MNIT-FIRE-NER-HINDI

Report for NER Track at FIRE 2013

Lokesh Kumar Sharma[1], Dr. Namita Mittal[2]

[1]PhD Scholar, [2]Asst. Professor
MNIT Jaipur
lokesh.gbu@gmail.com, mittalnamita@gmail.com

Abstract. Our system is tagging named entities given in test files. For this we are training our system using previously given training dataset. We have recognized all possible NER tags and trained our system for these tags. Finally we collect all NER tags of test file in tab-separated format.

1 Introduction

Named entity recognition (NER) or tagging is the task of finding names such as organizations, persons, locations, etc. in text. Our system is tagging the named entities from the annotated training files. There are several other methods and algorithms for named entity recognition such as a simple semi-supervised algorithm for named entity recognition [1] and ranking algorithms for named–entity extraction: boosting and the voted perceptron [2].

2 Methodology

We have used the supervised learning, as we are given few training files. We used this training dataset to train our system for tagged named entities and kept these tags in a tab separated files. Later on our system uses these tags for named entity tagging.

2.1. Algorithm and flowchart

There were 80 test files for named entity tagging. We designed our system in such a way so it can process all 80 files together but it was making system performance very poor so we took 10 files at a time for tagging.

Hence, we need to process 8times for tagging all given test files.
Elementary idea was to train our system for all possible named entities for this we

used the provided training dataset. The dataset is primary focus for our training. Following flow-chart is showing the basic flow of our system in detail.

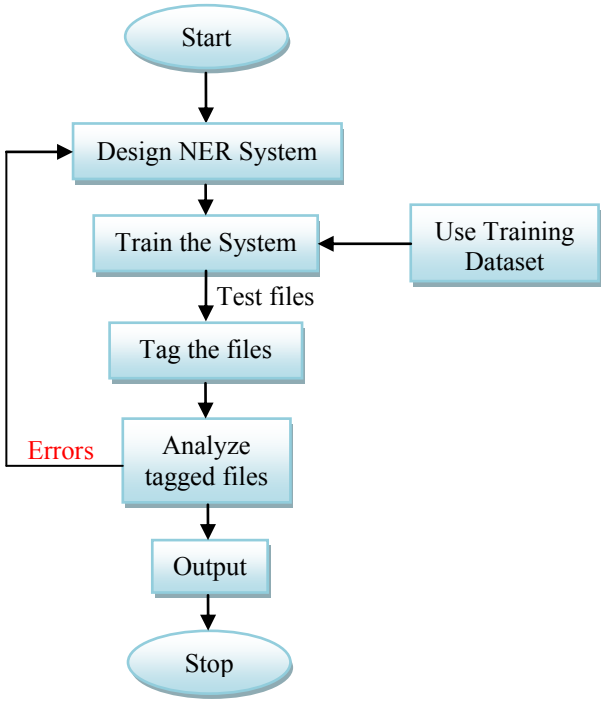

Figure 1: Basic flowchart of our system working

The final tagged results are in following format. First three are taken from the same test file and the last tag is taken by comparing the name with the training dataset. i.e.

भारत	NNP	I-NP	B-LOCATION
एक	QC	B-NP	B-COUNT
सदी	NNPC	I-NP	I-PERIOD
प्रदेश	NNP	I-NP	I-LOCATION

2.2. Technology used

We have used **JDK 1.6** for developing our system and we have also used **NetBeans 6.8 IDE** as a supporting tool for designing our system. Running platform was **Windows 7 home basic-64 bit**.

References

[1]. Wenhui Liao and Sriharsha Veeramachaneni. "A Simple Semi-supervised Algorithm For Named Entity Recognition", *Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing,* pages 58–65, Boulder, Colorado, June 2009.

[2]. A. M. Di Sciullo. "Ranking Algorithms for Named–Entity Extraction: Boosting and the Voted Perceptron", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* Philadelphia, July 2002, pp. 489-496.