

氏名	おおの ゆきこ 大野 由紀子
学 位 の 種 類	博士（薬学）
学 位 記 番 号	博士甲 第 号
学位授与の日付	2025年3月10日
学位授与の要件	学位規則第4条第1項該当
学 位 論 文 題 目	自然言語処理技術を用いた医療記録からの症状抽出方法に関する 研究 ～患者の主観的情報に焦点を当てて～
論 文 審 査 委 員	（主査）教授 大谷 壽一（博士（薬学）） （副査）教授 漆原 尚巳（博士（医学）） 教授 堀 里子（博士（薬学））
慶應義塾大学学位規程第 11 条に基づく審査委員会実施日： 2025年2月4日	

論文内容の要旨

【研究の背景】

自然言語とは、私たちが日常で用いる言語である。詳細で自由度の高い内容が大量に作成されることから医療分野においても利活用が期待される情報源であるが、構造化されておらず、また意味に曖昧性を含むことからそのままではコンピュータによる解析が困難であった。この課題を乗り越えて、自然言語をコンピュータに解析させるために、目的に応じた自然言語処理技術が必要である。自然言語処理技術は、ロコミの解析、機械翻訳、検索エンジン、チャットボットといった多様な活用がされている。また、医療分野においても、自然言語処理の研究は、疾患・副作用症状・副作用症例の検出、疾患の自動診断、精神疾患や神経認知疾患患者の言語パターン解析、辞書の構築、文書分類など、多岐に渡って行われている。

自然言語処理技術の一つである固有表現認識（Named entity recognition; NER）を用いることで、医療テキストから病名や症状といった臨床情報を収集することができる。医療テキストの中でも、患者が発した生の情報（Subjective 情報; S 情報）に注目が集まっている。医師は患者の症状を過小評価・過少報告する傾向にあることが知られており、一方で患者の自己報告アンケートの有用性が認められている。また、主観的な症状を最初に気づくのは患者であることから、S 情報をモニタリングすることで有害事象のモニタリングや分析が期待できる。患者が発した生の情報源として、ソーシャルメディア投稿や医療記録に含まれる S 情報が挙げられるが、特に、薬剤師は薬学的観点から患者指導を行うため、薬剤師記録中の S 情報に有害事象情報が豊富に含まれる。しかしこれまで、薬剤師記録中の S 情報を対象とした病名や症状の NER システムは、ルールベースのものや、限定的な患者層のデータ、仮想患者のデータを用いたものとどまる。そこで本研究では、有害事象モニタリングの実現可能性を指向する汎用性の高

いシステムの構築を目標とし、次の 4 つの検討を行った。

第一章：医師記録を学習済みの病名や症状を収集する自然言語処理システムの薬剤管理指導記録への適用

第二章：医師記録を学習済みの自然言語処理システムを用いたバンコマイシン塩酸塩による急性アレルギー性皮膚症状の症例検出

第三章：薬剤師記録中の患者の訴えから病名や症状を抽出する症状抽出システムの性能向上：学習量増加による効果

第四章：薬剤師記録中の患者の訴えから病名や症状を抽出する症状抽出システムの性能向上：他施設データ追加学習による効果

【第一章】医師記録を学習済みの病名や症状を収集する自然言語処理システムの薬剤管理指導記録への適用

医師記録を用いて機械学習（fine-tuning）した病名抽出システムであり、病名や症状の NER および文脈からその発現の有無を判定（肯定否定判定）できる MedNER-J を薬剤管理指導記録に適用できるか検証し、機械学習ベースの薬剤管理指導記録分析のベースラインの性能を提示することとした。

【結果および考察】

本検討は慶應義塾大学医学部倫理委員会（承認番号：20200067）より承認を受けて実施した。

2018 年 4 月～ 2019 年 3 月に慶應義塾大学病院でセファゾリンナトリウムを投与された患者 49 名、60 件の SOAP 形式の薬剤管理指導記録中の S 情報、Objective (O) 情報、Assessment (A) 情報、Plan (P) 情報に MedNER-J を適用し、病名や症状の NER、抽出した症状の発現の有無を判定する肯定否定判定を行った。研究者は目視で同様の作業を行い、NER の性能と NER および肯定否定判定の性能を、適合率、再現率、および両者の調和平均である F1-score を用いて評価した。

その結果、S、O、A、P 情報でそれぞれ、NER の F1-score は 0.46、0.70、0.76、0.35、肯定否定判定を含めた F1-score は 0.28、0.39、0.64、0.077 であった。

