
一般口演

一般口演5 知識工学

2017年11月21日(火) 08:30 ～ 10:00 G会場 (10F 会議室1006-1007)

[2-G-1-OP5-3] 誤嚥性肺炎予防の為に、観察記録解析における文章ベクトル化技法の有用性の検討

小牧 祥太郎^{1,3}, 村永 文学², 岩穴口 孝², 宇都 由美子³, 熊本 一朗³ (1.鹿児島医療技術専門学校, 2.鹿児島大学病院 医療情報部, 3.鹿児島大学大学院医歯学総合研究科医療システム情報学)

【背景及び目的】

医療機関においては誤嚥性肺炎の予防が求められ、入院時スクリーニング等で初回評価を行っている。しかし、患者状況は刻一刻と変化し、初回評価で問題とならない患者の発症も懸念される。我々は以前、テキストマイニング技法の TF-IDF を用いて看護師の観察記録の解析より、発症が危惧される患者の特徴的な語彙の抽出が行えたがそれに留まっている。今回、観察記録文章を数値ベクトル化する手法を用いて、入院患者の誤嚥性肺炎を予防する為の方法と、その有用性について検討を行った。

【方法】

2011年から2012年に鹿児島大学病院に入院した患者より、誤嚥性肺炎を発症した症例（誤嚥性肺炎群）22例と、対照として未発症の症例（非誤嚥性肺炎群）22例を抽出。双方の群それぞれ、機械学習用データ17例と評価用データ5例に分割した。その後、MeCabを用いて観察記録の形態素解析を実施（誤嚥性肺炎群については、発症日までの記録を使用）。gensimライブラリの Doc2vec を用いて機械学習用データに各評価用データを1例ずつ追加し、評価用データの cosine 類似度の算出を行った。また、解析にあたりベクトル次元数、window 幅等のパラメータも変更し適切な値の検証を行う。

【結果】

機械学習用データに誤嚥性肺炎評価用データを適用した結果、誤嚥性肺炎群では0.95以上、非誤嚥性肺炎群では0.95以下の cosine 類似度を示し、その値を閾値としたフィッシャーの正確確率検定において有意に識別が行えた（ $p < 0.01$ ）。また、パラメータの検証において、ベクトル数を抑え window 幅を広げた場合に誤嚥性肺炎症例と非誤嚥性肺炎症例の cosine 類似度の乖離幅が増大した。

【考察】

Doc2vec は近年注目されているディープラーニングと同様、ニューラルネットワークを使用し高速に文章ベクトルを算出する。TF-IDF は特徴的な語彙抽出に留まり、文章全体を評価する場合は主観的な部分が介在するが、cosine 類似度による識別は文章全体の定量的で客観的な評価が可能である。

誤嚥性肺炎予防の為に、観察記録解析における文章ベクトル化技法の有用性の検討

小牧 祥太郎^{*1,3}、村永 文学^{*2}、
宇都 由美子^{*3}、岩穴 孝^{*2,3}、熊本 一朗^{*3}

*1 鹿児島医療技術専門学校、*2 鹿児島大学病院 医療情報部、
*3 鹿児島大学大学院医歯学総合研究科医療システム情報学

Prevention of aspiration pneumonia by sentence vectorization technique

Shotaro Komaki^{*1,3}, Fuminori Muranaga^{*2},
Yumiko Uto^{*3}, Takashi Iwaanakuchi^{*2,3}, Ichiro Kumamoto^{*3}

*1 Kagoshima Medical Professional College, *2 Medical Informatics, Kagoshima University Hospital
*3 Medical Informatics Science, Kagoshima University Graduate School of Medical and Dental Sciences

