International Journal of Computer and Organization Trends
ISSN: 2249 – 2593 / https://doi.org/10.14445/22492593/IJCOT-V14I3P301
Volume 14 Issue 3, 1-7, Sep-Dec 2024

Review Article

The Role of Named Entity Recognition (NER): Survey

Girma Yohannis Bade[1], Olga Kolesnikova[2], Jose Luis Oropeza[3]

[1,2,3]*Centro de Investigación en Computación(CIC), Instituto Politécnico Nacional (IPN), Miguel Othon de Mendizabal, Ciudad de México. México.*

[1]*Corresponding Author : girme2005@gmail.com*

Received: 16 June 2024 Revised: 23 July 2024 Accepted: 14 August2024 Published: 31 August 2024

Abstract - *Named Entity Recognition (NER) is an Information Extraction (IE) building block. Though the information extraction process has been automated using various techniques to find and extract relevant information from unstructured documents, the discovery of targeted knowledge still poses many research difficulties because of Web data's variability and lack of structure. NER, a subtask of IE, came to exist to smooth such difficulty. It deals with finding the proper names (named entities), such as a person's name, country, location, organization, dates, and event in a document. It categorises them as predetermined labels, an initial step in IE tasks. This survey paper presents the roles and importance of NER to IE from the perspective of different algorithms and application area domains. Additionally, it summarizes how researchers implemented NER in particular application areas like finance, medicine, defense, business, food science, archeology, etc. It also outlines the three NER sequence labeling algorithms types: feature-based, neural network-based, and rule-based. Finally, the state-of-the-art and evaluation metrics of NER were presented.*

Keywords - *NER, Information Extraction (IE), Sequence labeling algorithms, Application area.*

1. Introduction

Information extraction involves finding and extracting the input's subsequence corresponding to the information of interest. Though the extraction process has been automated using various techniques, automated discovery of targeted knowledge material still poses many research difficulties because of the variability and lack of structure in Web data (Tang et al., 2007).

Named Entity Recognition (NER), a subtask of IE, came to exist to smooth such difficulty. Finding the proper names or named entities, such as the name of the person, country, location, organization, dates and event in a document, is the initial step in the majority of IE tasks, and this process is known as NER.

It is a useful task of Information Extraction (IE), which deals with the issue of locating and categorizing present concepts in a particular domain. To label them into predetermined classes (labels) that reflect concepts of interest in a specific domain it seeks to extract words or phrases from the text (Popovski et al., 2020).

The Conference on Computational Natural Language Learning (CoNNL) framework has thoroughly evaluated issues of NER and outlined the techniques used from rule-based to machine learning(Khalid et al., 2008).

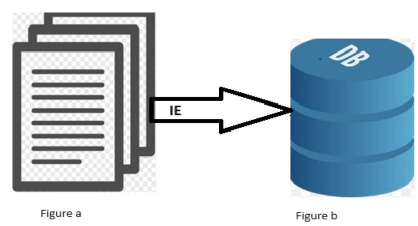

Fig. 1 a and b show how IE converts unstructured documents into structured relational forms using NER

This study aims to investigate the role of NER. This paper intends to explain how NER helps convert unstructured documents into structured forms, highlighting its significance in different application areas.

It also overviews the Natural Language Processing (NLP) tasks requiring sequence labeling and their respective algorithms.

2. Demand Area of NER

The identified entity serves as the text's primary information carrier and is employed to convey the text's core idea.

Implementing several natural language processing techniques, such as information extraction, retrieval, machine translation, and question-and-answer systems, depends on accurately detecting their contents(Bade, 2021).

2.1. Business

Named entities in business documents include significant data that can greatly impact the firm. For instance, winning information from a tender is extremely important to the business. For businesses, gathering much tender text, sifting it, and finding its key information is crucial to business intelligence(Guan et al., 2019).

Financial institutions subject to strict regulation, such as banks, insurance companies, and investment firms, closely monitor the financial activity of their clients. These institutions have established a Know Your Customer (KYC) process. Today's KYC process is time-consuming; gathering information through declarative questionnaires and watch list lookups is time-consuming. This has prompted financial institutions to gather as much information as they can about the financial activity of their clients. Automated solutions can assist financial institutions in this regard by enabling the continual updating of data that may affect the evaluation of risk profiles. The knowledge Base (KB) of the clients' financial relationships, which gains population through automated information extraction, can be used to accomplish this purpose (Jabbari et al., 2020).

