医療分野における自然言語処理の研究発表や研究論文に関する一考察

大塚敬義
清水洋生

1．自然言語処理の概略、そして医療情報学分野への浸透の経緯

コンピュータのソフトウェアを開発する際に用いる、Java（ジャバ）やCといったプログラミング言語は非自然発生的な人工的言語であるのに対し、コンピュータの登場以前から人間が意思疎通に用いてきた日本語や英語といった自然発生的な言語を自然言語（natural language）とよぶ。

自然言語処理を専門としない方々も読者層に含まれると考え、自然言語処理の概略を述べたい。

コンピュータはその黎明期においては、元来の開発目的が数値計算であったこともあり、比較的簡単な記号類や数字を取り扱うまでが限界であった。コンピュータの出自は英米両国に由来していることもあり、使用者にとってみれば自然言語の表示に必要な文字群は、小文字・大文字の違いを区別しなければ英語の表記に必要なAからZまでの26種類の文字を取り扱えればそれで充分であった。

日本において1970年代にまだコンピュータがマイコンとよばれていた時代、市民の間に流通していた機種では漢字は勿論、ひらがなの表示も難しい機種も存在し、当時はカタカナが取り扱えればコンピュータは商品として成立した。

やがてコンピュータの性能向上につれて、ひらがな、漢字の表示は当然のごとく標準装備の機能として実現されさらに時代が下ると、漢字を入力する際には、文字コード表に従い10進数の4桁の数字で記載された句点コードを1文字分ずつ指定する方式が不便とされ、ひらがな文字列から直接漢字文字列に変換する「かな漢字変換」の機能が発明された。たとえばかな文字列「ちばけん」をキーボードから入力し変換キーを押下すると、演算結果をもとに最適候補として漢字文字列「千葉県」に変換される。このように自然言語をコンピュータ上で処理する目的で花開いた学問分野が自然言語処理（Natural Language Processing）の源流に当たる。1990年代前半に米国のクリントン・ゴア政権の下で民間に対して商用目的でのインターネット利用が開放されると[1]、ウエブと呼ばれるインターネット空間からウエブページを探し出すGoogleやYahooといった検索エンジンが社会で存在感を帯び始めた。

検索エンジンの根幹技術を為す自然言語処理の技術は、かな漢字変換の機能を搭載したワープロの製造や、Amazonなどにおける商品検索など産業分野において早い時期から実用面での利活用が行われてきた。これに対し医療分野においては、病院内の各部門に情報技術の導入を進めてきてはいたが、導入対象は医事会計システム（診療点数を計算するソフトウェアあるいはそれに特化したハードウェア）や、オーダリングシステム（看護師や薬剤師など医療技術職に対して行う指示内容を直接コンピュータに入力して正確にかつ迅速に各部門へ伝達するシステム）[2]にとどまることがあり、医師が手書きで患者の病状を記述してきたカルテ（診療録）を電子化して保存・参照する電子カルテシステムの導入段階までは未到達となっていたことがあった。

このため医療情報学と呼ばれる研究分野では従前、各疾病の平均在院日数などの、定量的に扱える数値デー

タを分析対象とする研究が比較的多数行われてきた。

近年ではようやく電子カルテシステムの普及が進み、カルテの記録方法が紙媒体から電子媒体へと変化しつつある。記録方法が電子媒体となれば、それに格納された文字列はデジタル情報であるから分析対象のデータとしてみなされるようになった。とはいえ自然言語の文字列を中心とするデータは定量的ではなく定性的なデータ部分が存在するため、数学的な計算対象とはなりにくい。そのため、記録媒体が紙からハードディスクに移行した後も、解析手法の乏しさや、個人情報保護といった法律上の問題が存在してきたことから、自然言語処理の技術発展の恩恵を受けにくかった。

