分散表現と文字情報を用いた自由入力病名の表記揺れ解消

Orthographic Normalization of Disease Names Using Word Embedding and Syntactic Information

田代尚己 [1]　今井健 [2]

Naoki Tashiro[1], Takeshi Imai[2]

[1, 2] 東京大学大学院医学系研究科

[1,2] Graduate School of Medicine, The University of Tokyo

Abstract : Orthographic normalization is significant for the acquisition of knowledges from sentences, such as medical case reports. Many Japanese orthographic variations exist in medical sentences. To identify terms meaning the same disease, serious costs of human resources are required. Various preceding methods are studied to solve this problem, but some problems are still remains. Therefore, this paper proposes integrated model for orthographic normalization of disease names using word embedding and syntactic information.

1 はじめに

症例報告文章や電子カルテ、退院サマリー、人口動態調査の死亡表などの医療分野における自由入力文章には、標準病名マスターなど既存の標準用語集に基づく Post Coordination では表現できない病名の表記揺れが多数存在する。そのため、病名の分類や検索、医用文章からの知識抽出などの処理において表記揺れ病名とその正しい標準病名を同一視できない問題が生じる。よって、自由入力病名の表記揺れ解消は解決すべき重要な課題である。

表記揺れ解消における先行研究では表記揺れ用語の派生元用語を推定する transliteration による手法が日本語[1], [2]のみならず中国語[3]や韓国語[4]などの言語で提案されてきた。しかし、元用語が特定できない場合も存在するため、表記揺れが生じる知識を学習することで表記揺れを解消する手法が提案された[5], [6]が、未知の表記揺れには対応できない課題があった。他にも表記揺れ関係にある用語同士はスペルが近いという前提で編集距離の類似度を用いて表記揺れを解消する手法が提案されている[7]。しかしながら、漢字・平仮名・アルファベット表記の差異は十分に扱われていなかった。また、編集距離の類似度を用いた手法ではスペルが全く異なる同義語に適用できないことも問題である。

そこで、本研究では漢字・平仮名・アルファベット表記の差異や同義語も扱える自由入力病名に対する表記揺れの解消を目的とし、表記揺れ用語が有する意味の情報と文字の情報を用いた表記揺れ解消手法を提案する。そして、用語の意味情報から標準病名を推定する手法と文字情報から推定する手法をどのように組み合わせればより精度を高めることができるのか複数の統合方法を比較する。ここで、本稿で述べる表記揺れ解消とは、表記揺れ用語を標準病名マスターに登録されている標準病名の代表表記あるいは ICD10(2013)に紐づけることを意味する。

2 症例報告文章内の表記揺れ用語

本研究では AMED の臨床研究等 ICT 基盤構築研究事業「人工知能による総合診療診断支援システムの開発」で作成された CaseMap データベースを借用して研究の材料に用いた。CaseMap データベースは内科学会の 13,637 症例報告文章から疾患とその疾患が誘発する病態や症状所見を抽出したデータで構成されており、症例報告文章に記述された病態の遷移を捉えている。CaseMap データベースに登録されている用語数は 322,944 であり、用語種類数は 46,592 である。表 1 に CaseMap データベース内の病態遷移の例を示した。

表 1 CaseMap データベース内の病態遷移の例

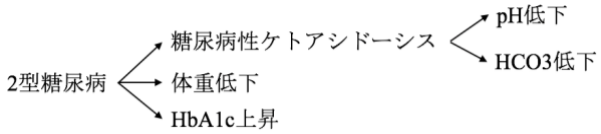

表 1 の 2 型糖尿病などある用語に着目した際にその上流に用語を持たない用語は病名であることが多いと考え、条件を満たす用語を集計し、標準病名マ

スターの代表表記、同義語、修飾語テーブルを用いてどの程度標準病名で被覆できるのか調査した。その結果、39,136 用語のうち 58%となる 22,776 用語が被覆できず、用語の種類別では 3,694 用語のうち 67%となる 2,471 用語が被覆できなかった。表 2 にそれらの被覆できなかった用語とそれらの名寄せ先となる標準病名の例を示した。表記揺れは一般に文字上の差異を意味すると思われるが本稿では名寄せ先標準病名と同義語関係にある被覆できなかった用語も表記揺れ用語として扱った。

