The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

語彙の可視化を目的としたライフログ解析による単語の理解有無の推定

Estimation of understanding of words by lifelog analysis for vocabulary visualizing

野澤 一真[*1] 望月 典樹[*1] 増山 岳人[*2] 中村 壮亮[*3]
Kazuma Nozawa Noriki Mochizuki Gakuto Masuyama Sousuke Nakamura

[*1] 法政大学大学院理工学研究科電気電子工学専攻
Department of Science and Engineering, Hosei University Graduate School

[*2] 名城大学理工学部電気電子工学科
Faculty of Science and Technology, Meijo University

[*3] 法政大学理工学部電気電子工学科
Faculty of Science and Technology, Hosei University

In this study, we made an estimate understanding of words for visualization of individual vocabulary. We have hypothesized that each of words which a person understand and does not are concentrating on certain word concept. Therefore, we proposed a method for estimating understanding of words with hierarchical neural network and word2vec, and succeeded in estimating the understanding of words by accuracy of over about 96%. From these results, it has been shown the validity of the hypothesis and that it is possible to estimate the understanding of word by life log analysis.

1. はじめに

近年，機械学習，IoT 等の技術が進み，社会的課題を解決するためにビックデータを利活用するデータ主導社会の実現が図られている[総務省 2017]. 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提供[大井 2015], SNS での投稿履歴の利用による生活パターンの推定[潘 2013], 生体情報の利用によるストレス軽減[前田 2016]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）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.

知能補助サービスの例を挙げる. [片山 2013]で提案されているような，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，特定個人と同じ語彙能力を持ったアンドロイドの実現などが挙げられる.

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.

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，「理解している単語」と「理解していない単語」の特徴を学習器に学習させることで理解有無の推定が可能であると考えられる.

2. 関連研究

文書データや生体情報から単語の理解有無の測定や推定が行われている. [四方 2014]は認知症患者によるブログから語彙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. しかしながら，[四方 2014]は語彙の「測定」であり，ブログに出現しない語彙を扱う「推定」までには至っていない. 本研究では単語の「測定」と「推定」を区別している. これはブログ等の記録媒体では使用単語の偏りもあるなど，全て理解している単語が出現するわけではないため，「測定」した単語の理解有無に基づき全単語に対する理解有無を「推定」することが肝要となるためである.

一方で，理解有無の推定に関する研究も行われている. 一般的な単語の推定[天野 1996]の研究が多いが[片山 2013]では専門用語の単語の知っているか否かの推定を行なっている. しかしながら，推定している単語が 1 つの分野の用語約 400 語に限定して評価を行なっているため，他の専門分野の単語や日常会話レベルの単語の推定も必要である.

また，[江原 2010]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. 個人がクリックした単語をライフログとして記録し，このライフログ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，英語母語話者が人手で 12 段階の難易度をつけた語彙リストを使用している. 難易度リストを用いた推定は人の主観が入ることや人的コストがかかる.

連絡先：野澤一真，
法政大学大学院理工学研究科電気電子工学専攻
〒184-0002 東京都小金井市梶野町 3-7-2
090-6477-3817, kazuma.nozawa.0703@gmail.com

3. 提案手法

本研究では関連研究を踏まえて，難易度リストを使用せず，個人の「理解している単語」と「理解していない単語」で学習させることで理解有無を推定する手法を提案する．word2vec で表現された単語ベクトルをニューラルネットワークに入力し，被験者にヒアリングで収集した理解有無を出力させるように学習させた単語理解度識別モデルを使用す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 1 に示す．

3.1 word2vec を用いた単語の低次元表現

機械学習でテキストを扱うためには，単語を単語そのものではなく，何らかの数値に置き換える必要がある．設計が容易な単語のベクトル表現に One-Hot 表現があるが，ベクトルの各要素に単語が対応するため，単語数と等しい膨大な次元数となるため適切ではない．そこで，word2vec[Mikolov 2013], [Rong 2016]と呼ばれる Mikolov 氏が提案した単語の分散表現を利用する．

word2vec とは単語の分散表現(以下，単語ベクトルと呼ぶ)を獲得する手法である．単語のベクトル表現が可能であるため，類似語はベクトル空間でグループ化され，意味特徴を捉えたものになる．

