Research Trends for Named Entity Recognition in Hindi Language

Chapter · January 2020

DOI: 10.1007/978-3-030-25797-2_10

4 authors:

Dr Arti Jain
Jaypee Institute of Information Technology
34 PUBLICATIONS **272** CITATIONS
SEE PROFILE

Devendra Kumar Tayal
IGDTUW
57 PUBLICATIONS **567** CITATIONS
SEE PROFILE

Divakar Yadav
Indira Gandhi National Open University (IGNOU)
136 PUBLICATIONS **990** CITATIONS
SEE PROFILE

Anuja Arora
Jaypee Institute of Information Technology
99 PUBLICATIONS **1,213** CITATIONS
SEE PROFILE

Research Trends for Named Entity Recognition in Hindi Language

Arti Jain, Devendra K. Tayal, Divakar Yadav and Anuja Arora

Abstract Named Entity Recognition (NER) is a process of identification and classification of names into pre-defined categories- person, location, organization, date, time and others. NER serves as one of the most valuable application tools for varying languages and domains. Despite its popularity and successful deployment in English, this area is still exploratory for the Hindi language. NER in Hindi is a challenging task due to the scarceness of language resources and complex morphological structure. An initial impetus to write this chapter constitutes concise research on prevailing NER systems in the Hindi language. To achieve this objective, all-inclusive review and analysis of the research trends for NER in Hindi starting from the year 1999 to till date is conducted from varied articles. These articles include full-length articles, survey papers, dissertations, and guidelines which exploit decision analysis of the NER with respect to six vital aspects. All these aspects are collaborated and visualized in details within the chapter. The future research directions, challenges, and open issues in the field of NER are also presented for keen researchers. One can then design a unified, coherent NER system which can be applied to vivid language processing tasks for the Hindi and other Indian languages.

Keywords Named entity recognition · Hindi language · Training corpora · Ner techniques · Gazetteer lists · Evaluation measures

A. Jain (✉) · A. Arora
CSE, Jaypee Institute of Information Technology, Noida, UP, India
e-mail: arti.jain@jiit.ac.in

A. Arora
e-mail: anuja.arora@jiit.ac.in

D. K. Tayal
CSE, Indira Gandhi Delhi Technical University for Women, Delhi, India
e-mail: dev_tayal2001@yahoo.com

D. Yadav
CSE, Madan Mohan Malvia University of Technology, Gorakhpur, UP, India
e-mail: dsycs@mmmut.ac.in

J. Hemanth et al. (eds.), *Data Visualization and Knowledge Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 32,
https://doi.org/10.1007/978-3-030-25797-2_10

1 Introduction

Named Entity Recognition (NER) [2, 31, 39, 44, 56, 57, 68, 79–81] is a sequence labelling task which seeks identification and classification of Named Entities (NEs). An identification of named entity represents the presence of a name- word or term or phrase as an entity within a given text. Classification of named entity denotes the role of an identified NE such as person name, location name, organization name, date, time, distance, and percent. In general, named entities are distributed into three universally [5] defined tags (Fig. 1)—ENAMEX, TIMEX, and NUMEX which are detailed below:

- **ENAMEX Tag**: ENAMEX tag is used for names entities e.g. person, location, organization, facility, locomotive, artifact, entertainment, material;
- **TIMEX Tag**: TIMEX tag is used for temporal entities e.g. time, day, month, period, date, special day, year;
- **NUMEX Tag**: NUMEX tag is used for numerical entities e.g. distance, quantity, money, count, percent, measurement.

In addition, depending upon the specific user needs there are marginal NE types viz. book title, phone number and email address [91], medical emails, scientific books and religious text [54], job title [8], scientist and film [27], project name and research area [92], device, car, cell phone, currency [64].

NER is recommended in situations where NEs have more importance than the actions they perform. Over the decades, NER is proved successful in vivid natural language processing application areas such as information retrieval [47], text summarization [35], co-reference resolution [11], word sense disambiguation [55], question answering [63], machine translation [90] and so on. NER has a profound impact on the society as it enables crowdsourcing [4] which is rapidly growing as social media

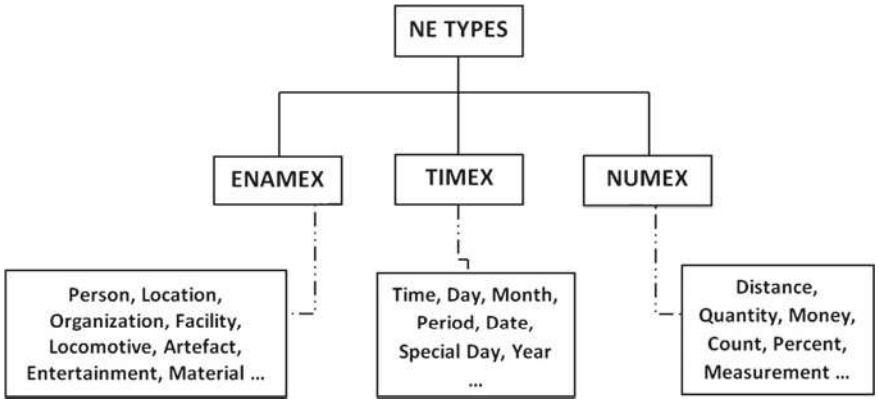

Fig. 1 NE categories and their sub-categories

content through plug-in of GATE[1] crowdsourcing and DBpedia[2] linking. In actuality, NER is a first step towards extraction of structured information from an unstructured text- newspaper articles, web pages and many more.

For more than a decade, NER for Indian languages, especially Hindi is thriving as a promising research topic. Several researchers have contributed their work for NER in Hindi using statistical [82, 88], rule-based [76] and hybrid systems [36, 77]. Some well-known statistical Hindi NER techniques are- Conditional Random Field (CRF) [52, 82], Maximum Entropy (MaxEnt) [67], Hidden Markov Model (HMM) [7], Expectation Maximization (EM) [9], Support Vector Machine (SVM) [20, 73], Genetic Algorithm (GA) [21, 22], Long Short Term Memory (LSTM) [1], Hyperspace Analogue to Language (HAL) [41, 42]. Some rule-based Hindi NER techniques are Association Rule Mining (ARM) [40], rule list-lookup [46], memory-based learning [76]. Some hybrid Hindi NER techniques are- combination of MaxEnt, rules and gazetteers [66], dimensionality reduction- CRF and SVM [74], Multi-Objective Optimization (MOO) [23, 75], K-Nearest Neighbour (KNN) and HMM [77], CRF and LSTM [87], HAL and CRF [41, 42]. Such a situation has probably contributed to the fact that the Hindi language has a wider perspective, a wider range of content, and so deserves more attention towards language-based research. The Hindi language is in the Devanagari [34] script and comes at the fourth position as the world's most-spoken first language. Also, it is characterized by the highly inflectional, morphologically rich, suffix-based, and word re-ordering language. A commendable source for the Hindi language processing is the Hindi WordNet[3] (HWN).

