22世紀医療センター
22nd Century Medical and Research Center

講座名　医療AI開発学講座
英文講座名 Department of Artificial Intelligence in Healthcare

演題名：希少・難治性疾患を対象とした症例報告テキストコーパス

演者名：河添 悦昌、篠原 恵美子、柴田 大作、嶋本 公徳、平林 真衣

はじめに： 希少・難治性疾患の診断において，カルテに記録される症状や所見など表現型の情報は重要となる．これまで、テキストから表現型を抽出するNLP技術開発のため，指定難病の症例報告テキストを材料とし50種の固有表現タグと35種の関係により，テキストの表現型を網羅的にアノテートする基準を開発してきた[1]（図１）．

目的： 本発表ではより具体的なコーパスの構築方法とコーパスの内訳について述べる．

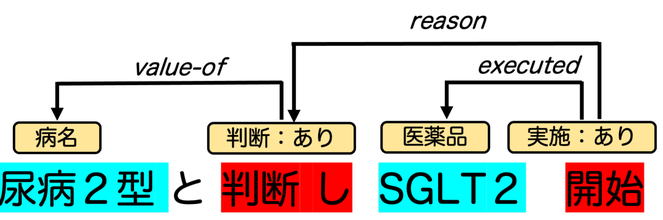
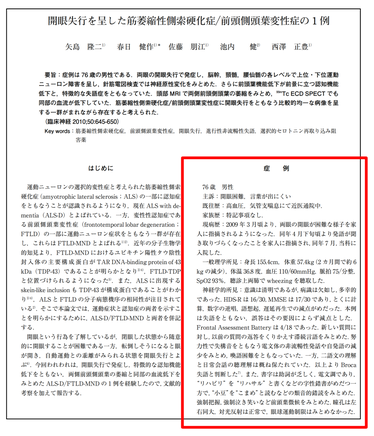


図1 固有表現タグと関係によりテキストを構造化する

方法：

症例報告の収集： 指定難病333疾患を対象とし、タイトルに疾患名を含む症例報告をJ-STAGEで検索して得た362症例報告（151疾患）を対象．症例提示部分をテキスト化して利用．


開眼失行を呈した筋萎縮性側索硬化症/前頭側頭葉変性症の１例

はじめに

症　例

図2 症例提示部分

アノテーション基準： 以下の基本方針のもと、アノテーションの実施と基準の修正を繰り返すことで構築．

1. テキストに含まれる情報を網羅的にアノテート．
2. 医学・医療的な意味を損なわない範囲でタグの範囲を詳細化．
3. 専門的な知識ではなく，文字を頼りに判断できる情報をアノテート（例：原因-結果関係は「呼吸困難による睡眠障害」の”による”のように，文字を頼りに判断可能な場合にアノテート）

アノテーション実施環境： アノテーション支援ツールBratをクラウド上に設置し，アノテータが在宅で作業できる環境を構築。固有表現タグと関係のインスタンスに与えられる一意のURLをコミュニケーションツールで共有することで，内容に疑義が生じた箇所を迅速に参照・修正できるようにした．

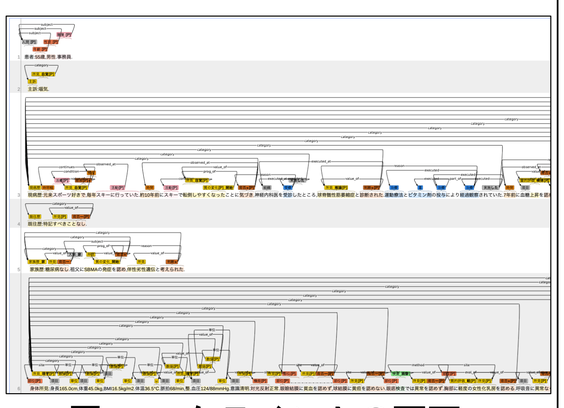


図3 Bratクライントの画面

AIを併用したアノテーション修正： 膨大な数の固有表現タグと関係をアノテートする必要から，機械学習モデルによるタグと関係の推論結果をアノテータに提示し，必要に応じて修正することで，タグ・関係のつけ忘れや，タグ選択範囲の微細な違いを統一した．

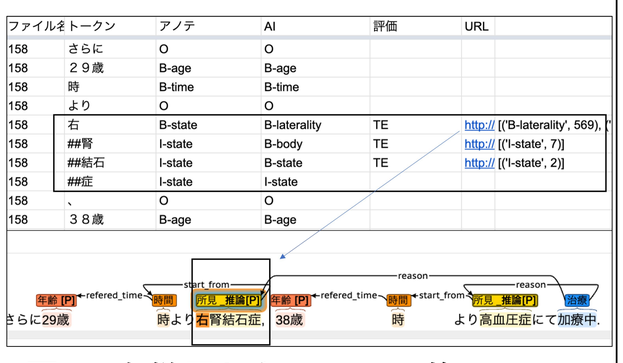


図4 AIを併用したアノテの修正．この例ではAIの推論結果が正しい．

結果： 収集した362症例報告のうち，179症例（102疾患）について掲載元よりテキストの再配布許可を得た．以下，この179症例報告を本コーパスとして述べる．

参考文献：

1. 篠原 恵美子ら．医療テキストに対する網羅的な所見アノテーションのためのアノテーション基準の構築．第25回日本医療情報学春季学術大会．
2. Yada, S., et.al. Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases. Proceedings of The 12th Language Resources and Evaluation Conference, 4567-4574.
3. MedTxt-CR:症例報告 (Case Reports) コーパス．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．https://sociocom.naist.jp/medtxt/cr/ (cited 2021-Aug-30)

