Named Entity Recognition of Indian Origin Names in English Documents

Chaitanya Gupta, Deepanshu Sood, Mahua Bhattacharya

ABV-Indian Institute of Information Technology & Management
Morena Link Road Gwalior 474010, India
Corresponding author e-mail: bmahua@hotmail.com

Abstract: Named Entity Recognition (NER) is the task of identifying and classifying all proper nouns in a document as person names, organization names, location names, date & time expressions and miscellaneous. There has been a growing interest in this field since early 1990s. Earlier, work has been done on NER taking English language as the medium. Apart from that some researchers have also tried their hands on Hindi and regional languages such as Telugu. The objective of our project is to identify names of Indian origin in English documents. The idea is to cover cross-linguistic aspects of text while performing NER on Indian names. The proposed project mainly distinguishes persons, organizations, locations and contact numbers in a document. The approach adopted is mainly unsupervised learning based on the feature space. Gazetteers are also used to improve the results of the experiment. The application is developed in C#.NET using the IDE of Visual Studio 2008.

Keywords: NER, Indian names, Text mining, Hindi names in English documents.

1. INTRODUCTION

The objective of NER is to classify all tokens in a text document into predefined classes such as person, organization, location, miscellaneous. In evaluations at the Message Understanding Conferences of the 1990s, it became clear that in order to reasonably extract information from documents, it is useful to first identify certain classes of information referred to in the text. They therefore established the Named Entity Task, where systems attempted to identify dates, times, numerical information and names \[1\]. At the time, MUC was focusing on IE tasks wherein structured information on company and defense-related activities are extracted from unstructured text, such as newspaper articles. In defining IE tasks, people noticed that it is essential to recognize information units such as names including person, organization, and location names, and numeric expressions including time, date, money, and percentages. Identifying references to these entities in text was acknowledged as one of IE’s important sub-tasks and was called “Named Entity Recognition (NER).” Before the NER field was recognized in 1996, significant research was conducted by extracting proper names from texts. A paper published in 1991 by Lisa F. Rau \[2\] is often cited as the root of the field. Named Entity Recognition has remained an essential component of Information Extraction (IE) and related NLP tasks. NER also finds application in question answering systems and machine translation. NER is an essential subtask in organizing and retrieving biomedical information \[3\]. NER can be treated as a two step process

- identification of proper nouns.
- classification of these identified proper nouns.

A large number of techniques have been developed to recognize named entities for different languages. Some of them are Rule based and others are Statistical techniques. The rule based approach uses the morphological and contextual evidence of a natural language and consequently determines the named entities. This eventually leads to formation of some language specific rules for identifying named entities. The statistical techniques use large annotated data to train a model (like Hidden Markov Model) and subsequently examine it with the test data.

In its canonical form, the input of an NER system is a text and the output is information on boundaries and types of NEs found in the text. This work is about the creation of an autonomous NER system, which based on some rules and feature space will be able to recognize Indian origin names in English documents.

We may list some tasks related to NER. These tasks revolve around the notion of rigid designation, whereby the direct goal is not to recognize the named things from documents. We also thoroughly survey fifteen years of research—from 1991 to 2006—in a systematic review published in a special issue of
**Personal name disambiguation** is the task of identifying the correct referent of a given designator. For example, it may consist of identifying whether Sachin Bansal is the race driver, the film editor, or the Flipkart founder in a given context. Corpus-wide disambiguation of personal names has applications in document clustering for information retrieval. In the work of Mann and Yarowski, it is used to create biographical summaries from corpora.

**Named entity translation** is the task of translating NEs from one language to another.

**Analysis of name structure** is the identification of the parts in a person name. For example, the name “Doctor Saurav R. Sharma” is composed of a person title, a first name, a middle name, and a surname. It is presented as a preprocessing step for NER and for the resolution of co-references to help determine, for instance, that “APJ Abdul Kalam” and “President Kalam” are the same person, while “APJ Abdul Kalam” and “Shahid Kalam” are two distinct persons.

**Acronym identification** is described as the identification of an acronym’s definition (e.g., “ACM” stands for “Association for Computing Machinery”) in a given document. The problem is related to NER because many organization names are acronyms (GE, NRC, etc.). Resolving acronyms is useful, again, to build co-reference networks aimed at solving NER. On its own, it can improve the recall of information retrieval by expanding queries containing an acronym with the corresponding definition.

