［別添 3］

厚生労働科学研究費補助金 政策科学総合研究事業

(臨床研究等 ICT 基盤構築・人工知能実装研究事業 総括研究報告書)

カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築，及び，自動構造化機能を有した入力機構の開発

研究代表者　荒牧英治 奈良先端科学技術大学院大学 研究推進機構

研究要旨

電子カルテは患者情報が全て記録されているものの，非文法的かつ断片化した表現が多く自然言語処理を応用した利活用は困難であった．これを二次利用するため申請者等は電子カルテから医療用語の自動抽出及び自動コーディングを行う研究に従事してきた．本研究は，この研究を発展させ，カルテ文章の解析において，実用化可能な精度を達成し，大規模な医療用語辞書の構築，及び，実際に電子カルテ入力において寄与するシステムを構築する．

若宮翔子（奈良先端科学技術大学院大学 研究推進機構･特任助教）

河添悦昌（東京大学大学院医学系研究科･特任准教授）

A.研究目的

これまで，医療医学用語をまとめる試みは多く，多くの医学大辞典が出版されてきた．しかし，これまでの多くの用語リソースは，トップダウン的なアプローチで専門家が定義したものであり，医療の臨床現場で実際に使用されている用語と乖離している場合もある．このため，カルテ入力をサポートするシステムを作る際に，実際に入力したい用語が収載されていないことも起こりえた．

そこで，本研究では，自然言語処理により用語を抽出する機構を開発する．その結果，得られた用語を精査して辞書にする．最後に，これをベースに入力支援アプリケーションを開発する．これは，実際に臨床現場で記述された用語を材料に，高次のシステムを作るため，ボトムアップのアプローチといえる．一見，自然で簡単なアプローチに思えるが，これを実際に行うためには，自然言語処理の利用が必須となる．例えば，次の例を考える：

その後，むかつきがあり受診

の表現の下線部「むかつき」は嘔吐･悪心（ICD10 コード R11）に相当すると考えられ，このような表現をカルテから抽出したいと考える．しかし，この表現を抽出するためには，「むかつき」が症状であることが辞書に記載されている必要がある．すなわち，用語を抽出するためには一般には辞書が必要であり，辞書に掲載されていない用語を抽出するためには，辞書に頼らない抽出方法を用いることが必須となる．

このジレンマを解決するために，本研究では，医療従事者が記載した電子カルテや退院サマリから症状や病名に関連する用語を辞書を用いない自然言語処理手法を用いて抽出し，そのデータを精査して「万病辞書」[1]として辞書化し公開している．本稿では，「万病辞書」のファイル構成や統計について報告する．

B.研究方法

B-1. 2010 年 1 月 1 日から 2016 年 12 月 31 日の期間を対象として， 東京大学医学部附属病院の電子カルテに記載された診療記録を抽出した．

B-2. B-1 で抽出した診療記録を入力として，奈良先端大学のソーシャル・コンピューティング研究室で開発した病名抽出ツール（mednlp parser v006）で処理を施し，症状・所見・疾患を抽出した．

（倫理面への配慮）

研究の実施に際しては，奈良先端科学技術大学院大学情報学系の倫理承認（承認番号 2016-I-30）および東京大学大学院医学系研究科の倫理承認（承認番号：11446）を得て行った．

C.研究結果

結果として，退院サマリから 22,434 病名，診療記録から 18,691,219 病名が抽出された．ただし，これらすべてが本当の病名でなく，解析エラーも含まれるため抽出された表現には病名として不適切なものも存在する．これらを整理し「万病辞書」として公開している．

本研究では，ICD-10 対応標準病名マスターの病名（ICD10 対応標準病名マスター V4.04，2018 年 4 月 1 日改訂 [2] を利用）を含み，それに加えて医療現場で得られる症状や病名を備えた「万病辞書」を作成している．2019 年 3 月末時点で，3 つの施設から抽出・精査した 362,866 件の病名用語(うち,25,678 件が標準病名)を収載している．同時点で,33,239 件の高頻出の病名表現について医療従事者（最大 3 名）によるコーディング [3, 4] が施されており，残りについては機械学習などにより自動的に結果を付与している [5, 6]．なお，コーディングの信頼度を明示するために，標準病名マスターに記載されているもの，人手でコーディングされたもの，機械により自動コーディングされたものなどをそれぞれ区別している．また，人手でコーティングされたものについては，1 名がコーディングしたものと 2 名以上がコーディングしたものを区別し，さらに，後者についてコーディング結果の一致度を考慮した区別を行い，辞書リソース化している．さらに，日本語形態素解析器として代表的な Mecab 用辞書も作成して提供している．

