Using Artificial Intelligence Chains for SNOMED CT entity extraction from free text clinical notes

Alejandro Lopez Osornio | SNOMED International
Kevin Gao | Undergraduate, Imperial College London
Rory Davidson | SNOMED International
Yongsheng Gao | SNOMED International

Abstract

In this project, we use LLMs (Large Language Models) to extract clinical concepts in SNOMED CT from free text. We apply the concept of Chaining LLM steps and terminology server queries together, where a task is split into subtasks which the LLM completes in order, each time building on the output of the previous step[1]. Using an LLM Chain, terms undergo multiple layers of LLM-driven simplification until they can be matched to a concept by name using the terminology server APIs.

The two main applications are post-processing of clinical notes, where concepts are extracted from a free text record without human interaction, and real-time processing, where the concepts in a description of a patient's current conditions or procedures are compared against electronic patient records.

Chain components

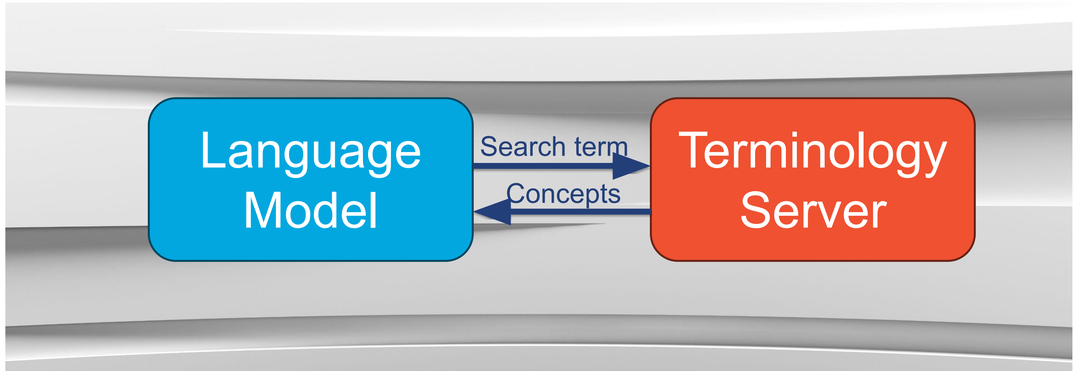
Large Language Models (LLMs)

The language model performs all entity extraction, simplification and semantic similarity rating steps in the chain. We used a generic framework that can run with different language models:

- Llama-2 (Meta) – 13 Billions parameters version
- gpt-3.5-turbo, gpt-4 (OpenAI) (using chat completions)
- Bard (Google) (attempted using a web scrapper, no API was available)

Terminology Server

The terminology server is used match clinical terms and retrieve SNOMED Concepts.



Interaction between the Language Models and the Terminology Server

Introduction

Global terminology enabling quality information exchange
SNOMED CT EXPO 2023
OCTOBER 26 - 27 2023
ATLANTA USA & ONLINE

ePosters sponsored by: termMed

Using Artificial Intelligence Chains for SNOMED CT entity extraction from free text clinical notes

Alejandro Lopez Osornio | SNOMED International
Kevin Gao | Undergraduate, Imperial College London
Rory Davidson | SNOMED International
Yongsheng Gao | SNOMED International

