情報処理学会論文誌　デジタルプラクティス　Vol.4　No.3　69–73 (July 2023)

発話者の潜在ニーズ予測とその可視化 Word2Vec モデルを用いた機械学習モデルの精度改善に関する検討

種村 菜奈枝[1,a)]　町井 湧介[2]　佐々木 剛[3]　荒木 通啓[1]　佐藤 淳子[4]　千葉 剛[1]

受付日 2022年7月21日, 採録日 2023年3月16日

概要：近年，市民参画型の必要性は健康政策のみならず，社会における場でも議論されている．しかし，日本はハイコンテクスト文化であり，一般市民がコンテクストに頼らずに意見を明確な言葉として表現するコミュニケーションには不慣れであり，一般市民の声を政策等へ反映することは容易ではない．本研究では，口語テキストから発話者の潜在的ニーズを予測するための機械学習モデル構築，およびニューラルネットワークを用いて単語をベクトル変換する手法である Word2Vec モデルを用いて機械学習モデルの精度改善を検討した．予備検討では，機械学習モデルの精度比較を行い，最適なモデルを選択した．本調査では，Word2Vec モデルを用いて同義語辞書を作成し，この辞書を使用して同一の特徴量に変換し学習を行う新手法を検討した．新手法の適応の有無で機械学習モデルの精度比較を行った．予備検討でのモデル選定実験の結果，モデル精度は xgboost で F 値 0.54 と最も高く，本調査では，モデル精度は同義語辞書ありで F 値 0.61，なしで F 値 0.54 であり，Word2Vec モデルを用いた同義語辞書の適応が機械学習モデルの精度改善に寄与した．

キーワード：潜在ニーズ，機械学習，可視化，ニューラルネットワーク，Word2Vec

Prediction and Visualization of Latent Needs: Improving the Accuracy of Machine Learning Models using the Word2Vec Model

Nanae Tanemura[1,a)]　Yusuke Machii[2]　Tsuyoshi Sasaki[3]　Michihiro Araki[1]　Junko Sato[4]　Tsuyoshi Chiba[1]

Received: July 21, 2022, Accepted: March 16, 2023

Abstract: The need for public engagement has been deliberated in recent years. However, because of Japan's high context culture, Japanese people are not accustomed to communicating their opinions in clear. Therefore, it is not easy to reflect the voices of the general public in policy making. In the process of building a machine learning model for predicting the latent needs from spoken text, this study examined how to improve the accuracy of the model using the Word2Vec model, that uses a neural network to transform words into vectors. In this preliminary study, we compared the accuracy of machine learning models and selected the best model. We examined a new method that uses the Word2Vec model to create a synonym dictionary to convert the word clusters for identical features for learning. We compared the accuracy of machine learning models with and without adaptation of the dictionary. The results of model selection showed that xgboost had the highest model accuracy with an F value of 0.54. The model accuracy was 0.61 with the dictionary and 0.54 without. It showed that the adaptation of the synonym dictionary using the Word2Vec model can improve the accuracy of the model.

Keywords: latent needs, machine learning, visualization, neural network, Word2Vec

[1] 医薬基盤・健康・栄養研究所
National Institutes of Biomedical Innovation, Health and Nutrition, Settsu, Osaka 566–0002, Japan
[2] 独立研究者
Independent Researcher
[3] 千葉大学医学部附属病院
Chiba University Hospital, Chiba 260–8677, Japan
[4] 医薬品医療機器総合機構
Pharmaceuticals and Medical Devices Agency, Chiyoda, Tokyo 100–0013, Japan
[a)] n-tanemura@nibiohn.go.jp

1. はじめに

一般市民が，社会における科学技術のあり方や政策決定する場に参画する動きは世界的にも検討されてきた．一般市民は知識がないという欠如モデルに基づく科学コミュニケーションではなく，一般市民は具体的な状況に関しては専門家よりも知識が多く，それらは「智恵（folk knowl-

edge)」と認識されている一方で，科学知識を典型とする知識と比べて正当な評価を受けることは少なかった．その後，アメリカで開発された「コンセンサス会議」で，科学技術に関わる政策的議論の場に一般市民が関わる可能性が模索され，現在では一般市民がそのような場に参画することが日常になってきた [1].

