一般口演6
用語/正規化/テキストマイニング

2020年11月19日(木) 14:00 ～ 15:23 H会場 (研修センター2階・音楽工房ホール)

[2-H-2-03] 日本語病名正規化システムの開発

*氏家 翔吾[1]、矢田 竣太郎[1]、若宮 翔子[1]、荒牧 英治[1] （1. 奈良先端科学技術大学院大学）
*Shogo Ujiie[1], Shuntaro Yada[1], Shoko Wakamiya[1], Eiji Aramaki[1] （1. 奈良先端科学技術大学院大学）

キーワード：Disease Name Normalization, Word Sense Disambiguation, Natural Language Processing, Electronic Health Records, Named Entity Normalization

【背景・目的】診療録や症例報告（医療文書）などに記述される病名の抽出，及び正規化は，医療文書からの知識獲得や検索システムにおいて重要である．英語においては病名の抽出・正規化の標準的ツールが整備され，医療文書解析の基礎となっている．日本語においても病名抽出システムは開発されてきたが，病名の正規化は辞書検索や編集距離による正規化に止まっており，表現が似通う病名しか扱えない問題がある．本研究では，日本語の病名に対して病名正規化手法の定量的評価を行い，病名正規化ツールの開発・公開を行う．【方法】病名に対して， TF-IDFを用いたランキング学習により標準病名マスター収載の標準病名に紐付けることで正規化を行った．実験材料として，１）診療録や退院サマリから抽出した病名と標準病名の組み30415件，２） NTCIR11 shared task MedNLP2で使用された模擬診療録49件内の病名2121件を用いて性能評価を行った．２については，病名抽出から正規化までの一貫した精度評価も行った．【結果・考察】１における標準病名への紐付けにおいて精度0.69，２における病名抽出を含む評価において F値0.459の精度を得た．ひらがなと漢字などの異なる文字種間の変換や英語の混在などの問題も見られ，日本の医療文書の言語的特徴を考慮した手法の有用性が示唆された．本発表では開発したツールを公開し，今後の発展の可能性について議論する．

日本語病名正規化システムの開発

氏家 翔吾*1、矢田竣太郎*1、若宮 翔子*1、荒牧 英治*1

*1 奈良先端科学技術大学院大学

Developing Japanese disease normalization system

Shogo Ujiie*1 , Shuntaro Yada*1 ,Shoko Wakamiya*1 , Eiji Aramaki*1

*1 Nara Institute of Science and Technology

[Background] Extraction and normalization of the disease names appearing in medical documents, such as medical records and case reports, are important for knowledge acquisition and retrieval systems from text inputs. In English, standard tools for disease name extraction and normalization have been available. In Japanese, however, such tools are limited to rule-based primitive implementation, which can only handle the names of diseases with similar expressions. This study adopted a normalization method for Japanese disease names. [Methods] We used a TF-IDF-based ranking method to normalize disease names by linking them to standard disease names defined in the Standardized Disease Master. We evaluated the performance of 1) 30,415 pairs of standard names of diseases that are extracted from medical records and discharge summaries, and 2) 2,121 disease names in 49 simulated medical records used in the MedNLP2, a shared task of NTCIR11. [Results and discussions] We obtained an accuracy of 0.69 in the normalization of disease names, and an F value of 0.459 in the extraction and normalization of disease names. Conversion between different Japanese characters and the mixture of English, make the normalization difficult, which implies usefulness of a method that takes into account the linguistic characteristics of Japanese medical documents. We published the developed tools and discussed the possibility of their future development.

Keywords: Disease Name Normalization, Word Sense Disambiguation, Natural Language Processing, Electronic Health Records, Named Entity Normalization.