O、A 情報からの NER の F1-score はそれぞれ 0.70、0.76 であり、S、P 情報の F1-score (0.46、0.35) よりも高かった。これは O、A 情報に MedNER-J の教師データである症例報告と同様に専門用語の病名・症状が多く含まれることによると考えられた。一方、肯定否定判定を含む F1-score が A 情報のみで高かった (F1-score=0.64) のは、A 情報の記載様式と記載内容が、症例報告と類似することに起因すると考えられた。S 情報では、MedNER-J の不抽出によるエラーが最も多く、動詞・形容詞・副詞

のエラーが最多であった。このことから、MedNER-J は患者が用いるような一般用語、口語体の表現の分析に適さないことが分かった。

MedNER-J は薬剤管理指導記録から NER および肯定否定判定を実施でき、特に A 情報において医師記録の分析に近い性能を示した。一方で、A 情報以外の記録の分析には課題があり、MedNER-J を薬剤管理指導記録に適用するためには、S 情報を中心に教師データを増強する必要があることが判明した。

【第二章】医師記録を学習済みの自然言語処理システムを用いたバンコマイシン塩酸塩による急性アレルギー性皮膚症状の症例検出

前章にて MedNER-J は A 情報以外の分析に課題があることが示されたが、A 情報には比較的高い性能を示したことから多様な医療記録の分析に利用できる可能性がある。また、医療テキストから病名・症状を収集できる病名抽出システムを活用することにより、有害事象モニタリングの自動化が期待される。そこで、実際にバンコマイシン塩酸塩 (VCM) によるレッドネック症候群 (RMS) が一時的に多発した事例をポジティブコントロールとすることとした。VCM を投与された患者の看護記録を MedNER-J による情報抽出とルールベースによるアレルギー症状の特定を組み合わせた手法で分析することにより、VCM による RMS を含めた急性アレルギー性皮膚症状 (アレルギー症状) を発現した症例を検出し、発現割合および発現が増大した期間とそれ以前の期間における発現割合の推移を特定できるか調査した。

【結果および考察】

本検討は慶應義塾大学医学部倫理委員会 (承認番号: 20221114) および慶應義塾大学薬学部 人を対象とする研究倫理委員会 (承 240228-4) より承認を受けて実施した。

2018 年 4 月 2 日～2018 年 6 月 29 日と RMS が通常より多く報告された 2019 年 3 月 25 日～2019 年 3 月 29 日の 2 つの期間に、慶應義塾大学病院にて VCM を点滴投与された患者の看護記録を MedNER-J に適用した結果、NER 単体、肯定否定判定を含む F1-score はそれぞれ 0.65, 0.60 であった。

また、研究者は論文、ICD-10、ICH 国際医薬用語集を参考にアレルギー症状に該当する症状を一覧とした用語リストを作成した。MedNER-J により発現があった症状を収集し、用語リスト内の症状と完全一致または、抽出語が用語リスト中の症状を完全に内包する形で部分一致した場合、アレルギー症状と判定するルールベースの手法により、アレルギー症例の検出を試みた結果、症例検出性能は F1-score = 0.81-0.82 であった。さらに 2 つの期間におけるアレルギー症例の発現割合を確認したところ、2 つの期間における症状発現割合の変動は MedNER-J とルールベースを組み合わせた方法と研究者による目視確認のいずれでも 0.05 と等しい値であった。

MedNER-J による看護記録分析の性能は、期間による差はほぼなく、先行研究と概

ね同等、または先行研究を上回った。さらに、アレルギー症例検出性能は高く (F1-score = 0.82、0.81)、研究者による目視と同等の発現割合の変動を検出することができた。以上より、VCM によるアレルギー症状のモニタリングに提案手法が利用できる可能性が示唆された。

【第三章】薬剤師記録中の患者の訴えから病名や症状を抽出する症状抽出システムの性能向上：学習量増加による効果

前章にて NER システムの有害事象モニタリングへの利用可能性が部分的に示されたが、有害事象をより早期に特定するためには、S 情報の分析が不可欠である。しかし、MedNER-J の S 情報分析性能が低かったことから、新たに S 情報を教師データとすることで S 情報を分析対象として病名や症状の NER および肯定否定判定を行う症状抽出システムの構築を目指した。システムをより高性能なものにするため、システムの学習量を段階的に増やし、これに伴う性能の推移を明らかにすることで学習量増加がシステムの性能に与える影響を調査した。