１０２疾患の一覧と症例報告のタイトル例： コーパスに含まれる１０２疾患の一覧（表１）と症例報告のタイトル例（表２）を示す．タイトルは主となる指定難病と合併した疾患とで表されることが多いことから，本文テキスト中には指定難病だけでなく，一般的な疾患の症状や治療の情報が幅広く含まれると推測された。

表1 コーパスに含まれる102疾患

	指定難病名	件数							
あ	IgA腎症	2	さ	左心低形成症候群	1	は	肺静脈閉塞症／肺毛細血管腫症	1
	IgG4関連疾患	3		再生不良性貧血	1		肺動脈性肺高血圧症	2
	アッシャー症候群	1		再発性多発軟骨炎	1		肺胞低換気症候群	1
	アラジール症候群	1		三頭酵素欠損症	1		パーキンソン病	4
	アレキサンダー病	2		シェーグレン症候群	2		バージャー病	2
	悪性関節リウマチ	1		シャルコー・マリー・トゥース病	2		ヒルシュスプルング病	2
	遺伝性鉄芽球性貧血	1		スティーヴンス・ジョンソン症候群	1		皮膚筋炎／多発性筋炎	2
	ATR－X症候群	1		修正大血管転位症	4		肥大型心筋症	1
	HTLV－1関連脊髄症	1		神経線維腫症	2		非典型溶血性尿毒症症候群	1
	遠位型ミオパチー	1		進行性核上性麻痺	3		非特異性多発性小腸潰瘍症	3
	黄色靭帯骨化症	1		進行性骨化性線維異形成症	1		ファロー四徴症	1
か	家族性地中海熱	2		進行性多巣性白質脳症	2		ファンコニ貧血	2
	海馬硬化を伴う内側側頭葉てんかん	1		重症筋無力症	4		プリオン病	2
	完全大血管転位症	4		脊髄空洞症	1		封入体筋炎	3
	肝型糖原病	1		脊髄小脳変性症	2		副甲状腺機能低下症	1
	急速進行性糸球体腎炎	1		脊髄性筋萎縮症	1		ベーチェット病	2
	球脊髄性筋萎縮症	2		先天性横隔膜ヘルニア	3	ま	マルファン症候群	1
	巨細胞性動脈炎	3		先天性気管狭窄症／先天性声門下狭窄症	1		慢性炎症性脱髄性多発神経炎／多巣性運動ニューロパチー	3
	筋ジストロフィー	2		先天性赤血球形成異常性貧血	1		慢性血栓塞栓性肺高血圧症	2
	筋萎縮性側索硬化症	4		先天性無脾無汗症	2		慢性特発性偽性腸閉塞症	2
	筋型糖原病	2		全身性アミロイドーシス	3		ミトコンドリア病	4
	偽性副甲状腺機能低下症	2		全身性エリテマトーデス	1		無脾症候群	3
	グルコーストランスポーター1欠損症	1		全身性強皮症	1		もやもや病	1
	クロウ・深瀬症候群	3		総排泄腔外反症	1	ら	ライソゾーム病	2
	クロンカイト・カナダ症候群	2	た	高安動脈炎	2		リンパ管腫症/ゴーハム病	1
	結節性多発動脈炎	1		多系統萎縮症	2		リンパ脈管筋腫症	1
	血栓性血小板減少性紫斑病	2		多発性硬化症／視神経脊髄炎	1		両大血管右室起始症	1
	原発性抗リン脂質抗体症候群	3		多発性嚢胞腎	2		類天疱瘡	1
	限局性皮質異形成	1		多脾症候群	3		レーベル遺伝性視神経症	2
	好酸球性多発血管炎性肉芽腫症	2		単心室症	2			
	拘束型心筋症	1		弾性線維性仮性黄色腫	1			
	甲状腺ホルモン不応症	3		中毒性表皮壊死症	3			
	骨形成不全症	1		天疱瘡	1			
	混合性結合組織病	2		特発性後天性全身性無汗症	1			
	強直性脊椎炎	1	な	22q11．2欠失症候群	2			
				嚢胞性線維症	1			
				脳腱黄色腫症	2			

