

要 約

報告番号	① 乙 第	号	氏 名	百 田 友 紀
主 論 文 題 名 Language patterns in Japanese patients with Alzheimer disease: A machine learning approach (日本におけるアルツハイマー型認知症患者の発話パターンに関する機械学習を用いた研究)				
(内容の要旨) アルツハイマー型認知症 (Alzheimer disease : AD) の正確な診断および追跡評価は日常診療や疾患修飾薬の開発において重要であるが、バイオマーカーを用いた検査はコストや侵襲性などの課題がある。このため臨床場面で取得しやすいデータを診断および評価に活用できれば有用である。ADの主徴は記憶障害であるが比較的早い病期から言語障害を呈することが指摘される。発話の特徴として、名詞などの言葉の想起が困難で、発話量に比し情報量の少ない「空疎な発話」 (“empty speech”) が報告されている。このような発話特徴を評価するため用いられてきた手法の一つが絵画説明課題であるが、評価手続きが煩雑で検査者の主観を要する場合もあり、臨床場面での実用性は十分ではなかった。一方、近年は機械学習などの手法を用いて発話内容を客観的に分析し、AD患者の発話特徴や、AD患者と健常者の識別精度を調べる研究も報告されている。しかし、日本語話者を対象にAD患者と健常者の発話を識別し、その根拠となる発話特徴を示した研究は、知る限り報告がない。本研究の目的は、自然言語処理および機械学習を用いて日本語話者のAD患者と健常者の発話を識別し、AD患者の発話特徴 (パターン) を特定することである。 対象は50歳以上のAD患者42名および対照健常者52名で、絵画説明課題における1分程度の発話データ計276例を用いた。自然言語処理には構文解析ソフトのGiNZAを用いた。GiNZAは多言語共通の尺度による構文解析を目指して開発されたUniversal Dependenciesに基づいており、広く普及したプログラミング言語のPythonに対応している点で再現性に寄与する。発話内容は、意味を持つ最小単位である形態素に分割してタグを付与し、次に形態素の係り受け関係にタグを付与し、各々を形態素および構文の特徴量とした。機械学習には勾配ブースティングツリーアルゴリズムのeXtreme Gradient Boosting (XGBoost) を用い、学習過程での各特徴量の貢献度から識別 (予測) の根拠を推定した。 各タグの出現確率と隣接するタグペアの出現確率に基づく機械学習の結果、識別精度は総合的評価指標のf1スコアが0.84、Area Under the Curveは0.90と示された。識別への貢献度が高い10個の特徴量のうち7個は形態素の特徴量で、特に名詞、動詞などの内容語に関連していた。Box plotではAD群の発話において内容語に関連する特徴量が減少し、発話の停滞に関連する特徴量が増加することが示され、「空疎な発話」の特徴に一致すると考えられた。 本研究は、比較的多数の日本語話者のデータを用いた機械学習により、高い精度でAD患者と健常者を識別し、臨床的に有用な客観的評価指標となり得る発話特徴を特定した。今後の展望として、サンプル数の増大、言語の意味論や語用論を考慮した高次の自然言語処理等により精度向上の可能性はある。また、認知症の診断は病理との関連が重視される傾向にあり、アミロイドβなどのタンパク質の蓄積との関連を考慮した分析が有用と考えられる。				