Named Entity Recognition in NLP For Hindi and Arabic: A Comparative Study

Mohammed Nasser Jaber

Department of Computer Science and Engineering, Vivekananda Global University, Jaipur, India

abunasserip@gmail.com

Abstract. This work is a first-of-its-kind study in its subject. It undertakes a comparative investigation into Named Entity Recognition (NER) across two common linguistical low-resource languages, Hindi and Arabic, to elucidate language-specific challenges and efficacious methodological innovations within the Natural Language Processing (NLP) domain. Through a systematic evaluation of varied NER models and algorithms, incorporating linguistic feature analysis, data annotation, and machine learning techniques, the research identifies distinct challenges pertaining to the script, morphology, and contextual dependencies inherent to each language. By highlighting the efficacy of language-specific model adaptations, the study offers novel insights into optimizing NER performance for underrepresented languages, thereby contributing a foundational framework for advancing cross-linguistic NLP research and technology development.

Keywords: Hindi Language, Arabic Language, NLP, Named Entity Recognition, Comparative Study, Linguistical Challenges

1 Introduction

NLP is a modern field that draws on multiple sciences to allow computers to understand human language. It's recognized as one of the more challenging fields in artificial intelligence, as it often requires a strong comprehension of the world and be able to use what one learns to better apply it. NLP draws on a number of sciences, including mathematics, computer sciences, information sciences, artificial intelligence, and robotics [11]. One practical application of NLP is Named Entity Recognition (NER): the ability to identify and classify key elements within text, such as names of people, organizations, and locations [2]. NER refers to a building block subtask of the macro–NLP domain and serves as a key technique in information extraction, automatically capturing and classifying named entities contained in non-structured text [1]. It enables robust analysis of textual data by identifying key words, phrases, themes, emotional nuances, medical codes, time expressions, quantities, and monetary value, facilitating applications such as sentiment analysis, content summarization, and information retrieval [2].

N. Vyas et al. (eds.), *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA 2025)*, Advances in Intelligent Systems Research 189,
https://doi.org/10.2991/978-94-6463-700-7_34

The NER challenge, first articulated during the Sixth Message Understanding Conference (MUC-6) in 1995 [1], has significantly shaped the development of systems for identifying named entities in text. It has developed the standard for systems with named entities. Primarily, the task of NER entails the identification of the limits of entities in a text and then classifying them according to pre-defined classes. A popular approach to address NER is to conceptualize it as a sequence-labelling-task [3]. For example, in the expression, "Mahatma Ghandhi visited Aden in September 1931," an ideal NER system would identify " Mahatma Ghandhi " as a PERSON, "ADEN" as a PLACE, and "September 1931" as a DATE. The competence and effectiveness with which such entities can be identified underpin a broad spectrum of the NLP applications.

The advantages of NER have been appreciated for high-resource languages, particularly English, for quite some time. The phenomenal growth of the digital sphere underscores a vital and urgent necessity on a national scale, the development of effective NER and multilingual NLP for low-resource languages. Recent research reveals the challenges and potential solutions for NER tasks in low-resource languages, as multilingual learning approaches have shown promise in addressing data sparsity issues by leveraging features from closely related languages. These methods can improve NER performance through increased named entity vocabulary, cross-lingual sub-word features, and regularization effects [8]. Transfer learning techniques, such as unsupervised truth inference and limited supervision, have also demonstrated effectiveness in modulating transfer from multiple source languages to low-resource target languages [9]. Further, data augmentation strategies, including labelled sequence translation and generation-based methods, can enhance cross-lingual NER performance [10][11][8]. These approaches offer promising avenues for improving NER in low-resource languages. Hindi and Arabic, while spoken by mostly one billion within the world, consider as low-resource languages and experience a shortfall of labelled data, and properly advanced NLP tools compared to English. Such an imbalance engenders a discernible digital divide, significantly impeding both access to information and the trajectory of technological development for these linguistic communities. Conversely, within this disparity lies substantial untapped potential, given the recognized utility of NER in facilitating access to information, preserving cultural heritage through digital archives, and accommodating the development of inclusive language technologies. Consequently, this comparative study recognizes this pivotal juncture, positing that the advancement of NER for Hindi and Arabic should be considered not merely an academic endeavour, but rather an instrumental step toward fostering a more equitable and interconnected global digital landscape.

This research examines NER in Hindi and Arabic languages that still represent significant uncharted territory within the scope of NLP research, especially against resource-rich languages such as English. Various reasons warrant language selection. Firstly, both Hindi and Arabic bolster a larger population of users. Hindi, a language primarily spoken in India, is among the most-spoken languages in the world, with more than 500 million speakers [12]. Likewise, Arabic is the official language of 22 countries located in the Middle East and North Africa; more than 400 million people speak it over the world [12]. Secondly, the digital footprint of both languages is now growing rapidly, as internet access improves combined with the proliferation of the growth of digital

content uploaded in these regions. The raise in online content in Hindi and Arabic also demands effective NLP tools, such as advanced NER systems, to glean valuable information from this ever-growing reservoir of linguistic data.

Besides their demographic and digital significance, morphologically rich languages (MRLs) like Arabic and Hindi bring significant complications for computational processing and language acquisition. Arabic's complex root-and-pattern morphology and high degree of ambiguity make it particularly hard for NLP systems [13]. Likewise, Hindi has greater morphological complexity compared to English, particularly in its case marking and inflectional systems, along with issues like word sense disambiguation and polysemy. This complexity poses challenges for machine translation and information retrieval [14][15]. This linguistic complexity provides a significant challenge to automatic linguistic analysis and the NER task. A rich system of prefixes, suffixes, and infixes drastically alters the meaning and grammatical function of the word. For example, a single verb root in Arabic might give rise to several forms for conjugation, based mainly on tense, aspect, mood, voice, gender, and number, which lead to numerous surface forms [13]. Similarly, Hindi has a complex system of noun declensions and verb conjugations [15]; thus, this complexity further enriches the task of entity boundary detection and classification.

In addition to other issues, a variation in the script brings another level of difficulty to NER. In contrast to English, where the Latin alphabet is used, Hindi employs the Devanagari script, characterized by a conjunct consonant system in which two or more consonants may combine into one character [14][15]. This may present some challenges in automated word segmentation and identification of entity boundaries. Whenever it comes to Arabic script, it presents unique challenges in typography and word recognition due to its right-to-left writing direction, connected letters, and diacritical marks. Diacritics play a crucial role in Arabic text usability, affecting homogeneity, clarity, and legibility, which are frequently omitted in written text but able to substantially alter the meaning of a word [13][17]. Such script-specific characteristics give rise to the challenge of creating precise preprocessing engagements as well as models that handle such orthographic discrepancies when seeking the facilitation of accurate NER for Hindi and Arabic. These linguistic and orthographical characteristics add immense complexity to the proper construction of NER systems for Hindi and Arabic. This research aims to address these complexities and perform a comparative analysis based on their impact on NER performance.

