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　５ 年　 ５ 月 １６ 日現在

機関番号：１２５０１
研究種目：若手研究
研究期間：2021～2022
課題番号：２１Ｋ１７８４８

研究課題名（和文）学術誌抄録から学習した疾患名の分散表現は疾患同士の距離を表現しうるか

研究課題名（英文）Can Distributed Representations of Disease Learned from Academic Journal Abstracts Represent the 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）　　3,600,000円

研究成果の概要（和文）：疾患を診断する方法(診断推論)の習得を支援したり代替するシステムのうち、実際の医療の現場で使用できるものはない。作成には疾患同士の類似度(疾患間距離)の計算が有用と考えた。医学中央雑誌の1842156件の抄録をWord2Vecにより学習し、疾患を表す単語の分散表現(疾患ベクトル)を獲得し、距離の算出に成功した。
ICD-10(国際疾病分類第10版)と一致する疾患ベクトルは8031個（ICDコードは3915種類）だった。コーフェン相関係数の最大値は0.7748だった。ICDコードとの調整ランド指数、正規化相互情報量、調整相互情報量の最大値は0.3208、0.8783、0.4242だった。

研究成果の学術的意義や社会的意義
疾患間距離を医学文書から計算することで、主観的な経験則に基づく疾患想起から、客観的な距離に基づく疾患想起への転換が可能になる。医師の経験は収斂され分散表現へと置き換わり、普遍的で汎用性が高い距離データを得られる。これは診断推論の「疾患想起のプロセス」をAIに置き換え、自動診断システムや診断支援システムの開発への第一歩となる。これらのシステムは、自宅で病院へ行くべきか判断する際のサポートツールとなり、医療費削減や医療の質の均質化に貢献し、悲劇的な誤診や見逃しを削減する可能性がある。日本語に基づいたAI診断支援システムを作成するためには、日本独自の研究が必要で、本研究はその基盤となる。

研究成果の概要（英文）：Presently, no systems can be applied in actual medical settings to aid or replace the acquisition of diagnostic reasoning methods. I hypothesized that calculating disease similarity (inter-disease distance) could be instrumental in creating such systems. By training on 1,842,156 abstracts from the Central Medical Journal with Word2Vec, I procured distributed representations (disease vectors) of words denoting diseases and successfully calculated their distance.
I found 8,031 disease vectors corresponding to the International Classification of Diseases, 10th Edition (ICD-10), which match 3,915 ICD codes. Cohen's correlation coefficient reached a maximum value of 0.7748. Adjusted Rand index, standardization, normalized mutual information, and adjusted mutual information for ICD codes achieved maximum values of 0.3208, 0.8783, and 0.4242, respectively.

研究分野：自然言語処理

キーワード：診断推論　自然言語処理　Word2Vec　分散表現　埋め込みベクトル　疾患間距離　抄録　医中誌

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

(1)　診断推論と支援システムの現状

診断推論とは患者の病歴や身体診察、検査により患者が患っている疾患名を診断することである。診断推論は複雑で、習得には多くの時間を要するため、習得を支援する、あるいは代替するシステムの開発が望まれる。診断推論に最も寄与する「病歴」に基づいて駆動する自動診断システムや診断支援システムは、欧米を中心に試用され始めており約 20 種類が報告されているが、科学的なピアレビューを受けたものは少なく、実臨床に耐えうるものはない。また日本語に対応したものはほとんど無い。これは日本語が持つ自然言語処理の難しさに由来する部分と、エキスパートを模倣したアルゴリズムに多くを頼っているためと考えられる。

(2)　診断推論の一般的な方略と疾患間距離の役割

診断推論過程で必要な病歴には、性別・年齢・主訴という基本情報と、その疾患の病態的特徴を示す情報がある。一般的な診断推論は、基本情報をもとに最も事前確率（疾患頻度を元にした経験則）が高い疾患を想起することで始まる。次に病歴聴取を行い病態的特徴との合致点／非合致点を確認し、事後確率を変化させていく。このとき同時に、想起している疾患の対立仮説として鑑別診断を複数想起し、それらの事後確率も同時に変化させ比較する。

