川崎医療福祉学会誌　Vol. 28　No. 1　2018　167 − 178

原　著

単語の分散表現を用いた文書分類

田　中　昌　昭*1

要　　約

文書分類は自然言語処理の代表的な研究課題のひとつで，トピック分類，評判分析，フィルタリングなどに応用されている．文書分類では従来，文書の特徴量として単語の出現頻度が用いられてきた．しかしながら，単語そのものが持つ情報からは単語間の類似度や関連性を計算することは難しい．そこで，特徴量として単語の分散表現を用い，分類性能の向上を目指した．まず，医学論文情報データベースである医中誌 Web から重複を除いた7,881件の抄録を抽出して学習コーパスとした．次に，skip-gram モデルを使って単語のベクトル表現（分散表現）を獲得した．得られた単語ベクトルの重心および合成ベクトルを特徴量に用いて，抄録を5つの疾患に分類する実験を行った．評価のため，単語の出現頻度を用いる従来の方法で分類した結果と比較した．その結果，本手法による分類の正確度は0.770となり，従来の方法（0.807）を上回ることはできなかったが，それに匹敵する分類性能を得ることができた．本手法による分類性能が従来の手法よりも低かった原因として，単語ベクトルの品質，単語の多義性，特徴選択の問題などが考えられた．なかでも獲得した情報の大部分を利用しないで捨ててしまう特徴選択には改善の余地が残された．

1．緒言

近年，機械学習の一手法であるディープラーニング（深層学習）の発展は，従来困難とされてきた多くの問題に著しい進展をもたらしている[1)]．医学領域においても例外ではなく，バイオインフォマティクス，医用画像処理，パーベイシブ・センシング，医療情報学，そして公衆衛生学への応用が進められている[2)]．中でも特筆すべきは医用画像処理分野で，ディープラーニング技術を用いたコンピュータ支援画像診断システムも実際の臨床現場に登場している[3)]．また，大規模な EHR（Electronic Health Record）のデータベースから自動的に患者の特徴を推測し，糖尿病，統合失調症，がんなど特定の疾患の発症確率を予測する研究も報告されている[4)]．

今日のディープラーニングの隆盛は，ネット上の大量の画像から猫の特徴を自発的に学習した“Google の猫”として知られる研究[5)]に端を発している．これを契機に，音声認識，画像認識，物体検出の分野で飛躍的に性能が改善した．音声認識では，A/D 変換された音声信号が処理対象となり，画像認識では，色の3原色を表す3組の整数値が処理対象となる．これらは，主観の入る余地のない物理的な情報であるため，画像間の，あるいは音声間の類似度を計算することはたやすい．これが，初期のディープラーニングにおいて音声認識や画像認識の領域で集中的に研究が進んだ理由であろう．一方，文字や単語といった記号を扱う自然言語処理の分野では，記号間の類似度や関連性を記号それ自体が持つ属性から直接計算することができないため，音声認識や画像認識のように簡単にはいかない[6)]．

たとえば文書分類の問題を考えてみよう．カルテに「両側の唾液腺に腫脹が見られ，流行性耳下腺炎の疑いを認める」という記述があったとする．この記述を，「両方の耳下腺に腫れが見られ，おたふくかぜではないかと疑う」と同じカテゴリに分類できるだろうか．それを可能にするには，「流行性耳下腺炎」が「おたふくかぜ」であると知っていなければならない．その解決策として，シソーラスの利用が考えられるが，シソーラスの開発には多大な労力がかかるだけでなく，人間が作る以上，すべての用語をカバーすることは困難で，たとえできたとしても，それをどうやって利用するかという課題が残る．

*1 川崎医療福祉大学　医療福祉マネジメント学部　医療情報学科
（連絡先）田中昌昭　〒701-0193　倉敷市松島288　川崎医療福祉大学
E-mail : mtanaka@mw.kawasaki-m.ac.jp

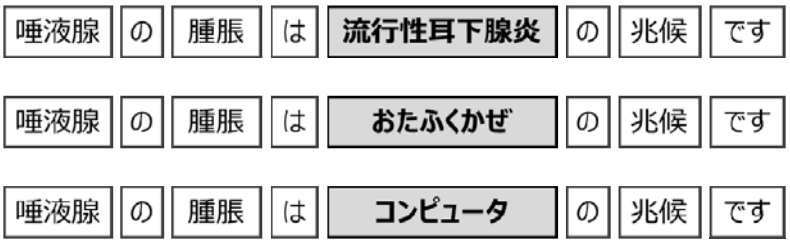

図1 分布仮説

これに対して，大量のデータから単語の意味を自動的に学習して，単語をベクトルで表現する手法が考案された[7]．単語をベクトルで表現できれば，コサイン類似度などを使って単語間の類似度や関係性を計算できる．このベクトル表現は，単語の分散表現と呼ばれ，ディープラーニングを自然言語処理に適用したニューラル言語モデルの副産物として誕生した[6]．

こうした背景のもとに，本稿では単語の分散表現を用いて文書の分類実験を行い，単語の頻度を用いた従来の方法と比較した．そして，単に文書分類にとどまらず，分散表現が医療にどのように応用できるかについても考察した．

2．理論

2．1 単語の分散表現

単語の分散表現のアイデアは，1950年代に提案された分布仮説[8]に遡る．これは，「単語の意味はその単語の出現した際の周囲の単語（文脈）によって決まる」という仮説である．たとえば，図1に示すように上段と中段はいずれも違和感のない自然な文章なので，「流行性耳下腺炎」や「おたふくかぜ」は「唾液腺」，「腫脹」など周囲の単語（文脈）と高い相関で現れることが予想される．そのため，分布仮説の帰結として「流行性耳下腺炎」と「おたふくかぜ」は類似した意味を持つものと期待される．一方，図1下段の文章はいかにも不自然であり，「コンピュータ」が「唾液腺」や「腫脹」といった単語と高い相関で現れることはないはずである．そのため，「コンピュータ」は「流行性耳下腺炎」や「おたふくかぜ」と意味的に類似していないことが推測される．

