膨大な医学知識を用いたカルテ入力誤り文検出

古賀　貴士† 　矢田竣太郎† 　若宮　翔子† 　荒牧　英治†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916-5

E-mail: †{t-koga,s-yada,wakamiya,aramaki}@is.naist.jp

あらまし　近年導入が進む電子化された診療録（電子カルテ）は，医師による患者動態の網羅的な追跡を可能とし，臨床業務や医学研究の効率化に貢献してきた．一方で，医師が電子カルテに記録する際に重要な病名等の誤入力が発生すると，最悪の場合，致命的な医療事故に発展しうる．そこで本研究では，病名に対する入力誤り文の検出を目指す．そのために，医学テキストをもとに擬似誤り文を生成し，誤り文を検出するモデルを構築する．具体的には，まず，擬似誤り文は，症例報告論文の文中に出現する病名をランダムに置換することで，擬似誤り文を生成した．生成した誤り文に対して，擬似誤り文のみを入力とする手法，擬似誤り文と標準病名を入力とする手法，擬似誤り文と近傍事例の総数を入力とする方法，擬似誤り文と近傍事例を入力とする手法を提案し，誤り文の検出を行なった．実験では，擬似誤り文と標準病名を入力とする手法が最も精度が高く，F1-score が 0.74 であった．

キーワード　誤り文検出，電子カルテ，テキスト分類，医療情報システム，自然言語処理

1　はじめに

1.1　背　　景

近年，情報通信技術を活用した電子化された診療録（電子カルテ）の実用化に伴い，医学・診療情報の蓄積が急速に発展してきている．現在，電子カルテは大規模病院のほとんどで導入され，中小規模病院や診療所でも約半数が導入されている [1]．電子カルテは医師による患者動態の網羅的な追跡を可能とするため，新しい医学的知見や既存薬の新しい効能の発見，診療行為の質の評価，まれな疾患や薬物の有害事象の調査など，臨床業務や医学研究の効率化に貢献してきた．これまでに記録されてきた膨大な診療録を利活用することで，診断支援や類似症例検索といった医療情報サービスの開発も進められている．

一方で，医師や看護師による電子カルテへの記録はフリーテキストで記載されることが多く，入力ミス（以下，誤入力）による医療事故が報告されている [2]．誤入力に関する医療事故の大部分は患者確認をおろそかにすることが原因である．しかし，紙カルテと同様に，人間が電子カルテを入力するため，ヒューマンエラーを避けては通れない．誤入力に関する事例として，患者確認ミスによるものの他に，漢字や文節の変換ミス，読み間違いや用語の意味理解の誤りなどによる誤字などが観測されている [3]．

臨床現場において誤変換や誤字による誤入力は軽微な問題であると考えられるが，用語理解を誤った誤入力は重大な医療事故を引き起こしかねない．例えば，医師が電子カルテに記録する際に重要な病名等を誤入力すると，最悪の場合，致命的な医療事故に発展することが容易に想像できる．この問題を解決するためには，電子カルテに記入されたテキストに入力誤りが存在するか否かを検知する必要があると考えられる．

1.2　目的とアプローチ

本研究では，電子カルテに記入されたテキストから，病名の誤入力がみられる文を検出するモデルの構築をする．入力誤り文のモデルを構築することにより，医療事故の削減に寄与する他，電子カルテに記述された内容の更なる信頼性向上につながることが期待できる．

一般に誤り文の検出は，誤りがあるかないかを分類するため，文書分類タスクとして捉えられる．文書分類タスクの代表的なアプローチの一つとして，Bidirectional Encoder Representations from Transformers (BERT) [4] が挙げられる．BERT は大規模な汎用的な言語モデルであり，文書分類を含む 11 のタスクで発表当時の SoTA を記録した．現在では日本語の事前学習済みモデルも複数公開されており，ファインチューニングを行うだけで高精度な分類が可能である．しかし，基本的な BERT モデルは Wikipedia などの一般ドメインの文書をコーパスとしているため，医療など特定のドメインでは性能が低くなることが報告されている [5]．日本語の医療ドメインにおいては，BERT の事前学習済みモデルとして東京大学大学院医学系研究科が UTH-BERT [6] を公開している．UTH-BERT は，1 億 2000 万行の日本語診療記録から事前学習されており，医療ドメインに特化した BERT である．

本研究では UTH-BERT をファインチューニングし，カルテ入力誤り文を検出するモデルの構築を行う．カルテ入力文に含まれる誤入力の判断などといった，医学的な意味解析を行うためには，膨大な医学的知識が必要と考えられる．UTH-BERT は医療文書に特化した事前学習モデルであるが，更なる医学的知識を補うべく，本研究では症例報告論文を外部知識として利用し，入力誤り文を検出するモデルの構築を行う．症例報告論文は，患者の診断名・転帰，入院時の症状や初見，治療後の経過など，患者情報が要約された医学文書であり，医師の教育や類似症例の参考のために用いられる．症例報告論文は学会に報

告される文書であることから，電子カルテとは異なるが，多くの医学用語が含まれるため，有用な外部知識であると考え，本研究の材料として用いた．

本研究の提案モデルが検知する入力誤り文は，アノテーション済みの医療文書から機械的に生成した．この擬似的な誤り文（以降，擬似誤り文）は，病名や症状に対してアノテーションされた病名タグを，異なる病名タグに置換した，病名誤りの擬似誤り文である．この擬似誤り文と置換操作を行わなかった原文に対して，本研究では誤り文を検出する文書分類モデルを提案する．

