[46]難病・希少疾患の症例報告を用いた、言語資源作成における課題

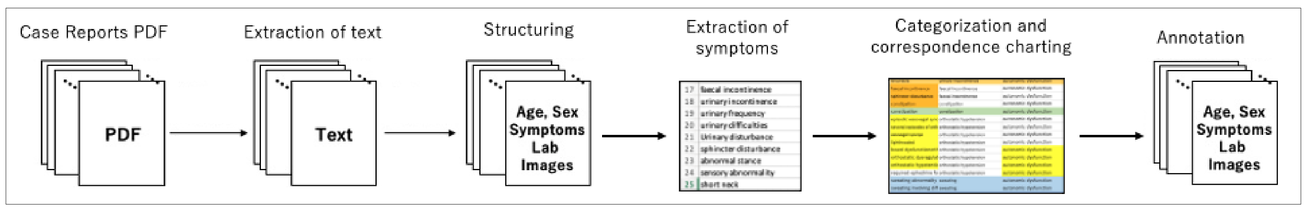
土肥栄祐[1]、建石由佳[2]、藤原豊史[3]、山本泰智[3]

1：国立精神神経医療研究センター　2：国立研究開発法人科学技術進行機構 NBDC 事業推進部 3：ライフサイエンス統合データベースセンター

● Abstract

症例数が少ない難病・希少疾患は、医療者が十分に経験し難く診断に時間を要する。診断支援ツールの性能向上のためにも症例ベースのコーパスの重要性は増している。私たちは日本語症例コーパス作成に取り組んでおり、ここで明らかとなった課題に関し報告する。
コーパス作成において、①PDFからのテキスト抽出、②PDFの病名の誤記・表記揺れ、③日本語病名リストの不足、④症例データの構造化、⑤情報の抽出とアノテーション、⑥HPOの把握の困難さ、これらの課題が明らかとなった。
PDFからのテキスト抽出にChatGPT（GPT-4）は有効であった。難病・希少疾患の病名は、疾患によっては15を超える誤記・表記揺れを認めた。臨床データの構造化には、CaseSharing（臨床データ管理システム）により効率化ができた。情報の抽出とアノテーションに関しては、ChatGPTにより症状・所見の有無、時系列での抽出は可能であったが、症状・所見へのHPOによるアノテーションは困難であった。HPOは全体像の把握が困難であり、HPOに対応語句の有無の判断が困難であった。そこで症例データより作成した対応表をGPTsに情報として与えると、原文へのアノテーションは十分可能となった。
患者側、医療者側の双方にとって理解しやすい症例コーパスは今後ますます重要になると考えられる。残された課題はあるが、技術の進歩により今後の効率化が期待できる。

方法



症例のテキストデータは、症例報告のPDFから抽出され、症状や所見は抽出されたのちに、より高位の概念への対応表を作成。
その後、テキストデータにアノテーションを行った。テキスト抽出や、症状・所見の抽出、またアノテーションなどの手順には効率化のために、生成AI（ChatGPTなどのLLM）を使用。対応表の作成は手動で施行。

結果

1. 病名の表記揺れによる、PDF選択への影響の可能性

```
re.findall(r'[A-z][-A-z ]*病', title)
```

（Pythonの正規表現検索）をやると

```
'Hirochsprung 病',
'Hirsch sprung病',
'Hirschisprung 病',
'Hirschsprnug 病',
'Hirschsprunfi 病',
'Hirschsprung 病',
'Hirschsprung 病',
'Hirschsprung病',
'Hirschsprung病',
'Hirschspruug 病',
'Hirschspurung病',
'Hirschsrung病',
```

日本の指定難病、小児慢性特定疾患の病名オントロジーであるNANDO（Nanbyo Disease Ontology）を用いた。
PDFのタイトル内にある病名と突合させ完全一致するものを抽出したが、病名の表記揺れが、疾患によっては15種類ものバリエーションを呈するものも認め、PDF抽出に影響を与えている可能性が考えられた。

2. 日本語症例報告の、難病・希少疾患のカバー率

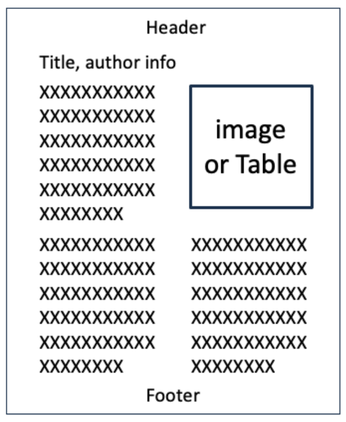
	PDFの数
J-STAGE収載の日本語医学文献	204195
タイトルがユニーク	178148
タイトルに、[例]を含む	45070
タイトルにNANDOの病名を含む	11468
タイトルが番号で始まっていない	2770
PDFからテキスト抽出が可能	2183

