Embedding Models for Addressing Linguistic Challenges

Mentor(s): Sheng Wong and Scott Barnett

Background and Problem Statement

State-of-the-art text (SOTA) embedding models on Massive Text Embedding Benchmark (MTEB) show promising results on various tasks such as retrieval, classification and pair classification. However, current SOTA embedding models still have problems handling and identifying complex linguistic challenges such as negation and spatial awareness, as shown in our earlier evaluation study. These limitations stem from the models' inability to effectively capture contextual information or perform reasoning on relationships between entities within a sentence. This limitation significantly impacts the models' performance in real-world applications where a nuanced understanding of language is crucial.

Aim

To enhance the capacity of SOTA text embedding models to effectively process and interpret complex linguistic phenomena, particularly in areas such as negation, and spatial awareness.

Objectives

1. Conduct a comprehensive literature review and data survey to identify and collect existing datasets relevant to complex linguistic tasks.
2. Develop and implement a methodology for generating high-quality, task-specific data using LLMs for each linguistic phenomenon such as spatial, hierarchical or temporal.
3. Evaluate existing embedding models on a larger dataset for each linguistic task.
4. Implement a fine-tuned pipeline for existing embedding models to handle multiple linguistic tasks.

Expected Outcome

1. Comprehensive evaluation results of embedding models on complex linguistic tasks using an expanded dataset.
2. A set of fine-tuned embedding models demonstrating improvements in handling complex linguistic phenomena, compared to baseline performance
3. A presentation showing the final results of the work and a tutorial on how to use the scripts created as part of this work.
4. Scripts for data generation, evaluation and fine-tuning pipeline for embedding models.

Technical/ Skills

- Strong Python programming skills
- Knowledge of NLP and Deep Learning
- LLMs Prompting will be helpful

Resources

X2 unpaid internships.