表 2 標準病名で被覆できなかった用語とその正解標準病名の例

違いのパターン	被覆できなかった用語	標準病名
英語と日本語	骨Paget病	骨パジェット病
カタカナとひらがな	ツツガムシ病	つつが虫病
大文字と小文字	Klinefelter症候群	KLINEFELTER症候群
修飾語の有無	劇症型抗リン脂質抗体症候群	抗リン脂質抗体症候群
補足の有無	全身性エリテマトーデス(SLE)	全身性エリテマトーデス
文字の省略	リステリア髄膜炎	リステリア性髄膜炎
文字の差異	Creutzfeldt-Jakob病	CREUTZFELDT JACOB病
同義語	内臓幼虫移行症	トキソカラ症

3 手法

本研究では 2 章で取得した標準病名で被覆できなかった表記揺れ病名に対してその名寄せ先となる標準病名あるいは ICD10(2013)を 1347 種類の標準病名から正しく選択できるか試みた。表 1 における 2 型糖尿病や糖尿病性ケトアシドーシスのような下流に病態や症状所見を持つ用語群に存在する 1508 種類の標準病名に対して同義語をその代表表記と同一概念とみなし、1347 種類の標準病名を取得した。また、ICD10 で同一概念とみなすと 901 種類の ICD10 が存在した。

研究の流れとして、まず意味情報を用いて標準病名を推定する手法 4 種類と文字情報を用いて推定する手法 2 種類をそれぞれ比較し、精度の高かった手法を 1 つずつ選択した。その後、選択した手法を組み合わせた 3 種類の手法を比較検討した。

3.1 コーパス作成

表 1 に示した病態遷移の情報を以下に示したように Word2Vec に用いるコーパスとして作成した。

2型糖尿病 糖尿病性ケトアシドーシス pH低下 2型糖尿病 糖尿病性ケトアシドーシス HCO3低下 2型糖尿病 体重低下 2型糖尿病 HbA1c上昇

Word2Vec で異なる病態遷移データの用語の学習を防ぐためにハイパーパラメータである Window サイズの値と同数のカンマで異なる病態連鎖間を繋いだ。事前調査において、症例報告文章を形態素分割したコーパスよりも症例報告文章から抽出された病態遷移の情報をコーパスに用いた方が語の類似度判定の成績が良かったため、以下本稿では病態遷移の情報をコーパスとして用いた。

3.2 データセット作成

以下、3.3 節以降で述べるトレーニングデータセットにおいても共通して CaseMap データベースの病態遷移の情報を用いた。前処置として CaseMap データベースにおける用語の出現頻度で上位 500 用語を確認し、白血球と WBC のような症状・所見項目の正規化を行った。

テストデータセットには表記揺れ病名とその名寄せ先となる標準病名のデータセットを用いた。標準病名マスターで被覆できなかった 2,471 種類の用語のうち、CaseMap データベース内の出現頻度順 454 用語について、名寄せ先の標準病名を人手で正解付けした。今回、提案する手法は病名の意味情報を用いたため、表記揺れ病名の名寄せ先となる標準病名が上記の 1347 種類内に存在していなければ解くことができない。そのため、名寄せ先に正解付けした標準病名が 1347 種類内に存在する 246 データセットをテストデータセットに用いた。

