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. Find out this project on GitHub[1].

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.[2]

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.[3] 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 core task in Natural Language Processing (NLP) that involves identifying and categorizing named entities such as persons, locations, organizations, dates, and numerical values. According to Mansouri et al., 2008, NER is essential for various applications, including question answering, information retrieval, machine

[1] https://github.com/AshenWELI/Development-and-Training-of-a-Sinhala-si-Named-Entity-Recognition-NER-Model-for-SpaCy-Framework

[2] https://en.wikipedia.org/wiki/Sinhala_language

[3] https://mode.gov.lk/

translation, and text summarization. The Message Understanding Conference (MUC-6) (Zhou and Su, 2002) introduced the concept of Named Entities, providing a benchmark for NER systems.

For humans, entity recognition appears intuitive due to capitalization, word patterns, and contextual clues. However, for computational models, challenges arise due to semantic ambiguity and domain dependency. For instance, the phrase ”The White House” could refer to a location or an organization, depending on the context. In Sinhala, similar ambiguities exist, particularly due to inflections, compound words, and script complexities.

The NER Process (Ekbal and Bandyopadhyay, 2008) can be broken down into the following steps:

- **Text Preprocessing**
  - Tokenization: The input text is split into individual tokens (words or subwords).
  - Normalization: Handling variations in text such as case folding and stemming.
  - Part-of-Speech (POS) Tagging: Assigning syntactic labels (e.g., noun, verb) to words.
- **Feature Extraction**
  - Lexical Features: Character-level patterns (e.g., capitalization for names).
  - Syntactic Features: POS tags that provide context (e.g., nouns frequently appearing as entities).
  - Contextual Features: Neighboring words influence classification.
  - Word Embeddings: Pre-trained representations like Word2Vec, FastText, or contextual embeddings to capture relationships between words.
- **NER Model Processing**
  - The input text is passed through a Machine Learning-based model, such as:
    - ∗ Hidden Markov Model (HMM)
    - ∗ Conditional Random Fields (CRF)
    - ∗ Neural Networks (BiLSTM, CNN, Transformer-based models like BERT and XLM-R)
  - BIO Tagging Scheme is applied, where each token is labeled as :
    - ∗ B-PER (Begin of Person Name)
    - ∗ I-PER (Inside Person Name)
    - ∗ O (Outside Named Entity)
- **Post-Processing and Evaluation**
  - The extracted named entities are evaluated based on Precision, Recall, and F1-score.
  - If necessary, rule-based corrections are applied to improve accuracy.

Given Sinhala’s status as a low-resource language, existing rule-based approaches lack adaptability, and machine learning-based models require substantial annotated data. This project focuses on developing an efficient NER system for Sinhala using a machine learning-based approach within the SpaCy framework.

2.2 Technologies

The development of a Sinhala NER model leverages a combination of modern NLP techniques and annotation tools.

- **Annotation and Data Preprocessing:**
  - The dataset was manually annotated using BIO (Begin, Inside, Outside) format, ensuring high-quality labels. (In this project, I have not used BIO annotations.)
  - Online entity categorization tools were used for annotation validation. Additionally, tools like Doccano, Brat, and INCEpTION can be used for large-scale labeling.
- **NER Model Architecture:**
  - SpaCy’s CNN-BiLSTM Model: SpaCy’s NER pipeline uses transition-based parsing with Convolutional Neural Networks (CNNs) and a BiLSTM (Bidirectional Long Short-Term Memory) network to learn entity dependencies efficiently.

  - Word Embeddings: The model incorporates word embeddings for contextual understanding, improving its ability to distinguish between ambiguous entities.

- **Implementation Details:**
  - Programming Language: The model was implemented in Python, leveraging SpaCy's robust NER capabilities.
  - Data Storage: Annotated data is stored in JSON format, ensuring compatibility with different NLP frameworks.

