Review of Name Entity Recognition in Marathi Language

Ms. Jayshri Arjun Patil[1], Ms. Poonam Bhagwandas Godhwani[2]
[1, 2] Department of Computer Science & Technology
[1, 2] UTU-Bardoli (Gujarat), India

Abstract- *Named Entity Recognition (NER) is a subtask of Information Extraction. Named Entity Recognition (NER) is a task to discover the Named Entities (NEs) in a document and then categorize these NEs into diverse Named Entity classes such as Name of Person, Location, River, Organization etc. This paper gives a brief introduction to Name Entity Recognition. We also discusses about different approaches, issues and challenges in Marathi Language Name Entity Recognition.*

I. INTRODUCTION

NER can be defined as a two stage problem:- Identification of Proper Noun and classification of the Proper Noun into a set of classes such as Person names, Location names(cities, countries etc.), Organization names(Companies, government Organizations, committees etc.), Miscellaneous(date, time, percentage, monetary expression, number expression and measurement expression). Named Entity Recognition (NER) is an important tool in almost all of the Natural Language processing applications such as Information Retrieval, Information Extraction, Question Answering, Machine Translation and Automatic Summarization etc.

II. DIFFERENT APPROACHES OF NAME ENTITY RECOGNITION

A) Linguistic Approach
B) Machine learning based Approach.
C) Hybrid approach

A. Linguistic Approach/Rule based Approach:

Rule based approach mainly concerned with manual rules written by the linguistics. Many rules based NER contains:

a) Lexicalized Grammar
b) Gazetteer list
c) List of triggered words

The main disadvantages of these rule-based techniques are:

- They require huge experience and grammatical knowledge on the particular language or domain.
- The development is generally time-consuming.
- Changes in the system may be hard to accommodate.
- These systems are not transferable, which means that one rule-based NER system made for a particular language or domain cannot be used for other languages or domains.

B. Machine Learning Based- approach

Machine learning based approach concerned with some pre-defined methods. These are:
a) Hidden Markov Models(HMM)
b) Decision Trees
c) Maximum Entropy Models(ME)
d) Support Vector Machines(SVM)
e) Conditional Random Fields(CRF)

C. Hybrid Approach

It is an approach where more than two approaches are used in order to improve the performance of the NER system, so the hybrid approach may be a combination of HMM model and CRF model or CRF and ME approach or Gazetteer method with HMM approach etc.

III. PROBLEM FACED IN INDIAN LANGUAGE

Since for English Language lots of NER system has been built. But such NER system for Indian Language cannot be used because of the following reason:
a) Detection of NEs in raw information
b) It is not easy in Indian languages because Indian languages do not have capitalization.
c) Indian names are ambiguous and this issue makes the recognition a very difficult task.
d) Indian languages are relatively free-order languages [9].
e) Indian language is inflectional and morphologically rich [9].
f) Non- availability of large Gazetteer

Work on NER in Indian languages is a difficult and challenging task and also limited due to scarcity of resources.

IV. ISSUE AND CHALLENGES IN MARATHI NAMED ENTITY RECOGNITION

1) Ambiguities in named entity classes

Ambiguities in names where the words have multiple interpretations while analyzing the text containing words with NEs making NER process difficult. Word Sense Disambiguation (WSD) is used to resolve the ambiguities in the text to determine the classification of a named entity.

2) Abbreviations and non-local dependencies

Multiple tokens can be written in different ways such as abbreviations or long form, usually first instance with descriptive long formulation followed by instances with short forms or aliases. Such tokens sometimes require same label assignments or require cross referencing. This ability of a system refers to non-local dependency. External knowledge is required to deal with non-local dependencies. Construction of external knowledge including names lists and expansive lexi consis not easy since domain lexicons and names are continuously expanding. Named entities can be composed of single or multiple words chunk of text. Parsing prediction or name chunking model is required to predict whether consecutive multiple words belong to same entity[7].