【結果および考察】

本検討は慶應義塾大学医学部倫理委員会（承認番号：20200067）より承認を受けて実施した。

2018 年度の慶應義塾大学病院における薬剤管理指導記録に含まれる S 情報 6,559 症例、12,004 件を情報源として教師データを作成し、10 段階にかけて教師データを増量しながら 10 分割交差検証により、機械学習モデルである BERT-CRF (bidirectional encoder representations from transformers-conditional random field) を fine-tuning した結果、システムの F1-score は 0.67 から 0.82 へと上昇した。学習量 3,600 件で F1-score は 0.78 となり、以降、学習量に対する性能向上の効率が減少した。性能の向上に伴い、システムの誤抽出によるエラーが最も減少し、適合率の上昇に寄与した。また、汎用性評価のために構築したシステムを症例報告に適用した結果、学習量 1,200 件から 12,004 件にかけて、症例報告分析の F1-score は 0.34 から 0.41 へ増加した。さらに NER が部分一致の場合を含めた性能をレーベンシュタイン距離の類似度を判定基準として評価した結果、S 情報に対する F1-score は 0.70 から 0.84、症例報告に対する F1-score は 0.38 から 0.46 へ向上した。いずれの F1-score も完全一致のみの場合の F1-score より高い値を示したものの、その差は大きくなかった。

学習量 3,600 件以降、学習量の増大に伴う性能の向上には飽和傾向がみられたものの、大量のデータを fine-tuning することにより、薬剤管理指導記録の S 情報を分析対象とした高性能な症状抽出システムを構築することができた。また、学習量増加により、特にシステムの誤抽出を改善できた。

【第四章】薬剤師記録中の患者の訴えから病名や症状を抽出する症状抽出システムの性能向上：他施設データ追加学習による効果

前章では単施設のデータを教師データとしたが、施設によって患者が有する疾患や治療の侵襲度は異なり、S 情報に違いが見られると考えられた。そのため、薬局を含む他施設においてシステムは異なる性能を示す可能性がある。ある施設でシステムを導入したい場合、既に構築したモデルをベースにすることは自然な考えである。その場合、既存モデルをそのまま利用できる、あるいは、新たに作成する教師データを減らせる可能性がある。理想的には新しいモデルがどの施設に対しても高いパフォーマンスを発揮することが期待される。しかし、これまで、追加の **fine-tuning (2nd fine-tuning)** が NER を行うシステムの性能に与える純粋な影響は不明であった。そこで、病院の薬剤管理指導記録中の S 情報で **fine-tuning** したモデルに、薬局薬歴中の S 情報を **2nd fine-tuning** し、その効果を明らかにすることで、他施設への応用を踏まえた効率的なシステムの性能向上方法を検討した。

【結果および考察】

本検討は慶應義塾大学医学部倫理委員会（承認番号：20200067）および慶應義塾大学薬学部 人を対象とする研究倫理委員会（承 240228-3）より承認を受けて実施した。

病院の薬剤管理指導記録に含まれる S 情報 2,794 症例、3,600 件（病院データ）のみを **fine-tuning** したモデル（ベースモデル）に薬局薬歴に含まれる S 情報 2,250 症例、2,400 件（薬局データ）を **2nd fine-tuning** したモデル（追加学習モデル）を構築し、病院データと薬局データに対する性能を評価した結果、追加学習モデルはベースモデルと比較して、薬局データに対する性能がわずかに上昇した。（追加学習モデルの **F1-score = 0.78**、ベースモデルの **F1-score = 0.74**）一方、病院データに対する性能は大幅に低下した（追加学習モデルの **F1-score = 0.62**、ベースモデルの **F1-score = 0.78**）。追加学習モデルの薬局データに対する性能は、追加学習量が一定量に達するまでベースモデルの性能を下回った。また、一般的にシステムは最後に **fine-tuning** した内容に重みづけされるため、**fine-tuning** の順序がどの程度システムの性能に影響するか調査した。新たに薬局データ、病院データの順に **fine-tuning** したモデル（逆順モデル）を構築し、ベースモデルおよび追加学習モデルの性能と比較した。逆順モデルは追加学習モデルでなくベースモデルに近い性能を示したことから、システムの性能は最後に **fine-tuning** した教師データに大きく依存することが示された（病院データに対する **F1-score = 0.77**、薬局データに対する **F1-score = 0.76**）。また、病院データと薬局データに含まれる症状の数を比較した結果、記録 1 件あたりの症状の数は病院データが多かった（病院データ 1.77、薬局データ 1.43）一方、100 字あたりに含まれる症状の数は薬局データの方が多く（病院データ 2.44、薬局データ 3.43）、病院データの方が複雑

