Study and Implementation of Named Entity Recognition in English Language

Poonam Kumari¹ and Dr. Mahesh Yadav²

Abstract—In this paper, it is tried to introduce the development of NER System for English. It also tried to deal with various NER issues related to language. A List Look-Up Approach is used in the development of the NER System and Named Entity Tag set is associated with this NER System. This paper results in NER open source, NEP, for NEO(Named Entity Organization-English) for NEL(Named Entity Location-English), NEM(Named Entity Movie), NEBAT(Named Entity Banking-Terms).

Keywords—NER, NEP, NEO, NEL, NEM, NEBAT.

I. INTRODUCTION

English is a West Germanic language originally spoken in England, and is now is the most extensively used as a primary language by majority of the citizens of several nations, including the United Kingdom, the United States, Canada, Australia, Ireland and New Zealand. It has become third commonly used native language in the world, after Mandarin Chinese and Spanish [1]. It is widely accepted as a second language and is an official language of the European Union, many Commonwealth countries and the United Nations, as well as in many world organizations [2].

English is spoken by approximately 375 million people as their first language across the world. It is the third largest language by number of native speakers in today’s world, after Mandarin Chinese and Spanish [3].

Named Entity Recognition started with various Rule based approaches which uses dictionaries, lexicons and grammar for recognizing entities. With passing time, to develop accurate NER systems, the need of machine learning approaches was felt extensively. Hidden Markov Model [5] is used by NER systems as a very effective form of generative model that defines some joint probability distribution with observation and label sequences. Various independent assumptions are made for the words in advance in HMM models. One more approach termed as Maximum Entropy [4] is also used in NER systems. The Approach plays a very important role in the development of NER system. The accuracy of Named Entities (NE) depends largely on approach used in the process.

System Description

The Developed System is using XAMPP (i.e. for windows) which is an open source and free cross-platform web server solution stack package. It consists of MySQL database, Apache HTTP Server, and interpreters for scripts written in the PHP and Perl programming languages [6].

Figure 1 Logo of XAMPP

XAMPP’s name is an acronym for:

- X (to be read as "cross", meaning cross-platform)
- Apache HTTP Server
- MySQL
- PHP
- Perl

The program is released under the terms of the GNU General Public License and is a free web server able to serve dynamic pages. XAMPP is available for various Operating Systems Like Microsoft Windows, Linux, Solaris, and Mac OS X, and is extensively used for web development projects while creating dynamic webpages using programming languages like PHP, JSP, Servlets [7].

XAMPP setup needs only one zip, tar, 7z, or exe file to be downloaded and executed, with little or no configuration of the different components that make up the web server is required. XAMPP gets regular updates to get in tune with the latest releases of Apache/MySQL/PHP and Perl [8]. It is also available with a number of other modules like OpenSSL and phpMyAdmin.

Self-contained, multiple instances of XAMPP can run on a single system, and any given instance can be easily transferred to another machine. It is available in both full, standard version and a smaller version with limited facilities.

Officially, XAMPP’s development team designed it for use as a development tool only, to help web designers and developers to analyse their effort on their own PC, with or without Internet[9]. To keep this simple, a lot of important security features are disabled by default. In practice, it is often used to actually serve web pages on the Internet. A specializes used to provide support to create and manipulate databases in MySQL and SQLite. Once XAMPP is installed localhost can be treated like a remote host by connection using an FTP client. Using a program like FileZilla has many advantages when installing a content management system (CMS) like Joomla. Localhost can also be connected via FTP with your HTML editor.

Poonam Kumari, Computer Science Engineering Department, MRK IET, Rewari, Haryana, INDIA.
Dr. Mahesh Yadav, Computer Science Engineering department, MRK IET, Rewari, Haryana, INDIA.
The default MySQL user is "root" while there is no default MySQL password.

II. RESULT AND DISCUSSIONS

The NER system developed in PHP for English and Hindi is the result, which is as shown in the below figure. The developed system for English will recognize the Named Entities written in English.

The developed NER System for English can be shown in the below figure.

The result obtained for NEP (Named Entity Person-English) is shown in the below figures.

The result obtained for NEO (Named Entity Organization-English) is as below.

The result obtained for NEL (Named Entity Location-English) is as below.

The result obtained for NEM (Named Entity Movie) is as below.

The result obtained for NEBAT (Named Entity Banking-Terms) is shown in the below figures.
III. CONCLUSION

In this research work, the Named Entity Recognition for English has been explored. A lot of new tag sets have been identified for particularly recognizing Named Entities. The Gazetteers list is maintained which contains data from lot of areas which can be maintained or updated easily and the data set will be extended to many more fields and it can be updated easily. Its scope can be extended to develop the Gazetteers list which can uniquely identify the Named Entities used in different fields like to identify the Medical Related Entities, Business & Marketing Related Entities.

REFERENCES