Chapter 2

LITERATURE SURVEY

2.1 Introduction

To facilitate the conference for message understanding (MUC-6) during the year 1996[GS96] the term Named Entity (NE) was introduced. With the inception of NE, a lot of inquisitiveness surfaced among the educationalists of scientific society [MG98]. The reason was the inclusion of NE as a comprehensible (understanding) task for evaluation during the two message understanding conferences MUC-6, MUC-7 and also in the evolutionary of first and second Multilingual Entity task MET-1 & MET–2 respectively[SC98] , [HG98], [IC95].

Objectivity: The objective of Named Entity Recognition and Classification (NERC) was designed for easy identification and classification of words under a document into some earmarked (Predefined) categories viz. Person-name, Location-name, Organization/Institution-name, Miscellaneous–name briefing the details of Date, Time, Measure and Monitority aspects usage and none of the above. The process of locate and detection of Named Entities (NEs) involve and carry a line of difficulties that is construed as a challenging task. In the process of carrying the NERC the task certainly would prevail as hassles at different stages/levels. Usage of a number of words is quite possible and they can be used both as NE and as a common noun based on the situation that suggests a right place of common utility for the above. Therefore it becomes pertinent to ensure that a decision be made at the primary level to decide whether the entity is an NE or not, whereas the secondary level is concerned with
deciding the type of the ambiguous entity. A distinct tag would help to indicate an appropriate context. Therefore NERC is vital task in almost all NLP application areas.

2.2 Available Named Entity Tag-Set for Indian Languages

The conference on Natural Language Processing IJCNLP brought out a refined (a fine grained NE tag set) NE Tag set which was elaborated containing 12 tags. The 3rd International Joint Conference held in the year 2008 Shared Task on Named Entity Recognition for South and South East Asian Languages–(NER SSEAL), where at the outcome of a well defined NE tag set occurred. This has got more number of tags than the CONLL–2003 shared task on NERC having four tags to its credit [WD02]. The usage of this fine NE tag set is meant for using NERC system in various NLP applications, particularly in Machine Translation (MT) to determine and classify the maximal NEs as well as the nested NEs[KH98]. The constituent parts of the larger NE, is the main objective of the shared task [AS08], [AR08], [CB08]. As per the periodical necessity, rather frequently parts of the entities got to be translated, while the rest can be at the best get transliterated. It has assumed importance since the Nested entities can be useful for MT systems.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEP</td>
<td>Person name</td>
<td>K.K.Reddy/NEP, Prasad/NEP</td>
</tr>
<tr>
<td>NEL</td>
<td>Location name</td>
<td>Nellore Petta sriram Jilla/NEL, Hyderabad/NEL</td>
</tr>
<tr>
<td>NEO</td>
<td>Organization name</td>
<td>Acharya Nagarjuna University/NEOAPSRTC/NEO</td>
</tr>
<tr>
<td>NED</td>
<td>Designation</td>
<td>Director/NED, Professor/NED</td>
</tr>
<tr>
<td>NEA</td>
<td>Abbreviation</td>
<td>BHEL/NEA, NRSA/NEA</td>
</tr>
<tr>
<td>NEB</td>
<td>Brand Names</td>
<td>Philips/NEB, Bang/NEB</td>
</tr>
<tr>
<td>--------</td>
<td>--------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>NETP</td>
<td>Title Persons</td>
<td>Mr. /NETP, Dr/NETP</td>
</tr>
<tr>
<td>NETO</td>
<td>Title Objects</td>
<td>Miss. Universe/NETO, Padmasri/NETO</td>
</tr>
<tr>
<td>NEN</td>
<td>Number</td>
<td>Two/NEN, Twenty/NEN</td>
</tr>
<tr>
<td>NEM</td>
<td>Measurement</td>
<td>1gr/NEH, 10Mtr/NEM</td>
</tr>
<tr>
<td>NETE</td>
<td>Term Expressions</td>
<td>Hindustan computers Ltd/NETE Natural language Processing/NETE</td>
</tr>
<tr>
<td>NETI</td>
<td>Time Expressions</td>
<td>02/10/1968/NETI, 15.08.1947/NETI</td>
</tr>
</tbody>
</table>

**Table 2.1 Named Entity Tag-Set**

The management of IJCNLP–08 NERSSEAL shared task brought out certain guidelines, which are common in application to explain and contextualize the data with the above 12 NE tags formulated vividly and being unique of their analogy. As an integral constituent of the shared task the NE annotated corpuses had been facilitated for the five Indian languages viz. Hindi, Bengali, Urdu, Telugu and Oriya.

