Building a Khmer NER Benchmark from Health News Data Towards Event Extraction

Cheaminh Chiep, Natt Korat, Vathna Lay, Rottana Ly
Cambodia Academy of Digital Technology
cheaminh.chiep@student.cadt.edu.kh

Abstract

Khmer Named Entity Recognition (NER) is a sub-task in Khmer Natural Language Processing (NLP) that extracts information to locate and classify named entities in Khmer text into predefined categories, such as names of persons, organizations, and locations. As a low-resource language, there is a lack of high-quality datasets for Khmer NER. This study addresses the lack of an NER dataset for the Khmer language, particularly in the public health domain. The Khmer Health Event Extraction Dataset (KHEED) was introduced. It comprises 525 annotated articles (5,980 sentences) from Khmer health news, covering eight entity types: Disease, Pathogen, Location, Human Count, Organization, Symptom, Medication, and Date. To evaluate the performance of Khmer NER on the proposed dataset, five Khmer-compatible NLP models were selected, and from the experimental results, XLM-RoBERTa Base achieved the best performance with a moderate F1-score of 0.7646. The KHEED Dataset will be publicly available and serve as the foundation and benchmark for future Event Extraction (EE) Datasets.

Keywords: *Khmer NLP, Khmer NER, Event Extraction, Health News Data, KHEED*

1 Introduction

Cambodia faces two major health issues: non-communicable diseases (NCDs), which cause 64% of deaths [1], and ongoing communicable diseases, especially in rural areas [1]. Khmer-language health news, provides important real-time information on outbreaks and health policies. However, this information is often unstructured and hard to access due to challenges in processing the Khmer language [2].

Named Entity Recognition (NER) is the process of highlighting text spans of important information, which are known as entities [3]. In the context of public health in Cambodia, these entities are utilized to represent diseases, pathogens, locations, human counts, organizations, dates, and medications, which are critical in monitoring health trends such as disease outbreaks. Event Extraction (EE), on the other hand, is a task of identifying event types, triggers, and arguments. NER and EE automation help accelerate the analysis of health trends and summarize Khmer health news, eliminating the need for time-consuming work by human annotators [4]. Developing NER and EE systems for the Khmer language presents numerous challenges due to the language's complexity and the scarcity of resources, including well-annotated corpora and pre-trained models [5]. In addition, Khmer writing does not use spaces like other languages, which means that an accurate word tokenization tool is required, as this is crucial in Natural Language Processing (NLP) [6].

In this paper, the contributions are:

1. Introduction of a novel domain-specific dataset: Creation of the Khmer Health Event Extraction Dataset (KHEED), comprising a 525-annotated corpus of named entities from Khmer news articles in public health.
2. Extension of general NER capabilities: Provision of a domain-specific dataset extending the general NER dataset for familiar entities.
3. Baseline performance evaluation: Evaluation of five existing Khmer-compatible NLP models on the NER task using the KHEED dataset.
4. Foundation for event extraction (EE): A fundamental step toward supporting the future creation of the EE dataset.

The paper is organized as follows. In Section 2, related work is presented, followed by the preparation of the KHEED Dataset in Section 3. Section 4 outlines the experiments of the selected Khmer-compatible models on the KHEED. Last but not least, a Discussion is provided in Section 5 before the conclusion and future work discussed in Section 6.

2 Related Work

This section reviews the previous work of NER and EE in the health domain, including existing datasets and models, as well as common methodologies employed.