This chapter presents a comprehensive understanding of NER in the Hindi language and related issues. To achieve this objective, all-inclusive review and analysis of the research trends for NER in the Hindi starting from the year 1999 to till date is conducted from varied articles. These articles include full-length articles, survey papers, dissertations, and guidelines while exploiting NER decision analysis w.r.to six vital aspects. These research aspects are, namely- datasets, state-of-art techniques, training corpora, NE types, gazetteer lists, and evaluation measures. All these Hindi NER aspects collaborate among each other that envisage researchers to answer the following research questions:

- What are the existing datasets that are available for the Hindi NER?
- What are the different techniques that are applied for the Hindi NER?
- What are the varied training corpora that are accessible for the Hindi NER?
- What are the current and emerging named entity types for the Hindi NER?
- What are the several gazetteers lists that are available for the Hindi NER?
- What are the performance evaluation metrics to evaluate Hindi NER?
- What are the challenges, future directions and open issues for the Hindi NER?

[1] https://gate.ac.uk/wiki/crowdsourcing.html.

[2] https://github.com/dbpedia/links.

[3] http://www.cfilt.iitb.ac.in/wordnet/webhwn/.

In order to respond to the above questions, relevant research articles on the named entity recognition in the Hindi language are to be mined while laying down the following criteria: -

- **Level of relevance**: The research articles are given the highest to the lowest level of relevance based upon the SCImago Journal Ranking, author level metric/H-index, impact factor, citation and so on. For example, an article on the Hindi NER which is published in an International Journal with high impact factor and H-index has a higher relevance.
- **String-based search**: The string-based search is performed mainly on the titles of the research articles. For example, <NER for South Asian Languages>, <NER for South and South Asian Languages>, <NER for Indian Languages>, <Indian Language NER>, <NER in Indian Languages>, <NE in Under-Resourced Language>, <Hindi Language NER>, <Hindi NER>, <NER in Hindi>, <NER for Hindi>, <Hindi NE Annotation>.
- **Research timeline**: The timeline of the online Hindi NER research articles that starts from a novice study [9] to the latest [62] are been taken care of. NER for the Indian languages have come to limelight during 2007–2008 and since then research in this field is marching ahead.

Rest of the chapter is organized as follows. Section 2 describes named entity tagset for the Hindi NER. Section 3 describes state-of-art for NER in Hindi. Section 4 illustrates the contemporary challenges in Hindi NER. Section 5 concludes the chapter.

2 Named Entity Tagset for the Hindi Language

Named entity tagset for the Hindi language is standardized using the tagset guidelines[4] that are prepared at the AU-KBC Research Centre, Chennai, India. According to these guidelines, there are three prime NE categories for the Hindi NER viz. ENAMEX, NUMEX, and TIMEX which have a certain resemblance to the English NER guidelines.[5] Further, each of these NE categories contains various NE types that are described here using XML@ script.[6]

2.1 ENAMEX Tag

ENAMEX entity for the Hindi language includes various name expressions such as person, location, organization, facility, locomotive, artefact, entertainment, cuisine, organism, plant, disease NEs. All such NEs are listed here one by one.

[4] www.au-kbc.org/.

[5] https://cs.nyu.edu/faculty/grishman/mu6.html.

[6] https://www.w3schools.com/xml/.

2.1.1 Person NE

Person entity is considered as a human being, who is an individual or a set of individuals/group. It also includes fictional characters from the story, novel, etc. A person name includes person first name, middle name, and last name, along with titles such as Mr., Mrs., Ms., and Dr.

E.g. <ENAMEX TYPE = "PERSON"> जशोदाबेन नरेंद्र मोदी</ENAMEX>

English Translation: जशोदाबेन नरेंद्र मोदी(Jashodaben Narendra Modi)

2.1.2 Location NE

Location entity is considered as a geographical entity such as geographical area, landmass, water body and geological formation. It includes address such as PO Box, street name, house/plot number, area name, and pin code. It also includes name of village, district, town, metropolitan city, cosmopolitan city, state or province capital, state, the national capital, nation, highway, continent, any other place such as religious place, etc. In addition, it includes the nickname given to a city or state such as Jaipur-Pink City. Further, fine-grained refinement of few locations based NEs also persist such as water body includes the name of pond, lake, stream, river, sea, and ocean. The landscape includes mountain, mountain range, valley, glacier, desert, forest, ghat, and wetland and so on.

E.g. <ENAMEX TYPE = "LOCATION"> पुरानी दिल्ली</ENAMEX>

English Translation: पुरानी दिल्ली (Old Delhi)

2.1.3 Organization NE

Organization entity is considered as limited to a corporation, agency, and another group of people that are defined by an established organizational structure. It includes the name of the village panchayat, municipality, road corporations, travel operator, advertisement agency, and TV channels such as news channel; political party/group, the militant/terrorist organization, professional regulatory bodies such as Indian Medical Association, IEEE; charitable, religious board and many more.

E.g. <ENAMEX TYPE = "ORGANIZATION"> माता वैष्णो देवी तीर्थ बोर्ड</ENAMEX>

English Translation: माता वैष्णो देवी तीर्थ बोर्ड(Mata Vaishno Devi Shrine Board)

2.1.4 Facility NE

Facility entity is considered as limited to building, man-made structure, and real-estate such as hospital, institute, library, hotel, factory, airport, railway station, police station, fire station, harbor, public comfort station, etc. The hospital can include the name of a hospital, clinic, dispensary, primary health unit. Institute can include the

name of an educational institution, research center, and training center. The library can include the name of a public library, private library. The hotel can include the name of a hotel, restaurant, lodge, fast food center. Factory can include the name of a factory, refinery, chemical plant, sewage treatment plant. The airport can include the name of the domestic airport, international airport, air force base station and so on.

E.g. <ENAMEX TYPE = "FACILITY"> इन्दिरा गाँधी अन्तरराष्ट्रीय हवाईअड्डा </ENAMEX>

English Translation: इन्दिरा गाँधी अन्तरराष्ट्रीय हवाईअड्डा(Indira Gandhi International Airport)

2.1.5 Locomotive NE

The locomotive entity is considered as a physical device that is designed to move, carry, pull or push an object from one location to another. It includes the name of flight, train, bus and many more.

E.g. <ENAMEX TYPE = "LOCOMOTIVE"> वंदे भारत एक्सप्रेस</ENAMEX>

English Translation: वंदे भारत एक्सप्रेस(Vande Bharat Express)

2.1.6 Artefact NE

Artefact entity is considered as an object which is shaped by human craft. It includes tool, ammunition, painting, sculpture, cloth, medicine, gem and stone. The tool can be an object that can help the human to do some work such as a hammer, knife, crockery item, vessel, and computer. Ammunition can be a weapon or a bomb that can be used in war. Painting can be an artwork such as modern art. Sculpture can be the name of a sculpture or a statue. The cloth can be a variety of textile such as silk, cotton, rayon. Gem and stone can be a marble stone, granite, diamond and many more. Medicine can be a kind of ayurvedic, allopathic and homeopathic medicine to prevent and cure disease.

E.g. <ENAMEX TYPE = "ARTEFACT"> त्रिफला चूर्ण </ENAMEX>

English Translation: त्रिफला चूर्ण(Triphala Powder)

2.1.7 Entertainment NE

Entertainment entity is considered as an activity which diverts human attention, gives pleasure and amusement. It includes the performance of some kind such as dance, music, drama, sport, and event. Dance can have different forms such as folk dance-bhangra, ballet. Music can have different forms such as western, Indian classical. Drama can have different forms such as theatre art, cinema, and film. Sport can have different forms such as outdoor- cricket, football; indoor- chess. An event can have

different forms such as ceremony, show, conference, workshop, symposium, seminar and many more.