Our study aims to compare Hindi and Arabic to fill these gaps in the literature and contribute to the growth of NER task in NLP for linguistically heterogeneous languages. First, to rigorously analyze and compare the structural and grammatical features in Hindi and Arabic that are of concern in NER tasks specifically pointing to morphological complexities, syntactic variation, and script-specific characteristics that may affect NER performance [28]. Second, to provide detailed descriptions of linguistic challenges faced while performing NER for Hindi and Arabic, pointing out possible problems that distinguishes one language from another. This research seeks promising directions for future investigation that will improve NER performance in Hindi and Arabic.

Following on from our objectives, the core of this study pillared on several key questions that we aim to unpack:

- What are the primary structural and grammatical differences between Hindi and Arabic that play a substantial influence on NER performance?
- What are the unique linguistic challenges and potential pitfalls that might be encountered when performing NER in Hindi and Arabic? how do these challenges differ in nature and severity between the two languages?
- Looking ahead, what are the most promising future research directions for advancing the SOTA in NER for both Hindi and Arabic?

This paper is organized as follows: first we give a brief review of the related work done (Section 2). Next, we present the detailed background of Hindi language, Arabic Language, NER task and existing NER approaches (rule-based, ML-based and hybrid-based approaches) in NLP landscape (Section 3). Next, we discuss the main linguistical challenges in Hindi and Arabic that have high impact on NER (Section 4). And last, we wrap up by summarizing our findings and suggesting where future research efforts might be best directed to further advance NER for these languages (Section 5 and 6).

2 Literature Review

While the landscape of Named Entity Recognition (NER) research encompasses a substantial body of work dedicated to individual languages, including both Hindi and Arabic, a notable lacuna exists concerning direct comparative analyses of the linguistic challenges they present within this domain. Numerous studies have independently explored the intricacies of NER within Hindi and, separately, within Arabic, leading to significant advancements in language-specific methodologies and resource development. However, the explicit comparison of their inherent linguistic characteristics and the consequential impact on NER performance remains a relatively uncharted territory. This absence of direct comparative investigation underscores both the novelty and the potential significance of the current study, which seeks to bridge this gap by directly examining the commonalities and divergences in the linguistic hurdles encountered when performing NER in these two typologically distinct languages. This endeavour should not be interpreted as a critique of existing research, but rather as a timely opportunity to synthesize existing knowledge from parallel linguistic contexts to foster a more nuanced understanding of cross-lingual NER challenges.

Given this absence of direct comparative investigations, this literature review adopts a strategic approach, synthesizing insights from two distinct yet relevant streams of research: those focusing explicitly on NER for Hindi and those dedicated to NER for Arabic. By examining the methodologies, challenges, and findings reported within each language-specific body of literature, we aim to construct a comparative framework that illuminates the shared and unique linguistic obstacles influencing NER performance.

The subsequent sections of our review are designed to provide a thorough understanding of the current state of research. At first, we will examine the existing literature pertaining to NER in Hindi, followed by a targeted analysis of research focusing on NER in Arabic. Therefore, we will integrate the findings from these individual language analyses by making connections to relevant literature in theoretical linguistics that elucidates the structural and grammatical features of both Hindi and Arabic. This integrated approach will pave the way for a more informed discussion of the methodological approaches employed in NER for these languages and the potential avenues for future research.

2.1 Named Entity Recognition for Hindi

In the last couple of decades, Named Entity Recognition research targeted at Hindi languages has made great strides, echoing broader advancements in the field of NLP. The traditional approaches relied on hand-crafted linguistic features and gazetteers through rule-based systems. With the ever-increasing popularity of machine learning techniques, supervised learning, especially Conditional Random Fields (CRFs), emerged as a dominant model for Hindi NER since their abilities with sequence labelling tasks demonstrated their efficiency for Hindi NER [18][19][20][21][23], obtaining competitive results by incorporating a rich set of features such as morphological, contextual, and dictionary-based information. More recently, deep learning architectures, including Bi-LSTM and RNN-LSTM, and transformer-based models (e.g. BERT and XLM-R), have been increasingly explored for Hindi NER, demonstrating their ability to learn intricate patterns from raw text and achieve state-of-the-art performance [22][24][25][26].

2.2 Linguistic Challenges in Hindi NER

The relatively rich morphology of Hindi shows agglutination and inflectional processes [16]. This combination leads to the formation of word forms, which may be very complex because of combining several morphemes together into one and makes it difficult for one to identify the root form and extract the relevant features. For example, inflectional morphology, especially nominal and verbal inflections on case, number, gender, and tense, adds further complication to NER. Therefore, the authors have devised ways to incorporate morphological analyzers and stemmers into their NER systems. The [27] paper described the tool of morphological analyzer and generator with a paradigm-based approach. In other instances, rules have been explored to generate inflectional and derivational words from a base/root word to model the complexities posed [29] and detect spelling error [32] from morphology. The authors of [30][31][33] addressed critical issues and challenges during the anaphora resolution, limitation of availability of labelled corpora for the Hindi language were discussed along with emphasis on its relevance for various NLP tasks, and unique features of Hindi pronouns were highlighted that create challenges for anaphora resolution. Also, the becoming of challenges like

resource scarcity and script-related issues have been addressed in such studies like [34] paper.

2.3 Named Entity Recognition for Arabic

Named Entity Recognition task for Arabic language comes with unique challenges due to the Arabic's complex morphology and orthography. In fact, several approaches, including rule-based, machine learning, hybrid methods, and deep learning, have been developed to address these challenges and improve performance. Rule-based systems, such as NERA [38], have handled specific challenges in Arabic NER, including the recognition of 10 important named entity categories and the development of resources like whitelists and grammars. In statistical approaches, conditional random fields (CRFs) with optimized feature sets have shown promising results on the ANERcorp dataset [35]. Additionally, the [42] paper evaluated CRF, SVM, and Stochastic Gradient Descent (SGD) for scientific text Arabic NER, with SGD outperforming the others. A hybrid approach combining rule-based and machine learning techniques [41] has demonstrated superior performance compared to individual methods [39][36][40]. Recently, artificial neural networks (ANNs), Bi-LSTM, and transformer-based approaches have also been successfully applied to Arabic NER [37][44][45], and achieving state-of-the-art performance. These diverse approaches highlight the ongoing efforts to improve Arabic NER performance and overcome language-specific obstacles.

2.4 Linguistic Challenges in Arabic NER

Arabic Named Entity Recognition experiences such complicated challenges due to the language's intrinsic features. For instance, the lack of capitalization, which is significant in English NER, makes Arabic NER inherently more difficult [46]. Additionally, Arabic's highly morphological nature and the ambiguity between proper nouns and common nouns/adjectives further complicate the task [46][47][40]. As the authors of [49] highlighted that the existence of Classical, Modern Standard, and various colloquial dialects presents additional layers of complexity in NER, since models must be trained to recognize entities across these forms. That being so, various approaches have been employed to address these challenges, including rule-based systems [48] and corpus-based methods [46]. The [43][44] papers proposed ML-based techniques to overcome Arabic's inflection, ambiguity issues and morphological, and spelling challenges.

3 Background

Language is a basic structure of communication among humans, an organ that allows people to convey and produce meanings of messages through structured and conventional means [56]. This multi-layered system of communication includes verbal and written means, often supplemented by both the formal and informal non-verbal cues of communication meant to help establish understanding and expression. Language consists, on their most basic level, a set of spoken or written symbols, each associated with

a specific set of phonemes, the smallest unit of sound that serves as a distinguishing feature in meaning [57]. The diversity of languages around the world is like other properties of languages; it is characterized by distinct phoneme inventories and phonological systems, which result from such diverse causes as culture, geography, and social context.

