MedNormJ：日本語医療テキストにおける病名正規化のための文脈付きデータセットの構築

田代勇希 [1] 清水 聖司 [1] 西山 智弘 [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
tashiro.yuki.ty3@naist.ac.jp
{shimizu.seiji.so8,nishiyama.tomohiro.ns5,wakamiya,aramaki}@is.naist.jp

概要

医療テキスト中における病名正規化は，臨床データの二次利用において重要な基盤技術である．しかし，医療ドメインの高度な専門性とアノテーションの複雑さ等が障壁となり，日本語における標準的な評価データセットは未だ整備されていない．そこで本研究では，病名正規化を目的とした初の日本語データセット **MedNormJ** を構築し，公開する．本データセットは，症例報告と放射線読影レポートの計 96 件から抽出された 397 件の医療用語とその正規形のペアで構成されている．また，構築したデータセットを用いて既存の正規化手法の比較実験を行い，日本語医療テキストにおける医療用語の正規化の現状と課題を明らかにする．

1 はじめに

近年，電子カルテに蓄積されたリアルワールドデータの利活用により，新薬開発や臨床研究を加速させる社会的要請が世界的に高まっている [1]．特に，医療テキストは，患者の状態を詳細に記録した高密度な情報源である一方，その多くは非構造化データゆえに同義語，略語，誤記といった多様な表記揺れを含んでおり，大規模なデータ解析を妨げる要因となっている．この課題を解決する基盤技術が**医療用語正規化**である．これは，図 1 に示す医療テキスト中の「高分化腺癌」や「高度進行 S 状結腸癌」といった病名（以下，**出現形**）を，標準的な用語セット（以下，**オントロジー**）の中の適切な概念（以下，**正規形**）へマッピングするタスクとして定式化される．例えば，病名マスター [2] をオントロジーとして用いる場合，出現形「高度進行 S 状結腸癌」に対して，収載された約 2 万語の標準病名から「S 状結腸癌」を選択する処理がこれに該当する．また，SNOMED CT [3] や ICD-10 [4] のように，各用語が固有のコードを有するオントロジーへの紐付けは，特にコーディングとも称されるが，本稿では，出現形を標準的な概念へと紐付ける処理を総称して正規化と呼ぶ．

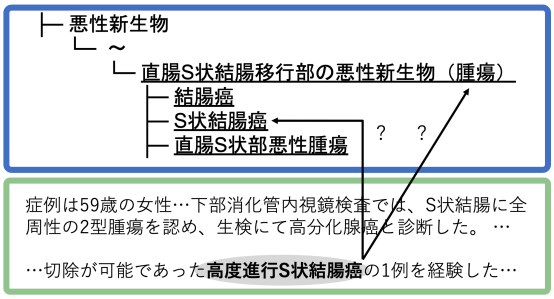

図 1 病名正規化タスクの概観．医療テキスト中の出現形（例：高度進行 S 状結腸癌）を，標準病名マスターに含まれる正規形候補（オントロジー）へ対応付ける．同一出現形が複数の候補に紐づき得るため，文脈に基づく曖昧性解消が必要で，本例では S 状結腸癌が正規形である．

このように正規化はタスクとしては明確であるものの，医療ドメインにおける正規化データセットの構築は難易度が高い．これには主に二つの要因がある．一つ目は，**アノテーションコストの高さ**である．難解な病名を正確に扱うには，医学的知見に精通した専門家の協力が不可欠であり，一般ドメインと比較して，アノテーターの確保や作業コストの面での制約が大きい．二つ目は，**正規化粒度の決定における不確実性**である．医学概念はオントロジー上で深い階層構造を有しているため（図 1），出現形をどの階層の概念に紐付けるべきかという判断が困難である [5]．例えば，「高度進行 S 状結腸癌」という出現形に対し，部位情報を保持した「S 状結腸癌」を選択すべきか，あるいはより広義の「結腸癌」や「腫瘍」に集約すべきかという正規化粒度の決定は，後続の解析目的や文脈に依存する．つまり，病名正規化は，専門家を確保しにくいというリソースの課

