NAMED ENTITY RECOGNITION FROM BENGALI NEWSPAPER DATA

Shamima Parvez

Lecturer, Department of Computer Science and Engineering, Premier University, Chittagong, Bangladesh

ABSTRACT

Due to the dramatic growth of internet use, the amount of unstructured Bengali text data has increased enormous. It is therefore essential to extract event intelligently from it. The progress in technologies in natural language processing (NLP) for information extraction that is used to locate and classify content in news data according to predefined categories such as person name, place name, organization name, date, time etc. The current named entity recognition (NER), which is a subtask of NLP, plays a vital rule to achieve human level performance on specific documents such as newspapers to effectively identify entities. The purpose of this research is to introduce NER system in Bengali news data to identify events of specified things in running text based on regular expression and Bengali grammar. In so doing, I have designed and evaluated part-of-speech (POS) tags to recognize proper nouns. In this thesis, I have explained Hidden Markov Model (HMM) based approach for developing NER system from Bengali news data.

KEYWORDS

Hidden Markov Model, Named Entity Recognition, Artificial Intelligence, part-of-speech

1. INTRODUCTION

The advances in the field of Artificial Intelligence (AI) over the last 60 years are mentionable; the ultimate goal of making machines to understand human language is still an insufficient. People speak different languages around the world. Natural language processing (NLP) is the branch of artificial intelligence that deals with the interaction between computers and natural language like human language. NLP is concerned with human-computer interaction [1][2]. It enables the capability of a computer program to understand human speech as it is spoken. The challenges of NLP are teaching computers to understand the way of human learn and use language.

Named entity recognition (NER) is the process of identifying and classifying phrases in text. The purpose of NER is to classify every term in a document into predefined categories such as person name, place name, organization name, miscellaneous name like date, time, percentage and monetary expressions etc. [3]. NER has too many applications in NLP including machine translation, automatic question answering, information retrieval etc. [4]. The current named entity recognition (NER), which is a subtask of NLP, plays a vital rule to achieve human level performance on specific documents such as newspapers to effectively identify entities. The purpose of this research is to introduce NER system in Bengali news data to identify events of specified things in running text based on regular expression and Bengali grammar. In so doing, I have designed and evaluated part-of-speech (POS) tags to recognize proper nouns. In this thesis, I have explained Hidden Markov Model (HMM) based approach for developing NER system from Bengali news data.

DOI: 10.5121/ijnlc.2017.6304

In section 2, I delineate the existing NER system and their application and in section 3 I have presented the proposed NER system for Bengali news text data. In section 4, simulation setting, POS tagging used for the developed system, implementation results and performance evaluation of the research and finally, the features of the proposed NER system are described. Section 5 draws the conclusion and future work of the research.

2. LITERATURE REVIEW

One of the primary topics for Artificial Intelligence (AI) in NLP as indictment of determining and extracting named entity. NER designates the task of correctly identifying words or phrases in textual documents that express named entities such as persons, organizations, locations, etc. During the last decades, NER has been widely studied and the best NER approaches now-a-days produce near-human level recognition accuracy for generic domain such as news article [5]. Most of the research works pertaining to NER systems have been on the issues related to the NER system of a specific language. On the other hand, NER system for Bengali news data in Bangladesh aspect is new arena of NER system. Thus, providing an efficient NER system for Bengali language becomes a significant mission. Although NER system has been continued to study for about forty years, NER systems for Bengali news text data are still at the infant stage for research. In this section, the discussion of major NER systems proposed to date are explained and highlight the advantages and performance issues of each system, the current issues and research areas of NER systems. In many language, NER system use statistical Conditional Random Fields (CRFs). That enables different contextual information of the words along with the variety of features that are very helpful in predicting the different types of named entity classes [6]. But in my system I use HMM based NER system for Bengali news data.

