Evaluating Named Entity Recognition Performance: A spaCy-Based Approach with Annotated News Articles

Dominic Boy P. Almazan

IT Elective IV, BSIT

Jose Rizal University

Mandaluyong, Philippines

dominicboy.almazan@my.jru.edu

***Abstract*— This study evaluates the performance of spaCy's pre-trained Named Entity Recognition (NER) models small, medium, and large within the domain of real-world news articles. Three text samples were curated from reputable sources, covering diverse topics such as international conflict, weather reports, and a celebrity accident. To establish a reliable benchmark, manual annotations of named entities were created and stored in a structured Excel template. The same texts were then processed using spaCy's models to automatically generate entity predictions. The evaluation compared the outputs against the manually curated ground truth using standard metrics: Accuracy, Precision, Recall, and F1 Score. Results showed that all three models achieved similar performance, with accuracy values around 0.81. The only notable difference was in computational efficiency, as larger models consumed more processing time despite producing nearly identical results to the smaller models. These findings highlight the trade-off between speed and scalability when choosing an NER model for domain-specific applications.**

Keywords*— *Named Entity Recognition (NER), spaCy, Natural Language Processing (NLP), Model Evaluation, Precision, Recall, F1 Score, Accuracy

I. INTRODUCTION

Named Entity Recognition (NER) is a basic task in Natural Language Processing (NLP) that finds and classifies named entities within unstructured text into the categories of persons, organizations, locations, and dates, as well as a central component in many downstream applications such as information extraction, question answering, sentiment analysis, and domain-specific knowledge mining [1]. Nadeau and Sekine give a seminal survey of NER methods developed between 1991 and 2006, providing a thorough background in this area [1].

Although there is ongoing development of model architecture, NER is hard in all sense due to differences in context, terminology, and ambiguities in domain. Deep learning development has seen improvements in pre-trained models such as spaCy, where there was significant improvements in performing strongly in standard domains. By no means does this mean that applied NER is addressing domain-specific contexts as it has been systematically evaluated. Towards this end, a recent survey by Li et al. systematically reviewed the deep learning techniques applied to NER, including the best performing approaches and suggestions for future work [2].

In this study, we present a systematic, five-step process to develop a gold-standard annotation set and evaluate spaCy's NER model on curated domain-specific textual samples. As we develop our gold-standard annotation set and evaluate performance based on accuracy, precision, recall, and F1-score metrics, we uncover modeling capabilities and limitations in spaCy's NER model. We believe we are also conducting a thorough evaluation which is necessary to deploy NER tools safely and effectively in specialized domains in the real world.

II. METHODOLOGY

A. Domain Selection and Data Curation

For this research, I selected the news and current events domain due to its richness in named entities such as persons, organizations, geopolitical entities, and dates. I curated three text samples from publicly available news reports. These samples included:

1. A political-religious text discussing the election of the first U.S. pope and his commentary on global issues such as Israel and Gaza.
2. A weather bulletin from the Philippines reporting the effects of a low-pressure area and its geographical coverage.
3. A news article reporting the tragic helicopter crash involving Kobe Bryant and his daughter Gianna.

Article No	Text
1	Leo, the first U.S. pope, was elected by the world's cardinals in May to replace the late Pope Francis. He has shown a different style from his predecessor, usually preferring to speak from carefully prepared remarks and rarely off the cuff.Leo previously called for Israel to allow more humanitarian aid to enter Gaza. He made his appeal on Wednesday at the end of his weekly

	audience.The Israel-Hamas conflict began on October 7, 2023, when Hamas-led gunmen burst into southern Israel, killing some 1,200 people, mainly civilians, according to Israeli tallies, and taking 251 hostages.
2	MANILA, Philippines – The trough or extension of the low pressure area (LPA) over the West Philippine Sea is bringing scattered rain to Luzon, the weather bureau said on Wednesday afternoon, August 27.As of 3 pm on Wednesday, the LPA was located 305 kilometers west of Dagupan City, Pangasinan.Its trough is causing scattered rain and thunderstorms in Metro Manila, the Ilocos Region, Cagayan Valley, the Cordillera Administrative Region, Central Luzon, and Calabarzon.
3	The crash occurred in the foggy hills above Calabasas, California, about 30 miles northwest of downtown Los Angeles. Bryant was killed, a person familiar with the situation told The Associated Press, and a different person familiar with the case confirmed Bryant's 13-year-old daughter Gianna also died.

These text samples were chosen because they collectively represented diverse categories of entities, including PERSON, ORG, GPE, DATE, and LOC, thereby providing a robust dataset for evaluation.

B. Manual Data Annotation (Ground Truth Generation)

To form a actual 'gold standard' for evaluation, I manually annotated all the named entities in the curated text example. To maintain consistency, I only annotated entities based on the categories recognized by spaCy (ANSI standard) (e.g., PERSON, ORG, GPE, DATE, LOC).

