

LUCÍA SANZ-VALDIVIESO AND BELÉN LÓPEZ-ARROYO

Human vs. ChatGPT corpus annotation: Data augmentation using LLM fine-tuning

1. Introduction

The relevance of training and fine-tuning datasets to develop text production tools seems apparent, so curation and synthesis of linguistic data are now popular means to obtain quality material. Manual approaches seem to have become outdated, and some authors defend that technology has “paved the way for automation using AI techniques and tools to analyse and extract information from documents, trying to emulate what human beings are capable of doing with a limited volume of text data” (Pillai & Tedesco 2024: 3).

In this chapter, we explore the potential of Pretrained Large Language Models (PTMs) to serve as research and data enhancement tools towards the future development of multilingual text production applications. We aim to test ChatGPT’s ability to provide automatically annotated high-quality data compared to human manual annotation. The focus of the experiment is on rhetorical annotation for two reasons: first, data availability, since the CLANES project already had manually annotated comparable data; and second, the close relationship between rhetorical annotation and other Natural Language Processing (NLP) tasks of interest, such as Named Entity Recognition or sentiment analysis, which are complex semantic tasks that cannot rely on language form exclusively. The result could be fine-tuned datasets for the subsequent development of specialised neural tools for language generation. Our long-term purpose is to leverage low-data availability contexts and the undeniable need for language tools in all sectors of human activity.