2.2. Healthcare

The number of scientific papers in the literature that contain an understanding of the biomedical, health, and clinical sciences is consistently increasing. Since the findings are not currently automatically archived, most of this data is still hidden in textual details that are difficult to access for subsequent use or analysis. For this reason, information extraction from such publications uses text mining and Natural Language Processing (NLP) techniques. For example, to determine the interactions between proteins (medications or genes) and diseases, NER) and Relationship Detection (RD) are used (Perera et al., 2020). This data can be included in networks to condense extensive details on a specific biological or clinical issue, making it simple to handle the data and do further analysis.

2.3. Defense

Processing and interpreting massive amounts of online data has become more and more challenging due to the recent decade's information explosion. To this end, creating a new technology like information extraction is critical, and NER is an invaluable tool for its creation. Research and development on information extraction systems powered by the Advanced Research Projects Agency (ARPA) have focused on evaluating systems performing various application tasks in the top-secret domain (Okurowski, 1993).

2.4. Archaeology and Geo-Spatial

Archaeological literature is increasing quickly. These data were previously only available through metadata search. Domain-specific elements like places, eras, and artifacts are crucial in archaeological IR. In this regard, the study (Brandsen et al., 2022) developed a text retrieval engine for a sizable archive of archaeological texts. It annotated the entire collection with archaeological named entities, and a Named Entity Recognition (NER) model was created. Gazetteers are extensive listings of geographic entity names that are typically enhanced with additional details, such as the category (such as town, river, dam, etc.), size, and location (i.e. about some relative or absolute coordinate system, such as longitude and latitude)(Leidner et al., 2003).

2.5. Food science

The large amount of food-related information supplied as heterogeneous textual data makes it beneficial to extract such information using computer-based methods automatically. Utilizing NER techniques, frequently employed in computer science for information extraction, is one way to achieve this. The field of food science continues to be under-resourced despite the abundance and sophistication of NER approaches in the biomedical realm. The publication (Popovski et al., 2020) presents many food-related NER techniques that can automatically extract food-related information from text.

2.6. Significances and Challenges of NER for IE

NER plays a critical role in Information Extraction (IE) by identifying events and establishing relationships between key entities or "players." Moreover, it supports a range of language processing tasks, such as sentiment analysis and question-answering systems, by mapping unstructured text to structured data within knowledge sources.

However, NER faces significant challenges, particularly due to segmentation ambiguity (G. Bade et al.,2024), which requires determining what qualifies as an entity and delineating its boundaries.
Another complexity arises from type ambiguity, where the same entity can have multiple interpretations based on context. For example:

1) Washington[PER]: Refers to an individual, "Washington was born into slavery on James Burroughs' land."
2) Washington[ORG]: Denotes an organization, "In the four-game series, Washington had a two-game lead."
3) Washington[LOC]: Identifies a location, "Blair arrived in Washington for what might be his final state visit" (Pollock et al., 2010).

3. Sequence Labeling and Relation Extraction

This section overviews the extensively researched basic sequence labeling tasks: Named Entity Recognition (NER), text chunking, and Part-of-Speech (POS) tagging. In addition, relation and event extraction, which are tasks closely similar to NER, are outlined. Sequence labeling is one of the core NLP techniques used to recognize and name the words or phrases that make up a sentence. It is an essential preprocessing step for many NLP applications, such as sentiment analysis, retrieving information, and machine translation (Yohannis Bade, 2018).

3.1. Sequence Labeling

3.1.1. Named Entity Recognition

Named entity recognition is a well-known classical sequence labeling task. It recognizes named entities from texts (sentences) and classifies them into pre-defined classes. These classes typically include three main categories: entity, time, and numeric, with seven sub-categories: person-name, organization, location, time, date, currency, and percentage(He et al., n.d.).The main categories of sub-category pairs are entity(person-name, organization, location), time expression(time, date), and numeric expression (currency/money, percentage). The BIOES system is often the most extensively used tagging scheme in NER; a word labeled "B" (Begin), "I" (Inside), or "E" (End) indicates that it is the first, middle, or last word of a named entity phrase, respectively. The word labeled "0-" (Outside) denotes that it is the outside word that represents an entity, and "S-" (Single) implies that it is the only word that belongs to any specified entity phrase. A fundamental component of many high-level applications, including search engines, question and answer systems, recommendation systems, and translation systems, utilize NER as a crucial task in natural language processing.

