International Journal of Computer Applications (0975 - 8887)
Volume 142 - No.8, May 2016

Named Entity Recognition in Assamese

Padmaja Sharma
Department of CSE
Tezpur University
Assam, India 784028

Utpal Sharma
Department of CSE
Tezpur University
Assam, India 784028

Jugal Kalita
LINC Lab,Department of Computer Science
University of Colorado at Colorado Springs
Colorado, USA 80918

ABSTRACT

Named Entity Recognition is a process through which a program extracts proper nouns in texts and associates them with a proper tag. NER has made significant progress in European languages, but in Indian languages due to the lack of effort as well as proper resources, it remains a challenging task. Recognizing ambiguities and assigning the correct tags to the names is the main goal of NER. Thus NER can be defined as identification of the proper nouns and the classification of these nouns into classes such as person, location, organization and miscellaneous including date, time and year. The main aim of this work is to develop a computational system that can perform NER in text in Assamese, which is a resource poor Indo-Aryan language. This article present an overview of NER and its issues in the context of Assamese, and also the work done in Assamese using various approaches. ■

1. INTRODUCTION

Natural Language Processing (NLP) is the branch of computer science focused on developing systems that enable computers to communicate with people using everyday language. The Internet-focused world one lives in generates a large amount of textual data every-day and access to such data has changed the way one live and work. This abundant data would be of use only if suitable techniques were available to process the data and extract knowledge from it. One of the most important efforts in this regard was the Message Understanding Conference [5] whose main goal was to identify entities which can be considered names from a set of documents and classify them into predefined categories. This is called Named Entity Recognition (NER). A Named Entity (NE) is an element in text that refers to the name of a thing such as that of a person, organization or location. Recognition and tagging of Named Entities in text is an essential component of tasks such as Information Extraction (IE), Question Answering (QE), and Automatic Summarization (AS). In the Message Understanding Conferences (MUC) [11], it became clear that it is necessary to first identify certain classes of named entities in order to extract information from a given document. Later the conference established the Named Entity Recognition task [3] . Systems were asked to identify names, dates, times and numerical information. NER can be defined as a two stage problem - identification of proper nouns and the further classification of these proper nouns into a set of classes such as person names, location names (e.g., cities, and countries), organization names (e.g., companies, government organizations, and committees), and miscellaneous names (e.g., date, time, number, percentage, monetary expressions, number expressions and measurement expressions). NER can be treated as a tagging problem where each word in a sentence is assigned a label indicating whether it is a part of a named entity and the entity type. A few conventions for tagging Named Entities were established in the MUC conferences [3]. These include ENAMEX for names (organization, person, location), NUMEX for numerical entities (monetary, percentages) and TIMEX tags for temporal entities (time, date, year). For example consider the sentence:

Mr. Alex attended the conference in U.K which was held in July 2015.
Using an XML format, it can be marked up as follows:

> <ENAMEX TYPE="PERSON">Mr. Alex</ENAMEX> attended the conference in <ENAMEX TYPE="LOCATION">U.K</ENAMEX> which was held in <TIMEX TYPE="DATE">July 2015 </TIMEX>.

Here, the markups show the named entities in the document.

2. PROBLEMS IN NAMED ENTITY RECOGNITION

The task of NER in general, faces the challenge of ambiguity. Consider the sentence

> *Rose rose to put rose roes on her rows of roses.*[1].

In this sentence *rose* can either be a person name or a common noun, making it difficult for the computer to resolve the ambiguity between the two. A program is forced to use domain or linguistic knowledge, possibly in the form of rules, such as that the name of a person (which is a proper noun) usually begins with a capital letter, in order to resolve the issues. Thus, the correct NE annotation of the above sentence can be:

> <ENAMEX TYPE="PERSON">Rose</ENAMEX> rose to put rose roes on her rows of roses.

Domain or linguistic knowledge can come in other ways also, e.g., statistics. Another form of ambiguity is that frequently there are overlaps among classes of NEs. Ambiguity is one of the main challenges in NER. The different types of ambiguity that occurs in NER are as follows:

[1] http://www.wikipedia.org/wiki/List_of_linguistic_example_sentences

—*Person vs. location*:- In English, a word such as *Washington* or *Cleveland* can be the name of a person or a location. Similarly, in Indian English (or in an Indian language when written in the appropriate script), words such as *Kashi* can be a person name as well as a location name.

—Words or word sequences also exist such as *Thinking Machines* (a company), *Gates* (a person), that can occur in contexts where they do not refer to NEs.

—*Common noun vs. proper noun*:- Common nouns sometimes occur as a person name such as *Surya* which means sun in Sanskrit, thus creating ambiguities between common nouns and proper nouns. Another example is *Rose* (a person) in the above example.

—*Organization vs. person name*:- *Amulya* may be the name of a person as well as that of an organization, creating ambiguity. An English example may be *Trump*, which can be the name of a person as well as the name of a company or a brand.

