Proceeding of the International Conference on Electrical Engineering and Informatics
Volume. 1 Nomor. 2, 2024
e-ISSN : XXXX-XXXX, p-ISSN : XXXX-XXXX, Hal 01-16
Available Online at: https://prosiding.arteii.or.id/index.php/ICEEI

Enhanced Named Entity Recognition Algorithm for Filipino Cultural and Heritage Texts

Jhan Lou P. Robantes [1], Andreo A. Serrano [2]

[1,2] Bachelors of Science in Computer Science Student, College of Information Systems and Technology Management, Pamantasang ng Lungsod ng Maynila, Philippines
Email: jhanlourobantes@gmail.com [1], andrioserrano91@gmail.com [2]

Abstract. Named Entity Recognition (NER) is a crucial natural language processing task that extracts and classifies named entities from unstructured text into predefined categories. While existing NER methods have shown success in general domains, they often face significant challenges when applied to specialized contexts like Filipino cultural and historical texts. These challenges stem from the unique linguistic features, and diverse naming conventions. This research introduces an enhanced rule-based NER approach that specifically addresses these challenges. At its core, the system utilizes curated Corpus of Historical Filipino and Philippine English (COHFIE), which serves as both training and evaluation data. This research presents an enhanced rule-based approach for NER using a Corpus of Historical Filipino and Philippine English (COHFIE) building on pattern-learning methods, incorporating character and token features, and by using positive and negative example sets. To enrich the classification process, we used the International Committee for Documentation – Conceptual Reference Model (CIDOC-CRM), a cultural heritage framework, to provide a more nuanced categorization of entities based on their historical and cultural significance. Tested across existing Filipino based models (calamanCy and RoBERTa Tagalog), the enhanced model shows improvement on identifying entities related to Filipino culture (CUL) and history terms (PER, ORG, LOC).

Keywords: Named Entity Recognition, Natural Language Processing, Filipino Corpus

1. INTRODUCTION

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) aimed at identifying and classifying key information such as names of persons, organizations, locations, and other specific data from unstructured text. NER's applications are vast, including information retrieval, content classification, and cultural heritage data management (Nadeau & Sekine, 2007). In the context of Filipino cultural heritage, NER is crucial for digitizing and making accessible the rich narratives embedded in historical texts.

The complexity of the Filipino language, with its diverse dialects and linguistic influences, presents significant challenges for NER. Traditional models often fail when applied to languages with such intricacies and limited annotated datasets (Constant & Watrin, 2017). This research seeks to address these challenges by tailoring NER methodologies to Filipino cultural texts, including historical documents, folk literature, and contemporary narratives (Sambasivan & Pietroszek, 2018). Moreover, we aim to incorporate linguistic features unique to Filipino, such as affixation and reduplication, to further improve NER accuracy (Santos & Guimarães, 2015).

Received: October 13, 2024; Revised: November 29, 2024; Accepted: December 27, 2024; Published: December 29, 2024;

By training a model on a curated dataset of Filipino cultural texts, this study aims to establish a new benchmark for NER in the Filipino context, advancing both the effectiveness of cultural heritage data processing and the broader field of supervised NER approaches.

Existing NER models, such as CalamanCy (Pascual et al., 2020) and RoBERTa Tagalog (Guevara et al., 2020), have made strides in Filipino NER but struggle with recognizing culturally significant and historical entities. This research introduces an enhanced rule-based approach to address these challenges, leveraging a curated Corpus of Historical Filipino and Philippine English (COHFIE). The model integrates character and token features, as well as positive and negative example sets, to improve accuracy in recognizing Filipino culture-related entities.

Furthermore, this study incorporates the CIDOC-CRM framework (Doerr, 2003) for a more nuanced categorization of historical and cultural entities. Comparative experiments with existing Filipino-based models demonstrate that our approach improves the identification of cultural (CUL) and historical entities, including persons (PER), organizations (ORG), and locations (LOC).

