29 難病・希少疾患症例コーパス構築の課題と効率化に向けたLLM・ツールの活用

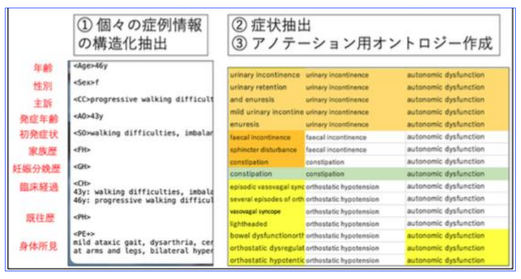
土肥栄祐[1)]、金進東[2)]、早川格[3)]、松旗知康[4)]、高月照江[2)] 、建石由佳[2)] 、藤原豊史[2)]、山本泰智[2)]

1) NCNP　神経研究所　疾病研究第三部、2) DBCLS、3) NCCHD　神経内科、4)　広島大学神経内科

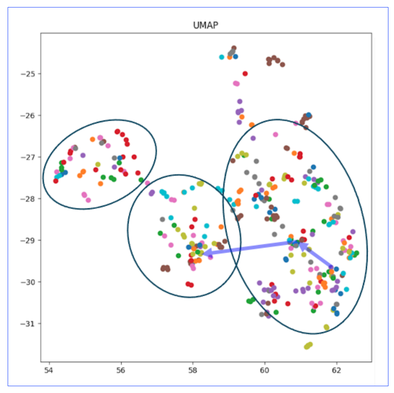
難病・希少疾患は約1万種存在し、症例数の少なさから診断まで時間を要し、特に非典型例では時間がかかるため、質の高い症例コーパス構築やその構築の手法開発が望まれている。本研究では、日本語症例報告に疾患名・症状名をタグ付けした高品質コーパス構築において、大規模言語モデル（LLM）とWebベースのアノテーション管理・編集ツールを組み合わせ、効率化を図った。J-STAGEに収載されている医学文献から、NANDO辞書と形態素解析を用い、難病・希少疾患の疾患名に基づき症例報告を抽出した。テキスト正規化後、LLMによる症状・所見のアノテーションを実施し。JSON形式出力、入力のチャンク化、テキスト部省略によるトークン節約により、安定した結果を得た。アノテーションはPubAnnotationで管理し、GUIベースのTextAEで専門家が評価・修正を行った。Human-in-the-loopにより精度向上と作業負担軽減を実現し、専門家参画を促進した。今後は用語オントロジーの適合性向上など、さらなる課題解決が必要である。

これまでの可視化の検証

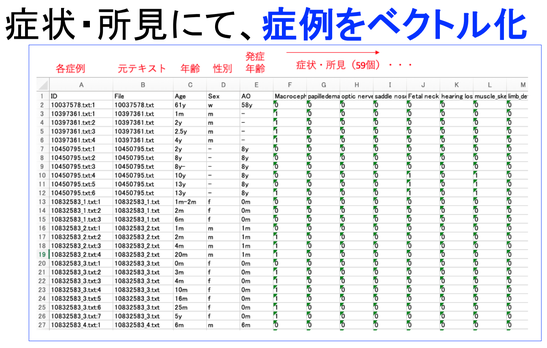
- アレキサンダー病の症例情報（200例）の**構造化**
- 938個の**症状・所見** → 59個の症状所見に**アノテーション**



