Literature Survey on Named Entity Recognition (NER)

Named Entity Recognition (NER) is a sub-task of Information Extraction (IE) that focuses on identifying and classifying entities from text into predefined categories such as person names, organizations, locations, dates, etc.

Evolution of NER
The development of NER has evolved through various phases:

1. **Rule-Based Approaches (1990s):** Early NER systems relied on hand-crafted rules and dictionaries. These systems were effective but lacked scalability.

2. **Statistical Models (2000s):** With the introduction of statistical models like Hidden Markov Models (HMM) and Conditional Random Fields (CRF), the performance of NER improved significantly.

3. **Machine Learning Models (2010s):** Machine learning algorithms such as SVM, Decision Trees, and Random Forest were utilized to identify entities based on features like word context, capitalization, and word embeddings.

4. **Deep Learning Models (2015+):** Recent advancements in deep learning, especially with the advent of Recurrent Neural Networks (RNNs) and Transformers, have revolutionized NER. Models like BERT, GPT, and RoBERTa can recognize entities with minimal hand-crafted features.

5. **Pre-trained Language Models:** Huggingface Transformers, SpaCy, and other pre-trained models have significantly simplified the NER task by providing pre-trained embeddings and token

classifications.

Modern Techniques

Modern NER systems leverage transformers like BERT, which process text with contextual embeddings, allowing more accurate and flexible entity extraction.

Challenges in NER

1. **Ambiguity:** Words can have different meanings in different contexts, making it challenging to classify entities.
2. **Out-of-Vocabulary Words:** NER systems often struggle with unseen or uncommon words.
3. **Multilingual Texts:** Handling multiple languages remains a significant challenge.
4. **Data Scarcity:** Lack of annotated data in low-resource languages affects model performance.

Conclusion

The field of NER has witnessed significant evolution from rule-based to transformer-based approaches. The use of pre-trained language models like BERT has substantially improved accuracy and generalizability. However, challenges like ambiguity, data scarcity, and multilingualism still persist, warranting further research.