

Twitterを用いた4大認知症の自動分類に関する検討

宇都宮和希[†] 坂野 遼平^{††}

[†] 工学院大学情報学部情報通信工学科 〒192-0015 東京都八王子市中野町 2665-1

^{††} 工学院大学大学院工学研究科情報学専攻 〒192-0015 東京都八王子市中野町 2665-1

E-mail: [†]tj019047@ns.kogakuin.ac.jp, ^{††}banno@cc.kogakuin.ac.jp

あらまし 日本では少子高齢化が社会問題となっており、2040年には高齢者の数が約4000万人になると予測されている。それに伴い、認知症患者が増えると予想される。主要な認知症として4大認知症が知られており、それぞれ適切なケア方法が異なる。また、症状を抑えるためには早期発見が重要である。このため、認知症の種類を手軽に見定めることができる手段が求められる。本研究では4大認知症についての自動分類手法の提案を行う。クラウドソーシングを用いてTwitter上から4大認知症の症状を抽出し、BERTを用いて学習器を作成し、精度を測定した。

キーワード 医療・ヘルスケア, 機械学習, テキスト分類, 情報抽出

1 はじめに

日本では少子高齢化が社会問題となっており、2040年には高齢者の数が約4000万人[1]になると予測されている。それに伴い、図1に示すように今後認知症患者が増えると予想される[2]。主要な認知症として4大認知症が知られており、それぞれ適切なケア方法が異なる。また、症状を抑えるためには早期発見が重要である。このため、認知症の種類を手軽に見定めることができる手段が求められる。そこで近年、言語能力に基づいた認知症スクリーニングが注目されているが、少子化による医療従事者減少や病院の不足などの問題から人手によるスクリーニングが難しいケースがある。既存研究ではサイトから情報抽出して分類する手法[3]や説明課題を行い、機械学習のモデル構築と評価を行う手法[4]があり、人手不足の解消に資することが期待されているがデータ収集が困難であることが指摘されている。

本研究ではBERTを用いることでスクリーニングの人手不足の解消とTwitterからテキストを集めることでデータ不足を解決できる可能性に着目し、Twitterと機械学習を用いた4大認知症の自動分類手法を提案する。これにより早期発見による症状緩和につなげることを目的とする。

1.1 4大認知症について

認知症という一括りで考えられがちだが、その発症原因や症状、症状に対する対応方法は多岐にわたる。しかし近年までの研究によって、それらにはいくつかの傾向があることが分かかってきており、認知症全体の約9割を大きく分けて4種類に分類することができ、これらは4大認知症[5]と呼ばれている。ここでは4大認知症の特徴的な症状とそれぞれの主なケア方法について記載する。

- アルツハイマー型認知症

記憶を司る海馬を中心に、広範囲で脳が委縮することで引き起こる。主に記憶障害、妄想などが出現する。その後、症状が

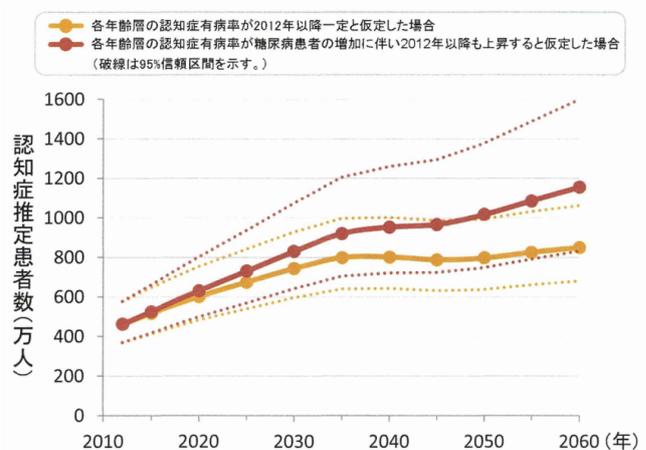


図1 認知症患者の将来推計 ([2]より引用)

進行すると幻覚や見当識障害が顕著になる。アルツハイマー型認知症になると同じ話を何度もしたり、物の名前が一致しなかったり、同居している家族等はストレスを感じる事が多くなるが、否定せずに本人が過ごしやすい環境を整えることが重要になる。

- 血管性認知症

血管性認知症には3つの特徴的な関連性がある。1つめは認知症状態が見られ、2つめに脳血管疾患があり、3つめで認知症状態と脳血管疾患発症との間に時間的関連性があるものである。特徴的な症状として実行機能障害が多い。男性患者が多く、症状は段階的に早く進行する。身体機能が低下して寝たきりになることもあるので、リハビリなどを行い、体のケアに努めることが症状の改善に繋がる。