この分布仮説に基づいて大量の文書データ（コーパス）から単語のベクトル表現を学習するWord2Vecと呼ばれる手法が提案された[9, 10]．次節ではWord2Vecがどのようにして単語ベクトルを獲得するかについて説明する[6, 11, 12]．

2．2 Word2Vec

まず，学習コーパス（文書例）を単語列 $w_1, w_2, \cdots, w_T$ とする．ここで，Tは学習コーパスに含まれる単語数である．次に，ある位置 t で出現する単語 w_t に対して，その前後 δ 個の単語列を文脈 $C_{w_t} = (w_{t-\delta}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+\delta})$ とする．文脈 C_{w_t} から単語 w_t を予測する条件付確率分布関数 $p_\theta(w_t|C_{w_t})$ を定義し，式(1)に示す対数尤度関数 L を最大化するように $p_\theta(w_t|C_{w_t})$ を学習する[11]．

$$\mathcal{L} = \sum_{t=1}^{T} \log p_\theta(w_t|C_{w_t}) \quad \cdots (1)$$

これがWord2Vecの基本的な考え方である．なお，ここでは，確率分布 $p_\theta(w|C_w)$ の母数（パラメタ）θ を最尤推定することになるが，このパラメタ θ が単語の分散表現になる．

2．3 CBoWモデルとskip-gramモデル

Word2Vecは，CBoWとskip-gramという2つのモデルの総称として用いられている用語である[6]．本節では，この2つのモデルについて説明する．

skip-gramでは，文脈 C_{w_t} に含まれる語は互いに独立であると仮定して，$\log p_\theta(w_t|C_{w_t})$ を，文脈単語 c から対象とする単語 w_t を予測する条件付き確率分布関数 $p_\theta(w_t|c)$ の積に分解する．よって，式(1)は次式に変形できる．

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{c \in C_{w_t}} \log p_\theta(w_t|c) \quad \cdots (2)$$

そして，条件付き確率分布関数を次のように対数双線形モデルを用いて定式化する．

$$p_\theta(w_t|c) = \frac{\exp(\boldsymbol{v}_c \cdot \tilde{\boldsymbol{v}}_{w_t})}{\sum_{w' \in V} \exp(\boldsymbol{v}_c \cdot \tilde{\boldsymbol{v}}_{w'})} \quad \cdots (3)$$

ここで，v_c は文脈内にある単語 c のベクトル表現，$\tilde{v}_w$ は予測単語 w のベクトル表現，そして V は

コーパス全体の語彙集合である．

式(3)の右辺の分母には，全語彙集合 V に対する内積と指数関数の計算が現れるため，大規模なコーパスに対しては膨大な計算量となる．そこで，$|V|$ に比例しない計算量でこの計算を近似する様々な手法が考案されている[6]．なかでも skip-gram は，負例サンプリングという手法を用いて大幅に計算量を削減している．

負例サンプリングでは，学習データに現れる単語・文脈ペア $\langle w_t, c\rangle$ ごとにランダムに K 個の擬似負例単語 $\langle \tilde{w}_t, c\rangle$ を生成し，それらを識別するように学習する．具体的には，正例 $\langle w_t, c\rangle$ に対しては1，負例 $\langle \tilde{w}_t, c\rangle$ に対しては0を予測するロジスティック回帰モデルで近似する．

これに対して CBoW モデルでは，式(2)のように文脈 C_{w_t} に対する条件付確率分布関数を文脈語 c に対する条件付確率分布関数の積に分解せず，文脈内にある単語 c のベクトル表現 ν_c の和 $\nu_{C_t} = \Sigma_{c \in C_t} \nu_c$ を用いて次式のように条件付き確率分布関数 $p_\theta(w_t | C_{w_t})$ を定式化する[11]．

$$p_\theta\left(w_t | C_{w_t}\right) = \frac{\exp\left(\boldsymbol{\nu}_{C_t} \cdot \tilde{\boldsymbol{\nu}}_{w_t}\right)}{\sum_{w' \in V} \exp\left(\boldsymbol{\nu}_{C_t} \cdot \tilde{\boldsymbol{\nu}}_{w'}\right)} \quad \cdots (4)$$

これ以降の計算は skip-gram モデルと同様に行う．

ところで，1単語だけからなる文脈 $C_{w_t} = (c)$ を考えると，式(4)は，式(3)と同じになる．すなわち，CBoW モデルと skip-gram モデルは同じモデルになる．したがって，skip-gram モデルは CBoW モデルの特別な場合と考えることができる[6]．

2．4　ニューラル言語モデルと単語ベクトルの関係

式(3)で唐突に単語ベクトル $\nu_C, \tilde{\nu}_w$ が出てきたが，先述したように，これがコーパスを使って最尤推定したいパラメタ θ である．そして，これらの単語のベクトル表現は入力層，隠れ層，出力層からなるニューラル言語モデルの重み行列になっている[12]．

図2は，Word2Vec のニューラル言語モデルを概念的に描いた図である．この図において，$|V|$はコーパス全体の語彙数で，N は隠れ層のニューロン数を表している（$N \ll |V|$）．また，$W_{|V| \times N}$ は入力層と隠れ層の間の重み行列，$W'_{N \times |V|}$は隠れ層と出力層の間の重み行列である．入力層から文脈単語 c の one-hot ベクトルが入力され，重み行列 $W_{|V| \times N}$ を使って隠れ層（中間層）への入力値が計算される．ここで，one-hot ベクトルとは，(0,1)を要素とする $|V|$次元ベクトルで，単語番号（全語彙 V の各単語に1から$|V|$までの番号を割り振ったもの）の要素のみが1で，それ以外の要素はすべて0とするベクトルである．次いで，隠れ層の出力と重み行列 $W'_{N \times |V|}$ を使って出力層への入力値を求める．最後に，出力層に Softmax 関数を適用することにより，予測対象単語 w の one-hot ベクトルが得られる．こうして得られた単語が目的の単語でなかった場合，つまり，ニューラルネットが間違った答えを出した場合，誤

