Effectiveness of Cross-linguistic Extraction of Genetic Information using Generative Large Language Models

Milindi Kodikara
School of Computing Technologies
RMIT University
Melbourne, Victoria, Australia
s3667779@student.rmit.edu.au

Karin Verspoor
School of Computing Technologies
RMIT University
Melbourne, Victoria, Australia
karin.verspoor@rmit.edu.au

ABSTRACT

There is a persistent need for organised genetic information to support advancements in scientific discovery and personalised healthcare. Typically, this organisation process involves extraction and storage of key entities and their relationships from vast amounts of biomedical literature into databases by biocurators. This is an arduous and time consuming manual task. With the exponential growth of literature, efforts have been directed towards automating this process with natural language processing. Recently, methods based on generative AI have shown promising results for biomedical information extraction (IE) [1, 2, 3].

This project aims to explore the effectiveness of a general-domain generative large language model (LLM) on IE within the domain of genetics. We specifically focus here on the challenge of Named Entity Recognition (NER) of genes, genomic variants, and associated diseases from Spanish-language scientific text, in the context of the GenoVarDis competition which is a part of the IberLEF 2024 campaign. This competition provides an annotated dataset in the Spanish language, with which the NER task can be explored.

Our key aim in this work is to explore the effectiveness of utilising a general, primarily English-language, LLM for IE in this challenge. We explore the impact of providing English-language instructions with the Spanish-language target text (cross-linguistic setting) cf. matching the instruction language to the text, and experiment with a range of instruction strategies, including zero-shot and few-shot prompting.

Our method involves the creation of an end-to-end IE pipeline with a manually crafted library of prompts. These are sent to a generative LLM via an API (GPT-3.5) to perform the IE tasks, and outputs are post-processed to conform to the required format. Performance of the system is measured by comparison with a gold standard dataset. Industry standard metrics, namely Precision, Recall, and the F1 score, are used to evaluate the performance. Preliminary tests resulted in a F1 score of 0.5 for instructions written in English for each Spanish target text, under the one-shot paradigm. Further experiments are underway to improve these results.

We hope to gain insight into the utility of generative LLMs for this highly specialised biological IE task, with the additional complexity of addressing it in a language with substantially more limited use in the scientific literature.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
BIOKDD'24, August, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s).

CCS CONCEPTS

• Applied computing → Life and medical sciences;
• Computing methodologies → Artificial intelligence;

KEYWORDS

Natural Language Processing, Generative large language models, Computational Biology, Generative AI, Bioinformatics

ACM Reference Format:

Milindi Kodikara and Karin Verspoor. 2024. Effectiveness of Cross-Linguistic Extraction of Genetic Information using Generative Large Language Models. In *Proceedings 23rd International Workshop on Data Mining in Bioinformatics (BIOKDD 2024). Barcelona, Spain, 1 page*

REFERENCES

[1] Goel, A. et al. 2023. LLMs Accelerate Annotation for Medical Information Extraction. In *Proc 3rd Machine Learning for Health Symposium*, PMLR 225:82-100.

[2] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A.S., Ceder, G., Persson, K.A., & Jain, A. 2024. Structured information extraction from scientific text with large language models. *Nature Communications* 15, 1418 (2024).

[3] Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D.A. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of EMNLP 2022*, p.1998–2022, ACL, Abu Dhabi, UAE.