表 1　病名正規化におけるガイドラインと対応するアノテーションの例

ガイドラインの説明（抜粋）		例	
		文脈（太字：出現形）	正規形（✓ 採用例, ✗ 不採用例）
原則	アノテーション対象の病名について，文脈上，特定の部位（臓器など），形質などを判断できる場合は，文脈情報を用いて**可能な限り最も詳細な粒度の病名**に正規化する．また，文中で同一概念が異なる粒度で繰り返し記述されている場合は，文脈から判定できる最も詳細な病名を用いる．	胃底部に…**隆起性腫瘍** を認めた．	✓ 胃腫瘍 ✗ 腫瘍
例外 1	ある病態の複数の所見が記述されており，かつ，それらをまとめる上位概念に相当する適切な病名が存在する場合には，その病名を用いる．	胸部 X 線写真で…**小結節**…**すりガラス影**…**腫瘤影** が見られた．	✓ 胸部異常陰影 ✗ 結節 小結節 すりガラス影
例外 2	「癌」「肉腫」「腺癌」など明らかに悪性を示す場合は「癌」または「悪性腫瘍」に正規化する．一方，「腫瘍」「腫瘤」など良悪性が不明な場合は「腫瘍」とする．	画像上，**腫瘤性病変** を認め，悪性の可能性が示唆された．	✓ 癌 ✗ 腫瘍

題と，専門家であっても正解の基準が揺らぎうるという定義の課題を併せ持っている．これらの課題が障壁となり，日本語医療ドメインにおいては，標準的な評価データセットが未だ確立されていないのが現状である．

そこで，本研究では，前述の 2 つの課題を解決し，日本語の病名正規化評価データセット **MedNormJ** を構築する．構築にあたっては，まず専門家による予備アノテーションを実施し，アノテーター間で生じた解釈の不一致を分析した．分析結果に基づき，正規化の粒度を統一するためのルールを定めたアノテーションガイドラインを作成した．具体的には，文脈から部位・性状・病期などを特定できる場合には**可能な限り下位概念へ詳細化する**ことを原則として，判断基準を定めた（表 1）．そのガイドラインをもとにアノテーションを行い，正規化の既存手法で評価を行なった．構築したデータセットは研究用として一般公開する[1)]．

2　関連研究

英語圏においては，医療用語正規化タスクの研究開発を支援する標準的なデータセットがいくつか公開されている．例えば，NCBI Disease Corpus [6] は，病名正規化研究におけるデファクトスタンダードとして広く利用されている．他にも，化学物質および疾患に関するデータセットである BC5CDR [7] や多様な UMLS 概念が紐づけられたデータセットである MedMentions[8] などが構築されてきた．これらのリソースの存在は，DNorm [9] や SapBERT [10] などの正規化手法の開発に貢献してきた．

対照的に，日本語医療ドメインでは，ベンチマークとして広く共有可能な病名正規化データセットが存在していない．いくつかの研究で独自のデータセットが構築されているものの [11, 12]，正規化に関する情報を含まない，あるいは含まれていたとしても ICD-10 に基づく付与にとどまっていた．しかし，ICD-10 は本来，統計や行政目的のための疾病分類であり，医療テキストに出現する多様な病名表現を臨床的に十分な粒度で捉えられないことがある．

3　データセット構築

日本語病名正規化データセット **MedNormJ** の構築プロセスについて述べる．

3.1　材料

MedNormJ のベースとして，既存の固有表現抽出用データセットである MedTxt[2)] を採用した．MedTxt は，症例報告を収録した MedTxt-CR [13] と，放射線読影レポートを収録した MedTxt-RR [14] の 2 つのサブセットで構成され，それぞれ 50 件と 46 件を対象としている．これらの統計情報は，付録・表 6 に示す．症例報告は希少な症例を報告する学術論文であり，一方で読影レポートは放射線診断専門医が CT や MRI などの医用画像を読影し，そこから所見や疾患をまとめた報告書である．本研究では，これらの既存データセットに含まれる病名の出現形に対し，標準病名を正規化先として新たに付与した．