1. 諸論

診療録や症例報告などの医療文書に記述された病名の抽出は，医療文書からの知識獲得や検索システムの構築などの重要な要素技術の一つである．しかし，抽出された病名は同一の疾患であっても異なる表記で記述される場合が多くある．例えば「疼痛」という症状は「痛み」「激痛」など様々な表記を持ち，特別な処理を施さなければこれらの表記は異なる概念として扱われる．この結果，語彙数をいたずらに増加させ，検索システムなどの最終的なアプリケーションにおいて，分類などの精度を低下させてしまう．そのため，自由記述された病名を，統制された語彙や ICD-10 などの疾病分類に紐付け，正規化することが重要となる．この処理は，正規化，表記揺れ吸収，自動標準化，エンティティーリンキングなどと呼ばれることがあるが，本稿では正規化と呼ぶ．病名の抽出や正規化は多くの医療文書を用いたシステムの基盤となる技術のため，簡単に扱えるツールが整備されている必要がある．

英語の病名に対する正規化手法は古くから提案されており，規則を用いた手法がよく用いられてきた．すなわち，大規模な類義語辞書や略語展開辞書，正規表現などに従って正規化を行う．MetaMap[1)]では類義語による病名の拡張や略語展開などにより，入力文字列から医療表現をメタシソーラスである UMLS へ紐付けた．Souza らは，完全一致や部分一致などの規則を注意深く設計することで，MetaMap を大きく上回る精度で正規化した [2)]．規則による正規化は予測の正確性が高い一方で，人手ですべての病名を網羅することは難しい．また，作成された規則は大きく言語に依存するため，大きな作成・管理コストがかかる．

そこで，規則を用いた手法の網羅性を改善するため，分類器を用いた手法が提案されている．これは入力病名を Bag of Words などのベクトルに変換し，統計的モデルに学習させることで入力病名を標準語彙へ分類する手法である．Limsopatham らは，深層学習を用いて患者によって表現された病名を SNOMED-CT や SIDER などの標準語彙に正規化した [3)]．さらに，近年では，大規模テキストでの学習により文脈に応じて単語をベクトルに符号化できる Bidirectional encoder representations from transformers（BERT）を用いた病名正規化も行われている [4)]．BERT は様々な自然言語処理タスクで高い精度を示しており，病名の正規化においても有用であると示された．これらの分類器を用いた手法は，規則による正規化よりもノイズに頑健である一方で，低頻度の病名については学習が行われないため，十分な学習データが必要となる．

このような問題に対応するため，近年多く提案されている手法として，類似度によるランキングを用いた正規化手法がある．これは，入力病名と標準語彙をそれぞれ固定長ベクトルに符号化し，それぞれのベクトル間の類似度を元に正規化する手法である．学習時は，入力病名と正解となる標準病名との類似度が高くなるようにパラメータが決定され，入力病名と類似度が最も高い標準語で正規化される．ベクトル間の類似度には余弦類似度がよく用いられる．DNorm は，TF-IDF ベクトルの変換行列を学習し，PubMed の論文概要に含まれる病名をアノテーションした NCBI disease corpus[5)]において高い精度を示した [6)]．近年では，大量の医療文書を用いて単語の意味的な情報をベクトルに埋め込む単語分散表現を用いる手法 [7)8)] や BERT の分散表現を用いた手法 [9)]が提案されており，高い精度を達成している．類似度によるランキングは，分類器を用いた手法に比べて低頻度語について頑健である特徴があり，現在はこの手法が盛んに研究されている．

英語においては，これらの手法を元に病名の抽出・正規化を行うツールの整備が進んでおり，MetaMap[1)]や cTakes[10)]，NormCo[11)]などさまざまなソフトウェア及びソースコードが公開されている．

日本語においても病名正規化の研究は行われており，荒牧らはサポートベクターマシンを用いて病名間の類似度を学習し，病名の正規化を行なった [10)]．また，ツールの整備に関