提案モデルは大きく分けて，文レベルの誤り検出モデルと文書レベルの誤り検出モデルの２つからなる．1 文を入力とした誤り検出はコンテキスト情報が少ないため，誤り検出のコストが小さい反面，短い文に対して誤りを検出することが困難である．そこで本研究では，文レベルの誤り検出に加えて文書レベルの誤り検出も行う．

2 関 連 研 究

誤り検出における主な研究は，文法誤り訂正や文章構成などのタスクが挙げられる [7–10]．しかし，これらのほとんどは文法的な誤りやタイプミスを扱った研究であり，本研究のような専門知識を要する意味誤りを検出する研究はほとんどされていない．文中の語意を解析する意味解析等の研究は古くからされてきたが，医学的意味誤りを検出するには，医学的知識をベースとしたナレッジグラフなどの外部知識が必要となる．このように本研究と関連が深いのは，外部知識，その中でも特に，近傍事例を用いた機械学習である．そこで，本章では近傍事例を用いた既存研究についての説明を行う．

ニューラル言語モデルが学習する際，すべての世界知識を訓練データのみで学習することは，パラメータ増加などのニューラル言語モデルの肥大化や学習コストの増加など様々な問題が挙げられる．そのため近年では，訓練データでモデルを学習するだけでなく，推論時に入力と関連のあるテキスト（近傍事例）を用いることで，ニューラル言語モデルの性能を高める研究が様々な言語処理タスクにおいて注目されている．例として，テキスト生成，機械翻訳，質問応答モデルなどが挙げられる．

関連のある文書の例として，Wikipedia の記事全体で “tiger” という単語の出現頻度は 0.0037%であるのに対し，トラに関する記事の中では “tiger” が出現頻度は 2.8%にも上ることが報告されている [11]．このように，ある文書で出現した単語が，再度，その文書で出現する確率は高まる傾向にある．Grave らはこの特徴に着目し，学習済みの言語モデルにキャッシュモデルを導入した言語生成モデルを構築した．具体的に，近傍事例ベクトルを用いて計算した単語予測分布を，言語モデルの単語予測分布と線形補完することで，言語モデルの perplexity の改善を達成した．近傍事例を用いる際のアルゴリズムは線形補完の他に，k-NN アルゴリズムを用いた kNN-LM [12] などが提案されている．

機械翻訳モデルでは，k-NN をベースとした Nearest Neighbor Machine Translation (kNN-MT) [13] が研究されている．加えて，kNN-MT モデル近傍事例検索の改善に関する研究 [14] や近傍事例を格納するデータストアのコンパクト化を行う研究 [15]，知識の蒸留を用いた研究 [16] なども進められている．

また近傍事例を用いたオープンドメイン質問応答タスクでは，Open-Retrieval Question Answering (ORQA) [17] や Retrieval-Augmented Language Model Pre-Training (REALM) [18]，言語生成モデル T5 をベースとした Fusion-In-Decoder (FiD) [19] などが研究されている．

表 1: データセットの詳細

	文書数	文の総数	病名や症状を含む文の総数
MedTxt-CR-JA	224	3,444	2,069
JST 症例データセット	94,711	520,357	306,857

3 材　　料

3.1 データセット

本研究では 2 つのデータセットを用いる．一つ目は，奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室が公開している症例報告論文テキストのコーパス **MedTxt-CR-JA** である．MedTxt-CR-JA は，J-STAGE[1]でオープンアクセス公開されている症例報告論文 PDF から OCR で抽出したテキストのコーパスである．一般的な病名について実際の出現頻度にそって文書数がバランスされており，2010 年以降かつ本文が 1500 字以内の症例報告に対してアノテーションがされている．本研究では，このデータセットを用いて，擬似誤り文を生成する．

二つ目のデータセットは，J-STAGE に収録された症例報告論文のアブストラクトを抽出した **JST 症例データセット**を使用している．JST 症例データセットは 1975 年以降に収録された論文のうち，タイトルに “症例” を含む論文のアブストラクトを抽出したデータセットである．本研究では，このデータセットを文書分類時の外部知識として利用する．

3.2 擬似誤り文の生成

本研究の目的であるカルテ入力誤り文の検出は，文中の病名や症状に誤りがあるかないかの二値分類タスクとする．本タスクを解くにあたり誤り文が必要となるが，実際の誤り文を手に入れにくい他，人手による擬似データの生成では多くの時間や労力を要する．そのため，擬似誤り文を機械的に生成し，擬似誤り文の評価を行う．

擬似誤り文の生成は，文法誤り訂正のタスクでしばしば行われる．擬似誤り文の生成手法として，逆翻訳を用いた手法 [7]，言語学習者の誤り傾向を考慮した手法 [8]，置換や挿入など擬似誤りを直接生成する手法 [9] などが挙げられる．

病名の入力誤り文を検出するためには，文法的な誤りではな

1：J-STAGE https://www.jstage.jst.go.jp/

ID	SEX	AGE	CATEGORY	DATE
JP0900-1	FEMALE	77	変形性膝関節症	-1

行	本文
1	【背景およびプロフィール】
2	D(+) 腸脛靭帯摩擦症候群 はランニングの A 足部 時(MISC) 接地後 、 A 膝関節 屈曲３０°弱で A 大腿骨外側上顆と腸脛靭帯（ＩＴ ｂａｎｄ以下ＩＴＢ）の後方 で摩擦が生じることにより C 発症 する。
3	D(+) 疼痛 部位としては、 A 膝外側部、大腿骨外側上顆付近 が挙げられる１）。
4	時(DATE) 今回 、 腸脛靭帯摩擦症候群 を疑った患者に対し、評価とクリニカルリーズニングを経て R(+) プレーティングアプローチ を行い結果が良好であった症例について報告する。
5	対象は 時(AGE) ７７歳 女性。