3.2　予備アノテーション

まず，予備アノテーションとしてそれぞれの文章から 50 件をランダムに取得した計 100 件の出現形に対して，2 名の医療従事者による独立したアノテーションを実施した．

その結果に対してアノテーター間一致率（Inter-

1）https://github.com/sociocom/mednormj

2）https://sociocom.naist.jp/medtxt/

Annotator Agreement; IAA）を算出した．表 2 に示すように，全体の一致率は 0.570 であり，43 件においてアノテーター間の不一致が確認された．

その不一致例を分析したところ，主に概念粒度の違いに起因する系統的な不一致が確認された．例えば，「肺野異常陰影」に対し，「胸部異常陰影」や「肺部異常陰影」といった異なる同義語が割り当てられる表記の揺れや，「腫瘤」と「肺腫瘤」のように文脈補完の有無による粒度の差が顕著であった（付録・表 5）．このように，医療用語における正規化というタスクは曖昧性が高く，難しいことが分かる．

3.3 ガイドラインの作成

予備アノテーションでの分析結果をもとに，一貫した正規化品質を担保し，かつ判断基準を体系化するために，アノテーションガイドラインを作成した．主なアノテーションルールは表 1 に示す．詳細は付録に記載する．アノテーション原則としては，文脈から部位・性状・病期などを特定できる場合は，可能な限り下位概念へ詳細化することとした．

しかし，予備アノテーションの不一致には，この「詳細化」原則だけでは一意に定めにくいパターンが存在した．第一に，読影レポートでは複数の所見が並列表現で記述されることが多く，個々の所見を個別に正規化するとアノテーター間で正規化先がばらつきやすい．実際に，予備アノテーションにおける MedTxt-RR の一致率は 0.360 であり，MedTxt-CR の 0.780 と比べて低かった（表 2）．そこで，複数所見が同一病態として解釈でき，それらを包括する妥当な上位概念が存在する場合には，上位概念へ統合して正規化する（例外 1）と定めた．

第二に，「腫瘍／腫瘤」など良悪性が確定しない語と，「癌／腺癌」など悪性を明示する語では，正規化先（癌・悪性腫瘍・腫瘍）が分岐し，原則だけでは選択基準が明確にならない．そのため，悪性を示す語が明示される場合は「癌」または「悪性腫瘍」，良悪性が不明な場合は「腫瘍」とする個別ルール（例外 2）を設けた．

3.4 本アノテーション

作成したガイドラインに基づき，全件で最終的な評価用データセットを構築した．構築されたデータセットの統計情報を表 6 に示す．本データセットは合計 96 件の医療文書を対象としており，正規化された出現形の総数は 515 件であった．また，正規化先のユニークな正規形の数は 259 件であった．

表 2 アノテーター間一致率（IAA）．Pilot は予備アノテーション，Main は本アノテーションを指す．

フェーズ	指標	MedTxt-CR	MedTxt-RR	全体
Pilot	Accuracy	0.780	0.360	0.570
	Cohen's κ	0.776	0.346	0.565
Main	Accuracy	0.761	0.784	**0.768**
	Cohen's κ	0.759	0.722	**0.762**

構築したデータセットに対して再度，IAA を行なった．その結果，MedNormJ 全体での一致率が 0.570 から **0.768** へと改善した．ここで予備アノテーション（n=100）と本アノテーション（n=515）で，件数が異なるため，正確な比較はできないことに注意する必要がある．Landis らの基準 [15] に照らすと，これらの値は **相当な一致（Substantial Agreement）** に該当し，構築したデータセットが一定の信頼性を有しているといえる．

