平成２９年度厚生労働科学研究費補助金
臨床研究等 ICT 基盤構築事業・人工知能実装研究事業

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に関する研究

平成２９年度 総括・分担研究報告書

研究代表者 荒牧　英治

平成３０（2018）年３月

[別添 2]

目次

I. 研究総括報告
カルテ情報の自動構造化システムとその機能を有した入力機構の開発に関する研究
…… 1

II. 分担研究報告
病名自動抽出のための辞書リソースに関する研究（若宮担当分）
…… 6

カルテ文章からの病名自動抽出に関する研究（河添担当分）
…… 11

III. 研究成果の刊行に関する一覧
…… 13

[別添 3]

平成 29 年度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（申請者荒牧及び分担者河添が所属する研究室主宰者の大江ら）は，2008年から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．その成果は，日本内科学会の症例報告検索システムなどとして実用化され，現在も用いられている．本研究は，電子カルテの二次利用のさらなる実用化に向けて問題となる次の２つの課題を解決する．

（課題１）実用化可能な解析精度の達成

（課題２）電子カルテに組み込み可能な実装の開発

若宮翔子（奈良先端科学技術大学院大学 研究推進機構･博士研究員）

河添悦昌（東京大学医学部附属病院 企画情報運営部･講師）

A. 研究目的

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は，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ため，既存の入力方法だけでは，ユーザーに過度に負担を強いることも少なくない．このため，本研究では，病名など複雑な固有名詞を多く含む医療文書作成を支援する知的なテキスト入力支援ツールの構築を目的とし，入力支援ツールを開発した．

B. 研究方法

B-1. 病名のサジェスト表示による入力支援

本稿で述べるテキスト入力支援ツール(図1)は，IMEによる仮名漢字変換後の入力された単語，文字の接頭辞から推測される病名，症状名をサジェストし，提示された選択肢から項目を選ぶことで病名の入力を補完する機能を提供する．サジェストの表示は,IMEでの仮名漢字変換後に行うため，基本的にどのIMEとも共存して利用可能である．提案するツールはテキストエディタとしての利

用を想定し，編集後にテキスト文書として保存または，クリップボード経由で他のアプリケーションへデータ転送する利用形態を想定している．

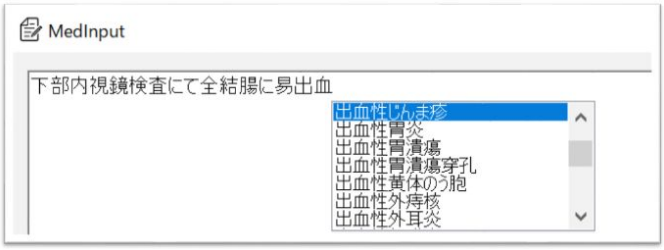

図1 予測病名リストの表示

B-2.病名辞書の構築

辞書作成のソースデータは病名の一覧が記載されたテキストファイルである．病名の一覧データはMEDIS（医療情報システム開発センター）から提供されているものを用いた．このテキストデータからトライ構造による文字ベースの辞書を構築した．トライは，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である．病名検索では，トライの根節点から葉節点に向かって，検索文字列の各文字を先頭から1回たどるだけで，入力文字列の先頭から始まる全ての接頭辞を探索することができる．従ってトライ全体に格納されている見出し語つまり病名の総数に関係なく，検索文字列の長さに比例した計算時間で検索が終了する．

B-3. 入力サジェストの提示

病名のサジェストは，図1に示すリストボックスウィンドウがポップアップで表示される．ユーザーはリストから，入力しようとしている病名を選択することで，その該当部分に選択された病名が挿入される．該当病名がない場合はEscキーにより，ポップアップを閉じることができる．なお，ポップアップはモードレスで動作しているため，ポップアップ表示中も文字入力を継続して行うことが可能である．

B-4. 入力サジェストの表示制御

入力サジェストの表示はテキストが更新される毎に判断を行う．判断には，キャレットの直前にある文字列を取得して，その文字列をキーワードとして，トライ辞書の検索を行う．