一方，これら場面において，一般市民が参画していくにあたり，諸外国との文化的背景や国民性の違いに留意することが重要である．日本はハイコンテクスト文化であり，相手に伝える努力をしなくても分かり合えるという側面が強い．一方，欧米のローコンテクスト文化の場合は言語に依存したコミュニケーションである [2]．この特徴を踏まえると，諸外国のように，一般市民の「個人」の潜在ニーズを政策等の場に迅速に反映することは難しいと思われるが，現状，一般市民の潜在的なニーズを自動抽出する方法論やそれらを政策決定の意思決定の場に声として反映させるためのシステムは存在しない [3], [4].

日本のハイコンテクスト文化を踏まえると，対象者に配慮した環境での対面でのインタビューやアンケート等の文字テキストを通じた意思伝達は容易であると考えられる．そこで本研究では，口語テキストから発話者の潜在的ニーズの自動抽出をするための機械学習モデル構築，および口語テキストに含まれる膨大な単語から潜在ニーズを精度高く予測するために，ニューラルネットワークを用いて単語をベクトル変換する手法である Word2Vec モデルを用いて同義語辞書を作成し，近い意味の単語に同一の特徴量を付与する新手法の有用性を検討した．

Word2Vec とは，単語の意味は周辺単語分布によって形成されるという分布仮説に基づき，単語の分布傾向をニューラルネットワークによって学習し，前後に出現する分布傾向が近い単語同士を近いベクトル表現へ変換するモデルを構築可能な手法である [5].

2. 方法

まず初めに予備検討では，発話者の潜在ニーズ予測のための機械学習モデル構築において最適なモデルの選択を行うための実験を行った．次に，本調査では，前述の予備検討で選択したモデルでの機械学習モデルの精度向上を目的に，新手法（Word2Vec モデルを用いた同義語辞書の作成）の適応別で精度比較を行い，新手法の有用性を評価した．

2.1 データセット

本研究では，特定非営利活動法人　健康と病いの語りディペックス・ジャパンより提供を受けた，「健康と病いの語り」データのうち，ウェブサイトで一般公開されている語りウェブページの口語テキストを使用した [6].

この選択理由は，近年，認知症患者の増加に伴いその家族を含めた潜在的ニーズが多くある一方で市民の声を政策等の場への反映の必要性が高い領域と考え選択した．

インタビュー者の問い 1 つに対して，対象者が回答した文章群を 1 レコードとし，全 1107 レコードを本研究の分析対象とした．次に，各文章群において期待や改善して欲しいといった対象者の潜在ニーズが含まれている可能性が高い関連表現として「あるといい」「いい」「やっぱり」「やっぱ」「やはり」という表現を 1 つ以上含む文章群に対して注目すべき意味を含む文章群と判定した [7]．その後，独立した 2 名が，これら文章群に対して，対象者の潜在ニーズの有無を判定した．このニーズ判定にあたっては，2 名の意見が分かれた文章群に対しては協議を行い，疑義がなくなるまで繰り返し検討を重ねた．

2.2 予備検討

分析対象の口語テキストに対して，形態素解析を行った後，入力データは，単語（名詞），立場（本人/家族等）とし，出力データは，潜在ニーズの有無とした．形態素解析エンジンは，形態素解析器 MeCab（mecab-python3 v1.0.5）を，形態素解析用辞書は，辞書の定期更新を含むメンテナンスが定期的になされており，新語や固有表現に強いことから mecab-ipadic-NEologd（2022 年 2 月時点公開版）を選択した．予備検討では，最適な機械学習モデル選択のため，解釈性の高い木構造モデルである決定木，ランダムフォレスト，xgboost の 3 つのモデルで精度比較を行った．

2.3 本調査

2.3.1 Word2Vec モデルを用いた機械学習モデル構築

① 形態素解析

分析対象の口語テキストに対して，2）予備検討と同様の手順で形態素解析にて名詞抽出を行った．

② 同義語辞書作成

予備検討では，形態素解析後に BoW で抽出した名詞を特徴量化したが，単語の意味の近さの程度によらず同等の特徴量を単語に付与するため，予備検討のモデル選択だけでは，高精度の機械学習をさせることは難しいと予想した．そこで，予備検討で構築した機械学習モデルの精度向上を目的に，近い意味の単語に対して，同一の特徴量を付与するための新たな手法（同義語辞書の作成）を考案した．

