ISSUES AND CHALLENGES IN MARATHI NAMED ENTITY RECOGNITION

NITA PATIL, AJAY S. PATIL AND B. V. PAWAR

School of Computer Sciences, North Maharashtra University, Jalgaon

ABSTRACT

Information Extraction (IE) is a sub discipline of Artificial Intelligence. IE identifies information in unstructured information source that adheres to predefined semantics i.e. people, location etc. Recognition of named entities (NEs) from computer readable natural language text is significant task of IE and natural language processing (NLP). Named entity (NE) extraction is important step for processing unstructured content. Unstructured data is computationally opaque. Computers require computationally transparent data for processing. IE adds meaning to raw data so that it can be easily processed by computers. There are various different approaches that are applied for extraction of entities from text. This paper elaborates need of NE recognition for Marathi and discusses issues and challenges involved in NE recognition tasks for Marathi language. It also explores various methods and techniques that are useful for creation of learning resources and lexicons that are important for extraction of NEs from natural language unstructured text.

KEYWORDS

Named Entity Recognition, Information Extraction, Issues, Challenges, Marathi, Techniques & Resources

1. INTRODUCTION

Named Entity Recognition (NER) is one of the important sub tasks of Information Extraction (IE). Named Entities (NEs) are noun phrases in the natural language text. Natural language text is a sequence of sentences, where sentence in turn is a sequence of words and punctuations combined to add semantic to the text. Further, a word is a character sequence. In NER the aim is to distinguish between character sequence that represent noun phrases and character sequence that represents normal text. A proper noun in the text that is used to refer person, company, location etc. is called as Named Entity (NE). The main types of NEs are shown in Fig. 1

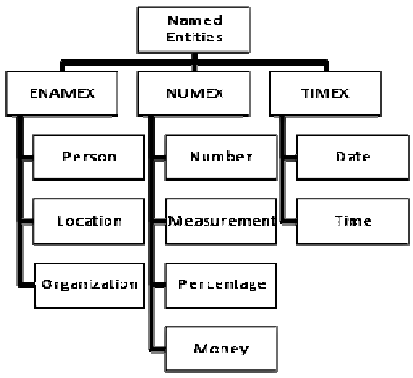

Figure 1. Types of Named Entities

DOI: 10.5121/ijnlc.2016.5102

Proper nouns play vital role in discovery of semantics hidden in the text. Identification of proper nouns in the raw text and classification of identified proper nouns in appropriate category is an important sub task of information extraction called as Named Entity Recognition (NER). Any entity such as person, organization, geographical location, government agency, event that has a name is treated as a named entity. Expressions such as money, percentage, phone numbers, date, time, URLs, addresses are also treated as named entities even if they are not exactly like traditional proper nouns [1]. Marking NEs in natural language text is significant pre-processing step useful for NLP applications like business intelligence, summarizations, machine translations and question answering systems etc. NER is a significantly difficult task. Example given below depicts many complexities involved in NER task.

[PER: Washington's] daughter and [ORG: GM] chief [PER: Mary Barra] apologized for the [LOC: US] automaker's failure in [LOC: Washington] Metropolitan Area in [LOC: United States].

PER, ORG and LOC labels in above sentence represent Person, Organization and Location entities respectively. This sentence contains six instances of named entities that show many complexities related to NER problem. The example sentence includes first instance of the name Washington which is used to refer person whereas second instance Washington is a location name, which is ambiguous. GM is an acronym for the General Motors, which is a well known car manufacturer in US. Mary Barra is a two word chunk that indicates a single person. Further, US and United States are two different chunks that should be identified with same label LOC.

Indian languages have originated from Sanskrit whereas Indic scripts have originated from Brahmi. In ancient India, Sanskrit was spoken by Brahmins and Prakrit was the language used in general by common people. Marathi has originated from Prakrit. Marathi is the official language of Maharashtra state in India and is recognized as one of the twenty-two official languages of the country by Constitution of India. Very limited Indian population can either read or write English. Over 90 percent population is normally far from taking benefits of English based Information Technology. As International boundaries are becoming narrower day by day, language should not be barrier for the common people to seek information. Importance of natural language processing is dramatically growing. A big enhancement in information technology is agreed to involve semantic search. Similar to other Indian languages (Bengali, Hindi, Telugu) NLP tool development is necessary to enrich the Marathi language technology.

2. RELATED WORK

Research in NER has been surveyed by various researchers from time to time from various aspects. Significant work in NER began after 1987. Survey of the progress in NER during 1991 to 2006 can be found in David et. al.[2]. Their study has reported various NE types, domains, and genres, various learning methods of NER and performance evaluation techniques in detail. Their work has also explored word-level, dictionary-level and corpus-based features of text in natural language. Khaled Shaalan [3] describes the increase in interest and progress made in Arabic NER research, the importance of the NER task, the main characteristics of the Arabic language, the aspects of standardization in annotating named entities, linguistic resources, approaches, features of common tools and standard evaluation metrics used in Arabic NER. Sasidhar et. al. [4] studied named entity recognition and classification (NERC) and presented various NERC system development approaches and NERC system evaluation techniques with respect to Indian languages. They highlighted that research in NER is difficult for Indian languages. Padmaja et. al.[5] had presented a overview of NER , issues in the Indian languages, different approaches

used in NER and research in NER in different Indian languages like Bengali, Telugu, Hindi, Oriya and Urdu along with the methodologies used. The objectives of the survey mentioned in this paper are to assimilate resources useful for development of NERC system for Indian language Marathi, to explore challenges in named entity extraction from Marathi text, to study techniques used by other researchers and to investigate steps involved in development of resources required to pinpoint suitable technique for Marathi language NERC system.

