別添4

令和元年度　厚生労働科学研究費　補助金
（政策科学総合研究事業（臨床研究等ICT基盤構築・人工知能実装研究事業）
（分担）研究報告書

症例報告書からの副作用に特徴的な素性の自動抽出および抽出された素性に基づく副作用の自動評価

研究分担者　潮田　明　国立研究開発法人産業技術総合研究所・人工知能研究センター・招聘研究員

研究要旨

【目的】今年度は症例報告書からの副作用に特徴的な素性（特徴量）の自動抽出の枠組みを構築し、昨年度開発した素性・素性値ペアのセットから副作用の自動判定を行うモジュールと統合して、全行程を通じてAIにより副作用判定を行うことのできるモデルを構築し、判定精度の試行的評価を行う。またエラー分析を通じて、特徴量改良のための調査を行う。

【方法】1410件のSJSの副作用症例報告書および200件の中毒性表皮壊死融解症（TEN）の副作用症例報告書に対して副作用に特徴的な素性（特徴量）のアノテーションを施し、素性抽出のためのトレーニングデータとした。アノテーションされたデータのうち、昨年度副作用自動判定の評価に用いたSJSの副作用症例報告書200件についてはテスト用データとして用い、残りの1410件のうち、981件を学習用データとして、429件をバリデーションデータとして用いた。またBertを用いて素性アノテーションのための機械学習モデルを作成した。Bert用の日本語pretrainedモデルは医療・医薬品分野のテキスト約6千万文を用いて構築した。作成した素性アノテーションモデルと、素性・素性値ペアのセットから副作用の自動判定を行うモジュールとを統合して、全行程を通じてAIにより副作用判定を行うことのできるモデルを構築して、判定精度の試行的評価を行った。

【結果】Maximum Entropy Classifier(MEC)を学習器として用いた評価において81.5%の副作用判定精度が得られた。専門家の精度領域に近づいたことで手法の有効性が確認できた。

【考察】現場でAIを活用しようとした場合、81.5%の精度ではまだ安心して使えるレベルだとは言えない。しかしながら全行程をAIが行った場合でもAIが負例と判定した症例の場合、AIの確信度が上位37%の症例に関してはすべてAIが正解判定を行っている。従って負例判定症例の場合、AIを用いた評価対象の自動絞り込みによる作業効率化の可能性が示されたと言える。また今回素性アノテーションに用いたトレーニングデータに対しては確認や検証はまだ行っておらず、データにはフォーマットエラーを含む文が多く含まれているため、トレーニングデータのクリーニングが、副作用自動判定の精度向上のための有効な手段と考えられる。素性の改良に関しては、人手による判定の難しい症例におけるエラー分析を通じて、現状の素性ではまだ捉えられていない要素が浮き彫りになった。特に何が記載されているかと同時にどのように記載されているか、すなわち「書きぶり」も大きな要素であることが分かり、素性の改良のための指針として活用が期待される。

A．研究目的

副作用を迅速かつ客観的に評価するための人工知能を活用した副作用症例報告評価技術の開発を本研究課題の最終目標として、本分担研究においては、人工知能による汎用性が高い副作用評価を行うための基盤整備として症例報告書からの自動アノテーションを活用した副作用に特徴的な素性（特徴量）の抽出を行う前半部と、得られた素性と素性値から機械学習により副作用判定を行う後半部に分けて研究を行っている。昨年度は当初前半部を遂行する予定であったが、個人情報管理の実務的制約が厳しく、副作用症例報告書そのものの内容を扱う前半部よりも個人情報保護への配慮がより軽減される後者の方を先に進めた方がより効率的に研究を進められると判断し、副作用評価のための機械学習用トレーニングデータの作成およびトレーニングデータを用いた試行的副作用評価を先行して実施した。その結果副作用判定精度に関しては、Maximum Entropy Classifier(MEC)を学習器として用いた評価において86.0%の精度を達成した。

今年度はまず前半部に相当する、症例報告書からの副作用に特徴的な素性（特徴量）の自動抽出の枠組みを構築し、次いで前半部と後半部を結合して全行程を AI で学習・推論できるモデルを構築し、試行的精度評価を行った。また、副作用判定の正解データの作成を行った PMDA の専門委員とともに行った AI によるに副作用判定のエラー分析を通じて、今回機械学習用の素性として選択した副作用自動判定のための特徴量について改良のための調査を行った。