In view of the wide coverage and command on the major parts of India these languages were selected as priority languages. Besides, they comprise the characteristics of almost all the language groups (families) which are by and large spoken across the country whereas “Hindi” is accredited as the National language and occupy a unique position in the world as the 3rd widely spoken language. Bengali, being the second one spoken by and large in the eastern part is followed closely by “Telugu” which occupies the third position as a well acquainted language in the country and is spoken widely in southern India.
2.3 Learning Methods

Learning methods are divided into two types in NERC systems, manual and automatic.

**First type:** The overall Test data besides tagging is duly prepared by the user whereas all the cited items are generated by themselves in the automatic system [PS08]. The latter further has two basic accesses the first one being Rule-based and the second one statistical (Machine Learning).

![Diagram for Learning Methods of NER](image)

**Fig 2.1 Diagram for Learning Methods of NER**

Making the unknown entities known has been enabled by the NERC systems. Therefore such ability invariably relies on the rules of recognition and classification systematized from specific characters (or features) connected and identified with
positive and negative precepts. The manual or manmade provision of the early research has been given departure and the advanced versions employ well organized machine learning to automatically pursue the Rule-based systems or sequence labeling algorithm learning with a set of illustrations connected with training. This has reference to the research fraternity carried through an established manifestation where five out of eight systems are among those Rule-based in the MUC-7 conference, sixteen of the systems were launched at CONLL–2003, an institution committed to acquire and refine techno skills [AH98]. In the absence of training equations are not available as resources, manually incepted provisions hold good for a selective technique. This has been displayed in S.Sekine and Nobata(2004) [SN04] who developed a NERC system for 200 entity types.

2.4 Early work and NER in English

2.4.1 Proper Noun Identification System in English

Research in computation works has been aimed at instant identification of named entities in textual forms, a large and heterogeneous saddle of traits, methods and representations. Lisa F Rau in the year 1991 presented his first research paper in this direction. The venue was the 7th IEEE conference on artificial intelligence. Rau’s paper got accredited for his, “Extract and recognize company name”. It was dependant on heuristics (Self learning) and handcrafted rules. Various works of research are listed below in the direction viz.

MUC – 6 (R. Grishman & Sundheim 1996)
HUB – 4 (N. Chinchor et al. 1998)
MUC – 7 and MET – 2 (N. Chinchor 1999)
IREX (S. Sekine & Sahara 2000)
CONLL (E – Tjong Kim Sang 2002)
E. Tjong Kim Sang & De Meulder 2003
ACE (G. Doddington et al. 2004)
HAREM (D. Santos et al. 2006)

A substantial quantum of work in NERC research is focused to the studies in English, however language independence and multilingualism constraints were given comparatively more attention (By the NERC research). German is well studied in CONLL–2003. In CONLL–2002 much emphasis has laid on Spanish and Dutch. Japanese has been studied in the MUC – 6. Petasis et al. 2001, Poibeau 2003, Greek (S. Boutsis et al. 2000) and Italian (W.Black et al. 1998, A. Cucchiarelli & Velardi 2001) and many other languages received attention as well.

Basque (C. Whitelaw & Patrick 2003), Bulgarian (J.Da Silva et al. 2004), Hindi (S. Cucerzan & Yarowsky 1999, J. May et al. 2003), Korean (C. Whitelaw & Patrick 2003), Polish (J. Piskorski 2004), Romanian (S. Cucerzan & Yarowsky 1999), Russian (B. Popov et al. 2004), Swedish (D. Kokkinaki 1998) and Turkish (S. Cucerzan & Yarowsky 1999), Portuguese was examined by D. Palmer & Day 1997. Arabic (F. Huang 2005) began to receive a lot of attention in large-scale projects such as Global Autonomous Language Exploitation (GALE).

