様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ４ 月 ２２ 日現在

機関番号：１２５０１
研究種目：若手研究
研究期間：2019～2020
課題番号：１９Ｋ１６９４１

研究課題名（和文）Word2Vecによる医学用語の分散表現は疾患間の数学的距離を定量的に表現するか

研究課題名（英文）Does the Word Embeddings of Medical Terms by Word2Vec Quantitatively Represent the Mathematical Distance between Diseases?

研究代表者
横川　大樹（Yokokawa, Daiki）

千葉大学・医学部附属病院・特任助教

研究者番号：８０７７９８６９

交付決定額（研究期間全体）：（直接経費）　　1,000,000円

研究成果の概要（和文）：本研究では2013年から２０１９年に千葉大学医学部附属病院 総合診療科に受診した患者さんの診療録(電子カルテのテキストデータ)より、深層学習の技術の一つであるWord2Vecと Doc2Vecを用いて、単語や文章の分散表現を得ることができました。Word2Vecではのべ10578020語を用いて、単語と単語の近さや関係などを深層学習しました。その結果、「痛み」と「疼痛」、「咳嗽」と「咳」、「花粉症」と「アレルギー性鼻炎」などが似ている単語として示されました。またDoc2Vecによる診療録の埋め込みベクトルと診断名をペアにして深層学習し、診断名の予測を試みましたが、精度は50%に留まりました。

研究成果の学術的意義や社会的意義
疾患や症状がベクトルで数学的に表現でき、日本語の医学用語として正しい結果と解釈できる場合、疾患と疾患の類似度が表現でき、疾患同士の距離（近さや遠さ）と解釈することができます。医師は臨床診断をするときに疾患同士の距離をイメージしますが、これまでは医師の経験に大きく頼らざるを得ない状況でした。疾患同士の距離が私達が研究で得たベクトルにより定量的に数字で表現できるれば、病名の想起し忘れなどがないよう助けるシステムが構築できる可能性があります。医師も人間である以上、悲劇的な誤診を避けられず、誤診の削減は我々の大きな目標であり、今後は個人の努力だけでなくシステムとしてサポートできる可能性が期待できます。

研究成果の概要（英文）：In this study, we used Word2Vec and Doc2Vec, two deep learning techniques, to obtain distributed representations of words and sentences from the medical records (text data of electronic medical records) of patients who visited the Department of General Medicine, Chiba University Hospital from 2013 to 2019. In Word2Vec, a total of 10578020 words were used for deep learning of word-to-word proximity and relationships. As a result, "consultation" and "referral," " cough" and "nasal discharge," and "hay fever" and "allergic rhinitis" were shown to be similar words. We also tried to predict diagnosis names by deep learning using Doc2Vec by pairing embedding vectors of medical records with diagnosis names, but the accuracy was only 50%.

研究分野：自然言語処理、診断推論

キーワード：診療録　自然言語処理　分散表現　埋め込みベクトル　疾患間距離　症状間距離　Word2Vec　Doc2Vec

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1)　臨床推断と支援システムの現状

臨床診断とは、患者の病歴や身体診察、検査により患者が患う疾患名を特定することである。病歴聴取は臨床診断にあたり最も重要で強力な技術であり、習得を支援し、あるいは代替するようなシステムの開発が望まれている。また病歴に基づいて駆動する自動診断システムや診断補助システムとしては、欧米を中心に試用され始めており約 20 のアプリケーションが報告されているが、日本語に対応したものはほとんど無く、内容についても科学的なピアレビューを受けたものは非常に少ない。これは日本語が持つ自然言語処理の難しさが一因であり、同時に診断推論エンジンが医師の経験にもとづく人工的なアルゴリズムに頼らざるを得ないためと考えられる。