[Background/Purpose] We examined methods for early detection of patients who are at risk of onset of aspiration pneumonia and methods for prevention of onset. [Methods] The nursing history records of patients in whom onset of aspiration pneumonia occurred during hospitalization in 2011 were used as machine learning data. We investigated whether prediction would be possible from the high degree of cosine similarity using Doc2Vec of GenSim library of the Python language. In searching for the cosine similarity threshold, we verified the goodness of fit by Fisher's exact test. Appropriate parameter values of Doc2Vec size, window and min_count were verified. [Results] A total of 11 patients who had aspiration pneumonia in 2011 was used as subjects. The data of 11 patients with aspiration pneumonia in 2012 and 11 control patients were used as evaluation data. The variation of each parameter value was 160. Threshold with a significant difference was found in the 14 parameter pattern values. [Discussion] It seems that in the evaluation record of the present study, the size of the Doc2Vec parameter of 50–100, window lower than 3–4, and min_count lower than 3 are appropriate. In the future, we would like to perform further examination by increasing the number of patients and leveling the number of characters used in the evaluation records.

Keywords: Dysphagia, Aspiration Pneumonia, Natural Language Processing, Neural Network, Incident

1. 緒論

医療機関において、入院中の患者に発生する誤嚥性肺炎は、転倒・転落や、患者の取り違え等のインシデントと同様、その発生を予防しなければならない責務がある。また厚生労働省が発表する人口動態統計によれば、主な死因別死亡率として肺炎が第3位となっており¹⁾、誤嚥性肺炎は社会的にも問題視されている。

現在、誤嚥性肺炎の予防については、入院時スクリーニング等で初回評価を行い、患者状態の把握を行っている。しかし、患者状況は入院開始時点から刻一刻と変化しているため、初回評価で問題が認められなかった患者にも、誤嚥性肺炎は発生しうる。入院時スクリーニングだけでは、誤嚥性肺炎のリスク評価は不可能であると思われる。

我々は以前、看護師の看護経過記録の解析より、誤嚥性肺炎の発症が危惧される患者の予見が可能であるか検討を行った²⁾。その結果、テキストマイニング技法のTF-IDFを用いて、誤嚥性肺炎が危惧される患者の特徴的な語彙の抽出までは行えたが、誤嚥性肺炎発症例か否かを識別するまでには至らなかった。TF-IDFでは、同義語の調整が必要であったり、抽出した語彙にノイズが含まれた場合に研究者の主観によるノイズの除去が必要となる等の問題点があった。

近年、ディープラーニング等の多層ニューラルネットワークに対する機械学習により、画像解析や自然言語解析の分野で飛躍的な技術革新が見られている。自然言語解析手法の一つに、テキストデータを文章ベクトル化する手法があり、例

えばLeらが提案したParagraph Vectorがある³⁾。Paragraph Vectorは、ニューラルネットワークを用いて文章を数値ベクトル化し、文章の類似性判断等に活用することが出来るツールであり、Mikolovが開発した単語を数値ベクトルとして表現する手法であるWord2vecの応用版である⁴⁾。

そこで我々は、TF-IDFより客観的な分析方法として、看護経過記録を数値ベクトル化する手法を用いて誤嚥性肺炎発症の予見が可能か試行してみた。具体的にはParagraph Vectorの実装の一つであるGenSimライブラリ⁵⁾のDoc2vecを用いて、看護経過記録を文章ベクトル化し、誤嚥性肺炎の発症が危惧される患者の早期発見に繋げ、発症を予防するための方法と、その有用性について検討を行った。

2. 目的

GenSimライブラリのDoc2vecを用いて、入院患者の看護経過記録を文章ベクトル化し、誤嚥性肺炎を発症した患者の記録との類似性を測定することで、誤嚥性肺炎の発症の予見を行えるか調査する。また、解析パラメーターの最適値についても模索する。

3. 方法

3.1 対象(学習データ・評価データ)

誤嚥性肺炎学習用データとして2011年に鹿児島大学病院に入院した患者を対象とし、全ての入院診療科から、入院中における誤嚥性肺炎を発症した症例を抽出した。評価デ

ータとして、2012 年も同様の手法により、誤嚥性肺炎の症例の抽出を行った（以下、症例群）。また、対照群としては、交絡による影響を可能な限り排除するために、症例群の診療科と年齢群が同様の構成となるように、2012 年のデータから対照群として抽出した。

