LLMを用いた呼吸器内科SOAPデータの分類、特徴抽出及び構造化

張 孜恒[1]、韓 豊[1]、唐 明輝[2]、平田 健司[2]、北井秀典[3] 、小笠原 克彦[4]、工藤 興亮[2]

1. 北海道大学大学院医学院　画像診断学教室
2. 北海道大学大学院医学研究院　画像診断学教室
3. 北海道大学大学院医学研究院　呼吸器内科学教室
4. 北海道大学大学院保健科学研究院　健康科学分野

Introduction &Objective

電子カルテ（EMR）の普及により、SOAP（Subjective, Objective, Assessment, Plan）形式の診療記録が標準的に使用されている[1]。しかし、SOAPは記述の自由度が高く、情報の標準化や自動処理が難しいという課題がある[2]。

例えば、同じ診断・症状でも記載方法が異なる場合が多く、「主訴 (CC)」や「現在の病気の履歴 (HPI)」の境界が曖昧になったり、「診断 (Assessment)」の記述が治療計画 (Plan) に含まれることもある。また、医師ごとの表記の違いにより、自動処理やデータ解析の妨げとなる[3] 。そのため、SOAPフォーマット内の情報を、**統一的なルールに基づいて詳細なカテゴリ（第一層）とその細分類（第二層）に分け、標準化・構造化する**ことが重要である

本研究では、**日本語の呼吸器内科SOAPデータを対象に、大規模言語モデル（LLM）を用いた自動アノテーション手法を開発・評価し、診療記録の情報抽出と構造化の精度を向上させることを目的とする。**

Methods

1. データセット

200例の呼吸器内科SOAPデータ
（医師による確認済み、ゴールドスタンダードデータとして使用）

- トレーニング用データ160例
- 評価用データ40例

	S(症状)	昨日よりずいぶん体が楽である。本日、浣腸後排便あり、とても楽になった。以前は毎朝8時30分に排便あったが、ここ1週間は排便なかった。少しからだを動かそうとして室内を歩いたら、看護師さんの止められて、今日は安静にしようと思う。
	O(所見)	BT37.8→36.9度　P110　BP130/81　SpO2　98%(room air) BW40.50
症例数: 11万2420 病名: 1.肺気腫 2.肺腫瘍 3.気管支喘息 4.肺線維症 5.肺炎		L/D 01 白血球数　5.6　緊 02 赤血球数　3.13 L 緊 03 ヘモグロビン　8.5 L 緊 …… 30 カルシウム　8.1 L 緊　ZT CXp (ポーター臥位) コントラストの差の影響もあるが8/3と比較し両肺野の透過性は改善
	A(判断)	昨日、入院時と比較し呼吸状態、全身状態の悪化なし 貧血は赤血球輸血2単位によりHb8.5まで改善 低Na進行注意 廃用症候群注意→明日以降、付き添いで棟内フリー等を御検討いただく
	P(計画)	アザクタム、ダラシンによる加療継続 明日、CXp、L/Dにて再評価

Figure 1: Example of SOAP data

2. アノテーション方式

- 15の**第一層ラベル（カテゴリ）**（例：「生命兆候 Vital Signs」「身体検査 Physical Examination」）
- 58の**第二層ラベル（詳細ラベル）**（例：「血圧 Blood Pressure (BP)」「脈拍 Pulse (HR)」）
- **階層的アノテーション**：
  Step 1: 第一層ラベル（カテゴリ）を適用
  Step 2: 第一層の範囲内で第二層ラベルを適用

Figure 2: Annotation Labels

3. モデル・手法

- LLaMA3.3: 70b
- **Few-shot Prompting（5例） + Prompt Engineering + Retrieval-Augmented Generation (RAG)** (160例のゴールドスタンダードデータ)

Future

1.未登録ラベルが生成される問題 → **Promptの最適化が必要**
2. 第二層ラベルの適用に課題あり → **Prompt Tuningが必要**
3. **評価を行い**
   - Precision（適合率）、Recall（再現率）、F1-score
   - LLMの出力と医師のアノテーションを比較し、精度を評価
   - 以下の2つの状況は正しいと判定
     ✓ 完全に合わせる
     ✓ 完全に合わせないが、80%以上のテキスト類似性が持つ場合

Results

✅ 良い結果

1. Vital Signs（生命兆候）など、**定型的な情報は比較的正確に分類される**
   - 例：「SpO2 96%, BP 135/80, HR 80」は適切に「生命兆候」として分類
   - 第二層のラベル（「血圧」,「脈拍」,「SpO2」）も適切に付与

2.**第一層（カテゴリ分類）は概ね意図したラベルで識別される**
   - 「生命兆候（Vital Signs）」、「身体検査（Physical Examination）」のカテゴリで特に高精度

```
{
  "text": "O) SpO2 94-96%, Bp146/96, P124, BT 37.8℃",
  "labels": ["生命兆候 Vital signs"]
},
{
  "text": "SpO2 94-96%",
  "labels": ["経皮的動脈血酸素飽和度 Oxygen Saturation (SpO2)"]
},
{
  "text": "Bp146/96",
  "labels": ["血圧 Blood Pressure"]
},
{
  "text": "P124",
  "labels": ["心拍数 Heart Rate"]
},
{
  "text": "BT 37.8℃",
  "labels": ["体温 Body Temperature"]
},
```

Figure 3: 良い結果の例（モデルの一部出力）

⚠ 誤りの傾向

1.**未登録ラベルの生成**：事前に定義されていないラベルが生成された
   - 例：「症状」や「赤血球数」など、定義されていないラベルが出現）

2. **第二層ラベルの適用が不完全**
   - 例：「*胸部写真：もともと右上肺野の透過性が低下しているが、その隙間の透過性が低下*」が「画像診断」には分類されるが、第二層の「X線検査」のラベルが付与されなかった

```
{
  "text": "胸部写真: もともと右上肺野の透過性が低下しているが、その隙間の透過性が低下",
  "labels": ["画像診断 Imaging Studies"]
},
{
  "text": "A) 気管支肺炎の疑い",
  "labels": ["診断 Diagnosis"]
},
```

Figure 4:　第二層ラベルの適用が不完全の例

Discussion &Conclusion

本研究では、SOAPデータの階層的ラベリングをLLMを用いて自動化する試みを行った。結果として、第一層ラベルの適用は概ね良好であるが、第二層ラベルの適用にばらつきが見られた。特に以下の点が課題として浮かび上がった：

1. 未登録ラベルの出力：**ほとんどは過剰推論のため、Prompt をさらに強化する必要あり**
2. 第二層ラベルの適用が不完全： **Few-shotの例数が足りないのが原因と考えされるが、これ以上増やすと文字数が多すぎるため、Prompt Tuningが必要**

これらの問題は、**モデルがラベルの適用ルールを完全には理解できていない**こと、および**入力テキストの構造的なばらつき**に起因すると考えられる。

[1] Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes . I.B. Wilson et al. 1995
[2] LungDiag: Empowering Artificial Intelligence for Respiratory Disease Diagnosis Through Electronic Health Records. Hengrui Liang. Et al. 2023
[3] Large language models as data preprocessors . Zhang H . et al. 2023

＊ 開示すべきCOI: 利益相反なし。