3.1.2. Part-of-speech Tagging (POS)

POS is a typical sequence labeling task that tries to correctly identify each lexical item (also known as a word), such as a noun (NN), verb (VB), and adjective (JJ). It is well-accepted by both academics and industry(He et al., n.d.). POS tags are automatically assigned to words in a phrase (or grammatical tags). An essential natural language processing tool for machine translation, word sense disambiguation, question-answering parsing, and other applications is point-of-service (POS) tagging. Because manually assigning POS tags to words is a time-consuming, costly, and tiresome activity, there is growing interest in automating the tagging process(Chiche & Yitagesu, 2022).

3.1.3. Text Chunking

The text chunking task separates the text into syntactically related, non-overlapping phrase types, such as noun and verb phrases. The task can be viewed as simply a sequence labeling problem where words in sentences are given specific labels(He et al., n.d.),(Akhundov et al., 2018).

Under the general heading of Natural Language Understanding (NLU), tasks like semantic slot filling and shallow parsing are resolved by marking significant text segments.

This type of work is typically approached as a sequence labeling issue, in which an IOB-based (Inside-Outside-Beginning) label is assigned to each word in a sentence. In the line "But it could be much worse," for instance, we classify "could" as B-VP, "be" as I-VP, and "it" as B-NP, but "But" is a member of an artificial class O. This marking suggests that the sentence fragment "could be" is a Verb Phrase (VP) in text chunking task(Zhai et al., 2017).

3.2. Relation and Event Extraction

3.2.1. Relation Extraction

Relation extraction involves identifying and categorizing semantic relationships between text items. These relationships, such as child-of, employment, part-whole, and geographic, are frequently binary. Relation extraction and relational database population are closely related processes. For example: "Barack Obama was the president of the USA"(Pollock et al., 2010).

It has numerous real-world advantages in information discovery and natural language processing. Knowledge graphs can be produced using relation extraction in machine learning for precise information presentation, service suggestions, and query responses (Of & In, 2019).

3.3. Event Extraction

Event extraction involves inferring precise information about happenings mentioned in texts. Event extraction is a technique that can be used on various written texts, including blogs, manuscripts, and (online) news bulletins. For instance, we extract temporal expression from texts, consisting of timings like 3:00 PM and days of the week, to determine when events occurred(Pollock et al., 2010). IE systems may benefit in many ways from event extraction over unstructured data. Events, for instance, could improve the effectiveness of tailored news systems since news items could be chosen more carefully based on user preferences and detected subjects (or events). Events can also be helpful in risk analysis software, monitoring systems, and decision-support tools (Hogenboom et al., n.d.).

4. State of the Art

This section presents various techniques and tools applied to the NER function for information extraction. We begin by giving a summary of the works in Table 1. This summary comprises various named-entity recognition tasks, their respective techniques (method), the data sources, and the performance indicators.

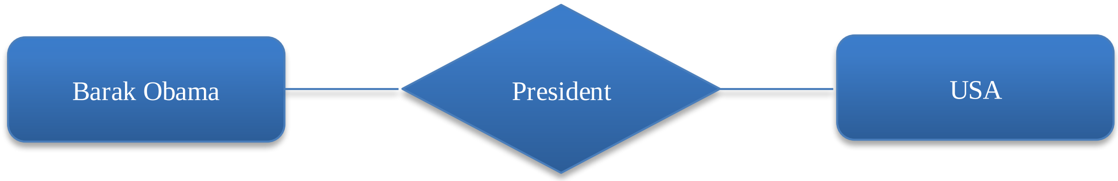

Fig. 2 shows how the relation extraction extracts the relation for the above example. It creates the binary meaningful relation for sentences.

