Entity Recognition on Border Security

GEORGE SUCIU*, Beia Consult Int, Romania
MARI-ANAIS SACHIAN, Beia Consult Int, Romania
RAZVAN-ALEXANDRU BRATULESCU, Beia Consult Int, Romania
KOCI KEJSI, Beia Consult Int, Romania
GRIGOR PARANGONI, Beia Consult Int, Romania

Entity recognition, also known as named entity recognition (NER), is a fundamental task in natural language processing (NLP) that involves identifying and categorizing entities within text. These entities, such as names of people, organizations, locations, dates, and numerical values, provide structured information from unstructured text data. NER models, ranging from rule-based to machine learning-based approaches, decode linguistic patterns and contextual information to extract entities effectively. This article explores the roles of entities, tokens, and NER models in NLP, detailing their significance in various applications like information retrieval and border security. It delves into the practices of implementing NER in legal document analysis, travel history analysis, and document verification, showcasing its transformative impact in streamlining processes and enhancing security measures. Despite challenges such as ambiguity and data scarcity, ongoing research and emerging trends in multilingual NER and ethical considerations promise to drive innovation in the field. By addressing these challenges and embracing new developments, entity recognition is poised to continue advancing NLP capabilities and powering diverse real-world applications.

Additional Key Words and Phrases: Entity, Recognition, Machine Learning, Border Security, Travel, NER, Frameworks, RNNs, SVM

ACM Reference Format:
George Suciu, Mari-Anais Sachian, Razvan-Alexandru Bratulescu, Koci Kejsi, and Grigor Parangoni. 2024. Entity Recognition on Border Security. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Entity recognition, also known as named entity recognition (NER), is a natural language processing (NLP) task that involves identifying and classifying entities (objects, terms, or "named entities") within a text. These entities can be predefined categories such as names of people, organizations, locations, dates, numerical values, and other specific types of information. The goal of entity recognition is to extract structured information from unstructured text data. In the context of NLP, an "entity" typically refers to a real-world object that can be categorized into a specific class. Entity recognition is a crucial component in various applications, including information retrieval, question answering systems, chatbots, and more. It helps in automatically identifying and categorizing important pieces of information within a document, enabling machines to better understand and process human language. Named Entity Recognition (NER) serves as a crucial element in Natural Language Processing (NLP), un- raveling structured information from unstructured text. It revolves around identifying and categorizing entities—named objects like people, places, and dates. Tokens, the building blocks obtained through tokenization, play a pivotal role in this process. NER models, ranging from rule-based to machine learning-based, decode language intricacies, making sense of entities within text. This

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

Fig. 1. Entities Parameters

synergy forms the backbone for applications spanning information retrieval to intelligent question answering systems. This exploration delves into the roles of entities, tokens, and NER models in the dynamic landscape of NLP. Definition: Entities refer to specific, named objects or concepts that can be categorized into predefined classes. Examples include names of people, organizations, locations, dates, and other types of information. Role in NER: Named Entity Recognition (NER) is concerned with identifying and classifying entities within a given text. The output of an NER system includes the recognized entities and their corresponding categories. Tokens: Definition: In the context of NLP, a token is a unit of text that has been extracted from a larger body of text during the process of tokenization. Tokens can be as short as a single character or as long as an entire word. Role in NER: NER models typically operate on a tokenized version of the input text. During tokenization, the text is divided into individual tokens, which serve as the basic units of analysis for the NER system. Each token is then analyzed to determine if it corresponds to a named entity. NER Models: Definition: NER models are machine learning models or algorithms designed to perform Named Entity Recognition. These models are trained on labeled datasets that include examples of text with annotated entities, allowing the model to learn patterns and associations between words and entity categories [5]. Types of NER Models: Rule-based NER: These models use predefined rules and patterns to identify entities in text. Rules may be based on linguistic patterns, regular expressions, or other heuristics. Machine Learning-based NER: These models use statistical and machine learning techniques, such as Conditional Ran- dom Fields (CRF) or more recently, deep learning approaches like Recurrent Neural Networks (RNNs) and Transformers. These models learn to recognize entities from labeled training data[11]. The rest of the paper is organized in the following manner. Section 2 describes how NER models work, section 3 describes tNER Technologies and machine learning approaches. Section 4 contains information about the frameworks and pre-trained NER models. Section 5 and 6 discuss about ethical risks and concerns as well as challenges in NER model trainning. Further on sector 7 and 8 the usage of NER models in border security is analysed and finally section 9 concludes this research and explains future trends.