しても，荒牧らによって開発された病名抽出器であるMedEX/J[12]などがあるが，病名の正規化は辞書による完全一致や編集距離などの規則に基づいた処理にとどまっていた．日本語病名の正規化では，英語ほど言語資源がなく，深層学習などの大量の医学論文や診療録を前提とする手法は同様の精度で正規化が行えない可能性がある．そのため，日本語においては大量テキストによる事前学習を必要とする手法よりも，TF-IDF 値を用いた DNorm などの手法が有用である可能性がある，そこで，本研究では正規化がなされた日本語病名を材料とした DNorm の有効性を検証し，医療文書に広く適用できる病名正規化ツール DNorm-J の開発，公開を行なったので報告する．

2. 方法

まず，任意の病名の正規化先語彙を定める必要がある．本研究では，標準病名マスターに登録された標準病名（以下，標準病名）を使用した．つまり，本研究では，自由記述された病名を入力とし，あるひとつの標準病名に紐付けることを病名の正規化と呼ぶ．

2.1 実験材料

病名正規化の実験データセットとして，1 つの病名について 1 つ以上の標準病名が紐付けられたデータセットが必要である．本研究では，以下の 2 つを使用した．

1. 万病辞書
   万病辞書は，電子カルテや退院サマリから自動抽出された病名が，ICD-10 対応標準病名マスター記載の標準病名に紐付けられている．各病名への標準病名の紐付けは，人手や機械による自動付与で行われており，紐付けの方法により信頼度が付与されている，本研究では人手により紐付けられた信頼度 A から C の病名を使用した．
2. MedNLP2
   NTCIR-11 の MedNLP2[13]において使用されたデータセットで，模擬診療録と過去の状態に関する質問表から構成されており，学習データとして 104 文書，テストデータとして 49 文書が公開されている．これらに対して病名，及び病名に対応する ICD-10 が紐付けられている．

これらのデータをどのように扱うかは 2.4 節で述べる．

2.2 手法

本研究では正規化手法として，低頻度語に頑健であり，大量の言語資源を必要としない DNorm [2]を用いた．DNorm は，TF-IDF ベクトルを入力としてランキング学習を行い，病名を正規化する手法である．m を入力病名の TF-IDF 値，n を標準病名の TF-IDF 値とすると，入力病名と標準病名の類似度を

$$score(m,n) = m^T W n$$

として定義する．ここで W は学習パラメータである重み行列である．予測時には，入力病名に対して，全ての標準病名に対する類似度を計算し，類似度最大の標準病名で正規化を行う．学習時は，正解となる標準病名（n^+）との類似度と，不正解となる標準病名（n^-）との類似度の差（マージン）が1以上となるように学習を行う．

$$W = argmin_W \Sigma_m \Sigma_{n^+} \Sigma_{n^-} \max(0, 1 - m^T W n^+ + m^T W n^-)$$

Wは確率的勾配降下法で学習を行い，各イテレーションにおけるWの更新式は以下のように表される．ただし，λは学習率であり，Wの更新幅を表す．

$$W = W + \lambda(m(n^+)^T - m(n^-)^T)$$

ここで，n^+は実験材料において明示的に与えられているが，n^-は不正解となる標準病名の中から任意の標準病名を抽出する必要がある．Leaman らは，各イテレーションにおいて，不正解となる標準病名の中から最も類似度が高い標準病名をn^-として抽出した [6]．そのため，本研究においても同様にしてn^-の抽出を行った．

2.3 ツール

ツールは Python を用いて実装を行った．Python のバージョンは 3.6.1 を使用した．また，病名正規化は病名抽出と合わせて行われることが多いことを考慮し，すでに公開された BERT の単語分散表現を用いた病名抽出ツールである MedNER-J（https://sociocom.naist.jp/medner-j/）と容易に統合できる形で開発を行った．このツールは Github において公開している（https://github.com/sociocom/DNorm-J/）．

2.4 実験設定

病名をモデルに入力する際に，前処理として，略語辞書 [14]を用いて略語展開を行った．その後，日本語の形態素解析器である MeCab を用いて病名を形態素系列に変換した．実験では，学習率は 0.01 に設定し，5分割交差検定を行った．交差検定において，検証データの正解率が最も大きかった重みを用いてテストデータの評価を行った．