単語 *word* の単語ベクトル v は以下の式で求められる．

$$v = word2vec(word) \quad (1)$$

ただし，$word2vec(word)$は word2vec で表現される *word* の分散表現である．

また，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 の記事を使用し，各単語をベクトル化をする word2vec のモデルを生成した．

3.2 単語理解度識別モデルの生成

単語の理解度の推定には入力層，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(図 1)を使用した．

ニューラルネットワークは，word2vec で表現した単語ベクトル(x_1, x_2, ⋯, x_{200})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にすることによる．

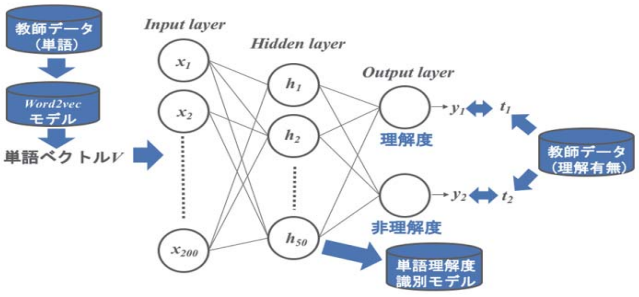


図 1 ニューラルネットワークによる理解有無の推定

4. 評価実験

4.1 データの収集方法

3 人の被験者(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，教師データを収集した．実用化の際は音声認識等を用いて，自動で収集することが望ましいが，音声をテキストに変換する際にノイズ等が混入することと，本研究の本質は「理解している単語」と「理解していない単語」が推定可能であるかを確かめることであることからヒアリングを行うこととした．

提示する単語リストは wikipedia 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形態素(言語で意味をもつ最小単位)に分割し，名詞のみ抽出した．そこから，wikipedia 内で頻度が 6 以上かつ文字数が 10 文字以下のみを抽出して，単語リストを生成した．

4.2 word2vec の学習条件

word2vec の学習には自然言語処理ライブラリ gensim [Radim Rehurek 2018]を使用した．学習させるモデルは CBOW を用い，word2vec の出力を 200 次元とした．ウィンドウサイズは 15，頻度が 5 未満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．

4.3 ニューラルネットワークの学習条件

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．word2vec で表した単語ベクトル 200 次元をニューラルネットワークへの入力とした．隠れ層のノードは 50 個とし，活性化関数は ReLU 関数とした．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 関数とした．

損失関数 E は推測値と目標値(教師データ)の二乗和誤差とした．損失関数 E を(2)式に示す．

$$E = \frac{1}{2N}\sum_{n=1}^{N}\sum_{k=1}^{2}(y_k^n - t_k^n)^2 \quad (2)$$

ただし，N は学習に使用する教師データの数とする．

ニューラルネットワークのパラメータ w は誤差逆伝播法で最適化し，勾配更新には SGD と Adam[Kingma 2014]を用い，それぞれの場合で精度を比較した．SGD を(3)式に示す．

$$w^{t+1} \leftarrow w^t - \eta\nabla E(w^t) \quad (3)$$

ただし，学習率 η は 0.01 とした．

次に Adam を(4)式から(8)式に示す．

$$m_{t+1} = \beta_1 m_t + (1-\beta_1)\nabla E(w^t) \quad (4)$$

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)\nabla E(w^t)^2 \quad (5)$$

$$\hat{m} = \frac{m_{t+1}}{1-\beta_1^t} \quad (6)$$

$$\hat{v} = \frac{v_{t+1}}{1-\beta_2^t} \quad (7)$$

$$w^{t+1} \leftarrow w^t - \alpha\frac{\hat{m}}{\sqrt{\hat{v}}+\epsilon} \quad (8)$$