2 HOW NER MODELS WORK

Named Entity Recognition (NER) models work by leveraging linguistic patterns, contextual information, and, in the case of machine learning-based models, learned associations from labeled training data. The process can be broadly outlined as follows: Tokenization: The input text is divided into individual units called tokens. Tokens can be words, subwords, or characters, depending on the level of granularity desired. Feature Extraction: Features are extracted from the tokens to provide the model with relevant information. These features may include word embeddings, part-of-speech tags, syntactic information, and contextual representations. Rule-Based Approaches: In rule-based NER models, predefined rules and patterns are applied to identify entities based on linguistic and syntactic cues. For example, a rule might specify that two consecutive capitalized words are likely to be a person's name.

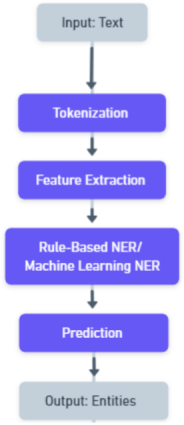

Fig. 2. Flowchart related to Named Entity Recognition (NER)

3 MACHINE LEARNING-BASED APPROACHES:

In machine learning-based NER models, the system is trained on labeled datasets that include examples of text with annotated entities. The training data consists of input text and corresponding entity annotations, indicating the start and end positions of entities and their types (e.g., person, organization)[9]. Features extracted from the tokenized text are used as input to the model. These features are fed into a machine learning algorithm, which learns to associate patterns in the features with the presence and types of entities. Common machine learning algorithms for NER include Conditional Random Fields (CRF), Support Vector Machines (SVM), and more recently, deep learning architectures such as Recurrent Neural Networks (RNNs) and Transformers. Prediction: Once trained, the NER model can predict the entities in new, unseen text. The model analyzes the features of each token and predicts whether it belongs to an entity and, if so, the type of entity[9]. Output: Entities: The final output includes identified entities and their types. Overall, the effectiveness of an NER model relies on the quality and diversity of the training data, the relevance of the features extracted, and the chosen algorithm's ability to generalize patterns learned during training to new, unseen text. The continuous evolution of machine learning techniques, especially with the advent of deep learning, has significantly enhanced the accuracy and robustness of NER models in recent years.

3.1 NER Models and Technologies

Entity recognition, with its pivotal role in various natural language processing (NLP) applications, relies on a variety of tools, techniques, and frameworks to achieve optimal performance. Here, we explore some of the essential components and considerations in the realm of NER models and technologies.

4 POPULAR NER LIBRARIES AND FRAMEWORKS:

An excess of NER libraries and frameworks have emerged, catering to diverse needs and preferences within the NLP community. Among the most widely used are:spaCy: Renowned for its ease of use and high performance, spaCy offers robust support for NER tasks. It provides pre-trained models and customizable pipelines, making it a popular choice for both beginners and seasoned practitioners.NLTK (Natural Language Toolkit): As one of the oldest and most comprehensive NLP libraries, NLTK offers a suite of tools for entity recognition, including classifiers and chunkers.

While its capabilities may require more manual configuration compared to newer libraries, NLTK remains a valuable resource for NLP enthusiasts. Stanford NER: Developed by the Stanford NLP Group, this library offers state-of-the-art NER models trained on large-scale datasets. Known for its accuracy and reliability, Stanford NER is frequently employed in research and industry applications[12].

4.1 Pre-trained NER Models:

Pre-trained NER models provide a convenient starting point for NLP projects, offering pre-trained weights and parameters that can be fine-tuned or utilized directly for entity recognition tasks[6]. These models are trained on extensive corpora and exhibit varying degrees of performance across different domains and languages. Notable examples include: BERT (Bidirectional Encoder Representations from Transformers): Leveraging the power of transformer architectures, BERT has revolutionized NLP tasks, including NER. Pre-trained BERT models, such as BERT-Base and BERT-Large, offer strong performance on a wide range of NLP benchmarks [6]. GPT (Generative Pre-trained Transformer): While primarily known for its text generation capabilities, GPT models can also be adapted for NER tasks. By fine-tuning pre-trained GPT models on labeled NER datasets, researchers have achieved competitive results in entity recognition [13].

