原著論文

特定個人に対する単語の理解有無の推定における 単語概念の有用性に関する調査†

野澤 一真*¹・望月 典樹*¹・増山 岳人*²・中村 壮亮*³

本研究では個人の語彙の可視化に向けて、単語の理解有無の推定を行った。我々は、理解している単語や理解していない単語がそれぞれ特定の単語概念（単語が指示する概念）に集中しているとの仮説を立てている。そこで、word2vec で表現される単語概念と階層型ニューラルネットワークを用いて単語の理解有無を推定する手法を提案し、約 72% を超える正解率での理解有無の推定結果を得た。結果から、仮説の妥当性及び単語概念が理解有無推定に有用であることが示唆された。難易度などの別指標との併用により、実用水準での推定への期待も持てる結果といえる。

キーワード：語彙推定、ニューラルネットワーク、word2vec、単語概念

1. はじめに

近年、インターネット利用の増大と IoT の普及により、大量のデジタルデータの生成、収集、蓄積が進みつつある。それらデータの機械学習による分析結果を業務処理の効率化、最適なアドバイスの提供、効率的な機械の制御等に活用することで、新たな価値を創造し、社会的課題を解決するデータ主導社会の実現が図られている [1]。また、2017 年 5 月に施行された改正個人情報保護法により、個人情報利用の障壁が下がった事を受けて、個人に関するデータを記録したライフログを利活用するビジネスや研究への期待が高まっている。ライフログを活用した研究の例として、購入履歴の利用による消費者の嗜好・ニーズに合致した商品やサービスの推薦 [2]、SNS での投稿履歴の利用による行動パターンの推定 [3]、生体情報の利用によるストレス計測 [4] といったものが挙げられる。

このような社会的情勢を受けて、我々は個人の語彙（使いこなせている単語の集合）に基づいた知能補助サービスの検討を始めている。具体的には発話ログや文書データ等のライフログを解析して個人の語彙を可視化し、それをサービスへと応用する展開を考えている。本研究では語彙の可視化を、ライフログに記録されていない単語の理解有無についてもライフログに記録された単語を元に推定することで、全ての単語の理解有無を推測することと定義する。

知能補助サービスの例を挙げる。片山ら [5] が提案している検索システムにおいて個人の知らない知識に対して注釈をつけ

るといった情報提示サービスが考えられる。その他にも、聞き手の語彙に応じたロボットの発話生成、語彙能力劣化診断による認知症の早期認知などが挙げられる。

これらのサービスを確立するうえで、語彙の可視化が必要不可欠となる。そこで、本研究の目的を、個人の語彙の可視化に向けた「単語の理解有無の推定機能の開発」とする。

本研究における「理解している単語」とはその単語の意味を一つ以上説明できることとする。単語を使いこなすことはその意味を理解することから始まる。一部には理解しているが使いこなせない単語があるが、その割合は小さいと考えられるため、理解有無を推定すれば語彙の可視化につながると考えられる。従って、個人の語彙を把握するために、本研究では単語の理解有無の情報を利用することとした。

類似研究を紹介する。四方らの研究では認知症患者によるブログから語彙量が時間経過とともにどのような推移をたどるのかを調査している [6]。また、Segbers らの研究では言語コーパスと被験者の理解有無に基づいて、児童の語彙量の変化を測定している [7]。しかしながら、これらは語彙量の「測定」であり、ブログやヒアリングで出現しない語彙の「推定」までには至っていない。

ブログ等の記録媒体では理解していない単語の収集が出来ず、理解していない単語も含めた個人の語彙を網羅することは困難である。さらには、使用単語の偏りもある。一方で、ヒアリングでは理解していない単語もデータベース化出来るが、多くの単語の理解有無をデータベース化することは時間の制限から限界がある。以上より、全数測定せずに一部のサンプルから、個人の語彙を広く把握するため、機械学習等を用いた語彙の推定が有用と考えられる。

そのような単語の理解有無の推定に関する研究も行われているので紹介する。江原らは英文 web ページに対して、知らない英単語を個人の英語力に応じて訳を予め付与する読解支援システムを提案している [8, 9]。個人がクリックした単語を解析することによって、知らない英単語を項目応答理論を用いて推定している。推定の際には基本的な語彙 12000 語に対し、英語

† Survey on the Usefulness of Word Concepts in Estimating the Understanding of Words for Specific Individuals
Kazuma NOZAWA, Noriki MOCHIZUKI, Gakuto MASUYAMA, and Sousuke NAKAMURA