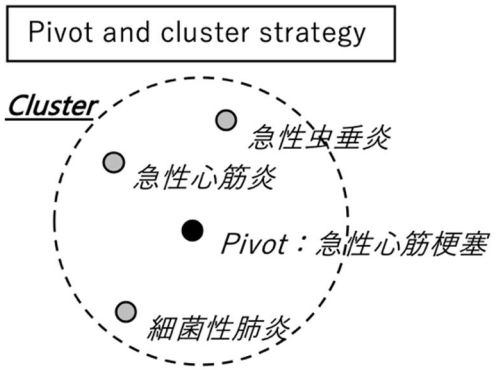

図 1 Pivot and cluster の模式図。
中心にある急性心筋梗塞（Pivot）のまわりに、同時に想起された疾患群（Cluster）がある。例示した疾患間距離は仮想のものである。

「疾患想起」を行う際、直感を元に想起した一つの診断（Pivot）と、その近傍に位置する「似ている疾患群（Cluster）」を同時に想起する手法を Pivot and cluster strategy という（図 1）。意識的にこの手法を行うことは教育に有用であり、誤診を減らす可能性がある。しかし、Cluster の想起は医師の主観・経験に頼る部分が大きく、初学者では実行が難しい。さらに Cluster において、疾患と疾患がどれくらい似ているのか（異なるのか）、すなわち「疾患間距離」が定量的に示されたものはこれまでなく、機械的な計算や提案も難しかった。

似た病態生理に基づく疾患群は、疾患同士が持つ病歴の特徴も似ていることも事実である。例えば「梗塞性疾患」は血管が閉塞することで生じる疾患の一群であり、発症様式が突然発症となる共通点がある。他にも「感染症疾患」「変性疾患」「内分泌疾患」など病態生理学的に発症機序が類似した異なる疾患群はいくつか存在し、病歴に共通点が存在する。また異なる病態をもつ疾患群であっても傷害される臓器が同一であれば、同様の病歴を呈することがある。これらが疾患間距離の表現にどのように影響しているかはわかっていない。

(3)　疾患間距離獲得の試みによって明らかになったこと

研究代表者は病歴情報から疾患間距離を計算し明らかにするために、数年前より Word2Vec を用いた単語の分散表現（単語埋め込み）の獲得を試みている。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、特定の次元に圧縮されたベクトルで単語を表現する事ができる（単語埋め込みベクトル）。ベクトルの演算を通して単語に内在する意味を数学的に理解することや、類似度の計算が可能になる。これまでの研究では電子カルテの総合診療科の診療録記載をコーパスに Word2Vec によって分散表現を得ることができていた。しかし、診療録の記載を利用したため、含まれる情報の偏りや表記ゆれの影響が大きく、一様な疾患間距離空間を抽出するのは難しいと考えた。また記載内容が文章として安定している学術コーパス（医学中央雑誌刊行会 医中誌 Web）を用いた研究において、「梗塞」を含む疾患に関する疾患間距離の算出に成功した。疾患間距離はクラスタリングにおいて解剖学的差異や時間的差異を表現できる特徴を内包している可能性があった。しかしながら「梗塞性疾患」という単一ドメインに限

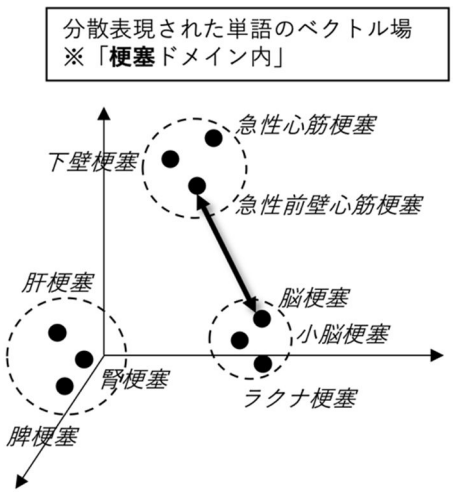

図 2 「梗塞」ドメイン内の疾患間距離の模式図
臓器や病態毎に cluster を形成していた。

定した結果であった（図 2）。
学習に使用するコーパスの質や量によって得られる分散表現が異なるため、汎用的な診断推論の支援システムを構築するためには、より大きく全体を包括するコーパスにおいて学習させ、疾患数（単語の種類）を増やす工夫が必要と考えた（図 3）。しかしその時に、これまでの研究と同様に疾患間距離が得られるかどうかはまだ不明であった。

