Jackalope Plus: AI-Enhanced Solution for Mapping Unmappable Concepts

Denys Kaduk [1,3], Marta Vikhrak [1], Polina Talapova [1,2], Eduard Korchmar [1,4] , Inna Ageeva [1], Max Ved [1]
1 IT company SciForce, Kharkiv, Ukraine
2 Kharkiv National Medical University, Kharkiv, Ukraine
3 V. N. Karazin Kharkiv National University, Kharkiv, Ukraine
4 Goldsmiths, University of London, London, United Kingdom

Background

Conversion of medical data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) relies on mapping the source terms to standard concepts in the OMOP Standardized Vocabularies. However, this process is limited by the availability of standard concepts at a given time. While this limitation may not be significant for most source data, there are situations where direct mapping results in the loss of relevant clinical details, especially when dealing with complex meanings or uncommon local variations of medical expressions.

Until a year ago, there were no available methods to preserve the semantic meaning of such cases in a reusable way. In 2022, we introduced the open-source Jackalope software[1] as a supplementary pipeline, which utilizes the sophisticated post-coordination model provided by SNOMED CT to capture, process, utilize, and preserve the full semantic meaning of unmappable source data. The tool creates custom SNOMED-like concepts with attributive and hierarchical relationships and incorporates them as new standards into an OMOP CDM instance. Nevertheless, the current implementation still requires skilled manual creation of post-coordinated expressions (PCE) for unmappable concepts, which hampers user compliance.

In this paper, we propose an AI-enhanced version of Jackalope to eliminate the need for manual processing of source terms without standard equivalents in the OMOP Vocabulary.

Methods

To enhance the efficiency of the Jackalope pipeline, we have incorporated Natural Language Processing and Machine Learning components focused on advancing the semantic search and parent concepts suggestion processes. These components utilize an embedding model to reduce the pool of concepts, ultimately streamlining the handling of clinical and observational data mapping.

The primary elements include:

- *Semantic Search*: Our system leverages semantic search powered by an embedding model. This approach significantly narrows down the pool of concepts by identifying the top N concepts most relevant to the input concept. The service performs a search with a state-of-the-art speed without compromising on results. This semantic search is instrumental in ensuring that the system works with precision, reducing manual effort, and improving the accuracy of the mapping process.
- *Final Parent Concepts Suggestion*: Once the semantic search has refined the list of relevant concepts, we employ this reduced list to suggest final parent concepts. This step is crucial in providing users with concise and effective recommendations, further enhancing the efficiency of the mapping process.

By integrating these elements into our pipeline, we aim to create a seamless and *intelligent system* that not only minimizes manual effort but also maximizes the accuracy and relevance of clinical and observational data mapping.

Results

Our comprehensive approach to testing the pipeline involved the creation of a diverse dataset. This dataset was crafted, comprising both real-world ICD-10-CM terms and synthetically generated terms developed by our team of medical experts. This combination allowed us to thoroughly assess the pipeline's performance across various scenarios and challenges.

By incorporating genuine medical terminology alongside expert-generated content, we ensured that our testing process encompassed a wide spectrum of clinical and healthcare contexts. This approach not only validated the pipeline's capability to handle real-world data but also its adaptability to specialized healthcare scenarios.

Once post-coordinated expressions are generated, they can be seamlessly transferred to the Jackalope interface through a POST request. This process establishes a new vocabulary containing new concepts within the OMOP CDM instance, with SQL or CSV serving as the backend. Additionally, the development of a user-friendly interface has made Jackalope more accessible during the communication process and more useful when adding pre-existing post-coordinated expressions to existing vocabularies.

Using the model we obtained results with semantic similarity scores for the comparison as shown in Table 1 and final parent concept suggestions shown in Table 2.