4.2 Custom NER Model Development:

While pre-trained models provide a convenient solution for many NLP tasks, custom NER model development remains essential for domain-specific applications and specialized requirements [11]. Custom models can be tailored to specific datasets, domains, or languages, enabling finer control over model architecture and training objectives. Key considerations in custom NER model development include: Data Annotation: Annotated datasets are crucial for training custom NER models, as they provide labeled examples of entities within text data. Manual annotation or semi-automated labeling techniques may be employed to create high-quality training data. Model Architecture: Depending on the complexity of the task and available computational resources, various architectures can be explored, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models. Training Strategies: Training a custom NER model involves optimizing hyperparameters, selecting appropriate loss functions, and managing computational resources effectively. Techniques such as transfer learning and domain adaptation may also be employed to improve model performance[11].

5 ETHICAL RISKS OF USING NER SYSTEMS IN BORDER SECURITY

The deployment of Named Entity Recognition (NER) systems in border security applications introduces a range of ethical concerns that must be carefully examined. While these systems offer the potential to enhance security and streamline processes, their use also poses significant risks, particularly in terms of privacy violations and the potential for misuse by state actors.

5.1 Privacy Violations

NER systems, by design, are trained to identify and extract personal data from text. In the context of border security, this capability can be used to analyze and monitor vast amounts of personal information from various sources, such as social media, emails, travel documents, and communications[15].This raises serious privacy concerns:

Surveillance Overreach: The ability to continuously monitor and analyze personal communications can lead to an intrusive level of surveillance. This could infringe on individuals' rights to privacy and free expression, creating a

chilling effect where people feel they must censor themselves to avoid scrutiny. Data Misuse: There is a risk that the data collected and processed by NER systems could be misused for purposes beyond their original intent. For instance, information extracted from individuals' communications might be used to profile and target specific groups based on their ethnicity, religion, or political beliefs [15]. Data Security: The storage and handling of sensitive personal data pose security risks. If the data collected by NER systems is not adequately protected, it could be susceptible to breaches and unauthorized access, potentially leading to identity theft and other forms of exploitation.

5.2 Ethical Concerns of State Actors

The use of NER systems by state actors in border security brings additional ethical dilemmas: Discrimination and Bias: NER systems are trained on large datasets that may contain inherent biases. When used in border security, these biases can lead to discriminatory practices, such as unfairly targeting individuals based on their nationality, race, or ethnicity. This can exacerbate existing social inequalities and reinforce systemic discrimination. Lack of Transparency and Accountability: State actors may not always be transparent about how NER systems are used, what data is collected, and how it is processed. This lack of transparency can prevent proper oversight and accountability, making it difficult to address potential abuses and ensure that the systems are used ethically and lawfully [3]. Human Rights Violations: The intrusive nature of NER systems, combined with their potential for misuse, can lead to human rights violations. For instance, activists, journalists, and dissidents could be targeted and monitored, suppressing dissent and freedom of expression. Additionally, the information extracted could be used to detain or deport individuals unjustly, based on automated decisions without adequate human oversight [3].

6 CHALLENGES IN NER MODEL TRAINING:

Despite the advancements in NER technologies, several challenges persist in the training and deployment of NER models: Data Scarcity: Annotated NER datasets are often limited in size and diversity, especially for specialized domains or languages. This scarcity of training data can hinder the development of accurate and robust NER models. Ambiguity and Variability: Entity recognition tasks may involve ambiguous or variable entities, such as names with multiple interpretations or entities with context-dependent meanings [4]. NER models must effectively handle such ambiguity to ensure accurate identification and classification. Evaluation Metrics: Assessing the performance of NER models requires appropriate evaluation metrics that account for factors such as entity boundaries, entity types, and overlapping entities. Common metrics include precision, recall, and F1-score, but selecting the most suitable metric depends on the specific task and requirements. Domain Adaptation: NER models trained on generic datasets may struggle to perform effectively in domain-specific contexts due to differences in terminology, style, and entity characteristics. Domain adaptation techniques, such as fine-tuning or multi-task learning, can help mitigate these challenges [4]. Addressing these challenges requires ongoing research and innovation in NER model development, data annotation, and evaluation methodologies. By addressing these considerations, NER technologies can continue to evolve and meet the growing demands of real-world applications in NLP and beyond.