3.3 意味情報モデルと文字情報モデル

A. 意味情報モデル 1 (SEM1)

3.1 で作成したコーパスを用いて Word2Vec で用語の分散表現を獲得し、コサイン類似度から表記揺れ病名に対する標準病名及び ICD10 を推定した。

B. 意味情報モデル 2 (SEM2)

3.1 で作成したコーパス内の各用語に対して形態素分割処理を行ったコーパスを用いて Word2Vec で用語の分散表現を獲得し、コサイン類似度から表記揺れ病名に対する標準病名及び ICD10 を推定した。形態素分割は MeCab に IPA 辞書、表記揺れ病名、標準病名、LiLak 症状所見辞書を追加して行った。

C. 意味情報モデル 3 (SEM3)

下流に何らかの病態や症状所見を持つ用語群に含まれる標準病名に対し、その疾患が誘発する下流の病態や症状所見を Bag of Words 形式で取得し、その疾患の分散表現ベクトルとみなした。ベクトルの次元数は 17,769 であり、全症例で 1 度しか出現しなか

った用語は考慮しなかった。表 1 を例にとると 2 型糖尿病、体重低下、HbA1c 上昇、糖尿病性ケトアシドーシス、pH 低下、HCO3 低下が 1、その他の病態・症状所見が 0 となった分散表現ベクトルとなる。

獲得した分散表現ベクトルを入力とし、1347 の標準病名を出力とする 3 層のニューラルネットワークを構成した深層学習モデルを用いて学習した。テストとして表記揺れ病名についても同様のデータを作成し、標準病名及び ICD10 を推定した。ただし、モデルの条件は中間層 200 次元、中間層の活性化関数 Relu、勾配降下法 (SGD)、学習率 0.01、バッチサイズ 10、エポック数 300 に設定した。

D. 意味情報モデル 4 (SEM4)

C で獲得した疾患が誘発する下流の病態や症状所見の用語に対して形態素分割処理を行った上で形態素を Bag of Words 形式で取得し、これを標準病名の分散表現ベクトルとみなした。ベクトルの次元数は 7,366 であり、1 度しか出現しなかった形態素は考慮しなかった。形態素分割は B と同条件で行った。

テストとして表記揺れ病名についても同様のデータを作成し、C と同条件の深層学習モデルで標準病名及び ICD10 を推定した。

E. 文字情報モデル 1 (SYN1)

各 246 表記揺れ病名に対して対象となる 1347 標準病名との編集距離の類似度を求め、類似度の高さで標準病名を推定した。編集距離の類似度を計算した式を以下に示した。

$$\text{編集距離の類似度}(t_1, t_2) = 1 - \frac{Edit\ Distance(t_1,\ t_2) \times 2}{len(t_1) + len(t_2)}$$

編集距離(Edit Distance)は 2 つの文字列間において一方の文字列をもう一方の文字列に変換するための文字の置換、削除、挿入の数である。[8] 式中の $len(t_n)$ は用語 t_n の文字数である。編集距離の類似度は-1 から 1 の値を取りうる。

F. 文字情報モデル 2 (SYN2)

標準病名に対して構成文字を分散表現として Bag of Words 形式で取得し、これを標準病名の文字情報ベクトルとみなした。また、表記揺れ病名の正解となる標準病名が標準病名マスターの同義語テーブルに同義語を持つ場合はその同義語も学習に用いた。文字情報ベクトルの次元数は 826 となった。

テストとして表記揺れ病名についても同様のデータを作成し、C と同条件の深層学習モデルで標準病名及び ICD10 を推定した。

3.4 統合モデル

用語の正解数を指標にした時に高い精度を示した手法は意味情報では SEM4、文字情報では SYN2 であったため、それらのモデルを統合モデルに用いた。

A. 統合モデル 1 (INT1)

入力に用いた分散表現ベクトル同士を統合した。次元数は 8,192 となった。

テストとして表記揺れ病名についても同様のデータを作成し、3.3 節の C と同条件の深層学習モデルで標準病名及び ICD10 を推定した。

B. 統合モデル 2 (INT2)

表記揺れ病名ごとに解くモデルを変えることでより精度が向上すると予測した。そこで、各 246 表記揺れ病名に対して対象となる 1347 標準病名との編集距離の類似度を求めた際に最も類似度の高い値が設定した閾値よりも高い値を持つ表記揺れ病名グループ、それ以外のグループに分けた。なお、閾値を 0.1 ずつずらして比較し、より精度が高くなった閾値を調べた。