同義語辞書の作成にあたっては，オープンソースである Python 3.7.12 ライブラリとして Gensim 3.6.0 の主なアルゴリズムのうち，Word2Vec モデルを用いて同義語辞書を作成した．本研究では，前後に出現する単語の分布傾向が近い単語同士は同義語の関係にあると仮定し，分析対象の口語テキストに対して Word2Vec モデルを用いて文脈情報から単語同士の意味的な距離関係を学習させ，単語をベ

クトル化し，「ウォード法」で近い意味の単語をクラスタリングすることで同義語辞書を作成した．

「ウォード法」とは，凝集型クラスター分析の手法の1つで，分類感度が高いため，一般的によく使用される手法である．よって，本研究でもウォード法を採用した．この手法は，それぞれのデータと平均値の差を二乗した値の和を求め，平方和が小さなものからクラスターを作成していく手法であり，最終的に1つのクラスターになる．なお，最適なクラスター数が未知の場合，クラスター数を変化させ比較することで最適な基準を定めるとされている [8]．本研究では，最適なクラスター数を決定するために，研究者がデンドログラムにて各クラスターの特徴を目視で確認したうえで，複数の閾値を適応し，複数の同義語辞書を作成した．

③　同義語作成

形態素解析後に BoW で抽出した名詞を特徴量化させる際，上述の 3）(1) ②で作成した同義語辞書 1 から 3 を用いて，近い意味の単語に対しては同一の特徴量を付与した．

④　精度評価

同義語辞書の有無別で，予備検討で構築したモデルの精度比較を行った．

2.3.2　潜在ニーズの可視化

機械学習モデルの予測に寄与した特徴量を可視化させるための手法である SHapley Additive exPlanations（SHAP）解析を用いることで機械学習モデルの解釈が可能となるため [9]，本研究では，分析対象の口語テキストから潜在ニーズの予測に寄与した特徴量である単語群（クラスター）を明らかにすることによって，モデルの有用性を評価した．

具体的には，SHAP 解析後，特徴量ごとの SHAP 値を図へプロットのうえ，予測値に対して大きな影響を与えている特徴量（単語のクラスター番号）を上から順に図に示した．

その後，高頻出の SHAP 値が大きい潜在ニーズ予測に対して，寄与度が高いクラスターの単語を含む文脈を確認したうえで，発話者の潜在ニーズの有無および解釈を行った．

表 1　潜在ニーズ予測モデル選択のための精度比較モデル
Table 1　Comparison of accuracy for selecting a latent needs prediction model.

モデル	F 値
決定木	0.50
ランダムフォレスト	0.01
xgboost	0.54

3.　結果

3.1　予備検討

予備検討の結果，最も精度が高ったモデルは xgboost であり，F 値 0.54 であった（**表 1**）．そのため，機械学習モデルとしては xgboost を選択した．

3.2　本調査

3.2.1　Word2Vec モデルを用いた機械学習モデル構築

同義語辞書の有無別に予備検討で構築したモデルの精度比較を行った．複数の同義語辞書作成のための閾値は，研究者がデンドログラムにて各クラスターの特徴を目視で確認したうえで，12.5，15，17.5 の 3 種の閾値を設定した．その結果，同義語辞書なしの場合は F 値 0.54 であったのに対して，同義語辞書 2 の場合，F 値 0.61 と最も高かった（**表 2**）．

3.2.2　潜在ニーズの可視化

最後に SHAP 解析で潜在ニーズ予測に寄与したクラスターの出現頻度とそれらが潜在ニーズの有無のどちらの予測に対してどの程度，寄与したのかを可視化した（**図 1**）．

潜在ニーズ予測に対して，寄与度が高いクラスター順に，［不安］，［病名，行動］，［意識，正常］，［場合］と続いた．また，［病名，行動］クラスターで，SHAP 値が大きく，かつクラスター出現頻度も高かった．たとえば，［病名］を含む，口語テキストでの前後の文脈としては，『(以下，転載）ところが，あの，現実，その病名を出さなきゃいけないので，レビー小体型認知症ということで，…（略）…，ほかの認知症とはやっぱりちょっと違う．だから，身体症状とか，それから，どうしてそれが起きるのかっていうことはきちっと勉強してもらわないと，やっぱり，ケアするにしても，気を付けなきゃいけないことがたくさんあるんですよね.』であった．研究者は，この発話者の潜在ニーズとして，「介護者に対して，本人自身との関わりでは，事前に疾患の特徴を理解して欲しい」というニーズが含まれていたと解釈した．

予測値に対して大きな影響を与えている特徴量（クラスター番号）を上から順に示した．各プロットは，クラスターが含まれるレコードで，クラスター出現頻度を低（青

表 2　Word2Vec モデルを用いた機械学習モデル予測精度
Table 2　Prediction accuracy of machine learning model using the Word2Vec model.