図 1: 病名タグを含む MedTxT-CR-JA の文書例 [2]

図 2: 病名タグの出現回数の分布

く医学的意味誤りを含む文を生成する必要がある．そこで本研究では，病名や症状に対してアノテーションした病名タグを，ランダムに置換を行うことで擬似誤り文を生成した．置換操作の対象間において品詞は同じであるため文法的には正しいが，医学的に文意の通らない擬似誤り文を効率的に生成できる．擬似誤り文の生成元として MedTxt-CR-JA を使用し，以下の手順に従って生成した．Med-TxT-CR-JA における文書例を図 1 に示す．

（1） MedTxt-CR-JA の各文書を文単位に分割し，病名タグが出現する文を抽出する（2,069 文）

（2） 抽出した 2,069 文のうち，擬似誤り文の生成対象として 1,034 文（半数）をランダムに選択する

（3） 擬似誤り文の生成対象の各文にて，出現する病名タグをランダムに 1 つ選択する

（4） 選択した病名タグを，MedTxt-CR-JA 内で同程度の頻度で出現する病名タグに置換する

病名タグの出現回数を図 2 に示す．横軸は MedTxt-CR-JA 内に出現する各病名タグの出現回数を示し，縦軸は各出現回数における病名タグの総数を示している．病名タグの出現回数はロングテール分布となっている（図 2）．重複を除いた病名タグの総数は 2,294 単語であり，ほとんどの病名タグにおいて MedTxt-CR-JA 内で出現する回数が 3 回以下であった．そのため，単語の出現頻度を考慮せず病名タグをランダムに置換すると，稀にしか出現しない病名タグに置換される可能性が高くなる．そこで，病名タグの出現頻度に応じたクラス分けを行い，同一クラス内で置換を行なった．

2：MedTxt-CR: 症例報告 (Case Reports) より引用 https://sociocom.naist.jp/medtxt/cr/

表 2: 人手による擬似誤り文の分類結果

	Precision	Recall	F 値
医療従事者 (n=2)	0.88	0.91	0.89
非医療従事者 (n=1)	0.60	0.24	0.34

3.3 擬似誤り文の評価

擬似誤り文は文中の病名タグのみを置換していることから，文法的な誤りは発生せず，医学的に文意が通らない文章が生成されているはずである．そこで，擬似誤り文が医学的に文意が通らない文章となっているかを医療従事者 2 名が人手評価した．比較対象として非医療従事者 1 名による評価も行なった．全ての擬似誤り文に対して評価することは非常にコストが大きくなる．そのため，ここでは，そのため，ここでは，生成した擬似誤り文（1,034 文）からランダムにサンプリングした 50 文と，原文（1,035 文）からランダムにサンプリングした 50 文の計 100 文を評価した．正しく擬似誤り文と判断できた場合を True Positive (TP)，擬似誤り文と判断したが実際は擬似誤り文でなかった場合を False Positive (FP)，正しく擬似誤り文でないと判断できた場合を True Negative (TN)，擬似誤り文でないと判断したが実際は擬似誤り文であった場合を False Negative (FN) として，Precision，Recall，F1-score を算出した．医療従事者の Precision，Recall，F1-score は算術平均で表している．

表 2 において，非医療従事者は医学的知識を有していないため，どのスコアも低い値となった．また，Precision，Recall のスコアから，誤りが存在する文に対して正確に分類できていないことが分かる．一方で，医療従事者による分類結果は，全ての評価指標で高いスコアとなっている．このことから，本手法で生成される擬似誤り文は，高確率で医学的に文意が通らない文になっているといえる．

医療従事者が誤って分類した文の傾向を調べるため，擬似誤り文および原文から抽出できる標準病名の個数を調査した．標準病名とは電子カルテにおける病名表現の違いを無くすために，1 つの疾患につき割り当てられた 1 病名表現のことを指す．標準病名の抽出には，形態素解析ツール mecab，mecab-ipadic-NEologd 辞書 [3]，万病辞書 [4]を用いた．その結果，医療従事者が誤って分類した文の約半数が，標準病名の個数が 0 個あるいは 1 個であった．このことから出現病名の少ない文は，周囲のコンテキスト情報から誤り文かどうかを判断できない可能性が考えられる．

4 提案手法

提案手法は大きく分けて，文レベルの誤り検出モデルと文書レベルの誤り検出モデルの 2 つからなる．文レベルの誤り検出

3：mecab-ipadic-NEologd https://github.com/neologd/mecab-ipadic-neologd

4：万病辞書 https://sociocom.naist.jp/manbyou-dic/

表 3: 医療従事者が分類を誤った例

	擬似誤り文	正解ラベル	予測ラベル	病名 入れ替え前	病名 入れ替え後
(1)	【はじめに・目的】ＣＫ７患者に対し，高負荷筋力増強訓練により筋力やＡＤＬが向上するとの報告が散見されるが，一方で筋炎患者では低筋機能のため高負荷運動の遂行が困難な症例も多い.	誤りあり	誤りなし	筋炎	ＣＫ７
(2)	岡（２００７）によると肉芽組織内の筋線維芽細胞はお互いを引き付けるような収縮作用をもつため、肉芽組織が収縮するようになり**リンパ節腫脹**部の収縮、上皮形成へと繋がっていくとされている。	誤りあり	誤りなし	創	リンパ節腫脹
(3)	再度，経皮的生検を行い、x 特殊染色を含む病理学的検索にて、腫瘍は比較的明るい細顆粒状の胞体を有する異型類円型細胞の蜂巣状、シート状の増殖からなり、ｓｉｇｎｅｔ ｒｉｎｇ ｃｅｌｌもみられた。	誤りなし	誤りあり	-	-
(4)	ＰｅｒｓｏｎａｌｉｚｅｄＭｅｄｉｃｉｎｅの強皮症での運用方法を検討するため，イマチニブもしくはＣＤＤＯが６名の強皮症患者皮膚由来の線維芽細胞のコラーゲン産生に及ぼす効果について検討した.	誤りなし	誤りあり	-	-

