Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach

Conference Paper · June 2019

DOI: 10.1109/ICSCC.2019.8843661

4 authors, including:

Amit Kumar Das
East West University (Bangladesh)
39 PUBLICATIONS **290** CITATIONS

SEE PROFILE

Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach

Mah Dian Drovo, Moithri Chowdhury, Saiful Islam Uday, Amit Kumar Das
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
E-mail: mahdian.drovo@gmail.com, moithrichowdhury@gmail.com, udayusq@gmail.com, amit.csedu@gmail.com

***Abstract*— Named Entity Recognition (NER) is the subtask of Natural Language Processing (NLP) which tries to achieve human level on a specific domain (e.g. newspaper) to identify named entities. It seeks to locate and classify named entities (Person Name, Location, Organization names etc.), which is the most vital step of Information Extraction (IE). In many cases Machine Learning (ML) is mostly used to perform NER. Apart from that, another method is applied which is known as Rule Base approach. This paper presents a method which is using both ML and Rule Base approach together for NER basing on Bengali language. Mainly the rule based approach has been merged with ML. For ML Hidden Markov Model (HMM) and for rule base approach Regular Expression has been used. A Named Entity (NE) tagged corpus has been developed by using Bengali newspaper, which consists of 10k words that has been manually annotated with seven tags. This paper concludes with experimental results which shows two distinctive ways of our proposed model.**

Keywords-Name Entity Recognition; Hidden Markov Model; Corpus; Rule Base; Information Extraction; Named Entity; Regular Expression

I. INTRODUCTION

In 1995 at Message Understanding Conference (MUC-6), NER was first inaugurated and in 1996 at a similar conference named MET-1 saw its first NER in non-English text. NER is the most vital part of NLP, which can be used to make relation between words in a given text by Parts of Speech (POS) tagging and several application can be performed like IE, Summarization, Translations etc. There are so many languages in the world and they have different kinds of characteristics which make them ambiguous. These kinds of complexity make NER so tough. One example can be shown regarding the ambiguity level. For example, two sentences can be compared. "উদয় বিশ্ববিদ্যালয়ে যায়।" (Uday goes to university) and "সূর্য উদয় হচ্ছে।" (The sun is rising). Now in case of Bengali vocabulary, in the two sentences, the term "উদয়" is portraying two distinct meaning. In the first sentence the term "উদয়" (Uday) is a noun which describes only a name. But in the second sentence the term "উদয়" (rise) is a verb which describes its denotative meaning. Although these two terms describe two forms of POS, but still they are used in different sentences without any visual differences. In English language, names or nouns are supposed to be capitalized thus making it easy to differ from nouns to any other kinds of parts of speech. But in Bengali language, there is no scope for capitalization, which makes NER for Bengali language complicated.

Mainly, two approaches can be applied for NER: the first one is Rule Base approach and second one is ML approach. In Rule Base approach grammar based NER algorithms are applied to classify name entities. The problem of this approach is it is not robust because it can only be applied for a separate language. If one of the rules gets altered, the algorithm might need to be modified [4]. Second one is currently the trending approach where different ML methods are being applied like Hidden Markov Model (HMM) [2], Conditional Random Fields (CRFs) [1], Maximum Entropy [3]. These algorithms are very popular because ML can be applied in different languages and the maintenance of machine learning system is rather cheaper than the rule based one [1].

The proposed model is designed to use both Rule base and ML approach's characteristics to achieve better NER which will classify seven entities, which are Person's Name, Organization's Name, Location, Date, Time, Email and others (every entities except the mentioned ones) for Bengali Language. For ML approach HMM is used and Rule base approach has been merged with HMM. Mainly HMM is working to classify entities and Rule base approach is helping HMM to increase accuracy and to acquire better NER. This paper has explained HMM based Rule Base approach of our developed NER model.

II. RELATED WORK

NER on Bangla language has rarely been touched upon in previous studies. In prior studies, it can be observed that, NER is basically done in two methods. First one is ML and second one is Rule base. In one of the previous studies the researchers have used CRF model for Bengali language [1] which is one of the statistical model of ML. There are several other statistical methods like Support Vector Machines (SVMs), HMM [2, 8] Maximum Entropy Approach [3] etc. that are used for NER.

978-1-7281-1557-3/19/$31.00 ©2019 IEEE

978-1-7281-1557-3/19/$31.00 ©2019 IEEE

In another study, NER was also done by Rule base approach for Arabic language where it was explained that Rule based approach can only be applied upon individual language [4]. Rule Base approach is not only stagnated in NER, but is also used in IE. Another approach for Bengali language can be mentioned which is called Knowledge Base approach [5]. In this approach, gazetteer is used which contains categorized entities.

In NER it is observed that, ML is mostly used. There are few reasons for this method and primary one is it can be used for any language which is not applicable for Rule Base approach. Right after NER, rest of the parts of NLP can be performed.

