2022

NLP@VCU: Crop Characteristic Extraction Framework

Cora Lewis
Macalester College

Bridget McInnes
Virginia Commonwealth University

Getiria Onsongo

Cora Lewis
Bridget McInnes
Getiria Onsongo
Department of Computer Science - Natural Language Processing Lab

NLP@VCU: Crop Characteristic Extraction Framework

Introduction

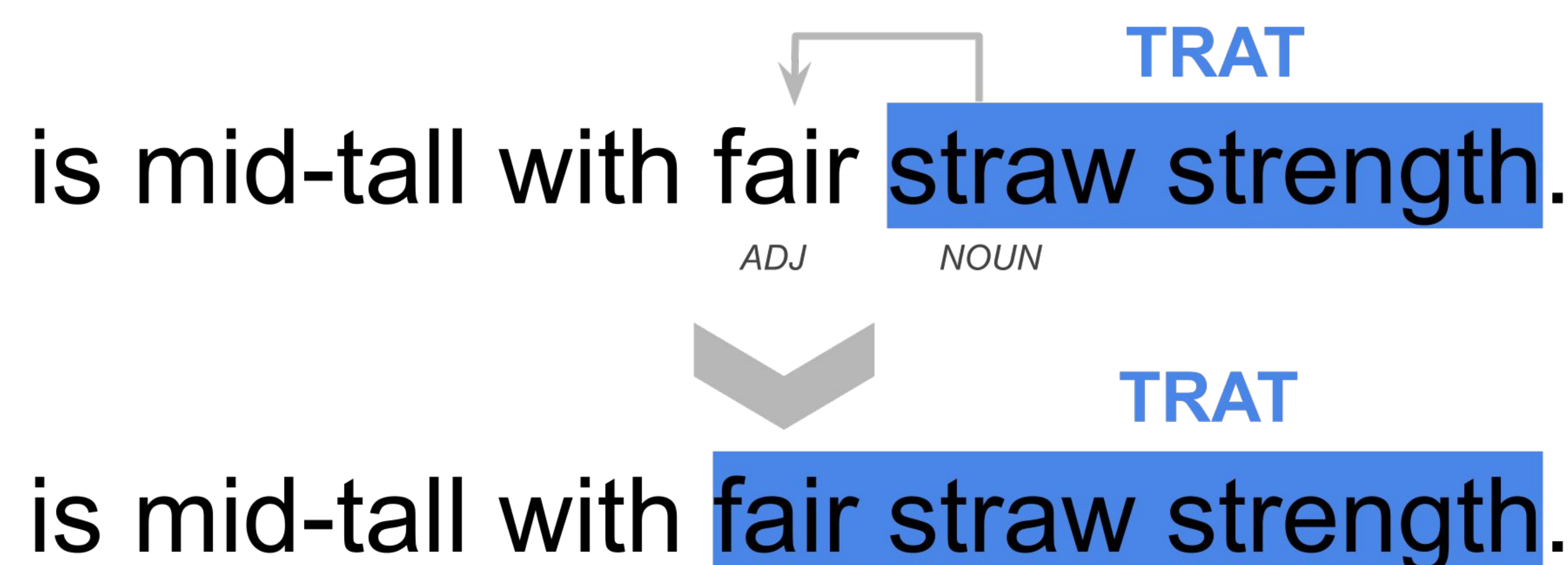IaaAgDataNER is an annotation tool for named entity recognition. A custom trained SpaCy model suggests entities to the annotators, and our aim was to build out that model's infrastructure and improve its predictive abilities.
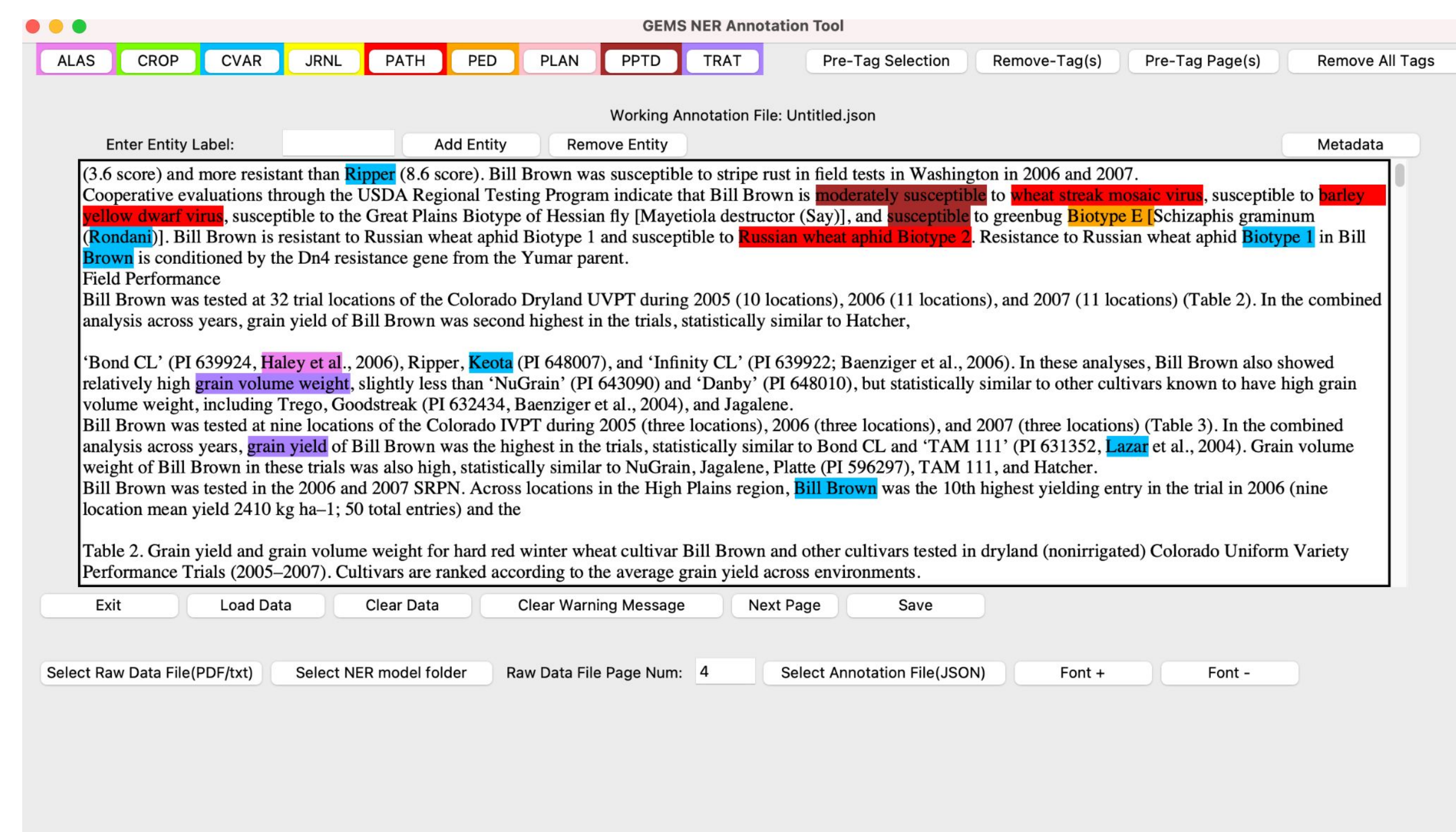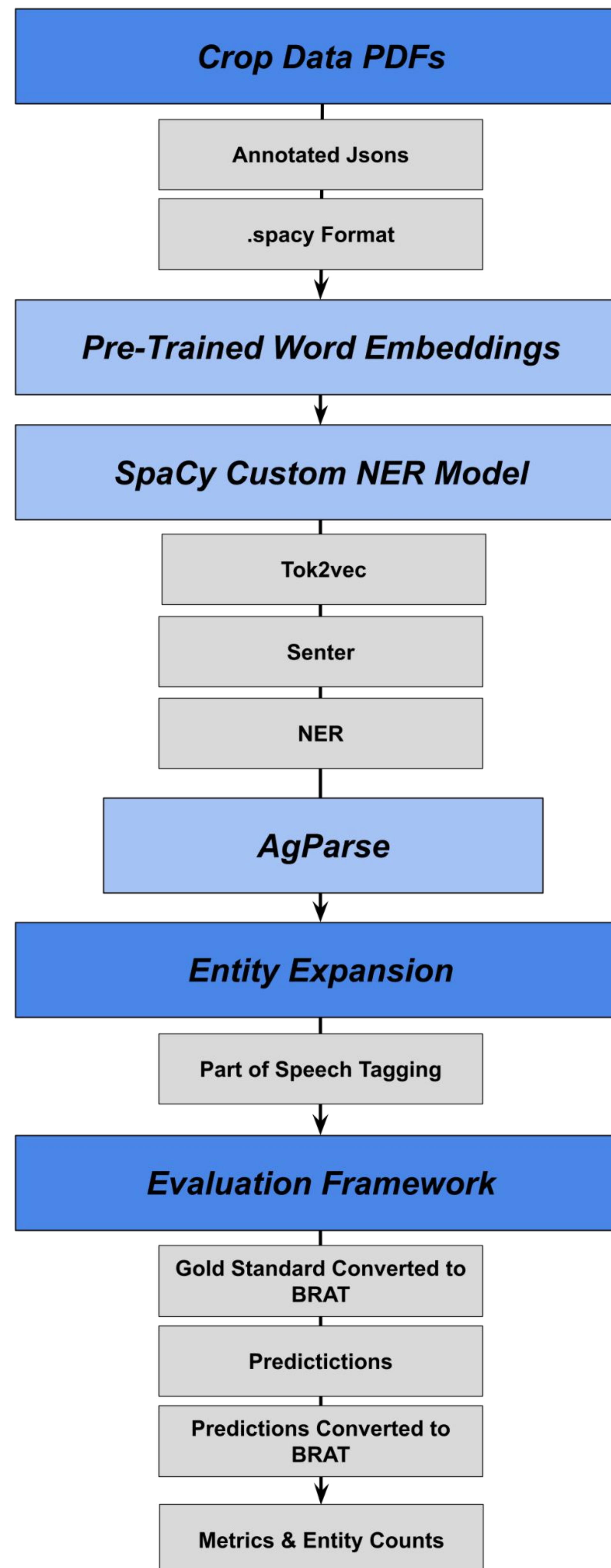
Entity Expansion

