Long-Range Named Entity Recognition: A Comprehensive Survey

Ronit Ghode[1], Hariom Ingle[2], Ishwari Gondkar[3], Jidnyasa Harad[4], Ravindra Murumkar[5], Raviraj Joshi[6]

[1,2,3,4]Department of Information Technology, Pune Institute of Computer Technology, Pune, Maharashtra, India

[5]Indian Institute of Technology Madras, Chennai, Tamil Nadu, India

DOI: https://doi.org/10.51584/IJRIAS.2026.110200157

Received: 06 March 2026; Accepted: 11 March 2026; Published: 23 March 2026

ABSTRACT

The exponential growth of unstructured digital text has created a pressing need for sophisticated Natural Language Processing (NLP) methods to extract meaningful information. Named Entity Recognition (NER), the task of identifying and classifying named entities in text, is a cornerstone of this effort. While traditional NER has achieved remarkable success on short, self-contained texts, its application to long-form docu-ments—such as legal contracts, clinical records, and scientific literature—presents formidable challenges. This survey provides a comprehensive analysis of the state-of-the-art in Long-Range Named Entity Recognition. We trace the evolution from classical statistical models to the rise of Transformers, detailing the inher-ent quadratic complexity of models like BERT that limits their scalability. We conduct an in-depth exploration of the primary architectural paradigms designed to overcome this bottleneck: efficient Transformers that employ sparse attention mechanisms, and graph-based approaches that model explicit relational struc-tures within documents. Furthermore, we investigate critical challenges, including the data scarcity problem in specialized domains and unique linguistic complexities in multilingual con-texts. Drawing from recent analyses, we synthesize persistent open problems in document-level information extraction, focusing on long-distance coreference resolution and the need for robust, multi-step reasoning. Finally, we chart a course for future research, postulating that the next generation of solutions will be found in hybrid architectures that synergistically combine the strengths of deep sequential encoders with structured reasoning frameworks.

Indexterms—Named Entity Recognition, Long-Range NER, Transformers, Longformer, BigBird, Graph Neural Networks, Document-Level Information Extraction

INTRODUCTION

The Age of Information and the Need for Extraction

We live in an era of unprecedented data creation. From legal archives and corporate reports to electronic health records and scientific publications, the volume of unstructured text data is expanding at an exponential rate. This data deluge presents both an immense opportunity and a significant chal-lenge. Buried within these documents is a wealth of valuable information that, if extracted and structured, could drive dis-covery, inform decisions, and automate complex workflows. Information Extraction (IE) is the field of NLP dedicated to this task, transforming human language into machine-readable data.

The proliferation of digital documents across various do-mains has amplified the need for automated information extraction. Legal firms manage millions of contracts, hos-pitals maintain extensive electronic health records, financialinstitutions process countless regulatory filings, and academic researchers navigate ever-growing corpuses of scientific liter-ature. Manual processing of these documents is not merely time-consuming—it is often practically impossible at scale.

Defining Named Entity Recognition

At the heart of information extraction lies Named Entity Recognition (NER). Formally, NER is a sequence labeling task that seeks to identify and classify spans of text—or entities—into predefined categories such as persons, organi-zations, locations, dates, and medical codes. For example, in the sentence On Monday, Dr. Anya Sharma of the Apollo Hospital announced a new treatment in Mumbai, an NER system would identify Dr. Anya Sharma as a PERSON, Apollo Hospital as an ORGANIZATION, Monday as a DATE, and Mumbai as a LOCATION. This process is the first crucial step in understanding the who, what, where, and when of a text.

NER serves as a foundational component for numerous downstream applications including question answering sys-tems, knowledge graph construction, semantic search engines, and domain-specific applications such as adverse drug event detection in clinical texts and contract analysis in legal docu-ments.

