Interpreting Readability Evaluation with Explainable AI

Nouran Khallaf, Centre for Translation, Interpreting and Localisation Studies, School of Languages, Cultures and Societies, University of Leeds- N.khallaf@leeds.ac.uk

Project Overview

This project investigates how explainability methods and LLM-based evaluation can be used as an evaluation framework for readability and text simplification systems. The work requires **HPC resources** because it involves running transformer models, attribution methods, and LLM-based analyses over large aligned corpora with repeated GPU-based inference. Although LLMs and BERT classifiers are increasingly used to analyse sentence complexity, it is often unclear which linguistic signals they rely on. Rather than assessing models only through metrics such as accuracy or F1 score, the project examines whether their explanations align with linguistic transformations observed in human-authored simplifications. Using aligned corpora such as **Wikipedia–Vikidia**, the project will investigate how and why models classify sentences as complex, identify **complexity triggers,** the tokens or phrasal structures that make a sentence difficult, and test whether model explanations correspond to human simplification operations. By combining explainability techniques, alignment analysis, simplification strategy annotations, and **LLM-as-judge evaluation**, the project aims to develop interpretable evaluation methods for complexity models and assess whether predicted complexity or simplification decisions are linguistically justified.

Research Questions: The project addresses three main questions:

1. Which tokens or linguistic features drive model predictions of sentence complexity?
2. Do explanation methods identify the same complexity triggers that human editors modify during simplification?
3. Can LLM-as-judge evaluation help identify errors or biases in readability predictions?

Data and Methods: The project will use the **Wikipedia–Vikidia** and **iDEM** corpora, which provide aligned complex and simplified texts, word-level simplification links, and strategy annotations across several languages. Experiments will run on **HPC resources** using open-source NLP tools including **Hugging Face Transformers, PyTorch, Captum, and spaCy**. The analysis will combine attribution methods such as **Integrated Gradients**, **faithfulness evaluation**, **alignment- and strategy-based analyses** comparing model explanations with human simplification operations, and **LLM-as-judge evaluation** using structured prompts and scoring criteria.

Expected Outputs: The project will produce an explainability- and LLM-based evaluation framework, an analysis of complexity triggers in readability prediction, and a reproducible research codebase with documentation, supporting interpretable and reproducible NLP research using HPC resources.