2.1 Existing NER and EE for English and other Languages

Numerous datasets are published for English and other widely spoken languages. They are crucial for researchers to develop NLP models for NER and EE tasks. *CoNLL-2003* is a common gold standard NER dataset, a shared task of annotations of news wire articles with four familiar entities [7]. *OntoNotes 5.0* is another well-known annotation of a large multilingual corpus with 18 named entities [8]. More recently, *MultiNERD* was introduced as a large silver standard NER dataset. This dataset consists of a wide range of 10 languages, including English, Chinese, and Dutch, covering an extensive 15 NER categories, such as familiar entities (e.g., Person, Location, Organization) and less common entities (e.g., Biological entities, Mythological entities) [9]. MultiNERD addresses the work of manual annotation by automatically detecting entities using a combination of *mBERT + BiLSTM + CRF* and benchmarking against gold standard datasets, achieving high scores. For the EE task, the *2005 Automated Content Extraction (ACE)* is a standard collection of annotated entities, events, and relations at the sentence level for English, Chinese, and Arabic [10]. *MAVEN*, another EE dataset, is a general domain event detection dataset designed to expand upon previous EE datasets, covering a wide range of 168 event types and 118,732 event mentions, which is significantly larger than the ACE dataset [11]. However, these general domain datasets are not specialized in the health domain and lack support for the Khmer language.

2.2 Previous Work in Health Event Extraction

There are many ongoing efforts to support health EE due to its applications in disease monitoring. One of the key contributions in 2013 is the *GENIA* dataset, which is composed of biomedical articles with annotations of biomolecular entities and events [12]. In recent times, the *BAND* dataset was introduced to address the lack of publicly available surveillance on health news in the English Language [13]. This dataset comprises 1,508 samples sourced from news and emails. In addition, it offered several evaluation tasks, including NER, EE, and Question Answering (QA), developed by epidemiology experts. Similarly, the *SPEED++* dataset, a multilingual EE dataset comprising four languages collected from social media, features seven event types and 20 argument roles, aiming to provide early warnings of health hazards [14]. While these datasets are valuable for a standard health EE dataset, they may not be suitable for the Khmer news reporting style.

2.3 Existing Khmer NER and EE Resources

Publicly available resources for Khmer NER and EE tasks are minimal [15]. Currently, Khmer researchers are developing models and datasets [16]. However, an elementary open-source dataset was released to foster the development of NER for the Khmer Language [17]. This dataset provided labeling of the persons and location entities in Khmer text for over 47,700 sentences. In terms of Khmer-compatible models, Transformer-based models were often fine-tuned for downstream tasks. Models like mBART50 [18] were experimented with in Khmer NLP research and have achieved moderate results [16]. Despite these advances, datasets for various domains with complex annotation depths are still necessary.

2.4 Standard Annotation Schema and Tools

Effective NER and EE rely on well-defined annotation schemas and tools. A standard scheme is the BIO (Beginning, Inside, Outside) tagging scheme, the entities are broken down into tokens or sub-words, the first token is labeled as ’B-’ and the consequence tokens are labeled as ’I-’ while tokens that are not part of an entity are labeled as ’O [19]. An extension to this scheme

is the BIOES (Beginning, Inside, Outside, End, Single) tagging scheme, where similar principles are applied, but the ending token is labeled as 'E-' and single-token entities are labeled as 'S-' [19]. For event extraction, the scheme includes an event type supplement with event triggers and event arguments.

Various tools are available to support annotation tasks. A popular open-source labeling tool is *Label Studio*, a customizable interface for manual annotation workflow [20]. Another well-known annotation tool is *Docanno*, which provides support for text classification tasks and other tasks, such as sequence labeling and sequence-to-sequence tasks [21]. In addition, *Prodigy* is a modern annotation tool that enables the efficient development of AI systems [22].

2.5 Standard Benchmark and Evaluation Metrics

Evaluation of NER and EE can be done at three different levels. A standard evaluation for NER is the entity-level evaluation, which involves the correct and exact matching of predicted entities against their proper tags. The metrics used at this level are Precision, Recall, and F1-score [23]. Precision measures the percentage of correct predictions among all predicted entities. Recall measures the proportion of correct predictions among all actual entities. F1-score the the harmonic mean of the two metrics. A more precise evaluation is the event-level or tuple-level evaluation, which assesses the entire event extraction structure, including the event type, triggers, and argument roles, against a manually annotated gold standard dataset [24]. The metrics used at this level are more complex and can vary depending on their real-world application. The most advanced level is the document-level event evaluation, where models at this level learn the context within an entire document rather than just sentences [25].

