ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 8, February 2014

Named Entity Recognition for Nepali language: A Semi Hybrid Approach

Arindam Dey, Abhijit Paul, Bipul Syam Purkayastha

Abstract: ***Named entity recognition (NER) is a process of specific Information Extraction (IE) from a text document or paragraph. Named Entity has some specific classes like Name of Person, Location, Number, Organization, Currency, Quantifier etc. Our target language is Nepali. This language is one of the Indian language as well as the national language of Nepal. In this paper the main focus is on development of stemming tool, POS (Part of Speech) tagging tool and finally NER detection tool using a semi hybrid approach(Using HMM and some rule based approaches. Accuracy of these tools is also determined.***

Keywords: **Named Entity Recognition (NER), Named Entity (NE), Part Of Speech (POS), Information Extraction (IE), Hidden Markov Model (HMM), Information Retrieval (IR), Machine Translation (MT).**

I. INTRODUCTION

Named entity recognition (NER) is a process of Information Extraction (IE), Information Retrieval (IR) and Machine Translation (MT) from a given text document. It is an Identification or Classification of Named Entities (NEs) in a given text document. NEs in a document are the Names of Person, Location, Number, Organization, Currency, Quantifier etc. These NEs are proper nouns, common nouns and also there combinations.NER tool find these NEs and tag them as **PERSON, LOCATION, NUMBER, ORGANIZATION, CURRENCY, Quantifier**. Depending upon their classes. Before using the NER tool, stemming of document is required for finding the proper root words. Without the root words proper NEs cannot be determined.

Example: Consider a Nepali Sentence with Tagged NER.

छवर्षसम्म"टि. मार्शलहानजुनियर"PERSON

लेभद्ररशान्त"जर्जबुस"PERSON शैलीमानिगमअभिग्रहणगरे।

In the above sentence the NER system determines the NEs and then categorized into different Named Entities Classes. In the above sentence टि.मार्शलहानजुनियर and जर्जबुस refers to Person name.

II. APPROACHES OF NERS

There are three approaches of NERs. They are (i) Rule based approach and (ii) Statistical Approach and (iii) Hybrid Approach. [2][3][4][5] The Rule Based Approach can either be List lookup Approach or a Linguistic Approach. For NER detection using lookup approach or linguistic approaches, a lot of human effort is required. A large Gazetteer list has to be built for different Named Entity classes under lookup approach. Then, search operations are performed to find that the given word in the corpus is under which category of the Named Entity Classes. In a linguistic approach, a linguist set the rules and algorithms to determine NEs in a corpus and also classifies these NEs into respective Named Entity Classes.[1][6][7][8] In Statistical Approach very less amount of human labour is required. It is an automated approach. It is of following types:

A. Hidden Markov Model(HMM)
B. Maximum Entropy Model(MEM)
C. Conditional Random Field(CRF)
D. Support Vector Machine(SVM)
E. Decision Tree(DT)[1][2]

In Hybrid Approach two approaches can be merged together. It improves the performance of NER system. It can be the combination of Linguistic and Statistical models like Gazetteer list and HMM, HMM and CRF or CRF and MEM etc.

III. CURRENT STATUS IN NER FOR INDIAN LANGUAGES (ILS)

Although a lot of work has been done in English and other foreign languages like Spanish, Chinese etc. with high accuracy but regarding research in Indian languages is at initial stage only. Accurate NER systems are now available for European Languages especially for English and for East Asian language. For south and South East Asian languages the problem of NER is still far from being solved. There are many issues which make the nature of the problem different for Indian languages.

For example:- The number of frequently used words (common nouns) which can also be used as names (Proper nouns) is very large for European language where a large proportion of the first names are not used as common words.

