大規模医療コーパス開発に向けて

荒牧英治* 岡久太郎* ** 矢野憲* 若宮翔子* 伊藤薫*

*奈良先端科学技術大学院大学 研究推進機構
**京都大学大学院 人間・環境学研究科

{aramaki,taro-o,yanoken,wakamiya,kito} @is.naist.jp

1. はじめに

高度な医療人工知能研究のためには，その材料となるデータが必須である．しかし，医療，特に臨床に関わる分野では，そのデータが使いやすい形で存在しない．データとして考えられるリソースは，医学教科書，論文や電子カルテ文章など様々であるが，いずれも，相当な部分は自然言語文である．特に， 電子カルテ文章については，大学病院規模では一ヶ月に20万以上もやり取りされるともされ，膨大な量があるものの，記述内容の具体については，医療従事者に一任されてきた面が多い．このため，カルテ文章を二次利用するためには，非文法的かつ断片化したカルテ文章から医療情報，例えば傷病名や愁訴，検査，治療に関する記述を抽出し，標準化する処理が必要となる．この標準化のためには，自然言語処理による自動化が必須となる．

一方，近年の自然言語処理は，コーパスベースの手法が中心であり，システムを開発するためには アノテーション付きコーパスが必須となる．しかし，現在，利用可能なアノテーション付きコーパスは，我々の知るかぎりでは GSK 模擬診療録テキスト・データ[1]や NTCIR MedNLP コーパス[2][1]などいずれも数十文章と量が希少であり，コーパスベースの研究が困難な状態である．

このような背景の中，我々は，医療分野の数万文章の大規模なコーパスを構築すべく，開発を行っている．材料として，カルテと類似した論文である症例報告を扱う．タグ付け内容は，もっとも重要な情報の 1 つと考えられる病名や症状(以降，本稿では単に**病名**と呼ぶ)についてその出現の有無とする．

本稿では，進行中のこのコーパス構築のプロジェクトについて述べるとともに，医療テキストへのタグ付けで発生する諸問題とその解決法を議論する．

[1] GSK2012-D 模擬診療録テキスト・データ http://www.gsk.or.jp/catalog/gsk2012-d/
[2] MedNLP ワークショップ http://mednlp.jp/

2. 関連研究

海外では，2006 年から i2b2 NLP Challenge がコーパスの整備を進めつつある[2]．i2b2 NLP Challenge は，タスクを変えながらコーパスをリリースしており，これまで，匿名化と喫煙歴(2006)，肥満(2008) ，投薬(2009)，関係抽出(2010) ，カルテ中の照応関係(2011)などが扱われてきた．

本邦においては，GSK 診療録コーパスにて診療録コーパスが，NTCIR MedNLP コーパスにて退院サマリコーパスが，小規模ながら作成されている．両コーパスは，時間表現，病名，医薬品，個人情報，および，病名についてはモダリティ（事実性）が付与されている．

本コーパスは，これを病名に絞った．これは，医薬品情報は医薬品オーダーに記載されており，また， 日付情報もファイルのタイムスタンプとして記録されているなど，他の情報を利用することでより正確な情報が入手可能なためである．そこで，もっともカルテから抽出が必要な症状のみに絞った．また，事実性ももっとも重要な分類である陽性と陰性の 2 値に簡略化した．このように簡略化したものの，かつてない大規模なデータ構築に挑んでいる．

3. アノテーション

3.1 材料:症例報告

症例報告とは，一人の患者を対象とした報告形式の論文である．分類としては論文の一種であるものの，対象となる患者が 1 名であること，手法部分に相当するものがなく，観察した結果と考察が主な内容となることなどから，サマリと呼ばれる形式のカルテと類似している．そのため，将来的にカルテへの適応拡大が可能であると思われる．

3.2 方法

医療テキストには専門用語が多く含まれており，正確なアノテーションのためには医学知識が必須となる．例えば，「DIC あり」における「DIC」が＜**播種性血管内凝固**＞という疾患を指すといった判断は，前後の文脈を用いても非医療従事者にとっては容易

ではない．一方，医療従事者は，DIC は頻出するため，容易な作業である．そこで，すべての作業を医療従事者によって行うのが理想的である．しかし，実際にはコスト的にも人材的にも困難である．最も人数が多い医療従事者は看護師であるが，慢性的に人手不足であり，また，雇用単価も高い（時給換算で 2000 円～2300 円）．また，そもそも，看護師の本来の職務とタグ付け作業は大きく乖離しており，熱意を持ってタグ付けに従事可能な人材の確保は容易でない．