しかしながら現今、自然言語処理の技術を応用した解析ツールが巷間に幅広く頒布される環境が整い始めた。また研究機関が、匿名化された診療録データを倫理遵守の条件付で研究者らに貸与する制度の登場も相まって、医療分野においても自然言語処理技術を適用した研究論文や研究発表の件数が増加した。ただしそのような「医療言語処理」（医療分野に特化した自然言語処理の研究）の専門書は刊行点数が現時点では多くない。少数ながら書籍の形で和文の専門書も登場するようにはなりつつも、初版の刊行や改訂版の登場からいくらか年月が経過している場合は、医療分野における自然言語処理の研究動向を把握しにくいのが実情である。

また最も活発的な部類に属する医療言語処理の専門的研究会が存在しても、そこで発表される研究報告は論文が英文で記述されているため、日本国内の医療従事者にとっては研究報告の内容を把握しにくいのが難点である。

そこで著者らは信頼性の高い専門書や、発表された新しい研究報告などに対し、Google 等による検索のみを用いた手法でなく、学会年次大会で配付された DVD-ROM 現物媒体や冊子体の原典を入手し、Google 等のみでは入手不可能な情報を、正確にかつ各論文全体について研究内容を把握できる方法を採る手法により複数調査し、医療言語処理の研究発表や研究論文について、本稿で報告する。

2． 各ドメインにおける学界動向

2．1 図書、特に学術専門書

医療言語処理に焦点を絞って扱う専門書として和文で執筆された奥村、荒牧による「医療言語処理」[3]がある。初版刊行時期は 2017 年 8 月と比較的新しく、それまで医療言語処理の各研究報告などは、各種学会の学会誌や学会年次大会の予稿集等に散発的に掲載されがちであったが、本書 [3] は医療言語処理の全体像を幅広い視点から俯瞰し、体系的に整理して記載している点が非常に優れている。自然言語処理の初学者にも読みやすくなるよう配慮がなされ、適宜脚注において解説がなされている。本稿執筆時点で最良の専門書である。

また永年、医療情報学の分野において専門図書の刊行実績のある篠原出版新社から『医療情報』シリーズの改訂版 [4] [5] [6] が 2016 年に出版された。『医療情報』は全 3 分野から構成されており、「医療情報システム編」[4]、「医学・医療編」[5]、「情報処理技術編」[6] がある。医療機関に特化した情報システムに関する知識を深めるのであれば [4] [6] が関連書籍として適している。

2．2 学会誌に掲載の研究報告

情報技術の必要性が叫ばれ始めた頃、比較的早い時期から医療分野において存在感を発揮している医療情報学の専門誌である「医療情報学」が存在する。情報処理学会が毎月刊行する会誌である雑誌「情報処理」にも時折、医療情報学分野の記事は掲載される。しかし「情報処理」は情報処理教育や情報工学全般の話題を取り扱う一般紙としての性格が強く、1 解説記事当たりのページ数が少ない。さらに医療情報学を専門としない読者層に

も膾炙すべく、多くのページ数を要する詳細な研究報告や、医療情報学の分野でも特殊性の高い研究テーマを扱う記事の掲載に不向きな傾向がある。これに対し日本医療情報学会が刊行する論文誌「医療情報学」は、前述の「情報処理」が有する傾向に該当しないので、掲載上の割り当てページ数や、掲載内容が特定の研究領域に細かく焦点を当てていても掲載対象から除外されにくい。したがって論文誌「医療情報学」は、雑誌「情報処理」と比較すると、掲載内容がより専門性が高い傾向を持つと言えるだろう。学界の動向を知るためにはサーベイ対象に含めることが有益である。2018 年 1 月付刊行の論文誌には、柴田、若宮、木下、荒牧らによる医療言語処理の研究論文[7]があり、認知症者の発話における感情表現を、彼ら独自の日本語感情表現辞書（JIWC）を利用し調査することにより、アルツハイマー型認知症群では「嫌悪感」、「怒り」、「不安」に分類された感情表現の使用割合が有意に増加し、認知症および軽度認知障害の群では「不安」に分類された感情表現の使用割合が有意に増加していたとする、認知症者のスクリーニングを試みている。