また，病名の正規化は通常，病名の抽出と合わせて行われることが多く，病名の抽出を含めた一連の処理について評価することもアプリケーションの評価に役立つと思われる．そのため，本研究では以下の 2 つの実験設定で評価を行う．

1. 実験 1:
   病名を入力として，標準病名を予測する．MedNLP2 については，あらかじめ正解データとなる病名を抽出し，抽出された病名に対して正規化を行う．
2. 実験 2:
   病名を含む文を入力として，病名抽出器を用いて病名を抽出し，抽出された病名に対して正規化を行う．万病辞書は病名のみしか与えられていないため，MedNLP2 のみを用いて評価を行う．病名抽出器は，すでに公開されている病名抽出器 MedNER-J を用いた．MedNER-J は BERT の単語分散表現を用いた条件付き確率場によって病名の抽出を行っている．

2.5 評価方法

実験 1 では，入力病名に対して正しい標準病名を付与した場合に正解とし，正解率で評価を行った．1 つの出現形に複数の標準病名が正解として付与される場合は，予測された標準病名がその中に含まれる場合正解として扱った．また，模擬診療録では，標準病名でなく ICD-10 へ紐付けられているため，予測した標準病名を ICD コードに変換する必要があるため，入力病名に対して予測された標準病名を，標準病名マスターを用いて ICD コードへの変換し，ICD-10 の一致によって評価を行なった．標準病名のうち，複数の ICD-10 に紐付けられているものは，そのうちのいずれかが正解となる ICD コードの場合に正解として扱った．

実験 2 では，適合率，再現率，及び F1 値を用いて評価を行なった．ここで，適合率，再現率は以下のように定義される．F1 値はその調和平均として表される．病名の抽出は，抽出された範囲が完全に一致した場合のみ正解とした．

$$適合率 = \frac{システムが正しく抽出・正規化した病名}{システムが抽出した病名}$$

$$再現率 = \frac{システムが正しく抽出・正規化した病名}{データセットに含まれる病名}$$

2, 5 比較手法

比較手法として. 以下の手法を用いた.

1. ルールベース
   以下の規則を逐次適用し, 正規化を行った. 各ステップで正規化が行われた場合, そこで処理を終了した.
   STEP1 完全一致: 入力病名と完全一致した標準病名で正規化した.
   STEP2 修飾語削除: 標準病名マスターの修飾語テーブルに従って, 入力病名から前置および後置修飾語を削除した. 削除後の病名と完全一致した標準病名で正規化した. 修飾語の削除は再帰的に行い, 削除を行ったすべての状態に対して一致をみた.
   STEP3 部分一致: 入力病名の文字集合を I, ある標準病名の文字集合をNとすると, $I \in N$もしくは$N \in I$を満たす標準病名で正規化した. ただし, 上記を満たす標準病名が複数ある場合は, 文字数の差が最小となるような標準病名を選択した. 文字数の差が同数だった場合, その中から無作為に選択した.
2. 編集距離
   入力病名と標準病名の編集距離に従い, それぞれの文字数を考慮した類似度を計算し, 類似度最大の標準病名で正規化した. m,n はそれぞれ入力病名, 標準病名であり, |m|, |n|はそれぞれの文字数を表す.

$$score(m,n) = 1 - \frac{2EditDistance(m,n)}{|m| + |n|}$$

3. Encoder
   入力病名, 及び全ての標準病名を固定長のベクトルに符号化し, 入力病名とすべての標準病名との余弦類似度を求め, 余弦類似度が最大の標準病名で正規化を行った. その際, 正解となる標準病名との余弦類似度が大きくなるように符号化器(Encoder)の学習を行った. 符号化器は一般的に, 深層学習の一種であり, 系列データを扱うことのできる Long short-term memory (LSTM)や畳み込みニューラルネットワーク(Convolutional neural network: CNN), Wikipedia などで学習された BERT などが用いられる. 本研究では, NCBI disease corpus において高い精度が報告されている CNN と BERT を符号化器として用いた. CNN への入力には通常, 大量のテキストで学習した単語分散表現を使用するが, 本研究では, 内科症例報告 50 万文を用いて学習した 200 次元の単語ベクトルと, ランダムに初期化したベクトルの 2 つを使用した. BERT は日本語 wikipedia1 億 2000 万文で事前学習を行ったモデル [14]を使用した.