ただし，$m_0 = 0$，$v_0 = 0$，$\alpha = 0.001$，$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 10^{-8}$とした．

また，重みの初期値は ReLU 関数と相性が良い「He の初期値」[He 2015]を用いた．

ニューラルネットワークの学習条件を表 2 に纏める．

表 1 被験者の詳細

被験者名	性別	年齢	大学での専門
A	男	21	教育学
B	男	22	電気電子工学
C	男	20	デザイン工学

表 2 ニューラルネットワークの学習条件

入力層	次元	200
	活性化関数	線形写像
隠れ層	次元	50
	活性化関数	ReLU関数
出力層	次元	2
	活性化関数	softmax関数
バッチサイズ		50
epoch		50
損失関数		二乗和誤差
重みの初期値		Heの初期値

4.4 評価方法

未知データに対する認識率を評価するために交差確認法を用いた. 交差確認法の手順を以下に示す.

1. 全単語をグループ X_1, X_2, X_3, X_4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同じにする.

2. X_iを除いた 3 グループで学習し, X_iを用いて正解率を算出する. このとき, X_iを検証データとし, それ以外を教師データとする.

3. 各試行の正解率 C_iを算出する. 正解率 C_iは[片山 2013]に倣って, (9)式を用いた. 被験者が理解していると答えた単語の集合を T, 被験者が理解していないと答えた単語の集合を F, モデルが理解していると推定した単語の集合を P, モデルが理解していないと推定した単語の集合を N とすると正解率 C_i は以下の式で表される.

$$C_i = \frac{|P \cap T| + |N \cap F|}{|P| + |N|} \qquad (9)$$

4. この手順を全ての $i(1<i<4)$について行い, 4 回の正解率の平均をモデルの正解率 C とする.

5. 結果

エポック数が 50 の時の各試行における正解率と全試行の平均を表 3 に示す. 表 3 の結果より, 最適化手法が Adam のときにはどの被験者においても正解率は 96[%]を超えた. テストデータは 200 個であったので, およそ 190 個は正解したこととなる. 一方で SGD の時の正解率はおよそ 70[%]に留まった. これらのことから重みを更新する際, Adam を適用した方が正解率が高いことが示された.

図 2 にエポック数に対する正解率をプロットしたグラフを示す. どの被験者においても学習を進めるにつれて正解率が増加していることがわかる. Adam の場合はエポック数が 5 回目までは急激に精度が向上し, それ以降は緩やかに増加している. 一方で, SGD の場合は初めの正解率が低く, 学習を重ねても Adam に比べて大きな正解率の向上は見られなかった. 一般的に SGD より Adam の方が素早く学習が可能であり, 精度が高いことは[Kingma 2014]で示されており, 今回の実験でも同様な結果となった.

また, 被験者ごとの正解率の差は Adam の場合においては最大 1[%]と, 小さかった. 理解している単語は個人ごとに違うにも関わらず, 同じ手法で理解有無の推定が可能であることが確認できた.