E.g. <ENAMEX TYPE = "ENTERTAINMENT"> अभिज्ञानशाकुन्तलम्</ENAMEX>

English Translation: अभिज्ञानशाकुन्तलम्(Abhijñānashākuntala)

2.1.8 Cuisine NE

Cuisine entity is considered as a variety of food types that are prepared in different manners such as Chinese, South-Indian, North-Indian foods, etc. It includes different dishes, food recipes such as idly, soup, ice-cream and many more.

E.g. <ENAMEX TYPE = "CUISINE"> चपाती</ENAMEX>

English Translation: चपाती(Chapati/Bread)

2.1.9 Organism NE

Organism entity is considered as a living thing which has the ability to act or function independently such as animal, bird, reptile, virus, bacteria, etc. It also includes human organs. Here, the plant is not considered since it is classified as a separate entity.

E.g. <ENAMEX TYPE = "ORGANISM"> छिपकली</ENAMEX>

English Translation: छिपकली(Lizard)

2.1.10 Plant NE

Plant entity is considered as a living thing which has photosynthetic, eukaryotic, multi-cellular organism characteristics of the kingdom Plantae. It contains chloroplast, has a cellulose cell wall, and lacks the power of locomotion. It includes the name of the herb, medicinal plant, shrub, tree, fruit, flower and many more.

E.g. <ENAMEX TYPE = "PLANT"> नीम का पेड़ </ENAMEX>

English Translation: नीम का पेड़(Neem Tree)

2.1.11 Disease NE

Disease entity is considered as a disordered state or incorrect functionality of an organ, body part, body system which occurs because of genetic factors, developmental factors, environmental factors, infection, poison, toxin, deficiency in nutrition value or its imbalance, etc. It can cause illness, sickness, ailments such as fever, cancer; and comprises of disease names, symptoms, diagnosis tests, and treatments, etc.

E.g. <ENAMEX TYPE = "DISEASE"> कुष्ठरोग </ENAMEX>

English Translation: कुष्ठरोग(Leprosy)

2.2 NUMEX Tag

NUMEX[7] entity for the Hindi language includes various numeric expressions based NEs such as count, distance, money, quantity NEs. All such NEs are listed here one by one.

2.2.1 Count NE

Count entity indicates the number or count of item or article or thing etc.
For example, <NUMEX TYPE = "COUNT"> दस आम </NUMEX>
English Translation: दस आम(10 Mangoes)

2.2.2 Distance NE

Distance entity indicates distance measures such as- miles, kilometre etc.
For example, <NUMEX TYPE = "DISTANCE"> 2.5 मील </NUMEX>
English Translation: 2.5 मील(2.5 Miles)

2.2.3 Money NE

Money NE: Money entity indicates different units of money.
For example, <NUMEX TYPE = "MONEY"> पाँच सौ रुपये </NUMEX>
English Translation: पाँच सौ रुपये(Five Hundred Rupees)

2.2.4 Quantity NE

Quantity entity indicates a measure of volume, weight, etc. It also includes expression that conveys some quantity such as- percentage, little, some, etc.
For example, <NUMEX TYPE = "QUANTITY"> २ लिटर </NUMEX>
English Translation: २ लिटर(2 l)

2.3 TIMEX Tag

TIMEX entity for the Hindi language includes various time-related expressions based NEs such as time, date, day, period NEs. All such NEs are listed here one by one.

[7] ltrc.iiit.ac.in/iasnlp2014/slides/lecture/sobha-ner.ppt.

2.3.1 Time NE

Time entity refers to the expression of time such as hour, minute and second.
For example, <TIMEX TYPE = "TIME"> सुबह पांच बजे </TIMEX>
English Translation: सुबह पांच बजे(At Five in the Morning)

2.3.2 Date NE

Date entity refers to the expression of date in different forms such as month, date and year.
For example, <TIMEX TYPE = "DATE"> 15 अगस्त 1947 </TIMEX>
English Translation: 15 अगस्त 1947 (15 August 1947)

2.3.3 Day NE

Day entity refers to the expression of the day such as some special day, days that are weekly, fortnightly, quarterly, biennial, etc.
For example, <TIMEX TYPE = "DAY"> स्वतंत्रता दिवस </TIMEX>
English Translation: स्वतंत्रता दिवस(Independence Day)

2.3.4 Period NE

Period entity refers to the expression of the duration of time, time interval, or time period, etc.
For example, <TIMEX TYPE = "PERIOD"> चौदह साल </TIMEX>
English Translation: चौदह साल(14 Years)

2.4 Other-Than-NE

Other-than-NE considers those names which are mentioned in datasets such as the International Joint Conference on Natural Language Processing (IJCNLP). However, these NEs are fine-grained entity types which are clubbed within certain other NEs to result in coarse-grained NE [15] categorization for an efficient NER system. For example, NEs- abbreviation, brand, designation, a technical term, title-person, and title-object are grouped under the other than NEs.

2.5 *Newfangled NEs*

Among all the above stated NE types, most of the authors have published work mainly for a person, location, and organization NEs. Recently, authors Jain et al. [43], Jain and Arora [41, 42] have published work on Hindi NER for disease, symptom, consumable, organization and person NEs in the health care domain. Since the health care industry is emerging as one of the largest industries in the world which has a direct impact on the quality of everyone's life and forms an enormous part of our country economy. To work in this direction, the chosen health NEs are related to diagnosis, treatment, and prevention of disease; and is delivered by care providers such as medical practitioners, nursing, pharmacy, and community health workers. For this, Hindi Health Data (HHD) corpus is taken into consideration which is available at the Kaggle dataset[8] and contains more than 105,500 words. The HHD corpus is a quite beneficial resource for the health-based Natural Language Processing (NLP) research.

3 State-of-the-Art for NER in Hindi

This section presents a holistic analysis of the Hindi NER while integrating almost all the major studies in this field. The relevant research articles include full-length articles, survey papers, dissertation, thesis, Hindi NER guidelines that are examined from the six rationale aspects. These Hindi NER research aspects are- datasets, techniques, training corpora, named entities, gazetteers and evaluation measures. A comprehensive review of the rationale aspects is presented here.

3.1 *Hindi NER Datasets*

The Hindi NER datasets include Message Understanding Conference (MUC) dataset, Translingual Information Detection, Extraction, and Summarization (TIDES) dataset, NLPAI, Shallow Parsing for South Asian Languages (SPSAL) dataset, IJCNLP NER Shared Task for South and South East Asian Languages (NERSSEAL) dataset, Forum for Information Retrieval Evaluation (FIRE) dataset, International Conference on Natural Language Processing (ICON) dataset, Central Institute of Indian Languages (CIIL) dataset, Dainik Jagaran- popular Hindi newspaper, Gyaan Nidhi corpus- collection of various books in Hindi, web source- tourism domain, HHD corpus- health domain, and varied sources- topics include social sciences, biological sciences, financial articles, religion but not a news corpus. Apart from the above corpora, several other corpora are also encountered such as Language Technologies Research Centre (LTRC) IIIT Hyderabad, tweet corpora- health tweets,

[8] https://www.kaggle.com/aijain/hindi-health-dataset.

