Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Advanced Developments in Arabic Named Entity Recognition: A comprehensive Study

Mahdi Ahmed Ali1[1] , Ahmed Bahaaulddin A. Alwahhab [2] , Yaghoub Farjami [3]

1- Middle Technical University, Technical College of Engineering\ Baghdad . maaah @ mtu.edu.iq

2- Middle technical University , Technical college of management , Information Technology management, ahmedbahaaulddin@mtu.edu.iq

3-University of Qom , Department of computer and IT , farjami@qom.ac.ir

* corresponding author: Mahdi Ahmed Ali

ARTICLE INFO

Received: 14 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Why has extracting important and essential information from Arabic texts become so useful and necessary? The answer to this question lies in the fact that the frequent appearance of Arabic words and texts on the Internet has led to interest in this topic. Named entity recognition (NER) refers to a fundamental task that is an integral part of many natural language processing (NLP) functions, such as information retrieval and machine translation, which are tasks of information extraction. When we review previous research and studies, we see that they relied on the recognition of known entities (NER) from large sources of knowledge and the application of hand-made features. This approach takes a long time and is no longer sufficient for languages with scarce resources, such as Arabic. Recently, the process of recognizing named entities in the Arabic language (NER) has begun to attract attention. The features and characteristics of the Arabic language, as a member of the Semitic language family, have become major challenges to recognizing named entities. The performance of the Arabic NER component positively impacts the overall performance of the NLP system. Many researchers have improved the methods to extract a variety of entities from different text types and languages. Additionally, there has been a push in the research community to update, develop, and implement new strategies that are considered more modern and innovative for extracting diverse and different entities that contain useful names in various natural language applications. In this paper, we provide an overview of the research advancements and progress made in classification studies and Arabic Named Entity Recognition.

Keywords: Arabic, Information Extraction, Named Entity Recognition, Natural Language Processing, Linguistic Resources

INTRODUCTION

In the last century, specifically in the 1990s, and in conferences held on the subject of understanding messages, Named Entity Recognition (NER) was mentioned as a new subject that was presented for the first time. The research community considered it very useful and necessary for extracting information.

Several levels will be used, including levels of letters and words, to be recognized. This gives rise to the concept of overlapping named entities. To achieve accurate identification, the authors drafted a model that combines information from both levels. As a result, compared to the traditional methods used, this model is effective in dealing with overlapping entities.[1]

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

There are several ways to identify overlapping entities. One of these methods is a modern approach that uses the bidirectional hierarchical model—conditional random field—long short-term memory (Bi-LSTM-CRF). As a result, the authors focus on leveraging this structure to handle relationships between different entities. We observe a significant improvement in the accuracy of identifying these entities.[2]

In NER, named entities (NEs) cover proper nouns but also include various numeric expressions, such as monetary amounts, time expressions, and other types of entities. Proper nouns encompass three classic classes: organizations, persons, and locations.[3]

The English language was the first to be addressed by some early articles about such systems, followed by publications on other languages, such as Indian, Dutch, German, Spanish, Japanese, and so on. Applications for the Arabic language were not launched until 2005.[4] This indicates that a significant amount of research in this domain has been conducted in various languages, achieving a high performance level comparable to that of English.

Over the past 10 years, significant progress has been made in Arabic Named Entity Recognition (ANER), and the presented systems have utilized different methods to recognize various types of NEs that could be challenging to categorize. Methods based on rule-based, multiple strategies, as well as machine learning (ML) strategies, have become more useful as the system can be easily adapted and trained with several different linguistic domains.[5]

Despite this progress, ANER systems have witnessed remarkable advancements in the past decade, but developing these powerful systems is still considered a challenge when compared to other languages.

The Arabic language has several characteristics that make the NER task more complex, such as:

- Lack of capitalization: The Arabic text lacks the use of capital letters, as is the case in other languages, which makes identifying entities more difficult.
- Grammatical complexity: The Arabic language requires addressing grammar, morphology, and word meanings. Due to the similarity in some of these words, careful treatment of the rules of grammar and morphology, as well as the use and deletion of diacritics and pronouns, is necessary to determine the accuracy of the meanings of these words.
- Ambiguity and spelling differences: There are also similar words that have the same letters but differ in meaning. In addition, the ambiguity in the text leads to some spelling differences. We will explain this in detail in the following paragraphs.[6]

Scarcity of resources: The lack of tagged and organized data hinders the development of NER and makes it ineffective. Many NLP techniques have been known in the Arabic language, especially in recent years, and have achieved good results. Examples include the Bi-LSTM conditional random field (CRF) technique and the BERT technique, through which it is possible to obtain an accurate concept and meaning of a word or phrase. Table 1 shows recent surveys and reports in Arabic and presents the development of methodologies. This will be explained more fully in the following paragraphs.[7]

Table 1. Recent Surveys and Reports on Arabic NER

Year	Type	Title
2023	Survey	A Survey on Arabic Named Entity Recognition: Past, Recent Advances, and Future Trends
2023	Report	WojoodNER 2023: The

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

		First Arabic Named Entity Recognition Shared Task
2022	Survey	A Critical Survey on Arabic Named Entity Recognition and Diacritization Systems
2021	Survey	A Survey On Deep Learning Approaches For Named Entity Recognition
2020	Survey	A Recent Survey of Arabic Named Entity Recognition on Social Media
2019	Report	Arabic Named Entity Recognition: What Works and What's Next
2019	Report	Arabic Named Entity Recognition Using Deep Learning Approach

Present a comprehensive review of Arabic NER, particularly the latest advances in deep learning (DL) and pre-trained language models. Jarrar et al. [8]

Present a successful solution to ANER challenges, incorporating various tailored methods such as sequence labeling and ensemble learning. Rateb et al. [9]

Explore cutting-edge toolkits for processing Arabic, including Madamira, Cameltools, and Farasa.[10]

Provide a comprehensive overview of current deep learning techniques for Named Entity Recognition. They begin by introducing NER's resources, such as its tagged kits and off-the-shelf tools. In addition, in 2020, Ali et al. [11]

Tests were conducted using English datasets during the Named Entity Recognition study in Arabic on social media. Several reports that introduce the competition methodology or system design are included in Table I.