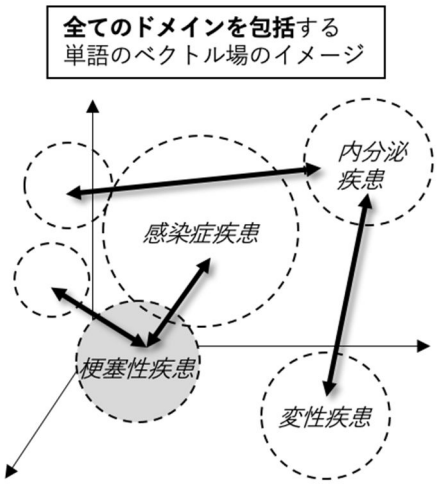

図 3 すべてのドメインの疾患間距離の模式図
他ドメインの疾患同士の疾患間距離を示すことができる。

２．研究の目的

本研究では学術的な医学文書である医中誌データに含まれる全抄録に対して Word2Vec を用いることで、疾患を表す単語の分散表現（疾患ベクトル）を得ることを第一の目的とした。

さらに、得られた疾患ベクトル同士の数学的な距離である「疾患間距離」を計算する。疾患間距離による階層的クラスタリングの実行とその評価を行うことを第二の目的とした。

３．研究の方法

(1) データの前処理、学習、分散表現と疾患ベクトルの獲得

図 4 に研究概念図を示す。研究に用いるデータは、医学中央雑誌刊行会（医中誌）が保持している医中誌データ全年分（2020 年分は 4/16 更新の 8 号データまで、約 180 万件）の抄録テキストデータである。この抄録テキストデータを前処理し単語列へ変換し、Word2Vec を用いて単語の分散表現を学習させた。得られた分散表現より、特に疾患名についてのベクトル値を得て、疾患ベクトルとした。

(2) 疾患間距離の計算および、階層的クラスタリングの実行と評価

得られた疾患ベクトルと、異なる疾患ベクトルとの「コサイン距離」等を計算することで、疾患間距離を得た。このような距離をもとに、疾患ベクトルの集合に対して階層的クラスタリングを行った。手法と距離定義の組み合わせごとに、内的妥当性尺度コーフェン相関係数を、外的妥当性尺度として ICD-10（国際疾病分類 第 10 版）コードとの Adjusted Mutual Information（調整相互情報量）、Normalized Mutual Information（正規化相互情報量）、Adjusted Rand Index（調整ランド指数）をそれぞれ計算し評価した。

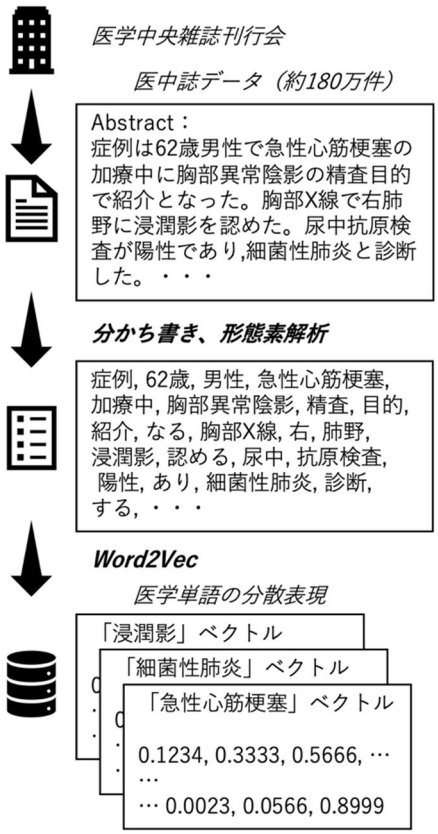

図 4 研究概念図
図内の抄録は架空のものである。疾患名ごとに 200 次元のベクトルとする。

なお解析には AMD EPYC 7402P (24core/48thread 2.80GHz)および 128GB のメモリが搭載された端末を使用した。
前処理と学習には Python(3.6.9)および scikit-learn(0.24.2)、gensim Word2Vec(3.8.3)を用いた。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。

４．研究成果

(1) データの基礎値

1,842,156 件の抄録を対象とした。一つの抄録に含まれる平均単語数は 323.55 字（標準偏差 190.25 字）であった。前処理としての形態素解析と分かち書きでは、名詞・動詞・形容詞・副詞を抽出し、数は除外した。一つの抄録を変換した一単語列に含まれる平均単語数は 92.55 字（標準偏差 51.28 字）であった。
図 5 に頻度順のヒストグラムを示す。する、れる、られるといった一般コーパスでも多用される助動詞の頻度が多かったが、例、群、患者、症例、検討といった、医学の学術コーパスで多用される単語を多く認めた。この点は、梗塞ドメインに絞った過去の研究と同様の所見である（未発表）。本研究で使用したコーパスに含まれる総のべ語数は 170,490,448 語、総異なり語数は