IV. CHALLENGES IN NER

Named Entity Recognition was first introduced as part of Message Understanding Conference (MUC-6) in 1995 and a related conference MET-1 in 1996 introduced named entity recognition in non-English text. In spite of the recognized importance of names in applications, most text processing applications such as search systems, spelling checkers, and document management systems, do not treat proper names correctly. This suggests proper

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 8, February 2014

names are difficult to identify and interpret in unstructured text. Generally, names can have innumerable structure in and across languages. Names can overlap with other names and other words. Simple clues like capitalization can be misleading for English and mostly not present in non-western languages like Nepali. The goal of NER is first to recognize the potential named entities and then resolve the ambiguity in the name. There are two types of ambiguities in names, structural ambiguity and semantic ambiguity. Wacholder et al. (1997) describes these ambiguities in detail. Non-English names pose another dimension of problems in NER e.g. the most common first name in the world is Muhammad, which can be transliterated as Mohammed, Muhammad, Mohammad, Mohamed, Mohd and many other variations. These variations make it difficult to find the intended named entity. This transliteration problem can be solved if the name Muhammad is written in Arabic script as .محمد

V. NEPALI LANGUAGE

Nepali or Nepalese (नेपाली), is a language in the Indo-Aryan languages. It is the official language and de facto lingua franca of Nepal and is also spoken in Bhutan. Nepali has official language status in the formerly independent state of Sikkim and in West Bengal's Darjeeling district as well as Assam. Nepali developed in proximity to a number of Indo-Aryan languages, most notably Pahari and Magahi, and shows Sanskrit influences. However, owing to Nepal's geographical area, the language has also been influenced by Tibeto-Burman. Nepali is mainly differentiated from Central Pahari, both in grammar and vocabulary, by Tibeto-Burman idioms owing to close contact with the respective language group. Nepali language shares 40% lexical similarity with Bengali language. Historically, the language was first called the Khas language (Khaskurā), and then Gorkhali or Gurkhali (language of the Gorkha Kingdom) before the term Nepali (Nepālībhāṣā) was taken from Nepal Bhasa. Other names include Parbatiya ("mountain language", identified with the Parbatiya people of Nepal) and Lhotshammikha (the "southern language" of the Lhotshampa people of Bhutan). According to the 2011 national census, 44.6 per cent of the population of Nepal speak Nepali as a native language. The Ethnologic website counts more than 17 million (2007) and 42 million (2012) speakers worldwide, 17 million within Nepal (from the 2001 census).

VI. RELATED WORKS

Although over the years there has been considerable work done for NER in English and other European languages, the interest in the South Asian languages has been quite low until recently. One of the major reasons for the lack of research is the lack of enabling technologies like, parts of speech taggers, gazetteers, and most importantly, corpora and annotated training and test sets. One of the first NER study of South Asian languages and specifically on Urdu was done by Becker and Riaz (2002) who studied the challenges of NER in Urdu text without any available resources at the time. The by-product of that study was the creation of Becker-Riaz Urdu Corpus (2002). Another notable example of NER in South Asian language is DARPA's TIDES surprise language challenge where a new language is announced by the agency to build language processing tools in a short period of time. In 2003 the language chosen was Hindi. Li and McCallum (2003) tried conditional random fields on Hindi data and reported f-measure ranging from 56 to 71 with different boosting methods. Mukund et al. (2009) used CRF forUrdu NER and showed f-measure of 68.9%. By far the most comprehensive attempt made to study NER for South Asian and South East Asian languages was by the NER workshop of International Joint Conference of Natural Language Processing in 2008. The workshop attempted to do Named Entity Recognition in Hindi, Bengali, Telugu, Oriya, and Urdu. Among all these languages Urdu is the only one that has Arabic script. Test and training data was provided for each language by different organizations therefore the quantity of the annotated data varied among different languages. Hindi and Bengali led the way with the most amounts of data; Urdu and Oriya were at the bottom with the least amount of data. Urdu had about 36,000 thousand tokens available. A shared task was defined to find named entities in the languages chosen by the researcher. There are 15 papers in the final proceedings of NER workshop at IJCNLP 2008, all cited in the references section, a significant number of those papers tried to address all languages in general, but resorted to Hindi, where the most number of resources were available. Some papers only addressed specific languages like Hindi, Bengali, Telugu and one paper addressed Tamil. There was not a single paper that focused on only Urdu named entity recognition. The papers that tried to address all languages, the computational model showed the lowest performance on Urdu. Among the experiments performed at Named Entity Workshop on various Indic languages and Urdu, almost all experiments used CFR with limited success.