3 KHEED Dataset

This section describes in detail the creation of the KHEED dataset.

3.1 Data Collection

We collected 28,057 health news articles from 2020 to 2025 from nine Khmer news outlets: VOA Khmer, RFA Khmer, Fresh News, Camboja, Khmer Times, Kampuchea Thmey Daily, DAP News, Cambodia Express News, and Khmer Breaking News (Table 1). Articles were scraped and thoroughly cleaned by removing HTML tags, after which they were segmented into sentences.

Table 1. Data Collection Summary

Source	Articles
VOA Khmer	1,120
RFA Khmer	1,907
Fresh News	15,018
Camboja	59
Khmer Times	106
Kampuchea Thmey Daily	1,657
DAP News	3,803
Cambodia Express News	1,775
Khmer Breaking News	2,612
Total	**28,057**

3.2 Annotation Schema

We annotate eight entity types: Disease, Pathogen, Location, HumanCount, Organization, Symptom, Medication, and Date. Table 2 provides names and examples for each type. We use these canonical labels throughout the paper and dataset.

These entities capture the essential details in a health news report. The distribution shows that Organization, Disease, and Location are the most frequent entities in the dataset, accounting for 73.4% of all entities. Date, HumanCount, Pathogen, Symptom, and Medication are less common; this creates an imbalance in the dataset, which could potentially reduce model performance. Figure 1 shows the distributions of each entity type.

3.3 Annotation Process

To improve the efficiency of manual annotation as a single annotator, a pre-annotation Python script was developed to assist in capturing entities in the articles. First, word segmentation was applied using the Khmer-NLTK library [26] in Python. Next, the script utilizes a predefined dictionary lookup, prefixes and suffixes with subsequent tokens, and regular expressions to automatically annotate entities. These pre-annotations are then imported into Label Studio for manual verification and correction, while also

Table 2. Named Entity Types and Examples

Entity	Examples
DIS	ជំងឺអេដស៍ (AIDS)
PAT	វីរុសកូរ៉ូណា (Coronavirus)
SYM	ក្អក (Cough)
MED	ថ្នាំអង់ទីប្យូទិច (Antibiotic)
HUM	៣០នាក់ (30 people)
DAT	ថ្ងៃទី១៥ ខែមករា ឆ្នាំ២០២៥ (January 15, 2025)
LOC	ខេត្តកំពង់ចាម (Kampong Cham Province)
ORG	មន្ទីរពេទ្យកាល់ម៉ែត (Calmette Hospital)

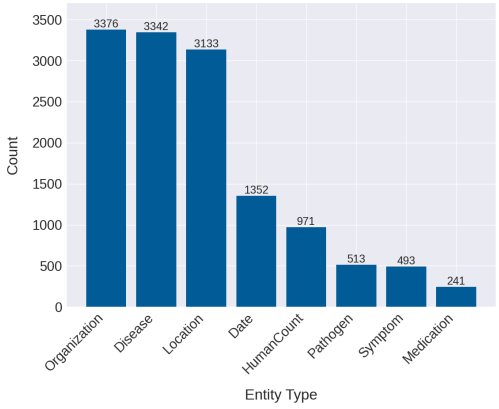

Figure 1. Distribution of Entity Types

annotating additional entities. Finally, verified annotations are exported as JSON, structured with text, entity spans, and labels.

Several challenges were encountered during annotation, including ambiguous entities, where organization names could be either specific names or general categories (e.g., ”royal government”). Another challenge was the dual-category entities, which are entities that could mean different things depending on the context (e.g., ”school” as an institution vs. physical location).