3.2 観察記録の形態素解析

MeCab⁶⁾を用いて看護経過記録の形態素解析を実施した（誤嚥性肺炎症例については、発症日までの記録を使用）。なお、形態素解析用辞書として、MeCab に付属されている ipadic 辞書、Comejisyo⁷⁾（1文字単語は削除）に加え、鹿児島大学病院独自のユーザー辞書を使用した。ユーザー辞書には、症例群・対照群のデータとは関連のない、過去の看護経過記録を MeCab で分かち書きを行い、うまく分かち書きされないような用語（固有名詞等）を抽出し、ユーザー辞書として登録した。

3.3 分析・評価

評価方法として、2011 年の誤嚥性肺炎発症データを学習用データとして学習させ、次に 2012 年の症例群データ、および対照群データから 1 例のデータを選択し、Doc2Vec で算出されるコサイン類似度により判定を行った。つまり、過去の誤嚥性肺炎発症の記録に、評価用の各データがどの程度類似しているかを判断させるデザインである。

なお、今回使用した Doc2vec ライブラリにおいては、評価データに対して、学習データの個々の記録全てに対するコサイン類似度が出力される為、出力されたコサイン類似度の平均を算出し、その結果を本研究における評価データのコサイン類似度とした。それぞれの患者のコサイン類似度において一定の値を示した症例を、診断基準上の誤嚥性肺炎が危惧される症例とし、それを下回る症例を診断基準上の誤嚥性肺炎のリスクが少ない症例として、感度・特異度について評価を実施した。

本研究では、size（ベクトル次元数）、window（使用する近隣単語数）、min_count（除去する単語の出現回数）の 3 つのパラメータについて最適値を模索した。なお、各パラメータの調査範囲については表 1 の通りとする。

表 1 各パラメータの調整値

size	50, 100, 200, 300 (4 パターン)
window	3, 4, 5, 6, 7, 8, 9, 10 (8 パターン)
min_count	1, 2, 3, 4, 5 (5 パターン)

4. 結果

4.1 学習データと評価データの基本属性

学習データとなる 2011 年の誤嚥性肺炎患者は 11 例抽出された。評価データとなる 2012 年の症例群も 11 例抽出されたため、例数を統一するため対照群の評価データについても 11 例抽出を行った。今回の対象となった患者の記録における基本語数（単語数、文字数）を表 2 に、分析日数を表 3 に示す。また、対象患者の DPC データに登録された病名集計（主病名）の結果を表 4 に示す。

表 2 学習データ、対象群（症例群、対照群）の基本語数

	学習データ		評価データ (2012 年症例群)		評価データ (2012 年対照群)	
	単語数	文字数	単語数	文字数	単語数	文字数
1	4,839	13,362	4,127	11,395	4,561	12,481
2	6,309	17,526	49,277	138,062	73,634	206,164
3	29,984	82,321	3,892	10,758	44,195	121,777
4	399	1,097	2,775	7,584	10,036	27,476
5	1,379	3,864	4,576	12,570	9,818	27,054
6	4,869	13,719	4,155	11,548	1,408	3,900
7	19,716	54,460	8,988	24,827	5,759	15,759
8	1,778	4,964	2,117	5,838	20,312	55,467
9	1,208	3,316	12,121	33,373	16,400	45,865
10	6,817	18,897	7,334	20,484	1,303	3,553
11	7,871	21,775	4,844	13,514	8,022	22,317

表 3 学習データ、対象群（症例群、対照群）の分析日数

	学習データ	評価データ (2012 年症例群)	評価データ (2012 年対照群)
分析日数	412 日 (1 人平均 37 日 ± 64)	377 日 (1 人平均 34 日 ± 35)	822 日 (1 人平均 75 日 ± 75)

