Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach

Arindam Dey
Research Scholar
Department of Computer Science
Assam University Silchar, India

Bipul Syam Prukayastha
Professor
Department of Computer Science
Assam University Silchar, India

ABSTRACT
Named Entity Recognition (NER) is a task to discover the Named Entities (NEs) in a document and then categorize these NEs into diverse Named Entity classes such as Name of Person, Location, River, Organization etc. Area of concentration is the performance of NER in the Indian languages (IL).Nepali is the target language. In this paper different technique of NER and a brief introduction of Gazetteer method and Hidden Markov Model especially n-gram technique has been described. Different types of problem faced in handling Nepali Grammar are also described.

Keywords
Named Entities (NEs), Named Entity Recognition (NER), Indian Languages (ILs), and Hidden Markov Model (HMM).

1. INTRODUCTION
Named Entity Recognition (NER) is an important tool in almost all of the Natural Language processing applications such as Information Retrieval (IR), Information Extraction (IE), Question Answering (QA), Machine Translation (MT) and Automatic Summarization (AS) etc. NER can be defined as a two stage problem:- Identification of Proper Noun and classification of the Proper Noun into a set of classes such as Person names, Location names(cities, countries etc), Organization names(Companies, government Organizations, committees etc), Miscellaneous(date, time, percentage, monetary expression, number expression and measurement expression). These words are collectively defined as "ENAMEX" by MUC-6[15]. Thus NER can be said as the process of identifying and classifying the tokens into the above predefined classes. A hybrid approach of Gazetteer method and N- Gram technique of Hidden Markov Model to capture NER from a given text are also introduced.

2. DIFFERENT APPROACHES OF NER

2.1 Rule Based Approach
Rule based approach mainly concerned with manual rules written by the linguistics. Many rules based NER contains:

a) Lexicalized Grammar

b) Gazetteer list

c) List of triggered words

2.2 Machine Learning Based- approach
Machine learning based approach concerned with some pre-defined methods. These are:

a) Hidden Markov Models(HMM)

b) Decision Trees

c) Maximum Entropy Models(ME)

d) Support Vector Machines(SVM)

e) Conditional Random Fields(CRF)

2.3 Hybrid Approach
It is an approach where more than two approaches are used in order to improve the performance of the NER system. So the hybrid approach may be a combination of HMM model and CRF model or CRF and ME approach or Gazetteer method with HMM approach etc.

3. PROBLEM FACED IN INDIAN LANGUAGE (ILS)
Since for English Language lots of NER system has been built. But such NER system for Indian Language cannot be used because of the following reason:

a) There is no capitalization on leading character words in Indian language.

b) Indian names are ambiguous and this issue makes the recognition a very difficult task.

c) Indian languages are relatively free-order languages.

d) Indian language is inflectional and morphologically rich.

e) Non- availability of large Gazetteer [11].

4. INFLECTIONAL LANGUAGE
In grammar, inflection or inflexion is the modification or marking of a word (or more precisely lexeme) to reflect grammatical (that is, relational) information, such as gender, tense, number or person. Languages that have some degree of inflection are synthetic languages. There can be highly inflected language, such as Latin, or weakly inflected language, such as English. Languages that are so inflected that a sentence can consist of a single highly inflected word (such as many American Indian languages) are called polysynthetic languages. Nepali is an Inflectional Language.

5. NEPALI NER

In English and some other languages, capitalization features play an important role as NEs are generally capitalized for these languages. Unfortunately this feature is not applicable for Nepali. Also Indian person names are more diverse. Lots of common words having other meanings are also used as person names. These make difficult to develop a NER system on Nepali. Li and McCallum (2004) used the entire word text, character n-grams (n = 2, 3, 4), word prefix and suffix of lengths 2, 3 and 4, and 24 Hindi gazetteer lists as atomic features in their Hindi NER. Kumar and Bhattacharyya (2006) used word features (suffixes, digits, special characters), context features, dictionary features, NE list features etc. in their MEMM based Hindi NER system. No work on Nepali has done on till date.

6. GAZETTEER METHOD

Gazetteer Method is the creation of different gazetteer classes (list) for different Named Entities and then applies search operations to classify the names [11]. This method needs two types of input rather collection of gazetteer, one for each named entity classes of interest and second for other class that give example of entities that are not been extracted. To build such type of gazetteer classes a very large corpus is needed. But it fails to resolve ambiguities in a given document. For example if in a document a name Ganga exists then according to Nepali Language Ganga may be in the list of person name and in the list of river name. So there is an ambiguity. And it is difficult task for gazetteer method to correctly identify or tag the name Ganga.