B.研究方法

PMDA において副作用評価を行っている複数の専門家からのヒアリング結果および「重篤副作用疾患別対応マニュアル」をもとに昨年度作成した、副作用自動判定のための機械学習用の素性は表１に示す通りである。昨年度の取り組みにおいては、副作用症例報告書から人手により素性と素性値を抽出し、その結果を用いて AI による副作用判定のためのトレーニングと推論を行った。今年度はこれらの素性の抽出も機械学習を用いて行うために、素性抽出のためのトレーニングデータの作成およびトレーニングデータを用いた素性抽出のモデルを構築した。

具体的には、表１の各素性のうち、素性番号１～５は副作用症例報告書の自由記載のテキストから抽出しなければならないため、素性番号１～５について素性抽出のためのトレーニングデータの作成を行った。素性番号６と７については、今回の研究利用に先立って行った副作用症例報告書の匿名化の過程で検査結果がマスキングされているために、テキストから抽出することができない。しかしながら、副作用症例報告書にはテキスト欄とは別に DSLT などの検査結果を表形式で記載する欄があるため、そこから抽出することが可能である。そこで今回はマスキングされた副作用症例報告書とは独立に管理された検査結果表から検査結果（陰性／陽性）を抽出し、素性番号１～５の抽出結果と足し合わせて機械学習に用いることとした。素性番号８に関しては昨年度の調査において、MEC で学習されたモデルにおけるそれぞれの素性に対応する重みを比較してどの素性が副作用判定においてより重要であるかを調査した結果９つの素性の中で最低であったこと、また実際に今年

度の予備実験により素性番号８は除外した方が副作用判定精度が高くなるという結果が得られたため、除外することとした。一方素性番号９は上述の素性の重要性の比較において 2 番目に高い位置にあるため、採用した。表 2 に今回アノテーションの対象とした素性のセットを示す。

素性抽出のためのトレーニングデータの作成

素性抽出のためのトレーニングデータは、形態素解析を施した副作用症例報告書のテキストに対して副作用に特徴的な素性（特徴量）のアノテーションを施すことにより作成した。アノテーション作業は大学病院における約 10 年の臨床経験のある正看護師 1 名が行った。図 1 にアノテーションの 1 例と簡単な手順を示す。

アノテーターが記号に煩わされずにテキストの表現に集中できるように、アノテーション作業は 2 段階に分けて行った。まずテキストに対して、副作用判定において重要な用語（主に単語）の、素性に応じた色付けを行った(a)。例えば、「顔面の紅斑」という表現があった場合、これは皮膚の汎発性の紅斑（やや軽度）を表す表現であるので、対応する色（ここでは水色）で重要な用語を色付けする。「顔面」は皮膚の比較的広範囲の領域を指す用語であるので色付けの対象となる。

一旦このような色付けを行った後に、色付け結果を参照しながら、形態素解析を施したテキストに対して形態素ごとに IOB2 フォーマットを用いた素性のアノテーションを行った(b)。

アノテーションは、昨年度副作用自動判定の評価に用いた SJS の副作用症例報告書 200 件を含む 1410 件の SJS の副作用症例報告書および 200 件の中毒性表皮壊死融解症（TEN）の副作用症例報告書に対して行った。TEN の副作用症例報告書においても素性番号１～５の素性に相当する重要表現は頻出するため、トレーニングデータとして活用することにした。

アノテーションされたデータのうち、昨年度副作用自動判定の評価に用いた SJS の副作用症例報告書 200 件についてはテスト用データとして用い、残りの 1410 件のうち、981 件を学習用データとして、429 件をバリデーションデータとして用いた。

素性アノテーション用学習モデル

素性アノテーション用学習モデルには Bert を用いた。素性アノテーションはテキストの各要素（ここでは形態素）に IOB2 のタグを付与するオペレーションであるので様々な自然言語用タグ付けモデルが活用可能であるが、その中で Bert は現時点で自然言語の様々な分野において最も高い精度が報告され、また個別の NLP タスクにファインチューニング可能なモデルである。

Bert を素性アノテーションに用いるためには、アノテーション用トレーニングデータの他に Bert 用の日本語 pretrained モデルが必要である。京都大学の黒橋・村脇研究室が公開している日本語 pretrained モデルは研究用にも広く活用されているが、モデル構築用のテキストに Wikipedia（約 1,800 万文）が用いられており、副作用症例報告書とは分野も語彙も大きく異なると考えられる。そこで、SJS に関連する論文を含む広範囲にわたる医学雑誌の論文、JAPIC が提供する医薬品添付文書および独自開発の自動検索装置により WEB から

