一般口演
# 一般口演21
# 医療データ分析９（機械学習・テキストマイニング２）
2018年11月25日(日) 10:40 〜 12:10 D会場 (4F 413+414)

[4-D-2-4] 文章ベクトル化における調整パラメーター値とカットオフ値の検討 ー 看護観察記録を用いた誤嚥性肺炎発見手法においてー

○小牧 祥太郎[1], 村永 文学[2], 宇都 由美子[3], 岩穴口 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）

【背景・目的】　医療機関において誤嚥性肺炎の発症を予防する責務があるが、刻一刻と変化する患者状況において入院時の初回評価のみでは不十分である。今回、機械学習における文章ベクトル化手法を用いて発症が危惧される患者の早期発見に繋がる手法を検討した。【方法】　2011年に誤嚥性肺炎を発症した症例の看護観察記録を学習用データとして用いた。評価用データとして、2012年の誤嚥性肺炎症例の記録と、2011年、2012年における誤嚥性肺炎未発症例の記録を用いた。 Python言語の GenSimライブラリの Doc2Vecを用いて、学習用データに対する評価用データのコサイン類似度より適合性を評価し予見可能か判定を行う。評価にあたり、Doc2Vecの Size（以下、 S）、 Window（以下、 W）、 Min_count（以下、 M）の適切なパラメーター値と、カットオフ値となるコサイン類似度の値について検証を行う。【結果】　学習用データは7例、評価用データにおいて誤嚥性肺炎症例が10例、誤嚥性肺炎未発症例が18例抽出された。統計学的に最も確実性の高い識別が行えたパラメーター条件は、① S=10,W=6,M=2② S=10,W=8,M=1③ S=10,W=9,M=4の3パターンとなった。カットオフ値を0.9992と定めた場合、①感度90〜100%、特異度33〜39%②感度100%、特異度28%③感度50〜80%、特異度61〜67%であった。【考察・まとめ】　パラメーター条件として、 Sizeは10が統計学的に最も確実性が高いと考えられたが、 Window、 Min_countは組み合わせ条件で変化が生じることが確認された。カットオフ値は、0.9991〜0.9995付近が適切と考えられ、最終的にパラメーターの組み合わせ条件により最も適切な値を判定可能と考える。本手法は誤嚥性肺炎の早期発見への有効性が示唆された。今後、例数の増加により精度の向上に努め、他の疾患の発症予測へも応用したいと考える。

文章ベクトル化における調整パラメーター値とカットオフ値の検討
―看護観察記録を用いた誤嚥性肺炎発見手法において―

小牧 祥太郎[*1,3]、村永 文学[*2]、
宇都 由美子[*3]、岩穴口 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

Examination of adjustment parameter value and cut-off value in document vectorization in the aspiration pneumonia discovery method

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

[Introduction]In order to promote early detection of aspiration pneumonia that may occur during hospitalization, we explored natural language processing by machine learning. We analyzed nursing observation records, which are unstructured records, using Doc2Vec of the Gensim library, and evaluated their suitability. [Purpose]We examined the refinement of the adjustment parameter value and the discrimination ability of the machine vector learning method in the nursing observation records of patients who were concerned about the onset of aspiration pneumonia. [Method]From 2013 to 2017, through random sampling, we established a control group by extracting patients with aspiration pneumonia during hospitalization. Conformity of learning data is verified by cosine similarity obtained when inputting evaluation data. In the analysis, after setting the number of iterations and the amount of learning data required for calculation, each adjustment parameter value was changed and the appropriate value was verified. [Results]Ninety-eight cases of aspiration pneumonia were extracted, and 228 cases of aspiration pneumonia were randomly extracted. Based on the adjustment parameter value, when the point of size 40, window=2, min_count=1 was the most suitable and the cut-off value of cosine similarity was set at 0.999613, the rate of sensitivity was 81.6% and the rate of specificity was 57.9%. [Consideration]It was possible to confirm the effectiveness of analysis by Doc2Vec and to refine effective parameter value in medical texts. In the future, we would like to apply these findings to the prediction of the onset of other diseases.