522,430 語であった。この内、頻度が 5 より大きい単語を学習に使用した。学習に使用したのべ語数は 169,875,236 語、学習に使用した異なり語数は 189,785 語あった。

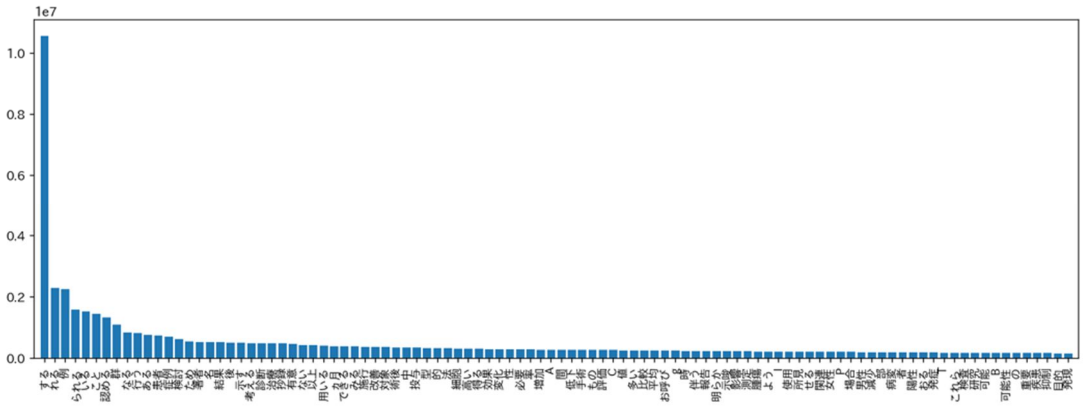

図 5 単語ヒストグラムの一部
例、群、患者、症例、検討など、医学学術コーパスに特徴的な単語が高頻度である。

(2) モデルの学習
gensim Word2Vec による学習を行った。ハイパーパラメータは size=200, min_count=5, window=5, iter=100, sg=1 とした。本研究では Skip-gram を用いた。学習時間は 311 分 55 秒だった。モデルのファイル容量は 14,055,811 バイトであった。

(3) 内的妥当性の検証
内的妥当性の検証はコーフェン相関係数を用いて行った。階層的クラスタリングを実行するときの更新方法（method; average, centroid, complete, median, single, ward, weighted）と距離定義（metric; braycurtis, canberra, chebyshev, cityblock, correlatioin, cosine, euclidean, sqeuclidean）のすべての組み合わせで計算し、最大値および最大値を取る方法と距離定義を調べた。
まず得られたすべての単語ベクトルを用いて、コーフェン相関係数を計算したところ、method=centroid、metric=euclidean のときコーフェン相関係数が最大値 0.8123 をとった（図 6）。

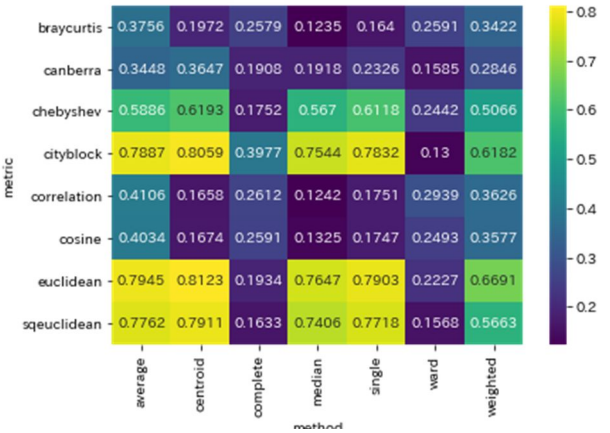

図 6 全単語でのコーフェン相関係数

次に、ICD-10 の病名と完全一致した単語に限定し検証した。この単語ベクトルを疾患ベクトルと呼ぶ。ICD-10 と一致した疾患ベクトル 8031 個あった。その時、同じ ICD-10 コードをもつ疾患が存在するため、ユニークな ICD-10 コードは 3915 種類あった。コーフェン相関係数は、method=centrid、metric=euclidean の 組み合わせのとき最大値 0.7748 をとった（図 7）。