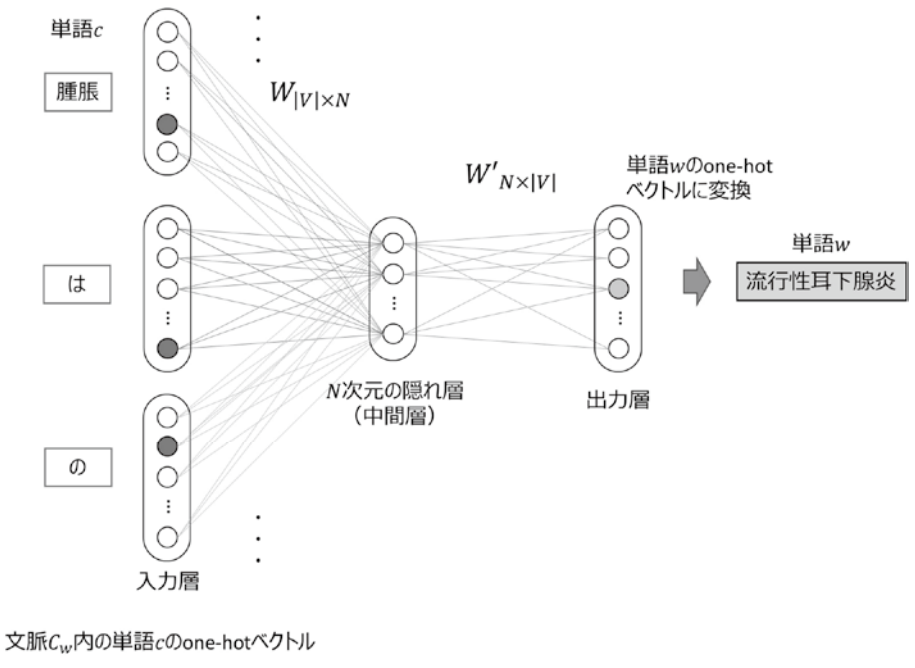

図2　Word2Vec のニューラル言語モデル

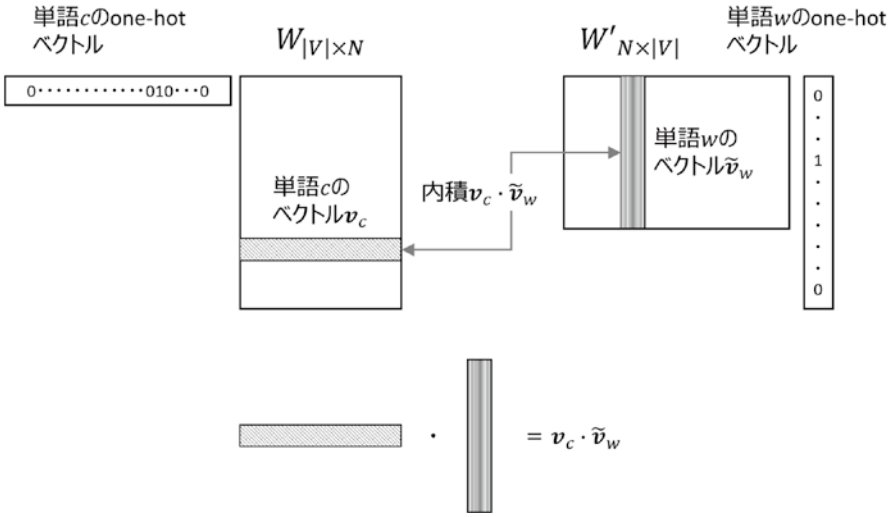

図3 ニューラルネットと分散表現

差を逆伝搬させて重み行列 $W_{|V|\times N}$, $W'_{N\times|V|}$を更新する.

図3は，これまで述べてきたことを図で表したものである．図に示すように，入力層と隠れ層の間の重み行列 $W_{|V|\times N}$ の行ベクトル ν_c が文脈単語 c の単語ベクトルで，隠れ層と出力層の間の重み行列 $W'_{N\times|V|}$の列ベクトル $\tilde{\nu}_w$ が予測対象単語 w の単語ベクトルになる.

3．方法

3．1 文書分類

一般に分類問題は，入力データ x（特徴量あるいは素性と呼ばれる）からカテゴリ $C\in\{C_1,\cdots,C_{|C|}\}$ への写像 $y=f(x;\lambda)$ とみなすことができる（$|C|$ はカテゴリの数）．ここで，λ は写像のパラメタで，学習データ (x_i,y_i), $i=1,\cdots,M$ から決定する（M は学習データの数）．学習によって λ が決まると，未知の入力データ x^* に対して写像 $f(x^*;\lambda)$ を適用すれば，x^* に対応するカテゴリを求めることができる.

自然言語処理で行われる文書分類タスクでは，特徴量として文書に含まれる単語の出現頻度が用いられることが多い．すなわち，$x=(f_1,\cdots,f_{|V|})$ で，f_i は i 番目の単語がこの文書に出現した回数である．本研究では，特徴量として，単語そのものではなく，前章で述べた単語のベクトル表現を用いる．具体的には，文書 D に含まれる単語の単語ベクトルを ν としたとき，

① 合成ベクトル：$x\equiv\nu_S=\Sigma_{\nu\in D}\nu$

② 重心ベクトル：$x\equiv\nu_G=\nu_S/|D|$

の2パターンの特徴量を用いた分類実験を行った．ここで，$|D|$ は文書 D に含まれる単語数である．その際，ベースラインとして単語の出現頻度を特徴量とした分類実験を行い，それらの間で分類性能を比較した．なお，分類にはロジスティック回帰モデルを用い，scikit-learn の LogisticRegression を使った.

