日本語カルテをアノテートする

荒牧英治 * **　四方朱子 *　島本裕美子 *　久保圭 *
宮部真衣 *　大熊智子 ***　狩野芳伸 ** ****　森田瑞樹 *****
*京都大学　**科学技術振興機構 さきがけ *** 富士ゼロックス
****国立情報学研究所 ***** 東京大学
mednlp.office@gmail.com

1　はじめに

言語処理は，新聞や辞書中のテキストなどから，Web など扱う対象を広げつつある．これにともない，特許文章などドメインに特化した研究も活発になっている[1]．本研究では，医療コーパスに焦点を当てる．医療コーパスは，日常業務のための文章であり，多くの他のドメインのように公開を前提としておらず，効率化のための略記や非文法表現が含まれている．また，専門性も高い．このような文章を処理するためにはシステムの頑健さ，高度さなどこれまで以上の精度が必要であり，言語処理の新たな課題であると思われる．さらに，電子カルテを処理することは，超高齢化社会を迎える本邦にとって社会的重要性も高い（２章）．

以上のような背景から，我々は研究に利用することが可能な日本語のアノテーション済み医療文書を構築し，また，これを用いた解析タスクと共に，研究コミュニティに提供している[2]．本稿では，このコーパスの構築方法を述べ（３章），カルテの特徴を分析したのち（４章），アノテーション方針や問題について議論する（５章，６章）．

2　背景

2.1. 情報学における背景

これまで扱われてきたドメインは特許文章[1]，法令文[3]，旅行対話[4]など数多くあるが，これらはいずれも編集または校正された文章という点では共通している．一方，カルテ文章は，医療業務という実業務に用いられる文章であり，医師間の情報伝達，医師個人のための忘備録メモとしての側面がある．このような日常業務での文章は，非常に量が多いもののこれまで言語処理で扱われることはまれであった．カルテ文章の特徴は以下である．

- **効率的**： 記述時間を減らすために，略記，省略が頻繁に起こる．
- **非文法的**：誤字，脱字などが含まれる．
- **専門的**：専門用語が多用される．また，背景知識を前提にしているので言語化されていない事実が多い．

他のドメインにおいても旅行会話（非文法的な文も含まれる），特許文章（専門的な文も含まれる）など，上記の性質を部分的に持っているコーパスは存在する．しかし，すべてを持っている文章は，営業日報など機密性が高い場合が多く，公開される機会が少ない．本研究で提供する日本語カルテは，頑健な言語処理を開発するための材料となりうる．

2.2. 医学における背景

カルテ文章を扱う意義は社会的にも高い．近年，従来の紙を媒体とするカルテが急速に電子化されつつある．これに伴い，紙カルテの時代には事実上不可能であった大規模な医療情報の利活用が期待されている．これを実現するためには，病名，医薬品名などの専門用語や，それに対応するコードの標準化が必須となり，この自動化のために言語処理技術の利用が注目されている．実際，海外では 1960 年代から医療分野における言語処理研究が盛んに進められている[5,6]．しかしながら，他の分野と比べて進展が遅いという問題点も指摘されている[7]．その問題の原因としては，主に次の２点が挙げられる:

- 入手可能なコーパスが不足している．
- コーパスにおける仕様（アノテーション）の方針が統一されていない．

我が国においても，上記の問題に対する懸念は同様であり，日本語で書かれたアノテーショ

ン済みの医療文書を研究者が共有できるようなシステムや体制が望まれる．

3　カルテの収集方法

カルテ文章は，多くの個人情報が含まれ，その公開は容易ではない．この問題を緩和するために，我々は模擬病歴報告と国家試験問題という２つの模擬文章を利用した．

3.1 模擬病歴報告

本研究では，病歴報告 （medical report）からコーパスを構築した．病歴報告とは，入院患者が退院する際や，現在の患者を他の医師に紹介する際に，第三者がその患者の症例を理解することを目的として書かれる文書で，自然言語を用いて記述される．我々は，この病歴報告の書式を模して書き起こした「模擬病歴報告」を収集した．この模擬病歴報告の作成にあたっては，現実的な患者像を反映させるため，医師免許を取得している臨床医に依頼した．

3.2 医師国家試験の臨床実施問題

医師国家試験問題においては，患者の状況から医学的な知識を問う設問（以下，**臨床実施問題**と呼ぶ）がしばしば設けられており，その状況説明の形式は前述の病歴報告に類似する．そこで，厚生労働省のホームページ[8]で公開されている過去の医師国家試験問題から臨床実施問題に該当するものを抽出し，以下の手順で一部の記述に修正を施し，実際のカルテの文章に類似するものとした．

- **Step1**：臨床実施問題の症例のうち，現病歴や既往歴，家族歴が含まれる箇所を抽出する．
- **Step2**：Step1 によって抽出した文章から，質問部分や図解部分を除き，診断病名を加える．