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

1. 緒論

現在、高齢化が進む日本においては肺炎は主な死因別死亡率の第 3 位であり、社会的にも問題となっている。また、肺炎の発症により、入院中の死亡率の上昇や、誤嚥性肺炎による入院日数も増加 [1,2]がみられるなど、医療機関においても、誤嚥性肺炎の発症は入院費用の観点からも非効率 [2,3]である。

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おり、初回評価で低リスクの患者においても誤嚥性肺炎は発生しうる為、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困難であると思われる。

そこで、入院中に発症しうる患者の早期発見に繋げるために、我々は機械学習による自然言語処理として、Le らが提案した Paragraph Vector[4]を用いたテキストデータを文章ベクトル化する手法に着目した。これは、ニューラルネットワークを用いて文章の数値ベクトル化により文章の類似性判断を行うことが出来るツールである。その実装の一つに GenSim ライブラリ[5]の Doc2Vec が存在する。Doc2Vec は 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 の文章応用版である[6]。Word2Vec[7]は、単語の one-hot ベクトルを入力層とし、中間層にて重み付けを行い、出力層より中間層へ誤差伝播を行った際に更新される値を単語ベクトルとして出力する手法である。

我々はこれまで、非構造化記録である看護観察記録を Doc2Vec にて解析する事で、誤嚥性肺炎発症の予見が可能か試行を行った[8,9]。その結果、最も適切とされるパラメーター値の確定には至らなかったが、多くのパラメーター範囲において誤嚥性肺炎症例を有意に識別することが出来た[8,9]。しかしながら、学習用データ・評価用データとなる記録量の統一がなされていなかった事、学習用データとなる例数が少ないことが懸念された。そのため、本研究では、記録量の統一と例数を増加し、前回の研究で得た有効パラメーター範囲を参考にパラメーター調査を行い、適合性評価を実施する。そのなかで、カットオフ値を定め、感度・特異度について検証を行い、モデルの識別能力の精度検証を行う。

2. 目的

誤嚥性肺炎の発症が危惧される患者の看護観察記録において、機械学習の文章ベクトル化手法における調整パラメーター値の精緻化と識別能力について検証を行う。

3. 方法

3.1 解析記録について

誤嚥性肺炎学習用データ、評価用データともに、2013 年から2017 年において鹿児島大学病院に入院した患者を対象とする。誤嚥性肺炎患者群（以下、症例群）においては、当該期間において、誤嚥性肺炎の診断がついた患者を用いた。

なお、共同研究者の医師とカルテ記録及び薬歴・検査結果を詳細に調査し、入院中における全ての入院診療科から誤嚥性肺炎と診断された症例より、処方として抗生剤の使用が確認された症例を症例群として定義した。また、対照群として、2018 年 6 月 1 日～7 月 20 日の間に入院した患者を抽出し、症例群の年齢構成より 95%信頼区間に相当する年齢に該当した症例をランダムに抽出し対照群とした。

3.2 解析記録の詳細と前処理について

解析対象とした看護観察記録は、1 記録あたり 40 文字(80Byte)の制限が設けられおり、簡素な表現で患者容体を逐次記録するように設計されている。看護師の観察記録は各症例により記録量は異なる。本解析においては、4 日間の記録量において検証を行う。なお、症例群は発症日より記録を遡り記録日数の調整を行い、対照群については該当期間中の入院により時間軸に沿って記録日数の調整を行った。解析の前処理として、Doc2Vec ではテキスト記録を事前に形態素解析しておく必要がある。今回、MeCab[10]を用いて看護師の観察記録の形態素解析を実施した。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、Comejisyo[11]（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。ユーザー辞書の作成には、症例群・対照群のデータとは関連のない、過去の看護観察記録を用いた。MeCab で分かち書き後に、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 分析手法

抽出された症例群を学習用データと評価用データにランダムに 2 分割し機械学習を行う。学習用データに対して、評価用データ（症例群、対照群）から 1 例ずつ算出されるコサイン類似度により判定を実施。出力されたコサイン類似度の平均を算出し、本研究における評価用データのコサイン類似度とした。

なお、Doc2Vec をはじめニューラルネットワークを用いた解析では、計算処理の効率より乱数を用いて重みづけの調整が行われる [12]。そのため、得られるコサイン類似度は毎回の計測に当たり微妙に変化が生じる。今回、解析対象とする医療文章においては、記述の特性上、ある程度似通った文章になる事が予想され、コサイン類似度は近い値になると考えられる。そのため、算出コサイン類似度は、バラツキを考慮して数回算出し、その平均とする。その際の繰り返し回数として、F 検定により分散が有意水準内に収まる試行数を検証し、繰り返し回数の決定を行う。

