Natural Language Processing

Assignment 1 - Aryan Yadav & Samvit Jatia

GITHUB REPOSITORY

Section A

Q3) Analyze the NER and POS performances of the three models. You should record the following insights:

(a) The comparative strengths of each model

(b) The comparative weaknesses of each model

Note down any patterns you observe in the relative strengths or weaknesses. (20 marks)

Answer 3)

To compare spaCy, NLTK, and Stanford NER based tasks of POS tagging and NER, lets look at the strengths and weaknesses of each model.

(a) Comparative Strengths of Each Model

spaCy - Overall strength - spaCy performs particularly well due to its efficiency and processing speed, making it highly suitable for applications requiring rapid text analysis across large datasets. It was able to parse through the entire dataset at a very good speed making it a good candidate for applying in practical applications.

NER performance - With an accuracy of 0.76, spaCy has an above average NER capability.
POS tagging - With a POS tagging accuracy of 0.53, spaCY indicates weaker performance compared to other models.

NLTK - NLTK excels in detailed analysis, reflected in its excellent performances in both NER and POS tagging. Its strength lies in the precision and depth of analysis, making it ideal for academic and research applications where detailed insights are necessary.
NER performance - With an accuracy of 0.83 , NLTK is highly effective at named entity recognition
POS tagging - NLTK shows the highest accuracy for POS tagging with an accuracy score of 0.958. Highlighting its incredible capability to understand and analyze complex sentences.

StanfordNER - Built with the backing of advanced machine learning algorithms , Stanford ner demonstrates excellent performance in linguistic modeling and excels in tasks requiring high accuracy.
NER performance - With an NER accuracy of 0.83, Stanford NER matches NLTK in its ability to accurately identify and classify named entities across various text types and domains.

POS tagging - The POS tagging accuracy of 0.95 nearly matches that of NLTK, indicating that Stanford NER is also highly reliable for POS tagging

(b) The comparative weaknesses of each model

spaCy - The most notable weakness for spaCy is its POS tagging performance, with an accuracy of 0.5392. This suggests that spaCy's algorithms or training data may not be as finely tuned for the granular details of POS tagging as they are for other tasks.

NLTK - Although NLTK excels in performance quality, its most glaring weakness lies in its processing efficiency and speed, especially when handling large datasets or in real-time applications. Its strength in accuracy might come at the cost of speed, making it less ideal for time-sensitive or large-scale processing tasks.

StanfordNER - similar to NLTK, Stanford NER's processing speed can be a weakness, especially in comparison to spaCy. In addition to that as it is also not very user friendly, as space and NLTK have installable python libraries which can be easily integrated, however integration of Stanford is a lot more complex and requires external installation.

Patterns Noticed:

The major pattern seen across comparing all the models was primarily the Tradeoff between Speed and Quality. A common pattern among the models is the trade-off between processing speed and the granularity of analysis. spaCy, while fast and efficient, shows lower accuracy in POS tagging, indicating that its streamlined processing may come at the cost of detailed linguistic precision. In contrast, NLTK and Stanford NER, which provide high accuracy in linguistic tasks, do so with potentially slower processing speeds.

Section B

Q5) Run the above POS and NER models on the newly created dataset. Discuss the following:
(a) Are the levels of performance the same as those recorded earlier?
(b) Discuss the results in detail, providing results to justify your observations.
(10 marks)

Answer 5)
(a) Levels of Performance Comparison
The performance levels of the Part-of-Speech (POS) tagging and Named Entity Recognition (NER) models from the spaCy, NLTK, and StanfordNER libraries show significant variation when applied to the provided dataset compared to our newly created dataset. Here's a brief comparison:

- spaCy:
    - POS Tagging: The accuracy dropped from 0.5392 on the given dataset to 0.3746 on our dataset.
    - NER: The accuracy slightly decreased from 0.7683 to 0.7488.
- NLTK:
    - POS Tagging: There was a drastic decrease in accuracy from 0.9589 on the professor's dataset to 0.36 on our dataset.
    - NER: Accuracy decreased marginally from 0.83 to 0.77.

- StanfordNER:
    - POS Tagging: The accuracy saw a significant decrease from 0.95 on the professor's dataset to 0.33 on our dataset.
    - NER: The accuracy remained constant at 0.77.

From these observations, it's evident that the levels of performance are not the same as those recorded earlier, with a noticeable decrease in POS tagging accuracy across all three libraries on our dataset, while NER accuracy shows less variation.

(b) Detailed Discussion and Justification of Observations

- spaCy:

    - POS Tagging: The substantial decrease in accuracy could be attributed to the smaller size and potentially unique linguistic features present in our dataset that spaCy's model, trained on more generalized data, struggled to accurately interpret. Even though the model's accuracy on the given dataset was also quite low which could also be because of how we have programmed our code to implement this model.
    - NER: The slight decrease in accuracy suggests that spaCy's NER model is somewhat robust to the changes in our dataset, although it still faced challenges possibly due to similar reasons mentioned above.
- NLTK:
    - POS Tagging: The drastic drop in accuracy for POS tagging with NLTK is notable. It's possible that our dataset contained linguistic structures or annotations that did not align well with NLTK's expectations, leading to poor performance.
    - NER: The marginal decrease in NER accuracy with NLTK indicates a somewhat stable performance, suggesting that while NLTK's NER capabilities are not as affected by the dataset's peculiarities as its POS tagging, there is still an impact.
- StanfordNER:
    - POS Tagging: The significant decrease in accuracy for StanfordNER's POS tagging could indicate that the model may not generalize well to datasets with unique or less common linguistic features, similar to spaCy.
    - NER: The consistency in NER accuracy is interesting and suggests that StanfordNER's NER model may be particularly well-suited to handling a variety of data types and annotations without a significant loss in performance.

The differences in performance across both datasets for each library can be attributed to several factors:

- Size and Diversity of Data: Our dataset's small size and potentially less diverse linguistic features compared to the professor's dataset could impact the models' ability to accurately predict POS and NER tags.
- Annotation Quality: Since we annotated our dataset manually using Doccano, there might be inconsistencies, biases, or errors in our annotations compared to professionally annotated datasets.
- Difference In Tagging: POS tags tend to be more detailed as compared to NER tags as if we notice in any sentence annotated using NER tags we will see a lot of 'O' tags.