な文章から構成されることが分かった。このことから、複雑な病院データを **fine-tuning** することで、複雑な病院データと比較的簡潔な薬局データの双方に対応できるシステムとなったと推察された。

病院データのみを用いて **fine-tuning** したベースモデルは薬局データを **2nd fine-tuning** した追加学習モデルと比較して、双方の施設のデータに対して安定して高い性能であった。このことから、複数施設のデータを混合せず、順番に教師データとして用いる場合、システムの性能は教師データの施設数でなく、最後に用いた教師データの内容と学習量に依存する可能性が示唆された。

【総括】

本研究では、既存の医師記録を教師データとしたシステムの薬剤師記録分析性能を調査し、薬剤師記録分析のベースラインを明らかにした。また、実際に薬剤師記録の S 情報を教師データとして、患者の発した情報から多様な症状を高性能に抽出するシステムを構築し、教師データの増量と他施設データの **2nd fine-tuning** による学習効果を初めて報告した。さらに、既存システムの看護記録からの有害事象抽出への利用可能性を提示した。このことから、本研究で構築した S 情報に特化したシステムを用いて S 情報を分析することで、今後、患者の発した情報を元にした報告バイアスの少ない有害事象モニタリングの実現が期待される。

また、本研究では医療従事者が作成した記録を分析対象としてきたが、記録を作成する時点で医療従事者により情報が取捨選択され、報告バイアスが生じている可能性がある。そこで、今後は患者の発話情報を自動で書き起こし、収集されたデータを分析対象とすることで、報告バイアスを最小化したリアルタイムでの有害事象モニタリングが期待される。このような取り組みを通して本システムが患者の健康増進に寄与することが期待される。

論文審査結果の要旨

1. 審査概要

パワーポイントを用いたプレゼンテーションの後 (25 分)、副査との質疑応答を行った (49 分)。各副査からの質問に対して適切に応答が行われた。主な質疑応答内容は以下の通りである。

【堀里子教授による質問】

S 情報から抽出されるエンティティは多様であるため、正規化しないと有害事象モニタリングはできないのではないかと問われ、指摘の点についても検討を行ったこと、す

なわちこれまでに教師データを作成する際にアノテーションした症状を収集し、人力でコーディングを行うことで、標準化のための用語集の作成を検討していること、システムと用語集を組み合わせて使用することでモニタリングに使用できると考えていることを回答した。

第一章で使用したデータは限局的であったが、その結果から機械学習モデルを用いた薬剤師記録分析のベースラインを示したと言えるのかと問われ、これ以前に機械学習モデルを用いて薬剤師記録を分析されたことはなく、初めて性能を評価したという点からベースラインと言えることを回答した。

学位論文の表題に含まれる「患者の発話」とは何を意図したものであるか問われ、「S 情報」を意図したものであることを回答した。それに対して、「S 情報」は表題に含まれる「医療記録」の一部であるため、「患者の発話」と「医療記録」を並列に記すことの適切性について指摘があった。意味の混同を防ぐため、発表内で本研究における患者の発話の定義を明示する、または、論文の表題を変更するといった対応を検討することを回答した。

部分一致の定義について、第一章では 1 文字以上の一致を、第三章ではレーベンシュタイン距離の類似度が一定値を超えた一致を基準として用いていたが、部分一致の中に意味のない抽出が含まれたのではないかと問われた。レーベンシュタイン距離の類似度については、予備調査の結果を参考に意味のない抽出が増える直前を基準値としたため、その懸念は少ないが、1 文字以上の一致については、意味のない抽出がどの程度発生したか解析していなかったため、追加で解析することを回答した。

第二章で比較対象として最新の先行研究を使用しなかったのは研究内で用いている評価指標が異なるためであるか問われ、第二章の内容を初めて発表した学会時点のデータを使用していたため最新の先行研究でなかったこと、確認の上、比較が可能であれば最新の先行研究に差し替えることを回答した。

【漆原尚巳教授による質問】

第一章、第三章、第四章のエラー分析にて、適合率・再現率・F1-score があれば十分であるにもかかわらずエラー率を算出した意義、エラー率を算出する際に分母に True Negative を含めておらず算出方法が適切であったか、True Negative を含めないことにより考えられるバイアスについて問われた。本研究では、4 種類のエラーを一度に確認したかったため、エラー率を算出したこと、True Negative は数値化が困難であるため、True Positive, False Positive, False Negative の和を分母としたことを回答した。

