ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

STUDY AND IMPLEMENTATION OF NAMED ENTITY RECOGNITION IN HINDI LANGUAGE USING PHP

Poonam Kumari*, Dr. Mahesh Yadav
Department of Computer Science Engineering[1,2], MRKIET, Haryana, India

ABSTRACT

NER Technique for HINDI language and a lot of NER issues relating to language has been developed and studied in this paper. This System includes a List Look-Up Approach for Named Entity Tagset associated with NER System. This paper results in 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.*

INTRODUCTION

Standard Hindi, or more precisely Modern Hindi, also known as Manak Hindi (Devanagari: मानक हिन्दी; meaning "Standard Hindi"), High Hindi, Nagari Hindi, and Literary Hindi, is a standardised and sanskritised chapter of the Hindi-Urdu language which is based on the Khariboli dialect of Delhi and Western UP. It is the official language of the Republic of India [1].

Hindi is mutually intelligible with another Hindustani language called Urdu. Mutual intelligibility decreases in literary and specialized contexts which rely on educated vocabulary [2]. Because of religious nationalism and communal issues, speakers of both languages at times assert that thesee are distinct languages, despite the fact that native speakers cannot distinguish the colloquial languages apart. The combined population of Hindi-Urdu speakers in the world is the fourth largest. However, the exact number of native speakers of Hindi is unclear. According to the 2001 census, 258 million people in India informed their native language to be "Hindi". However, it contains large numbers of speakers of Hindi dialects besides Standard Hindi; as of 2009, the best figure find for Khariboli dialect was a 1991 citation of 180 million.[3]

NER ISSUES IN RELATED LANGAUGE

There are many issues which make the nature of the problem different for Indian languages. Some of these issues are:

No Capitalization: Hindi scripts don't have Capitalization feature, which acts as an important indicator for NER.

Ambiguities: The ambiguity problem in NER occurs when one token represents more than one entity[4].

For example the names of the Organization vs. Location.

eSa mrj izn"k esa jgus------(Loc)
mrj izn"k f"k{kk cksMZ-------(Org)

Spelling Variation: one more important language related issue is the variation in the spellings of person names. For ex: the same person name
uhjt xqIrk, MkW-xqIrk, MkW-,u-xqIrk

and so on. This results in increase in number of words in the List[5].

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.78

PHP IMPLEMENTATION

PHP is a general-purpose server-sidescripting language which was originally designed for Web development. It is one among the first developed server-side scripting language to be embedded into an HTML source document rather than calling an external file for processing data. The code is first interpreted by a Web server with a PHP processor module by which resulting webpage is generated. It also has a command-line interface and can be used in standalone graphical applications. PHP can be used with most Web servers and also as a standalone shell with almost every operating system and platform irrespective of charge. Its a competitor to Microsoft's Active Server Pages (ASP) server-side script engine and similar languages, as per a study PHP is installed on more than 20 million Web sites and 1 million Web servers. Software that uses PHP includes Joomla, Wordpress, MyBB, and Drupal.
PHP was originally created by Rasmus Lerdorf in 1995[6]. The main implementor of PHP, The PHP Group serves as the formal reference to the PHP language. PHP is a free software which was released under the PHP License, which is not compatible with the GNU General Public License (GPL) due to restrictions on the usage of the term *PHP*.

Figure 1Logo of "PHP : Hypertext Preprocessor"

While PHP originally stood for "Personal Home Page", it is now said to stand for "PHP: Hypertext Preprocessor", PHP generally runs on a web server [7]. Any PHP code in a requested file can be executed by the PHP runtime, usually to create dynamic web page content or dynamic images which can be used on Web sites or elsewhere. It can also be used in command-line scripting and client-side graphical user interface (GUI) applications. PHP can be deployed easily on most Web servers available today, many operating systems and platforms, and can be effectively used with many relational database management systems (RDBMS).

Table 1 : PHP Details

Designed By	Rasmus Lerdorf
Developer	The PHP Group
Influenced By	C,Perl,Java,C++
Implementation Language	C
OS	Cross Platform
Usual Filename Extension	.php,php3,.php4,.php5,.phtml
Website	www.php.net

PHP is primarily a filter, which takes input from a file or stream containing text and/or PHP instructions and gives output as another stream of data; most commonly in HTML form. Since PHP 4, the PHP parsercompiles input to produce bytecode for processing by the Zend Engine, because of which it have improved performance over its interpreter predecessor [8-9].
Originally it was designed to create dynamic Web pages, PHP is now used mainly for server-side scripting, and it is very much similar to other server-side scripting languages that provide dynamic content from a Web server to a client, such as Microsoft's ASP.NET, and mod_perl, Sun Microsystems' JavaServer Pages. PHP has also attracted the developers of many frameworks that provide building blocks and a design structure to promote rapid application development (RAD) [10]. Some of these include CakePHP, Symfony, CodeIgniter, Yii Framework, and Zend Framework, offering features similar to other web application frameworks.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.78

RESULTS

The developed NER system in PHP for Hindi can be shown in the below figure. The developed system for HINDI will recognize the Named Entities written in Hindi.