7 ENTITY RECOGNITION PRACTICES

In the dynamic landscape of border security, the ability to efficiently process and analyze legal documents submitted by travelers is paramount. Named Entity Recognition (NER) emerges as a transformative tool, offering border security agencies a sophisticated solution to streamline document analysis procedures and bolster security measures at border checkpoints.

7.1 Border Security Scenario:

An international airport or a busy land border crossing, where travelers from diverse backgrounds converge, each presenting a multitude of legal documents, including passports, visas, and entry permits. Amidst the flurry of activity, border security personnel face the daunting task of scrutinizing these documents swiftly and accurately to verify travelers' identities, assess their eligibility for entry, and ensure compliance with immigration regulations.

7.2 Implementation of NER in Legal Document Analysis:

Passport Information Extraction: In implementing Named Entity Recognition (NER) for passport information extraction from scanned or photographed passport images, developers follow a structured approach [18]. They begin by preprocessing the images using Optical Character Recognition (OCR) to convert them into machine-readable text. Following this, a labeled dataset of passport images is collected and annotated for data preparation. This dataset serves as the foundation for training a suitable NER model architecture, fine-tuned to accurately identify and extract critical information such as the traveler's full name, date of birth, nationality, passport number, and expiration date. Once trained, the NER model is seamlessly integrated into the passport processing system, where it efficiently processes incoming passport images, extracting the required information. To ensure accuracy and completeness, error handling mechanisms and validation checks are implemented. Furthermore, the performance of the NER model is continually optimized through iterative testing and evaluation, enabling border security agencies to expedite document review procedures and enhance the efficiency of passport processing at border crossings. This automated extraction process not only expedites document review procedures but also allows border control officers to focus their attention on assessing potential security risks and ensuring the integrity of travel documents [13]. Visa Details Recognition: To implement Visa Details Recognition using Named Entity Recognition (NER) from a developer's standpoint, the process involves several stages. Initially, the scanned or digital copy of the visa document is received, often in formats like PDF or image files. Optical Character Recognition (OCR) technology is then utilized to convert the document into machine-readable text, preserving its structure and formatting. The text is tokenized into individual words or phrases to facilitate analysis. Developers parse through the tokenized text to pinpoint relevant sections or fields containing visa details, such as the issuing country, visa type, validity period, and any specific conditions or restrictions. A pre-trained NER model is then applied or custom NER models are developed specifically for recognizing visa details. This NER model is employed to identify and extract named entities corresponding to visa details, such as country names, date ranges, visa types, and specific conditions or restrictions. Validation of the extracted entities follows, ensuring accuracy and completeness. Checks may include verifying country names against a list of valid countries and validating date ranges for consistency and correctness. Cross- referencing the extracted visa details with external databases or reference sources further validates the authenticity and validity of the visa. Extracted visa details are then structured into a standardized format or data model, such as key-value pairs or structured objects, and organized into meaningful categories like issuing country, visa type, validity period, and specific conditions. Additionally, error handling mechanisms are implemented to address cases where visa details cannot be reliably extracted due to document quality issues, OCR errors, or ambiguous text. Exception handling routines are developed to address edge cases and handle unexpected scenarios, ensuring robustness and reliability in the Visa Details Recognition process. Entry Permit Parsing: First, the system must be designed to intake entry permits from travelers, either physically or electronically. This involves creating a mechanism for securely receiving and storing the documents. Next, the text of the entry permits undergoes preprocessing to standardize formatting and enhance readability. Text normalization techniques are applied to handle variations in

layout, language, and structure. An appropriate NER model or library is selected for entity extraction. This could involve choosing pre-trained models like those provided by spaCy or training custom models using labeled datasets specific to entry permit parsing. The selected NER model parses through the text of the entry permits, extracting relevant entities such as the purpose of travel, duration of stay, and permitted entry dates. Entity categories and labels are defined accordingly. Extracted entities are converted into structured data representations, such as dictionaries or JSON objects, for further processing [18]. Additional features or metadata may be included to enrich the extracted information. Validation checks are implemented to ensure the accuracy and completeness of the extracted entities. Verification against reference sources, such as immigration databases, is performed to confirm the legitimacy of the entry permits. The parsed entry permit information is integrated with decision support systems used by border control officers. User-friendly interfaces or dashboards are provided for officers to access and review the extracted information. Alerting mechanisms are implemented to flag discrepancies or anomalies detected during parsing. Error handling ensures that errors are handled gracefully, with informative messages and logging for troubleshooting.