The Leap from Sentences to Documents

For decades, NER research primarily focused on sentence-level or paragraph-level analysis, often using news articles or encyclopedia entries as benchmark datasets. However, the most valuable information is often found in long-form documents where context is distributed across many pages. Consider a legal contract where a company, Innovate Corp, is defined on page 1 with full legal specifications, but is subsequently referred to as the Company, the Corporation, or simply it for the next 50 pages. A sentence-level model, processing each sentence in isolation, would fail to connect these subsequent mentions to the original entity, leading to incomplete and inaccurate extraction. This necessity to un-derstand context across vast distances—potentially spanning thousands of tokens—defines the field of Long-Range NER.

Core Challenges in Long-Range NER

The transition to document-level analysis introduces a con-fluence of challenges:

Computational Complexity: The self-attention mechanism that powers modern Transformers scales quadratically with input length ($O(n^2)$), making it prohibitively expensive for long documents. A standard BERT model is limited to 512 tokens.

Long-Distance Dependencies: Models must resolve coref-erences, abbreviations, and relationships between entities men-tioned thousands of tokens apart. This requires selective atten-tion to relevant information across vast spans.

Domain Specialization: Long documents are often highly technical. Models must understand domain-specific jargon, entity types, and contextual conventions, requiring specialized training data.

Data Scarcity: Expert annotation of long, specialized doc-uments is incredibly time-consuming and expensive, creating a severe data bottleneck.

Scope and Structure of this Survey

This survey provides a structured overview of Long-Range NER. Section II establishes the historical and foundational context. Section III delves into architectural paradigms for handling long sequences. Section IV examines data challenges and domain adaptation methods. Section V synthesizes unre-solved challenges. Section VI discusses future directions, and Section VII concludes.

FOUNDATIONAL CONCEPTS AND EVOLUTION

Early Methodologies: Rule-Based and Statistical Models

The earliest NER systems were based on handcrafted rules using gazetteers and regular expressions. While effective for narrow domains, they were brittle and not scalable. The field then shifted to statistical models like

Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), which could learn patterns from annotated data. CRFs became a standard for sequence labeling tasks due to their ability to model dependencies between adjacent labels while considering rich feature sets.

The Deep Learning Revolution

With the advent of deep learning, Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, became state-of-the-art. A typical Bi-LSTM-CRF architecture, which processes text in both forward and backward directions and uses a CRF layer for decoding, remained the dominant NER model for several years. However, RNNs process text sequentially, making them slow to train and difficult to parallelize, with limited ability to capture long-range dependencies.

The Transformer Era: BERT and Self-Attention

The introduction of the Transformer architecture [21], and particularly the BERT model [6], marked a seismic shift in NLP. Instead of processing text sequentially, BERT's self-attention mechanism allows every token to directly relate to every other token in the input. This parallel processing and deep contextualization led to unprecedented performance on a wide range of tasks, including NER. Domain-adapted versions, like ClinicalBERT, further pushed boundaries by pre-training on specialized corpora.

The Scalability Bottleneck

The power of BERT comes at a cost. The self-attention mechanism requires computing a pairwise attention score between every token pair in the sequence. If a sequence has n tokens, this results in an $n \times n$ attention matrix. Both the memory required to store this matrix and the computation needed to calculate it scale quadratically with sequence length, i.e., $O(n^2)$. Doubling the sequence length to 1024 would quadruple the computational cost, and processing a 4096-token document would be 64 times more expensive, rendering it infeasible on standard hardware. This quadratic bottleneck is the single greatest barrier to applying standard Transformers to long documents.

ARCHITECTURAL PARADIGMS FOR LONG-RANGE

NER

To break the quadratic barrier, researchers have pursued two main strategies: making the attention mechanism more efficient through sparsity, or augmenting it with external structures that model relationships explicitly.

Efficient Transformers: Taming Self-Attention

This class of models retains the Transformer architecture but replaces full self-attention with sparse attention mechanisms, reducing complexity from $O(n^2)$ to approximately linear $O(n)$.