2.4.2 Autonomous Language Exploitation (GALE)

A proper noun is generally expressed or denoted as a noun that is capitalized. Since the proper noun is capitalized, it will not pose any ambiguity because the beginning of a sentence is affected with a noun or a proper noun. Learning of words should be practiced as to which word is a proper noun.
2.4.3 Rule Based Approach

This is a Rule-based system that needs emphasis on a lot of grammatical and linguistic parameters to draft rules. It has been noticed that Rule-based approach has paid the dividends with required gazetteers lists, features blended with language and conditions as rules for languages [HI01]. Named Entities are unrestricted open class words. It is a trend set that with every update on a new word being appended to languages and gazetteers updating the list are a ceaseless task as observed from the factual findings [BR91]. This has paved the way to enable the gazetteers to accommodate as many words as they could by dividing them into finite tests like prefix, suffix and contextual words [WV95].

All rule based approaches are language dependent. It cannot be thrust upon a implementation of language independent [CY02] NER system for our national languages since language independent rules differ from one language to another.

A Rule-based NER system was developed (1995) by Ralph Grishman. The system verifies and pursues certain unique name dictionaries comprising nomenclature of countries, larger cities, companies and common first names [GR95]. A text was brought out under Rule-based approaches, where a set of rules are so defined as patterns in helping out to identify the named entities in a text. The other NER system, a Rule-based one was developed during 1996, which makes use of several gazetteers like individual names of people, names of organizations, locations, titles etc...
The set back to these rule-based techniques is that the systems need to have expertise in grammatical knowledge of a particular or specific language or domain. These systems are not transferable to other languages.

### 2.4.4 Machine Learning approach

Machine Learning is a branch of artificial intelligence, is concerned with the construction and study of systems that can learn from data. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders.

The core of machine learning deals with representation and generalization. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory.

There is a wide variety of machine learning tasks and successful applications. Optical character recognition, in which printed characters are recognized automatically based on previous examples, is a classic example of machine learning.

#### 2.4.4.1 Supervised Learning

The advantage of supervised learning is to understand the features of positive and negative forms of NE collected from the annotated sources of texts and rules pertaining to design that hold the presentation of a given type [DN07]. Supervised Learning has a set back on the requirement of a big amount of annotated corpus
As a matter of fact the absence of such resources and the cost factor involved in realizing them gave into two successive learning methods, the semi supervised and unsupervised learning (UL).

2.4.4.1.1 Hidden Markov Model

The supervised learning approach requires a labeled data to bring out a statistical model. However supervised approaches are costlier when compared to their counter parts (un-supervised approaches) in terms of the time spent to pre process the training data. Good performance can be achieved by supervised approaches provided; large amount of high quality training data is available [RA89]. Supervised approaches are rather expensive, as it incurs time whereas it is consummated to preprocess the training data [FZ03], [VI67]. HMM, Decision Tree models and Conditionals Random Fields have been used.

The HMM based system has been formulated by Bikel et al. (1998) [BN98] which enables to recognize, classify the nomenclature, dates, times, and quantity in alpha numerical method. English and Spanish languages were used to evaluate the model, and the news on air (broadcast) as the speech input. They have given a demo depicting the impact of training set size on performance. This has given a data indication where 100,000 words for training data is sufficient to get an outcome of performance around 90% on news wire.

In furtherance, Bikel et al. (1999) [BD99] developed an NE locating system (recognition system) with a slightly amended, a modified version of a HMM, known as “Nymble”. They projected it with a simple proto type of a model, locating names, and numerical entities as per the specified MUC tasks that could enable “near human performance”, frequented to an F-measure of 90 or more such a system can be result
oriented when provided with correct and consistent answer keys duly marked. It can be trained on languages, which are new and foreign to the system trainer.

NERC system is built a from out of a HMM based chunk tagger proposed by Zhou & Su (2002) [ZS02] the system is built to recognize and classify names, times and numerical quantities. The system enables to apply and integrate four types of internal and external evidences, viz. Simple deterministic internal feature of the words, such as capitalization and digitalization internal semantic features of important triggers, internal gazetteers feature and external macro context feature. Evaluation of the system was done on MUC-6 and MUC–7 English NE tasks and achieved F–measures at 96.6% and 94.1% respectively.

