Computer Science • 23(1) 2022 https://doi.org/10.7494/csci.2022.23.1.3977

Arti Jain
Divakar Yadav
Anuja Arora
Devendra K. Tayal

NAMED-ENTITY RECOGNITION FOR HINDI LANGUAGE USING CONTEXT PATTERN-BASED MAXIMUM ENTROPY

Abstract *This paper describes a named-entity-recognition (NER) system for the Hindi language that uses two methodologies: an existing baseline maximum entropy-based named-entity (BL-MENE) model, and the proposed context pattern-based MENE (CP-MENE) framework. BL-MENE utilizes several baseline features for the NER task but suffers from inaccurate named-entity (NE) boundary detection, misclassification errors, and the partial recognition of NEs due to certain missing essentials. However, the CP-MENE-based NER task incorporates extensive features and patterns that are set to overcome these problems. In fact, CP-MENE's features include right-boundary, left-boundary, part-of-speech, synonym, gazetteer and relative pronoun features. CP-MENE formulates a kind of recursive relationship for extracting highly ranked NE patterns that are generated through regular expressions via Python@ code. Since the web content of the Hindi language is arising nowadays (especially in health care applications), this work is conducted on the Hindi health data (HHD) corpus (which is readily available from the Kaggle dataset). Our experiments were conducted on four NE categories; namely, Person (PER), Disease (DIS), Consumable (CNS), and Symptom (SMP).*

Keywords context patterns, gazetteer lists, Hindi language, Kaggle dataset, maximum entropy, named-entity recognition, feature extension

Citation Computer Science 23(1) 2022: 81–115

1. Introduction

Natural language processing (NLP) [59] serves as an important field in the computer science area for analyzing a variety of texts. NLP escalates different fields; for example, information retrieval [6], information extraction (IE) [58], machine translation [89], question answering [90], social media [62,63,66], and so on. Among these, IE is defined as a task for extracting structured information from unstructured or semi-structured text that provides valuable information for users. One of the vital sub-tasks of IE is to identify names or named entities (NEs) within a text and classify them into pre-defined categories like person, location, organization, date, time, etc. The process of identifying and classifying named entities is generally known as named-entity recognition (NER) [10, 17, 82]. So far, several NER systems have been successfully deployed for varied languages, such as the English [80, 139], French [50, 94], Spanish [79, 130], Greek [2, 127], Arabic [3, 120], Chinese [16, 40], German [57, 95], and Indian languages [1, 133]. In addition, NER systems are built for various domains (such as newswire [19, 92], financial [39, 140], clinical [136, 141], biomedical [72, 135], etc.) with the use of various techniques. NER techniques are classified as machine-learning, rule-based, and hybrid techniques. Machine learning-based NER needs a large training corpus and considers models such as maximum entropy [111, 116], expectation maximization [104], perceptron [13], naïve Bayes [132], voted co-training [75], decision tree [131], bootstrapping [103], hidden Markov [88], latent semantic analysis [46], support vector machine [25], conditional random field [60], graph [51], clustering [53], deep learning [81, 138], and many more. The rule-based approach for NER [18, 38, 67, 71] looks into a set of rules that are defined by experts for extracting NEs. Such an approach generates pattern sets that have grammatical, syntactical, and orthographical features along with supported dictionaries. However, the manual computation of the rules is quite labor-intensive and extremely costly, and it requires substantial language and domain expertise. The hybrid approach for NER [60, 68, 73, 86, 100] accompanies the combination of two or more NER strategies for a significant improvement in an NER's performance.

Research on NER enlightens several well-developed systems for resource-rich languages like English for domain-independent and domain-specific tasks [15, 74, 85, 128] with high scores regarding the evaluation metrics. However, the construction of an NER system for resource-poor Indian languages (such as Hindi) is quite challenging [5, 34, 61, 123, 125]. The Hindi language uses the Devanagari script [47], and it is an official language of the Indian government (along with English). Hindi is characterized by its inflectional and morphological richness and is a suffix-based language. In the present scenario, the emerging demand of smart health applications that are based on the Hindi language trigger health domain-based named entities for Hindi. Some NEs such as "disease" and "symptom" do not have well-defined NE nomenclatures. However, health-related NEs are comprised of long compound words, short abbreviations, wide variations in spelling, and the cascading of one NE into another. In addition, numerous new health NEs continue to evolve, while there is no complete

dictionary to incorporate health NEs in Hindi. Thus, the NERs in Hindi for the health domain are quite perplexing, which makes it necessary to somehow be solved.

In this paper, the Hindi Health Data (HHD) corpus is considered from Kaggle datasets. An HHD corpus looks for four NE categories; namely, Person (PER), Disease (DIS), Consumable (CNS), and Symptom (SMP) NEs. In order to perform an NER task on the HHD corpus, a context pattern-based maximum entropy named-entity (CP-MENE) framework is proposed that is an extension of an existing baseline MENE (BL-MEME) [11]. BL-MENE serves as a flexible statistical model that has diversified features without looking into hand-coded patterns for the NER task. Still, some NEs are left out from the corpus due to being partially recognized or mistakenly classified by BL-MENE. This happens because the words in isolation within BL-MENE may contribute to the multi-faceted potential for the meaning of an utterance. To solve this, it is necessary to assign the respective word-sense to the use of contextual patterns and differentiate one sense from another. In order to resolve the NE based misclassification errors, CP-MENE discovers new patterns set and attaches an appropriate meaning to Hindi words using abstract behavior of words. CP-MENE also contains an extension of features such as right-boundary, left-boundary, part-of-speech, synonym, gazetteer, and relative pronoun features for resolving the NE's boundary-detection errors. Thus, CP-MENE incorporates a baseline MENE, pattern set, and extensive features for handling the NE boundary-detection and its misclassification errors. The performance of BL-MENE is, thus, reclassified by using the CP-MENE, and a significant improvement in the Hindi NER is achieved by using the CP-MENE over other NER approaches (like an ontology-based NER) [64].

The rest of the paper is organized as follows. Section 2 discusses the literature that is related to the maximum entropy and NER for the Hindi language. Section 3 describes a BL-MENE for an NER in Hindi, which includes a baseline MENE model, its significance, and its related features. Section 4 details a CP-MENE system that is comprised of the context pattern-based MENE model, its significance, and its extensive features. Section 5 illustrates the Hindi NER's experimental results as compared to CP-MENE and other NER approaches. Section 6 concludes the paper.

2. Related literature

This section briefly discusses the literature that is related to the maximum entropy (ME) model and named-entity recognition in the Hindi language. In order to understand the jargon that is used here, researchers are directed to go through Appendix B.

2.1. Maximum entropy

The maximum entropy framework supports widespread applications with a wider timeline (1998–present day), as is detailed in the literature below.

Borthwick et al. [11] applied ME for the NER task at the MUC-7 conference; it was there where the term "MENE" was coined. They discussed the fact that MENEs

can work better when combined with rule-based approaches. Osborne [96] stated that ME exploited categorical features to a greater extent as compared to naïve Bayes and showed the better performance of sentence extraction for document summarization and for question-answering systems. Bender et al. [8] detailed MENEs by using a Gaussian prior for effective smoothing over a larger feature set. Han et al. [52] described ME performance for estimating the probability of English articles (a/an, the, or zero article) for noun phrases that use local-context features; these have proven to be beneficial for speakers of Chinese, Russian, Korean, Japanese, and other languages that do not contain articles. Benajiba et al. [7] built an NER system for the Arabic language (named ANERsys) while using the ME model. Ekbal et al. [32] proposed an ME-based part-of-speech (POS) tagger that outperforms HMM-based taggers for the Bengali language. Saha et al. [111] provided a study on word clustering- and word selection-based feature reduction for NER using the ME classifier. Ekbal et al. [33] proposed an ME-based NER under a multi-objective optimization framework for the resource-constrained Bengali language and achieved an F–measure of 77.11%.

In addition, there are several other NLP applications that use the ME model, such as prepositional phrase attachment [105], parsing [14], word morphology [134], relational extraction [69], phrase reordering [137], word-sense disambiguation [12], phonotactic learning [56], text classification [37], and many more.

ME is characterized by the fact that it maintains a reasonable performance even when there is little available data for the training purpose. Also, ME is highly portable for other languages and domains for sufficient and appropriate training corpora; e.g., ME has achieved good results with the Hindi, Gujarati, and Bangla languages for the FIRE-2013 shared task [43]. In the biomedical domain, Raychaudhuri et al. [106] applied ME to gene ontology tags, where it routinely outperforms the naïve Bayes and k-nearest neighbor methods. Pakhomov [97] discussed ME usage for acronym and abbreviation normalization within medical texts. Mora et al. [87] showed that the ME model is useful for antibody diversity. Asti et al. [4] mentioned that the ME model is beneficial for predicting antigen-antibody affinity.

2.2. NER in Hindi

NERs in the Hindi language represent works that are primarily in the news, health, and web-source domains that are discussed here. Table 1 summarizes the Hindi NER literature according to a chronological order of the years of publication, benchmark datasets, Hindi NER methodologies, NEs, and F-measure values. Along with the commonly stated literature, there are some more related studies such as Saha et al. [114], who have worked on the sports corpus, and Patel et al. [98], who have observed inductive logic programming (ILP) through the WARMR and TILDE-C4.5 methodologies. Several researchers have performed detailed surveys on Hindi NERs; e.g., [65, 70, 99, 101, 119], and [121].