(2)　疾患を診断する一般的な方略

診断過程で必要な情報である病歴は半構造化された質問内容に則り診療録（カルテ）に記載されていく。臨床診断の方略の一つには、以下のものがある。1. 基本情報より最も事前確率（疾患頻度や経験則による）が高い疾患を想起する。2. 病歴聴取で疾患に合致する点としない点を確認する。3. 確認した情報により確率を更新する（事後確率を変化させる）。1.で示した疾患想起の際に用いる方略として Pivot and cluster strategy がある。これは直感を元に一つの鑑別診断をまず想起し、同時にその近傍に位置する鑑別診断を想起する方法である。近傍に位置するということは仮に疾患と疾患の類似度に基づく距離空間が定義できた場合、距離が近い疾患同士と考えることができる。この類似する疾患群では、病態生理が共通していたり、診療録に記載される病歴の特徴が似ている可能性がある。しかしこの cluster は医師の経験以外に定量的に示されたデータはこれまでないため機械的にレコメンドすることも難しい。定量的に示すためには、まず疾患や症状自体を定量的に示す必要があり、単語のベクトル化がその入口にある。

(3)　Word2Vec と Doc2Vec の概要

近年自然言語処理における単語のベクトル化には次元圧縮された分散表現（埋め込みベクトル）を用いることが一般的である。単語が埋め込みベクトルで表現されると、単語に内在する意味を数学的に理解することや、類似度の計算（コサイン距離など）や加算減算（$word\ \ analogy$: $king - man + woman = queen$ など単語の意味を考慮した演算）が可能であり、コンピュータ・サイエンスとの親和性が高い。Word2Vec とは、ニューラルネットワークを利用した教師なし学習器で、単語の分散表現を計算する方法である。また Doc2Vec はそれを文章単位へ拡大した概念で、ある文章と文章がどれくらい似ているのかを類似度として表示できる。

(4)　医学用語の分散表現によるベクトル場の表現

このように Word2Vec や Doc2Vec を用いて診療録の情報を用いて医学用語の分散表現が得られると、ある疾患とある疾患がどれくらい似ているのかという「疾患間距離」や「症状間距離」が計算できる。疾患ベクトルや症状ベクトルによる表現は、疾患想起の際に役立つことが予想される。例えば急性心筋梗塞と狭心症は血管性病変を共通の特徴にもつ疾患であり、急性虫垂炎は腸管の感染症が主病態である。このような疾患と疾患の類似度が定量的に示すことができる様になる（図 1）。

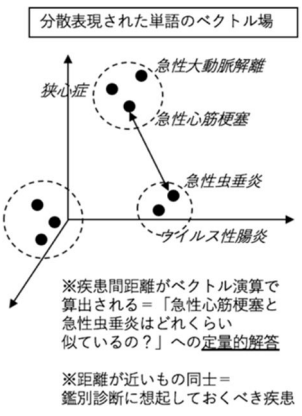

図 1 疾患間距離のイメージ図

２．研究の目的

本研究では、千葉大学医学部附属病院 総合診療科（以下当科） 初診外来を受診した匿名化された患者の電子カルテデータをコーパスとした、分散表現を獲得し、得られた埋め込みベクトルを利用した疾患間距離を表現することを第一の目的とする。さらに文章の分散表現による診断名文類が可能かどうかを検証する。

３．研究の方法

(1)　データの洗浄と前処理

研究概要図を図 2 に示す。総合診療科の医師が記載した電子カルテデータの 2 号用紙より SOAP 形式（Subjective：患者の主観的記述、Objective：医師による客観的記述、Assessment：医師の考察、Plan：今後の方針）の記述をテキストデータで抽出した。テキストデータに対して分かち書きおよび形態素解析を行い、標準形へ変換された単語列を作成した。分かち書きおよび形態素解析については MeCab を使用し、辞書には mecab-ipadic-NEologd および ComeJisyo を使用した。形態素解析の結果、品詞が名詞、形容詞、副詞、動詞である単語を抽出した。名詞のうち数は除外した。

