Strategies for Annotating Clinical Case Reports on Intractable Rare Diseases Using Large Language Models (LLMs)

◯Eisuke Dohi, Jing-Dong Kim[1], Itaru Hayakawa[2], Tomoyasu Matsubara[3], Toyofumi Fujiwara[1], Yuka Tateishi[4], Yasunori Yamamoto[1]

1. Research Organization of Information and Systems, Database center for Life Science
2. National Center for Child Health and Development, Department of Neurology
3. Tokushima University, Department of Neurology
4. Japan Science and Technology Agency, Department of Information Infrastructure Office of NBDC Program

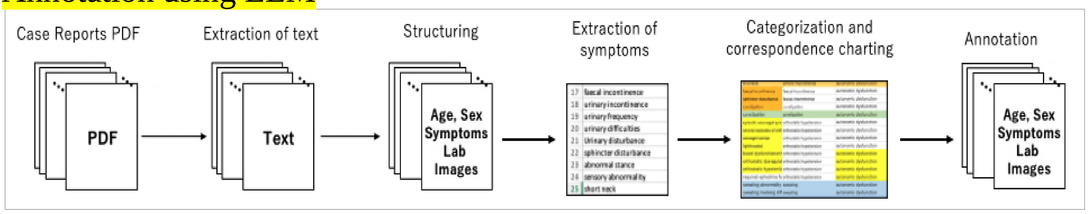
Introduction

The number of intractable and rare diseases is estimated to be around 10,000. Due to the limited number of cases, it is difficult for medical professionals to gain sufficient experience, and it is said that an average of 7 to 8 years is required to reach a diagnosis. In response to this challenge, the application of artificial intelligence is being explored, and there is a growing demand for the development of high-quality case corpora. However, the creation of such corpora involves considerable effort, including the complexity of the annotation work and the verification of the terms used for annotation. Furthermore, in the process of constructing corpora based on case reports, the cooperation of clinicians and medical students with adequate clinical expertise is required. From the perspectives of inefficiency and the substantial effort involved, securing personnel becomes a significant obstacle. Against this background, we are streamlining the process by utilizing large-scale language models (LLM) in the development of a corpus in which disease names and symptom names are tagged in Japanese case report texts, in combination with web tools (PubAnnotation and TextAE) for managing and editing annotations.
In this paper, we introduce the challenges inherent in corpus creation and share our implementation experience with Human in the loop in case corpus creation using LLMs and these tools, while discussing the functions necessary for an efficient workflow.

Papers targeted for text extraction

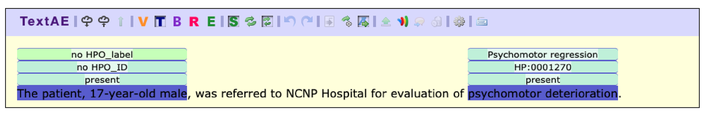
The study targeted 204,195 medical literature items from J-STAGE using an agreement with JST. It extracted case reports on intractable and rare diseases by filtering metadata and using the NANDO disease name ontology. Through morphological analysis with a custom MeCab dictionary, 11,468 papers were initially identified; after excluding non-case reports and conference abstracts, 2,770 papers remained. Text was extracted from PDFs using ChatGPT (4o), and approximately 1,000 case report texts were collected after manual review. Finally, the texts were normalized by standardizing section headings (e.g., [Chief Complaint], [Physical Findings]) to enhance clarity and facilitate integration across specialties.

