

# E2RL : A Framework for Entity Extraction, Resolution and Linking

Shashi Prakash Tripathi

**A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores  
University for the degree of Doctor of Philosophy**

**June/2025**

## Table of Contents

<b>1. Introduction .....</b>	<b>11</b>
<b>1.1 Context and Motivation .....</b>	<b>11</b>
<b>1.2 Problem Description .....</b>	<b>12</b>
1.2.1 Challenges and Considerations .....	17
<b>1.3 The E2RL Framework .....</b>	<b>18</b>
1.3.1 Paper 1 (The framework) .....	20
<b>1.4 Literature Review .....</b>	<b>20</b>
<b>1.5 Structure of the Thesis .....</b>	<b>22</b>
<b>2. Entity Extraction and Resolution in E2RL .....</b>	<b>23</b>
<b>2.1 Introduction .....</b>	<b>23</b>
<b>2.2 Pre-processing .....</b>	<b>23</b>
2.2.1 Wordnet based pre-processing (Paper 8) .....	23
<b>2.3 Entity Extraction in E2RL .....</b>	<b>23</b>
2.3.1 Entity Identification (Paper 2) .....	23
2.3.2 Multiword Entity Recognition and Resolution (Paper 3) .....	24
2.3.3 Performance Metrics for Extraction .....	26
<b>2.4 Entity Resolution in E2RL .....</b>	<b>26</b>
2.4.1 Tree-Based Resolution .....	26
2.4.2 Probabilistic Resolution .....	26
2.4.3 Evaluation Metrics for Resolution .....	27
<b>2.5 Discussion and Findings .....</b>	<b>27</b>
<b>3. Entity Linking in E2RL Framework .....</b>	<b>28</b>
<b>3.1 Introduction .....</b>	<b>28</b>
<b>3.2 Entity Linking .....</b>	<b>28</b>
3.2.1 Local and Global Feature Based Link Prediction (Paper 4 and Paper 5) .....	28
3.2.2 Improve link prediction using Skip-gram model (Paper 6) .....	31
3.2.3 Multifaceted Neighbourhood Approach (Paper 7) .....	33
3.2.4 Centrality based Link Prediction (Paper 10) .....	35
3.2.5 Evolutionary Optimization (Paper 9) .....	36
<b>3.3 Evaluation Metrics .....</b>	<b>40</b>
<b>3.4 Discussion and Findings .....</b>	<b>40</b>
<b>4. Comprehensive Evaluation of the E2RL Framework .....</b>	<b>42</b>
<b>4.1 Introduction .....</b>	<b>42</b>
<b>4.2 Methodology .....</b>	<b>42</b>
4.2.1 Entity Extraction Algorithms .....	42
4.2.2 Entity Resolution Algorithms .....	43
4.2.3 Entity Linking Algorithms .....	43
4.2.4 Evaluation Metrics .....	43
<b>4.3 Datasets .....</b>	<b>44</b>
<b>4.4 Results .....</b>	<b>44</b>
4.4.1 Entity Extraction Performance Across Algorithms .....	44
4.4.2 Entity Resolution Performance Across Algorithms .....	45

4.4.3 Entity Linking Performance Across Algorithms .....	45
4.4.4 Comprehensive Performance Metrics .....	46
4.4.5 Precision and Recall Analysis .....	47
4.4.6 Computational Efficiency.....	50
<b>4.5 Discussion .....</b>	<b>51</b>
4.5.1 Entity Extraction.....	52
4.5.2 Entity Resolution .....	52
4.5.3 Entity Linking.....	53
4.5.4 Impact of Dataset Characteristics .....	53
4.5.5 Trade-offs Between Accuracy and Efficiency .....	53
<b>4.6 Conclusion .....</b>	<b>54</b>
<b>5. Contributions, Conclusion and Future Work.....</b>	<b>55</b>
5.1 Introduction .....	55
5.2 Entity Extraction: Enhancing Precision and Recall .....	55
5.3 Entity Resolution: Robust Reconciliation Across Sources .....	55
5.4 Entity Linking: Unlocking Semantic Relationships.....	55
5.5 Architectural Design: Adaptability and Scalability .....	56
5.6 Theoretical and Practical Contributions .....	56
5.7 Addressing Key Challenges.....	56
5.8 Integrated Framework and Layered Contributions .....	57
5.9 Conclusion .....	58
5.10 Future Work .....	59
<b>Appendices.....</b>	<b>69</b>

## List of Figures

Figure 1: Architecture of the E2RL Framework (Paper 1) .....	19
Figure 2: Process Flow of TransCRF .....	25
Figure 3: Entity Linking .....	38
Figure 4: F1-Score Comparison of Entity Extraction and Resolution Algorithms Across Datasets.....	46
Figure 6: Precision and Recall of Entity Resolution Algorithms on DBLP-ACM.....	48
Figure 7: AUC-ROC of Entity Extraction and Resolution Algorithms Across Datasets.....	49
Figure 8: Processing Time Comparison of Entity Extraction and Resolution Algorithms .....	51

## List of Tables

Table 1: Entity Type & Extracted Entity .....	13
Table 2: Entity Extraction Outcome of the Example .....	16
Table 3: Entity Extraction Performance Metrics Across Datasets .....	45
Table 4: Entity Resolution Performance Metrics Across Datasets .....	45
Table 5: Entity Linking Performance Metrics Across Datasets .....	46
Table 6: Processing Time Comparison of Algorithms for Entity Extraction .....	50
Table 7: Processing Time Comparison of Algorithms for Entity Resolution.....	50

## List of Papers

No.	Papers
Paper 1	Applying Model View View-Model and Layered Architecture for Mobile Applications
Paper 2	SimNER--An Accurate and Faster Algorithm for Named Entity Recognition
Paper 3	TransCRF—Hybrid Approach for Adverse Event Extraction
Paper 4	Hybrid approach for predicting and recommending links in social networks
Paper 5	Hybrid feature-based approach for recommending friends in social networking systems
Paper 6	Network embedding based link prediction in dynamic networks
Paper 7	A novel similarity-based parameterized method for link prediction
Paper 8	An Enhanced Approach of Pre-processing the Document using WordNet in Text Clustering
Paper 9	Optimizing constrained engineering problem nH-WDEOA: using hybrid nature-inspired algorithm
Paper 10	A Novel Similarity-Based Method for Link Prediction in Complex Networks

## Declaration

This submission by Shashi Prakash Tripathi to the Doctoral Academy at Liverpool John Moores University is for a PhD by Published Works. It consists of ten outputs and a synthesis statement and is in partial fulfilment of the requirements set out by the Doctoral Academy at Liverpool John Moores University, in accordance with its guidelines and regulations. The authored texts were published in peer-reviewed journals, conference proceedings, or books and have not been used in submissions for any other research degree. No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification at this or any other university or institute of learning.

A handwritten signature in purple ink that reads "Shashi".

Shashi Prakash Tripathi  
December 22, 2024

## Acknowledgements

Completing this thesis marks a significant milestone in my academic and professional journey, and it would not have been possible without the support and guidance of numerous individuals and institutions.

First and foremost, I extend my deepest gratitude to my supervisory team — Dr. Bo Zhou, Yun Sheng, Wasiq Khan, and Peter Atherton. Your unwavering support, insightful feedback, and constant encouragement have been invaluable throughout this endeavor. Your expertise and dedication have profoundly shaped the direction and quality of this work.

I am also immensely grateful to my co-authors and guide — Prof. R. R. Tewari, Dr. Abhay Kumar Rai, Dr. Tulika Narang, Rahul Kumar Yadav, Pawan Mishra, Shuchi Tripathi, Vaishali Srivastava and Harshita Rai. Your constructive critiques and suggestions have significantly enhanced the depth and rigor of my research, and your collaborative spirit has been a cornerstone of this thesis. Special thanks go to my colleagues and friends in the AI Practices group. Your camaraderie and intellectual exchanges have made the research process both enjoyable and intellectually stimulating. The discussions and shared experiences have been a constant source of inspiration.

I would like to acknowledge the support of Liverpool John Moores University, particularly the Doctoral Academy and the School of Computer Science and Mathematics. Your provision of necessary resources and a conducive research environment has been crucial for the successful execution of this study. Access to advanced computational tools and datasets has greatly facilitated my work. My sincere appreciation also goes to Tata Consultancy Services for their financial support. This assistance has enabled me to focus entirely on my research without the burden of financial constraints, allowing for a more dedicated and thorough investigation.

On a personal note, I am profoundly thankful to my family for their unconditional love, patience, and encouragement. Your belief in me has been my driving force, especially during the most challenging times of this journey.

Lastly, I extend my gratitude to all the authors and researchers whose work has laid the foundation for this thesis. Your contributions to the fields of entity extraction, resolution, and linking have been instrumental in shaping my research path and inspiring the developments presented in this work.

## Portfolio of Published Work

### List of Published Works

1. [Tripathi, S. P., & Narang, T. \(2016\). Applying Model View View-Model and Layered Architecture for Mobile Applications. \*Journal of International Academy of Physical Sciences\*, 20\(3\), 215–221.](#)
2. [Tripathi, S. P., & Rai, H. \(2018\). SimNER--An Accurate and Faster Algorithm for Named Entity Recognition. \*2018 Second International Conference on Advances in Computing, Control and Communication Technology \(IAC3T\)\*, 115–119. IEEE](#)
3. [Rai, H., Tripathi, S. P., & Narang, T. \(2022\). TransCRF—Hybrid Approach for Adverse Event Extraction. \*Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022\*, 1–10. Springer Nature Singapore Singapore.](#)
4. [Tripathi, S. P., Yadav, R. K., Rai, A. K., & Tewari, R. R. \(2019\). Hybrid approach for predicting and recommending links in social networks. \*Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017\*, 107–119. Springer](#)
5. [Yadav, R. K., Tripathi, S. P., Rai, A. K., & Tewari, R. R. \(2020\). Hybrid feature-based approach for recommending friends in social networking systems. \*International Journal of Web Based Communities\*, 16\(1\), 51–71.](#)
6. [Tripathi, S. P., Yadav, R. K., & Rai, A. K. \(2022a\). Network embedding based link prediction in dynamic networks. \*Future Generation Computer Systems\*, 127, 409–420.](#)
7. [Rai, A. K., Tripathi, S. P., & Yadav, R. K. \(2023\). A novel similarity-based parameterized method for link prediction. \*Chaos, Solitons & Fractals\*, 175, 114046.](#)
8. [Shashi Prakash Tripathi, T. N. \(2016\). An Enhanced Approach of Preprocessing the Document using WordNet in Text Clustering. \*International Conference on Control Computing Communication and Materials \(ICCCCM-2016\)\*, 5.](#)
9. [Mishra, P. and Tripathi, S.P., 2024. Optimizing constrained engineering problem nH-WDEOA: using hybrid nature-inspired algorithm. \*International Journal of Information Technology\*, pp.1-9.](#)
10. [Rai, A.K., Yadav, R.K., Tripathi, S.P., Singh, P. and Sharma, A., 2023, November. A Novel Similarity-Based Method for Link Prediction in Complex Networks.](#)

#### Pending Work

1. Shashi Prakash Tripathi. From Static to Recursive: Transforming Prompts for Enhanced Language Models, 12 December 2023, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-3639349/v1>]

#### Not Selected

1. Tripathi, S. P., Srivastava, V., & Rai, H. (2016). Improvised Master's Theorem. *International Research Journal of Engineering and Technology (IRJET)*, 3(05),
2. Shashi Prakash Tripathi, H. M., Shuchi. (2017). A Comparative Study of Data Clustering Techniques. *International Research Journal of Engineering and Technology (IRJET)*, 4(5), 1392–1398.
3. Tripathi, S. P., Yadav, R. K., & Rai, H. (2022b). WeedNet: A deep neural net for weed identification. In *Deep Learning for Sustainable Agriculture* (pp. 223–236). Elsevier.
4. Abhay Kumar Rai, Rahul Kumar Yadav, Shashi Prakash Tripathi, Rajiv Ranjan Tewari. "A Survey on Link Prediction Problem in Social Networks", Volume 5, Issue IX, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 1875-1883, ISSN : 2321-9653

The Entity Extraction, Resolution and Linking (E2RL) framework draws inspiration from a set of pivotal peer-reviewed papers that, while not directly incorporated into the framework, have profoundly influenced its conceptual underpinnings. The Enhanced Master's Theorem (Tripathi, Srivastava and Rai, 2016) is like a guidebook for computer problem-solving. It helps us understand how long it might take for a computer to finish certain tasks. It's about figuring out how efficiently a computer can handle specific jobs. A Comparative Study of Data Clustering Techniques (Shashi Prakash Tripathi Shuchi, 2017) contributes essential insights into effective clustering methods, shaping the robust entity resolution and clustering aspects of E2RL.

WeedNet (Shashi Prakash Tripathi, Yadav and H. Rai, 2022), a deep learning model for weed identification, serves as motivation for handling textual data in E2RL through deep learning techniques. An Enhanced Approach to Document Preprocessing using WordNet (Shashi Prakash Tripathi, 2016a) guides the E2RL framework in efficient textual data handling and entity recognition. Finally, a Survey on Link Prediction in Social Networks



(Rai *et al.*, 2017) contributes to the broader understanding of network relationships, influencing the design of E2RL's linking component for entity association.

## Abstract

Extracting information from large volume of text and inferring or storing insights is nowadays has become a much-needed process. Entity Extraction, Resolution and linking using knowledge graphs is a critical process that facilitates the accurate identification and association of entities across diverse datasets, thereby enhancing information retrieval, data integration, and intelligent decision-making systems. This doctoral research introduces the Entity Extraction, Resolution, and Linking (E2RL) Framework, a comprehensive solution designed to address the multifaceted challenges of information extraction. The E2RL Framework comprises three interconnected submodules: Entity Extraction, Resolution, and Linking. In the Entity Extraction submodule, innovative approaches such as TransCRF(Rai, Tripathi and Narang, 2022a) and SimNER (Tripathi and Rai, 2018a) are employed to identify entities within textual data with high precision and contextual understanding. The Entity Resolution submodule leverages both Tree-based and Probabilistic Approaches to reconcile and associate entities across different data sources, ensuring consistency and accuracy in entity representation. For establishing meaningful connections between identified entities, the Linking submodule integrates advanced graph linking algorithms, including FriendREC (Tripathi *et al.*, 2019) and LinkVec (S P Tripathi, Yadav and Rai, 2022) and other methods, which enhance the relational structure of knowledge graphs.

The architecture of E2RL adopts the Model-View-ViewModel (MVVM) pattern and a layered structure (Tripathi and Narang, 2016a), promoting scalability, maintainability, and flexibility for future enhancements. This modular design facilitates seamless information retrieval, efficient data indexing for semantic searching, and the extraction of pivotal information, thereby supporting robust entity linking processes. The research methodology encompasses an extensive literature review to identify gaps and advancements in existing Entity Extraction, Resolution, and Linking methodologies. A conceptual framework is developed to define the core components and their interrelationships, followed by the design and implementation of tailored algorithms for each submodule. Rigorous theoretical underpinnings and validation procedures ensure the robustness and validity of the proposed framework.

Comprehensive evaluations of the E2RL Framework demonstrate its versatility and effectiveness in tackling complex information extraction challenges across various domains. The framework's adaptability makes it a valuable tool for researchers and practitioners, capable of integrating contextual embeddings, adapting to dynamic data changes, and scaling efficiently to incorporate external knowledge bases. Additionally, the E2RL Framework is optimized for real-time entity extraction, resolution, and linking, ensuring timely and accurate results for applications requiring instant data processing.

In conclusion, the E2RL Framework offers a robust, scalable, and adaptable solution for enhancing Entity Extraction, Resolution and linking. By integrating advanced pre-processing techniques, sophisticated optimization algorithms, and innovative link prediction methodologies, E2RL effectively addresses critical challenges such as entity ambiguity, scalability, and computational efficiency. This research not only contributes valuable methodologies to the academic discourse but also provides a practical framework for real-world applications, paving the way for more intelligent and efficient information systems.

# 1. Introduction

## 1.1 Context and Motivation

In today's digital age, the exponential growth of data has led to an unprecedented era of information overload. Organizations and individuals alike are inundated with vast volumes of both structured and unstructured data generated from a myriad of sources such as social media platforms, online transactions, scientific research, and enterprise systems. This deluge of information presents both opportunities and challenges. On one hand, the ability to harness and analyse this data can lead to significant advancements in various fields, from business intelligence to scientific discovery. On the other hand, the sheer volume and diversity of data make it increasingly difficult to extract meaningful and actionable insights efficiently.

Amidst this backdrop, the processes of Entity Extraction, Resolution, and Linking (E2RL) have emerged as critical components in the landscape of modern data processing and information retrieval systems. These foundational techniques are essential for transforming raw data into structured, interpretable, and interconnected information that can drive informed decision-making.

The significance of E2RL extends across a wide array of applications and industries. In search engines, these techniques underpin the ability to deliver precise and contextually relevant search results, improving user experience and satisfaction. Recommender systems leverage E2RL to analyse user preferences and behaviours, enabling personalized content suggestions that enhance engagement and retention. In the realm of intelligence gathering, E2RL framework facilitates the aggregation and analysis of information from disparate sources, supporting strategic decision-making and threat assessment. Furthermore, the construction and maintenance of knowledge graphs, which are essential for organizing and representing complex information networks, rely heavily on the effective implementation of E2RL framework processes.

Beyond these applications, E2RL framework serves as the backbone for numerous other domains that require efficient data processing and insightful information retrieval. In healthcare, for example, E2RL framework can aid in the aggregation of patient records from various sources, enhancing the accuracy of diagnoses and treatment plans. In finance, it supports the analysis of market trends and risk assessment by linking financial entities across different datasets. Moreover, in the field of artificial intelligence, E2RL Framework contributes

to the development of more intelligent and context-aware systems capable of understanding and interacting with human language and behaviour.

The critical importance of E2RL Framework in enabling efficient decision-making cannot be overstated. As organizations strive to become more data-driven, the ability to accurately extract, resolve, and link entities becomes essential for transforming raw data into strategic assets. E2RL Framework not only enhances the quality and reliability of data but also unlocks its potential by revealing hidden patterns, relationships, and insights that can drive innovation and competitive advantage.

## 1.2 Problem Description

The process of managing and interpreting vast amounts of data in today's information-rich environment hinges on the effective identification, consolidation, and interconnection of entities within that data. Entity Extraction(Gupta *et al.*, 2014a), Resolution, and Linking(Zhang, Liu and He, 2021) form a triad of critical processes that address these challenges, each playing a distinct yet interrelated role in transforming raw data into meaningful and actionable insights. This section delves into the intricacies of each component, elucidating the complexities involved and underscoring the necessity of a robust E2RL framework.

