

# 機械学習を利用した 医療現場の退院時サマリ作成業務軽減に向けた試み

早川 龍雄<sup>※1</sup> 内野 詠一郎<sup>※2</sup> 玉田 嘉紀<sup>※2</sup> 奥野 恭史<sup>※3</sup>

※1 富士通株式会社 ソフトウェア事業本部 AI ソフトウェア開発事業部

※2 京都大学大学院医学研究科 医療情報 AI システム学講座

※3 京都大学大学院医学研究科 人間健康科学系専攻 ビッグデータ医科学分野

電子カルテには患者に関する膨大な医療情報が蓄積されており、これらのデータを利活用することで先制医療や個別化医療につながることを期待されている。医療現場でも過去の患者への治療情報から医師・看護師の作業効率化へ応用し、働き方改革を実現するニーズが高い。中でも退院時サマリ作成は患者の入院期間に記録された自由記載テキストを含む膨大な診療録からポイントとなる治療経過を抽出する作業となる為、時間がかかる作業であり医師の作業負担となっている。本研究は、退院時サマリ作成支援を医師の知見（暗黙知）に基づいた診療録からの重要文抽出問題と捉え、自然言語処理を用いて実現する技術を確立することで医師の作業負担を軽減することを目的とした。要素技術としては LSTM により単語の出現パターン予測により重要文かどうかを判定し、医師の判断の観点をモデル化する試みを実施し、良好な結果が得られることを確認した。

キーワード: 自然言語処理, LSTM, BERT, 電子カルテ (EHR), 退院時サマリ, 働き方改革

## 1. はじめに

2019年4月に働き方改革関連法案の一部が施行され、様々な企業で働き方の見直しや効率化に対する検討が進んでいる。医療現場でも医師の労働時間等の環境改善の必要性が問われているが、進んでいない状況となっている [1]。実際の診療行為だけでなく事務書類の作成業務などがあり、さらに大学病院の場合は研究の実施や論文作成など多岐にわたる作業が存在していることがその要因である。中でも患者の退院時サマリ作成業務は大量の電子カルテの患者入院期間に記載されたテキスト記載情報から必要な情報を要約し記録として残すという作業である。診療報酬における加算（診療録管理体制加算1）に必要な作業で、患者退院後2週間以内の迅速

な作成が必要となる [2]。電子カルテシステムによる IT を活用した作成支援機能の実現は医療現場医師の作業負担を大きく軽減することが期待できる機能であると考えられる。

著者らは、電子カルテシステム開発の知識を活用しながらアカデミアと共同で電子カルテ診療録のテキストデータの解析の為の研究を行っている。京都大学と富士通は、新たな診療支援や創薬の実現など AI を活用した次世代の高度医療化に向けた研究開発を目指し、共同研究講座を設立した [京都大学医学部附属病院 医の倫理委員会 承認番号: R1408] [3]。研究で開発した要素技術やノウハウを応用し、社会に成果を広く還元していくことを目指している。その中で電子カルテのテキスト記載情報から退院時サ

マリを作成する上での医師がどのような知見（暗黙知）で重要な記載であると判断するかという問題について NLP 技術を利用して研究を進めている。以降、実際に LSTM や BERT などの深層学技術習を用いて実施した電子カルテテキスト情報からの重要文抽出処理の研究内容と結果について論じる。

## 2. 電子カルテのテキスト記載情報の特徴と課題

電子カルテ診療録には、患者の様々な情報が記載されており、患者の身体所見・既往歴・家族構成、患者の SOAP（主観的データ・客観的データ・評価・計画）などが存在する。電子カルテ診療録は患者の治療の為の情報を記載するシステムであり、後にデータを活用するように考慮されていない為にほぼ構造化されておらず、文書と検査項目が混在し、口語表現・書き言葉を多く含んでいる。用語の使い方も統一されておらず、略語表記も多用され、あまり第3者が見やすいような考慮は行われていないのが現状である。退院時サマリはそれらの記載から医師が患者の入院期間の治療過程でポイントとなる文書を要約して記載している。退院時サマリを作成する為の書き方ルールは標準化されておらず、記載する医師によっては文書から必要なセンテンスをコピー&ペーストして編集したり、記載内容を参考に新しい文書を作成したり、など様々なパターンが存在する。患者の入院期間の長短でも違いがあり、長期入院患者の場合は、文書を独自に作る傾向が強くなっている。このようなデータからまず重要文抽出処理の為にデータプレパレーションとアノテーション付与の処理を実施した。