(2)　Word2Vec および Doc2Vec の学習
得られた単語列を gensim のパッケージを用いて学習を行った。ハイパーパラメータは Word2Vec では Skip-gram を、Doc2Vec では PV-DBOW を用いた。他、ベクトル次元を 200 次元、window size は 5、最小カウントは 5 回、反復回数は 100 回とした。

(3)　分散表現の評価
先の前処理によって得られた SOAP 形式の単語列に対して、Word2Vec による学習を行った。Word2Vec により得られた分散表現を元に、一般用語、疾患を示す医学用語、症状を示す医学用語について、類似度（コサイン類似度を使用）が高い「類似語」を示し定性的な評価を行う。また単語ベクトルによる内的妥当性尺度の評価のため、全単語よりランダムに 3000 単語を取り出し階層的クラスタリングを行った際のコーフェン相関係数を算出し、100 回繰り返した後の平均値と標準偏差を求めた。またコーフェン相関係数が最大となる距離尺度および更新方法を求めた。
また外的妥当性尺度として、ICD-10（国際疾病分類第 10 版）との整合性を評価した。研究で用いた単語のうち、ICD-10 に含まれる単語を抽出し、ユニーク数をクラスタ数とした。そのクラスタ数になるように埋め込みベクトルを複数の距離および更新方法により階層的クラスタリングを行い、ICD-10 コードとの Adjusted Rand Index (ARI)、Normalized Mutual Information (NMI)、Adjusted Mutual Information (AMI)を計算した。ARI は−1 から 1 の値を、NMI と AMI は 0 から 1 の値をとり、値が高いほうが 2 つの集合の類似性が高いとされている。

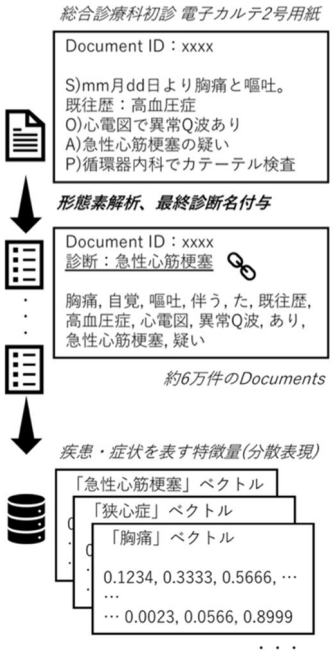

図 2 研究概要図

(4)　文章ベクトルと識別機による疾患群文類
最後に Doc2Vec での文章分散表現を用いた文章分類を行う。当科では診療録とは別に患者 ID 毎に最終診断名および ICPC（国際プライマリケア疾病分類）コードのデータを保持している。その最終診断データベースと本研究で獲得した診療録データベースとを左外部結合したデータベースを作成し文書分類を試す。結合したデータベースより、SOAP 記載がないレコード、診断が複数ついているレコードを削除し、SOAP 形式のテキストを抽出、前処理、単語列への変換を行った。この単語列に Doc2Vec を用いた学習を行った。
学習の結果得た文章ごとの分散表現に対して、最終診断名の ICPC コードにおけるアルファベットをラベルとして紐付け、分類器を用いて多クラス分類における精度評価を行った。分類器には標準的なマルチレイヤーパーセプトロン（ReLu による活性化層、Dropout(0.2)、ReLu による活性化層、Dropout(0.2)、ソフトマックスによる活性化層）を用いた。Loss には categorical_crossentropy を用いた。バッチサイズは 128、epoch は 3000 とした。
ICPC コードのアルファベットは A：全身、B：血液、D：消化器など異なる系を示している。ICPC コードラベルごとの適合率、再現率、F1 値、個数、それぞれのマクロ平均および精度を計算する。
本研究は、CPU Inter Core i9-9960X 3.10GHz、メインメモリ 64.0GB を搭載した Windows 10 Pro Insider Preview Build 21322 を用いて実行された。機械学習のフレームワークには、python (3.6.9)、scikit-learn (0.23.2)、gensim (3.8.3)を用いた。研究解析は研究代表者によって全て行われた。

