Chapter 3

Rule-based and Gazetteer-based NER for Assamese

3.1 Introduction

Different types of approaches exist in NER such as rule-based and ML-based approaches. In this chapter, we discuss NER in Assamese using rule-based and gazetteer-based approaches. Rule-based methods are seen to work well provided the rules are carefully prepared and cover all 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. We test NER in Assamese with rules as well as with gazetteer lists approach. Our rule-based approach involves the identification of the root word from an inflected or a derivational form; the process is known as stemming. Handcrafted rules are also used to identify different classes of named entities. The second approach of this chapter involves the NER using a gazetteer list of persons, locations and organizations.

3.2 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. Therefore, removal of suffixes from root words by automatic means is an important task. 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 *cats* it is *cat*. 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(for example Frakes and Yates[45]) to improve performance. When a user enters the word *running*, it is most likely that it will retrieve the word *run* in the documents. In highly inflectional languages such as Assamese, identification of the root form of words is more important. In other languages such as English and Indian languages like Assamese and Bengali, there are words which are not NEs, but their root words are NEs. For example the words ভাৰতীয় [bʱaɾɒtiyo] [E:Indian], or মনিপুৰী [mɒnipuɾi] [E:Manipuri], which are adjective, are not named entity, but the root words ভাৰত [bʱaɾɒt] [E:India], or মনিপুৰ [mɒnipuɾ] [E:Manipur] are.

3.2.1 Different Approaches to Stemming

There are several types of stemming algorithms that differ with respect to performance and accuracy. The following are some of the algorithms used in stemming[1].

[1]http://en.wikipedia.org/wiki/Stemming

1. Brute-Force approach.
2. Suffix-Stripping approach.
3. Suffix-Substitution approach.
4. Lemmatization approach.
5. Affix-Removal approach.
6. Production approach.
7. Matching approach.

Brute-Force approach:- In the brute force approach, we employ a look-up table or a database table which contains the relations between the root words and the inflected words. To stem a word, the table is searched to find a matching inflection. If a matching inflection or derivation is found, the root form is returned. For example, if we enter the word *cats*, it searches for the word *cat* in the list. When a match is found, it displays the word. The accuracy of the algorithm depends on the size of the database. The more the entries in the database, the higher the accuracy.

Suffix-Stripping approach:- It does not rely on a look-up table that consists of inflected forms and root form relations. This approach can also be called a rule-based approach. A list of rules is created and stored. These rules provide a path for the algorithm, to find the root form of a given input word form. For example, if the word ends in *es*, it remove *es*.

Suffix-Substitution approach:- This technique is an improvement over suffix stripping. The substitution rule replaces a suffix with an alternate suffix. Instead

of removing the suffix, it replaces it with some other suffix. For example, *ies* is replaced with *ey*. To illustrate, the algorithm may identify that both the *ies* suffix stripping rule applies as well as the suffix substitution rule. In suffix substitution *friendlies* becomes *ly* instead of *friendly*. Thus, *friendliness* becomes *friendly* through substitution and becomes *friendl* through stripping.

Affix-Removal approach:- This approach is a generalization of suffix stripping. The term affix refers to either a prefix or a suffix. This approach performs stemming by removing word prefixes and suffixes. It removes the longest possible string of characters from a word according to a set of rules. For example, given the word *indefinitely*, it identifies that the leading "in" is a prefix that can be removed.

Production approach:- Rather than trying to remove suffixes, the goal of a production algorithm is to generate suffixes. For example, if the word is *run*, then the algorithm might automatically generate the form *running*.

Matching Approach:- Such algorithms use a stem database. To stem a word the algorithm tries to match it with stems from the database, applying various constraints such as one based on the relative length of the candidate stem within the word, for example, the word *brows* in *browse* and in *browsing*. On the other hand, the short prefix *be*, which is the stem of such words as *been*, and *being* would be considered as the stem of the word *beside* as well. This is a deficiency of this method.