III. Named Entity In Bengali Overview

NER in Bengali is not that much well known like English. As a result, developing a better model for Bengali Language it is quite challenging. There are so many prior researches already done for English language and the resources are so rich.[18, 19, 20] But resources for Bengali language are close to unobtainable which makes it challenging to develop a better NER. For this reason, the NE tagged corpus has been developed from available Bengali online newspaper named Prothom Alo. There was a tool developed which can manually tag NE for training data by a user. This is not any special tool rather with this tool users can tag words through interface and give an output to corpus. This corpus has 10k words which are annotated with seven tags. The developed corpus' tags and details are demonstrated in the Table I.

TABLE I. Named Entity Classifier Details

NER Tag Name	**Meaning**	**Example**
Per	Single Worded Person's Name	Drovo/Per
Loc	Single Worded Location Name	Dhaka/Loc
Org	Single Worded Organization Name	Unilever/Org
Mail	Single Worded Email Address	mahdian.drovo@gmail.com/Mail
Time	Single or Multi Worded Time	12:30/Time
Date	Single or Multi Worded Date	31.12.1996/Date
Other	Every Single Worded entities except the mentioned ones	happiness/Other

A. Named Entity Corpus

A corpus was developed which fitted the proposed model. In the corpus there are sequenced words in sentences which is annotated with Name Entity (NE) tags because HMM require this kind of corpus [8]. The developed corpus contains one word in a row which is annotated with two other values. That means there are three values in a row. First one is the number of the sentence, second one is the word itself and third one is the NE tag. The sentence number is added to calculate start probability for HMM which will be described in Section IV. The corpus has 10k words which is extracted from the combination of 690 sentences.

B. Rule Base Approach

Rule base approach can be applied for specific language based on its grammatical form. But it becomes so much complicated to maintain [1] and it is also specific for that particular language [4]. But in some ways it can be so much powerful to detect NE. There are some NE which has some specific patterns like Date, Email, Time etc. [10, 11, 12]These kinds of entities have some specific ways to be presented in a text otherwise there are no other ways. If date can be taken as an example, it can be represented in form of 20 January 2017, 20.01.2017 or 01/20/2017 as well as few other types of forms which are very limited. This representation system can be taken as feature what can be detected by Regex. Regex stands for Regular Expression which is used to match character in a string.[13, 14] This property is very powerful to detect character combination. So this entities are detected by using Rule Base approach which is written by Regex. For this kind of NE Rule Base approach shows better accuracy than ML approaches [6]. If these kinds of entities were detected by using Rule base approach there is no need to detect them by ML. These can be classified directly by Rule base approach and give the result to HMM and HMM will treat it as it has been classified by Rule Base. This technique was used because there is no need for ML to predict those which already had a pattern. As ML's prediction might be faulty where few entities got their own pattern which won't be incorrect if we detect those with regular expression.

IV. Proposed NER System And Hidden Markov Model (HMM)

HMM is a sequence classifier. That means it will make calculation over training set with respect to its sequence and will provide classifier on test set. For this reason the developed corpus is sequence of words in sentences. The calculation of HMM is basically based on three different components [8, 9]. Those are Initial Probability, Transition

Probability and Emission Probability. When these values are measured based on training corpus, the best sequence of states can be calculated. This is called decoding and it can be solved using the Viterbi algorithm [7, 8, 9].

A. *Start Probability*

Start probability is denoted by π. It demonstrates the probability of a word being start position of a sentence. For this purpose in the developed corpus, sentence numbers were put in front of the word. It helps to measure the start probability. For start probability over state π_i is the probability that the Markov chain will start in state i. There could be some state have $\pi_i = 0$ which means they cannot be initial state.

B. *Transition Probability*

Transition probability which is denoted by A is estimation based on a corpus of transition probability between words in a sentence which means for each A_{ij} the probability of moving from state i to state j. So it can easily estimate the probability of transition among words by counting occurrences between words.

The estimation is:

$$A_{ij} = \frac{|q_i q_j|}{|q_i|}$$

C. *Emission Probability*

Emission Probability is a measurement which indicates the probability of a vocabulary being emitted under a specific class which is denoted by B. The probability of a word w being emitted under a class q_i is calculated being divided by the total number of words is emitted under that particular class.

The estimation of emission probability $B_i(w_t)$ is:

$$B_i(w_t) = \frac{|w_t q_i|}{|q_i|}$$

D. *Viterbi Algorithm*

The Viterbi Algorithm is a dynamic programming which tries to find most likely sequence of hidden states what is called Viterbi path. This path is based on Hidden Markov Model's (HMM) components. This algorithm generate a path $X = (x_1,x_2,\ldots,x_T)$ where different hidden state are there. In this part rule base approach was merged by giving a value for a state's Emission Probability what is detected by Rule Base approach. That implies that for that known stage $B_i(o_t) = 1$ and by multiplication with transition probability it will be taken as that selected stage.

In the proposed model HMM and Rule base approach were merged which means the both of them will perform concurrently which has been described in Figure 1. There is another way when only HMM is run at first than Rule base approach take place which has been described in Figure 2. The result for every mentioned way is shown in Section V.

