電子カルテテキストを自動臨床データベース化する要約システムの開発

Development of Summarization System for Electronic Health Record

柴田大作 [1]　若宮翔子 [1]　伊藤薫 [1]　荒川豊 [1]　吉江智秀 [2] 荒牧英治 [1]
Daisaku Shibata[1]　Shoko Wakamiya [1]　Kaoru Ito [1]
Yutaka Arakawa[1]　Tomohide Yoshie[1] Eiji Aramaki[1]

[1] 奈良先端科学技術大学院大学
[1] Nara Institute of Science and Technology
[2] 聖マリアンナ医科大学
[2] St. Marianna University

Abstract: [Background] To introduce Artificial Intelligence (AI) to medical filed has drawn much attention because advances in AI are remarkable. [Objective] We develop a summarization system for Electronic Health Record (EHR) to support building clinical database tasks. [Method] This study challenges to extract information of mainly diabetes and smoking states from EHR. [Result] Our model achieved macro F-measure of 0.80 in diabetes and 0.73 in smoking states. [Conclusion] The preliminary experimental results show that our model could extract information of diabetes and smoking states with the practical accuracy.

1　はじめに

人工知能 (Artificial Intelligence: AI) の発展に伴い，AI の医療への導入は飛躍的に進歩している．米国では画像診断を中心とした AI ソフトが認可[1]されており，日本でも病理診断や製薬に用いる AI などが研究されている．しかしその一方で，医療現場の医師が必要としている AI システムとの乖離がある．我々の調査として現役の医師 37 名にアンケートを実施したところ，最も必要とされている AI システムは画像診断や病理診断に関するものではなく，もっと単純な「事務作業を補助するシステム」や「研究に必要なデータを自動で収集するシステム」であることが確認された．サンプル数は少ないが，臨床現場で実際に働く医師の貴重な意見と考えられる．

本研究の目的は研究に必要なデータを自動で収集・データベース化するシステムの開発，つまり医師の臨床データベース構築作業の支援である．医師の代表的な事務作業の一つである「電子カルテから治験や臨床研究に必要な情報を抽出し，リスト化する作業」の補助を行う．具体的には電子カルテを入力として，身長，体重や年齢などの基本事項の抽出と，糖尿病・喫煙に関する情報の抽出を行い，その結果をデータベース化するソフトウェアである Pheno Encoder (PE) を開発する．表記揺れが少ない (どの医師が書いても同様の表現になる可能性が高い) 身長や体重など項目はルールベースで，反対に表記揺れが大きいと考えられる糖尿病疾患や喫煙歴に関する記載は機械学習により抽出する．糖尿病・喫煙は，本研究における対象分野が循環器であり，循環器において重要な項目であるため抽出項目としている．

一般的に上記のような情報抽出では，大きく分けて，3 種類の手法 (ルールベース，機械学習，深層学習) がある．ルールベースは最も単純で解釈が容易であるが，表記揺れやドメインによってはルールが膨大となり，メンテナンスが困難になる可能性がある．一方で機械学習と深層学習では表記揺れに対してある程度ロバストであるが，機械学習では特徴量エンジニアリングが必要であり，深層学習では膨大な学習データが必要になる．本研究では表記揺れが少ないと考えられる身長や体重などの基本項目はルールベースで，医師ごとに表記の方法が異なり，表記揺れが多いと考えられる糖尿病と喫煙の識別に関しては機械学習で抽出する．

本論文では，提案するソフトウェア，学習データの構築方法と機械学習モデルの構築・評価について説明

[1] https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm596575.htm

する．また，機械学習の学習データとして日本内科学会の症例報告 (n = 6,918 件) を使用した評価実験において，糖尿病ではマクロ F-measure が最大 0.80，マイクロ F-measure が 0.96，喫煙ではマクロ F-measure が最大 0.73，マイクロ F-measure が 0.88 と一定の精度で情報を抽出できることを確認したので報告する．