3.2.2 Previous Work on Stemming

A considerable amount of work exists in English stemming. There are a number of stemming algorithms for English such as those discussed by Paice[91], Lovins[73], Dawson[29], Krovertz[101] and Porter[75]. Most of the work was done using rule-based approaches[[75], [73]]. Among these, the Porter stemmer is the most prevalent one since it can be applied to other languages as well. It is a rule-based stemmer with five steps, each of which applies a set of rules. Indian-language work on stemming includes Larkey[70], Ramanathan and Rao[8], Chen and Grey[3], Dasgupta and Ng[28] and Sharma et al.[116]. Work on Assamese can also be found in Sharma et al.[117] where the authors developed a lexicon and a morphology rule base. Larkey[70] developed a Hindi lightweight stemmer using a manually-constructed list of 27 common suffixes that include gender, number, tense, and nominalization. A similar approach was used by Ramanathan and Rao[8] in which they used 65 of the most common inflectional suffixes. Chen and Grey[3] describe a statistical Hindi stemmer for evaluating the performance of a Hindi information retrieval system.
Sharma et al.[116] focus on word stemming of Assamese, particularly in identifying the root words automatically when a lexicon is unavailable. Similar work has been described in Dasgupta and Ng[28] for a Bengali morphological analyzer. Work on a morphological analyzer and a stemmer for Nepali was described in Bal and Shrestha[10]. Sharma et al.[118] discussed issues in Assamese morphology where the presence of a large number of suffixal determiners, sandhis, samas, and suffix sequences make almost half of the words inflected. Sarkar and Bandyopadhyay[110] present the design of a rule-based stemmer for Bengali. Paik and Pariu[53] used an n-gram approach for three languages, Hindi, Bengali and Marathi. Majgaonker and Siddique[76] describe a rule-based and unsupervised Marathi stemmer. Similarly, a suffix-stripping-based morphological analyser for Malayalam was described in Rajeev

et al.[95]. Kumar and Rana[67] describe a stemmer for Punjabi language that uses a brute force technique.

3.2.3 Assamese Language and its Morphology

Assamese is the principal language of Assam. It belongs to the Indic branch of the Indo-European family spoken by around 30 million people in Assam, Arunachal Pradesh and other north-eastern Indian states. A large amount of literary work can be found in Assamese language. Assamese has derived its phonetic character set and behavior from Sanskrit. It is written using the Assamese script. The history of the Assamese language can be classified in three different eras: Early Assamese, Middle Assamese and Modern Assamese. The way of writing Assamese is the same as that of English, i.e., from left to right. There is no upper case or lower case in this script. Bengali and Manipuri scripts are similar to the Assamese script except that the Assamese consonant ৰ (ra) is distinct, and Assamese ৱ wabo) is not found in Bengali.

The Assamese script consists of 11 vowels, 34 consonants and 10 digits. In addition to these, it also includes conjunct words (*juktakhars*), i.e. a combination of consonants. The punctuation marks used in Assamese are the same as those of Roman except the full stop which uses the symbol । (*daari*). In this language the basic differences exist among the following characters.

- Hraswa-i (ই) and dirgha-i (ঈ) and their corresponding operators (ি) and (ী).
- Hraswa-u (উ) and dirgha-u (ঊ), and their corresponding operators (ু and(ূ).
- Pratham-sa (চ) and dwitiya-sa (ছ),
  murdhanya-ta (ট) and dantya-ta (ত),

murdhanya-tha (ঠ) and dantya-tha (থ),
murdhanya-da (ড) and dantya-da (দ),
murdhanya-dha (ঢহ) and dantya-dha (ধ),
murdhanya-na (ণ) and dantya-na (ন),
talabya-sa (শ), murdhanya-sa (ষ) and dantya-sa (স),

Assamese is a free-word order language, which means that the position of a word in a sentence can change without changing the overall meaning. For example the sentence

মই আজি ঘৰলৈ যাম। [mɒi aji gʱɒɾɒloi jam] [E:I will go home today].

can be written in any of the following forms given below.