One such report is the Challenge of NER, organized by Topcoder.com. These publications outline a functional system, but do not provide A comprehensive examination of how NER appears in Arabic. The linguistic peculiarities and unique difficulties of Arabic are not taken into account in these studies. There is no comprehensive analysis of Arabic NER techniques currently in use. We expect this evaluation to help in Organize and systematically identify entities named in Arabic techniques. It is worth noting It is difficult to discover information of high value through a language of less richness, such as Arabic which is also a great opportunity for the NLP community.[12]

The structure of this paper is as follows:

- **Provides background on the features of the Arabic language.**
- **Covers Arabic Language Characteristics.**
- **Key issues in Arabic NER.**
- **Reviews the existing literature on NER, focusing on advancements in deep learning approaches.**
- **Practical Applications.**
- **Addresses and Issues in the NER Task.**
- **Concludes the study and offers directions for future research.**

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Arabic Language Characteristics

The application of Named Entity Recognition (NER) and Natural Language Processing (NLP) presents unique challenges when dealing with the Arabic language due to its distinct characteristics. The following features significantly impact the effectiveness of Named Entity Recognition (ANER) systems: [13]

Absence of Capitalization: Unlike Latin-based languages, Arabic script does not utilize capitalization to denote proper nouns.

Here is a live example of a text in English and another in Arabic that illustrates the use of capital letters in the English text and its lack in the Arabic text.

"محمد رمى الكرة في ملعب كرة القدم, اعطى الكرة الى احمد, لعبوا كرة القدم"

Text With Capitalization:

"**Mohammed** threw the ball in the football field, he gave the ball to Ahmed, they played football".

Text without Capitalization:

"Mohammed threw the ball in the football field, he gave the ball to ahmed, they played football".

In the example above, names like "**Mohamed**" and "**Ahmed**" were capitalized to recognize proper nouns, which helped identify entities.

In summary, if the Arabic language used capital letters, it would enhance the clarity of the Arabic text and help identify entities more effectively.

This absence complicates entity identification, as the standard practice of capitalizing named entities cannot be applied. However, transliterated English equivalents may sometimes serve as useful indicators in this context.[14]

Agglutinative Nature: Arabic exhibits a high degree of agglutination, where words can incorporate various prefixes, suffixes, and inflections. This complexity contributes to a rich morphological structure that can complicate the identification of named entities. For instance, the root system in Arabic allows for numerous derivations, which can obscure the original entity.[15]

Optional Short Vowels: Diacritics (short vowels) are essential for disambiguation and accurate pronunciation in Arabic. However, they are frequently omitted in modern standard texts, leading to one-to-many ambiguities where a single word form can correspond to multiple meanings based on context. For example, the word "طرق" can mean "ways" or "to knock," and it is pronounced (torok ,taraka) depending on the diacritical markings that are not explicitly present.

The presence of diacritical marks, such as the hamza (ء), introduces further complexity, as it can be represented in multiple forms ("ا" "إ" "أ," or "آ"). This linguistic richness, while a strength, poses significant challenges for ANER systems, which often struggle with the lack of comprehensive linguistic resources.

Spelling Variants: Variability in spelling is common in Arabic script; a single word may be spelled in several ways while retaining a similar meaning. For example, the phrase "ولتكملوا" translates to "You can finally complete." Furthermore, Standard Arabic often omits diacritics—essential for accurately determining word meanings. It can also have different meanings despite having the same letters for the same word, such as the word (سائل) which comes with one meaning derived from the question and the other derived from the cases of the material, which is the liquid material**.**

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Additionally, ambiguity in Arabic script can lead to spelling variations, particularly with borrowed words. For instance, the word "تلغرام" can also be written as "تلگرامThis phenomenon creates many-to-one ambiguities, complicating the recognition and classification of named entities.

Lack of Linguistic Resources: There is a notable scarcity of linguistic resources available for Arabic, particularly annotated corpora suitable for training ANER systems. Many existing datasets lack sufficient named entity annotations, making them inadequate for effective NER tasks. The difficulty that researchers face in understanding and explaining their linguistic resources for evaluation and training due to the lack of Arabic dictionaries. .

Different dialects: Each of the different dialects has its own pronunciation, grammar, and vocabulary, which increases the complexity and difficulty of generalizing ANER, which is trained in Modern Standard Arabic (MSA). Therefore, it has become necessary to create advanced systems to accommodate special features of each dialect or models. This difference requires the development of systems that can accommodate dialect-specific features or the creation of models that contain classical dialect data with data from other dialects.

Switching of codes: The topic of code switching between Arabic and other different languages such as English, for example, especially in digital contexts of modern Arabic usage, is a common topic, through identifying the entities contained in different languages and the possibility of processing them within the same text, which can make this topic, which is the mixing of languages, a challenge added to the other challenges for ANER.

Due to the presence of these characteristics, it became necessary when developing ANER to address many of the obstacles and concerns that may be faced, including:

1- Morphological complexities

2- The ambiguity inherent in the language

3- Access to linguistic resources

4- And the typographical differences that the writer usually uses

To create an effective and powerful text in Arabic texts, we must address these challenges, to make the possibility of identifying and classifying the mentioned entities for NER systems very accurate.

Example:

Conversation:

"ليلى: "كيف كان فطارك في الصباح؟

مازن: "كان لذيذا It was very delicious!"