*1 法政大学大学院理工学研究科

Graduate School of Science and Engineering, Hosei University

*2 名城大学理工学部

Faculty of Science and Engineering, Meijo University

*3 法政大学理工学部

Faculty of Science and Engineering, Hosei University

母語話者が人手で 12 段階の難易度をつけた難易度リストを使用している。また、片山らの研究ではトピック適応を用いて、個人ごとに難易度をカスタマイズし、専門用語の単語に対して理解有無の推定を行なっている [5]。

ここで、難易度は人手で決めた指標と単語の生起確率に基づいて算出した指標に大きく分けられる。前者は人の主観が入ることや人的コストがかかる。一方で後者は個人が感じる難易度と生起確率に基づいた難易度に差が生じるといわれている [5]。さらに、難易度を用いた推定はある法則で決められた仮想的なスケール上で判断しているため、個人が獲得している単語概念の集合（個人の理解している領域、専門領域）においても、決められた難易度が高ければその単語を理解していないと誤判定してしまう。このように、難易度は有効な指標ではあるもののそれのみでは個人の語彙推定には不十分である。そこで、難易度とは独立した有効な指標の導入による推定精度の向上が考えられる。

本研究では単語が指示する概念（単語概念）に着目した。我々は、興味、嗜好および生活環境に従って、理解している単語や理解していない単語がそれぞれ特定の単語概念に集中しているとの仮説を立てている。この仮説が正しければ、個人の理解情報が既知である単語概念を学習器に学習させることで未知の単語の理解有無を推定が可能と考えられる。最終的には、関連研究にある難易度の指標と本稿で扱う単語概念の指標を併用することでさらなる精度向上が見込めると考えている。

以上より、本研究では単語概念（word2vec で表された単語ベクトル¹⁾）を用いた単語の理解有無推定手法を提案し、未知の単語に関して個人の理解有無の推定がどの程度可能であるかを調査する。

2. 提案する単語の理解有無の推定手法

2.1 概要

単語概念によって個人の単語の理解有無の推定を行うために word2vec を利用した。word2vec で表現された単語ベクトルを入力とし、被験者にヒアリングで収集した理解有無を出力とした階層型ニューラルネットワーク（図 1）で学習させることで単語の理解有無の推定を実現する。提案手法を図 2 に示す。また、以下で記述する。

2.2 word2vec を用いた単語のベクトル表現

単語の概念を数値で表現するために word2vec を使用した。word2vec[10, 11] はニューラルネットワークを用いて単語の分散表現（単語の概念を表すベクトル表現）を獲得する手法である。学習手法として、Skip-gram モデルと CBOW モデルが提案されている。Skip-gram モデルはある単語が与えられた際に

図 1 階層型ニューラルネットワーク

その周辺に出現する単語を予測し、分散表現を獲得する。一方、CBOW モデルは周辺にある単語から対象単語を予測し、分散表現を獲得する手法である。獲得した単語ベクトルは単語概念に近いものになり、類似語はベクトル空間で近い関係となる。そのため、近年では、単語ベクトルは機械翻訳や文書分類等にも応用されている。

本研究では word2vec で生成した単語概念を捉えた単語ベクトルを利用して、単語の理解有無を推定した。なお、word2vec は大量のテキストデータを学習させる必要があるため、本研究では wikipedia[12] の記事を使用し、各単語をベクトル化するための word2vec のモデルを生成した。

2.3 ニューラルネットワークを用いた単語の理解有無推定

単語の理解度推定のための学習モデルとしては、入力層、隠れ層及び出力層からなる階層型ニューラルネットワークを使用した。ニューラルネットワークは人の脳神経系のニューロンを模範したネットワーク構造であり、入力層から特徴量を伝搬し、出力を決定する。本研究では、word2vec で表現した単語ベクトル $(x_1, x_2, \dots, x_{200})^T$ を入力として、理解度 y_1 及び非理解度 y_2 を出力とした。そして、被験者に対して理解の有無をヒアリングした単語に対して、その word2vec による単語ベクトルを入力とし、出力を $t_1 = 1, t_2 = 0$ （理解している場合）または $t_1 = 0, t_2 = 1$ （理解していない場合）としたものを教師データとした。教師データとしては被験者の「理解している単語」と「理解していない単語」の両方をバランス良く用意することで「理解している単語の特徴」と「理解していない単語の特徴」を識別するモデルが生成できると期待される。

これは、理解している単語の集合と理解していない単語の集合は word2vec で表される単語ベクトル空間上で複雑に交錯していると予想されるがニューラルネットワークは複雑な非線形分離を可能とすることによる。