The model would often miss adjectives describing nouns, so using SpaCy's part of speech tagger, we created a post-processing entity expansion step.
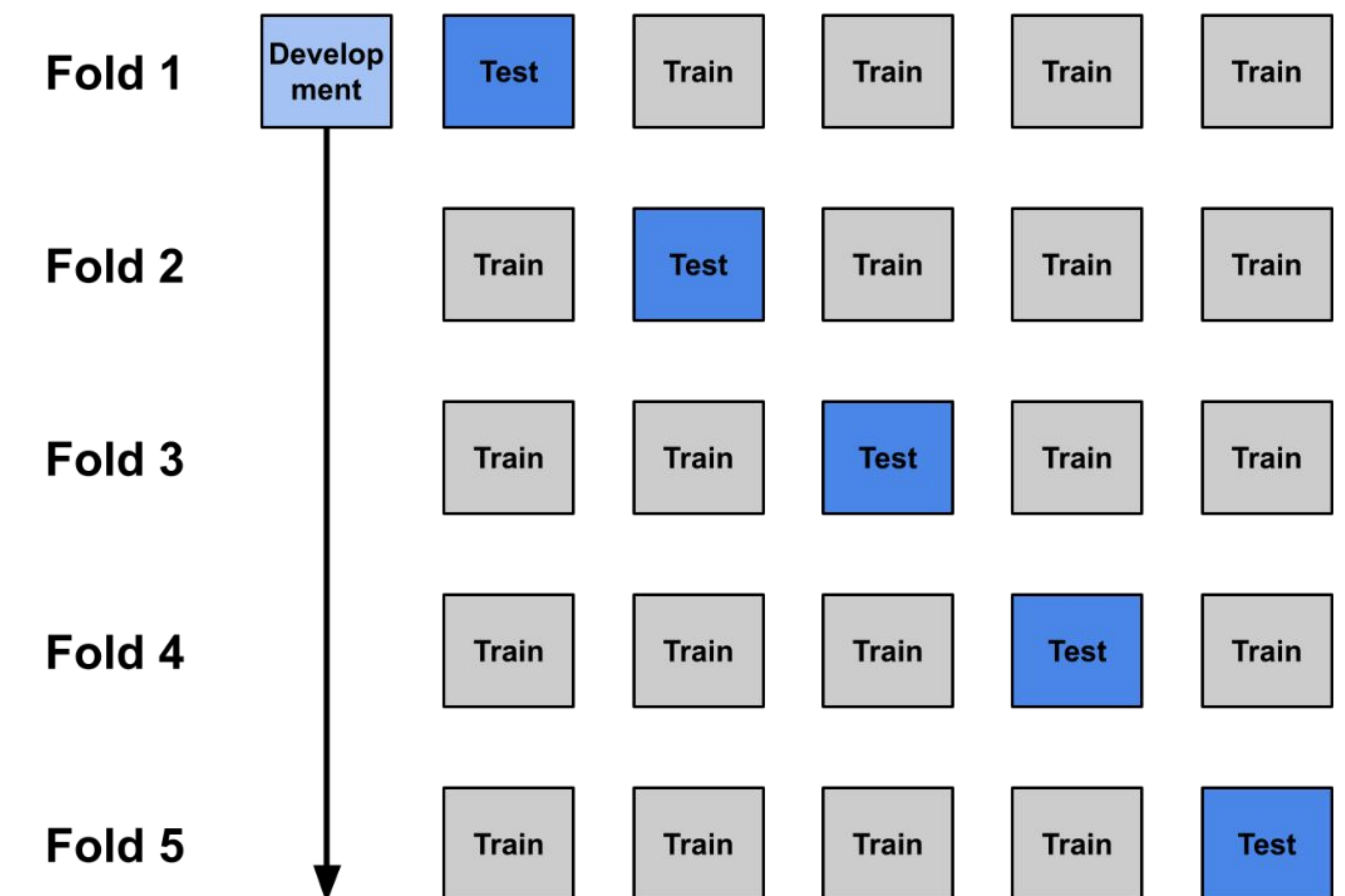
TRAT