মই আজি ঘৰলৈ যাম। [mɒi aji gʱɒɾɒloi jam]
আজি মই ঘৰলৈ যাম। [aji mɒi gʱɒɾɒloi jam]
ঘৰলৈ আজি মই যাম। [gʱɒɾɒloi aji mɒi jam]
যাম আজি মই ঘৰলৈ। [jam aji mɒi gʱɒɾɒloi]

Here though the ordering of the words is changed the meaning remains the same. The predominant word order is subject-object-verb.
Example: ৰামে ভাত খালে [ɾamɛ bʱat kʱalɛ] [E:Ram ate rice].
ৰাধাই কিতাপ পঢ়িছে [ɾadʱai kitap pɒɾhisɛ] [E:Radha is reading book].
Assamese is highly inflectional in nature and requires morphological analysis of the input words for further processing. Morphological analysis deals with the identification of the internal structure of the words of a language. We find a large number of words in Assamese which are morphologically transformed from the root

word. Derivation and inflection through three types of transformations can be found, viz., prefix, suffix, and compound forms.

An example of a word with a prefix is অ [ɒ] + জ্ঞানী [gyani] = অজ্ঞানী [ɒgyani] [E:Fool].
An example of a word with suffix is বিলাত [bilat] + ঈ [ii] = বিলাতী [bilati] [E:Foreigner].
An example of a compound words is অসম [ɒsɒm] + বাসী [basi] = অসমবাসী [ɒsɒmbasi] [E:People of Assam].
In Assamese, a single root word may have different morphological variants. For example ৰামৰ [ɾamɒr], ৰামক [ɾamɒk], and ৰামলৈ [ɾamɒloi] [E:Ram's, for Ram, to Ram] are morphological variants of the word ৰাম [ɾam] [E:Ram].

Different types of named entities may occur with different affixes. Some of the suffixes found in Assamese for location NEs and organization NEs are shown in the Table 3.1.

Root Word	Suffixes	Surface form
মনিপুৰ [mɒnipuɾ] [E:Manipur]	ঈ [ii]	মনিপুৰী [mɒnipuɾi] [E:Manipuri]
ভাৰত [bʱaɾɒt] [E:India]	ঈয় [iiyo]	ভাৰতীয় [bʱaɾɒtiyo] [E:Indian]
ভাৰত [bʱaɾɒt] [E:India]	ৰত্ন [ɾɒtnɒ]	ভাৰতৰত্ন [bʱaɾɒtɾɒtnɒ] [E:Bharat Ratna]
কলেজ [kalɛz] [E:College]	ঈয়া [iya]	কলেজীয়া [kalɛziya] [E:Of College]
স্কুল [skul] [E:School]	ঈয়া [iya]	স্কুলীয়া[skuliya] [E:Of School]

Table 3.1: Example of suffixes found in locations and organizations

Like other Indian languages, Assamese also follows a specific pattern of suffixation, which is:
<token> = <root/stem word> + <inflection>.

For example, কিতাপৰ [kitapɒr] [E:From book] = কিতাপ [kitap] [E:Book] + ৰ [r], and কলমটো [kɒlɒmtu] [E:The pen]= কলম [kɒlɒm] [E:Pen] + টো [tu]

3.2.4 Assamese Corpora

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

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.

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 articles.

3.2.5 Our Approach

We have reported in Table 3.1 that 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 our approach is to generate the root word from a given input word resulting in a location NE. Some examples are given below in the Table 3.2. To obtain the root words, we

Table 3.2: Examples of location named entities

Root	**Surface form**
অসম [ɒsɒm] [E:Assam]	অসমীয়া [ɒsɒmiya] [E:Assamese]
নেপাল [b^{n}ɛpal] [E:Nepal]	নেপালী [nɛpali] [E:Nepali]
তেজপুৰ [tezpuɾ] [E:Tezpur]	তেজপুৰীয়া [tezpuɾiya] [E:People of Tezpur]
ভাৰত [b^{ɦ}aɾɒt] [E:India]	ভাৰতীয় [b^{ɦ}aɾɒtiyɒ] [E:People of India]
ভাৰত [b^{ɦ}aɾɒt] [E:India]	ভাৰতৰত্ন [b^{ɦ}aɾɒtɾɒtnɒ] [E:BaratRatna]
অসম [ɒsɒm] [E:Assam]	অসমবাসী [ɒsɒmbasi] [E:People of Assam]
যোৰহাট [zɒɾhat] [E:Jorhat]	যোৰহটীয়া [zɒɾhtiyɒ] [E:People of Jorhat]
বঙালকুছী [bɒngalkusi] [E:bongalkusi]	বঙালকুছীয়া [bɒngalkusiya] [E:People of bongalkusi]