ここで、本研究におけるモデルの汎化性能と過学習についての考えを説明する。一般的に汎化性能は未知データに対して正解する能力であるが、本研究では「モデルが学習した対象被験者の未知データに正解する能力（個人内の汎化能力）」と「対象ではない被験者の未知データに正解する能力（被験者間の汎化能力）」と分けて考える。本研究で向上したい汎化性能は後者の被験者をまたいだ被験者間の汎化性能ではなく、前者の対象被験者において、データベースに記録されていない単語まで理解有無を推定する個人内の汎化性能である。また、過学習は

1 word2vec で表された単語ベクトルは、単語の多義性を考慮できないなど、完全には単語概念を表しているとはいえない。しかしながら、単語の意味に基づいた演算が可能など、類似語は単語ベクトル空間で近い関係になるという点は単語の意味を表現しようとするもので、単語概念に近いものがある。本論文では単語概念を用いた理解有無推定というコンセプトの実現可能性を探ることに主眼を置いたため、厳密さにはこだわり過ぎず、単語概念に類する数値表現として word2vec を利用することとした。

図2 提案する理解有無推定手法のネットワーク構造

「モデルが学習した対象被験者の一部の単語の理解有無のみを捉えていて、単語全体の理解有無を捉えていない状態」である。

3. 評価実験

3.1 データの収集方法

被験者 8 人（表 1）に単語リストから無作為に単語を提示し、理解している/いないのヒアリングを行うことで、データを収集した。被験者としては 20 代前半の大学生 6 名及び大学院生 2 名を対象とした。本研究は初めに述べた「理解している単語」や「理解していない単語」がそれぞれ特定の単語概念に集中しているという仮説の検証が重要であるため、文理や専攻をまたいで被験者を募った。被験者にはデータ収集当日に収集方法を説明し、順次提示される単語に対して、意味を一つ以上説明できる場合に「理解している単語」、それ以外を「理解していない単語」に分類するよう指示した。理解している単語と理解していない単語がそれぞれ 500 個になるまで単語を提示した。なお、単語の意味がうろ覚えで曖昧な場合には、回答せずに飛ばすように被験者に注意した。

提示する単語は wikipedia[12] の記事から抽出した。公開されている wikipedia には学習には必要のない XML のコンテナ部分や独特な記号を含むため、それらを削除し、形態素解析を用いて文章を単語に分割し、名詞のみ抽出した。そこから、wikipedia 内で頻度が 50 以下かつ文字数が 10 文字以上の単語を削除し、計 20 万単語を単語リストに保存した。形態素解析には MeCab[13] を使い、辞書として、mecab-ipadic-NEologd[14] を用いた。mecab-ipadic-NEologd は、週 2 回以上更新され、固有表現に強く、語彙数が多いため、最新の造語への対応が期待できる。

3.2 word2vec の学習条件

word2vec の学習には自然言語処理ライブラリ gensim[15] を使用した。学習させるモデルは CBOW を使い、word2vec の出力を 200 次元とした。ウィンドウサイズは 5、頻度が 5 未満

表 1 被験者詳細

被験者	専攻	文理	年齢	性別
A	電気電子工学	理系	23	男
B	電気電子工学	理系	22	男
C	電気電子工学	理系	21	男
D	電気電子工学	理系	22	女
E	システムデザイン工学	理系	20	男
F	生命科学	理系	20	男
G	経営学	文系	21	男
H	心理学	文系	21	女

表 2 word2vec の学習条件

オプション	説明	パラメータ
size	単語ベクトルの次元	200
sg	学習モデル	CBOW
window	文脈の最大単語数	5
min_count	破棄する単語の登場回数	5
sample	単語を無視する頻度	0.001
negative	ネガティブサンプリングに用いる単語数	5

の単語は学習に使用しなかった。他の設定は gensim のデフォルト値を使用した。また、学習に用いた wikipedia は被験者に提示した wikipedia と同じ更新日のものを使用した。word2vec の学習条件を表 2 に示す。

3.3 ニューラルネットワークの学習条件

単語の理解有無推定の学習モデルであるニューラルネットワークの構築には tensorflow をバックエンドとした keras を使用した。ニューラルネットワークには word2vec で変換した単語ベクトルを入力、その単語の意味理解の有無を出力とする教師データを学習させた。