3．2 分類対象文書

国内医学論文情報データベースである医中誌 Web に，代表的な女性疾患である「子宮内膜症」,「子宮頸癌」,「子宮体癌」,「子宮筋腫」,「卵巣腫瘍」の5つの疾患名を検索キーワードとして入力し，抽出された文献の抄録を分類対象の文書，検索キーワードを分類カテゴリとした．なお，検索の際，絞り込み条件は「原著論文」,「症例報告」,「抄録あり」とした．また，異なる疾患名で検索したにも拘わらず，同じ文献がヒットした場合は，分類対象から除外した.

3．3 単語ベクトルの学習

単語のベクトル表現は，前節で述べた抄録を学習コーパスとして用い，skip-gram モデルによって獲得した．ただし，獲得する単語ベクトルは名詞に限定した．また，単語のベクトル表現のドメイン依存性を調べるために，2017年7月20日の Wikipedia 日本語全文データ（ファイルサイズ7.1GB）[21] を学習コーパスとして単語のベクトル表現（以降，Wiki ベクトルと呼ぶ）を獲得し，抄録を学習コーパスとした場合（以降，抄録ベクトルと呼ぶ）と比較した．なお，単語ベクトルの獲得には gensim 社の word2vec を利用した.

３．４　分類実験と評価方法

分類実験は3.2節で述べた抄録の3/4を学習データに用いて分類モデルを構築し，残りの1/4を検証データに用いてモデルの性能を評価した．抄録から単語を切り出すために形態素解析ツール MeCab を用いた．ただし，単語は名詞のみを対象とした．また，形態素解析による単語の過分割を防ぐために，KHCoder の「複合語の検出」機能に備わっている専門用語自動抽出システム「TermExtract」を用いて47,674語の複合語を抽出した．抽出した複合語で MeCab のユーザ辞書を作成し，形態素解析実行時に参照した．

また，形態素解析処理で切り出された単語の中に正解である疾患名やそれに近い単語が含まれるのを防ぐために，ストップワードとして「子宮内膜症，子宮頸がん，子宮頸癌，子宮体がん，子宮体癌，子宮筋腫，卵巣腫瘍，がん，癌，子宮，膜，頸，体，筋腫，卵巣，腫瘍」の16語を指定した。

文書分類の性能評価には，全検証データのなかで正しく分類されたデータの割合として定義される正確度（Accuracy）を用いた．また，分類結果を混同行列の形に整理し，分類モデルの網羅性を示す指標である分類カテゴリ（疾患名）ごとの再現率（Recall）や分類モデルの正確性を表す適合率（Precision）を算出した．

４．結果

４．１　抄録データの抽出結果

表1に医中誌 Web の検索結果を示す．抽出した抄録は全部で11,037件あり，そのうち3,156件は重複していたため除外し，残った7,881件で抄録コーパス（抄録を用例として集めたもの）を作成した．ファイルサイズは約7.2MB だった．重複の除外後，抄録数が最も多かった疾患名は卵巣腫瘍で3,368件，最も少なかったのが子宮体癌の743件で両者の間に約5倍の違いがあった．このようにカテゴリごとのインスタンス数が大きく違う場合，正確度が必ずしも良い性能指標とはならないことに注意して結果を解釈する必要がある．

４．２　分類結果

表2に分類結果を示す．表中最下段の「ベースライン」は，単語の出現頻度を特徴量として分類した結果である．

使用した gensim 社の word2vec には単語ベクトルの次元数 N（隠れ層のニューロン数），文脈の単語数 δ，そして学習コーパスに出現する単語の出現頻度の下限 m（出現回数が m 未満の単語はベクトル化しない）をパラメタとして指定できる．表2は，N=500, δ=50, m=10 として分類を行った結果である．また，表中の項目「正則化」は，ロジスティック回帰モデルの正則化パラメタ C である．C を0.1

表1　医中誌 Web の検索結果

疾患名	抄録数		
	重複なし	重複あり	合計
子宮筋腫	1, 409	640	2, 049
子宮体癌	743	777	1, 520
子宮内膜症	1, 071	490	1, 561
子宮頸癌	1, 290	205	1, 495
卵巣腫瘍	3, 368	1, 044	4, 412
合計	7, 881	3, 156	11, 037

表2　分類結果

Word2Vec学習コーパス	特徴量	正確度	正則化
抄録コーパス	合成	0.770	C=10000
	重心	0.770	C=10000
Wikipedia	合成	0.619	C=0.1
	重心	0.630	C=0.1
ベースライン（単語の出現頻度）		0.807	C=0.1

表3 分類結果の混同行列

疾患名	分類器の予測						
	子宮頸癌	子宮筋腫	子宮内膜症	卵巣腫瘍	子宮体癌	合計	再現率
子宮頸癌	**237**	20	4	39	12	312	0. 760
子宮筋腫	19	**240**	26	67	7	359	0. 669
子宮内膜症	5	31	**202**	41	5	284	0. 711
卵巣腫瘍	21	27	21	**745**	12	826	0. 902
子宮体癌	26	20	7	44	**93**	190	0. 489
合計	308	338	260	936	129	1971	
適合率	0. 769	0. 710	0. 777	0. 796	0. 721		0. 770