モデルでは，擬似誤り文の 1 文に対して誤り検出を行い，文書レベルの誤り検出モデルでは連続した複数の擬似誤り文に対して誤り検出を行った．文レベルの誤り検出モデルへの入力文例を表 4 の S_1 から S_4 に，文書単位の誤り検出モデルへの入力文例を表 4 の S_5 に示す.

4.1 文レベルの誤り検出

文レベルの誤り検出のために，事前学習済みモデル UTH-BERT を用いた 4 つの手法を提案する．**BERT-BASE**, **BERT-Jmed** は MedTxt-CR-JA から生成した擬似誤り文データセットと文中から得られる情報を入力文としており，**BERT-JSTJmedNum**，**BERT-JSTJmedNum** は擬似誤り文データセットに加えて，外部知識として JST 症例データセットを用いる.

- **BERT-BASE**: 擬似誤り文のみを入力文として学習する．**BERT-BASE** の入力文の形式を図 3(a) に示す.
- **BERT-Jmed**: 擬似誤り文と擬似誤り文に出現する病名の標準病名を [SEP] で連結し，入力文として学習する．標準病名は，擬似誤り文の評価の際と同様に，mecab+mecab-ipadic-NEologd+万病辞書により抽出した（以降，提案モデルにおける標準病名の抽出には，同様の手法を用いる）．**BERT-Jmed** の入力文の形式を図 3(b) に示す.
- **BERT-JSTJmedNum**: 擬似誤り文に出現する病名の標準病名が全て含まれる JST 症例データセットの文数を算出し，擬似誤り文と JST 症例データセットの文数を [SEP] で連結し，入力文として学習する．**BERT-JSTJmedNum** の入力文の形式を図 3(c) に示す.
- **BERT-JSTJmedTxt**: 擬似誤り文と擬似誤り文に出現する病名の標準病名が全て含まれる JST 症例データセットの文を [SEP] で連結し，入力文として学習する．擬似誤り文と JST 症例データセットの文との組合せは，擬似誤り文 1 文につき 30 セットまでとし，最大セット数を超える場合はランダムに JST 症例データセットの文を選定した．予測時は，1 つの擬似誤り文に対応した 30 セットの予測結果を多数決し，その値を最終的な予測結果とした．**BERT-JSTJmedTxt** の入力文の形式を図 3(d) に示す.

4.2 文書レベルの誤り検出

文書レベルの誤り文の検出のために，事前学習済みモデル UTH-BERT を用いた 2 つの手法を提案する．文レベルの誤り検出では，コンテキスト情報が少ない時，入力文に含まれている病名が誤りかどうか判断できない場合がある．本節では複数の連続した文を入力した，文書レベルの誤り検出モデルを提案する．しかし，コンテキスト情報多いほどが複雑になり，誤り検出するコストが大きくなる傾向がある．したがって，文書レベルの誤り検出モデルでは擬似誤りデータセットのみを入力文として使用し，文書の最初の 5 文を入力とした分類モデルを作成した.

- **$\mathbf{BERT_{doc}}$-BASE**: 擬似誤り文データセットのみを入力文として学習する．**$\mathbf{BERT_{doc}}$-BASE** の入力文の形式を図 4(a) に示す.
- **$\mathbf{BERT_{doc}}$-Jmed**: 擬似誤り文と擬似誤り文に出現する病名の標準病名を [SEP] で連結し，入力文として学習する．**$\mathbf{BERT_{doc}}$-Jmed** の入力文の形式を図 4(b) に示す.

5 実　　験

5.1 設　　定

本研究の目的は電子カルテに記載された病名の入力ミスを検出することであるため，擬似誤り文が医学的に文意の通らない文となるケースは除外する必要がある．本研究では文レベルの誤り検出と文書レベルの誤り検出を行うが，文レベルの誤り検出では文中から抽出した標準病名の個数が 2 つ存在する文（519 文）を対象に実験を行うこととする．なお，文書レベルの誤り検出では，文書中に標準病名が必ず複数個含まれるため，特に対象を制限せず，224 文書に対して実験を行う.