3. ISSUES AND CHALLENGES IN MARATHI NER

Implementation of Marathi NER system is challenging because of difficulties introduced in recognition. NER is very challenging task. It becomes much more difficult and challenging for Marathi language. The issues elaborated here depict the major challenges in Marathi NER.

3.1. Ambiguities in named entity classes

Natural language text is simply a series of language character set. It is very hard to distinguish between the characters that represent proper nouns and characters that represent normal text. It is also challenging to classify the similar sequence of words representing different classes in different instances. Ambiguities in names containing words that infer multiple meanings or a chunk that can be part of different types of NEs makes NER difficult. Word Sense Disambiguation (WSD) is required to provide inference ability to a system to determine that a chunk is actually a named entity or to determine the classification of a named entity. The following example depicts the overlap of person and organization entity class. The words दिनानाथ मंगेशकर overlaps both in the person and organization entity class.

[PER: दिनानाथ मंगेशकर] यांच्या स्मृती प्रित्यर्थ [ORG: दिनानाथ मंगेशकर मेमोरीयल] मध्ये आयोजीत कार्यक्रमात [PER: लता मंगेशकर] यांनी गीत गायले.

In program sponsored by [ORG: Dinanath Mangeshkar Memorial] [PER: Lata Mangeshkar] sang a song in memory of [PER: Dinanath Mangeshkar.]

3.2. Abbreviations and non local dependencies

Multiple tokens can be written in different ways such as abbreviations or long form, usually first instance with descriptive long formulation followed by instances with short forms or aliases. Such tokens sometimes require same label assignments or require cross referencing. This ability of a system refers to non local dependency. External knowledge is required to deal with non local dependencies. Construction of external knowledge including names lists and expansive lexicons is not easy since domain lexicons and names are continuously expanding. Named entities can be composed of single or multiple words chunk of text. Parsing prediction or name chunking model is required to predict whether consecutive multiple words belong to same entity. The following example depicts the issues mentioned above regarding Marathi text.

[MISC: आंतरविद्यालय जलतरण स्पर्धेत] [NUM: चार] [Full form of token: जिल्हे] सहभागी झाले. [Multiword PER: अनिकेत अजय पाटील] याने [MEASURE: प्रथम] पारितोषिक पटकावले. [Alias PER: अनिकेत] हा सेंट जोसेफ [Short form of word: जि.] जळगाव शाळेचा विद्यार्थी आहे.

[NUM: Four] districts participated in [MISC: interschool swimming competition]. [PER: Aniket Ajay Patil won [MEASURE: first] prize. [PER: Aniket] is student of [ORG: St. Joseph School, Jalgaon]

3.3. Foreign words

Foreign words appear in Marathi texts which are spelled in Devanagari script. Table 1 shows some instances of person, organization, location and miscellaneous names that are English words spelled in Devanagari script. It is very challenging to recognize such words. It is very difficult to create gazetteers that include such names because they are not limited.

Table 1. Named entities containing foreign words

Person	**Organization**	**Location**	**Miscellaneous**
डेव्हिड धवन (David Dhawan)	स्कूल ऑफ आर्ट्स (School of Arts)	यूपी (UP)	वन वर्ल्ड फेलोशिप स्कीम (One World Fellowship Scheme)
ए. के. पुरवार (A. K. Purvar)	स्टुडिओ लिंक (Studio Link)	कॅनडा (Canada)	न्युक्लिअस बजेट योजना (Nucleus Budget Yojana)
विल्फ्रेड डिसूझा (Wilfred D'souza)	एनएफडीसी (NFDC)	इंग्लंड (England)	बॉम्बे टू मॉरिशस (Bombay to Mauritius)

3.4. Agglutinative and inflectional nature of Marathi

Marathi is agglutinative language. Unlike English prefixes and suffixes are added to root words in Marathi to form meaningful contexts. Sometimes in Marathi when suffixes or prefix are added to root word it changes it's semantic also. The word forms such as अबोल, निंबोल are formed when prefix is added to word बोल (to talk). The agglutinative nature of Marathi language can be seen in table 2 that shows some variations of word बोल (from the IIT Bombay FIRE-2010 corpus) after adding suffix/prefix to it.

