Comparative Methods for Building Chatbots: Open Source, Hybrid, and Fully Integrated Large Language Models

Kristen "Kit" Lewers [‡]

‡ University of Colorado Boulder, Boulder, United States of America

Corresponding author: Kristen "Kit" Lewers (krle4401@colorado.edu)

Abstract

In the complex and dynamic realm of biodiversity informatics, the accessibility and comprehension of standards and vocabularies are pivotal for, but not limited to, effective data management, research, policy, regulation, and education. Biodiversity Information Standards (TDWG) provides a suite of standards crucial for the interoperability and consistency of biodiversity data applied to petabytes of data aggregated at GBIF (Global Biodiversity Information Facility). Among these, Darwin Core (DwC; Darwin Core Task Group 2009) and its extensions, stand out as foundational frameworks that guide the mapping and sharing of biodiversity information. However, the richness and depth of these standards, while essential for biodiversity data interoperability, often present challenges for stakeholders, especially as a barrier to entry for those newly introduced to the field. This paper presents a comparative study of four distinct methods for building chatbots aimed at enhancing the accessibility and understanding of these biodiversity informatics standards.

This project introduces an innovative approach to mitigating the complexities inherent in navigating TDWG standards. The project aims to create specialized, conversational interfaces by leveraging different methods, including fully open-source solutions without large language models (LLMs), fully open-source solutions using LLMs, hybrid approaches leveraging OpenAI's API, and fully integrated solutions using GPT (Generative Pre-trained Transformer) models. These interfaces are designed to facilitate easier querying and interpretation of the nuanced aspects of biodiversity standards. This could be especially helpful for individuals who are new to the world of biodiversity standards and are not sure where to start. The implementation would allow for individuals to engage with standards on their own time and own terms, if members of the organization were unavailable. The urgency and importance of this project are underscored by the accelerating pace of biodiversity loss and the critical role of data standards in supporting research efforts. By enhancing the accessibility of TDWG

standards, this project directly contributes to improving data management practices, thereby supporting the broader objectives of biodiversity informatics.

The methodology begins with a comprehensive data collection phase, targeting both the structured documentation of TDWG standards and the community-generated content on GitHub*[1]. This dual-source approach ensures that the chatbot training dataset not only covers the foundational aspects of the standards but also captures the current discussions, updates, and real-world applications as reflected by the community of practice. Following data collection, a rigorous preprocessing phase is undertaken to optimize the dataset for model training. This involves normalization, data augmentation, noise reduction, and careful annotation, all aimed at creating a consistent, relevant training set. The subsequent model training phase utilizes various techniques depending on the method employed:

1. **Fully Open Source without Large Language Models:** Utilizing tools like spaCy and NLTK (Natural Language Tool Kit), we build chatbots based on traditional NLP (Natural Language Processing) techniques without leveraging large language models.
2. **Fully Open Source with Large Language Models:** Incorporating transformers from Hugging Face to enhance the chatbot's capabilities.
3. **Hybrid Approach:** Using OpenAI's API to generate structured question and answer pairs and training an open-source model (like Rasa) with this data.
4. **Fully Integrated Large Language Models:** Employing davinci-002, an outdated but cheap GPT; GPT-3 as a more affordable option, in terms of tokens; or GPT-4, to leverage the latest and most robust model, for the entire chatbot solution, leveraging its extensive capabilities for natural language understanding and generation.

For this project, the choice of models was driven by both budget constraints and the need for accurate, detailed responses. For example, the older OpenAI davinci-002 model, despite its affordability, yielded results that were less than satisfactory, even though a GPT product, highlighting the trade-offs between model capabilities and cost. The comparative analysis of the four methods is based on criteria such as performance, cost, ease of implementation, flexibility, and scalability with testing and iteration still on-going.

Developing a specialized chat model for biodiversity informatics standards is a complex and multi-step process that involves careful data collection, preparation, and iterative model training. Each method brings its own set of challenges and benefits, and the choice of method can significantly impact the chatbot's effectiveness and user satisfaction. Despite the complexities, the proofs of concept thus far have demonstrated promising results and will continue to be refined with the goal of enhancing the tool’s accuracy and user-friendliness. User testing and feedback with a variety of experience levels regarding TDWG standards are the next steps in the project. This project represents a confluence of cutting-edge artificial intelligence and community-sourced expertise aimed at bridging gaps in the field of biodiversity informatics. By making the TDWG standards more accessible and understandable, this initiative aims to enhance support for biodiversity informatics workflows, improve data management practices, and foster a deeper engagement with biodiversity data standards.

Keywords

NLP, machine learning, GPT, biodiversity informatics, Darwin Core, TDWG, biodiversity informatics workflows, user engagement, standards accessibility

Presenting author

Kristen "Kit" Lewers

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Darwin Core Task Group (2009) Darwin Core. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/450

Endnotes

*1 https://github.com/kllewers/tdwg_scraper