閾値よりも高い値を持つグループには文字情報モデル、低いグループには意味情報モデルを適用し、それぞれのグループごとに標準病名及び ICD10 を推定した。

C. 統合モデル 3 (INT3)

深層学習モデルは出力層で 1347 種類の標準病名に対して何らかの値を出力し、その後、最も値が高かった標準病名を推定病名として選択している。そこで、SEM4 と SYN2 がそれぞれ最終層で出力した 1347 標準病名に対する値を合計してから最も高かった値で標準病名及び ICD10 を推定するモデルを考えた。

合計する際に各モデルの値の比率を 0%から 100%の間で 10%ずつずらして寄与率を変化させ、より精度が高くなった寄与率を調べた。

4 実験結果

4.1 意味情報、文字情報モデルの結果

表 3 に各意味情報モデルと各文字情報モデルのテストにおける標準病名の正解数とその正解率、ICD10 の正解数とその正解率、標準病名の TOP10 正解数とその正解率を示した。標準病名の TOP10 正解数はモデル推定後に対象となる標準病名を類似度の高かった順に並べたとき上位 10 位以内に名寄せ先

となる標準病名が入っていた数である。いずれにおいても正解数は 246 表記揺れ病名のうちいくつ正解したかを表す。

表 3 実験結果 1

	標準病名 正解数	標準病名 正解率[%]	ICD10 正解数	ICD10 正解率[%]	標準病名 TOP10 正解数	標準病名 TOP10 正解率[%]
SEM1	20	8.1	23	9.3	39	15.9
SEM2	39	15.9	44	17.9	92	37.4
SEM3	66	26.8	73	29.7	152	61.8
SEM4	80	32.5	93	37.8	168	68.3
SYN1	70	28.5	90	36.6	129	52.4
SYN2	91	37.0	99	40.2	152	61.8

SEM1, 2 と 3, 4 では標準病名の正解率に倍程度の差が見られ、3, 4 の方が良い精度を示したことから、Word2Vec で病名の分散表現を獲得し、コサイン類似度による類似度評価よりも、疾患が有する病態や症状所見の Bag of Words 形式の分散表現を獲得し、3 層のニューラルネットワークで類似度評価を行った手法の方が病名の意味を捉えやすいことがわかった。また、SEM1 よりも 2、3 よりも 4 の正解率の方が高かったことから、用語の形態素分割が精度向上に寄与することがわかった。

SYN1 と 2 では 2 の正解率の方が高かったことから、編集距離の類似度を用いるよりも Bag of Character として構成文字の情報を用いる方が精度向上に寄与することがわかった。

ICD10 を推定する結果ではいずれのモデルにおいても標準病名の正解率よりも ICD10 の正解率の方が高かった。このことは、1 位には正解とした標準病名は出現しなかったが同じ ICD10 を持つ標準病名が 1 位に出現していたことを意味する。

4.2 統合モデルの結果

表 4 に INT1 から INT3 における結果を示した。

表 4 実験結果 2

	標準病名 正解数	標準病名 正解率[%]	ICD10 正解数	ICD10 正解率[%]	標準病名 TOP10 正解数	標準病名 TOP10 正解率[%]
INT1	102	41.5	110	44.7	154	62.6
INT2	108	43.9	118	48	188	76.4
INT3	138	56.1	146	59.3	210	85.4