Table 2. Variations of word बोल

बोलू	बोलता	बोलही	बोलण्याने	बोललात	बोलणाऱ्याला	बोलवायला
बोल	बोलती	बोलीन	बोलण्याला	बोललास	बोलण्यातील	बोलविण्यात
बोलकी	बोलते	बोलून	बोलण्यास	बोलवणे	बोलण्यातून	बोलविलेली
बोलके	बोलतो	बोलणार	बोलतच	बोलवत	बोलण्यामुळे	बोलविल्या
बोलक्या	बोललं	बोलणारे	बोलतांना	बोलवावी	बोलण्यावर	बोलल्याबद्दल
बोलणं	बोलला	बोलणाऱ्या	बोलतात	बोलविता	बोलतानाही	अबोलपणानं
बोलणी	बोलली	बोलण्याचा	बोलताना	बोलविली	बोलपटात	बोलण्यावरुन
बोलणे	बोलले	बोलण्याची	बोलताही	बोलविले	बोलल्यावर	बोलता-बोलता
बोलतं	बोललो	बोलण्याचे	बोलबाला	बोलवून	बोलवण्यात	अबोल
बोलत	बोलल्या	बोलण्यात	बोलभांड	बोलणाऱ्यांची	बोलवायचं	निंबोल

It is very difficult to use gazetteers, dictionaries, similarity measurement and pattern matching techniques to recognize Marathi names. Dictionaries or gazetteers contain entities without any suffix added. In Marathi suffixes are added to words in order to create the meaningful context.

Table 3. Named entities instances and actual gazetteer Record

Named Entity	**Gazetteer Entry**
मधुबालाच्या (Madhubala's)	मधुबाला
फ्लिंटॉफकरवी (from Flintop)	फ्लिंटॉफ
खानदेश एज्युकेशन सोसायटीचं (of Khandesh Education Society)	खानदेश एज्युकेशन सोसायटी
राष्ट्रवादी काँग्रेसमधील (in Rashtravadi Congress)	राष्ट्रवादी काँग्रेस
अथेन्समध्ये (in Athens)	अथेन्स
वाशीहून (from Washi)	वाशी

Table 3 illustrates the difficulties found in use of list lookup and pattern matching technique for Marathi text. A well written stemmer is required for morphologically rich language Marathi to separate the root from the suffix in order to compare the word forms with gazetteer or dictionary entries. It cannot be claimed that stemming will solve the problem completely because adding suffixes to roots may change the grammatical category of the root word, which may result in wrong entity recognition. Table 4 depicts how entity class changes after adding suffix to the tokens.

Table 4. Named entities class before and after stemming

Entity class before stemming	**Entity class after stemming**
गरवारे महाविद्यालयाजवळच (near Garware College)	गरवारे महाविद्यालय
थाई ऍडव्हर्सरी क्लबवर (on Thai Adversary Club)	थाई ऍडव्हर्सरी क्लब
कोपरी प्रभाग कार्यालयाजवळील (near Kopari Divisional Office)	कोपरी प्रभाग कार्यालय
पुणे कॅन्टोन्मेंटमधून सरळ (straight to Pune Cantonment)	पुणे कॅन्टोन्मेंट
दामोदर हॉलच्यामागे (behind Damodar Hall)	दामोदर हॉल

3.5. Spelling variations

In Marathi, words containing the four vowels इ (i), vowel sign (ि), or ई (ī), vowel sign (ी), उ (u), vowel sign (ु), ऊ (ū), vowel sign(ू) do not make phonetic difference but differs in writing and spellings. The words such as आणि (grammatically correct) and आणी (grammatically incorrect) or पाणी (grammatically correct) and पानी (grammatically incorrect) are interchangeably used in many writings. Marathi grammar may not be followed completely in free style text writing. This affects text recognition systems. There is lack of uniformity in writing spellings in Marathi. Some Marathi words shown in Table 5 can be written in either or other form in text which may be used alternately to point same thing.

Table 5. Lack of uniformity in writing

Word form1	Word form2
एनक्विस्टकडून	एनक्वीस्टकडुन
भुतियाखेरीज	भुतीयाखेरिज
मुंबई टाइम्स	मुंबई टाईम्स
काश्मिरी	काश्मीरी
हंगेरीच्या	हंगेरिच्या

3.6. Encoding related issues

Each language uses its own character set in legacy encodings. Characters written using one native character encoding may not be displayed correctly by another encoding system [6]. Marathi text is encoded using various fonts and systems. If a document is opened on computer that does not support the font or system using which the document is written, then text in document is displayed with unreadable characters and becomes unusable. Sometimes document created using one computer with a particular operating system cannot fully display the text on computer with same configuration but with another operating system. In such situation some characters are readable but some are displayed using hollow circles or wrongly spelled. This is called as partly corrupted text. Table 6 shows some instances of text typed on one computer that appeared different on another.

Table 6. Problems with charter encodings

Incorrect version	Correct version
कैंद	कैंद्र
प्रश्ान्	प्रश्न
विलेपालेर्	विलेपार्ले
इंग्लंडतफेर्	इंग्लंडतर्फे
मंुबई	मुंबई

3.7. Dialects

Marathi is spoken using many dialects such as standard Marathi, Warhadi, Ahirani, Dangi, Vadvali, Samavedi, Khandeshi, and Malwani in various regions of India. There are specific words used in each dialect to express the text. Words from different dialects also appear in Marathi text. Some regional words, similar word used in standard Marathi and its meaning in English are illustrated in Table 7.