6. 結論

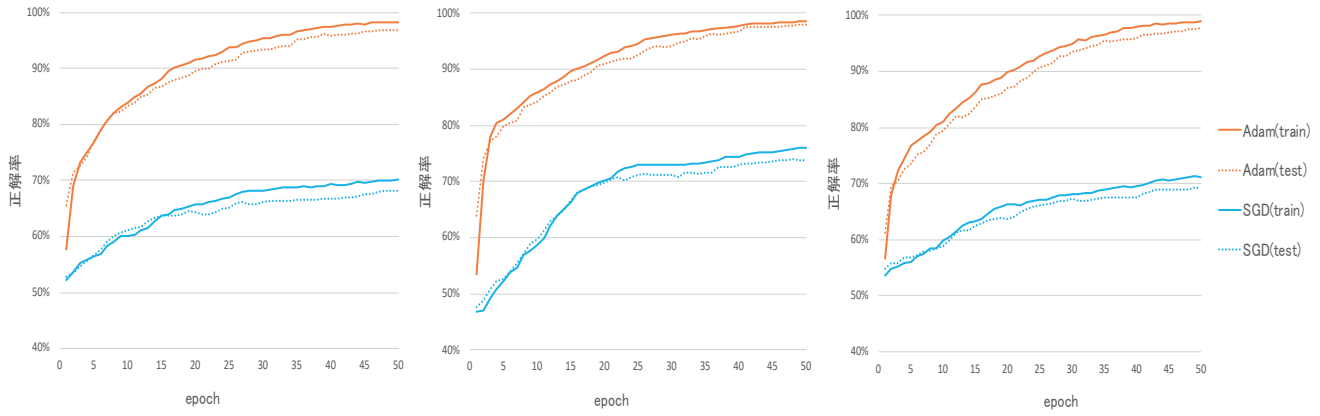
6.1 まとめ

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. 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 96[%]超える正解率で識別することに成功した.

成功した理由として,「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. 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

表 3 実験結果

被験者	A		B		C	
最適化手法	Adam	SGD	Adam	SGD	Adam	SGD
1回目	95.5%	65.5%	98.0%	70.5%	99.0%	66.0%
2回目	97.5%	72.5%	98.0%	76.0%	96.5%	73.5%
3回目	97.5%	65.5%	98.5%	74.5%	97.0%	66.5%
4回目	97.0%	69.5%	97.0%	74.5%	98.5%	70.5%
平均	96.9%	68.3%	97.9%	73.9%	97.8%	69.1%



(a) 被験者 A　(b) 被験者 B　(c) 被験者 C

図 2 エポック数に対する正解率の変化

という仮説の妥当性を示すものでもあり，ライフログを解析して個人の単語の理解有無の推定が可能であることが示唆された．

6.2 今後の展望

本研究は会話ログやチャットといったライフログから個人の語彙を推定することが最終目標である．そのため，ヒアリングを行うのではなく，SNS（Twitter，Facebook 等）やチャットツール(LINE 等)や会話ログといったライフログをソースとして，本手法を適用する必要がある．

また，word2vec は他言語での学習も可能であるため，英語等の言語の理解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

[総務省 2017] 総務省: 平成 29 年度版 情報通信白書 データ主導経済と社会変革.

[大井 2015] 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，65(4),pp.286-293, 2015-03-10.

[潘 2013] 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，28(6), pp.829-834, 2013-11-01.

[前田 2016] 前田直樹，平部裕子，荒川豊，安本慶一: オフィスワーカ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,情報処理学会論文誌，2016, pp.1788-1793, 2016-07-06 .

[片山 2013] 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，27，pp.1-4，2013.

[四方 2014] 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，pp.1126-1129，2014-3.

[天野 1996] 天野成昭: 頭の中に単語はいくつある?情報処理学会，00037(00004), pp.351-354，1996-04-15.

[江原 2010] 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，1，pp.1-7，2010-5-20 .

[Mikolov 2013] T. Mikolov, W. Yih, G. Zweig: Linguistic Regularities in Continuous Space Word Representations, NAACL-HLT-2013, pp.746–751, May 27, 2013.

[Rong 2016] Xin Rong: word2vec Parameter Learning Explained, arXiv:1411.2738v4 ,cs.CL,Jun. 2016.

[Radim Rehurek 2018] Radim Rehurek: genism topic modelling for humans https://radimrehurek.com/gensim/ .

[Kingma 2014] Diederik P. Kingma, Jimmy Ba: Adam: A Method for Stochastic Optimization, 22 Dec 2014.

[He 2015] Kaiming He，Xiangyu Zhang，Shaoqing Ren，Jian Sun: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, arXiv:1502.01852v1, cs.CV, 6 Feb 2015.