４．研究成果
千葉大学医学部附属病院総合診療科の医師が記載した電子カルテの SOAP 形式の 2 号用紙を、2013 年から 2019 年にかけて 26565 件抽出した。抽出したデータは個人情報を削除した。電子カルテシステムの更新の関係で、2013 年以前の診療録は抽出できず、予定より数が少なくなった。SOAP 形式のうち、S の記載を含む診療録は 25066 件、O は 17397 件、A は 20670 件、P は 20877 件あった。SOAP すべてが記載されたものは 15267 件であった。のべ語数および異なり語数は、S で 6490937 語（一診療録あたり 259 語）および 74635 語、O で 1123776 語（同 65 語）および 25888 語、A で 2557568（同 124 語）および 41272 語、P で 509389 語（同 24 語）および 17795 語であった。総合診療科では他の医療機関から診断困難例の紹介を受けるため、一診療録あたりの記載料が多い傾向にある。単語出現回数を図 3 に示す。名詞では“症状”、“痛み”、“受

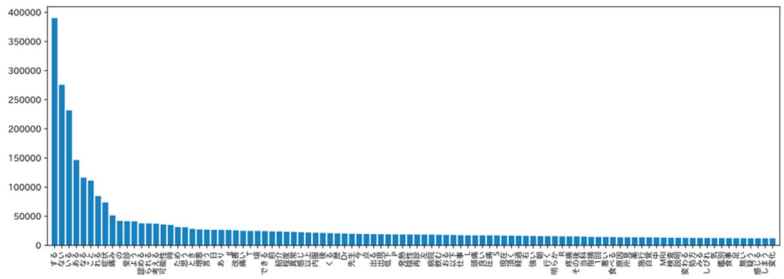

図 3 単語出現回数

診”、“可能性”という単語の出現回数が多かった。

SOAP 形式のすべてのテキストを用いた単語列を作成し、Word2Vec による学習を行った。最小カウントを 5 としたため、学習に用いたのべ語数および異なり語数は 10578020 語および 36517 語であった。得られた分散表現をもとに上記で得た単語の類似度を計算し表 1 に示す。また、疾患や症状を示す単語も同様に類似度の計算を行った。表 2 および表 3 にその一例を示す。

表 1 分散表現より得た類似語とコサイン類似度の例

症状		痛み		受診		可能性	
類似語	類似度	類似語	類似度	類似語	類似度	類似語	類似度
痛み	0.5492	痛む	0.7559	紹介	0.6028	低い	0.6610
腹部症状	0.4803	疼痛	0.7019	紹介受診	0.5963	否定	0.6600
違和感	0.4648	激痛	0.6808	近医	0.5938	鑑別	0.6521
疼痛	0.4577	痛い	0.6749	かかりつけ	0.5794	考える	0.6277
こと	0.4384	鈍痛	0.6572	外来受診	0.5547	否定的	0.6176

表 2 疾患名における類似語とコサイン類似度の例

咳嗽		悪寒		鼻汁		失語	
類似語	類似度	類似語	類似度	類似語	類似度	類似語	類似度
咳	0.7816	悪寒	0.8721	鼻閉	0.8539	失書	0.5628
鼻汁	0.6892	悪寒戦慄	0.7392	鼻水	0.7733	流暢性	0.5600
喀痰	0.6797	寒気	0.7322	後鼻漏	0.7400	失認	0.5207
痰	0.6587	発熱	0.7038	咳	0.7184	側頭葉	0.5080
咽頭痛	0.6599	寝汗	0.6981	咽頭痛	0.7127	失文法	0.5043

表 3 症状名における類似ごとコサイン類似度の例

肺炎		菌血症		花粉症		脳梗塞	
類似語	類似度	類似語	類似度	類似語	類似度	類似語	類似度
誤嚥性肺炎	0.5769	感染性心内膜炎	0.6185	アレルギー性鼻炎	0.7179	心筋梗塞	0.6784
気管支炎	0.5681	IE	0.6164	スギ	0.7013	脳出血	0.6449
肺結核	0.5498	抗菌薬	0.5891	鼻炎	0.6846	梗塞	0.5864
器質化	0.5341	血液培養	0.5754	花粉	0.6409	クモ膜下出血	0.5611
細菌性肺炎	0.5331	aglactiae	0.5715	ハウスダスト	0.6341	脳卒中	0.5560