表2 症例報告のタイトル例

症例報告タイトル	学会誌名
周期的な筋原性酵素の上昇と血糖コントロールの悪化を認めた**糖尿病**合併**球脊髄性筋萎縮症**の1例	糖尿病
呼吸筋麻痺と**四肢麻痺**の時期が乖離し診断に苦慮した呼吸筋型**筋萎縮性側索硬化症**の1例	日本集中治療医学会雑誌
超低体温下弓部**大動脈人工血管置換術**後に発症した**進行性核上性麻痺類似症候群**の1例	臨床神経学
Charcot-Marie-Tooth diseaseに合併した**食道神経鞘腫**の1例	日本消化器外科学会雑誌
内分泌障害を伴ったPOEMS**症候群**の1例	日本内科学会雑誌

固有表現タグの内訳： 50種のタグの総出現数は69,807であった．うち，出現個数が多い上位10種のタグを表３に記す．

表3 出現頻度が高い上位10種の固有表現タグ

	名称	付与基準	出現数	タグ付けされるテキストの例
1	state	患者の状態．いわゆる，症状（患者の訴え），所見（医療者の観察結果），病名をすべて含む．	12,859	吐き気 萎縮症, 糖尿病, 口渇感
2	body	解剖用語として特定の名前がつく人体部位．	7,047	手足, 身体, 眼瞼結膜, 球結膜, 眼底
3	item	患者の状態を表すために参照される項目．	5,538	血糖, 血糖値, ＨｂＡＩｃ, 食欲, 身長
4	PN_Positive	患者の状態がある（肯定する）ことが明示される表現．	5,469	気づき, 認め, し, 示す, 認める
5	value	検査値など身体や検体を測定し得られる数値．	4,960	7.5, 20, 1, 5, 165.0, 45.0, 16.5
6	unit	数値との組で表される単位．	4,121	%, mg/日, 行, 回, cm, kg, kg/m2
7	time	時間軸の特定の場所に位置する時点や時区間	3,491	元来, 約10年前, 7年前, その後, 直後
8	clinical_test	臨床検査．itemとの違いは計測方法を含むか否か．	3,154	検査, 神経学的検査, 徒手筋力検査
9	treatment	一定の手順から成る治療法．	2,933	運動療法, 経口投与, コントロール
10	execute_Done	検査や治療などの行為が行われたことが明示される表現に付与するタグ．	1,992	した, 行い, 遵守, おこない, 施行した

関係の内訳： 35種の関係の総出現数は71,047であった．出現個数が多い上位10種の関係を表4に記す．

表4 出現頻度が高い上位10種の関係

	名称	関係の概略	出現数	関係の例
1	value_of	sourceがtargetの値である．	21,243	身長（target）は170 (source) cmであり
2	category	sourceがtargetの一種である．	8,617	筋原性酵素（target）は，CK（source）416IU/L
3	site	sourceがtargetの部位である．	6,944	四肢（target）の筋力(source)低下
4	method	sourceがtargetの（方法に）より得られる．	4,812	聴診（target）上，異常(source) なし
5	unit	sourceはtargetの単位である．	4,433	身長は170（target）cm（source）であった
6	executed	targetの実施が（source)によって明示．	4,032	経口投与（target）が開始された（source）
7	reason	sourceと判断された根拠はtargetである．	2,582	蜂窩織炎(target)から菌血症（source）を疑われ
8	referred_time	sourceの基準時間はtargetである．	2,420	開始（target）4日目（source）に頭痛が出現
9	executed_at	source（行為）が実施された場所/時間はtargetである．	2,083	総合病院（target）に緊急入院した(source)
10	observed_at	source（状態）が観察された場所/時間はtargetである．	1,993	2002年（target）に前医で振戦を認めた（source）

先行研究コーパスとの比較： 奈良先端科学技術大学院大学で開発された医療分野の日本語コーパス（MedTxt-CR[2,3]）との諸元の比較を示す．本コーパスは，1) 関係を利用している，2) 固有表現タグの種類が多い，3) 1文書の文字数が多い，4) 1文書あたりの固有表現タグ数が多い，が特徴としてあげられる．

		MedTxt-CR	本コーパス
文書数		224	179
文書種類		症例報告	症例報告
固有表現タグの種類		13	50
関係の種類		-	35
1文書毎	平均文字数（S.D）	696（320）	1,915（696）
	平均単語数（S.D）	417（195）	972（330）
	平均固有表現タグ数（S.D）	63（25）	394（129）
	平均関係数（S.D）	-	387（127）

結語： 希少・難治性疾患を対象とした179の症例報告からなるテキストコーパスを構築．本コーパスは次の特徴を持つ．

1. 指定難病102疾患を含み，幅広い診療科や疾患領域をカバー．
2. 詳細なタグ粒度でテキストを網羅するｱﾉﾃｰｼｮﾝ基準に基づく．
3. NLPタスクとして定式化し，機械学習モデルの開発に利用可．
4. テキストの再配布許可を得て，研究利用のために公開．

https://ai-health.m.u-tokyo.ac.jp/home/research/corpus

Entity Linkingのアノテーションを追加することが今後の課題．