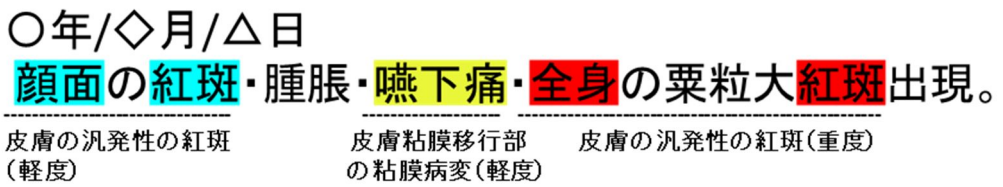
表１　昨年度の評価実験で用いた素性のセット

素性番号	素性	素性値	判断基準
1	皮膚粘膜移行部の広範かつ 重度な粘膜病変	2	有り
		1	有り（やや軽度）
		0	なし
2	皮膚の汎発性の紅斑に伴う表皮の 壊死性障害に基づくびらん・水疱	2	有り
		1	有り（やや軽度）
		0	なし
3	発熱	1	有り
		0	なし
4	病理組織所見	1	有り
		0	それ以外
5	皮膚科によるSJS診断	1	有り
		0	なし
6	被疑薬のDSLT検査	1	検査有り（陽性）
		0	検査有り（陰性）
		-	検査なし
7	併用薬のDSLT検査	1	検査有り（陽性）
		0	検査有り（陰性）
		-	検査なし
8	報告書テキスト内に被疑薬投与後の症状発現が記載	1	有り
		0	なし
9	報告書テキスト外テーブルより 被疑薬投与後の症状発現と判断	1	できる
		0	できない

表２ アノテーションの対象となる素性のセット

素性番号	素性	素性値（TAG）	判断基準
1	A 皮膚粘膜移行部の広範かつ 重度な粘膜病変	2 (AA2)	有り
		1 (AA1)	有り（やや軽度）
		0	なし
2	B 皮膚の汎発性の紅斑に伴う表皮の 壊死性障害に基づくびらん・水疱	2 (BB2)	有り
		1 (BB1)	有り（やや軽度）
		0	なし
3	C 発熱	1 (CC)	有り
		0	なし
4	S 病理組織所見	1 (SS)	有り
		0	それ以外
5	H 皮膚科によるSJS診断	1 (HH)	有り
		0	なし

a) 素性に応じてテキストを色付け



b) IOB2 フォーマットを用いた素性アノテーション

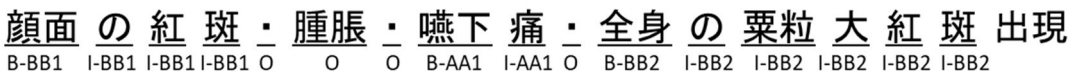


図１　IOB2 フォーマットを用いた素性アノテーションの例

収集された医療分野のテキストからなる総計 6 千万文の日本文テキストを用いて Bert 用 pretrained モデルを構築して用いた。

全行程を通じた AI による副作用自動判定

学習済み Bert モデルを用いて 200 件のテスト用副作用症例報告書に素性アノテーションを施し、その結果を用いて昨年度と同じ副作用自動判定の評価を行った。昨年度との違いは、昨年度は副作用判定用の素性と素性値が人手で付与されたのに対して、今回は機械学習により自動付与された素性と素性値を用いる点にある。Bert により識別される素性・素性値のタグはテキスト中の各文に対して付与されるものであるが、症例報告書中の文に付与された素性・素性値はそのまま症例報告書に付与されたものとして扱うこととした。

C. 研究結果

自動アノテーションおよび副作用自動判定の精度を表 3 に示す。素性のアノテーションでは素性番号１～５に対応するタグの他に、どの素性にも対応しない形態素に付与する 0-tag(0ther)を用いている。0-tag を含むすべてのタグ付けのテストデータにおる精度は 96.5%、0-tag を除いた精度は 69.4%であった。 またこの自動アノテーションの結果をもとに副作用自動判定を行った結果 81.5%の判定精度が得られた。判定精度評価は昨年度と同様に MEC を用いて 5 分割交差検証により行った。素性を人手で付与した場合の副作用自動判定の精度(86.0%)からは 4.5 ポイント低下しているが、当初目標としていた 80%の精度を上回る結果が得られた。