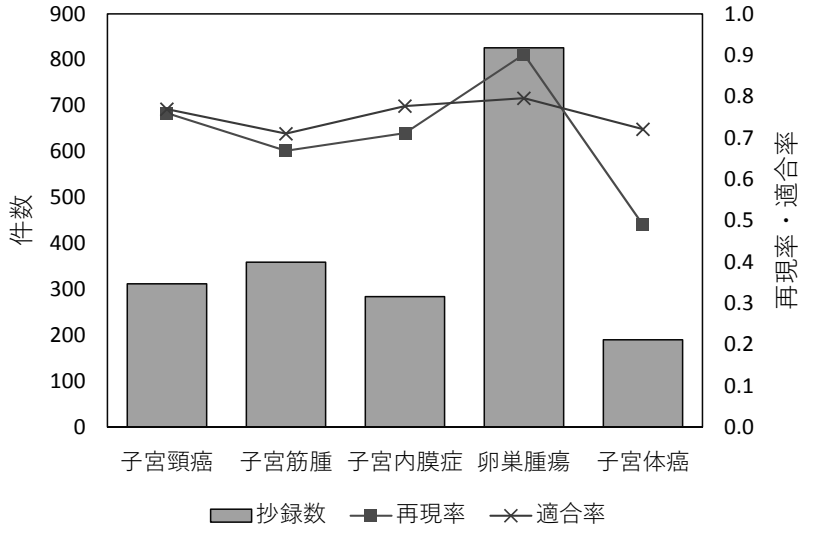

図4 再現率と適合率

～10000まで10倍ずつ変えながら分類を行い，正確度が最も高い値とそのときの正則化パラメタを示してある．

抄録ベクトルを特徴量として行った分類結果はWikiベクトルの場合に比べて約0.14～0.15高かったが，ベースラインに比べて約0.04低かった．また，単語ベクトルの合成ベクトルと重心ベクトルでは，抄録コーパスもWikipediaもほとんど違いはなかった．

表3は，抄録ベクトルの重心を特徴量に用いて行った分類実験の混同行列である．そのときの再現率・適合率を各疾患名の抄録件数とともにグラフ表示したものを図4に示す．

4．3 分類性能のパラメタ依存性

word2vecのパラメタが分類性能にどのような影響を及ぼすかを調べた．図5は，単語の出現頻度の下限 m を10に固定し，文脈の単語数 δ =10, 20, 30, 40, 50 のそれぞれに対して，分類性能が単語ベクトルの次元数 N によってどのように変わるかを描いたものである（ N=100, 200, 300, 400, 500 ）．特徴量には単語ベクトルの重心を用いている．また，正則化パラメタは C=10000 とした．なお，skip-gramモデルは負例サンプリングの際，ランダムに擬似負例単語を生成しているため，モデル構築の都度，異なる単語ベクトルが獲得され，その結果，分類結果も変わってくる．そこで，図5を描くにあたって同じパラメタで5回の分類実験を行い，得られた正確度の平均値をプロットした．

5．考察

5．1 分類性能の比較

表2に示すように，特徴量として単語ベクトルを使うよりも単語の出現頻度を使った方が分類性能は高かった．しかし，出現頻度による分類は単語の類

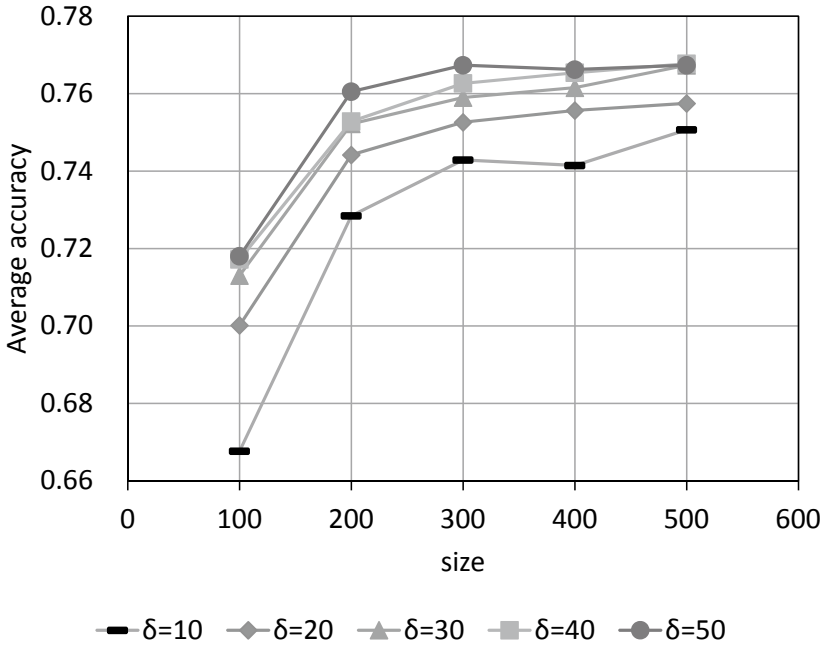

図5　分類性能の次元数と文脈語数依存性

表4 「子宮筋腫」に類似した単語

	抄録ベクトル		Wikiベクトル	
	類似単語	類似度	類似単語	類似度
1	卵巣腫瘍	0.825	気管支喘息	0.936
2	卵巣嚢腫	0.816	ナルコレプシー	0.931
3	粘膜下筋腫	0.717	クローン病	0.927
4	子宮腺筋症	0.711	膀胱炎	0.927
5	筋腫	0.698	潰瘍性大腸炎	0.925

似度を全く考慮していない．たとえば,「疾患」と「病気」のように表層形が異なれば全く違う特徴量とみなす．にもかかわらず，単語ベクトルを使った分類性能が低かった原因として次のようなことが考えられる．

① 単語ベクトルの品質の問題
② 単語の多義性，文脈依存性の問題
③ 対義語の問題
④ 特徴選択（feature selection）の問題
⑤ Word2Vec の手法自体の問題