NANDO（Nanbyo Disease Ontology）

指定難病	227/1031（22%）
小児慢性特定疾患	220/1681（13%）

日本語の症例報告では、NANDOの疾患全てをカバーできていない可能性

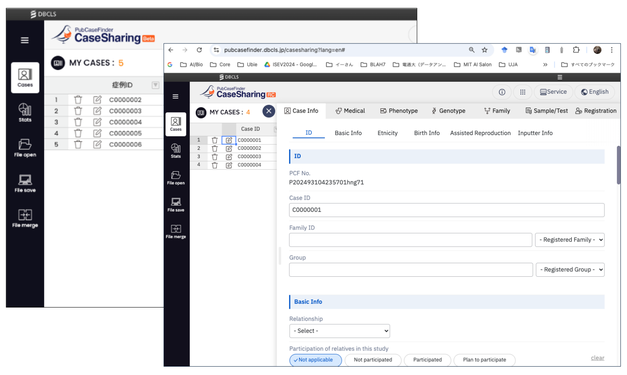
3. 症例報告PDFからのテキスト抽出



日本語症例報告の特徴
・テキストが2カラム形式で配置
・画像やTableデータ
・間にキャプションが挿入
・Header、Footerあり

OCR、LLMによる抽出し、人の確認を加え、効率化と正確性を担保。

4. 症例情報データの構造化

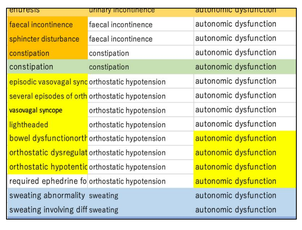


https://pubcasefinder.dbcls.jp/casesharing?lang=en#

Pubcasefinderに実装された症例登録システム「CaseSharing」は、構造化された臨床情報の表示・管理・操作が可能。JSONデータによるデータ抽出やデータ統合が可能なため、症例情報の活用が促進される。

5. 症例情報データの構造化

対応表



faecal incontinence	faecal incontinence	autonomic dysfunction
sphincter disturbance	faecal incontinence	autonomic dysfunction
constipation	constipation	autonomic dysfunction
constipation	constipation	autonomic dysfunction
episodic vasovagal sync	orthostatic hypotension	autonomic dysfunction
several episodes of orth	orthostatic hypotension	autonomic dysfunction
vasovagal syncope	orthostatic hypotension	autonomic dysfunction
lightheaded	orthostatic hypotension	autonomic dysfunction
bowel dysfunctionorth	orthostatic hypotension	autonomic dysfunction
orthostatic dysregulat	orthostatic hypotension	autonomic dysfunction
orthostatic hypotentic	orthostatic hypotension	autonomic dysfunction
required ephedrine fo	orthostatic hypotension	autonomic dysfunction
sweating abnormality	sweating	autonomic dysfunction
sweating involving diff	sweating	autonomic dysfunction

LLM

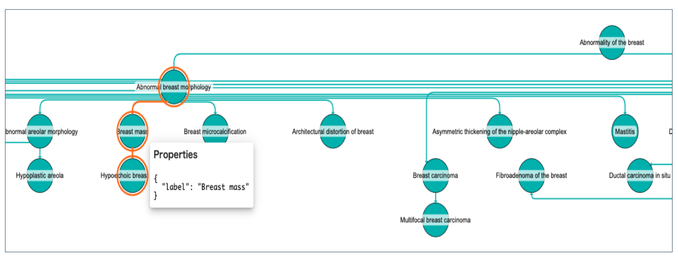
アノテーション後のテキストデータ

A 13-year-old boy exhibited cervical kyphosis and deteriorating cognition<"demetia">. He was born normally and achieved developmental milestones without delay. The family history was noncontributory. At the age of 2, he had 2 simple febrile seizure<"seizure">s. At the age of 8, he had a generalized febrile convulsion that lasted about 10 min. Subsequent electroencephalography (EEG) revealed bilateral frontal spikes during sleep. Valproate was started under the diagnosis of epilepsy. Since then, his school performance slowly began to deteriorate<"psychomotor regression">.

LLMのみでは十分なアノテーションは困難

症状・所見の対応表をLLMにより十分なアノテーションが可能に

6. HPOの可視化による、HPOへの理解の促進



Abnormality of genitourinary_system_HP_0000119_shortest_paths_with_labels_checked

Term	Label	Shortest Path to Root	leaf
http://purl.obolibrary.org/obo/HP_0000119	Abnormality of the genitourinary system	0	
http://purl.obolibrary.org/obo/HP_0000078	Abnormality of the genital system	1	
http://purl.obolibrary.org/obo/HP_0000079	Abnormality of the urinary system	1	
http://purl.obolibrary.org/obo/HP_0007379	Neoplasm of the genitourinary tract	1	
http://purl.obolibrary.org/obo/HP_0012620	Cloacal abnormality	1	
http://purl.obolibrary.org/obo/HP_0100589	Urogenital fistula	1	
http://purl.obolibrary.org/obo/HP_0000080	Abnormality of reproductive system physiology	2	

Dohi E., et al., BioHackathon 2024

アノテーションを評価する際、エンドユーザーの臨床医に、オントロジーに適切な語句が含まれるかの判断は難しい。

各カテゴリー内の数や、親子関係、重複の度合いなど、HPOの可視化と、これを通した新たな利活用が重要。

● Conclusion and Future plan

PDFからのテキスト抽出の種類や数には制限があるが、患者側、医療者側の双方にとって理解しやすい症例コーパスはますます重要になると考えられ、コーパス作成の効率化は重要な課題である。
オントロジーに起因する課題も明らかとなったため、今後は、エンドユーザーに理解しやすいオントロジーの可視化も進める。

COI

発表者には、潜在的な利益相反はないことを宣言します。

Licensed under CC-BY 4.0 ©2024 EISUKE DOHI (NCNP)