評価結果の分析

自動アノテーションの結果をもとに行った副作用自動判定において、専門員により副作用として認められた報告書（正例）に対する自動判定精度は 85%、副作用と認めるのに十分な所見あるいは情報がないと判断された報告書（負例）に対する自動判定精度は 78%であった。85%と

78%はそれぞれ感度と特異度に該当する。すなわち、判定エラーに関しては、false negative よりも false positive の方が多いことが分かる。素性を人手で付与した場合の副作用自動判定の感度と特異度はともに 86%であることから、false positive の方がより多く増加した原因は素性の自動アノテーションのエラーにあると考えられる。

素性アノテーションのエラー分析

素性アノテーションのトレーニングデータは６０万形態素以上にも及ぶが、一人のアノテーターの作業結果に対して確認や検証はまだ行っていない。しかしながら、テストデータ中の 100 件の正例症例についてアノテーションのフォーマットチェックを行ったところ、100 件中 13 件の症例報告書についてフォーマットエラーを起こす文が含まれていることが分かった。フォーマットエラーとは IOB2 フォーマットの規約から外れたアノテーションを指し、例えばエンティティの先頭は B-の付くタグで始めなければならないが、いきなり I-SS などのタグで始まるケースなどを指す。データを機械学習にかける際にはフォーマットエラーを起こす文は自動的に除外している。文中の素性はすべてその文を含む症例報告書の素性として扱われるため、１つの文のフォーマットエラーは症例報告書の素性の欠損に直結する可能性が高い。従って、素性アノテーション用トレーニングデータのクリーニングは、副作用自動判定の精度向上のための有効な手段と考えられる。

表 3 自動アノテーションおよび副作用自動判定の精度

素性のアノテーション	テストデータの アノテーション精度 (%)
全形態素の精度	96.5
O-tag 付与語を除く全形態素の精度	69.4

素性のアノテーション	テストデータの副作用自動 判定の精度 (%)
人手によるアノテーション	86.0
AI によるアノテーション	81.5

D. 考察

全行程を通じたAIによる副作用自動判定により81.5%の判定精度を達成した。3名の専門員の判定精度の概算値が80%～90%であることから、専門家の精度領域に近づいたと言える。しかしながら現場でAIを活用しようとした場合、81.5%の精度ではまだ安心して使えるレベルだとは言えない。そこで、AIを現場で役立てて行くための方法の１つとして、AIの確信度、すなわち「AIが自分自身の判定結果に付与する確率」の活用が考えられる。

図2はAIの確信度と負例判定症例の判定精度の関係を示したものである。横軸はAIが負例と判定した症例について確信度の高い順に症例を並べて、特定の確信度以上の症例を選択したときの、選択された症例の数の割合を示し、縦軸は選択された症例に対するAIの判定精度を示す。青線は素性を人手で付与した場合、オレンジ線は全行程をAIが行った場合のグラフである。後者の精度は前者よりやや減少しているが、全行程をAIが行った場合でもAIの確信度が上位37%の症例に関してはすべてAIが正解判定を行っている。従って負例判定症例の場合、AIを用いた評価対象の自動絞り込みによる作業効率化の可能性が示されたと言える。

一方図3はAIの確信度と正例判定症例の判定精度の関係を示したものである。このケースでは全行程をAIが行った場合AIの確信度が上位10%以内であってもAIの判定精度が70%を下回る場合があり、このままでは評価対象の自動絞り込みは難しい。ここからもfalse positive改善の課題が浮かび上がってくる。

副作用判定の正解データの作成を行ったPMDAの専門委員とともに、今回機械学習用の素性として選択した副作用自動判定のための特徴量について改良のための調査を行った。3名の専門員による正解データの作成において、当初作成した副作用自動判定のための素性セットの素性の多くについて高い素性値が示されているにも関わらず3名とも負例と判断するケースが散見された。そのようなケースにおいてまだ捉え切れていない特徴量として何が挙げられるか考察を行った。その結果、正例負例の判定を行う上で、何が記載されているかと同時にどのように記載されているか、すなわち「書きぶり」も大きな要素であることが分かった。こういった要素は今回用いた素性ではほとんど捉えられていない。図４に特に判定の難しい症例におけるエラー分析を通じて浮かび上がった素性セット改良のためのヒントを頻度とともに示す。最も頻度が高かったのは、症例報告書中に主要な文言は揃っているものの、病状の記述に具体性が欠けているために信頼性が低いと判断される症例報告の事例である。また、病状の経過が時系列で追えるケースや薬剤投与量と投与期間が明確に記載されているケースは信頼性が高いと判断される傾向にある。また副作用症例報告書の中には報告者が直接関わったものではなく、文献からピックアップしたと思われる事例の報告もあり、そのような場合多くの重要な素性が含まれていても負例と判断されることがある。