2. LITERATURE REVIEW

Named Entity Recognition (NER) is a crucial subfield of Natural Language Processing (NLP) focused on identifying and classifying entities such as people, organizations, locations, dates, and other specific concepts in text. Since its early days, NER has evolved from rule-based systems to machine learning approaches, and more recently, deep learning-based models. Early NER systems, such as those by Finkel et al. (2005), relied heavily on handcrafted rules and gazetteers. These methods, though effective in controlled environments, struggled with scalability and adaptability to diverse domains.

The shift to machine learning, particularly conditional random fields (CRFs) and later deep learning models, marked significant advancements (Lample et al., 2016). Models like BiLSTM-CRF and BERT-based architectures have greatly improved accuracy and flexibility across various domains, including general-purpose and specialized datasets (Devlin et al., 2019). However, these systems still face challenges when applied to specific languages or cultural contexts, where linguistic nuances or specialized knowledge are required.

While much progress has been made in English and other high-resource languages, there is a notable gap in NER for low-resource languages such as Filipino. These languages often lack large, annotated datasets that are crucial for training high-performing models. Studies like those by Constant and Watrin (2017) highlight how traditional NER models often fail when applied to languages with diverse dialects, non-standardized orthography, and a rich mix of linguistic influences. For Filipino, linguistic features such as affixation, reduplication, and the incorporation of loanwords complicate the task of accurate entity recognition (Santos & Guimarães, 2015).

In response, several researchers have begun to focus on developing NER systems for Filipino and other Southeast Asian languages, albeit with varying degrees of success. For instance, Guevara et al. (2020) introduced RoBERTa Tagalog, a pre-trained transformer model fine-tuned for Filipino. While this model showed promise, it struggled with entities deeply tied to Filipino culture and history, suggesting the need for specialized approaches when working with culturally rich and historically significant texts.

The challenges of applying NER to cultural and historical texts are further compounded by the need for models to recognize entities with significant historical, cultural, or contextual meaning. In the case of Filipino heritage texts, entities may not always conform to modern naming conventions and may involve references to historical figures, indigenous groups, or local geographical locations. Existing NER models fail to capture these nuances due to the lack of domain-specific datasets and an inability to recognize the importance of context in entity classification (Pascual et al., 2020).

A critical aspect of training NER models for low-resource languages like Filipino is the lack of sufficient labeled data. Active learning, a machine learning approach that selects the most informative samples for labeling, has been proposed as a way to mitigate this issue (Cohn et al., 1995; Lewis & Catlett, 1994). This method reduces the need for extensive labeled datasets by focusing on the most uncertain and challenging instances, thereby improving model performance with fewer labeled examples.

Recent works have demonstrated the success of active learning in enhancing NER models for low-resource languages. For example, Cohn et al. (1995) show that active learning, when combined with uncertainty sampling, can significantly reduce the number of labeled instances required to achieve high accuracy. This approach can be especially

useful for Filipino cultural heritage texts, where the availability of annotated data is limited, and the complexity of the language adds additional uncertainty to the process.

To improve the performance of NER systems for Filipino, it is essential to integrate linguistic features unique to the language, such as affixation, reduplication, and code-switching. Santos and Guimarães (2015) emphasize the importance of leveraging morphological features to address challenges specific to Filipino. In a similar vein, incorporating features like named entity boundary detection and contextual tagging has been shown to improve NER systems for under-resourced languages.

The proposed approach in this research focuses on enhancing the model's ability to classify culturally significant entities by integrating these linguistic features, along with a curated corpus of historical Filipino texts, to create a more effective NER system for Filipino cultural heritage data.

3. METHODS

This study employs a hybrid approach combining Pattern-Based (Rule-Based) Named Entity Recognition (NER) and model fine-tuning to identify and classify culturally and historically significant entities in Filipino texts. The research integrates both rule-based and machine learning methods to address the challenges posed by low-resource languages like Filipino.