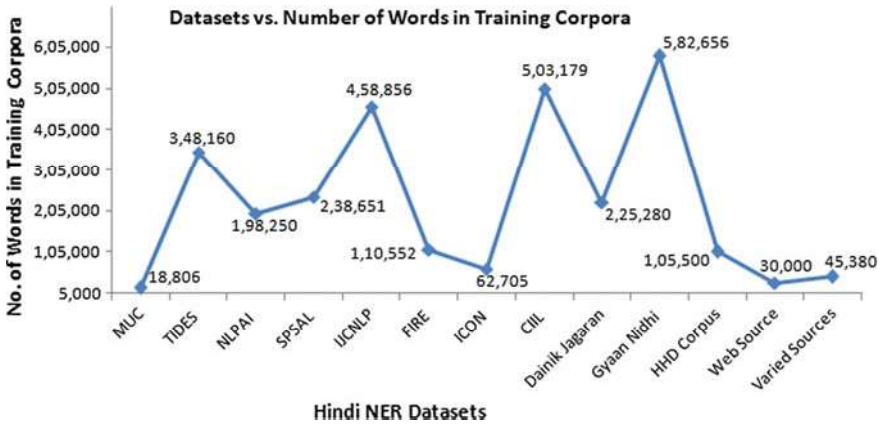

Fig. 2 Hindi NER datasets versus number of words in training corpora

Entity extraction in Social Media text track for Indian Languages (ESM-IL), Code Mix Entity Extraction for Indian Languages (CMEE-IL). It is noted that the NER in Hindi is performed mainly upon these datasets. Figure 2 shows Hindi NER datasets along with the distribution of a total number of words in each of them. Among them, Gyaan Nidhi corpus comprises the highest number of words in the training corpus, followed by CIIL corpus and others.

3.2 Hindi NER Techniques

Several Hindi NER techniques are worked upon by the researchers' over varied datasets w.r.to F-measure that are detailed in Table 1.

In the above table, few techniques result in more than one F-measure values. This is due to certain variant factors such as choice of NE types, distinguished feature sets, gazetteer lists, and datasets that are taken care of by the researchers. To exemplify the same, both [14, 15] consider the CRF technique over the IJCNLP-08 NERSSEAL shared task dataset. Certain features are common to them such as context word, word prefix, word suffix, NE information, and digit information. Also, both contain gazetteers such as first name, middle name, last name, measurement expression, weekdays and month name. However, [14, 15] work for different NE types and few distinguishing features which give different F-measures (Table 2). Other NER techniques and their F-measures as in Table 1 can be justified on the same lines.

Table 1 Hindi NER techniques over datasets are evaluated w.r.to F-measure

Datasets	Hindi NER techniques	F-measure (%)
MUC-6	EM style bootstrapping [9]	41.70
SPSAL 2007	MaxEnt, HMM [3]	71.95
	HMM [12]	78.35
NLPAI 2007	CRF [30]	58.85
BBC and EMI documents	CRF [52]	71.50
Gyaan Nidhi corpus	CLGIN [33]	72.30
Hindi newspapers	ARM [40]	77.81
	Rule-based, list lookup [46]	95.77
Health tweets	HAL, CRF [41]	40.87
Dainik Jagaran	CRF, word clustering, cluster merging [62]	85.83
	MaxEnt, gazetteers, context patterns [67]	81.52
	MaxEnt, gazetteers [69]	81.12
	MaxEnt, word clustering, word selection [70]	79.85
	MaxEnt, semi-supervised learning [72]	78.64
	SVM [73]	83.56
HHD Corpus	HAL, CRF [42]	89.14
	OntoHindi NER [43]	69.33
LTRC Corpus	Bi-directional RNN, LSTM [1]	77.48
ICON 2013	Memory-based, backward elimination [76]	78.37
	HMM [29]	75.20
IJCNLP-08 NERSSEAL	HMM [7]	87.14
	MaxEnt [13]	82.66
	CRF [14]	36.75
	CRF [15]	78.29
	MaxEnt, CRF, SVM [16]	92.98
	SVM [17]	77.17
	GA based classifier ensemble [18]	86.03
	GA-weighted vote based classifier ensemble [19]	72.60
	SVM [20]	80.21
	MOO [21]	92.80
	GA based classifier ensemble- 3 classifiers [22]	92.20
	MOO- MaxEnt, SVM, CRF [23]	93.20
	SVM [24]	89.81
	Ensemble learning [25]	87.86
	Ensemble-based active learning [26]	88.50
	CRF, heuristic rules [28]	50.06
	MaxEnt [37]	82.66

(continued)

Table 1 (continued)

Datasets	Hindi NER techniques	F-measure (%)
	GA [38]	80.46
	HMM, CRF [51]	46.84
	MaxEnt, rules, gazetteers [66]	65.13
	MaxEnt, heuristic rules, context patterns, bootstrapping [71]	96.67
	MOO- 7 classifiers [75]	94.66
	Differential evolution [85]	88.09
	CRF, MaxEnt, rule-based [89]	80.82
FIRE 2010	Bisecting K-means clustering [50]	76.20
FIRE 2013	CRF [83]	96.00
FIRE 2015	CRF [53]	57.59
Twitter	CRF, LSTM [87]	72.06
	DT, CRF, LSTM [88]	95.00
CMEE-IL	SVM [10]	54.51
ESM-IL	CRF [59]	57.55
	CRF [65]	61.61
Web sources	Phonetic matching [58]	65.23
	CRF [32]	64.42
	CRF [84]	45.48
Unknown sources	Rule-based heuristics, HMM [6]	94.61
	CRF, SVM [48]	47.00
	MaxEnt [49]	79.17
	CRF [82]	70.45
	WARMR, TILDE [60]	–
	KNN, HMM [77]	–

Table 2 F-measure for varying NE types and features on CRF technique over IJCNLP-08

Reference	NE Types	Distinguished features	F-measure (%)
Ekbal et al. [14]—CRF	12-person, location, organization, abbreviation, brand, title-person, title-object, time, number, measure, designation, term	word frequency, rare word	36.75
Ekbal and Bandyopadhyay [15]—CRF	5-person, location, organization, miscellaneous, and other-than-NE	first word, infrequent word, word length, POS information	78.29

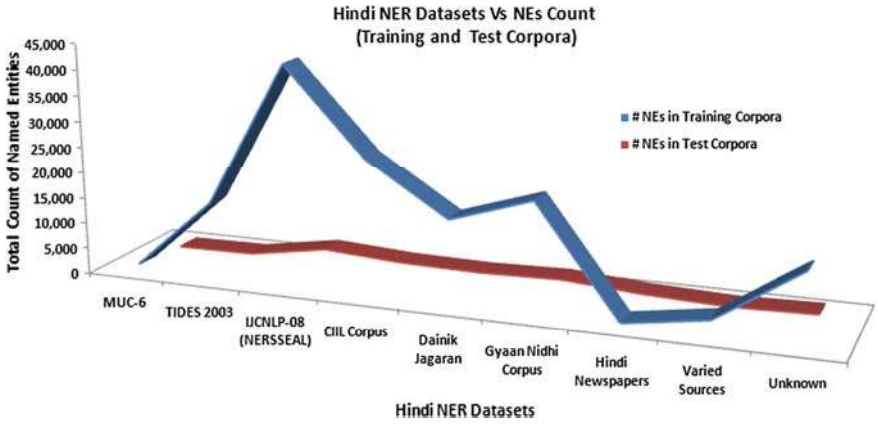

Fig. 3 Hindi NER datasets versus named entities count

3.3 Dataset Versus NE Count

Hindi NER datasets contain training and test corpora having a certain number of named entities. The total count of the number of words as NEs which occur in both training and test corpora are shown in Fig. 3 for the Hindi NER datasets. For example, IJCNLP-08 has the highest number of NEs count in training (43,021) and test (3,005) while CIIL corpus has a number of NEs count in training (26,432) and test (2,022) respectively. TIDES 2003 has a number of NEs count in training (15,063) and test (476) while Hindi Newspapers- Punjab Kesari, Navbhart Times, and Hindustan has a number of NEs count in training (1,030) and test (687) respectively. In addition, some unknown datasets (missing in literature) have a number of NEs count in training (13,767) and test (892), etc.

