

Development and Training of a Sinhala (si) Named Entity Recognition (NER) Model for SpaCy Framework

Thalgamuwe Gedara Ashen Akalanka Weligalle

thawe276@student.liu.se

Linköping University

732A81 - Text Mining

Abstract

Named Entity Recognition (NER) for low-resource languages like Sinhala poses significant challenges due to limited annotated data and the complexity of the language structure. This paper presents the development of a custom NER model for Sinhala using the spaCy framework. The dataset, comprising 6418 tokens and over 2000 manually annotated named entities across 18 categories, was meticulously curated from diverse news sources. I utilized a combination of statistical and rule-based approaches for annotation validation and implemented a custom spaCy pipeline customized to handle the linguistic complexities of Sinhala. The model achieved an impressive F-score of 98.04%, reflecting its accuracy and robustness. This work provides a valuable foundation for further research and development of NLP tools for Sinhala and other low-resource languages.

1 Introduction

Sinhala, the national language of Sri Lanka, is spoken by approximately 23 million people worldwide, making it one of the top 100 most widely used languages. Sinhala's influence extends beyond Sri Lanka, with vibrant communities in countries like Australia, Italy, Canada, and New Zealand actively using the language. As the digital age progresses, the demand for Sinhala language processing tools has grown, driven by the increasing need for inclusivity in global and local technological advancements.¹

This project aligns with Sri Lanka's 2025 resolutions, which emphasize digitization and the establishment of a paper-free environment to improve efficiency and accessibility across

public and private sectors.² A critical component of this transformation is the development of robust Natural Language Processing (NLP) tools tailored for Sinhala, including Named Entity Recognition (NER) systems.

The project develops a small-scale NER model for Sinhala using a dataset of over 6,000 annotated words. Despite its modest size, the dataset enables exploration of Sinhala's rich morphology, unique script, and context-sensitive grammar. While challenging due to linguistic intricacies, this NER system holds significant potential for applications in document processing, information retrieval, and digital communication.

As an extension of this work, I aim to develop medium- and large-scale NER models to enhance usability and performance, addressing a wider range of real-world applications. Furthermore, I plan to seek my university's recommendation to contribute this model to the official SpaCy repository, enabling broader adoption and fostering further advancements in Sinhala NLP. This initiative represents a significant step toward bridging the gap in language technology resources for Sinhala and supporting Sri Lanka's vision of a digital future.

2 Theory and Technology

2.1 Background

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that involves identifying and classifying entities in a text into predefined categories, such as persons, locations, organizations, and others. For Sinhala, a low-resource language with limited annotated datasets and tools, NER poses unique challenges. This has led researchers to explore various approaches,

¹https://en.wikipedia.org/wiki/Sinhala_language

²<https://mode.gov.lk/>

including statistical models, hybrid techniques, and rule-based methods, to address these challenges.

$$P = \frac{|\text{NewENTS} \cap \text{predictedENTS}|}{|\text{predictedENTS}|} \quad (1)$$

According to the above equation 1, the model is trained to match external input labels with the corresponding labels in the trained data. If the probability of a given entity aligns closely with the probability of a new token, it is categorized as belonging to the corresponding entity class. This process is iteratively applied to enhance the accuracy of the predicted entity probabilities. Fundamentally, the model was trained on labeled data and subsequently evaluated on unseen data to assess its performance and labeling accuracy.

2.2 Technologies

The development of a Named Entity Recognition (NER) model for Sinhala leverages a combination of advanced technologies and methodologies to achieve accurate results. To secure data accuracy and annotation, an online entity categorization tool was used. You can also use tools like Doccano, Brat, and INCEpTION if you are comfortable with them. The model leverages the SpaCy framework, utilizing its support for tokenization, lemmatization, and custom pipeline creation tailored to Sinhala processing. Python was used for implementation, and JSON format for efficient dataset storage and processing.

2.3 Related Literature

Dahanayaka and Weerasinghe, 2014 conducted foundational research on Sinhala NER by evaluating the performance of Maximum Entropy (ME) and Conditional Random Fields (CRF) models using a manually annotated corpus. Their work primarily focused on identifying named entity boundaries without performing classification into specific categories such as persons, locations, or organizations. Despite the limited scope, they concluded that the CRF model outperformed the ME model. However, the limited availability of training data constrained their ability to make comprehensive comparisons between tools.

Ranathunga et al., 2024 present a comprehensive multi-way parallel corpus for English, Tamil, and Sinhala, annotated with Named Entities (NEs). This corpus significantly contributes to multilingual Named Entity Recognition (NER) by facilitating the evaluation and development of models across linguistically diverse languages. The study highlights the challenges of implementing BIO (Begin, Inside, Outside) tagging schemes in languages with complex morphosyntactic structures and varying word orders.