この結果からは一般用語だけではなく、疾患名や症状名などの場合でも、類似度が高いものには意味が似ている単語が並んでおり、本分散表現が、医学用語の意味を学習している可能性が示唆された。一方で“可能性”と“低い”や“咳嗽”と“鼻汁”のように必ずしも類義語ではないにもかかわらず類似度が高い背景には、Word2Vec が持つ共起性の影響があると考えられる。すなわち本文中で近くに登場する単語は、似たような意味を持つという特性である。この特性のため逆の意味を示す単語などの類似度が高く計算されており、これはこの方法における限界である。また Word2Vec は計算の過程でランダム性を内包しており、学習のたびに結果が異なることも一つの限界である。

次に、Word2Vec の分散表現における階層的クラスタリングの内的妥当性についての検討結果を示す。ランダムに取り出した 3000 語のコーフェン相関係数についての平均値は、0.8004 (±0.0042) であり、最大値を示した距離は 88%でユークリッド距離、分類手法は 100%で群平均法であった。コーフェン相関係数は 0 から 1 の値を取り、値が大きいと妥当性が高いことを示している。クラスタリングの評価では客観性の担保が難しく、絶対的な評価基準は存在しないものの、コーフェン相関係数が 0.8004 であるということは一定の妥当性を示すものと考えられた。

また外的妥当性の評価として、本分散表現によるクラスタリングと ICD-10 が持つ階層構造との一致を検討した。Word2Vec の学習に用いた単語のうち、ICD-10 の病名に合致する単語は 2307 語あった。ARI、NMI、AMI を、距離および更新方法の組み合わせ別に図 4 に示す。
ICD−10 には階層構造があり、68000 以上のコードを持つ。しかしすべての疾患が同じレベルの粒度を持っているわけではなく、オントロジー的に離れている ICD コードは、一緒にグループ化される可能性も低いことが知られている。本研究では診療録を元にした埋め込みベクトルを ICD-10 に紐付けてはいるが、AMI や ARI は非常に低く、この埋め込みベクトルをそのまま ICD-10 の連続空間への写像と見ることはできないだろう。また内的妥当性尺度を最大にする距離および更新方法であるユークリッド距離および群平均方法は、必ずしも外的妥当性尺度を最大にするものではなかった。この結果となったことは、ICD-10 がこの分散表現から得た疾患間距離

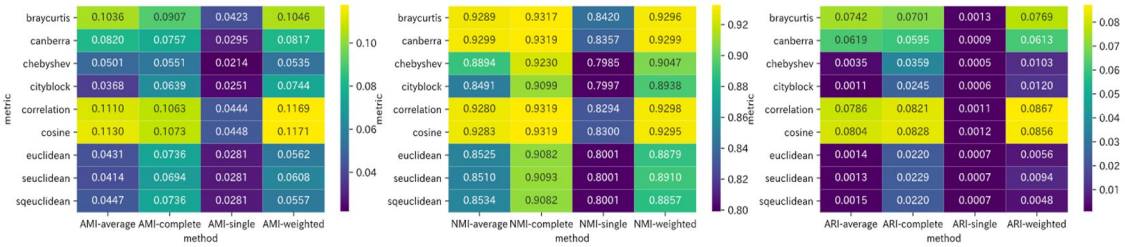

図 4 ICD-10 との外的妥当性評

の表現とは異なる表現を持つということを示している。

最後に Doc2Vec での書類分散表現の獲得と、文章分類の結果を示す。最終診断データベースと診療録とを左外部結合したデータベースより、SOAP 記載がないレコード、診断が複数ついているレコードを削除したところ、7229 個のレコードが残った。この 7229 個のレコードに対して Doc2Vec を行った。図 5 にエポックごとの学習ロスを示す。
500epoch あたりで精度および学習ロスはほぼ横ばいとなった。