2 先行研究

本研究と最も関連する研究の一つとして Informatics for Integrating Biology to the Bedside (i2b2) の Deidentification and Smoking Challenge[1] が挙げられる．これは退院サマリから喫煙に関する情報を抽出するタスクの精度を競うワークショップで，参加チームは 11 チームである．退院サマリを入力として，現在の喫煙，過去の喫煙，時制不明の喫煙，喫煙の記述なしのいずれかを識別するモデルをそれぞれ構築する．訓練データは 398 個，テストデータは 104 個と小規模のデータセットであるが，退院サマリという貴重なデータであること，2008 年に開催されたことを踏まえれば十分なデータ数であると考えることができる．

機械学習を用いて電子カルテから患者のライフスタイルに関する情報を抽出する研究もある．Ma ら [2] は電子カルテから肥満，喫煙，飲酒に関する情報を抽出する研究において，単語ベクトルを作成する際に，あるキーワードの前後 n 単語のみを使用し，かつ不均衡データに対してオーバーサンプリングを行うことで，i2b2 の Deidentification and Smoking Challenge と比較して高い精度で識別できることを報告した．

そこで本研究では臨床データベース作成作業の負担軽減を目的とした支援システムの開発を行う．先行研究 [2] では，肥満，喫煙や飲酒の有無にのみ焦点を当てているが，我々は機械学習による疾患情報の抽出だけでなく，レガシーなルールベースを併用し，身長や体重などの一般的な情報を抽出可能なシステムを構築する．また将来的には医師の負担をどれだけ軽減できるのかといったユーザビリティの面まで考慮したシステム開発を目指す．

3 PhenoEncoder

開発するシステムである PhenoEncoder (以下，PE) の入力画面を図 1 に，出力結果を図 2 に示す．

PE は入力文から表 1 に示す項目を抽出して表示するアプリケーションであり，身長，体重，性別，年齢，HbA1c，CRP と薬の投与歴はルールベースで，糖尿病，喫煙の有無は機械学習によって抽出する．抽出項目は，本研究の対象分野が循環器であることから，循環器分野において重要であると考えられるいくつかの項目を医療従事者との相談で決定した．

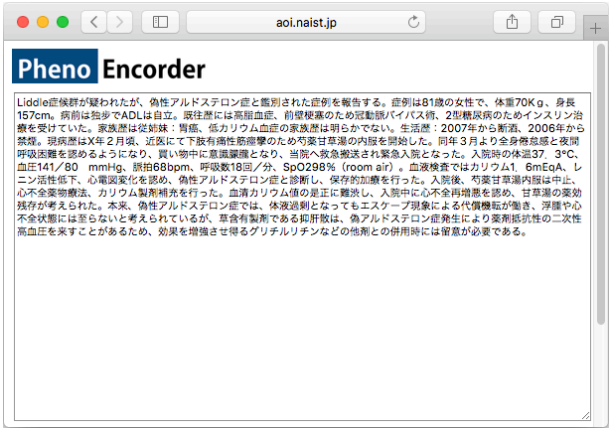

図 1: PhenoEncoder の入力画面

表 1: 抽出項目

対象	説明	手法
性別	性別 (カテゴリ) を抽出	ルールベース
身長	身長 (数値) を抽出	
体重	体重 (数値) を抽出	
年齢	年齢 (数値) を抽出	
HbA1c	HbA1c (数値) を抽出	
CRP	CRP (数値) を抽出	
薬の投与歴	薬の投与歴を抽出	
血圧	血圧 (数値) を抽出	
糖尿病	糖尿病疾患の有無を抽出	機械学習
喫煙	喫煙歴の有無を抽出	

本研究ではデモのために Web アプリケーションとして実装しているが，今後はオフラインで使用可能なソフトウェアを開発する予定である．

4 コーパス

4.1 材料

本研究で使用する電子カルテは症例報告と呼ばれるテキストである．症例報告とは，学会に提出される患者情報の要約であり，患者の診断名・転帰，入院時の症状および所見，治療後の経過などが簡潔に記載され，医師の教育や類似した症例の参考のために照会される．このため症例報告は患者に対して記述する診療録よりも高い可読性で記載される傾向にある．症例報告は学会への報告であるため，記録先は異なるが入院時の患者情報の要約であるという点に置いては退院サマリと類似している．先行研究 [1] と最も異なる点は，言語 ([1] は英語，本研究は日本語)，[2] と異なる点はデータセット ([2] は電子カルテ，本研究は症例報告) と対象項目である．

