Comparative Analysis of Models with Manually Created Knowledge Graphs

Group 9: Mayank Tamakuwala, Nithin Bhat, Agasti Mhatre

1. Description

   ➔ Identify the optimal POS algorithm for tagging news articles. We aim to validate this by manually generating a knowledge graph through the dataset, then training and testing each algorithm/model to produce the same knowledge graph on the same dataset, allowing us to compare the performance of the models in relation to ground truth.

2. Dataset

   ➔ https://www.globenewswire.com/newsroom. We will manually review 300 news articles to label and annotate them. The dataset includes 207 different categories of articles, so we will likely select 300 news articles from various categories to diversify our samples.

3. Methodology and Expected Results:
   ➔ The major external tools we plan to use include Scikit-Learn, spaCy (for NLP), Transformers, Matplotlib, and Seaborn for visualization. We are currently deciding on the models that can generate the knowledge graphs for us. These models will likely include the Viterbi POS Algorithm and BERT variants like ALBERT or RoBERTa to perform predictions and compare their training times and performance.

   ➔ The main results we want to have, are the labels to the words showing the annotated pairs for an inputted dataset.

4. Timeline:
   ➔ Week 1 and Week 2: Finish manually labeling 300 news articles to create a ground truth and train/testing datasets for the model.

- ➔ Week 3: Train and test different models and finetune/optimize their performance.
- ➔ Week 4: Create and compare the performance visualizations for the different models.

5. Responsibilities:
   - ➔ Our team has three members: Mayank Tamakuwala, Nithin Bhat, and Agasti Mhatre. Each member is going to handle around 100 news articles to create the knowledge graph dataset manually. Then, all of us are going to handle the creation of at least one algorithm that generates a knowledge graph and then the evaluation of those models.