6.1 Advantages of Gazetteer Method

a) The Gazetteer method gives very fast result of NER.

b) The accuracy of Gazetteer method depends on completeness of the Gazetteer used.

c) Creating the gazetteer manually is effort-intensive, error-prone and subjective.

d) But the problem is how to automatically create a gazetteer with less effort, in less time and with high accuracy using a given document.

6.2 Disadvantages of Gazetteer Method

a) Ambiguity resolution is difficult.

b) Since the words are created repeatedly. So keeping a gazetteer list for these words up-to-date is challenging.

c) Without ambiguity resolution the precision is low.

7. HIDDEN MARKOV MODEL

A Hidden Markov Model (HMM) is a statistical Markov Model in which the system being modeled 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 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'.[3]

Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics.

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation are related through a Markov process rather than independent of each other.[10]

7.1 N-Gram Technique under HMM

N-grams have been widely investigated for a number of text processing and retrieval applications. An n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus [2].

An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram", size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n, e.g., "four-gram", "five-gram", and so on.

An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of an $(n-1)$ order Markov model. N-gram models are now widely used in probability, communication theory, computational linguistics (for instance, statistical natural language processing), computational biology (for instance, biological sequence analysis), and data compression. The two core advantages of n-gram models (and algorithms that use them) are relative simplicity and the ability to scale up by simply increasing n. A model can be used to store more contexts with a well-understood space time tradeoff, enabling small experiments to scale up very efficiently [17].

This paper concerned with word sequences and we referred up to five-gram (here n = 5 words), since few name of persons could be up to five words (mehboob hasan ben seraj mazarbhuiya) and also few organizations are there whose name consists of five to six words.

8. RELATED WORKS

Saha et.al(2008) [5] describes the development of Hindi NER using ME approach. The training data consists of about 234k words, collected from the newspaper “Dainik Jagaran” and is manually tagged with 17 classes including one class for not name and consists of 16,482 NEs. The paper also reports the development of a module for semi-automatic learning of context pattern. The system was

evaluated using a blind test corpus of 25K words having 4 classes and achieved an F-measure of 81.52%.

Goyal (2008) [6] focuses on building a NER for Hindi using CRF. This method was evaluated on test set 1 and test set 2 and attains a maximum F1-measure around 49.2% and nested F1-measure around 50.1% for test set 1 maximum F1-measure around 44.97% and nested F1-measure around 43.70% for test set 2 and F-measure of 58.85% on development set.

Saha et.al(2008) [7] has identified suitable features for Hindi NER task that are used to develop an ME based Hindi NER system. Two-phase transliteration methodology was used to make the English lists useful in the Hindi NER task. The system showed a considerable performance after using the transliteration based gazetteer lists. This transliteration approach is also applied to Bengali besides Hindi NER task and is seen to be effective. The highest F-measure achieved by ME based system is 75.89% which is then increased 81.2% by using the transliteration based gazetteer list.

Li and McCallum (2004) [1] describes the application of CRF with feature induction to a Hindi NER. They discovered relevant features by providing a large array of lexical test and using feature induction to construct the features that increases the conditional likelihood. Combination of Gaussian prior and early-stopping based on the results of 10-fold cross validation is used to reduce over fitting.

Gupta and Arora (2009) [3] describes the observation made from the experiment conducted on CRF model for developing Hindi NER. It shows some features which makes the development of NER system complex. It also describes the different approaches for NER. The data used for the training of the model was taken from Tourism domain and it is manually tagged in IOB format.

David Nadeau *et al.* [12]. proposed a named-entity recognition (NER) system that addresses two major limitations frequently discussed in the field. First, the system requires no human intervention such as manually labeling training data or creating Gazetteers. Second, the system can handle more than the three classical named-entity types (person, location, and organization). They propose a named-entity recognition system that combines named entity extraction with a simple form of named-entity disambiguation. They use some simple yet highly effective heuristics, to perform named-entity disambiguation.

Deepti Chopra *et al*. [14]. have discussed about NER, Challenges in NER in the Indian languages, Performance Metrics and finally the methodology and the results. They have obtained F-Measure and accuracy of about 88.4% by performing NER in Punjabi using Hidden Markov Model (HMM).