図４に記された項目の多くは単純なパターンマッチングでは捉えられず、それぞれ個別に自動識別手法を開発していく必要がある。

副作用自動判定のための特徴量の改良

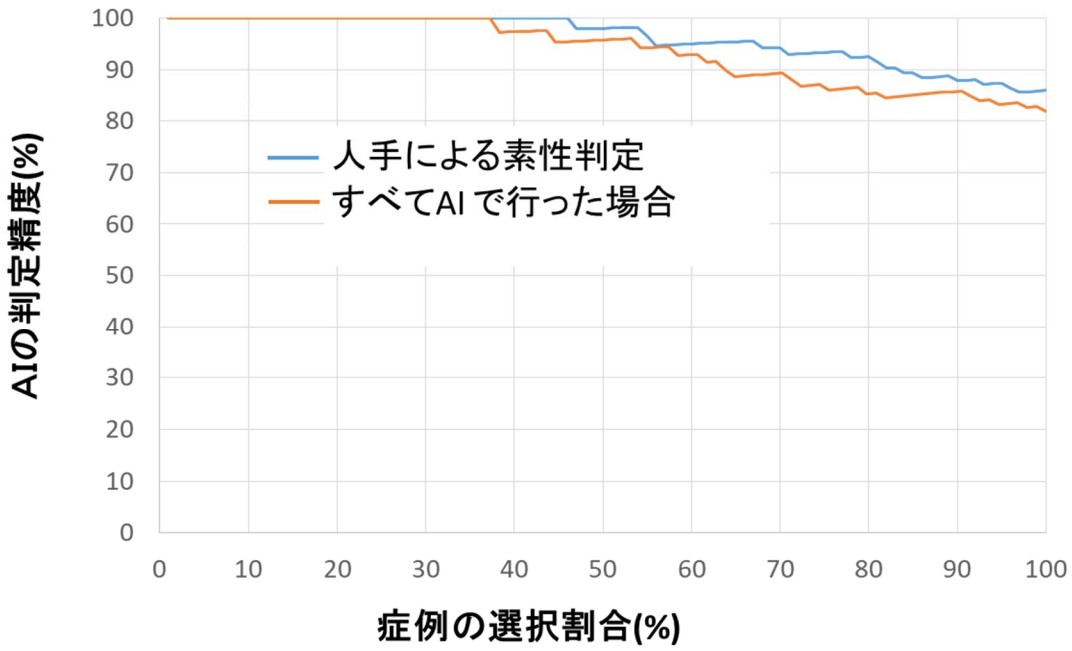


図 2　確信度上位の負例判定症例の判定精度

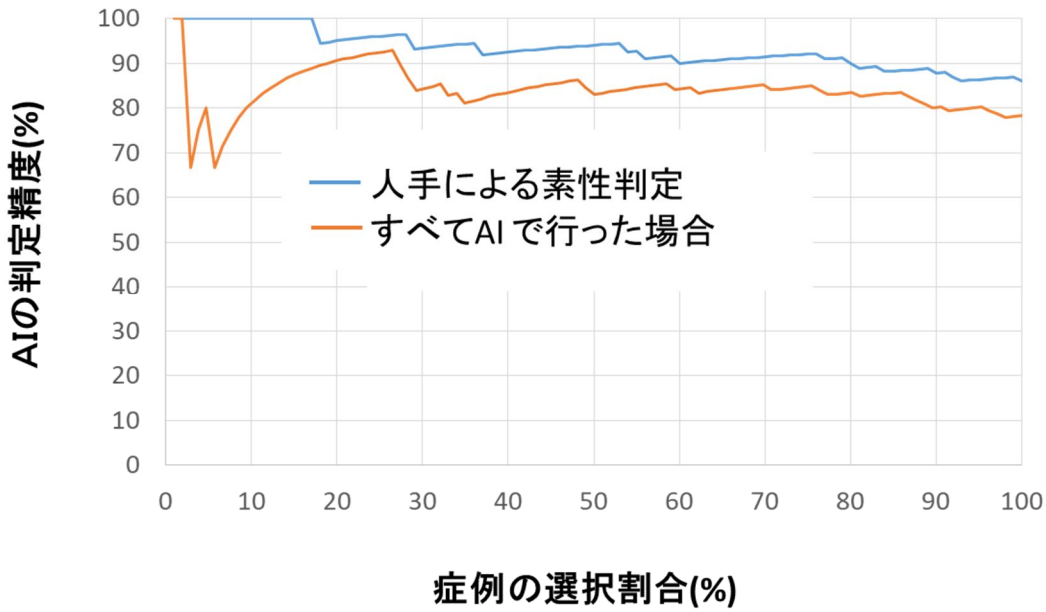


図 3　確信度上位の正例判定症例の判定精度

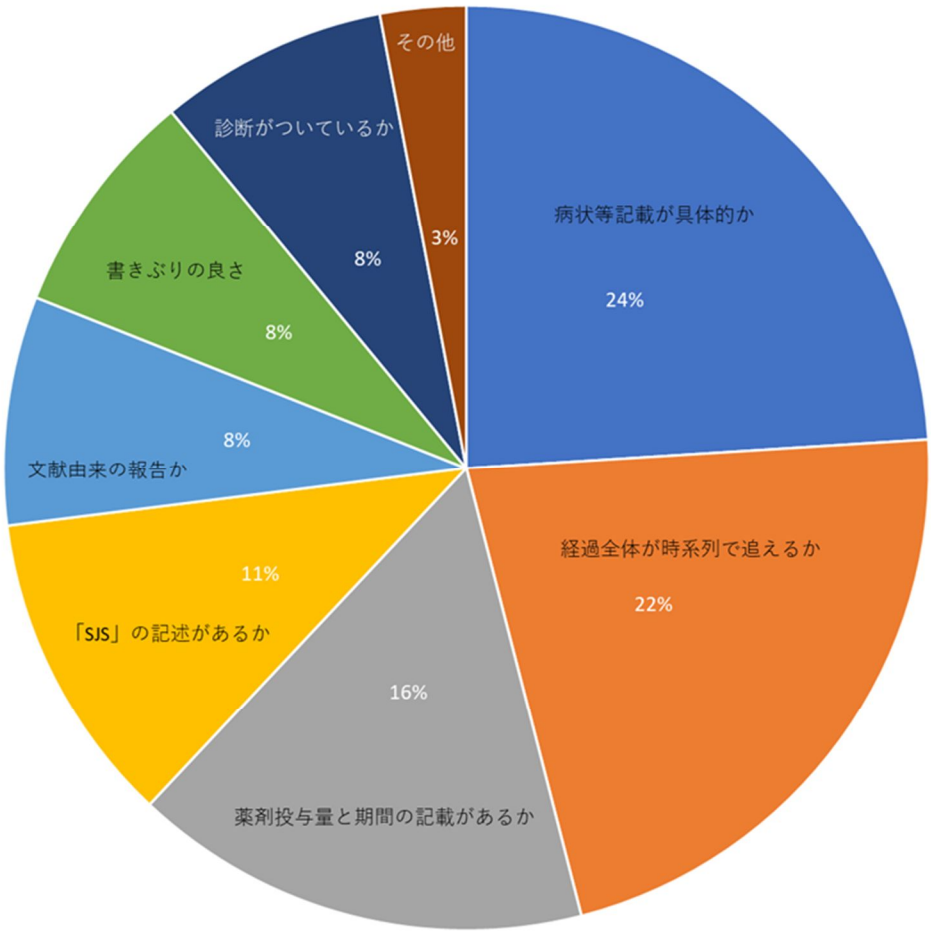


図４　素性セット改良のためのヒント

E．結論

症例報告書からの副作用に特徴的な素性（特徴量）の自動抽出の枠組みを構築し、昨年度開発した素性・素性値ペアのセットから副作用の自動判定を行うモジュールと統合して、全行程を通じてAIにより副作用判定を行うことのできるモデルを構築し、判定精度の試行的評価を行った。その結果当初目標の80%を上回る81.5%の判定精度を達成した。3名の専門員の判定精度の概算値が80%～90%であることから、専門家の精度領域に近づいたと言える。

本研究を通じて開発した素性抽出のためのトレーニングデータおよび副作用評価のための機械学習用トレーニングデータはいずれもSJSに特化したものであるが、機械学習用のモデルおよびNLPの要素技術は他の副作用にも適用可能であり、今後の拡充が期待される。

F．健康危険情報

該当なし

G．研究成果

1．論文発表

該当なし

2．学会発表

該当なし