ليلى: "الحمد لله! How was the weather?"

مازن: "كان جيدا، especially the rain!"

Key Issues in Arabic NER

(ANER) Arabic Named Entity Recognition systems encounter several challenges related to the unique characteristics of the Arabic language. The primary issues include[16]:

The complexity of Arabic Script: One of the features of the Arabic text is the style of writing that is joined or attached to words and texts, which makes spelling difficult and different from that found in other languages, such as the English language, for example, which leads to a lack of understanding and

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

identifying named entities, and makes this topic one of the challenges facing ANER. Arabic calligraphy has characteristics, features, and types that we can discuss briefly.

There are several types of Arabic calligraphy, and each type has its own features and characteristics, such as Thuluth , Naskh, Diwani, and others. Arabic calligraphy also has aesthetics, decoration, and art used in architecture, mosques, and others, which expresses coordination and aesthetics. And the arts and religious expression in it.

Complex morphology: Complicated Morphology: The morphological complexity in the Arabic language refers to the way words are formed, varied, and derived from a single root. The following is a brief overview of these derivations:[17]

The root is three letters, معلم, معلم, علم, علم, م, ل, ع.

It also uses specific and diverse forms to produce new words such as the verb: علم (its various verbs علم, يعلم)

The active participle: معلم (that which he taught)

The direct object: معلوم (that which he is known)

The inflection is also done according to the verbs (past, present, imperative) (درس, يدرس, أدرس)

New words can also be formed by adding letters such as the possessive letter as in (سيارتي) the letter كي أدرس ل (لادرس) (كي, ( )).

Therefore, we see that this morphological complexity in this language expresses the richness of this language and adds importance and flexibility to its use by the general educated public**.**

Lack of Resources: There is a general shortage of resources to test NER systems in Arabic , and only a few corpora have been created by individual researchers. Some of these corpora are publicly available, while others are accessible under license agreements.[18]

Capitalization Issue: Arabic orthography does not use capital letters to identify the initial letters of proper names, as in other languages like English. This makes the identification of named entities, whether expressed in single words or word sequences, challenging.

The presence of capital letters in the Arabic language causes some problems and difficulties for non-native speakers. Examples of the presence of capital letters are in (words such as names, titles, days (**الاثنين**) and months (**اذار**), abbreviations, religious and cultural terms(**القران**)).

Named Entity Inherent Ambiguity: Like other languages, Arabic NLP systems face the issue of ambiguity between two or more named entities, which poses even more critical challenges for modeling Arabic NLP systems.

There is ambiguity in some words that are difficult to identify and understand and come in the following forms:

1- As names (ذهب أحمد الى المدرسة)..(Ahmed went to school)..... Who is Ahmed?

2- Place (سأذهب الى النادي).(I will go to the club)... Which club?

3- Event (فاز أحمد بالمدالية في السباق).(Ahmed won the medal in the race)... Which medal and which race?

4- Company name (اشتريت من امازون جهاز الكتروني). (I bought an electronic device from Amazon)... Which Amazon?

5- Story, movie or book (زار الامير بيتنا). (The prince visited our house)... Which prince?

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

We see in these examples the presence of ambiguity in understanding the text and the difficulty in identifying the named entity, so it was necessary to include or add supporting words that help complete the text, understand its meaning and identify the named entity for NER.

LITERATURE REVIEW

Researchers at NER adopt four types of approaches and methods: hybrid approach, rule-based approach, and machine learning approach.

RULE-BASED APPROACHES

The techniques based on rules for NER in Arabic were among the first to be presented, as they demonstrate the development of simplified grammatical laws. Such techniques generally depend on the specific expert knowledge on language presented by gazetteers/human linguists and also need manually created attributes for being expanded, particularly for named entities recognition. NER systems based on rules rely on manually compiled local grammatical rules which the experts extracted and confirmed in linguistics. Gazetteers and lexicons (dictionaries) are used as the structural rules in a text that the named entities display. Initially, many gazetteer vocabularies are applied as trigger words for aiding the recognition of named entities in the presented context.

Knowing Arabic names: In [19], the authors presented a technique for identifying Arabic person names using an Arabic entity type lexicon. Such techniques covered the dictionaries with a statistical model given the rules for model extraction that could diagnose the appearance of person names. Moreover, four rules were created for name identification given the linguistic data of naming. Such a strategy was performed in 3 domains: politics, sports, and economics. This presents great outcomes in every domain in terms of f-measure: 90.43% for politics, 92.04% for economics, and 92.66% for sports

Extract Named Entity from Crime Documents: In [20], the authors performed rule-based techniques for ANE extraction from crime documentaries. They built and shaped numerous syntactical rules as well as ANER model general indicator lists, using morphological data and an ANE glossary corpus from the crime domain. Such rules and models could identify and group named entities. These techniques show great performance and obtained 90% accuracy.

Grammar Analysis Based on Heuristic: In [21], the authors analyzed Arabic sentence grammar and looked for NE indicators applying heuristic-based rules. Another successful strategy was the use of POS morpho-syntactic tags, which could be used to recognize NE boundaries in sentences

The Examination of Classical Arabic Document: In [22], the authors described a rule-based NER technique that could be applied to Classical Arabic documentaries. The presented technique depended on trigger blacklists, patterns, rules, vocabularies, and gazetteers, created using linguistic information on ANE. This technique operates in 3 steps: operational, preprocessing, and rule application..