4　カルテ文章の特徴

4.1. 効率的

執筆時間を短くするために，カルテ文章には略語が頻出する[9].また，以下の例のように様々な助詞や主語が省略される．

格助詞の省略

精査目的にて当科（φ）紹介入院となった。

格（「医師」や「患者」など）の省略

腹部単純 X 線写真で胸部異常影を（φ）指摘され当院紹介受診となる

4.2. 非文法的

誤表記がみられることがある．

表 1: 診療分野別の分野数

診療分野	模擬病歴報告	臨床実施問題
消化管・腹壁・腹膜疾患	6	11
肝・胆・膵疾患	3	10
内分泌・代謝・栄養疾患	11	9
腎・泌尿器疾患	4	9
免疫・アレルギー性疾患・膠原病	7	6
血液・造血器疾患	3	7
感染症	6	9
呼吸器・胸壁・縦隔疾患	17	11
合計	72	82

*全身が**向くんで**歩けない*

*指に**後半**が出現。*

*電子顕微鏡結果**をを**待っている。*

*ＨｂＡ１ｃ　９．１％と**コントロー**不良の糖尿病のため*

また，一文内で動作主が異なる動詞句が見られる．

2006 年 10 月 2 日（φ=患者は）自覚症状がなく、白血球増多、…の上昇を（医療者が）認め、慢性骨髄性白血病を（φ=医療者が）疑った

これらの非文法的な表現はカルテ文章の特徴であると考えられるため，コーパスではそのまま残した．

4.3. 専門的

医療文書において，「発熱なし」などの否定表現 （negation） は多く出現し，症状に関する表現においては，30%が否定表現となっている[9]．一般には，否定表現が文脈なく使用されるのは不自然であるが，医療文書においては予想されるべき症状が「ない」という記述により，医師が何を考えて／疑っていたかが分かる場合があり，重要な情報になる場合がある．また，「ない」ことは少なくとも検査をしたことを意味し，診断ミスがあったかどうかの判定にも用いられる．このため，コーパスには，症状が「ある」のか「ない」のかを示すモダリティ情報を付与した．

また，非医療従事者には理解できない専門的な表現が含まれこともある．

波動を触れる

肝・脾を触知しない

「*波動を触れる*」については，腹水などを調べるために，医師が患者の腹部に手をあて，水分を多く含んでいる柔らかい感触を得ることを示している．「*触知しない*」については，肝・脾

を触知しようと医師が患者の腹部に手を当てたが患者のこれらを触知できなかったことを示す．これらの専門用語については，情報処理研究者の理解を助けるために解説を用意した（表２）．全文はウェブから閲覧可能である[1]．

5　アノテーション

本研究では，カルテには患者のイベントが時系列に記述されるということを考慮し日付や時間に関わる表記（以後，日時とよぶ）と「症状と診断表現」に関わる表記に対してタグを付与した．さらに，「症状と診断表現」に対しては，表記ゆれを吸収するために疾病コード属性を，事実性の有無をマークするために「モダリティ」属性を付与した．

5.1.時間表現タグ：<t>

日時にタグ<t>を付与した．この作業は以下の方針に基づいて行った．

数詞の「時間」単位にタグを付与する．数詞を含まない場合でも，前後の文脈より日付または時間のいずれかの特定が可能な場合にはタグを付与する．

<t>４月８日</t>再度上部消化管内視鏡施行し，<t>同日より</t>三分粥の潰瘍食開始

絶対時間だけでなく，相対時間も多くみられ．これらを対象とした．

発症から<t>３時間以内</t>である

ただし，病状や治療に関連がないと思われる情報は扱わないものとした．

40 年来の専業農家.

5.2. 症状と診断表現タグ： <c>

病状，患者の症状，病名等にタグ<c>を付与した．この作業は以下の方針に基づいて行った．

複合名詞はまとめてタグを付与した．

<c>両側肺門リンパ節腫大</c>

「結核菌」のように，菌が存在することで特定の病状が決定される場合は，その有無に関してもタグを付与した．

喀痰Ｇｒａｍ染色で，<c>Ｇｒａｍ陽性双球菌</c>の白血球による貧食像を多数認める.

単なる検査表現であっても，それ自体で疾病コードが割り当てられている場合はタグを付与した．

<c>ツベルクリン反応陽性</c>.

治療や手術の名称に含まれる病名にはタグを付与しない．

左下肢静脈瘤手術を行った.

5.3 モダリティ属性

また，症状と診断表現に関しては，モダリティを属性として記述した．本研究では，以下の４種類のモダリティを扱う．

Positive	その症状が実際に認められた	３７℃台の<c>微熱</c>を主訴に来院した.
Negation	その症状が実際には認められなかった	<c modality="negation">圧痛 </c> や <c modality="negation">発赤</c> なし.
Suspicion	病名についての推測や疑い	当初 <c modality="suspicion">両側丹毒</c>が疑われた
Family	家族歴としての病名	９人兄弟で姉２人が<c modality="family">糖尿病</c>.

