NER Model Evaluation and Asessment using spaCy

WW – ITE Elective

Shan Hiro Rosario   
Computer Studies and Engineering  
Mandaluyong City, Philippines   
shanhiro.rosario@my.jru.edu

***Abstract*—Named Entity Recognition (NER) classifies spans of texts, words or sentences into its labeled categories based on the NER Model such as Person, ORG, GPE and etc. For further improvement on NLP methodology and approach for the related practitioners, an evaluation and assessment are made to provide an emphasis on the task of implementing NER using spaCy in regards to future Natural Language Processing pipeline and provide insights into the preparation of a dataset and the methodology itself**

***Keywords*—Natural Language Processing, Named Entity Recognition, spaCy, NER Evaluation.**

# **Introduction**

Natural Language Processing involves many fields, its broad and interdisciplinary industry spans several complex domain, such as Sentiment Analysis, Multimodal NLP, Conversational AI, Speech Processing and many more. The ability to handle such advanced studies and its practical applications requires the practitioner of the field to enhance their understanding of foundational knowledge in the field—In the context of NLP, these are the text representation methods such as Bag of Words, TF-IDF and Word Embeddings. However, these techniques does not directly understand the context and nuances behind the formation of a sentence or each words in the dataset provided for the NLP Task. Named Entity Recognition then, becomes the foundation to identify and classify certain words and texts into predefined categories built within the model to provide the foundational information-extraction ability for enabling the analysis of large unstructured volumes of social-media using the approach of sentiment analysis where such large amounts of unstructured data are processed within the NER model using accurate entity extraction to improve NLP capabilities on later pipelines.

1. **Methodology**

The proponent of the study utilizes three sample data in the domain of Book Reviews or Literary Analysis of specifically “To Kill a Mockingbird”, “1984” & “Pride and Prejudice”. Due to the The automatic labeling of spaCy’s NER model, these book reviews then, becomes an excellent platform to evaluate and assess its capabilities on determining the entities automatically, of which its accuracy, precision, recall, and f1 are presented as the evaluation metrics.

A close-up of a text

AI-generated content may be incorrect.

Figure 1. Sample Review

The following figure above shows the sentence or text of the first review sample as there are a total of 3. This review represents both information on the book itself and external information and geographical places such as the State of Alabama. Showcasing excellent mix of fictional/literary and realistic settings, locations, geopolitical entity to analyze the accuracy of SpaCy / NER label’s ability to automatically determine the named entities in the context of the review.

A screenshot of a phone

AI-generated content may be incorrect.

Figure 2. Gold Standard

A gold standard are created is then manually created to provide a ground truth for model comparison, specifically considering consistency and standardization where, each entity within the default labels of spaCy are tagged manually and accurately defined in order to serve as the benchmark for evaluation. All three samples are provided with a correct label for its named entities.

A close-up of a list

AI-generated content may be incorrect.

Figure 3. SpaCy Annotation

The Figure 3 above shows the results of spaCy’s automatic labeling of the named entities of the three sample data provided. It must be noted that the study used the “en\_core\_web\_sm” library for these available NER Categories and was not changed nor used a custom label for the application of the Gold Standard benchmark for consistency.

A screenshot of a computer code

AI-generated content may be incorrect.

Figure 4. Python Codeblock

The python code used in the study are provided by the instructor in the classwork modules. The execution depended on showcasing the available NER Labels and entering a text to process and detect the named entities of the input. This is the basis for the SpaCy Annotation in the figure 3 above.

# **Results**

We must first discuss our findings in spaCy’s NER Labeling and this can be seen in the figure 5 below.

A computer screen shot of a computer code

AI-generated content may be incorrect.

Figure 5. Sample 1 spaCy

The results shown in the figure 5 above are the text that are automatically detected by the library of spaCy itself, including its label and explanation of what each labels mean.

It must be noted that the title itself “To Kill A Mockingbird” which are explicitly written in the review are missing within the output completely missing the context of the sentence that the nature in itself is a book-review. This is in addition to the duplication and misidentified Atticus Finch which is a fictional character in the book and the Jim Crow are supposed to be NORP.

A white text on a black background

AI-generated content may be incorrect.

Figure 6. Sample 2 spaCy

The entities of British, Orwell, Newspeak, Twitter, and Oxford are incorrectly placed as they are either inaccurate or missing a the complete words such as the British Library are the complete text and it would be classified as an Institution, then an ORG label by principle, in addition to Orwell Newspeak being a fictitious language in the book.

A computer screen shot of a white text

AI-generated content may be incorrect.

Figure 7. Sample 3 spaCy

The exact text-matching of label of the Joe Wright’s has an excess letter of ‘s’, that gives false insights in preparation for the NLP task later on the pipeline, therefore it should be considered to further preprocess this mistake and provide appropriate removals of excess tokens that ruins the analysis of the sentiment. This is in addition to inaccurate labels of Chatsworth House, Longbourn, and Regency.

The correct labels for the whole texts of the three sample reviews are provided in the excel spreadsheet alongside this document report. It includes the Gold Standard, SpaCy’s labels, raw Data and Evaluation Metrics.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 8. Evaluation

The figure 8 above shows the results of accuracy, precision, recall, and f1 on each sample reviews in addition to the overall evaluations of the whole dataset. We can closely observe that the highest accuracy between the total correctly classified words divided by total words in dataset is the sample 1 whereas it has an overall 65% accuracy in comparison to the lowest amount of 35% accuracy in the third sample. However, these are all determined using the gold standards provided by manually tagging each named entities of the sentence and comparing it to spaCy’s automatic labeling.

##### **V. Conclusion**

The results has shown that spaCy’s default labeling regarding the context of reviews or comments that consists of fictitious and non fictitious entities proves to be unreliable with low evaluations of less than 50% accuracy in average using the en\_core\_web\_sm model. This low performance may be due to the domain-specific factors such a the reviewer’s creative language, prose that are either ambiguous or generic, contains inside information in the book such as quoted excerpts that are literary-specific or reviewer slang including but not limited to references among multiple book titles, movies, actors, and acronyms that are not readily available within the features of spaCy.

Despite the poor NER accuracy, it must be noted that the application of these literary reviews proves to be useful in sentiment analysis, ratings, and opinions, and a wide amount of information to be gained on the potential ofdata-driven results. To ensure reliable and highly accurate results on such domain, it is necessary to manually handle the semantics between each inaccurate contexts that are captured by spaCy or provide custom NER labels for further improvement. In addition to pretrained models by adding custom and domain-specific terms, labels, and fine tuning based on historical reviews, comments or data in the domain.

‌