8 TRAVEL HISTORY ANALYSIS:

In the workflow of Travel History Analysis, the process unfolds systematically, each step meticulously designed to extract, analyze, and interpret data from travelers' documents. At the outset, documents, such as passports or travel records, are received and processed using Optical Character Recognition (OCR) software. This initial step converts scanned or photographed documents into machine-readable text, laying the foundation for subsequent analysis. The extracted information undergoes Named Entity Recognition (NER), a pivotal process where specific entities like locations, dates, and visa details are identified and extracted. This step is crucial as it enables the system to discern pertinent information from the text, shaping the subsequent analysis. Following entity extraction, the data undergoes normalization to ensure consistency and uniformity in formats, facilitating seamless aggregation and analysis. This step is essential for aligning extracted entities to standardized formats, enhancing the accuracy and reliability of subsequent analysis. With normalized data in hand, the system aggregates the extracted entities to generate comprehensive travel history profiles for each traveler. This step provides insights into travel patterns, destinations, and compliance with immigration regulations, forming the basis for further analysis. Pattern recognition and anomaly detection algorithms come into play at this stage, scrutinizing travel history profiles to identify patterns, trends, and irregularities. By analyzing travel behavior, these algorithms can flag suspicious activities or potential security risks, aiding in decision-making. Decision support systems are integral to the workflow, providing border security personnel with actionable insights and recommendations based on the analyzed travel history data. These systems empower officers to make informed decisions regarding entry permissions, enhancing border security measures. Reports and visualizations summarize the findings of travel history analysis, presenting trends, anomalies, and insights in a clear and comprehensible format. This step facilitates communication of findings to stakeholders and aids in understanding the implications of travel patterns. Integration with existing border security systems ensures seamless data exchange and interoperability, enabling travel history analysis to complement and enhance existing security infrastructure [2]. Document Verification: NER technology plays a crucial role in verifying the authenticity of submitted documents by cross-referencing extracted information with databases of valid passports, visa records, and watchlists of individuals with restricted entry status. This automated verification process enhances border security measures by identifying travelers with suspicious or irregular documentation and mitigating potential security threats. NER algorithms are instrumental in identifying contraband, such as alcohol and drugs, at border crossings by analyzing textual descriptions provided in customs declarations, cargo manifests, and traveler documentation [7]. For instance, when a traveler declares items in their

possession, NER algorithms parse through the declaration to identify mentions of substances like alcohol or drugs. Additionally, NER systems can analyze cargo manifests to detect shipments containing suspicious descriptions or quantities of controlled substances. By automatically flagging such mentions or descriptions, NER technology assists border security agencies in targeting high-risk individuals and shipments for further inspection, thereby enhancing the detection of contraband at border crossings. However, NER primarily relies on the information provided in textual descriptions, such as declarations, documents, and cargo manifests, to identify entities. Therefore, if contraband items like alcohol, drugs, guns, or dangerous weapons are not mentioned in these documents, NER may not be able to detect them directly. In scenarios where contraband items are not explicitly mentioned in declarations or documents, Named Entity Recognition (NER) still contributes to the detection of illicit activities by flagging inconsistencies or anomalies in the provided information. For instance: At a border crossing, a traveler submits a customs declaration form stating that they are carrying personal belongings and gifts for friends. The declaration appears routine, with no explicit mention of contraband items such as drugs or firearms [2]. However, upon closer examination, NER algorithms detect discrepancies in the declaration. NER algorithms analyze the textual descriptions provided in the declaration, parsing through the language to identify entities and extract key information. While the declaration may not explicitly mention contraband items, NER technology identifies linguistic cues or patterns that raise suspicion. For example, vague descriptions, ambiguous statements, or unusual phrasing may indicate potential discrepancies. NER-enabled systems compare the information provided in the declaration against established norms or expected patterns. Inconsistencies between the declared contents and typical travel behavior may trigger alerts. Leveraging contextual information, NER algorithms discern potential risks. For instance, if a traveler's declaration indicates a short trip but includes excessive quantities of certain items, it may warrant further scrutiny. Based on the detected anomalies, NER technology generates alerts or flags the declaration for secondary inspection by border security officers. Border security officers conduct additional checks, which may include physical inspection of belongings, questioning the traveler, and examining travel history records. If suspicious items are discovered during the secondary inspection, border security agencies initiate further investigation, potentially leading to the interception of contraband and the apprehension of individuals involved in illicit activities. Through the proactive use of NER technology, border security agencies uncover potential security risks that may have otherwise gone unnoticed. By flagging inconsistencies in declarations and documents, NER contributes to the prevention and detection of contraband smuggling, thereby safeguarding national borders and protecting public safety.