モダリティを２種類以上伴う場合は，カンマで区切り併記した．

<c modality="negation,family">大腸癌</c>の家族歴なし

5.4 コード属性

症状と診断表現に関しては，ICD（疾病及び関連保健問題の国際統計分類）コードを属性として付与した．本研究において使用したのは、2003 年に改訂された第 10 版（ICD-10）である．ICD-10 は疫学調査など国家レベルでの国際的調査を想定して作成されたものであり，発展途上国でも使用できる配慮のもと，複雑な診断基準は設けられていない．コードは 1 文字のアルファベット（大分類）と３～４桁の数字（細分類） により構成されている．

6　アノテーションの問題と対策

前章までの方針でアノテーションを行うと，実際には様々な問題や境界例や例外が存在する．これらのうち，主要なものとその対策を以下に記述する．

6.1. 専門知識や文脈から推測されるアノテーション

「腫瘍」などを始め，多くの ICD コードは，その発生部位のよってコードを異にする．一方，カルテ中において，部位は一度記述されると諸略されやすい．コーパスでは，文脈から部位を判断し，その部位に即したコードを付与した．

[1] http://mednlp.jp/ntcir11/words20131113.pdf

表 2：専門用語の解説（抜粋）

用語	説明
清明	意識などがはっきりしている様子
造影剤	画像診断の際、画像に明暗の差を付けたり、特定の組織を強調して撮影するために患者に投与される医薬品
奏功	目標どおりの成果があがること
治験	医薬品もしくは医療機器の製造販売に関して、薬事法上の承認を得るために行われる臨床試験
肺野	肺の外側の面積約 50 %の領域のこと。
頻回	回数が多いこと。また、多くの回数。
不穏	周囲に対する警戒心が強くなり、興奮したり暴力をふるったりしやすくなる状態
緩病	患者に対する全身的な医学管理の中心となっていない疾患
膨隆	皮膚や粘膜などの局部的なふくらみ
予後	手術後の患者の状態や、病気・けがの将来的な状態に関する見込み
ラ音	肺または気管支の疾患の際に聴診器で聴かれる、呼吸に伴う肺部の雑音
レジメン	がん治療で、投与する薬剤の種類や量、投与期間、手順などを時系列で示した計画書

6.2.名詞句を超えた表現

名詞句の範囲を超えて症状が表現されることがある．例えば，「*熱がでる*」にみられるように，症状が名詞句におさまらず，動詞句を含む構造を持って現れる場合は，これをアノテーションしなかった．これは下記のように文全体で倦怠感が表現されている場合など，アノテーションを行う箇所を特定することが困難であるためである．

日中はだるいのに，夜はなかなか寝つけない，とのこと．

6.3. 患者症状でない疾患表現

患者の症状ではない疾患表現がある．例えば，下記の例のける「リウマチ」は施設名の一部であり，これは患者の症状を指すものではない．このような表現についてはアノテーションの対象外とした．

かかりつけの■■■内科リウマチ科クリニックより当院紹介受診となった．

上記の他にも様々な境界例や言語学的に興味深い例が存在する．詳細はウェブを参照されたい[2]．

7　まとめ

本研究では，日本語のアノテーション済み医療コーパスの構築方法および，その使用について述べた．今後は，ひきつづき，本コーパスを用いた解析タスクと共に，研究コミュニティに提供を予定している．

謝辞:本研究は，JST 戦略的創造研究推進事業（さきがけタイプ），NII NTCIR 共同研究費及び，科研費補助金(若手研究 A)による．本論文を書くにあたって有益な議論をいただいた岡田真穂氏に感謝いたします．

[2] http://mednlp.jp/ntcir11/

参考文献

[1] Shoichi Yokoyama: Machine Translation Summmit XIV Workshop Proceedings on Patent Transltion

[2] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, Eiji Aramaki: Overview of the NTCIR-10 MedNLP task, In Proceedings of NTCIR-10, 2013. (2013/06/18, Tokyo, Japan)

[3] 岩本秀明，野村浩郷: 法律文の自然言語処理について，1991，37 (1991-NL-083).

[4] Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, Eiichiro Sumita: Multilingual Spoken Language Corpus Development for Communication Research, Computational Linguistics and Chinese Language Processing. Vol. 12, No. 3, September 2007, pp. 303-324.

[5] SNOP (Systematized Nomenclature of Pathology) by College of American Pathologists, 1965.

[6] SNOMED-CT, 1974.

[7] Wendy Chapman, Prakash Nadkarni, Lynette Hirschman, Leonard D'Avolio, Guergana Savova, Ozlem Uzuner. Over coming barriers to NLP for clinical text: the role of sharedtasks and the need for additional creative solutions. JAmMed Inform Assoc,18, 540–543, 2011.

[8] 第 107 回医師国家試験の問題および正答について（http://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp130723-01.html）

[9] Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, Kazuhiko Ohe: TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification, Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing}, BioNLP '09, 185-192.

[10] Emiko Shinohara, Eiji Aramaki, Takeshi Imai, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kazuhiko Ohe: An easily implemented method for abbreviation expansion for the medical domain in Japanese text: A preliminary study, Methods of Information in Medicine 2013; 52 (1): 51-61.