VII. HIDDEN MARKOV MODEL

A Hidden Markov Model (HMM) is a statistical Markov Model in which the system being modelled is assumed to be a Markov process with unobserved (*hidden*) states. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a Hidden Markov model the state is not directly visible, but the output dependent on the state is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 8, February 2014

sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; even if the model Parameters are known exactly, the model is still 'hidden'. [2].

A. N-gram technique

An N-gram is an ordered collection of N elements of the same kind, usually presented in a large collection of many other similar N-grams. The individual elements are commonly natural language words, though N-grams have been applied to many other data types, such as numbers, letters, genetic proteins in DNA, etc. Statistical N-gram analysis is commonly performed as part of natural language processing, bioinformatics, and information theory.

N-grams may be derived for any positive integer N. 1-grams are called "unigrams," 2-grams are called "bigrams," 3-grams are called "trigrams," and higher order N-grams are simply called by number, e.g. "4-grams". N-gram techniques may be applied to any kind of ordered data. Metadata such as end-of-sentence markers may or may not be included.

In the proposed system the collection of up to 5 elements can be done by using 5-gram technique.E.g.श्री टि. मार्शल हान जुनियर is tagged as person. Here five different entities are combined together by using 5-gram technique.

VIII. PROPOSED SYSTEM DESIGN.

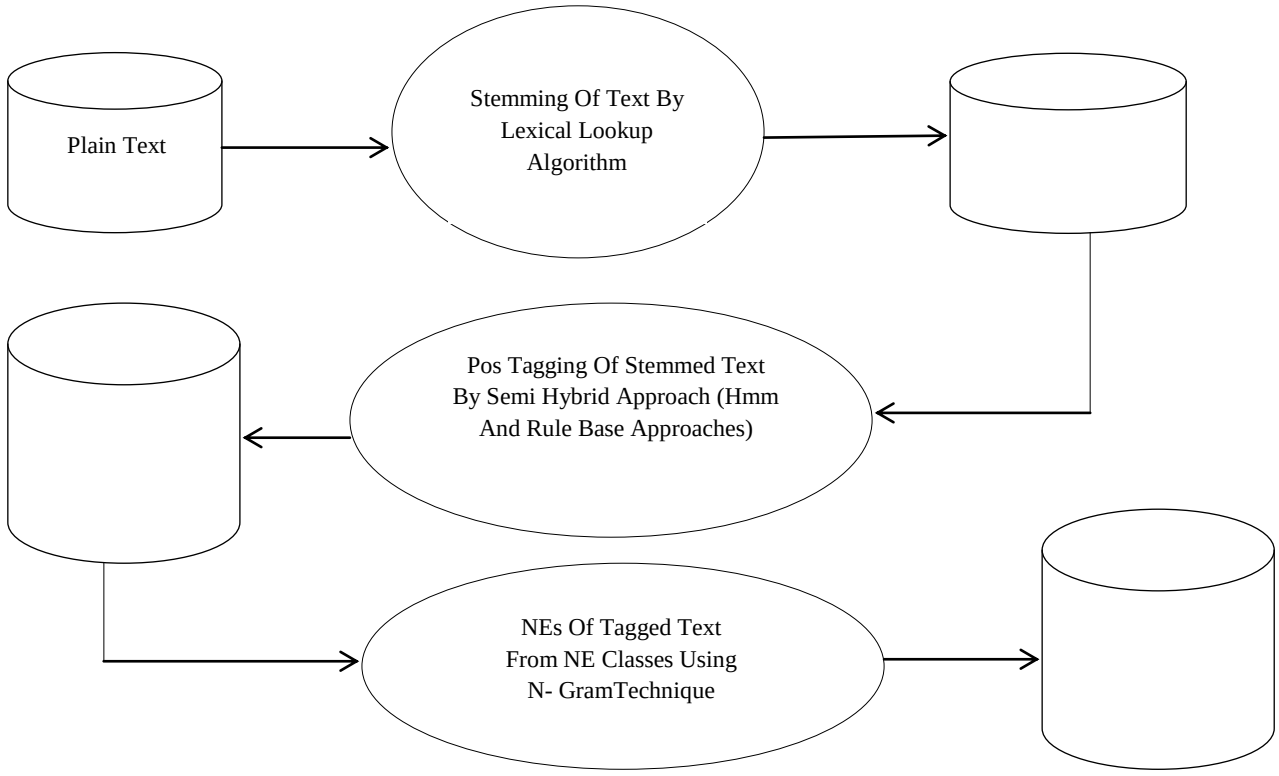

Fig: 1- Architecture of NER tools

B. How it woks

Example: Before stemming.