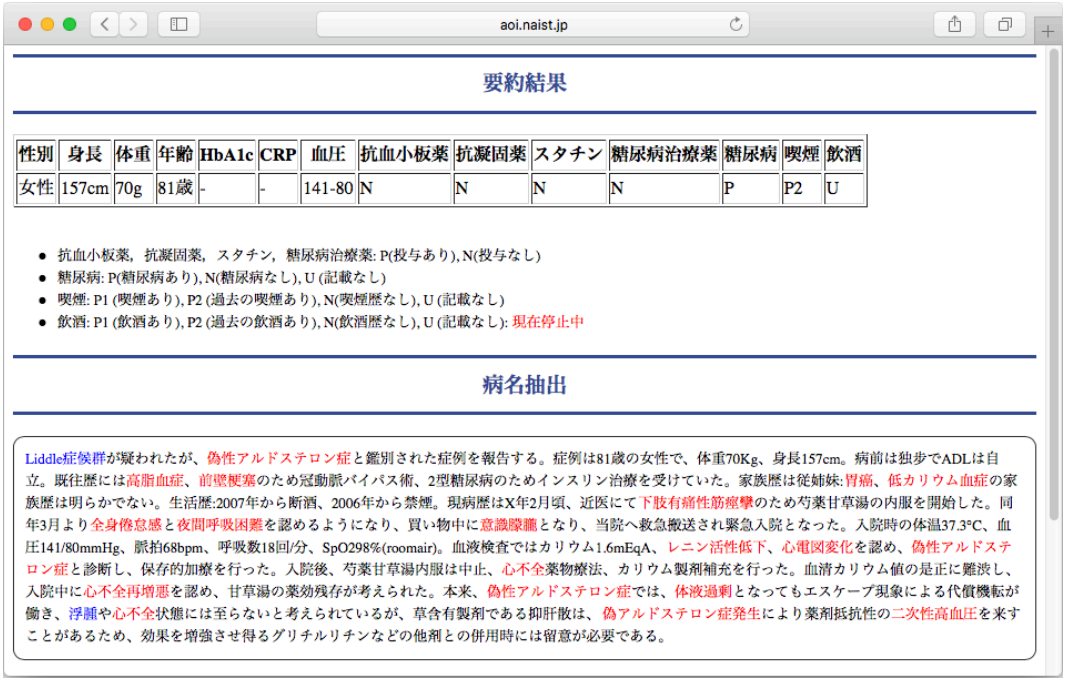

図 2: PhenoEncoder の入力画面

本コーパスの材料であるテキストは日本内科学会に報告された症例報告 [7] である．これは 2004 年以降に報告された全症例であり，11,866 施設，26,235 人の医師からなる．またカバーする診療科は内科領域全域である．なお，本研究では，循環器分野に焦点を当てた研究を行うため，使用するデータは分野が循環器 (6,982 件) であるものに限定する．

表 2: 類似語一覧．下線はキーワードを示す．

対象	キーワード
糖尿病	未治療糖尿病, 性糖尿病, 膵性糖尿病, 糖尿病群, 非糖尿病, 境界型糖尿病, 二次性糖尿病, 高齢者糖尿病, 合併糖尿病, 糖尿
喫煙	喫煙,喫煙歴,タバコ, 生活歴, 嗜好, 煙草, 嗜好品, 嗜好歴, 飲酒歴, 飲酒

4.2 アノテーション: タグ付け

本研究におけるアノテーションは教師データを作成する作業である．また，タグ付けとは，症例報告の各文に対して，糖尿病の有無や喫煙歴の有無のタグを付与する作業であり，その手順を以下に示す．

Step 1: 症例報告の文分割
症例報告を文ごとに分割する．文の境界は句点とし，分割を行った．文境界にカンマを使用している文章はあらかじめ削除した．この結果，約 6 万件の文が抽出された．