・クラスタリング
・時系列の可視化
・非典型例の推定



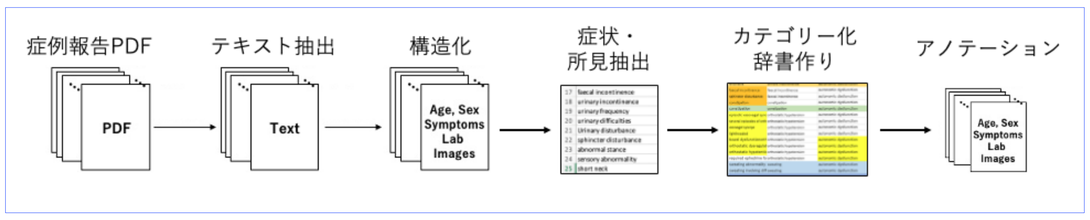
症状・所見にて、**症例をベクトル化**



しかし、これらの工程から明らかとなった課題として、

1: 工程に時間と労力がかかる
2: ツールや、オントロジーは理解が難しく、コーパス作成に携わるエキスパート（医療者）が、参入し難い
→

各工程の自動化やツールの導入を試みた



1：テキスト抽出

・症例報告PDFを、LLM（ここではCloud版のChatGPT）にuploadし、プロンプトエンジニアリングにてテキスト抽出し、マニュアルで確認を行った。

2：構造化

・症例報告における、年齢・性別・主訴・現病歴などのカテゴリーを、大きく分類しするためのPythonコードを作成。抽出したテキストを活用し、内部の構造を把握し統計情報を取得したのちに、構造化されたテキスト形式で出力。

【症例】43歳、女性

【現病歴】患者は 33 歳頃に四肢の線維束収縮を自覚した。36歳時には右下肢の脱力が出現した。四肢・体幹の筋力低下・筋萎縮が進行したため、37 歳時には杖歩行となり、左上肢および両上肢にも脱力が拡大、痰喀出も困難となった。このため精査目的に 37 歳時に当科に入院した。

【既往歴】30 歳時に小麦アレルギー。

【家族歴】母親および母方の祖父がALS。母親は SOD1 遺伝子の変異（Cys6Phe）が確認され、家族性ALSと診断されている（既報告）。

【入院時現症】入院時は頸部および四肢に徒手筋力テストで 3〜4レベルの右下肢優位の筋力低下をみとめ、遠位優位の四肢筋萎縮と腱反射低下、舌の線維束収縮、歩行不安定をみとめた。髄液検査、神経伝導検査、頭部・頸髄・腰髄MRI では明らかな異常をみとめなかった。針筋電図上 3髄節領域で慢性および急性の神経原性変化をみとめ、家族性ALSと診断した。診断確定時には舌の線維束収縮、嚥下障害をみとめ、肺活量 1,770ml、%肺活量 61%と呼吸機能が低下していた。感覚障害や高次機能障害、膀胱直腸障害はいずれもみとめなかった。

【経過】退院後、38 歳時（発症 2年後）には歩行不能、その後も嚥下・呼吸障害が進行したため、39 歳時（発症 3年後）に胃瘻を造設、さらに非侵襲的陽圧人工呼吸（NIPPV）を開始した。40 歳時（発症 4年後）にはさらに呼吸障害が進行したため、気管切開術を施行し終日人工呼吸管理となった。気管カニューレはプラスチック製の内圧自動調節カフ・エバックチューブ付きの 27Fr の口径のものを使用した。それに相前後して腹部膨隆、重度貧血、発熱、および低アルブミン血症がみられ再度入院し、以後中心静脈栄養となった。これらの発熱等の原因は胃瘻部からの局所感染による腹壁膿瘍であり前播種性血管内凝固症候群（Pre DIC）まで合併したが、抗生剤投与で軽快した。この頃には四肢麻痺も進行し、臥床状態となった。41 歳時（発症

