Mini Project Presentation

IIIT Allahabad,Aug-Dec,2018

DATE:20/11/18

NAMED ENTITY RECOGNITION FOR HINDI

UNDER THE GUIDANCE OF PROF. **UMA SHANKAR TIWARY**

Group Members:-

- DEEPANSHU GOYAL(IIT2016037)
- SUBHAM RAJ(IIT2016010)
- ANKIT KUMAR(IIT2016024)

Problem Statement

- The Aim of the project is to collect Hindi data and to identify Named Entity in it and then classify it into categories:-
  - Person
  - Location
  - Organization
  - Other

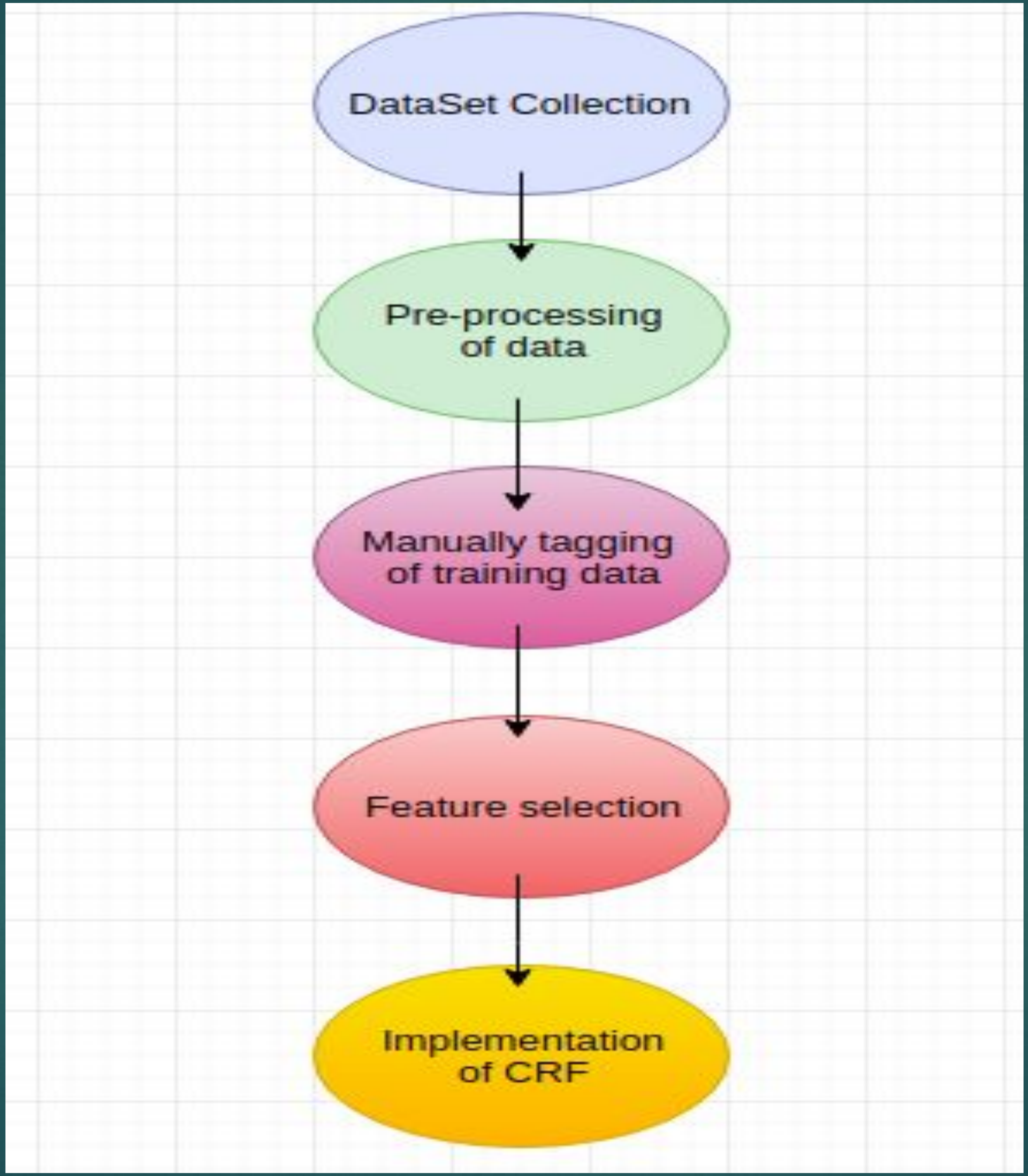
Motivation

- Although a lot of work has been done in English and other foreign languages like Spanish, Chinese etc. with high accuracy but the research regarding NER for Hindi is still in its infancy . So we decided to work in the field Named Entity Recognition for Hindi.