—*Nested entities*:- Nested entities such as *New York University*, also create ambiguity because they contain two or more proper nouns.

Such phenomena are abundant in Indian languages as well. These ambiguities in names can be categorized as structural ambiguity and semantic ambiguity. [12] describe such ambiguities in detail. A number of additional challenges need to be addressed in South Asian languages such as Hindi, Bengali, Assamese, Telugu, Urdu and Tamil. The key challenges are briefly described as follows. Although the examples are in specific languages, similar phenomena occur in all Indian languages and Assamese in particular.

—*Lack of capitalization*:- Capitalization plays a major role in identifying NEs in English and some other European languages. However, Indian languages do not have the concept of capitalization.

—*Ambiguity*:- In Indian languages, the problem of ambiguity between common nouns and proper nouns is more difficult since names of people are usually dictionary words, unlike Western names. For example, [akas] and [zun] mean *sky* and *moon*, respectively, in Assamese, but also can indicate person names. In fact most people's names are dictionary words, used without capitalization.

—*Nested entities*:- Indian languages also face the problem of nested entities. Consider, in Assamese [tEzpuR bisHobidyaloi] [E:[2]Tezpur University]. It creates a problem for NER in the sense that the word [tEzpuR] [E:Tezpur] refers to a location, whereas [bisHobidyaloi] [E: University] is a common noun and thus [tEzpuR bisHobidyaloi] [E:Tezpur University] is an organization name. Thus it becomes difficult to retain the proper class.

—*Agglutinative nature*:- Agglutination adds additional features in the root word to produce complex meaning. For example, in Assamese, [guwahati] [E:Guwahati] refers to a location named entity whereas [guwahatiya] [E:Guwahatiya] is not a named entity as it refers to the people who live in Guwahati.

—*Ambiguity in suffixes*:- Indian languages can have a number of postpositions attached to a root word to form a single word. In Assamese the word [monipuR] [E:Manipur] is a place name, but when the suffix [ee] is attached, it gives a different meaning compared to the original one which means the people of Manipur.

—*Resource constraints*:- NER approaches are either rule based or machine learning(ML)-based. In either case, a good-sized corpus of the language under consideration is required. Such corpora of significant size are still lacking for most Indian languages. Basic resources such as parts of speech (POS) taggers, or good morphological analyzers, and name lists, for most Indian languages do not exist or are in research stages, whereas a number of resources are available in English.

[2]E: English meaning

3. APPROACHES TO NAMED ENTITY RECOGNITION

Techniques for NER can be classified into three:

(1) Rule-based approaches,
(2) Machine Learning approaches, and
(3) Hybrid approaches.

Rule-based NER focuses on the extraction of names using human-made rules. A rule-based system requires a human expert to define rules in which the person needs to be a domain expert and have good programming skills. This method is easier to develop and interpret than statistical methods. In general, the rule-based approach consists of a set of patterns using grammatical, syntactic and orthographic features. This approach lacks portability and robustness. One needs a significant number of rules to maintain optimal performance, resulting in high maintenance cost. There are several rule-based NER systems for English providing 88%-92% F-measure [[13], [11]]. [2] proposed a name identification system called FASTUS. LASIE by [7] and LASIE II by [6] used the concept of a look-up dictionary and grammatical rules to identify the NEs. The main attraction of the ML approach is that it is trainable and can be adapted to different domains. The maintenance cost is also less than that of the rule-based approach. The ML approach identifies proper names by employing statistical models of classification. ML models can be broadly classified into three types: Supervised, Unsupervised and Semi-supervised.

In supervised learning, the training data include both the input and the correct output. In this approach, the construction of proper training, validation and test sets is crucial. This method is usually fast and accurate. As the program is taught with the right examples, it is "supervised". A large amount of training data is required for good performance of this model. Several supervised models used in NER are: Hidden markov Model (HMM) [[9],[10],[14]]; Conditional Random Field(CRF) [8]; Support Vector Machine(SVM) [4]; and Maximum Entropy(ME) [1].

4. WORK ON NER IN ASSAMESE

This section discusses NER in Assamese using a rule-based approach, a gazetteer-based approach and ML approaches. Rule-based methods are seen to work well provided the rules are carefully prepared and cover all the possible cases. Similarly, for well known proper nouns that occur frequently in texts, looking up a gazetteer list containing such nouns works well for NER. NER in Assamese is tested with rules as well as a gazetteer-list approach. The rule-based approach involves the identification of the root word from an inflected or a derivational form, which is known as stemming. Handcrafted rules are also used to identify the different classes of named entities.