しかし，文字（単語）が入力される毎に，辞書検索を行いヒットがあった場合にサジェストを行うことには問題がある．例えば，辞書には以下のように「た」で始まる病名が含まれているが，入力文が　「～した」など動詞の末尾で終了した場合などに，最後の助動詞“た”を病名の開始と判断し，以下の病名がヒットしてしまうという課題がある．

■たこつぼ型心筋症

■たこ壺型心筋障害

上記のように，入力された文字（単語）が病名の可能性が低い場合，入力サジェストの提示を抑止する必要がある．

B-5.入力された単語（文字）が病名の一部であるかの判定

入力サジェスト提示の判断を行うため，提案手法は，既存の文字ベースの病名抽出ツール「MedEX/J」を利用する．つまり，編集中の文を病名抽出ツールに適応し，入力された単語（文字）が病名の一部かを判定する．予備調査の結果，文が未完結の場合でも病名抽出を高い精度で行えることを確認した．

なお，本手法は，病名（標準病名のみ）を扱い人名やIDは扱わないため，個人情報を不用意に削除・保管することはない．また，機械学習モデル内部においても，単語レベルで情報を保持しており，特定の個人に結びつくこともない．

C．研究結果

C-1.　病名の一括抽出と病名の標準化

病名抽出機能を入力された文章全体に対して適応することで，文章中に含まれる全ての病名の確認と，病名の標準化を行うことができる．図2に病名の一括抽出と事実性の判定を行った例を示す．抽出された陽性の病名は赤で，陰性の病名は青で表示される．処理方法としは，初めに，文章を文に分割し，1文ごとに前述の病名抽出ツールにより病名を抽出している．抽出された病名は同時に自動的に陽性か陰性の極性判断がなされる．

C-2.　抽出された病名の標準名への変更

マウス操作により，抽出された赤字または青字の病名を右クリックすると，その病名に対する標準名およびICD-10コードが表示される(図3)．ユーザーはここでサジェストされた標準病名に置き換えることが可能である．

図 2 病名の一括抽出と，陽性（赤）/陰性（青）の判断

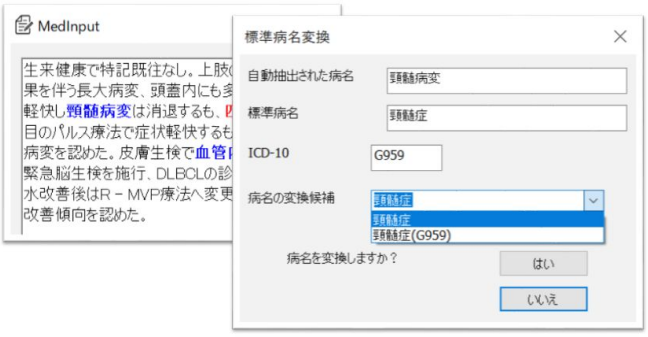

図 3 病名の ICD-10 コードの確認と標準病名への変換

図 4 カルテ入力パレットのインターフェイス

C-3. Web版入力支援ツールの開発

本研究では，さらに入力支援ツールの操作性を高めるために，Web版の入力支援システム（以下，カルテ入力パレット）を構築した（図4）．

図４に示すとおり，Web上の入力枠に医療テキストを入力し，解析ボタンを押下することで極めて簡易に使用することができる．

図4の入力例では，糖尿病に関する症例報告を入力している．傷病名と判断した箇所が信頼度に対応した色で強調表示され，当該箇所にマウスポインタをあわせることで，ICD10コードに対応した標準病名が提示される．提示された標準病名をマウス選択することで，容易に傷病名を変更することが可能である．なお，入力枠の右側に簡易的な人体解剖図が示されており，入力テキスト解析により特定された傷病名に対応した部位が点滅するという直感的な可視化機能を備えている．

D. 考察

本ソフトウエアのユーザービリティーの評価として，数名の被験者にこのツールを使用した時の医療テキストの入力時間を計測した．その

結果，このツールを用いた場合とそうでない場合では，入力時間の違いに有意差は見られなかった．しかし，テスト時には事前に被験者によるツールの使い方の練習は全く行わなかったので，ツールの使い方に慣れてくれば，ある程度入力時間の短縮は期待できると予想している．

E．結論

本研究では病名支援入力ツールを開発した．入力時の支援だけではなく，本ツールはバッチ処理により，文章から一括して病名を抽出する機能も有し，抽出された病名のICD-10コードの確認と病名変更の機能も提供している．今後は，医療従事者により，本システムを評価を実施する予定である．

F. 健康危険情報

特になし.