3.4 NE Tagset Versus NE Count

The hierarchical NE tagset for Hindi NER versus count of each NE type is shown in Fig. 4. Here, the count of NE determines the total number of full-length articles which takes that particular NE type into consideration, and its value is represented as follows. ENAMEX tag comprises of person (54), location (54), organization (53), artefact (8), disease (5), symptom (2), consumable (3), entertainment (10), facility (4), locomotive (10), organism (6), plant (6), and cuisine (4). NUMEX tag comprises of count (4), distance (1), money (6), and quantity (6). TIMEX tag comprises of time (26), date (18), day (4), and period (4). Other-than-NE (12) contains abbreviation, brand, designation, a technical term, title-object, title-person NEs.

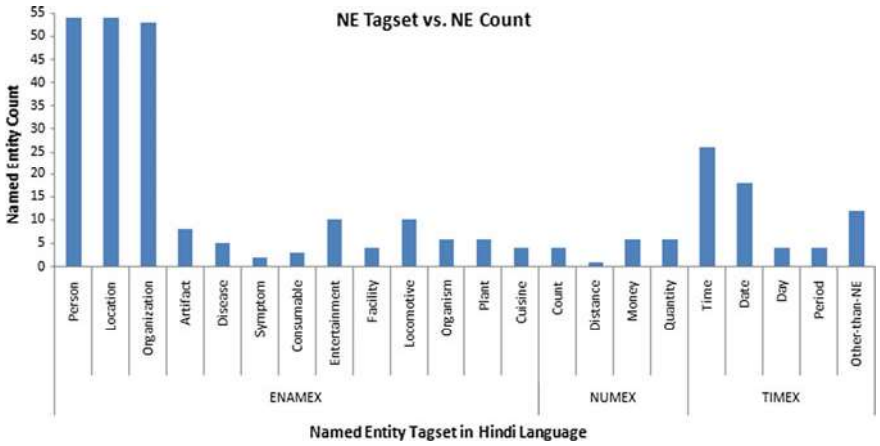

Fig. 4 Hindi NE tagset versus named entities count

3.5 Hindi NER Gazetteers

Gazetteer lists are entities based dictionaries which are important for performing NER effectively [43, 68] as they support non-local model to resolve multiple names of the same entity. Such lists are neither dependent upon previously discovered tokens nor on annotations. Each of these list intakes raw textual input and finds a match based upon its information. Table 3 shows a list of gazetteers which are used by researchers in their Hindi NER related work. For example, person gazetteer comprises of designation, title-person, first name, middle name, last name or surname. Location list comprises of common location, location name, etc. Organization list comprises of the organization name, organization end words. However, certain gazetteers have no further sub-divisions- Jain and Arora [42] have discussed person, disease, symptom, and consumable gazetteers. Ekbal and Bandyopadhyay [15] have discussed month names, weekdays, function words, and measurement expressions gazetteers.

3.6 Datasets Versus Evaluation Measures

The three standard evaluation metrics for Hindi NER- precision, recall, and F-score are taken care of w.r.to different datasets (Fig. 5). For example, IJCNLP-08 has the highest F-score, of 96.67% along with precision as 97.73%, and recall as 95.64% which is followed by FIRE 2013 with F-value as 96.00%, and others.

Table 3 Gazetteers in Hindi NER

Reference	Person name	First name	Middle name	Last name	Person prefix	Person suffix	Designation	Location	Organization name	Organization suffix	Disease
Biswas et al. [3]		Y	Y	Y							
Cucerzan and Yarowsky [9]		Y		Y				Y			
Devi et al. [10]	Y							Y	Y		
Ekbal and Bandyopadhyay [13]		Y	Y	Y							
Ekbal et al. [14]		Y	Y	Y							
Ekbal and Bandyopadhyay [15]		Y	Y	Y							
Ekbal and Bandyopadhyay [17]	Y							Y	Y		
Ekbal and Saha [18]		Y	Y	Y							
Ekbal and Saha [23]		Y	Y	Y							

(continued)

Table 3 (continued)

Reference	Person name	First name	Middle name	Last name	Person prefix	Person suffix	Designation	Location	Organization name	Organization suffix	Disease
Gali et al. [28]											
Gupta and Bhattacharyya [33]	Y				Y	Y		Y	Y		
Hasanuzzaman et al. [37]		Y	Y	Y							
Jain and Arora [41]	Y								Y		Y
Jain and Arora [42]	Y										Y
Jain et al. [43]	Y										Y
Saha et al. [66]		Y	Y	Y				Y			
Saha et al. [67]		Y	Y	Y	Y			Y	Y	Y	
Saha et al. [68]		Y		Y				Y			
Saha et al. [69]		Y	Y	Y				Y		Y	
Saha et al. [71]					Y	Y	Y	Y		Y	
Saha et al. [72]							Y	Y		Y	
Sarkar and Shaw [76]	Y							Y	Y		
Sharma and Goyal [82]	Y				Y			Y	Y		

(continued)

Table 3 (continued)

Reference	Consumable	Symptom	Entertainment	Function words	Season name	Month name	Week day	Time expression	Numerals	Measurement
Biswas et al. [3]						Y				Y
Cucerzan and Yarowsky [9]										
Devi et al. [10]			Y							
Ekbal and Bandyopadhyay [13]				Y		Y	Y			Y
Ekbal et al. [14]						Y	Y			Y
Ekbal and Bandyopadhyay [15]				Y		Y	Y			Y
Ekbal and Bandyopadhyay [17]										
Ekbal and Saha [18]				Y		Y	Y			Y
Ekbal and Saha [23]				Y		Y	Y			Y
Gali et al. [28]								Y	Y	Y
Gupta and Bhattacharyya [33]										
Hasanuzzaman et al. [37]				Y		Y	Y			Y
Jain and Arora [41]	Y									
Jain and Arora [42]	Y	Y								
Jain et al. [43]	Y	Y								
Saha et al. [66]						Y	Y			
Saha et al. [67]						Y	Y			
Saha et al. [68]										

(continued)

Table 3 (continued)

Reference	Consumable	Symptom	Entertainment	Function words	Season name	Month name	Week day	Time expression	Numerals	Measurement
Saha et al. [69]						Y	Y			
Saha et al. [71]					Y	Y	Y			Y
Saha et al. [72]										
Sarkar and Shaw [76]				Y		Y	Y			Y
Sharma and Goyal [82]										