3.1 HINDI

Hindi is an Indo-Aryan language that is the second lingua franca of India, spoken by millions of people from a considerable number of states, such as Madhya Pradesh, Delhi, Uttar Pradesh, Uttarakhand, Bihar, Rajasthan, Chhattisgarh, Haryana, Himachal Pradesh, and Jharkhand. Belonging to a more extensive family of Indo-European languages, Hindi is developed through a stupendous range of factors, historical, cultural, and social, entailing its role as an instrument of communication among members of the language-speaking community and providing linguistic identification for them [59]. The dynamism of Hindi is ensured by the dialectal fabric marking its linguistic landscape, wherein each dialect persists in its unique cultural and socio-historical context reflecting its region of occurrence.

The fourth most spoken language in the world, Hindi occupies a major position in the language scene. According to the 2011 census [60], Hindi has 528 million speakers reportedly situated in India, establishing its demographic importance. The lexicon and phonology, while clearly under strong influence from Sanskrit and Persian, have seen notable evolution in Hindi, which makes use of the Devanagari script and has managed to remain strongly attached to its Sanskrit roots. Its importance reaches beyond the borders of India and becomes synonymous with a worldwide presence, second only to Mandarin Chinese, Spanish, and English [58].

3.2 ARABIC

Arabic, the language of the Arab world, happens to be one of the top ten most common languages globally [61], around 420 million native speakers give it a head start. This is also corroborated by the existence of over 25 different dialects which contribute to the overall linguistic diversity among Arabic speakers [63]. Arabic is set apart from many other languages by the fact that it has such a complex morphology, extremely sophisticated systems of inflection and derivation. Such a linguistic framework is sufficient to promote nuanced expression but significant challenges for learners and scholars.

Arabic has several layers, classified broadly as either Classical Arabic or Modern Standard Arabic (MSA), plus a number of regional dialects. Coupled with MSA and dialects of Arabic [63], it sets up a diglossia context with substantial influence on the Arabic sociolinguistics. Since independence in India (after 1947), the government has shown renewed interest in Arabic [62], and several Arabic language departments have set up at various universities and colleges around the country.

3.3 NAMED ENTITY RECOGNITION TASK

NER task refers to identifying and classifying character sequences corresponding to noun phrases against general text. Named entities are specific real-world objects that denote an individual, organization, or place. These entities are classified into different classes like (Person, Organization, Location, and so forth,) as shown in Table 1, to help build a structure for information extraction from unstructured text.

Table 1. Entity annotations of the NER dataset CoNLL2003 [55].

Entity	Description
PERSON	Individual human beings, whether real or fictional.
NORP	Groups defined by nationality, religion, or political affiliation.
FAC	Physical structures such as buildings, airports, highways, and bridges.
ORG	Organizations including companies, agencies, and institutions.
GPE	Geopolitical entities like countries, cities, and towens.
LOC	None-GPE locations such as mountain ranges and water bodies.
PRODUCT	Items such as vehicles, weapons, and foods.
EVENT	Named occurrences like hurricanes, battles, and sports events.
WORK_OF_ART	Titles of creative works including books and songs.
LAW	Legal documents that have been enacted into law.
LANGUAGE	Recognized languages.
DATE	Specific dates or time periods, either absolute or relative.
TIME	Time expressions referring to durations shorter than a day.
PERCENT	Percentage figures, including the "%" symbol.
MONEY	Monetary amounts, specifying the currency unit.
QUANTITY	Measurements indicating attributes like weight or distance.
ORDINAL	Terms denoting order, such as "first" or "second."
CARDINAL	Numerals that are not classified under other types.

NER takes a collection of documents/texts and automatically identifies and labels named entities within them. Here's the process:

- Step 1. It starts feeding NER system with a collection of documents/texts in various formats (e.g. .pdf, .html, .docx). These texts serve as the input for the annotation process.
- Step 2. The input texts go through a cleaning and preparation stage to make them easier to analyze. This involves:

- Step 2.1. Input Validation: The system first checks if the input text is in a language it's designed to handle.
- Step 2.2. Tokenization: Each document is broken down into individual words, symbols, and punctuation marks.
- Step 2.3. Stop Word Removal: Common words that don't contribute much to entity recognition are removed.
- Step 2.4. Stemming: The words within given documents are reduced to their root form.
- Step 2.5. Morphological Analysis: In this phase of pre-processing, the system analyses the internal structure of words, which is especially important for languages like Hindi and Arabic that have complex word formation rules.

– Step 3. This step represents the core of NER system, where entities are identified and classified:

- Step 3.1. Feature Extraction: The system looks for features in the text that suggest a word might be part of a named entity. These features can be things like capitalization, surrounding words, or specific word endings.
- Step 3.2. Entity Identification and Classification: Using a pre-defined algorithm (which could be rule-based, ML-based models, or a combination of both), the system identifies and labels the named entities based on the extracted features.

– Step 4. The system then outputs the original documents with the identified named entities automatically annotated. Fig. 1 represents the named entity recognition system.

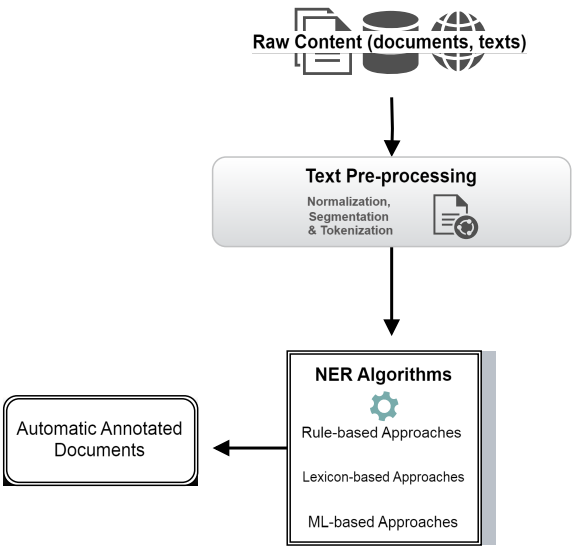

Fig. 1. Named Entity Recognition System

For example:

Hindi

Input text:

हिंदी और अरबी दोनों भाषाएँ अपनी समृद्ध शब्दावली और विविध रचनाओं के लिए जानी जाती हैं

Tokenization:

हिंदी | और | अरबी | दोनों | भाषाएँ | अपनी | समृद्ध | शब्दावली | और | विविध | रचनाओं | के | लिए | जानी | जाती | हैं |

Stemmer:

हिंदी और अरबी दोनों भाषा अपन समृद्ध शब्दावल और विविध रचना के लिए जान जात हैं

Morph Analysis & POS tagging output:

हिंदी /NNP - और /CC - अरबी /NNP - दोनों /DEM - भाषाएँ /NN - अपनी /PRP - समृद्ध /JJ - शब्दावली /NN - और /CC - विविध /JJ - रचनाओं /NN - के /PSP - लिए /PREP - जानी /JJ - जाती /VM - हैं /VM