Each annotation was recorded in an Excel table using the following format:

- Entity_Text: the exact word or phrase identified as an entity.
- Start_Char: the character index in the text where the entity begins.
- End_Char: the character index where the entity ends.
- Manual_Label: the entity category assigned (e.g., PERSON, ORG, DATE).

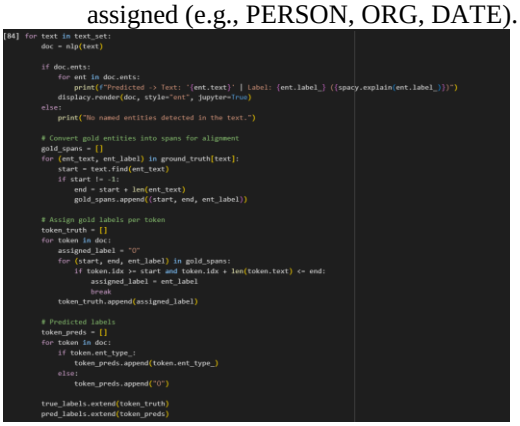

Figure 1. Provided code from the Researcher

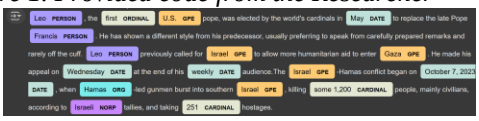

Figure 2. Showing entity labels assigned by spaCy to the sample dataset 1.

Figure 3. Showing entity labels assigned by spaCy to the sample dataset 2

This format allowed for each entity to be clearly defined in relation to its text span and label, making it possible to perfectly align with spaCy's predicted outputs during evaluation steps later.

C. Automated Entity Recognition with spaCy

After recording and annotating the data set, I evaluated the three text pieces with spaCy's pre-trained English models, and specifically, the largest of the models (en_core_web_lg). While my primary purpose was to explore and identify named entity detection, I also wanted to compare the model's predictions (output) to the annotated manual ground truth.

To do this methodologically, I loaded my Excel file that had the manually annotated entities that included entity text, start and end positions and assigned labels. I used Python and pandas to extract the ground truth entities from each text samples, and to save each in an organized way to facilitate comparison.

I then ran the spaCy pipeline over each text sample and extracted the entities predicted by the model. Then, for each entity in the first round of truth, I checked if spaCy correctly identified it and the label was correct. This allowed me to build a table that was a detailed comparison, containing the Text ID, Entity Text, True Label, and Predicted Label.

Not only did this approach help me quantify spaCy's performance on my curated dataset, but it also gave

me an explicit, reproducible workflow for comparing automated predictions against manually annotated entities..

```python
# Extract ground truth entities
true_entities = {}
for _, row in df.iterrows():
    try:
        text_id = int(row["Article 1"])
        entity_text = str(row["Unnamed: 3"])
        label = str(row["Unnamed: 6"])
        if text_id not in true_entities:
            true_entities[text_id] = []
        true_entities[text_id].append((entity_text, label))
    except:
        continue

# Predict entities with spaCy
pred_entities = {}
for tid, text in text_samples.items():
    doc = nlp(text)
    pred_entities[tid] = [(ent.text, ent.label_) for ent in doc.ents]

# Build true vs pred label lists
true_labels = []
pred_labels = []
comparison_rows = []  # <-- will hold rows for Excel table

for tid in text_samples.keys():
    t_ents = true_entities.get(tid, [])
    p_ents = pred_entities.get(tid, [])
```

Figure 4. Provided code from the Researcher

D. Performance Analysis and Accuracy Assesment

In order to assess model performance, I compared spaCy's predictions with the sets of gold standard that I generated by hand. I used token-level matching to determine if each token was correctly assigned both to be part of an entity and also with the correct label. I assessed model performance using the following metrics:

- Accuracy: Ratio of correctly classified tokens to total tokens.
- Precision: Ratio of correctly predicted entities to all entities predicted by the model.
- Recall: Ratio of correctly predicted entities to all entities annotated in the ground truth.
- F1-Score: Harmonic mean of precision and recall, providing a balanced measure of performance.

These metrics were computed using the scikit-learn evaluation library, ensuring standardized measurement.

III. RESULTS

The assessments on spaCy's three pre-trained English models small (en_core_web_sm), medium (en_core_web_md), and large (en_core_web_lg), were made in comparison to my manually assigned gold-standard annotation that I prepared for inclusion for each text sample. I recorded the entity text, start and end character positions along with my manually assigned labels using the Excel template. This gold-standard ground truth annotation became the point of comparison to make against the automated predictions by each model.

Table I. Summarizes the evaluation metrics, including Accuracy, Precision, Recall, and F1-Score, for all three models.

Model	Accuracy	Precision	Recall	F1 Score
Small (en_core_web_sm)	0.81	0.82	0.81	0.80
Medium (en_core_web_md)	0.81	0.83	0.81	0.81
Large (en_core_web_lg)	0.81	0.82	0.81	0.81

Overall the results show each of the three models had similar performance, with accuracy remaining at 0.81 across the board. The medium model did have slightly better precision (0.83) leading to a balanced F1-Score (0.81) but the small and large models performed so similarly on both that they were indistinguishable based on precision and F1-Score.