4.1 Assamese Corpora

A corpus is a collection of text in a single or in multiple languages. Annotation is an important task that can be performed on a corpus for linguistic research. It is the process of adding a label or tag to each word or some other component. A corpus can be obtained from different sources such as newspapers, articles and books. Several large corpora are available in English and Indian languages. But most Indian languages are low-resource. In Assamese, the number of corpora available is quite small compared to other languages. Throughout the work, Assamese text encoded in Unicode which ranges from U0980-U09F is used. The following corpora of Assamese are used for the work.

(1) *EMILLE/CIIL Corpus*: EMILLE (Enabling Minority Language Engineering) developed jointly by Emille Project, Lancaster University, UK, and the CIIL (Central Institute of Indian Languages), India; consisting of 2.6 million wordforms [3].

(2) *Asomiya Pratidin Corpus*: This corpus was obtained by downloading articles from the website of newspaper during 2000-01 by Utpal Sharma at Tezpur University. It consists of nearly 372,400 wordforms. The articles includes general news, sports, news, editorials, etc.

(3) *Tezu Assamese Corpus*: It is a collection of Web and news articles from online newspapers and electronic magazines which consists of 2,950 news articles, literature, science and Medicine. It consists of 1,060,550 word forms.

4.2 Rule-based approach

NER requires morphological analysis of the input words, i.e., analyzing how the input words are created from basic units called morphemes. Generally, identification of the root form of a word, i.e., stemming is required. Stemming is the process of reducing inflected and derived words to their stems or base or root forms. For example, the stem of the word *governing* is *govern*, of *cycling* it is *cycle*. The study of stemming in Indian languages is quite limited compared to European, Middle-Eastern and other Eastern languages. Apart from NER, stemming is widely used in information retrieval to improve performance. When a user enters the word *happiness*, it is most likely that it will retrieve documents with the word *happy*. In highly inflectional languages such as Assamese, identification of the root form of words is crucial. In languages such as English and also in Indian languages like Assamese and Bengali, there are words which are not NEs, but their root words are NEs. For example the words [bHaRotiyo] [E:Indian], or [monipuRi] [E:Manipuri], which are adjective, are not named entity, but the root words [bHaRot] [E:India], or [monipuR] [E:Manipur] are. Different suffixes are attached to the root word to form different words with different meanings. There are root words that represent location NEs, whereas the surface words are not NE. The main aim of the approach is to generate the root word from a given input word resulting in a location NE. To obtain the root words, the suffix stripping approach is used. Suffix stripping algorithms do not rely on a lookup table instead rules are stored to find its root/stem form. It is a fast process as the search is done only on the suffix. Some examples are given below in the Table 1.

In this experiment a part of Asomiya Pratidin corpus of size 20K words is used. The suffix stripping is use for those words whose root words represent location NEs. The statistics of the training data and the effectiveness of the method is given in the Table 2. The approach gave an f-measure of 89% accuracy. The performance of this stemmer degrades for those words which do not have suffixes attached to the root words and the last character of the root word matches with the suffix list. Example of such words are:

Table 2. Test of suffix stripping experiment for location named entity

Feature	Data
Total words	20,000
Total NE present (T)	475
Total NE retrieved (K)	565
Total correct NE retrieved (S)	465
Precision	82.3
Recall	97.8
F-measure	89

— [mazuli] [E:Majuli].
— [dHubuRi] [E:Dhubri].
— [tinsukiya] [E:Tinsukia].
— [goRoimaRi] [E:Goroimari].
— [tEtEliya] [E:Teteliya].

For example, word [mazuli] [E:Majuli] = [mazul] [E:Majul] + . Here the last character matches with the suffix list which on applying the stripping approach becomes [mazul] [E:Majul] which is not a location named entities.

Hand coded rules are also derived to identify different classes of NE. These rules are based on detailed analysis of the three named classes using different Assamese corpora available locally and data from the Internet. A person name can be a single-word or multi-word entity e.g., [Ram] [E:Ram] and [Ram kumaR] [E:Ram Kumar]. In the rule-based approach, a person name is determined based on its preceding and succeeding words, their associated attributes, like POS tags such as verb, common noun, etc., and certain clue words present in a name. To determine a single-word person name, the surrounding context is consider, as single-word names normally do not contain any clue word. In case of a multi-word person name, there are certain clue words associated with it. These words provide valuable information in determining if it is a person's name. Clue words can be broadly categorized into three classes, viz., title, surname, and middle word. A title like [sRimoti]or [sRiman] [E:This is used to identify a person name] and such a title normally marks the beginning of a person name and a surname normally signifies the end, although at times multiple surnames within a single name can be seen e.g., [Ram kumaR dEka] [E:Ram Kumar Deka].

Following are some of the rules to identify different classes of NE:

(1) If the previous and succeeding words are verbs, the current word is most likely to be a person name.

Example:
— [bHat khai Ram kHeliboloi goise] [E:Ram went to play after eating rice]. Here [Ram] [E:Ram] is NE, previous word [khai] [E:eat] and succeeding word [kHeliboloi] [E:play] are bot verbs.

