平成２９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

副作用症例報告の評価支援のための自動アノテーションおよび辞書作成に関する研究

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】本研究は、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に先立ち、副作用症例報告テキスト中の重要情報を過不足・曖昧性なく機械に伝えるための自動アノテーション技術を開発することを目的としている。今年度は自動アノテーションを機械学習を用いて行うための学習用データの作成、および表記ゆれ解消のための機械学習用辞書の構築を目的としている。【方法】アノテーションには、日本語にも対応したテキストアノテーションツール「brat rapid annotation tool」を用いることとした。表記ゆれ解消のための辞書作成に関しては、本研究が対象とする皮膚障害に限らず副作用報告書全般に適用可能な表記ゆれ対策を視野に入れて、Web から収集した大量テキストから類似表現を検出する手法の検討を行った。【結果】現在 PMDA にて実施されている副作用評価に必要な情報として、医療用語（医薬品名、病名、症状名、医師による処置名など）、投与量、投与期間、投与頻度をはじめ、薬剤の種類と投与量の変化、患者の状態の変化などの「変化」や薬と症状との関係性などを中心にアノテーション対象を選定し、それらを configuration ファイルにまとめてアノテーション作業を開始した。辞書作成に関しては、標準的病名約２万５千語をベースにしたニューラルネットワークによる類義表現抽出実験を行った。Web から収集した標準病名を含む大量のテキストをもとに分散意味表現に基づく類似度評価器を作成した結果、評価器と人手によるチェックを組み合わせることで特定の医療関連用語とほぼ同義かあるいは置換可能なレベルで類似した用語を効率よく収集することが可能であることが分かった。【考察】今回試みた分散意味表現を用いる手法は、用語の内部構成ではなく用語のテキスト中の文脈の近さをもとに類似度を評価するものであり、同じ用語のカタカナ表記と漢字表記の類似性を検出できるなど大きな利点を有するものと考えられる。

A. 研究目的

今年度は、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に先立ち、副作用症例報告テキスト中の重要情報を過不足・曖昧性なく機械に伝えるための自動アノテーションを機械学習を用いて行うための学習用データの作成、および表記ゆれ解消のための機械学習用辞書の構築を目的とした。

B.研究方法

アノテーション作業

副作用症例報告の経過欄のテキストデータを対象にアノテーション作業を開始した。アノテーションには、日本語にも対応したテキストアノテーションツール「brat rapid annotation tool」を用いることとした。アノテーションの作業画面の例を図１に、アノテーション結果を格納したアノテーションファ

イルの例を図２に示した。アノテーションは、専門家がテキストから読み取れる重要な内容を過不足・曖昧性なく機械に伝えるために必須である。現在 PMDA にて実施されている副作用評価に必要な情報として、医療用語（医薬品名、病名、症状名、医師による処置名など)、投与量、投与期間、投与頻度をはじめ、薬剤の種類と投与量の変化、患者の状態の変化などの「変化」や薬と症状との関係性などを中心にアノテーション対象を選定し、それらを configuration ファイルにまとめてアノテーション作業を開始した。手動でのアノテーション結果は今後２つの用途で使用する。１つ目の用途は、アノテーションの自動化のための学習データである。テキスト中のそれぞれの用語に付与されたタグ（上記アノテーション対象に付与された医薬品名、病名などの名称）の付け方を用語の前後の文脈とともに機械学習させることにより、アノテーション作業の自動化を図る。２つ目の用途は副作用評価のための機械学習用データを作成する際の元データとしてである。本機械学習においては、アノテーション付きのテキストを入力として、副作用評価結果を出力する機能を学習する。

機械学習用辞書作成

機械学習用辞書作成に関しては、「ICD10 対応標準病名マスター」に収載された標準的病名約２万５千語（以下「ベース用語」）をベースにしたニューラルネットワークによる類義表現抽出実験を行った。この類義表現抽出実験は本研究が対象とする皮膚障害に限らず副作用報告書全般に適用可能な表記ゆれ対策を視野に入れて手法の可能性を検証する目的で行ったものである。

大量のテキストから類似した用語を抽出するクラスタリング技術に関してはこれまで様々な手法が提案されているが、従来アプローチである統計的手法も最近著しく発展してきているニューラルネットワークによる手法も、大量のテキスト（コーパス）の存在が前提となっている。しかしながら医療関係のテキストに関しては、研究用に利用できる日本語のリソースはごく限られている。本研究においても、ある程度まとまった量の副作用報告症例データの入手を前提に計画を進めており、それらに対するアノテーション作業完了後には、図３に示すようにアノテーション機械学習用データから表記ゆれの吸収を学習する手法を開発する計画である。しかしながら同時に適宜他の言語リソースによる補完も進める必要がある。初年度である今年度は、副作用報告症例の使用において、個人情報保護などの観点から使用可能なデータの精査と匿名化等のデータの加工を行う必要があり、症例データの入手が予定よりも大幅に遅れたため、その対応策として、Web からテキストを収集するアプローチを試みることとした。