**Record linkage** is the task of matching named entities across databases. It involves the use of clustering and string matching techniques in order to map database entries having slight variations (e.g., Sachin Tendulkar and S. Tendulkar). It is used in database cleaning and in data mining on multiple databases.

**Case restoration** consists of restoring expected word casing in a sentence. Given a lower case sentence, the goal is to restore the capital letters usually appearing on the first word of the sentence and on NEs. This task is useful in machine translation, where a sentence is usually translated without capitalization information.

1.3 Earlier Research: Computational research aiming at automatically identifying NEs in texts forms a vast and heterogeneous pool of strategies, methods, and representations. One of the first research papers in the field was presented by Lisa F. Rau at the 7th IEEE Conference on Artificial Intelligence Applications. Rau’s paper describes a system to “extract and recognize [company] names.” It relies on heuristics and handcrafted rules. From 1991 to 1995, the publication rate remained relatively low. It accelerated in 1996, with the first major event dedicated to the task: MUC-6. It has not decreased since, with steady research and numerous scientific events: HUB-4, MUC-7 and MET-2, IREX, CONLL, and HAREM.

A good proportion of work in NER research is devoted to the study of English, but a possibly larger proportion addresses language independence and multilingualism problems. German is well studied in CONLL-2003 and in earlier works. Similarly, Spanish and Dutch are strongly represented, and were boosted as the focus of a major conference: CONLL-2002. Japanese has been studied in the MUC-6 conference, the IREX conference, and other works. Chinese is studied in abundant literature, and so are French, Greek, and Italian.

**PROBLEM STATEMENT**

As discussed in earlier section, a lot of work has been done on English and other languages such as German as well. Hindi in its pure form witness a lot of challenges as Hindi is a kind of unstructured language where subject can come early or later to predicate.

People have considered phoneme-based approach for finding named entities in Hindi language in past. However, the case we consider in this thesis is when a document covers Hindi names and places in an English document, i.e., addressing cross-linguistic issues while extracting information from a document.

English is the third most spoken language in the world and most of the countries have adopted it and created their own form of spoken English. In India also, most of the print media have adopted Hinglish (Hindi+English) as a common notion of information sharing.

Since most of the text and information on internet and print media is available in English language, it is important to come up with an approach that can effectively extract information available in Hindi language from those documents.

Our main focus would be to extract names, locations, organizations and contact numbers from the documents.
3. METHODOLOGY

While early studies were mostly based on handcrafted rules, most recent ones use supervised machine learning (SL), as a way to automatically induce rule-based systems or sequence labeling algorithms, starting from a collection of training examples.

3.1 Supervised Learning: The current dominant technique for addressing the NER problem is supervised learning. SL techniques include Hidden Markov Models (HMM), Decision Trees, Maximum Entropy Models (ME), Support Vector Machines (SVM), and Conditional Random Fields (CRF). These are all variants of the SL approach, which typically feature a system that reads a large annotated corpus, memorizes lists of entities, and creates disambiguation rules based on discriminative features.

3.2 Semi Supervised Learning: The term “semi-supervised” (or “weakly supervised”) is relatively recent. The main technique for SSL is called “bootstrapping” and involves a small degree of supervision, such as a set of seeds, for starting the learning process. For example, a system aimed at “disease names” might ask the user to provide a small number of example names. Then, the system searches for sentences that contain these names and tries to identify some contextual clues common to the five examples. Then, the system tries to find other instances of disease names appearing in similar contexts. The learning process is then reapplied to the newly found examples, so as to discover new relevant contexts. By repeating this process, a large number of disease names and a large number of contexts will eventually be gathered.

3.3 Unsupervised Learning: The typical approach in unsupervised learning is clustering. For example, one can try to gather NEs from clustered groups based on context similarity. Basically, the techniques rely on lexical resources, on lexical patterns, and on statistics computed on a large corpus.

The approach we would be adopting is a blend of Semi-supervised and Unsupervised learning with the help of a feature space. Firstly, machine analyses the document based on patterns or features present. If features clearly indicate word to be name or place, then it is categorized as such, else gazetteer would be looked into and based on list look-up for a particular word, rules would be derived.