Table 7. Variations in dialects in Marathi

Dialect	Text Dialect	Standard Marathi	Meaning in English
Ahirani	सैपाक	स्वंयपाक	to cook
Warhadi	घेउन घ्या	घ्या	take
Warhadi	खिलविले	खाऊ घातले	feed
Warhadi	काहून?	का?	why?
Warhadi	लोचा	समस्या	problem
Konkani	उबी फूटपट्टी	अनुलंब मापनी	vertical ruler
Khandeshi	वावरात	शेतात	in farm

Annotated corpus, name gazetteers, dictionaries, morphological analyzers, POS taggers etc. are not available in Marathi in required quantity and quality. Marathi is relatively free order language.

4. METHODOLOGIES AND TECHNIQUES

A mechanism that can lookup, analyze and extract patterns required to distinguish noun phrases from the normal text is needed. Every word has characteristics that describe and designate it. Such characteristics are called as features. Features describing word can be used for identification and classification of NE's. The main techniques for named entity recognition are word-level feature based, list or dictionary lookup and corpus based recognition. Name finding is based on one of the following approaches:(1) hand-crafted or automatically acquired rules or finite state patterns (2) look up from large name lists or other specialized resources (3) data driven approaches exploiting the statistical properties of the language (statistical models) [7].

NER involves three significant subtasks. First is tokenization that is segmentation of text into tokens. Second is assignment of appropriate tag to segmented token. This task may result in ambiguous assignment of tags to the tokens. Third subtask is selection of correct tag for a token by resolving ambiguity introduced by second step. Tokenization includes identification of sentence and word form boundaries, segmenting text into tokens, marking the word boundaries and resolving ambiguities while segmenting the text into tokens. Next to tokenization is lexicon creation. Lexicon is a token with potential features and properties. Lexicons can be full formed, morphological or with statistical information. Most preferable way of lexicons for NE tagging generation is either using morphological analysis or derived from large corpus. Rule based NER uses morphological lexicons whereas machine learning techniques need statistical probabilities associated with token and tag pairs.

4.1. Word-level feature based recognition

Word level features describe attributes of individual tokens. These properties further can be used to identify token if it is NE and classify it into more appropriate NE class. Typical word level features useful for NER tasks are shown in fig. 2.

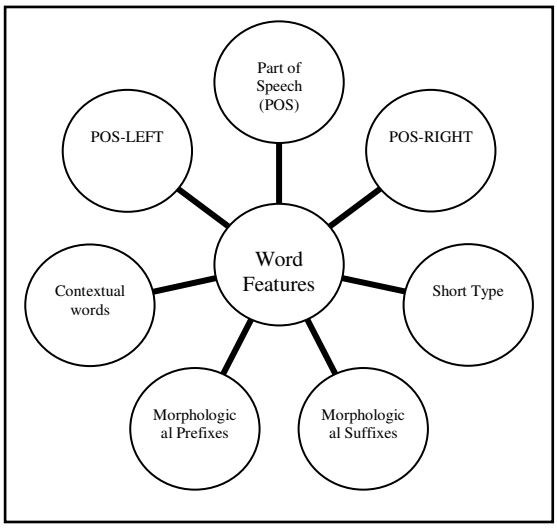

Figure 2. Word Level Features for NER

Lexicons used for rule based NER can be developed using morphological analysis and POS tagging. Following are the techniques that mainly focus on tokens, describes them with features that are helpful for NERC task.

4.1.1. Regular expressions (RE) lookup

RE lookup is mostly useful for segmentation of text into tokens and to describe patterns that match numerical entities such as monetary expression, dates, email address, phone numbers such as ([०-९]+[,])*[०-९]([.] [०-९]+)?. RE lookup can be applied for Marathi named entity recognition to identify numerical patterns, date and time expressions etc.

4.1.2. Morphological analysis

Morphological analysis is the process of decomposing words into its constituents [8]. It is used to build NER lexicons that are represented by finite state networks. NER lexicons consists of word form followed by its optional lemmas, morphological feature and NE tag. Heuristic are derived from morphology and semantic of input. Morphological analysis can give information about word form inflections that is beneficial for NE tagging [9]. Named Entity recognition techniques that are used to focus on tokens and their associated features are illustrated in fig.3.

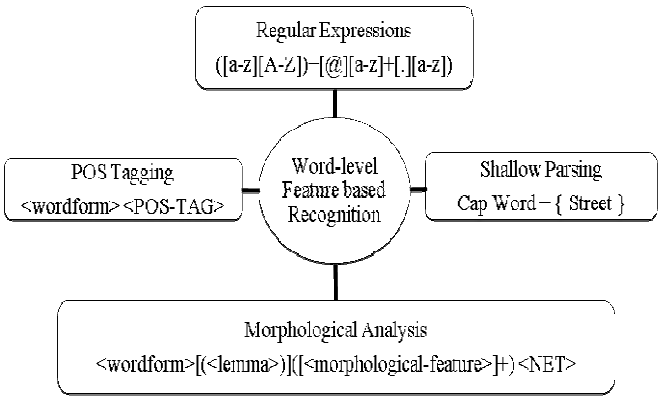

Figure 3. Word level Named Entity Recognition

4.1.3. POS tagging

Part of speech of a word (Table 8) plays vital role in NER [10]. POS patterns highlights word forms such as proper names that can be named entities, verbs, nouns etc. POS of token can mark it as a noun in Marathi text. Further nouns can be classified as common noun or proper noun. All the extracted proper nouns will be the candidates for classification into appropriate NE class such as people, location organization etc.