3.4 Data Processing

During processing, sentences containing fewer than five tokens were removed to focus on informative sentences, resulting in a final dataset of 5,980 sentences. The annotated data were then BIO tagged using KhmerNLTK [26] for tokenization and stored in JSON format. The data was then split into a ratio of 80:10:10 for training, validation, and testing, using random sampling to reduce bias and improve generalization.

4 Experiments

This section outlines the fine-tuning and evaluation of five Khmer-compatible NLP models on the KHEED dataset.

4.1 Model Selection

The models for evaluation were chosen based on their multilingual capabilities and pre-training in the Khmer language: XLM-RoBERTa-Khmer-Small, BERT-Khmer-Small-Uncased, PrahokBART-Base, XLM-RoBERTa-Base, and BiLSTM-CRF. Table 3 summarizes the specifications in each model.

4.2 Experimental Setup

The models were trained on a server equipped with an NVIDIA GeForce RTX 4070 Ti SUPER GPU and 16GB of memory. Hyperparameters were tuned on the validation set, with a fixed seed (42) for reproducibility. Early stopping was implemented to prevent the model from overfitting, halting training when the validation loss stopped improving. The best model was selected based on validation loss. Table 4 presents the hyperparameter settings for each model.

4.3 Model Performance

The evaluation was performed at the entity level, following standard NER evaluation metrics, as calculated using the formulas below.

- **Precision**: The ratio of correctly predicted positive tokens to the total predicted

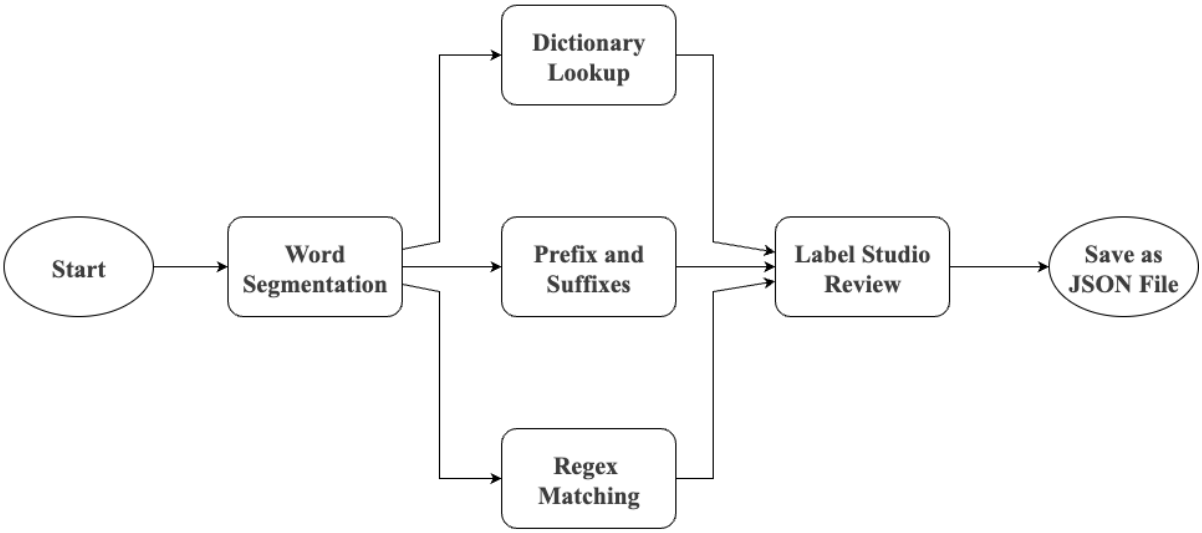

Figure 2. Annotation Process Flowchart

Table 3. Model Architecture Specifications

Model	Type	Architecture	Parameters
XLM-RoBERTa-Khmer-Small	Encoder	Transformer	49M
BERT-Khmer-Small-Uncased	Encoder	Transformer	29.1M
PrahokBART-Base	Enc-Dec	Transformer	62M
XLM-RoBERTa-Base	Encoder	Transformer	125M
BiLSTM-CRF	Sequential	LSTM+CRF	N/A