各層のノードには入力した値 x に重み w が掛けられ、次の層のノードに伝搬され、各ノードの出力は

図3 評価方法（被験者2人の場合）

$$z_j = h \left(\sum_{i=1}^M w_{ji} x_i \right) \quad (1)$$

となる。ただし、 $i(1, 2, \dots, M)$ は前の層のノード、 $j(1, 2, \dots, K)$ は出力 z のノード、 h は活性化関数とする。

隠れ層のノードは予備実験において正解率が良い、200個とし、活性化関数は Rectified Linear Unit (ReLU) とした。ReLU は

$$relu(u) = \max(0, u) = \begin{cases} 0 & (u < 0) \\ u & (u \geq 0) \end{cases} \quad (2)$$

と定義される。

また、出力層のノードは2個（理解している/していない）とし、活性化関数は Softmax とした。Softmax は

$$softmax(u) = \frac{\exp(u_k)}{\sum_{j=1}^K \exp(u_j)} \quad (3)$$

と定義される。ただし、 K はクラス数 ($K = 2$)、 k は出力層の各ノードとする。

損失関数 E は推測値 y と目標値 t (教師データ) のクロスエントロピー誤差とした。損失関数 $E(w)$ を (4) 式に示す。

$$E(w) = -\frac{1}{N} \sum_{n=1}^N \sum_{k=1}^2 t_{nk} \log y_{nk} \quad (4)$$

ただし、 $n(1, 2, \dots, N)$ は学習に使用する訓練データとする。

ニューラルネットワークの重み w は誤差逆伝播法で最適化し、勾配更新には Adam[16] を用いた。Adam を (5) 式から (9) 式に示す。

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1) \nabla E(w^t) \quad (5)$$

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2) \nabla E(w^t)^2 \quad (6)$$

$$\hat{m} = \frac{m_{t+1}}{1 - \beta_1^t} \quad (7)$$

$$\hat{v} = \frac{v_{t+1}}{1 - \beta_2^t} \quad (8)$$

$$w^{t+1} = w^t - \alpha \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon} \quad (9)$$

ただし、 $m_0 = 0$ 、 $v_0 = 0$ 、 $\alpha = 1.0 \times 10^{-5}$ 、 $\beta_1 = 0.9$ 、 $\beta_2 = 0.999$ 、 $\epsilon = 1.0 \times 10^{-8}$ とした。

表3 ニューラルネットワークの学習条件

入力層	次元	200
	次元	200
隠れ層	活性化関数	ReLU
	Dropout 率	0.5
出力層	次元	2
	活性化関数	Softmax
バッチサイズ		50
最大 epoch		1000
損失関数		クロスエントロピー誤差
重みの初期値		He
最適化手法		Adam

また、重みの初期値は ReLU 関数と相性が良い「He の初期値」[17] を用いた。

過学習を避けるためにランダムに隠れ層のノードを無視する Dropout[18] を適用した。Dropout 率を 50[%] とし、隠れ層のノードをランダムに消した。Dropout を適用することでモデルの汎化能力向上が期待される。

今回、隠れ層が1層と2層の場合で比較した。隠れ層が1層のときの各層のノード数は入力層:200、隠れ層:200、出力層:2 とし、隠れ層が2層のときの各層のノード数は入力層:200、1 番目の隠れ層:200、2 番目の隠れ層:200、出力層:2 とした。word2vec で 200 次元に圧縮しているため、情報量を保つために、隠れ層のノードは word2vec と同じ次元数にした。

ニューラルネットワークの学習条件を表3に纏める。

3.4 ニューラルネットワークの評価方法

未知データに対する認識率を評価するために 5-fold cross validation を利用した。評価方法の概要を図3に示す。

各被験者から収集した単語の理解有無のデータは、学習モデルの教師データと評価用のテストデータに分類される。ここで、教師データと同じ被験者のテストデータで学習モデルを評価する個人内の交差検証、生成されたモデルの個人差を検証するために教師データと異なる被験者のテストデータで評価する被験者間の交差検証の2種類の評価方法で評価した。評価手順を以下に示す。

1. 被験者 S(A~H) の収集した全単語（1000 単語）をグルー

プ X_{S1} , X_{S2} , X_{S3} , X_{S4} , X_{S5} に分割する. なお, 分割したグループには理解している単語と理解していない単語の数をそれぞれ同じにする.