2.1. Existing NER System

In the recent few years, a good number of NER systems have been developed in order to provide efficient identification of entities. It has not been possible to develop this type of system for all the languages. I have introduced some known NER systems. These NER systems are based on different types of languages. These systems are discussed in these subsections.

2.1.1. English Based NER System

At first, English is the most popular language factor to research NER system. At the core of any NLP task there is the important issue of understanding natural language. The process of building computer programs that understand natural language includes three major problems: the first one associated with the thought process, the second one with the representation and meaning of the linguistic input, and the third one to the world knowledge. Thus, NLP system may start at the word level – to determine the morphological structure, nature such as part-of-speech, meaning etc. of the word and they may move on to the sentence level in order to determine the word order, grammar, meaning of the entire sentence, etc. and then to the context and the overall environment or domain [7]. English is the most popular language to research NER. Many researches are done to reads text in English and assigns part-of-speech to each word or other token, such as noun, verb, adjective, etc.

2.1.2. Hindi Based NER System

A language independent NER system for Hindi language by using contextual and morphological evidences for five languages such as English, Greek, Romanian, Turkish, and Hindi have developed by Ccerzan and Yarowsky (1999) [8]. The performance of Hindi NER system is very poor and has F-measure of 41.70% with very poor 27.84% recall and nearly 85% precision.

Various systems have been designed to change the training corpus in a file that can make the system portable to a new Indian language [3]. Although the resource for Indian language is poor and gazetteers are not sufficient [9].

2.1.3. Urdu Based NER System

NER can detect entities in a text and can classify them into predefined categories. The NER systems for Urdu language also faced several problems due to its rich morphology and like Bengali language; Urdu language is at infancy stage. A corpus of 2200 Urdu documents has been developed [8]. The sub module of NER for information extraction by using two models: Maximum Entropy (ME) and Conditional Random Field (CRF) based NER for Urdu has been developed where ME have 55.3%, F-measures and CRF based module for NER F-measure value of 68.9%.

2.1.4. Garman Based NER System

Tagging a German corpus manually and semi-automatically with semantic roles in order to drive a large domain independent lexical semantic resource is included in the SALSA project [10]. A 1.5-million-word corpus of newspaper text with manually annotated syntactic structure and the semantic annotation is performed using semantic roles. Another problem is raised because of the development of large role-annotated corpora. It is necessary to represent these corpora, a standard multi-level annotation format, which integrates semantic role annotation with other linguistic annotation levels.

2.1.5. Other NER Systems

Japanese texts are written without blank spaces. The morphological analysis of Japanese NER has close bond, and due to this fact it is necessary to perform activities on this language. The NER system of this language relies merely on local context. The Japanese NER system proposed in, which achieved the highest F-measure among conventional systems introduced the bunsetsu (a commonly used linguistic unit in Japan) feature in order to consider wider context, considers only adjacent bunsetsus [11]. Support vector machine (SVM) base NER system was introduced for Japanese language. To achieve human-level accuracy a recently in Chinese language, a topic sentence extraction is to employ machine learning methods. For example, trainable classifiers have been used in to select sentences which are based on features such as cue phrase, location, sentence length, word frequency and title, etc [12][13][14][15]. The Japanese FrameNet(JFN) research project includes researchers to corpus contains currently one million sentences, taken from Kyoto University Annotation Text Corpus 10].

2.2. Current issues and Research Area of NER System

While a NLP has gained a lot of momentum in the past several decades, much of this research effort has been focusing on only a handful of politically-important language such as English, Chinese, and Arabic [16]. However, Bengali NER is complicated due to the scarcity of annotated data and lack of an accurate part-of-speech (POS) tagger, but the demonstration of a baseline tagger and the Wikipedia-related features significantly improve a baseline POS named entity recognizer is significantly improved. The essential part of NER systems is the ability to recognize previously unknown entities. Such ability hinges upon recognition and classification rules that are triggered by distinctive modelling features associated with positive and negative examples.