一方で，特定の用語において体系的な不一致も確認された（表 7）．例えば，「筋緊張」を含む表現に対し，アノテーター A は「筋緊張性障害」，アノテーター B は「筋痙縮」を割り当てる傾向があった．また，「無気肺」についても，「下葉無気肺」のように，文脈に基づいて部位などを特定した下位概念まで正規化するか，表層表現どおり「無気肺」とするかについて，アノテーター間で判断が分かれる事例が見られた．これらの結果は，臨床的解釈が分かれやすい表現に対して，より詳細なガイドラインの作成，あるいは複数ラベルの許容といった柔軟な評価体系の必要性を示唆している．

4 実験

MedNormJ に対する正規化タスクの難易度を評価するため，3 つの手法を用いてベースライン実験を行なった．具体的には，(1) Exact-match に基づく手法，(2) Levenshtein 距離を用いた文字列類似度ベースの手法（以下，Levenshtein），(3) TF-IDF ベースのベクトル表現を用いる DNorm-J を比較した．

4.1 ベースライン性能

実験結果を表 3 に示す．Exact-match は Acc (A & B) で 0.476 と比較手法の中で最も低い性能を示した．完全一致というシンプルな手法でも一定の性能を達成したものの，辞書との厳密な一致に頼るだけでは，医療文書における表記揺れや略語の吸収は困難であることがわかる．一方，Levenshtein では

表 3　各モデルによる正規化性能比較（Acc@1）．Acc (A), Acc (B) はそれぞれアノテーター A, B の結果を正解としたときの精度，Acc (A & B) は両アノテーターの結果が一致したものを正解としたときの精度である．

Model	Acc (A) (n=515)	Acc (B) (n=515)	Acc (A & B) (n=397)
Exact-match	0.398	0.410	0.476
Levenshtein	0.484	0.491	0.564
DNorm-J	0.451	0.466	0.547

表 4　Levenshtein 距離に基づく正規化手法における代表的なエラー例．分類 1・2 は本文中の分類にリンクする．

分類	エラー例	出現形	正規形
(1)	腫瘤	腫瘤	肺腫瘤
(1)	中心結節	結節影	胸部異常陰影
(2)	癌	低分化型腺癌	上行結腸癌

0.564 を達成し，単純な編集操作で解決可能な表記揺れに対してある程度有効であった．

4.2　エラー分析

実験で最も高い精度を示した Levenshtein に対して，エラー分析を行った．代表的なエラー例を 表 4 に示す．これらのエラーは，大きく二つの傾向に分類できる．

第一に，**文脈情報や標準病名の概念階層を考慮できないエラー**である（分類 (1)）．例えば，「腫瘤」という出現形に対して，文脈中に肺に関する記述が存在する場合であっても，出現形そのものが正規形候補にあるために選択され，適切な正規形の「肺腫瘤」に正規化が行われないケースが確認された．同様に，「結節影」という出現形に対しても，本研究で正解と定義する上位概念「胸部異常陰影」ではなく，結節に関連する用語が選択される事例が見られた．これらは，「結節」のように文字列距離に基づく手法が，文脈に基づく部位情報の補完や，標準病名が持つ概念階層を適切に扱えないことに起因する．

第二に，**抽象度の高い概念へ過度に集約されるエラー**である（分類 (2)）．例えば，「低分化型腺癌」という具体的な出現形に対して，正規形である「上行結腸癌」ではなく，より抽象的な概念である「癌」が選択されるケースが確認された．この傾向は，距離ベース手法の特性に起因すると考えられる．

以上より，Levenshtein は，単純かつ高い性能を示す一方で，文脈理解や概念階層の扱いを必要とする正規化に対しては，本質的な限界を有することが明らかとなった．

5　議論

専門家によるアノテーションを分析し，日本語医療テキストにおける病名正規化では，文脈や概念粒度の解釈に起因する揺れが不可避に生じることが確認された．付録・表 7 に示すように，特定の表現に対して体系的な不一致があり，病名正規化における一意な正解の定義自体が困難であることが明らかになった．したがって，今後の評価設計で，より詳細なガイドラインの拡充や，複数ラベル許容といった柔軟な枠組みが有用だと考えられる．