positive tokens, calculated as:

$$P = \frac{TP}{TP + FP} \tag{1}$$

- **Recall**: The ratio of correctly predicted positive tokens to all actual positive tokens, calculated as:

$$R = \frac{TP}{TP + FN} \tag{2}$$

- **F1-Score**: The harmonic mean of Precision and Recall, providing a balanced measure of performance:

$$F1 = \frac{2 \times (P \times R)}{P + R} \tag{3}$$

Table 5 present the overall performance of each model.

XLM-RoBERTa Base achieves the highest F1 Score of 0.7646, demonstrating its effectiveness on Khmer text. BERT-Khmer Small follows closely with a 0.7048 F1-score, showing that small pre-trained models are also effective. The BiLSTM-CRF model significantly underperformed compared to transformer-based models. Table 6 shows the scores of each model by entity type.

The results reveal a moderate performance in Date, HumanCount, Location, Pathogen, and Disease, benefiting from the patterns of these entities in news articles. A slightly lower performance for the Organization is due to excessively long text spans, which are challenging for the models to capture accurately. For medication entities, the presence of pathogen and disease names in the medication entities confused the models. Regarding symptoms, the models recognized them but failed to capture the whole entities.

Table 4. Model Hyperparameter Configurations

Model	Epochs	Batch Size	Learning Rate
XLM-RoBERTa-Khmer-Small	5	16	0.00002
BERT-Khmer-Small-Uncased	7	16	0.00002
XLM-RoBERTa-Base	8	16	0.00002
PrahokBART-Base	15	4	0.00002
BiLSTM-CRF	23	32	0.001

Table 5. Entity-Level Performance Results

Model	Precision	Recall	F1-Score
XLM-RoBERTa Base	**0.7006**	**0.8414**	**0.7646**
BERT-Khmer Small	0.6575	0.7595	0.7048
XLM-RoBERTa Khmer-Small	0.5943	0.7793	0.6744
PrahokBART Base	0.5732	0.6641	0.6153
BiLSTM-CRF	0.5147	0.5609	0.5368

5 Discussion

In this section, we will discuss about the current state of Khmer compatible NLP models, their limitation and capabilities.

5.1 Error Analysis

We defined four errors types, which include false positives, where the model mistakenly over predict entities, missed annotations could also cause false positives, boundary errors arising from segmentation issues, and type confusions, including cases where hospitals were tagged as Location instead of Organization due to contextual cues like នៅ (*at*). While tagging hospitals as Location in phrases like នៅមន្ទីរពេទ្យ (*at the hospital*) is contextually valid, consistently tagging them as Organization could improve model learning by reducing ambiguity. Figure 3 compares the errors of the predictions made by the models on the test set.

These errors reflect our challenges in the Khmer NER task, such as overfitting, inconsistent annotations, and opportunities for schema improvement. The following examples illustrate these issues:

- **False Positive**: The word ដង់ស៊ីតេ (*density*) in អត្រាកើតមានជំងឺថ្មី (ដង់ស៊ីតេ) (*density of new disease cases*) was incorrectly predicted as a Disease entity, but it is not an entity. This error occurred due to its proximity to the keyword ជំងឺ (*disease*).
- **Missed Annotation**: The word ខែមករា (*January*) in នៅខែមករាឆ្នាំនេះ (*in January this year*) was predicted as a Date entity, but the gold standard incorrectly labeled it as a non-entity, indicating annotation inconsistency.
- **Boundary Error (Under-Segmentation)**: The entity ជំងឺសរសៃឈាមបេះដូង (*heart disease*) was partially tagged as ជំងឺសរសៃឈាម (*vascular disease*) as a Disease entity, missing បេះដូង (*heart*). The model failed to recognize the correct entity boundary.
- **Boundary Error (Over-Segmentation)**: The entity ប្រទេសថៃ (*Thailand*) was correctly predicted as a Location entity, but the gold standard incorrectly annotated only ថៃ (*Thai*) as Location, indicating an annotator error.
- **Missed Entity**: The entity ជំងឺទឹកនោមផ្អែម (*diabetes*) in ក្តីព្រួយបារម្ភអំពីជំងឺទឹកនោមផ្អែម (*concern about diabetes*) was not predicted