3. THE PROPOSED NER SYSTEM FOR BENGALI NEWS

In this section, the NER system for Bengali news has been presented in detail. HMM has been applied in NER system. HMM is a generative model to assign the joint probability to paired

observation and level sequence [17]. Then, the parameters are trained to maximize the joint likelihood of training sets. The evaluation of HMM is higher compared to other models because the HMM has better ability of capturing the locality of phenomena, which indicates names in text. In this research, HMM is used as HMM based NER has the following benefits that are useful for identifying entities from Bengali language:

- HMM is independent language so it can be used for Bengali language domain.
- The HMM based NER is easy to understand, implement and analyze. In order to make the system scalable, it can be used for any amount of data.
- Sequence of leveling can be solved effectively by it.
- The states are not fixed so one can use it according to their requirements and domain of interest. In other words, it is very dynamic in nature.
- No language expertise is essential for NER system that depends on HMM.

HMM is determined as three groups of parameters: the state transition probability $A=\{a_{mn}\}$, where $a_{mn}=P(q_{x+1}=j|q_x=i)$; the observation word probability $\mathbf{B}=\{b_j(k)\}$, where $b_j(k)=P(o_x=v_k|q_x=j)$; and the initial state distribution $\pi=\{\pi_i\}$, where $\pi_i=P(q_1=i)$. Here q_x is the state time at x, v_k is the distinct observation symbols in observation space and o_x is the observation vector in time x. To make it simpler I used $\lambda = (A, B, \pi)$ as model parameter [18]. In our case, the observation space is the news collected from different newspaper and I need to split them to a finite number of vectors to use in HMM.

3.1. Overview of the Proposed NER System

Proposed system provides easy to use method which requires less effort for NER in Bengali language. People only need to tag their corpus and sentences are tested by using the system. The following steps are applied as:

- Data Processing
- Training
- Testing

3.1.1. Data Processing

Initially, it is needed to create a part-0f speech (POS) tagging list. This is the source for tag news data. The data is collected from any source of electronic news document for training that containing text. So, so as to make these file in trainable form, I have converted raw text into tagged one. Initially I decompose each word in the sentence based on Bengali grammar and tokenize the words. I also perform chunking if required. Then tagged (named entity) the words by using experiences. Finally, the corpus is trained to use.

3.1.2. Training

HMM training process specially find the name of people and place to identify an event. To train data I have used people, organization, place, time as states in different classes to learn the model structure from training data. In the training data, each and every word belongs to its own state, where it follows the transitions to the state of the word. Each state is related to its label of class of its word token. To get the HMM parameters I have found states, calculate start probability (π), transition probability (A), Word probability (B).

Procedure to find states

I simulate a Java code that take news documents of interest as input and produce output that stored in a vector space. In this phase I take an annotated text file according to POS

.

Procedure to find start probability

Start probability is the probability that the sentence starts with particular tag. In this procedure, the frequency of a tag as starting tag can be determined

Start probability (= $\frac{(Number\ of\ sentences\ start\ with\ particular\ tag)}{(Total\ number\ of\ sentences\ in\ corpus)}$

Procedure to find transition probability

To determine the transition probability, I consider that there are two pair of tags called T_a and T_b then transition probability is the probability of occurring of tag T_a after T_b.

Transition Probability (A) = $\frac{(\text{Number of sentences from Ta to Tb})}{(\text{Total number of Tb})}$

Procedure to find word probability

The probability that is assigned to particular tag to the word is called word probability of the corpus or document.

Word Probability = $\frac{(\text{Total number of occurances of word as a tag})}{(\text{Total occurances of that tag})}$

3.1.3. HMM Testing

Named entities are fined after testing sentences with all these parameters. Entity identification from text benefited classification, document search, integration, or summarization. In this research, the tasks of NER take the advantage of HMM model and Bengali grammar. I proposed the NER system based on HMM for Bengali news to identify entities from text documents that contain news about Bangladesh. In this work, I have directly applied the methods in real–time electric Bengali news paper like Ittefaq, BDNews, Prothom Alo, and other local news paper those contain news about Bangladesh. This approach may add value in the field of information extraction would probably yield higher effective values for the task I consider.