Annotation using LLM



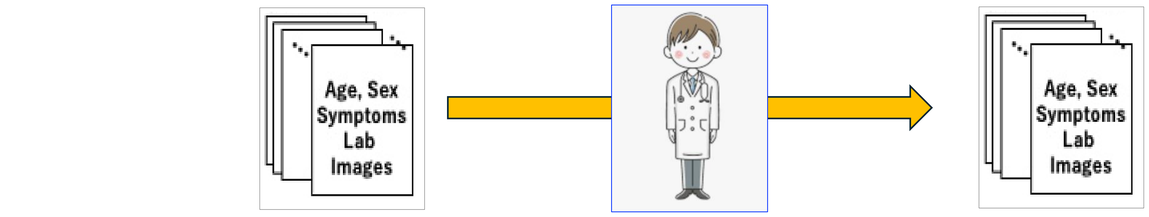
JSON Format Management: PubAnnotation stores, manages, and edits both text and annotation data in JSON format.
Instruction Prompt Design: A prompt was created for LLM-based annotation to ensure outputs strictly adhere to the JSON structure.
Annotation Labeling: Labels and IDs were primarily assigned to disease-related items (per [Robinson 2008]) with provisions for clinically significant items not covered by HPO.
LLM Utilization: OpenAI's ChatGPT API using the "o1-mini" model was employed, with prompt refinement through trials to achieve stable JSON annotations.
Token Management:
・Input tokens were chunked to prevent premature termination
・output tokens were conserved by omitting raw text data during annotation.
Seamless Integration: Post-output, the raw text is reintegrated to produce JSON data compatible with PubAnnotation, ensuring smooth annotation management.

Fig1. Display in TextAE, where intuitive operations are possible through the GUI. Examples of JSON format

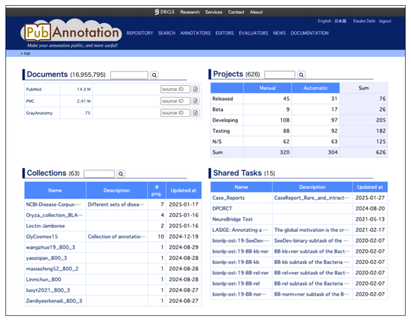


By specifying the database and document ID, project management becomes possible. When using LLM for annotation, omitting the text part saves output tokens.
The text is shown here for illustration, but the target text is "The patient, 17-year-old male, was referred to NCNP Hospital for evaluation of psychomotor deterioration.".
Three annotations—a HPO label, a HPO ID, and its presence or absence—are assigned to a single term or phrase.

Workflow

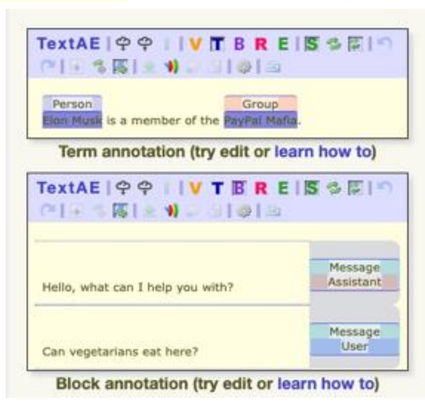


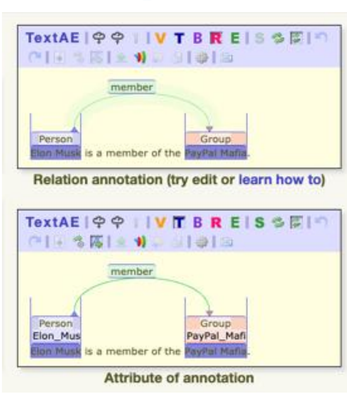
PubAnnotation : (https://pubannotation.org/)



Web tool ：managing and editing annotation projects.
Multiple projects：Any document for diverse annotation perspectives and comparisons.
Direct annotation, editing and bulk upload： JSON-formatted text and annotation data.
GUI-based TextTAE： Intuitive annotation editing.

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





Display Features:
・Shows the document's text with highlighted annotated sections.
・Clearly displays annotation tags.
Multifunctional Capabilities for Annotation:
① Individual words/phrases ② Relationships between words
③ Sentence or paragraph blocks ④ images
Efficiency Enhancements:
・Annotations can be performed via intuitive drag-and-drop and shortcut keys.
・Project-specific instruction guides facilitate stress-free learning of the tool.
Accessibility and Integration:
・Standalone web tool for learning annotation methods in advance.
・Enables editing of pre-annotated documents registered in PubAnnotation via the GUI.
・Integrates document management and annotation on a single platform.

Conclusion and perspectives

For creating a high-quality corpus, ensuring quality through experts remains important. In fact, the use of LLM-based annotation has dramatically improved the efficiency of corpus creation. Additionally, the development and enhancement of ontologies continue to be key issues, and plans are in place to leverage AI to further improve these aspects.