Table 1
NER literature survey for Hindi language

References	Benchmark Dataset	Hindi NER Methodology	Named Entities	F-measure [%]
Cucerzan & Yarowsky, 1999 [21]	MUC-6	EM-style Bootstrapping	PER, LOC	41.70
Li & McCallum, 2006 [83]	TIDES 2003	CRF	PER, LOC, ORG	71.50
Kumar & Bhattacharyya, 2006 [77]	–	MEMM	–	79.1
Ekbal & Bandyopadhyay, 2008 [22]	IJCNLP-08 (NERSSEAL)	ME	PER, LOC, ORG, MSC	82.66
Gali et al., 2008 [41]		Hybrid (CRF + Heuristics)	PER, LOC, ORG, MSN, TME, DEG, ABB, TTP, TTO, NUM, BRD, THT	50.06
Goyal et al., 2008 [45]		CRF		58.85
Kumar & Kiran, 2008 [102]		Hybrid (HMM + CRF)		46.84
Saha et al., 2008 [108]		Hybrid (ME + Rules + Gazetteers)		65.13
Nayan et al., 2008 [93]	English Phonetic Transliteration	Rule-based	PER, LOC, ORG	64.24
Saha et al., 2008 [115]	Dainik Jagran	ME	PER, LOC, ORG, DTE	81.25
Saha et al., 2008 [113]		Hybrid (ME Gazetteers)		83.05
Saha et al., 2008 [109]		Hybrid (ME + Word Cluster + Selection)		79.85
Shishtla et al., 2008 [124]	Varied Sources	CRF	–	45.48
Singh et al., 2008 [126]	CIIL Corpus	ME	PER, LOC, ORG, MSN, TME, DEG, ABB, TTP, TTO, NUM, BRD, THT	73.99

Table 1 cont.

References	Benchmark Dataset	Hindi NER Methodology	Named Entities	F-measure [%]
Ekbal & Bandyopadhyay, 2009 [23]	IJCNLP-08 (NERSSEAL)	CRF	PER, LOC, ORG, MSC	78.29
Ekbal & Bandyopadhyay, 2009 [24]		ME, CRF, SVM		76.35
Hasanuzzaman et al., 2009 [54]		ME		82.66
Krishnarao et al., 2009 [76]		CRF, SVM	PER, LOC, ORG, MSN, TME, DEG, ABB, TTP, TTO, NUM, BRD, THT	47.00
Saha et al., 2009 [117]		Hybrid (ME + Heuristic + Context Patterns + Bootstrap)		96.67
Saha et al., 2009 [111]	Dainik Jagran	Semi-Supervised using ME	PER, LOC, ORG	78.64
Gupta & Arora, 2009 [48]	Web Source	CRF		58.00
Biswas et al. 2010 [9]	SPSAL 2007	Hybrid (ME + HMM)		71.95
Ekbal & Bandyopadhyay, 2010 [25]	IJCNLP-08 (NERSSEAL)	SVM	PER, LOC, ORG, MSC	77.17
Ekbal & Saha, 2010 [27]		ME + GA		89.65
Ekbal & Saha, 2010 [28]				72.60
Hasanuzzaman et al., 2010 [55]				80.46
Gupta & Bhattacharyya, 2010 [49]	Gyaan Nidhi Corpus	NGI + S-MEMM	PER, LOC, ORG, BOK, PLY	82.90
Saha et al., 2010 [112]	Dainik Jagran	SVM	PER, LOC, ORG	83.56

Table 1 cont.

Ekbal & Bandyopadhyay, 2011 [26]	IJCNLP-08 (NERSSEAL)	SVM	PER, LOC, ORG, MSC	80.21
Ekbal et al., 2011 [29]		MOO		92.80
Ekbal & Saha, 2011 [30]		GA		92.20
Srivastava et al., 2011 [129]		Hybrid (CRF + ME + Rules)	PER, LOC, ORG, MSN, TME, DEG, ABB, TTP, TTO, NUM, BRD, THT	82.95
Kumar et al., 2011 [78]	FIRE-2010	Bisecting k-means Clustering	PER, LOC, ORG	71.00
Chopra et al., 2012 [20]	Hindi Newspapers	Hybrid (Heuristics + HMM)	PER, LOC, ORG, TME, MNH, VEH, SPR, RIV, QSM	94.61
Ekbal & Saha, 2012 [31]	IJCNLP-08 (NERSSEAL)	MOO-ME, CRF, SVM	PER, LOC, ORG MSC	93.20
Ekbal et al., 2012 [35]		SVM		89.81
Ekbal et al., 2012 [36]		SVM + CRF		87.87
Sikdar et al., 2012 [125]		Diferential Evolution		88.09
Saha et al., 2012 [110]	Dainik Jagran	Dimensionality Reduction	PER, LOC, ORG	85.31
Gayen & Sarkar, 2013 [42]	ICON 2013	HMM	PER, LOC, ORG, ART, ETN, FCT, LMT, MAT, OSM, PLN, CNT, DST, MNY, QNT, DTE, DAY, PIO, TME, YER	75.20
Saha & Ekbal, 2013 [107]	IJCNLP-08 (NERSSEAL)	MOO	PER, LOC, ORG, MSC	94.66
Sharnagat & Bhattacharyya, 2013 [123]	FIRE-2013	CRF	PER, LOC, ORG,	96.00
Jain et al., 2014 [67]	Hindi Newspapers	Association Rules	PER, LOC, ORG	77.81

Table 1 cont.

References	Benchmark Dataset	Hindi NER Methodology	Named Entities	F-measure [%]
Nanda et al., 2014 [91]	–	–	PER, LOC, ORG, DTE, NUM	62.40
Kaur & Kaur, 2015 [71]	Hindi Newspapers	Hybrid (Rule-based + List lookup)	PER, LOC, ORG, DTE, MNY, MSN, TNS, ANM, DRE	95.77
Athavale et al., 2016 [5]	ICON 2013	Bi-directional RNN-LSTM	PER, LOC, ORG, DIS, ETN, FCT, ART, LIV, LMT, PLN, MAT	77.48
Ekbal et al., 2016 [34]	IJCNLP-08 (NERSSEAL)	Active Learning: SVM, CRF	PER, LOC, ORG, MSC	88.50
Jain et al., 2018 [64]	Kaggle (HHD Corpus)	OntoHindi NER	PER, DIS, CNS, SMP	78.77
Jain & Arora, 2018 [60]		Hybrid (HAL + CRF)		90.69
Jain & Arora, 2018 [61]	Health Tweets		PER, DIS, ORG, CNS	69.41
Sharma et al., 2020 [122]	TDIL	Deep Neural Network	PER, ETN, LOC, ORG, LIV, DST, CNT, PIO, MNH	70.00

The main contributions (RC#) of this research work are stated as follows:

RC1: experiment with baseline features for NER task using BL-MENE.

RC2: explore extensive features and recursive relationship for extracting highly ranked NE patterns.

RC3: propose novel CP-MEME method for dealing with boundary detection, misclassification errors, and partial recognition of NEs.

RC4: compare CP-MENE with respect to other existing Hindi NER approaches.

3. Baseline MENE

To perform the NER task, the baseline maximum entropy-based named-entity method [11] is applied. BL-MENE serves as a statistical framework that is flexible enough for incorporating diversified knowledge resources without looking into hand-coded patterns at all. BL-MENE includes lexical features that are binary-valued (either 0 or 1) in nature; it neither acquires human intrusion nor references to external knowledge sources and serves as an important contributor to the bulk of MENE's power.

3.1. BL-MENE model

The BL-MENE model undergoes a training phase where an HHD training data set is tokenized and each token is labeled with an NE outcome that is based on the following criteria. The set of $N(= 4)$ tags are based on the four considered NE categories (Person, Disease, Consumable, and Symptom), and a particular tag (XXX) from the N tags is in one of four states (Begin [B-XXX], Continue [I-XXX], End [E-XXX], and Unique [U-XXX]) with an additional tag (Other [O]) to indicate an unnamed entity. In other words, an NER problem can be reduced to a problem of assigning one of the possible tags in an overall outcome space (F) that is comprised of a tag-set that has a maximum limit of $4N + 1(= 17)$ tags. Given input sentence (S_{in}) from the HHD corpus, for example, its NE outcome (S_{ne}) is stated as below:
Input (S_{in}):

सेब के सिरके को एक कप पानी में मिलाएं (mix apple cider vinegar into one cup water)

Output (S_{ne}): B-CNS I-CNS E-CNS O O O U-CNS O O

In other words, NE annotation on S_{in} is represented as सेब/B-CNS के/I-CNS सिरके/E-CNS को/O एक/O कप/□ पानी/U-CNS में/O मिलाएं/O

Here, $F_s = S_{ne}$ = B-CNS, I-CNS, E-CNS, O, O, O, U-CNS, O, O, and $F_s \subset F$

F= {B-PER, I-PER, E-PER, B-DIS, I-DIS, E-DIS, B-CNS, I-CNS, E-CNS, B-SMP, I-SMP, E-SMP, U-PER, U-DIS, U-CNS, U-SMP, O}

To generate an outcome space F_s for the test phase, BL-MENE is linked with the feature set (Section 3.3 – BL-MENE Features) and labeled as a training corpus. BL-MENE computes the conditional probability as can be seen in Equation (1).

$$P(t|h) = \frac{\prod_i \alpha_i^{f_i(h,t)}}{z_\alpha(h)} = \frac{\prod_i \alpha_i^{f_i(h,t)}}{\sum_t \prod_i \alpha_i^{f_i(h,t)}} \tag{1}$$

where:

t – NE tag ($t \in F$),
F – overall outcome space,
h – history (condition data to make decision among F from H,
H – space of possible histories,
$P(t/h)$ – probability for any tag t, for every history h,
f_i – i^{th} feature,
α_i – real-valued weight parameter of f_i,
$Z_\alpha(h)$ – normalization factor,
$\prod_i \alpha_i^{f_i(h,t)}$ – product of weightings for all features active on history for tag t.

Here, the weight parameter (α_i) is optimized by using the conjugate gradient descent (CGD)-based optimization technique [84]. CGD converges faster and is numerically more stable than other optimization techniques such as generalized iterative scaling (GIS) and improved iterative scaling (ITS) [44, 96].

3.2. BL-MENE significance

BL-MENE concentrates on finding the features that characterizes an NER problem and leaves an assignment of the feature weights behind for an estimation routine. It is observed that the BL-MENE estimation routine guarantees that, for each feature, the expected value of f_i (Equation (2)) must equal an empirical expectation of f_i (Equation (3)).

$$f_i^{expected} = \sum_{(h,t)\in(H,F)} P(h,t).f_i(h,t) \tag{2}$$

$$f_i^{empirical_expected} = \sum_{(h,t)\in observed(H,F)} f_i(h,t) \tag{3}$$

Baseline MENE assigns a new piece of text with the appropriate NE tags after training with the proper weight assignment for each feature. BL-MENE aims to mark correct tags for NEs and does not allow any invalid tag sequences to occur. For instance, tag sequence [B-PER, I-DIS] is invalid since it does not contain an ending token; moreover, these two tokens are not of the same NE-tag category.