4. PERFORMANCE AND EVALUATION

4.1. Examples of POS tagging

For the experiment I used a POS tagger that contains 56,196 Bengali words with tag. In the following figure I have enlisted some of these tagged Bengali words:

হ্রাসপ্রাপ্তADJ
হাসিমুখ ADJ
হাসপাতাল NN
হামলা NN
স্মৃতি NN
ক্ষতি NN
কাজআদায়করা VM
কাজচলা VM
নির্দিষ্ট ADJ
নির্দিষ্ট করা VM
নির্দেশ NN
সবান্ধবADJ
সডাক ADJ
উপায়-উপকরণ NN
উপায়ক্ষম ADJ
কাজকরা VM
উপায়বিহীন ADJ

Figure 1. POS tag of Bengali Words

Tagging Example

Untagged input text from Bengali newspaper:

সুন্দরবনেরযাতেআরক্ষতিনাহয়,সেজন্যশ্যালাওপশুরনদীতেছড়িয়েপড়াফানেসঅয়েলসরিয়েনেয়ারকাজত্বরান্বিত করতেসংশ্লিষ্টকমকর্তাদেরনির্দেশদিয়েছেনপ্রধানমন্ত্রীশেখহাসিনা।

Tagged output:
সুন্দরবনের =NN
যাতে =null
আর =ADV
ক্ষতি =NN
না =PRP
হয় =null
সেজন্য =null
শ্যালা =NN
ও =INT
পশুর =NN
নদীতে =NN
ছড়িয়ে=VM
পড়া =ADJ
ফানেস =NN
অয়েল =null
সরিয়ে =null

নেয়ার =VM
কাজ =NN
ত্বরান্বিত =ADJ
করতে = VM
সংশ্লিষ্ট =ADJ
কমকর্তাদের =NN
নির্দেশ =NN
দিয়েছেন =null
প্রধানমন্ত্রী =NN
শেখ =NN
হাসিনা =NN

But according to this POS tagger I only can decompose each word with its POS tag. But my proposed system is for identifying entity. As I research with identify entity using NER system based on HMM from Bengali news. In order to doing so, I use Bengali grammar (এ, অ, তে, কে, রে, এরে, তে, গনেরমধে, দিগেরমধে, র, এর, কার,কের, য়, য়ে, এতে) for chunking entity. And regular expression is also used to decompose word based on Bengali letter and punctuations such as, , , ; ,!, ?, : etc. After using এ, অ, তে, কে, রে, এরে, তে, গনেরমধে, দিগেরমধে, র, এর, কার,কের, য়, য়ে, এতে, I can then decompose and chunk the entity.

For instance, if the input text is:

হানিফসিদ্দিকি,চট্টগ্রামবিশ্ববিদ্যালয়েরকম্পিউটারবিজ্ঞানওপ্রকৌশলবিভাগেরএকজনঅধ্যাপক, the sentence is now decomposed with chunk as follows:

হানিফসিদ্দিকি
চট্টগ্রামবিশ্ববিদ্যালয়ের
কম্পিউটারবিজ্ঞানওপ্রকৌশলবিভাগের
একজন
অধ্যাপক

4.2. Experiment and Result Analysis

Table 1. Tagged and non-tagged words in training sentences

Sentence No.	**Total number of words in the training sentence**	**Total number of tagged word**	**Total number of non-tagged word**
1	27	21	6

F-Measures the accuracy of the test by considering precision P and recall R in order to compute score of a system, determined the overall quality performance as well. This measure facilitates the capability of a system to extract relevant information and reject irrelevant [19]. F-Measure is the weighted average of the **precision and recall** measurements and it is displayed as a percentage value [1]. In this case precision is evaluated as the retrieval of desired words as true positive

prediction, false positive. Finally, F-measure can be defined as a harmonic mean of precision, P and recall, R:

$$F - measure = 2.\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision can be defined as

$$Precision = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

and

Recall can be defined as

$$Recall = \frac{\text{True Positive}|}{(\text{True Positive} + \text{False Negative})}$$