以降，これらの問題について考察を行う．

５．１．１　単語ベクトルの品質

当然のことながら，単語ベクトルの品質は学習コーパスの品質に依存する．学習コーパスの品質の問題には，それがどのような領域あるいはテーマを扱っているかというドメイン依存性の問題と，辞書あるいは教科書のような体系的な文書から構成されているかどうかという問題がある．

前者については，たとえば政治・経済について書かれた新聞記事から獲得した単語ベクトルを用いて医学文書の文書分類を行っても有効な結果は得られないだろう．そのことは表2の結果にも表れている．Wikipedia は何か特定のテーマを扱っているコーパスではないため，それを学習して得られた単語ベクトルは汎用的なものになることが予想される．それに対して，医学論文の抄録から学習した単語ベクトルは，今回の分類対象文書である女性疾患に関する論文の抄録が扱うテーマに特化した単語ベクトルが獲得されるであろう．そのため，表2に示すように後者の方が高い分類性能を示した．

gensim 社の word2vec には，指定した単語に類似する単語を求めるメソッドがある．ここでいう「類

似」とは，ベクトル間のコサイン類似度に基づくものである．表4に「子宮筋腫」という単語に類似した単語を抄録ベクトルとWikiベクトルとの間で比較したものを示す．なお，いずれの単語ベクトルも100次元で，文脈単語数が5，出現頻度下限値が5語で作成している．

抄録ベクトルの場合では「子宮筋腫」に最も類似している語として「卵巣腫瘍」を出力している．それに対してWikiベクトルでは「気管支喘息」を筆頭に挙げている．しかも類似度は0.936とかなり高い．確かに「気管支喘息」も疾患名であることに変わりないが，疾患名としては「卵巣腫瘍」の方がより近い概念であろう．このように，獲得される単語ベクトルにはドメイン依存性が顕著に表れる．

次に，辞書や教科書のように極力曖昧性を排除した文書に比べて，省略や暗黙の知識を前提にした文書では文脈から単語の意味を正確に捉えることが難しい．これは，アルゴリズムは，いわゆる「行間を読む」ことができないためである．たとえば「原因究明のため，基底部から組織を採取し・・・」という文章において，その文書が胃について書かれたものか，脳について書かれたものか，あるいは足について書かれたものかによって，「基底部」の意味が変わる．抄録のように，専門知識を前提に書かれた文書ほど，こうした問題が発生する可能性が高い．

５．１．２　単語の多義性と文脈依存性

ひとつの単語が複数の意味を持つ場合がある．これを単語の多義性という．たとえば，「勉強する」には「学習する」と「値引きする」という2つの意味がある．また，「頭」には「頭部」の意味だけでなく「先端部」や「首領」の意味もある．さらに，日本語はひらがな表記が可能なので，「参加」，「傘下」，「酸化」，「賛歌」などすべて「さんか」と表記できる．このような同音異義語は単語ベクトルの学習に際してノイズになる可能性が高い．すなわち，「さんか」に対応する単語ベクトルが意図したものにならない，あるいは，焦点の定まらないものになる可能性がある．

５．１．３　対義語

「良性」と「悪性」のように反対の意味を持つ語や，「正中面」と「前頭面」のように対照的な関係になっている語を対義語という．「喉頭部に良性の腫瘍が認められる」と「喉頭部に悪性の腫瘍が認められる」のように対義語は類似した単語と共起することが多いため，反対の意味にもかかわらず，似たようなベクトル，つまり，コサイン類似度の高い単語ベクトルになってしまう．そのため，症例報告を「良性腫瘍」に関するものと「悪性腫瘍」に関するものに分類するといったタスクでは単語ベクトルが有効に機能しない．単語の出現頻度を特徴量とする分類手法ではこのようなことは生じない．

５．１．４　特徴選択

今回の実験では単語ベクトルの重心ベクトルまたは合成ベクトルを特徴量として文書の分類を行った．しかしながら，重心ベクトルにしても合成ベクトルにしても，文書内の単語ベクトルが持つ情報を切り捨て，たった1つのベクトルで文書を特徴づけようとしている．すなわち，単語をベクトル表現することによって得られた豊富な情報を文書分類に有効に活用していない．このあたりにも，単語の出現頻度を特徴量とした場合に比べて分類性能が低くなっている原因があると考える．せっかくの単語ベクトルを有効に活用する特徴量の考案が望まれる．

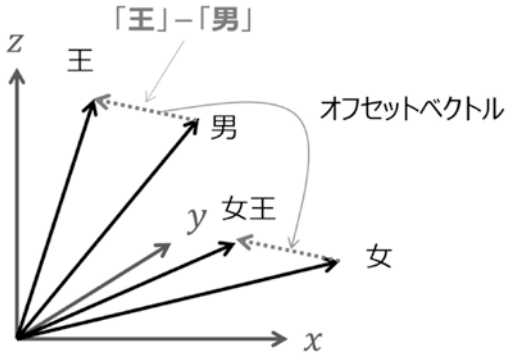

図6　Word2Vec の加法構成性

表5　抄録ベクトルの加減算の例

子宮癌−癌＋筋腫		子宮筋腫−筋腫＋癌	
得られた単語	類似度	得られた単語	類似度
粘膜下筋腫	0.717	子宮体癌	0.833
子宮筋腫	0.705	子宮癌	0.816
筋腫分娩	0.691	胃癌	0.791
創部	0.671	子宮内膜癌	0.790
瘢痕	0.631	子宮頸癌	0.786

５．１．５　Word2Vec の問題

Word2Vec の良く知られた性質に加法構成性がある[14]．有名な例として「$\overrightarrow{\text{王}}$のベクトルから$\overrightarrow{\text{男}}$のベクトルを引き，$\overrightarrow{\text{女}}$のベクトルを足すと$\overrightarrow{\text{女王}}$のベクトルに近づく」というのがある．図6はそれを3次元空間で説明したものである．