छवर्षसम्मटि.

मार्शलहानजुनियरलेभद्ररशान्तजर्जबुसशैलीमानिगमअभिग्रहणगरे।

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 8, February 2014

In the above Nepali sentence two underlined words are not root word. So they are to be stemmed.

Example: After Stemming

छवर्षसम्मटि.

मार्शलहानजुनियरलेभद्ररशान्तजर्जबुसशैलीमानिगमअभिग्रहणगरे।

In the above sentence the underlined words are stemmed and the root words are determined.

After stemming part of speech tagging is required for only proper nouns and common nouns in the document. Other text remains untagged.

Example: POS tagging of document.

छवर्षसम्मटि.PC मार्शलPC हानPC जुनियरPCलेभद्ररशान्तजर्जPC बुसPCशैलीमानिगमअभिग्रहणगरे।

In the above document PC (Proper Common)is tagged to all the NEs in the document. This will determine the presence of proper noun and common noun in the sentence.

After the POS tagging the NER can be easily determined.

Example: NER detection.

छवर्षसम्म"टि. मार्शलहानजुनियर"PERSON लेभद्ररशान्त"जर्जबुस"PERSON शैलीमानिगमअभिग्रहणगरे।

XI. RESULT ANALYSIS

A. Table: 1-Results of NER in NEPALI Language. [14]

Total No Of Sentence	750		
	Total Tags	Total no of correctly observer text	Percentage of Accuracy
Person	101	86	85.15%
Location	89	81	91.01%
Number	45	39	86.67%
Organization	77	74	96.10%
Currency	6	6	100%
Quantifier	26	26	100%
Unknown Words	0	32	

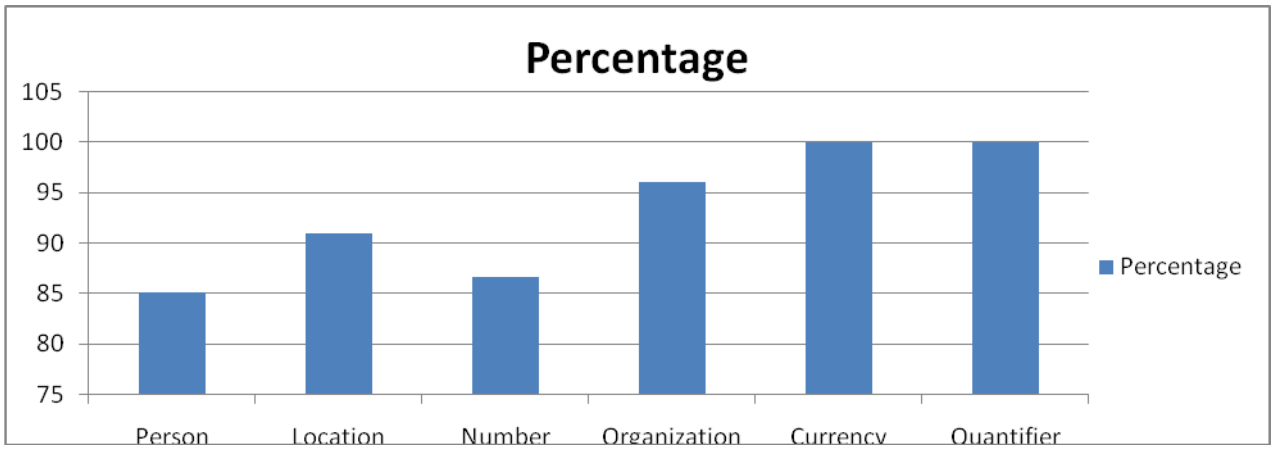

Fig 1: Result showing accuracy of the individual tags.[14]

X. CONCLUSION

Area of focus is on a new SEMI-HYBRID approach where HIDDEN MARKOV MODEL will be combined with LOOK-UP algorithm and RULE based approach. In this proposed approach Look-Up algorithm and few rules are used for some ambiguous words and rest are done in Hidden Markov Model. Since most of the work is done in HMM approach and very few amount of work

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 8, February 2014

is done in rule based approaches so it is named as Semi Hybrid approach. The proposed system can be upgraded by solving word sense disambiguation problem. This is the future work for the system which will increase the accuracy up to 100%.