The HMM based NER system for Bengali news text has been trained and tested for Bengali news. The tagged and non-tagged words are presented in table 1 for a Bengali sentence. The F-measure for the tested sentence is shown in the following Table 2:

Table 2**.** Performance of NER system for POS tagged words

Sentence	**Precision**	**Recall**	**F-measure**
1	18/(18+3)=0.857	18/(18+1)=0.947	90%

Now, the performance of NER system for Bengali news is discussed based on only two sample news sentence. My purpose is to chunk words and here I emphasize on person name such as**প্রধানমন্ত্রী শেখ হাসিনা।** The developed system searches all the news documents to locate **প্রধানমন্ত্রী শেখ হাসিনা** and retrieved all the information that is relevant with this name.

Example 1

Input Text:

সুন্দরবনেরযাতেআরক্ষতিনাহয়,
সেজন্যশ্যালাওপশুরনদীতেছড়িয়েপড়াফানেসঅয়েলসরিয়েনেয়ারকাজত্বরান্বিতকরতেসংশ্লিষ্টকমকর্তাদেরনির্দেশ দিয়েছেনপ্রধানমন্ত্রীশেখহাসিনা।

Output after chunk is:

সুন্দরবনের
যাতে
আর ক্ষতি না হয়,
সেজন্য শ্যালা ও পশুর
ছড়িয়ে
পড়া ফানেস অয়েল সরিয়ে
কাজ ত্বরান্বিত করতে
সংশ্লিষ্ট কমকর্তাদের
নির্দেশ দিয়েছেন প্রধানমন্ত্রী শেখ হাসিনা।

Example 2

Input Text:
হানিফ সিদ্দিকি, চট্টগ্রাম বিশ্ববিদ্যালয়ের কম্পিউটার বিজ্ঞান ও প্রকৌশল বিভাগের একজন অধ্যাপক

Output after chunk is:
হানিফসিদ্দিকি
চট্টগ্রাম বিশ্ববিদ্যালয়ের
কম্পিউটার বিজ্ঞান ও প্র□□□□ □□□□□
□□□□
□□□□

The HMM based NER system for Bengali news text has been trained and tested with two example news sentence. The chunked and non-chunked person names are presented in table 3.

Table 3. Precision/ Recall of NER system for chunked person name

No. of sentence	Total person name	Chunked person name	Non-chunked name
2	2	2	1

The F-measures for the tested sentences for chunking is shown below:
Precision=0.7
Recall=1
So, F-measure=0.82.

3. CONCLUSIONS AND FUTURE WORK

This paper proposed a methodology to search entity in Bengali newspaper for information extraction based on HMM. The purpose of this work is the research will contribute in the field Named Entity Recognition (NER) for Bengali language. I have explained when and how NER is used in applications and enlisted the core obstacle of NER system for Bengali language, and also its benefits. The sets of tag in NER are discussed, and each tag also has detailed explanation in the aspect of grammar. In this research NER system for Bengali language is evaluated. This procedure contains HMM based NER for identifying entities. Through the evaluation results, it is easy to find that HMM is more standardized.

The method of evaluation has limitation. Using this system, I have calculated the result of precision and recall. As I have only evaluated some of sentences, the accurate value may not be shown. It is imperative to develop a perfect NER system for Bengali language in future. In the developed NER system, I cannot chunk all the person names according to Bengali grammar. Although, there are different NER systems exist in the world, but work that support Bengali NER system is really too poor and so require more research and the corpus is limited in Bengali language. Therefore, to find a new better evaluation methodology to identify any words that are used in Bengali language can be a key point in the future research.