表 4 学習データ、対象群（症例群、対照群）の病名集計

学習データ		テストデータ (2012 年症例群)		テストデータ (2012 年対照群)	
分類	数	分類	数	分類	数
腫瘍系疾患	16	腫瘍系疾患	7	腫瘍系疾患	10
循環器系疾患	13	循環器系疾患	9	循環器系疾患	10
廃用症候群関連	5	廃用症候群関連	6	糖尿病性疾患	3
泌尿器系疾患	3	呼吸器系疾患	5	廃用症候群関連	3
消化器系疾患	2	眼系疾患	3	眼系疾患	2
眼系疾患	1	泌尿器疾患	3	呼吸器系疾患	2
運動器系疾患	1	精神系疾患	2	消化器系疾患	2
その他	6	消化器系疾患	2	運動器系疾患	2
		内分泌疾患	2	精神系疾患	1
		神経変性疾患	1	その他	2

4.2 評価データへの適合度とパラメータの探索について

各評価用データについて、学習用の誤嚥性肺炎症例データで機械学習させたオブジェクトに対し、コサイン類似度を算出し、その平均値を求めた。表 5 は、パラメータを size=50, window=3, min_count=1 と設定した場合の例である。求められたコサイン類似度の大きい順に並べ替えている。

次に、各コサイン類似度を閾値として、2×2 テーブルを設定し、フィッシャーの正確確率検定にて適合度を検定した。例えば表 5 において、コサイン類似度の閾値を 0.997830 と設定した場合に、誤嚥性肺炎症例（症例）8 例及び非誤嚥性肺炎症例（対照）2 例が採用されるため、2×2 テーブルは、(8,3,2,9) より p=0.03 と算出される。

表 5 各評価データの学習データ(肺炎症例)に対する
コサイン類似度・検定結果

(size=50,window=3,min_count=1 の場合)

評価データ	コサイン類似度	p 値
★症例_06	0.998303	1.0000
対照_01	0.998181	1.0000
★症例_10	0.998110	1.0000
★症例_04	0.998108	0.5865
★症例_05	0.998004	0.3108
★症例_01	0.997969	0.1486
★症例_08	0.997957	0.0635
対照_06	0.997889	0.1827
★症例_03	0.997873	0.0805
★症例_11	0.997830	0.0300
★症例_09	0.997770	0.0089
対照_10	0.997684	0.0300
対照_05	0.997589	0.0805
対照_08	0.997537	0.1827
対照_07	0.997361	0.3615
★症例_07	0.995560	0.1486
対照_11	0.995331	0.3108
対照_04	0.990321	0.5865
★症例_02	0.956099	0.2143
対照_03	0.944767	0.4762
対照_02	0.936698	1.0000
対照_09	0.910369	1.0000

size,window,min_count の各調整パラメーター値を変更し
た計 160 パターンに対して、同様の処理を行った。このうち有
意差が得られたパターンの詳細を表 6 に示す。

表 6 有意差の得られた各パラメーター値

size	window	min_count	閾値 (コサイン類似度)	p 値
50	3	1	0.9977	0.008
50	3	2	0.9959	0.001
50	4	5	0.992	0.035
100	3	3	0.9946	0.029
100	4	2	0.9985	0.029
100	6	4	0.9986	0.012
100	7	2	0.9995	0.035
200	4	2	0.9992	0.029
200	4	5	0.9978	0.012
200	7	2	0.9997	0.008
200	8	2	0.9997	0.023
300	3	3	0.9982	0.023
300	6	2	0.9997	0.023
300	8	1	0.9999	0.023

※(各パターン、最も p 値が低く得られた閾値を示す)