2.4.4.1.2 Maximum Entropy Markov Model

Based on the probabilities, Borthwick developed a ML based system during 1999 which is the Maximum Entropy Markov Model based system [BS98]. About Eight dictionaries were referred for authenticity and record ML based techniques for NER and were employed to utilize a large amount of NE annotated training data to acquire high level language techniques by using gazetteer lists[AB99], [BO03]. Using both the “Learning concepts” by employing the machine learning techniques that enabled a lot of research work is carried on NER for English. In English language identifying NE is easier since the capitalization of names has become effected. Moreover, the unsupervised learning approaches does not need labeled training data i.e., as it necessitates few seed lists and large un-annotated corpora [BD96]. The set goal or objective of an unsupervised learning is to receive or incorporate the representations from data.
2.4.1.3 Support Vector Machine

The Support Vector Models are based on differentiate approach and use both positive and negative examples to make a difference between the two classes (Vapnik, 1995) [VV95]. The SVMs handle big featured sets and develop models that magnify their appearance as Generalizability [KM01], [TH99]. Generalizability in SVMs forms the basis of statistical learning theory and the observation that it is useful is sometimes beneficial to classify some of the training data so that the margin between other training points is maximized (Cortes and Vapnik 1995) [CV95]. This is particularly useful for real world data sets containing inseparable data points. Secondly, although the training is generally regarded as slow, the resulting model is usually small and runs quickly because only the support vectors need to be retained. The pattern helps define the function which separates positive from negative examples. Thirdly, SVMs being binary classifiers, it is necessary to combine SVM models to obtain a multiclass classifier.

Yamada et al. (2001) [TY01] proposed the SVM based NERC system for the Japanese. This is an extension of Kudo’s Chunking system which exhibited one of the best performances at CONLL–2000 shared tasks. Viz. Takenchi and Collier, 2002 [TC02] and Masayuki and Matsumoto 2003 [MM03], are the two other works, with which the other SVM based NERC system can be found. The use of SVMs for an extended NE task has been explored by Takeuchi and Collier in the year 2002. The application of NE to non-news domains facilitates extension of the NEs so that is can take over categories, are presumed as examples of conceptual divisions as well as individuals. They verified (for authenticity) the identification and classification of technical terms in molecular biology domain and made a contrast with the outcome.
secured for NE recognition on MUC–6 data sets [AK95]. They have presented that the SVM utilizing a resourceful feature set of a context window of previous three and following two words and POS features (MUC–6 only) had a significant performance advantage on both the MUC–6 and molecular biology data.

### 2.4.4.1.4 Conditional Random Fields

During the year 2004 [SC04] Sarawagi & Coehn studies the semi–Markov CRFs (Semi–CRFs), a conditionally updated version of semi–Markov chains. NERC experiments were shown that semi–CRFs work par excellence[HW04].

In order to improve NERC system performance, Cohen and Sarawagi have proposed a method in the direction in the year 2004. It is presumed difficult as most high performance NERC systems operate by sequentially classifying words whether or not they participate in any entity name. But the pressing necessity suits the most useful similarity measures score entire candidate names. In order to correct the unsuitability they have tried to formalize a semi-Markov extraction process, which finds a basis in sequentially classifying segments of several adjacent words instead of signal words [LM01]. Additionally, a natural way of coupling high performance NERC methods is necessitated through certain formalism facilities with the direct use of other entity–level features and enables a formulation of NERC problem which is natural than sequential word classification[DS97]. They experimented in multiple domains and demonstrated that their proposed model can substantially improve extraction performance over a strong baseline NERC, using Collins perception based algorithm (Collins, 2002) [CO02] for training a state of the art, multi level encoding meant for dictionary information.
They are the variation of the supervised approach that reads a large annotated corpus, memorizes lists of entities, creating disambiguation Rules-based on discriminative features.

2.4.4.1.5 Unsupervised learning

The unmonitored learning has a clustering type approach observed in unsupervised learning [EO05]. Taking example from which one can gather named entities from clustered groups as per the similarity of context primarily [NT06], the techniques depend on lexical resources, patterns and cumulative worked on a large un-annotated corpus.