加えて、本稿執筆時点で「医療情報学」の最新号第 38 巻・第 1 号に、日本医療情報学会の秋の年次大会である第 37 号医療情報学連合大会（第 18 回日本医療情報学学術大会）にて開催された大会企画シンポジウムの抜粋[8]が掲載され、医学領域での自然言語処理について研究成果を挙げてきた 3 名の研究者らが、現状の到達点とこれからの研究の展望について語っている。

申し添えておくと学会誌「医療情報学」の冊子体に掲載されている各論文の全文を、誰でもが単純に Google 等で検索するのみでは閲覧することは 2019 年 1 月 20 日現在、不可能である。ウエブ上からのアクセスのみで論文の全文を閲覧するには、医療情報学会の会員資格を得たのち、会員専用ページからログインする必要がある。会員専用ページを介さずに各論文の全文を閲覧することは原則として不可能である。

2．3 口頭発表やポスター発表等の学会発表

各学会の年次全国大会において発表された研究報告を、学会ごとに新しいものから順に取り上げたい。

2018 年 3 月に岡山県で開催された言語処理学会第 24 回年次大会（NLP2018）にて発表された各研究報告のうち、医療言語処理に関連の深い発表が 2 件存在する。香川ら[9]は、医師が診断に至る思考過程に関連する情報を付与したカルテコーパスの作成と利用が診療支援に有用であるとし、病名以外の情報の利用が必要であり、よってそのデータに情報を付与する必要性に言及している。矢野ら[10]は、医療の現場などで作成される文書の日本語入力に関して触れ、入力内容に非文法的な表現、低頻度の複雑な複合名詞、省略型の多用などの特徴があることを指摘した上で、辞書に登録する各見出し語の共通接頭辞を併合することにより構築される木構造（トライ構造）による文字ベースの辞書を構築し、独自の医療向け入力支援ツール「MedInput」が実用的であり、MedInput の利用が推測変換機能の強化につながることを示唆した。

次に 2017 年 11 月に大阪府で開催された第 37 回医療情報学連合大会（第 18 回日本医療情報学会学術大会）にて発表された研究報告に着目する。

知識工学のセッションにおいて 6 件の発表があり、松尾ら[11]は医療文書を用いてデータマイニングを行う際の前処理として医療文書の特性に対応しつつ、自然言語処理技術により単語の重み付けをすることが重要であると述べ、医療文書内の単語を分析目的に応じて、階層的に分類する手法を提案している。加えて、データマイニング・テキストマイニングのセッションにおいて 6 件の発表があり、山ノ内ら[12]は、自然言語のような非構造化データを解析対象とする際、特に単語辞書や類語辞典に新語を追加し続ける更新作業などの困難さを軽減すべく、自然言語処理の弱点を改善したオープンエンド型発見手法「iKnow」および深層学習を用いる方法で、外来初診時の候補病名を予測することを試みている。

2.4 特化型ワークショップ NTCIR における成果発表例

国立情報学研究所（NII）は、研究プロジェクト「NTCIR」（エンティサイル、NII Testbeds and Community for Information access Research：国立情報学研究所が用意した、コンピュータープログラムによる検証試験の場および情報検索の研究に関する情報交換の場）を主催している。NTCIR では、多数の研究者が大規模な評価基盤（「テストコレクション」とよばれる実験用データ、後述）を共有・利用し、研究者同士が相互に各システム間の性能比較や知見の共有を行う[13][14]。