5. 考察

5.1 パラメーターの設定について

今回、計 160 パターンのパラメーター値について検証を行
い、そのうち 14 パターンにおいて有意差が認められた。その
なかでも、p 値が特に小さく出現したパラメーター値を考慮す
ると、ベクトル次元数(size)が 50 の地点が挙げられる。
Mikolov は、特徴ベクトルの次元が倍になるたびに学習用の
単語量も倍にする必要があると述べている⁴⁾。本研究の検証
が必要な点として、学習データ数の問題が挙げられる。学習
データは 11 症例分の看護経過記録であり、学習用の単語数
として十分な量とは言い難い。そのため、今回の検証におい
てはベクトル次元数を抑えることは有効であったと考えられる。
また、ベクトルの調整として今回の解析データである看護経
過記録は、医療文章として共通した点が多い事が想定され
る。そのため、今回のデータにおいては豊富なデータ量を用
意しない限り、多次元のベクトルを反映する事は難しいと考
える。

また、window 幅については、その語彙の近隣単語を拾う
ものである。看護経過記録は「簡潔」さが求められ、シンプ
ルな表現を用い冗長な記述は多くはみられない事が想定され
る^{8),9)}。抜く記録の性質上、こちらもあまり幅を広くとることなく、
4~5 分節程度が適当であると考えられる。小野は、
Word2Vec は元々英語の文章解析を対象として作られたもの
であり、日本語は分かち書きをする必要があるため、文章か
らの単語取得精度が低下せざるを得ないと述べている¹⁰⁾。
本研究でも、分かち書き用の辞書として、医療現場で用いら
れる語彙が適切に認識されるよう、Comejisyo⁷⁾を導入した。
しかしながら、「ML」などの量を示す単位が、「M」、「L」のそ
れぞれの単位に分かち書きされてしまうなど、適切に語彙を
認識出来なかった為、今回の研究においては、1 文字の単
語については削除を行った。window 幅については、解析対
象となるテキストの性質においてどの程度の幅で行うかは、
今後、辞書の整備と合わせて十分な検討が必要であると考
える。

また、min_count については、数値を大きくすると、記録の
なかでも頻回に使われる語彙のみに限定される可能性がある。
著者らは、誤嚥性肺炎患者と非誤嚥性肺炎患者における、
TF-IDF 値による語彙の出現頻度の検討において、誤嚥
性肺炎患者では、非誤嚥性肺炎患者の語彙には認められ
ない ADL 状況の低下(「オムツ交換」、「酸素」等)を示すよ
うな語彙の確認が行えた²⁾。しかし、今回のデータにおいては、
語彙数が様々であり、語彙数が少ない文章であれば、誤嚥
性肺炎患者の特徴的なサインとなりえる語彙が消えてしま
う事も懸念される。よって、min_count はあまり増加させない事
が適切であると考えた。

よって、今回の検証においては size は 50~100 程度、
window は 3~4 以下、min_count は 3 以下の付近が適正で
はないかと示唆された。

5.2 分類精度の向上について

学習データの量を増加させ、パラメーターの size を大きく
することにより、多次元の距離を表せる事が可能となる。しか
しながら、文書量の差がコサイン類似度に現れる可能性も懸
念される。今回は単語数や解析日数は表 1 にあるように、
個々の学習データにおいて様々である。誤嚥性肺炎症例の
学習データや評価データにおいては、誤嚥性肺炎の診断が
付く 1 日前までを解析対象としている。場合によっては、誤嚥
性肺炎の徴候がほとんどみられない入院当初も解析に入っ
ていることが考えられる。記録をベクトル化した際には、ノイズ

となる情報が大量に存在する場合、誤嚥性肺炎の症例であっても、コサイン類似度は乖離する事も懸念される。その為、今後は記録量を調整し解析を行っていく必要も伺える。

また、今回の対照群(非誤嚥性肺炎群)の評価データとして、誤嚥性肺炎の症例と同じ、診療科・年齢群より抽出している。その為、誤嚥性肺炎の診断はなくとも、原文参照を行うと肺炎に繋がる可能性の高い症例が同えた(対照群評価データ1の語彙として、「酸素」、「口腔ケア」など)。純粋に学習データの質を上げるためには、誤嚥性肺炎症例のデータ量を増やしていくことはもちろんの事、明確に誤嚥性肺炎とならない症例のデータを抽出し、コサイン類似度がどの程度乖離するか検証を行い、明確な閾値の設定を行う必要がある。