(2) If two words in sequence are both verbs, the previous word is most likely to be a person name.

Example:

[3] www.lancaster.ac.uk/fass/projects/corpus/emille/

Table 1. Examples of Location Named Entities.

Root	Surface form
[osom] [E:Assam]	[osomiya] [E:Assamese]
[nEpal] [E:Nepal]	[nEpali] [E:Nepali]
[tezpuR] [E:Tezpur]	[tezpuRiya] [E:People of Tezpur]
[bHaRot] [E:India]	[bHaRotiyo] [E:People of India]
[osom] [E:Assam]	[osombasi] [E:People of Assam]
[zoRhat] [E:Place of Assam]	[zoRhtiya] [E:People of Jorhat]

— [komol douRi ahise] [E: Kamal came running]. Here [komol] [E:Kamal] is NE and [douRi ahise] [E: came running] are both verbs.

(3) If there exist a word like [nogoR] [E: town], [zila] [E:district], [sohoR] [E:city], [paRot] [E:Lane]. The previous word represents a location named entity.
Example - [kamRup zila] [E:Kamrup district], [sonitpuR zila] [E:Sonitpur district].

(4) If the current word is a number and the next word represents a unit of measurement such as [kilo] [E:Kilo] , [gRam] [E:Gram]; etc it represent a Measurement NE.
Example - [E:1 Kilo].

(5) If the current word is a digit and the following word is a month name, it represents a date NE.
Example - [E:1 June].

(6) If the current word is a number and the next word is a month name followed by a digit, it represents a date NE.
Example - [E:5 June 2011].

(7) If the current word is a digit followed by a word like [bojat], [minit], [ghonta], [sekend] [E:second, hour, minute], it represents time NE.
Example - , [E:3 mins].

(8) If there exists a month name preceded by a digit word list, it represents date NE.
Example - - [E:6-7 June].

(9) If there exists a digit followed by a word [son], [bosoR], it represents a date NE.
Example - [E:In 1992] , [E:10 year].

(10) If there exists a digit followed by a word [sonR] [E:year], a digit and a month name, it represents a date NE.
Example - [E:1980 year 23 May].

(11) If a month name is followed by a word like , it represents a month NE.
Example - [E:May month].

(12) If a digit exists in a range followed by a month, it represents a date NE.
Example - [E:1-4 June].

(13) If a dot exists between each consecutive letter, it is most likely to be an Organization NE.
Example - . . [b.j.p].

4.3 Gazetteer-based Approach

A traditional gazetteer is a dictionary or directory that contains information about geographical names that are found in a map. Such information may include physical features and social statistics of the place associated with the name. Since NER is the process of labeling of proper nouns into different categories, viz., person, location, organization, and miscellaneous, and gazetteers containing reference entity names that are labeled by human experts in pre-defined categories relevant to the task, gazetteers are useful for NER. For example a location gazetteer list may be used as a source of background knowledge to label location NEs. A gazetteer list of persons, locations and organizations are prepared for use in NER from various sources, including the Internet. The gazetteers are simply lists of names of the appropriate type. These lists are dynamic in nature as more names can be added to them later. The main merit of building a gazetteer list is that high accuracy can be obtained, depending on the size of the list. Common disadvantages of the gazetteer-based approach list include the following.

—The gazetteer list has to be updated regularly.

—Ambiguity exists among the words.

Three gazetteer lists are prepared, viz., titles, surnames, and middle names to accommodate these words and use them while deriving the rules. Maintaining such lists is relatively easy as the distinct number of clue words is limited. Since person names can have multiple words, it is necessary to give a proper labeling to a person name with a start and end tag.It is often hard to derive a specific set of rules to identify a location name without using any clue word list as location names hardly follow any specific pattern. A gazetteer for such clue words e.g., [nogoR], [zila], [kusi] etc. [Words in this lists are used for identifying place names comparable to English words like Town in Morgantown, ville as in Huntsville and Ton as in Edmonton.] are also prepared which helps to identify the location. Further, location names can also be found in combination with person name such as [maHatma ganDhi Rod] [E:Mahatma Gandhi Road].
A list of 700 organization names are collected and have analyzed that organization names also follow a specific pattern. An organization name always ends with an organization clue word such as [osomiya soNgotHon]] [E:Assamese Organization], [bHaRot soRkaR] [E:Indian Government], but, several consecutive organization clue words are normally not seen in a single name. Organization names like [osom bidyaloi bisHobidyaloi] [E: Assam School University] are not normally seen. So, if an organization clue word is found it can be marked as the end of an organization name.
Most organization names can be seen with middle clue words such as [bElguRi madHomik bidyaloi] [E:Belguri Secondary School]. In a single organization one or multiple middle clue words can exist. So, a list of middle clue words for organizations is also prepared.
There is huge number of organizations whose names are those of famous persons or derived from names of famous persons such as [kali cHaran dAs kolez] [E:Kali Charan Das College]. And if a middle clue word exists in an organization name, a person name normally comes before such clue words e.g., [kali cHaran dAs balok bidyaloi] [E:Kali Charan Das Boys School] but we do not normally see an organization name like [usHotoR

Table 3. Sizes of Gazetteer List

LIST	DATA
Surnames	6000
Locations	12000
Organization Clue Words	37
Organization Middle Words	24
Location Clue Words	29
Pre-Nominal Words	120
Organization names	800
Person names	9,6000

Table 4. Results obtained using gazetteer list.