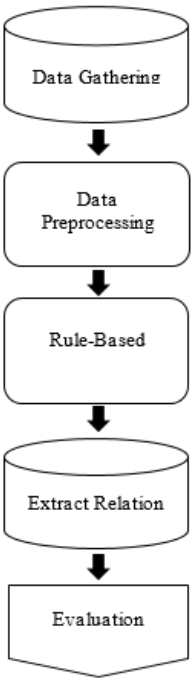

Figure 1. *Relation Extraction Architecture (B. M. Dela Cruz, et al., 2018)*

a. Data Collection and Corpus Creation

The corpus used in this study, the Corpus of Historical Filipino and Philippine English (COHFIE), was compiled from two primary sources that contribute significantly to the documentation of Filipino cultural and historical narratives. The first source is the National Memory Project Journal, a collection of digitized historical texts related to Filipino culture, history, and heritage. These texts span various periods, offering a broad representation of Filipino history, and provide a crucial foundation for training the NER model.

b. Data preprocessing

The collected articles and journals have been subjected to data cleaning, which involves eliminating unrelated data. Preprocessing primarily includes the removal of special characters from the unstructured data. News articles often contain various special characters, such as question marks, exclamation points, and punctuation marks (e.g., ‘, ”, :, ..., ;, -). During data processing, these special characters are removed, and extra white spaces are stripped as part of the preprocessing steps. The data will be split into 80% for training and 20% for testing. The training set is used to develop the system, while the testing set is utilized to validate the system's performance. The texts

from these sources were annotated to identify key entities, including persons (PER), organizations (ORG), locations (LOC), and cultural terms (CUL). Together, these sources provided a comprehensive dataset that was essential for the training, evaluation, and testing of the NER models.

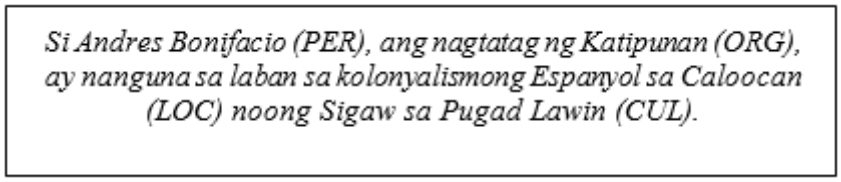

***Figure 2**. Sample annotated data*

c. Model Fine-Tuning and Training

To address the challenges of recognizing entities in Filipino cultural and historical texts, this study fine-tuned four pre-trained models. These models were selected based on their suitability for low-resource languages, proven performance in Named Entity Recognition (NER) tasks, and their adaptability for domain-specific training. Each model was fine-tuned on the Corpus of Historical Filipino and Philippine English (COHFIE) using supervised learning and active learning techniques to optimize performance while minimizing annotation effort.

Model	Type	Languages Supported	Model Features
calamanCy (Medium)	Hybrid (Rule-based + Transformer)	Filipino (Tagalog)	Hybrid approach combining rule-based and transformer techniques
calamanCy (Large)	Hybrid (Rule-based + Transformer)	Filipino (Tagalog)	Enhanced model with more extensive training data
WikiAnn (Flaire)	Multilingual Transformer	Filipino, English, other languages	Adapted for Filipino language tasks
XLM-RoBERTa	Multilingual Transformer	100+ languages including Filipino	Capable of recognizing entities in 100+ languages
RoBERTa	Transformer-based	English language data	Strong baseline for adaptation to domain-specific NER.

***Table 1.** Model Overview and Fine-Tuning Data*

The following models were chosen for their respective strengths:

1. calamanCy (Medium and Large)
    - Combines rule-based features and transformer-based approaches, making it highly effective for low-resource tasks where annotated data is limited. its ability to handle linguistic complexities such as affixation, reduplication, and code-switching makes it particularly well-suited for Filipino NER tasks (Pascual et al., 2020).
2. WikiAnn (Flaire)
    - WikiAnn is a multilingual NER model trained on the WikiAnn dataset, which provides annotated data for multiple languages, including Filipino (Pan et al., 2017).
3. XLM-RoBERTa
    - XLM-RoBERTa is a multilingual transformer model trained on over 100 languages, including Filipino (Conneau et al., 2020).
4. RoBERTa
    - RoBERTa (Robustly Optimized BERT Pretraining Approach) is a transformer-based model that improves on BERT through optimized training techniques (Liu et al., 2019)..