3.4 Feature Space: Features are describers or characteristic attributes of words designed for algorithmic consumption. The system has two types of rules:

- a recognition rule (for example, capitalized words are entity candidates)
- a classification rule (for example, the type of entity candidates of length greater than or equal to 3 words is organization)

The features that we will be using in identifying names, location, organizations are listed as follows:

| TABLE 1: Feature space for identifying names (PER) |
| --- | --- |
| Case | Name begins with Capital letter |
| Length | More than or equal to 3 characters |
| Titles | Dr., Mr., Mrs. |
| Part of Speech | Use of ‘he’, ‘she’, ‘I’ relates to a person |
| Punctuations | Presence of apostrophe s (‘s) |
| Grammar | Next character such as ‘is’ denotes an entity. |
| Frequency of occurrence | A Person’s name does not occur too frequently in a document. |

| TABLE 2: Feature space for identifying locations (LOC) |
| --- | --- |
| Case | Name begins with Capital letter |
| Length | More than or equal to 3 characters |
| Morphology | Common ending. Examples: ‘ore’ in Bangalore and Mangalore. |
| Punctuations | Presence of apostrophe s (‘s) |
| Grammar | Use of ‘in’ or ‘at’ before the entity refers to a location |
| Frequency of occurrence | A Place’s name does not occur too frequently in a document. |

| TABLE 3: Feature space for identifying organizations (ORG) |
| --- | --- |
| Case | Name begins with Capital letter, All letters capital or mixed case |
| Length | More than or equal to 3 characters |
| Frequency of occurrence | An organization’s name does not occur too frequently in a document. |
| Punctuations | Use of ‘-’or ‘.’ In between or at the end of the entity or special characters such as ‘&’. |

There are many challenges that might come across in this model. For example, after encountering an entity with initial capital letter, we mark it as PER, what if other entity starts from the second word again with a capital letter. It actually
denotes that the first entity didn't end and second entity is a part of the first entity like a last name of a person. So we use B-PER for first entity encountered and E-PER to denote that it is a part of previously found B-PER entity. Similarly, we can resolve the same issue for organizations by using B-ORG and I-ORG.

Based on the above feature-space discussed, we will differentiate if an entity is a person, organization or a location.

3.5 Proposed algorithm

1. Application fulfills the purpose of Named Entity Recognition in two ways: by simply entering the text on home screen or by uploading a word document.
2. The text is read from the textbox or from the document and read word by word.
3. Let's say we pick up a word, X at a time.
4. We check if X starts from a number or ‘+’? If yes, then jump to next step else go to step 6.
5. Call a routine to check if X fulfills the criteria to be called a contact number.
6. Check if X has first letter capital? If no, then pick up next word and go back to step 4, else go to next step.
7. Check if X is not an article or interrogative forms (The, These, They, Are, Is, Was, When, Why etc.)? If yes, go to next step else pick up next word and go back to step 4.
8. Check if X is a title (Mr., Mrs., Dr. etc.)? If yes, then mark X as ‘PER’ and look for B-PER in next word and I-PER next to that, else go to next step.
9. Check for X in all possible feature space, i.e. for Name, Place and Organization.
10. For whichever feature space, X satisfies the highest ratio of features classifies into that feature-set.

3.6 Flowchart

The flowchart shows various modules of the application and the flow of the application. The entities once recognized are checked for features against different feature-space of name, place or organization/miscellaneous. Contact numbers are identified separately in the beginning only depending on the occurrence of the digits in a word.
4. OBSERVATIONS AND RESULTS

The two basic parameters to judge the output are: Precision and Recall.

\[
\text{Recall} = \frac{\text{Number of NEs detected by the system}}{\text{Number of NEs present in the gold standard test set}} \times 100\%
\]

\[
\text{Precision} = \frac{\text{Number of detected NEs that are correct}}{\text{Number of NEs detected by the system}} \times 100\%
\]

We will test the application on some random excerpts of data and calculate the recall and precision for them.

Some screenshots have been pasted for example:

Recall = 80%, Precision = 75%, f-score = 77.42%

Recall = 75%, Precision = 83%, f-score = 78.797%

On an average, including many other test cases performed on this application, the Recall came to be more than 83% and Precision is around 90%.

Gazetteer is being improvised to enhance the results of the application.

5. CONCLUSION AND FUTURE WORK

The limitation with this model is that it never gives 100% accurate result but the results can be improved. Most of the previous works have achieved a precision of about 80%. To improve the result, A Gazetteer might also be used but will be an over-head on the complexity of the system and processing time.

Till now, based on the results obtained while testing the application, it can be said that application is able to give a satisfactory performance but efforts are being made to improve the result and reach a Recall and Precision of almost 100% if not exactly 100%.

Also, once named entities have been recognized from a given text, relations can be derived based on the context. This can be the further step in the project once best results from this application have been achieved in extracting named entities.
REFERENCES


