[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

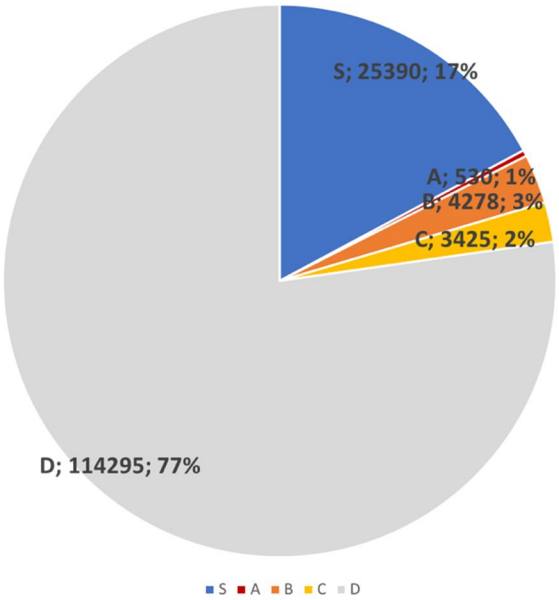

図 2. 信頼度 LEVEL ごとの件数. データラベルは, 信頼度 LEVEL; 件数; パーセンテージ

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)