These representatives are to be supplied for data compression, classifying, decision making and to invoke many other purposes. As most of the unsupervised learning is devoid of suitability, the approach for NER and the others employs unsupervised learning [MG99].

This has been illustrated by Collins et al. [CN99] on NE classification with the aid of unlabelled examples of data. In support of the above, Kim et al. [KK02] proposed an unsupervised NE classification model, used a small scale NE dictionary and an unlabelled corpus for classifying named entities [YL02]. Whereas the process involved in implementing supervised learning is associated with a program used to classify furnished examples that would be set apart as part of the exercise in the direction.

2.4.5 Hybrid Approach

The custom of using more than two different approaches for achieving a particular goal is what is referred to as Hybrid approach. The application of Hybrid Approach in NLP is the
combination of both Rule-based and Statistical approach. Rule-based approach does not require training data however requires a large amount of language independent rules. Intensive knowledge of the language is required in a Rule-based approach and resolution of any form of uncertainty is more complex. In order to achieve this lot of rules must be generated. In statistical approach, the accuracy of the system depends on training data. The more advanced the training data the higher the accuracy. The system is therefore not generalized. Rule-based approach is applied first followed by the application of Statistical techniques [PR08]. This approach is more suitable for resource poor languages and rating, ranking of data is thereby established [SC00].

- HMM approach and Rule Based approach
- CRF approach and Rule Based approach
- MEMM approach and Rule Based approach
- SVM approach and Rule Based approach

2.5 Related Work on NER in Indian Languages

2.5.1 Hidden Markov Model

During the year 2007 [EB07], the development of HMM based NER system was brought out by Ekbal and Bandyopadhyay. Manually it was tested over a corpus compromising 34 Million word forms developed from the online news papers of Bangali language. Partly the tagged news corpus compromising 1,50,000 word forms is used to undertake training of the NER system, through HMM–based parts of speech tagged with 26 different POS tags the training set so secured, is a corpus tagged with 16 NE tags. The test (experimental) outcome of the 10–fold cross valuation gets an average Recall, Precision, and F-score values of 90.26%, 79.48% and 84.5%. Subsequently, the HMM–based NER system is also subject to training and tested with
Hindi data to show the post effect of the language unique features (36), whereas the Hindi NER results show an average Recall, Precision, and F-score values of 82.5%, 74.6% and 78.35% respectively.

In 2008 [PP08] Pandian et al. demonstrated a hybrid 3 stage approach for Tamil NER. The HMM algorithm was involved to locate the best sequential order for the rest of two phases and then modified to suit the free word order problem. The POS and NER tags were together used undisclosed variables in the algorithm. As a result, the system came out with an F-measure of about 72.72% for various types of the entities.

In the year 2011 B.Sasidhar et al. developed a Hidden Markov Model approach for the development of NER on Telugu language [SV11]. In their work they used HMM model for handling unknown words. They collected the corps from various news articles and other web resources. Finally they achieved the Precision of 77% for Person, 77% for Location and 81% for Organization, 80% for others and Recall of 70% for Person, 77% for Location,75% for Organization, 78% for others.

2.5.2 Maximum Entropy Markov Model

In 2009 [HE09], Hasannuzaman et al. traced the development of NER System in Bengali and Hindi using ME frame work with 12 NE tags to a coarse grained NE tag set of 4 tags namely, person name, location name, organization name and miscellaneous name[CN99], [MF00]. The system uses different contextual information of the words along with the variety of orthographic word–level features that help in predicting the four NE classes. Ten-fold cross validation test results in the
average recall, precision and f-measure of 88.01%, 82.63%, 85.22% for Bengali and 86.4%, 79.23% and 82.66% for Hindi.

Based on MEM approach, Raju et al. [RS10] formulated a NER system for Telugu. The corpus was collected from the Eenadu, Vaarta newspapers and Telugu Wikipedia. Manually tagged test data was prepared to evaluate the system. The system made use of different contextual information of words, gazetteer list was also prepared manually or semi automatically from the corpus and came out with an F-measure of 72.07% for person, 6.76%, 68.40%, and 45.28% for organization, location and others.

Suitable features for Hindi NER task was identified by Goyal, and they developed a ME based Hindi NER system. A two phased transliteration methodology was utilized to make the English lists useful in the Hindi NER task. The system demonstrated extensive performance after making use of the transliteration–based gazetteer lists. This approach is also applied to Bengali besides Hindi NER task and it was presumed to be proactive.