Classes	Size	F-measure(%)	Size	F-measure(%)
Person	2298	74	6000	82.4
Location	8951	70.5	12000	78
Organization	500	78.8	800	80

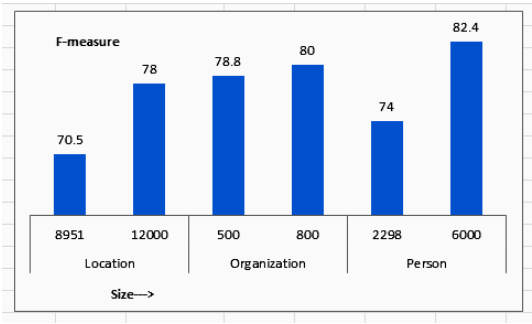

Fig. 1. Graphical Representation of Tagging of NE using Gazetteer list.

madHomik boRuwa kalEz] [E:Higher Secondary Barua College]. Two gazetteer lists are prepared, viz., organization clue words and middle clue words to accommodate these words and use them while deriving the rules.

The sizes of the different gazetteer lists prepared for NER are shown below in Table 3.

A test corpus of 100K wordforms is tagged with three labels of NEs, viz., person, location, and organization considering the gazetteer list for defined sets of classes. The results obtained for different classes of NEs are shown in Table 4. The same corpus of 100K wordforms is once again tested with the increase in the length of the gazetteer list which shows an improvement of 5-10% in accuracy.

Fig1 shows the graphical representation of tagging of NE using gazetteer list.

5. NER IN ASSAMESE USING MACHINE LEARNING

For a resource-rich language like English, NER has shown high accuracies. This section discusses work on NER using CRFs and HMM. The work on NER in Assamese using CRFs and HMM is the first such work. The ML approach has advantages over the rule-based approach in that it is adaptable to different domains and has robustness, whereas the rule-based approach is labor-intensive and time-consuming. Hence, ML approach is use, E.g., CRFs and HMM, which results in an accuracy of 75%-85%. Finally, a hybrid approach is propose which shows an improvement over both CRF and HMM.

5.1 Features used in Named Entity Recognition

Different types of contextual information along with a variety of other features are used to identify NEs. Prefixes and suffixes of words also play important roles in NER. The features used may be language-independent or dependent. Different language independent features that help in identifying NER include contextual information, prefixes and suffixes of all the words, NE tags of previous and following word(s), whether it is the first word, length of the word, whether the current "word" is a sequence of digits, whether the current word is infrequent and the POS of the current and surrounding words. In contrast, language dependent features include the set of known suffixes; clue words that help identify person, location and organization names; and designation words that help to identify person names which are described below.

Language independent features used in NER include the following.

(1) *Context word features*: Surrounding words, such as the previous and the next word of a particular word serve as important features when finding NEs. For example, a word like [zila], [puR] or [paRa] indicates the presence of a location. These words are used to identify location names. Similarly, [ustad] [E:Expert], [kriRabid] [E:Sportsman] and [kobi] [E:Poet] denote that the next word is a person name.

(2) *NE information*: The NE tag information for the previous and the following words are important features in deciding the NE tag of the current word. For example, [Ram osomoloi gol] [E:Ram went to Assam]. In this example, [Ram] is a person NE which helps identify that the next word is likely to be again a NE. Similarly in Bengali, [Ram asame giyesil] [E:Ram went to Assam] can also help to identify the person NE.

(3) *Digit features*: Different types of digit features have been used in NER. These include whether the current token is a two-digit or four digit number, or a combination of digits and periods and so on. For example, [5 jun 2011].

(4) *Organization suffix word list*: Several known suffixes are used for organizations. These help identify organization names. For example, if there exists a word like *Ltd* or *Co*, it is likely to be a part of an organization's name. Similarly for Indian languages also, there are some suffixes used for organization names such as [got] [E:Group], [soRkaR] [E:Govt].

(5) *Length of words*: It is often seen that short words less than 3 characters are not usually NE. But there are exceptions, e.g., [Ram] [E:Ram], [sita] [E:Sita], [Ron] [E:Ron].