Table 8. Marathi Word with its POS tag and NE class

Marathi Word	POS tag	NE class
श्रुती	Proper Noun	Person
जळगाव	Proper Noun	Location
नदी	Common Noun	No entity
पुस्तक	Common Noun	No entity
निरमा	Proper Noun	Organization

4.1.4. Shallow parsing

A position of the word phrase in sentence helps in understanding its role and meaning in construction. Shallow parsing at syntactic level can analyze arrangement of words in construction. Proper names often have some implicit structure that can be used as internal evidence to identify its entity type. Names are also often surrounded by predictive words that can be used as external evidence in NER. Marathi language also has lot of predictive words that surrounds NE can act as external evidence for shallow parsing. For instance the word appearing before the word 'यांनी' (yanee), 'यांचे' (yanche) will definitely be a person name. Similarly the word appearing before 'येथे' (yethe), 'इथून' (ithoon) will be location, the token appeared before the word 'रोजी' (roji) will be date expression etc. Title person such as 'डॉ.' (Dr.), 'श्री.' (shri), 'सौ.' (sau), 'प्रा.' (pra) or designating words such as 'प्रधानमंत्री' (pradhanmantree), 'राज्यपाल' (Rajyapal), 'गृहमंत्री' (gruhmantri) also can be used to recognize people in the Marathi text. The expression <ता.> + {२ मार्च} + <रोजी> can be used to recognize date expressions in the text.

4.2. List lookup based recognition

Lists of named entities also called as gazetteers, dictionary or lexicon are compressive lists that includes names of organizations, people, continents, countries, capitals of countries, states, cities, geographical regions, government, celebrities, airlines, events, month names, days of week [2] etc. Word phrases are compared with entities in various gazetteers. For instance, if word phrase is found in list of countries then probability of that word with entity type location becomes higher. The list lookup can be done using either exact full string matching, or using filtering operations such as stemming or lemmatization, or using phonetic matching techniques such as soundex or editex or using approximate matching techniques which calculate distance between strings to be matched [11]. The advantages of this technique are that it is simple, fast and gives accurate results. But the downside is generation and expansion of lists is expensive. Lists cannot deal with variance, also cannot resolve ambiguity [12]. Words in sentences of the natural language text can have variations, inflections or derivations. Therefore techniques that resolve such variations are needed so that string matching can be done. Some useful techniques that can be used for matching words in test text against words in various lists, gazetteers or dictionaries are shown in fig. 4,

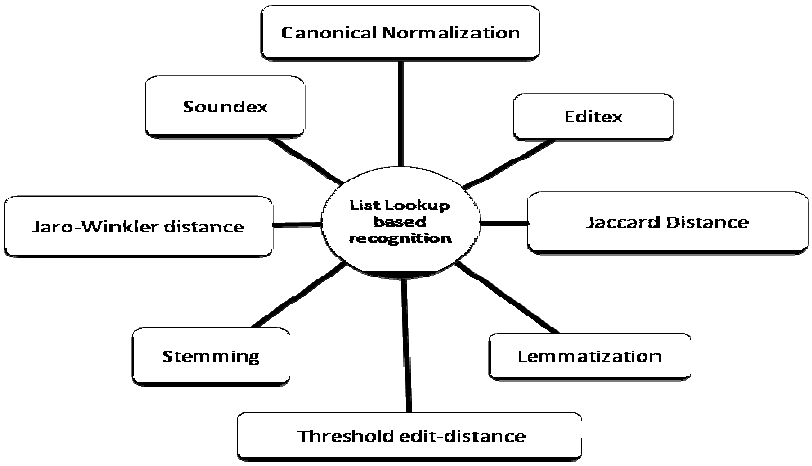

Figure 4. List lookup Techniques for Named Entity Recognition

4.2.1. Stemming

Stemming is the technique used to remove affixes from words. The part of the word which is common to all variants is called stem. Candidate words are trimmed out for inflectional and derivational suffixes as shown in Table 9. These stemmed words are further matched with words in lists to find its correct entity type.

Table 9. Stemming of Marathi Words

Inflected Word Form	Stemmed Word Form
सावरकरांच्या	सावरकर
सेहवागसह	सेहवाग
काँग्रेसने	काँग्रेस
राष्ट्रवादीतील	राष्ट्रवादी
दिल्लीत	दिल्ली

4.2.2. Lemmatization

Lemmatization is the technique that finds a match between the word and an entry in the language vocabulary or knowledge base. If the word gets partially matched then it generates list of possible lemma. Lemmatization includes basic word variations like singular vs. plural, thesaurus or synonym matching. It normalizes inflections. Table 10 shows some Marathi language instances which show need of lemmatization.