use the suffix stripping approach. It is a fast process as the search is done only on the suffix. We have listed some suffixes that identify a location such as [বাৰী [baɾi], নগৰ [nɒgɒr], পাৰা [paɾa], পুৰ] [puɾ] [E: These suffixes are used to identify location names] and are described in Appendix A. These suffixes are normally attached to the location root word to represent a location NE. Example of such NEs are বিৰুবাৰী [biɾubari] [E:Birubari] and খানাপাৰা [k^{ɦ}anapara] [E:Khanapara]. We also see that some suffixes apply to both as person names as well as locations such as [ৰ [ɾ], ক [k], লৈ [loi]] [E:These suffixes are used for both person names and location names]. For

example: [ভাৰতক [bʰaɾɒtɒk] , হীৰেনক [hiɾenɒk]],[অসমলৈ [ɒsɒmɒloi], ৰামলৈ [ɾamɒlɔi]].
Assamese uses derivations to add additional features to a root word to change its grammatical category. Examples of such derivation are in Table 3.2. These words are actually formed by adding features like ীয়া [iiya],ী [ii],ীয় [iiyɒ]], etc., to location names after which they are no longer NEs. But when these words are stemmed using the suffix stripping approach, they produce NEs which fall under the location category.
Below are the steps used in our approach. The input is an Assamese word which occurs in a given text. The key file contains all possible suffixes such as:

ীয়া [iiya],ী [ii],ীয় [iiyɒ], বাসী [basi], ৰত্ন [ɾɒtnɒ] [E:These are some suffixes when used with location names changes its meaning].

that combine with the location NEs resulting in the change of the meaning of the word. Finally the output is the root word that falls under location NE category.

- Input the word.
- Compare the input word with the keyfile[ীয়া,ী,ীয়, বাসী, ৰত্ন].
- If the last character of the input word matches with an entry in the key file, then
- Remove the suffix of the input word that matches with the keyfile from R.H.S.
- Exit.

The flow chart of our approach is shown in Fig 3.1.