Arabic

Input text:
" تتمتع اللغتان العربية والهندية بثراء لغويّ فريد من حيث المفردات والتراكيب "

Tokenization:
تتمتع | اللغتان | العربية | و | الهندية | بثراء | لغويّ | فريد | من | حيث | المفردات | و | التراكيب

Stemmer:
تمتع | لغة | عرب | و | هند | ثراء | لغو | فريد | من | حيث | مفرد | و | تركيب

Morph Analysis & POS tagging output:
تتمتع /VERB - اللغتان /NOUN - العربية /ADJ - و /CONJ - الهندية /ADJ - بثراء /NOUN - و /CONJ - التراكيب /NOUN - من /PREP - حيث /CONJ - المفردات /NOUN - و /CONJ - فريد /ADJ - لغويّ /ADJ

3.4 Existing NER Approach

NER methodologies are primarily categorized into three main approaches:

Rule-based approach

Rule-based NER approaches rely on proposed rules and lexical patterns devised by domain experts [4]. These systems other than that apply hand-constructed semantic and syntactic rules for recognition of the entities [50] [51] [52] [53]. However, these methods are often driven by the specifics of the area in demand and thus are limited in their generalization, normally resulting in high precision but low recall. One example is in the biomedical domain, where researchers in 2005 [54], used pre-processed synonym dictionaries to identify protein mentions and possible genes in texts. The development of large rule sets to cater to various complexities of expressions is an ardent task quite daunting. Yet still, for those tasks that regular methods have been applied on, the rule-based methods rule out high precision but are labelled with low recall.

To counter these shortcomings, an increase in machine learning-based approaches to NLP is under foot, which applies itself to automatically learn patterns from extensive databases and offers improved scalability and adaptability to multiple domains.

Supervised approach

Supervised learning is a big evolution in NER systems, as it relies on feature engineering. Basically, it means that it builds an annotated dataset upon which features to represent each data sample will be carefully extracted. The machine learning algorithms then learn from these features the higher-level patterns and apply the learned methods on unseen data, deciding which category or "bucket" a given record belongs to in terms of classification tasks. Common supervised learning algorithms for NER include:

- Decision Trees: These are Prediction models which makes predictions by deriving a simple set of decision rules from data features.
- Hidden Markov Models (HMM): A statistical model representing the probability distribution of sequences, are commonly used for sequential data, e.g., text.
- Support Vector Machine (SVM): A very powerful classification algorithm that finds the boundary separating the entity classes in a very high dimensional feature space.
- Conditional Random Fields (CRF): A probabilistic model for structured prediction tasks such as identifying a sequence of entities in the text.

It is known that supervised NER systems, including those that use deep learning, require a large quantity of annotated data in order to be trained effectively. Meanwhile, such annotation is both tedious and costly. For example, the cost of developing a high-quality training dataset may run very high, depending on the complexity of the data and the requirements for annotation. Given these challenges, unsupervised approaches of Active Learning are proposed.

Unsupervised approach

As mentioned before, the annotation of data for Named Entity Recognition (NER) is quite resource-consuming on account of requirement of multi-stage pipelines and highly trained annotators. Active learning provides a genuine solution within the ambit of actively selecting the most informative data samples to be labelled. Where most supervised learning approaches draw examples randomly and assign labels, active learning algorithms intelligently select the examples that should be labelled in a way that maximizes the learning potential. The corner objective of active learning is to identify that small subset of data containing the most information to be thus utilized in the labelling process so that there is a degree of cut down on the annotation work. The main limitation of this approach lies in the determination of which data points bear the most information, followed by ways of getting the active learner to know this beforehand from existing knowledge.

Applications of NER

The significance of NER finds its way through a wide array of NLP tasks, strongly transforming the methods through which textual data can be processed and knowledge can be inferred. Within information extraction, the role of NER is integral since it facilitates the population of databases and knowledge graphs with structured information abstracted from huge corpora of text [2][5][6]. In question answering, NER plays a vital role by identifying key entities on which information is being requested by the user, thus allowing for the provision of more precise and relevant answers [7]. In machine translation, NER can help preserve meaning in the translated text, especially via its often-complicated translation of named entities that may have very different characteristics from a linguistic and cultural standpoint across various languages [2]. These examples represent only a small portion of the extensive potential applications of Named Entity Recognition (NER), whose scope encompasses numerous fields such as text summarization, sentiment analysis.

4 Linguistic Construction of Hindi and Arabic

4.1 Agglutinative Nature

Hindi

Morphology is a basic field of linguistics that studies how words can be organized internally and formed into various contexts and constructions. There are mainly two subfields to it: derivational morphology, which is the process of forming new words by means of prefixes and suffixes, and inflectional morphology, which works on the modifications of words in order to mark grammatical features like tense, number, or case. Morphological analysis and generation are thus essential to Natural Language Processing (NLP) task like NER. Morphological analysis refers to the process of breaking a word into its root or stem and affixes. The core of a word is the basis for interpreting the meaning of the word and its syntactic role in a sentence. Such analysis acquires

further importance in languages like Hindi, where, like most Indo-Aryan languages, a rich heritage of inflectional morphology exists to present many shades of meaning. The complexity of such morphological systems implies their robust analysis is needed for the proper deciphering and processing in this NLP domain itself.

Hindi has an abundant inflectional morphology especially in nouns, pronouns, and adjectives. Suffixes will indicate case, gender, and number. Words are combined into compound nouns. Such agglutinative nature results in numerous forms of a singular lemma. For instance, the masculine noun 'horse' is written thus 'घोड़ा' in direct singular, while its oblique singular is 'घोड़', and is mostly used together with post-positions to denote other complements, as in 'घोड़ को' (dative singular). For plural forms, the equation runs 'घोड़' in direct case and 'घोड़_' in oblique case. Hindi adjectives can be inflected as well as uninflected. Uninflected adjectives stay unchanged before all nouns and in all the circumstances, just as with English adjectives (e.g., 'सदर', 'beautiful'). All inflected adjectives generally end in 'आ'(e.g., 'काला', 'black'), and their inflection depends upon gender and case of the noun they qualify (e.g., for the masculine noun 'काला घोड़ा', 'black horse', 'काल घोड़', 'black horses'; or for the feminine noun in 'काली बिल्ली', 'black cat', and 'काली बिल्ल्लयाँ', 'black cats'). The word 'घर' (ghar, "house") may take different forms like "घरों" (gharõ, "houses"-plural), "घर से" (ghar se, "from the house"), "घर में" (ghar me, "in the house"), and "घरवाला" (gharwala, "house owner" or "resident"). In NER, it means that this system must be able to classify 'दिल्ली' (dillī, "Delhi") and 'दिल्ली में' (dillī mẽ, "in Delhi") [64] as referring to the same entities. Although almost all English nouns have only 7 or 8 inflected forms, Hindi nouns can display as many as 40 or more distinct inflected forms, which is suggestive of the rich inflectional morphology of Hindi.

Furthermore, derivational morphology refers to the processes by which new lexemes are formed out of existing ones, mainly through the presentation of existing roots with affixes. In most cases, along with the change in some degree of meaning, this process tends to change the syntactic category of the word. In Hindi, for instance, "म" (ma) and "मेरा" (mera) develops into "ममेरा" (mamera) [64], putting a pronoun in the adjectival category.