5.3 TF-IDF との違いについて

文章の特徴を表現する手法として、従来、TF-IDF 等による単語の出現頻度を用いる方法が使われてきた。しかし、語彙が抽出される TF-IDF では、特徴語として挙げられる語彙が同義語である事も多い。その為、語彙を統一するにあたり、主観的な要素が含まれやすい。また、挙げられた特徴語について、その文章を表すには解釈に困難な語彙もしばしば見受けられる。そのため、特徴語彙から文章の判断を行う事は、判定者により差が生じることも懸念される。一方、Doc2vec は、ニューラルネットワークに学習させ、その中間層の値を文章の特徴ベクトルとして出力している。そのため、定量的であり、客観性の高い評価が行えると考えられる。

また、TF-IDF は上述しているように語彙の出現頻度のみが抽出されるため、その特徴語彙の前後までは考慮されていない。よって、「むせ込み」の語彙の後に続く、「あり(むせ込みあり)」、「なし(むせ込みなし)」等の判断は行えない。観察者が「むせ込み」への着目から、懸念している様子の推測に留まる。その点、Doc2vec は window 幅により、症状の有無を近隣単語として捉える事が可能であり、ベクトル表現として評価が行える。

5.4 閾値の設定とエラーについて

今回のコサイン類似度により降順に並び替えた結果から、誤嚥性肺炎と非誤嚥性肺炎の症例が明確に識別されていない。対照群において、高いコサイン類似度として算出された症例については、誤嚥性肺炎の発症に至らなくても、発症が危ぶまれる患者であることが懸念される。よって、そのような症例についてはリスクが高いものとして扱っても、医療管理上の問題は少ないと考える。今後、最適な感度、特異度となるような予測閾値のカットオフ値を決定し、判定基準をどこで設けるか検討を行う必要がある。

6. 結論

本研究において、Doc2Vec による文章ベクトル化は TF-IDF と比較し、誤嚥性肺炎の予見に有用であることが示唆された。ベクトル化において、看護経過記録に関する Doc2Vec の最適パラメータ値は size は 50~100 程度、window は 3~4 以下、min_count は 3 以下の付近が適正と判定を行った。今後、例数の増加による学習用パラメータの充実と、学習データとなる記録の量の統一方法について検討が必要である。

7. 文献

- 1) 厚生労働省.平成 28 年(2016)人口動態統計の年間推計.
[<http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/suikei16/dl/2016suikei.pdf>].

- 2) 小牧祥太郎,村永文学,川原浩ら.診療記録の記載事項より誤嚥性肺炎の発症を予見する取り組みとその評価ーデータマイニングを用いてー. 第 34 回医療情報学連合大会論文集 2015;35:750-753.
- 3) Quoc Le, Tomas Mikolov. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014 pp. 1188-1196.
- 4) Mikolov Tomas,Chen Kai,Corrado Greg et al. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781,2013,pp.1-12
- 5) Rehurek Radim,Sojka Petr. Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.
- 6) MeCab. <http://taku910.github.io/mecab/>.
- 7) ComeJisyo. <https://ja.osdn.net/projects/comedic/>.
- 8) 村松洋,渡部勇,大崎千恵子.看護記録のテキストマイニング.情報処理学会論文誌 2010;No.3:112-122.
- 9) 平木久美子. 看護実践の一連の過程が見える看護記録ー質的監査と看護必要度監査ー. 看護きろく看護過程 2015;vol.25 no.4:26-29.
- 10) 小野良太,川村秀憲.ディープラーニングによるイベント情報分類器に向けた Word2Vec の活用検討. SIG-SAI, 25(6),2016,pp 1-6.