統合モデル 2 において編集距離の類似度の閾値を 0.2 に設定したときに正解率が最も高かった。

統合モデル 3 において意味情報モデルの出力値を 50%、文字情報モデルの出力値を 50%として統合したときに正解率が最も高かった。

モデル間に着目すると INT3 で用いた手法が他と比較して最も病名の表記揺れ解消に適していることがわかった。

4.3 病名の表記揺れが解消できた例

表 5 に SEM4 が SYN2 より優位な結果であった例例を、表 6 に SYN2 が SEM4 より優位な結果であった例をそれぞれ 10 ペアずつ示した。

表 5 SEM4 が優位な結果であった例

SEM4でTOPに名寄せ先標準病名が出現した例	
表記揺れ病名	標準病名
肝原発悪性リンパ腫	悪性リンパ腫
成人型Still病	成人スチル病
Crowned_Dens_Syndrome	偽痛風
Guillan-Barre症候群	ギラン・バレー症候群
成人T細胞性リンパ腫	成人T細胞白血病リンパ腫
Churg-strauss症候群	好酸球性多発血管炎性肉芽腫症
Fibrillary_glomerulonephritis	糸球体腎炎
IgD型多発性骨髄腫	多発性骨髄腫
Multicentric_CastlemanDisease	キャッスルマン病
Henoch-Schoenlein紫斑病	IgA血管炎

表 6 SYN2 が優位な結果であった例

SYN2でTOPに名寄せ先標準病名が出現した例	
表記揺れ病名	標準病名
びまん性大細胞型リンパ腫	びまん性大細胞型B細胞性リンパ腫
非閉塞性腸間膜虚血症(NOMI)	非閉塞性腸間膜虚血
インフルエンザウイルス肺炎	インフルエンザ肺炎
Langerhans細胞組織球症	ランゲルハンス細胞組織球症
抗基底膜抗体型急速進行性腎炎	急速進行性糸球体腎炎
緩徐進行性1型糖尿病	緩徐進行1型糖尿病
亜急性連合変性症	亜急性連合性脊髄変性症
CML_lymphnodalblast_crisis	慢性骨髄性白血病
Diffuse_Large_BCellLymphoma	びまん性大細胞型B細胞性リンパ腫
肝型ウイルソン病	ウイルソン病

SEM4 を用いることで Crowned_Dens_Syndrome と偽痛風や Guillain_Barre_Syndrome_Variant とギラン・バレー症候群など編集距離の類似度や構成文字では解くことができない病名の表記揺れを解消することができたことからアルファベット・漢字・カタカナ間の表記揺れにも対応できることがわかった。

文字情報モデルが正解した例からほとんどは文字列上の差異を捉え、表記揺れを解消していることがわかったが、CML_lymphnodalblast_crisis と慢性骨髄性白血病は一見構成文字では推定できないと思われる。しかし、標準病名マスターに代表表記である慢性骨髄性白血病を示す CML が同義語として登録されているため、構成文字で正解できたことがわかった。

5 考察

いずれのモデルにおいても形態素分割を行ったモデルの方が行っていないモデルよりも良い精度を示した理由は、形態素分割が分散表現ベクトルの次元数の抑制に寄与したからであると考えられる。(実際に SEM3 の次元数は 17,769 であり、SEM4 の次元数は 7,366 であった。)

編集距離の類似度を用いたモデルよりも構成文字を用いたモデルの方が精度が高かった理由は、日本語の漢字には癌や炎など文字自体に意味があるため、この情報が表記揺れ病名の解消に貢献したと推察される。

いずれのモデルにおいても ICD10 を推定したときの正解率が標準病名を推定したときの正解率を上回った結果から、ICD10 分類としては正解していても文字列上における標準病名の差異を扱うための病態遷移の情報が CaseMap データベースに存在していなかったためにその差異を捉えられなかったのではないかと示唆される。

本来、意味情報が十分に存在していれば表記揺れ解消に文字情報を用いる必要はないはずであるが、統合モデルの方が意味情報モデルよりも良い精度を示したことから、CaseMap データベース内の病態遷移の情報には類似疾患を区別するのに十分な情報量が含まれていなかったため、意味情報と文字情報の組み合わせによって相乗効果が生じたと考えられる。