レーベンシュタイン距離の類似度の設定に AUC (ROC 曲線) を用いなかった理由を問われた。申請者は ROC 曲線を用いた検討も行っており、少なくとも抽出語の文字数の半分以上の部分一致を望ましく考えていたところ、ROC 曲線ではそれを下回る類

似度が指定されてしまったため、別の方法により類似度を設定したことを回答した。

第一章、第二章の分析対象データの条件に含まれるセファゾリン（第一章）とバンコマイシン（第二章）投与後の日数の設定根拠を問われた。第一章では、薬剤管理指導記録を使用しており、薬剤管理料算定の上限回数を参考に、記録が最低1度は作成される期間を設定したことを回答した。また、第二章では、毎日記録が作成されると考えられる看護記録を分析対象としたことに加え、急性症状が検出対象であることから投与後すぐの発現が想定されるが、夜間に投与して翌日に記録が作成される可能性を考慮し、投与当日及び翌日の2日間を対象としたことを回答した。

本研究においては、医薬品安全性情報のスクリーニングに有用なツールの構築を目指しているが、実際に抽出されているのは(傷)病名であり、薬剤投与との因果関係、関連性については検証されていないのではないかと問われ、指摘の通りであり、この点について論文を修正する旨、回答した。

本研究の中で MedDRA を集計に使用しているか問われ、第一章の研究者の判定基準に MedDRA を参考にする旨は記載しているが、参考にただけで集計は行っていないことを回答した。

【判定結果】

質疑応答ののち当該学生を退室させ、主査・副査による可否判定が行われた。判定会議では、本研究に対する当該学生の理解度、論理的思考および研究遂行能力、さらに本研究内容の新規性ならびに有用性について議論を行った。審査委員会は、申請者の研究は博士（薬学）の学位に相当する内容を含むと考えたが、論文題目については再考の必要があること、論文において結果の重要な部分が十分に示されていないことなどが指摘された。このため、申請者に対して、論文題目の再考を求めるとともに、研究内容を仔細に考察するに十分な結果や具体的なデータを論文に掲載するよう求めることとした。また質疑応答で不明確であった部分、すなわち第一章において一文字以上の部分一致を用いたことによる不適切な抽出の程度については、追加で解析を実施させることとした。また、医薬品投与との関連を考慮した有害事象の特定に応用するにはまだ課題があることなども考慮し、審査時の指摘に沿って論文本文及び発表スライドの適切な修正等を求めることとした。

以上の審査委員会における指摘事項を踏まえて、申請者は論文題目を「自然言語処理技術を用いた医療記録および患者の発話からの症状抽出方法に関する研究」から「自然言語処理技術を用いた医療記録からの症状抽出方法に関する研究～患者の主観的情報に焦点を当てて～」に修正した。さらに、論文への詳細な結果やデータの加筆、内容及び発表スライドの修正を行なった。続けて申請者は、令和7年2月18日に実施された学位論文公聴会において、論文の内容を発表した。

発表後、研究科委員より、本研究で構築したシステムによって有害事象症状の重症度判定や有害事象と原疾患の症状の判別が可能であるか、生成 AI との差異および本システムの強み、教師データの選択方法によりシステムの性能へどのような影響があるか、追加学習によりシステムの性能が低下する理由、などについて質問がなされ、申請者はいずれの質問に対しても適切に回答した。

以上の過程を通して、学位論文には科学的に新規かつ独創的な内容を含んでいること、申請者は専門領域および関連領域における十分な知識を有していることが確認されたことから、申請者は、博士（薬学）の学位を授与されるにふさわしいと判断する。

論文目録

1. **Yukiko Ohno**, Riri Kato, Haruki Ishikawa, Tomohiro Nishiyama, Minae Isawa, Mayumi Mochizuki, Eiji Aramaki, Tohru Aomori. Using the Natural Language Processing System Medical Named Entity Recognition-Japanese to Analyze Pharmaceutical Care Records: Natural Language Processing Analysis. *JMIR Formative Research*. 2024; **8**: e55798.
2. **Yukiko Ohno**, Tohru Aomori, Tomohiro Nishiyama, Riri Kato, Reina Fujiki, Haruki Ishikawa, Keisuke Kiyomiya, Minae Isawa, Mayumi Mochizuki, Eiji Aramaki, Hisakazu Ohtani. Performance Improvement of the Natural Language Processing Tool to Extract Patient Narratives Related to Medical States From Japanese Pharmaceutical Care Records by Increasing the Amount of Training Data: Natural Language Processing Analysis. *JMIR Medical Informatics*. 2025; in press.