The highest F-measure achieved by ME-based system is 75.89% which was enhanced to 81.2% by using transliteration based gazetteer list, Li and Mc Callum (2003) [LM03].

In the year 2006 Kumar and Bhattacharya made an achievement by yielding an F-measure of 79.7% using a maximum Entropy Markov model [KB06]. Among other Indian languages by and large, Punjabi trails behind in this area. There was much larger focus given for the research on NER for Punjabi.
2.5.3 Conditional Random Fields

A two phase/stage approach in a CRF based NERC system was provided by Krishnan and Manning (2006) [KM06], [PN03] using local futures to prediction and train one more CRF that uses both the local information and predictions of the future and then train another CRF which uses both local information and features extracted from the output of the first CRF. The other NERC systems could be located in Buneser and Mooney (2004) [BN04], Finkel et al. (2004) [FG04], Chen et al. (2006) [CZ06], Feng et al. (2006) [FS06] and Jingh Zhon and Houfeng (2008) [JH08].

In the year 2008 [GO08], Goyal planned to develop building an NER for using CRF. This method was evaluated on test set 1 and test set 2 and achieved a maximum F1-measure about 49.2% and nested F1-measure about 50.1% for test set 1, Maximum F-measure about 44.97% and nested F1 measure 43.70% for test set 2 and F-measure of 58.85% on development set.

Part of the LERC-UOH Telugu corpus was used by Srikanth P and Murthy K.N. in the year 2008 [SM08]. Here, CRF based Noun tagger is built using 13,425 words manually tagged data and experimented on a test information or data set of 6,223 words and came out with an F-measure of 91.95%. Subsequently they planned a Rule-based NER system consisting of 72,152 words that include 6,268 named entities with which they located some issues pertaining to Telugu NER and later developed a CRF dependent NER system for Telugu and secured an overall F-measure between 80% and 97% in various experiments.

Ekbal et al. [EB09] emphasized on the improvement of Bengali NER system using the statistical CRF. NE classes would be identified by the system utilizing
different contextual sources as information of the words along with various kinds of features. The training set will have 150k word form annotated manually with 17 tags.

The outcome experimented results of 10-fold cross validation tests exhibit the pro-activeness of the planned CRF based NER system with a recall of an out and out average, precision and f-score values of 93.8% 87.8% and 90.7% respectively.

A learning based NE-recognizer for Bengali was presented by Hasen et al. (2009) [HR09]. However it did not rely on manually constructed gazetteers. They planned two designs for the NER system. The corpus compromising 77,942 words is tagged with one of the 26 tags in the tag set defined by IIT, Hyderabad where they used CRF++ to train the POS tagging model. Evaluation results show that the finder achieved an improvement of 7.5% in F-measure over a base line recognizer.

A CRF based NER system for Telugu was developed by B.Sasidhar et al. in the year 2010 [SY10]. They experimented on three tags (Person, Location, Organization) using Features like context features, context patterns, parts of speech features, suffix features, morphological analyzer and CRF using gazetteers. They achieved Precession, Recall and F-measures of 81%, 85% and 82% respectively.

2.5.4 Support Vector Machine

NER system for Hindi and Bengali with SVM was developed by Asif Ekbal and Bandyopadhay in the year 2008 [AB08]. An annotated corpus of 122, 467 tokens of Bengali 502, 974 tokens of Hindi were used and tagged with 12 NE classes. The system NER has been tested against the gold standard test sets of 35K, and 60k token for Bengali and Hindi. Evaluation results have shown to demonstrate the recall,
precision, and f-score of 88.61%, 80.12% and 84.15% for Bengali, 80.23%, 74.34% and 77.17% for Hindi.

Different contextual messages information of the words along with variety of features are being used by the system and are helpful in predicting the Named Entities. An incompletely (partially) NE tagged Bengali new corpus has been involved to make a training set created, the training set for the experiment and the training set comprises of 150k word forms that is manually tagged with 17 tags. Tested or experimental results of the 10 fold cross validation test indicated the soundness of the proposed SVM based NER system with the corresponding average recall, precision and F-score relating of 94.3%, 89.4% and 91.8% respectively.