Table 1. The overview of different NER tasks with different methods and results

Ref	Descriptions	Method	Data source	F1-measure
(Jabbari et al., 2020)	Financial institutions must look for automated information extraction tools to track their customers' financial activity.	SpaCy	official reports, news articles, and social media	0.87
(Perera et al., 2020)	Determines how proteins and medications or genes and diseases interact.	Rule-based, Dictionary-based, Machine Learning based, and Hybrid models	Publications from the web of science	---
(Brandsen et al., 2022)	Archaeology IR, domain-specific objects like places, eras, and artifacts	BERT	Research plans, appendices, maps and data descriptions	0.735
(Guan et al., 2019)	Business documents include a wealth of knowledge that can be valuable to the company.	Combined BiLSTM and CRF, combined n-gram features and character features	bidding documents	BiLSTM= 0.87 BiLSTM+CRF = 0.863
(Zhang et al., 2021)	The recognition of Chinese-named entities in the field of apple diseases and pests, including different entity categories, entities with aliases or acronyms, and the challenge of recognizing rare entities	character-based BiLSTM-CRF	apple disease and pest control books	0.921
(Zhong & Chen, 2021)	Using a straightforward pipelined technique, establish the new state-of-the-art using common standards for entity and relation extraction.	LSTM, BERT-base	newswire and online forums.	0.877

5. Methodology

5.1. Source for Survey

The methodology used in this paper involves surveying various literature sources related to NER and IE. The authors interpret and present the gathered information concisely and conveniently. The data for this study is collected through a systematic survey of literature sources, including research papers, conference proceedings, and other relevant publications.

5.2. Algorithms Used to Develop NER

The word-by-word sequence labeling task that makes up the standard algorithm for named entity recognition allows the supplied tags to capture the border and the type simultaneously. To identify specific types of named entities in a text, tokens in the text must tagged by sequence labeling algorithms such as the Maximum-entropy Markov Model (MEMM), Conditional Random Field (CRF), bidirectional long short-term memory(bi-LSTM). This algorithm can be broadly categorized as feature-based, neural network-based, and rule-based(Pollock et al., 2010; Yigezu et al., 2023).

5.3. Feature-based Algorithm

This method identifies *wi* neighboring words, embeddings for *wi* embeddings for neighboring words part of *wi*, part of speech of neighboring words base-phrase syntactic chunk label of *wi*, and neighboring words the presence of *wi* in a gazetteer.

A gazetteer is a list of place names, often providing numerous entries for locations with detailed geographical and political information.

CRF is a well-known illustration of feature-based sequence labeling. It trains using L2 regularization and a variety of variables, including word frequency, part-of-speech information, local context, chunk information, and suffix and prefix characters (Sikdar, 2017). The first approach is to extract features and train a MEMM or CRF sequence model.

5.4. Neural network-based Algorithms

Describing entity properties, domain knowledge, entity context, and linguistic features in feature-based models takes time and requires labor-intensive feature engineering (G.Bade et al.,2024). Therefore, using neural network models is suggested to reduce the requirement for feature engineering (Approach, 2019).

According to (He et al., n.d.), there are three axes: embedding module, context encoder module, and inference module, which are used in the deep learning approach to categorize word sequences. In the case of an embedding module, words are mapped into their distributed representations in the initial step.

The context encoder module captures contextual data, and the inference module makes predictions about labels and produces the best label sequence as a model's output. The standard neural network algorithm for NER is based on the Bi-LSTM. In this particular model, word and character embeddings are computed for input word *wi*.

W={w1,w2,w3,..,wn}. For instance, "Mark Wetney visits Mars".
W={ 'Mark', 'Wetney', 'visits Mars' }→wi.

These words are passed through a left-to-right LSTM and a right-to-left LSTM, whose outputs are concatenated to produce a single output layer at position *i*. Since named entity tagging follows a greedy approach and it does not allow us to impose the strong constraints, neighboring tokens have on each other, the decoding process is insufficient. Instead, a CRF layer is normally used on top of the Bi-LSTM output, and the Viterbi decoding algorithm is used to decode(Bade & Afaro, 2018; Pollock et al., 2010).