そこで，本研究では，タグ付けに熟練し，かつ，事務作業とも親和性の高い，医療事務経験者を中心に雇用し，タグ付けを行う．ただし，医療事務経験者は看護師と比較し，そもそも人数が多くない．そこで，タグ付けのプロセスを専門知識が必要な箇所とそうでない箇所の 2 つに分け，必要な部分のみ，医療従事者が担当する．

- **STEP1：病名特定作業**
  医療テキストから，病名に相当する部分をタグ付けする．病名かどうか分からない場合は，すべて病名扱いとする．この際，患者に認められる病名には陽性タグ（<P>タグ）を，患者に認められない病名には陰性タグ（<N>タグ）を付与する．非医療従事者が担当する．
- **STEP2：病名コーディング作業**
  タグ付けされた病名をコーディングし，病名コードと標準病名を属性として付与する．コーディング不可能なもの（例えば，曖昧すぎる病名「副作用」「異常」や病名でないもの「STAGE I」）はこの段階で削除される．医療従事者が担当する．

以下に例を示す：

> *<P code="R104: 腹痛">腹痛</P>により来院，急性の<N code="A09: 感染性胃腸炎">胃腸炎症状</N>を疑う*

タグ（陽性<P>，陰性<N>）の付与が病名特定作業である．ICD コードの分類体系に沿って病名コードを付与するのが病名コーディング作業である．

4. STEP1：病名特定

4.1 原則

病名に対するタグ付けでは，以下の 3 つの目的に適うものを目指している．

(i) 非医療従事者がタグ付けを行いやすい基準を設ける．
(ii) 症例報告の患者に関する情報を整理する．
(iii) 病名コードが付与できる表現を出来る限り多く採取する．

これらの目的を達成する上で問題となってくるのは，(i) においては複数の離れた語が病名と対応する場合にタグ付けの範囲決定が難しくなること，(ii)においてはデータに含まれる全ての病名にタグ付けする場合，症例報告の患者と無関係のものも抽出されてしまうこと，(iii)においては非医療従事者には病名コードが付与可能かどうかの判断が難しいことである．これらの問題を解決するために，以下の 7 つの原則を設定した．

(I-a) 名詞（名詞または複合名詞）で表現される病名に対してのみタグ付けを行う．
(I-b) 記号（または，それに類する表現）・程度表現は，病名を表す複合名詞の一部となっている時のみタグ付けを行う．

(II-a) <N>タグを付与するのは，患者に生じていない病名（陰性所見）とする．
(II-b) 治癒を表す表現が伴っている病名については，症状や疾患が完全に消失したことが明示されている場合のみ<N>タグを付与する．
(II-c) 複合名詞に含まれた病名は，患者の疾患・症状ではない場合，タグ付けを行わない．
(II-d)データの末尾にしばしば記述される考察や一般論には<SKIP>タグを付与し，<SKIP>以降の病名は<P>タグや<N>タグの付与対象としない．
(III)「陽性」「陰性」の表記に関しては，複合名詞として特定の病名（ないしはその否定）を表す場合を想定し，タグ付けの対象とする．

4.2 各原則の詳細

4.1 節で述べた原則の詳細を具体例とともに示す．

4.2.1 原則 (I-a) について

タグ付け範囲を明確化させるため，タグ付け対象は名詞に限る．そのため，(1) のような単独の名詞，(2) のような複合名詞，(3) のような英語名，その略称にはタグ付けするが，(4) のような動詞や (5) のような修飾句を伴う名詞句全体にはタグ付けしない．

(1) *<P>肺小細胞癌</P>と診断された*
(2) *<P>両側肺門リンパ節腫脹</P>を認め*
(3) *<P>carcinoid tumor</P>，<P>ly1</P>，<P>v1</P>であった。*
(4) *右膝が<u>腫れてきた</u>*
(5) *<u>ポリープ状の</u><P>腫瘍</P>を認め*

ただし，(6,7) のようなサ変動詞の語幹については，それ単体で疾患・症状を表す名詞と認定できるものに限ってタグ付け対象としている．

(6) *<P>狭窄</P><u>している</u>と考えられた*
(7) *<P>腎機能低下</P><u>する</u>*

4.2.2 原則 (I-b) について

タグ付け範囲を統一するために，タグ付けは可能な限り最長となるような名詞に対して行う．そのため，中黒点（・），スラッシュ（/），ハイフン（-），コンマ（，），読点（、）等の記号やそれらに類する「及び」のような表現で区切られたものについては，以下の 2 つ

の場合に分けた. (a) その記号の前後が独立した名詞として認定される場合は別々の名詞としてタグ付けする (8, 9). (b) 前部ないしは後部がもう一方と連結して複合名詞を作る場合は, 当該の記号を挟んでタグ付けする (10).