ベースライン実験では，Exact-match が 0.476 と低く，文字列の完全一致に依存する手法では表記揺れ・略語・誤記を吸収できないことが確認された．また，Levenshtein が 0.564 で最良であったことは，編集距離で解決可能な表記揺れが一定割合存在することを示す一方，依然として 0.6 未満にとどまる点から，タスクの難しさが文脈理解と概念階層の扱いに由来することが示唆される．実際にエラー分析では，(i) 文脈中の臓器・部位情報を補完できず上位概念に留まる誤り，(ii) 短く抽象的な語へ過度に一般化される誤り，(iii) 標準病名の階層構造を考慮できない誤りが確認された．これらは，単なる文字列類似度では解決できない文脈による判断が主要な難しい点であることを示し，これを解消すると性能が向上する可能性が高い．

本データセットは症例報告と読影レポートの 2 種類に限られ，網羅性には限界がある．文書種ごとに記載様式や含まれる文脈情報が大きく異なるため，今後は診療録や退院サマリなどの医療文書の種類を増やすことを通じて，より実用的な正規化研究へ接続することが課題である．

6　おわりに

本研究では，日本語医療テキストを対象とした高品質な病名正規化データセット **MedNormJ** を構築し，その詳細なアノテーションプロセス，およびベースライン性能について報告した．特に，正規化粒度と文脈解釈の一貫性を重視したアノテーションガイドラインを作成することで，高いアノテーター間一致率を達成した．本データセットが手法改善などの日本語医療言語処理のさらなる発展に貢献することを期待するとともに，今後の病名正規化データセットの拡充を目指す．

謝辞

本研究の一部は，JST CREST「リアルワールドテキスト処理の深化によるデータ駆動型探」（課題番号：JPMJCR22N1）および「戦略的イノベーション創造プログラム（SIP）」「統合型ヘルスケアシステムの構築」JPJ012425, JSPS 研究スタート支援 JP25K24412 の補助を受けて行なった．データセット構築にあたり，アノテーション作業に従事してくださった有森美紀子氏，藤牧貴子氏に感謝する．

参考文献

[1] Naoto Usuyama, Cliff Wong, Sheng Zhang, Tristan Naumann, and Hoifung Poon. Biomedical natural language processing in the era of large language models. **Annual Review of Biomedical Data Science**, Vol. 8, , 2025.

[2] 標準病名マスター作業班. 標準病名マスター作業班（病名検索サイト）. http://www.byomei.org/. Accessed: 2026-01-06.

[3] Eunsuk Chang and Javed Mostafa. The use of snomed ct, 2013-2020: a literature review. **Journal of the American Medical Informatics Association**, Vol. 28, No. 9, pp. 2017–2026, 2021.

[4] World Health Organization. **International Statistical Classification of Diseases and related health problems: Alphabetical index**, Vol. 3. World Health Organization, 2004.

[5] Dao Sy Duy Minh, Nguyen Lam Phu Quy, Pham Phu Hoa, Tran Chi Nguyen, Huynh Trung Kiet, and Truong Bao Tran. Dragon: Dual-encoder retrieval with guided ontology reasoning for medical normalization. In **Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association**, pp. 230–239, 2025.

[6] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. **Journal of biomedical informatics**, Vol. 47, pp. 1–10, 2014.

[7] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. **Database**, Vol. 2016, , 2016.

[8] Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with umls concepts. **arXiv preprint arXiv:1902.09476**, 2019.

[9] Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. Dnorm: disease name normalization with pairwise learning to rank. **Bioinformatics**, Vol. 29, No. 22, pp. 2909–2917, 2013.

[10] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4228–4238, June 2021.

[11] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. Overview of the ntcir-10 mednlp task. In **NTCIR**, 2013.

[12] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. Overview of the ntcir-11 mednlp-2 task. In **NTCIR**, 2014.