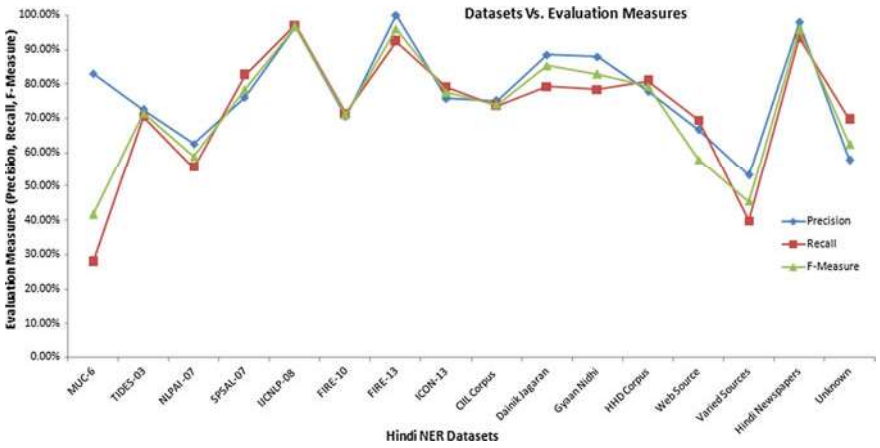

Fig. 5 Datasets versus evaluation measures

4 Contemporary Challenges in Hindi NER

Named Entity Recognition in the Hindi language is much more difficult and challenging as compared to the English. It is due to the wide range of problems that persist in the Hindi NER [45, 61, 78, 86] some of them are discussed here.

- The Hindi language lacks capitalization information which is an important cue for the named entity recognition in the English.
- Although Hindi has a very old and rich literary history, its technological development has a recent origin.
- Hindi names are more diverse and mostly appear as common nouns in the language.
- Hindi is a resource-constrained language- annotated corpus, dictionary, morphological analyzer, part-of-speech tagger and others in the required measure.
- Hindi is a relatively free-word order and highly inflectional language in nature.
- Although gazetteers are easily available for the English, they aren't readily available for the Hindi.
- The Hindi language lacks standardization and spelling of names which result in NE ambiguities.

In addition to the general Hindi NER problems, there are several other domain-specific challenges; some of them are discussed here for the health domain [42].

- NEs e.g. कुटकी चिरौता(Kutki Chirauta) is a consumable NE with rare occurrence within HHD corpus.
- Lack of standardization within abbreviations e.g. Dr. is an abbreviation of Doctor which is represented in Hindi as डॉ, डॉ, डा.
- Postposition marker such as में(in) is present or not with context e.g. सिर में दर्द, सिर दर्दboth represent (Headache).

- Variation in semantic information e.g. मधुमेह, इक्षुप्रमेह, डायबीटीज़ all represent disease NE- (Diabetes).
- Hindi doesn't use the capitalization concept e.g. Acquired Immune Deficiency Syndrome (AIDS) is represented in Hindi as अक्वायर्ड इम्युनोडेफिशिएंसी सिंड्रोम (एड्स).
- The ambiguity between NEs e.g. बुखार(Fever) and खांसी(Cough) can be disease NE or symptom NE.

5 Conclusion

Named entity recognition serves as a vital area of NLP as it has several useful applications. This chapter presents a holistic analysis of NER in the Hindi language while assembling major research works that are done so far in this regard. Major research articles on Hindi NER that are published from 1999 to 2019 are comprehensively reviewed. These articles are examined from six rationale aspects- datasets, techniques, training corpora, named entities, gazetteer lists, and evaluation measures. On unfolding these Hindi NER aspects, researchers gain a better insight into existing problems and proposed solutions for Hindi NE identification and classification. It is highlighted that the widely used Hindi NER datasets are- IJCNLP, Gyaan Nidhi, CIIL, TIDES, SPSAL, Dainik Jagaran, NLPAI, FIRE, ICON, and HHD. Among the NER techniques, a hybrid approach which combines MaxEnt, heuristic rules and bootstrapping gives the best result having F-measure as 96.67% over IJCNLP dataset. Also, the IJCNLP-08 has the highest number of NEs count in the training dataset (43,021 NEs) and in test dataset (3,005 NEs) with the most extensively used NEs as a person, location, and organization. In addition, several other domains specific NEs such as disease, symptom and consumable NEs have emerged for the health corpus with competitive F-measure.

Future Research Directions: Despite the importance of Hindi language, only a few language processing resources are available for researchers such as gazetteers, POS taggers, Hindi WordNet and others. In this regard, there is an abundant research opportunity to fulfill the demand for comprehensive investigation in the Hindi language, especially for the Hindi NER system. Such a task is even more challenging due to lack of appropriate tools and morphological structure of the Hindi. The advancement of Hindi processing tools such as NLP toolkit, well designed NER systems, extensive gazetteers, huge annotated corpora, and computational lexicons are some of the open areas for the research. These tools can facilitate researchers to apply existing approaches in the NER task, as well as many other areas of the Hindi language. Language independent and language dependent feature selection aspect can be investigated. In addition, efforts can be directed towards semi-supervised learning to reduce annotation task, and to provide robust performance for the NER task. Another challenge is to devise NER related models for Twitter and other micro-blogs to handle linguistically challenging contexts over there. While keeping in mind the importance of NER and its applications, there is dire attention for NLP researchers to explore

Hindi Named Entity Recognition in leaps and bounds. Other Indian languages- Bengali, Tamil, Oriya, Urdu, and Telugu also need to be examined.