テストコレクションは「実験用データの集合体」として3種類の役割を持つ。役割のまず1つめは、訓練用データや評価用データ等を蓄積したデータベースとしての性質を有する点である。2つめは、利用者の検索要求を記述した「検索課題」である。たとえばあるテキストに対し、年齢や性別を示す箇所にマークアップのタグを付与させる指示等を各種集めたものである。3つめは、検索課題を満たす「正解文書の網羅的なリスト」であり、いわゆる模範解答に該当するデータ群である。

NTCIR という場の最大の特徴は、評価ワークショップ形式を導入した点にある。すなわち国立情報学研究所が、テストコレクションおよび実験結果を評価する目的に即した共通の手順を用意し提供する。換言すればNTCIR において、各実験の企画や課題（タスク）を発案した主催者に相当するグループがデータを用意する。主催者側から発信された企画参加の呼びかけに応じタスクに参加する各グループは、各グループ独自の手法により研究および実験を行う。一般的に、情報検索やテキスト処理の研究において、実験に繰り返し利用可能な大規模な標準データセットが重要である。NTCIR はそのようなデータを提供し、かつ研究上の着想や技術の交換・移転を目的とする研究者の集会の場を提供することによって、研究の推進を企図している、新しいタイプの共同研究である[14]。

NTCIR はおよそ1年から1年半毎に回次が進む。2018 年 5 月時点で第 14 次のプロジェクト「NTCIR-14」が進行中である。NTCIR の最近の回次では、NTCIR の第 10 回次である NTCIR-10 以降、医療言語処理のタスクが実施されており、こちらは古いものから時系列に沿って記述するとタスクの発展や高度化を理解しやすい。

最初に NTCIR-10 で「MedNLP」とよばれるパイロットタスクが医療言語処理の特徴を帯びて登場した[15]。なお NTCIR-11 以降の他の医療言語処理タスクと区別する意味で、NTCIR-10 における医療言語処理タスクを「MedNLP-1」と表記する場合がある。MedNLP-1 では，日本語の医療文書から重要な情報（個人情報や医療情報）の抽出を行う[16]。設定されたタスクは主に匿名化タスク、症状と診断タスクなどから構成される。挙例すると、匿名化タスクでは、文章に対して年齢（age）を意味するタグ <a> や、同様に日時（time）のタグ <t>、病院名（hospital）のタグ <h>、場所（location）のタグ <l>、個人名（person）のタグ <p>、性別（sex）のタグ <x> などを付与する。2 種類目の症状と診断タスクでは、文章に症状（complaint）と診断（diagnosis）のタグ <c> を付与する。

その後の NTCIR-11 では、「MedNLP-2」[17] と名称が変わり、タスク内容がより高度になった。具体的にはテキストから症状を抽出する病名・症状抽出タスクのほか、症状に国際疾病分類（ICD）に準拠する病名コードを付与する病名・症状正規化タスクが設定され 12 件の研究報告がなされた。後者のタスクにおいては、入力データの原文「2025 年月 8 月 2 日（来院 5 日前）頃から腹痛が生じるとともに、食欲不振、嘔気・嘔吐出現した」に対するタグ付きの出力データ例として「<t>2025 年月 8 月 2 日（来院 5 日前）頃 </t> から <c icd="R104"> 腹痛 </c> が生じるとともに、<c icd="R630"> 食欲不振 </c>、<c icd="R11_"> 嘔気 </c> ・ <c icd="R11_"> 嘔吐 </c> 出現した」が例示されている[18]。

続く NTCIR-12 は「MedNLPDoc」（MEDNLP-3）と称し、9 件の研究報告がなされ、複数の病名コードを持ちうる診療データを扱うようになり、文章（document）に対するマルチ・ラベリング問題にも挑んでいる。前回との主な違いは、MedNLP-2 においては名詞単位で病名に病名コードを付与していたが、MedNLP-3 では一定量のテキスト全体の記述を総合的に判断し、システムにより病名コードを付与する。訓練データとして診療情報管理士向けのテキスト「ICD コーディングトレーニング第 2 版」の診療データを使用している。