- レビー小体型認知症

記憶を司る海馬から視覚を司る後頭葉までの広範囲で血流が悪くなり、機能が低下することによって起こる。初期症状は睡眠障害や幻視が現れることが多い。便秘や異常な発汗など自律神経症状も現れる。幻視などは本人にはそう見えているので、否定すると頭の中が混乱し、症状を悪化させる原因にもなる。

話を合わせるなどして悪化を防ぎ、症状緩和に繋げる。

- 前頭側頭型認知症

人格を司る前頭葉と、言語を司る側頭葉が萎縮することによって起こる。症状として人格変化や常同行動が生じる。進行すると自分本位な行動が多くなり、食の変化が顕著になる。人の声や動きに敏感になるので刺激の少ない静かな環境を整えることが重要になる。

2 既存研究

小杉ら [3] はちえのわ net という体験談の投稿サイトから投稿された認知症の行動・心理症状に関するテキスト情報を、機械学習を用いて 10 のカテゴリに自動分類する実験を行った。実験当時、認知症ちえのわ net には約 3300 件の認知症の行動・心理症状に関するテキスト情報が集積されており、そのうちの約 2,000 件を対象にした結果、分類精度はカテゴリごとに差はあるものの、全体として約 50 % の正解率を得ることができた。また同じデータを用いて深層学習を行ったところ、正解率を約 73 % まで向上させられることが分かったが、収集した体験談データに偏りがあったため、カテゴリごとに分類精度に差が出る結果となった。

柴田ら [4] はコーパス構築と認知症予備軍の分類研究の手法として画像、エピソード、アニメーション説明の 3 つの課題から発話データセットの構築を行い、それを用いた軽度認知症者と健常者の自動識別を行った。実験結果としてすべての課題で得られた特徴量の平均値を特徴量にした場合が識別性能が最も高く、複数課題を用いた場合は識別性能向上の可能性を示すことができた。しかし先行研究と比較するとサンプル数が 60 人と少なく、信頼性に問題がある点が挙げられている。

以上より、これらの既存研究ではデータ数が問題となっている。本研究では情報抽出先を Twitter とし、膨大な投稿からデータ数という課題を解決できる可能性に着目した。

3 提案手法

本研究では、Twitter の 4 大認知症に関するツイートから、BERT を用いた機械学習によりそれぞれの認知症を自動分類する手法を提案する。提案手法の概要を図 2 に示す。提案手法の流れとして、4 大認知症に関するツイートを収集し、形態素解析と符号化を行った後にツイートテキストから学習データとテストデータを作成する。作成した学習データを BERT に入力し、ファインチューニングを行い、ファインチューニングを行った学習モデルと事前に振り分けたテストデータを用いて精度評価を行う。

3.1 ツイートの取得とテキスト作成

4 大認知症に関するツイートの取得を行い、連結して 4 大認知症ごとの 4 つのテキストを作成する。テキストの前処理として形態素解析器を用いての単語の分割を行い、分割した単語は図 3 のように ipadic という形態素解析の辞書をもとに ID を割り当てて数値化し、BERT へ入力を行う。

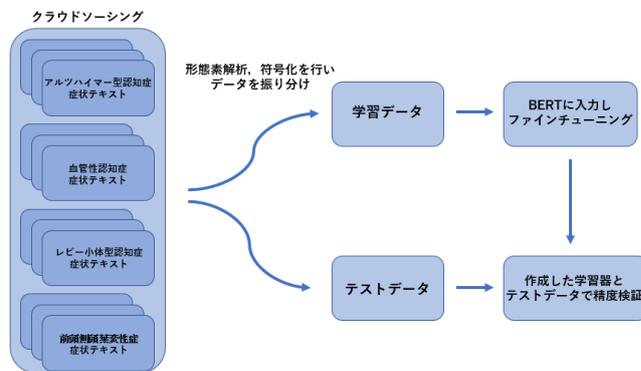


図 2 提案手法の概要

ID	トークン
5	の
6	、
7	に
8	。
9	は
10	を
...	...