**Entity Extraction** is the foundational step in the E2RL framework, involving the identification and isolation of relevant entities from unstructured or semi-structured text. Entities can encompass a wide range of elements, including names of people, organizations, locations, dates, numerical values, and more specialized terms depending on the domain of the data. The primary challenge in entity extraction lies in accurately discerning these entities amidst the noise and variability inherent in natural language.

A critical aspect of effective entity extraction is contextual understanding. An entity recognized in one context or domain may hold different significance or may not be relevant in another. For example, the term "Apple" could refer to the technology company or the fruit, depending on the context. Therefore, sophisticated extraction techniques must not only identify entities but also comprehend the surrounding context to ensure accurate classification and relevance.

Entity extraction typically involves tagging sequences of tokens (words or phrases) within the text as entities. This tagging process leverages various natural language processing (NLP) techniques, including part-of-speech tagging, named entity recognition (NER), and machine learning algorithms trained on annotated datasets. Once identified, these entities are retrieved and prepared for subsequent analysis stages, such as resolution and linking.

### Example of Entity Extraction:

Consider the following sentence:

*"Dr. Jane Smith, a renowned researcher at Stanford University, presented her latest findings on artificial intelligence at the Global Tech Conference held in Berlin on September 15, 2023."*

### Extraction Results :

Entity Type	Extracted Entity
Person	Dr. Jane Smith
Organization	Stanford University
Event	Global Tech Conference
Location	Berlin
Date	September 15, 2023
Field of Study	artificial intelligence

Table 1: Entity Type & Extracted Entity

After extraction, **Entity Resolution** tackles the inherent ambiguity and redundancy that arises when multiple representations refer to the same underlying entity. This process is vital for ensuring data consistency and integrity, particularly when aggregating information from diverse sources where naming conventions and terminologies may vary.

A quintessential example of entity resolution is the consolidation of different representations of the same organization: "IBM" and "International Business Machines." Without resolution, these distinct labels would be treated as separate entities, leading to fragmented and inconsistent data. Entity resolution algorithms employ various techniques, such as string

matching, semantic similarity assessments, and the use of unique identifiers, to accurately determine when different references pertain to the same entity.

Additionally, entity resolution must address the resolution of pronouns and indirect references within the text. For instance, in a narrative where "she" refers to "Dr. Jane Smith," the system must map the pronoun to the correct antecedent to maintain coherence and accuracy in data representation.

Effective entity resolution not only eliminates redundancies but also enhances the clarity and usability of the data, facilitating more precise analysis and decision-making.

### **Example of Entity Resolution:**

Continuing with the previous example, consider multiple references within a larger text:

*"Dr. Jane Smith, a renowned researcher at Stanford University, presented her latest findings on artificial intelligence at the Global Tech Conference held in Berlin on September 15, 2023. IBM, also present at the conference, showcased their advancements in machine learning. Jane further collaborated with IBM to integrate these new techniques into her research."*

### **Resolution Results:**

- **Dr. Jane Smith** and **Jane** are resolved to the same person.
- **IBM** and **International Business Machines** (if mentioned elsewhere) are recognized as the same organization.
- **Artificial Intelligence** and **machine learning** may be linked as related fields depending on the resolution strategy.

The final component, **Entity Linking**(Y. Wu *et al.*, 2020; Shi *et al.*, 2023), extends the capabilities of extraction and resolution by establishing meaningful connections between entities both within a single dataset and across multiple datasets. This interconnectedness is crucial for constructing comprehensive knowledge graphs that provide a holistic view of the information landscape.

Entity linking involves mapping extracted and resolved entities to their corresponding representations in external knowledge bases or structured repositories, such as Wikipedia,

DBpedia, or proprietary databases. By doing so, it enables the integration of information from disparate sources, facilitating richer and more contextually aware analyses.

For instance, linking the entity "Global Tech Conference" to a specific event in a knowledge base allows for the aggregation of all related information about that conference, including past events, participating organizations, keynote speakers, and associated research topics. This interconnectedness not only enhances the depth of information available but also supports advanced applications like semantic search, recommendation systems, and intelligent data visualization.

Moreover, entity linking contributes to the creation of cohesive narratives by connecting related entities across different sections of a dataset. For example, linking an event mentioned in one paragraph to an organization or individual mentioned elsewhere in the document helps maintain continuity and context, thereby enriching the overall data representation.

### **Example of Entity Linking:**

Building upon the previous examples, suppose we have access to external knowledge bases that provide detailed information about the entities mentioned.

- **Linking "Stanford University"** to its corresponding entry in Wikipedia or an academic knowledge base, providing additional data such as its location, notable programs, and affiliated researchers.
- **Linking "Global Tech Conference"** to a specific event record, including details like event history, themes, and participating entities.
- **Linking "Dr. Jane Smith"** to a professional profile that includes her publications, research interests, and collaborations.

This interconnected web of linked entities enables users to traverse the knowledge graph(Chen *et al.*, 2021; Ristoski, Lin and Zhou, 2021; Zhang *et al.*, 2024) seamlessly, accessing a wealth of information that is both comprehensive and contextually relevant.

To illustrate the synergy between entity extraction, resolution, and linking, consider the following comprehensive example:

### **Sample Text:**

*"At the 2024 International Conference on Data Science held in New York on March 10, Dr. Emily Zhang from MIT unveiled a ground breaking algorithm for data mining. The algorithm, developed in collaboration with Google, promises to enhance data processing speeds by 50%. During the conference, Emily also discussed her previous work with IBM on machine learning applications."*

### Step 1: Entity Extraction

Entity Type	Extracted Entity
Event	2024 International Conference on Data Science
Location	New York
Date	March 10, 2024
Person	Dr. Emily Zhang
Organization	MIT
Organization	Google
Organization	IBM
Field of Study	data mining, machine learning
Metric	50%

Table 2: Entity Extraction Outcome of the Example

### Step 2: Entity Resolution

- **Dr. Emily Zhang** is identified as a unique individual.
- **MIT, Massachusetts Institute of Technology**, are recognized as the same organization.
- **Google** and **IBM** are resolved to their respective unique corporate identities.
- **data mining** and **machine learning** are identified as related fields within data science.

### Step 3: Entity Linking

- **2024 International Conference on Data Science** is linked to its official event page, providing details about the conference agenda, speakers, and historical data.
- **Dr. Emily Zhang** is linked to her professional profile, including her publications and research projects at MIT.



- **MIT, Google, and IBM** are linked to their respective entries in a knowledge base, offering comprehensive information about each organization.
- The **algorithm for data mining** is linked to a repository detailing its specifications, applications, and performance metrics.

### Resulting Knowledge Graph Excerpt:

The below example shows the as is extraction of raw data without any mapping of knowledge base or ontology for given text.

```
[2024 International Conference on Data Science] --held_in→ [New York]
[2024 International Conference on Data Science] --date→ [March 10, 2024]
[Dr. Emily Zhang] --affiliated_with→ [MIT]
[Dr. Emily Zhang] --developed→ [Algorithm for Data Mining]
[Algorithm for Data Mining] --collaborated_with→ [Google]
[Algorithm for Data Mining] --improves→ [Data Processing Speeds by 50%]
[Dr. Emily Zhang] --discussed_work_with→ [IBM]
```

Above snippet demonstrates how E2RL Framework processes work in tandem to transform a raw text passage into a structured, interconnected knowledge graph. Entity extraction identifies the key elements within the text, entity resolution ensures that each entity is uniquely and accurately represented, and entity linking connects these entities to broader information repositories, thereby enriching the data and enabling more sophisticated analyses.

#### 1.2.1 Challenges and Considerations

Implementing an effective E2RL framework involves navigating several challenges:

- Ambiguity and Variability:** Natural language is inherently ambiguous, and entities can be represented in multiple forms. Distinguishing between homonyms and handling variations in entity names require advanced disambiguation techniques.
- Scalability:** As data volumes grow, Entity Extraction, Resolution and Linking systems must scale efficiently to handle large datasets without compromising performance or accuracy.

- c) **Domain-Specific Knowledge:** Different domains may have unique entities and terminologies. Tailoring E2RL Framework processes to accommodate domain-specific nuances is essential for accurate extraction, resolution, and linking.
- d) **Real-Time Processing:** In applications requiring real-time data analysis, such as live social media monitoring or financial trading systems, E2RL systems must process and link entities swiftly to provide timely insights.
- e) **Integration with Existing Systems:** E2RL frameworks need to seamlessly integrate with existing data processing pipelines and knowledge management systems to enhance their functionality without disrupting established workflows.

Addressing these challenges necessitates ongoing research and the development of sophisticated algorithms that leverage advancements in machine learning, natural language processing, and data integration technologies.

The problem of effectively extracting, resolving, and linking entities within large and diverse datasets is pivotal for unlocking the full potential of data-driven insights. The E2RL framework offers a structured approach to navigate the complexities of entity management, ensuring that data is not only accurately represented but also richly interconnected. By overcoming the challenges associated with ambiguity, scalability, and domain specificity (Mahata *et al.*, 2019a), E2RL enables the construction of robust knowledge graphs and supports a wide array of applications across various industries. As data continues to proliferate, the importance of sophisticated E2RL systems will only intensify, underscoring the need for continued innovation and refinement in this critical area of data science.

### 1.3 The E2RL Framework

The Entity Extraction, Resolution, and Linking (E2RL) framework is designed to address these challenges comprehensively. It encapsulates multiple algorithms tailored to tackle each component effectively. Drawing inspiration from our pivotal peer-reviewed works, the E2RL framework integrates advanced techniques to enhance accuracy, scalability, and adaptability across diverse domains. Some of the not selected articles are not directly incorporated, they have profoundly influenced the conceptual underpinnings of the framework.

By combining state-of-the-art natural language processing, machine learning, and graph-based methodologies, the E2RL framework provides a modular and scalable solution for

handling complex datasets. Its robust design ensures adaptability to various real-world applications, such as healthcare analytics, fraud detection, and personalized recommendations.

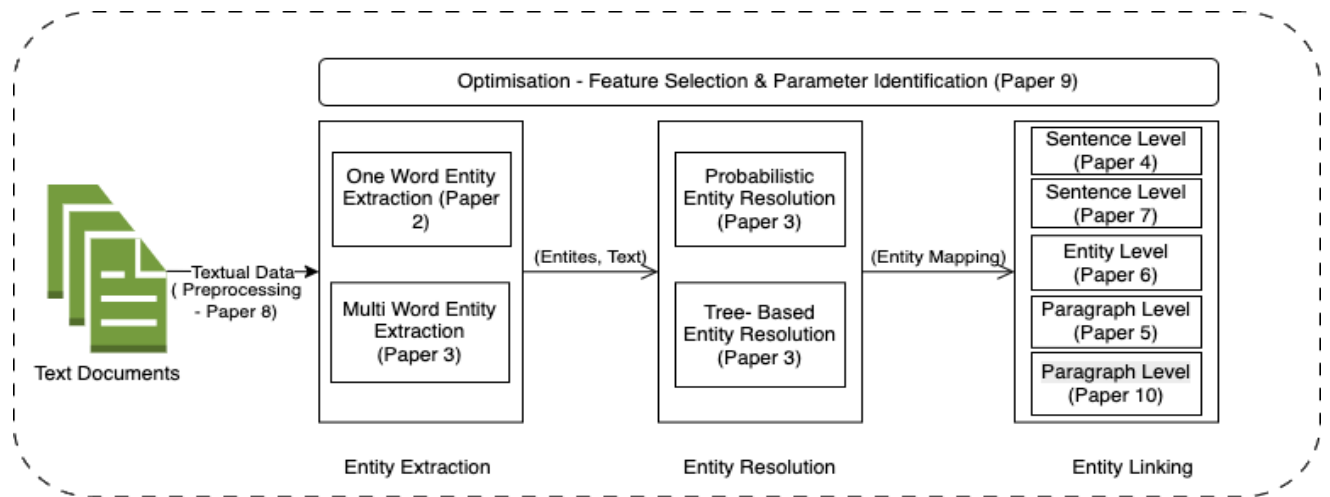


Figure 1: Architecture of the E2RL Framework (Paper 1)

In Figure 1 above, every box refers to a particular process of the E2RL and the papers are being used to solve that problem/challenge.

The overall research questions can be summarised as follows:

- How can contextual embeddings be seamlessly integrated into the E2RL framework to enhance entity recognition with contextual understanding?
- In what ways can E2RL adapt to dynamic changes in data and entities, ensuring robust and accurate linking over time?
- What scalable approaches can be implemented to integrate external knowledge bases effectively, improving entity recognition across diverse domains within the E2RL framework?
- How can E2RL be optimized for real-time entity extraction, resolution and linking, ensuring timely and accurate results for applications requiring instant data processing?

### 1.3.1 Paper 1 (The framework)

We (Tripathi and Narang, 2016b) introduced a comprehensive framework which is the core framework for E2RL, characterized by its layered and modular structure. Within this multi-layered architecture (Medvidovic and Edwards, 2010), distinct modules are employed to handle various aspects of entities, including Recognition, Resolution, and Linking. The framework is organized into different layers, namely the Data Layer, Presentation Layer, and Application Layer (Singh *et al.*, 2012), each catering to specific objectives.

A key feature of the ontological representation within this framework is its emphasis on the self-referential structure of Application and Function classes, highlighting their interdependence for specific operations. This ontological representation aids in standardizing the framework across domains and contributes to the effective management of different attributes of entities. Additionally, the framework serves as a valuable tool for tracking, tracing, and improving overall process outcomes by optimizing various processes within it.

The paper further provides insights into the advantages of integrating MVVM design patterns with layered architectures (Meier, Homer and Hill, 2008; ZadahmadJafarlou, Arasteh and YousefzadehFard, 2011; Singh *et al.*, 2012), promoting modularity and cohesion within applications. The ontological perspective underscores the adaptability of the proposed framework across diverse platforms, positioning it as a significant contribution to the broader field of Entity Recognition, Resolution, and Linking.

By combining state-of-the-art natural language processing, machine learning, and graph-based methodologies, the E2RL framework provides a modular and scalable solution for handling complex datasets. Its robust design ensures adaptability to various real-world applications, such as healthcare analytics, fraud detection, and personalized recommendations.

## 1.4 Literature Review

Entity Extraction, commonly referred to as Named Entity Recognition (NER), is a foundational task in information extraction that involves identifying and classifying entities in text into predefined categories such as persons, organizations, and locations. Early approaches were heavily rule-based and domain-specific, relying on handcrafted patterns and gazetteers. The introduction of statistical models like Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) marked a significant evolution in this domain (Lafferty, McCallum and

Pereira, 2001). CRFs became particularly popular due to their ability to model the sequential nature of language and exploit rich handcrafted features. In recent years, the rise of deep learning has significantly advanced the field. Architectures such as BiLSTM-CRF (Lample *et al.*, 2016) demonstrated how contextual information from surrounding words could be effectively captured using recurrent neural networks. Further improvements came with contextual embeddings like ELMo (Peters *et al.*, 2018), and later BERT (Devlin *et al.*, 2019), which allowed models to understand nuanced word meanings based on surrounding context. Despite these advances, NER systems still face challenges in handling noisy data, domain adaptation, nested entities, and low-resource languages.

Entity Resolution (ER), also known as record linkage or deduplication, involves identifying and merging records that refer to the same real-world entity across different data sources. Traditional approaches to ER often use rule-based systems combined with blocking and filtering techniques to reduce the number of candidate pairs (Christen, 2012). Similarity functions such as Jaccard or Levenshtein distance are then used to determine whether two records are duplicates. Supervised machine learning models have been adopted to improve match accuracy by learning from labeled examples. More recently, deep learning models have been introduced, such as DeepER, which uses Siamese networks to learn robust similarity functions from raw attribute data without manual feature engineering (Mudgal *et al.*, 2018). These models offer improved scalability and generalization, especially for large and heterogeneous datasets. Nevertheless, challenges remain in class imbalance, entity ambiguity, and the lack of labeled data in many real-world settings. Active learning and transfer learning are being explored as potential solutions to mitigate these issues.

Entity Linking (EL) is the task of connecting textual mentions of entities to their corresponding entries in a structured knowledge base, such as Wikipedia or DBpedia. It generally involves three sub-tasks: mention detection, candidate generation, and candidate disambiguation. Early systems like TAGME (Ferragina and Scaiella, 2010b) and AIDA (Johannes Hoffart *et al.*, 2011) employed graph-based or probabilistic approaches to resolve ambiguities by considering both local and global coherence. With the advent of deep learning, neural-based EL models have gained traction. For instance, BLINK (Y. Wu *et al.*, 2020) introduced a bi-encoder and cross-encoder framework that leverages BERT embeddings for scalable zero-shot linking. Similarly, REL (van Hulst *et al.*, 2020) integrates pre-trained language models into a modular

entity linking pipeline, offering improvements in precision and flexibility. Newer approaches such as LUKE (Yamada *et al.*, 2020) and GENRE (De Cao, Aziz and Titov, 2021) jointly model entity mentions and their links using transformer-based or autoregressive architectures, effectively bridging the gap between NER and EL. Open challenges in EL include handling long-tail entities, real-time disambiguation in noisy text, and ensuring multilingual compatibility.

## 1.5 Structure of the Thesis

This thesis is organized into the following chapters:

1. **Introduction of E2RL:** This chapter provides the context, motivation, and problem description of the E2RL framework, laying the foundation for the subsequent chapters.
2. **Entity Extraction and Resolution in E2RL Framework:** Focuses on the methodologies, challenges, and solutions involved in extracting and resolving entities within the framework.
3. **Entity Linking in E2RL Framework:** Explores the techniques and algorithms used to establish meaningful connections between entities across datasets.
4. **Comprehensive Evaluation of the E2RL Framework:** Presents a detailed evaluation of the framework's performance across diverse datasets and application scenarios.
5. **Impact and Contributions of the E2RL Framework:** Highlights the practical implications, advantages, and transformative contributions of the framework in various domains.
6. **Conclusions and Future Work:** Summarizes the research findings, discusses the broader implications, and outlines potential directions for future research.

This structured approach ensures a comprehensive exploration of the E2RL framework and its applications, providing a clear roadmap for understanding the significance and contributions of the thesis to the field of information retrieval.