3, 結果

実験 1 の実験結果を表1に示す. 万病辞書, MedNLP2 のいずれのデータセットに対しても DNorm の正解率が最も高かった.

表1 病名正規化における正解率

Method	万病辞書	MedNLP2
ルールベース	0.436	0.404
編集距離	0.356	0.463
DNorm	0.691	0.565
CNN(内科症例報告)	0.394	0.423
CNN(random)	0.614	0.454
BERT	0.444	0.480

実験 2 の実験結果を表 2 に示す. 病名抽出からの一連の処理での評価においても, DNorm が最も精度が高かった.

表2 病名抽出・正規化の一連した実験結果

Method	Precision	Recall	F1
ルールベース	0.369	0.299	0.330
編集距離	0.417	9,339	9,374
DNorm	0.508	0.412	0.519
CNN(内科症例報告)	0.373	0.303	0.334
CNN (random)	0.453	0.367	0.405
BERT	0.432	0.350	0.387

4. 考察

4. 1 比較手法に関する考察

ルールベースや編集距離による正規化と比較して, DNorm の正規化がいずれのデータセットに対しても正解率が上回ることが示された. このことから, 日本語においても機械学習を用いた手法のルールベースや編集距離といった手法への優位性が示唆された. DNorm は, 重みの学習によって単語ごとの重要度を考慮することで, 単語ごとの重要度を一律に扱うルールベースや編集距離との精度の差に影響したと考えられる. さらに, MedNLP2 の精度において正解率が高いことから, 小さいデータセットに対する DNorm の優位性が示唆された.

次に, CNN と BERT と比較しても DNorm の正解率が上回ることが示された. 英語においては CNN や BERT を用いた手法の有効性が示されている一方で, 日本語病名に対する精度が低いことから, 医療分野における日本語の単語ベクトルの質の問題が示唆される. CNN は, 大量の医療分野のテキストで事前学習された公開された単語分散表現 [13]を前提としている. 本研究で事前学習に使用した内科症例報告は約 50 万文だが, 英語での事前学習に用いられるデータ量は約 2000 万報の PubMed 論文の概要などであり, 十分に意味的情報が単語ベクトルに埋め込まれていない可能性がある. また, BERT は Wikipedia を用いて事前学習されており, 医療分野のテキストに対応できていない可能性がある. そのため, 医療分野のテキストで学習された BERT を用いることで, 精度が向上する可能性がある. さらに, BERT の事前学習は文中のマスクされた単語を前後の単語列から予測する Masked Language Model と, 2 つの文が隣接する分解中を予測する Next Sentence Prediction の 2 つで行われている. これらは文を入力としており, 単語内の現象である病名正規化には適していない可能性もある.

提案手法の精度は，既報のどの手法よりも高いが，69%程度であり，一見，十分に高い精度には見えない．しかし，現実には多くの表記ゆれは病名マスターや万病辞書などの既存辞書に収載されているため，これらにマッチしなかった場合にみ，提案手法による突合が行われることになる．この利用形態を考慮すると，表記ゆれの問題自体は一定の解決を見れるものと考えている．

4.2 予測誤りについての考察

万病辞書における交差検証の各イテレーションにおいてDNorm が予測を誤った病名から無作為に 100 例抽出し，分析を行った．正規化の誤りは，1) 上位語，2) 下位語，3) 略語，英語，4) その他，5) アノテーションミスの5種類に分類された．表3に各分類誤りの説明と，例数を示す．