9 CHALLENGES AND FUTURE DIRECTIONS

9.1 Addressing Ambiguity in Entity Recognition:

One of the key challenges in Named Entity Recognition (NER) is dealing with ambiguity in text. Ambiguity arises when a single word or phrase can have multiple meanings or interpretations, leading to confusion for NER models. To address this challenge, researchers are exploring context-aware approaches that consider the surrounding context to disambiguate entities. Techniques such as deep learning architectures, attention mechanisms, and contextual embeddings are being investigated to improve the accuracy of entity recognition in ambiguous contexts. Additionally, the integration of domain-specific knowledge and ontologies can provide additional context cues to resolve ambiguity more effectively[16].

9.2 Multilingual Entity Recognition:

As the need for multilingual NER continues to grow, researchers are focusing on developing models capable of recognizing entities in multiple languages[5]. Multilingual NER poses unique challenges due to language-specific variations in syntax, morphology, and named entity conventions [14]. Recent advancements in cross-lingual transfer learning and multilingual pre-trained models have shown promise in addressing these challenges by leveraging shared representations across languages . Additionally, domain adaptation techniques are being explored to fine-tune NER models for specific languages or domains where labeled data may be limited [5].

9.3 Ethical Considerations in NER:

With the increasing use of NER in various applications, including surveillance, content moderation, and personalization, ethical considerations have become paramount. Concerns related to privacy, bias, fairness, and consent arise when processing sensitive information through NER systems. To mitigate these risks, researchers and practitioners are advocating for transparent and responsible AI practices. This includes implementing privacy-preserving techniques such as differential privacy, designing fair and unbiased NER models, and ensuring informed consent and user control over data usage. Ethical guidelines and regulations are also being developed to govern the responsible deployment of NER technology.

9.4 Emerging Trends in Entity Recognition:

Looking ahead, several emerging trends are shaping the future of entity recognition. One trend is the integration of multimodal information, combining text with other modalities such as images, audio, and video, to enhance entity recognition in diverse data sources [17]. Another trend is the adoption of self-supervised learning techniques, where NER models learn from unlabeled data to improve performance in low-resource scenarios. Additionally, advancements in zero-shot and few-shot learning are enabling NER models to generalize to new entity types and domains with minimal labeled data. Collaborative efforts between academia, industry, and policymakers are driving innovation in NER, paving the way for more accurate, robust, and ethically sound entity recognition systems in the future[8].

ACKNOWLEDGMENTS

The work presented in this paper has received funding from the following projects: 1.The European Research Executive Agency (REA) programme,HORIZON-CL3-2021-BM-01 FLEXI-CROSS project under grant agreement No 101073879. 2. The HORIZON-EUSPA-2021-SPACE research and innovation programme CHRISS project under grant agreement No. 101082440 3.The European Union's Horizon 2020 Research and Innovation programme, VITAL-5G project under grant agreement No 101016567.

