自然言語処理技術の最前線と医療応用の可能性

松村泰志[*1]、鳥澤健太郎[*2]、篠原恵美子[*3]、
鈴木隆弘[*4]、荒牧英治[*5]
*1 大阪大学大学院医学系研究科医療情報学、
*2 情報通信研究機構ユニバーサルコミュニケーション研究所、
*3 東京大学医学部附属病院企画情報運営部、
*4 千葉大学医学部附属病院企画情報部、
*5 奈良先端科学技術大学院大学情報科学研究科

The front line of natural language processing and its possibility for medical use

Matsumura Yasushi[*1], Torisawa Kentaro[*2], Shinohara Emiko[*3]
Suzuki Takahiro[*4], Aramaki Eiji[*5]
*1 Osaka University Graduate School of Medicine. Medical Informatics,
*2 National Institute of Information and Communications Technology, Universal Communication Research Institute,
*3 Tokyo University Hospital. Department of Healthcare Information Management,
*4 Chiba University Hospital, Medical Informatics and Management
*5 Nara Institute of Science and Technology Graduate School of Information Science. ,

Because of prevalence of electronic medical records and increase of digitalized medical contents being able to be accessed via network, massive volume of medical information becomes to be existing in digital form. However, because most of them are unstructured data, they cannot be able to be processed by computer system without natural language processing (NLP). Although the research for NLP have a long history, it has remarkably progressed in recent years. On the other hand, its application for medical use is not easy, because a vast volume of specialized terminology are used in medical field. In this organized session, firstly Torisawa make a keynote speech about current trend of NLP technology and its possibility for medical use. Secondary, Shinohara make a speech about a framework for analyzing medical record and how to use NLP. Thirdly, Suzuki talks about method for searching similar cases using discharge summary or case reports as an application of NLP. Lastly, Aramaki presents his 3 projects which used NLP; a detection of side effects from discharge summaries, diagnosis support system based on case reports, and infection surveillance using social media.

Keywords: natural language processing, unstructured data, electronic medical record, case report

1. 本セッションの趣旨

医学の進歩は著しく、日々多くの論文、ガイドライン、書物が発行され、多くはコンピュータでアクセス可能となっている。また、電子カルテシステムが普及し、大量の診療データがデータベースに蓄積されるようになった。このように、今日では、日々大量の医療情報がデジタルデータとして発生している。しかし、これらのデータは、自然言語で記録された非構造化データのものが殆どである。特に電子カルテのデータは、医療におけるビッグデータと期待されてはいるものの、重要なデータは非構造化データの中に含まれている。これらの電子化された大量の医療情報に対し、処理を加え、診療支援、臨床研究、意思決定支援等で利用するためには、自然言語処理が必須となる。

自然言語処理は、古くから期待され、研究されてきた領域である。従来からの形態素解析、構文解析、文脈解析、意味解析に至る方法の研究に加え、最近では統計的自然言語処理の技術も加わり、大きく進歩してきており、多くの実用化事例が登場するようになった。一方、医学領域の自然言語処理については、多くの専門用語で構成され、決して容易な領域ではない。

本企画では、自然言語処理技術の最前線の技術を、情報通信研究機構ユニバーサルコミュニケーション研究所の鳥澤健太郎先生よりレクチャーしていただく。また、医学領域での自然言語処理について研究成果を挙げてこられた東京大学医学部附属病院企画情報運営部の篠原恵美子先生、千葉大学医学部附属病院企画情報部の鈴木隆弘先生、奈良先端科学技術大学院大学情報科学研究科の荒牧英治先生に、現状の到達点と、これからの研究の方向性・可能性、そのために解決すべき課題について解説頂く。

2. NICTにおける自然言語処理研究と医療応用の可能性（鳥澤 健太郎）

近年、自然言語処理技術は劇的な進歩を遂げている。情報通信研究機構(NICT)は総務省傘下の国立研究開発法人であり、過去数十年にわたり自然言語処理の研究開発を実施してきており、こうした進歩の一翼を担ってきたと自負している。具体的にNICTでは、スマートフォン上の多言語音声翻訳システムVoiceTra[1)]や、大規模災害の被災状況をSNSにおける一般ユーザの発信を元に分析する対災害SNS情報分析システムDISAANA/D-SUMM[2)]、40億ページのWeb上の情報を元に様々な質問に回答するWISDOM X[3)]等の一般公開も行なっており、ビジネス化、実用化への動きも加速しているところである。現在では、さらに次世代のメディア技術として、一般ユーザと音声で対話を行う対話エージェント「WISDOMちゃん」（仮称）の開発を進めているが、これは現在実用化が

行われているいわゆる AI スピーカーとは異なり、膨大な Web 情報と深層学習を用いて、話題を限らず、多様な対話が可能な対話エージェントである。本講演では、これらのシステムの解説を通して、自然言語処理の現状、特に機械学習の一種である深層学習の進歩による研究スタイルの変化や、それがもたらす可能性と限界について述べる。一般の方々の多くは自然言語処理の究極の姿として、ユーザが考え入力した質問に、テキストに書かれた適切な回答を返す質問応答システムを思い浮かべると思われるが、現在の自然言語処理研究のスコープはそうした質問応答システムの域を超え、有用な質問をユーザの入力を待たずに自動生成してしまう試みや、テキストに書かれていない仮説を導き出すといった試みも始まっており、より高度なユーザ支援の実現を狙い始めている。本講演ではそうした高度なユーザ支援を実現するための研究戦略や、そうした技術の医療応用の試みや可能性、そして医療自然言語処理を実現する上での障害であるデータの量についても議論する。