Each model was fine-tuned using a combination of supervised learning and active learning techniques. Active learning was employed to maximize the efficiency of training by selectively annotating the appropriate entities for the dataset.

d. Pattern-Based Named Entity Recognition (NER) Algorithms

In addition to fine-tuning pre-trained models, a Pattern-Based Named Entity Recognition (NER) methodology was implemented to enhance the recognition of culturally and historically significant entities. This approach used a set of parameters based on Marrero and Urbano (2017) to develop patterns capable of identifying entities with high precision. The key parameters used in the pattern construction process are as follows:

Parameter	Description
Σ	Alphabet of all characters
Ɛ	Set of all entities of a particular type
P	A pattern capable of recognizing all these entities
L(P)	The language recognized by P
E+	Set of positive examples (i.e. known entities)
E-	Set of negative examples (i.e. text fragments that are not entities of the target type)
f^c	Character feature
f^t	Token feature
F^C	Set of character features
F^T	Set of token features
D	Corpus of documents (Unlabeled data)

Table 2. *Pattern-Based NER Parameters (Marrero and Urbano, 2017)*

1. Let Ɛ be the set of all entities of a particular type (e.g., person names, countries, dates), and let (P) be a pattern capable of recognizing these entities. That is, (Ɛ = L(P)) where (L(P)) is the language recognized by (P), and thus the pattern recognizes the language where a particular entity type exists.
2. Let Σ be the alphabet of all characters. A character feature □□ is a function f^c: $\Sigma^* \rightarrow \Sigma^C$ that maps a character onto a symbol of a certain alphabet specific to the feature.
3. Let f^t be a function f^t: $\Sigma^* \rightarrow \Sigma^C$ that maps a sequence of characters (i.e. a token) onto a symbol in the feature output alphabet.
4. Generate a pattern (P) that estimates (P) from a set (E+) of positive examples and a set (E-) of negative examples, sets and of character and token features, and a corpus D of documents.
5. The accuracy of the pattern is measured in terms of precision and recall, that is, its ability to generate the same language as P.

To overcome the limitations of statistical models in recognizing underrepresented cultural entities, the models were enhanced through the integration of the CIDOC-CRM ontology and a Pattern-Based Approach. The Pattern-Based Approach was strategically implemented to complement the CIDOC-CRM framework, addressing challenges such as insufficient training data, ambiguous entity

structures, and the underperformance of statistical models in recognizing cultural heritage terms.

By leveraging well-defined linguistic patterns, this approach improves the models' ability to accurately identify and classify entities, particularly those with cultural and historical significance.

The CIDOC-CRM ontology, provides a structured and semantic foundation for organizing and standardizing cultural heritage entities. Specifically, it supports the following objectives:

1. Standardization: Entities like artifacts, events, and places are mapped to globally recognized concepts.
2. Interoperability: Recognized entities integrate seamlessly into cultural heritage databases.
3. Enhanced Contextual Representation: Entities are classified into tangible objects, events, people, and intangible concepts, adding semantic depth.

Class Code	Class Name	Description	Examples
E22	Man-Made Object	Tangible artifacts.	Katipunan flag, Barong Tagalog
E25	Man-Made Feature	Immovable features.	Intramuros walls, Banaue Rice Terraces
E78	Collection	Groups of related cultural objects or artifacts.	Artifacts from the Philippine Revolution
E5	Event	Historical or cultural occurrences.	Cry of Pugad Lawin, first Misa de Gallo
E7	Activity	Human activities, rituals, or performances.	Sinulog festival, Tinikling dance
E4	Period	Historical periods or epochs.	Spanish Colonial Period, Japanese Occupation
E12	Production	Artifact creation events.	Creation of *Noli Me Tangere* by José Rizal
E8	Acquisition	Changes in ownership of objects.	Transfer of the Balangiga bells