図に示すように，ある特定の関係を共有するすべての単語対（$\overrightarrow{\text{王}}$と$\overrightarrow{\text{男}}$，$\overrightarrow{\text{女王}}$と$\overrightarrow{\text{女}}$）は，ベクトル空間内で同じオフセットベクトル（図6の点線で示すベクトル）で関連付けられるように埋め込まれる．このように，Word2Vec は単に単語の類似性を獲得するだけでなく，セマンティックスも捕捉しているとされている．実際に抄録ベクトルで行った加減算の例を表5に示す．ここで，使用したパラメタは表4と同じである．

この表を見ると，もっともらしい結果のように思えるが，果たしてこれで妥当な文書分類が可能かという疑問が生じる．たとえば，表7の右側から「子宮筋腫－筋腫＋癌＝子宮体癌」という関係が得られるが，これは「子宮筋腫＋癌＝子宮体癌＋筋腫」と同等である．したがって，単語ベクトルの合成ベクトル（つまりベクトルの和）を特徴量とした場合，「子宮筋腫」と「癌」だけから構成される文書は，「子宮体癌」と「筋腫」だけからなる文書と見分けがつかない．これが文書分類タスクにプラスに働くかマイナスになるかは文脈に依存する．単語の出現頻度を特徴量とする分類ではこのようなことは起きない．

５．２　次元数依存性

図5から，単語ベクトルの次元が200次元あたりまでは次元数の増加に伴って急激に分類性能が向上するが，その後は緩やかに増え続けるものの，ほぼ飽和状態に達しつつあることがわかる．これは，単語の分散表現が元の単語空間の次元圧縮に相当しているからである[11]．

m_{ij} をコーパス中のある単語 i の周辺に文脈語 j が出現した回数とすると，m_{ij} を要素とする $|V|\times|C|$ 行列 M を単語文脈行列という．ここで，$|C|$ は文脈語集合 C の数である．行列 M を特異値分解して，特異値を d 個（$d<|C|$）で打ち切って低ランク近似することにより，単語ベクトルの次元を $|C|$ 次元から d 次元に圧縮することができる．これが特異値分解による低次元圧縮であるが，skip-gram モデルで求めた単語ベクトルは，これとほぼ等価であることが示されている[15]．

今回の実験に用いた抄録コーパスの語彙数は6,917個なので，word2vec によって6917次元の語彙ベクトル空間が約500次元の単語ベクトル空間に次元圧縮されたとみなすことができる．

５．３　文脈語数依存性

図5から文脈語数を10から段階的に上げていくと分類性能が系統的に上昇しているのがわかる．しかし，文脈語数が増えるにしたがって，分類性能の上昇が鈍る傾向が見られる．文脈語数は，単語の意味の決定に影響を及ぼす周辺単語の範囲を表しているので，これは単語間の意味的相関距離とも解釈できる．したがって，分類性能が鈍り始める δ ～30あたりが抄録の単語間意味的相関距離と考えられる．

分布仮説は，単語の意味は周辺の単語によって決まるという仮説であるが，筆者の知る限り，その範囲を調査した研究はない．もちろん，文書が扱う領域や文書の種類（論文，新聞記事，小説，会話文，SNS，・・・），著者などにも依存するだろうが，間接的であっても単語の分散表現を用いた文書分類から単語間意味的相関距離を求めることができるのは興味深い．

５．４　医療への応用

本研究は，文書データから単語ベクトルを獲得し，それを文書分類という自然言語処理に応用するというものであった．最後に視点を変えて分散表現の医療への応用について考察する．

自然言語処理以外の Word2Vec を用いた研究として，商品購買履歴から商品ベクトルを学習し，類似するユーザや購入が期待できる商品などを予測する研究がある[16]．この研究では，Word2Vec における単語を商品とみなし，1回の買い物で同時に購入した商品（これをトランザクションと呼ぶ）を文脈とみなして商品ベクトルを獲得する．いったん商品をベクトル化してしまえば単語の場合と同様に商品の類似性や商品間の関連を求めることができる．また，ユーザごとのトランザクションを集めてそれを文書とみなし，本研究で行ったのと同じように文書分類を行うことによって類似するユーザに分けることができる．

そもそも分散表現の対象は単語である必要はなく，離散オブジェクトであれば何でもよい[6]．分散表現は離散オブジェクトを計算機上で扱うための道具である．それをうまく利用したのが前述の研究である．

これと全く同じ発想で，購買履歴を1回の診療（外来診療や入院から退院までの診療行為），商品を診療行為とみなすことにより，診療行為のベクトル表現を獲得できる．そうすれば電子カルテやレセプトから処方や検査などの診療行為を収集して類似する診療行為の類似度や関連性を調べたり，類似する患者を探したりすることができるかもしれない．さら

に一連の診療行為が表す文脈の中に疾患や転帰などのラベルを付与して文書分類ならぬ診療行為分類を行えば，現在行っている診療行為から患者の予後を予測できるかもしれない．

6．結語

本論文では，単語の分散表現を用いて文書分類を行った．その際，特徴量として単語ベクトルの重心ベクトルと合成ベクトルを用いた．その結果，単語の出現頻度を特徴量に用いる従来の方法を上回ることができなかったものの，それに匹敵する分類性能を示すことができた．分類性能が下回った原因として，単語ベクトルの品質，単語の多義性，文脈依存性，対義語，特徴選択，Word2Vec の手法自体の問題などが考えられた．なかでも折角獲得した情報の大部分を利用しないで捨ててしまう特徴選択には改善の余地が残された．これについては，複数の分散表現モデルのベクトルを平均したり結合させたりすることで分類性能が向上したという報告がある[17, 18]．また，リカレントニューラルネットワークの一種である Long Short-Term Memory (LSTM) に Word2Vec で獲得した単語ベクトルを系列として投入して文書分類を行う事例[19]や Word2Vec を改良して文書ベクトルを生成して感情分析や情報検索を行う手法[20]が提案されており，いずれも単語ベクトルだけを特徴量として用いる場合よりも分類性能が向上したと報告している．今後はこれらの方法も試してみたい．