訓練データは箇所ごとに <text> タグで区分されており、属性値として「現病歴」、「手術」、「手術後経過」が付与されている。現病歴から手術後経過までの一連の経過が記述された後に、<icd> タグが訓練データの末尾に属性名および属性値が正解例として示されている（<icd code="J931"> </icd> と同時に <icd code="Z720"></icd>）。

MedNLP2 では、「食欲不振」「嘔気」といった名詞ごとにそれぞれ病名コードを付与していたが、MedNLP-3 では現病歴、手術、手術後経過といった一連の経過に記述された語句群を手がかりにして最終的にシステムが ICD コードを推定するので、単純な文字列パターン照合のみでは不可能であるから、共起語などの概念を熟知していることが必要となり、タスクに参加するハードルが上がった。[19]。

そして NTCIR-13 では、医療言語処理のタスクとしては通算第 4 次に当たる「MedWeb」が設定され [20]、10 件の研究報告がなされた。そして、Twitter 上のツイートに焦点を合わせて、8 つの病気または症状（インフルエンザ、花粉症ほか）に罹患しているか否かを、陽性（Positive:p）または陰性（Negative:n）のいずれかに判定してラベルを付与するマルチラベル分類タスクが行われた [21]。対象となるツイートの言語は日本語、英語、中国語の 3 カ国語にわたる点も特徴の一つである。

2.5 その他　AI（人工知能）と医療

人工知能の性能向上が近年めざましく、「第三次 AI ブーム」と評されている。本稿の執筆日程の都合上、ウエブ上にアップロードされている資料のみの紹介に留まり、かつ紹介件数が下記のように少数であることは著者の不徳の致すところである。

例として Google において「医療 自然言語処理 AI」のようにして検索すると、自然言語処理を医療分野に応用していることに言及した荒牧 [22] による発表資料を閲覧できる。

また慶應義塾大学医学部と富士通が共同し、AI による診療支援を実現する技術を開発した。これは放射線科医が読影した画像検査報告書に、自然言語処理と機械学習が可能な AI 技術を適用し、入院などの要否を分類する機能を有する [23]。

3．今後の展望

本稿では最新の研究報告に目を向け、医療情報処理の学界動向をサーベイした。ICT（情報通信技術）の発展に伴い、研究者らが解析対象とするデータは、実在する医療機関から匿名化処理を経たものの他に、ウエブ上のツイートにまで拡がっていることもわかった。また学会発表もサーベイ対象とし、医療系かつ情報系に絞り込んで動向調査を行ったが、日本国内の英文の研究報告だけでも 2006 年頃と比較すると激増しており、隔世の感がある。今後は情報処理学会の年次全国大会のほか、MEDINFO などの海外の国際学会なども調査対象に加え、学界動向をさらに精度の高いものにしたい。

当該分野の学術図書としての専門書も、刊行ペースが早まり続々と刊行点数が増加していくと予想する。著者らは当論文を礎とし、医療情報学の分野において自然言語処理の知見を適用した各種研究を紹介する調査を

さらに続けてゆきたい。それにより、たとえば医療情報技師といった職業に従事する方々に対して、医療と自然言語処理の具体的な関わりを知るための「道しるべ」の役割を果たしたい。

[註釈]

[1] 西垣 通，伊藤 守：『よくわかる社会情報学』，ミネルヴァ書房（2015）.

[2] 千葉県匝瑳市ウエブページ：「オーダリング・電子カルテとは」，［http://www.city.sosa.lg.jp/index.cfm/18,23828,c,html/23828/20120717-142843.pdf（cited 2018-May-10)］.

[3] 奥村学，荒牧英治：『医療言語処理』，コロナ社（2017）.

[4] 日本医療情報学会医療情報技師育成部会 編：『医療情報　第 5 版　医療情報システム編』，篠原出版新社（2016）.