Named Entity Recognition in Various Domains: In [23] a strategy was proposed for recognizing named entities, particularly individual names, across sports, economics, and politics. The method demonstrated a precision of 92.25% and recall of 91.25%, with potential applicability to other fields such as religion and medicine. In Table 2 explain the comparison of approaches which used for Rule-based NER in this paper.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2. Comparison of approaches used for Rule-based ANER

Method	Dataset	Evaluation Results	Advantages
Rule-based ANER systems using dictionaries	ANERcorp	Precision= 92.29%, Recall= 72.75%, F-score= 81.36%	Effective in addressing ambiguity in Arabic personal names
Rule-based method for crime documents	Arabic crime documents	Accuracy=90%, Precision=91%, Recall=89%, F-score=89.46%	Considers morphological and POS information
Heuristic-based with keywords	500 news articles from Aljazeera	Precision=88%, Recall=90%, F-score=89%	Utilizes structured laws for accurate name recognition
Rule-based method for Classical Arabic	CANERCorpus	Precision=90.2%, Recall=89.3%, F-score=89.5%	Addresses specific challenges in Classical Arabic texts
Rule-based approach for named entity types	In-house corpus from online Arabic newspapers	Precision=92.25%, Recall=91.25%, F-score=91.71%	Adaptable to other fields like religion and medicine

MACHINE LEARNING APPROACHES

Various ML methods have been confirmed to be efficient for linguistic tasks, and many researchers have extended ANER techniques by applying them. Typically, these techniques rely on learning algorithms to obtain relationships among vocabularies and recognize named entities based on statistical computations.

In [24] the authors used a feature selection strategy applying a genetic algorithm to identify an optimized set of features.

Then, the best optimal feature set integration was applied for recognizing and grouping ANEs using support vector machines.

In [25] the authors presented an ML-based NER system that used CRF and SSVM as ML components. They used FARASA-based classifiers for the preprocessing step, which allowed them to address various orthographic and morphological complexities in Arabic. Additionally, they investigated the influence of location, POS, person, pre-name percolation whitelist, organization gazetteers, and bag-of-words features.

In [26], the authors used SVM recursive feature elimination (SVM-RFE) techniques for feature selection to identify an optimal set of features. Then, the optimal feature set was integrated and applied for named entity recognition classification using SVMs.

In [27], the authors defined a technique called NAMERAMA using a Bayesian Belief Network (BBN) pattern for extracting Arabic medical text signs, illness names, and treatment detection.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In [28], the authors presented ensemble architecture for the crime named entity identification task.

The primary goal was to effectively combine sets of features with classification algorithms for proper classification synthesis. First, three popular text classification algorithms (SVM, Naïve Bayes, and K-Nearest Neighbor) were used as main classifiers for each feature set. Second, a weighted voting ensemble technique was applied to combine the three classifiers.

These strategies can achieve great results based on high-quality annotated datasets for training. However, the disadvantage is the high cost and time-consuming nature of creating such annotated datasets, which requires expertise in this field. .In Table 3 explain the comparison of approaches which used for Machine Learning NER in this paper.

Table 3. Comparison of approaches used for Machine Learning ANER

Method	Dataset	Evaluation results	Advantages
Arabic named entities (NEs) based on support vector	**randomly chosen tweets from May 3-12, 2012, from November 23, 2011, to November 27, 2011**	**Precision= 88.66%** **Recall= 65.80%** **F-score= 76.79%**	**This significantly reduce features' number required for system training and simultaneously develop named entity identification performance.**
two machine learning algorithms CRF and SSVM	**ANERCrop4**	**Precision= 86.86%** **Recall= 79.38%** **F-score= 82.76%**	**NER task based on SSVM needed less time and obtained better performance than NER task CRF based for clinical entity identification applying similar attributes.**
the support vector machine recursive feature elimination (SVM-RFE)	**Darwish's dataset**	**Precision= 78.41%** **Recall= 56.34%** **F-score= 65.57%**	**Decreased attributes performed better than associated complete sets of feature.**
Bayesian Belief Network (BBN) model	**27 articles describing different types of cancer,**	**Precision= 96.60%** **Recall= 90.79%** **F-score= 93.60%**	**BBNs are really flexible. BBN approach obtained great NEs precision in illness and treatment levels' techniques.**

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Naïve Bayes, Support Vector Machine and K-Nearest Neighbor	**the data of the corpus were collected from the Malaysian National News Agency (BERNAMA)**	**Accuracy= 89.48%** **F-score= 93.36%**	**Presented model was important to recognize crime kind also associated named entities extraction from crime documentaries.**

EMERGING TRENDS AND FUTURE DIRECTIONS

Recent trends in ANER research include the integration of contextual embeddings, such as those from BERT-based models, which have shown promise in enhancing entity recognition by considering the surrounding context. Moreover, the exploration of transfer learning and multilingual models could provide significant advancements in handling the complexities of Arabic NER.

In conclusion, while rule-based methods have established a strong foundation in ANER, the integration of machine learning and hybrid approaches represents a promising direction for future research. Continued exploration of these methodologies, alongside the development of richer linguistic resources, will be vital for improving the accuracy and robustness of Arabic NER systems.

DEEP LEARNING APPROACHES

Currently, deep learning (DL) techniques have been widely applied for ANER tasks. Compared to traditional, manually-engineered feature-based strategies, DL techniques can learn more powerful attributes and reduce the effort required for handcrafting features. Additionally, gradient-based DL can train models in an end-to-end fashion. Leveraging these advantages, researchers have been able to design more sophisticated ANER systems..

In [29], the authors applied the Madamira external resource to create additional vocabulary attributes. They also evaluated the effect of adding them to the traditional representations of words and features to perform the tasks of NER on Modern Standard Arabic (MSA) texts. By using the Bidirectional Long Short-Term Memory architecture and Conditional Random Fields (BiLSTM-CRF) the NER model was implemented. They also introduced various syntactic, morphological, and grammatical features of the vocabulary representations to train the patterns.