This methodology ensures high accuracy, adaptability to Sinhala-specific linguistic structures, and scalability for future improvements. Future work includes expanding the dataset and integrating the model into official SpaCy repositories to enhance Sinhala language 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 Sinhala[4]. 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 tool[5] 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 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.
```

[4] https://www.bbc.com/sinhala/topics/cg7267dz901t

[5] https://arunmozhi.in/ner-annotator/

```
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

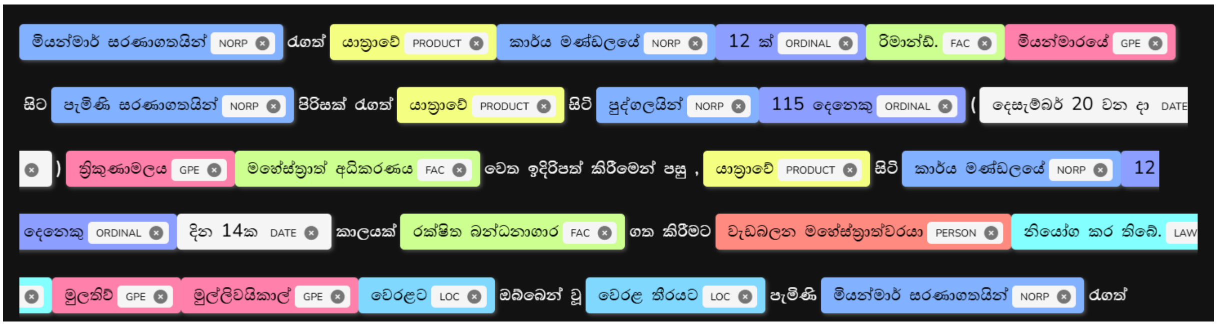
The model's performance was evaluated on a Sinhala NER (Named Entity Recognition) task. The classification report indicates varying performance across different entity types. Figure 3 indicates that 10 out of 45 entities were incorrectly recognized, resulting in an accuracy of 70.78% for the selected data according to the human evaluation but in the confussion matrix (Figure 2), the model was 64% on the selected dataset, with significant discrepancies in the recognition of certain entities.The difference between the human evaluation and statistical evaluation can be attributed to factors such as semantic ambiguity and domain dependency in Sinhala. Additionally, misclassifications such as ”Unknown” tags were mapped to the ”O” (non-entity) class [6], ensuring that ambiguous or unrecognized words did not disrupt the classification process. This step was crucial for reducing confusion during the recognition phase. The detailed classification report is as follows Table 1

As examples of misclassifications,

* ශ්‍රී ජින් පිං (Xi-Jinping) 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.

[6] https://dulaj.medium.com/confusion-matrix-visualization-for-spacy-ner-9e9d99120ee9



මියන්මාර් සරණාගතයින් NORP රැගත් යාත්‍රාවේ PRODUCT කාර්ය මණ්ඩලයේ NORP 12 ක් ORDINAL රිමාන්ඩ්. FAC මියන්මාරයේ GPE

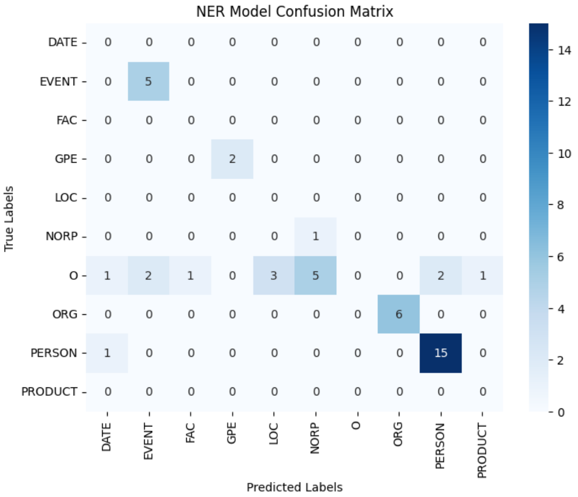
සිට පැමිණි සරණාගතයින් NORP පිරිසක් රැගත් යාත්‍රාවේ PRODUCT සිටි පුද්ගලයින් NORP 115 දෙනෙකු ORDINAL (දෙසැම්බර් 20 වන දා DATE