3.3. BL-MENE features

For the success of any machine-learning approach, an appropriate feature selection is quite critical. In this research, a BL-MENE-based classifier is considered, as it is capable enough to utilize various features while computing conditional probabilities for NE types. This section details the baseline MENE features for the Hindi NER as follows:

- **Head Noun:** A head noun is usually defined as a major noun or noun phrase of an NE that describes its function or property.
- **Word Window:** A word window feature represents the previous words and the next words of a current HHD corpus word as a feature. A word window with a size of five is considered; i.e., w_{i-2}^{i+2}=w_{i-2}, w_{i-1}, w_0, w_{i+1}, w_{i+2}. Here, w_0 is the current word, w_{i-2}, w_{i-1} are the two previous words, and w_{i+1}, w_{i+2} are the two following words.
- **Root Word:** A root word feature represents the root forms of HHD corpus words by using a Hindi morphological analyzer (http://sampark.iiit.ac.in/hindimorph/web/restapi.php/indic/morphclient). This checks the root words, as Hindi is morphologically rich and its corpus words are highly inflected in different forms based on number, case, tense, and gender.

- **Word Suffix:** A word suffix feature represents the suffix of current words and those that surround the HHD corpus that is considered as a feature. Two- to four-character suffixes are used. Table 2 represents sample suffixes and their corresponding examples from the HHD corpus.

Table 2
Word suffixes

Suffix	Examples
-दर्द	सिरदर्द, पेटदर्द, गलादर्द, कमरदर्द
-हट	अकुलाहट, मिचलाहट, झनझनाहट, सरसराहट, खरखराहट, घबराहट
-पन	दुबलापन, गंजापन, चिपचिपापन, चिड़चिड़ापन, भारीपन
-इटिस	टेन्टीनाइटिस, बरसाइटिस, अर्थराइटिस, आस्टियोअर्थराइटिस
-त्सक	चिकित्सक, मनोचिकित्सक, दंतचिकित्सक
-पान	धूम्रपान, खानपान

- **Word Length:** This represents the word length of HHD corpus words as a feature. The length of HHD corpus words occurs within a range of 3–25 letters that belong to a certain NE class. Table 3 represents the HHD corpus word, word length, and corresponding NE type.

Table 3
Word-length feature

HHD Word	Word Length	NE Type
मां (Mother)	3	PER
गैस्ट्रोइंटेस्टाइनल_कैंसर (Gastrointestinal Cancer)	25	DIS

- **NE Information:** The NE information feature represents the NE tags of the previous words as a dynamic feature.
- **Frequent Words:** A list of the most frequently occurring words in the HHD corpus is prepared, and those words that occur more than 15 times are considered to be frequent words.
- **Shallow Parsing:** The shallow parsing feature represents chunk information (https://www.nlpworld.co.uk/nlp-glossary/c/chunking/) that is useful for knowing constituents such as the noun group, verb, verb group, etc. of the HHD corpus by using a Hindi shallow parser (http://ltrc.iiit.ac.in/analyzer/hindi/).
- **N-Gram:** The n-gram feature currently extracts uni-grams, bi-grams, and tri-grams from the HHD corpus due to limitations in the corpus size.
- **Stop Words:** The reemoval of stop words from the HHD corpus and the generation of word pairs from the rest of the HHD corpus formulates these word pairs as an extremely crude syntax approximation. This may be useful for removing those word pairs that consist solely of stop words. However, incorporating the stop words within the word pairs of the HHD corpus improves the BL-MENE system (which results in better NER performance).

4. Context pattern-based MENE

Baseline MENE can find NEs, but some are partially recognized NEs and are mistakenly classified as NEs. To resolve such a crucial issue, a context pattern-based MENE is needed for NER. The CP-MENE methodology is garnered with an effort to boost the baseline MENE performance while additionally incorporating a pattern set and an extensive feature set. These features include right-boundary, left-boundary, part-of-speech, synonym, gazetteer (single- or multi-word-entity dictionaries), and relative pronoun features. CP-MENE is well-suited for NE tagging decisions, as it significantly improves NER's performance over BL-MENE.

4.1. CP-MENE model

The context pattern-based MENE model has a sequence of token representations that possess a strong meaning as a unit and are independent of the individual words that are treated separately. CP-MENE characterizes the indices of patterns in which the words are included; otherwise, it treats them as zero whenever no pattern match is found. CP-MENE formulates a kind of recursive relationship to extract highly ranked NE patterns that are generated based on the regular expressions via Python@ code; these are exemplified in Equations (4) through (10).

$$< pattern\_list >:=< pattern > [\text{“|”} < pattern\_list >] \tag{4}$$

$$\begin{gathered} < pattern >:=< token\_list > [\text{“,”} < pattern >] | \text{“(”} < token\_list > \text{“)”} \\ [\text{“,”} < pattern >] \end{gathered} \tag{5}$$

$$\begin{gathered} < token\_list >:=< token\_expression > [\text{“|”} < token\_list >] \\ | \text{“(”} < token\_expression > \text{“)”} [\text{“|”} < token\_list >] \end{gathered} \tag{6}$$

$$\begin{gathered} < token\_expression >:=< token\_name > [< postposition\_marker >] \\ [< name\_constraint >] \end{gathered} \tag{7}$$

$$\begin{gathered} < token\_name >:=< gaz\_list > [\text{“}name\text{(”} < gaz\_list > \text{“)”}] \\ | \text{“}name\text{(”} < gaz\_list > \text{“)”} [< token\_name >] \end{gathered} \tag{8}$$

$$\begin{gathered} < gaz\_list >:=< element\_name > [\text{“,”} < gaz\_list >] \\ | \text{“(”} < element\_name > \text{“)”} [\text{“,”} < gaz\_list >] \end{gathered} \tag{9}$$

$$\begin{gathered} < name\_constraint >:=< Bv\_sG > [\text{“,”} < name\_constraint >] \\ | \text{“(”} < Bv\_sG > \text{“)”} [\text{“,”} < name\_constraint >] \end{gathered} \tag{10}$$

The CP-MENE model observes that <*pattern_list*> represents a list of patterns that are generated from the HHD corpus using <*pattern*>; <*pattern*> represents a list

of tokens that are built up using *<token_list>*; *<token_list>* represents a series of token expressions that are constructed using *<token_expression>*; *<token_expression>* represents the token name(s) from the HHD corpus using *<token_name>*, the optional *<postposition_marker>*, and *<name_constraint>*; *<token_name>* represents tokens from gazetteers that are incorporated from gazetteer lists; *<gaz_list>* represents a gazetteer list for names using *<element_name>*; *<element_name>* represents single or multiple words within the gazetteers' NEs dictionaries that are extended by using the Hindi WordNet synset *<HWN_synset>*; and Hindi WordNet (HWN) is a lexical database for the Hindi language that was developed at IIIT Hyderabad, India. HWN is useful for grouping various Hindi words into sets of synonyms; this is also known as a synset (http://www.cfilt.iitb.ac.in/wordnet/webhwn/).

In addition, Hindi grammar contains a postposition tag that represents short words that are placed after nouns and are very similar to prepositions in English. This tag is known as **सम्बन्धबोधक** (postposition) in Hindi, which is transliterated as *<saMbaMXa_boXaka>* and acronym as *<sM_bX>* [47]; within the CP-MENE model, this is represented as *<postposition_marker>*. On the same lines, the "Abstract Noun" tag is a noun type that is defined in terms of aspect, concept, experience, feeling, idea, state of being, trait, quality, or another entity that is not experienced with the five human senses (sight, touch, hearing, taste, and smell). This tag is known as **भाववाचक-संज्ञा** (abstract noun) in Hindi, which is transliterated as *<BAv_vAcaka_saMGYA>* and acronymized as *<Bv_sG>* [47]; within the CP-MENE model, this is represented as *<name_constraint>*.

Consider example *<pattern_list>* **डॉक्टर की सलाह** (the advice of a doctor) as input into the CP-MENE model (as can be seen in Figure 1).

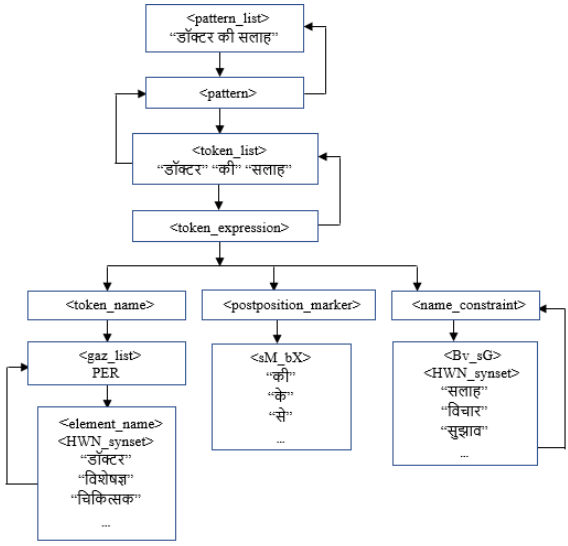

Figure 1. CP-MENE Model

<pattern_list> undergoes recursive *<pattern>*, *<token_list>*, and *<token_expression>* phases; *<token_expression>* is broken down into *<token_name>*, *<postposition_marker>*, and *<name_constraint>*; and *<token_name>* provides *<gaz_list>* (such as a PER gazetteer) that carries a recursive *<element_name>*. The PER's *<element_name>* generates words such as डॉक्टर (doctor), विशेषज्ञ (specialist), and चिकित्सक (doctor) using the Hindi WordNet synset *<HWN_synset>*; *<postposition_marker>* extracts सम्बन्धबोधक *<sM_bX>* such as की (of), के (to), and से (from); *<name_constraint>* extracts भाववाचकसंज्ञा textit<Bv_sG> such as सलाह (advice), विचार (opinion), and सुझाव (suggestion), which are extended using *<HWN_sysnset>*. Thus, context patterns are generated in a recursive manner from the HHD corpus using the CP-MENE model.

