Research Plan for Named Entity Recognition for Long Sequence

Aim and Background:

Named Entity Recognition (NER) is one of the essential tasks in Natural Language Processing (NLP). It identifies and classifies named entities like names, locations, organizations, time, etc., from texts, and this structured data can then be used to build more complexed NLP systems like information retrieval, question answering, chatbots, and more. Deep learning-based NER algorithms have gained much attention in academia and the industry because they require less manual feature engineering compared to traditional machine learning methods. Rresearchers are also actively exploring NER applications in both academic settings (Ehrmann; Ehrmann 1-47) and industries like finance and healthcare (Catelli 213).

Current NER systems excel at handling short text datasets, but a shortcoming lies in their approach to handling long sequences. It is shown that information from other sentences can improve NER models' performance (Chang 77; Chen). In real-world scenarios, NLP applications often expected to process long sequences. Consider the banking industry, where documents like loan applications, business proposals, are lengthy. Extracting specific information often requires reviewing entire documents for just a few keywords. An NER system that handles long sequences would significantly reduce processing time, allowing people to focus on core business decisions.

This research aims to improve current methods for modeling features from document-level long sequence and use long-range context to enhance the performance of NER models. Our primary goal is to develop a better architecture to model document-level features and improve the identification and classification of named entities in long sequences. The succeed of this research will improve methods for performing NER on entire documents.

Literature Review:

According to current research, deep learning approaches typically frame NER as a sequence labeling task. These models (e.g., BERT + BiLSTM + CRF) often follow a three-step process:

1. A pre-trained language model to generate contextualized word representations for the input text (e.g., BERT);
2. A context encoder to processes the PLM outputs to capture dependencies within the sequence (e.g., BiLSTM)
3. A tag decoder that assigns entity labels to each token based on the encoded context (e.g., CRF).

Such architecture has achieved good performance in data sets where texts are of short and middle length sequence. However, there are challenges in long sequence NER:

1. Many models limit the max input length due to computational and memory limitations.
2. It is challenging to capture relationships between words while filtering out irrelevant information in a long sequence.
3. The same expression may be different entities under different context.

Ongoing research to solve these challenges come in the following directions:

1. Improved PLMs: These research focuses on enhancing the ability of pre-trained models to capture long-range dependencies. (Hu)
2. Enhanced Encoders: Expanding the context window will include more sentences around the target word; and combining different feature levels will generate better representations. (Pakhale; Wang)

3. Advanced Tag Decoders: Improving methods for assigning entity labels will also improve the NER results. (Vu; Khandelwal)
4. Alternative System Architectures: Architectures like Graph Neural Networks (GNNs), generative models, or prompt learning can approach NER problems differently from sequence labeling. (Chen; Krishnan)

Methodology:

This research builds upon previous literature to improve NER architectures for long sequence texts. We will explore these improvements in 4 directions mentioned above:
1. We will investigate different network architectures for the context encoder to enhance feature extraction from long sequences.
2. We will explore methods to improve the tag decoders in sequence labeling tasks.
3. We will use alternative approaches for NER that are beyond sequence labeling.
4. We will pretrain and finetune language models that are suitable for long sequences and explore large language models (LLMs)'s capabilities for long sequence.

The datasets we will use are long sequence text documents from publicly available datasets like DWIE (Zaporojets). This is a dataset of English news articles for multiple learning tasks like NER and document-level information extraction. We will also consider constructing our own datasets of document-level long sequences. This corpus can be built by ethically crawling web sources, utilizing academic papers for pre-training, and collaborating with corporations to get access to their internal data.

We will use different evaluation metrics to evaluate our NER models and compare them with benchmark models, include standard NLP metrics like precision, recall, and F1-score. We will also define specific metrics for long sequence to evaluate the model's ability to utilize long-sequence features and capture long-range dependencies within the document.

Furthermore, we will do ablation studies. We will remove each component, re-evaluate the model, and analyze the component's impact to understand the contribution of it in our model architecture.

References:

1. Ehrmann, Maud, et al. "Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents." *CEUR Workshop Proceedings*. No. 3180. CEUR-WS, 2022.
2. Ehrmann, Maud, et al. "Named entity recognition and classification in historical documents: A survey." *ACM Computing Surveys* 56.2 (2023): 1-47.
3. Catelli, Rosario, et al. "Combining contextualized word representation and sub-document level analysis through Bi-LSTM+ CRF architecture for clinical de-identification." *Knowledge-Based Systems* 213 (2021): 106649.
4. Chang, Jun, and Xiaohong Han. "Multi-level context features extraction for named entity recognition." *Computer Speech & Language* 77 (2023): 101412.
5. Chen, Jiawei, et al. "Learning in-context learning for named entity recognition." *arXiv preprint arXiv:2305.11038* (2023).
6. Hu, Qingfeng. "Research on Named Entity Recognition Technology based on pre-trained model." *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. IEEE, 2022.
7. Pakhale, Kalyani. "Comprehensive overview of named Entity Recognition: Models, Domain-Specific applications and challenges." *arXiv preprint arXiv:2309.14084* (2023).
8. Wang, Yudi, et al. "Efficient Named Entity Recognition Based on Broad Learning System and BERT." *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*. IEEE, 2022.
9. Vu, Thanh, Dat Quoc Nguyen, and Anthony Nguyen. "A label attention model for ICD coding from clinical text." *arXiv preprint arXiv:2007.06351* (2020).
10. Khandelwal, Urvashi, et al. "Sample efficient text summarization using a single pre-trained transformer." *arXiv preprint arXiv:1905.08836* (2019).
11. Chen, Xiang, et al. "LightNER: A lightweight tuning paradigm for low-resource NER via pluggable prompting." *arXiv preprint arXiv:2109.00720* (2021).
12. Krishnan, Prashant, et al. "Towards Few-shot Entity Recognition in Document Images: A Graph Neural Network Approach Robust to Image Manipulation." *arXiv preprint arXiv:2305.14828* (2023).
13. Zaporojets, Klim, et al. "DWIE: An entity-centric dataset for multi-task document-level information extraction." *Information Processing & Management* 58.4 (2021): 102563.