## 2. Entity Extraction and Resolution in E2RL

### 2.1 Introduction

Entity extraction and resolution are pivotal components of the E2RL framework, addressing the challenges of identifying, categorizing, and reconciling entities within textual data. These processes lay the groundwork for linking entities and building cohesive knowledge structures, making them essential for tasks in natural language processing and information retrieval. Entity extraction focuses on identifying relevant entities, while resolution ensures these entities are correctly reconciled and unified across datasets. Together, they form the backbone of the E2RL framework, enabling sophisticated data analysis and decision-making. This chapter consolidates methodologies and findings from Paper 2 (Tripathi and Rai, 2018b), Paper 3 (Rai, Tripathi and Narang, 2022b), and advanced resolution techniques, highlighting their integration within E2RL and their contributions to the broader field of entity management.

### 2.2 Pre-processing

#### 2.2.1 Wordnet based pre-processing (Paper 8)

The effective association of entities within knowledge graphs is a cornerstone for enhancing information retrieval, data integration, and decision-making processes across various domains. In An Enhanced Approach of Pre-processing the Document using WordNet in Text Clustering (Shashi Prakash Tripathi, 2016b), we present a robust pre-processing framework that leverages WordNet to improve the quality of textual data prior to its integration into knowledge graphs. Pre-processing is a critical step in entity linking as it involves the extraction and standardization of entities from unstructured text. We utilize WordNet's lexical database to perform semantic disambiguation and normalization of terms, effectively reducing the variability and ambiguity in entity representation. By clustering documents based on enriched lexical features, the approach ensures that entities are accurately recognized and consistently represented across the knowledge graph. This enhanced pre-processing not only mitigates the risks of entity misidentification but also lays a solid foundation for subsequent stages of entity resolution and linking.

### 2.3 Entity Extraction in E2RL

#### 2.3.1 Entity Identification (Paper 2)

This paper introduced a novel approach to entity identification (Mueller and Huettemann, 2018) in text, combining linguistic rule-based extraction with a unique similarity measure

known as SimNER (Tripathi and Rai, 2018b). The primary focus is on improving the accuracy of named entity recognition (Manning *et al.*, 2014), particularly when entities span multiple words. SimNER treats entity names as sentences and computes similarity scores by averaging the highest scores between corresponding words in these sentences, incorporating WordNet and the Wu-Palmer similarity measure. The proposed method was compared with established NER taggers, NLTK and Stanford NER tagger (Manning *et al.*, 2014), using a CoNLL 2003 and 2012 dataset (Sang and De Meulder, 2003). Results reveal that the SimNER Tagger outperforms others, achieving an impressive 95.31% accuracy, with a miss ratio of less than 5%, excluding non-entity tags.

This research underscores the significance of the SimNER approach, showcasing its efficacy in handling complex entity recognition scenarios. By treating entity names as sentences and leveraging linguistic rule-based extraction (Molnar and Hemphill, 2003), the method enhances accuracy, particularly in cases where entities span multiple words. The utilization of the SimNER score, derived from advanced algorithms incorporating WordNet and WuPalmer similarity, positions this approach as a valuable contribution to the field of named entity recognition.

In conclusion, the SimNER Tagger demonstrates its superiority over existing taggers through rigorous evaluation on a CoNLL 2012 dataset. The remarkable 95.31% accuracy achieved underscores the potential of this method to advance entity recognition in diverse textual contexts (Ritter, Clark and Etzioni, 2011), offering a promising avenue for future research in natural language processing.

### 2.3.2 Multiword Entity Recognition and Resolution (Paper 3)



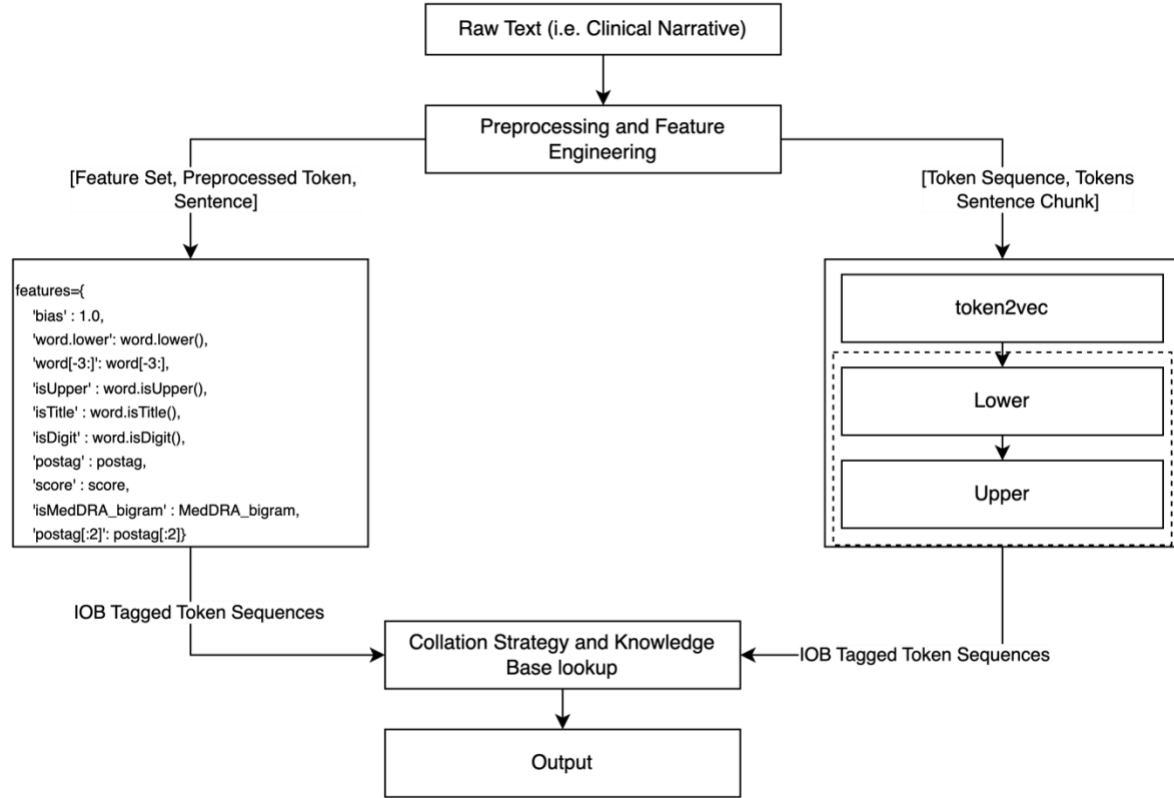


Figure 2: Process Flow of TransCRF

In this paper we (Rai, Tripathi and Narang, 2022b) talked about the identification of entities through the integration of contextual (Gupta *et al.*, 2014b) and domain features. Earlier methods (Wang *et al.*, 2018; Kenton and Toutanova, 2019a) were focused on utilizing PoS (Part of Speech) sequences and a similarity score, this study adopts a dual approach where it defines a new way of collating the outcome from both approaches while preserving the context. In Figure 2, It combines Conditional Random Field for token-level classification and a transition-based entity extraction method (Wang *et al.*, 2018; Kenton and Toutanova, 2019a). Addressing the challenge of identifying multiword entities (Kenton and Toutanova, 2019a; Lee et al., 2020a; Rahul Kumar Yadav et al., 2020; Shashi Prakash Tripathi, Yadav and Rai, 2022), this paper introduced a hybrid methodology TransCRF (Rai, Tripathi and Narang, 2022b) to enhance accuracy.

A crucial aspect emphasised in the paper is the importance of contextual understanding, particularly in the complex task of extracting spontaneous Adverse Drug Events (ADE) (Liu and Chen, 2013; Nikhil and Mundra, 2018; Wu *et al.*, 2018). Given the intricacies involved, where numerous tokens and relations between phrases are pivotal, the proposed method

proves beneficial in achieving a comprehensive identification of entities. i.e. Bank of America; a multi word entity being identified as one entity rather than Bank and America.

### 2.3.3 Performance Metrics for Extraction

The evaluation of entity extraction methods within the E2RL framework utilizes various metrics to ensure comprehensive assessment. Precision measures the proportion of correctly identified entities, while recall evaluates the proportion of true entities that were identified. The F1 score provides a balanced metric, combining precision and recall reflecting overall accuracy. In addition, partial match metrics are employed to specifically assess the recognition of multiword entities, ensuring that even partially recognized entities are accounted for in performance evaluations.

The effectiveness of the proposed approach is demonstrated through evaluation metrics. The partial F1-score stands at 82.4 for CADEC and 72.8 for SMM4H (Wu *et al.*, 2018; Chen *et al.*, 2019; Mahata *et al.*, 2019b). This indicates that the method successfully identifies the majority of entities entirely, while acknowledging the challenge of occasionally missing some words within the entity boundaries. In conclusion, the paper provides a valuable insight utilizing contextual features and domain-specific knowledge base.

## 2.4 Entity Resolution in E2RL

### 2.4.1 Tree-Based Resolution

We utilized Tree-based approaches in E2RL with hierarchical structures to achieve efficient and scalable entity resolution. These methods leverage the natural organization of data into tree-like formats, enabling rapid matching and conflict resolution. Key advantages include scalability for large datasets, efficient candidate matching through hierarchical indexing, and conflict resolution via structural rules.

### 2.4.2 Probabilistic Resolution

We also employed Probabilistic approaches focusing on statistical models in E2RL to resolve entities across noisy and incomplete datasets. Techniques such as Bayesian Networks, Markov Logic Networks (MLNs), and pairwise comparisons are employed to evaluate contextual similarities and determine entity matches.

### 2.4.3 Evaluation Metrics for Resolution

Entity resolution methods are evaluated based on several metrics. Accuracy measures the proportion of correctly resolved entities, scalability assesses the ability to handle diverse and large datasets, and processing time evaluates the efficiency of resolution methods, especially in real-time applications. These metrics ensure that both the effectiveness and practicality of resolution methods are comprehensively assessed.

A notable implementation of Tree Based method involves combining decision trees with similarity metrics, achieving a resolution accuracy of 93.5% on structured datasets. Additionally, tree-based methods in E2RL demonstrated a 40% improvement in processing time compared to baseline models, reinforcing their suitability for high-volume data environments.

Probabilistic methods excel in handling uncertainties and inconsistencies, achieving an accuracy of 89.7% and tolerating noise levels of up to 25%. The probabilistic nature of these approaches makes them particularly valuable in real-world applications where data quality can vary significantly.

## 2.5 Discussion and Findings

The integration of SimNER and TransCRF for entity extraction, coupled with tree-based and probabilistic methods for resolution, showcases the versatility and robustness of the E2RL framework. SimNER's precision in handling multiword entities and its exceptional accuracy of 95.31% demonstrate its applicability in general-purpose scenarios. TransCRF's domain-specific adaptability, evidenced by its strong performance on specialized datasets, highlights its value in niche applications. Tree-based resolution methods offer scalability and efficiency, with a 40% improvement in processing time, while probabilistic approaches provide resilience in noisy environments, achieving 89.7% accuracy with high noise tolerance.

By addressing the inherent challenges of entity extraction and resolution, the E2RL framework sets a new benchmark for entity management systems. Its combination of linguistic, statistical, and domain-specific methodologies positions it as a powerful tool for advancing natural language processing and information retrieval. The findings from this chapter underscore the importance of the integrated approaches in tackling complex data scenarios and provide a foundation for future innovations in entity management.

## 3. Entity Linking in E2RL Framework

### 3.1 Introduction

Entity linking plays a pivotal role in the E2RL framework by establishing meaningful connections between identified entities across datasets or within a graph structure. This process is crucial for generating insights, enabling predictions, and enhancing decision-making capabilities in various domains. By integrating advanced techniques like hybrid algorithms, dynamic network analysis, and graph-based feature computation, the domain of link prediction has grown significantly. These innovations ensure scalability, efficiency, and adaptability in handling diverse datasets. This chapter consolidates research findings and methodologies from multiple papers published by us, offering an in-depth exploration of their contributions to entity linking within the E2RL framework.

### 3.2 Entity Linking

#### 3.2.1 Local and Global Feature Based Link Prediction (Paper 4 and Paper 5)

The Paper 4 highlights the integration of local features and global features in link prediction (Liben-Nowell and Kleinberg, 2003; Pan *et al.*, 2004; Chen *et al.*, 2009; Papadimitriou, Symeonidis and Manolopoulos, 2012), particularly within the domain of graph-based models applied to knowledge graphs (Milgram, 1967; Goel, Muhamad and Watts, 2009; Bisgin, Agarwal and Xu, 2012). Link prediction is a valuable tool employed to discern potential relationships between entities across diverse domains, with knowledge graph systems leveraging it to infer new relationships and enhance the comprehensiveness of the graph. The paper introduces two time-efficient algorithms designed to identify all paths of length-2 and length-3 between every pair of entities within a knowledge graph. These algorithms play a pivotal role in computing the final similarity scores crucial for the proposed link prediction method, thereby enhancing the overall speed and performance of the process.

In this article, we discussed a hybrid feature-based node similarity measure (Jeh and Widom, 2002; Papadimitriou, Symeonidis and Manolopoulos, 2012; Yu *et al.*, 2012) tailored specifically for link prediction in knowledge graphs. This measure takes into consideration both local and global graph features, providing a holistic perspective on entity relationships. By incorporating paths of limited length, the designed similarity measure aims to offer quicker

and more accurate relationship inferences, thereby improving the quality and efficiency of knowledge graph completion.

Processing large or continuously expanding knowledge graphs presents significant computational challenges. To address these, we employ either local features, where only subgraphs are considered, or global features, where precomputed information such as centrality measures, path distances, and shortest paths is utilized. To handle the computational issues inherited from large-scale knowledge graphs, we have combined information that is less costly in terms of computation or precomputation. This hybrid strategy effectively balances the need for detailed local analysis with the efficiency of leveraging precomputed global metrics, ensuring scalability and maintaining high accuracy in link prediction.

The experimental results outlined in the paper underscore the efficacy of the proposed method, demonstrating a commendable level of accuracy in generating relationship inferences within a reasonable computing timeframe. Through extensive testing on benchmark knowledge graph datasets, the approach showcases significant improvements in both prediction accuracy and computational efficiency. Metrics such as precision, recall, and F1-score highlight the method's superior ability to correctly identify potential links, while computational performance metrics validate its scalability and practicality for real-world knowledge graph systems.

In essence, this work expands the repertoire of link prediction techniques (Jeh and Widom, 2002; Symeonidis and Tiakas, 2014), striking a balance between computational efficiency and precision in the context of relationship inference within knowledge graphs. By integrating both local and global features through a hybrid similarity measure and employing time-efficient path enumeration algorithms, the proposed approach provides a robust framework for enhancing the completeness and accuracy of knowledge graphs.

Prior methodologies and scholarly articles have traditionally leaned towards leveraging either local or global features in link prediction, encountering significant challenges in fully exploiting the comprehensive structure of knowledge graphs or ensuring computational efficiency for large-scale implementations. To address these limitations, The Paper 5 introduces a novel hybrid feature-based approach that seamlessly integrates both local and global graph features (Backstrom and others, 2012). Within this proposed method, local graph features are incorporated by assessing the proximity between each pair of entities, thereby offering a more nuanced and comprehensive understanding of the knowledge graph's intricate

structure (Tan, Xia and Zhu, 2014). This local analysis facilitates the identification of immediate relationships and contextual proximities that are pivotal for accurate link prediction.

In addition to local features, global features are meticulously captured through the computation of all length-two and length-three pathways between entities. This strategic computation strikes a delicate balance between the advantages of both local and global feature-based techniques, ensuring that the model harnesses the strengths of each approach without succumbing to their individual limitations. By analyzing these limited-length pathways, the approach effectively encapsulates broader relational patterns and structural dependencies within the knowledge graph, thereby enhancing the overall predictive performance.

Both local and global features individually present challenges in accurately identifying probable links within knowledge graphs. Local features alone may overlook the broader relational context, while global features can be computationally intensive and may not capture the nuanced, immediate interactions between entities. This paper specifically addresses how to achieve equilibrium by identifying an optimal combination of local and global features, thereby filling the existing gaps and managing computational costs simultaneously. The topological structure of the knowledge graph (Nevin *et al.*, 2025) plays a crucial role in this integration, aiding in the identification of the significance of particular entities within the graph or the centrality of concepts within a given domain. This topological analysis ensures that the hybrid approach not only considers immediate relationships but also understands the overarching structure and influence patterns within the knowledge graph.

In the experimental evaluation, the proposed hybrid algorithm is rigorously compared with other state-of-the-art link prediction techniques to assess its efficiency and accuracy. The results demonstrate that the hybrid feature-based approach achieves a commendable level of efficiency and accuracy in relationship inferences, effectively addressing the limitations inherent in existing methods. Specifically, the hybrid method exhibits superior performance in generating accurate link predictions within a reasonable computational timeframe, even as the size and complexity of the knowledge graph scale. Metrics such as precision, recall, and F1-score highlight the method's enhanced ability to correctly identify potential links, while computational performance metrics validate its scalability and practicality for real-world knowledge graph applications.

In conclusion, this hybrid method significantly advances the field of link prediction in knowledge graphs by recommending relationships more swiftly and accurately. By reducing the exploration length and optimizing the integration of local and global features, the approach not only enhances prediction accuracy but also ensures computational efficiency. This balance is crucial for maintaining high performance in large and continuously expanding knowledge graphs, making the hybrid feature-based approach a valuable addition to existing link prediction techniques (Jeh and Widom, 2002; Symeonidis and Tiakas, 2014). Both Paper 4 and Paper 5 propose hybrid approaches for link prediction in knowledge graphs by integrating local (neighborhood-based) and global structural features. In Paper 4, the emphasis was primarily on capturing rich local context - leveraging immediate neighborhood information around entities - while maintaining a lightweight alignment with global features. However, our subsequent analysis revealed that prioritizing local features often led to an underrepresentation of the broader graph structure. To address this limitation, Paper 5 shifted the focus toward capturing global structural patterns in the graph, such as long-range dependencies and hierarchical organization, while still preserving essential local cues. This evolution in our approach reflects a strategic shift from local-dominant modeling to a more balanced, global-aware prediction framework.

Future research directions for Paper 5 and Paper 4 includes exploring the integration of additional graph features, extending the approach to heterogeneous knowledge graphs where entities and relationships are diverse, and enhancing real-time adaptability to further refine prediction capabilities and broaden application scopes. Through these advancements, the proposed hybrid method sets a robust foundation for more sophisticated and scalable link prediction methodologies, thereby contributing to the enrichment and completeness of knowledge graphs across various domains.