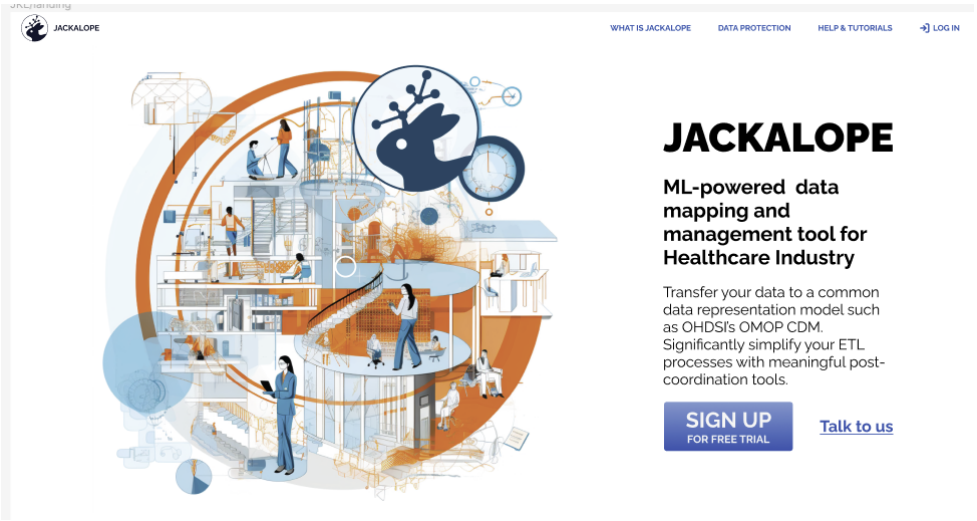
Table 1. Semantic similarity score generated by model for search

Name	Score
Laceration of greater saphenous vein at lower leg level, left leg, initial encounter	
Injury of greater saphenous vein at lower leg level	0.94
Injury of lesser saphenous vein at lower leg level	0.94
Laceration of left lower leg	0.93
Cocaine dependence with intoxication delirium	
Cocaine intoxication delirium	0.97
Cocaine induced delirium	0.96
Cocaine delirium	0.96
Cocaine intoxication	0.94
Cocaine dependence	0.93

Table 2. Final parent concept suggestions

Source Name	Concept Name	Score
Aphasia following nontraumatic subarachnoid hemorrhage	Aphasia due to and following non-traumatic subarachnoid hemorrhage	0.99
Ataxia following nontraumatic subarachnoid hemorrhage	Ataxia due to and following non-traumatic subarachnoid hemorrhage	0.98
Laceration of greater saphenous vein at lower leg level, left leg, initial encounter	Injury of greater saphenous vein at lower leg level	0.94
	Laceration of left lower leg	0.93
Cocaine dependence with intoxication delirium	Cocaine dependence	0.93
	Cocaine intoxication delirium	0.97

These scores indicate that the model demonstrates an understanding of the context, as reflected in the semantic similarity results. The user interface (UI) in our project is designed with a strong emphasis on user-friendliness and accessibility. It features intuitive navigation and a visually appealing layout to ensure that users, whether they are healthcare professionals or researchers, can seamlessly interact with the system, fostering a positive and productive experience.



Conclusion

The integration of AI service into Jackalope's data processing and transformation pipeline offers numerous benefits. By leveraging its Natural Language Processing capabilities, versatility, and automation features, the efficiency of the ETL (Extract, Transform, Load) process is enhanced. Additionally, the extraction of attributes and parents contributes to a more comprehensive understanding of the data. Through the utilization of the Jackalope interface and standardized vocabularies, the overall experience of working with post-coordinated expressions is streamlined, providing an efficient and user-friendly solution. The potential next steps include conducting further beta-testing of the tool with a group of adopters who will receive an access to evaluate the performance of the AI-enhanced Jackalope in generating post-coordinated expressions for

unmappable concepts. This can be followed by refining the algorithms for parsing source data and improving the automation process. We would also like to encourage people to join our journey.

References

1. Jackalope (https://www.ohdsi.org/2022showcase-41/)