Table 6. F1-Score Performance by Entity Type

Entity Type	XLM-RoBERTa Khmer-Small	XLM-RoBERTa Base	BERT-Khmer Small	BiLSTM-CRF	PrahokBART Base
Date	0.85	**0.86**	0.83	0.62	0.65
Disease	0.74	**0.78**	0.77	0.67	0.74
HumanCount	0.74	**0.84**	0.77	0.47	0.49
Location	0.73	**0.83**	0.77	0.56	0.66
Medication	0.30	**0.46**	0.00	0.08	0.12
Organization	0.53	**0.69**	0.59	0.43	0.50
Pathogen	0.63	0.73	0.65	0.55	**0.78**
Symptom	0.51	**0.58**	0.46	0.37	0.55

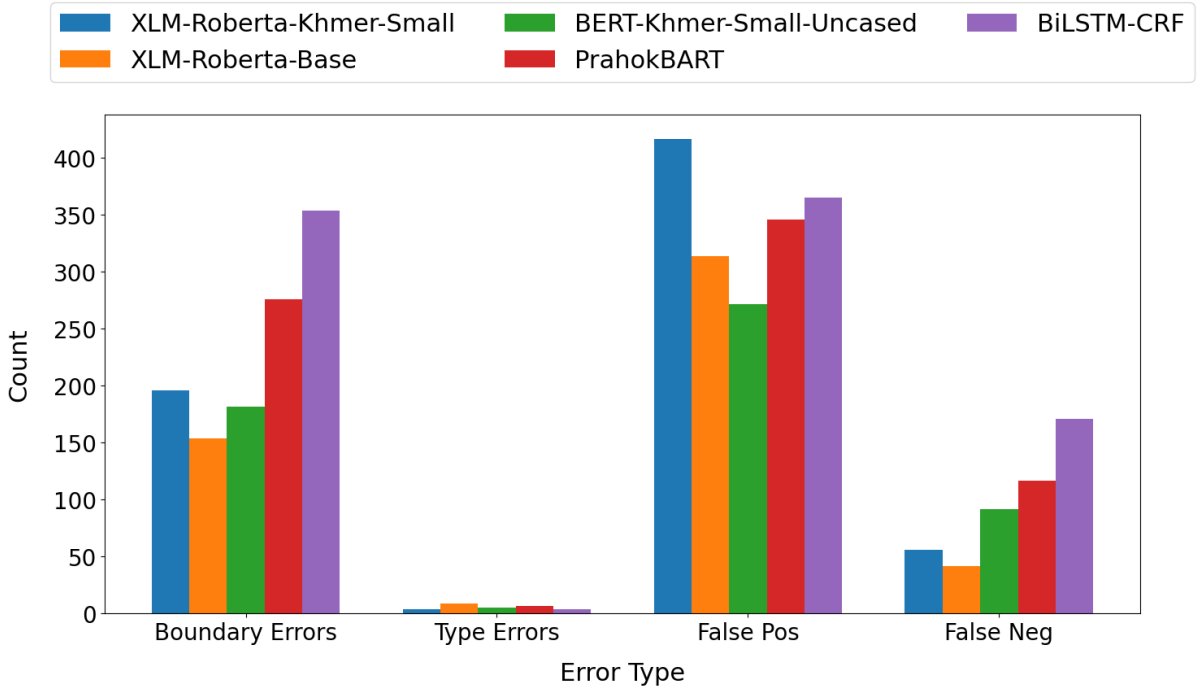

Figure 3. Model Performance by Error Type

as a Disease entity, due to segmentation issues.