- X_{Si} を除いた 4 グループで学習し, X_{Si} を用いて正解率を算出する. このとき, X_{Si} をテストデータとし, それ以外を教師データとし, その際に生成されるモデルを M_{Si} とする.
- 各試行の正解率 C_{Si} を算出する. 正解率 C_{Si} は

$$C_{Si} = \frac{TP+TN}{TP+FN+FP+FN} \quad (10)$$

と定義される. ただし, 被験者 S が理解していて, モデル M_{Si} も理解していると判定した数を TP, 被験者 S が理解してなくて, 語彙モデル M_{Si} が理解していると判定した数を FN, 被験者 S が理解していて, 語彙モデル M_{Si} が理解していないと判定した数を FP, 被験者 S が理解してなくて, 語彙モデル M_{Si} が理解していないと判定した数を TN とした. 正解とモデルの予測の関係である混同行列を表 4 に示す.

- この手順を全ての $i(1 \leq i \leq 5)$ について行い, 5 回の正解率の平均を個人内交差検証の正解率 C_S とする.
- また, 正解率 C_S と同様に理解している/していないのそれぞれの再現率 (*Recall*:理解している単語の再現率, *Specificity*:理解していない単語の再現率) を求めて評価した. 今回, 評価した再現率を (11), (12) 式に示す.

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$Specificity = \frac{TN}{FP+TN} \quad (12)$$

- モデルの個人差を検証するために, 被験者間の交差検証を行う. 被験者 S の回答結果である検証データ X_{Si} を語彙モデル M_{Si} を除く, 他の被験者の語彙モデル $M_{\bar{S}i}$ に分類させる. 今回, 被験者が 8 人であるため, 1 つのモデルに対して, 他の被験者のテストデータ 7 個を分類させた. その際の評価指標は個人内の交差検証と同様に (10) 式を用いた.
- 他の試行 $i(1 \leq i \leq 5)$ についても行い, 5 回の正解率の平均をモデル $M_{\bar{S}i}$ の正解率 $C_{\bar{S}}$ とする.
- これを全ての被験者に対して行う.

3.5 単語ベクトルの評価方法

単語ベクトル空間上における理解している単語と理解していない単語の特徴を分析するため, 単語ベクトルの分散と k-means 法で評価する.

3.5.1 単語ベクトルの分散による評価

単語の理解有無ごとに単語ベクトルの分散を求めて, 比較し評価する. 手順を以下に示す.

- 被験者 $S(A \sim H)$ の単語ベクトルを最大値 1, 最小値 0 になるように正規化する.
- 単語ベクトルの成分ごとに理解している単語と理解していない単語の分散を求める.
- 2 で求めた分散の合計をそのクラスの分散とする.

表 4 モデル検証のための混同行列

		モデルの予測	
		理解している	理解していない
正解 (被験者の回答)	理解している	TP	FN
	理解していない	FP	TN

- 3 の結果を大小比較し, 大きい方を分散が大きいクラスとする.

3.5.2 k-means 法による評価

クラスごとに k-means 法を用いてクラス数と残差平方和 (SSE 値) で評価する. SSE 値は

$$SSE = \sum_{n=1}^{500} \sum_{k=1}^K r_{nk} (\mathbf{x}_n - \mu_k)^2 \quad (13)$$

と定義される. ただし, $\mathbf{x}_n$ は n 番目の単語ベクトル, μ_k はクラス k の重心 (セントロイド) である. また, r_{nk} は

$$r_{nk} = \begin{cases} 1 & \left(k = \arg \min_k (\mathbf{x}_n - \mu_k)^2 \right) \\ 0 & (otherwise) \end{cases} \quad (14)$$

である. 手順を以下に示す.

- 被験者 $S(A \sim H)$ の理解している単語の集合をランダムに K 個のクラスに割り当て, 初期クラスを決定する. ただし, k-means 法は初期クラスに結果が依存しやすいため, 異なる初期クラスを用い, 10 回試行する. その中で, 最も SSE 値が小さいモデルを最終モデルとして選択する.
- 各クラスについて重心を計算する. 重心は

$$\mu_k = \frac{\sum_{n=1}^N r_{nk} \mathbf{x}_n}{\sum_{n=1}^N r_{nk}} \quad (15)$$

と定義される.

- 単語ベクトル $\mathbf{x}_n$ を各クラス k の重心 μ_k との距離 $\mathbf{x}_n - \mu_k$ を最小にするクラス k へ割り当てる.
- 前の反復とクラスに変化がないか反復数が最大イテレーション回数を超えたら終了する. そうでなければ 2 に戻る. 最大イテレーション回数は 300 とした.
- 理解していない単語も同様に行う.
- これを全ての被験者の回答した単語に対して行う.