Table 10. Lemmatization

Singular	Plural		Word	Synonym
खुर्ची	खुर्च्या		वर्षा	पाऊस
बॅग	बॅगा		सुर्य	रवी
बाटली	बाटल्या		चंद्र	शशी
पाकळी	पाकळ्या		रात्र	निशा
फुलपाखरू	फुलपाखरे		आनंद	हर्ष

4.2.3. Canonical normalization

This technique can use Unicode normalizer. Diacritics are additional glyphs added to letters. They are mainly used to change the sound value of the letter in the word. For string matching ignoring diacritics is useful. Canonical normalization identifies characters that have the same meaning but different appearance and replaces them with a matching character. It is useful for lists matching to extract correct entity type of the word. e.g. स्वता -> स्वत:

4.2.4. Threshold edit-distance

This technique uses search and extension based algorithms. Edit distance technique is used to deal with typographic and spelling mistakes and to determine syntactic closure of two strings. The distance between two strings is defined as the minimal number of edits required to convert one into the other. Character level operations used are insertion, deletion, replacement and transpose. To find closest in a large dictionary entries are organized in trie, partitioned by length by assuming small edit distance called as threshold. e.g. केंद -> केंद्र.

4.2.5. Jaccard distance

Jaccard distance [13] measures dissimilarity between sample sets. It returns a distance between two items to measure how close or similar items altogether. The strings to be compared are first tokenized. Jaccard distance is computed by dividing the number of tokens shared by the strings to the total number of tokens. The comparisons are either character level, sentence level or n-gram level. e.g. पावसाळ्यात नदीकाठी असलेली हिरवळ मनमोहक असते. -> मनमोहक असलेली हिरवळ पावसाळ्यात नदीकाठी असते.

4.2.6. Jaro-Winkler distance

Distance between strings is based on vector similarity using the cosine measure and weighted term frequencies. Two strings are more similar if they contain many of the same tokens with the

same relative number of occurrences of each. Tokens are weighted more heavily if they occur in few documents [13].

4.2.7. Soundex

Soundex converts a word into its uttered sound. Basic idea is similarly spelled or pronounced words are decoded such that the words build the same code [14]. e.g. आणी -> आणि.

4.2.8. Editex

It is an enhancement of Levenshtein edit distance technique. Editex groups similar phonemes into equivalent classes. Editex straightforwardly combines Levenshtein distance with 'letter groups' (aeiouy, bp, ckq, etc.) such that letters in a similar group frequently correspond to similar phonemes [15]. e.g. मुंबई टाइम्स -> मुंबई टाईम्स

4.3. Machine learning using corpus based recognition

Word level and list look up based recognition gives accurate and high performance but needs exhaustive linguistic resources and grammar expertise. Systems developed with these approaches cannot be portable to other domains. Therefore many researchers have given preference to corpus based recognition. It is also known as stochastic or statistical machine learning. Fig. 5 explores data driven approach (Corpus based Recognition) for named entity recognition. Data driven approach is simply using supervised; semi supervised or unsupervised learning methods. Various statistical models used to design named entity recognition and classification systems are explained as follows,

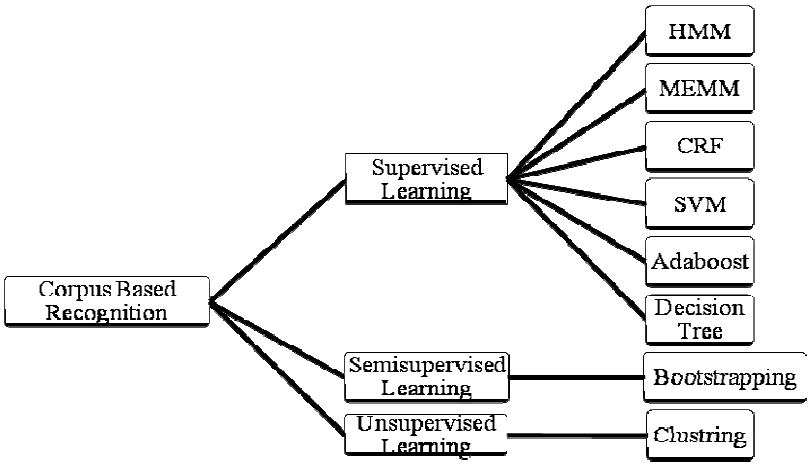

Figure 5. Machine learning approaches for Named Entity Recognition

4.3.1. Hidden Markov Models (HMM)

HMM is a statistical language model that computes the likelihood of a sequence of words by employing a Markov chain, in which likelihood of next word is based on the current word. In this language model words are represented by states, NE classes are represented by regions and there is a probability associated with every transition from the current word to the next word. This model can predict the NE class for a next word if current word and its NE class pair is given. The

Viterbi algorithm is used to find the sequence of NE classes with highest probability. श्री. रतन टाटा हे भारतातील एक य़शस्वी व्य़क्तीमत्व होय., in this sentence if, P ([रतन]|[श्री.], PER-NAME) > P ([रतन]|[श्री.], ORG-NAME) then, the word रतन will be person NE as its probability of being a person NE is greater than being an organization in a sentence. HMM is supervised learning algorithm. To develop a system that can recognize named entities using HMM needs tokenizer, large named entity tagged training corpus, N-GRAM language models, implementation of Viterbi algorithm for tagging and implementation of Baum-Welch algorithm to improve the parameters of HMM for unmarked texts.

4.3.2. Maximum Entropy Models (MEMM)

Maximum Entropy Model computes probability distribution based on maximum entropy that satisfies the constraints set by training examples. Entropy is measure of uncertainty and randomness of the event [16]. NER system implementation using Maximum Entropy needs tagged training corpus, tokenizer, language parser with at least a minimal amount of parsing technology to the system, morphological analyzer, ME model for tagging of unknown words.