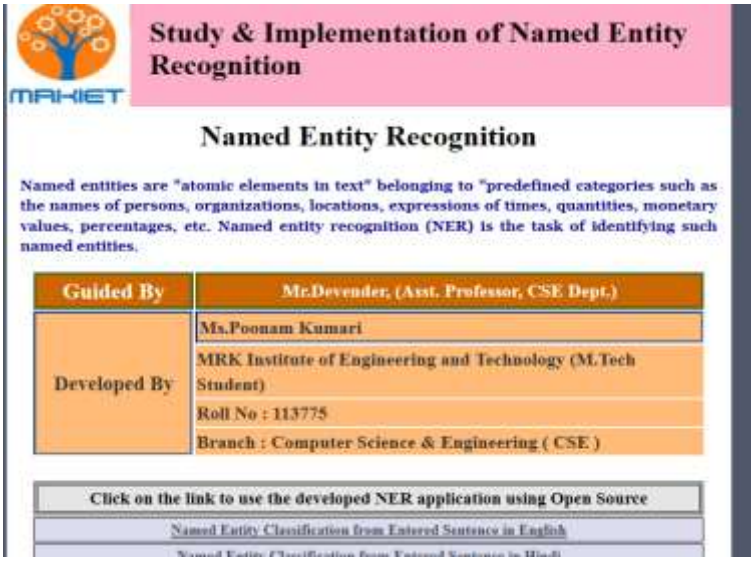

Figure 2 Gateway of NER Application using Open Source

The developed NER System for Hindi can be shown as below.

Figure 3 Developed NER application for Hindi

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.78

The result obtained for NEPH(Named Entity Person-Hindi) is shown as below.

Figure 4 Result of NEPH (Person-Hindi)

The result obtained for NEOH(Named Entity Organization-Hindi) is shown in the below figures.

Figure 5 Result of NEOH (Organization-Hindi)

The result obtained for NELH(Named Entity Location-Hindi) is shown in the below figures.

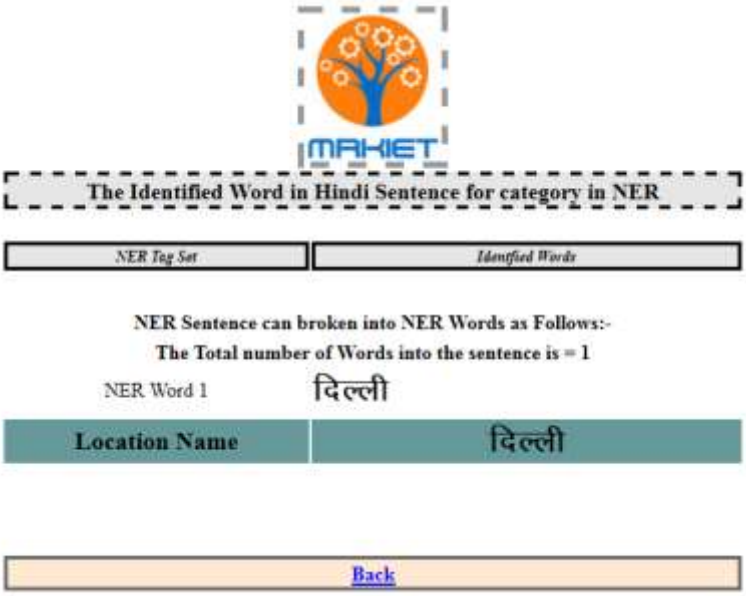

Figure 6 Figure 6.9 Result of NELH (Location-Hindi)

CONCLUSION

In this paper Hindi language and various issues regarding NER are identified and different methodology used to evaluate an NER system for related language in Hindi. This approach is very fast, language independent & easy to retarget, very cost effective. Using this approach the named entities (NE) are identified in the different category and that category is classified with specific color for each class.

REFERENCE

[1]. Hindi Language, http://en.wikipedia.org/wiki/Standard_Hindi, accessed on June 06, 2012.
[2]. Kex, http://www.hgc.ims.u tokyo.ac.jp/service/tooldoc/KeX/intro.html, accessed on June 06, 2012.
[3]. MinorThird, http://minorthird.sourceforge.net/, accessed on June 06, 2012.
[4]. R. Grishman, "The NYU System for MUC-6 or Where's the Syntax", In Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan-Kaufmann Publishers, Columbia, Maryland , November 1995, pages 167-175.
[5]. R. Grishman, Beth Sundheim, "Message Understanding Conference - 6: A Brief History", In the Proceedings of the 16th International Conference on Computational Linguistics (COLING), Morgan Kaufmann publishers, Center for Sprogteknologi, Copenhagen, Denmark , August 5-9, 1996, Pages 466 – 471.
[6]. R. Sirhari, C. Niu, W. Li, "A Hybrid Approach for Named Entity and Sub-Type Tagging", In the Proceedings of the sixth conference on Applied natural language processing, Association for Computational Linguistics Publishers, Seatttle, Washington, USA, April 29-May 4, 2000, pages 247 – 254.
[7]. R. Vijayakrishna and L. Sobha. 2008. "Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields", In the Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Asian Federation of Natural Language Processing Publishers, IIIT, Hyderabad, India, January 12, 2008, pages 59-66.
[8]. Carabao MorphoLogic, http://www.digitalsonata.com/demo.aspx? Component = morphologic, accessed on June 06, 2012.
[9]. Oscar3, http://sourceforge.net/projects/oscar3-chem/, accessed on June 06, 2012.
[10]. Support Vector Machines :-http://en.wikipedia.org/wiki/Support_vector_machine/, accessed on June 06, 2012.