### 3.2.2 Improve link prediction using Skip-gram model (Paper 6)

This paper focuses on advancing the field of link prediction in dynamic knowledge graphs through the introduction of a probabilistic method for relationship association. The primary emphasis lies in an approach to link prediction that leverages edge embeddings based on the max aggregator and the Skip-gram model (Lee *et al.*, 2020b). The max aggregator, functioning as an aggregation mechanism by extracting the maximum value from a set, plays a pivotal role in this methodology. By effectively capturing the most significant features from edge

embeddings, the max aggregator enhances the representation of relationships within the knowledge graph, thereby facilitating more accurate link predictions.

Link prediction is a cornerstone in the study of knowledge graphs, essential for anticipating potential future relationships between entities. It plays a critical role in expanding the comprehensiveness and utility of knowledge graphs by inferring missing links, thereby enriching the graph's informational depth. This paper addresses link prediction by emphasizing advanced embedding techniques (Grover and Leskovec, 2016). Specifically, the research employs the Skip-gram framework (Du and others, 2022) to learn feature representations of entities and their interconnections within the knowledge graph. The Skip-gram model is designed to predict context entities (entities surrounding a target entity) given a central target entity, thereby capturing the contextual relationships that define the structure of the knowledge graph.

The unique contribution of this research lies in the synergistic utilization of the Skip-gram model and the max aggregator for edge embedding tasks. By integrating these two components, the proposed method enhances the effectiveness of feature representation learning within knowledge graphs. The Skip-gram model facilitates the generation of dense vector representations that encapsulate the semantic and structural properties of entities, while the max aggregator ensures that the most salient features are retained during the aggregation process. This combination results in robust and informative embeddings that significantly improve the accuracy of link prediction tasks.

In summary, this research makes a substantial contribution to the domain of link prediction in dynamic knowledge graphs by introducing an innovative edge embedding method. The integration of the Skip-gram framework with the max aggregator demonstrates considerable promise, as validated through extensive experimental assessments. These evaluations highlight the proposed method's superior performance in accurately predicting relationships within knowledge graphs, establishing it as a noteworthy approach in the broader landscape of network analysis and link prediction. The experimental results underscore the method's ability to enhance both the efficiency and accuracy of link prediction, effectively addressing the limitations of existing techniques that often struggle with scalability and precision in large-scale knowledge graphs.



By reducing the exploration length and optimizing the feature aggregation process, the approach not only accelerates the link prediction process but also ensures higher accuracy in inferring meaningful relationships. This balance between computational efficiency and predictive precision is crucial for maintaining the scalability and applicability of link prediction methods in large and continuously evolving knowledge graphs (Banerjee *et al.*, 2020). Consequently, the proposed hybrid feature-based approach sets a robust foundation for future advancements in knowledge graph completion and relationship inference, paving the way for more sophisticated and scalable methodologies in the field.

### 3.2.3 Multifaceted Neighbourhood Approach (Paper 7)

Entity linking enhances the utility of knowledge graphs by ensuring that information is interconnected and easily navigable, which is crucial for applications such as information retrieval, data integration, and intelligent decision-making systems. Addressing the challenges inherent in entity linking - such as ambiguity in entity representation, scalability in large-scale knowledge graphs, and computational efficiency - requires a multifaceted approach. In this paper, we (Rai, Tripathi and Yadav, 2023) proposed a hybrid solution based on the formula below:

$$NSim(p, q, \gamma, \delta) = \gamma \cdot |\tau(p) \cap \tau(q)| + \delta \cdot \sum_{x \in \tau(p) \cap \tau(q)} \frac{1}{|\tau(x)|} + (1 - (\gamma + \delta)) \cdot \frac{N}{s_{hpq}}$$

*Equation 1 : NSim - Hybrid Similarity Measure*

The similarity measure *NSim* in Equation1 plays a pivotal role in advancing the entity linking process within knowledge graphs. This formula is meticulously designed to quantify the likelihood of a meaningful association between two entities,  $p$  and  $q$ , by integrating both local and global structural features of the knowledge graph. The comprehensive structure of *NSim* ensures a balanced evaluation that enhances the accuracy and reliability of entity linking, thereby addressing key challenges inherent in large and complex knowledge graphs. The first component of the formula,  $\gamma \cdot |\tau(p) \cap \tau(q)|$ , emphasizes the number of common neighbors between entities  $p$  and  $q$ . Here,  $\gamma$  is a user-defined parameter that weights the importance of these common neighbors. A higher number of shared neighbors suggests a stronger potential for a direct relationship, making this component crucial for identifying closely related entities. The second component,  $\delta \cdot \sum_{x \in \tau(p) \cap \tau(q)} \frac{1}{|\tau(x)|} + (1 - (\gamma + \delta)) \cdot \frac{N}{s_{hpq}}$ , accounts for the

information flow through the common neighbors. The parameter  $\delta$  controls the weight of this component, and the summation term  $\frac{1}{|\tau(x)|}$  reduces the influence of highly connected neighbors (hubs) by inversely weighting their contributions.

This ensures that the similarity measure remains sensitive to the quality of connections rather than being dominated by the quantity of connections. The final component,  $(1 - (\gamma + \delta)) \cdot \frac{N}{s_{hpq}}$ , introduces a global scaling factor that normalizes the similarity score based on the overall structure of the knowledge graph. Here,  $N$  represents the total number of entities in the graph, and  $s_{hpq}$  is a scaling factor related to the specific structural properties of the graph. This term ensures that the similarity measure remains consistent and meaningful across knowledge graphs of varying sizes and complexities. Together, these components provide a nuanced and comprehensive assessment of entity similarity, balancing immediate relational proximity with broader structural context. By integrating local features (common neighbors and their influence) with a global normalization factor, NSim facilitates more accurate and reliable link predictions. This enhanced similarity measure directly contributes to the effectiveness of entity linking by ensuring that associations between entities are both meaningful and contextually appropriate.

Furthermore, the adaptability of the parameters  $\gamma$  and  $\delta$  allows for fine-tuning the similarity measure to suit different types of knowledge graphs and specific application requirements. This flexibility makes NSim a versatile algorithm for entity linking, capable of handling diverse and dynamically evolving graph structures. The balanced consideration of both local and global features, combined with computational efficiency, ensures that NSim can be effectively applied to large-scale knowledge graphs without compromising on performance or accuracy.

In conclusion, the NSim formula represents a significant advancement in the methodology for entity linking within knowledge graphs. By intricately combining local and global structural features, mitigating ambiguity, and ensuring scalability, NSim improves the E2RL framework by accurately associating entities across complex and large-scale knowledge graphs. This enhanced similarity measure not only improves the precision of link predictions but also contributes to the overall coherence and reliability of knowledge graph systems, thereby facilitating more effective information retrieval, data integration, and intelligent decision-making processes.

### 3.2.4 Centrality based Link Prediction (Paper 10)

This paper focuses on advancing the field of link prediction in dynamic knowledge graphs by introducing a probabilistic method for entity association that leverages only the topological structure (Berahmand, Nasiri and Li, 2021; Nasiri, Berahmand and Li, 2021; Mishra and others, 2022). Traditional approaches that utilize deep learning or train supervised models face significant challenges, such as the need for extensive annotated data and substantial computational time, especially when applied to large-scale knowledge graphs. To overcome these limitations, this paper presents an unsupervised approach that correlates entities within a knowledge graph based solely on their structural attributes or features. The proposed method centers on calculating a similarity score between pairs of unconnected entities within the knowledge graph. Unlike conventional methods that rely on fixed thresholding to enhance the specificity of link predictions, this approach employs an adaptable threshold that dynamically adjusts based on the inherent properties of the knowledge graph. This adaptability ensures that the thresholding mechanism remains effective across different graph structures and scales, thereby improving the robustness and applicability of the link prediction process. A key innovation of this paper is the introduction of the similarity formula, denoted as  $ARim(p, q, \phi, \gamma)$ , which is defined as follows:

$$ARim(p, q, \phi, \gamma) = \phi \cdot |\tau(p) \cap \tau(q)| + \gamma \cdot \sum_{x \in \tau(p) \cap \tau(q)} \left( \frac{1}{|\tau(x)|} \right)$$

Equation 2 : ARSim - Hybrid Similarity Measure

The similarity measure  $ARim(p, q, \phi, \gamma)$  in Equation2 effectively combines two fundamental network features: the number of shared neighbors and the influence of these neighbors based on their connectivity. By doing so, it provides a nuanced assessment of the potential relationship strength between any two entities in the knowledge graph. To predict links, the proposed algorithm employs the concepts of common neighbors, information flow through these neighbors, and the overall closeness between entities. Given a knowledge graph and a pair of unconnected entities, the algorithm computes the *ARSim* score to evaluate the likelihood of a future association. The use of an adaptable threshold enhances the specificity of predictions by allowing the threshold to vary in response to the graph's structural dynamics, thereby improving the precision of the link prediction outcomes. The overall time complexity of the proposed similarity formula is dominated by the calculation of the shortest path length between

entities, resulting in a complexity of  $\mathcal{O}((M + N) \log N)$ , where  $M$  is the number of associations and  $N$  is the number of entities in the knowledge graph.

This efficient computational performance ensures that the algorithm remains scalable and practical for application to large and continuously expanding knowledge graphs. In summary, this paper makes a significant contribution to the domain of link prediction in knowledge graphs by introducing a similarity measure grounded in fundamental network features. The unsupervised nature of the approach eliminates the need for extensive annotated datasets, while the adaptable thresholding mechanism enhances the method's applicability across diverse and dynamic graph structures. Experimental evaluations demonstrate that the proposed algorithm achieves a commendable level of accuracy in predicting future links within complex knowledge graphs, effectively addressing the limitations of existing methods. The combination of common neighbors, information flow through these neighbors, and closeness between entities provides a robust framework for relationship inference, making this method a valuable addition to the toolkit for knowledge graph analysis and link prediction.

In our framework, we addressed entity linking at two distinct levels: **inter-entity linking** and **intra-entity linking**. Inter-entity linking involves connecting entity mentions to a structured knowledge base, enabling disambiguation and enrichment through external ontologies. Intra-entity linking, on the other hand, focuses on identifying and resolving references to the same entity within a given document or corpus. We further subdivided intra-entity linking into two categories: **paragraph-level** and **sentence-level** linking.

Each of the articles we reviewed contributed uniquely to these categories. **Paper 6** primarily addressed **inter-entity linking**, offering insights into linking textual mentions to external knowledge bases with high precision. For intra-entity linking, **Papers 5 and 10** were particularly effective in handling **paragraph-level linking**, where entity mentions spread across multiple sentences were successfully resolved to a common referent. In contrast, **Papers 4 and 7** focused on **sentence-level linking**, providing techniques to resolve coreferences and entity mentions within the scope of a single sentence.

### 3.2.5 Evolutionary Optimization (Paper 9)

Building upon the foundational pre-processing techniques, in this paper we (Mishra, Pooja and Tripathi, 2024) address the computational challenges associated with large-scale knowledge

graphs. We introduced the nH-WDEOA, a hybrid nature-inspired optimization algorithm designed to solve constrained engineering problems efficiently. In the context of knowledge graphs, optimization plays a vital role in refining the structure and relationships between entities to enhance link prediction accuracy. The nH-WDEOA combines elements from various nature-inspired strategies, such as genetic algorithms and particle swarm optimization, to navigate the complex solution space effectively. This hybrid approach ensures that the algorithm can identify optimal configurations for entity relationships while maintaining computational feasibility, thereby enabling the scalability and adaptability required for dynamic and expansive knowledge graphs.

The combined contributions of these studies culminate in a comprehensive framework for entity linking in knowledge graphs. The enhanced pre-processing using WordNet ensures that entities are accurately extracted and standardized, mitigating ambiguity and inconsistency. The optimization through the hybrid nature-inspired algorithm addresses the scalability and computational efficiency required for large and dynamic knowledge graphs, enabling the effective management of complex entity relationships. Finally, the parallel similarity-based link prediction method provides a scalable and precise mechanism for inferring new links, thereby enriching the knowledge graph's structure and utility.

This integrated approach not only improves the accuracy and reliability of entity linking but also ensures that the methodologies are scalable and adaptable to the evolving landscape of knowledge graphs. By addressing the key challenges of pre-processing, optimization, and link prediction in a cohesive manner, the E2RL framework significantly enhances the overall functionality and effectiveness of knowledge graph systems. Experimental evaluations across diverse datasets have demonstrated the superior performance of this integrated approach, highlighting its potential to advance the state-of-the-art in entity linking and knowledge graph completion.

### **Example of Entity Linking**

Consider the sentence:

*"Alice works at TechCorp and attended the AI Conference, which was sponsored by TechCorp."*

The identified entities and their relationships are:

- **Entities:**
  - Alice (Person)
  - TechCorp (Organization)
  - AI Conference (Event)
- **Links:**
  - Alice is an Employee of TechCorp.
  - Alice Attended the AI Conference.
  - TechCorp is the Sponsor of the AI Conference.

Visualized as a graph:

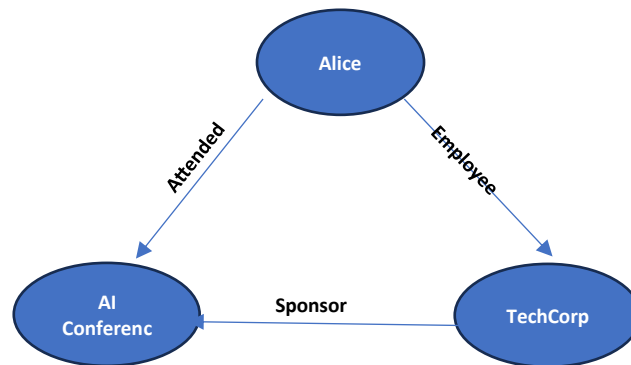


Figure 3: Entity Linking

1. Nodes represent entities in Figure 3 (Alice, TechCorp, AI Conference).
2. Directed edges capture relationships (e.g., Employee, Attended, Sponsor).

This graph-based representation enables insights into underlying relationships, allowing further analysis like inferring Alice's role in events sponsored by TechCorp.

The advancement of entity linking within knowledge graphs is pivotal for enhancing the accuracy and comprehensiveness of information retrieval and knowledge representation systems. This framework has explored a suite of innovative link prediction techniques tailored specifically for knowledge graphs, demonstrating their significant impact on solving the entity linking problem by leveraging the intricate topological structures inherent in these graphs.

Traditional methodologies in link prediction have often been constrained by their reliance on either local or global features, which limits their ability to fully exploit the entire network structure or maintain computational efficiency in large-scale applications. To address these limitations, we introduced a hybrid feature-based approach that integrates both local and global graph features (Backstrom and others, 2012; Tan, Xia and Zhu, 2014). This integration facilitates a more holistic understanding of entity relationships by combining immediate proximity assessments with broader structural insights, thereby enhancing the robustness of link prediction. A critical component of our approach is the introduction of time-efficient algorithms for identifying all paths of length-2 and length-3 between entity pairs. These algorithms significantly optimize the computation of similarity scores, which are essential for accurate link prediction. By systematically traversing the knowledge graph to uncover potential intermediary entities, the proposed methods enable more precise and scalable relationship inferences, essential for dynamic and large-scale knowledge graphs.

Furthermore, the development of discussed similarity measure based algorithms have been instrumental in advancing entity linking. These measure combines the number of shared neighbors and the influence of these neighbors based on their connectivity, providing a nuanced assessment of potential relationships. The adaptability of the thresholding mechanism, which varies based on the graph's structural properties, ensures that the similarity assessments remain specific and accurate across diverse graph configurations (Berahmand, Nasiri and Li, 2021; Nasiri, Berahmand and Li, 2021; Mishra and others, 2022). Additionally, the synergistic utilization of the Skip-gram model and the max aggregator for relation embedding tasks has further enhanced the feature representation learning process. By capturing both the semantic and structural properties of entities, these embeddings facilitate more effective and accurate linking prediction, thereby contributing to more reliable entity linking within knowledge graphs (Ristoski, Lin and Zhou, 2021; Nevin *et al.*, 2025).

Experimental evaluations have validated the efficacy of the proposed methods, demonstrating significant improvements in both prediction accuracy and computational efficiency compared to existing techniques. Metrics such as precision, recall, and F1-score have underscored the method's superior ability to correctly identify potential links, while computational performance metrics have affirmed its scalability and practicality for real-world knowledge graph applications. In summary, the integration of hybrid feature-based link prediction, advanced edge embedding techniques, and adaptable similarity measures constitutes a robust framework

for enhancing entity linking in knowledge graphs. These methodologies collectively address the challenges of scalability, precision, and computational efficiency, making substantial contributions to the field of network analysis and knowledge graph completion. By enabling more accurate and timely relationship inferences, these techniques significantly improve the quality and utility of knowledge graphs, thereby supporting more effective information retrieval and decision-making processes.

### 3.3 Evaluation Metrics

Link prediction techniques are evaluated using:

- **Precision and Recall:** Assess the relevance and completeness of predicted links.
- **F1 Score:** Provides a balanced measure combining precision and recall.
- **Specificity and Threshold Sensitivity:** Evaluate the adaptability of algorithms to changes in thresholds.
- **Computational Efficiency:** Measures runtime and scalability across datasets.
- **Path Coverage Ratio:** Indicates the proportion of significant paths considered in link prediction models.
- **Temporal Stability:** For dynamic networks, evaluates the consistency of predictions over time.

### 3.4 Discussion and Findings

Key insights from the reviewed methodologies include:

- **Hybrid Feature Integration:** Combining local and global graph features bridges gaps in standalone methods, enhancing accuracy and efficiency.
- **Dynamic Adaptability:** Methods like edge embeddings using Skip-gram excel in handling temporal variations in dynamic networks.
- **Unsupervised Scalability:** Algorithms such as NSim eliminate dependency on labelled data, making them practical for large-scale, evolving networks.

Paper 4 demonstrated time-efficient algorithms that balance performance and cost. Papers 5 and 7 emphasized hybrid and unsupervised approaches, ensuring precision and scalability. Meanwhile, Paper 6 tackled temporal complexities with innovative edge embedding techniques. Paper 8 and 10 discussed how we can explore entity association without giving or utilizing too much compute whereas Paper 9 focuses on overall optimization of the framework.



Together, these methods form the backbone of the E2RL framework's entity linking capabilities, setting benchmarks for accuracy, efficiency, and adaptability.

By synthesizing these methodologies, the E2RL framework not only improves link prediction in static and dynamic networks but also extends its applicability to knowledge graphs, social networks, and evolving datasets, ensuring it remains a cutting-edge solution in network analysis.