表3 システムの出力の分類

種類	分類	説明	数
1	上位語	より抽象的な病名に正規化した誤り	26
2	下位語	より具体的な病名に正規化した誤り	8
3	略語，英語	略語の展開誤り，及び英語病名の展開誤り	9
4	その他	全く異なる語に正規化した誤り	40
5	アノテーションミス	予測した正規化がより適切な場合	17

1，2 は，大まかな疾患の分類には成功しているものの，標準病名の微細な違いを捕らえられなかった例である．1 としては「右前頭葉出血」に対して「出血」（正解は「脳出血」），2 としては「左腎癌」に対して「腎頭部癌」（正解は「腎癌」）が例である．

3 は，英語で記された病名を正しく日本語に展開できず，正規化を誤った例である．「Wenckebach 型房室ブロック」は「Wenckebach」が略語展開できず「房室ブロック」に正規化が行われたが，正解は「ウェンケバッハ型第2度房室ブロック」である．このように，略字辞書に掲載のない略語，英語については日本語への変換ができず，正規化が正しく行われない問題がある．

4 は全く異なる語に正規化された例である．例えば「ツツガムシ感染症」は「つつが虫病」に正規化されるべきであるが，DNorm は「ドノバニア感染症」に誤って正規化した．これは，システムが「ツツガムシ」と「つつが虫」を別の単語だと認識してしまったために起こったと考えられる．このように，病名を Bag of Words ベクトルで表現する際には漢字，ひらがな，カタカナはすべて別の概念として扱われる問題がある．

5 は，アノテーションの正解ラベルが誤っている例である．機械学習を用いて病名を正規化する場合，学習データの量・質がモデルの性能を大きく左右する．そのため，性能改善にはモデルの改良などに加えて，アノテーションの質の向上，及び学習データ量の増強が必要である．

5. 結論

本研究では，医療テキストに出現する日本語病名の正規化を行った．その結果，Bag of Words で TF-IDF ベースの機械学習モデルである DNorm による病名の正規化が最も正解率が高いことが示された．また，本研究で検証を行った病名正規化手法を用いたツール DNorm-J を開発し，公開した．病名の正規化は多くのアプリケーションの精度に関わる重要な処理であり，今後も略語辞書の拡充や学習データの増補など継続して開発を行うことで，日本語での医療言語処理に貢献していく．

参考文献

1) Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001 : 17-21.

2) D'Souza J, Ng V. Sieve-Based Entity Linking for the Biomedical Domain. Proc 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015 : 297-302.

3) Limsopatham N, Collier N. Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016 : 1014-23.

4) Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)-Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. JMIR Med Inform. 2019 ; 9 ; 7(3) : e14830.

5) Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform. 2014 ; 2 ; 47 : 1-10.

6) Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013 ; 11 ; 29(22) : 2909-17.

7) Li H, Chen Q, Tang B, et al. CNN-based ranking for biomedical entity normalization. BMC Bioinformatics. 2017 ; 18(S11) : 385.

8) Mondal I, Purkayastha S, Sarkar S, et al. Medical Entity Linking using Triplet Network. Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 : 95-100.

9) Sung M, Jeon H, Lee J, Kang J. Biomedical Entity Representations with Synonym Marginalization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 : 3641-50.

10) Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 ; 17(5) : 507-13.

11) Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013 ; 29(22) : 2909-17.

12) Aramaki E, Yano K, Wakamiya S. MedEx/J: A One-Scan Simple and Fast NLP Tool for Japanese Clinical Texts. Stud Health Technol Inform. 2017 ; 245 : 285-8.

13) Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Overview of the NTCIR-11 MedNLP-2 Task. Proc the 11th NTCIR Conference. 2014 : 147-154.

14) 清水尊. カルテ&レセプト略語 16000. 医学通信社, 2008.

15) 柴田知秀，河原大輔，黒橋禎夫. BERT による日本語構文解析の精度向上. 言語処理学会 第 25 回年次大会.2019 : 205-208.