UTH-BERT モデルの fine tuning には，誤り文であるか否

表 4: 提案モデルへの入力文例（S_1 から S_4 が文レベルの誤り検出モデル，S_5 が文書レベルの誤り検出モデルへの入力文例）

ID	文例	出現病名	標準病名	材料
S_1	６月中旬から上肢，体幹，大腿に痒性紅斑出現。	痒，性紅斑	そう痒，紅斑症	擬似誤り文
S_2	糖日光にあたるとかゆみを伴う紅斑と膨疹が出現した日光じん麻疹の症例について述べた。	かゆみ，紅斑，膨疹，麻疹	そう痒，紅斑症，じんま疹，麻疹	JST 症例データセット
S_3	２３歳の男性は頸の両側に掻痒性紅斑発疹を生じた。	掻痒，性紅斑，発疹	そう痒，紅斑症，発疹	JST 症例データセット
S_4	症例は７９歳の男性で，掻痒を伴う全身の紅斑および頸部リンパ腫の膨脹を主訴に受診した。	掻痒，紅斑，頸部リンパ腫	そう痒，紅斑症，リンパ腫	JST 症例データセット
S_5	症例は７０歳の女性。２週間前から血便が続いていたが放置。その後突然の早期食道癌があり救急車で来院した。来院時は意識清明、体温３６．９度、血圧８０／５１mmHg、HR１２２回／分。左下腹部に限局した圧痛、反跳痛があったが筋性防御は認めなかった。	圧痛, 血便, 血圧, 筋性防御, 早期食道癌	血圧異常, 圧痛, 血便, 筋性防御, 早期食道癌	擬似誤り文

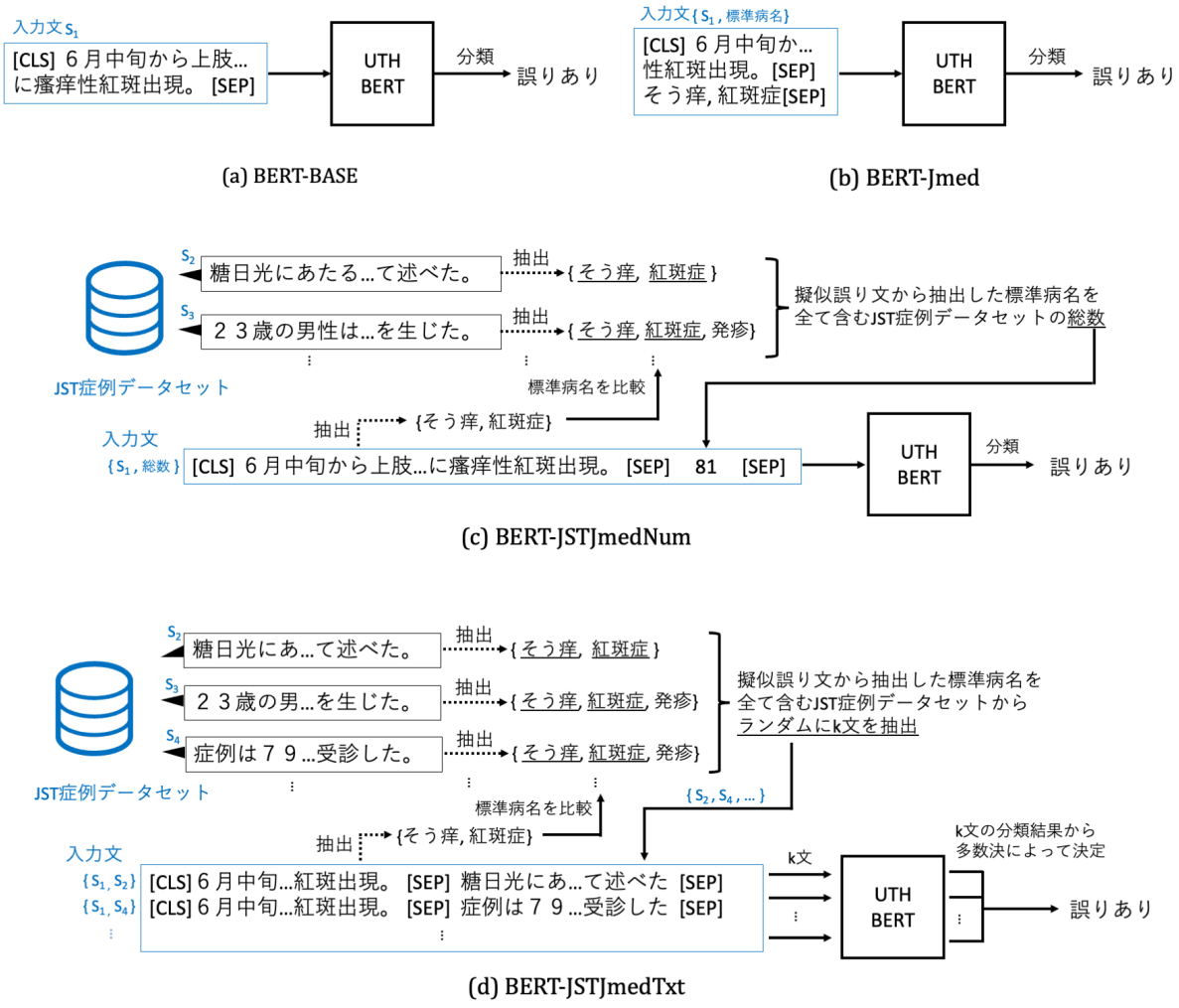
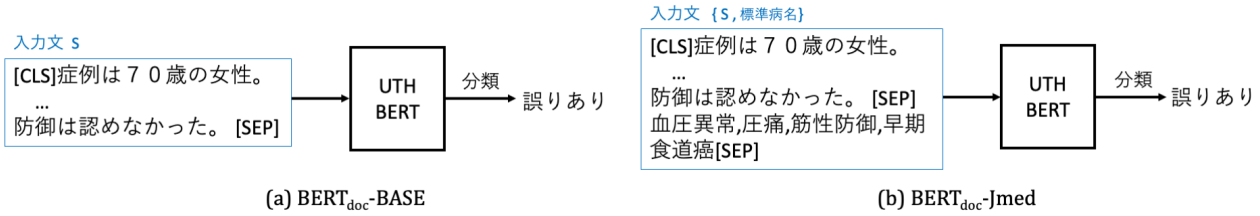