In [30], to address the ANER model task, the authors introduced a new neural network operating style. In various natural language processing (NLP) applications, the proposed strategy has benefited from the recent success of deep neural networks (DNNs). They have applied different integrated DNN operating styles by using Convolutional Neural Networks (CNNs), BERT and Long Short-Term Memory (LSTM) with the feature representation set to generate rich semantic and syntactic vocabulary representation vectors.

In [31], to address the ANER topic, the authors introduced a special strategy with multiple representations. They also evaluated this strategy using the AQMAR dataset. Considering the specific challenges of Arabic, the authors and the Arabic NER transformers (BERT) investigated the performance of the combined contextual representations as well as the bidirectional encoding representations. It uses

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

pre-trained words and combined contextual representations from the beginning to the end. The proposed technique is a DL model, with the integration of the BERT model.

In [32] the authors presented bidirectional encoder-decoder architecture to address the ANER task, where the encoder and decoder are bidirectional LSTM. In addition to the word-level and feature-class representations, they integrated them through a feature-class attention mechanism. This model can dynamically determine which information to use from a word/feature-class element through the attention mechanism.

In [33], the authors classified named entities in Arabic contexts and presented a new strategy for recognizing these entities. Using a semi-supervised learning mechanism called joint training; they proposed a deep joint learning strategy, which they adapted to a deep learning model.

To train the deep joint learning strategy (NER), they first extended a Wikipedia-based classifier using LSTM DNN, and to create a semi-labeled dataset for the Arabic NER task, this classifier was applied.

These methods are generally complex and involve many different steps, such as model training, feature engineering, and classification. To handle linguistic tasks, there are different network frameworks that can be applied to handle linguistic tasks, including LSTMs, RNNs, CNNs, and others. . In Table 4 explain the comparison of approaches which used for Deep Learning NER in this paper.

Table 4. Comparison of approaches used for Deep Learning ANER

Method	Dataset	Evaluation Results	Advantages
BiLSTM-CRF	ANERcorp, AQMAR	Precision= 87.77%, Recall= 85.66%, F-score= 86.71%	Enhanced performance through additional vocabulary attributes
CNN, LSTM, BERT	ANERcorp	Precision= 93.77%, Recall= 93.60%, F-score= 93.68%	Comprehensive integration of morphological, orthographic, and semantic features
BERT with pooled contextual embeddings	AQMAR	Precision= 79.53%, Recall= 75.79%, F-score= 77.62%	Ability to create embeddings for previously unseen vocabulary
Bidirectional LSTMs	ANERCorp, AQMAR	Precision= 93.52%, Recall= 90.54%, F-score= 92.01%	Dynamic information assignment via attention mechanism
LSTM DNN	TWEETS, AQMAR, NEWS	F-score= 74.1%	Significant improvement through semi-labeled data

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

HYBRID APPROACHES

Hybrid approaches combine the strengths of both machine learning and rule-based techniques, leveraging predefined rules while allowing for model adaptation through training on annotated datasets. These methods aim to enhance performance across diverse domains and tackle the challenges of Arabic NER more effectively.[34]

In ANER hybrid methods take advantage of the strengths of both rule-based and machine learning techniques, combining various strategies to enhance named entity identification. This method often involves using simpler techniques alongside more complex ones to optimize performance.

Combination of Rule and ML Methods: In , a hybrid system was developed that combined a rule-based approach with decision tree classifiers. This integration led to an F-score improvement of 8% to 14% compared to traditional rule-based systems and pure ML methods.[35]

Integration of Multiple Observed Methods: The work in [36] improved ANER systems by integrating rule-based strategies with logistic regression, decision trees, and SVMs. This multifaceted approach yielded superior performance against new ANER systems on standard test sets.

Statistical Models with Rule-Based Systems: In [37], a novel NER system was proposed that combined rule-based and machine learning strategies to improve ANER performance. This system successfully identified three types of named entities: organizations, persons, and locations.

Dependency Parsing and Clustering: The study presented in [38] utilized a combination of dependency parsing and clustering algorithms to achieve fine-grained NER. This approach addressed the limitations of traditional window-based representations, significantly enhancing recognition accuracy. In Table 5 we explain the Comparison of approaches used for Hybrid ANER in this paper.

Table 5. Comparison of approaches used for Hybrid ANER

Method	Dataset	Evaluation Results	Advantages
Decision tree with rule-based systems	ANERcorp, ACE Newswire, ACE Broadcast News	Precision= 94.9%, Recall= 90.78%, F-score= 92.8%	Effective tagging across diverse corpora
Rule-based with SVMs, Decision Trees	ANERcorp	F-score= 94.4%	Capable of identifying multiple entity types
Rule-based and SVM	ANERcorp	Precision= 97.01%, Recall= 96.30%, F-score= 96.65%	Robust identification of key entity types
Rule-based with statistical models	1,423 tweets	Precision= 85.40%, Recall= 65.17%, F-score= 73.92%	Transferable to other fields and languages
Dependency parsing and clustering	WikiFANE	Precision= 84.98%, Recall= 65.00%, F-score= 73.66%	Tackles limitations of window-based representations

In summary, deep learning and hybrid approaches provide promising paths for developing Arabic named entity recognition. While deep learning provides powerful tools for automatic feature extraction and

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

representation, hybrid approaches facilitate the integration of diverse techniques to improve performance. Future research should focus on improving these methodologies, exploring new architectures, and developing richer and more annotated datasets to enhance the capabilities of Arabic named entity recognition systems. In figure 1 A diagram showing the rule methods used for Hybrid ANER and Relationship between rule methods and Precision, Recall, F-score in this paper.