3) Agglutinative and inflectional nature of Marathi

Marathi is agglutinative language. Unlike English prefixes and suffixes are added to root words in Marathi to form meaningful contexts[7]. Sometimes in Marathi inclusion of suffixes or prefix to root word leads to change in semantic. So a difficult and critical situation is raised to use gazetteers, dictionaries, similarity measurement and pattern matching techniques to recognize Marathi names. Dictionaries or gazetteers contain entities without any suffix added. In Marathi suffixes are added to words in order to create the meaningful context. A well written stemmer is required for morphologically rich language Marathi to separate the root from the suffix in order to compare the word forms with gazetteer or dictionary entries. Next, it cannot be claimed that stemming will solve the problem completely because adding suffixes to roots may change the grammatical category of the root word, which may result in wrong entity recognition.

4) Spelling variations

In Marathi, words containing the four vowels इ (i), vowel sign (ि◌◌), or ई (ī), vowel sign (◌◌ी), उ(u), vowel sign (◌◌ु), ऊ (ū), vowel sign(◌◌ू) do not make phonetic difference but differs in writingand spellings. The words such as आ_ण (grammatically correct) and आणी (grammaticallyincorrect) or पाणी (grammatically correct) and पानी (grammatically incorrect) are interchangeablyused in many writings [7]. It is not necessary to completely follow Marathi grammar in free style textwriting which may affects text recognition systems.

5) Foreign words

Foreign words in some instances of person, organization, location and miscellaneous names that are English words appear in Marathi texts which are spelled inDevanagari script. The real challenge lies in recognition of such foreign words. It is very difficult to create gazetteers that include such names because they are not limited.

6) Dialects

Marathi is spoken using many dialects such as standard Marathi, Warhadi, Ahirani, Dangi,Vadvali, Samavedi, Khandeshi, and Malwani in various regions of India. There are specific wordsused in each dialect to express the text. Words from different dialects also appear in Marathi text [7].

V. CONCLUSION

Named Entity Recognition (NER) is an important tool in many NLP applications such as Information Retrieval, Information Extraction, Question Answering, Machine Translation and Automatic Summarization etc. NER is difficult and challenging for Marathi language because it has various issues like ambiguity, Agglutinative and inflectional nature, Abbreviations and non-local dependencies, Spelling variations etc. In this paper many approaches have been discussed which may be used to develop NER Marathi system.

REFERENCES

[1] SudhaMorwal, NusratJahan "Named Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013, ISSN: 2277 128X

[2] Nita Patil, Ajay S. Patil, B.V. Pawar " Survey of Named Entity Recognition System respect to Indian and Foreign Languages" International Journal of Computer Applications (0975 – 8887) Volume 134 – No.16, January 2016

[3] SudhaMorwal, NusratJahan “Named Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages” International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 4, April 2013ISSN: 2277 128X

[4] Deepti Chopra, NusratJahan, SudhaMorwal “Hindi Named Entity Recognition by Aggregating Rule Based Heuristics And Hidden Markov Model” International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.6, November 2012

[5] NusratJahan , SudhaMorwal and Deepti Chopra “Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach” NusratJahan et al./ International Journal of Computer Science & Engineering Technology (IJCSET)

[6] SujanSaha, Sanjay Chatterji, SandipanDandapat “A Hybrid Approach for Named Entity Recognition in Indian Languages” Proceedings of the IJCNLP-08 Workshop on NER for South East Asian Languages, Pages 17-24, Hydrabad, India January 2008

[7] Nita Patil, Ajay S. Patil and B.V. Pawar “ISSUES AND CHALLENGES IN MARATHI NAMED ENTITY RECOGNITION” International Journal on Natural Language Computing (IJNLC) Vol. 5, No.1, February 2016

[8] GauriDhopavkar 1,4 , Manali Kshirsagar2 , Latesh Malik3 “EXPLOITING RULES FOR RESOLVING AMBIGUITY IN MARATHI LANGUAGE TEXT” International Journal of Research in Engineering and Technology, eISSN: 2319-1163 | pISSN: 2321-7308

[9] Padmaja Sharma, Utpal Sharma, JugalKalita May 2011, “Named Entity Recognition: A Survey for the Indian Languages”.