References

1. Athavale V, Bharadwaj S, Pamecha M, Prabhu A, Shrivastava M (2016) Towards deep learning in Hindi NER: an approach to tackle the labelled data scarcity. arXiv:1610.09756
2. Balyan R (2002) Dealing with Hinglish named entities in English corpora. In: Speech & natural language processing lab, CDAC, Noida, India
3. Biswas S, Mishra MK, Acharya S, Mohanty S (2010) A two stage language independent named entity recognition for Indian Languages. Int J Comput Sci Inf Technol (IJCSIT) 1(4):285–289
4. Bontcheva K, Derczynski L, Roberts I (2017) Crowdsourcing named entity recognition and entity linking corpora. In: Handbook of linguistic annotation. Springer, pp 875–892
5. Chinchor N, Robinson P (1997) MUC-7 named entity task definition. In: Seventh conference on message understanding, vol 29, pp 1–21
6. Chopra D, Jahan N, Morwal S (2012) Hindi named entity recognition by aggregating rule based heuristics and Hidden Markov model. Int J Inf 2(6):43–52
7. Chopra D, Joshi N, Mathur I (2016) Named entity recognition in Hindi using Hidden Markov model. In: 2nd International conference on computational intelligence & communication technology (CICT), pp 581–586. IEEE
8. Cohen WW, Sarawagi S (2004) Exploiting dictionaries in named entity extraction: combining Semi-Markov extraction processes and data integration methods. In: 10th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 89–98
9. Cucerzan S, Yarowsky D (1999) 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, pp 90–99
10. Devi RG, Veena PV, Kumar, A. M., Soman, K. P.: AMRITA-CEN@ FIRE 2016: Code-mix entity extraction for Hindi-English and Tamil-English tweets. In: CEUR workshop proceedings, vol 1737, pp 304–308
11. Durrett G, Klein D (2014) A joint model for entity analysis: coreference, typing, and linking. In: Transactions of the association for computational linguistics, vol 2, pp 477–490
12. Ekbal A, Bandyopadhyay S (2007) A Hidden Markov model based named entity recognition system: Bengali and Hindi as case studies. In: International conference on pattern recognition and machine intelligence. Springer, pp 545–552
13. Ekbal A, Bandyopadhyay S (2008) Named entity recognition in Indian languages using maximum entropy approach. Int J Comput Process Lang 21(03):205–237
14. Ekbal A, Haque R, Das A, Poka V, Bandyopadhyay S (2008) Language independent named entity recognition in Indian languages. In: IJCNLP-08 workshop on named entity recognition for South and South East Asian Languages, pp 33–40
15. Ekbal A, Bandyopadhyay S (2009) A Conditional random field approach for named entity recognition in bengali and Hindi. Linguist Issues Lang Technol 2(1):1–44
16. Ekbal A, Bandyopadhyay S (2009) A multiengine NER system with context pattern learning and post-processing improves system performance. Int J Comput Process Lang 22(02n03):171–204
17. Ekbal A, Bandyopadhyay S (2010) AEkbalSBandyopadhyay2010Named entity recognition using support vector machine: a language independent approach. Int J Electr Comput Syst Eng 4(2):155–170
18. Ekbal A, Saha S (2010) Classifier ensemble selection using genetic algorithm for named entity recognition. Res Lang Comput 8(1):73–99
19. Ekbal A, Saha S (2010) Weighted vote based classifier ensemble selection using genetic algorithm for named entity recognition. In: International conference on application of natural language to information systems. Springer, pp 256–267

20. Ekbal A, Bandyopadhyay S (2011) Named entity recognition in Bengali and Hindi using support vector Machine. Lingvist Investig 34(1):35–67
21. Ekbal A, Saha S (2011) A Multiobjective simulated annealing approach for classifier ensemble: named entity recognition in Indian languages as case studies. Expert Syst Appl 38(12):14760–14772
22. Ekbal A, Saha S (2011) Weighted vote-based classifier ensemble for named entity recognition: a genetic algorithm-based approach. ACM Trans Asian Lang Inf Process (TALIP) 10(2):1–37
23. Ekbal A, Saha S (2012) Multiobjective optimization for classifier ensemble and feature selection: an application to named entity recognition. Int J Doc Anal Recognit (IJDAR) 15(2):143–166
24. Ekbal A, Saha S, Singh D (2012) Active machine learning technique for named entity recognition. In: International conference on advances in computing, communications and informatics. ACM, pp 180–186
25. Ekbal A, Saha S, Singh D (2012) Ensemble based active annotation for named entity recognition. In: 3rd international conference on emerging applications of information technology (EAIT). IEEE, pp 331–334
26. Ekbal A, Saha S, Sikdar UK (2016) On active annotation for named entity recognition. Int J Mach Learn Cybern 7(4):623–640
27. Etzioni O, Cafarella M, Downey D, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A (2005) Unsupervised named-entity extraction from the web: an experimental study. Artif Intell 165(1):91–134
28. Gali K, Surana H, Vaidya A, Shishtla P, Sharma DM (2008) Aggregating machine learning and rule based heuristics for named entity recognition. In: IJCNLP-08 workshop on named entity recognition for South and South East Asian Languages, pp 25–32
29. Gayen V, Sarkar K (2014) An HMM based named entity recognition system for Indian languages: the JU system at ICON 2013. arXiv:1405.7397 (2014)
30. Goyal A (2008) Named entity recognition for South Asian Languages. In: IJCNLP-08 workshop on named entity recognition for South and South East Asian Languages, pp 89–96
31. Goyal A, Gupta V, Kumar M (2018) Recent named entity recognition and classification techniques: a systematic review. Comput Sci Rev 29:21–43
32. Gupta PK, Arora S (2009) An approach for named entity recognition system for Hindi: an experimental study. In: ASCNT–2009, CDAC, Noida, India, pp 103–108
33. Gupta S, Bhattacharyya P (2010) Think globally, apply locally: using distributional characteristics for Hindi named entity identification. In: 2010 Named entities workshop, association for computational linguistics, pp 116–125
34. Gupta JP, Tayal DK, Gupta A (2011) A TENGRAM method based part-of-speech tagging of multi-category words in Hindi language. Expert Syst Appl 38(12):15084–15093
35. Gupta V, Lehal GS (2011) Named entity recognition for punjabi language text summarization. Int J Comput Appl 33(3):28–32
36. Gupta, V.: Hybrid multilingual key terms extraction system for Hindi and Punjabi text. In: Progress in systems engineering. Springer, pp 715–718
37. Hasanuzzaman M, Ekbal A, Bandyopadhyay S (2009) Maximum entropy approach for named entity recognition in Bengali and Hindi. Int J Recent Trends Eng 1(1):408–412
38. Hasanuzzaman M, Saha S, Ekbal A (2010) Feature subset selection using genetic algorithm for named entity recognition. In: 24th Pacific Asia conference on language, information and computation, pp 153–162
39. Hiremath P, Shambhavi BR (2014) Approaches to named entity recognition in Indian languages: a study. Int J Eng Adv Technol (IJEAT), 3(6):191–194, ISSN:2249-8958
40. Jain A, Yadav D, Tayal DK (2014) NER for Hindi language using association rules. In: International conference on data mining and intelligent computing (ICDMIC). IEEE, pp 1–5
41. Jain A, Arora A (2018) Named entity system for tweets in Hindi Language. Int J Intell Inf Technol (IJIIT), 14(4):55–76 (IGI Global)
42. Jain A, Arora A (2018) Named entity recognition in hindi using hyperspace Analogue to Language and conditional random field. Pertanika J Sci Technol 26(4):1801–1822