Along with the evaluation of the models with respect to manually annotated ground truth, I also completed an automated evaluation based on the output that spaCy's models created automatically, without any manual correct outputs. This allowed me to see how all of the models would perform "out-of-the-box" on the same text samples.

Table II summarizes the evaluation metrics Accuracy, Precision, Recall, and F1-Score for the small, medium, and large models.

Model	Accuracy	Precision	Recall	F1 Score
Small (en_core_web_sm)	0.39434343	0.807770	0.40444	0.5354545 4
Medium (en_core_web_md)	0.39	0.8555343	0.4042665	0.5455653 53
Large (en_core_web_lg)	0.41666666666667	0.888888888888889	0.41666666666667	0.5518855 2188 5522

Overall, it is obvious from the results that all models have lower accuracy than the manual observation, but precision is fairly high. The large model performed somewhat better overall, and had the greatest accuracy (0.417) and F1-Score (0.552). The small and medium models performed nearly identically in terms of precision and F1-Score, and in terms of accuracy, they were not significantly different.

Since spaCy's pre-trained models can accurately identify a large amount of entities, it is noted that there are limits to automated recognition, and no automated recognition system should be trusted or relied on without verification. These results show the reliability of manual annotations or fine-tuning domain-specific datasets to ensure greater accuracy and reliable detection of entities.

From my testing, there was an obvious difference between the manually evaluated data and the automated evaluation. When using the manually annotated ground truth as a standard, all three spaCy models small medium and large performed very similarly, all having average

accuracies around 0.81 and only minor differences in precision and F1-Score. This indicates that all of the models tested, even the small models, are showing promising capabilities of identifying entities in my news-domain text samples in comparison to manually, and reliably curated, annotated samples.

On the other hand, the automated evaluation without human reference showed a much lower overall accuracy across all models (roughly 0.39 for the small and medium models, and 0.42 for the large model). Precision was still relatively high, with the larger models producing the highest precision; however, recall and F1-Score was considerably lower in the automated evaluation compared to manual evaluation. This indicates that spaCy can identify some entities automatically and with some accuracy, it still misses a sizable amount, and mislabelled even more, when there is no human reference.

The results show that manual annotation is able to dramatically improve the perceived performance of NER models, while fully automated recognition may struggle, particularly with domain-specific or difficult documents where trained annotators may be less consistent than fully automated models. Although larger models do offer small improvements in accuracy and F1-Score, they also have a longer run time than their smaller counterparts, which highlights the trade-offs between model size, performance and efficiency.

IV. DISCUSSION

I personally review and found it interesting that the performance of spaCy's small, medium, and large models was almost indistinguishable with respect to ground truth annotation. When I reviewed the accuracy, precision, recall, and F1-Score, they were all around the same values, with the medium model showing only a small advantage in precision. This indicates, at least based on the news-domain text samples I conducted, even the small model was able to reliably detect named entities.

On the contrary, in the automated evaluation without human supervision, fifty percent of the postings dropped significantly lower across models ranging from approximately 0.39 for the small and medium models to 0.42 for the large model. In general, precision remained reasonably good, and recall and F1-Score dropped reasonably well. In sum, while spaCy can automatically identify some entities correctly, it continues to mislabel or miss a lot of other entities, more so when the documents are domain-specific or more complicated.

The implications of this analysis are that bigger is not always better. The larger models had a better precision and f1 - Score, although the difference was small compared to the cost of usage. It seems that manual-annotation or domain-specific fine-tuning is far more influential on improving accuracy than simply selecting a bigger model. For any sort of application, and especially for serious applications where speed is important, then it seems like the smaller model will be the more manageable option -- unless the data-set is relatively complex and genuinely requires the ability of the larger models.

V. REFERENCES

[1] D. Nadeau and S. Sekine, “A Survey of Named Entity Recognition and Classification,” *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007 Available: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

[2] J. Li, A. Sun, J. Han, and C. Li, “A Survey on Deep Learning for Named Entity Recognition,” IEEE Transactions on Knowledge and Data Engineering, Jan. 2022.Available: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[3] Reuters, “Pope Leo calls for ceasefire, hostages’ release in Israel-Hamas conflict,” Reuters, Aug. 27, 2025. [Online]. Available: https://www.reuters.com/world/europe/pope-leo-calls-ceasefire-hostages-release-israel-hamas-conflict-2025-08-27/

[4] NBA, “Kobe Bryant dies in helicopter crash,” NBA.com, Jan. 26, 2020. [Online]. Available: https://www.nba.com/news/kobe-bryant-dies-helicopter-crash

[5] Rappler, “LPA and southwest monsoon update, August 27, 2025, 5 PM,” Rappler, Aug. 27, 2025. [Online]. Available: https://www.rappler.com/philippines/weather/lpa-southwest-monsoon-update-pagasa-forecast-august-27-2025-5pm/