## 3. データプレパレーション処理

今回の検証を行うにあたり、京都大学医学部附属病院の腎臓内科の約 2000 名の患者の匿名加工処理済み電子カルテ内の SOAP データと退

院時サマリデータを利用した。データプレパレーションとして電子カルテ診療録のテキスト構造に合わせて以下の様な処理を実施している。

- 辞書に合わせた全角・半角の規則統一
- 数値表記（ローマ数字・丸数字）変換
- 括弧表記の統一化
- 検査結果等を含むスペース・タブ等除去
- 文書のセンテンス分割（「。」や改行等）
- コピー&ペーストで作成された重複センテンスの削除

センテンス単位に分割されたデータの中には、プロブレムと呼ばれるタイトルとして記載されている内容や、検査結果項目のみの表記、覚書のように単語のみで記載されている内容なども存在している。今回は LSTM によるセンテンス内の単語の系列情報のパターン識別を目的としており、これらに該当するセンテンスは除外対象とした。

退院時サマリには、家族歴や既往歴、医師の所感、検査結果、画像診断レポートなど入院時の電子カルテ診療録以外の情報を含めて参照されるデータも存在する為、今回は電子カルテを参照して作成される治療経過記録のみを抽出した。さらに入院期間が 28 日を超える患者データについても除外対象とした。

## 4. アノテーション付与処理

センテンス単位に分割されたデータに対して退院時サマリに記載されるかどうかの分類アノテーションを付与した。膨大なデータに対して機械的に分類を行う為に Levenshtein 編集距離で SOAP データと退院時サマリの特定の患者の同じ日時で編集距離が近いセンテンスにアノテーションを付与する方針とした。Levenshtein 編集距離は単語ではなく文字の編集距離となるので、形態素解析による分かち書きを行った結果に対して編集距離を出力した。

退院時サマリにはコピー&ペーストした後に

単語を置き換え、複数文書を結合するなどの加工を行うケースも多く、Levenshtein 編集距離だけではアノテーション付与が不足する為、電子カルテ診療録と退院時サマリから Word2Vec モデルを作成し、その上で文書をベクトルの分散表現として類似性比較する Word Mover's Distance (WMD) を利用した[4]。WMD だけでは Word2Vec で学習された特徴単語で類似単語を考慮してスコアを算出する為、医療的な観点で全く異なる意味でもスコアが低くなる（類似度が高くなる）傾向があったので、Levenshtein 編集距離と WMD の組合せを行った。

アノテーション付けする編集距離と WMD のそれぞれの閾値は少量データに対して臨床医や学生が分類した内容の精度に近い値を採用した。

## 5. 学習処理と評価

アノテーションが付与されたデータセットに対して、LSTM で 2 クラス分類での重要文抽出を実施した。退院時サマリは電子カルテ上で医師の判断で必要となる情報を抜粋・加工して記載される為、要約文書の生成ではなく退院時サマリに記載する候補を重要文として抽出する方針とした。評価方法としては重要文としての漏れを最小限とする（逆に重要文でないセンテンスが候補となってもある程度許容する）こととし、recall 値による判断を行った。

当初、1 センテンスに 1 アノテーションで学習処理を行っていたが、対象のセンテンスの前後のセンテンスを含めて学習を行うことで精度を向上させることが確認できた [表 1]。

	1センテンスに対して 正解アノテーションを付与  【例】 センテンスA : 1 (重要文) センテンスB : 1 (重要文) センテンスC : 0 (重要文でない) センテンスD : 0 (重要文でない) センテンスE : 0 (重要文でない)	前後を含めた3アノテーションに対して 中央の正解アノテーションを付与  【例】 センテンスABC : 1 (重要文) センテンスBCD : 0 (重要文でない) センテンスCDE : 0 (重要文でない)	前後を含めた3アノテーションに対して それぞれのセンテンスのアノテーションの OR条件でアノテーションを付与  【例】 センテンスABC : 1 (重要文) センテンスBCD : 1 (重要文) センテンスCDE : 0 (重要文でない)
accuracy	0.615 ± 0.009	0.798 ± 0.032	0.706 ± 0.024
recall	0.776 ± 0.040	0.728 ± 0.068	0.864 ± 0.012

表 1：複数センテンスの組合せ検証

LSTM のネットワーク内で、1 センテンス用と前後を含めた 3 センテンス用（アノテーションは OR 条件）の Embedding の結合を試みた [図 1]。電子カルテ診療録の記載で医師が重要文と判断するのは、センテンスごとの記載内容の重要性と、そのセンテンスの前後のセンテンスに含まれる原因 (cause) や結果 (outcome) などの関係性を示す単語が含まれている可能性を学習結果に反映しようとした。

少量データ（患者 200 人）でモデルのパラメータの比較検討を行い、以下の設定値にて実証を行った。

- learning rate : 0.0001
- loss 関数 (Cross Entropy) の weight : 正解データの比率に合わせ 1:3
- dropout : 全層 (選出確率 0.2)
- batch size : 64