Arabic

Arabic has a very complex and fascinating morphological structure too that is non-concatenative in nature and is built on roots and patterns. This highlights that a language like Arabic revolves around a root that can take a lot of derived forms which makes tasks such as NER extremely difficult. There are more through discussion examples with reference later in this paper. Central to the idea of Arabic morphology is the idea of root, which is most often tri-literal or sometimes quadrilateral geniuses of stem that supposedly is able to retain the basic meaning of any stem built onto it. That root has words with analogous sense formed by vowel and sometimes additional consonant slots filled in as well. An example could be taken from "ك-ت-ب" (k-t-b) which transfers a

basic idea of to write. When its base is established, one can then go ahead and build off it by applying different patterns on its base:

— كَتَبَ (kataba): "He wrote" (verb, perfective aspect, active voice)
— كُتِبَ (kutiba): "It was written" (verb, perfective aspect, passive voice)
— يَكْتُبُ (yaktubu): "He writes" (verb, imperfective aspect, active voice)
— يُكْتَبُ (yuktabu): "It is being written" (verb, imperfective aspect, passive voice)
— كِتَاب (kitāb): "Book" (noun)
— كُتُب (kutub): "Books" (noun, plural)
— كِتَابَة (kitāba): "Writing" (verbal noun/infinitive)
— كَاتِب (kātib): "Writer" (noun, active participle)
— مَكْتُوب (maktūb): "Written" (adjective, passive participle)
— مَكْتَب (maktab): "Office/Desk" (noun)
— مَكْتَبَة (maktaba): "Library/Bookstore" (noun)

In addition to this, Arabic employs a system of extensive affixation using prefixes, suffixes, and circumfixes that render alterations in meaning and/or to the grammatical function of words. Clitics, such as prepositions and conjunctions can also be attached to words. It doesn't end there. The combination of clitics with a morpheme has the rather strange effect of rendering a simple concept into many convoluted forms. Some examples of these as clitics are: the definite article "ال" (al-, "the") that is affixed as a prefix to nouns; for example, the word "كتاب" (kitāb, "book") becomes "الكتاب" (al-kitāb, "(the) book"). Then, there are prepositions like "بِـ" (bi-, "with/in"), "لِـ" (li-, "to/for"), "مِن" (min, "from"), and conjunctions like "وَ" (wa, "and") which may be clitics and attach to words, as shown in Table 2:

— بِالكِتَاب (bil-kitābi): "With the book"
— لِلكَاتِب (lil-kātibi): "To the writer"
— وَالكِتَاب (wal-kitābu): "And the book"

Table 2. Case Makers examples in Hindi and Arabic

Case	Case Maker Hindi	Hindi Example	Case Maker Arabic	Arabic Example	Translation
NOMINATIVE	Void	वह बाजार के लिए चला गया।	-	ذهب إلى السوق	He went to the market
ERGATIVE	ने	लड़के ने एक कलम खरीदा।	Usually unmarked in the ending, except for tanween	اشترى الولدُ قلما	The boy bought a pen.

			in indefinite nouns		
DATIVE	को	उसने लड़के को कलम दिया।	لِ or إلى	أعطى القلمَ للولدِ. / أعطى القلمَ إلى الولد.	He gave the pen to the boy.
ACCUASITIVE	को	राम ने उस को मारा।	ـه	أوقعهُ رام	Ram beat him.
ABLATIVE	से	लड़के से पुस्तक लिखी गई।	ْمِن	الكِتابُ كُتِبَ من قِبَلِ الولد.	The book was written by the boy.
LOCATIVE	में	छात्र कक्षा में है।	في	الطالبُ في الصفّ.	The student is in the classroom.
INSTRUMENTAL	से	लड़की ने कलम से लिखा।	بِـ	كتبت البنتُ بالقلم.	The girl wrote with the pen.

4.2 No Capitalization

Hindi

The Hindi language, comprised in the Devanagari script, has not adopted the practice of capitalizing the letters of the first word in a sentence and proper nouns. This absence deprives Hindi semiconductors of a valuable orthographic cue that NER systems usually depend on in English. For instance, in English, "Apple" (the company) is clearly different from "apple" (the fruit) owing to capitalization. In Hindi scripts, both of them are given similar pronunciation, and a script "सेब" (seb) will be used and the NER system would have to depend on surrounding text to locate the location between them.

Arabic

In the case of Hindi, the Arabic language is also written without using capitalization [65]. This implies that proper nouns, which usually comprise most of the named entities, are visually indistinguishable from other nouns as the first word of the sentence. For example, the name "خالد" (Khālid) and "كتاب" (kitāb, meaning 'book') when used do not differ in form on account of capitalization. The NER system has to deduce the kind of entity from other linguistic features and contextual methods.

4.3 Spelling Variants

Hindi

Hindi has its characters altered as a result of dialects, colloquial or even different accepted pronunciations of Hindi words. As an instance, for the word 'World,' it could be written in different forms, including "दुनिया" (duniyā) and "दूनीया" (dūnīyā). Such variation necessitates an adequate level of strength within the NER system. There is a need for the NER system to be able to distinguish between the two spelling variants as variants that refer to the same entity. So does "मुंबई" (mumbaī), or "बंबई" (bambahī) – both are used for Mumbai.

The body of Hindi writing and spelling data resources relies largely on the guidelines provided by the Government of India on Standard Modern Hindi. On the other hand, there are various standardised and legible variants that are used by non-Indian speakers of Hindi around the world. While there is a formal system of romanisation of Hindi language, its speakers are rather loose in informal romanisation. This use is largely of a phonetic character in view of the various regional accents seen in India.

- पूजा (worship) → Puja, Poojae, Pujaa, Puuja
- करता (doing) → Kartha, Karta, Kaarta, Kaerta, Kearta
- दोस्त (dost, "friend") → Dost, Doste, Doost, Doust, Daust
- प्यार (pyār, "love") → Pyar, Pyaar, Piar, Peyar, Pyaer
- आशीष (blessings) → Ashish, Ashheesh, Asheesh, Ashhish, Aashish
- राही (traveller) → Raahi, Rahi, Rahee, Raahee

This level of orthographic inconsistency raises problems for NER.

Arabic

Arabic does not only possess regional dialects but actually a different spelling execution of terms and in some cases, there is an attached sound but there are no diacritics on the spelling. For instance, the name "عبد الله" ('Abd Allāh) is spelled differently across the world, see Table 3. Varied pronunciations such as these pose a threat especially when Arabic words or names are written in other languages due to the impact a dialect can have over the word. This creates problems for NER, where different spellings ought to belong to one entity. Muhammad could also have a different pronunciation where two vowels are differently used or a shadda is placed above it in the Arabic language "محمد" (Muḥammad).

Table 3. Spelling Variants examples in Hindi and Arabic.