G. 研究発表

1. 論文発表

- 該当なし

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.

H. 知的財産権の出願・登録情報

特になし

[別添 4]

平成 29 年度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)

分担研究報告書

病名自動抽出のための辞書リソースに関する研究

研究分担者：若宮翔子 奈良先端科学技術大学院大学 研究推進機構

A.研究目的

医療文書から病名を抽出する処理は，これまで医療言語処理分野の研究で盛んに行われてきた．ほとんどの病名抽出処理においては，ICDのような標準規格で規定された病名が用いられている．しかし，実際の医療現場では，正式名称ではなく略記や英語名を用いることが少なくない．そのため，定型的な病名コードだけでは，症状や病名に関する情報をすべて抽出したいといった要望には応えることが難しい．このような課題を解決するために，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語を幅広く抽出し，そのデータを「万病辞書」として辞書化し公開している[1]．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

本研究では，ICD-10 対応標準病名マスターの病名(まもなく公開予定の最新版は 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．特定の病院のカルテ文章を調査したところ，延べ 45 万の病名表現（種類数約 6.2 万種類）が得られた．そのうちの 28.3%（種類数約 1.7 万種類）が，標準病名のみではカバーされていないことが分かった．この標準病名のみではカバーされていない病名表現のうち，高頻度のものから順に，医療従事者（最大 3 名）によりコーディングを行っている [3, 4]．2018 年 3 月末の時点で，8,233 の病名表現について人手でのコーディングが施され

ており，残りについては機械学習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する．

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

（倫理面への配慮）

本研究については以下の課題名で，奈良先端科学大学院大学情報学系の倫理審査に申請し，申請が受理されている．

C.研究結果

万病辞書の抜粋を図1に示す．図1のように，万病辞書は以下の5つの項目から構成されている．

(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）．

(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ICD10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．

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．

(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5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

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

よみがな，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．

・よみがな：

- 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：「万病辞書 よみがなくん」[7]により自動付与（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3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：5件未満

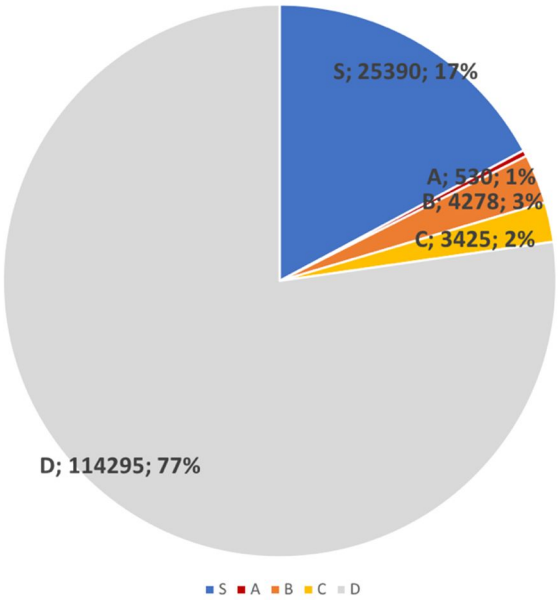

図 2. 信頼度 LEVEL ごとの件数. データラベルは, 信頼度 LEVEL; 件数; パーセンテージ

D.考察

2018年3月末の時点で，8,233の病名表現について人手でのコーディングを行ったが，図2から分かるように，完了したのは全体に占める割合は6%ほどであった. ただし, 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．また，頻度が低い病名表現の中には，実際に希少な疾患である場合もあれば，ノイズとなるような表現が誤抽出されている場合もあるため，後者のようなノイズについては人手でフィルタリングしていく必要がある．さらに，これまでのコーディング結果（信頼度LEVELがS, A, B, Cのデータ）を学習データとして用いて機械学習のモデルを学習し，信頼度LEVELがDのデータに結果を自動的に付与し直し，それを人手により精査することにより，コーディングの信頼度および作業効率の向上を目指す.