(8) <P>血痰</P>・<P>下血</P>も出現した
(9) <P>全身性エリテマトーデス</P>（<P>ＳＬＥ</P>）と診断され
(10) <P>ウイルス性，細菌性肺炎</P>疑いあり
【ウイルス性肺炎＋細菌性肺炎】

また, (11,12) の例のように, 症例報告では「↑」(上昇, 増加), 「↓」(下降, 減少), 「(+)」(陽性), 「(−)」(陰性) といった記号が用いられるが, これらは当該の文脈において対応する単語に置き換えた上で, (I-a) の基準に照らし合わせてタグ付けの判断を行う. なお, 「(+)」「(−)」に関しては, 原則 (III) も参照されたい.

(11) <P>皮膚ツルゴール↓</P>
(12) <P>項部硬直(+)</P>

程度表現については, 「...程度」「...傾向」「...stage I」のように疾患・症状を表す名詞に後続し, 1 つの複合名詞をなす場合は, タグ内に含める (13-15). なお, 程度表現と病名との間に助詞が存在する場合は, 原則 (I-a) により, 程度表現にはタグを付与しない (16).

(13) <P>血痰程度</P>の軽度の症状
(14) 体幹下肢の<P>皮膚病変痂皮傾向</P>
(15) <P>胃癌 stageIV</P>の診断
(16) Stage IVb の<P>進行食道癌</P>

4.2.3 原則 (II-a) について

＜N＞タグを付与するのは, 「見られない」「認められない」のような否定表現が用いられている場合 (17), または「危惧される」「危険性が高い」のような, 発生が予見されるがまだ生じていない事態を表す表現が用いられている場合 (18) である.

(17) 気管や大動脈への<N>浸潤</N>は見られなかった
(18) <N>肺塞栓症</N>の発症が危惧された

なお, 「疑う/疑われる/疑い」は文脈によっては, (19) のように, 当該の事態が実際には否定される場合が存在するため, この場合に限り, ＜N＞タグを付与している.

(19) 当初<N>胆道感染症</N>を疑い抗生剤(SBT/CPZ、MEPM)使用するも効果認めなかった。

4.2.4 原則 (II-b) について

治癒を表す表現が伴う病名については, 当該の疾患・症状が完全に消失したことが明示されている場合に限り＜N＞タグを付与し, それ以外は全て＜P＞タグを付与している. (20-24) の例では, (20) のみが＜N＞タグ付与の対象であり, それ以外は＜P＞タグ付与の対象となる.

(20) <N>リンパ腫所見</N>は消失しており
(21) <P>腫瘤</P>はほぼ消失していた
(22) <P>発疹</P>は消退傾向となり
(23) <P>心不全症状</P>の軽快を認めた
(24) <P>腹痛</P>は改善した

4.2.5 原則 (II-c) について

タグ付けの対象は, 患者に関する病名であるため, (25,26) のような検査名, 施設名等の一部となっている名詞にはタグ付けを行わない.

(25) 梅毒定性は陰性であった
(26) ■■■内科リウマチ科クリニックより

なお, (27,28) のように, 厳密にはそれ自体が疾患・症状名ではない複合名詞であっても, 患者の疾患・症状を表していると思われるものにはタグ付けしている.

(27) <P>急性肝炎様</P>に<P>AIH</P>を発症
(28) <P>壊疽部</P>に痛みが残る

4.2.6 原則 (II-d) について

3 節で述べた通り, 今回扱っている材料は症例報告であるため, 多くのデータがその末尾に報告の意義や考察を含んでいる. これらの文章は, 実際にその報告内で取り上げた患者の病名というよりも, 一般論としてその疾患や症状を記述している性格を強く有している. 我々は, 症例報告の対象患者の疾患や症状についての情報の整理を目指しているため, このようなデータ末尾にある総括的文章はタグ付けの対象外となる. このタグ付け対象外の範囲を表すために, ＜SKIP＞タグを設けた. ＜SKIP＞タグ以降の文章は解析対象とならない. (29) はあるデータにおける末尾の 2 文であり, 最後の 1 文の前には<SKIP>タグが付与される.

(29) 摘出標本の病理組織では、嚢胞壁に絨毛上皮と軟骨、平滑筋の存在を認め、<C>気管支性嚢胞</C>と診断された。<SKIP>気管支性嚢胞は先天性嚢腫の一つで、後腹膜由来のものは、我々が医中誌で検索した限りでは本邦で約50例と比較的まれであり、若干の文献的考察を交え報告する。

なお, ＜SKIP＞タグは, それ以後の文章をタグ付け対象外とするため, データの末尾よりも以前に登場した一般論等に関しては, (30) のように＜N＞タグを付与することで実際に患者に認められた病名とは区別している.