C．研究結果

上記２万５千語のベース用語を Web から医療関係のテキストを収集するための検索キーワードとして使用した。まず Web ボットを用いてそれぞれのベース用語を含む合計約 280 万の Web サイトからテキスト情報を抽出し、タグ等を排除して約 32 億語のテキストを収集した。次にこのテキストを学習データとして、用語を高次元のベクトルで表現（分散意味表現）するためのニューラルネットワーク（word2vec）を

学習させ、ベクトルの近さで用語の類似度を評価する類似度評価器を作成した。図４に類似度評価器により検出された病名(A)および症状(B)の類義語の例を示す。図４の例は、見出し語との類似度が最も高い３００語から人手により類義語／表記ゆれ等を抽出した結果である。少数サンプルの抽出による調査を行ったところ、本評価器と人手によるチェックを組み合わせることで特定の医療関連用語とほぼ同義かあるいは置換可能なレベルで類似した用語を効率よく収集することが可能であることが分かった。

D．考察

副作用症例報告の経過欄に記載される病名や患者の症状は、診療録中の記載同様非常に多様な表現で記述されることが多い。一方でこれらのテキスト情報を機械で処理する場合、一字でも異なる用語は全く別の用語として処理されるため、表記ゆれの吸収は極めて重要な課題である。同じ病名でも漢字だけの表記、カタカナだけの表記、アルファベットだけの表記、漢字とカタカナを組み合わせた表記、アルファベットと漢字を組み合わせた表記など様々な表記形態が存在し得る。図４（A）に示される例からも推測できる通り、カタカナ表現に対しては非常に多くのバリエーションや誤記表現が用いられる可能性がある。また一部の漢字表現も同様と考えられる。これまで表記ゆれの対策においては、用語中の文字配列の近さ、すなわち文字の置換・挿入・削除といったオペレーションによる用語の相互遷移の容易さを基準に類似度を評価する手法が主流であったが、「アナフィラキシー」の例にも見られるように誤記までを対象に考えた場合、オペレーションの組み合わせは限りなく存在し、必要となる学習データの量を考えた場合、用語中の文字配列と言った局所的特徴のみに頼るには限界があると考えられる。今回試みた分散意味表現を用いる手法は、用語の内部構成ではなく、用語のテキスト中の文脈の近さをもとに類似度を評価するものであり、同じ用語のカタカナ表記と漢字表記の類似性を検出できるなど大きな利点を有するものと考えられる。

E．結論

副作用症例報告中の表記ゆれを吸収するためには、アノテーション機械学習用データをなるべく多く作成することが有効であることは間違いないが、人手によるデータ作成のみに頼るのではなく、外部の言語リソースによる補完を行うことも有効である可能性が示された。今回はWebから収集した標準病名を含む大量のテキストをもとに分散意味表現に基づく類似度評価器を作成した結果、評価器と人手によるチェックを組み合わせることで特定の医療関連用語とほぼ同義かあるいは置換可能なレベルで類似した用語を効率よく収集することが可能であることが分かった。