Figure 3.1: Flow Chart Of The Process

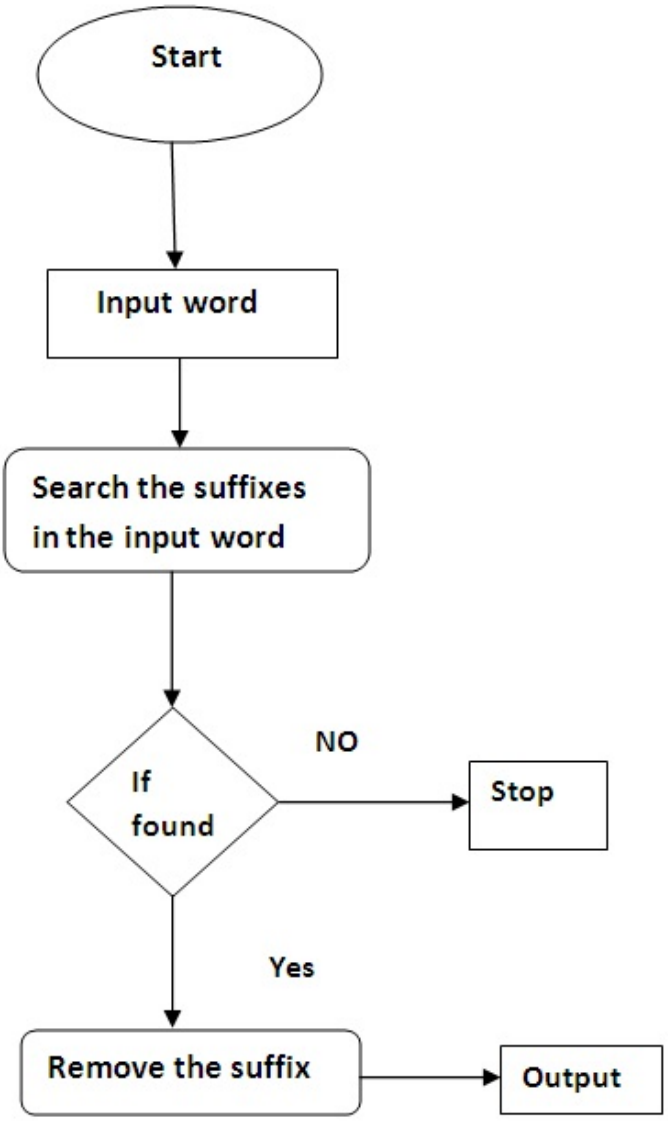

Table 3.3: Suffix stripping experiment for location named entity

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

3.2.6 Results and Discussions

In this experiment, we use a part of the Asomiya Pratidin corpus of size 20K words. We use suffix stripping for those words whose root words represent location NEs. The statistics of the training data and the effectiveness of the method are given in Table 3.3. The f-measure of 89% obtained in this experiment shows the effectiveness of the proposed approach. We also observe some errors in our experiment. 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:

- মাজুলী [mazuli] [E:Majuli].
- গুৱাহাটী [guwahati] [E:Guwahati].
- ধুবুৰী [dʰubuɾi] [E:Dhubri].
- তিনচুকীয়া [tinsukiya] [E:Tinsukia].
- গৰৈমাৰী [gɒɾɒimaɾi] [E:Goroimari].

- তেতেলীয়া [tɛtɛliya] [E:Teteliya].

We see that some input words themselves are root words, but the last character matches a suffix from the suffix list such as [মাজুলী [mazuli], গুৱাহাটী] [guwahati], etc. In these cases the stemmer removes the last characterী [ii] from the word [মাজুলী [mazuli], গুৱাহাটী] [guwahati], producing wrong stems [মাজুল [mazul], গুৱাহাট [guwahat]], which are no longer location NEs whereas the original words themselves represent location NEs. Thus, in such cases the stemming approach should not be applied to a word, which itself represents a location NE. To remove these errors, we have created a dictionary or gazetteer of location names where the most frequently occurring roots are kept and each word has to be checked against the list which is described in Section 3.3.

We have also derived some hand coded rules to identify the different classes of NE. These rules are based on detailed analysis of the three named classes using the Assamese corpora available locally and data from the Internet. A person name can be a single-word or multi-word entity, e.g., ৰাম [ɾam] [E:Ram] and ৰাম কুমাৰ [ɾam kumaɾ] [E:Ram Kumar]. In our rule-based approach, a person name is determined based on its preceding and succeeding words, their associated attributes, like verb, common noun, etc., and certain clue words present in a name. To determine a single-word person name, we consider the surrounding context, 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 a person's name. Clue words can be broadly categorized into three classes, viz., title, surname, and middle word. A title like শ্ৰীমতি [sɾimɒti] or শ্ৰীমান [E:These are used to identify a person name] etc normally marks the beginning of a person name and a surname normally signifies the end, although at times we

see multiple surnames within a single name e.g., ৰাম কুমাৰ ডেকা [ɾam kumaɾ dɛka] [E:Ram Kumar Deka]. We have prepared three gazetteer lists, viz., title, surname, and middle name 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. We have prepared a gazetteer for such clue words e.g., নগৰ [nɒgɒɾ], জিলা [zila], কুচি [kusi] etc, [E: These lists are used for identifying place name] which help identify location names described in Appendix A. Further, location names can also be found in combination with person name such as ৰাধা গোবিন্দ বৰুৱা ৰোড [ɾadʱa gɒbindɒ bɒɾuwa ɾɒd] [E:Radgha Govindo Baruah Road].

We have made a list of 700 organization names and have analyzed that organization names also follow a specific pattern. An organization name always ends with an organization clue word such as অসমীয়া সংগঠন [ɒsɒmiya sɒŋgɒtʱɒn]] [E:Assamese Organization], ভাৰত চৰকাৰ [bʱaɾɒt sɒɾkaɾ] [E:Indian Govt], but, no consecutive organization clue words are normally seen in a single name. Organization names like অসম বিদ্যালয় বিশ্ববিদ্যালয় [ɒsɒm bidyalɒi bisʱɒbidyalɒi] [E: Assam School University] are not normally seen. So, if we find an organization clue word it can be marked as the end of an organization name.

Most organization names can be seen with middle clue words such as বেলগুৰী মাধ্যমিক বিদ্যালয় [bɛlguɾi madʱɒmik bidyalɒi] [E:Belguri Secondary School]. In a single organization one or multiple middle clue words can exist. So, we have prepared a

list of middle clue words for organizations.