結果として、recall 値で 0.805 の結果を出力することができた。電子カルテ診療録のそれぞれのセンテンス内の出現単語に重要文判定として特徴的な単語が次のセンテンスに出現するというパターンが LSTM の系列パターンとして特徴をよりとらえたことが可能となったと考察している。

同じデータセットを用いて BERT との精度比較についての検証も実施した。BERT を利用するにあたり、京都大学 黒橋・河原研究室の BERT 日本語 Pre-trained モデルを利用し、形態素解析エンジンとしては JUMAN++V2 を利用した[5]。特に LSTM の時の様なデータセットに対する工夫を行わずにファインチューニング

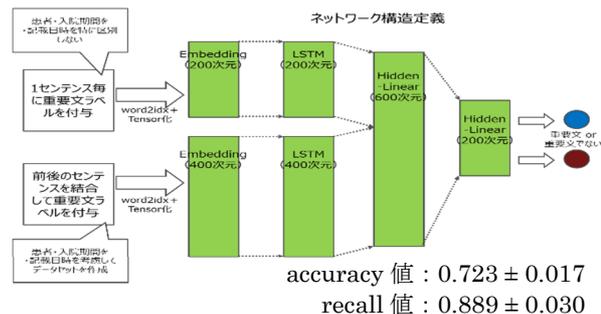


図 1：ニューラルネットワーク構造図

した結果、AUC は 0.84 となり閾値を 0.05 以上で重要文と判断することで recall 値が 0.725 という結果となった [図 2][表 2]。

## 6. 今後の課題

作成したモデルを利用し、実際の電子カルテ診療録のサンプル患者データに適用した結果、正解率に大きな差があった。原因としてはモデルの Embedding で利用した単語群に含まれない未知語の出現数であり、未知語が多いと正解率が低下する傾向となった。退院時サマリ作成における重要文抽出としては未知語が多いケースでの対策が重要となる。データが増加するごとにファインチューニングして語彙をアップデートすることが望ましい。暫定的な方法として、対象患者のみの少量データで Word2Vec モデルを作成し、類似性の単語の中から LSTM で作成

したモデルを構築する際に既知語となっている単語に置き換えて学習するという手法としているが、データ量が少ない患者ではあまり効果がなく、今後対応策を検討していく必要がある。

## 7. おわりに

本論文では、電子カルテ診療録から医師が退院時サマリ作成の負荷軽減を目的とした深層学習技術の利用についての研究結果を述べた。

まだ研究段階であり、退院時サマリ作成支援機能としての実用化は今後さらに検討を行う必要があるが、医療テキストデータから重要な部分を抽出する為の技術として有意義であると考えている。今後診療現場における医師の働き方改革を支援する技術として応用していくことに期待している。

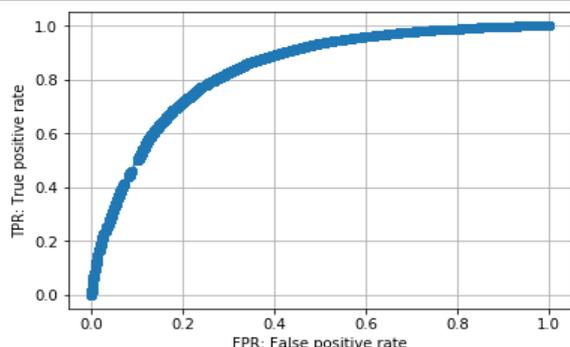


図 2 : BERT での出力結果 (ROC 曲線)

	BERTの出力結果に対する重要文判定の閾値					
	0.05	0.1	0.2	0.3	0.4	0.5
accuracy	0.790	0.855	0.913	0.932	0.937	0.938
precision	0.189	0.234	0.306	0.392	0.481	0.555
recall	0.725	0.583	0.309	0.164	0.071	0.042
f1	0.300	0.334	0.307	0.231	0.123	0.078

表 2 : 重要文判定の閾値

## 参考文献

- 1) 総務省 第 1 回 医師の働き方改革に関する検討会 資料 3  
「医師の勤務実態等について」, 平成 29 年 8 月 2 日
- 2) 退院時要約等の診療記録に関する標準化推進合同委員会(日本医療情報学会・日本診療情報管理学会)  
「退院サマリ作成に関するガイダンス」, 2019 年 9 月 25 日
- 3) 富士通プレスリリース, 「京都大学と富士通、高度医療化に向けた AI 活用の共同研究講座を開設」  
[<http://pr.fujitsu.com/jp/news/2018/01/24.html>] (2019/12/19 アクセス)
- 4) Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, Kilian Q. Weinberger  
Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130  
「From Word Embeddings To Document Distance」
- 5) 京都大学 黒橋・河原研究室, 「BERT 日本語 Pretrained モデル」  
[<http://nlp.ist.i.kyoto-u.ac.jp> > BERT 日本語 Pretrained モデル] (2019/12/19 アクセス)