2.5. 5. Rule-Based Approach

In the year 2008 [CB08] Chowdari and Bhattacharya tested an automatic detection of Named Entities in Bengali. A three phased approach was involved viz. the dictionary based for NE rules for NE and left–right co-occurrences statistics. The corpus of Anad bazer Patrika has been used for the period, 2001-2004. Manual tagging was taken up by the linguistics based on global knowledge. Tested as experiments with outcome as results pointed out the average recall, precision and F-measure as 85.50%, 94.24% and 89.51% respectively.

In the year 2011 B.Sasadhar et al. used a Named Entity Recognition in Telugu using Language Dependent Features and Rule based Approach [SG11]. In this approach they used Person Gazetteer with 30,000 names including first name, last name and contexts of people. Location Gazetteer with 27,000 names of villages, mandals, cities, and districts. Organizations gazetteer with 2000 names of companies,
political party names. They collected nearly 40,000 words manually from CP Brown, to eliminate closed class words. In this approach they achieved more than 90% of accuracy in identification of named entities.

2.6 Hybrid Approach

“A Hybrid approach for named entity recognition in Indian languages” [SS08] was proposed by Sujan Kumar Saha et al. (2008) which involved using maximum Entropy Markov Model. Language dependent rules and gazetteers have been taken into account with twelve divisions of NER on Hindi, Bengali, Oriya, Telugu and Urdu. As an effort to enhancement in the direction, a text from Shakthi standard format was given conversion into IOB format and experimented with more than 5,00,000 of Hindi, 1,60,000 of Bengali, 93,000 of Oriya, 64,000 of Telugu and 36,000 Urdu words have been involved for usage. The validation reported F-score of 65.13% for Hindi, 65.96% for Bengali, 44.65% for Oriya 18.75% for Telugu and 35.47% for Urdu respectively.

2.7 Analysis of All Methods

Bengali is the seventh popular language in the world, second in India and the national language of Bangladesh. In the year 2009 [EB09], the development of NER in the language was reported by Ekbal and Bandyopadhyay. He tried to get through the task successfully by combining the output of the classifier like ME, CRF and SVM (15) respectively. The training set comprises of 150 k word formation for tracing the 4 NE Tags Viz. person, location, organization and miscellaneous objects. In order to enhance the performance of the classifier, about three million word forms were used, extracted from lexical context pattern generated from an unlabelled
Bengali corpus. Evaluation results of 30k word forms have shown that altogether, precision and f-score values as 87.11%, 83.61% and 85.32%. This indicates an improvement of 4.66% in f-score over the best performing SVM based system and 95% in f-score over the ME based system.

A report on the development of Bengali news corpus from the web comprising of 34 million word forms was propounded by Ekbal and Bandyopadhyay in 2008 [BE08]. Part of it, about 150k word forms, is manually tagged with 16 NE and with one non-NE tag, besides, 30k word forms are tagged up with a tag set of 12 NE tags explained and defined for the IJCNLP–08 NER shared task for SSEAL. A change in tag (conversion) routine has been fared to convert the 16 NE tagged corpus of 150k word forms to the corpus tagged with IJCNLP– 08, 12 NE tags where the former has been used to develop the Bengali NER system suing HMM ME, CRF, SVM respectively. The output (evaluation) results of 10 fold cross validation experiments give the F-score of 84.5% for HMM, 87.4% for ME and 90.7% for CRF and 91.8% for SVM.

Ekbal and Bandhopadhyay in 2008 [EB08] reported on the development of NER system in Bengali combining the outputs of the classifier like ME, CRF, and SVM. The corpus consisting of 250k word forms is manually tagged with four NEs namely person, location, organization, and miscellaneous. The system makes use of different contextual information of words along with a variety of features that help in identifying the NES experimental results and indicates the effectiveness of the proposed approach with overall average recall, precision and f-score values of 90.78%, 87.35% and 89.03% respectively. This shows an improvement of 11.8% in
f-score over the best performing SVM based baseline system and an improvement of 15.116 in f-score over the least performing ME based system.