4.2. CP-MENE significance

CP-MENE is capable of resolving the boundary-detection problem by extending the boundary of the partially recognized NEs and considering the maximal NE classification. Once the boundary error correction is performed, the NER results of BL-MENE are reclassified. CP-MENE allows for flexible feature selection; i.e., new features are added to MENE so that there is no need to reformulate the model, and its estimation routine automatically calculates new weight assignments. CP-MENE considers features in a categorical manner; it yields significant results when accompanied with extremely informative features but does not make unnecessary feature independence assumptions. CP-MENE exploits beneficial features, ignores irrelevant features, integrates knowledge sources (such as Hindi WordNet), and is supported with optimized priors that are potentially useful for the NER task.

4.3. CP-MENE features

The CP-MENE features for the NER in Hindi includes right-boundary, left-boundary, synonym, part-of-speech, gazetteer list, and relative pronoun features; each of these is detailed as follows:

- **Right-Boundary ($\mathbf{F}_{RH}$):** the right-boundary feature is computed when an NE is followed by another NE or a head noun of the same type. Table 4 depicts the right-boundary feature in detail.
- **Left-Boundary ($\mathbf{F}_{LB}$):** The left-boundary feature is computed when a Not-Named Entity (NNE) is followed by head noun based Named Entity (NE)-type then it results in the formulation of an extended Named Entity. Table 5 depicts the left-boundary feature in detail.
- **Gazetteer Lists ($\mathbf{F}_{GL}$):** Gazetteers lists (or simply gazetteers – GAZ) are entity dictionaries that are important for effectively performing NER [118]. These are dependent on neither previously discovered tokens nor on annotations; they only expect raw text as input and then find matches in the HHD corpus based on its contents. The dictionary feature contains single- or multi-word dictionaries and are an important component of the CP-MENE system. Table 6 describes the

gazetteers for each NE type and the total count of NEs in each gazetteer, and it provides illustrative examples. For example, the gazetteer list for a person NE contains 3726 named entities and other details.

- **Synonym ($\mathbf{F}_{SS}$):** The synonym feature extends synonyms of the HHD corpus words using the Hindi WordNet synset (as can be seen in Table 7).
- **Relative Pronoun ($\mathbf{F}_{RP}$):** The relative pronoun (pronoun type) feature is used to introduce a relative or dependent clause in a sentence or as a standalone subject or object of the sentence. This is represented in Hindi grammar as सम्बन्धबोधक-सर्वनाम, which is transliterated as *<saMbaMXa_boXaka_srvnAma>* and acronymized as *<sM_sr>* [47].
  The relative pronoun feature in Hindi represents special words such as कौन (who), किसे (whom), कौन सी (which), किसको (whom), किसका (whose), किसकी (whose), and जिसे (whomever). In addition, their answers are those previously seen nouns in the HHD corpus that belong to the NE types.
- **Part-of-Speech ($\mathbf{F}_{PS}$):** The part-of-speech feature represents the part-of-speech markers or linguistic category of the HHD corpus words (which are also known as the lexical or grammatical categories) while using two taggers. These two taggers are the POS tagger from IIIT Hyderabad and the TENGRAM method (both are detailed here). Table 8 shows an example of an HHD corpus sentence that is tagged using the two taggers.

  – **POS TAGGER (IIIT HYDERABAD):** The POS tagger for Hindi, which was developed at IIIT Hyderabad (https://bitbucket.org/sivareddyg/hindi-part-of-speech-tagger/), India, is comprised of a variety of POS tags that are based on English grammar. Such tags include noun tags (NN – singular nouns; NNS – plural nouns; NST – nouns based on time and space; NNP – proper nouns), verb tags (VM – main verbs; VAUX – auxiliary verbs), adjective tags (JJ), adverb tags (RB), postposition tags (PSP), conjunction tags (CC), and punctuation tags (PUNC).
  – **POS TAGGER (TENGRAM):** The POS tagger using the TENGRAM [47] method is comprised of a variety of POS tags that are based on Hindi grammar. Such tags (Appendix A) include *sG*: संज्ञा (noun), *vN*: विशेषण (adjective), *ky*: क्रिया (verb) (*mu_ky:* मुख्यक्रिया (main verb) and *amu_ky:* अमुख्यक्रिया (auxiliary verb)), *sM_bX:* सम्बन्धबोधक (postposition), and *vr_cn:* विराम चिह्न (punctuation). The संज्ञा (noun) tag is further classified as *vv_sG:* व्यक्तिवाचक संज्ञा (proper noun), *jv_sG:* जातिवाचक संज्ञा (common noun), *sv_sG:* समुदायवाचक संज्ञा (collective noun), *xv_sG:* द्रव्यवाचक संज्ञा (material noun), and *Bv_sG:* भाववाचक संज्ञा (abstract noun). Among these diverse संज्ञा (noun) tags are व्यक्तिवाचक संज्ञा (proper noun), जातिवाचक संज्ञा (common noun), and समुदायवाचक संज्ञा (collective noun); these are beneficial for PER's NE recognition. द्रव्यवाचक संज्ञा (material noun) is advantageous for CNS NE recognition, and भाववाचक संज्ञा (abstract noun) is helpful in CP-MENE-based NE pattern recognition.

Table 4
Right-boundary CP-MENE feature

NE Type	HHD Phrase	BL-MENE Tag	CP-MENE Tag (F_{RH})
PER	रोगी व्यक्ति (Sick Person)	रोगी/U-PER व्यक्ति/U-PER	[रोगी व्यक्ति]/PER
DIS	दमा रोग (Asthma)	दमा/U-DIS रोग/U-DIS	[दमा रोग]/DIS
CNS	जौ का पानी (Barley Water)	जौ/B-CNS का/I-CNS पानी/E-CNS	[जौ का पानी]/CNS
SMP	घबराहट और खुजली (Nervousness and Itching)	घबराहट/U-SMP और/O खुजली/U-SMP	[घबराहट और खुजली]/SMP

Table 5
Left-boundary CP-MENE feature

NE Type	HHD Phrase	BL-MENE Tag	CP-MENE Tag (F_{LB})
PER	दंत चिकित्सक (Dentist)	दंत/O चिकित्सक /U-PER	[दंत चिकित्सक]/PER
DIS	अग्नाशय के कैंसर (Pancreatic Cancer)	अग्नाशय/O के/O कैंसर/U-DIS	[अग्नाशय के कैंसर]/DIS
CNS	काली मिर्च (Black Pepper)	काली/O मिर्च//U-CNS	[काली मिर्च]/CNS
SMP	फेफड़ों में कमजोरी (Lung Weakness)	फेफड़ों/O में/O कमजोरी/U-SMP	[फेफड़ों में कमजोरी]/SMP

Table 6
Description of gazetteer lists

GAZ	Count	Examples
PER	3726	मरीज़ (patient), चिकित्सक (doctor), एक्सपर्ट्स (experts)
DIS	1094	दमा (asthma),चेचक (whooping cough), कालीखांसी (chickenpox)
CNS	1988	शराब (wine), लहसुन (Garlic), ग्लूकोज (glucose), अनाज (grain)
SMP	762	सूजन (swelling), मतली (nausea), इंफेक्शन (infection), थकावट (tiredness)

Table 7
Synonym feature-based examples

NE Type	HHD Word	Synonym Examples
PER	बच्चा (Child)	नवजात_शिशु, नवजातक, लड़का, बालक, छोकड़ा, छोरा, छोकरा, लौंडा, वत्स, नन्हा-मुन्ना, नन्हा_मुन्ना, पुत्र, बेटा, सुत, शिशु
DIS	गठिया (Arthritis)	संधिवात, संधिशोथ, सन्धिवात, सन्धिशोथ, संधि_शूल, डमरुआ, डबरुआ, पवन-व्याधि, आर्थ्राइटिस, आर्थराइटिस
CNS	खाना (Food)	खाद्य_वस्तु, खाद्य_पदार्थ, आहार, खाद्य, भोज्य_पदार्थ, खाद्य_सामग्री, अन्न, आहर, फूड, भोजन, रसोई, रोटी, डाइट
SMP	दर्द (Pain)	तकलीफ, दरद, पीड़ा, तक़लीफ़, पीर, हूक, उपताप, उत्ताप, पीरा, वेदना, बेदना, क्लेश, व्यथा, अनुसाल

Table 8
POS features using taggers

Unannotated HHD Sentence	Tagged HHD Sentence	Tagger
पीड़ित व्यक्ति को दर्द है। (The victim is suffering)	पीड़ित/JJ व्यक्ति/NN को/PSP दर्द/VM है/VAUX ।/PUNC	IIIT HYDREBAD
	पीड़ित/vN व्यक्ति/jv_sG को/sM_-bX दर्द/mu_ky है/amu_ky ।/vr_cn	TENGRAM

5. Evaluation setup and results

Section 5.1 discusses HHD dataset statistics. Section 5.2 determines CP-MENE-based context patterns for Hindi NER. Section 5.3 evaluates performance-evaluation measures (precision, recall, and F-measure) for BL-MENE and CP-MENE. Section 5.4 compares CP-MENE with respect to BL-MENE and other NER approaches [64].

5.1. Dataset statistics

For experimental purposes, the HHD corpus is crawled from four well-known Hindi health domain-based Indian websites: The Traditional Knowledge Digital Library (http://www.tkdl.res.in/), the Ministry of Ayush (http://ayush.gov.in/), the University of Patanjali (https://www.patanjaliayurved.net/), and the Linguistic Data Consortium for Indian Languages (http://www.ldcil.org/). The HHD corpus is uploaded from the Kaggle dataset (https://www.kaggle.com/aijain/hindi-health-dataset/) and is comprised of the following dataset statistics: the corpus contains a total number of 193 MS Word pages that contain 5236 paragraphs, 9483 lines, 105,058 words,

411,462 characters (with no spaces), and 517,847 characters (with spaces). The corpus provides in-depth information of more than 100 diseases and their descriptions along with detailed symptoms, causes, treatments, and home remedies. The corpus is formed from unstructured data, from which four named-entity types are identified and classified: Person, Disease, Consumable, and Symptom. For training purposes, 80% of the corpus was considered; the unseen remaining 20% of the corpus was chosen as the test corpus.

5.2. CP-MENE patterns for Hindi NER

CP-MENE promises the longest NE annotation over BL-MENE-based individual word NE annotations. The longest NE annotations are formulated through context patterns that are beneficial for recognizing the specified NE types. Table 9 identifies the CP--MENE-based context patterns for each NE type (including the pattern count and exemplified patterns).