is mid-tall with fair straw strength.

ADJ   *NOUN*

TRAT

is mid-tall with fair straw strength.

Model

- **Crop Data PDFs**
- Annotated Jsons
- .spacy Format
- **Pre-Trained Word Embeddings**
- **SpaCy Custom NER Model**
  - Tok2vec
  - Senter
  - NER
- **AgParse**
- **Entity Expansion**
  - Part of Speech Tagging
- **Evaluation Framework**
  - Gold Standard Converted to BRAT
  - Predicitions
  - Predictions Converted to BRAT
  - Metrics & Entity Counts

Evaluation Framework

We implemented a 5-fold cross validation wrapper for SpaCy custom model training to take full advantage of the small dataset

Fold 1	Development	Test	Train	Train	Train	Train
Fold 2		Train	Test	Train	Train	Train
Fold 3		Train	Train	Test	Train	Train
Fold 4		Train	Train	Train	Test	Train
Fold 5		Train	Train	Train	Train	Test

Conclusion

The preliminary results show that using pre-trained word embeddings and entity expansion improves the custom spaCy model. Glove 6B trained on Wikipedia and Gigaword appears to be the best pre-trained word embedding for the model. However, more data is required to fully analyze the system. Future work includes:

- Annotate more data to evaluate model
- Consider more post-processing steps
- Configure a CRF layer