Additionally, the research explores integrating advanced language models, such as XLM-R and mBERT, to leverage multilingual learning for NER tasks. These advancements highlight the importance of robust annotated datasets, as language-specific behaviors alone are insufficient to guarantee improved model performance. In the context of this project, the insights from Ranathunga et al., 2024 underscore the importance of robust language-specific resources and methodologies, even though the potential of BIO schemes could not be fully explored due to time constraints.

3 Data

To generate the dataset, I utilized news paragraphs sourced from ongoing stories published on BBC News Sinhala³. The selected topics included significant and timely issues such as the illegal immigration of Rohingya individuals in Sri Lanka, the Sri Lankan President's visit to India, and the potential impact of Donald Trump's policies on Sri Lanka. This process resulted in a dataset containing a total of 6,418 tokens.

To annotate the dataset, I leveraged an online entity categorization tool⁴ and successfully identified over 2,000 named entities. These entities were then organized into 18 categories, consistent with the naming conventions used in the spaCy repository. This categorization approach ensured alignment with established standards for Named Entity Recognition (NER). The categories (as mentioned below) included a diverse range of entity types, covering persons, locations, organizations, and

³<https://www.bbc.com/sinhala/topics/cg7267dz901t>

⁴<https://arunmozhi.in/ner-annotator/>

other relevant groups commonly used in natural language processing tasks.

Listing 1: List of Entities

PERSON:	People, including fictional
NORP:	Nationalities or religious or political groups.
FAC:	Buildings, airports, highways, bridges, etc.
ORG:	Companies, agencies, institutions, etc.
GPE:	Countries, cities, states.
LOC:	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:	Objects, vehicles, foods, etc. (Not services.)
EVENT:	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART:	Titles of books, songs, etc
LAW:	Named documents made into laws.
LANGUAGE:	Any named language.
DATE:	Absolute or relative dates or periods.
TIME:	Times smaller than a day.
PERCENT:	Percentage, including “%”.
MONEY:	Monetary values, including unit.
QUANTITY:	Measurements, as of weight or distance.
ORDINAL:	“” first, “”second, etc.
CARDINAL:	Numerals that do not fall under another type.

As illustrated in Figure 1, the dataset was manually categorized through a meticulous, time-intensive, and fully human-driven process. This manual effort ensured a high level of accuracy and contextual understanding, which is crucial for Named Entity Recognition (NER) tasks in a complex language like Sinhala. After completing the categorization, the dataset was converted into a structured JSON format, capturing both the textual data and its corresponding entity annotations.

Finally, the JSON dataset was utilized to create a SpaCy-compatible object. This step involved integrating the annotated data into SpaCy’s processing pipeline, enabling efficient model training and validation. The conversion and integration process ensured that the dataset adhered to SpaCy’s format requirements, facilitating seamless experimentation and deployment of the custom NER model.

4 Method

To develop the Sinhala NER model, we followed these steps:

• Dataset Preparation:

- Collected news articles from BBC News Sinhala.
- Annotated 6418 tokens manually and created over 2000 named entities across 18 categories using an online entity categorization tool.

• Data Formatting:

- Stored the annotated data in JSON format for efficient processing.

• Model Development:

- Used Python and spaCy framework for NER model development.
- Built a custom pipeline leveraging spaCy’s tokenization, lemmatization, and NER functionalities.

• Training:

- Trained the model using the annotated dataset and evaluated its performance.

Listing 2: Model train step by using training and validation data

```
!python -m spacy train config.cfg --
    output ./model --paths.train ./data/
    si_core_web_sm_train.spacy --paths.
    dev ./data/si_core_web_sm_valid.
    spacy
```

5 Results

Figure 2 indicates that 10 out of 34 entities were incorrectly recognized, resulting in an accuracy of 70.59% for the selected data. As examples,

* ෂී ජින් පිං (Shi-Jin-Pin) should show as a PERSON but in output the entity type shows as a FAC.

* රාජ්‍ය තාන්ත්‍රිකයින් (Government officers) shows as a ORG, But it should NORP.

* ක්ෂේත්‍ර ගණනාවක් (Several fields) shows as Product. But it should not categorize under any entity.

Some trained special words were not recognized by the model. Ex-

- චීන (China) - GPE
- අග්‍රාමාත්‍ය (Vice President) - PERSON



Figure 1: Train data set and manual entities.



Figure 2: Random testing paragraph and NER outputs

Event	F-score	Precision	Recall
Early Performance	4.94	3.23	10.47
Significant Improvement	65.86	67.78	64.04
High Accuracy Achieved	94.40	94.73	94.07
Final Metrics	98.04	98.38	97.70

Table 1: Summary of NER model metrics

* The model starts with an F-score of 4.94 after the initial iteration, indicating a low initial performance, as expected.

* By iteration 5 (400 updates), the F-score jumps to 65.86, showing that the model quickly learns key patterns in the data.