Hindi Word	Hindi Variants	Arabic Word	Arabic Variants
दिल्ली (Delhi)	दिल्ली, देहली, दिल्ही	صنعاء (Sanaa)	صنعا, صنعاء
भारत (India)	भारत, ভারতবর্ষ, भारता, भरत, भारत	الهند (India)	لهند, ألهند, لِهند, ألْهِنْد

मुंबई (Mumbai)	मुंबई, बंबई, मुम्बई, बॉम्बे	دبي (Dubai)	دُبيّ, دُبَي, دِبِي
रविवार (Sunday)	रविवार, रवीवार	الأحد (Sunday)	يوم الأحد, يوم الاحد, الأحد

4.4 Lack of Resources

Hindi

Even if the scenario is changing, Hindi still lacks in terms of number of large annotated high-quality corpora that are specifically suited for NER. Such a situation is interspersed with difficulty in developing robust machine learning models. This situation is made worse when annotation is lapses in basic standards.

Sometimes researchers have to use smaller corpora or use approaches such as transfer learning or unsupervised approaches. There have been attempts to develop datasets such as FIRE shared tasks plus the IIIT-H Hindi NER corpus, but the need is great for more such resources.

Arabic

Like Hindi, Arabic NER studies are also affected by a lack of large sized or well-annotated corpuses. There are some resources that are present such as the ANERCorp but these tend to be fewer in number and may not be comprehensive enough to include all needed types or areas of entities as compared to the resources for English. This was however an impetus for creating methods that can be learned with lesser data or utilised using multilingual information.

4.5 No Short Vowels

Hindi

Hindi's Devanagari has markings for vowels (matras) but it should be noted that short vowels assuming a seat in consonant letters do not require any ‘matra’ to change that. This is not exactly the case with Arabic languages as diacritics are completely absent at times and less endowed with too few regards. Nevertheless, short vowels in Hindi would be also problematic with regard to disambiguation even in case where one would like to make use of more words type or named entities but not frequently. An instance is the word “कल” (kal) which could mean yesterday if some vowel sounds were to be envisaged or a machine part.

Arabic

In contemporary Arabic writings, short vowels including diacritical markers which are requisite for orthography as well as unambiguous representation are frequently ignored. Consequentially, this creates a bad effect in overall clarity since one form in writing can represent several meanings. E.g. The word "علم" (ʿlm) could means, without diacritics, 'flag', 'science' and 'he knew' depending on the consonants used. This ambiguity is difficult for Arabic NER for the system has to predict the correct meaning and the right kind of entity that the context refers to.

5 Discussion

This comparative analysis of linguistic issues in developing NER systems for Hindi and Arabic reveals both a number of commonalities and one or two very striking differences in the obstacles faced by NER in these morphologically rich and typologically diverse languages.

5.1 Commonalities in Hindi and Arabic:

Hindi and Arabic are characterized as very challenging for NER because they share similar features of complex morphological systems, lack of capitalization, abundant spelling variations, and an overall lack of large-scale, annotated corpora.

- Morphological Challenges: The rich morphology of both languages, albeit through different mechanisms, poses one of the primary challenges. In Hindi, agglutinative systems and Arabic follow the root-and-pattern method, which leads to unequal inflected forms from a single lemma. Therefore, it will be an arduous task for the NER systems to successfully identify and map these various surface forms into one underlying entity since sometimes the inflectional paradigms may overlap. Thus, morphological analysis of advanced nature is required in order to sufficiently understand the language intricacies possessed by different language systems.
- Absence of Capitalization: The absence of capitalization as an orthographic cue in Devanagari and Arabic scripts is simply put. Considering the lack of this orthographical clue in Hindi and Arabic, the NER-referenced system will have to base its proper noun identification totally on contextual and lexical features. This finding is in accord with other such findings where their performance in other not to mention Chinese poses many challenges.
- Spelling Variations: Variability is a challenge that each language has to deal with. Whether differences in dialect, verbosity in transcription, or a result of Arabic language constraints in dropping diacritics are to blame, variations will continue. Dispel this ambiguity; hence, different surface forms will now map to a canonical entity representation. For instance, multiple surface forms for "Delhi" in Hindi are список имен (दिल्ली, देहली, दिल्ही) or "Dubai" in Arabic (دبي, دُبَي, دُبيّ). There are some advances at the level of models which have emerged as a viable solution in the recent past, taking character-based neural models

that have attempted to reduce this difficulty by considering the subword structure.

- Resource Scarcity: Hindi and Arabic have relatively fewer large-scale collections of well-annotated corpora - as compared to resource-rich languages such as English. This limits the extent of development and thus an evaluation for data-hungry machine learning models. While resources such as the Arabic Named Entity Corpus (ANERCorp) and the IIIT-H Hindi NER corpus have seen some light, to this day, they aren't sufficient: the small size and scope warrant the applications of semi-supervised, unsupervised, and transfer learning techniques to appropriately exploit unlabelled data that might be in hand, or derive transferable knowledge from resource-rich languages.

5.2 Differences in Linguistic Features

The similarities notwithstanding, a deeper probe unveils significant differences between Hindi and Arabic NERs in the particulars of their linguistic challenges:

- Morphological Operations: While Hindi is an agglutinative language, indulging in suffixation and compounding, Arabic follows a morphological system based on a root-and-pattern model, characterized by non-concatenative forms. The nature of Arabic morphology is considerably more complex, given the vast array of methods through which a single root can interact with prefixes, suffixes, infixes, and internal vowel patterning. Such contrasts warrant varying approaches to their computational treatment, namely that Arabic NER often draws on elaborate morphological analysers like the MADAMIRA toolkit, while Hindi NER leans toward stemmers and suffix stripping.
- Short Vowels and Diacritics: The frequent omission of short vowels (diacritics) in modern Arabic texts adds to the already great level of ambiguity as compared to Hindi. While Hindi's Devanagari script has inherent short vowels in consonant letters, in Arabic, this results in many-to-one correspondence between written forms and their possible pronunciations and meanings; thus, Arabic NER involves advanced context-driven ambiguity resolution methods. Recent studies concerning the integration of diacritization models into Arabic NER pipelines report some encouraging results.
- Word Order: Hindi's relatively flexible SOV word order, whereas Arabic relies more on the VSO word order (though still varies), leads to differing challenges in interpreting syntactic relations and establishing entity boundaries. This difference would impact the design of features dependent on word order and syntax dependencies. Dependency parsing information was beneficial for Hindi NER, whereas for Arabic NER, it depended more on establishing verb-subject-object patterns.
- Challenges Specific to the Script: The Devanagari and Arabic scripts present their challenges in dealing with issues involved with character representation, tokenization, and script-specific features like conjunct consonants in Hindi and connected letter forms in Arabic. This results in specific approaches to the problems requiring

some preprocessing along with maximizing the overview of tokenization and normalization processes particular to each script.

6 Conclusion and Future Directions

In summary, while Hindi and Arabic present serious challenges to a NER system, given their complicated natures, those challenges are different in quite a considerable number of aspects. While they share some similarities-like the lacking capitalization and morphological richness-the particularities in Arabic root-and-pattern morphology, omnipresence of ambiguity due to missing diacritics, and the variations in syntactic structures necessitate different techniques and language resources.