また、分類器の調整としては、評価用データに対し、学習用データの個々のコサイン類似度が出力されるが、その際の計算に要する学習用データ数(topN)について、MAP（Mean Average Precision）を指標に検証し、学習用データ数を決定する。

その後、探索パラメーターとして、size（ベクトル次元数）、window（使用する近隣単語数）、min_count（n 回未満登場する単語を破棄）についてパラメーター毎の AP（Average Precision）を基準に最適パラメーターの模索を行う。パラメーター値の調査範囲については、我々の過去の研究 [8,9]をもとに本解析記録である文章の性質を考慮し、min_count の調査範囲を決定後、size、window の調査範囲を定めた（表 1）。なお、min_count の調査範囲を 3 以下に限定した理由については考察で述べる。

表 1 各パラメーターの調整値

Size	10,20,30,40,50,60,70,80,90,100
window	4,5,6,7,8,9,10,11,12,13,14,15
min_cont	1,2,3

4. 結果

4.1 学習データ・評価データ数と基本属性

症例群は 98 例抽出された。対照群は 228 例抽出された。観察記録の個々の記録容量は症例群の記録量において、2～16kbyte、対照群にて 1～10kbyte であった。

分析に使用した記録における、語彙種類の総数、語彙数、各記録における語彙の出現頻度の割合を表 2 に示す。

表 2 基本語彙数・出現語彙の頻度

	症例群			対照群		
	A.V	max	min	A.V	max	min
語彙種類総数	337	675	120	162	451	31
語彙数	844	2208	217	343	1284	48%
1 語	63%	77%	51%	66%	83%	47%
2 語	16%	22%	10%	15%	30%	5%
3 語	7%	11%	3%	7%	15%	1%
4 語	4%	7%	0%	4%	11%	0%
5 語	2%	5%	0%	2%	12%	0%

また、症例群より 5 例をランダムに抽出し、記録中の単語における TF-IDF 値を算出した。上位 10 語を表 3 に示す。

表 3 症例群 5 例の記録における TF-IDF 値上位の語彙

症例 1		症例 2		症例 3	
語彙	語	語彙	数	語彙	数
L	11	嘔吐	8	吸引	7
分	14	上田	3	褥瘡	5
酸素	10	SPO2	8	瞳孔不同	4
5	13	多量	6	多量	6
/	14	L	6	白色痰	4
なし	29	酸素	6	.	7
.	10	末梢確保	3	発語	3
滴	4	低下	5	(+)/	2
ラコール	3	ノロウイルス	2	ガーゼ上層	2
金属音	3	巻きなおし	2	処置	3

症例 4		症例 5	
単語	数	単語	数
ロペミン	2	エネーボ	6
熱	3	大声	6
低値	3	入眠	15
配薬	2	多量	10
下痢	3	ホリゾン	4
効果	2	缶	5
ふらつき著明	2	覚醒	10
発汗	2	白色粘稠痰	5
ない	4	興奮状態	4
なし	11	A	7

4.2 分散の収束試行数について

繰り返し施行における分散のバラツキを以下に示す(表 4)。8 回から 9 回の時点にて分散が一定の有意水準範囲に収束した。

表 4　繰り返し回数ごとの分散と前試行との p 値

繰り返し回数	分散	前試行数との p 値
1	2.72695E-13	
2	1.79959E-14	3.9626E-17
3	2.97634E-14	0.042273374
4	2.20691E-15	4.57481E-16
5	6.67313E-15	9.86282E-05
6	5.66698E-16	7.50483E-15
7	2.08676E-15	6.88494E-06
8	8.24187E-16	0.000825608
9	6.72263E-16	0.241484009

(検証用パラメーターに size50,window10,min_count2 を使用)

4.3 計算に要する学習用データ数

topN として、1～20、49(学習用データ全例)において、計算に要する学習用データ数の MAP を図 1 に示す。

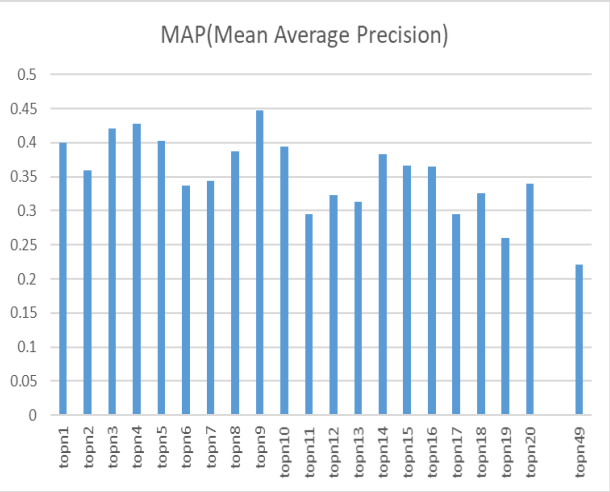

図 1　計算に要する学習データ数ごとの MAP

(検証用パラメーターに size10-40、window4-15、min_count1-5 を使用)

