Named Entity Recognition

Aayush Birla, Himanshu Kandpal, Srish Kulkarni

Named Entity Recognition (NER) is

one of the most popular applications of Natural Language Processing. In astronomy alone, more than 41,000 new articles are published every year and the vast majority are available either via an open-access model or via pre-print services. It is difficult to navigate in this ocean of data.

NER helps us extract key information from scientific papers which can help search engines to better select and filter articles.

Data and Labels

Datasets with text fragments from astrophysics papers, provided by the NASA Astrophysical Data System with manually tagged astronomical facilities and other entities of interest.

Datasets[1] are in JSON Lines format (each line is a json dictionary). "tokens": the list of tokens (strings) that form the text of this sample. Must be included in the predictions.

"ner_tags": the list of NER tags (in IOB2 format)

References

[1]https://huggingface.co/datasets/fgrezes/ WIESP2022-NER

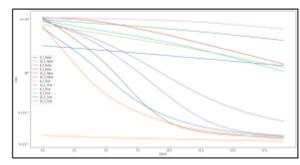
Model

1] Tokenization

We did multiple forms of tokenization like training word embeddings and using the keras tokenizer for our problem.

2] Modelling

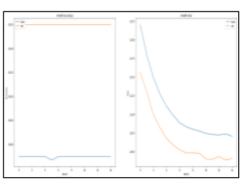
We tried multiple models using SimpleRNN, LSTM and GRU layers from keras and plot their training curves. We used SOTA models like DistilBERT & BERT to predict the entities after customizing the relevant layers.



3] Web-App

We created a web-based application that would take sentences as input and display the entities present.

Results





Conclusion and Future Work

- We can use this model to optimize search engines as we can run a NER model on the articles once and store the entities associated with them permanently and use those entities to search for articles.
- NER can be used in developing algorithms for recommender systems that make suggestions based on our search history or on our present activity. This is achieved by extracting the entities associated with the content in our history or previous activity and comparing them with the label assigned to other unseen content.