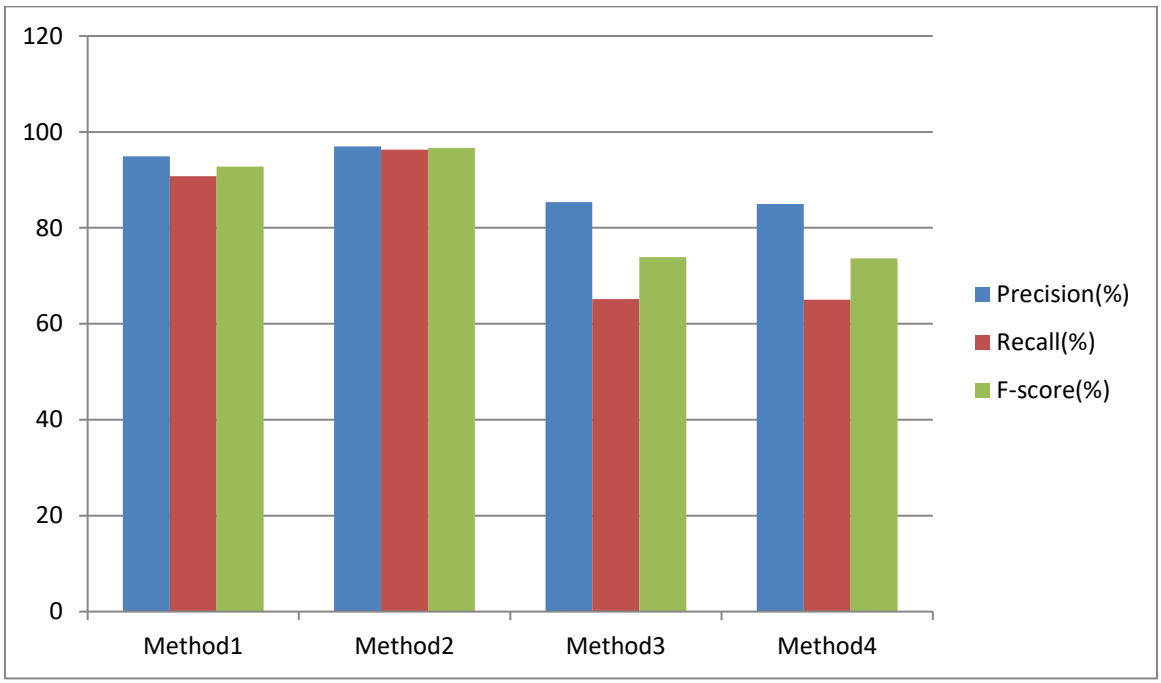

Fig.1Relationship between rule methods and Precision ,Recall ,F-score

Method1: Decision Tree with Rule-Based Systems

Method2: Rule-Based with SVMs, Decision Trees

Method3: Rule-Based and SVM

Method4: Rule-Based with Statistical Models

Practical Applications

NER systems are used in many different fields, such as:

Media data analysis: NER can be used to analyze media texts and extract important entities such as people and organizations

NER in media data analysis in academic journals such as: Journal of Information Science, Natural Language Engineering.[39]

Example: Imagine that you have a set of news articles about political events. Using NER, you can analyze these articles to extract important entities such as names, organizations, and places.

Analysis steps: Data entry

A news article such as: "التقى الرئيس الامريكي ترامب الرئيس الفرنسي جو ماكرون في البيت الابيض"

NER application: NER system analyzes the text and identifies the entities

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

People: ترامب , جو ماكرون

Place: البيت الابيض

(NER) in extracting medical information from health records in academic journals such as:

Journal of Biomedical Informatics ,International Journal of Medical Informatics

Example: Scenario for using NER in Health records

Input data: Text from a medical record

" تم تشخيص حالة المريض محمد بانسداد شرياني ويحتاج إلى علاج بالأسبرين"

NER application: NER system analyzes the text and identifies entities

Persons: محمد

Diseases: انسداد الشريان

Treatments: الاسبرين

We conclude that doctors and researchers can use this information to extract data about the number of patients with a certain disease, or to know the types of treatments commonly used.[40]

Challenges and Issues in the NER Task

The NER task faces several challenges that complicate its implementation and effectiveness. Key concerns include accessibility of resources, handling nested entities, text ambiguity, and the quality of training data annotations. Addressing these issues is crucial for developing robust NER systems. The following challenges are highlighted:

Nested Entities

Nested entities refer to the phenomenon where entities are contained within other named entities. This situation poses significant challenges for NER systems, as it complicates entity identification and classification. To mark the sector as a potential solution, the traditional approach was put forward, enabling researchers to better manage nested structures. [41]

Text Ambiguity

Text ambiguity arises when a named entity can be interpreted in multiple ways, depending on the context. For example, the term "Jordan" may refer to a geographical location or a person's name. Effective disambiguation is essential for accurate entity recognition. A decisive role in resolving the issue of text ambiguity Contextual information plays a decisive role in resolving this ambiguity, guiding systems in determining the correct entity type [42]. Advanced disambiguation techniques, including context-aware models, are needed to enhance the accuracy of NER systems.

Training Data Annotation

In NER and its related supervised learning techniques, annotated data can be a vital element. However, we can say that the process of annotating data requires real expertise and effort.

In many cases, this leads to a shortage of high-quality annotated datasets. To mitigate this issue, researchers are exploring semi-supervised and unsupervised learning methods, which can leverage small amounts of labeled data as seed samples for training.[43]

These approaches can help improve NER performance in data-scarce scenarios.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Resource Shortages

The effectiveness of NER systems heavily relies on the availability of large annotated corpora and linguistic resources such as gazetteers, morphological analyzers, and POS taggers. Unfortunately, many languages, including Arabic, Indonesian, and various South Asian languages (e.g., Urdu, Hindi, Bengali), lack adequate resources, making the NER task even more challenging.[44]

Efforts to develop and share linguistic resources in these languages are crucial for advancing NER research.