また，より辞書リソースとしての利便性を向上させるために，ICDコードに対応するMedDRA/Jコード [8]の付与を行う予定である．なお，MedDRA/J とはヒトに使用される医療用製品のための国際的な規制情報の共有を促進するための高品質で特異性が高い標準化された医学用語集の日本語版である.

E.結論

医療文書から実際の医療現場で用いられるような幅広い病名表現の抽出を可能にするために，医療従事者が記載した電子カルテや退院サマリから抽出した病名表現に対応するICDコードや標準病名をコーディングし，そのデータを「万病辞書」として辞書リソース化している．本稿では，「万病辞書」のファイル構成や統計について報告し，今後の課題について整理した.

[参照文献]

[1] 万病辞書．http://mednlp.jp/DIC/index.html

[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）

[7] 万病辞書 よみがなくん．http://mednlp.jp/yomiganakun.html

[8] MedDRA/J．https://www.meddra.org/how-to-use/support-documentation/japanese

F.健康危険情報

該当なし

G.研究発表

1．論文発表

該当なし

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)

H. 知的財産権の出願・登録情報

該当なし

平成 29 年度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)
分担研究報告書

カルテ文章からの病名自動抽出に関する研究

研究分担者：河添悦昌 東京大学医学部附属病院 企画情報運営部

A．研究目的

東大病院の電子カルテに記載された診療記録から症状・所見・疾患に関する単語を抽出する.

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，奈良先端大学の荒牧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.

B-3. 研究の実施に際しては，東京大学大学院医学系研究科の倫理承認（承認番号：11446）得て行った.

URL:http://www.m.u-tokyo.ac.jp/medinfo/wp-content/uploads/2013/08/ethics-20170208.pdf

C．研究結果

C-1. 合計約 1870 万件の診療記録を対象とした.病名抽出ツール（mednlp parser v006）の処理に要す.

C-2. 表 1 に抽出結果の要約を示す.診療記録の総件数は，2010 年から 2015 年にかけて一定の割合で増加傾向にあるが，2016 年には急増していた.この原因として診療記録のテンプレートが細分化したことにより，見た目上の件数が増えたなどの原因が考えられた.

C-3. 2016 年を除き，1 診療記録あたりの病名単語数（重複あり）は増加傾向にあるものの，病名単語数（重複なし）は一定の割合を保っていることから，1 診療記録あたりの記載量は増加しているが，疾患に関するトピックが増えているわけではないと考えられた.

データ抽出過程のため特になし.

D. 考察

データ抽出過程のため特になし.

E. 結論

データ抽出過程のため特になし.

F. 健康危険情報

G. 研究発表

データ抽出過程のため特になし.

H. 知的財産権の出願・登録情報

該当なし

表１：抽出結果の要約

	診療記録総件数	病名単語数（重複あり）		病名単語数（重複なし）	
		全件	1診療記録あたり	全件	1診療記録あたり
2010	2,314,455	8,493,913	3.67	356,726	0.15
2011	2,301,886	8,766,392	3.81	359,430	0.16
2012	2,540,405	10,217,999	4.02	389,905	0.15
2013	2,640,369	10,907,097	4.13	395,737	0.15
2014	2,712,801	11,830,890	4.36	403,289	0.15
2015	2,720,191	12,512,910	4.60	408,914	0.15
2016	3,461,112	13,518,290	3.91	415,599	0.12
合計（平均）	18,691,219	76,247,491	(4.07)	(389,943)	(0.15)

[別添5]

研究成果の刊行に関する一覧表

論文

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
該当なし					