条件	F 値
同義語辞書なし	0.54
同義語辞書あり	
同義語辞書 1（閾値 12.5）	0.59
同義語辞書 2（閾値 15）	0.61
同義語辞書 3（閾値 17.5）	0.58

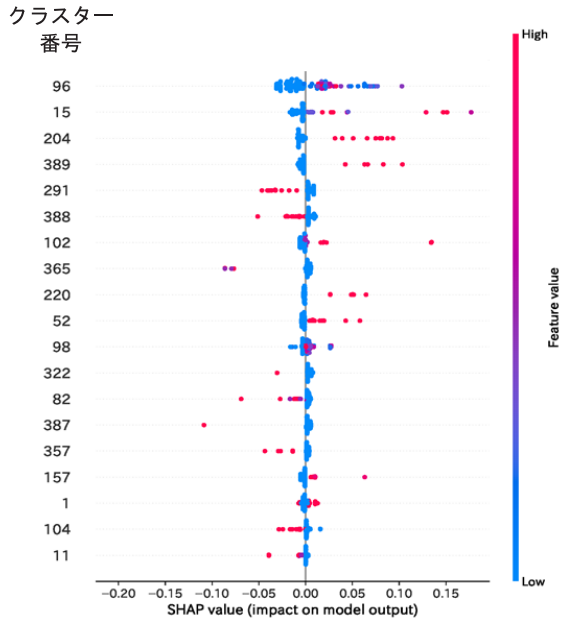

図 1　潜在ニーズ予測

Fig. 1　Latent needs prediction.

色）から高（赤色）で示した．横軸に潜在ニーズ有無のどちらの予測に対して寄与したのかを数値で示した．SHAP 値 0 起点に正がニーズあり，負がニーズなしへの寄与度を示す．

4.　考察

本研究は，口語テキストから発話者の潜在ニーズを自動抽出するための機械学習モデル構築とその有用性の検討，およびニューラルネットワークを用いて単語をベクトル変換する手法である Word2Vec モデルを用いて作成した同義語辞書の適応による機械学習モデルの精度改善を検討した．

予備検討の結果，3 つの木構造モデルの中でも xgboost を使用した場合が最も良い予測精度であった．xgboost とは，勾配ブースティングと呼ばれるアンサンブル学習と決定木を組み合わせた手法で非常に高い汎化能力があり，一般的に決定木よりも予測精度が高いと言われている [10]．そのため，本研究においても xgboost で最も予測精度が高かったと思われる．一方，ランダムフォレストで最も予測精度が低かったが，その理由としてバギングと呼ばれるアンサンブル学習と決定木を組み合わせた手法であった点が考えられた．バギングとは，学習データからランダムサンプリングしてサブ学習データを複数作成のうえ，それぞれのサブ学習データを用いて木構造モデルを作成する．それぞれの木構造モデルの予測結果を集約し最終的な結果を出力させる手法である．つまり，同義語辞書を用いて単語をクラスタリングせずに，学習データからサブ学習データを作成したことで，特徴量の分布の偏りが大きいサブ学習データが作成されたことから，予測精度が低くなったと考えられた．

表 3　潜在ニーズ予測に寄与した特徴量

Table 3　Features contributing to latent needs prediction.