(6) *POS*: Part-of speech is an important feature in identifying the NEs. For example, if two words in sequence are both verbs, the previous word is most likely to be a person name. Example: [komol douRi ahise] [E:Kamal came running]. Similarly in Bengali one can say as [komol kHeye gHumaise] [E:Kamal slept soon after having food].

Language dependent features used in NER include the following:

(1) *Action verb list*: Person names generally appear before action verbs. Examples of such action verbs in Assamese are [koisil] [E:told], [goisil] [E:Went]. [kothatu rame koisil] [E:Ram told it].
[shihotor ghoRot hoRi goisil] [E:Hari went to their home].

(2) *Word prefix and suffix*: A fixed-length prefix or suffix of a word may be used as a feature. It has been seen that many NEs share common prefix or suffix strings which help identify them. For

example, in Assamese [dada] [E:Older Brother], [baidEu] [E:Older Sister] are used identify person NEs. Similarly in Bengali, [dada] [E:Older Brother], [didi] [E:Older Sister] are used to identify a person NEs.

(3) *Designation words*: Words like Dr, Prof etc often indicate the position and occupation of named persons, serving as clues to detect person NEs. For example, in Assamese words like [profEsor dAs] [E:Professor Das], [montRi borai koi] [E:Minister Bora said].

5.2 NER in Assamese Using HMMs

An HMM is a statistical model that can be used to solve classification problems that have an inherent state sequence representation. The model can be visualized as a collection of a set of states. These states are connected by a set of transition probabilities, which indicate the probability of traveling between two given states. A process begins in some state, and moves through new states as dictated by the transition probabilities. In an HMM, the exact sequence of states that the process generates is unknown (i.e., hidden), hence it is a hidden model. The output of the HMM is a sequence of output symbols. A Markov chain assumes that the probability of a tag being the next state depends on the previous tag. For example, consider the sentence
Ram is playing cricket.
In this sentence after the verb *playing*, it is most likely that the next word will be a noun or preposition, with certain probabilities for each. The main aim of the Hidden Markov Model in tagging is to find the highest probability of a particular tag sequence for a given word sequence.
NER may be viewed as a classification problem, where every word is either part of some name or not part of any name.
The bigram statistical model is used to obtain the Name-Class (NCs) which is dependent on the previous word. For the purpose of name-finding, given a sequence of words (W), to find the most likely sequence of NC [9] ,i.e.,

$$max\,\Pr(NC|W). \tag{1}$$

By using Bayes theorem;

$$\Pr(NC|W) = \frac{\Pr(W, NC)}{\Pr(W)}. \tag{2}$$

Now, as Pr(W), the unconditioned probability of any word sequence, does not change with the different values of NC, the main aim is to maximize the numerator, i.e., the joint probability of the word sequence and the name-class sequence. The HMM approach for attaining this joint probability is based on the below three components: start probability, transition probability and observation probability. The entire space of all possible name-class assignments are searched, using the Viterbi algorithm (Viterbi, 1967), and maximizing the numerator Pr(W, NC)
The standard 3-fold cross-validation experiment is conducted. The test data and the training data are the same as used in CRF approach. Thus the f-measure for HMM approach is shown in Table5.

5.3 NER using CRF Approach

CRFs[8] are a type of discriminative probabilistic model used for labeling and segmenting sequential data such as natural language text or biological sequences. CRFs represent an undirected graphical model that define a single non-linear distribution over the joint probability of an entire label sequence given a particular observation sequence. CRFs can incorporate a large number of arbitrary, non-independent features and are used to calculate the conditional probabilities of values on designated output nodes, given the values on designated input nodes.
The conditional probability of a state sequence $S = (s1, s2..sT)$ given an observation sequence $O = (o1, o2, o3...ot)$ is calculated as

$$P(s|o) = \frac{1}{Z_o}\exp(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(S_{t-1}, S_t, o, t))$$

where Z_o is a normalization factor over all state sequence.

$$Z_o = \sum\exp(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(S_{t-1}, S_t, o, t))$$

and $f_k(S_{t-1}, S_t, o, t)$ is a feature function whose weight λ_k is to be learned via training.
When applying CRFs to the NER problem, an observation sequence is a token of a sentence or document of text and the state sequence is its corresponding label sequence. The library called Stanford NER is used, which is a simple, customizable, and open-source Java implementation of CRF for segmenting or labeling sequential data. For the Stanford NER, the training file should be in a tab-separated column, i.e. words in column 0 and the corresponding label in column 1. For this purpose the corpus is tokenized into words per line and is annotated with the three labels, viz., person, location, and organization. The standard 3-fold cross-validation experiment is conducted. In each fold, there is training data and test data. Then in each folder a learning model is created based on the training data. Out of .2 million wordforms, a set of 130K wordforms have been manually tagged with four tags namely person, location, organization and miscellaneous. This is used as the training set for the CRF based NER system and the rest 70K wordforms are considered test data. The words which were unseen during the training phase are assigned the class 0. The f-measure for CRF approach is shown in Table 6.