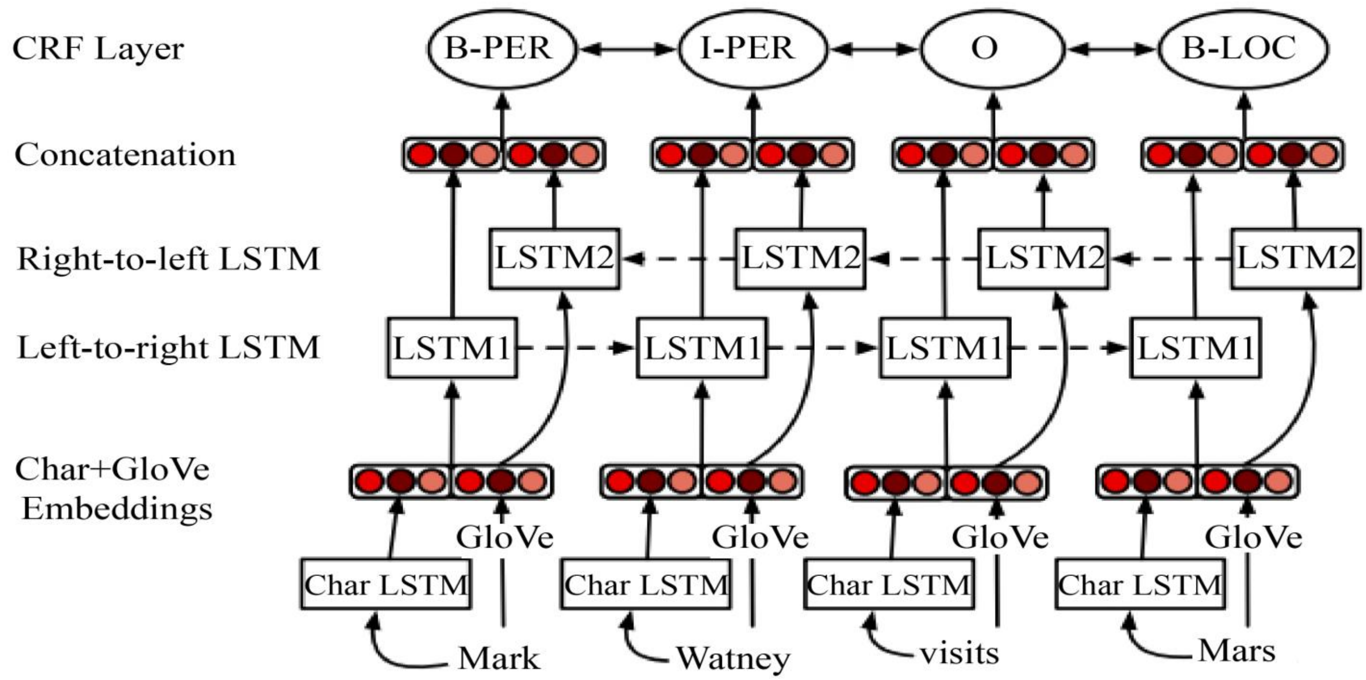

Fig. 3 Character embeddings and words together in a bi-LSTM sequence model*(Pollock et al., 2010)*.

5.5. Rule-based Sequence Labeling algorithm

While machine learning or neural sequence models are the norm in academic research, commercial approaches to NER are often based on pragmatic combinations of lists and rules with a smaller amount of supervised machine learning.

This algorithm specifies complex declarative constraints for tagging tasks in a formal query language that includes regular expressions, dictionaries, semantic constraints, NLP operators, and table structures, making the system an efficient extractor(Pollock et al., 2010). One of the common approaches is to make repeated rule-based passes over a text, allowing the results of one pass to influence the next. The stages typically involve using rules with extremely high precision but low recall.

6. NER Evaluations

Like other machine learning models (G.Y.Bade et al.,2018), NER is also measured with familiar evaluation metrics such as recall, precision, and F1 measures used to evaluate NER systems (Hkiri et al., n.d.).

6.1. Precision Measurement

Precision measurement is defined by the percentage of entities found by the system and which are correct. In formal terms, precision in the context of entity recognition can be expressed as follows:

$$\text{Precision} = \frac{\textit{Number of correctly identified entities}}{\textit{Total number of identified entities}}$$

The numerator (Correct Entities Found) represents the number of entities correctly identified by the system, matching the ground truth regarding type and boundary. The denominator (Total Entities Identified) represents all entities the system marked as correct, including both true and false positives. This metric evaluates the system's accuracy in retrieving only relevant entities, making it a crucial measure for assessing the quality of a named entity recognition (NER) or information extraction system.