- **Type Confusion**: The entity អាជ្ញាធរជាតិប្រយុទ្ធនឹងជំងឺអេដស៍ (*National Authority for Combating AIDS*) was partially tagged as ជំងឺអេដស៍ (*AIDS*) as a Disease entity instead of the correct Organization entity. This reflects confusion caused by disease terms within organizational names.
- **Schema Ambiguity (Type Confusion)**: The entity នៅមន្ទីរពេទ្យបង្អែកខេត្តកំពង់ធំ (*at Kampong Thom Provincial Hospital*) was predicted as an Organization entity, but the gold standard labeled it as Location. This confusion in the model indicates that it is unable to distinguish between organization and location in different contexts.

5.2 Current State

Our results demonstrate that the models are at a practical state, but still require significant advancement to compare with state-of-the-art models for other languages.

6 Conclusion

In closing, we'll summarize the contributions of this study and look closer into the extensibility of the dataset.

6.1 Limitations

Several limitations temper these results: severe entity class imbalance (e.g., Medication and Symptom constitute <5% of annotations) significantly degrades performance on rare types (F1 $\leq$ 0.46 for Medication); single-annotator labeling risks inconsistency, as evidenced by boundary errors and type confusion (e.g., hospitals misclassified as Location); domain restriction to formal news text limits applicability to clinical records or conversational Khmer; pre-annotation biases from dictionary and regex heuristics may introduce systematic errors; and moderate overall performance (best F1 Score $\sim$0.76) remains far below English NER benchmarks ($\sim$0.90+), highlighting the need for larger, higher-quality training resources.

Future work should focus on multi-annotator refinement, balanced sampling or loss reweighting, cross-domain data integration, advanced Khmer tokenization, and larger pre-trained models to improve robustness, generalization, and clinical utility. KHEED lays a critical foundation for advancing health NLP in Khmer.

6.2 Primary Contributions

This work presents two significant contributions to Khmer NLP community. First, we introduced the KHEED dataset, comprising of 525 articles with eight entity type, this dataset provide an automation to monitor health events from Khmer heatlh news. Second, we have evaluated the performance of Khmer compatible NLP models accommodated by the analysis about their limitations and key challenges for future research. We commit to publicly release the dataset, models, and documentation to facilitate similar research. The dataset, splits, tokenization scripts, and training configurations will be released under a CC BY-NC 4.0 license on Github (`https://github.com/CADT-LLM/kheed.git`), fostering reproducibility.

6.3 Future Research

Several promising research avenues emerge from this work. Enhancing KHEED to a large-scale health corpus that incorporates medical literature, clinical notes, and patient forums could improve the model's understanding, potentially eliminating boundary detection and data imbalance. Building upon entity recognition, studying the relationship between entities and the linking of events would allow for a more sophisticated analysis of health news. We plan to develop an EE schema, annotating triggers and arguments for events such as Outbreak (Disease, Location, Date, HumanCount), which will enable more sophisticated health trend analysis. Entity relation extraction and document-level event modeling are also promising areas of research.

References

[1] World Health Organization and Ministry of Health Cambodia. Prevention and control of noncommunicable diseases in cambodia: Progress report 2018. *WHO UniATF*, 2018. Accessed: 2025-11-11.

[2] H. Kaing, S. Deth, S. Uth, S. Sorn, S. Sar, S. Sam, and S. Nop. Towards tokenization and part-of-speech tagging for khmer. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–22, 2021.

[3] K. Pakhale. Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv.org*, 2023.

[4] C. Yu, Z. Cao, P. Zhao, D. D. Zeng, T. Luo, and J. Wang. Named entity recognition for epidemiological investigation in covid-19. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6, 2023.

[5] S. Jiang, S. Fu, N. Lin, and Y. Fu. Pretrained models and evaluation data for the khmer language. *Tsinghua Science and Technology*, 27(4):709–718, 2022.