3.6 結果・考察

全被験者の正解率とその平均を表 5 に示す. また, 最高正解率時の全被験者の再現率を表 6 に示す. 隠れ層が 2 層の場合の方が正解率は高く, 216[epoch] 時は平均で 73[%] であった. 一人につき評価した単語数は 1000 個であったため, 平均で 730 個は正解したこととなる. 最高正解率は 216[epoch] 時の被験者 B の 79.5[%] であった.

ラベルごとの再現率は隠れ層が 1 層と 2 層共に理解している単語より理解していない単語の方が高い結果であった. 被験者ごとの理解している単語と理解していない単語の分散を表

表 5 全被験者の正解率と平均

被験者		A		B		C		D		E		F		G		H		average	
隠れ層		1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
epoch	10	68.4%	69.3%	67.8%	67.0%	64.2%	64.2%	62.4%	63.2%	62.1%	65.5%	64.4%	66.3%	62.4%	60.9%	65.0%	63.8%	64.6%	65.0%
	100	74.9%	74.2%	78.3%	78.5%	72.6%	72.8%	69.4%	71.1%	70.7%	70.3%	74.1%	74.1%	70.4%	71.5%	69.6%	69.3%	72.5%	72.7%
	121	75.4%	73.8%	79.0%	78.7%	73.0%	73.0%	69.5%	71.1%	70.5%	70.6%	74.8%	74.7%	70.6%	71.5%	70.5%	69.1%	72.9%	72.8%
	216	75.1%	73.4%	78.4%	79.5%	72.6%	73.7%	69.8%	71.9%	71.1%	69.9%	74.8%	75.6%	70.7%	70.7%	69.3%	69.2%	72.7%	73.0%
	1000	74.3%	74.0%	77.8%	78.3%	71.0%	71.7%	69.9%	70.0%	69.0%	68.7%	72.7%	72.9%	68.3%	70.5%	68.0%	67.6%	71.3%	71.7%

表 7 理解している単語と理解していない単語の分散

被験者	理解している単語	理解していない単語
A	2.538	1.056
B	2.322	0.609
C	2.671	0.827
D	2.412	1.039
E	2.185	0.668
F	2.398	0.881
G	2.356	0.807
H	2.462	0.951

表 6 最高正解率時の性能評価

隠れ層	epoch	再現率		正解率
		理解している	理解していない	
1 層	121	69.6%	76.3%	72.9%
2 層	216	71.6%	74.4%	73.0%

図 4 クラスタ数による SSE 値の変化

7, 各クラスの単語のクラスタ数による SSE 値の変化を図 4 に示す。図 4 より、どのクラスタ数においても理解していない単語より理解している単語の方が高い SSE 値を示した。SSE 値は各クラスタの重心からの単語ベクトルの距離の総和であるため、SSE 値が高いということは単語ベクトル空間上で分散している可能性があるといえる。さらに、表 7 より、理解している単語が理解していない単語より、単語ベクトルの分散がどの被験者も大きかった。これらより、理解していない単語の再現率が高かった理由として、理解していない単語は単語ベクトル空間上で偏って分布しており、一方で理解している単語は単語ベクトル空間上で全体に分散しているため、識別が困難となった可能性が高いと考えられる。

モデルの個人差を検証するために、被験者の回答結果であるテストデータを他の被験者のモデルに分類させた結果を図 5、個人内の交差検証（教師データと同じ被験者のテストデータで評価した正解率）と被験者間の交差検証（教師データと違う被

験者のテストデータで評価した正解率）の平均の比較を図 6 に示す。図 5、図 6 共にニューラルネットワークは隠れ層が 1 層、121[epoch]とする。図 5 より、個人内交差検証の正解率が被験者間交差検証のどの正解率よりも高かったのは被験者 A, B, C, F, H であった。いずれの被験者においても、訓練データと同じ被験者のテストデータで評価した個人内交差検証が最も高くなることを予想していたが、D, E, G の被験者ではそのようにはならなかった。しかしながら、個人内交差検証の正解率が被験者間交差検証の正解率を下回った被験者に着目すると、被験者 D は 4 番目、被験者 E と被験者 G は 2 番目であり、1 番高くはならなかったが比較的上位の正解率の高さであった。このため、図 5 より、個人内交差検証の正解率と被験者間交差検証の平均正解率を比較した結果は、図 6 に示す通り、個人内交差検証の正解率の方が高い結果となった。一方で、被験者間交差検証で違う被験者のモデルにも関わらず、一定以上、正解した要因として、被験者間の汎化性能がある程度あるといえる。こ