Table 9
CP-MENE identified context patterns

NE Type	Count of Context Patterns	Example Context Patterns	Elaborative Examples
PER	125	<PER >से सलाह (Advice from <PER>)	डॉक्टर से सलाह एक्सपर्ट से सलाह चिकित्सक से सलाह
DIS	101	<DIS >के निदान (Diagnosis of <DIS>)	कैंसर के निदान पथरी के निदान डायबिटीज के निदान
CNS	375	<CNS >का सेवन (Intake of <CNS>)	त्रिफला का सेवन तम्बाकू का सेवन दूध का सेवन
SMP	242	<SMP >का महसूस होना (Feeling of <SMP>)	कमजोरी का महसूस होना थकान का महसूस होना घबराहट का महसूस होना

5.3. Evaluation measures for CP-MENE

In order to evaluate the NER models, precision, recall, and F-measure are considered. Table 10 determines the results that were obtained by the two classifiers (BL-MENE and CP-MENE), both of which being further broken down by the respective entity types.

Table 10
Comparison of BL-MENE- and CP-MENE-based NER

NE TYPE	BL-MENE			CP-MENE		
	Precision [%]	Recall [%]	F-Measure [%]	Precision [%]	Recall [%]	F-Measure [%]
PER	78.34	75.49	76.89	79.89	79.47	79.68
DIS	66.32	64.39	65.34	72.25	72.75	72.50
CNS	55.35	56.64	55.99	70.25	67.37	68.78
SMP	52.88	53.65	53.26	70.29	64.43	67.23

Among all four NE types, PER and NE were the best performers, followed by DIS NE, CNS NE, and SMP NE. CP-MENE performed quite significantly as compared to BL-MENE. There was a slight increase in the F-measure result for the PER NE type (2.79% from BL-MENE to CP-MENE), while the greatest increase in the F-measure result for SMP NE was 53.26% (for BL-MENE) to 67.23% (for CP-MENE). For DIS NE, however, there was an increase of 7.16% in the F-measure result from BL-MENE to CP-MENE. And for CNS NE, the F-measure result was 55.99% for BL-MENE and 68.78% for CP-MENE. These results indicate an overall increase in F-measure results for the NE types in CP-MENE due to better context patterns, gazetteers, and extensive features as compared to BL-MENE.

5.4. Comparison of CP-MENE with other Hindi NER approaches

CP-MENE is compared to other Hindi NER approaches with respect to the F-measure evaluation metric (as is detailed below). Figure 2 compares the CP-MENE approach with respect to BL-MENE and an ontology-based Hindi NER approach [64]. The results demonstrate the effectiveness of CP-MENE over the other NER approaches. It can be observed that, for all of the four chosen NE types (PER, DIS, CNS, and SMP), CP-MENE's performance was slightly better when compared to OntoHindi NER and much better when compared to BL-MENE.

The OntoHindi NER [64] methodology maps ontology for health data in order to recognize NEs while maintaining the hierarchical information of ontological categories using Hindi WordNet (HWN). This methodology is comprised of five stages when building gazetteer lists: data pre-processing, feature engineering, string matching, concept hierarchy-based mapping (CHM), and concept selection and aggregation (COSA). For building gazetteer lists, initial seed lists are chosen that are further extended by using HWN. For data pre-processing, tokenization, phonetic transformation, and stop-word removal are considered. For feature engineering, frequency and part-of-speech tag filters are applied. For string matching, the Levenshtein distance and the linguistic-based improved Lin's matcher are applied. CHM extracts the ontological concepts of the corpus words and provides hierarchical and ontological chain structures for dendogram generation with respect to each NE category. Furthermore,

COSA helps improve the CHM-generated clustering results by allowing optimized selection and k-means aggregation, which results in fine-grained NE clusters.

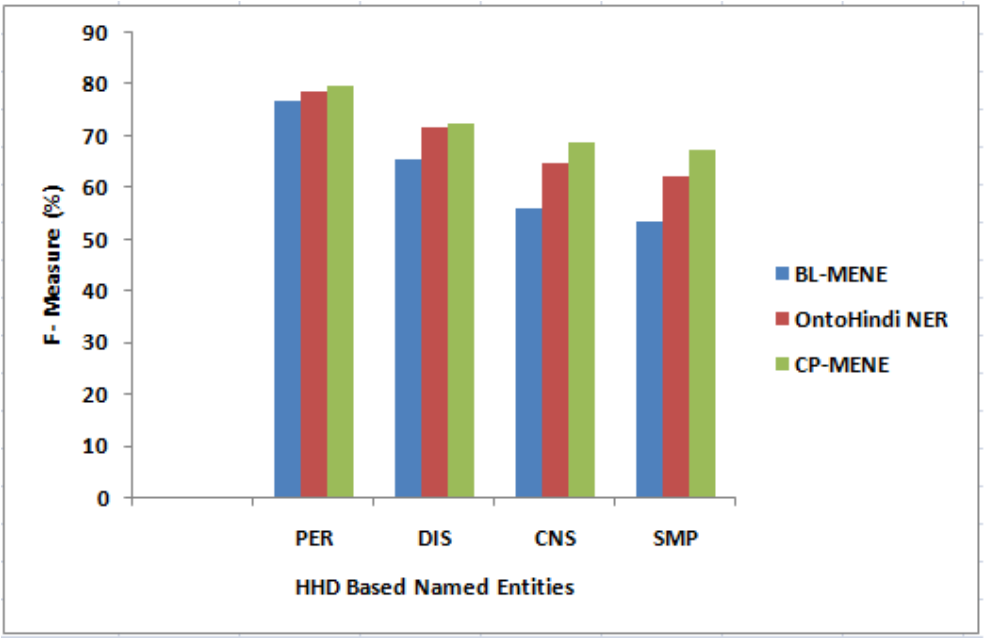

Figure 2. HHD-based comparison of different Hindi NER approaches

OntoHindi NER achieved the following F-measure results: PER (78.77%), DIS (71.57%), CNS (64.59%), and SMP (62.37%); however, CP-MENE-based NER achieved the following F-measure results: PER (79.68%), DIS (72.50%), CNS (68.78%), and SMP (67.23%). This shows that CP-MENE outperformed OntoHindi NER because of the fact that OntoHindi NER works with a single-word ontology while CP-MEME considers the longest NE annotations.

6. Conclusion

Named entities (NEs) make up one of the most important indexing elements in a given text for information extraction and other mining tasks. The construction of a named--entity-recognition system (NER) becomes quite challenging if proper resources for a language (such as Hindi) are not available. To solve this issue, the Hindi health domain (HHD) corpus was chosen from the Kaggle dataset, and four NE categories (Person [PER], Disease [DIS], Consumable [CNS], and Symptom [SMP] NEs) were considered. In order to perform NER on the HHD corpus, two methodologies were applied: baseline maximum entropy-based named-entity (BL-MENE), and context pattern-based MENE (CP-MENE) for NER. BL-MENE is an existing method that serves

as a flexible statistical framework and uses diversified features without looking into hand-coded patterns. The proposed CP-MENE method incorporates baseline MENE, pattern set, and extensive features. For each NE type, CP-MENE identifies varied context patterns and incorporates distinguished features. These features include right-boundary, left-boundary, part-of-speech, synonym, gazetteer list, and relative pronoun features. The part-of-speech feature is applied by using two taggers: the IIIT Hyderabad tagger (which uses English grammar), and the TENGRAM method (which uses Hindi grammar). POS tags that use Hindi grammar give more-detailed NE types as compared to English grammar. Even though BL-MENE performs the NER task, some NEs that are partially recognized or mistakenly classified by the baseline MENE are still left out in the HHD corpus. CP-MENE handles these boundary-detection and NE misclassification errors while exploiting beneficial features, ignoring irrelevant features, and integrating knowledge sources such as Hindi WordNet. In addition, CP-MENE is supported with an optimized prior that is potentially useful for the NER task. Thus, CP-MENE is observed to be more significant as compared to other Hindi NER approaches (OntoHindi NER [64], and BL-MENE). The findings of this research may serve as a beneficial tool – especially for those health care professionals who work in the northern rural areas of India (where the patients and their relatives do not understand English but are familiar with the Hindi language). These professionals investigate patient reports, diagnose diseases and their symptoms, counsel the eating habits of their patients, and comfort their patient's friends and family – all while needing to be understandable in Hindi. These professionals can also facilitate people through text summarizations and question-answering systems (i.e., giving verbal summaries of investigated reports, answering questions that are related to diseases, etc.) while applying the proposed Hindi NER system.

Limitations: The proposed NER system does not involve a deep-learning task. The NER system is applied only for the Hindi language and was tested over four NE categories.

In the future, a CP-MENE-based NER system can be mapped with a deep-learning strategy to further upgrade the results. Also, the proposed NER system can be applied to other Indian languages. In addition, some other NE types (such as treatment, quantity, and food) can be considered with more training corpora and enhanced sophisticated features that can further improve the NER system's performance.

References

[1] Abinaya N., John N., Ganesh B.H., Kumar A.M., Soman K.: AMRITA_CEN@FIRE-2014: Named Entity Recognition for Indian Languages using Rich Features. In: *Proceedings of the Forum for Information Retrieval Evaluation*, pp. 103–111, 2014. doi: 10.1145/2824864.2824882.

[2] Al-Rfou R., Kulkarni V., Perozzi B., Skiena S.: POLYGLOT-NER: Massive Multilingual Named Entity Recognition. In: *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 586–594, SIAM, 2015.

[3] Alsaaran N., Alrabiah M.: Classical Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and BERT, *IEEE Access*, vol. 9, pp. 91537–91547, 2021.

[4] Asti L., Uguzzoni G., Marcatili P., Pagnani A.: Maximum-entropy models of sequenced immune repertoires predict antigen-antibody affinity, *PLoS Computational Biology*, vol. 12(4), p. e1004870, 2016.

[5] Athavale V., Bharadwaj S., Pamecha M., Prabhu A., Shrivastava M.: Towards Deep Learning in Hindi NER: An approach to tackle the Labelled Data Scarcity, *arXiv preprint arXiv:161009756*, 2016.