クラスター番号	クラスター [単語群]
96	不安
15	病名，行動
204	意識，正常
389	場合
291	タイプ，自分のために，意外，受け入れ
388	本格的，仲，兄弟，尊敬
102	僕
365	夜
220	なに，姿，返事，動き，どん，攻撃，ズボン，シャツ，要求，二人，がさ，身体的，生き方，よろしくお願いします，しながら，帰宅，歩行，ごめんね
52	形
98	一人，ほんと
322	大学病院，通院，半年，看病，看護，疲れ，影響，女房
82	一緒
387	副作用，過敏，穏やか，日々，親せき，関西
357	的
157	申請，看護婦，ソーシャルワーカー，ケアマネジャー
1	何度，確認
104	嫌
11	お母さん

予測値に対して影響が高い順にクラスター番号を示した．

本調査において，Word2Vec モデルを用いた同義語辞書の適応が精度改善に寄与しており，本アプローチの有用性を確認することができた．同様の先行研究においては，単語の共起関係に基づく機械学習による文書分類を検討した福本らの報告があり，シソーラスの分類語を用いて，単語の特徴ベクトルである共起行列を生成する手法を提案されていたが，モデルとしてはランダムフォレストで正識別率 88.7% と高精度であったことから [11]，本アプローチは一定の有用性があるものを考えられた．

最後に，SHAP 解析を用いた潜在ニーズの可視化では，寄与度が高くかつ出現頻度が高い［病名，行動］クラスター文脈において，潜在ニーズとして解釈可能な口語テキストが含まれていたことが確認できた．精度が高い機械学習モデルを構築したとしても，その予測の根拠を十分に説明できない場合，その後，実用化につなげることは困難であるため，SHAP 解析を用いた潜在ニーズの可視化とその意味解釈には意義があったと言える．そのため，本研究で検討した新手法は，今後，社会実装可能な技術であったと示唆された．

研究の限界として，小規模なデータセットでの検討であったため，今後，さらに大規模なデータセットでの検討により，さらなる機械学習モデルの精度向上が見込まれる．

5. おわりに

口語テキストから発話者の潜在的ニーズを予測するための機械学習モデルに対して，ニューラルネットワークを用いて単語をベクトル変換する手法である Word2Vec モデルによる同義語辞書の適応はモデルの精度改善に寄与した．

謝辞　調査にあたっては，医薬基盤・健康・栄養研究所の藤浪淑子氏の協力に感謝の意を表する．本研究は，科学研究費助成事業（学術研究助成基金助成金）基盤研究（C）（一般）（課題番号 20K07206）の助成を受け実施された．

参考文献

[1] 金森　修，中島秀人：科学論の現在，勁草書房 (2002).
[2] エドワード・T・ホール：文化を超えて，TBS ブリタニカ (1993).
[3] Tanemura, N. et al.: Real World Survey of Patient Engagement Status in Clinical Research: The First Input from Japan, The Patient - Patient-Centered Outcomes Research, Vol.13, pp.623–632 (2020).
[4] 現代日本の生命医科学における疾患当事者の研究参画の研　究：〈https://kaken.nii.ac.jp/ja/file/KAKENHI-PROJECT-26750094/26750094seika.pdf〉（参　照 2022-5-24）.
[5] Mikolov, T. et al.: Efficient estimation of word representations in vector space, arXiv preprint arXiv: 13013781 (2013).
[6] 語りデータシェアリング：〈https://www.dipex-j.org/outline/data-sharing〉（参照 2020-10-01 日）.
[7] 白田由香利，橋本隆子：Web の口コミ情報からの潜在的事前期待の発見：@ コスメにおけるマスカラの評判分析，学習院大学 経済論集，Vol.52, No.1, pp.1–14 (2015).
[8] 志津綾香，松田眞一：クラスター分析におけるクラスター数自動決定法の比較，南山大学紀要『アカデミア』情報理工学編，Vol.11, pp.17–34 (2011).
[9] Lundberg SM, and Lee S-I: A unified approach to interpreting model predictions, Advances in neural information processing systems, Vol.30 (2017).
[10] Tianqi, C. and Carlos, G: Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp.785–794 (2016).
[11] 福元伸也，渕田孝康：単語の共起関係に基づく機械学習による文書分類，研究報告データベースシステム（DBS）, Vol.2014-DBS-160, No.28, pp.1–5 (2014).

種村 菜奈枝（非会員）

医薬基盤・健康・栄養研究所.

町井 湧介（非会員）

独立研究者.

佐々木 剛（非会員）

千葉大学医学部附属病院.

荒木 通啓（非会員）

医薬基盤・健康・栄養研究所.

佐藤 淳子（非会員）

医薬品医療機器総合機構.

千葉 剛（非会員）

医薬基盤・健康・栄養研究所.