INTRODUCTION

- **Named-entity recognition** (**NER**) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Work Flow


DataSet Collection
Pre-processing of data
Manually tagging of training data
Feature selection
Implementation of CRF

Data Set Collection

- Dataset contains approximately 3,000 hindi sentences having 50,000 words related to Indian religions,sports,politics taken from Hindi Monolingual Text Corpus ILCI-II.
- Dataset contains approximately 5000 named entity featuring.

- ## Preprocessing Of Data:-
  - Sentence Breaking

Sentence Breaking

- In Hindi we denote the end of a sentence by what is called a 'poorna viraam' which is denoted by '।'. Our goal here was to identify that and taking care of all the scenarios, break the entire text into sentences.

Sentence Breaking(continued)

In order to implement this,we had written a java code which works as follows:-

- Firstly we split the text as per whitespaces in each line.After this ,our dataset had word and it’s POS tag in individual lines.
- Then we split according to ‘\’ and give whitespaces between word and it’s POS tag.
- After completion of every sentence ,we had given linespace.

Tagging the Training Data

- We manually tagged the training data in CoNLL format whose attributes are words,it’s POS and named entity.
- If it’s POS is not noun(N_NNP),then it’s named entity is set as others(O).
- Example of CoNLL format
  - विदेह N_NNP Location
  - की PSP O
  - राजधानी N_NN O
  - मिथिला N_NNP Location

Approach for Named Entity Recognition:-

- **CONDITIONAL RANDOM FIELD :-**

It is aim to develop a standalone system based on CRF approach. The purpose of analysis is two fold. First, finding out the useful features for NER task and second, find out the optimum feature set for the task in hand.

In crf,we are using **L-BFGS** algorithm.

Approach for Named Entity Recognition(continued):-

FEATURES EXTRACTION :-

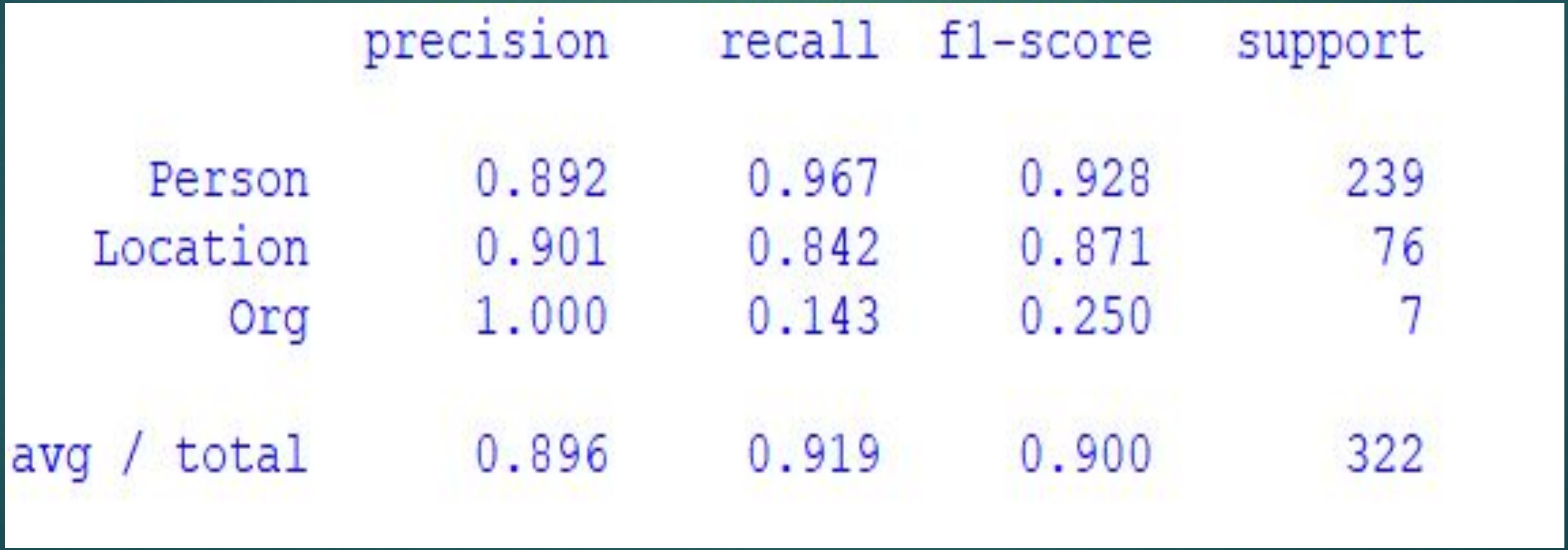
Features are the heart of CRF model. As inputs are not feeded directly rather it is in computed form for which we define features to it.