[6] Banawan K., Ulukus S.: The Capacity of Private Information Retrieval from Coded Databases, *IEEE Transactions on Information Theory*, vol. 64(3), pp. 1945–1956, 2018.

[7] Benajiba Y., Rosso P., Benedíruiz J.M.: ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 143–153, Springer, 2007.

[8] Bender O., Och F.J., Ney H.: Maximum entropy models for named entity recognition. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pp. 148–151, 2003.

[9] Biswas S., Mishra M., Acharya S., Mohanty S.: A two stage language independent named entity recognition for Indian languages, *(IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 1(4), pp. 285–289, 2010.

[10] Bontcheva K., Derczynski L., Roberts I.: Crowdsourcing named entity recognition and entity linking corpora. In: *Handbook of Linguistic Annotation*, pp. 875–892, Springer, 2017.

[11] Borthwick A., Sterling J., Agichtein E., Grishman R.: Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In: *Sixth Workshop on Very Large Corpora*, 1998.

[12] Carpuat M., Wu D.: Improving statistical machine translation using word sense disambiguation. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 61–72, 2007.

[13] Carreras X., Màrquez L., Padró L.: Learning a perceptron-based named entity chunker via online recognition feedback. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pp. 156–159, 2003.

[14] Charniak E.: A maximum-entropy-inspired parser. In: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

[15] Chatterjee N., Kaushik N.: RENT: Regular expression and NLP-based term extraction scheme for agricultural domain. In: *Proceedings of the International Conference on Data Engineering and Communication Technology*, pp. 511–522, Springer, 2017.

[16] Chen C., Kong F.: Enhancing Entity Boundary Detection for Better Chinese Named Entity Recognition. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Vol. 2: Short Papers)*, pp. 20–25, 2021.

[17] Chinchor N., Marsh E.: Muc-7 information extraction task definition. In: *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, pp. 359–367, 1998.

[18] Chiticariu L., Krishnamurthy R., Li Y., Reiss F., Vaithyanathan S.: Domain adaptation of rule-based annotators for named-entity recognition tasks. In: *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 1002–1012, 2010.

[19] Chiu J.P., Nichols E.: Named entity recognition with bidirectional LSTM-CNNs, *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016.

[20] Chopra D., Jahan N., Morwal S.: Hindi named entity recognition by aggregating rule based heuristics and hidden markov model, *International Journal of Information*, vol. 2(6), pp. 43–52, 2012.

[21] Cucerzan S., Yarowsky D.: Language independent named entity recognition combining morphological and contextual evidence. In: *1999 joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, 1999.

[22] Ekbal A., Bandyopadhyay S.: Named Entity Recognition in Indian Languages Using Maximum Entropy Approach, *International Journal for Computer Processing of Languages (IJCPOL)*, vol. 21(3), pp. 205–237, 2008. doi: 10.1142/S1793840608001913.

[23] Ekbal A., Bandyopadhyay S.: A conditional random field approach for named entity recognition in Bengali and Hindi, *Linguistic Issues in Language Technology*, vol. 2(1), pp. 1–44, 2009.

[24] Ekbal A., Bandyopadhyay S.: Voted NER system using appropriate unlabeled data. In: *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pp. 202–210, 2009.

[25] Ekbal A., Bandyopadhyay S.: Named entity recognition using support vector machine: A language independent approach, *International Journal of Electrical, Computer, and Systems Engineering*, vol. 4(2), pp. 155–170, 2010.

[26] Ekbal A., Bandyopadhyay S.: Named entity recognition in Bengali and Hindi using support vector machine, *Lingvisticæ Investigationes*, vol. 34(1), pp 35–67, 2011.

[27] Ekbal A., Saha S.: Classifier ensemble selection using genetic algorithm for named entity recognition, *Research on Language and Computation*, vol. 8(1), pp. 73–99, 2010.

[28] Ekbal A., Saha S.: Weighted vote based classifier ensemble selection using genetic algorithm for named entity recognition. In: *International Conference on Application of Natural Language to Information Systems*, pp. 256–267, Springer, 2010.

[29] Ekbal A., Saha S.: A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies, *Expert Systems with Applications*, vol. 38(12), pp. 14760–14772, 2011.

[30] Ekbal A., Saha S.: Weighted vote-based classifier ensemble for named entity recognition: a genetic algorithm-based approach, *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 10(2), pp. 1–37, 2011.

[31] Ekbal A., Saha S.: Multiobjective optimization for classifier ensemble and feature selection: an application to named entity recognition, *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 15(2), pp. 143–166, 2012.

[32] Ekbal A., Haque R., Bandyopadhyay S.: Maximum Entropy Based Bengali Part of Speech Tagging, A. Gelbukh (ed.), *Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal*, vol. 33, pp. 67–78, 2008.

[33] Ekbal A., Saha S., Hasanuzzaman M.: Multiobjective approach for feature selection in maximum entropy based named entity recognition. In: *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, vol. 1, pp. 323–326, IEEE, 2010.

[34] Ekbal A., Saha S., Sikdar U.K.: On active annotation for named entity recognition, *International Journal of Machine Learning and Cybernetics*, vol. 7(4), pp. 623–640, 2016.

[35] Ekbal A., Saha S., Singh D.: Active machine learning technique for named entity recognition. In: *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pp. 180–186, 2012.

[36] Ekbal A., Saha S., Singh D.: Ensemble based active annotation for named entity recognition. In: *2012 Third International Conference on Emerging Applications of Information Technology*, pp. 331–334, IEEE, 2012.

[37] El-Halees A.M.: Arabic text classification using maximum entropy, *IUG Journal of Natural Studies*, vol. 15(1), 2015.

[38] Farmakiotou D., Karkaletsis V., Koutsias J., Sigletos G., Spyropoulos C.D., Stamatopoulos P.: Rule-based named entity recognition for Greek financial texts. In: *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, pp. 75–78, Citeseer, 2000.

[39] Flood M., Grant J., Luo H., Raschid L., Soboroff I., Yoo K.: Financial entity identification and information integration (FEIII) challenge: the report of the organizing committee. In: *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*, pp. 1–4, 2016.

[40] Fu R., Qin B., Liu T.: Generating Chinese named entity data from parallel corpora, *Frontiers of Computer Science*, vol. 8(4), pp. 629–641, 2014.

[41] Gali K., Surana H., Vaidya A., Shishtla P.M., Sharma D.M.: Aggregating machine learning and rule based heuristics for named entity recognition. In: *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[42] Gayen V., Sarkar K.: An HMM Based Named Entity Recognition System for Indian Languages: The JU System at ICON 2013, *CoRR*, vol. abs/1405.7397, 2014. http://arxiv.org/abs/1405.7397.

[43] Gella S., Sharma J., Bali K.: Query word labeling and Transliteration for Indian Languages: Shared task system description. In: *Working Notes – Forum for Information Retrieval Evaluation (FIRE) 2013 Shared Task*, 2013.

[44] Goodman J.: Sequential conditional generalized iterative scaling. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 9–16, 2002.

[45] Goyal A.: Named entity recognition for South Asian languages. In: *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[46] Guo H., Zhu H., Guo Z., Zhang X., Wu X., Su Z.: Domain adaptation with latent semantic association for named entity recognition. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 281–289, 2009.

[47] Gupta J., Tayal D.K., Gupta A.: A TENGRAM method based part-of-speech tagging of multi-category words in Hindi language, *Expert Systems with Applications*, vol. 38(12), pp. 15084–15093, 2011.

[48] Gupta P.K., Arora S.: An approach for named entity recognition system for Hindi: an experimental study, *Proceedings of ASCNT–2009, CDAC, Noida, India*, pp. 103–108, 2009.

[49] Gupta S., Bhattacharyya P.: Think globally, apply locally: using distributional characteristics for Hindi named entity identification. In: *Proceedings of the 2010 Named Entities Workshop*, pp. 116–125, 2010.

[50] Hamdi A., Linhares Pontes E., Boros E., Nguyen T.T.H., Hackl G., Moreno J.G., Doucet A.: A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2328–2334, 2021.

[51] Han A.L.F., Zeng X., Wong D.F., Chao L.S.: Chinese named entity recognition with graph-based semi-supervised learning model. In: *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pp. 15–20, 2015.

[52] Han N.R., Chodorow M., Leacock C.: Detecting Errors in English Article Usage with a Maximum Entropy Classifier Trained on a Large, Diverse Corpus. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2004. http://www.lrec-conf.org/proceedings/lrec2004/pdf/695.pdf.

[53] Han X., Kwoh C.K., Kim J.j.: Clustering based active learning for biomedical named entity recognition. In: *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 1253–1260, IEEE, 2016.

[54] Hasanuzzaman M., Ekbal A., Bandyopadhyay S.: Maximum entropy approach for named entity recognition in Bengali and Hindi, *International Journal of Recent Trends in Engineering*, vol. 1(1), p. 408, 2009.

[55] Hasanuzzaman M., Saha S., Ekbal A.: Feature subset selection using genetic algorithm for named entity recognition. In: *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pp. 153–162, 2010.

[56] Hayes B., Wilson C.: A maximum entropy model of phonotactics and phonotactic learning, *Linguistic Inquiry*, vol. 39(3), pp. 379–440, 2008.

[57] Hennig L., Truong P.T., Gabryszak A.: MobIE: A German Dataset for Named Entity Recognition, Entity Linking and Relation Extraction in the Mobility Domain, *arXiv preprint arXiv:210806955*, 2021.

[58] Ionescu B., Müller H., Villegas M., Arenas H., Boato G., Dang-Nguyen D.T., Cid Y.D., Eickhoff C., de Herrera A.G.S., Gurrin C., *et al.*: Overview of ImageCLEF 2017: Information extraction from images. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 315–337, Springer, 2017.

[59] Jain A.: *Named Entity Recognition for Hindi Language Using NLP Techniques*, Ph.D. Thesis. Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India, 2019. http://hdl.handle.net/10603/241558.

[60] Jain A., Arora A.: Named Entity Recognition in Hindi Using Hyperspace Analogue to Language and Conditional Random Field, *Pertanika Journal of Science & Technology*, vol. 26(4), pp. 1801–1822, 2018.