Longformer: Sliding Windows and Global Attention: The Longformer model [1] employs a combination of two attention patterns. Sliding Window Attention allows each token to attend only to its immediate neighbors in a fixed-size window, capturing local context. Global Attention allows a small number of pre-selected tokens to attend to the entire sequence and be attended to by all tokens, acting as aggre-gators and broadcasters of document-level information. This hybrid pattern allows Longformer to build representations that are both locally rich and globally informed, scaling linearly with sequence length up to 4096 tokens and beyond.

Clinical-Longformer, pre-trained on clinical notes, signifi-cantly outperformed ClinicalBERT on tasks requiring reason-ing across multiple sections of medical records [2].

BigBird: Block Sparse Attention: BigBird [3] offers a more generalized sparse attention mechanism composed of three components: global tokens that attend to the entire sequence, local window attention where each token attends to neighbors, and random attention where each token attends to a small number of

random tokens. The random component ensures that information can travel between any two positions in logarithmic steps, maintaining theoretical completeness. BigBird's theoretical guarantees prove that sparse attention can approximate any function that full attention can approximate.

Other Notable Approaches: Other methods include Re-former [22], which uses locality-sensitive hashing to group similar tokens, and Linformer, which projects the attention matrix into a lower-dimensional space. While these models have different theoretical properties, Longformer and BigBird have emerged as the most empirically successful for long document NER tasks.

Graph-Based Methods: Modeling Explicit Structure

While efficient Transformers extend the reach of attention, they still process information largely sequentially. Graph-based approaches offer a more explicit way to model the relational structure of a document.

From Text to Graph: Construction Strategies: The first step is converting a document into a graph G = (V, E). Nodes can represent different granularities: individual tokens, entity mentions, sentences, or paragraphs. Edges define re-lationships and can be based on proximity, lexical similarity, semantic similarity using embeddings, syntactic dependencies, discourse relations, or explicit coreference links [26].

The choice of graph construction strategy significantly im-pacts model performance. Dense graphs provide rich connec-tivity but increase computational cost, while sparse graphs are efficient but may miss important connections.

Reasoning with Graph Neural Networks: Once the graph is constructed, a Graph Neural Network (GNN) performs reasoning. Common architectures include:

Graph Convolutional Networks (GCNs) iteratively update node representations by aggregating information from neigh-bors. After k layers, each node's representation incorporates information from its k-hop neighborhood.

Graph Attention Networks (GATs) [7] extend GCNs by learning attention weights for edges, allowing the model to focus on the most relevant connections.

The key advantage is explicit long-range connectivity. A pronoun node can directly receive information from its an-tecedent node hundreds of sentences away, provided they are connected in the graph, enabling powerful long-range reasoning that purely sequential models struggle with.

The Role of Large Language Models

Recent advancements in Large Language Models (LLMs) like GPT-4 have introduced a new paradigm. Through in-context learning, these models can perform NER in zero-shot or few-shot settings by being prompted with instructions and examples. While powerful, their application to long-range NER is hampered by context window limitations, high computational costs for inference, and challenges in ensuring structured, reliable output for enterprise-scale extraction tasks.

ADDRESSING DATA SCARCITY AND DOMAIN SPECIFICITY

Even with perfect long-range architecture, performance depends on high-quality training data availability—a major bottleneck in specialized domains.

The Data Annotation Bottleneck

Manually annotating a 50-page legal contract or detailed clinical record for named entities requires significant do-main expertise and can take hours per document. Annotators must understand domain-specific terminology, recognize non-standard entity formats, maintain consistency across long documents, and resolve ambiguous cases. This makes creating large-scale, gold-standard datasets practically infeasible for many domains.

Weakly Supervised NER: Learning with Less

To address data scarcity, researchers have developed meth-ods that learn from weaker forms of supervision.