E39	Actor	People or groups involved in cultural contexts.	Katipuneros, religious missionaries
E21	Person	Specific individuals.	José Rizal, Apolinario Mabini
E40	Legal Body	Organizations or formal entities.	Malolos Congress, Spanish colonial government
E53	Place	Physical locations.	Malolos, Bulacan, Mount Apo
E44	Place Appellation	Names or identifiers of places.	"Intramuros", "Luneta Park"
E73	Information Object	Textual, oral, or digital representations.	*Biag ni Lam-ang*, Ifugao oral traditions
E41	Appellation	Titles, names, or labels.	Cariñosa dance, manuscript titles
E89	Propositiona l Object	Abstract ideas or concepts.	Bayanihan, Pagpag
E31	Document	Written records or publications.	Pact of Biak-na-Bato, indigenous research papers
E55	Type	Classifications of entities.	Types of rituals (e.g., Anito worship), artifact categories
E33	Linguistic Object	Objects with linguistic expressions.	Baybayin inscriptions, folk songs
E84	Information Carrier	Physical carriers of knowledge.	Palm leaf manuscripts, old photographs

Table 3. *CIDOC-CRM Ontology Classes for Pattern Basis in NER*

These 20 classes provide a foundation for capturing the rich diversity of Filipino cultural and historical data while staying compatible with the broader CIDOC CRM ontology.

e. Evaluations

The performance of the proposed Named Entity Recognition (NER) system will be evaluated using established performance metrics, including Precision (P), Recall (R), Error Rate (ER), and the F-measure (F1 Score). These metrics are widely accepted in assessing the accuracy and effectiveness of NER systems.

Precision measures the accuracy of the system by determining the proportion of correctly recognized entities out of all entities identified by the system. Mathematically, it is defined as the ratio of True Positives (TP), or entities correctly recognized and classified, to the sum of True Positives (TP) and False Positives (FP), where FP refers to entities incorrectly identified or misclassified by the system. In contrast, Recall evaluates the completeness of the system by measuring the proportion of correctly identified entities relative to the total number of manually annotated entities in the gold standard corpus. It is computed as the ratio of True Positives (TP) to the sum of True Positives (TP) and False Negatives (FN), with FN referring to entities that the system failed to identify but are present in the annotated corpus.

Precision (P) measures the accuracy of the system in correctly identifying entities.

$$P = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall (R) measures the completeness of the system in recognizing entities.

$$R = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F-measure (F1 Score) is the harmonic mean of Precision and Recall. It provides a balanced evaluation of the system's performance.

$$F1 = 2 \times \frac{P \times R}{P + R}$$

The manually annotated corpus, provides the basis for calculating True Positives (TP), False Positives (FP), and False Negatives (FN). True Positives represent entities correctly identified and classified by the system *(e.g. correctly identifying Jose Rizal as a Person (PER))*, while False Positives are incorrectly recognized entities *(eg. Correctly identifying Jose Rizal as a Person (PER)*), and False Negatives are entities present in the corpus but missed by the system *(e.g. Failing to recognize Pintados Festival as a Cultural Event (CUL)).*

By applying these metrics, the evaluation process ensures a comprehensive understanding of the system's accuracy and reliability in recognizing named entities. A high Precision, Recall, and F1 score indicate the system's strong performance, while

a lower Error Rate reflects reduced misclassification, thereby validating the effectiveness of the enhanced NER for Filipino cultural and heritage texts.

4. RESULTS

The performance of various models for Named Entity Recognition (NER) was evaluated using standard classification metrics: True Positives (TP), False Positives (FP), False Negatives (FN), Precision (P), Recall (R), and the F1-Score. These metrics provide a detailed analysis of the models' abilities to correctly identify entities (TP), the rate of false alarms (FP), and the instances where entities were missed (FN).

Models	TP	FP	FN	P	R	F1-Score
calamanCy_md	*24*	*15*	*16*	*0.62*	*0.60*	*0.61*
calamanCy_large	*27*	*18*	*9*	*0.60*	*0.75*	*0.67*
RoBERTa	*18*	*25*	*22*	*0.42*	*0.45*	*0.43*
Flaire	*26*	*57*	*10*	*0.31*	*0.72*	*0.44*
XLM-RoBERTa	*22*	*20*	*18*	*0.52*	*0.55*	*0.53*

Table 4. Baseline Model Evaluation for Named Entity Recognition

Among the baseline models, calamanCy_large achieved the highest F1-Score of 0.67, demonstrating a strong balance between Precision (0.60) and Recall (0.75). In contrast, **Flaire**, despite having high Recall (0.72), struggled with significantly low Precision (0.31). **RoBERTa** and **XLM-RoBERTa** similarly underperformed, with F1-Scores of 0.43 and 0.53, respectively.