In the year 2008 [KS08] Vijayakrishna and Sobha L brought out “Domain Focused–Named Entity Recognition for Tamil using conditional Random fields”, developed a model titled “Domain focused NE Recognizer for tourism Domain conditional Random Fields Approach on Tamil language”. They used 106 tag sets for tourism domain and five feature templates. About Ninety four thousand words corpus was collected in Tamil for this domain. NE annotations NP Chunking, POS tagging, Morph analysis are presented as to their performance manually on the corpus. It comprised of roughly 20,000 titled entities divided into two sets. Whereas the fore most formed the training data while the other the test data, constituting 80% and 20% of the total data respectively. A total of 4059 entities where taken on testing for experiment and got overall F–measure 80.44%

Development of Hindi NER using ME approach was elucidated by Saha et al. (2008) [SS08], [SG08]. About 234 k words were stated to have comprised as training data, collected from the news papers “Dainik Jagaran” which were manually tagged with 17 classes, with 16,482 NEs.

The development of a module was also reported in the paper about the semi-automatic learning of context pattern, using a blind test corpus of 25k the system was evaluated as having 4 classes and achieved an F-measure of 81.52%.

A detailed observation was made out by Gupta and Arova in 2009 [GA09] and the experiment conducted on CRF models for developing Hindi NER. It indicates some features making the development of NER system more complex. It narrates the
different approaches for NER. The information used for the training of the model was
taken from tourism domain which is manually tagged in 10B format.

Using the SVM system, in the year 2008 [EB08], [AR08] Ekbal and
Bandyopadhyay developed NER system for Bengali.

Further to the usage of appropriate unlabelled data in 2009, [EB08] Ekbal and
Bandyopadyay briefed about a voted NER system. This above procedure locates the
basis in supervised classifier, namely ME, SVM, CRF where SVM makes use of two
different systems known as forward parsing and backward parsing. It was tested for
Bengali comprising 35,143 news document and 10 million word forms and make use
of language independent features along with different contextual information of the
words. A the end, the models were combined into an ultimate system with an
arranged voting technique and the test results extended the effectiveness of the
proposed approach with the recall precision and f–score values of 93.81%,92.18% and
92.98% respectively.

A language independent NER in Indian languages [EH08] was developed by

The system utilized variety of contextual information of the words along with
different features that was supportive in forecasting (predicting) the various NE
classes in both the language dependent and language independent areas.

The latter was applied to Hindi, Bangali Oriya Telugu and Urdu and language
dependent features were applied to only Bengali and Hindi. The system was
experimented with Bengali. (1,22,467 tokens), Hindi (5,02,974 tokens) Telugu
(64,026 tokens), Oriya (93,173 tokens) and Urdu (35,447 tokens) and tested with
Bengali (30,505 tokens), Hindi (38708 tokens), Telugu (6, 356 tokens), Oriya (24,640 tokens) and Urdu (3,782 tokens), and found the maximal F-measure of 53.46% for Bengali whereas for Telugu F-measure was found as a very performer.

2.8 Summary

Several tests as experiments were conducted in order to develop the versatile systems of NERC including supervised and unsupervised learning procedures and type of methods on English and other Indian languages. It has been identified and found that only nominal work carried out in the direction particularly on those like Bengali, Hindi, Telugu and Urdu. The obvious reason for this scanty work was not having capitalization concept as in English. Lack of adequate or sufficient resources and annotated data for Indian languages constitute another cause in this regard. The machine learning approaches require larger training feedback (data) or availability of annotated corpus, to extend the methods of training in needy areas. Machine learning approach comprises of two phases, learning phase being the first followed by testing phase[FL03]. It is a known fact that creating manual annotated data is a difficult task. The draw backs or a short coming of manual creation on annotated information (or data) is barred of consistency, common errors effort through manual means, and time strife process in order to sustain or bring more perfection in terms of getting accurate data, to be complied with.

NE identification and classification is carried out by involving language independent or dependent rules which implies it happens without using any training data. Rule-based approach needs exact rules for identification and classification. These rules hold good for application of any data in the language. We may expect
more accuracy from the output of the system provided the rules have consistency. Rule-based systems training data is not necessary as it keeps basis and relies only on language dependent rules. The survey undertaken observed and reveals that different types of experiments have been conducted in various languages.