図 3 形態素解析辞書例

3.2 BERT

2018 年に Google の Jacob Devlin らの論文 [6] で発表された自然言語処理モデルである。翻訳、文書分類、質問応答など多様な自然言語処理分野で最高スコア記録しており、高い文脈理解性能で知られている自然言語処理モデルである。BERT の仕組みとして、まず単語データの並びのことを文章またはシーケンスと呼び、BERT は入力されたシーケンスから別のシーケンスを予測する。BERT は事前学習モデルであり、入力されたラベルが付与されていない。つまり名前がついていない分散表現を処理することによって学習する。双方向からの学習による文脈理解精度、転移学習による汎用性、そしてラベル付けが行われていないデータにも対応できるのでデータ不足の克服などにも使うことができる。

4 評価

ツイートの取得はクラウドソーシングを用いて Twitter 上から 4 大認知症のいずれかについて具体的な症状や体験談が記載されたものを収集した。なお複数種類の認知症について言及しているツイートは除外した。取得したものはユーザ ID、ツイート本文、それに付随するツイート ID である。収集対象期間は 2010 年 7 月 1 日から 2022 年 12 月 10 日の約 12 年分とした。内訳としてはアルツハイマー型認知症を 196 件、血管性認知症を 197 件、レビー小体型認知症を 189 件、前頭側頭型認知症を

171 件で合計 753 件のツイートを収集した。

抽出したテキストを事前処理として Google Colaboratory 上で対象となる言語の文法や単語の品詞情報をもとに、文章を形態素に分解する形態素解析を行った。形態素解析のツールとして日本語の形態素解析ツールの MeCab を Python から扱えるようにした fugashi を利用した。形態素解析後は符号化し、BERT で学習を行ってから、4 大認知症の分類精度を測定し、分類結果からマクロ平均法を用いて Recall (再現率) および Precision (適合率) を算出し、評価指標 F1 の算出を行った。

4.1 Google Colaboratory

Google Colaboratory とは、Google が機械学習の教育及び研究用に提供しているインストール不要かつ、すぐに Python や機械学習・深層学習の環境を整えることが出来る無料のサービスである。機械学習で必要になる Python の実行環境がブラウザから利用可能であり、必要なライブラリが最初からインストールされているので、環境構築を手間取らず、BERT の学習などで必要となる GPU が無料で扱える [7]。

4.2 マクロ平均法

マクロ平均法は各クラス毎に評価指標 (適合率・再現率など) を計算した後に平均を取る方法である。各クラスの評価指標の平均を取ることで、各クラスのサンプル数の偏りに影響を受けることなく評価指標が算出できる。例えば、クラス 1 は 10000 サンプル、クラス 2 は 1000 サンプル、クラス 3 は 1000 サンプルであると仮定し、クラス 1 は適合率 0.5、クラス 2 は適合率 0.8、クラス 3 は適合率 0.8 であったとする。この場合、マクロ平均法を適合率の集計方法として採用すると多クラス分類問題全体における適合率は $(0.5+0.8+0.8)/3 = 0.7$ となる。このようにマクロ平均は全てのクラスに平等に重み付けするため、最も出現するクラスラベルに対して過度な影響を受けない。

図 4 を例にすると、クラス A で再現率、適合率を以下の式から求め、求めた再現率、適合率から F 値を算出する。この際、クラス A を Positive、その他のクラスを Negative とみなし、クラス B、C も同様の方法で再現率、適合率、F 値を求める。

$$\text{再現率} = \frac{TP}{TP + FN}$$

$$\text{適合率} = \frac{TP}{TP + FP}$$

$$F \text{ 値} = 2 \cdot \frac{\text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

算出したクラス A、B、C の再現率、適合率、F 値をもとに以下の式からマクロ再現率、マクロ適合率、マクロ F 値を算出できる。ここでは i はクラスを表す。

$$\text{マクロ平均再現率} = \frac{\sum_{i=1}^n (\frac{TP}{TP+FN})_i}{i} = \frac{\sum_{i=1}^n \text{再現率}_i}{i}$$

$$\text{マクロ平均適合率} = \frac{\sum_{i=1}^n (\frac{TP}{TP+FP})_i}{i} = \frac{\sum_{i=1}^n \text{適合率}_i}{i}$$

$$\text{マクロ平均 F 値} = \frac{\sum_{i=1}^n (2 \cdot \frac{\text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}})_i}{i} = \frac{\sum_{i=1}^n F \text{ 値}_i}{i}$$

	A	B	C
A(300テキスト)	240	20	40
B(200テキスト)	5	180	15
C(100テキスト)	25	5	70



	再現率	適合率	F 値
A	0.800	0.889	0.842
B	0.900	0.878	0.889
C	0.700	0.560	0.622



	平均再現率	平均適合率	平均 F 値
	0.800	0.776	0.784

図 4 マクロ平均法の例

表 1 評価結果

トークン数	精度	マクロ再現率	マクロ適合率	マクロ F1 値
512	0.1667	0.3921	0.3480	0.3295
256	0.3333	0.2773	0.2152	0.2220
128	0.6667	0.4453	0.5517	0.4398