There are a large number of organizations whose names are those of famous persons or derived from names of famous persons such as ৰাধা গোবিন্দ বৰুৱা কলেজ [ɾadʱa gɒbindɒ bɒɾuwa kɒlez] [E:Radha Govindo Baruah College]. And if a middle clue word exists in an organization name, a person name normally comes before such clue words e.g., ৰাধা গোবিন্দ বৰুৱা বালক বিদ্যালয় [ɾadʱa gɒbindɒ bɒɾuwa balɒk bidyalɒi] [E:Radha Govindo Barua Boys School] but we do not normally see an organization name like উচ্চতৰ মাধ্যমিক বৰুৱা কলেজ [uɲɔtɒɾ madʱɒmik bɒɾuwa kalɛz] [E:Higher Secondary Barua College]. We have prepared two gazetteer lists, viz., organization clue words and middle clue words to accommodate these words and use them while deriving the rules. The organization clue words and the middle name clue words are listed in Appendix A. Based on these analyses, we have postulated some rules to tag the person, location, and organization. We have also derived rules to identify the date, time, and measurement NE.

1. *Rules for organizations*:

   A: Find organization clue words based on the common organization gazetteer list.

   B: If found, tag the word as the end word of an organization.

   C: Search the middle clue words in the organization names.

   D: If found, search for the previous word in the same gazetteer list else

   E: Search for the previous word in the surname gazetteer list.

   F: If found, search for the previous word in the same gazetteer list else

   G: Search for the previous word in the title list.

   H: If found, mark it as the beginning of an organization NE tag, else,

   I: Mark as a beginning in the next word.

Examples of Organization:

(a) বেলগুৰি উচ্চতৰ মাধ্যমিক স্কুল [belguri utʃotɒr madʱyɒmik skul] [E:Belguri Higher Secondary School] .

(b) অসম বিশ্ববিদ্যালয় [ɒsɒm bisʱobidyalɒi] [E:Assam University] .

(c) ৰাধা গোবিন্দ বৰুৱা কলেজ [radʱa gɒbindɒ bɒruwa kɒlez] [E:Radha Govindo Baruah College].

Example of Organization that fails:

(a) অসম কাজিৰঙা বিশ্ববিদ্যালয় [ɒsɒm kazirɒŋga bidyalɒi] [E:Assam Kaziranga University].

(b) শ্ৰী শ্ৰী শঙ্কৰদেব উচ্চতৰ মাধ্যমিক স্কুল [sri sri sɒŋkɒrdeb madʱyɒmik skul] [E:Sri Sri Shankardev Higher Secondary School].

2. *Rules for person names*:

- Rules for multiword person names:

  A: Find the surname based on the common surname gazetteer list.

  B: If found, tag the word as the end word of a person NE

  C: Search the previous word in the surname gazetteer list.

  D: If found, search the previous word in the same gazetteer list else

  E: Search the previous word in the title list.

  F: If found, mark it as the beginning of a person NE tag else,

  G: Mark it as the beginning in the next word.

  Example of multiword person name:

  (a) শ্ৰী হীৰেন কুমাৰ বৰুৱা [sri hirɛn kumar bɒruwa] [E:Sri Hiren Kumar Barua].

  (b) ৰাম শৰ্মা দেউ [ram sɒrma deb] [E:Ram Sharma Deb]

- *Rules for single word person names*:
  - If the previous and succeeding words are verbs, the current word is most likely to be a person name.
    Example:
    (a) ভাত খাই ৰাম খেলিবলৈ গৈছে [bʰat khai ɾam kʰelibɒloi gɔise] [E:Ram went to play after eating rice].
    (b) খাই উঠি ৰামে পঢ়ি আছে [kʰai utʰi ɾam pɒɾhi ase] [E: Ram is studying after eating].
  - If two words in sequence are both verbs, the previous word is most likely to be a person name.
    Example:
    (a) কমল দৌৰি আহিছে [kɒmɒl dɒuɾi ahise] [E: Kamal came running].
    (b) হৰিয়ে কিতাপ পঢ়ি আছে [hɒɾiye kitap pɒɾhi ase] [E:Hari is reading book].