6.2. Recall Measurement

Recall measurement is the ratio between the number of found correct entities and the total number of entities in the corpus.

$$\text{Recall} = \frac{\textit{Number of correctly identified entities}}{\textit{Total number of entities in the corpus}}$$

Recall measures the ability of a system to retrieve all relevant entities, focusing on minimizing the number of missed entities (false negatives). It is particularly important in applications where missing critical entities, such as medical or legal document analysis, can have serious consequences.

The F1-score is the harmonic mean of precision and recall, providing a single metric to evaluate a model's balance between precision and recall. It is formally defined as:

$$\text{F1-Score} = 2 \times \frac{\textbf{\textit{Precision}} \times \textbf{\textit{Recall}}}{\textbf{\textit{Precision}} + \textbf{\textit{Recall}}}$$

The F1 score combines precision and recall into a single measure, making it particularly useful when there is an uneven class distribution or when false positives and false negatives need to be considered equally important. This metric is widely used in tasks like Named Entity Recognition (NER) and other information extraction problems to ensure a balanced evaluation.

The above concept is visualized in the form of information retrieval as follows:

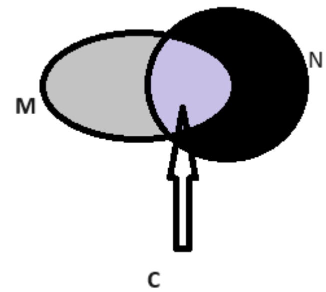

Fig. 3 The vein diagram representation of NER evaluation metrics. Precision = C/M, Recall=C/N and F-measure=2PR / (Precision + Recall)

7. Contribution and Importances

As a contribution, the paper presents the findings from the survey of literature sources, providing a comprehensive understanding of how NER is implemented in different application areas, such as finance, medicine, defense, business, food science, and archeology. It highlights the various algorithms used in NER and their contributions to the information extraction process. Additionally, the state-of-the-art techniques and evaluation metrics for NER are presented. As important, a paper adds value to the existing knowledge by collecting and making available different research findings related to NER and its applications. It emphasizes the importance of NER in information extraction and demonstrates the need for NER in specific domains.

8. Conclusion

In conclusion, this paper highlights the crucial role of NER in information extraction. It demonstrates the significance of NER in converting unstructured documents into structured forms, making targeted knowledge more accessible. The paper delves into different algorithms used in NER and their applications in various domains. It provides a comprehensive overview of NER's state-of-the-art techniques and evaluation metrics. Overall, this study contributes to the existing knowledge of NER and its relevance in information extraction tasks.

References

[1] Adnan Akhundov, Dietrich Trautmann, and Georg Groh, "Sequence Labeling: A Practical Approach," *arxiv*, pp. 1-10, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Abdelaziz Bouras, Jannik Laval, and Houssem Gasmi, "Information Extraction of Cybersecurity Concepts: An LSTM Approach," *Applied Sciences*, vol. 9, no. 19, pp. 1-15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Girma Yohannis Bade, "Natural Language Processing and Its Challenges on Omotic Language Group of Ethiopia," *Journal of Computer Science Research*, vol. 3, no. 4, pp. 26-30, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Girma Yohannis Bade, and Akalu Assefa Afaro, "Object Oriented Software Development for Artificial Intelligence." *American Journal of Software Engineering and Applications*, vol. 7, no. 2, pp. 22-24, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] Alex Brandsen et al., "Can BERT Dig It? Named Entity Recognition for Information Retrieval in the Archaeology Domain," *Journal on Computing and Cultural Heritage*, vol. 15, no. 3, pp. 1-18, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Alebachew Chiche, and Betselot Yitagesu, "Part of Speech Tagging: A Systematic Review of Deep Learning and Machine Learning Approaches," *Journal of Big Data*, vol. 9, no. 1, pp. 1-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yunfei Ji et al., "A Deep Learning Method for Named Entity Recognition in Bidding Document," *Journal of Physics: Conference Series*, vol. 1168, no. 3, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Zhiyong He et al., "A Survey on Recent Advances in Sequence Labeling from Deep Learning Models," *arxiv*, pp. 1-16, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Emna Hkiri, Souheyl Mallat, and Mounir Zrigui, "Improving Coverage of Rule Based NER Systems," *2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA)*, Marrakech, Morocco, pp. 1-6, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[10] Frederik Hogenboom et al., "An Overview of Event Extraction from Text," *Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), Workshop in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011)*, Bonn, Germany, pp. 48-57, 2011. [Google Scholar] [Publisher Link]