4.3.3. Conditional Random Fields (CRF)

CRF is relational learning model.NER using CRF is based on undirected graphical model of conditionally trained probabilistic finite state automata. CRF is used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. It incorporates dependent features and context dependent learning. It allows representing dependencies on previous classifications in a discourse. The basic idea is context surrounding name becomes good evidence in tagging another occurrence of the same name in a different ambiguous context. NER using CRF implementation needs feature vector consisting of language features and POS tags, morphological analyzer, gazetteers and NE annotated corpus.

4.3.4. Support Vector Machines (SVM)

SVM is binary classification technique used to classify named entities. Each sample is represented by a long binary vector. The SVM is trained on a large training corpus. Each sample is represented by a long binary vector. Each training sample is represented by a point plotted into a hyperspace. Then SVM attempts to draw a hyper plane between the two categories, splitting the points as evenly as possible. This self training supervised learning model needs annotated text for instance 500,000 words tagged with its POS and IOB tag, morphology analyzer and SVM implementation.

4.3.5. Adaboost

In this method recognition is done using binary classifiers to label the words. In this task the context of the current word is used and neighboring word is coded as feature along with the relative position. Then NE classification is done by using multiclass multilabel AdaBoost algorithm [20].

4.3.6. Decision Trees (DT)

Supervised learning system using decision tree uses information derived from previous, current and next word by asking sequence of questions about the history to determine possible output of the model which is NE tag for a word. Decision Trees model consider a language model which

attempts to determine the next word in the text given a history consisting of the two previous words. The idea is to build a tree by asking questions at every point reducing uncertainty about the set of features. NER system implementation using decision tree needs segmentation of text into tokens, POS tag for a token and prefix or suffix word dictionaries and list of words that are candidate NEs.

4.3.7. Bootstrapping

Bootstrapping refers to the process by which an initial kernel of language is acquired by explicit instruction [17]. In bootstrapping set of seeds are used to start the process. The system then searches the sentences that have these seeds and tries to identify contextual clues. Then system tries to find other instances of seeds found in similar context.

4.3.8. Clustering

This is unsupervised learning technique useful for NERC problem. Automatic generation of groups in data based on numerical distance or similarity between objects is called clustering. Important factors in clustering are feature selection, similarity function and set of constraints such as shape of cluster or cluster membership, number of clusters etc.

Table 11. Statistical techniques, features and reported F1 measure for NERC

Technique	**Features used**	**F1 Score**
HMM[18]	Dictionaries, Hand crafted rules for numeric entities	90.93
MEMM[19]	Lists derived from training data, Local features, Global features, Name lists	83.31
CRF[20]	Word level features, Probability features, Suffix features, Bigram & Trigram features	93.65
SVM[21]	Contextual Window of size 6, Prefix, suffix up to length 3, POS information, Gazetteers, Lists derived from training data	91.80
Adaboost [22]	Features used are Lexical, Syntactic, Orthographic, Affixes, Word Type Patterns, Left Predictions, Bag-of-Word, Trigger Word and Gazetteer	85.00
Decision Trees[23]	Syntactic properties, POS information, Gazetteers.	94.28
Boot-strapping[17]	Labeled data, a seed classifier, gazetteers	81.62
Clustering[24]	Permanency feature (ratio of frequency of word in corpus to the frequency of all occurrences of the word with case insensitive consideration), Standard deviation of probability & the length of word in name entity.	66.21

Important factors in unsupervised approach clustering that affects named entity recognition are ambiguity in terms, conditional probability of the context for a specified semantic class and syntactic construct of the terms and context. This technique needs categorization of proper names (Person, Location, and Organization names), assimilation of seed set and their feature vectors, classification of unknown words based on shared context or context similarity with seeds. Similarity measurements based on shared context words, similar hypernym in WordNet, congruence of the syntactical function of the name in the sentence [25]. Table 11 summarizes various techniques used for implementation of named entity recognition.

5. CONCLUSIONS

Named entity recognition is very important task of Information Extraction systems. This task is used to give structure to the raw unstructured information written in natural language. NLP tool development is necessary to enrich the Marathi language technology. NER is difficult for Indian languages and implementation of Marathi NER system is much more difficult and challenging because of various issues like the inherent agglutinative and inflectional nature of Marathi, ambiguities in named entity classes, non local dependencies, appearances of foreign words, spelling variations etc. This paper has explored various methodologies and techniques that may be used in designing Marathi named entity recognition system.

ACKNOWLEDGEMENTS

This research work is supported by grants under Vice Chancellor Research Motivation Scheme (VCRMS) of North Maharashtra University, Jalgaon and SAP (DRS-I), UGC New Delhi, India

REFERENCES

[1] "Text Analytics–Named Entity Extraction", Whitepaper, Lexalytics, (2012) Retrieved from http://www.angoss.com/wp-content/uploads/2012/10/Text-Analytics-Named-Entity-Extraction-White-Paper.pdf

[2] David Nadeau and Satoshi Sekine, (2007) "Survey of Named Entity Recognition and Classification", *Journal of Linguisticae Investigationes*, Vol. 30, No. 1, pp 1-20.