3. *Rules for Location*: If there exist a word like নগৰ [nɒgɒɾ], জিলা [zila], চহৰ [sɒhɒɾ], পাৰত [paɾɒt] [E:These words are used to represent a place name] etc., the previous word represents a location named entity.
   Example - কামৰূপ জিলা [kamɾup zila] [E:Kamrup district], শোণিতপুৰ জিলা [sɒnitpuɾ zila] [E:Sonitpur district].

4. *Rules for miscellaneous*:
   - If the current word is a number and the next word represents a unit of measurement such as কিলো[kilɒ] [E:Kilo] , গ্ৰাম[gɾam] [E:Gram] etc, it represent a Measurement NE.
     Example - ১ কিলো [E:1 Kilo].

- If the current word is a digit and the following word is a month name, it represents a date NE.
  Example - ১ জুন [E:1 June].
- 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 - ৫ জুন ২০১১ [E:5 June 2011].
- If the current word is a digit followed by a word like বজাত [bɒjat], মিনিট [minit], ঘন্টা [ghɒnta], ছেকেন্ড [sekend] [E:second, hour, minute], it represents a time NE.
  Example - ২ বজাত, ৩ মিনিট [E:3 mins].
- If there exists a month name preceded by a digit word list, it represents date NE.
  Example - ছয়-সাত জুন [E:6-7 June].
- If there exists a digit followed by a word চন [sɒn], বছৰ [bɒsɒɾ], it represents date NE.
  Example - ১৯৯২ চনত [E:In 1992] , ১০ বছৰ [E:10 year].
- If there exists a digit followed by a word চনৰ [sɒnɾ] [E:year], a digit and a month name, it represents date NE.
  Example - ১৯৮০ চনৰ ২৩ মে [E:1980 year 23 May].
- If a month name is followed by a word like মাহ it represents month NE.
  Example - মে মাহত [E:May month].
- If a digit exists in a range followed by a month, it represents date NE.
  Example - ১-৪ জুনত [E:1-4 June].
- If a dot exists between each consecutive letter, it is most likely to be an Organization NE.

Example - বি. জে. পি [b.j.p].

The above rules can also be represented as regular expression[121].

3.3 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. Gazetteers can be obtained for persons and organizations as well. Since NER is the process of labeling of proper nouns into different categories, viz., person, location, organization, and miscellaneous, and gazetteers contain reference entity names that are labeled by human experts in pre-defined categories relevant to the task, hence gazetteers are useful for NER. For example a location gazetteer list may be used as a source of background knowledge to label location NEs. We have prepared a gazetteer list of persons, locations and organizations for use in NER from different sources, including the Internet. Our 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 merits 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

We have encountered several issues while preparing the gazetteer lists such as ambiguity among different classes of NEs, common noun and proper noun and

spelling variations of a word. Some of the examples are listed below in the Table 3.4.

Table 3.4: Examples of issues in preparing gazetteer list

Spelling Variation	Person vs Location	Common noun vs Proper noun
সিং,সিঙ [singh]	কাশী [kas^{h}i]	জোন [jun]
দলগাৱ,দলগাওঁ [dalgaon]	বিশ্বনাথ [bis^{fi}nath]	পাঠক [pat^{fi}ok]
গলগাৱ,গলগাওঁ [galgaon]	মাৰঘেৰীতা [marg^{H}erita]	মালিক [malik]
ডেৰগাওঁ,ডেৰগাৱ [dergaon]	পশুপতি [dpas^{fi}pati]	বিশ্বাস [biswas]

We have manually tagged a corpus of 300K wordforms which can be used for training and validation of statistical approaches to NER. Besides preparing the gazetteer lists for person, location, and organization for Assamese, we have also prepared lists of clue words for identifying person, location, and organization which we use in the rule-based approach. The list is shown in Appendix A. The sizes of the different gazetteer lists we prepared for our purpose is shown in Table 3.5. The data collected for the different gazetteer lists are from various sources like the Internet, and newspaper articles. The surname gazetteer list consists of surnames used by people of different parts of Assam, and also some from foreign countries. Our list consists of 1000 Assamese surnames, 4000 from different parts of India, and 1000 from outside India. Some of the sources used for collection of the data are listed below.

1. http://en.wiktionary.org/wiki/Appendix:Indian-surnames.