本研究の意義は，単に文書分類に単語ベクトルを利用して従来の方法と比較しただけでなく，むしろ文書分類を評価尺度に利用して Word2Vec が生成する単語ベクトルの品質を評価した点にある．また，その副産物として，単語の意味的相関距離を求めることができたのも特筆すべき点である．

最後に医療への応用についてはデータの収集の問題，患者のプライバシー保護の問題，結果の評価方法など，実現に当たっては解決すべき課題は多いが，理論的には可能であり，興味深い知見が得られるのではないかと期待が高まる．医療情報（診療情報）の後利用に関しては，多くの関係者が関心を寄せていることでもあり，データさえ揃えば比較的手軽にできるこの手法を試してみる価値はあると考える．

謝 辞

抄録コーパスの作成に協力して頂いた片山裕加里氏，小山紘明氏，土屋拓海氏に謝意を表する．

文 献

1）LeCun Y, Bengio Y and Hinton G：Deep learning. *Nature*, **521**, 436-444, 2015.
2）Rav D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B and Yang GZ：Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, **21**(1), 4-21, 2017.
3）木戸尚治：ディープラーニング技術を用いたコンピュータ支援画像診断（CAD）. 臨床放射線 , **62**(10), 1223-1228, 2017.
4）Miotto R, Li L, Kidd BA and Dudley JT：Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Report*, **6**, 1-10, 2016.
5）Le QV, Ranzato MA, Monga R, Devin M, Chen K, Corrado GS, Dean J and Ng AY：Building high-level features using large scale unsupervised learning. *Proceedings of the 29th International Conference on Machine Learning (ICML'12)*, 507-514, 2012.
6）坪井祐太 , 海野裕也 , 鈴木潤：深層学習による自然言語処理 . 第1版 , 講談社 , 東京 , 2017.
7）Bengio Y, Ducharme R, Vincent P and Jauvin C：A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137-1155, 2003.
8）Harris ZS：Distributional structure. *Word*, **10**(23), 146-162, 1954.
9）Mikolov T, Chen K, Corrado G and Dean J：Efficient estimation of word representations in vector space. *International Conference on Learning Representations 2013 Workshop Proceedings*, 2013.
10）Mikolov T, Sutskever I, Chen K, Corrado G and Dean J：Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013.
11）岡崎直観：言語処理における分散表現学習のフロンティア（<特集>ニューラルネットワーク研究のフロンティア）. 人工知能 , **31**(2), 189-201, 2016.
12）Rong X：Word2vec parameter learning explained. **https://arxiv.org/abs/1411.2738**,[2014]. (2018.4.26 確認).
13）ウィキペディア日本語版のダンプ：

https://ja.wikipedia.org/wiki/Wikipedia: データベースダウンロード ,[2017]. (2017.7.20 確認)

14) Mikolov T, Yih SW and Zweig G：Linguistic regularities in continuous space word representation. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746-751, 2013.

15) Levy O and Goldberg Y：Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, **27**, 2177-2185, 2014.

16) Barkan O and Koenigstein N：Item2Vec: Neural item embedding for collaborative filtering. **https://arxiv.org/abs/1603.04259v3**,[2016]. (2018.2.13 確認)

17) Garten J, Sagae K, Ustun V and Dehghani D：Combining distributed vector representations for words. *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 95-101, 2015.

18) Yin W and Schütze H：Learning word meta-embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1351-1360, 2016.

19) 堅山耀太郎：Word Embedding モデル再訪（特集 自然言語処理と数理モデル）. オペレーションズ・リサーチ：経営の科学 , **62**(11), 717-724, 2017.

20) Le QV and Mikolov T：Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, **32**(2), 1188-1196, 2014.

（平成30年6月5日受理）

Document Classification using Distributed Representation of Words as Features

Masaaki TANAKA

(Accepted Jun. 5, 2018)

Key words : document classification, distributed representation, Word2Vec, skip-gram model, natural language processing

Abstract

Document classification is one of the representative research subjects of natural language processing and it has been applied to topic classification, reputation analysis, filtering, etc. In document classification, the word frequency has been used as features of a document. However, it is difficult to calculate the similarity and relevance between words from the information of the word itself. Therefore, the author aimed to improve the classification performance by using distributed representation of words as features. First, 7,881 abstracts excluding duplications were extracted from the ICHUSHI Web, which is a Japanese medical literature information database, and they were used as a corpus for machine learning. Next, vector representation of words was obtained using skip-gram model. Experiments were performed to classify the abstracts into five diseases using the centroids and synthetic vectors of the obtained word vectors as features. For the purpose of evaluation, the result was compared with the classification result by the conventional method using word frequency. As a result, the accuracy of classification by this method was 0.770, which was not able to exceed the conventional method (0.807), but it was able to obtain classification performance comparable to it. The reason why the classification performance by this method was lower than that of the conventional method was considered as the quality of the word vector, ambiguity of the word, problem of feature selection, and so on. Among them, there is room for improvement in feature selection which discards most of the acquired information without using it.

Correspondence to : Masaaki TANAKA

Department of Health Informatics
Faculty of Health and Welfare Services Administration
Kawasaki University of Medical Welfare
Kurashiki, 701-0193, Japan
E-mail : mtanaka@mw.kawasaki-m.ac.jp

(Kawasaki Medical Welfare Journal Vol.28, No.1, 2018 167－178)