Table 5. Average HMM Results

Classes	F-measure(%)
Person	82.33
Location	82.13
Organization	79.6
Miscellaneous	81.25

Table 6. Average CRF Results

Classes	F-measure(%)
Person	78.4
Location	78.65
Organization	80.03
Miscellaneous	78.8

5.4 NER using Hybrid approach

A hybrid approach is an approach where more than two approaches are used to improve the performance of an NER system. The performance of Assamese NER is improved to some extent by integrating the ML approach with the rule-based and gazetteer-based approaches to develop a hybrid system. To the best of

Table 7. Average Result for Hybrid Approach

Classes	F-measure(%)
Person	85.8
Location	85.25
Organization	85.8
Miscellaneous	88.06

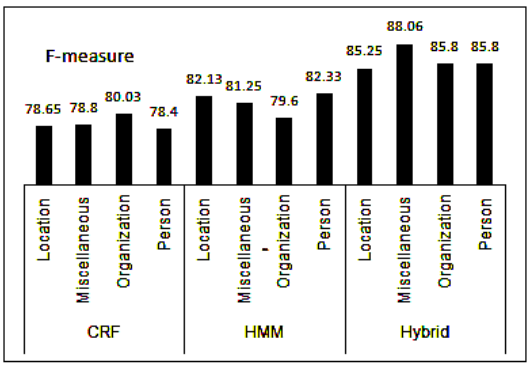

Fig. 3. Graphical Representation of HMM, CRF and Hybrid.

the knowledge, there is no work on hybrid NER in Assamese. Thus a hybrid NER system is develop that has the ability to extract four types of NEs. Each of the approaches has its own strengths and weaknesses. The processing goes through three main components: Machine-learning, rule-based, and gazetteer-based. The machine-learning component involves two approaches, CRFs, and HMM. Various NE features are used when implementing the two approaches. The rule-based approach involves the rules that is derived for different classes of NEs and the gazetteer-based approach involves the tagging of NEs using the look-up lists for location, person, and organization names.

The overall architecture of the hybrid approach is shown in Fig2. The results obtained after applying the hybrid approach are shown in Table 7.
The overall graphical representation of HMM, CRF and Hybrid approach is shown in Fig3.

6. DISCUSSION AND CONCLUSION

In natural language texts identification and classification of proper nouns is a challenging but useful task. While for many languages, it continues to be an active area of research, for many other languages, this work has barely started. The work NER in text in a resource-poor Indo-Aryan language, namely Assamese, has received little attention in computational linguistic research. NER is difficult as ambiguity exists among the different classes of NE. Various issues in NER in Indian languages and Assamese in particular is discussed. Assamese does not have the concept of capitalization, which is an important clue to identify the NEs in European languages. The suffix stripping is used for those words whose root words represent location NEs and has achieve an accuracy of 89%, but found that it produce errors when the last character of a word matches a suffix in the suffix list. Thus, in such cases the stemming approach should not be applied to a word, which itself represents a location NE. To remove these errors, a dictionary or gazetteer of location names is created where the most frequently occurring roots are kept and each word has to be checked against the list. Preparing a gazetteer list manually is a time-consuming process. Handcrafted rules are also derived for different classes of NER and found that handcrafted rules work well provided the rules are carefully prepared. The hand coded rules result in an accuracy of 70-75%. NER using a gazetteer list is also implemented which results in an accuracy of 75%-85% and have seen that the performance of the system increases as the size of the gazetteer list increases. Different ambiguities are encountered while preparing the gazetteer lists as the same name exists in different classes of NEs. Assamese being a highly inflectional language, special attention is required to handle morphological issues. Use of contextual information may be incorporated to reduce errors due to ambiguity. Two ML approaches namely CRFs and HMM which are existing approaches are implemented, but the work is for a new language where an annotated corpus was not available. Each of the approaches has its own merits and demerits. Various errors are encountered by the different approaches. CRF-based NER system encounters errors while labeling the NEs. Examples of such errors are: ( ) [mujomontRi potni huwar] [E:being the wife of Chief Minister], ( ) [Rajopal sRinibas] [E:Governor Srinivas] , Whenever the system find words like [mujomontRi], [Rajopal], the next word is tagged as a person name when obviously the are not. So, in such cases more careful rules need to be derived in order to avoid these errors. The only way to avoid these errors is to explore additional features besides the ones is used. Another way to improve the performance of the system is to increase the size of the training file and to explore some more features for each class. Some major issues are also encountered like ambiguity in names, unknown words when tagging a file with the HMM approach. Such errors are removed using the smoothing technique. Various language dependent and independent features are used when implementing the approaches. Both CRFs and HMM based NER systems perform well, but encounters problems, which are overcome to some extent using a hybrid approach which is a combination of the rule-based, gazetteer-based and ML approaches. The proposed hybrid system has achieved an overall improvement in Assamese NER performance. It is capable of recognizing four different types of NEs including person, location, organization, and miscellaneous which includes the date, time, and year. The experimental results show that the hybrid approach outperforms the pure rule-based approach, gazetteer-based approach, and the pure ML-based approach, with an F-measure of 80%-90%.