Limitations of Dictionary-Based Methods: A common approach uses domain-specific dictionaries to automatically label entity mentions. However, this has critical flaws: low coverage—dictionaries are inevitably incomplete, missing rare entities, spelling variations, and abbreviations; and ambigu-ity—dictionary terms often have multiple meanings depending on context, leading to false-positive labels.

High-Coverage Pseudo-Label Generation: Novel frame-works like HighGEN [5] address these limitations using phrase embedding search to find semantically similar phrases, cor-pus mining to automatically build high-coverage dictionaries, verification and filtering using multiple signals, and iterative refinement through bootstrapping. These methods can generate pseudo-labeled datasets an order of magnitude larger than manually annotated data while maintaining reasonable quality.

Distant Supervision: Distant supervision leverages exist-ing structured knowledge bases to automatically create training data. Advanced distant supervision methods use partial annota-tion, multi-instance learning, and explicit noise modeling [16].

Domain-Specific Pre-training

The performance of Transformer-based models is heavily influenced by their pre-training data. A model pre-trained on general text struggles with specialized vocabulary and conventions.

Case Study: The Clinical Domain: ClinicalBERT demonstrated the value of pre-training on clinical notes, outperforming general BERT on medical NER. However, it remained limited by the 512-token window. The introduction of Clinical-Longformer and Clinical-BigBird [2] combined long-range architecture with domain-specific knowledge by continuing pre-training from base checkpoints on large clin-ical corpora. These models consistently outperformed both ClinicalBERT (limited context) and base Longformer/BigBird (lacking clinical knowledge), achieving 5–10% absolute im-provement in F1-score on clinical NER tasks, with improve-ments exceeding 15% for tasks requiring long-range reasoning. This pattern has replicated across domains with Legal-Longformer for contract analysis, Financial-BigBird for regu-latory filings, and Scientific-Longformer for academic litera-ture.

PERSISTENT CHALLENGES AND OPEN QUESTIONS

Despite architectural advancements and improved training methods, comprehensive error analyses reveal that several deep challenges remain [4].

Long-Distance Reasoning and Coreference

This represents the most significant remaining hurdle. Even models with 4096+ token context windows can fail to correctly link entities across long distances when the relationship is complex or non-obvious.

Coreference Resolution: Connecting pronouns, abbrevia-tions, and definite descriptions to their antecedents remains difficult. A legal document might introduce Innovate Corp on page 1, abbreviate it as IC on page 3, refer to it as the Corporation on page 10, and simply as it on page 25. A model must recognize these as coreferent mentions and resolve ambiguity when multiple entities could be antecedents.

Multi-Hop Reasoning: Some entity recognition requires chaining multiple pieces of information [18]. Current attention mechanisms struggle with these scenarios where inference must proceed through multiple intermediate steps.

Linguistic Nuances and Cross-Lingual Adaptation

Challenges in Low-Resource Languages: Many lan-guages lack large pre-training corpora, annotated NER

datasets, standardized entity type taxonomies, and computa-tional resources for training large models. Transfer learning from high-resource languages shows promise but faces obsta-cles including different scripts, morphological systems, and cultural naming conventions.

Marathi and Morphologically Rich Languages: Marathi, an Indo-Aryan language spoken by over 80 million people, exemplifies challenges in morphologically rich languages [17]. It has complex morphology with extensive inflection, flexible word order, script complexity in Devanagari, and limited resources compared to English. Similar challenges exist for Arabic, Turkish, Finnish, Hungarian, and many other lan-guages.

Label Noise and Relation Transitivity

Annotation Inconsistency: Human annotators are more likely to make mistakes in long documents due to fatigue and difficulty maintaining consistency across pages.

Transitivity Issues: If a model learns that entity A relates to entity B, and B relates to C, can it infer a relationship between A and C? This type of transitive reasoning is crucial in legal documents, financial filings, and scientific literature. Current models lack explicit mechanisms for transitive reasoning.