図5 個人差評価（無地が個人内交差検証，斜線が被験者間交差検証）

図6 個人内交差検証の正解率と被験者間交差検証の平均正解率の比較

表8 welch の t 検定結果

被験者	p 値
A	6.0E-03
B	2.3E-02
C	2.0E-02
D	4.3E-01
E	7.9E-02
F	1.0E-08
G	6.8E-02
H	8.3E-02

れは被験者間で共通概念として獲得している単語もあるためと考えている。図6の個人内交差検証の正解率と被験者間交差検証の正解率の平均の大小関係が有意であるか Welch の t 検定を用いて検証した結果を表8に示す。被験者 A, B, C, F は被験者間交差検証より個人内交差検証の正解率の方が大きいことが片側で有意 ($p < 0.05$) であると認められた。また、被験者 E, G, H は有意であると認められないが片側で有意傾向 ($p < 0.1$) ではあった。以上より、学習モデルが被験者ごとの理解している/していないの識別能力をある程度獲得したといえる。

4. 結論

4.1 まとめ

本研究では、個人の語彙の可視化に向けて、単語概念を用いた単語の理解有無の推定手法を提案し、基礎実験の結果を報告した。実験結果より、理解している単語と理解していない単語を約 72% 超える正解率で識別が可能であった。また、ラベルごとの再現率は理解している単語より理解していない単語の方が高い結果であり、単語ベクトルの分散の比較と k-means 法より、理解している単語の方が単語ベクトル上で分散していることが示唆された。

このように一定の正解率が得られたことは、「理解している単語の集合」と「理解していない単語の集合」が word2vec で

表される単語ベクトル空間上で偏りがあったことによると考えられる。また、この偏りは個人差があることが被験者間の交差検証結果から示唆された。これは理解している単語や理解していない単語がそれぞれ特定の単語概念に集中しているという仮説の妥当性を示すものであり、単語概念による個人の単語の理解有無の推定の有用性が示唆された。

4.2 今後の展望

現在、単語概念を用いた推定において、いくつかの問題点が浮き彫りになっている。一つ目に同義語は単語ベクトル空間上において、距離が近いと、ニューラルネットワークでの理解有無の識別が難しい。二つ目に理解している単語は理解していない単語に比べて、ベクトル空間上で全体に分散している可能性が高いため、単語概念での識別精度が落ちる。しかし、関連研究で挙げた難易度を学習の際の指標として併用することなどで、正解率向上が見込めると予想している。そこで、今後は単語概念に関連研究で述べた難易度を取り入れ、精度向上に努めたい。

また、ニューラルネットワークにおける最適なパラメータの調査や、他のアルゴリズムにおける識別可能性の検討を行い、単語概念を用いた理解有無推定における最適なアルゴリズムを調査していきたい。

さらに, word2vec は他言語での学習も可能であるため, 英語等の他言語の理解有無の推定が可能かについても検証する必要がある. 日本人の英語学習者の支援や言語に依存しない語彙推定が可能となる可能性がある.

最後に, 語彙推定が確立した暁には, 知能補助サービスに取り入れ, 社会実装を図っていきたい.