7. REFERENCES

[1] Borthwick Andrew. A Maximum Entropy Approach to NER. In *Ph.D thesis,Computer Science Dept, New York University*, 1999.

[2] D Appelt, R Hobbs, J Bear, D Israel, M Kaymeyama, A Kehler, D Martin, K Myers, and M Tyson. SRI International FASTUS system MUC-6 Test Results and Analysis. In *Proceedings of the Sixth Message Understanding Conference*, pages 237–48, Columbia, Maryland, 1995.

[3] Nancy Chinchor, Eric Brown, Lisa Ferro, and Patty Robinson. Named Entity Recognition Task Definition. August-27 1999.

[4] Cortes and Vapnik. Support Vector Network. *Machine Learning*, pages 273–297, 1995.

[5] R. Grishman and B Sundheim. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 466–71, Copenhagen, Denmark, 1996.

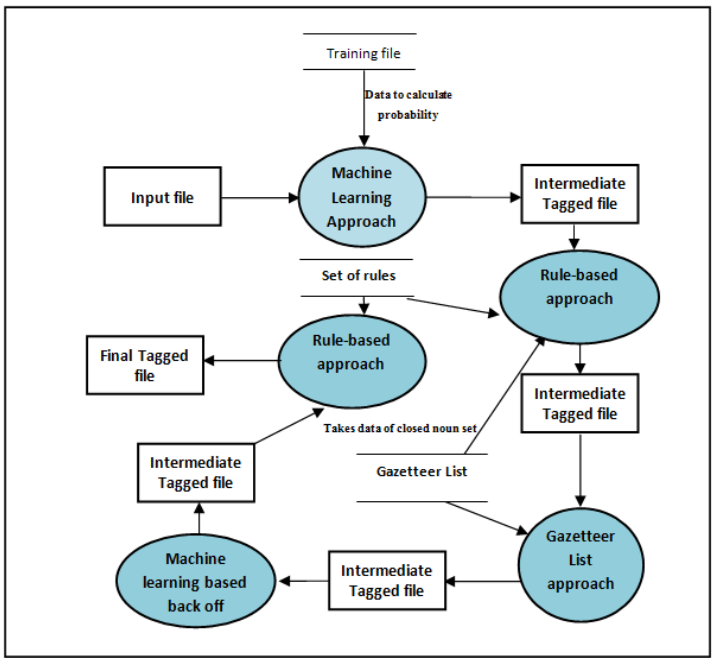

Fig. 2. Hybrid NER Architecture.

[6] Humphreys. K, Gaizauskas. R, Azzam. S, Huyck. C, Mitchell. B, Cunningham. H, and Wilks. Y. Description of the Lasie-ii System as Used for MUC-7. In *Proceedings of the 7th Message Understanding Conference*, Fairfax, VA, 1998.

[7] Kaufmann.M, Gaizauskas.R, Wakao. T, Humphreys. K, Cunningham.H, , and Wilks. Y. Description of the Lasie System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, pages 207–220, Columbia, Maryland, 1995.

[8] John Lafferty, Andrew McCallum, and Fernando Pereira. Probabilistic Models for Segmenting and Labelling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*, pages 282–289, Williams College, Williamstown, MA, USA, 2001.

[9] Bikel Daniel M, Miller Scott, Schwartz Richard, and Weischedel Ralph. A High Performance Learning Name-finder. In *Proceedings of the fifth Conference on Applied Natural language Processing*, pages 194–201, Washington, DC, USA, 1997.

[10] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, and the Annotation Group. BBN: Description of the SIFT System as Used for MUC-7. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*, pages 1–17, Fairfax,Virginia, 1998.

[11] Grishman Ralph. The New York University System MUC-6 or Where's the syntax. In *Proceedings of the Sixth Message Understanding Conference*, pages 167–175, Columbia, Maryland, 1995.

[12] Nina Wacholder, Yael Ravin, and Misook Choi. Disambiguation of Proper Names in Text. In *Proceedings of the Fifth Conference on Applied Natural Language*, pages 202–208, Washington Marriott Hotel, Washington, DC, USA, 1997.

[13] Takahiro Wakao, Robert Gaizauskas, and Yorick Wilks. Evaluation of an Algorithm for the Recognition and Classification of Proper Names. In *Proceedings of COLING-96*, pages 418–423, Copenhagen, Denmark, 1996.

[14] Shihong Yu, Shuanhu Bai, and Paul Wu. Description of the Kent Ridge Digital Labs System Used for MUC-7. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*, Fairfox, Virginia.