テキストの特徴を解析

複数症例がある場合→仕分け
カテゴリーの表記揺れ→標準化

＜除外基準＞
・100文字以内
・標準カテゴリーが2つ以下

段階を踏むことで構造化

3：LLMによるドラフトアノテーション

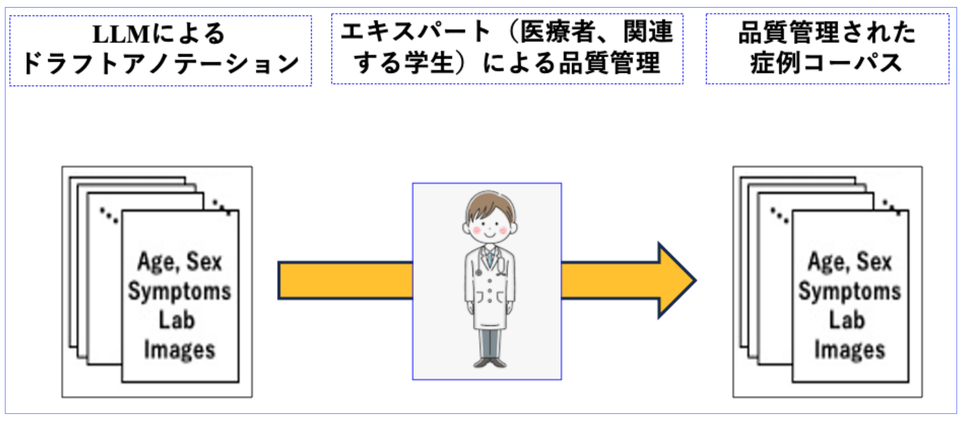
・プロンプトエンジニアリングを用い、ドラフトアノテーション

```
prompt_content = (
    "あなたは臨床家で、HPO(Human phenotype Ontology)に非常に長けたエキスパートです。"
    "与えられた症例報告のテキストを解析し、以下のルールに従ってアノテーションを行い、必ず有効なJSON形式で出力してください。\n\n"

    "【抽出・アノテーションのルール】\n"
    "1. テキスト中に登場する、医学・医療・症状・所見・病名・疾患・検査・データ・遺伝子検査の結果・健康状態などに関係する用語を漏れなく抽出してください。\n"
    "2. 抽出した用語がHPOに対応する場合は、HPOラベルとHPO IDを必ず記入し、対応しない場合はHPOラベルとHPO IDともに \"none\" としてください。\n"
    "3. HPOへの対応は文脈も考慮し、完全一致だけでなく、同義語や同じ意味を持つフレーズも積極的にアノテーションしてください。\n"
    "4. 一つも漏らさず、できる限り網羅的にアノテーションしてください（省略や抜け漏れがないように）。\n\n"

    "【出力フォーマット（JSON構造）】\n"
    "以下の構造を厳守してください。\n"
    "（※ 出力はJSONオブジェクトだけで、余分な文章やコードブロックは不要です。）\n\n"
    "{\n"
    "  \"text\": \"\",\n"
    "  \"denotations\": [],\n"
    "  \"config\": {\n"
    "    \"entity types\": []\n"
    "  },\n"
    "  \"attributes\": [],\n"
    "  \"relations\": [],\n"
    "  \"blocks\": []\n"
    "}\n\n"
    "- \"text\" フィールドは空文字で構いません。\n"
    "- \"denotations\" の各要素には以下の情報を含めてください:\n"
    "    - \"id\": \"T1\", \"T2\", ... のような通し番号\n"
    "    - \"span\": { \"begin\": (用語の開始インデックス), \"end\": (用語の終了インデックス) }\n"
    "    - \"obj\": 3種類の情報（HPOラベル、HPO ID、存在有無）を別々にオブジェクトとして記載（同じspanに対してオブジェクトが複数必要です）\n\n"

    "【denotations の作り方】\n"
    "1つの医学的用語につき、以下の3つの情報をそれぞれ別々のオブジェクトとして格納し、"
    "同じ span(begin, end) を使用してください。\n"
    "  - HPO ラベル（または \"-\"）\n"
    "  - HPO ID（または \"-\"）\n"
    "  - 存在するがんしない (\"+\" / \"-\")\n"
    "各オブジェクトには \"id\": \"Txx\", \"span\": {\"begin\": x, \"end\": y}, \"obj\": \"...\" が必須です。\n\n"

    "【config.entity types】\n"
    "必ず3種類のエンティティタイプを定義し、それぞれに固有のcolorを設定してください:\n"
    "  \"entity types\": [\n"
    "    { \"id\": \"HPO_label\", \"color\": \"#FF8888\" },\n"
    "    { \"id\": \"HPO_ID\",    \"color\": \"#8888FF\" },\n"
    "    { \"id\": \"presence\",  \"color\": \"#88FF88\" }\n"
    "  ]\n\n"

    "【最終的な出力】\n"
    "- 必ず有効なJSONオブジェクトのみを返してください。\n"
    "- コードフェンス（```）や追加説明、解説文は出力しないでください。\n\n"

    f"以下のテキストを解析して、上記仕様を満たすJSONを出力してください:\n\n{text}"
```