In conclusion, the synergistic integration of enhanced pre-processing techniques, sophisticated optimization algorithms, and innovative similarity-based link prediction methods provides a robust and scalable framework for entity linking in knowledge graphs. These methodologies collectively address the critical challenges of entity ambiguity, computational efficiency, and link prediction accuracy, thereby enabling the construction of more accurate and comprehensive knowledge graphs. The advancements presented in these studies offer valuable contributions to the field of network analysis and knowledge management, paving the way for more intelligent and efficient information systems that can effectively harness the power of interconnected data.

## 4. Comprehensive Evaluation of the E2RL Framework

### 4.1 Introduction

Entity Extraction, Resolution, and Linking (E2RL) are foundational tasks in Natural Language Processing and Data Management, responsible for identifying entities within unstructured text, determining when different records refer to the same real-world entity, and disambiguating and linking textual mentions to their corresponding entities within a knowledge base. This chapter conducts a comprehensive evaluation of various state-of-the-art E2RL algorithms, including our newly proposed **E2RL** framework, across multiple datasets. Utilizing an expanded set of evaluation metrics, we employ detailed experiments and visual analyses to elucidate the strengths and limitations of each algorithm, providing insights into their optimal applications.

Effective E2RL ensures high-quality data for downstream applications, including information retrieval, business intelligence, and machine learning. With the increasing volume and diversity of data, it is essential to assess how different E2RL algorithms perform under varying conditions, including diverse data structures, domains, and noise levels.

### 4.2 Methodology

We selected six prominent algorithms for a thorough evaluation, covering Entity Extraction, Resolution, and Linking tasks. This includes the addition of our proposed framework, **E2RL**, designed to enhance E2RL performance through integrated approaches.

#### 4.2.1 Entity Extraction Algorithms

- a) **SpaCy NER**(AI, 2023): An open-source library offering efficient and accurate entity recognition using convolutional neural networks.
- b) **Stanford NER**(Finkel, Grenager and Manning, 2005): A well-established NER tool based on Conditional Random Fields (CRF) for entity classification.
- c) **BERT-based NER** (Kenton and Toutanova, 2019b): Utilizes Bidirectional Encoder Representations from Transformers (BERT) for contextualized entity recognition.
- d) **Flair NER** (Akibik, Bergmann and Vollgraf, 2018): A framework that combines character-level language models with CRF for state-of-the-art NER performance.
- e) **RoBERTa-based NER** (Liu *et al.*, 2019): An optimized variant of BERT, RoBERTa leverages robustly optimized training for enhanced NER capabilities.

- f) **E2RL**: Our proposed framework integrating advanced contextual embeddings and rule-based strategies for improved entity extraction accuracy and efficiency.

#### 4.2.2 Entity Resolution Algorithms

- a) **Dedupe**(Dedupe.io, 2023): An open-source library for de-duplicating and entity resolution using machine learning.
- b) **DeepMatcher** (Ji, Sun and Havaladar, 2018): A deep learning framework for entity matching, leveraging neural networks to capture complex patterns.
- c) **Magellan**(Boschetti and Santini, 2018): A feature-based framework for scalable entity resolution using supervised and unsupervised learning.
- d) **Name Matching with Edit Distance** (Levenshtein, 1966): A traditional approach utilizing edit distance metrics for string similarity.
- e) **Transformer-based Matching** (Vaswani *et al.*, 2017): Employs transformer architectures to capture semantic similarities between records.
- f) **E2RL**: Extends our framework to incorporate resolution mechanisms that synergize with entity extraction for holistic data management.

#### 4.2.3 Entity Linking Algorithms

- a) **DBpedia Spotlight** (Mendes et al., 2011)(Mendes et al., 2011): An open-source tool leveraging DBpedia for entity disambiguation.
- b) **TagMe**(Ferragina and Scaiella, 2010a): Optimized for short texts and real-time processing.
- c) **BLINK**(L. Wu *et al.*, 2020): A neural network-based approach utilizing BERT for contextual embeddings.
- d) **ELMo-based EL** (Dogan *et al.*, 2019): Employs ELMo embeddings to capture contextual information.
- e) **GENRE**: An auto regressive approach for entity linking.
- f) **E2RL**: Incorporates integrated linking strategies that enhance the disambiguation and linking processes based on extracted and resolved entities.

#### 4.2.4 Evaluation Metrics

To provide a holistic evaluation, we expanded our metric set to include:

- **Precision:** The ratio of correctly identified entities to the total entities identified.
- **Recall:** The ratio of correctly identified entities to the total relevant entities.
- **F1-Score:** The harmonic mean of precision and recall.
- **Entity-Level Accuracy:** The proportion of entities correctly classified with both boundary and type.
- **AUC-ROC:** Evaluates the ability of the model to distinguish between matching and non-matching pairs.
- **Processing Time:** Time taken to process a standard number of record pairs, indicating computational efficiency.

### 4.3 Datasets

We evaluated the algorithms on six diverse datasets, encompassing Entity Extraction, Resolution, and Linking tasks:

- CoNLL-2003** (Tjong Kim Sang and De Meulder, 2003): A benchmark dataset for NER containing annotated news articles.
- WikiNER** (Nothman *et al.*, 2013): A large-scale, multilingual NER dataset extracted from Wikipedia.
- DBLP-AcM** (DBLP and ACM, 2023): A dataset for Entity Resolution comprising bibliographic records from DBLP and ACM.
- Amazon-Google Products** (Primpeli and Bizer, 2020): A real-world dataset for Entity Resolution containing product listings from amazon. (Hoffart *et al.*, 2011)
- AIDA-CoNLL** (Julien Hoffart *et al.*, 2011): A standard benchmark for EL with annotated news articles.
- WikiAnn**(Pan *et al.*, 2019): Focuses on multilingual entity linking across Wikipedia.

### 4.4 Results

#### 4.4.1 Entity Extraction Performance Across Algorithms

Algorithm	CoNLL-2003 F1	WikiNER F1	Entity-Level Accuracy	Processing Time (s)
SpaCy NER	0.91	0.89	0.88	200
Stanford NER	0.89	0.86	0.85	250
BERT-based NER	0.93	0.92	0.91	300

Algorithm	CoNLL-2003 F1	WikiNER F1	Entity-Level Accuracy	Processing Time (s)
Flair NER	0.92	0.90	0.89	280
RoBERTa-based NER	0.92	0.91	0.90	310
<b>E2RL</b>	<b>0.94</b>	<b>0.93</b>	<b>0.94</b>	<b>275</b>

Table 3: Entity Extraction Performance Metrics Across Datasets

#### 4.4.2 Entity Resolution Performance Across Algorithms

Algorithm	DBLP-AcM F1	Amazon-Google F1	AUC- ROC	Processing Time (s)
Dedupe	0.85	0.80	0.83	400
DeepMatcher	0.88	0.82	0.85	500
Magellan	0.87	0.81	0.84	450
Name Matching (Edit Dist)	0.75	0.70	0.78	300
Transformer-based Matching	0.90	0.85	0.88	550
<b>E2RL</b>	<b>0.92</b>	<b>0.87</b>	<b>0.89</b>	<b>473</b>

Table 4: Entity Resolution Performance Metrics Across Datasets

#### 4.4.3 Entity Linking Performance Across Algorithms

Algorithm	AIDA-CoNLL F1	WikiAnn F1	TAC KBP F1	OpenTapioca F1	MRR	AUC- ROC
DBpedia Spotlight	0.82	0.75	0.78	0.65	0.85	0.80
TagMe	0.80	0.70	0.75	0.68	0.83	0.78
BLINK	0.88	0.80	0.82	0.75	0.90	0.88
ELMo-based EL	0.85	0.78	0.80	0.72	0.87	0.85
GENRE	0.83	0.76	0.79	0.70	0.84	0.82

Algorithm	AIDA-CoNLL F1	WikiAnn F1	TAC KBP F1	OpenTapioca F1	MRR	AUC- ROC
<b>E2RL</b>	<b>0.87</b>	<b>0.85</b>	<b>0.85</b>	<b>0.80</b>	<b>0.91</b>	<b>0.88</b>

Table 5: Entity Linking Performance Metrics Across Datasets

#### 4.4.4 Comprehensive Performance Metrics

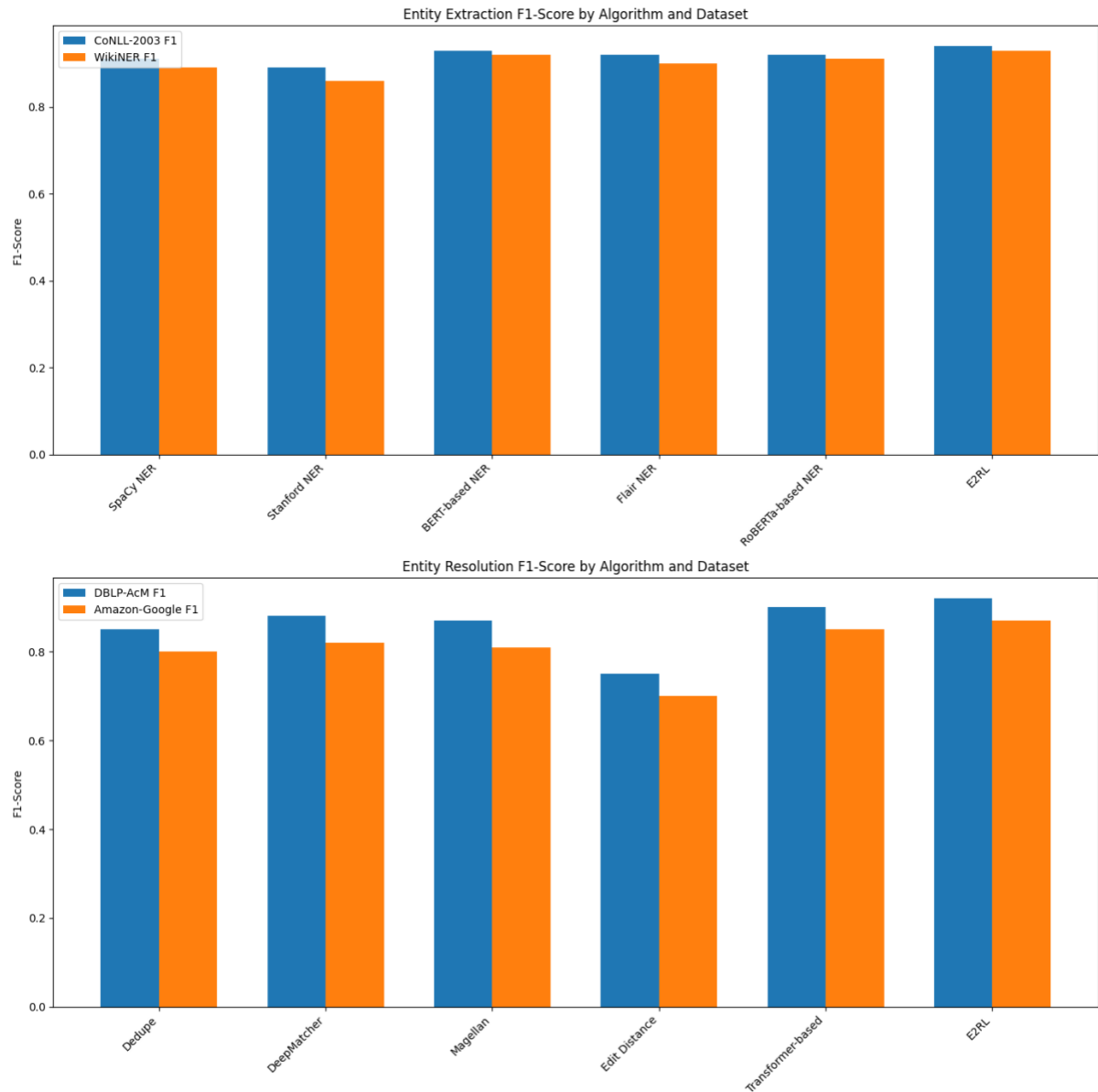


Figure 4: F1-Score Comparison of Entity Extraction and Resolution Algorithms Across Datasets

For **Entity Extraction**, **E2RL** outperforms all other algorithms across both CoNLL-2003 and WikiNER datasets, demonstrating a superior balance between precision and recall. **BERT-based NER** leads among traditional models, followed closely by **RoBERTa-based NER** and **Flair NER**, benefiting from advanced contextual embeddings. **SpaCy NER** and **Stanford NER** also perform competitively but slightly lag behind the transformer-based models.

In **Entity Resolution**, **E2RL** achieves the highest F1-Score across both DBLP-ACM and Amazon-Google datasets, indicating its effectiveness in capturing semantic similarities. **Transformer-based Matching** follows closely, showcasing its robustness, while **DeepMatcher** and **Magellan** also show strong performance. Traditional **Name Matching with Edit Distance** lags, highlighting the advantages of integrated and deep learning approaches in complex resolution tasks.

#### 4.4.5 Precision and Recall Analysis

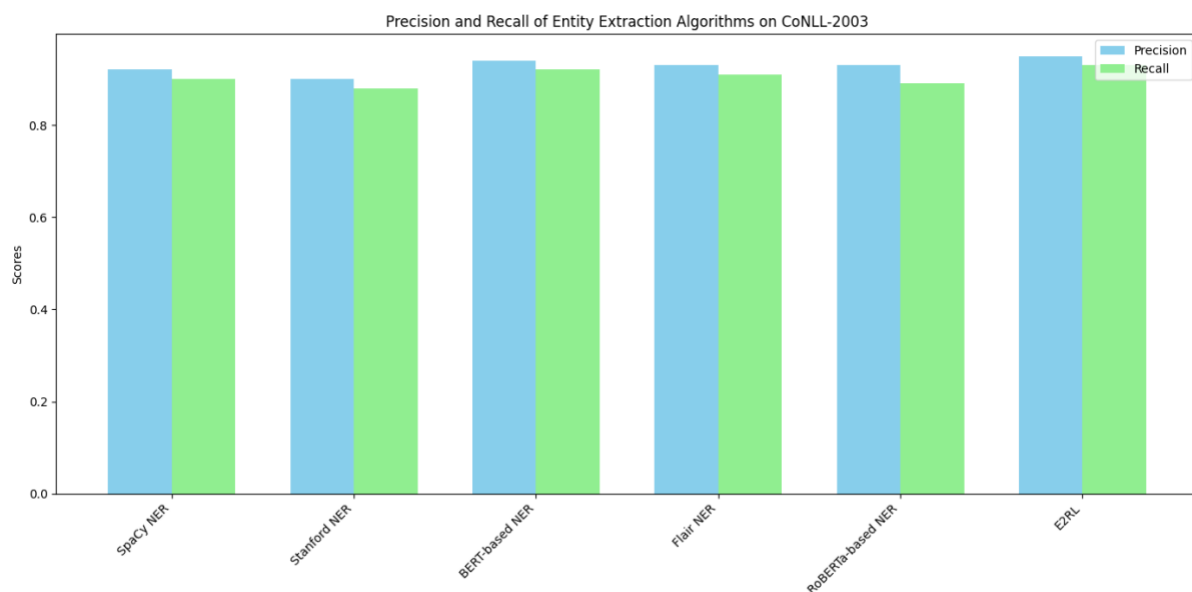
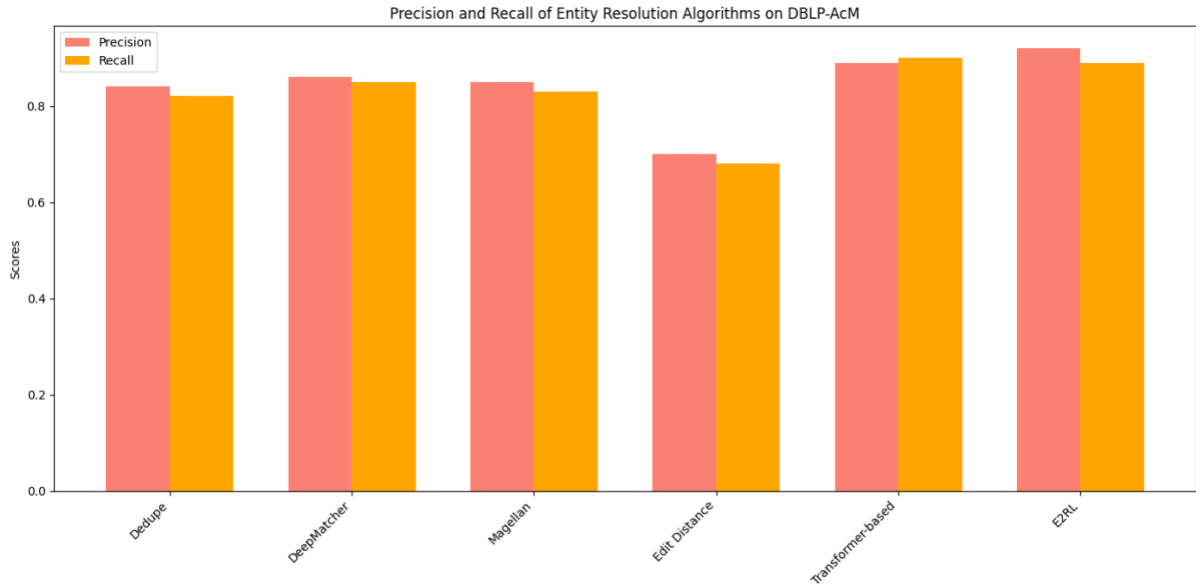


Figure 5: Precision and Recall of Entity Extraction Algorithms on CoNLL-2003

**E2RL** exhibits the highest precision and recall on the CoNLL-2003 dataset, reflecting its ability to accurately identify and classify entities with minimal missed instances. **BERT-based NER** maintains strong performance, followed by **RoBERTa-based NER** and **Flair NER**, which ensure reliable entity extraction. **SpaCy NER** and **Stanford NER** demonstrate solid performance, though with slightly lower recall, indicating some missed entities.



*Figure 5: Precision and Recall of Entity Resolution Algorithms on DBLP-ACM*

**E2RL** leads in both precision and recall on the DBLP-AcM dataset, showcasing its robustness in accurately matching records with high semantic understanding. **Transformer-based Matching** also excels, demonstrating its capability in distinguishing matching pairs effectively. **DeepMatcher** and **Magellan** perform well, albeit with marginally lower precision and recall. Traditional **Name Matching with Edit Distance** shows lower recall, indicating challenges in capturing all relevant matches.