[61] Jain A., Arora A.: Named entity system for tweets in Hindi language, *International Journal of Intelligent Information Technologies (IJIIT)*, vol. 14(4), pp. 55–76, 2018.

[62] Jain A., Gairola R., Jain S., Arora A.: Thwarting Spam on Facebook: Identifying Spam Posts Using Machine Learning Techniques. In: *Social Network Analytics for Contemporary Business Organizations*, pp. 51–70, IGI Global, 2018.

[63] Jain A., Gupta A., Sharma N., Joshi S., Yadav D.: Mining application on analyzing users' interests from Twitter. In: *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, pp. 26–27, 2018.

[64] Jain A., Tayal D., Arora A.: OntoHindi NER – An ontology based novel approach for Hindi named entity recognition, *International Journal of Artificial Intelligence (IJAI)*, vol. 16(2), pp. 106–135, 2018.

[65] Jain A., Tayal D.K., Yadav D., Arora A.: Research trends for named entity recognition in Hindi language. In: *Data Visualization and Knowledge Engineering*, pp. 223–248, Springer, 2020.

[66] Jain A., Tripathi S., Dwivedi H.D., Saxena P.: Forecasting price of cryptocurrencies using tweets sentiment analysis. In: *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pp. 1–7, IEEE, 2018.

[67] Jain A., Yadav D., Tayal D.K.: NER for Hindi language using association rules. In: *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pp. 1–5, IEEE, 2014.

[68] Jayan J.P., Rajeev R., Sherly E.: A hybrid statistical approach for named entity recognition for Malayalam language. In: *Proceedings of the 11th Workshop on Asian Language Resources*, pp. 58–63, 2013.

[69] Kambhatla N.: Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 178–181, 2004.

[70] Kaur D., Gupta V.: A survey of named entity recognition in English and other Indian languages, *International Journal of Computer Science Issues (IJCSI)*, vol. 7(6), pp. 239–245, 2010.

[71] Kaur Y., Kaur E.R.: Named Entity Recognition (NER) system for Hindi language using combination of rule based approach and list look up approach, *International Journal of Scientific Research and Management (IJSRM)*, vol. 3(3), pp. 2300–2306, 2015.

[72] Kocaman V., Talby D.: Biomedical named entity recognition at scale. In: *International Conference on Pattern Recognition*, pp. 635–646, Springer, 2021.

[73] Kongburan W., Padungweang P., Krathu W., Chan J.H.: Metabolite Named Entity Recognition: A Hybrid Approach. In: *International Conference on Neural Information Processing*, pp. 451–460, Springer, 2016.

[74] Konkol M., Brychcín T., Konopík M.: Latent semantics in named entity recognition, *Expert Systems with Applications*, vol. 42(7), pp. 3470–3479, 2015.

[75] Kozareva Z., Bonev B., Montoyo A.: Self-training and co-training applied to Spanish named entity recognition. In: *Mexican International conference on Artificial Intelligence*, pp. 770–779, Springer, 2005.

[76] Krishnarao A.A., Gahlot H., Srinet A., Kushwaha D.S.: A comparison of performance of sequential learning algorithms on the task of named entity recognition for Indian languages. In: *International Conference on Computational Science*, pp. 123–132, Springer, 2009.

[77] Kumar N., Bhattacharyya P.: Named entity recognition in Hindi using MEMM (Technical Report), IIT Mumbai, 2006.

[78] Kumar N.K., Santosh G., Varma V.: A language-independent approach to identify the named entities in under-resourced languages and clustering multilingual documents. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 74–82, Springer, 2011.

[79] Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C.: Neural architectures for named entity recognition, *arXiv preprint arXiv:160301360*, 2016.

[80] Leaman R., Lu Z.: TaggerOne: joint named entity recognition and normalization with semi-Markov Models, *Bioinformatics*, vol. 32(18), pp. 2839–2846, 2016.

[81] 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.

[82] Li P., Wang M., Wang J.: Named entity translation method based on machine translation lexicon, *Neural Computing and Applications*, vol. 33(9), pp. 3977–3985, 2021.

[83] Li W., McCallum A.: Rapid development of Hindi named entity recognition using conditional random fields and feature induction, *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 2(3), pp. 290–294, 2003.

[84] Lin S.B., Zhou D.X.: Distributed kernel-based gradient descent algorithms, *Constructive Approximation*, vol. 47(2), pp. 249–276, 2018.

[85] Liu S., Sun Y., Li B., Wang W., Zhao X.: HAMNER: Headword amplified multi-span distantly supervised method for domain specific named entity recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8401–8408, 2020.

[86] Meselhi M.A., Bakr H.M.A., Ziedan I., Shaalan K.: Hybrid named entity recognition – application to Arabic language. In: *2014 9th International Conference on Computer Engineering* & Systems (ICCES), pp. 80–85, IEEE, 2014.

[87] Mora T., Walczak A.M., Bialek W., Callan C.G.: Maximum entropy models for antibody diversity, *Proceedings of the National Academy of Sciences*, vol. 107(12), pp. 5405–5410, 2010.

[88] Morwal S., Jahan N., Chopra D.: Named Entity Recognition using Hidden Markov Model (HMM), *International Journal on Natural Language Computing (IJNLC)*, vol. 1(4), pp. 15–23, 2012. doi: 10.5121/ijnlc.2012.1402.

[89] Moussallem D., Wauer M., Ngomo A.C.N.: Machine translation using semantic web technologies: A survey, *Journal of Web Semantics*, vol. 51, pp. 1–19, 2018.

[90] Nakov P., Hoogeveen D., Màrquez L., Moschitti A., Mubarak H., Baldwin T., Verspoor K.: SemEval-2017 Task 3: Community Question Answering, *arXiv preprint arXiv:191200730*, 2019.

[91] Nanda M.: The Named Entity Recognizer Framework, *International Journal of Innovative Research in Advanced Engineering*, vol. 1(4), pp. 104–108, 2014.

[92] Nasar Z., Jaffry S.W., Malik M.K.: Named Entity Recognition and Relation Extraction: State of the Art, *ACM Computing Surveys (CSUR)*, vol. 54(1), pp. 1–39, 2021. doi: 10.1145/3445965.

[93] Nayan A., Rao B.R.K., Singh P., Sanyal S., Sanyal R.: Named entity recognition for Indian languages. In: *Proceedings of the IJCNLP-08 workshop on named entity recognition for South and South East Asian Languages*, 2008.

[94] Neudecker C.: An open corpus for named entity recognition in historic newspapers. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4348–4352, 2016.

[95] Nothman J., Ringland N., Radford W., Murphy T., Curran J.R.: Learning multilingual named entity recognition from Wikipedia, *Artificial Intelligence*, vol. 194, pp. 151–175, 2013.

[96] Osborne M.: Using maximum entropy for sentence extraction. In: *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pp. 1–8, 2002.

[97] Pakhomov S.: Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 160–167, 2002.

[98] Patel A., Ramakrishnan G., Bhattacharya P.: Incorporating Linguistic Expertise Using ILP for Named Entity Recognition in Data Hungry Indian Languages. In: *International Conference on Inductive Logic Programming*, pp. 178–185, Springer, 2009.

[99] Patil N., Patil A.S., Pawar B.: Survey of named entity recognition systems with respect to Indian and foreign languages, *International Journal of Computer Applications*, vol. 134(16), 2016.

[100] Plu J., Rizzo G., Troncy R.: A hybrid approach for entity recognition and linking. In: *Semantic Web Evaluation Challenges*, pp. 28–39, Springer, 2015.

[101] Prakash H., Shambhavi B.R.: Approaches to Named Entity Recognition in Indian Languages: A Study, *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 3(6), pp. 191–194, 2014.

[102] Praveen P., Ravi Kiran V.: Hybrid Named Entity Recognition System for South and South East Asian Languages. In: *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008. https://aclanthology.org/I08-5012.

[103] Putthividhya D., Hu J.: Bootstrapped named entity recognition for product attribute extraction. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1557–1567, 2011.

[104] Quasthoff U., Biemann C., Wolff C.: Named entity learning and verification: expectation maximization in large corpora. In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[105] Ratnaparkhi A., Reynar J., Roukos S.: A maximum entropy model for prepositional phrase attachment. In: *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8–11, 1994*, 1994.

[106] Raychaudhuri S., Chang J.T., Sutphin P.D., Altman R.B.: Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature, *Genome Research*, vol. 12(1), pp. 203–214, 2002.

[107] Saha S., Ekbal A.: Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition, *Data* & Knowledge Engineering, vol. 85, pp. 15–39, 2013.

[108] Saha S.K., Chatterji S., Dandapat S., Sarkar S., Mitra P.: A Hybrid Approach for Named Entity Recognition in Indian Languages. In: *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, pp. 17–24, 2008.

[109] Saha S.K., Mitra P., Sarkar S.: Word clustering and word selection based feature reduction for MaxEnt based Hindi NER. In: *Proceedings of ACL-08: HLT*, pp. 488–495, 2008.

[110] Saha S.K., Mitra P., Sarkar S.: A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition, *Knowledge-Based Systems*, vol. 27, pp. 322–332, 2012.

[111] Saha S.K., Mitra P., Sarkar U.: A semi-supervised approach for maximum entropy based Hindi named entity recognition. In: *International Conference on Pattern Recognition and Machine Intelligence*, pp. 225–230, Springer, 2009.

[112] Saha S.K., Narayan S., Sarkar S., Mitra P.: A composite kernel for named entity recognition, *Pattern Recognition Letters*, vol. 31(12), pp. 1591–1597, 2010.

[113] Saha S.K., Sarathi Ghosh P., Sarkar S., Mitra P.: Named entity recognition in Hindi using maximum entropy and transliteration, *Polibits*, (38), pp. 33–41, 2008.

[114] Saha S.K., Sarkar S., Mitra P.: Gazetteer preparation for named entity recognition in Indian languages. In: *Proceedings of the 6th Workshop on Asian Language Resources*, 2008.

[115] Saha S.K., Sarkar S., Mitra P.: A hybrid feature set based maximum entropy Hindi named entity recognition. In: *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[116] Saha S.K., Sarkar S., Mitra P.: Feature selection techniques for maximum entropy based biomedical named entity recognition, *Journal of Biomedical Informatics*, vol. 42(5), pp. 905–911, 2009.