万病辞書の抜粋を図 1 に示す．万病辞書の公開版は以下の 5 つの項目から構成されている．

出現形	ICDコード	標準病名	信頼度LEVEL	しゅつげんけい;icd=ICDコード/lv=信頼度LEVEL/freq=0;標準病名
皮疹	R21	発疹	A	ひしん;icd=R21/lv=A/freq=高頻度;発疹
嘔吐	R11	嘔吐症	A	おうと;icd=R11/lv=A/freq=高頻度;嘔吐症
痛み	R529	疼痛	A	いたみ;icd=R529/lv=A/freq=高頻度;疼痛
腹水	R18	腹水症	A	ふくすい;icd=R18/lv=A/freq=高頻度;腹水症
咳嗽	R05	咳	A	がいそう;icd=R05/lv=A/freq=高頻度;咳
骨髄抑制	D758	骨髄機能低下	A	こつずいよくせい;icd=D758/lv=A/freq=高頻度;骨髄機能低下
肝転移	C787	転移性肝腫瘍	A	かんてんい;icd=C787/lv=A/freq=高頻度;転移性肝腫瘍
しびれ	R208	しびれ感	A	しびれ;icd=R208/lv=A/freq=高頻度;しびれ感
肺転移	C780	転移性肺腫瘍	A	はいてんい;icd=C780/lv=A/freq=高頻度;転移性肺腫瘍

図 1. 万病辞書の抜粋

(1) 出現形

電子カルテや退院サマリから抽出された症状・病名である．すべて全角に変換済みである（例：１１β-水酸化酵素欠損症，１８常染色体異常など)．人手により，明らかな抽出ミスやノイズと考えられる語は除去している．一方，明らかに誤字であると思われる場合でも，その頻度が高い場合は，実際の医療現場で用いられている語として採用した．

(2) ICD コード

ICD10 対応標準病名マスター [2] に記載されている病名コードである．出現形が ICD10 対応標準病名マスターの標準病名と一致する病名については対応する ICD10 コードを割り当て（(4) 信頼度 LEVEL: S)，そうでない病名については，人手（(4) 信頼度 LEVEL: A から C)あるいは機械((4) 信頼度 LEVEL: D）により付与している．最終年度に新規に追加した語に関しては，それまでの万病辞書に類似する語があれば，その ICD10 コードを割り当て（(4) 信頼度 LEVEL: E)，そうでない場合には，暫定的に ICD10 コードなし（(4) 信頼度 LEVEL: F）とした．信頼度 LEVEL: D から F の語については，診療録における頻度順に人手により割り当てを見直し，信頼度 LEVEL の更新を行なっている．なお，下記に該当する病名については -1 を付与した．

・4 つ以上のコードが存在する場合（3 つまでは全て付与）
・出現形から判断が困難な場合（出現形がノイズである場合，その病名は除去）
・ICD コードが存在しない場合

(3) 標準病名

ICD10 対応標準病名マスターに記載されている標準病名である．出現形が ICD10 対応標準病名マスターの標準病名と一致する病名については対応する ICD10 コードを割り当て((4) 信頼度 LEVEL: S)，そうでない病名については，人手（(4) 信頼度 LEVEL: A から C）あるいは機械（(4) 信頼度 LEVEL: D）により付与している．(2) ICD コードと同様に，最終年度に新規に追加した語に関して

は，それまでの万病辞書に類似する語があれば，その標準病名を割り当て（(4) 信頼度 LEVEL: E），そうでない場合には，暫定的に標準病名なし（(4) 信頼度 LEVEL: F）とした．信頼度 LEVEL: D から F の語については，診療録における頻度順に人手により割り当てを見直し，信頼度 LEVEL の更新を行なっている．

(4) 信頼度 LEVEL

病名に対する ICD10 コードおよび標準病名のアノテーション方法に基づき信頼度を付与している．以下の 7 つの LEVEL を付与している．図 2 に信頼度 LEVEL ごとの件数を示す．

・S: ICD10 対応標準病名マスターに記載されている病名
・A: 2 名以上の医療従事者が同じコードを付与した病名
・B: 2 名以上の医療従事者が相談してコードを付与した病名
・C: 1 名の医療従事者がコードを付与した病名
・D: 計算機が自動的に割り当てた病名
・E: S から C の病名とのマッチングにより計算機が自動的に割り当てた病名
・F: S から C の病名とのマッチングにより計算機が自動的に割り当てられなかった病名

(5) しゅつげんけい;icd=ICD コード/lv=信頼度 LEVEL/freq=0;標準病名

出現形の読み仮名(しゅつげんけい)，ICD コード，信頼度 LEVEL，freq，標準病名から作成した複合文字列のラベルである．

・出現形の読み仮名：
- 信頼度 LEVEL が S の病名：ICD10 対応標準病名マスターに記載されている病名表記カナをもとに付与している（全角，アルファベットや数値はそのまま）
- 上記以外の病名：読み仮名を自動付与（アルファベットや数値も読みに変換）し，一部を人手により修正している．

・freq: 特定の病院における病名の頻度をもとに以下の 4 区分に分類している．
- 高頻度：50 件以上
- 中頻度：5 件以上 50 件未満
- 低頻度：3 件以上 5 件未満
- レア：3 件未満