[3] Shaalan, K., (2014) "A Survey of Arabic Named Entity Recognition and Classification", *Computational Linguistics*, Vol. 40, No. 2, pp 469-510, MIT Press.

[4] B. Sasidhar, P. M. Yohan, A. Vinaya Babu, A Goverdhan, (2011) "A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu", *International Journal of Computer Science Issues*, Vol. 8, No. 2, pp 438-443.

[5] Padmaja Sharma, Utpal Sharma and Jugal Kalita,(2011) "Named Entity Recognition: A Survey for the Indian Languages", *Journal of Language in India*, Special Volume: Problems of Parsing in Indian Languages, pp 35-40.

[6] McEnery, A. M., & Xiao, R. Z. (2005). Character encoding in corpus construction. In: Wynne, M. (ed.): Developing Linguistic Corpora: A Guide to Good Practice. Oxford: AHDS, 47-58.

[7] P.Srikanth, K. N. Murthy, (2008) "Named Entity Recognition for Telegu", *IJCNLP-08* Workshop on NER for South and South East Asian languages, Hyderabad, India, pp 41–50.

[8] Kemal Oflazer, (1999) "Morphological Analysis", *Syntactic Word class Tagging Text, Speech and Language Technology,* Vol.9, pp 175-205.

[9] Smruthi Mukund, Rohini Shrihari and Erik Peterson, (2010) "An Information- Extraction System for Urdu- A Resource Poor Language", *ACM Transactions on Asian Language Information Processing,* Vol. 9, No. 4, pp. 1-43

[10] H B Patil, A S Patil and B V Pawar (2014) "Part-of-Speech Tagger for Marathi Language using Limited Training Corpora", *IJCA Proceedings on National Conference on Recent Advances in Information Technology NCRAIT(4),* pp. 33-37.

[11] Artem Boldyrev,(2013) "Dictionary-Based Named Entity Recognition", *Master's Thesis in Computer Science*, University at des Saarlandes.

[12] Shaalan, K., and H. Raza, (2007) "Person Name Entity Recognition for Arabic", ACL 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic, Association for Computational Linguistics, pp. 17–24

[13] William W. Cohen , Sunita Sarawagi, (2004) "Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods", *International Conference on Knowledge Discovery and Data Mining, Seattle*, Washington, USA, pp 89-98.

[14] Hema Raghavan, James Allan, (2004) "Using Soundex Codes for Indexing Names in ASR Documents", W*orkshop on Interdisciplinary Approaches to Speech Index Retrieval*, at HLT-NAACL, Boston, Massachusetts, pp 22-27.

[15] Zobel, J., Dart, P.,(1996) "Phonetic String Matching: Lessons from Information Retrieval", *ACM SIGIR* , ACM, New York, pp 166-172.

[16] Chieu, H. L., Ng Hwee Tou, (2002) "Named Entity Recognition: A Maximum Entropy Approach using Global Information", *19th International Conference on Computational Linguistics* (COLING 2002), San Francisco, pp 190-196.

[17] Zornitsa Kozareva, (2006) "Bootstrapping named entity recognition with automatically generated gazetteer lists", *11th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (EACL '06), Stroudsburg, PA, USA, pp 15-21.

[18] Todorovic, B.T., Rancic, S.R., Markovic, I.M., Mulalic, E.H., Ilic, V.M.,(2008) "Named Entity Recognition and Classification using Context Hidden Markov Model," *Neural Network Applications in Electrical Engineering,* NEUREL 2008, pp 43-46.

[19] Hai Leong Chieu, Hwee Tou Ng, (2003) "Named Entity Recognition with a Maximum Entropy Approach", *7th Conference on Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, PA, USA, Vol. 4 pp 160-163 .

[20] Suxiang Zhang, Suxian Zhang, Xiaojie Wang, (2007) "Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields", *International Conference on Natural Language Processing and Knowledge Engineering* (NLP-KE 2007), Beijing, China, pp 229-233.

[21] Asif Ekbal, Sivaji Bandyopadhyay,(2008) "Bengali Named Entity Recognition using Support Vector Machine" *,Workshop on NER for South and South East Asian Languages* (IJCNLP-08), Hyderabad, India, pp 51–58.

[22] Xavier Carreras, Llu´ıs Marquez, Llu´ıs Padro, (2003) "A Simple Named Entity Extractor using AdaBoost", *7th Conference on Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, PA, USA, Vol. 4, pp 152-155.

[23] Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis, and Constantine D. Spyropoulos, (2000) "Learning Decision Trees for Named-entity Recognition and Classification", *ECAI Workshop on Machine Learning for Information Extraction*.

[24] Da Silva, Joaquim Ferreira, Kozareva, Z., Lopes, G. P., (2004) "Cluster Analysis and Classification of Named Entities", *4th International Conference on Language Resources and Evaluation (LREC 2004),* Lisbon, Portugal, pp 321-324

[25] Agichtein, Eugene, Luis Gravano, (2003) "Querying Text Databases for Efficient Information Retreival", *19th International Conference on Data Engineering*, IEEE computer Society, Columbia Univ., USA, pp 113-124.

[26] Steven Abnvey, (2007) "Semisupervised Learning for Computational Linguistics", Chapman & Hall/CRC.