(30) *<N>平滑筋肉腫</N>は、進行又は再発症例では有効な治療法が確立されておらず、その予後も不良であるのが現状である。今回我々はGemcitabine／Docetaxelの化学療法にて予後の改善が得られた後<P>腹膜原発平滑筋肉腫肝転移</P>の1例を経験したので報告する。…*

4.2.7 原則 (III) について

病名コードを有する病名の中には「ツベルクリン反応陽性」(ICD-10 コード: R761)のように，病名内に「陽性」というモダリティ表現を含むものが存在する．しかし，STEP1 の病名特定作業を行うのは非医療従事者であり，彼らは病名コードに関する知識を有していない．そこで，病名コードを持つ病名をより多く採取するために，「陽性」(「(+)」)，「陰性」(「(-)」)に関しては，(31-32)のように，複合名詞を構成する場合，まとめてタグ付けを行う．

(31) *<P>ツベルクリン反応陽性</P>を認めた*
(32) *<N>HBs 抗原陰性</N>、<P>HBs 抗体陽性</P>、<N>HBV-DNA 陰性</N>であった*

なお，(33)のように「陽性」「陰性」という表現が述部において用いられている場合は，原則(I-a)に基づき，タグ付けを行わない．

(33) *Tg 抗体は陽性、TSB 抗体は陰性であった。*

5. STEP2: 病名コーディング

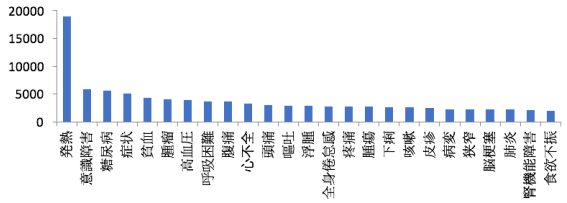

図 1: 病名の出現頻度．頻度が高いものから 25 病名．

STEP1 で特定された病名を正確に機械処理するためには表記ゆれを吸収する必要がある．これは症状にコードを振ることで対応する．図1に病名の出現頻度を示す．他の単語と同じく，病名の出現もベキ則に沿いロングテール状の頻度分布を示す．この性質から高頻度のものからコーディングを行うことで，着手時から高い被覆率を実現することができる．

コーディングに際しては既存の辞書リソース(標準病名マスター[3]を用いた)を可能な限り利用し，既存のリソースでカバーできないものについて複数人でコーディングする．

現状では，62,610 病名(異なり)が出現し，そのうち 12.5%(7,854 病名)のみがマスターにマッチし，それ以外の 87.5%(54,756 病名)がマッチしていない．

これらについて，医療従事者 3 名が以下の 3 つのフェーズに分けて作業している．

- **フェーズ１　合意形成コーディング**
  コーディング作業者 3 名が同一病名に対して，個別にコーディングし，差異があれば議論し，合意を得る．
- **フェーズ２　半独立コーディング**
  コーディング作業者は独立して，それぞれ別個のパートをコーディングするが，自信のない部分については，他の作業者を招集し，合意を得る．
- **フェーズ３　独立コーディング**
  コーディング作業者は独立して，それぞれ別個のパートをコーディングする．

我々は，フェーズ1に頻度 30 回以上の高頻度用語(5,600 病名)，フェーズ 2 に頻度 30〜10 回の中頻度用語(2300 病名)，フェーズ 3 に頻度 10 回未満の用語を割り当てている．現在(2017 年 1 月 8 日)はフェーズ 3 を行っている最中である．

6. おわりに

本稿では，医療テキストのタグ付け，コーディングの仕様と方法について述べた．本コーパス開発で扱うのは病名のみであり，直感的には容易に基準を定めることができるように思える．しかし，このような病名カテゴリについても，実際の医療コーパスにタグ付けを行うことはしばしば難しく，本研究が引き続き挑むべき課題となっている．なお，コーパスの構築および配布は今後の予定である．

謝辞

本研究の一部は，JSPS 科研費 JP16H06395, JP16H06399, JST, ACT-I, および厚生労働省科学研究費補助金(課題番号: H28-ICT-一般-008)の支援を受けたものです．

参考文献

1. Aramaki E, Morita M, Kano Y, Ohkuma T. Overview of the NTCIR-11 MedNLP-2 Task. NTCIR-112014.
2. Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:1252-3. PubMed PMID: 18998924.

[3] http://www.dis.h.u-tokyo.ac.jp/byomei/