LM-based Entity Linking Short Report

Yung-Ching Yang

1 Introduction

In this assignment, I choose two approaches to demo the given entity linking task. The first one is to integrate vector database with OpenAI's generative model, and the second approach is to fine tune BERT (Bidirectional Encoder Representations from Transformers) model for token classification. The reason of choosing these two approaches is to compare the the performance between a more state-of-the-art AI technique and a more conventional NLP method. Configuring vector database with OpenAI's generative model has the advantage of dynamic entity recognition, and it can be used on other vector search or generative search. The real-time data processing and adaptability to diverse data sets make it more flexible. However, this method could be constrained by its reliance on the model's accuracy and prompt design. On the other hand, the second approach, fine-tuning a BERT model for token classification, is grounded in established NLP practices. It promises stability and reliability, especially in handling structured tasks.

2 Challenges

2.1 Approach 1: Vector Database Integration with OpenAI's Generative Model

The main challenge of using the first approach is that this approach highly relies on GPT-3.5 for extracting company names introduces a dependency on the model's accuracy and the effectiveness of the prompt design. Inaccuracies of the prompt here could lead to incomplete or incorrect annotations, and the limitation of token count in article processing may lead to incomplete analysis of articles, potentially missing out on key information.

2.2 Approach 2: Fine-tuning BERT for Token Classification

One of the biggest challenges lies in the data augmentation part, wherein the GPT-3.5 chat completion API was utilized to generate mock articles as JSON objects. The quality of the generated data significantly depends on both the model and the prompt design, leading to variable and unstable outputs. As a result, articles with incorrect formatting and company annotations were omitted, culminating in a dataset comprising 70,925 sentences. The insufficiency of the data may also increase the risk of overfitting. The second limitation of this approach is that fine-tuning encoders with one layer or single task is less effective especially when the data is imbalanced, such as the varying frequency of company names in articles. This can be improved by adding sub classification tasks and use weighted losses. Although this approach may be more time and effort inefficient, particularly in preprocessing and fine-tuning phases, and may be highly task-specific, it affords increased flexibility in model customization. With adequate training data and sufficient time, it is assumed that this strategy could have more stability in entity linking tasks.