REFERENCES

[1] Kamaldeep Kaur, Vishal Gupta." Name Entity Recognition for Punjabi Language" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 .Vol. 2, No.3, June 2012.

[2] Arindam Dey, Bipul Syam Purkayastha "Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach" (IJCA) International Journal of Computer Applications, Vol. 84, 2013.

[3] B. Sasidhar, P. M. Yohan, Dr. A. Vinaya Babu3, Dr. A. overhand,."A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[4] AnimeshNayan,, B. Ravi KiranRao, PawandeepSingh,SudipSanyal and RatnaSanya "Named Entity Recognition for Indian Languages" .In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages ,Hyderabad (India) pp.97–104, 2008. Available at: http://www.aclweb.org/anthology-new/I/I08/I08-5014.pdf.

[5] Sujan Kumar Saha Sanjay ChatterjiSandipanDandapat. "A Hybrid Approach for Named Entity Recognition in Indian Languages".

[6] AsifEkbal, RejwanulHaque, Amitava Das, VenkateswarluPoka and SivajiBandyopadhyay "Language Independent Named Entity Recognition in Indian Languages" .In Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pages 33–40,Hyderabad, India, January 2008.Available at: http://www.mt-archive.info/IJCNLP-2008-Ekbal.pdf.

[7] Vishal Gupta, Gurpreet Singh Lehal "Named Entity Recognition for Punjabi Language Text Summarization" International Journal of Computer Applications (0975 – 8887) Vpl.33 No.3, Nov. 2011.

[8] S. Biswas, M. K. Mishra, Sitanath_biswas, S. Acharya, S. Mohanty "A Two Stage Language Independent Named Entity Recognition for Indian Languages" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (4) , 2010, 285-289.

[9] Darvinderkaur, Vishal Gupta. "A survey of Named Entity Recognition in English and other Indian Languages" .IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010.

[10] KhRajuSingha, BipulSyamPurkayastha and KhDhirenSingha,"Part of Speech Tagging in Manipuri with Hidden Markov Model"IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012.

[11] KhRajuSingha,Ksh Krishna BatiSingha ,BipulSyamPurkayastha,"Developing a Part of Speech Tagger for Manipuri" International Journal of Computational Linguistics and Natural Language Processing Vol 2 Issue 9 September 2013.

[12] TejBahadurShahi, Tank NathDhamala, BikashBalami,"Support Vector Machines based Part of Speech Tagging for Nepali Text" International Journal of Computer Applications, Volume 70– No.24, May 2013.

[13] KhRajuSingha, BipulSyamPurkayastha, KhDhirenSingha, "Part of Speech Tagging in Manipuri: A Rule based Approach"IJCA Journal Volume 51 - Number 14, Year of Publication: 2012.

[14] Deepti Chopra and Sudha Morwal,"Named Entity Recognition in Punjabi Using Hidden Markov Model " International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 3 No. 12 Dec 2012

AUTHOR'S PROFILE

ArindamDey received MCA degree from Skkim Manipal University in 2011.Currently he is pursuing his PhD degree in Computer Science Department from Assam University, Silchar. His research interests include Artificial Intelligence, Natural Language Processing, and Information Retrieval.

Abhijit Paul received MCA degree from Assam Engineering College, Guwahati in 2009.Currently he is pursuing his MPhil degree in Computer Science Department from Assam University, Silchar. His research interests include Artificial Intelligence, Natural Language Processing, Soft Computing.

Bipul Syam Purkayastha received PhD degree in Mathematics from North Eastern Hill University, Shillong in 1997.Currently he is a Professor in Computer Science Department in Assam University, Silchar. His research interests include Artificial Intelligence, Natural Language Processing.