今回のタスクでは標準病名を 1/1347、ICD10 を 1/901 で正確に当てられるかを解いているため、未学習時の正解率はそれぞれ 0.074%と 0.11%である。これを踏まえると統合モデル 3 における標準病名の正解率 56.1%、ICD10 の正解率 59.3%、標準病名の TOP10 正解率 85.4%は十分に高く、INT3 が病名の表記揺れ解消に実用的と考えられる。

今後より精度を向上させるためには INT3 の手法を考慮すると SEM4 と SYN2 のそれぞれを改善する必要があると考えられる。SEM4 が正解した表記揺れ用語とその標準病名、不正解であった表記揺れ用語とその標準病名の結果に差がないか調査したところ、正解した表記揺れ用語の標準病名はトレーニングデータセット内に平均 35 個存在したのにもかかわらず、不正解であった表記揺れ用語の標準病名はトレーニングデータセット内に平均 22 個しか存在しなかった。このことから、今後症例数の蓄積に伴い、学習が促進されると考えられる。

また、今後表記揺れ病名とその標準病名の正解データセットを多く作成し、学習データに表記揺れ病名を追加することで表記揺れの情報を学習できれば特に SYN2 における精度向上に寄与するのではないかと予想している。

6 おわりに

本稿では意味情報モデル SEM4 と文字情報モデル SYN2 の出力層の結果を統合した INT3 モデルを用いることで個々のモデルよりも多くの自由入力病名の表記揺れ解消ができることを示した。

本研究の限界として正解標準病名が CaseMap データベース内にないと表記揺れの解消はできない。そのため、今後未登録の標準病名を有する症例を加えていくことでより多くの表記揺れ用語に対応できると考えられる。

本研究は自然言語を用いたアプリケーションにおいて、病名の表記揺れを解消するタスクや ICD10 分類をするタスクへの適用を想定している。今回は疾患名のみに注目したが、今後は症状所見を対象とする予定である。表記揺れ用語を人手で個々に解消するタスクは人的コストが大きい。そこで、あらかじめ数個程度の候補を表示し、その中から選択し、なければ他を探すような形式で利活用されることを目指している。本研究が表記揺れ解消研究の礎となることを期待している。

参考文献

[1] Bilac S and Tanaka H. A hybrid back-transliteration system for Japanese: In Proceedings of The 20th International Conference on Computational Linguisics, COLING2004, pp. 597-603, 2004.

[2] Goto I, Kato N, Ehara T, and Tanaka H. Back transliteration from Japanese to English using English context: In Proceedings of The 20th International Conference on Computational Linguisics, COLING2004, pp. 827-833, 2004.

[3] Youn S, Kim K and Sproat R. Multilingual transliteration using feature based phonetic method: In Proceedings of the Annual Meeting of the Association of computational Linguistics, ACL2007, pp.112-119, 2007.

[4] Oh J and Choi, K. An English-Korean transliteration model using pronunciation and contextual rules: In Proceedings of the 19th International Conference on Computational Linguistics, COLING2002, pp.758-764, 2002.

[5] Aramaki E, Imai T, Miyo K, and Ohe K. Support vector machine based orthographic disambiguation: In Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation,

TMI2007, pp. 21-30, 2007.
[6] Bergsma Sand Kondrak G. Alignment-based discriminative string similarity: In Proceedings of the Association for Computational Linguistics, ACL-2007, pp.656-663, 2007.
[7] Aramaki E, Imai T, Miyo K, and Ohe K. Orthographic Disambiguation Incorporating Transliterated Probability: In Proceedings of the 3rd International Joint Conference on Natural Language Processing, ACL-2008, pp.48-55
[8] Levenshtein V. Binary codes capable of correcting deletions, insertions and reversals: Doklady Akademii Nauk SSSR, Vol. 163, No. 4, pp. 845-848, 1965.