Step 2: キーワード抽出
Step1 の結果の 6 万件全てを目視により確認し，タグを付与することは現実的ではない．そのため，特定の単語 (キーワード) を抽出し，それらを含む文のみを対象とする．キーワードは単語の分散表現を用いて決定する．分散表現は単語をベクトル化する手法であり，単語間の類似度を測定することが可能である．本研究では，全症例報告を用いて fastText[4] により分散表現の学習 (次元数: 100, 窓サイズ: 10) を行い，単語 (糖尿病，喫煙) の類似語を取得した．各単語において，上位 10 個の類似語を取得し，キーワードであると考えられる単語のみを抽出した．抽出された類似語とキーワードの一覧を表 2 に示す．

Step 3: タグの付与
Step2 で定めたキーワードを含む文を全て抽出し，タグの付与を行う．付与するタグの一覧を表 3 に示す．正確なタグの付与作業には医学知識が必要であるため，本作業は医療従事者によって実施された．タグの付与が困難であった文（2 文）と疾患に直接関係がないと考えられる文（7 文）は除外した．

表 3: タグ一覧

対象	タグ		意味
糖尿病	Positive	P	糖尿病の疾患あり
	Negative	N	糖尿病の疾患なし
喫煙	Positive1	P1	喫煙/飲酒あり
	Positive2	P2	過去の喫煙
	Negative	N	喫煙/飲酒なし

表 4: データ分布

状態	頻度 (%)	
	糖尿病	喫煙
N	27 (5.2)	23 (13.7)
P1 (P)	489 (94.8)	129 (76.8)
P2	-	16 (9.5)

Step1 から Step3 の作業により得られたデータの統計量を表 4 に示す.

5 実験

5.1 実験設定

機械学習による糖尿病疾患の有無と喫煙の有無を識別を行う．モデルには線形サポートベクターマシンを使用し，特徴量には Bag of Words (BoW), TF-IDF, Word Embedding (WE) を使用した．特徴量を作成する際の前処理として，日本語形態素解析器 MeCab による文の形態素解析を行なった．辞書には症状や病名に関連する語が幅広く登録されている万病辞書 [6] を使用し，高頻度で出現する記号 (不等号や#など) はストップワードとした.

5.1.1 特徴量

- Bag of Words (BoW)
  BoW はコーパスの全体集合から，単語から成り立つ語彙を作成し，各コーパスでの各単語の出現頻度を含んだ特徴ベクトルを構築するモデルである.

- TF-IDF
  TF-IDF は TF (単語頻度) と IDF (逆文書頻度) の積で定義される．以下に定義式を示す.

$$tf - idf(w_i, D_j) = tf(w, D) \times idf(w, D)$$

$$idf(w_i, D_j) = \log \frac{|D|}{1 + tf(w, D)}$$

  w_i: 単語.
  $tf(w, D)$: 文 D における単語 w の出現頻度.
  $idf(w, D)$: 逆文書頻度.

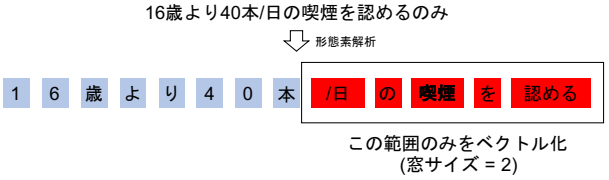

図 3: 特徴量 (B) の例

- Word Embedding (WE)
  分散表現はベクトルを密な固定長のベクトルに変換する手法であり，本研究では Facebook 社によって開発された fastText[2]を用いて WE を作成する．文に含まれる全単語の分散表現ベクトルの平均値をそのコーパスの単語ベクトルとし，節 4.2 で用いたモデルを使用する.

$$WE(x) = \frac{1}{n} \sum_{i=0}^{n} w_i$$

  w_i: 単語 i における分散表現.
  n: 文に含まれる単語数.