9. EXISTING WORK ON DIFFERENT INDIAN LANGUAGES

9.1 Table 1: Different Approaches and Their Accuracy

Author	Language	Approach	Words	Accuracy
[17]	Telugu	CRF	13,425	91.95%
[17]	Telugu	ME	-	50.00% aprx
[8]	Tamil	CRF	94,000	80.44%
[5]	Hindi	ME	25,000	81.52%
[6]	Hindi	CRF	-	60%
[1]	Hindi	CRF	-	-
[19]	Bengali	CRF	150,000	90.7%
[4]	Hindi	SVM	502,974	77.17%
[9]	Bengali	SVM	122,467	84.00% aprx
[9]	Hindi	ME	-	75.89%
[9]	Bengali	SVM	150,000	90.00% aprx

10. PROPOSED SYSTEM DESIGN

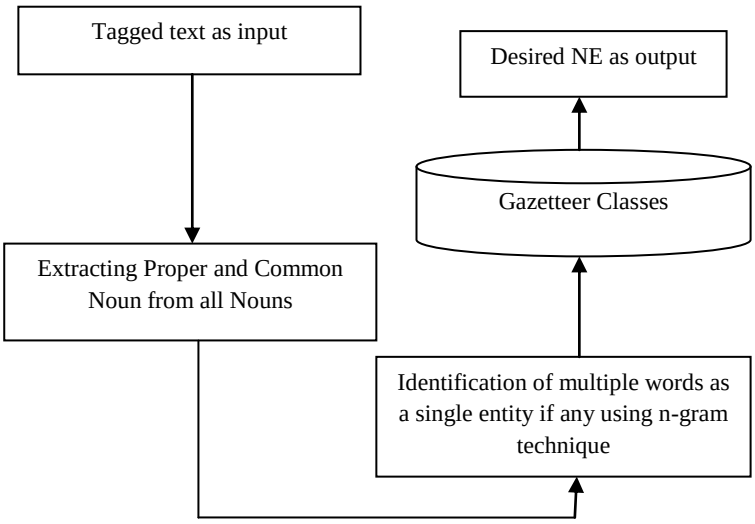
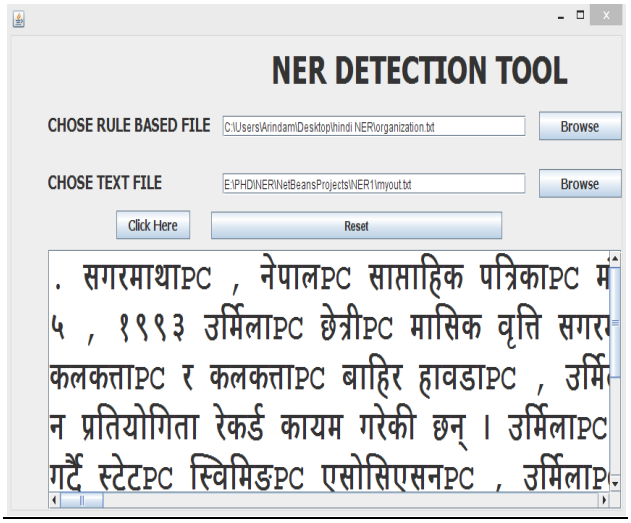


Fig: 1- Architecture of NER tool using Gazetteer method

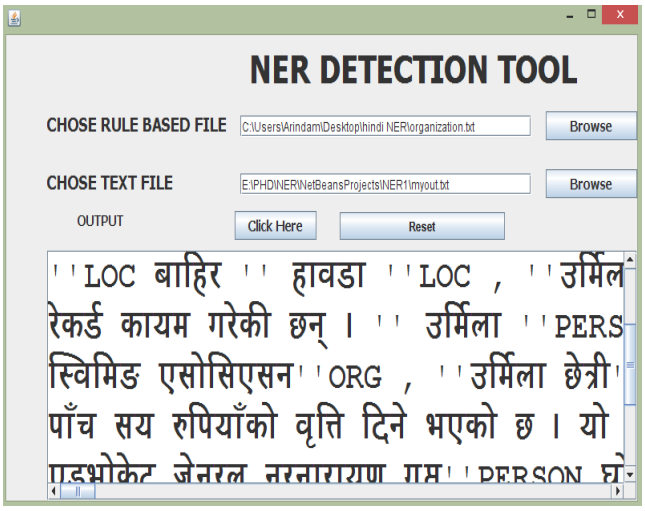
10.1 How it works

Loading the text in the machine



In the above tool the nouns which act as both proper and common noun have tagged as PC.

Named entity recognized like person (PERSON) and Location (LOC).



After detecting the "PC" tags the tagged document is again tagged with Gazetteer Classes and detects person names by "PERSON", organization names as "ORG", location as "LOC" and rest are remain untagged.

10.2 Design briefing

a) Tagged text is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children in the identification of words as nouns, verbs, adjectives, adverbs, etc. This text is collected.

b) Only Nouns (especially Proper and Common noun) are identified from the tag set.

c) Identification of more than one proper noun or common noun or combinations are placed one after another using n gram technique if any.

d) Comparing the resultant proper noun and Common noun with the Gazetteer classes to find the desired named entity.

11. RESULT ANALYSIS

Analysis of n-gram technique (here n=5) with gazetteer method on newspaper corpus which has about 1000 sentences has been done. The searching is done by HASH MAPPING technique which saves time.