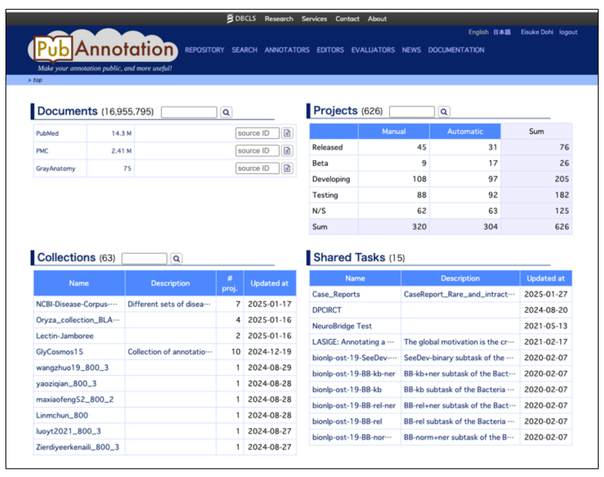
◯ チャンク化（カテゴリー、句読点）
◯ JSON出力（PubAnnotation対応）
◯ 症状・所見の有無
◯ 時間
◯ HPOの有無


```
{
"sourceDB":"hagshogo@Eisuke Dohi",
"sourceid":"2",

"text": "The patient, 17-year-old male, was referred to NCNP Hospital for evaluation of psychomotor deterioration.",
  "denotations": [
    {
      "id": "T1",
      "span": { "begin": 0, "end": 29 },
      "obj": "no HPO_label"
    },
    {
      "id": "T2",
      "span": { "begin": 0, "end": 29 },
      "obj": "no HPO_ID"
    },
    {
      "id": "T3",
      "span": { "begin": 0, "end": 29 },
      "obj": "present"
    },

    {
      "id": "T5",
      "span": { "begin": 91, "end": 116 },
      "obj": "Psychomotor regression"
    },
    {
      "id": "T6",
      "span": { "begin": 91, "end": 116 },
      "obj": "HP:0002170"
    },
    {
      "id": "T7",
      "span": { "begin": 91, "end": 116 },
      "obj": "present"
    }
  ],
  "config": {
    "entity types": [
      {
        "id": "no HPO_label",
        "color": "#8FF88",
        "default": true
      }
    ]
  }
}
```