図 3: 提案する文レベルの誤り検出モデル

かの 2 値分類を行なった．学習時のハイパーパラメータにおいて，学習時のバッチサイズには 32 を，学習率には 1.0×10^{-5}，学習のエポック最大数に 20 とし，Early stopping は 5 とした．最大入力トークン数は 512 とし，max_seq_length を超えたトークンは切り捨てた．そのほかのハイパーパラメータには初期値を使用した．

5.2 結　　果

正しく擬似誤り文と判断できた場合を TP，擬似誤り文と判断したが実際は擬似誤り文でなかった場合を FP，正しく



(a) BERT_{doc}-BASE　　(b) BERT_{doc}-Jmed

図 4: 提案する文書レベルの誤り検出モデル

表 5: 提案手法による入力誤り文検出の結果

	Precision	Recall	F 値
BERT-BASE	0.62	**0.78**	0.70
BERT-Jmed	**0.69**	**0.78**	**0.74**
BERT-JSTJmedNum	0.52	0.41	0.46
BERT-JSTJmedTxt	0.62	0.70	0.65
BERT_{doc}-BASE	0.53	0.64	0.58
BERT_{doc}-Jmed	0.49	0.01	0.15

擬似誤り文でないと判断できた場合を TN，擬似誤り文でないと判断したが実際は擬似誤り文であった場合を FN として，Precision，Recall，F1-score により評価した．各手法における擬似誤り文検出の結果を表 5 に示す．Precision および F1-score は **BERT-Jmed** が最も大きく，Recall は **BERT-BASE** と **BERT-Jmed** が最大となった．文に対して誤り検出を行うモデルのうち，JST 症例データセットを外部知識として利用した **BERT-Jmed** および **BERT-JSTJmedTxt** では，いずれのスコアも低い値となった．文書レベルの誤り検出を行うモデル，**BERT_{doc}-BASE** および **BERT_{doc}-Jmed** では，どちらのモデルでもスコアが低かった．

5.3 文レベルでの誤り分析

文レベルの誤り検出モデルにおいて，4 つ全てのモデルが同じ予測ラベルを出力した文例を表 6 に示す．表 6 の（1)〜(4) に示すような，全てのモデルにおいて正解と予測が一致しなかった文のうち，半数は，検査値や薬品値，日付等の時間表現などの数詞を含んでた．表 6 の（1)〜(4) に示すような，全ての誤り文検出モデルにて正解と予測が一致した文でも検査値等の数詞を含む文は観測できたが，正解と予測が一致しなかった文の数ほどではなかった．

5.4 文書レベルでの誤り分析

文書レベルの誤り検出モデルにおいて，4 つ全てのモデルが同じ予測ラベルを出力した文例を表 7 に示す．**BERT_{doc}-BASE** ならびに **BERT_{doc}-Jmed** の両モデルにおいて，文レベルの誤り文検出モデルと比べて，分類スコアが低いことから，コンテキスト情報をうまく活用できていない可能性がある．また，**BERT_{doc}-Jmed** ではほとんどの文書が負例と予測され，FN の数が増大した．このことから，文書レベルの誤り文検出モデルでは標準病名を入力として与えることはノイジーになることが分かる．

6 考　　察

6.1 外部知識利用法の改善

外部知識を用いた手法 **BERT-JSTJmedNum** および **BERT-JSTJmedTxt** は，予想に反してスコアが低くなる結果となった．**BERT-JSTJmedNum** のスコアが低くなった理由として，UTH-BERT の入力文として与えた JST 症例データセットの総数が，別の数字として解釈された可能性がある．UTH-BERT の事前学習時に出現する数詞は検査値や薬品値，日付等の時間表現など多岐にわたるため，入力文として与えた JST 症例データセットの総数が検査値や日付の一部として解釈され，誤り検出のノイズとなった可能性がある．ゆえに擬似誤りを含むかどうかの判断する材料に JST 症例データセットの総数を用いることは，ふさわしくないといえる．

次に，**BERT-JSTJmedTxt** のスコアが低くなった理由として，JST 症例データセットからサンプリングした文が UTH-BERT の入力文としてふさわしくなかったと推察される．外部知識を用いたモデルの一つとして挙げられる質問応答モデル ORQA [17] では，外部知識を入力文として選定する際に，質問文とのコサイン類似度を計算してサンプリングを行なっている．本研究では外部知識を入力文とするときにランダムに選定したため，入力文が誤り文であるかどうかを判断できる外部知識を適切に選定できていない可能性がある．以上のことから，外部知識を用いた手法には改善の余地があると思われる．

6.2 今後の展望

本研究において擬似誤り文を生成する際，病名の出現頻度を考慮してランダムに置換した．これは致命的な医療事故につながる恐れのある病名の入力誤りを防止することが目的であった．一方で医療現場では，ミスタイプや誤変換，電子カルテ上での患者の取り違えなどを中心とした多くの入力誤りが観測されている．したがって，致命的な医療事故につながる入力誤り文の検出だけでなく，医療テキストに特化した誤り訂正モデル（文章校正）を構築することも今後の医療現場では必要であり，ひいては医療言語処理において必要不可欠な事項であると考えられる．