Cascading Errors: In long documents, early errors can compound. If a model incorrectly identifies an entity on page 1, all subsequent references may also be misclassified.

Re-evaluating Evaluation Metrics

The standard metric for NER is token-level F1-score. However, this metric may inadequately capture long-range performance. A model might excel at identifying entities with strong local context while failing on entities requiring long-distance reasoning, with both contributing equally to F1-score and obscuring this distinction.

There is a need for more nuanced evaluation metrics and benchmark datasets specifically designed to probe a model's ability to handle long-range dependencies. Proposed improve-ments include coreference-aware metrics, reasoning-specific benchmarks, and error analysis frameworks that distinguish between local and global errors.

FUTURE DIRECTIONS AND THE VISION FOR HYBRID MODELS

Synergistic Hybrid Architectures

The limitations of unimodal approaches strongly suggest that the future of long-range NER lies in hybrid models [20]. The two dominant paradigms—efficient Transformers and GNNs—possess complementary strengths. A promising conceptual blueprint involves:

1. **Local Context Encoding:** An efficient Transformer like Longformer or BigBird processes the entire document to generate powerful, locally-aware token embeddings.

2. **Structural Scaffolding:** A document-level graph is con-structed with nodes representing candidate entity men-tions and sentences, and edges representing relationships like lexical similarity, syntactic dependency, or discourse coherence.

3. **Global Information Propagation:** A Graph Attention Network (GAT) operates on this graph, refining node representations by propagating information between dis-tant but related parts of the document.

4. **Feature Fusion and Decoding:** The locally-rich em-beddings from the Transformer and globally-aware em-beddings from the GNN are fused and passed to a final decoder to produce entity labels.

Such a hybrid approach could directly address the reasoning gap by allowing information to flow along structured relational paths, enabling the model to connect a pronoun to a distant antecedent via an explicit graph edge.

Reinforcement Learning for Information Extraction

Reinforcement Learning (RL) offers a paradigm where an agent can learn a policy for extracting entities, making deci-sions about where to look for information in a long document. This could be particularly useful for tasks that require multi-step reasoning or information-seeking behaviors. RL could learn to prioritize certain sections, follow cross-references, and adaptively allocate computational resources.

Interpretable and Explainable Document-Level NER

As models become more complex, understanding why they make a certain prediction becomes critical, especially in high-stakes domains like medicine and law [19]. Future research needs to focus on making these massive document-level mod-els more interpretable, perhaps by highlighting the evidence trails or relational paths used to identify an entity. Techniques include attention visualization, rationale extraction, counter-factual explanations, and human-in-the-loop systems.

Multi-Document and Cross-Document NER

Future systems will need to handle not just single long documents, but collections of related documents. This re-quires understanding how entities and information flow across document boundaries, maintaining consistent entity resolution across multiple sources, and aggregating information from heterogeneous document types.

CONCLUSION

The journey to solve Named Entity Recognition in long documents is a microcosm of the broader push towards true machine reading comprehension. We have progressed from brittle, rule-based systems to powerful but limited deep learning models, and now to architectures that can process thousands of tokens at once. This survey has charted that evolution, detailing the innovations in efficient Transformers and graph-based networks that define the current state-of-the-art.

We have also synthesized the critical remaining challenges, from the deep-seated problem of long-distance reasoning to the practical necessities of domain adaptation and learning from scarce data. By embracing hybrid architectures that synergize the best of sequential and structural modeling, and by devel-oping more nuanced methods for training and evaluation, the field is poised to unlock the vast stores of knowledge currently trapped in the world's long-form documents.

The integration of efficient Transformers with graph-based reasoning, coupled with advances in weakly supervised learn-ing and cross-lingual adaptation, promises to democratize access to sophisticated NER capabilities across domains and languages. The ultimate goal remains: enabling machines to understand and extract knowledge from documents with the same depth and nuance as human experts.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable feedback and suggestions that helped im-prove this survey.