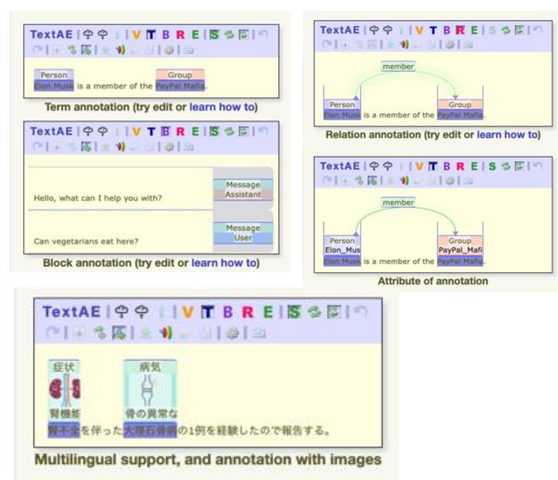

4：アノテーションの品質管理のための体制



PubAnnotation：（https://pubannotation.org/）



TextAE：（https://textae.pubannotation.org/）

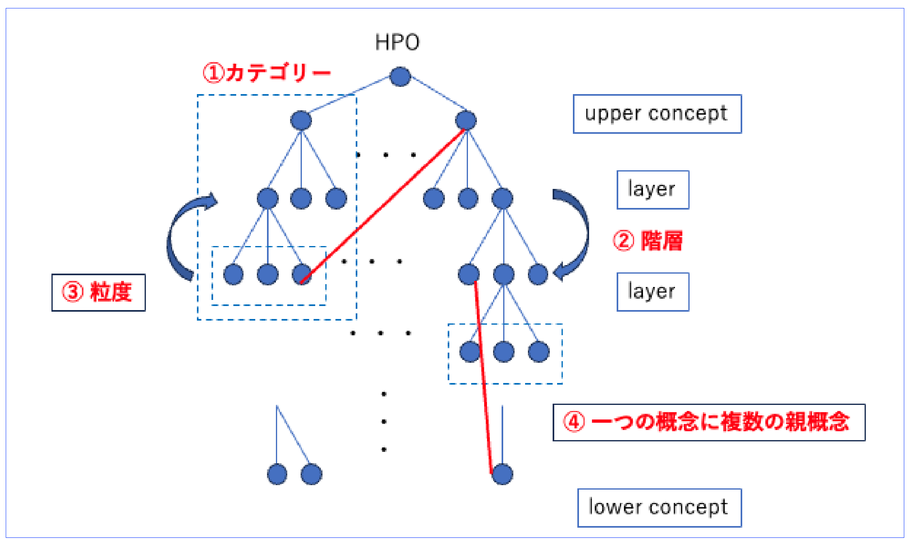


アノテーション管理・編集が可能なウェブツール
◯ 複数のアノテーションの管理が可能、アノテーター間での突合が可
◯ JSON形式で、アノテーションデータのアップロード
◯ GUIベースでアノテーションの確認・編集が可能（TextAE）

LLMにて、**PubAnnotationに対応したJSON形式のデータ出力。**ドラフトアノテーションと、ツールの導入により、エキスパートの参画はしやすい体制構築。

しかし、実際にアノテーションチェックの際に、
アノテーションの不備が、オントロジーに無いためか？
オントロジーの中で最適なものが選択されているのか？
この確認が困難であることが明らかとなった

5：オントロジーの視認性について



① **カテゴリー**：解剖（臓器）、システム（免疫、代謝etc）、時点など、... etc
② **階層**：カテゴリーの対象とするものの複雑性に依存して深さが異なる。
③ **粒度**：エンドフェノタイプを示すものが、別の階層に存在する。
④ **複数継承**：一つの概念に複数の親がある時
⇨ これが、俯瞰的な見方や全体像を掴むことを難しくしている

このDAGをわかり易く検索・利活用する可視化ツールを開発し、エキスパート（臨床医や医学生）のチェックの効率化を進めている

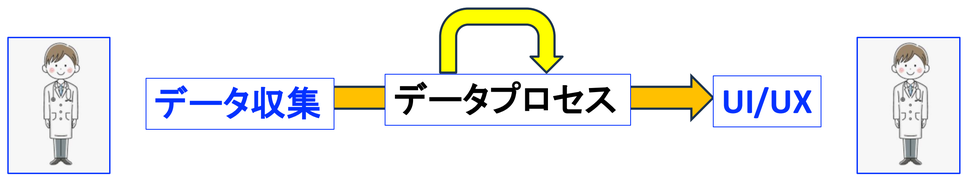
6：達成できた課題と、残る課題

- 実症例の蓄積＋生成AI
- 元データの質の底上げ
- 病名の表記揺れ・誤記
- 情報の構造化（自動化）
- アノテーションの効率化

- 現場の人が協力しやすいUI/UX
- 病名の表記揺れ・誤記
- HPOの整備・可視化 or 独自基準
- 時系列データの表現
- ローカルでの構築

7：まとめと展望

プロセスは技術で進歩してゆくため、生データの質、データの使い易さ・連結しやすさが、一層価値を増してくると考えている。



今後は、初期に行ったテキスト抽出などの自動化を再検討に加え、DAGの可視化・編集ツールを横展開し下記の可能性を展開してゆく。

① 医療情報の新規可視化法による、医療−患者間コミュニケーションの改善
② バックキャストによるデータ収集法の最適化
③ 施工過程の可視化による医療者の学習の効率化