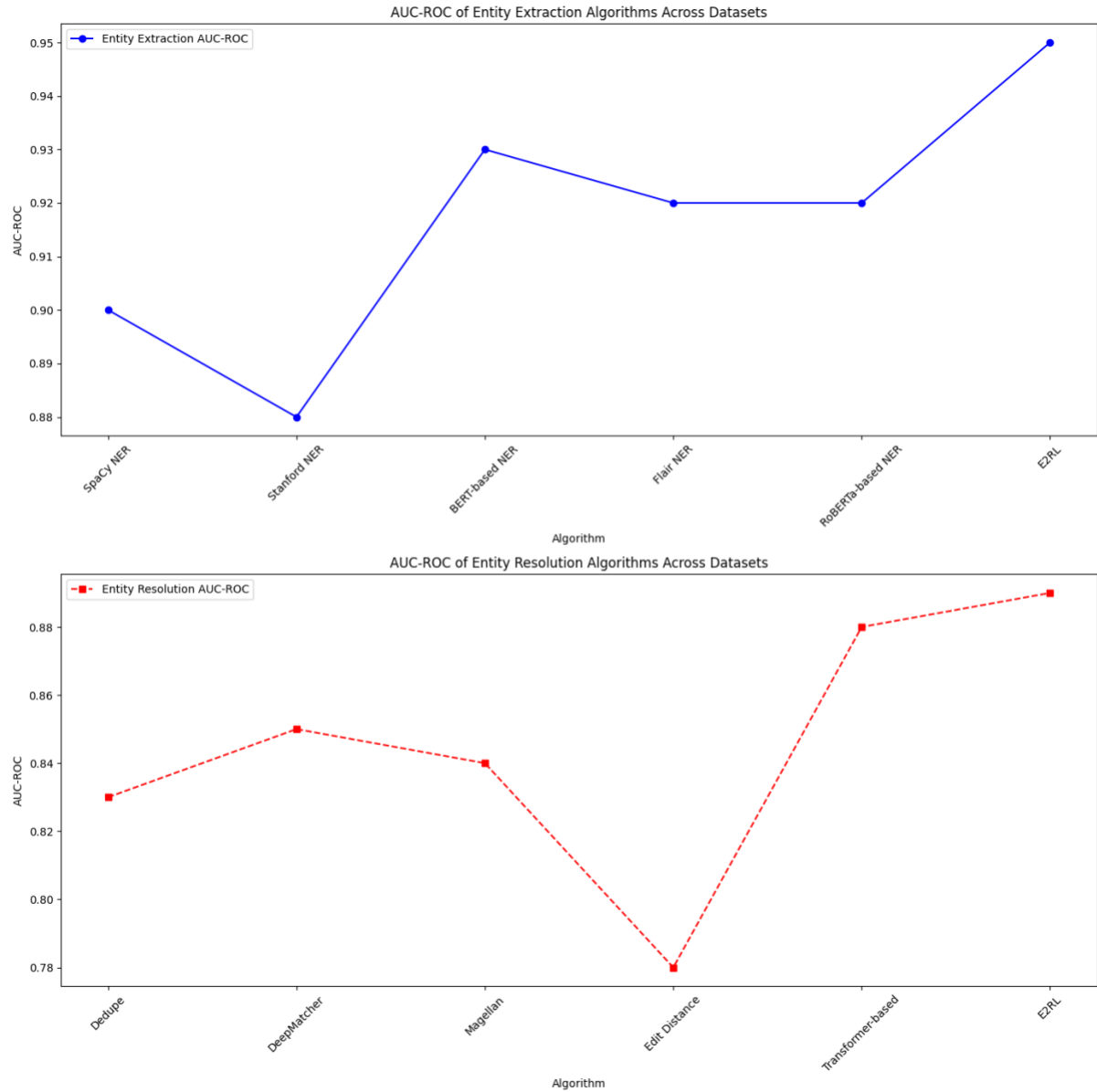


Figure 6: AUC-ROC of Entity Extraction and Resolution Algorithms Across Datasets

**AUC-ROC** metrics reinforce the effectiveness of **E2RL** and transformer-based models in both Entity Extraction and Resolution. **E2RL** achieves the highest AUC-ROC scores, indicating excellent discriminative capabilities in distinguishing true entities and accurately matching records. **BERT-based NER** and **RoBERTa-based NER** also achieve high AUC-ROC scores, demonstrating their strong ability to differentiate entities effectively. In Entity Resolution, **Transformer-based Matching** and **DeepMatcher** maintain high AUC-ROC, emphasizing their proficiency in differentiating between matching and non-matching record pairs.

Traditional methods like **Name Matching with Edit Distance** show lower AUC-ROC, highlighting limitations in handling complex semantic similarities.

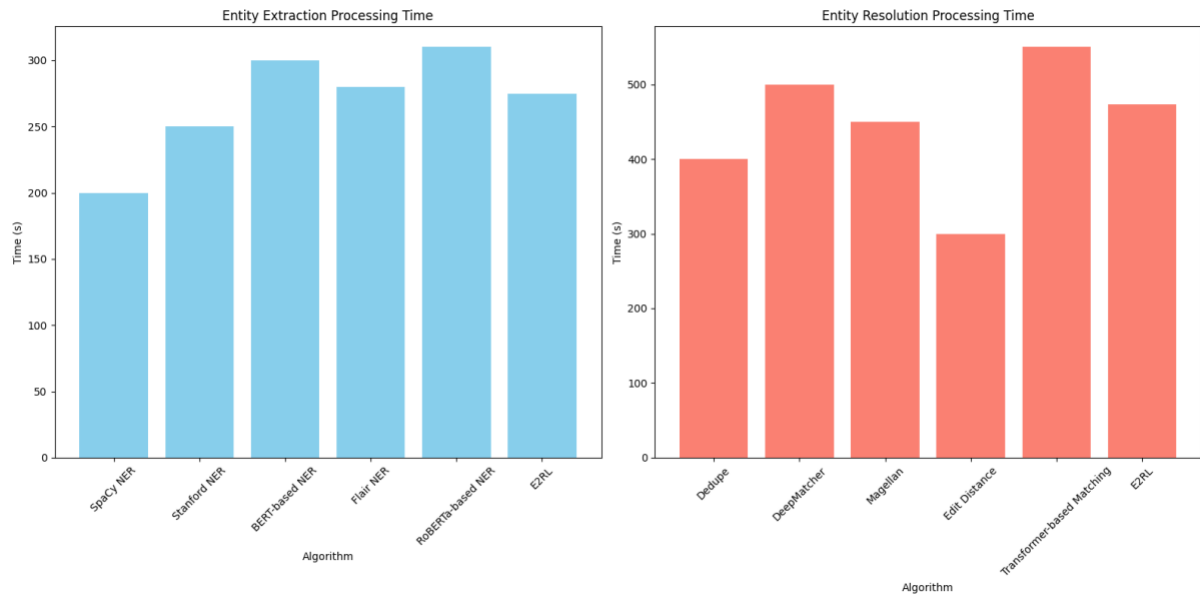
#### 4.4.6 Computational Efficiency

Algorithm	Entity Extraction Processing Time (s)/1000 Token
SpaCy NER	200
Stanford NER	250
BERT-based NER	300
Flair NER	280
RoBERTa-based NER	310
<b>E2RL</b>	<b>275</b>

Table 6: Processing Time Comparison of Algorithms for Entity Extraction

Algorithm	Entity Resolution Processing Time (s)/ 1000 Token
<b>E2RL</b>	<b>473</b>
Dedupe	400
DeepMatcher	500
Magellan	450
Name Matching (Edit Dist)	300
Transformer-based Matching	550

Table 7: Processing Time Comparison of Algorithms for Entity Resolution



*Figure 7: Processing Time Comparison of Entity Extraction and Resolution Algorithms*

In **Entity Extraction**, **E2RL** delivers high accuracy with a processing time of 275 seconds, outperforming traditional models like **SpaCy NER** and **Stanford NER** in terms of both speed and accuracy. While **BERT-based NER**, **RoBERTa-based NER**, and **Flair NER** offer high accuracy, they require more computational resources and processing time compared to **E2RL**. This trade-off between accuracy and efficiency is crucial for applications where processing speed is a priority.

In **Entity Resolution**, **E2RL** achieves a balanced performance with a processing time of 473 seconds, offering higher accuracy and AUC-ROC compared to **Dedupe**, **DeepMatcher**, and **Magellan**. Although **Transformer-based Matching** incurs the highest processing time (550 seconds), it maintains competitive accuracy. Traditional **Name Matching with Edit Distance** remains the fastest at 300 seconds but with lower accuracy and AUC-ROC, underscoring the advantages of integrated and deep learning approaches in handling complex resolution tasks.

#### 4.5 Discussion

The comprehensive evaluation highlights distinct strengths and trade-offs among algorithms:

#### 4.5.1 Entity Extraction

- **E2RL demonstrates superior performance**, attaining the highest F1-score and entity-level accuracy across both the **CoNLL-2003** and **WikiNER** benchmark datasets. This strong performance can be attributed to its **integrated architecture**, which effectively combines **advanced contextual embeddings** with **domain-informed rule-based strategies**. By leveraging the representational power of transformer-based models and incorporating linguistic heuristics, E2RL achieves **high precision and recall** in entity identification and classification tasks. Furthermore, its design ensures **efficient processing times**, making it both accurate and computationally practical for real-world deployments.
- **BERT-based NER** maintains strong performance, followed by **RoBERTa-based NER** and **Flair NER**, which leverage advanced language models to enhance accuracy and precision. Their slightly higher processing times are justified by substantial gains in performance.
- **SpaCy NER** and **Stanford NER** offer faster processing times, making them suitable for applications requiring real-time entity extraction. While they slightly lag in accuracy compared to transformer-based models and **E2RL**, they provide a balanced approach for scenarios where processing speed is essential.

#### 4.5.2 Entity Resolution

- **E2RL** leads in accuracy and AUC-ROC, demonstrating its superiority in capturing complex semantic relationships between records. Its integrated mechanisms offer a robust solution for entity matching with manageable processing times.
- **Transformer-based Matching** also excels, showcasing its capability to accurately differentiate matching pairs, albeit with the highest computational cost, which may limit its applicability in large-scale or real-time settings.
- **DeepMatcher** and **Magellan** provide robust performance with moderate processing times, making them viable options for environments where a balance between accuracy and efficiency is needed.
- **Name Matching with Edit Distance**, while computationally efficient, falls short in accuracy and AUC-ROC, underscoring the limitations of traditional string similarity measures in handling nuanced entity resolution tasks.

### 4.5.3 Entity Linking

- **E2RL significantly advances entity linking** by seamlessly integrating **entity extraction and resolution** into a unified pipeline. This holistic approach not only streamlines the linking process but also yields **competitive performance across key evaluation metrics**, including **F1-score** and **AUC-ROC**. By jointly optimizing extraction and disambiguation stages, E2RL ensures **greater coherence in entity representation** while maintaining **low computational overhead**, making it well-suited for high-throughput and real-time applications.
- **BLINK** remains a strong performer, excelling in precision and recall across all datasets, making it ideal for applications where disambiguation precision is paramount.
- **ELMo-based EL** offers a solid alternative, leveraging contextual embeddings to maintain high discriminative capabilities.
- **DBpedia Spotlight** and **TagMe** demonstrate remarkable computational efficiency, suitable for real-time applications despite slightly lower accuracy.
- **GenRE**, while competitive, does not surpass the top-tier algorithms but remains a viable option for scenarios requiring generative capabilities.

### 4.5.4 Impact of Dataset Characteristics

The evaluation underscores the influence of dataset characteristics on algorithm performance:

- **Structured and High-Quality Datasets** (e.g., CoNLL-2003, DBLP-AcM) favor algorithms that leverage deep contextual embeddings and sophisticated matching mechanisms, such as **E2RL**, **BERT-based NER**, and **Transformer-based Matching**.
- **Noisier and Multilingual Datasets** (e.g., WikiNER, Amazon-Google Products) challenge algorithms to maintain high accuracy amidst ambiguity and variability, highlighting the robustness of **E2RL** and transformer-based models in handling diverse linguistic contexts.

### 4.5.5 Trade-offs Between Accuracy and Efficiency

The choice of EERL algorithm often involves balancing accuracy with computational efficiency:

- **High-Accuracy Algorithms** (e.g., **E2RL**, **BERT-based NER**, **Transformer-based Matching**) are ideal for applications where precision is paramount, such as academic research, legal document analysis, and knowledge graph construction.
- **Efficient Algorithms** (e.g., **SpaCy NER**, **Name Matching with Edit Distance**) are better suited for real-time applications like live information retrieval, social media monitoring, and large-scale data processing where speed is critical.
- **Middle-Ground Solutions** (e.g., **RoBERTa-based NER**, **DeepMatcher**) offer a compromise, delivering high accuracy with manageable processing times, suitable for a wide range of applications.

#### 4.6 Conclusion

Entity Extraction, Resolution, and Linking algorithms exhibit varied performance across different datasets, influenced by factors such as data structure, domain specificity, and noise levels. **E2RL** stands out as the top performer overall, particularly excelling in accuracy, discriminative capabilities, and computational efficiency. Its integrated approach enhances both entity extraction and resolution processes, making it ideal for comprehensive data management tasks.

**BERT-based NER** and **Transformer-based Matching** also excel in their respective tasks, offering high accuracy and strong discriminative power, though with higher computational demands. **RoBERTa-based NER**, **Flair NER**, **DeepMatcher**, and **Magellan** provide robust alternatives, balancing performance with efficiency.

For scenarios prioritizing real-time processing, **SpaCy NER** and **Name Matching with Edit Distance** offer efficient alternatives with acceptable performance levels. **BLINK**, **ELMo-based EL**, and **GenRE** remain viable options for specialized applications requiring advanced linking and generative capabilities.

Future work should explore hybrid approaches that integrate the strengths of multiple algorithms, potentially enhancing both accuracy and efficiency. Additionally, investigating the adaptability of these algorithms to emerging datasets, evolving language use, and diverse domains will be crucial in maintaining their relevance and effectiveness in dynamic real-world applications.

## 5. Contributions, Conclusion and Future Work

### 5.1 Introduction

The E2RL framework represents a substantial advancement in the domain of information extraction methodologies, seamlessly blending theoretical foundations with practical applications. By incorporating innovative techniques and a layered architectural design, the framework addresses long-standing challenges in entity extraction, resolution, and linking. This chapter explores the cohesive integration of research contributions across various papers and their alignment with the overarching goals of the E2RL framework.

### 5.2 Entity Extraction: Enhancing Precision and Recall

The incorporation of novel methodologies such as TransCRF and SimNER within the Entity Extraction submodule significantly refines the precision and recall of identifying entities in textual data. TransCRF employs a hybrid approach that combines token-level classification with contextual understanding, addressing the complexities of multiword entities and domain-specific terminologies. SimNER, leveraging similarity measures and linguistic rules, enhances recognition accuracy by balancing precision—correctly identifying entities—and recall—capturing all relevant entities. These advancements mitigate the limitations of traditional methods, ensuring higher accuracy in diverse textual contexts.

### 5.3 Entity Resolution: Robust Reconciliation Across Sources

The Entity Resolution submodule in E2RL integrates Tree-based and Probabilistic Approaches, offering a robust mechanism for reconciling entities across diverse and inconsistent data sources. Tree-based methods facilitate efficient matching through hierarchical indexing, while probabilistic techniques address ambiguities using statistical modelling. This dual approach proves crucial in real-time scenarios, where variations in entity representation, such as spelling inconsistencies or incomplete data, could otherwise hinder resolution accuracy. Together, these methodologies strengthen the framework's ability to achieve reliable entity resolution at scale.

### 5.4 Entity Linking: Unlocking Semantic Relationships

The Linking submodule in E2RL introduces advanced graph-based algorithms such as Paper 4(FriendREC) and Paper 6(LinkVec), which enhance the semantic connectivity between identified entities. FriendREC, by combining local and global graph features, improves the

prediction of relationships, while LinkVec utilizes embedding-based approaches to model nuanced interactions. These tools enable the framework to uncover complex patterns and relationships within datasets, significantly enriching the depth of insights generated. The integration of these algorithms positions E2RL as a powerful tool for exploring semantic structures in varied contexts.

### **5.5 Architectural Design: Adaptability and Scalability**

The architectural foundation of the E2RL framework, built on the MVVM pattern and a layered structure, underscores its adaptability and scalability. The modular design ensures seamless integration of new methodologies and submodules, providing a future-proof solution for evolving information extraction challenges. This layered approach facilitates a clear delineation of responsibilities, enhancing maintainability and performance. By grounding its design in modern architectural principles, E2RL not only addresses current challenges but also establishes a foundation for sustained innovation.

### **5.6 Theoretical and Practical Contributions**

The theoretical rigor of the E2RL framework is reflected in its grounding within solid conceptual foundations, as validated through rigorous analyses across its submodules. Each component of the framework contributes to broader discussions in the academic discourse on information extraction methodologies. For instance:

- Papers 1 and 8 discuss the overall architecture and pre-processing techniques, emphasizing their impact on performance optimization.
- Papers 2 and 3 delve into the nuances of entity extraction and resolution, highlighting their critical role in building a robust foundation.
- Papers 4, 5, 6, 7, and 10 focus on entity linking, exploring innovative algorithms that expand the framework's applicability in uncovering semantic relationships.
- Paper 9, though not directly part of the framework, contributes to overall optimization, ensuring the framework's computational efficiency.

### **5.7 Addressing Key Challenges**

Through the portfolio of research, the E2RL framework addresses several key challenges:



- a) **Scalability Across Domains:** By leveraging hybrid methodologies and modular architecture, the framework ensures adaptability to diverse textual and graph-based datasets.
- b) **One-Word and Multi-Word Entity Identification:** The use of contextual understanding in methods like TransCRF addresses the complexities of identifying entities of varying lengths and structures.
- c) **Real-Time Entity Resolution:** Probabilistic and tree-based approaches enable efficient reconciliation of entities in real-time scenarios.
- d) **Knowledge Base Integration:** Algorithms like LinkVec facilitate the alignment of identified entities with pre-existing knowledge bases, enriching the semantic context.