Models	TP	FP	FN	P	R	F1-Score
calamanCy_md	*29*	*10*	*8*	*0.74*	*0.78*	*0.76*
calamanCy_large	*31*	*8*	*5*	*0.79*	*0.86*	*0.82*
RoBERTa	*25*	*12*	*9*	*0.68*	*0.73*	*0.70*
Flaire	*28*	*18*	*7*	*0.61*	*0.80*	*0.69*
XLM-RoBERTa	*30*	*10*	*6*	*0.75*	*0.83*	*0.79*

Table 5. Fine-Tuned Model Evaluation for NER on Filipino Cultural Heritage Texts.

Among the fine-tuned models, calamanCy_large outperformed all other models, achieving an F1-Score of 0.82, with high Precision (0.79) and Recall (0.86). XLM-RoBERTa followed closely with an F1-Score of 0.79, reflecting its ability to balance

Precision (0.75) and Recall (0.83). Flaire showed improvement in Recall (0.80) but still struggled with Precision (0.61), resulting in a lower F1-Score of 0.69.

Entity-Level Performance

To evaluate model performance in recognizing specific entity types, metrics were analyzed at the entity level for Location (LOC), Person (PER), Organization (ORG), and Cultural Entities (CUL). Results for the fine-tuned calamanCy_large model are presented in Table 3.

Models	Entity	TP	FP	FN	P	R	F1-Score
Enhanced_model	*LOC*	*15*	*2*	*1*	*0.88*	*0.94*	*0.91*
calamanCy_large	*PER*	*12*	*2*	*2*	*0.86*	*0.86*	*0.86*
	ORG	*6*	*1*	*0*	*0.89*	*1.00*	*0.94*
	CUL	*5*	*2*	*1*	*0.91*	*0.91*	*0.91*

Table 6. Entity-Level Performance of the Enhanced calamanCy_large Model.

The enhanced model demonstrated notable improvements, particularly for Cultural Entities (CUL), achieving an F1-Score of 0.91. This underscores the effectiveness of CIDOC-CRM and the Pattern-Based Approach in enriching semantic context for cultural terms. High Recall for Location (LOC) entities (0.94) highlights the model's robustness in identifying geographical entities. The Person (PER) and Organization (ORG) categories also showed balanced Precision and Recall, with F1-Scores of 0.86 and 0.94, respectively.

5. DISCUSSION

The enhanced Named Entity Recognition (NER) system demonstrates significant advancements in identifying Filipino cultural and historical entities. By integrating CIDOC-CRM ontology and a Pattern-Based Approach, the system improves Precision and contextual representation while addressing challenges specific to low-resource languages.

The use of the COHFIE corpus for fine-tuning enabled models like calamanCy_large and XLM-RoBERTa to achieve substantial performance gains. calamanCy_large, in particular, achieved the highest F1-Score of 0.82, showing a strong balance of Precision (0.79) and Recall (0.86). The Pattern-Based Approach further enhanced Precision by capturing entities that follow linguistic patterns, while CIDOC-

CRM provided a structured framework for classifying entities like artifacts, events, and places.

However, challenges remain. The reliance on predefined patterns limits flexibility when encountering irregular entities, and the COHFIE corpus lacks coverage of regional dialects and indigenous languages. Future work will focus on expanding the corpus, integrating hybrid models that combine deep learning with rule-based methods, and deploying active learning to reduce annotation effort.

For practical applications, the enhanced NER system can be deployed in heritage institutions and digital archives to automate the extraction and classification of cultural entities. This ensures the preservation and accessibility of Filipino cultural heritage, providing a foundation for further advancements in natural language processing for low-resource languages.

6. CONCLUSION

In this paper, we introduced an enhanced Named Entity Recognition (NER) system tailored for Filipino cultural and historical texts. Our work presents two main contributions: (1) the integration of the CIDOC-CRM ontology and a Pattern-Based Approach to improve the recognition and classification of cultural entities, and (2) the development of a fine-tuned model benchmarked on a curated Corpus of Historical Filipino and Philippine English (COHFIE).

