.605	0.477	**0.677**	0.560
	A		**0.375**	0.429	
	H		0.776	0.634	
4	F	**0.632**	**0.559**	0.613	**0.581**
	A		0.333	0.536	
	H		**0.849**	**0.667**	
5	F	0.583	0.548	0.567	0.523
	A		0.318	0.500	
	H		0.750	0.613	

4.3 FTLD/AD/HC の 3 群分類

表 3 に，3 群分類の分類性能を示す．表中の疾患 F は FTLD，A は AD，H は HC を指す．PV-DBOW による 150 次元の文書ベクトルを特徴量としたモデル 1 と，PV-DM による 150 次元の文書ベクトルを特徴量としたモデル 2 では，どちらも先行研究で用いられた 17 種の言語特徴量を用いたモデル 3 と比べ正診率および FTLD の再現率が低い結果となった．先行研究で用いられた 17 種の言語特徴量と PV-DBOW による文書ベクトルを合わせたものを特徴量としたモデル 4 が最も高い正診率を示したが，モデル 3 と比べ FTLD の再現率が低下している．以上より，PV-DBOW による文書ベクトルを特徴量として用いることで FTLD 患者の見逃しが増加することが分かった．2 群判別，3 群分類のどちらも，先行研究で用いられた 17 種の言語特徴量と PV-DBOW による文書ベ

クトルを合わせたものを特徴量としたモデル 4 が最も高い正診率を示したことから，PV-DBOW による文書ベクトルを特徴量として用いることの有用性が示唆された．しかし，判別が困難とされる FTLD 患者の見逃しが増加するため，3 群分類において PV-DBOW による文書ベクトルを特徴量として用いるには検討が必要である．また，2 群判別，3 群分類ともに，PV-DM により抽出した文書ベクトルを用いた場合と比べ，PV-DBOW により抽出した文書ベクトルを用いた方が高い正診率を示した．PV-DBOW は文書に含まれる単語を当てるように文書ベクトルを学習するため，多くの被験者の発話に共通の単語を含む自発話課題において有用であると考える．

5 おわりに

本研究では，専門知識がない人でも検査を可能にする認知症スクリーニングシステムの開発を目指し，発話音声による認知症の病型判別モデルを提案した．2 群判別，3 群分類ともに先行研究で用いられた 17 種の言語特徴量に PV-DBOW による文書ベクトルを加えることで正診率が向上した．このことから，PV-DBOW による文書ベクトルを特徴量として用いることの有用性が示唆された．しかし，3 群分類では PV-DBOW による文書ベクトルを用いることで FTLD 患者の見逃しが増加してしまうため，検討が必要である．そして，検討するにあたり判別に寄与する特徴量を求める必要があると考える．

今後は，更なる判別性能の向上を目指す．先行研究では花井ら [7] が，言語特徴量に加え音響特徴量，時間特徴量の計 406 種の発話特徴量を用いて判別している．そこで，本研究で用いた言語特徴量に加え，発話音声から得られる音響特徴量を用いて実験をする予定である．また，新たな判別モデルを検討する．本研究では判別モデルとして SVM を用いたが，今後は先行研究で Horigome ら [6] が使用した Deep Neural Network (DNN) などを実装し，認知症スクリーニングシステムにおける最適な判別モデルを検討する．

謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 JP24H00741），ならびに，国立研究開発法人情報通信研究機構委託研究の助成により行われた．

参考文献

[1] 公益社団法人 日本看護協会編：認知症ケアガイドブック, 照林社 (2016).
[2] Mendez, M. and Cummings, J.: Dementia: A Clinical Approach., Elsevier - Health Sciences Division (2003).
[3] Kato, S., Shimogaki, H., Onodera, A., Hiroki, U., Kenzo, O. and Ikeda, K.: Development of the revised version of Hasegawa's Dementia Scale (HDS-R) (1991).
[4] MF, F., SE, F. and PR, M.: "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician., J Psychiatr Res, Vol. 12, No. 3, pp. 189–198 (1975).
[5] Morris, J. C.: The Clinical Dementia Rating (CDR) current version and scoring rules, Neurology, Vol. 43, No. 11, pp. 2412–2412 (1993).
[6] 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. 1, p. 12461 (2022).
[7] 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 D, Vol. 104, No. 4, pp. 198–206 (2021).
[8] 杉下守弘訳：WAB 失語症検査, 医学書院 (1986).
[9] Le, Q. and Mikolov, T.: Distributed representations of sentences and documents, in International conference on machine learning, pp. 1188–1196PMLR (2014).
[10] MeCabhttps://taku910.github.io/mecab/, 最終アクセス 2024-6-12.
[11] gensimhttps://radimrehurek.com/gensim/, 最終アクセス 2024-6-12.
[12] Simpson, E. H.: Measurement of diversity, nature, Vol. 163, No. 4148, pp. 688–688 (1949).