本実験では全体のテキストに偏りが無いので、データセット全体を考慮して評価指標を算出するマイクロ平均法でも有効かと考えた。しかし、Google Colaboratory の環境上でデータセットを学習、テストデータに振り分けた際に学習データに特定の認知症が偏ってしまった場合、推論時に特定の認知症だけ予測結果が良く、ほかの認知症は結果が悪くなる可能性が考えられた。結果的に全体のテキストに偏りが無いのに、データ振り分けで偏りが出してしまうので出現するクラスに対して過度な影響を受けないマクロ平均法を採用した。

4.3 評価結果

本実験では学習データの batch size を 32、検証とテストデータの batch size を 256、epochs を 10 としてトークン数を変えながら検証を 3 回行い、モデルの精度測定と評価を行った。詳細な内訳としては学習、テストデータを 8:2 の比率に分割し、学習データでファインチューニングを行ってから、テストデータを用いて精度を測定した。トークン数ごとの結果を表 1 に示す。一定の精度が得られたが、トークン数を上げるほど精度は落ちたが、再現率が適合率を超える結果となった。

4.4 考察

実験結果の表 1 よりトークン数が 128 及び 256 ではランダムで分類した場合の期待値である精度 25 % を上回ったが、トークン数が 512 では下回った。原因として症状に類似性のある認知症がいくつかあり、似たような症状を学習したことにより、誤った分類をした可能性がある。

またトークン数を上げるほど精度が下がっている。本来トー

クン数を上げると精度も向上するが、本研究ではトークン数を上げるほど類似性のある症状が学習内容を複雑にしたため分類精度が落ちている可能性も示唆された。そしてトークン数を上げるほど再現率が適合率より良い結果を出すようになった。一般的に再現率と適合率はトレードオフの関係となっており、片方が増加すれば片方は減少する。トークン数を上げるほど再現率が上昇するという事は、分類器にとって正例と判断しやすい負例が多くなるが、負例は正しく判断できるようになっていると考えられる。

5 ま と め

本研究では、認知症における主要 4 つの認知症を Twitter 上にある関連ツイートから自動分類することを目的として BERT を用いた機械学習による手法を提案した。ツイートの収集にはクラウドソーシングを用いて、モデルの検証は学習データを用いてファインチューニングし、そのあとにテストデータを用いてモデルの精度測定と評価指標の計算を行った。ランダムで分類した場合の期待値精度を上回ったので、提案手法は一定の精度を得ることができたといえる。

評価結果としてトークン数を上げるたびに類似性のある症状テキストが学習内容を複雑にし、認知症を分類を困難にしていると考察できたので、今後の課題として類似する症状をもつ認知症にも対応できるように複数選択を許容しているマルチラベル分類法 [8] を用いた精度向上に取り組みたい。

文 献

- [1] 国立社会保障・人口問題研究所, “日本の将来推計人口（平成 29 年推計）”, 人口問題研究資料, No. 336, 2017.
- [2] 厚生労働科学研究費補助金厚生労働科学特別研究事業, “日本における認知症の高齢者の人口の将来推計に関する研究”, 総括・分割研究報告書, 2016.
- [3] 小杉尚子, 佐藤俊介, 吉山顕次, 数井裕光, “認知症ちえのわ net における認知症介護に関するテキスト情報の自動分類に関する研究”, In *DEIM*, C21-5, 2021.
- [4] 柴田大作, 伊藤薫, 若宮翔子, 荒牧英治, “対照群付き高齢者コーパスの構築とそれを用いた認知症予備軍スクリーニング技術の開発”, 人工知能学会論文誌, Vol. 34, No. 4, pp. B-J11.1-9, 2019.
- [5] 国立研究開発法人国立長寿医療研究センター, “認知症初期集中支援チームに必要な若年性認知症の知識”, 認知症初期集中支援チーム員研修会資料, 2018.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding”, *arXiv*, No. 1810.04805, 2019.
- [7] 近江崇宏, 金田健太郎, 森長誠, 江間見亜利, “BERT による自然言語処理入門”, ストックマーク株式会社, p. 44p, 2021.
- [8] 吉村卓亮, 馬場雪乃, 鹿島久嗣, “相互依存モデルによるマルチラベル分類”, 人工知能学会, No. 2K2-1in1, 2017.