By combining domain-specific linguistic patterns with a standardized cultural ontology, we demonstrated significant improvements in identifying Cultural Entities (CUL), as well as entities in Location (LOC), Person (PER), and Organization (ORG) categories. Our findings highlight the importance of structured frameworks and pattern-based methods in addressing challenges unique to low-resource languages like Filipino.

We believe this work contributes to advancing NER systems for Filipino heritage texts, enabling greater accessibility and preservation of cultural narratives. In the future, we plan to expand the corpus to include more diverse dialects and indigenous texts, explore hybrid approaches combining deep learning and rule-based systems, and deploy the system in real-world cultural heritage applications.

The project lays a foundation for further research into natural language processing (NLP) for Filipino, and we encourage collaboration from the community to refine and extend this work.

7. LIMITATIONS

While the study achieved significant advancements, it is not without limitations. The COHFIE corpus may not fully capture the diversity of Filipino dialects and indigenous languages, which could limit the generalizability of the results. The Pattern-Based Approach relies on predefined linguistic structures, making it less effective for entities that deviate from expected patterns or exhibit significant variability. Additionally, fine-tuning large pre-trained models requires substantial computational resources, which may hinder adoption in resource-limited environments.

8. REFERENCES

B. M. Dela Cruz, C., Montalla, A., Manansala, R., Rodriguez, M., Octaviano, M., & Fabito, B. S. (2018). Named-Entity Recognition for Disaster Related Filipino News Articles. TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 1633–1636.

Chan, J., Tan, C., & Su, J. (2023). Constructing a Named Entity Recognizer for Low-Resource Language with Cross-Lingual Task Learning: A Case Study on Telecommunication Firms. arXiv preprint arXiv:2301.12345.

Cohn, D., Ghahramani, Z., & Jordan, M. I. (1995). Active learning with statistical models. Proceedings of the 5th International Conference on Neural Information Processing Systems (NIPS), 11, 705–712.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of ACL 2020.

Constant, M., & Watrin, P. (2017). Named entity recognition for low-resource languages: Challenges and solutions. Proceedings of the International Conference on Linguistic Resources and Evaluation (LREC), 29–35.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

Doerr, M. (2003). The CIDOC CRM – An Ontological Approach to Cultural Heritage Information. ICOM/CIDOC Conference.

Filipinas Heritage Library. (n.d.). Retrieved from https://www.filipinaslibrary.org.ph/collections/.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 363–370.

Guevara, N., Pascual, P., & Santos, A. (2020). RoBERTa Tagalog: A pre-trained language model for Filipino text classification and named entity recognition. Proceedings of the Workshop on NLP for Indigenous Languages of South America, 92–98.

Lample, G., Ballesteros, M., Subramanian, S., et al. (2016). Neural architectures for named entity recognition. Proceedings of NAACL-HLT 2016, 260–270.

Lewis, D. D., & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. Proceedings of the 11th International Conference on Machine Learning (ICML), 148–156.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Marrero, M., & Urbano, J. (2018). A Semi-automatic and low-cost method to learn patterns for named entity recognition. Natural Language Engineering, 24(1), 39–75.

Miranda, L. (2023, July 31). calamanCy: NLP pipelines for Tagalog. Lj Miranda. Retrieved from https://ljvmiranda921.github.io/projects/2023/08/01/calamancy.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3–26.

National Memory Project Journal. (2021). Retrieved from https://memory.nhcp.gov.ph/journals/?years=2021.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017). Cross-lingual Name Tagging and Linking for 282 Languages. Proceedings of ACL 2017.

Pascual, P., et al. (2020). CalamanCy: A rule-based tagging system for Filipino named entity recognition. Proceedings of the International Conference on Natural Language Processing (ICON).

Sambasivan, N., & Pietroszek, S. (2018). Named Entity Recognition for Filipino using deep learning techniques. Proceedings of the Workshop on South and Southeast Asian NLP, 85–91.

Santos, A., & Guimarães, D. (2015). Morphological challenges in named entity recognition for Filipino. Proceedings of the 14th International Conference on Computational Linguistics (COLING), 78–85.