1. X. Luan et al., "Sparse Learning for Long Document Named Entity Recognition," in Proc. EMNLP, 2023.
2. Y. Yamada et al., "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention," in Proc. EMNLP, 2020.
3. X. Wang et al., "Named Entity Recognition for Long Text based on Multi-stage Attention," in Proc. NeurIPS, 2021.
4. X. Yao et al., "DocRED: A Large-Scale Document-Level Relation Extraction Dataset," in Proc. ACL, 2019.

5. R. Li et al., "Adaptive Attention Span for Document-Level Information Extraction," in Proc. NeurIPS, 2020.

6. J. Lin et al., "A Survey of Long Document Understanding," arXiv preprint arXiv:2107.06276, 2021.

7. K. Sahu et al., "Weak Supervision for Domain Adaptation in Clinical Named Entity Recognition," in Proc. AAAI, 2023.

8. S. Wang et al., "Cross-lingual Transfer and Adaptation for Marathi Named Entity Recognition," in Proc. COLING, 2024.

9. X. Zhang et al., "Multi-hop Reasoning for Document-level Machine Reading Comprehension and NER," in Proc. ACL, 2023.

D. Logan et al., "Explainability in Document-Level Information Extraction Models," in Proc. NAACL, 2022.

10. L. Chen et al., "Hybrid Architectures for Long-Document Named Entity Recognition," in Proc. EMNLP, 2024.

Vaswani et al., "Attention is All You Need," in Proc. NeurIPS, 2017.

11. N. Kitaev, L. Kaiser, and A. Vaswani, "Reformer: The Efficient Trans-former," in Proc. ICLR, 2020.

12. X. Shen et al., "Effective Weak Supervision for Medical Named Entity Recognition via a Multi-stage Label Refinement Framework," in Proc. ACL, 2023.

13. J. R. Uijlings et al., "Graph Neural Networks for Relation Extraction and Document-Level NER," in Proc. EMNLP, 2021.

14. F. Xia et al., "DocFormer: End-to-End Transformer for Document Understanding," arXiv preprint arXiv:2109.10260, 2021.

15. L. Cui et al., "Long-range Context Modeling with Graph Convolutional Networks for Document-level NER," in Proc. AAAI, 2023.

16. T. Wang et al., "Document-level Named Entity Recognition via Multi-granularity Attention," in Proc. COLING, 2022.

17. S. Chen et al., "Long Text Named Entity Recognition using Attention with Local and Global Context," in Proc. IJCAI, 2023.

18. Z. Chen et al., "Semi-supervised Learning for Long Document NER Leveraging Weak Labels," in Proc. ACL, 2024.

19. Y. Guo et al., "Exploring Reasoning Over Long Sequences for Document-Level Named Entity Recognition," in Proc. NeurIPS, 2023.

REFERENCES

1. I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020.
2. Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, "Clinical-Longformer and Clinical-BigBird: Transformers for Long Clinical Se-quences," Feinberg School of Medicine, Northwestern University, 2023.
3. M. Zaheer et al., "Big Bird: Transformers for Longer Sequences," in Proc. NeurIPS, 2020.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue II February2026

4. H. Zheng, S. Wang, and L. Huang, “A Comprehensive Survey on Document-Level Information Extraction,” Virginia Tech and University of California, Davis, 2024.
5. H. Kim, J. Yoo, S. Yoon, and J. Kang, “Automatic Creation of Named Entity Recognition Datasets by Querying Phrase Representations,” Korea University and Adobe Research, 2023.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL, 2019.
7. P. Velicˇkovic´ et al., “Graph Attention Networks,” in Proc. ICLR, 2018.
8. M. Peters et al., “Deep contextualized word representations,” in Proc. NAACL, 2018.
9. S. Xu et al., “GraphFormer: A Graph Transformer Architecture for NER,” in Proc. ACL, 2022.