[11] Ali Jabbari et al., "A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News," *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, pp. 2293-2299, 2020. [Google Scholar] [Publisher Link]

[12] Mahboob Alam Khalid, Valentin Jijkoun, and Maarten de Rijke, "The Impact of Named Entity Normalization on Information Retrieval for Question Answering," *Advances in Information Retrieval, 30th European Conference on IR Research*, Glasgow, UK, vol. 4956, pp. 705-710, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[13] Jochen L. Leidner, Gail Sinclair, and Bonnie Webber, "Grounding Spatial Named Entities for Information Extraction and Question Answering," *HLT-NAACL-GEOREF '03: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, vol. 1, pp. 31-38, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[14] Girma Yohannis Bade, "Four Basic Concepts of Object Technology : A Structuring Method, a Reliability Discipline, an Epistemological Principle and a Classification Technique," *Journal of Advancement in Engineering and Technology*, vol. 7, no. 3, 2019. [Google Scholar]

[15] Mary Ellen Okurowski, "Information Extraction Overview," *TIPSTER '93: Proceedings of a Workshop on Held at Fredericksburg, Virginia*, Fredericksburg, Virginia, pp. 117-121, 1993. [CrossRef] [Google Scholar] [Publisher Link]

[16] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib, "Named Entity Recognition and Relation Detection for Biomedical Information Extraction," *Frontiers in Cell and Developmental Biology*, vol. 8, pp. 1-26, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Rody Politt, Joy Pollock, and Elisabeth Waller, *Day-to-Day Dyslexia in the Classroom*, Routledge, 2nd Ed., pp. 1-224, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[18] Gorjan Popovski, Barbara Korousic Seljak, and Tome Eftimov, "A Survey of Named-Entity Recognition Methods for Food Information Extraction," *IEEE Access*, vol. 8, pp. 31586-31594, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Utpal Kumar Sikdar, and Bjorn Gamback, "A Feature-Based Ensemble Approach to Recognition of Emerging and Rare Named Entities," *Proceedings of the 3rd Workshop on Noisy User-Generated Text*, Copenhagen, Denmark, pp. 177-181, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[20] Jie Tang et al., *Information Extraction: Methodologies and Applications*, Emerging Technologies of Text Mining: Techniques and Applications, Information Science Reference, pp. 1-358, 2008. [Google Scholar] [Publisher Link]

[21] Mesay Gemeda Yigezu et al., "Multilingual Hope Speech Detection using Machine Learning," *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, Jaen, Spain, 2023. [Google Scholar] [Publisher Link]

[22] Girma Yohannis Bade, and Hussien Seid, "Development of Longest-Match Based Stemmer for Texts of Wolaita Language," *International Journal on Data Science and Technology*, vol. 4, no. 3, pp. 79-83, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[23] Feifei Zhai et al., "Neural Models for Sequence Chunking," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 3365-3371, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jiayu Zhang et al., "Chinese Named Entity Recognition for Apple Diseases and Pests Based on Character Augmentation," *Computers and Electronics in Agriculture*, vol. 190, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Zexuan Zhong, and Danqi Chen, "A Frustratingly Easy Approach for Entity and Relation Extraction," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 50-61, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Girma Yohannis Bade, and Hussien Seid, "Development of Longest-Match Based Stemmer for Texts of Wolaita Language," *International Journal on Data Science and Technology*, vol. 4, No. 3, pp. 79-83, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[27] Girma Bade et al., "Social Media Hate and Offensive Speech Detection Using Machine Learning Method," *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta, pp. 240-244, 2024. [Google Scholar] [Publisher Link]

[28] Girma Bade, et al., "Social Media Fake News Classification Using Machine Learning Algorithm," *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, St. Julian's, Malta, pp. 24-29, 2024. [Google Scholar] [Publisher Link]