) ත්‍රිකුණාමලය GPE මහේස්ත්‍රාත් අධිකරණය FAC වෙත ඉදිරිපත් කිරීමෙන් පසු , යාත්‍රාවේ PRODUCT සිටි කාර්ය මණ්ඩලයේ NORP 12

දෙනෙකු ORDINAL දින 14ක DATE කාලයක් රක්ෂිත බන්ධනාගාර FAC ගත කිරීමට වැඩබලන මහේස්ත්‍රාත්වරයා PERSON නියෝග කර තිබේ. LAW

මුලතිව් GPE මුල්ලිවයිකාල් GPE වෙරළට LOC ඔබ්බෙන් වූ වෙරළ තීරයට LOC පැමිණි මියන්මාර් සරණාගතයින් NORP රැගත්

Figure 1: Train data set and manual entities.



	DATE	EVENT	FAC	GPE	LOC	NORP	O	ORG	PERSON	PRODUCT
DATE	0	0	0	0	0	0	0	0	0	0
EVENT	0	5	0	0	0	0	0	0	0	0
FAC	0	0	0	0	0	0	0	0	0	0
GPE	0	0	0	2	0	0	0	0	0	0
LOC	0	0	0	0	0	0	0	0	0	0
NORP	0	0	0	0	0	1	0	0	0	0
O	1	2	1	0	3	5	0	0	2	1
ORG	0	0	0	0	0	0	0	6	0	0
PERSON	1	0	0	0	0	0	0	0	15	0
PRODUCT	0	0	0	0	0	0	0	0	0	0

Figure 2: Si NER confusion matrix

* 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

* The model starts with an F1-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 F1-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 F1-score stabilizes around 97–98%, indicating the model has converged and is performing robustly.

	precision	**recall**	**F1**	**sup**
DATE	0.00	0.00	0.00	0
EVENT	0.71	1.00	0.83	5
FAC	0.00	0.00	0.00	0
GPE	1.00	1.00	1.00	2
LOC	0.00	0.00	0.00	0
NORP	0.17	1.00	0.29	1
O	0.00	0.00	0.00	15
ORG	1.00	1.00	1.00	6
PERSON	0.88	0.94	0.91	16
PRODUCT	0.00	0.00	0.00	0
Accuracy			0.64	45
Macro avg	0.38	0.49	0.40	45
Weighted avg	0.57	0.64	0.60	45

Table 1: Summary of NER model matrics

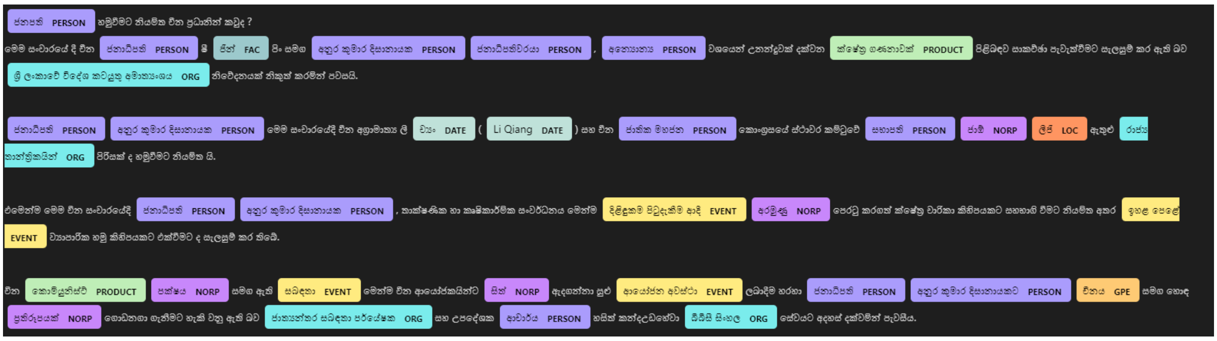
Event	**F1-score**	**Precision**	**Recall**
Early Performance	4.94	3.23	10.47
Significant Improvement	65.86	67.78	64.04
High Accuracy Achieved	94.40	94.73	94.07
Final Metrics	98.04	98.38	97.70