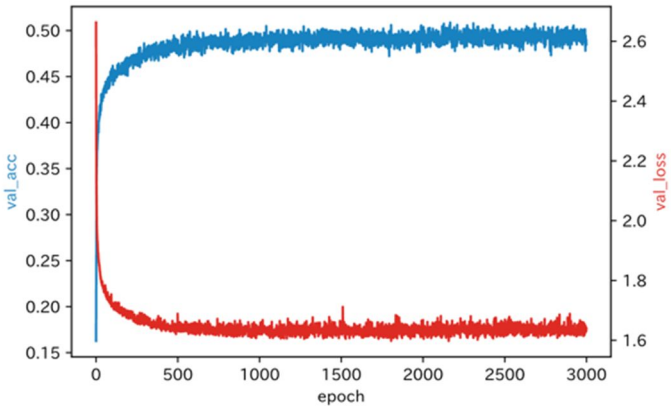

図 5 学習過程におけるロスと精度の推移

表 4 にテストデータでの ICPC コードラベルごとの適合率、再現率、F1 値、個数示す。精度は 0.50、適合率、再現率、F1 値のマクロ平均はそれぞれ 0.37、0.26、0.27 であった。

表 4 テストデータでの ICPC コードラベルごとの評価指標

	適合率	再現率	F1 値	個数
A	0.47	0.40	0.43	239
B	0.27	0.19	0.22	81
D	0.44	0.60	0.51	184
F	1.00	0.14	0.25	7
H	0.62	0.36	0.15	14
K	0.43	0.06	0.10	107
L	0.48	0.75	0.59	322
N	0.49	0.48	0.48	250
P	0.60	0.70	0.65	323
R	0.47	0.41	0.44	87
S	0.38	0.16	0.23	62
T	0.65	0.19	0.29	69
U	0.00	0.00	0.00	21
W	0.00	0.00	0.00	1
X	0.00	0.00	0.00	5
Y	0.00	0.00	0.00	3
Z	0.00	0.00	0.00	3

データベースに欠損値が含まれることや、先に述べたように想定よりも診療録数が少なかったため、学習の精度向上が望めなかった可能性が高い。また当科で診断がつく疾患の分布が一様ではなく、疾患によっては特に低い適合率や再現率となっている。診療録と診療録とが類似しているということは診療録に含まれる単語の種類や出現分布、その前後関係が似ていることを示しており、Doc2Vec はそれを学習している。当科の診療録は時代によっては「テンプレート」用いた記載を行っており、この場合診療録は自然言語のうち半構造化文章に近くなる。すると単語の順序に関する情報価値は薄れ、文類精度に影響した可能性がある。
本研究は 17 の多クラスへの分類であったが P：心理・精神についての適合率・再現率が最も高かった。心理・精神の患者の診療録では、社会生活に関する記述や本人の病の体験・解釈、日常への影響などの単語が、他のクラスより多く出現する可能性がある。また希死念慮や抑うつ気分など、他のクラスでは出現しないような特異的な単語も多いと予想される。また診療録数が多かったことも疾患精度が高いことに影響していたと考えられる。

本研究結果からは診療録を元にした分散表現は、単語の類似度をある程度学習しており、疾患ベクトル・症状ベクトルとして定量的解析に利用できる可能性があることがわかった。一方で外的妥当性尺度として ICD-10 を用いたが、本分散表現の構造は、ICD-10 の階層構造と一致しているとは言い難く、ICD-10 に置き換わるものではないことが言える。また文章ベクトルの学習については本データベースの特殊性と診療録数の制限のため、学習に用いる文章量としての拡充は望めないが、精度の向上はハイパーパラメータや前処理の調整により部分的に改善される余地はあると考えられた。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関