これら 3 つの特徴量 (A) に加えて，先行研究 [2] と同様にベクトル化する単語の範囲をキーワードの前後 n 文字に限定 (図 3) した特徴量 (B) も作成した．糖尿病の識別実験においては，キーワードの後ろ 10 文字のみを用いて単語ベクトルを作成し，喫煙の識別実験においてはキーワードの前後 10 文字を用いて単語ベクトルの作成を行った．喫煙では重要な単語が「喫煙」という単語の前後で出現する (例: 20 歳より喫煙 1 日 40 本) が，糖尿病では「糖尿病」という単語の後にのみ出現する (例: 糖尿病の既往歴あり)．そのため糖尿病ではキーワードの後ろのみ，喫煙ではキーワードの前後のベクトル化を行なった．窓サイズは一意に設定したが，今後はこの値を変更した際の精度についても調査する必要がある.

5.2 評価方法

10 分割交差検証を用いて，F-measure のマクロ平均とマイクロ平均による評価を行う．マクロ平均は、各クラスについて評価指標を計算し，それらの算術平均をとることで算出される値である．またマイクロ平均は，全クラスで結果を一つの分割表にまとめ，その表から評価指標を算出することで得られる値 [3] である.

[2]https://github.com/facebookresearch/fastText

表 5: 実験結果: 糖尿病

		N			P			Macro Average			Micro Average		
	Feature	P	R	F	P	R	F	P	R	F	P	R	F
(A)	BoW	0.70	0.26	0.38	0.96	0.99	0.97	0.83	0.63	0.68	0.95	0.95	0.95
	TF-IDF	0.40	0.13	0.20	0.95	1.00	0.97	0.67	0.56	0.59	0.95	0.95	0.95
	Word Embedding	0.39	0.33	0.32	0.96	0.96	0.96	0.67	0.64	0.64	0.92	0.92	0.92
(B)	BoW	0.80	0.53	0.61	0.97	0.99	0.98	0.89	0.76	**0.80**	0.96	0.96	**0.96**
	TF-IDF	0.60	0.27	0.36	0.96	1.00	0.98	0.78	0.63	0.67	0.95	0.95	0.95
	Word Embedding	0.43	0.47	0.43	0.97	0.97	0.97	0.70	0.72	0.70	0.94	0.94	0.94

表 6: 実験結果: 喫煙

		N			P1			P2			Macro Average			Micro Average		
	Feature	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
(A)	BoW	0.70	0.53	0.60	0.85	0.94	0.89	0.53	0.40	0.43	0.70	0.62	0.64	0.82	0.82	0.82
	TF-IDF	1.00	0.40	0.57	0.78	1.00	0.88	0.00	0.00	0.00	0.59	0.47	0.48	0.80	0.80	0.80
	WE	0.57	0.70	0.61	0.88	0.89	0.88	0.60	0.30	0.40	0.68	0.63	0.63	0.80	0.80	0.80
(B)	BoW	0.82	0.83	0.80	0.89	0.95	0.92	0.60	0.40	0.47	0.77	0.73	**0.73**	0.88	0.88	**0.88**
	TF-IDF	1.00	0.40	0.57	0.79	1.00	0.89	0.20	0.10	0.13	0.66	0.50	0.53	0.81	0.81	0.81
	WE	0.70	0.63	0.61	0.87	0.88	0.88	0.73	0.60	0.63	0.77	0.70	0.70	0.81	0.81	0.81

5.3 実験結果

糖尿病と喫煙の実験結果をそれぞれ表 5 と 表 6 に示す．糖尿病では，ベクトル化する単語の範囲を限定した BoW を特徴量としたときのマクロ F-measure が 0.80，マイクロ F-measure が 0.96 と最も高くなった．同様に喫煙でも，ベクトル化する単語の範囲を限定した BoW を特徴量としたときのマクロ F-measure が 0.73，マイクロ F-measure が 0.88 と最も高いことが確認された．

5.4 考察

糖尿病と喫煙の両方で，ベクトル化する単語の範囲を限定して作成した BoW を特徴量とした際に，精度が最も高いことが確認された．単語の範囲を限定することで簡易的に次元数を削減することができ，これが有効に作用したのではないかと考えられる．本研究では，範囲を一意に決定して実験を行ったが，今後この値を変化させたときに精度がどのように変化するのかを確認する必要がある．また今回は比較的識別が容易な項目について実験を行ったため BoW で一定の精度が得られたが，別の識別が困難であると考えられる項目においても同等の精度が得られるとは限らず，今後の検討が重要である．さらに，学習データが不均衡であり，一部のタグではデータ数が非常に少ないという問題があるため，継続したデータ収集が必要不可欠である．