Table 2: Total number of tags in the Corpus:

	Person(PERSON)	Organization(ORG)	Location(LOC)
Total	169	59	31

In the above table we have collected unique NER tagged entities from 1000 sentences.

11.1 Table 3: Accuracy is 79.54% from 1000 sentences using n-gram and gazetteer method.

	Person Tag		Organization Tag		Location Tag	
	Total Tag	Correct	Total Tag	Correct	Total Tag	Correct
	169	130	59	52	31	24
Accuracy	76.92%		88.14%		77.42%	

In the above table we have collected individual accuracy details of all the three tags of NER

12. CONCLUSION AND FURTHER WORK

In this paper, different approaches of Named entity recognition and their problems has been studied. Collection huge gazetteer classes are the main area of focus.

Above results shows more accurate and perfect recognition of Named Entities. Depending upon the collection of data this system will trace any named entities of almost all inflectional languages. Tracking of multiple proper and common noun words in a sequence up to five words in a row (using five gram technique) has been done succefully.

In the future, we will integrate named entity recognition into real applications, such as information retrieval, automatic summarization, and topic detection and tracking, so that we can further study and evaluate its influences to these systems. We are trying to track up to seven words in multiple proper and common noun words in a sequence

expression by implementing n gram technique. We will also try to remove ambiguity of WSD.

13. REFERENCES

[1] W. Li and A. McCallum, Sept 2003 "Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction(Short Paper)," ACM Transactions on Computational Logic.

[2] Suleiman H. Mustafa and Qasem A. Al-Radaideh 2004 "Using N-Grams for Arabic Text Searching" journal of the american society for information science and technology.

[3] Zubek, R. 2006. Introduction to Hidden Markov Models. In Rabin, S. (ed.), AI Game Programming Wisdom 3. Charles River Media, Hingham, MA.

[4] A. Ekbal and S. Bandyopadhyay 2008, "Named Entity Recognition using Support Vector Machine: A Language Independent Approach," International Journal of Computer, Systems Sciences and Engg (IJCSSE.

[5] S. K. Saha, S. Sarkar, and P. Mitra January 2008, "A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition," in Proceedings of the 3rd International Joint Conference on NLP, Hyderabad , India.

[6] A. Goyal , "Named Entity Recognition for South Asian Languages Jan 2008," in Proceedings of the IJCNLP-08 Workshop on NER for South and South-East Asian Languages, Hyderabad, India.

[7] S. K. Saha, P. S. Ghosh, S. Sarkar, and P. Mitra 2008, "Named Entity Recognition in Hindi using Maximum Entropy and Transliteration," Research journal on Computer Science and Computer Engineering with Applications.

[8] Asif Ekbal, Rajewanul Hague, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay 2008 "Language Independent Named Entity Recognition in Indian Languages" Proceedings of the IJNLP-08 Workshop on NER for South and South East Asian Languages, Hyderabad, India.

[9] M. Hasanuzzaman, A. Ekbal, and S. Bandyopadhyay, May 2009, "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi, "International Journal of Recent Trends in Engineering, vol. 1.

[10] Anastasia Rita Widiarti, and Phalita Nari Wastu 2009, "Javanese Character Recognition Using Hidden Markov Model"World Academy of Science, Engineering and Technology 33.

[11] Padmaja Sharma, Utpal Sharma, Jugal Kalita May 2011, "Named Entity Recognition: A Survey for the Indian Languages".

[12] David Nadeau, Peter D. Turney and Stan Matwin March 11 , 2011, "Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity" National Research Council Canada.

[13] Nusrat Jahan, Sudha Morwal and dipti Chopra 12 Dec 2012,"Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach".

[14] Deepti Chopra, Sudha Morwal Dec 12, 2012, "Named Entity Recognition in Punjabi Using Hidden Markov Model", "International Journal of Computer Science & Engineering Technology (IJCSET)".

[15] David Nadeau, Satoshi Sekine , "A survey of named entity recognition and classification" National Research Council Canada / New York University.

[16] Sujan Kumar Saha, Sudeshna Sarkar, Pabitra Mitra "Gazetteer Preparation for Named Entity Recognition in Indian Languages".Available at: http://www.aclweb.org/anthology-new/I/I08/I08-7002.pdf

[17] M. N. Karthik, Moshe Davis "Search Using N-gram Technique Based Statistical Analysis for Knowledge Extraction in Case Based Reasoning Systems" .

[18] B. Sasidhar#1, P. M. Yohan*2, Dr. A. Vinaya Babu3, Dr. A. Govardhan4," A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu", http://www.ijcsi.org/papers/IJCSI-8-2-438-443.pdf

[19] A. Ekbal, R. Hague, and S. Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field," in Proceedings of ICON, India, pp. 123–128.