Future research would include:

- Development of even finer morphological analysers and disambiguation processes to accommodate the specific challenges for these languages in the future, especially Arabic with complex morphology and diacritic representations.
- Creation of larger and more comprehensive annotated corpora for Hindi and Arabic NER that can help in building and evaluation of data-driven models.
- Testing out the effectiveness of cross-lingual transfer learning that can make use of resources from the resource-rich languages and thus make up for the lack of data to Hindi and Arabic.
- Investigation into the new neural architectures like transformers and pre-trained language models to tackle the tricky nature of these languages.
- Creating robust strategies for spelling variations, linking slip forms, and enabling canonical entity representations.

This ought, in sum, to lay the foundation for a more accurate and robust NER system for Hindi and Arabic, thereby establishing a way for the progression of NLP applications in those languages and providing value for their large speakers' community in these languages. This research highlights the critical need for investment in language-specific NLP R&D, particularly for lesser-resourced languages. People from developing countries paid on an unprecedented scale to bridge the digital divide and generated potential for language technologies that everyone could enjoy.

Acknowledgment. I would like to express my sincere gratitude to my supervisor, Prof. Dr. Surendra Yadav and co-supervisor, Dr. Nalin Choudhary, for their invaluable guidance, support, and constructive feedback throughout the development of this paper. Their expertise and mentorship have played a crucial role in shaping this work, and I deeply appreciate the time and effort they invested in reviewing and refining my research.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Monica, Munnangi. (2024). 2. A Brief History of Named Entity Recognition. doi: 10.48550/arxiv.2411.05057
2. Oleh, R., Staso., Nazarii, Burak. (2024). 6. Named entity recognition and its role in unstructured data analysis. doi: 10.15276/ict.01.2024.34
3. Shuhe, Wang., Xiaofei, Sun., Xiaoya, Li., Rongbin, Ouyang., Fei, Wu., Tianwei, Zhang., Jiwei, Li., Guoyin, Wang. (2023). 2. GPT-NER: Named Entity Recognition via Large Language Models. arXiv.org, doi: 10.48550/arXiv.2304.10428
4. Kalyani, Pakhale. (2023). 4. Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges. arXiv.org, doi: 10.48550/arxiv.2309.14084
5. Patil, A. V. (2024). Identifying specific details from text to populate databases and generate summaries using named entity recognition. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 08(05), 1-5. https://doi.org/10.55041/ijsrem33111
6. Borui, Zhang. (2024). 3. Getting to Know Named Entity Recognition: Better Information Retrieval. Medical Reference Services Quarterly, doi: 10.1080/02763869.2024.2335139
7. Wongso, R., Meiliana, & Suhartono, D. (2016). A literature review of question answering system using named entity recognition. 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 274-277. https://doi.org/10.1109/icitacee.2016.7892454.
8. Murthy, R., Khapra, M. M., & Bhattacharyya, P. (2018). Improving NER tagging performance in low-resource languages via multilingual learning. ACM Transactions on Asian and Low-Resource Language Information Processing, 18(2), 1-20. https://doi.org/10.1145/3238797.
9. Rahimi, A., Li, Y., & Cohn, T. (2019). Multilingual NER Transfer for Low-resource Languages. ArXiv, abs/1902.00193.
10. Liu, L., Ding, B., Bing, L., Joty, S., Si, L., & Miao, C. (2021). MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 5834-5846. https://doi.org/10.18653/v1/2021.acl-long.453.
11. Sabane, M., Ranade, A., Litake, O., Patil, P., Joshi, R., & Kadam, D. (2023). Enhancing low resource NER using assisting language and transfer learning. 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 1666-1671. https://doi.org/10.1109/icaaic56838.2023.10141204
12. https://www.ethnologue.com/language/
13. Soudi, A., Bosch, A. V., & Neumann, G. (2007). Arabic computational morphology: Knowledge-based and empirical methods. Springer Science & Business Media.
14. Goyal, V., & Lehal, G. S. (2008). Hindi morphological analyzer and generator. 2008 First International Conference on Emerging Trends in Engineering and Technology, 1156-1159. https://doi.org/10.1109/icetet.2008.11.
15. Kaye, A. S., Singh, R., & Agnihotri, R. K. (1999). Hindi morphology. Language, 75(3), 625. https://doi.org/10.2307/417098.
16. Sharma, A.R. (2020). Impact of Morphology on the performance of Search Engine while retrieving information in Hindi Language. International Journal of Advanced Trends in Computer Science and Engineering. https://doi.org/10.30534/ijatcse/2020/02942020.
17. Hssini, M., & Lazrek, A. (2011). Design of Arabic Diacritical Marks. ArXiv, abs/1107.4734.

18. Sharma, R., & Goyal, V. (2011). Name entity recognition systems for Hindi using CRF approach. Communications in Computer and Information Science, 31-35. https://doi.org/10.1007/978-3-642-19403-0_5
19. Chopra, D., Joshi, N., & Mathur, I. (2016). Named entity recognition in Hindi using conditional random fields. Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, 1-6. https://doi.org/10.1145/2905055.2905165
20. Sanglikar, M.A., & Kothari, D.C. (2011). Named Entity Recognition System for Hindi Language: A Hybrid Approach.
21. Li, W., & McCallum, A. (2003). Rapid development of Hindi named entity recognition using conditional random fields and feature induction. ACM Transactions on Asian Language Information Processing, 2(3), 290-294. https://doi.org/10.1145/979872.979879
22. Athavale, V., Bharadwaj, S., Pamecha, M., Prabhu, A., & Shrivastava, M. (2016). Towards Deep Learning in Hindi NER: An approach to tackle the Labelled Data Sparsity. ArXiv, abs/1610.09756.
23. Krishnarao, A. A., Gahlot, H., Srinet, A., & Kushwaha, D. S. (2009). A comparative study of named entity recognition for Hindi using sequential learning algorithms. 2009 IEEE International Advance Computing Conference, 1164-1169. https://doi.org/10.1109/iadcc.2009.4809179
24. Singh, S., Patel, S., Shah, Y., Nargunde, R., & Ramteke, J. (2021). Context-based deep learning approach for named entity recognition in Hindi. 2021 International Conference on Communication information and Computing Technology (ICCICT), 1-5. https://doi.org/10.1109/iccict50803.2021.9510090
25. Shelke, R., & Vanjale, S. (2022). Deep named entity recognition in Hindi using neural networks. Revue d'Intelligence Artificielle, 36(4), 575-580. https://doi.org/10.18280/ria.360409
26. Choure, A. A., Adhao, R. B., & Pachghare, V. K. (2022). NER in Hindi language using transformer Model:XLM-Roberta. 2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), 1-5. https://doi.org/10.1109/icbds53701.2022.9935841
27. R., Mahesh, K., Sinha. (2009). 10. Mining Complex Predicates in Hindi Using a Parallel Hindi-English Corpus. doi: 10.3115/1698239.1698247
28. Goyal, V., & Lehal, G. S. (2008). Hindi morphological analyzer and generator. 2008 First International Conference on Emerging Trends in Engineering and Technology, 1156-1159. https://doi.org/10.1109/icetet.2008.11
29. Kumar, D., Singh, M., & Shukla, S. (2012). FST Based Morphological Analyzer for Hindi Language. ArXiv, abs/1207.5409.
30. LalPal, T., Dutta, K., & Singh, P. (2012). Anaphora resolution in Hindi: Issues and challenges. International Journal of Computer Applications, 42(18), 7-13. https://doi.org/10.5120/5790-8056
31. , A., & Mohana, R. (2016). Anaphora resolution in Hindi: Issues and directions. Indian Journal of Science and Technology, 9(32). https://doi.org/10.17485/ijst/2016/v9i32/100192
32. Lakshmi, M., Gadhikar. (2023). 1. Vyakranly : Hindi Grammar & Spelling Errors Detection and Correction System. doi: 10.1109/ICNTE56631.2023.10146610
33. Joshi, R., Goel, P., & Joshi, R. (2020). Deep Learning for Hindi Text Classification: A Comparison (pp. 94–101). springer. https://doi.org/10.1007/978-3-030-44689-5_9
34. Ahmad, G. I., Singla, J., & Nikita, N. (2019). Review on Sentiment Analysis of Indian Languages with a Special Focus on Code Mixed Indian Languages. 352–356. https://doi.org/10.1109/icactm.2019.8776796