4.4 評価データへの適合性とパラメーターの検証

上記より、繰り返し回数を 8 回、計算に要する学習用データ数を top9 と定め、表 1 のパラメーター範囲において探索を行った。算出された各パラメーターの AP（Average Precision）上位 10 位を示す。

表 5　各パラメーター値の MAP における適合性上位 10 位

size	window	min_count	MAP
40	12	1	0.436205426
60	6	1	0.433256461
40	8	1	0.430546809
80	12	1	0.428175516
60	13	1	0.422276537
60	14	1	0.418351127
40	11	1	0.416334529
40	12	2	0.414355221
100	12	1	0.414306112
70	15	1	0.408291844

4.5 最適パラメーターにおける感度・特異度

パラメーター値において、AP が最も高い値を示した size40、window1、min_count1 における ROC 曲線を図 2 に示す。

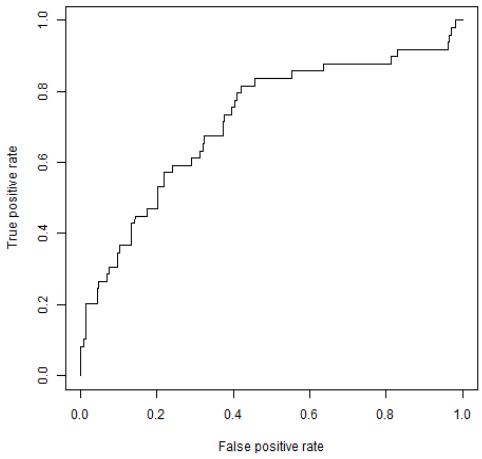

図 2　size40、window1、min_count1 における ROC 曲線

AUC は 0.718 であった。コサイン類似度のカットオフ値を 0.999613 と定めた場合、感度 81.6%、特異度 57.9%となった。

5. 考察

5.1 学習データ量の差異について

計算に要する学習用データにおいて、top1～20、49 において top9 が最も高い MAP を示し、学習用データ全例を計算する top49 は低い値を示した。よって、本解析において、計算を行う学習用データは全て使用するのではなく、上位に位置する数例での計算により識別能力が向上した。誤嚥性肺炎症例は食事の際の顕性誤嚥や、睡眠中などに生じる不顕性誤嚥例など誤嚥性肺炎に至るいくつかのパターンが存在する [13]。そのため、誤嚥性肺炎症例の記録も一律に近似したものにはならないことが考えられる。よって、本研究手法における算出方法では、誤嚥性肺炎症例記録の学習用データすべてを計算してしまう場合、様々なパターンが混同する事により、誤嚥性肺炎症例記録のコサイン類似度が高く算出されなかったことが懸念される。そのため、計算する学習データについては全例とせず、今回においては上位 9 位付近において最も識別率が高くなったと考えられた。

5.2 調整パラメーター値について

我々は以前の研究において [8,9]、調整パラメーターは size は 10～40 付近の低次元が妥当と考えられ、window に関しては、4～10。min_count は 1～4 が適当と考えた。

size はベクトル次元数を示す。Mikolov は特徴ベクトルの次元数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[6]。以前の我々の研究[9]において、size10～40 のなかで、size10 が最も有意に識別が行えていた。しかし、その際は学習データ数が 7 例のみであった。今回、40～100 の次元にて適合性が高くなった理由として、学習用データ数は 49 例に増加したことで、適合性の高いベクトル次元数は増加したと考えられた。

window は使用する近隣単語数である。以前、我々の研究においても window 幅の最適範囲の選定が困難なパラメーターである。こちらも表 5 より、6～15 の幅がみられたが、そのなかでも、11 以上の値が上位 10 位内に多くを占めた。これは、日本語の活用形の表現により形態素が細かく分割されたことが考えられる。例えば、「食べていなかった」を形態素解析すると、「食べ」、「て」、「い」、「なかっ」、「た」の 5 つの形態素に分けられる。よって、簡潔さが求められる文章である看護経過記録 [14,15]においても、表現によって分節数は多くなる事が予測される。その為、window は多くとることが効果的と考えられる。