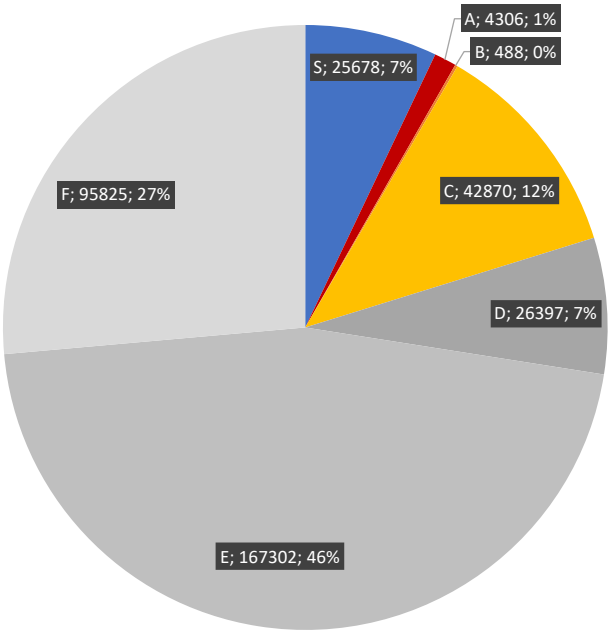

図 2. 信頼度 LEVEL ごとの件数（2019 年 5 月半ば時点）．データラベルは「信頼度 LEVEL; 件数; パーセンテージ」を表す．

D. 考察

2019 年 3 月末時点で，万病辞書には 3 つの施設から抽出・精査した 362,866 件の病名用語（うち，25,678 件が標準病名）が収載された．本事業終了後も用語精査作業は継続して実施しており，2019 年 5 月半ば時点で，標準病名または人手でのコーディングが行われた病名（信頼度 LEVEL: S, A から C）は 73,342 件であった．これは，全体に占める割合の 20%であるが，特定の病院の電子カルテや退院サマリにおいて頻出する病名表現については，概ねカバーできていると考えられる．今後，人手によりコーディングされた病名データを学習データとして用いて機械学習モデルを更新し，人手でのコーディングが行われていない病名に対するコーディング情報を更新することなどが期待される．

E. 結論

本研究では，これまで大規模な収集が困難であった病名や症状などの医学表現の収集を臨床文書から行った．この結果，構築された「万病辞書」は臨床現場で実際に使われる病名はほぼ網羅されたと期待される．本事業終了後も，各病名の出現形に付随する情報(ICD10, ICD11, MedDRA, HPO

など）の精査や追加を継続して進める予定である．また，これらの用語を活用した入力支援ツールについて試作を行った（図3）．この評価も今後の課題である．

[参照文献]

[1] 万病辞書. http://sociocom.jp/~data/2018-manbyo/index.html#abst

[2] ICD10対応標準病名マスター (V4.04 2018年4月1日改訂).

http://www2.medis.or.jp/stdcd/byomei/index.html

[3] 荒牧英治, 若宮翔子, 矢野憲, 永井宥之, 岡久太郎, 伊藤薫：病名アノテーションが付与された医療テキスト・コーパスの構築, 自然言語処理「言語処理の応用システム」特集号（技術資料）, 25(1), 2017. (2018/2/15)

[4] Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, Eiji Aramaki: J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage, In Proc. of International Conference on Language Resources and Evaluation (LREC), 2018. (2018/5/7, Miyazaki, Japan)

[5] Eiji Aramaki, Ken Yano, Shoko Wakamiya: MedEx/J: A One-scan Simple and Fast NLP Tool for Japanese Clinical Texts, Studies in Health Technology and Informatics, MEDINFO 2017: eHealth-enabled Health, Volume 245, 285-288, 2017.

[6] 矢野憲, 若宮翔子, 荒牧英治：医療テキスト解析のための事実性判定と融合した病名表現認識器, 言語処理学会 第23回年次大会, 2017. (2017/03/14, 筑波大学)

F. 研究発表

1．論文発表

E. Aramaki, K. Yano, S. Wakamiya: MedEx/J: A One-scan Simple and Fast NLP tool for Japanese Clinical Texts, Studies in health technology and informatics. 245:pp285-288, 2017

荒牧英治, 若宮翔子, 矢野憲, 永井宥之, 岡久太郎, 伊藤薫：病名アノテーションが付与された医療テキスト・コーパスの構築, 自然言語処理「言語処理の応用システム」特集号（技術資料）, 25(1), pp 119-152, 2018

2．学会発表

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, Eiji Aramaki: J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage, LREC 2018. (Miyazaki, Japan)

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, 2018.

矢野憲, 伊藤薫, 若宮翔子, 荒牧英治: 深層学習による医療テキストからの固有表現抽出器の開発とその性能評価, 人工知能学会全国大会（JSAI）, 2017

矢野憲, 若宮翔子, 荒牧英治: 医療テキスト解析のための事実性判定と融合した固有表現認識器, 言語処理学会年次大会, 2017

G. 知的財産権の出願・登録情報

該当なし

図 3 カルテ入力パレットのインターフェイス