REFERENCES

[1] D. Binu & P. Malathi , (2015) "Multi Model Based Biometric Image Retrival for Enhance Security", Indian Journal of Science and Technology, Vol 8(35), DOI: 1017485/ijst/2015/v8i35/81011.

[2] Rosan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu & Pavel Kuksa, (2011) "Natural Language Processing (Almost) from Scratch", Journal of Machine Research 12, page 2493-2537.

[3] Vivekananda Gayen & Kamal Sarkar, (2013) "An HMM based Named Entity Recognition Systems for Indian Languages", the JU System at ICON.

[4] Sudha Morwal, Nusrat Jahan & Deepti Chopra, (2012) "Named Entity Recognition using Hidden Markov Model (HMM)", International Journal on Natural Language Computing (IJNLC) Vol. 1, No.4.

[5] Roman Prokofyev, Gianluca Demartini & Philippe Cudre-Mauroux, (2014) "Effective Named Entity Recognition for Idiosyncratic Web Collections", WWW'14 Proceeding of the 23rd international conference on World Wide web, pages 397-408.

[6] Kazi Asif Ekbal, Rejwanul Haque & Sivaji Bandyopadhyay, (2008) " Named Entity Recognition in Bengali: A Conditional Random Field Approach", Proceeding of the Third International Joint Conference on Natural Language Processing: Volume-II.

[7] Gobinda G. Chowdhury,(2003) " Natural Language Processing", Annual review of information science and technology.

[8] Saeed Naz, Asif Iqbal Umar, Syed Hamad Shirazi & Sajjad Ahmad Khan, (2014) "Challenges of Urdu Named Entity Recognition: A Scarce Resourced language", Research journal of Applied Sciences, Engineering and Technology 8(10): 1272-1278.

[9] B. Sasidhar, P.M. Yohan, Dr. A. Vinaya Babu & Dr. A. Govardhan, (2011) "A Survey on Named Entity recognition in Indian Languages with particular references to Telugu", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2.

[10] Diana Trandabat, (2010) "Natural Language Processing using Semantic Frames", PhD thesis, University AI. I. Cuza lasi, Romanias.

[11] Ryohei Sasano & Sadao Kurohashi, (2007) "Japanese Named Entity Recognition Using Structural Natural Language Processing", International Joint Conference on Natural Language Processing.

[12] Eduard Hovy & Chin-Yew Lin, (1999) "Automated text summarization in SUMMARIST", I. Mani and M. Maybury, editors, Advance in Automated Text Summarization, pages 81-94, MIT Press.

[13] Julian M. Kupiec, Jan Pedersen & Francine Chen, (2001) "A Trainable document summarizer", SIGIR' 01, pages 349-357.

[14] Fan Li and Yiming , (2003) "A loss function analysis for classification methods in text categorization", ICML 03, pages 472-479.

[15] Simone Teufel & Marc Moens, (1997) "Sentence extraction as a classification task", ACL/EACL-97 Workshop on Intelligent and Scalable Text Summarization.

[16] Kazi Saidul Hasan, Md. Altafur Rahman & Vincent Ng, (2009) " Learning-Based Entity Recognition for Morphologically-Rich, Resource-Scarce Language", Association for Computational Linguistic.

[17] Tom Mitchell, (1997) "Pattern Classification and Scene Analysis", McGrawHill.

[18] Zhu Liu, Jincheng & Yao Wang, (1998) "Classification of TV Program Based on Audio Information Using Hidden", IEEE Second Workshop on Multimedia Signal Processing, 0 1998, pp. 27-32.

[19] Salton Gerard & Buckley Christopher, (1988) "Term-weighting approaches in automatic text retrieval", Information Processing and Management24, 5, 513–523.

AUTHOR

Shamima Parvez working as lecturer in the department of Computer Science and Engineering at Premier University, Chittagong Bangladesh. I have Completed Master degree from University of Chittagong, Bangladesh and also completed Bachelor of Science from the same university and department. I am working in the field of named entity recognition.