[6] H. Kaing, C. Ding, M. Utiyama, E. Sumita, S. Sam, S. Seng, K. Sudoh, and S. Nakamura. Towards tokenization and part-of-speech tagging for khmer: Data and discussion. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(6):104:1–104:16, 2021.

[7] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147,

2003.

[8] R. Weischedel, S. Pradhan, M. Palmer, W. Xiong, and H. Xiong. Ontonotes release 5.0. `https://catalog.ldc.upenn.edu/LDC2013T19`, December 2013. LDC2013T19.

[9] S. Tedeschi and R. Navigli. Multinerd: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States, July 2022. Association for Computational Linguistics.

[10] C. Walker et al. Ace 2005 multilingual training corpus, 2006.

[11] X. Wang, W. Lu, Z. Cao, Z. Yu, Y. Padigela, R. Deng, H. Chen, F. Sun, and K. Li. Maven: A massive general domain event detection dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online, November 2020. Association for Computational Linguistics.

[12] J.-D. Kim, Y. Wang, K. Takemoto, F. Berard-Dufresne, K. B. Cohen, M. Colosimo, J. Kim, J.-J. Kim, A. Névéol, S. Pyysalo, and J. Tsujii. The genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[13] Z. Fu, M. Zhang, Z. Meng, Y. Shen, D. Buckeridge, and N. Collier. Band: Biomedical alert news dataset. *arXiv preprint arXiv:2305.14480*, 2023.

[14] T. Parekh, J. Kwan, J. Yu, S. Johri, H. Ahn, S. Muppalla, K.-W. Chang, W. Wang, and N. Peng. Speed++: A multilingual event extraction framework for epidemic prediction and preparedness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12936–12965, Miami, Florida, 2024. Association for Computational Linguistics.

[15] S. Jiang, S. Fu, N. Lin, and Y. Fu. Pretrained models and evaluation data for the khmer language. *Tsinghua Science and Technology*, 27(4):709–718, 2021.

[16] R. Buoy, N. Taing, S. Chenda, and S. Kor. Khmer printed character recognition using attention-based seq2seq network. *HO CHI MINH CITY OPEN UNIVERSITY JOURNAL OF SCIENCE-ENGINEERING AND TECHNOLOGY*, 12(1):3–16, 2022.

[17] R. Buoy. Khmer ner dataset. `https://huggingface.co/datasets/rinabuoy/khmer-ner-dataset`, 2023. Accessed: 2023-08-04.

[18] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint*, arXiv:2008.00401, Aug. 2020.

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[20] V. Naik, P. Patel, and R. Kannan. Legal entity extraction: An experimental study of ner approach for legal documents. *International Journal of Advanced Computer Science and Applications*, 14(1), 2023.

[21] O. Irrera, S. Marchesin, F. Shami, and G. Silvello. Doctron: A web-based collaborative annotation tool for ground truth creation in ir. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

[22] C. Macri, I. Teoh, S. Bacchi, M. Sun, et al. Automated identification of clinical procedures in free-text electronic clinical records with a low-code named entity recognition workflow. *Journal of Medical Systems*, 46(7):1–10, 2022.

[23] R. Yacouby and D. Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pages 70–79, 2020.

[24] P. Liu, Y. Guo, J. Lei, and G. Li. Tagging

schemes can do more in named entity recognition: Take chinese as an example. *2022 International Conference on Natural Language Processing and Knowledge Engineering (NLPKE)*, pages 1–6, 2022.

[25] M. Tong, B. Xu, S. Wang, M. Han, Y. Cao, J. Zhu, C. Shi, et al. Docee: a large-scale and fine-grained benchmark for document-level event extraction. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3084–3096, 2022.

[26] P. V. Hoang. Khmer natural language processing tookit. `https://github.com/VietHoang1512/khmer-nltk`, 2020.