Domain-Specific Challenges

Different domains may present unique challenges for NER systems. For instance, medical terminology, legal language, and technical jargon often include specialized terms that standard NER systems may struggle to recognize. Tailoring NER approaches to specific domains by incorporating domain knowledge and specialized lexicons can enhance performance and accuracy.[45]

Scalability and Adaptability

As organizations increasingly require NER solutions for diverse applications, the scalability and adaptability of NER systems become critical. Systems must be able to handle large volumes of data across various contexts and languages. Developing flexible architectures that can be easily adapted to new domains and languages is an ongoing challenge in the field.[46]

CONCLUSION

NER is an emerging domain that is increasingly becoming a crucial integration In many different natural language applications. This paper aims to inform researchers about the key information related to ANER, its history, current developments, and future possibilities. The paper will help novice researchers gain a perspective on the problems and challenges associated with ANER classification. It provides a systematic overview of ANER strategies, as well as systematic overview of rule-based, deep learning (DL), machine learning (ML), and hybrid ANER systems.. Additionally, the paper presents a tabular comparison of the mentioned systems, providing helpful information on these strategies. The outcomes of various ANER systems are examined and discussed in detail. Finally, the paper highlights several NER task challenges that can assist researchers in improving ANER methods, thereby advancing the field of study.

REFERENCES

[1] Zhou, J., et al. (2019). "A Novel Approach for Nested Entity Recognition Using a Hierarchical Bi-LSTM-CRF Model." Journal of Natural Language Engineering.

[2]Lin, Y., et al. (2020). "Character-Level and Word-Level Representations for Named Entity Recognition." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics..

[3] Shaalan K. A survey of Arabic named entity recognition and classification. Computational Linuistics.2014 jun1;40(2):469-510.

[4] Ali BA, Mihi S, El Bazi I, Laachfoubi N. A Recent Survey of Arabic Named Entity Recognition on Social Media. Rev. d'Intelligence Artif.. 2020 May;34(2):125-35.

[5] Salah RE, binti Zakaria LQ. A comparative review of machine learning for Arabic named entity recognition. International Journal on Advanced Science. Engineering and Information Technology.2017 Apr;7(2):511-8.

[6] Alsaaran N, Alrabiah M. Arabic named entity recognition: A BERT-BGRU approach. Compute. Mater. Contin. 2021 Jan 1;68(1):471-85.

[7] Qu X, Gu Y, Xia Q, Li Z, Wang Z, Huai B. A survey on Arabic named entity recognition: Past, recent advances, and future trends. IEEE Transactions on Knowledge and Data Engineering. 2023 Aug 8.

[8] Jarrar M, Abdul-Mageed M, Khalilia M, Talafha B, Elmadany A, Hamad N, Omar A. WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. arXiv preprint arXiv:2310.16153. 2023 Oct 24.

[9] Rateb MN, Alansary S. A critical survey on arabic named entity recognition and diacritization systems. In2022 20th international conference on language engineering (ESOLEC) 2022 Oct 12 (Vol. 20, pp. 158-165). IEEE

[10] Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. IEEE transactions on knowledge and data engineering. 2020 Mar 17;34(1):50-70.

[11] Liu L, Shang J, Han J. Arabic named entity recognition: What works and what's next. InProceedings of the Fourth Arabic Natural Language Processing Workshop 2019 Aug (pp. 60-67).

[12] El Bazi I, Laachfoubi N. Arabic named entity recognition using deep learning approach. International Journal of Electrical & Computer Engineering (2088-8708). 2019 Jun 1;9(3).

[13] Shaalan K, Oudah M. A hybrid approach to Arabic named entity recognition. Journal of Information Science. 2014 Feb;40(1):67-87.

[14] Farber B, Freitag D, Habash N, Rambow O. Improving NER in Arabic Using a Morphological Tagger. InLREC 2008 May.

[15] AbdelRahman S, Elarnaoty M, Magdy M, Fahmy A. Integrated machine learning techniques for Arabic named entity recognition. IJCSI. 2010;7(4):27-36

[16] Salah RE, Zakaria LB. Arabic rule-based named entity recognition systems progress and challenges. International Journal on Advanced Science, Engineering and Information Technology. 2017 Jun;7(3):815-21.

[17] Saif A, Ab Aziz MJ, Omar N. Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features. Natural Language Engineering. 2017 Jan;23(1):53-91.

[18]Salah RE, binti Zakaria LQ. A comparative review of machine learning for Arabic named entity recognition. International Journal on Advanced Science, Engineering and Information Technology. 2017 Apr;7(2):511-8.

[19] Zayed OH, El-Beltagy SR, Haggag O. A Novel Approach for Detecting Arabic Persons' Names using Limited Resources. Res. Comput. Sci.. 2013;70:81-93.

[20] Asharef M, Omar N, Albared M, Minhui Z, Weiming W, Jingjing Z. Arabic named entity recognition in crime documents. Journal of Theoretical and Applied Information Technology. 2012 Oct 15;44(1):1-6.

[21] Elsebai A, Meziane F. Extracting person names from Arabic newspapers. In2011 International Conference on Innovations in Information Technology 2011 Apr 25 (pp. 87-89). IEEE.

[22] Salah R, Mukred M, Qadri binti Zakaria L, Ahmed R, Sari H. [Retracted] A New Rule-Based Approach for Classical Arabic in Natural Language Processing. Journal of Mathematics. 2022;2022(1):7164254.

[23] Aboaoga M, Ab Aziz MJ. Arabic person names recognition by using a rule based approach. Journal of Computer Science. 2013 Jul 1;9(7):922.

[24] M Rostami, K Berahmand, S Forouzandeh - Journal of Big Data, 2021 - Springer, A novel community detection based genetic algorithm for feature selection.