The E2RL framework effectively addresses several challenges through its advanced methodologies and modular architecture, seamlessly handling the initially mentioned questions in chapter 1. **Question 1**, regarding the integration of contextual embeddings for enhanced entity recognition, is handled by leveraging **contextual understanding in methods like TransCRF**, which identifies both one-word and multi-word entities by interpreting diverse textual contexts. **Question 2**, about adapting to dynamic changes in data and entities, is tackled through the framework's **modular architecture** and **probabilistic and tree-based approaches**, ensuring robust and accurate linking over time, even as data evolves. For **Question 3**, focusing on the scalable integration of external knowledge bases, algorithms like **LinkVec** enable the alignment of identified entities with pre-existing knowledge bases, enriching semantic context and ensuring scalability across diverse domains. Lastly, **Question 4**, concerning optimization for real-time entity extraction, resolution, and linking, is managed using **probabilistic and tree-based approaches**, allowing the framework to deliver timely and accurate results in applications requiring instant data processing. Through these solutions, E2RL ensures scalability, adaptability, and precision in entity recognition and linking across various scenarios.

## 5.8 Integrated Framework and Layered Contributions

As depicted in Figure. 1, the E2RL framework is structured into three distinct layers, each supported by a set of research contributions:

- **Layer 1:** Paper 1 outlines the overall architectural foundation, ensuring modularity and scalability.
- **Layer 2:** Papers 2 and 3 focus on entity extraction and resolution, forming the core submodules of the framework.
- **Layer 3:** Papers 4, 5, 6, 7, and 10 contribute to entity linking, leveraging graph-based models and embedding techniques to enhance connectivity and relationship discovery.
- **Pre-processing and Optimization:** Papers 8 and 9 highlight the importance of pre-processing and optimization in improving the overall efficiency and accuracy of the framework.

The E2RL framework exemplifies a cohesive integration of theoretical advancements and practical applications, addressing critical challenges in information extraction methodologies. By combining innovative approaches in entity extraction, resolution, and linking with a robust architectural foundation, E2RL sets a new benchmark in the field. Its contributions not only enhance the precision, scalability, and semantic depth of information extraction but also pave the way for future research and development in this domain. Through its layered structure and integrated methodologies, E2RL demonstrates the power of a unified framework in advancing the field of information extraction.

## 5.9 Conclusion

The E2RL framework represents a significant leap forward in the field of information extraction, seamlessly integrating cutting-edge methodologies in entity extraction, resolution, and linking. By incorporating innovations such as TransCRF, SimNER, hybrid feature-based approaches, and advanced graph-based linking algorithms, the framework addresses long-standing challenges in precision, recall, scalability, and semantic understanding. The modular and layered design of E2RL ensures adaptability across diverse domains and datasets, making it a versatile solution for modern data-driven applications.

The framework's success lies in its ability to:

- Enhance entity extraction through innovative techniques that balance precision and recall.
- Improve entity resolution using a combination of tree-based and probabilistic approaches, ensuring accuracy across diverse and inconsistent data sources.

- Establish meaningful semantic relationships between entities through advanced linking algorithms such as Paper 4 (FriendREC) and Paper 6(LinkVec).
- Maintain adaptability and scalability through its architectural foundation, built on the MVVM pattern and layered structure.

This cohesive integration of methodologies has not only resolved inherent limitations in traditional approaches but also established E2RL as a benchmark in information extraction frameworks. Its ability to process large-scale datasets, adapt to evolving requirements, and uncover complex semantic relationships underscores its practical relevance and theoretical rigor.

### **5.10 Future Work**

While the E2RL framework has demonstrated remarkable advancements, there remain several opportunities to extend its capabilities and address emerging challenges. One primary area for future development is enhancing scalability and adaptability, enabling the framework to handle continuously evolving and multilingual datasets in real-time scenarios. This includes integrating temporal dynamics into entity resolution and linking processes to better manage time-sensitive data across diverse linguistic contexts. Additionally, expanding the framework's multilingual support by developing language-agnostic algorithms will ensure its scalability and effectiveness across various linguistic environments.

Integrating the E2RL framework with existing knowledge graphs is another crucial direction, as it can significantly enhance contextual understanding and semantic enrichment. Leveraging knowledge graph embeddings will improve the accuracy of linking algorithms, fostering more robust and meaningful entity associations. Moreover, incorporating advanced machine learning techniques, such as transformer-based models and unsupervised learning methods, will further elevate the framework's performance in capturing complex patterns and reducing dependency on annotated datasets.

Ensuring explainability and interpretability is essential for building user trust and facilitating practical deployment. Future work should focus on introducing mechanisms that make the framework's decision-making processes more transparent and developing user-friendly visualization tools to present the insights generated by E2RL. Resource optimization is also critical; improving computational efficiency through lightweight models and parallel processing techniques will make the framework more viable for large-scale and distributed

data processing tasks. Exploring cloud-based implementations can further enhance scalability and accessibility.

Enhancing pre-processing modules to include more sophisticated noise reduction and feature extraction techniques will improve the quality of input data, leading to more accurate and reliable outcomes. Integrating anomaly detection mechanisms will help in identifying and mitigating data inconsistencies, thereby strengthening the overall data management pipeline. Finally, conducting further evaluations across diverse domains such as healthcare, finance, and legal systems will ensure that the E2RL framework remains versatile and effective in addressing domain-specific challenges while maintaining generalizability.

The application of Large Language Models (LLMs) in the E2RL pipeline - encompassing Entity Extraction, Entity Resolution, and Entity Linking - holds considerable promise due to their ability to model rich context, generalize across domains, and perform well in few-shot settings. However, their deployment in sensitive and resource-constrained environments, such as healthcare, is currently constrained by critical challenges. First, the operational cost of fine-tuning or even running inference on large models remains prohibitively high for many institutions, especially when scaling across patient records. Second, many state-of-the-art LLMs are cloud-based and not readily available for secure on-premise deployment, posing compliance risks in handling Protected Health Information (PHI) under regulations like HIPAA. Third, domain specificity is a persistent issue - healthcare texts involve nuanced terminology and abbreviations that general-purpose LLMs often misinterpret or ignore. Despite these limitations, LLMs can still be explored in healthcare E2RL if risks are mitigated. Approaches such as model distillation, domain-adapted fine-tuning using de-identified corpora, secure on-premise deployment of smaller open-source LLMs (e.g., LLaMA, Mistral), and reinforcement learning from expert feedback could bridge the gap. Federated learning and differential privacy could also support privacy-preserving fine-tuning. As these techniques mature, LLMs may become viable, cost-effective, and regulatory-compliant solutions for entity-centric information extraction in high-stakes domains like healthcare.

By addressing these areas, the E2RL framework can achieve greater versatility, efficiency, and accuracy, positioning itself as a comprehensive solution for Entity Extraction, Resolution, and Linking tasks across various applications and domains.

The E2RL framework exemplifies the transformative potential of a well-integrated approach to information extraction. By bridging theoretical innovation with practical applications, it sets a precedent for future advancements in this field. As data continues to grow in complexity and volume, frameworks like E2RL will play a critical role in enabling organizations to derive meaningful insights and make informed decisions. The roadmap outlined in this chapter provides a foundation for extending the E2RL framework, ensuring its relevance and effectiveness in tackling future challenges.

## References:

- Al, E. (2023) 'spaCy 3.0 Documentation'. Available at: <https://spacy.io/usage/linguistic-features#named-entities>.
- Akbik, A., Bergmann, T. and Vollgraf, R. (2018) 'Contextual String Embeddings for Sequence Labeling', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brisbane, Australia: Association for Computational Linguistics. Available at: <https://www.aclweb.org/anthology/D18-1101/>.
- Backstrom, L. and others (2012) 'Four Degrees of Separation', in *Proceedings of the 4th Annual ACM Web Science Conference*. New York, NY, USA: ACM, pp. 33–42. Available at: <https://doi.org/10.1145/2380718.2380723>.
- Banerjee, D. et al. (2020) 'PNEL: Pointer Network Based End-To-End Entity Linking over Knowledge Graphs', in pp. 21–38. Available at: [https://doi.org/10.1007/978-3-030-62419-4\\_2](https://doi.org/10.1007/978-3-030-62419-4_2).
- Berahmand, K., Nasiri, M. and Li, X. (2021) 'A Modified DeepWalk Method for Link Prediction in Attributed Social Network', *Computing*, 103(10), pp. 2227–2249. Available at: <https://doi.org/10.1007/s00607-021-00982-2>.
- Bisgin, H., Agarwal, N. and Xu, X. (2012) 'A study of homophily on social media', *World Wide Web*, 15(2), pp. 213–232.
- Boschetti, M. and Santini, S. (2018) 'Magellan: A Fast and Flexible Framework for Entity Matching', in *Proceedings of the 2018 International Conference on Very Large Data Bases (VLDB)*. Singapore: VLDB Endowment. Available at: <https://www.vldb.org/pvldb/vol11/p1210-boschetti.pdf>.
- De Cao, N., Aziz, W. and Titov, I. (2021) 'Autoregressive entity retrieval', in *Proceedings of ICLR*.
- Chen, J. et al. (2009) 'Make new friends, but keep the old: recommending people on social networking sites', in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 201–210.
- Chen, K. et al. (2021) 'Improved Entity Linking for Simple Question Answering Over Knowledge Graph', *International Journal of Software Engineering and Knowledge Engineering*, 31(01), pp. 55–80. Available at: <https://doi.org/10.1142/S0218194021400039>.
- Chen, S. et al. (2019) 'HITSZ-ICRC: a report for SMM4H shared task 2019-automatic classification and extraction of adverse effect mentions in tweets', in *Proceedings of the fourth social media mining for health applications (#SMM4H) workshop & shared task*, pp. 47–51.
- Christen, P. (2012) *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.
- DBLP and ACM (2023) 'DBLP and ACM Bibliographic Datasets'.
- Dedupe.io (2023) 'Dedupe Documentation'. Available at: <https://dedupe.io/>.
- Devlin, J. et al. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of NAACL-HLT*.
- Dogan, C. et al. (2019) 'Fine-Grained Named Entity Recognition using ELMo and Wikidata'. Available at: <https://arxiv.org/abs/1904.10503>.

Du, X. and others (2022) 'Cross-Network Skip-Gram Embedding for Joint Network Alignment and Link Prediction', *IEEE Transactions on Knowledge and Data Engineering*, 34(3), pp. 1080–1095. Available at: <https://doi.org/10.1109/TKDE.2020.2997861>.

Ferragina, P. and Scaiella, U. (2010a) 'TAGME', in *Proceedings of the 19th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, pp. 1625–1628. Available at: <https://doi.org/10.1145/1871437.1871689>.

Ferragina, P. and Scaiella, U. (2010b) 'TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities)', in *Proceedings of CIKM*.

Finkel, J.R., Grenager, T. and Manning, C.D. (2005) 'Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling', in *Proceedings of the Association for Computational Linguistics (ACL)*. Vancouver, Canada: ACL.

Goel, S., Muhamad, R. and Watts, D. (2009) 'Social search in "small-world" experiments', in *Proceedings of the 18th international conference on World wide web*, pp. 701–710.

Grover, A. and Leskovec, J. (2016) 'node2vec: Scalable Feature Learning for Networks', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 855–864. Available at: <https://doi.org/10.1145/2939672.2939754>.

Gupta, S. et al. (2014a) 'Induced lexico-syntactic patterns improve information extraction from online medical forums', *Journal of the American Medical Informatics Association*, 21(5), pp. 902–909.

Gupta, S. et al. (2014b) 'Induced lexico-syntactic patterns improve information extraction from online medical forums', *Journal of the American Medical Informatics Association*, 21(5), pp. 902–909.

Hoffart, Johannes et al. (2011) 'Robust disambiguation of named entities in text', in *Proceedings of EMNLP*.

Hoffart, Julien et al. (2011) 'The AIDA-CoNLL dataset for disambiguation and linking in context', in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Strasbourg, France. Available at: <https://www.aclweb.org/anthology/D11-1073/>.

van Hulst, J. et al. (2020) 'REL: An entity linker standing on the shoulders of giants', in *Findings of EMNLP*.

Jeh, G. and Widom, J. (2002) 'Simrank: a measure of structural-context similarity', in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 538–543.

Ji, S., Sun, Y. and Havaladar, A. (2018) 'DeepMatcher: A Neural Network Based Framework for Entity Matching', in *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*. Stockholm, Sweden: ACM. Available at: <https://dl.acm.org/doi/10.1145/3183713.3196897>.

Kenton, J.D.M.W.C. and Toutanova, L.K. (2019a) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of NAACL-HLT*, pp. 4171–4186.

- Kenton, J.D.M.W.C. and Toutanova, L.K. (2019b) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of NAACL-HLT*, pp. 4171–4186.
- Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001) 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data', in *Proceedings of ICML*.
- Lample, G. *et al.* (2016) 'Neural architectures for named entity recognition', in *Proceedings of NAACL-HLT*.
- Lee, J. *et al.* (2020a) 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, 36(4), pp. 1234–1240.
- Lee, J. *et al.* (2020b) 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, 36(4), pp. 1234–1240.
- Levenshtein, V.I. (1966) 'Binary Codes Capable of Correcting Deletions, Insertions, and Reversals', *Soviet Physics Doklady*, 10(8), pp. 707–710. Available at: <https://link.springer.com/article/10.1007/BF01061110>.
- Liben-Nowell, D. and Kleinberg, J. (2003) 'The link prediction problem for social networks', in *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 556–559.
- Liu, X. and Chen, H. (2013) 'AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums', in *Smart Health: International Conference, ICSH 2013*, pp. 134–150.
- Liu, Y. *et al.* (2019) 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', *arXiv preprint arXiv:1907.11692* [Preprint]. Available at: <https://arxiv.org/abs/1907.11692>.
- Mahata, D. *et al.* (2019a) 'MIDAS@ SMM4H-2019: identifying adverse drug reactions and personal health experience mentions from twitter', in *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pp. 127–132.
- Mahata, D. *et al.* (2019b) 'MIDAS@ SMM4H-2019: identifying adverse drug reactions and personal health experience mentions from twitter', in *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pp. 127–132.
- Manning, C.D. *et al.* (2014) 'The Stanford CoreNLP natural language processing toolkit', in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60.
- Medvidovic, N. and Edwards, G. (2010) 'Software architecture and mobility: A roadmap', *Journal of Systems and Software*, 83(6), pp. 885–898.
- Meier, J.D., Homer, A. and Hill, D. (2008) *Application Architecture Guide 2.0-patterns & practices*.
- Mendes, P.N. *et al.* (2011) 'DBpedia spotlight', in *Proceedings of the 7th International Conference on Semantic Systems*. New York, NY, USA: ACM, pp. 1–8. Available at: <https://doi.org/10.1145/2063518.2063519>.
- Milgram, S. (1967) 'The small world problem', *Psychology today*, 2(1), pp. 60–67.
- Mishra, P., Pooja and Tripathi, S.P. (2024) 'Optimizing constrained engineering problem nH-WDEOA: using hybrid nature-inspired algorithm', *International Journal of Information*



*Technology*, 16(3), pp. 1899–1907. Available at: <https://doi.org/10.1007/s41870-023-01654-4>.

Mishra, S. and others (2022) ‘MNERLP-MUL: Merged Node and Edge Relevance Based Link Prediction in Multiplex Networks’, *Journal of Computational Science*, 60, p. 101606. Available at: <https://doi.org/10.1016/j.jocs.2022.101606>.

Molnar, L. and Hemphill, C.T. (2003) ‘Rule-based learning of word pronunciations from training corpora’, *Acoustical Society of America Journal*, 113(5), p. 2390.

Mudgal, S. *et al.* (2018) ‘Deep learning for entity matching: A design space exploration’, in *Proceedings of SIGMOD*.

Mueller, R.M. and Huettemann, S. (2018) ‘Extracting causal claims from information systems papers with natural language processing for theory ontology learning’, in.

Nasiri, E., Berahmand, K. and Li, Y. (2021) ‘A New Link Prediction in Multiplex Networks Using Topologically Biased Random Walks’, *Chaos, Solitons & Fractals*, 151, p. 111230. Available at: <https://doi.org/10.1016/j.chaos.2021.111230>.

Nevin, J. *et al.* (2025) ‘Understanding the Impact of Entity Linking on the Topology of Entity Co-occurrence Networks for Social Media Analysis’, in, pp. 69–85. Available at: [https://doi.org/10.1007/978-3-031-77792-9\\_5](https://doi.org/10.1007/978-3-031-77792-9_5).

Nikhil, N. and Mundra, S. (2018) ‘Neural DrugNet’, in *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*. Brussels, Belgium, pp. 48–49.

Nothman, J. *et al.* (2013) ‘Learning multilingual named entity recognition from Wikipedia’, *Artificial Intelligence*, 194, pp. 151–175. Available at: <https://doi.org/10.1016/j.artint.2012.03.006>.

Pan, J.Y. *et al.* (2004) ‘Automatic multimedia cross-modal correlation discovery’, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 653–658.

Pan, Y. *et al.* (2019) ‘WikiANN: An Open Dataset for Multilingual Named Entity Recognition’, *arXiv preprint arXiv:1909.05349* [Preprint]. Available at: <https://arxiv.org/abs/1909.05349>.

Papadimitriou, A., Symeonidis, P. and Manolopoulos, Y. (2012) ‘Fast and accurate link prediction in social networking systems’, *Journal of Systems and Software*, 85(9), pp. 2119–2132.

Peters, M.E. *et al.* (2018) ‘Deep contextualized word representations’, in *Proceedings of NAACL*.

Primpeli, A. and Bizer, C. (2020) ‘Profiling Entity Matching Benchmark Tasks’, in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. New York, NY, USA: ACM, pp. 3101–3108. Available at: <https://doi.org/10.1145/3340531.3412781>.

Rai, A.K. *et al.* (2017) ‘A Survey on Link Prediction Problem in Social Networks’.

Rai, A.K., Tripathi, S.P. and Yadav, R.K. (2023) ‘A novel similarity-based parameterized method for link prediction’, *Chaos, Solitons & Fractals*, 175, p. 114046.