2. http://en.wikipedia.org/wiki/Category:Assamese-language-surnames.

3. http://www.list4everything.com/list-of-common-surnames-in-assam.html.

LIST	DATA
Surname	6000
Location	12000
Organization Clue Words	37
Organization Middle Words	24
Location Clue Words	29
Pre-Nominal Words	120
Organization names	800
Person name	9600

Table 3.5: Data for gazetteer list

4. http://en.wikipedia.org/wiki/List-of-the-most-common-surnames-in-Europe.

5. www.usgennet.org/.../ne/.../ne-s...foreign surname.

6. http://www.rong-chang.com/namesdict/100-last-names.htm.

7. http://familypedia.wikia.com/wiki/List-of-most-common-surnames.

Similarly, the location gazetteer list consists of the names of districts, cities, and villages of different parts of Assam, cities and villages of India, and some of foreign countries. Our list contains 4,309 places of Assam, 5,754 places of different parts of India, and 1,000 from outside India. Some of the sources used for collection of the data are listed below.

1. http://www.listofcountriesoftheworld.com/.

2. http://www.indianmirror.com/tourism/home/alphabetical-list.html List of cities in India.

3. http://en.wikipedia.org/wiki/List-of-districts-of-Assam.

4. http://en.wikipedia.org/wiki/Category:Cities-and-towns-in-Assam-district.

Besides the above sources a possible source of free gazetteer data is geonames(http://www.geonames.org). 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 3.6. 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.

Table 3.6: 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

Preparing a gazetteer list manually is a time-consuming and labor-intensive work. Moreover, it is not possible for these lists to be exhaustive. It is seen that the accuracy of a system increases as the lists increase in size. So, we prefer to prepare automatic gazetteers to make the process easier and faster.

We have prepared an automatic gazetteer list based on the tagged corpus. Our program will automatically read every word from the newly-tagged corpus and write it in three different gazetteer lists as per the tag, viz., person, location and organization. It is possible that a particular word may already exist in the gazetteer lists, so a counter is used which is associated with every entry in the list. For every new entry of a particular word the counter value will be always 1, and this is incremented by 1, for each subsequent entry. This list consists of words which falls under the open NC. The counter value gives a general idea of the occurrence of a

particular word as a person name, location name or organization name. This list is used to tag the remaining words of a corpus which are left untagged after applying the ML, rules-based and gazetteer-based approaches.

In the unigram approach if a particular word is present in more than one list, we consider the counter value to determine the highest probability of a word to be tagged. But if the counter value is the same, the word is tagged based on the precedence of person-location-organization. This precedence is based on our general study of words to be most likely on which category it falls.

Thus the pseudocode for preparing the automatic gazetteer list is as follows:

- Step1: Consider a word automatically tagged as a person from the tagged output file.
- Step2: Compare the tagged word of the output file in the person gazetteer list.
- Step3: If the word is present in the list, increment the counter i.e with each entry by one.
- Step4: Otherwise, append the word in the list with counter value 1.
- Step5: Repeat the above four steps for all the words tagged as person in the output file.
- Step6: Repeat steps 1 to 5 for organization and location.

3.3.1 Results and Discussions

The second approach of this chapter discusses the tagging of a corpus using gazetteer list. Our gazetteer list is simply a look-up table for three classes of NEs, namely, person, location, and organization. Ambiguity exists among the classes of NE while preparing the list. A test corpus of 100K wordforms is used and our system shows an improvement of 5-10% with increase in the length of the list.

3.4 Conclusion

This chapter discusses NER using the rule-based and gazetteer-based approaches. In NER, identification of the root word is an important task. We have tested a very simple stemming approach for location NER which is very fast and also performs reasonably well with an F-measure of 89%. Our experimental results also produce some errors as discussed in Section 3.2.3. One may also implement our process using finite state automata as well to improve efficiency, provided the list of suffixes is known during program coding. We have also derived some hand-crafted rules that help identify the person, location, organization, date, and time NEs. We have prepared different gazetteer lists manually and a corpus has been tagged using these lists and found that the accuracy of a system depends on the size of the gazetteer list. Since preparing a list manually is labor-intensive, we suggest a method for automatic gazetteer updation which helps increase the length of the list. We came across different ambiguities 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.