* After 800 iterations, the F-score reaches 94.40, with precision and recall closely aligned (94.73 and 94.07). This suggests the model is effectively balancing false positives and false negatives.

* Beyond 1,000 iterations, the F-score stabilizes around 97–98%, indicating the model has converged and is performing robustly.

* The final F-score is 98.04, with precision of 98.43 and recall of 97.64. These high values highlight the model’s strong ability to identify and classify entities in the Sinhala language dataset.

6 Discussion

Table 1 clearly shows that the model has been trained with high accuracy. When we examine the final metrics, the F-score achieved is 98.04%, and the precision is 98.38%. These results indicate the effectiveness of the approach. However, there is still room for improvement, particularly in scaling the model to handle the complexity of the Sinhala language. To achieve this, more data is needed to create a medium-scale model. For this study, I used approximately 2,300 sentences to train the model. According to discussions on StackExchange⁵, the amount of labeled data directly influences the performance of a Named Entity Recognition (NER) system. One of the most challenging aspects of this work was manually labeling the data. The task involved annotating 18 types of entities, requiring careful decisions for every single word. For instance, entities like චීන (China) need to be accurately identified and categorized based on their context.

When comparing this study to related work, such as the research by Dahanayaka and Weerasinghe, 2014 on Sinhala NER, the model

⁵<https://ai.stackexchange.com/questions/32501/how-much-labelling-is-required-for-ner-with-spacy>

presented here goes a step further by incorporating advanced tokenization techniques and leveraging the spaCy framework. However, similar to their findings, this study also highlights the limitations posed by the lack of large annotated datasets, which can impede the development of robust NER models for low-resource languages like Sinhala.

If we consider the research by [Ranathunga et al., 2024](#), they have attempted to build a multi-way parallel language model using models like XLM-R and mBERT. However, upon examining that research paper, it appears that while the study of language behavior is worthwhile, it has no constructive impact on achieving the results we hope to achieve through it. I also explored the potential of the BIO scheme for development in my project, but it was unsuccessful due to time constraints. I have mentioned that project here to record the consideration of the methods used in that project.

According to my study, various researchers have tried to theoretically create a Sinhala NER model, but it is unfortunate that it has not been developed as a functional project. The primary reason for this, in my view, is the lack of adoption of Sinhala digital interfaces in Sri Lanka’s service sector and the predominant use of English in public services. However, with the government’s New Year resolutions, which include a shift to digital interfaces and a reduction of throughput time, I believe such projects will become even more necessary in 2025.

Language-specific challenges, such as inflections, compound nouns, and context sensitivity, emphasize the need for an extensive dataset and domain-specific knowledge. While the current model achieves high accuracy, its reliance on a limited dataset suggests that future work should focus on expanding data sources and exploring semi-supervised or transfer learning approaches to reduce the need for manual annotation.

7 Conclusion

The analysis demonstrates that the developed Sinhala Named Entity Recognition (NER) model achieves a high level of accuracy, with a final F-score of 98.04% and precision of 98.38%. This indicates that the model effectively iden-

tifies and classifies named entities within the dataset. However, the project also highlights several challenges, including the complexity of the Sinhala language and the time-intensive nature of manual data annotation.

The project was successfully completed, and the future expectation is to develop this model to medium and large scales. The aim is to integrate this model into the official spaCy repository and release Sinhala NER functionality in its future versions. Additionally, I anticipate receiving university support and guidance to further enhance and expand this work.

While the model successfully addresses the stated problem by creating a functional NER system for Sinhala, its reliance on a relatively small dataset limits its scalability and robustness. The project underscores the need for larger, high-quality datasets and the exploration of advanced techniques such as transfer learning or semi-supervised learning to reduce dependence on manual annotation.

Overall, the project has made significant progress in filling a critical gap in Sinhala language processing, and it’s available on [gitHub public repository](#)⁶. It lays the foundation for future research and development, particularly as digital transformation efforts in Sri Lanka create a growing need for such language technologies.

Acknowledgments

I want to specially thank, Professor Marcel Bollmann for guiding and encourage me to conduct this mini research, and the assistance of ChatGPT⁷ in refining the document.

References

- JK Dahanayaka and AR Weerasinghe. 2014. Named entity recognition for sinhala language. In *2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 215–220. IEEE.
- Surangika Ranathunga, Asanka Ranasinghea, Janaka Shamala, Ayodya Dandeniya, Rashmi Galappaththia, and Malithi Samaraweera. 2024. A multi-way parallel named entity annotated corpus for english, tamil and sinhala. *arXiv preprint arXiv:2412.02056*.

⁶<https://github.com/AshenWELI/>

Development-and-Training-of-a-Sinhala-si-Named-Entity-Recognition-

⁷<https://chatgpt.com/>