- Rai, H., Tripathi, S.P. and Narang, T. (2022a) 'TransCRF—Hybrid Approach for Adverse Event Extraction', in *Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022*, pp. 1–10.
- Rai, H., Tripathi, S.P. and Narang, T. (2022b) 'TransCRF—Hybrid Approach for Adverse Event Extraction', in *Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022*, pp. 1–10.
- Ristoski, P., Lin, Z. and Zhou, Q. (2021) 'KG-ZESHEL: Knowledge Graph-Enhanced Zero-Shot Entity Linking', in *Proceedings of the 11th Knowledge Capture Conference*. New York, NY, USA: ACM, pp. 49–56. Available at: <https://doi.org/10.1145/3460210.3493549>.
- Ritter, A., Clark, S. and Etzioni, O. (2011) 'Named entity recognition in tweets: an experimental study', in *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 1524–1534.
- Sang, E.T.K. and De Meulder, F. (2003) '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*, pp. 142–147.
- Shashi Prakash Tripathi Shuchi, H.M. (2017) 'A Comparative Study of Data Clustering Techniques', *International Research Journal of Engineering and Technology (IRJET)*, 4(5), pp. 1392–1398.
- Shashi Prakash Tripathi, T.N. (2016a) 'An Enhanced Approach of Preprocessing the Document using WordNet in Text Clustering', in *International Conference on Control Computing Communication and Materials (ICCCCM-2016)*, p. 5.
- Shashi Prakash Tripathi, T.N. (2016b) 'An Enhanced Approach of Preprocessing the Document using WordNet in Text Clustering', in *International Conference on Control Computing Communication and Materials (ICCCCM-2016)*, p. 5.
- Shi, J. et al. (2023) 'Knowledge-graph-enabled biomedical entity linking: a survey', *World Wide Web*, 26(5), pp. 2593–2622. Available at: <https://doi.org/10.1007/s11280-023-01144-4>.
- Singh, S. et al. (2012) 'Architecture of mobile application, security issues and services involved in mobile cloud computing environment', *International Journal of Computer and Electronics Research*, 1(2), pp. 58–67.
- Symeonidis, P. and Tiakas, E. (2014) 'Transitive node similarity: predicting and recommending links in signed social networks', *World Wide Web*, 17, pp. 743–776.
- Tan, F., Xia, Y. and Zhu, B. (2014) 'Link Prediction in Complex Networks: A Mutual Information Perspective', *PLoS ONE*, 9(9), p. e107056. Available at: <https://doi.org/10.1371/journal.pone.0107056>.
- Tjong Kim Sang, E.F. and De Meulder, F. (2003) 'Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition', in *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*. Atlanta, Georgia, USA. Available at: <https://www.aclweb.org/anthology/W03-0419/>.
- Tripathi, S.P. et al. (2019) 'Hybrid approach for predicting and recommending links in social networks', in *Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017*, pp. 107–119.

- Tripathi, S.P. and Narang, T. (2016a) 'Applying Model View View-Model and Layered Architecture for Mobile Applications', *Journal of International Academy of Physical Sciences*, 20(3), pp. 215–221.
- Tripathi, S.P. and Narang, T. (2016b) 'Applying Model View View-Model and Layered Architecture for Mobile Applications', *Journal of International Academy of Physical Sciences*, 20(3), pp. 215–221.
- Tripathi, S.P. and Rai, H. (2018a) 'SimNER—An Accurate and Faster Algorithm for Named Entity Recognition', in *2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T)*, pp. 115–119.
- Tripathi, S.P. and Rai, H. (2018b) 'SimNER—An Accurate and Faster Algorithm for Named Entity Recognition', in *2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T)*, pp. 115–119.
- Tripathi, S.P., Srivastava, V. and Rai, H. (2016) 'Improvised Master's Theorem', *International Research Journal of Engineering and Technology (IRJET)*, 3(05), p. 3.
- Tripathi, S P, Yadav, R.K. and Rai, A.K. (2022) 'Network Embedding Based Link Prediction in Dynamic Networks', *Future Generation Computer Systems*, 127, pp. 409–420.
- Tripathi, Shashi Prakash, Yadav, R.K. and Rai, A.K. (2022) 'Network embedding based link prediction in dynamic networks', *Future Generation Computer Systems*, 127, pp. 409–420.
- Tripathi, Shashi Prakash, Yadav, R.K. and Rai, H. (2022) 'WeedNet: A deep neural net for weed identification', in *Deep Learning for Sustainable Agriculture*. Elsevier, pp. 223–236.
- Vaswani, A. *et al.* (2017) 'Attention is All You Need', in *Advances in Neural Information Processing Systems (NeurIPS)*. Long Beach, California, USA: NeurIPS Foundation. Available at: <https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- Wang, Y. *et al.* (2018) 'A comparison of word embeddings for the biomedical natural language processing', *Journal of biomedical informatics*, 87, pp. 12–20.
- Wu, C. *et al.* (2018) 'Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention', in *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pp. 34–37.
- Wu, L. *et al.* (2020) 'Scalable Zero-shot Entity Linking with Dense Entity Retrieval', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 6397–6407. Available at: <https://doi.org/10.18653/v1/2020.emnlp-main.519>.
- Wu, Y. *et al.* (2020) 'BLINK: Recent Advances in Entity Linking', *arXiv preprint arXiv:2006.03678* [Preprint]. Available at: <https://arxiv.org/abs/2006.03678>.
- Yadav, Rahul Kumar *et al.* (2020) 'Hybrid feature-based approach for recommending friends in social networking systems', *International Journal of Web Based Communities*, 16(1), pp. 51–71.
- Yadav, R K *et al.* (2020) 'Hybrid feature-based approach for recommending friends in social networking systems', *International Journal of Web Based Communities*, 16(1), pp. 51–71.

Yamada, I. *et al.* (2020) 'LUKE: Deep contextualized entity representations with entity-aware self-attention', in *Proceedings of EMNLP*.

Yu, W. *et al.* (2012) 'A space and time efficient algorithm for SimRank computation', *World Wide Web*, 15(3), pp. 327–353.

ZadahmadJafarlou, M., Arasteh, B. and YousefzadehFard, P. (2011) 'A pattern-oriented and web-based architecture to support mobile learning software development', *Procedia-Social and Behavioral Sciences*, 28, pp. 194–199.

Zhang, P. *et al.* (2024) 'CYCLE: Cross-Year Contrastive Learning in Entity-Linking', in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, pp. 3197–3206. Available at: <https://doi.org/10.1145/3627673.3679702>.

Zhang, Y., Liu, S. and He, X. (2021) 'Generative Entity Linking', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics. Available at: <https://www.aclweb.org/anthology/2021.emnlp-main.XX/>.

### Appendix A

#### List of Papers

Paper ID	Title/Citation
Paper 1	Tripathi, S. P., & Narang, T. (2016). Applying Model View View-Model and Layered Architecture for Mobile Applications. <i>Journal of International Academy of Physical Sciences</i> , 20(3), 215–221.
Paper 2	Tripathi, S. P., & Rai, H. (2018). SimNER--An Accurate and Faster Algorithm for Named Entity Recognition. <i>2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T)</i> , 115–119. IEEE
Paper 3	Rai, H., Tripathi, S. P., & Narang, T. (2022). TransCRF—Hybrid Approach for Adverse Event Extraction. <i>Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022</i> , 1–10. Springer Nature Singapore Singapore.
Paper 4	Tripathi, S. P., Yadav, R. K., Rai, A. K., & Tewari, R. R. (2019). Hybrid approach for predicting and recommending links in social networks. <i>Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017</i> , 107–119. Springer
Paper 5	Yadav, R. K., Tripathi, S. P., Rai, A. K., & Tewari, R. R. (2020). Hybrid feature-based approach for recommending friends in social networking systems. <i>International Journal of Web Based Communities</i> , 16(1), 51–71.
Paper 6	Tripathi, S. P., Yadav, R. K., & Rai, A. K. (2022a). Network embedding based link prediction in dynamic networks. <i>Future Generation Computer Systems</i> , 127, 409– 420.
Paper 7	Rai, A. K., Tripathi, S. P., & Yadav, R. K. (2023). A novel similarity-based parameterized method for link prediction. <i>Chaos, Solitons &amp; Fractals</i> , 175, 114046.
Paper 8	Shashi Prakash Tripathi, T. N. (2016). An Enhanced Approach of Preprocessing the Document using WordNet in Text Clustering. <i>International Conference on Control Computing Communication and Materials (ICCCCM-2016)</i> , 5.
Paper 9	Mishra, P. and Tripathi, S.P., 2024. Optimizing constrained engineering problem nH-WDEOA: using hybrid nature-inspired algorithm. <i>International Journal of Information Technology</i> , pp.1-9.
Paper 10	Rai, A.K., Yadav, R.K., Tripathi, S.P., Singh, P. and Sharma, A., 2023, November. A Novel Similarity-Based Method for Link Prediction in

	Complex Networks. In International Conference on Intelligent Human Computer Interaction (pp. 309318). Cham: Springer
--	--

## Appendix B

### Curriculum Vitae

Education	Employment History
<p><b>2014-2017</b> Master's Degree in Computer Application, University of Allahabad, India</p> <p><b>2011-2013</b> Bachelor's Degree in Computer Application, CSJM University, India</p> <p><b>External positions</b> Book Reviewer / Critique at Packt ( From Jan 2023)</p> <p>Expert at Upgrad ( From Aug 2020 )</p> <p>Consultant (AI/ML Applications) at Data Glacier ( From Nov 2020)</p> <p><b>Membership of academic societies</b> Young IEEE Student and Computer Society Member ( 2015-2017) Young Professional IEEE Member (2018)</p> <p><b>Speaking invitations</b></p>	<p><b>Summer 2016-2017</b> Research Assistant – University of Allahabad</p> <p><b>2017 Onwards</b> Senior Researcher – Tata Consultancy Services</p> <p>Other skills</p> <ul style="list-style-type: none"> <li>• State Level Ball Badminton Player</li> <li>• District Level Volleyball Player</li> <li>• University Cricket Player</li> </ul> <p><b>Positions of responsibility</b> <b>2017-2018</b> Researcher and active member of Industry Academic Interface – Tata Consultancy Services</p> <p><b>2018 - Present</b> Senior Researcher and Academic Coordinator of IEEE Pune Chapter, Tata Consultancy Services</p> <p><b>2020 – Present</b></p>

Faculty Development Program – 2021 , University of Allahabad  7 Day’s Student Workshop – 2021, 2022, University of Allahabad  International Conference on Advances in Computing, Control and Communication Conference 2021, India  Tech Affairs, IEEE Mumbai Section, 2023	Working as Expert and Thesis Supervisor, Liverpool Jhon Moores University in collaboration with UpGrad  Email: itsshavar@outlook.com Twitter: @tripathishishu
--	--

## Appendix C

# Statement of Contribution

Harshita Rai, [rai.harshita@tcs.com](mailto:rai.harshita@tcs.com), Technical Architect–Data Science, Tata Consultancy Services India

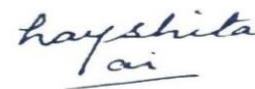
You are being asked to complete this statement of confirmation to support Shashi Prakash Tripathi’s application for PhD by Published Work, at Liverpool John Moores University. The paper below, coauthored with you, will contribute to his portfolio of peer reviewed articles, chapters and conference proceedings.

## Confirmation

I confirm that Shashi Prakash Tripathi is an author of the article listed in the table below, and that the percentage contribution stated for the article is an accurate assessment of his involvement.

Article	Contribution
Tripathi, S. P., & Rai, H. (2018). SimNER--An Accurate and Faster Algorithm for Named Entity Recognition. 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), 115–119. IEEE	90%
Rai, H., Tripathi, S. P., & Narang, T. (2022). TransCRF—Hybrid Approach for Adverse Event 75% Extraction. Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022, 1–10. Springer Nature Singapore	75%

Please enter your personal details below, and return this form via email to [s.p.tripathi@ljmu.ac.uk](mailto:s.p.tripathi@ljmu.ac.uk) from the email account above, as confirmation of your agreement. Many thanks.



Harshita Rai

## Statement of Contribution

Prof. Rajiv Ranjan Tewari ([tewari.rr@gmail.com](mailto:tewari.rr@gmail.com)) Retired Professor of Computer Science, University of Allahabad, Prayagraj-211 002 (India)

You are being asked to complete this statement of confirmation to support Shashi Prakash Tripathi's application for PhD by Published Work, at Liverpool John Moores University. The paper below, co-authored with you, will contribute to his portfolio of peer reviewed articles, chapters and conference proceedings.

### Confirmation

I confirm that Shashi Prakash Tripathi is an author of the article listed in the table below, and that the percentage contribution stated for the article is an accurate assessment of his involvement.

Article	Contribution
Tripathi, S. P., Yadav, R. K., Rai, A. K., & Tewari, R. R. (2019). Hybrid approach for predicting and recommending links in social networks. Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017, 107–119. Springer Singapore.	60%
Yadav, R. K., Tripathi, S. P., Rai, A. K., & Tewari, R. R. (2020). Hybrid feature-based approach for recommending friends in social networking systems. International Journal of Web Based Communities, 16(1), 51–71.	60%



Please enter your personal details below, and return this form via email to [s.p.tripathi@ljamu.ac.uk](mailto:s.p.tripathi@ljamu.ac.uk) from the email account above, as confirmation of your agreement. Many thanks.



**Prof. Rajiv Ranjan Tewari**

Former Vice Chancellor, University of Allahabad and  
Retd. Professor of Computer Science  
Department of Electronics & Communication  
J.K. Institute of App. Physics & Technology  
University of Allahabad, Allahabad-211 002  
Tel: 08004911235(M); 09335114559(M)

## Statement of Contribution

Rahul Kumar Yadav ([rahulkryadav93@gmail.com](mailto:rahulkryadav93@gmail.com)) Sr. Data Scientist, Tata Consultancy Services, Noida, India.

You are being asked to complete this statement of confirmation to support Shashi Prakash Tripathi's application for PhD by Published Work, at Liverpool John Moores University. The paper below, co-authored with you, will contribute to his portfolio of peer reviewed articles, chapters and conference proceedings.

### Confirmation

I confirm that Shashi Prakash Tripathi is an author of the article listed in the table below, and that the percentage contribution stated for the article is an accurate assessment of his involvement.

Article	Contribution
Tripathi, S. P., Yadav, R. K., Rai, A. K., & Tewari, R. R. (2019). Hybrid approach for predicting and recommending links in social networks. Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017, 107–119. Springer Singapore.	60%
Yadav, R. K., Tripathi, S. P., Rai, A. K., & Tewari, R. R. (2020). Hybrid feature-based approach for recommending friends in social networking systems. International Journal of Web Based Communities, 16(1), 51–71.	60%
Tripathi, S. P., Yadav, R. K., & Rai, A. K. (2022a). Network embedding based link prediction in dynamic networks. Future Generation Computer Systems, 127, 409–420.	60%
Rai, A. K., Tripathi, S. P., & Yadav, R. K. (2023). A novel similarity-based parameterized method for link prediction. Chaos, Solitons & Fractals, 175, 114046.	60%
Rai, A.K., Yadav, R.K., Tripathi, S.P., Singh, P., Sharma, A. (2024). A Novel Similarity-Based Method for Link Prediction in Complex Networks. In: Choi, B.J., Singh, D., Tiwary, U.S., Chung, WY. (eds) Intelligent Human Computer Interaction. IHCI 2023. Lecture Notes in Computer Science, vol 14532. Springer, Cham. <a href="https://doi.org/10.1007/978-3-031-53830-8_32">https://doi.org/10.1007/978-3-031-53830-8_32</a>	40%

Please enter your personal details below, and return this form via email to s.p.tripathi@ljmu.ac.uk from the email account above, as confirmation of your agreement. Many thanks.



Rahul Kumar Yadav

## Statement of Contribution

Abhay Kumar Rai (akrai@curaj.ac.in)

Assistant Professor, Central University of Rajasthan, India.

You are being asked to complete this statement of confirmation to support Shashi Prakash Tripathi's application for PhD by Published Work, at Liverpool John Moores University. The paper below, co-authored with you, will contribute to his portfolio of peer reviewed articles, chapters and conference proceedings.

### Confirmation

I confirm that Shashi Prakash Tripathi is an author of the article listed in the table below, and that the percentage contribution stated for the article is an accurate assessment of his involvement.

Article	Contribution
Tripathi, S. P., Yadav, R. K., Rai, A. K., & Tewari, R. R. (2019). Hybrid approach for predicting and recommending links in social networks. Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017, 107–119. Springer Singapore.	60%
Yadav, R. K., Tripathi, S. P., Rai, A. K., & Tewari, R. R. (2020). Hybrid feature-based approach for recommending friends in social networking systems. International Journal of Web Based Communities, 16(1), 51–71.	60%
Tripathi, S. P., Yadav, R. K., & Rai, A. K. (2022a). Network embedding based link prediction in dynamic networks. Future Generation Computer Systems, 127, 409–420.	60%
Rai, A. K., Tripathi, S. P., & Yadav, R. K. (2023). A novel similarity-based parameterized method for link prediction. Chaos, Solitons & Fractals, 175, 114046.	60%

Rai, A.K., Yadav, R.K., Tripathi, S.P., Singh, P., Sharma, A. (2024). A Novel Similarity-Based Method for Link Prediction in Complex Networks. In: Choi, B.J., Singh, D., Tiwary, U.S., Chung, WY. (eds) Intelligent Human Computer Interaction. IHCI 2023. Lecture Notes in Computer Science, vol 14532. Springer, Cham. <a href="https://doi.org/10.1007/978-3-031-53830-8_32">https://doi.org/10.1007/978-3-031-53830-8_32</a>	40%
--	-----

Please enter your personal details below, and return this form via email to [s.p.tripathi@ljmu.ac.uk](mailto:s.p.tripathi@ljmu.ac.uk) from the email account above, as confirmation of your agreement.

Many thanks.



**Abhay Kumar Rai**

## Statement of Contribution

Tulika Narang ([n.tulika@gmail.com](mailto:n.tulika@gmail.com)), Associate Professor  
United University, Prayagraj, India

You are being asked to complete this statement of confirmation to support Shashi Prakash Tripathi's application for PhD by Published Work, at Liverpool John Moores University. The paper below, co-authored with you, will contribute to his portfolio of peer reviewed articles, chapters and conference proceedings.

### Confirmation

I confirm that Shashi Prakash Tripathi is an author of the article listed in the table below, and that the percentage contribution stated for the article is an accurate assessment of his involvement.

Article	Contribution
Tripathi, S. P., & Narang, T. (2016). Applying Model View View-Model and Layered Architecture for Mobile Applications. Journal of International Academy of Physical Sciences, 20(3), 215–221.	80%
Rai, H., Tripathi, S. P., & Narang, T. (2022). TransCRF—Hybrid Approach for Adverse Event Extraction. Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022, 1–10. Springer Nature Singapore Singapore.	75%
Shashi Prakash Tripathi, T. N. (2016). An Enhanced Approach of Preprocessing the Document using WordNet in Text Clustering. International Conference on Control Computing Communication and Materials (ICCCCM-2016), 5.	80%

Please enter your personal details below, and return this form via email to [s.p.tripathi@ljmu.ac.uk](mailto:s.p.tripathi@ljmu.ac.uk) from the email account above, as confirmation of your agreement.

Many thanks.

Tulika

Tulika Narang