Table 2: Summary of NER model metrics

* The final F1-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 2 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 ap-



ජනපති PERSON හමුවීමට නියමිත චීන ප්‍රධානීන් කවුද ?
මෙම සංචාරයේ දී චීන ජනාධිපති PERSON ෂී ජින් FAC පිං සමග අනුර කුමාර දිසානායක PERSON ජනාධිපතිවරයා PERSON , අන්‍යෝන්‍ය PERSON වශයෙන් උනන්දුවක් දක්වන ක්ෂේත්‍ර ගණනාවක් PRODUCT පිළිබඳව සාකච්ඡා පැවැත්වීමට සැලසුම් කර ඇති බව
ශ්‍රී ලංකාවේ විදේශ කටයුතු අමාත්‍යාංශය ORG නිවේදනයක් නිකුත් කරමින් පවසයි.

ජනාධිපති PERSON අනුර කුමාර දිසානායක PERSON මෙම සංචාරයේදී චීන අග්‍රාමාත්‍ය ලී චියං DATE (Li Qiang DATE) සහ චීන ජාතික මහජන PERSON කොංග්‍රසයේ ස්ථාවර කමිටුවේ සභාපති PERSON ජාඕ NORP ලිජි LOC ඇතුළු රාජ්‍ය
තාන්ත්‍රිකයින් ORG පිරිසක් ද හමුවීමට නියමිත යි.

එමෙන්ම මෙම චීන සංචාරයේදී ජනාධිපති PERSON අනුර කුමාර දිසානායක PERSON , තාක්ෂණික හා කෘෂිකාර්මික සංවර්ධනය මෙන්ම දිළිඳුකම පිටුදැකීම ආදී EVENT අරමුණු NORP පෙරටු කරගත් ක්ෂේත්‍ර චාරිකා කිහිපයකට සහභාගී වීමට නියමිත අතර ඉහළ පෙළේ
EVENT ව්‍යාපාරික හමු කිහිපයකට එක්වීමට ද සැලසුම් කර තිබේ.

චීන කොමියුනිස්ට් PRODUCT පක්ෂය NORP සමග ඇති සබඳතා EVENT මෙන්ම චීන ආයෝජකයින්ට සිත් NORP ඇදගන්නා සුළු ආයෝජන අවස්ථා EVENT ලබාදීම හරහා ජනාධිපති PERSON අනුර කුමාර දිසානායකට PERSON චීනය GPE සමග හොඳ
ප්‍රතිරූපයක් NORP ගොඩනගා ගැනීමට හැකි වනු ඇති බව ජාත්‍යන්තර සබඳතා පර්යේෂක ORG සහ උපදේශක ආචාර්ය PERSON හසිත් කන්දුඋඩහේවා බීබීසී සිංහල ORG සේවයට අදහස් දක්වමින් පැවසීය.

Figure 3: Random testing paragraph and NER outputs

proximately 2,300 sentences to train the model. According to discussions on StackExchange [7], 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 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 identifies 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 expectations are to develop this model to medium and large scales and reduce the number of unknown entities. The aim is to integrate this model into the official spaCy repository and release Sinhala NER functional-

[7] https://ai.stackexchange.com/questions/32501/how-much-labelling-is-required-for-ner-with-spacy

ity 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. 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 ChatGPT [8] 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.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Improving the performance of a ner system by postprocessing and voting. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, SSPR & SPR 2008, Orlando, USA, December 4-6, 2008. Proceedings*, pages 831–841. Springer.

Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat. 2008. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344.

Surangika Ranathunga, Asanka Ranasinghea, Janaka Shamala, Ayodya Dandeniyaa, Rashmi Galappaththia, and Malithi Samaraweeraa. 2024. A multi-way parallel named entity annotated corpus for english, tamil and sinhala. *arXiv preprint arXiv:2412.02056*.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 473–480.

[8] https://chatgpt.com/