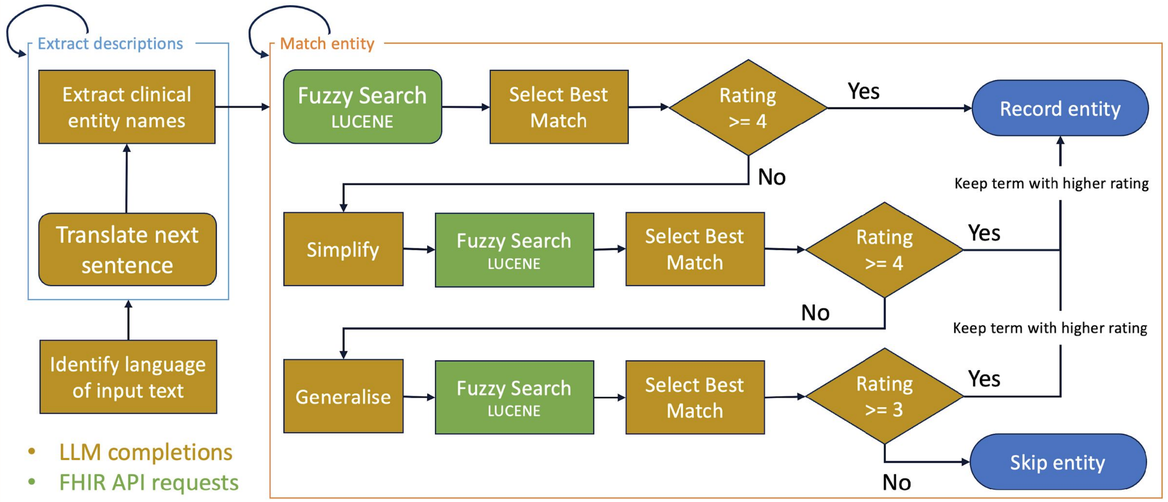
Methods

- The AI chain was implemented in Python, and relies on extensive prompt engineering techniques to define tasks for the LLMs
- The goal of the script is to translate the text to English if necessary, extract clinical entities, and then initiate an iterative process of refinement to identify the best SNOMED CT match for each entity
- In the “Rating” steps the LLMs evaluate the semantic similarity between the search term and a possible match, and decide whether it is necessary to attempt alternative matching strategies

Example LLM prompt

Semantic_similarity_rating_prompt = [{"role":"system", "content": """You are a clinical expert, that compares terms doctors write in clinical notes with SNOMED CT terms selected to represent the same meaning.
You will be given a clinician's term and a SNOMED term, and you need to assess how closely the terms are related **based on the given context**, on a scale from 1 to 5, where **1 means no meaningful relationship** and **5 means identical meaning**.
Do not include any commentary or explanation in your response. **Use this table for guidance:**
...

Chain Architecture



- Snowstorm Lite was selected as terminology server, a new simplified terminology server with a FHIR API and a Fuzzy Search feature
- A sample of phrases with clinical descriptions was extracted from case descriptions in clinical cases on peer reviewed journals
- A training set of 10 phrases was used to refine the chain implementation, and an evaluation set of 10 different phrases was reserved for human annotation and validation of the results

Using Artificial Intelligence Chains for SNOMED CT entity extraction from free text clinical notes

Alejandro Lopez Osornio | SNOMED International
Kevin Gao | Undergraduate, Imperial College London
Rory Davidson | SNOMED International
Yongsheng Gao | SNOMED International

Sample Llama-2 Execution

Paste link to video file here and we will embed. Please send separately.

Embed “llm-chain-output-eposter.mp4”

- Test hardware: Mac Studio M2 Max, 32gb RAM
- Total execution time is 772 seconds, 12.8 minutes, average 1.2 minutes per clinical phrase (average 25 words per clinical phrase)
- Llama-2 is slower than using OpenAI APIs, but preferred for the demonstration as an open source language model that can run locally

https://github.com/IHTSDO/llm-chain-entity-extraction

Results

- Executing the AI Chain with clinical phrases from the evaluation set, in comparison with the results provided by the human annotator.

	gpt-3.5	gpt-4	Llama-2
Coverage			
% of terms detected	100%	100%	92%
Precision*			
Identical terms	56%	70%	62%
Identical or broader	74%	93%	88%
Execution time	109 sec.	116 sec.	772 sec.

* Computed based on effective matches

- It was not possible to effectively use Google Bard with a web scrapper
- OpenAI models were more responsive to system prompts that define the parameters for the completions

Using Artificial Intelligence Chains for SNOMED CT entity extraction from free text clinical notes

Alejandro Lopez Osornio | SNOMED International
Kevin Gao | Undergraduate, Imperial College London
Rory Davidson | SNOMED International
Yongsheng Gao | SNOMED International

Discussion

- The AI Chains model presents an effective way to decompose a complex process like clinical coding into discrete tasks that can be executed by AI models and terminology servers
- Prompt engineering: an iterative process for testing prompts is required to identify the ones that provide the best results on each model, and this will probably require revisions with model updates
- The level of accuracy and coverage does not seem to be appropriate for clinical care use cases today, but for some other scenarios, like quality metrics or identifying candidates for clinical trials may be already applicable
- The use of closed models like OpenAI has better results but it may be limited by privacy and security requirements
- Local processing with Open-Source models like Llama-2 is private and secure, but it has a high demand of computer hardware for obtaining practical processing speeds

Meta AI OpenAI

Conclusions

- The use of pretrained **Large Language Models connected with terminology services** promises to be an effective model to implement clinical entity extraction pipelines, avoiding common pitfalls like hallucinations
- It is expected that with new versions of these models being released, accuracy and speed will improve rapidly

Future Directions

- Additional testing against human benchmarks
- Constraining matching by clinical domains
- Represent presence or absence with context attributes
- Use of standard Chains frameworks (LangChain)
- Formalizing input and output formats for batch processing

References

1. Wu, T., Terry, M., & Cai, C. J. (2022, April). AI chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In Proceedings of the 2022 CHI conference on human factors in computing systems (pp. 1-22).

Global terminology enabling quality information exchange
SNOMED CT EXPO 2023
OCTOBER 26 - 27 2023
ATLANTA USA & ONLINE

Discussion

ePosters sponsored by: termMed