5.5 予測誤りの分析

実験 (特徴量はベクトル化する単語の範囲を限定した BoW) において予測を誤った文の一例を表 7 に示す．

1 例目は糖尿病の識別実験において，本来は負例であるが，誤って正例だと識別された文である．これは文中に疾患を有することを示す (+) と疾患を有しないことを示す (-) が混在しており，語順を考慮しない BoW では表現をうまく捉えることができなかったことが原因だと考えられる．

2 例目は糖尿病の識別実験において，本来は正例であるが，誤って負例だと識別された文である．これは文中に「糖尿病」と「ない」が出現しているため，負例と識別されたと考えられる．

3 例目は喫煙の識別実験において，本来は N であるが，誤って P1 (喫煙あり) に，4 例目は本来は P2 であるが，誤って P1 だと識別された文である．3 例目では文中に「非」と「喫煙」 という単語が存在することから N に，4 例目では「禁煙」と「喫煙」という単語が存在することから P2 に識別されると考えられるが，学習データに類似する表現の文が少なかったため正しく識別されなかったと考えられる．

5 例目は喫煙の識別実験において，本来は N であるが，誤って P1 に誤って識別された文である．これは「嗜好品」という単語を含む文がデータセット中にこの 1 文だけであることから，3 例目，4 例目と同様に学習データが不足していることが原因であると考えられる．

表 7: 予測を誤った文一覧

No.	疾患	正解	予測	文
1	糖尿病	N	P	高脂血症 (+) 高血圧 (+) 糖尿病 (-) 喫煙 (+) 家族歴 (-)【現病歴】2009 年 3 月 24 日入眠中の午前 4 時頃突然頭痛が出現し覚醒した
2		P	N	2 型糖尿病の教育入院中に労作と関係のない胸部違和感やふらつきを認めた 80 歳代女性
3		N	P1	非喫煙高齢女性において Buerger 病類似の末梢動脈閉塞症を経験したので文献的考察を加えて報告する
4	喫煙	P2	P1	【冠危険因子】喫煙歴のみ (20 年前より禁煙)
5		N	P1	生来健康で生活習慣病家族歴嗜好品を含めた冠動脈危険因子は何も持っていなかった

6 まとめ

本研究では機械学習により糖尿病，喫煙の記述の有無を識別する実験と電子カルテを自動臨床データベース化するアプリケーションの開発を行なった．識別実験においては，ベクトル化する単語の範囲を限定した BoW を特徴量としたときの精度が糖尿病，喫煙ともに最も高く，有益な特徴量であることが確認された．また，実用的なシステムを作るには旧来通りのルールベースと機械学習を必要に応じて組み合わせることが重要である可能性が示唆された．

今後はルールベースなどのシンプルな手法との精度の比較や抽出項目の増設やアプリケーションの開発，ユーザビリティ評価などを行う予定である．

謝辞

本研究の一部は，国立研究開発法人日本医療研究開発機構 (AMED) の課題番号 JP18lk1010003，および，厚生労働省科学研究費補助金の課題番号 H28-ICT-一般-008 の支援を受けたものです．

参考文献

[1] Özlem Uzuner, Ira Goldstein, Yuan Luo, Isaac Kohane. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association, Volume 15, pp. 14-24, 2008.

[2] Xiaojun Ma, Emiko Shinohara, Hao Han, Masamichi Ishii, Takeshi Imai, Kazuhiko Ohe. Extracting Information on Lifestyle Issues from Clinical Narratives in EHR. 医療情報学 37(6), pp.313-321, 2018.

[3] 高村大也. 言語処理のための機械学習入門 (自然言語処理シリーズ). コロナ社, 2010.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606, 2016.

[5] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[6] 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.

[7] 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.