3. 医師の作成した診療録を解析するためのフレームワークおよび自然言語処理（篠原 恵美子）

診療録のテキストに含まれる情報は患者の症状や医師の判断など重要な内容を含み、診療情報の利活用という観点からこれを構造化する手法が必要とされている。診療情報の利活用にはさまざまな目的があり、中でも個々の医師・患者に適用される意思決定支援や処方に関する警告システム等においては解析誤りがインシデント・アクシデントを引き起こす可能性があり、非常に高い解析精度が求められる。自然言語処理研究では機械学習手法が主流となっているが、精度向上に必要な十分な量の学習データの準備だけでなく特定の解析誤りへの対応も困難であることから、医療現場への導入は慎重にならざるを得ない。一方、古くから研究されてきたフレームベース・知識ベースの手法は高い精度を達成した実績があるものの、解析のために必要な知識の作成・管理が容易でないという問題点があった。しかし現在では用語集やオントロジーなどのリソースが整いつつあり、また医学知識と自然言語の文法知識を分離することで知識管理の困難は大幅に減ると考えられる。特に診療録の中でも医師が記述するものは医学的内容が中心であり、医学領域に限定すれば網羅的な知識の記述は決して非現実的ではない。また、テキストからの情報抽出を困難にする要因の 1 つが「肺炎に注意」の記載から肺炎の有無を判定する等「行間を読む」ことであるが、我々は、医師が記述した診療録を対象とした場合に、自然言語処理が解決すべき課題を「書かれていることを文字通りに解釈し構造化すること」に限定し、それ以上の解釈は自然言語処理の後、他の構造化データ等と合わせて行うべきであると考えている。このような考えから、症状・所見・診断などの患者状態、および投薬・処置などの医療者の介入に関する情報を構造化し抽出する知識ベース手法の構築を目指しており、本シンポジウムではその概略を紹介する。

4. 医療テキスト解析の理想と現実（鈴木 隆弘）

千葉大学医学部附属病院では 1978 年より統一された形式で退院時サマリを作成してきた。このサマリは手書きの複写式で自由文と選択項目から成り、カルテへの添付と代行入力によるデータベース登録が行われ、病院情報システムからの検索サービスを提供していた。2000 年にはユーザの直接入力によるサマリの全文電子化を行った。これが当院で医療テキストを構造的に格納した最初のシステムである。2003 年には電子カルテの導入が始まり、医療テキストが大量に蓄積され始めた。

これらの蓄積されたテキストを利用して、隠れた知識を発見しようとする試みは当初から構想しており、2003 年には 14 疾患の退院時サマリを抽出して自動的に疾患を判定する試みを行い、良い成績を得た。このときテキストマイニング方法として用いたのがベクトル空間モデルで、重み付けにはシンプルな TF*IDF 法を採用した。以来、カルテ記載の本文や看護記録を対象に加え、診断や DPC 分類の自動判定などの試みを行ってきた。これらの成果は内科学会の類似症例検索システムとして実用化されている。現在は多施設と共同でサマリを比較検討し、医療用語辞書の作成にも協力している。

我々はこれまで、テキスト処理に高度な文法的解析は行っていない。大きな理由の一つは対象となるテキストそのものが文法を無視して書かれていることである。多忙な日常の臨床現場では入力に追われて文章を推敲する余裕が無いためか、カルテの記載は略語と誤変換で溢れている。

カルテは医療のアカウンタビリティの基本となるものなので、忙しいからといっていつまでも低品質のままでは許されない。診療情報管理士と協力して文章の改善に努めているが、今後はシステムの機能としても良質のカルテ記載をサポートするための技術開発が強く求められ、それが解析精度の向上にも繋がると考えている。

5. 自然言語処理の医療応用のこれまでとこれから：3つの開発事例をもとに（荒牧 英治）

医療現場で生成される多様なデータの相当な部分は自然言語文を含んでおり、今後もそれはただちに変わりそうにない。これまで医療データの二次利用は、健診データや DPC の診療報酬データなど、比較的構造化されたデータが主な材料であったが、最近では、より大規模かつ非構造化されたデータを扱う方向へ発展しつつある。その非構造化されたデータの代表が自然言語データである。我々は、この医療分野における自然言語データの利活用を以下の3つ動向に分類している。

まず、大きな動向は、診療録（電子カルテ）に代表される医師が日常診療で残すデータの利活用を目指す方向である。本発表では、我々が退院サマリからの副作用シグナルの自動抽出を紹介する。

もう一つの大きな動向は、論文やウェブ情報など公開されているデータを残す方法である。現在、医療人工知能の多くはこの動向に位置づけられる。本発表では、症例報告を用いた診断支援システムを紹介する。

最後に、この数年ほどの間に注目を集めているのが、患者がソーシャルメディアや患者会などを通じてやり取りするプライベートなデータを扱う方向である。本発表では、ソーシャルメディアを用いた感染症サーベイランスを紹介する。

以上、本発表では3つの具体的な研究事例を用いて、それぞれの到達点、共有できるリソースを紹介し、今後の動向を議論する。

参考文献

1) 多言語音声翻訳アプリ＜ボイストラ＞ VoiceTra, [http://voicetra.nict.go.jp] / (cited 2017-Sep-15)].
2) DISAANA/D-SUMM 災害情報要約システム, [https://disaana.jp/d-summ/]/ (cited 2017-Sep-15)].
3) WISDOM X, [http://wisdom-nict.jp] / (cited 2017-Sep-15)].