[117] Saha S.K., Sarkar S., Mitra P.: Hindi named entity annotation error detection and correction, *Language Forum*, vol. 35(2), pp. 73–93, 2009.

[118] Sahin H.B., Tirkaz C., Yildiz E., Eren M.T., Sonmez O.: Automatically annotated Turkish corpus for named entity recognition and text categorization using large-scale gazetteers, *arXiv preprint arXiv:170202363*, 2017.

[119] Sasidhar B., Yohan P., Babu A.V., Govardhan A.: A survey on named entity recognition in Indian languages with particular reference to Telugu, *International Journal of Computer Science Issues*, vol. 8(2), pp. 438–443, 2011.

[120] Shaalan K., Oudah M.: A hybrid approach to Arabic named entity recognition, *Journal of Information Science*, vol. 40(1), pp. 67–87, 2014.

[121] Sharma P., Sharma U., Kalita J.: Named entity recognition: A survey for the Indian languages, *Parsing in Indian Languages*, pp. 35–39, 2011.

[122] Sharma R., Morwal S., Agarwal B., Chandra R., Khan M.S.: A deep neural network-based model for named entity recognition for Hindi language, *Neural Computing and Applications*, vol. 32(20), pp. 16191–16203, 2020.

[123] Sharnagat R., Bhattacharyya P.: Hindi named entity recognizer for NER task of FIRE 2013, *FIRE-2013*, 2013.

[124] Shishtla P.M., Pingali P., Varma V.: A character n-gram based approach for improved recall in Indian language NER. In: *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[125] Sikdar U.K., Ekbal A., Saha S.: Differential evolution based feature selection and classifier ensemble for named entity recognition. In: *Proceedings of COLING 2012*, pp. 2475–2490, 2012.

[126] Singh A.K.: Named entity recognition for south and south east Asian languages: taking stock. In: *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[127] Smyrnioudis N.: *A Transformer-based Natural Language Processing Toolkit for Greek–Named Entity Recognition and Multi-task Learning*, Bachelor Thesis. Athens University of Economics and Business, Greece, 2021. http://www2.aueb.gr/users/ion/docs/smyrnioudis_bsc_thesis.pdf.

[128] Speck R., Ngomo A.C.N.: Ensemble learning for named entity recognition. In: *International Semantic Web Conference*, pp. 519–534, Springer, 2014.

[129] Srivastava S., Sanglikar M., Kothari D.: Named entity recognition system for Hindi language: a hybrid approach, *International Journal of Computational Linguistics (IJCL)*, vol. 2(1), pp. 10–23, 2011.

[130] Suárez-Paniagua V., Dong H., Casey A.: A multi-BERT hybrid system for named entity recognition in spanish radiology reports, *CLEF eHealth*, 2021.

[131] Szarvas G., Farkas R., Kocsor A.: A Multilingual Named Entity Recognition System Using Boosting and $C_{4.5}$ Decision Tree Learning Algorithms. In: *International Conference on Discovery Science*, pp. 267–278, Springer, 2006.

[132] Tanabe L., Xie N., Thom L.H., Matten W., Wilbur W.J.: GENETAG: a tagged corpus for gene/protein named entity recognition, *BMC Bioinformatics*, vol. 6(1), pp. 1–7, 2005.

[133] Thomas M., Latha C.: Sentimental analysis of transliterated text in Malayalam using recurrent neural networks, *Journal of Ambient Intelligence and Humanized Computing*, vol. 12(6), pp. 6773–6780, 2021.

[134] Uchimoto K., Sekine S., Isahara H.: The unknown word problem: a morphological analysis of Japanese using maximum entropy aided by a dictionary. In: *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001.

[135] Wang X., Yang C., Guan R.: A comparative study for biomedical named entity recognition, *International Journal of Machine Learning and Cybernetics*, vol. 9(3), pp. 373–382, 2018.

[136] Wang Y., Wang L., Rastegar-Mojarad M., Moon S., Shen F., Afzal N., Liu S., Zeng Y., Mehrabi S., Sohn S., Liu H.: Clinical information extraction applications: A literature review, *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, 2018. doi: 10.1016/j.jbi.2017.11.011.

[137] Xiong D., Liu Q., Lin S.: Maximum entropy based phrase reordering model for statistical machine translation. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 521–528, ACL, 2006.

[138] Yadav V., Bethard S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, *arXiv preprint arXiv:191011470*, 2019. doi: 10.48550/ARXIV.1910.11470.

[139] Yaseen U., Langer S.: Neural Text Classification and Stacked Heterogeneous Embeddings for Named Entity Recognition in SMM4H 2021, *arXiv preprint arXiv:210605823*, 2021. doi: 10.48550/ARXIV.2106.05823.

[140] Zhao L., Li L., Zheng X., Zhang J.: A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts. In: *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 1233–1238, IEEE, 2021. doi: 10.1109/CSCWD49262.2021.9437616.

[141] Zhou Y., Ju C., Caufield J.H., Shih K., Chen C., Sun Y., Chang K.W., Ping P., Wang W.: Clinical Named Entity Recognition using Contextualized Token Representations, *CoRR*, vol. abs/2106.12608, 2021. https://arxiv.org/abs/2106.12608.

Appendix A

It is important to note the following information regarding the applied POS tags in Hindi grammar [47].

- **Noun Tag:** N A noun tag is a part of speech that defines the name of a person, place, thing, idea, quality, etc. In Hindi grammar, a noun is considered to be संज्ञा, which is transliterated as *<saMGYA>* and acronymized as *<sG>*.
  - **Proper Noun:** A proper noun is defined as a unique entity that refers to the name of a specific person, place, or object and is represented in Hindi grammar as व्यक्तिवाचक संज्ञा, which is transliterated as *<v\yak\wi_vAcaka_saMGYA>* and acronymized as *<vv_sG>*.
  - **Common Noun:** A common noun is defined as the class of an entity that refers to the name of any individual person, place, or object and is represented in Hindi grammar as जातिवाचक संज्ञा, which is transliterated as *<jAwi_vAcaka_saMGYA>* and acronymized as as *<jv_sG>*.

  - **Collective Noun:** A collective noun is defined as a group or collection of people or things and is represented in Hindi grammar as समुदाय-वाचक संज्ञा, which is transliterated as *<samuxAya_vAcaka_saMGYA>* and acronymized as *<sv_sG>*.
  - **Material Noun:** A material noun is defined as the name of a substance that something is made of and is represented in Hindi grammar as द्रव्यवाचक संज्ञा, which is transliterated as *<xrav\ya_vAcaka_saMGYA>* and acronymized as *<xv_sG>*.
- **Adjective Tag:** An adjective tag is a part of speech that describes, identifies, or quantifies a noun or pronoun. In Hindi grammar, an adjective tag is considered to be विशेषण, which is transliterated as *<visheSNa>* and acronymized as *<vN>*.
- **Verb Tag:** A verb tag is a part of speech that describes an action, state, or occurrence and forms the main part of the predicate of a sentence. In Hindi grammar, a verb tag is considered to be क्रिया, which is transliterated as *<kriyA>* and acronymized as *<ky>*.
- **Punctuation Tag:** A punctuation tag is a part of speech that describes marks such as periods (full stops), commas, and brackets (which are used to separate sentences and their elements and to clarify meanings). In Hindi grammar, a punctuation tag is considered to be विरामचिह्न, which is transliterated as *<virAma_cinha>* and acronymized as *<vr_cn>*.

Appendix B

List of Acronyms & Abbreviations

ABB	Abbreviation NE
ANM	Animal NE
ART	Artefact NE
BL-MENE	Baseline MENE
BOK	Book NE
BRD	Brand NE
CIIL	Central Institute of Indian Languages
CRF	Conditional Random Field
CGD	Conjugate Gradient Descent
CNS	Consumable NE
CP-MENE	Context Pattern-Based MENE
CNT	Count NE
DTE	Date NE
DAY	Day NE
DEG	Designation NE
DRE	Direction NE
DIS	Disease NE
DST	Distance NE

ETN	Entertainment NE
EM	Expectation Maximization
FCT	Facility NE
FIRE	Forum for Information Retrieval Evaluation
GIS	Generalized Iterative Scaling
GA	Genetic Algorithm
HMM	Hidden Markov Model
ITS	Improved Iterative Scaling
ICON	International Conference on Natural Language Processing
IJCNLP	International Joint Conference on Natural Language Processing
LIV	Living Thing NE
LOC	Location NE
LMT	Locomotive NE
LSTM	Long Short-Term Memory
MAT	Material NE
ME	Maximum Entropy
MEMM	Maximum Entropy Markov Model
MSN	Measurement NE
MUC	Message Understanding Conference
MSC	Miscellaneous NE
MNY	Money NE
MNH	Month NE
MOO	Multi-Objective Optimization
NE	Named Entity
NGI	Named-Entity Identification Using Global Information
NER	Named-Entity Recognition
NERSSEAL	NER Shared Task for Southern and Southeastern Asian Languages
NUM	Number NE
OntoHindi NER	Ontology-Based Approach for Hindi NER
OSM	Organism NE
ORG	Organization NE
PCN	Percent NE
PIO	Period NE
PER	Person NE
PLN	Plant NE
PLY	Play NE
QNT	Quantity NE
RNN	Recurrent Neural Network
RIV	River NE
S-MEMM	MEMM-Based Statistical System
SPSAL	Shallow Parsing for Southern Asian Languages

SPR	Sport NE
SVM	Support Vector Machine
SMP	Symptom NE
THT	Technical-Term NE
TDIL	Technology Development for Indian Languages
TME	Time NE
TTO	Title-Object NE
TTP	Title-Person NE
TIDES	Translingual Information Detection, Extraction, and Summarization
TNS	Transport NE
VEH	Vehicle NE
YER	Year NE

Affiliations

Arti Jain
Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India
ajain.jiit@gmail.com

Divakar Yadav
NIT Hamirpur, Himachal Pradesh, India
divakar.yadav0@gmail.com

Anuja Arora
Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India
anuja.arora29@gmail.com

Devendra K. Tayal
Indira Gandhi Delhi Technical University for Women, New Delhi, India
dev_tayal2001@yahoo.com

Received: 14.09.2020
Revised: 01.10.2021
Accepted: 14.12.2021