[25] Y Zhang, X Wang, Z Hou, J Li - JMIR medical informatics, 2018 - medinform.jmir.org, Clinical named entity recognition from Chinese electronic health records via machine learning methods.

[26] Elsevier - N Rtayli, N Enneya - Journal of Information Security and Applications, 2020Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization.]

[27] eprints.staffs.ac.uk, S Alanazi - 2017 A named entity recognition system applied to Arabic text in the medical domain.

[28] HA Shabat, N Omar - Named entity recognition in crime news documents using classifiers combination, Middle-East Journal of Scientific ..., 2015, repository.atu.edu.iq.

[29] A Pasha, M Al-Badrashiny, MT Diab, A El Kholy Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. ... - Lrec, 2014 academia.edu.

[30] RJ Abrahart, F Anctil, P Coulibaly... Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting - Progress in ..., 2012 - journals.sagepub.com.

[31] S Ainsworth - Computers & education ,The functions of multiple representations 1999 - Elsevier.

[32] MNA Ali, G Tan - Bidirectional Encoder–Decoder Model for Arabic Named Entity Recognition - Computer Engineering and Computer Science,Published: 06 August 2019 Arabian Journal for Science and Engineering, 2019 - Springer.

[33] K Shaalan, H Raza - NERA: Named Entity Recognition for Arabic,First published: 22 April 2009 https://doi.org/10.1002/asi.21090 Journal of the American Society for ..., 2009 - Wiley Online Library.

[34] Abdallah S, Shaalan K, Shoaib M. Integrating rule-based system with classification for arabic named entity recognition. InInternational Conference on Intelligent Text Processing and Computational Linguistics 2012 Mar 11 (pp. 311-322). Berlin, Heidelberg: Springer Berlin Heidelberg.

[35] Oudah M, Shaalan K. A pipeline Arabic named entity recognition using a hybrid approach. In Proceedings of COLING 2012 2012 Dec (pp. 2159-2176).

[36] Deciphering Arabic question: a dedicated survey on Arabic question analysis methods, challenges, limitations and future pathways, Open access, 13 August 2024,Volume 57, article number 251, (2024).

[37] Sherief Abdallah, Khaled Shaalan & Muhammad Shoaib.Integrating Rule-Based System with Classification for Arabic Named Entity Recognition

Conference paper ,pp 311–322,Cite this conference paper,Computational Linguistics and Intelligent Text Processing,(CICLing 2012.

[38] Zara Nasar, Syed Waqar Jaffry, Muhammad Kamran Malik,Named Entity Recognition and Relation Extraction: State-of-the-Art,Authors Info & Claims,ACM Computing Surveys (CSUR), Volume 54, Issue 1,Article No.: 20, Pages 1 – 39,https://doi.org/10.1145/3445965, 11 February 2021 Publication History

[39] Elsebai A, Meziane F. Extracting person names from Arabic newspapers. In2011 International Conference on Innovations in Information Technology 2011 Apr 25 (pp. 87-89). IEEE.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[40] Huang, M., et al. (2020). "Named Entity Recognition in Electronic Health Records: A Survey." Journal of Biomedical Informatics, 103, 103-119.

[41] Soukaina MI, Ismail EL, LAACHFOUBI N. Arabic Named Entity Recognition on Social Media based on feature selection techniques using SVM-RFE. In2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS) 2020 Oct 21 (pp. 1-7). IEEE.

[42] Hatab AL, Sabty C, Abdennadher S. Enhancing deep learning with embedded features for Arabic named entity recognition. In Proceedings of the Thirteenth Language Resources and Evaluation Conference 2022 Jun (pp. 4904-4912).

[43] Nothman, J., et al. (2013). "Evaluating Named Entity Recognition Systems for Languages with Low Resources." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.

[44] Baker, P., & McDonald, C. (2018). "Challenges in Building Resources for Low-Resource Languages." Language Resources and Evaluation.

[45] Kumar, A., & Singh, S. (2020). "Domain Adaptation for Named Entity Recognition: A Survey." ACM Computing Surveys.

[46] Chen, Y., et al. (2021). "Scalable Named Entity Recognition for Big Data: A Review." Big Data Research.

MAHDI AHMED ALI received his Bachelor's degree from the Technical College of Management / Middle Technical University in Baghdad, Republic of Iraq, and his Master's degree from Ferdowsi University of Technology in the Islamic Republic of Iran. He is currently working as a professor at the Middle Technical University. He is currently studying for his Ph.D. in the Department of Information Technology at Qom University in the Islamic Republic of Iran. His research interests include artificial intelligence and named entity recognition in Arabic.Email:maaah@mtu.edu.iq,OrcidNumber:0000-0002-9760-2352.

Ahmed Bahaaulddin A.Wahhab .A lecturer in the Information Technology Management Department. ,. B.Sc. in computer science and education UOB. 2001, MSc in computer science from Iraqi Commission for Computer and Informatics 2005, lecturer at the Middle Technical University/Technical College of Management since 2005, teaching data structure, database, data compression, and programming. His research interests in artificial intelligence, natural language processing, and recommender systems .Email: ahmedbahaaulddin@mtu.edu.iq
OrcidNumber: 0000-0003-0965-4812.

Yaghoub Fargjami The ideal student for the master's stage at the Faculty of Mathematics, Sharif University of Technology 1370 • The first student to complete the expert course in three years in the 146-unit system at Sharif University of Technology • Selected as a student in the master's course at the Faculty of Mathematics, Sharif University of Technology in 2012 • Graduated with first class honors, Master's degree from the Faculty of Mathematics, Sharif University of Technology, academic year 1372-73• Distinguished graduate student in the doctoral course in mathematics, Faculty of Mathematics, Sharif University of Technology, 1377 AH. Email: farjami@qom.ac.ir, OrcidNumber: 0000-0003-1908-8826.