加えて，本研究では病名に対する入力誤り文を検出すること

表 6: 文レベルの誤り検出モデルの全てにおいて同一の予測が得られた文例

ID	擬似誤り文	正解ラベル	予測ラベル	病名 入れ替え前	病名 入れ替え後
(1)	もともとの精神発達遅滞とは異なる皮疹であり，CMZによる薬疹が疑われ，９月１７口より抗菌薬を中止された。	誤りあり	誤りなし	類天庖瘡	精神発達遅滞
(2)	病理組織所見では浸潤性乳管癌（充実腺管癌)、ｐＴｌｂ、ｎｕｃｌｅａｒｇｒａｄｅ１、脈管侵襲陰性で、センチネルリンパ節にＣＮはなく、乳腺切除断端は陰性であった。	誤りあり	誤りなし	転移	CN
(3)	Ａｘ手術時間（平均±ＳＤ）は乳癌１０９±２５分、ＣＮ１４４±２３分（ｐ＝０．０１)、Ａｘ出血量は乳癌４８±３７ｍ１、ＣＮ７９±３８ｍ１（ｐ＝０．０７）と、ＣＮ群での治療的郭清の影響が考えられた。	誤りなし	誤りあり	-	-
(4)	検査成績：ＧＯＴ３９２，ＧＰＴｌ２３，ＬＤＨ１８０３，ＣＰＫ９４００，アルドラーゼ８．６で，腹部ＣＴにより多数の腹腔内リンパ節腫大と心窩部に腫瘤を認めた。	誤りなし	誤りあり	-	-
(5)	４月１８日朝ＣＴ／ＭＲＩにて未分化癌を認め，当院脳神経外科転棟転科．同日リハビリテーション開始となった．	誤りあり	誤りあり	右中大脳動脈脳梗塞所見	未分化癌
(6)	【背景】本邦で開発されたバルーン閉塞下逆行性経静脈的塞栓術（ＢＲＴＯ）は嘔吐のみならず肝性脳症に対する低侵襲かつ効果的なＩＶＲ治療である．	誤りあり	誤りあり	胃静脈瘤	嘔吐
(7)	胆嚢炎に対する検査では、胆石と軽度胆嚢の壁肥厚を認めるが、他特記すべきことはなかった。	誤りなし	誤りなし	-	-
(8)	当初はラクツロースとＢＣＡＡ製剤のみでコントロール可能であったが，徐々にコントロール不良となり，リファキシミンおよびレボカルチンを追加し，更に週３回ＢＣＡＡ製剤の点滴を受けるも頻回に肝性昏睡に陥っていた．	誤りなし	誤りなし	-	-

ために，アプローチの一つとして症例報告論文から生成した擬似誤り文を検出するモデルの提案をしてきた．しかし実際の医療現場においては，どの文に誤りがあるかではなく，文章のどの部分に入力誤りがあるかについて提示できた方が好ましいと推測できる．本研究では医学文書における入力誤り文の検出を行なったが，今後の展望として，入力誤り文の検出結果に基づく誤り訂正モデルの構築が挙げられる．

また入力誤りであるかどうかを判断する際に，単文ではコンテキスト情報が少なく，判断できない可能性がある．このような入力誤り文については，より広い文脈を考慮したモデルの構築が必要となってくる．つまり，モデルへの入力において，連続する文中に入力誤りが存在するかを検出するモデルを構築する必要がある．このモデルを構築する場合は文書内に出現する全ての単語に対して，それぞれ入力誤りかどうかを判断しなければならないため，非常に難しい問題だと推察される．本研究でも文書単位を入力として構築したモデルの精度はあまり良くなかった．この問題を解決する方法の一つとして，まずは文レベルで誤り文の検出を行い，文レベルで学習しきれなかった例に対しては，文書レベルで誤り検出することで，効率よく誤り文の検出が行える可能性がある．

7 おわりに

本研究では，病名に対する入力誤り文を検出することを目的とし，そのアプローチとして症例報告論文の文中から生成した擬似誤り文を検出するモデルの提案を行った．入力誤り文を検出するモデルの構築にあたり，症例報告論文コーパス (MedTxt-CR-JA) を文単位に分割し，文中の病名タグをすることで擬似誤り文のデータセットを生成した．生成した擬似誤り文のデータセットに対して，医学的に文意の通らない文となっているかを医療従事者によって評価した．生成した擬似誤り文に対して，UTH-BERT と J-STAGE に投稿された症例報告論文のアブストラクトを外部知識として用いて，誤り文検出を行なった．本研究では文レベルの誤り検出と文書レベルの誤り検出をおこない，文レベルの誤り検出では擬似誤り文と擬似誤り文に出現する病名の標準病名を入力文としたモデルが最もスコアが高く，precision が 0.69，Recall が 0.78，F1-score が 0.74 であった．文書レベルの誤り検出では，precision が 0.53，Recall が 0.64，F1-score が 0.58 であった．今後の課題としては外部知識を入力として用いる際に，コサイン類似度を用いて選定するなどアルゴリズムの改善などが挙げられる．本研究では文レベルの誤り検出と文書レベルの誤り検出を行なってきたが，コンテキスト情報から誤り検出する方法について引き続き調査を進めていきたい．