参考文献

- [1] 総務省: 平成 30 年度版 情報通信白書 データ主導経済と社会変革, 2018.
- [2] 大井貴裕, 三川健太, 後藤正幸: “評価と購買の両履歴データの学習による確率的潜在クラスモデルの推定精度向上に関する一考察,” 日本経営工学会論文誌, Vol.65, No.4, pp. 286-293, 2015.
- [3] 潘睿, 松尾豊: “個人の生活管理のためのソーシャルデータからの行動パターンの発見,” 人工知能学会誌, Vol.28, No.6, pp. 829-834, 2013.
- [4] 前田直樹, 平部裕子, 荒川豊, 安本慶一: “オフィスワークのための自然なストレスモニタリングシステムの提案と設計,” マルチメディア, 分散協調とモバイルシンポジウム 2016 論文集, pp. 1788-1793, 2016.
- [5] 片山太一, 小林のぞみ, 牧野俊朗, 松尾義博: “トピック情報を利用したユーザの知識推定,” 人工知能学会全国大会, 2013.
- [6] 四方朱子, 荒牧英治: “言語能力検査としての言語処理: 長期間のブログ執筆を継続した認知症の 1 例,” 言語処理学会第 20 回年次大会, pp. 1126-1129, 2014.
- [7] J. Segbers and S. Schroeder: “How many words do children know? A corpus-based estimation of children’s total vocabulary size,” *Language Testing*, Vol.34, No.3, pp. 297-320, 2016.
- [8] 江原遥, 二宮崇, 清水伸幸, 中川裕志: “ユーザが知らない語を予測する読解支援システム SocialDict とそのリーダビリティ測定への拡張,” 情報処理学会研究報告, No.1, 2010.
- [9] 江原遥, 二宮崇, 清水伸幸, 中川裕志: “Web ページ中のユーザが知らない語を予測する読解支援システム,” 人工知能学会全国大会, 2010.
- [10] T. Mikolov, W. Yih, and G. Zweig: “Linguistic regularities in continuous space word representations,” *NAACL*, 2013.
- [11] X. Rong: “word2vec Parameter learning explained”, arXiv: 1411.2738, 2014.
- [12] Wikipedia dumps data: <https://dumps.wikimedia.org/jawiki/> [accessed May. 5, 2018]
- [13] MeCab: Yet Another Part-of-Speech and Morphological Analyzer: <http://taku910.github.io/mecab/> [accessed Apr. 30, 2018]
- [14] mecab-ipadic-NEologd: Neologism dictionary for MeCab: <https://github.com/neologd/mecab-ipadic-neologd> [accessed May. 5, 2018]
- [15] gensim topic modelling for humans: <https://radimrehurek.com/gensim/> [accessed Mar. 1, 2018]
- [16] D. P. Kingma and J. Ba: “Adam: A Method for Stochastic Optimization,” arXiv:1412.6980, 2014.
- [17] K. He, X. Zhang, S. Ren and J. Sun: “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” arXiv:1502.01852, 2015.
- [18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov: “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *J. of Machine Learning Research*, Vol.15, No.1, pp. 1929-1958, 2014.

(2018 年 9 月 23 日 受付)

(2019 年 3 月 1 日 採録)

[問い合わせ先]

〒184-8584 東京都小金井市梶野町 3-7-2

法政大学大学院理工学研究科

野澤 一真

E-mail: kazuma.nozawa.0703@gmail.com

著者紹介

のざわ かずま
野澤 一真 [非会員]

2018 年法政大学理工学部電気電子工学科卒業. 現在, 同大学大学院理工学研究科電気電子工学専攻修士課程在学中. 機械学習及び自然言語処理による特定個人の語彙推定に関する研究に従事.

もちづき のりき
望月 典樹 [非会員]

2013 年中央大学理工学部電気電子情報通信工学科卒業. 2015 年中央大学大学院理工学研究科電気電子情報通信工学専攻修士課程修了. 2017 年より法政大学大学院理工学研究科電気電子工学専攻博士課程に在籍中. VR インターフェース, 人間拡張に関する研究に従事.

まさやま たかくと
増山 岳人 [非会員]

2005 年名古屋大学工学部卒業. 2013 年東京大学大学院工学系研究科博士課程修了. 博士(工学). 中央大学理工学部助教を経て, 2017 年より名城大学理工学部准教授, 現在に至る.

なかむら そうすけ
中村 壮亮 [非会員]

2005 年 3 月東京大学工学部電気工学科卒業. 2007 年 3 月同修士課程修了. 2012 年 3 月同博士課程修了. 工学博士. 2012 年中央大学助教, 2016 年法政大学理工学部電気電子工学科専任講師, 2018 年同准教授, 現在に至る. 個人適応型の身体拡張システム, ウェアラブルデバイスへの無線電力伝送の研究に従事.

Survey on the Usefulness of Word Concepts in Estimating the Understanding of Words for Specific Individuals

by

Kazuma NOZAWA, Noriki MOCHIZUKI, Gakuto MASUYAMA, and Sousuke NAKAMURA**Abstract:**

In this study we estimated the understanding of words in order to visualize a person's individual vocabulary. We have hypothesized that words which people understand or, conversely, do not, tend to be concentrated on certain *word concepts* (concepts indicated by the word). Therefore, we proposed a method for estimating the understanding of words combining word concept created by word2vec and Hierarchical Neural Network, and succeeded in estimating the understanding of words with an accuracy of over 72%. From the results, it was suggested that the validity of the hypothesis and the word concept were useful for estimating the understanding of words. With implementation of additional indicators such as difficulty, estimation at a practical level can be anticipated.

Keywords: vocabulary estimation, neural network, word2vec, word concept

Contact Address: **Kazuma NOZAWA**

Graduate School of Science and Engineering, Hosei University

3-7-2, Kajinocho, Koganei-shi, Tokyo 184-8584, Japan

E-mail: kazuma.nozawa.0703@gmail.com