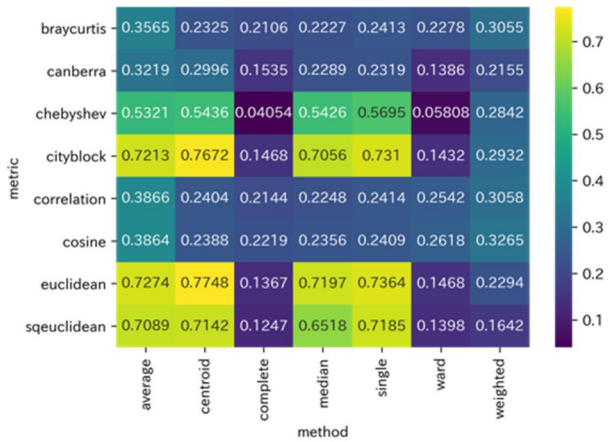

図 7 疾患ベクトルでのコーフェン相関係数
ICD-10 と完全一致した病名の単語ベクトルのみを用いて計算した。

(4) 外的妥当性の検証
ICD-10 は解剖学的に、あるいは病態毎に区分された階層構造をしている。例えば梗塞性疾患はアルファベットの I で示され、以下に続く数字 3 桁が、脳梗塞や心筋梗塞など特定の疾患を指す。ICD-10 の階層構造と、本研究で得た疾患ベクトルの階層的クラスタリングの結果が似ている場合、疾患ベクトル内に、解剖学的な関係や病態の情報が含有されている可能性がある。Adjusted Mutual Information、Normalized Mutual Information、Adjusted Rand Index の最大値と最大値を取るときのパラメータ（metric, method）はそれぞれ、0.4242（centroid, correlation/cosine）、0.8783（complete/ward, braycurtis）、0.3208（centroid, braycurtis）だった（図 8）。更新方

法と距離定義に一定の傾向はなかったが、内的妥当性尺度の結果とは合致しなかった。この傾向はこれまでの研究でも同様であり、梗塞ドメインに限定した計算結果よりも高い結果だった。このことから、梗塞ドメイン内の疾患ベクトルは他のドメイン内の疾患ベクトルに比較して ICD-10 との外的妥当性が低いことが示唆された。

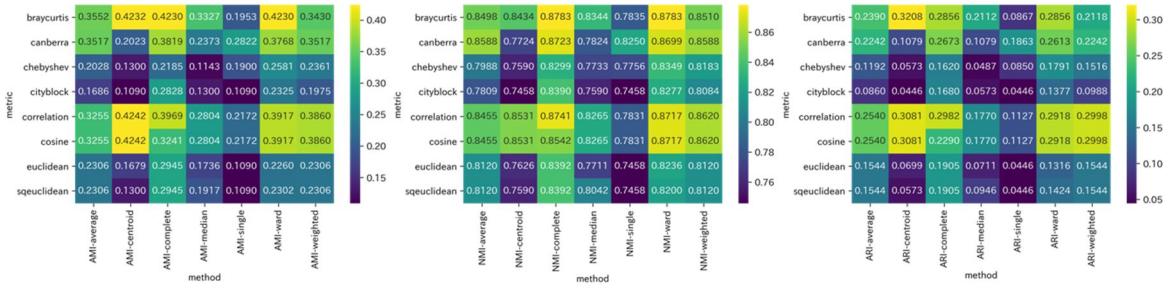

図 8 疾患ベクトルと ICD-10 との外的妥当性指標
Adjusted Mutual Information (AMI 左図), Normalized Mutual Information (NMI 中図), Adjusted Rand Index (ARI 右図)の結果を示す。

＜参考文献＞

Ikusaka M. Disease hypothesis generation: clinical hypothesis generation. Off J Japan Prim Care Assoc. 2011;34:77-9. https://doi.org/10.14442/generalist.34.77 (in Japanese).
Shimizu T, Tokuda Y. Pivot and cluster strategy: a preventive measure against diagnostic errors. Int J Gen Med. 2012;5:917-21. https://doi.org/10.2147/IJGM.S38805.
Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013. https://doi.org/10.48550/arxiv.1301.3781.
Mecab. https://taku910.github.io/mecab/.
mecab-ipadic-NEologd. https://github.com/neologd/mecab-ipadicneologd.
ComeJisyo. https://ja.osdn.net/projects/comedic/.
gensim-Word2Vec. https://radimrehurek.com/gensim/models/word2vec.html.
Yokokawa D, Noda K, Yanagita Y, et al. Validating the representation of distance between infarct diseases using word embedding. BMC Medical Informatics and Decision Making. 2022;22(1):322. https://doi.org/10.1186/s12911-022-02061-8

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