謝　辞

本研究は，JST CREST 課題番号：JPMJCR22N1 および JSPS 科研費 JP19H01118 の支援を受けたものである．

表 7: 文書レベルの誤り検出モデルの両方において同一の予測が得られた文例

ID	擬似誤り文	正解ラベル	予測ラベル	病名 入れ替え前	病名 入れ替え後
(1)	症例：２４才，男性．蕁麻疹の既往あり．母親がアトピー性皮膚炎．１９才より膀胱癌と肺癌の合併が出現．２０才より肺結核で２年間治療．喘息発作が改善しないため平成元年６月，当科を受診．ＷＢＣ９１００，好酸球２８．３%，ヒスタミン０．５２ｎｇ／ｍｌ，ＬＴＣ４・ＬＴＤ４ １０ｐｇ／ｍｌ以下，総ＩｇＥ ５２７５，ＩｇＥアスペルギルス（ＲＡＳＴ）３，アスペルギルス皮内反応即時型陽性（４７ｘ４５mm），粘液栓子喀出あり，ＦＥＶ１ １．７３，ＦＥＶ１% ５４．６，胸部X線上に移動性の手指状陰影が出没.	誤りあり	誤りなし	咳漱，喘鳴	膀胱癌と肺癌の合併
(2)	【はじめに】一般的に睡眠時無呼吸症候群（ＳＡＳ），チェーンストークス呼吸（ＣＳＲ），運動時周期性呼吸変動（ＥＯＶ）などの呼吸障害はいずれも予後不良因子として知られている。今回，心臓血管外科術後，種々の呼吸障害とそれに伴う意識レベルの低下により理学療法進行に難渋した患者への呼気ガス評価を実施することで，呼吸障害に対する早期の治療開始が可能となりＡＤＬ拡大，退院へと繋がった一例を経験したためここに報告する。【症例提示・経過】６０歳代男性。	誤りなし	誤りあり	-	-
(3)	【はじめに】科学的根拠に基づく褥瘡局所治療ガイドラインでは、光線療法（近赤外線あるいは紫外線）が嚥下障害の縮小に対して推奨度Ｃ１とされている。我々は第４２回日本理学療法学術大会において、低栄養状態の褥瘡患者に対する直線偏光近赤外線照射治療（以下、ＳＬ）による褥瘡の改善について報告した（椎名、２００７）。今回、新たに６症例を加え創の縮小に着目して検討したので報告する。【対象】。	誤りあり	誤りあり	創	嚥下障害
(4)	【背景・目的】キャンディン系抗真菌薬のミカファンギン（ＭＣＦＧ）はＣａｎｄｉｄａ属に対して殺菌的に作用し、幅広い抗真菌スペクトルを有する。また、投与後の組織への移行性が高く、既存の抗真菌剤と比較して副作用も少ないことから使用頻度が増加している。一方で、尿中排泄率か１%以下と低いことが報告されており、尿路感染症に対する使用報告は少ない。今回、Ｃａｎｄｉｄａ尿路感染症に対してＭＣＦＧを投与した２症例の臨床経過と血中ＭＣＦＧ濃度を報告する。【対象・方法】〔症例１〕９０歳代、男性。	誤りなし	誤りなし	-	-

文　献

[1] 厚生労働省, “医療分野の情報化の推進について,” https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryou/johoka/index.html,Accessed: 2022-12-16

[2] 医療事故防止事業部, “医療安全情報,” 公益財団法人 日本医療機能評価機構, 2019, No.154

[3] 相良かおる, “ComeJisyo の紹介と医療情報に含まれる誤字調査” 情報知識学会 第 22 回 (2014 年度) 年次大会, Vol.24, pp.204–209

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” In Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp.4171–4186, 2019

[5] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, Vol.36, pp.1234–1240,2019

[6] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, Kazuhiko Ohe, “A clinical specific BERT developed using a huge Japanese clinical text corpus,” PLoS ONE 16(11): e0259763., 2021

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Improving Neural Machine Translation Models with Monolingual Data,” In Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, Vol.1, pp. 86–89, 2016

[8] Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi, “Grammatical Error Correction Using Pseudo Learner Corpus Considering Learner’s Error Tendency,” In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp.27–32, 2020

[9] Wei Zhao, LiangWang, Kewei Shen, Ruoyu Jia, and Jingming Liu, “Improving grammatical error correction via pretraining a copy-augmented architecture with unlabeled data” In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp.156–165, 2019

[10] Hládek, Daniel and Staš, Ján and Pleva, Matúš, “Survey of Automatic Spelling Correction” Electronics, Vol.9(10), 2020

[11] Edouard Grave, Armand Joulin, and Nicolas Usunier, “Improving Neural Language Models with a Continuous Cache,” In Proc. of the 5th International Conference on Learning Representations, 2017

[12] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis, “Generalization through Memorization: Nearest Neighbor Language Models,” In Proc. of the International Conference on Learning Representations, 2021

[13] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke

Zettlemoyer, and Mike Lewis, "Nearest Neighbor Machine Translation," In Proc. of the International Conference on Learning Representations, 2021

[14] Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen, "Adaptive Nearest Neighbor Machine Translation," In Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp.368–374, 2021

[15] Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li, "Fast Nearest Neighbor Machine Translation," Findings of the Association for Computational Linguistics: ACL 2022, pp.555–565, 2022

[16] Yang Zhixian, Sun Renliang, and Wan Xiaojun, "Nearest Neighbor Knowledge Distillation for Neural Machine Translation," In Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.5546–5556, 2022

[17] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova, "Latent Retrieval for Weakly Supervised Open Domain Question Answering" In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pp.6086–6096, 2019

[18] Guu Kelvin, Lee Kenton, Tung Zora, Pasupat Panupong, and Chang Ming-Wei, "REALM: Retrieval-Augmented Language Model Pre-Training," In Proc. of the 37th International Conference on Machine Learning, pp.3929–3938, 2020

[19] Gautier Izacard, and Edouard Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," In Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp.874–880, 2021