min_count については、各記録における語彙の出現頻度は 1 語が約 50%以上となっている。よって、今回の文章サイズを鑑みた場合に、min_count によって単語を破棄するラインを高くしすぎると、症例群・対照群ともに文章自体が少なくなることが懸念される。また、表 3 にもあるように、TF-IDF 値によって記録の語彙を評価した際に、記録によっては 2 語～3 語のみの出現語彙であっても TF-IDF 値より重要度が示されている。よって、誤嚥性肺炎の特徴語も記録によっては出現頻度が必ずしも頻出しない事が考えられる。よって、min_count の設定を増加すると、重要な語彙を破棄してしまう事が懸念される。今回の検証からも、パラメーターの適合性における上位 10 位のうち 9 パラメーター値において min_count1 であり、今回の記録では語彙の破棄は行わない方が適切と考える。

5.3 カットオフ値の検討

表 3 における最適パラメーター値（size40、window12、min_count1）よりコサイン類似度のカットオフ値について検証を行った。評価用データ 277 例のコサイン類似度は、0.999838～0.99319233 の範囲に収まり、今回の解析データである看護観察記録は、医療文章として共通した点が多いことが想定された。そのため、コサイン類似度の範囲は、よほど記録が乖離していない限りは広がらないと予想される。今回の検証においては、小数点以下 3 桁からを有効値とみなすのが適切と考える。今回、カットオフ値を 0.999613 と定めた場合、感度 81.6%、特異度 57.9%となった。しかし、Doc2Vec は仕様上、計算させる度に僅かにコサイン類似度が変動する。また、調整パラメーターによっても変動する為、概ね内容が近似した文章を比較する場合においては、一律にコサイン類似度を固定する事は困難であると考える。

6. 結論

Doc2Vec における看護観察記録の解析において、我々のこれまでの研究において記録量の統一や例数が少ない事が懸念されていた。今回、それらを改善して検証を行った結果、Doc2Vec による解析の有効性を確認することが出来、医療文章における効果的なパラメーター値の精緻化を行うことが出来た。今後、他の疾患の発症予測へも応用したいと考える。

7. 文献

1) Ingeman, A., Andersen, G., Hundborg, H. H., Svendsen, M. L., & Johnsen, S. P. (2011). In-hospital medical complications, length of stay, and mortality among stroke unit patients. Stroke, 42(11), 3214-3218.
2) 小原仁. (2016). 入院後発症した誤嚥性肺炎の追加的医療費と在院日数: DPC データを用いた観察研究. 日本医療マネジメント学会雑誌= The journal of the Japan Society for Health Care Management, 17(3), 123-128.
3) Wilson, R. D. (2012). Mortality and cost of pneumonia after stroke for different risk groups. Journal of Stroke and Cerebrovascular Diseases, 21(1), 61-67.
4) 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.
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) 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
7) 西尾泰和. (2014). word2vec による自然言語処理. オライリー・ジャパン, May.
8) 小牧 祥太郎.村永 文学.宇都 由美子.岩穴口 孝.熊本 一朗 (2017). 誤嚥性肺炎予防の為の、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. 第 37 回医療情報学連合大会論文集;vol.37.380-383.
9) 小牧 祥太郎.村永 文学.宇都 由美子.岩穴口 孝.熊本 一朗 (2018). 看護観察記録の文章ベクトル化による誤嚥性肺炎発見手法の評価.第 22 回医療情報学会春季学術大会論文集;vol.22.
10) MeCab. http://taku910.github.io/mecab/.
11) ComeJisyo. https://ja.osdn.net/projects/comedic/.
12) 斎藤康毅. (2016). ゼロから作る Deep Learning: Python で学ぶディープラーニングの理論と実装. オライリー・ジャパン.
13) Marik, P. E. (2001). Aspiration pneumonitis and aspiration pneumonia. New England Journal of Medicine, 344(9), 665-671.
14) 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.
15) 平木久美子. 看護実践の一連の過程が見える看護記録 -質的監査と看護必要度監査-. 看護きろくと看護過程 2015;vol.25 no.4:26-29.