Our project highlights mainly 3 features which are as follows:

- Word Features:current word,previous word and next word
- POS tag of current word,previous word and next word
- suffixes of current word

Output & Screenshot:

- To predict and evaluate the accuracy of Named Hindi Recognition in Hindi.



```
             precision    recall  f1-score   support

     Person      0.892     0.967     0.928       239
   Location      0.901     0.842     0.871        76
        Org      1.000     0.143     0.250         7

avg / total      0.896     0.919     0.900       322
```

Output & Screenshot(continued):

1)Precision- Precision is ratio of correctly predicted values to the totally predicted values.

Precision=Correctly Predicted/Total Predicted

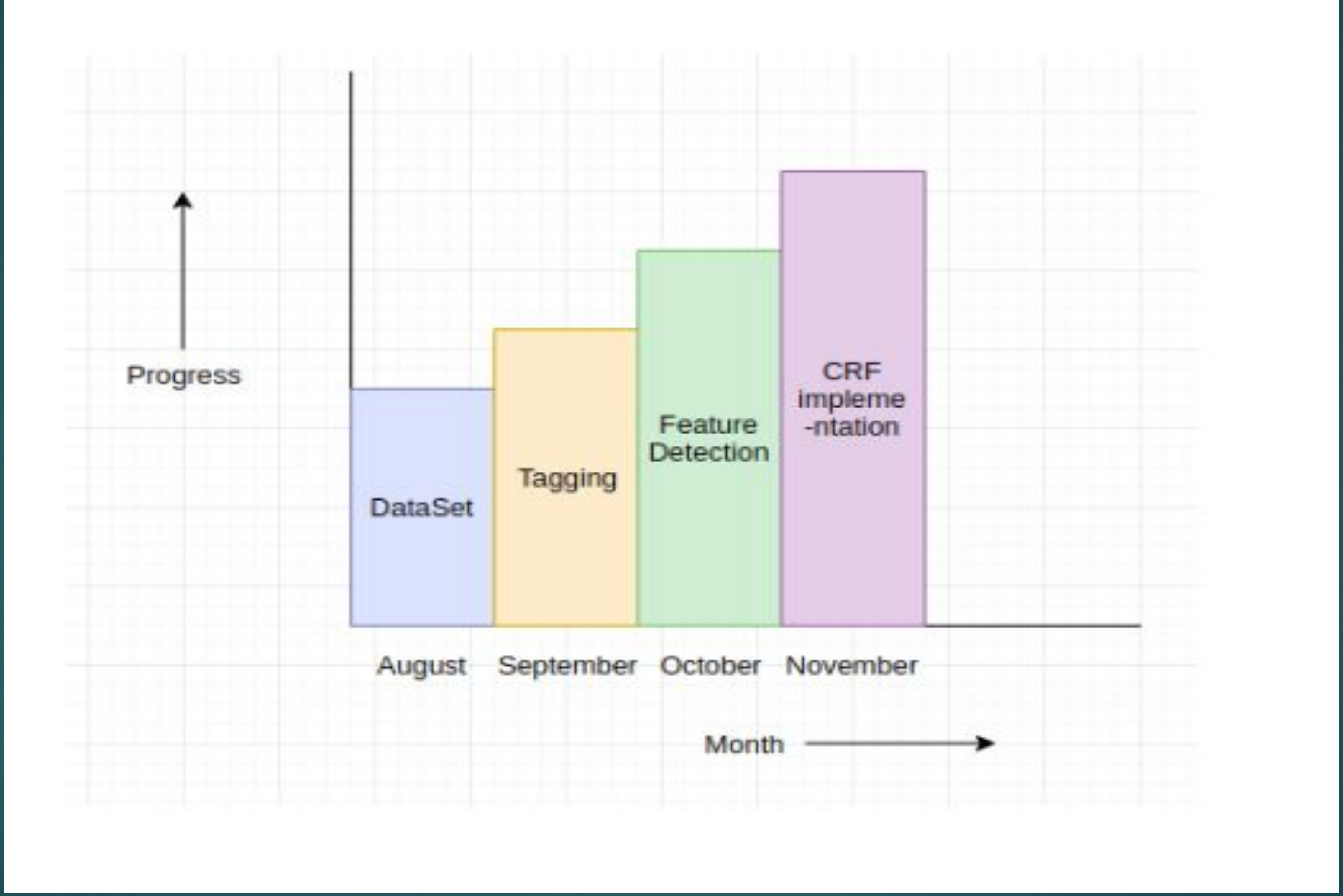
2)Recall- Recall is ratio of correctly predicted values to the all observations in actual class.

Recall=Correctly Predicted/Actual values.

3)F1-score-F1 score is the weighted average of precision and recall.

F1-score=2*(Recall*Precision)/(Recall+Precision)

Timeline:


Progress
DataSet
Tagging
Feature Detection
CRF impleme -ntation
August
September
October
November
Month

Thank You