[13] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. Real-mednlp: Overview of real document-based medical natural language processing task. In **Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies**, pp. 285–296, 2022.

[14] Yuta Nakamura, Shohei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Clinical Comparable Corpus Describing the Same Subjects with Different Expressions. **Stud Health Technol Inform**, Vol. 290, pp. 253–257, Jun 2022.

[15] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. **biometrics**, pp. 159–174, 1977.

A　付録

A.1　予備アノテーションでの不一致の分析

付録・表 5 に，予備アノテーションで確認されたアノテーター間の不一致例を示す．これらの不一致は，主に文脈補完の有無や概念粒度の解釈差に起因し，後続のガイドライン作成の動機となった．

A.2　MedNormJ の統計情報

付録・表 6 に，**MedNormJ**（A&B 一致サブセット）の統計情報を示す．本データセットは，症例報告・放射線読影レポートの計 96 件を対象とし，397 件の出現形と 206 件のユニークな正規形から構成される．

A.3　IAA で体系的な不一致が発生した具体例

付録・表 7 に，IAA 分析で確認された特定の表現に対して繰り返し生じた体系的な不一致例を示す．これは，同一の出現形に対し，アノテーター間で異なる正規形が一貫して選択された事例を抜粋したものである．

A.4　アノテーションガイドライン

以下に，本研究で作成したアノテーションガイドラインを示す．

原則：文脈に基づき可能な限り詳細化する

アノテーション対象の病名について，文脈上，特定の部位（臓器など），形質などを判断できる場合は，文脈情報を用いて**可能な限り最も詳細な粒度の病名**に正規化する．また，文書中で同一概念が異なる粒度で繰り返し記述されている場合には，文脈から判定できる最も詳細な病名を正規化先とする．

例：文中に「胃底部に隆起性腫瘍を認めた」と記述されている場合，「腫瘍」とするのではなく，「胃腫瘍」へ正規化する．

例：同一文書内に「腫瘍」と「低分化型腺癌」が併記されている場合は，より詳細な「低分化型腺癌」を正規化先とする．

表 5　予備アノテーションでのアノテーター間の一貫した不一致例

アノテーター A	アノテーター B	事例数
腫瘤	肺腫瘤	9
肺野異常陰影	胸部異常陰影	8
肺野異常陰影	肺部異常陰影	7

表 6　MedNormJ (A&B 一致サブセット) の統計情報（CR/RR は MedTxt-CR/-RR 別の値）

Statistic	CR/RR	MedNormJ
レポートの件数	50 / 46	96
出現形の数	270 / 127	397
ユニークな出現形数	234 / 62	295
ユニークな正規形数	187 / 24	206
出現形/正規形比	1.4 / 5.3	1.9

表 7　体系的な不一致が発生した具体例

出現形	正規形 A	正規形 B	件数
筋緊張（の亢進）	筋緊張性障害	筋痙縮	4
無気肺	下葉無気肺	無気肺	4

例外 1：複数所見を上位概念に一般化する

ある病態について，複数の所見が同時に記述されており，それらを包括する医学的に妥当な上位概念が存在する場合には，下位概念を個別に選択せず，上位概念に正規化する．

例：胸部 X 線写真や CT 検査において，「結節」「小結節」「すりガラス影」「腫瘤影」など複数の所見が併記されている場合，個々の所見を正規化するのではなく，「胸部異常陰影」へ正規化する．

例外 2：悪性・良悪性に関する個別ルール

悪性度に関して複数の選択肢が存在する場合には，以下の個別ルールを適用する．「癌」「肉腫」「腺癌」など，明らかに悪性を示す語が用いられている場合は，「癌」または「悪性腫瘍」に正規化する．一方で，「腫瘍」「腫瘤」など，良悪性が明示されていない場合には，「腫瘍」を正規化先とする．

例：画像所見として「腫瘤性病変」が記載され，確定的な悪性表現が用いられていない場合には，「腫瘍」と正規化する．