G．研究発表

１．　論文発表

該当なし

２．　学会発表

該当なし

H．知的財産権の出願・登録状況（予定を含む。）

１．特許取得

該当なし

２．実用新案登録

該当なし

3.その他

該当なし

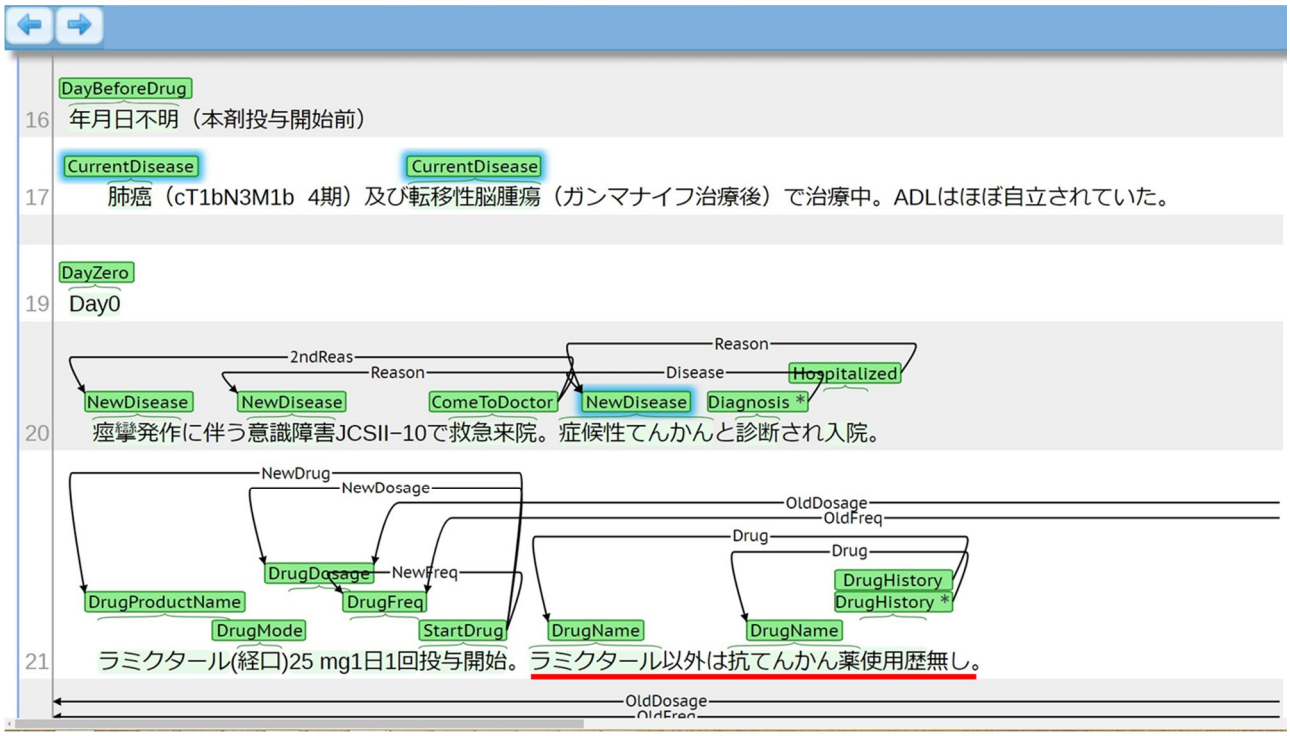

図 1 brat によるアノテーション例

912番目と916番目の文字境界に挟まれた文字列

Entity ID

テキスト中文字列

```
$ cat AdverseDrugReaction_14025805.ann

T1      NewDisease 912 916      意識障害
T2      DrugDosage 957 962      25 mg
T3      DrugDosage 1030 1035    50 mg
T5      CurrentDisease 859 865  転移性脳腫瘍
T6      NewDisease 930 937      症候性てんかん
T4      DrugDosageChange 1045 1047変更
E1      DrugDosageChange:T4 NewDosage:T3 OldDosage:T2 Drug:T13
T8      DrugFreq 962 966        1日1回
T9      DrugFreq 1035 1044      1日2回（朝、夕）
T10     DrugFreqChange 1045 1047  変更
E2      DrugFreqChange:T10 NewFreq:T9 OldFreq:T8 Drug:T13
T11     StartDrug 966 970       投与開始
E3      StartDrug:T11 NewDosage:T2 NewFreq:T8 NewDrug:T12
T12     DrugProductName 947 953  ラミクタール
T13     DrugProductName 1024 1030  ラミクタール
T14     DrugName 1082 1087      ステロイド
T15     StartDrug 1087 1089     投与
```

図 2 アノテーション情報ファイルの例

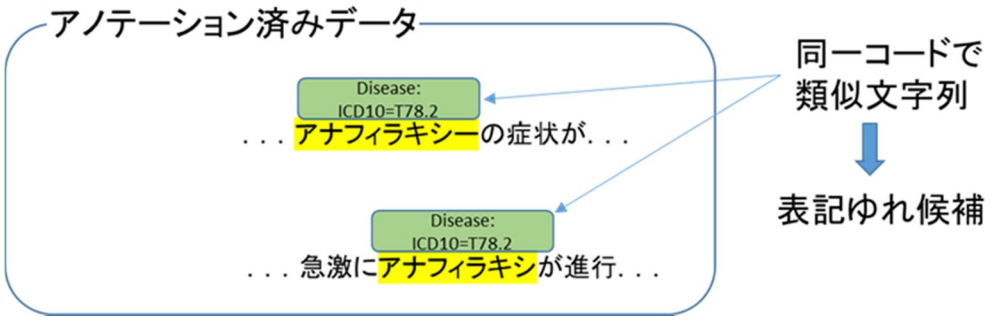

図 3 アノテーション機械学習用データからの表記ゆれ吸収学習

(A)

アナフィラキシー	変形性膝関節症	浮腫
Anaphylaxis アナフィラキシ アナフィラキシ- アナフィラキシーー アナフィラシキー アナフィラシーショック アナフィラキシ性ショック アナキラフィシー アナヂラキシーショック アナフェラキシーショック	変形性膝関節炎 変形性膝関節疾患 変形性膝関節足 変形性膝関節大腿 gonarthrosis gonarthrosis膝関節症	浮腫 ふしゅ エデーマ 畜水症 腫脹 ふしゆ 浮腫み 膨隆 膨瘤

(B)

吐き気		めまい	息苦しい	
嘔吐	悪心・嘔気	眩暈	息切れ	胸部圧迫感
悪心	はきけ	立ちくらみ感	動悸	呼吸難感
嘔気	おう吐	目眩	胸苦しい	動機息切れ
おうと	吐き気	身体揺らぎ	呼吸困難	息ぐるし
むかつき	胸つかえ	動揺感	窒息感	息切れ感
吐気	もよおす		呼吸困難感	胸圧迫感
嘔気	船酔い感		酸欠感	
嘔吐症	胸焼け			
嘔気症	げっぷ			

図 4 類似度評価器により検出された病名（A）および症状（B）の類義