[5] 日本医療情報学会医療情報技師育成部会 編：『医療情報　第 5 版　医学・医療編』，篠原出版新社（2016）.

[6] 日本医療情報学会医療情報技師育成部会 編：『医療情報　第 5 版　情報処理技術編』，篠原出版新社（2016）.

[7] 柴田大作，若宮翔子，木下彩栄，荒牧英治：「音声発話による認知症スクリーニング技術の開発　　感情表現辞書を用いた発話内容の質的分析」，医療情報学，2017 Vol.37，No.6（2018）.

[8] 松村泰志，鳥澤健太郎，篠原恵美子，鈴木隆弘，荒牧英治:「（夢）自然言語処理技術の最前線と医療応用の可能性」，医療情報学，2018 Vol.38，No.1（2018）.

[9] 香川璃奈，篠原恵美子，河添悦昌，今井健，大江和彦：「医師の暗黙知に基づくタグ付きカルテコーパス作成の要件の検討」，言語処理学会 第 24 回年次大会 発表論文集，pp.757-760，（2018）.

[10] 矢野憲，岩尾友秀，荒牧英治：「MedInput: 病名の自動予測補完による医療テキスト入力支援ツールの構築」，言語処理学会 第 24 回年次大会 発表論文集，pp.1039-1042，（2018）.

[11] 松尾亮輔，Ho Tu Bao，池田満，田中孝治，陳巍:「医療文書分析の目的限局性に基づく単語階層の構成法の検討」，第 37 回医療情報学連合大会論文集，一般口演 5:2-G-1-OP5-1，pp.370-374，（2017）.

[12] 山ノ内祥訓，廣瀬隼，宇宿功市郎：「オープンエンド型発見手法と深層学習を用いた外来初診時記録からの候補病名予測の検討」，第 37 回医療情報学連合大会論文集，pp.501-506，（2017）.

[13] 国立情報学研究所ウエブページ:「NTCIRについて - NTCIR プロジェクト」,［http://ntcir.nii.ac.jp/jp/about/（cited 2018-May-10)］.

[14] 国立情報学研究所ウエブページ：「NTCIR Project 概要」，［http://research.nii.ac.jp/ntcir/outline/prop-ja.html（cited 2019-Jan-19)］.

[15] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe and Eiji Aramaki: “Overview of the NTCIR-10 MedNLP Task”,［http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/01-NTCIR10-OV-MEDNLP-MoritaM.pdf（cited 2018-May-10)］.

[16]「NTCIR-10 医療言語処理（MedNLP）パイロットタスクとは」，［http://mednlp.jp/medistj-ja/（cited 2018-May-10)］.

[17] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano and Tomoko Ohkuma : “Overview of the NTCIR-11 MedNLP-2 Task”,［http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/OVERVIEW/01-NTCIR11-OV-MEDNLP-AramakiE.pdf（cited 2018-May-10)］.

[18]「NTCIR11 MedNLP 2」，［http://mednlp.jp/ntcir11/index-ja.html（cited 2018-May-10)］.

[19]「NTCIR MEDNLPDOC（MEDNLP-3)」，［https://sites.google.com/site/mednlpdoc/（cited 2018-May-10)］.

[20] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma and Eiji Aramaki : “Overview of the NTCIR-13: MedWeb Task”,［http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/01-

NTCIR13-OV-MEDWEB-WakamiyaS.pdf（cited 2018-May-10)］.
［21］「NTCIR13 MedWeb」,［http://mednlp.jp/medweb/NTCIR-13/index_ja.html（cited 2018-May-10)］.
［22］「自然言語処理の医療応用 - 総務省」,
［https://www.soumu.go.jp/main_content/000474414.pdf（cited 2019-Jan-20)］.
［23］「慶應義塾大学医学部と富士通、AI による診療支援を実現する技術を開発」,
［http://pr.fujitsu.com/jp/news/2018/07/31.html（cited 2019-Jan-20)］.