43. Jain A, Tayal DK, Arora A (2018) OntoHindi NER—An ontology based novel approach for Hindi named entity recognition. Int J Artif Intell 16(2):106–135
44. Kale S, Govilkar S (2017) Survey of named entity recognition techniques for various indian regional languages. Int J Comput Appl 164(4)
45. Kaur D, Gupta V (2010) A survey of named entity recognition in English and other Indian languages. Int J Comput Sci Issues (IJCSI) 7(6):239–245
46. Kaur Y, Kaur ER (2015) Named entity recognition system for Hindi language using combination of rule based approach and list look up approach. Int J Sci Res Manage (IJSRM) 3(3):2300–2306
47. Khalid MA, Jijkoun V, De Rijke M (2008) The impact of named entity normalization on information retrieval for question answering. In: European conference on information retrieval. Springer, pp 705–710
48. Krishnarao AA, Gahlot H, Srinet A, Kushwaha DS (2009) A comparative study of named entity recognition for Hindi Using sequential learning algorithms. In: International advance computing conference (IACC). IEEE, pp 1164–1169
49. Kumar N, Bhattacharyya P (2006) Named entity recognition in Hindi using MEMM. Technical Report, IIT Mumbai
50. Kumar NK, Santosh GSK, Varma V (2011) 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. Springer, pp 74–82
51. Kumar P, Kiran RV (2008) A hybrid named entity recognition system for South Asian Languages. In: IJCNLP-08 workshop on NER for South and South East Asian Languages, Hyderabad, India, pp 83–88
52. Li W, McCallum A (2003) Rapid development of Hindi named entity recognition using conditional random fields and feature induction. ACM Trans Asian Lang Inf Process (TALIP) 2(3):290–294
53. Mandalia C, Rahil MM, Raval M, Modha S (2015) Entity extraction from social media text Indian languages (ESM-IL). In: FIRE workshops, pp 100–102
54. Maynard D, Tablan V, Ursu C, Cunningham H, Wilks Y (2001) Named entity recognition from diverse text types. In: Conference on recent advances in natural language processing, Tzigov Chark, Bulgaria, pp 257–274
55. Moro A, Raganato A, Navigli R (2014) Entity linking meets word sense disambiguation: a unified approach. Trans Assoc Comput Linguist 2:231–244
56. Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvist Investig 30(1):3–26
57. Nanda M (2014) The named entity recognizer framework. Int J Innov Res Adv Eng (IJIRAE), 1(4):104–108. ISSN: 2349-2163
58. Nayan A, Rao BRK, Singh P, Sanyal S, Sanyal R (2008) Named entity recognition for Indian languages. In: IJCNLP-08 Workshop on named entity recognition for South and South East Asian Languages, pp 97–104
59. Pallavi KP, Srividhya K, Victor RRJ, Ramya MM (2015) HITS@ FIRE Task 2015: twitter based named entity recognizer for Indian languages. In: FIRE workshops, pp 81–84
60. Patel A, Ramakrishnan G, Bhattacharya P (2009) Incorporating linguistic expertise using ILP for named entity recognition in data hungry Indian languages. In: International conference on inductive logic programming. Springer, pp 178–185
61. Patil N, Patil AS, Pawar BV (2016) Survey of named entity recognition systems with respect to Indian and Foreign languages. Int J Comput Appl 134(16):21–26
62. Patra R, Saha SK (2019) A novel word clustering and cluster merging technique for named entity recognition. J Intell Syst 28(1):15–30
63. Przybyła P (2016) Boosting question answering by deep entity recognition. arXiv:1605.08675
64. Rahman A, Ng V(2010) Inducing fine-grained semantic classes via hierarchical and collective classification. In: 23rd international conference on computational linguistics (COLING 2010), association for computational linguistics pp 931–939

65. Rao PR, Malarkodi CS, Ram RVS, Devi SL (2015) ESM-IL: entity extraction from social media text for Indian languages@ FIRE 2015-an overview. In: FIRE workshops, pp 74–80
66. Saha SK, Chatterji S, Dandapat S, Sarkar S, Mitra P (2008) A hybrid approach for named entity recognition in Indian languages. In: IJCNLP-08 workshop on NER for South and South East Asian Languages, pp 17–24
67. Saha SK, Sarkar S, Mitra P (2008) A hybrid feature set based maximum entropy Hindi Named entity recognition. In: 3rd International joint conference on natural language processing, vol 1, pp 343–349
68. Saha SK, Sarkar S, Mitra P (2008) Gazetteer Preparation For Named Entity Recognition in Indian languages. In: 6th Workshop on Asian language resources, pp 9–16
69. Saha SK, Ghosh PS, Sarkar S, Mitra P (2008) Named entity recognition in hindi using maximum entropy and transliteration. Polibits 38:33–41
70. Saha SK, Mitra P, Sarkar S (2008) Word clustering and word selection based feature reduction for maxent based Hindi NER. In: ACL-08: HLT, association for computational linguistics, Columbus, Ohio, USA, pp 488–495
71. Saha SK, Sarkar S, Mitra P (2009) Hindi named entity annotation error detection and correction. In: Language forum, vol 35, no 2. Bahri Publications, pp 73–93
72. Saha SK, Mitra P, Sarkar S (2009) A semi-supervised approach for maximum entropy based Hindi named entity recognition. In: International conference on pattern recognition and machine intelligence. Springer, pp 225–230
73. Saha SK, Narayan S, Sarkar S, Mitra P (2010) A composite Kernel for named entity recognition. Pattern Recogn Lett 31(12):1591–1597
74. Saha SK, Mitra P, Sarkar S (2012) A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition. Knowl-Based Syst 27:322–332
75. Saha S, Ekbal A (2013) Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. Data Knowl Eng 85:15–39
76. Sarkar K, Shaw SK (2017) A memory-based learning approach for named entity recognition in Hindi. J Intell Syst 26(2):301–321
77. Sarkar K (2018) Hindi named entity recognition using system combination. Int J Appl Pattern Recogn 5(1):11–39
78. Sasidhar B, Yohan PM, Babu AV, Govarhan A (2011) A survey on named entity recognition in Indian Languages with particular reference to Telugu. Int J Comput Sci Issues 8(2):438–443
79. Sekine S, Ranchhod E (eds) (2009) Named entities: recognition, classification and use, vol. 19. John Benjamins Publishing
80. Sharma P (2015) Named entity recognition for a resource poor indo-aryan language. PhD Thesis, Department of Computer Science and Engineering School of Engineering, Tezpur University, India
81. Sharma P, Sharma U, Kalita J (2011) Named entity recognition: a survey for the indian languages. In: Parsing in Indian languages, pp 35–39
82. Sharma R, Goyal V (2011) Name entity recognition systems for Hindi using CRF approach. In: International conference on information systems for Indian languages. Springer, pp 31–35
83. Sharnagat R, Bhattacharyya P (2013) Hindi named entity recognizer for NER task of FIRE 2013. In: FIRE-2013
84. Shishtla P, Pingali P, Varma V (2008) A character n-gram based approach for improved recall in Indian language NER. In: IJCNLP-08 workshop on named entity recognition for South and South East Asian Languages pp 67–74
85. Sikdar UK, Ekbal A, Saha S (2012) Differential evolution based feature selection and classifier ensemble for named entity recognition. COLING 2012:2475–2490
86. Singh AK (2008) Named entity recognition for South and South East Asian Languages: taking stock. In: IJCNLP-08 workshop on named entity recognition for South and South East Asian Languages, pp 5–16
87. Singh K, Sen I, Kumaraguru P (2018) Language identification and named entity recognition in Hinglish code mixed tweets. In: ACL 2018 student research workshop, pp 52–58

88. Singh V, Vijay D, Akhtar SS, Shrivastava M (2018) Named entity recognition for Hindi-English code-mixed social media text. In: Seventh named entities workshop, pp 27–35
89. Srivastava S, Sanglikar M, Kothari DC (2011) Named entity recognition system for Hindi language: a hybrid approach. Int J Comput Linguist (IJCL) 2(1):10–23
90. Ugawa A, Tamura A, Ninomiya T, Takamura H, Okumura M (2018) Neural machine translation incorporating named entity. In: 27th international conference on computational linguistics, pp 3240–3250
91. Witten IH, Bray Z, Mahoui M, Teahan WJ (1999) Using language models for generic entity extraction. In: ICML workshop on text mining, pp 1–11
92. Zhu J, Uren V, Motta E (2005) ESpotter: adaptive named entity recognition for web browsing. In: Biennial conference on professional knowledge management/wissens management. Springer, pp 518–529