10 CONCLUSION

In conclusion, entity recognition stands as a cornerstone in the field of natural language processing (NLP), facilitating the extraction of structured information from unstructured text data. Through the identification and classification of entities such as names of people, organizations, locations, dates, and numerical values, NER enables machines to comprehend and process human language more effectively [10]. The landscape of NER is characterized by a diverse array of tools, techniques, and frameworks, ranging from popular libraries like spaCy and NLTK to cutting-edge pre-trained models such as BERT and GPT [1]. These technologies offer developers the flexibility to choose between pre-built solutions

and custom model development based on the specific requirements of their applications. However, challenges persist in the training and deployment of NER models, including data scarcity, ambiguity in entity recognition, and domain adaptation issues. Addressing these challenges requires ongoing research and innovation, with a focus on improving the quality and diversity of training data, enhancing model robustness, and developing effective evaluation methodologies. Looking ahead, emerging trends in entity recognition, such as multilingual NER and ethical considerations, are poised to shape the future of the field. Multilingual entity recognition promises to broaden the applicability of NER across diverse languages and cultures, while ethical considerations underscore the importance of responsible AI development and deployment. In summary, entity recognition continues to play a vital role in advancing NLP capabilities and powering a wide range of applications, from information retrieval to border security. By addressing current challenges and embracing emerging trends, the field of entity recognition is poised for continued growth and innovation in the years to come.

REFERENCES

[1] Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. German BERT model for legal named entity recognition. *arXiv preprint arXiv:2303.05388* (2023).

[2] Jinlan Fu, Pengfei Liu, and Qi Zhang. 2020. Rethinking generalization of neural models: A named entity recognition case study. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7732–7739.

[3] Marco Humbel, Julianne Nyhan, Andreas Vlachidis, Kim Sloan, and Alexandra Ortolja-Baird. 2021. Named-entity recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future. *Journal of Documentation* 77, 6 (2021), 1223–1247.

[4] Andrea Iovine, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2022. CycleNER: an unsupervised training approach for named entity recognition. In *Proceedings of the ACM Web Conference 2022*. 2916–2924.

[5] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering* 34, 1 (2020), 50–70.

[6] Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. NER-BERT: a pre-trained model for low-resource entity tagging. *arXiv preprint arXiv:2112.00405* (2021).

[7] Ioannis Mademlis, Marina Mancuso, Caterina Paternoster, Spyridon Evangelatos, Emma Finlay, Joshua Hughes, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Georgios Stavropoulos, Konstantinos Votis, et al. 2023. The invisible arms race: digital trends in illicit goods trafficking and AI-enabled responses. *Authorea Preprints* (2023).

[8] Natawut Monaikul, Giuseppe Castellucci, Simone Filice, and Oleg Rokhlenko. 2021. Continual learning for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13570–13577.

[9] Salman Naseer, Muhammad Mudasar Ghafoor, Sohaib bin Khalid Alvi, Anam Kiran, Shafique Ur Rahmand, Ghulam Murtazae, and Ghulam Murtaza. 2021. Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance. *Pakistan Journal of Multidisciplinary Research* 2, 2 (2021), 293–308.

[10] Arya Roy. 2021. Recent trends in named entity recognition (ner). *arXiv preprint arXiv:2101.11420* (2021).

[11] Hemlata Shelar, Gagandeep Kaur, Neha Heda, and Poorva Agrawal. 2020. Named entity recognition approaches and their comparison for custom ner model. *Science & Technology Libraries* 39, 3 (2020), 324–337.

[12] Sowmya Vajjala and Ramya Balasubramaniam. 2022. What do we Really Know about State of the Art NER? *arXiv preprint arXiv:2205.00034* (2022).

[13] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428* (2023).

[14] Zhili Wang, Yufan Wu, Pengbin Lei, and Cheng Peng. 2020. Named entity recognition method of brazilian legal text based on pre-training model. In *Journal of Physics: Conference Series*, Vol. 1550. IOP Publishing, 032149.

[15] Chanoksuda Wongvises, Assadarat Khurat, and Thanapon Noraset. 2022. Thai Privacy Notice Analysis Based On Named-Entity Recognition Technique. In *2022 26th International Computer Science and Engineering Conference (ICSEC)*. IEEE, 257–262.

[16] Apurwa Yadav, Aarshil Patel, and Manan Shah. 2021. A comprehensive review on resolving ambiguities in natural language processing. *AI Open* 2 (2021), 85–92.

[17] Jun Yang, Taihua Zhang, Chieh-Yuan Tsai, Yao Lu, and Liguo Yao. 2024. Evolution and emerging trends of named entity recognition: Bibliometric analysis from 2000 to 2023. *Heliyon* (2024).

[18] Ashwini V Zadgaonkar and Avinash J Agrawal. 2021. An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical & Computer Engineering (2088-8708)* 11, 6 (2021).