35. El bazi, I., & Laachfoubi, N. (2015). RENA: A named entity recognition system for Arabic. Lecture Notes in Computer Science, 396-404. https://doi.org/10.1007/978-3-319-24033-6_45
36. Meselhi, M. A., Bakr, H. M., Ziedan, I., & Shaalan, K. (2014). A novel hybrid approach to Arabic named entity recognition. Communications in Computer and Information Science, 93-103. https://doi.org/10.1007/978-3-662-45701-6_9
37. Chapram. (2012). Arabic named entity recognition using artificial neural network. Journal of Computer Science, 8(8), 1285-1293. https://doi.org/10.3844/jcssp.2012.1285.1293
38. Shaalan, K., & Raza, H. (2009). NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60(8), 1652-1663. https://doi.org/10.1002/asi.21090
39. Benajiba, Y., Rosso, P., & BenedíRuiz, J. M. (2007). ANERsys: An Arabic named entity recognition system based on maximum entropy. Lecture Notes in Computer Science, 143-153. https://doi.org/10.1007/978-3-540-70939-8_13
40. Shaalan, K., & Oudah, M. (2013). A hybrid approach to Arabic named entity recognition. Journal of Information Science, 40(1), 67-87. https://doi.org/10.1177/0165551513502417
41. Benajiba, Y., Diab, M., & Rosso, P. (2009). Arabic named entity recognition: A feature-driven study. IEEE Transactions on Audio, Speech, and Language Processing, 17(5), 926-934. https://doi.org/10.1109/tasl.2009.2019927
42. Marzouk, N., Nayel, H., & Elsawy, A. (2024). Advancing Arabic scientific text analysis: Evaluating machine learning models for named entity recognition. Benha Journal of Applied Sciences, 9(5), 45-48. https://doi.org/10.21608/bjas.2024.279914.1377
43. Dahan, F., Touir, A., & Mathkour, H. (2015). First order hidden Markov model for automatic Arabic name entity recognition. International Journal of Computer Applications, 123(7), 37-40. https://doi.org/10.5120/ijca2015905397
44. Ahmed, Abdou., Tasneem, Mohsen. (2024). 3. mucAI at WojoodNER 2024: Arabic Named Entity Recognition with Nearest Neighbor Search. doi: 10.48550/arxiv.2408.03652
45. Ait Benali, B., Mihi, S., Ait Mlouk, A., El Bazi, I., & Laachfoubi, N. (2022). Retracted: Arabic named entity recognition in social media based on BiLSTM-CRF using an attention mechanism. Journal of Intelligent & Fuzzy Systems, 42(6), 5427-5436. https://doi.org/10.3233/jifs-211944
46. AlGahtani, S. (2012). Arabic Named Entity Recognition: A Corpus-Based Study.
47. Shaalan, K. (2014). A survey of Arabic named entity recognition and classification. Computational Linguistics, 40(2), 469-510. https://doi.org/10.1162/coli_a_00178
48. Salah, R. E., & Zakaria, L. Q. (2017). Arabic rule-based named entity recognition systems progress and challenges. International Journal on Advanced Science, Engineering and Information Technology, 7(3), 815. https://doi.org/10.18517/ijaseit.7.3.1811
49. Yazeed, Al, Moaiad., Mohammad, Alobed., Mahmoud, Alsakhnini., Alaa, M., Momani. (2024). 1. Challenges in natural Arabic language processing. Edelweiss applied science and technology, doi: 10.55214/25768484.v8i6.3018
50. Habash, N. (2007). Arabic Morphological Representations for Machine Translation (pp. 263–285). springer netherlands. https://doi.org/10.1007/978-1-4020-6046-5_14
51. Goyal, V., & Lehal, G. S. (2008). Hindi Morphological Analyzer and Generator. 1156–1159. https://doi.org/10.1109/icetet.2008.11
52. Haspelmath, M. (2009). An Empirical Test of the Agglutination Hypothesis (pp. 13–29). springer netherlands. https://doi.org/10.1007/978-1-4020-8825-4_2
53. Litake, O., Patil, P., Joshi, R., Sabane, M., & Ranade, A. (2023). Mono Versus Multilingual BERT: A Case Study in Hindi and Marathi Named Entity Recognition (pp. 607–618). springer nature singapore. https://doi.org/10.1007/978-981-19-6088-8_56

54. Daniel, Hanisch, Katrin, Fundel., Heinz-Theodor, Mevissen., Ralf, Zimmer., Juliane, Fluck. (2005). 18. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics, doi: 10.1186/1471-2105-6-S1-S14
55. Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the conll-2003 shared task. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -, 4, 142-147. https://doi.org/10.3115/1119176.1119195
56. Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. Cognition, 141, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016
57. Marslen-Wilson, W. D. (2007). Morphological Processes in language Comprehension (pp. 175–194). oxford university. https://doi.org/10.1093/oxfordhb/9780198568971.013.0011
58. Saini, A. (2023). Hindi Language: Its Journey So Far. https://www.drishtiias.com/blog/hindilanguage-its-journey-so-far
59. Verma, K. (2024). Hindi Language: A Bridge between Cultural Heritage and Society. Mazedan Journal of Language and Literature, 4(1), 14-18.
60. Census (2011) Government of India. http://censusindia.gov.in
61. Abdeen, M. A., AlBouq, S., Elmahalawy, A., & Shehata, S. (2019). A closer look at Arabic text classification. International Journal of Advanced Computer Science and Applications,10(11). https://doi.org/10.14569/ijacsa.2019.0101189
62. Rahman, M. (2014). Arabic in India: Past, Present & Future.
63. Yonatan, Belinkov., Alexander, Magidow., Alberto, Barrón-Cedeño., Avi, Shmidman., Maxim, Romanov. (2019). 1. Studying the history of the Arabic language: language technology and a large-scale historical corpus. doi: 10.1007/S10579-019-09460-W
64. Singh, P., Kore, A., Sugandhi, R., Arya, G., & Jadhav, S. (2013). Hindi Morphological Analysis and Inflection Generator for English to Hindi Translation.
65. Buckwalter, T. (2004). Issues in Arabic orthography and morphology analysis. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages - Semitic '04, 31. https://doi.org/10.3115/1621804.1621813
66. (2022). 4. HiNER: A Large Hindi Named Entity Recognition Dataset. doi: 10.48550/arxiv.2204.13743