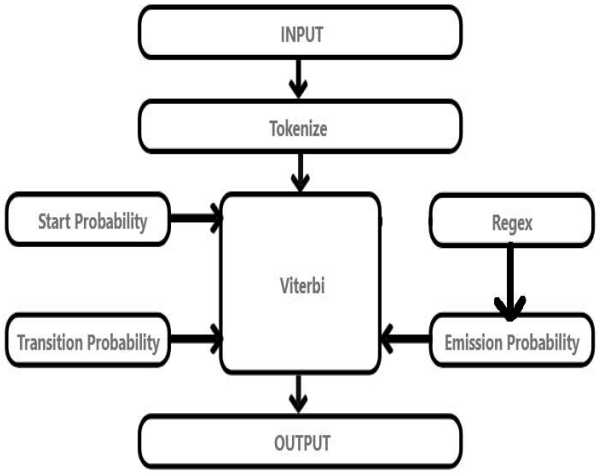

Figure 1. Process where HMM and Rule Based Approaches perform concurrently

The whole process of concurrent actions of HMM and Rule Base approach is demonstrated in figure 1. The process begins with the INPUT where texts are inserted and the obtained OUTPUT is classified with the same sequence of words in the text. In the field of Tokenize from Figure 1, probabilities are measured with the help of corpus as Viterbi requires Start Probability, Transition Probability as well as Emission Probability. Viterbi takes help from the mentioned probabilities and through its own calculation, it predicts any particular word's classifier name.[15, 16, 17] If Regex identify a particular word's class name (Time or Date or Email) before, then Viterbi doesn't predict that word. Viterbi takes that class name as absolute and progresses accordingly. By this joint process of HMM and Rule Base approach, words of the whole text are classified.

The Figure 2 is different from Figure 1 in the sense that, in Figure 2 Regex takes place after Viterbi is done predicting. Which means that, like the previous process, Viterbi predicts the classifier name from the words of the text. After Viterbi is done predicting, the words with the classifier name are sent to Regex, which then tries to identify the words according to its rules.

E. *Unknown Words*

In some instances there can be a word which might not exist in our corpus. In that case for that unknown word, w emission probability $B_i(w_t) = 0$. During this period HMM cannot identify that word's classifier name. To handle this particular situation, some processes are applied. One of them is Laplace Smoothing [9].

In Laplace Smoothing, the following equation is applied,

$$B_i(w_t) = \frac{|w_t, q_i| + 1}{|q_i| + |W|}$$

Here W indicates number of symbols.

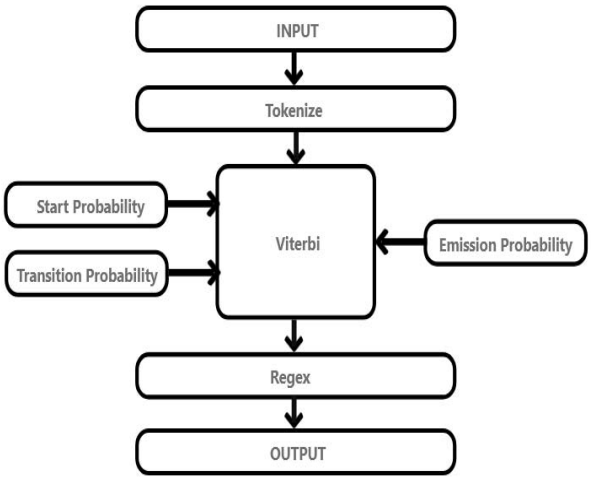

Figure 2. Process where HMM perform first than Rule Based Approach take place

V. Evaluation And Results

The developed corpus contains 690 sentences which includes 10k words that are taken from Bengali online newspaper named Prothom Alo available in web. The corpus was categorized using seven classifiers and while running trains, the model tried to predict them by using those same tags. The model has been merged by machine learning approach and rule based approach that has been presented in two separate ways. In the first way, both the machine learning approach and rule based approach run simultaneously. In the second way, ML predicts and upon finishing rule based approach take place.

10-fold Cross Validation test was performed and got 68.98% F-Score for the first way of which the information is demonstrated in the Table II.

TABLE II. 10 Fold Cross Validation For First Way

Test Set No.	Precision	Recall	F- Score (%)
1	75.02	69.83	72.34
2	68.03	65.97	66.98
3	79.64	73.31	76.35
4	81.01	74.81	77.79
5	37.27	23.37	28.73
6	73.10	69.38	71.19
7	68.43	81.43	74.37
8	87.49	77.36	82.12
9	60.47	76.20	67.43
10	72.69	72.30	72.49
Average	70.32	68.39	68.98

After applying the second way, it got 71.59% F-score of which the information is demonstrated in the Table III.

The equation used to measure F- Score:

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

This F-Score mainly depends on Corpus .For Named Entity Recognition it needs huge amount of data but for Bengali language it is not available, that's why the corpus was made. If corpus got abundant amount of data, the result could've been better [2].

In newspapers, time and emails are rarely prominent. That's why random texts containing time and email had to be added in corpus. Mainly, in this model, along with other entities it uses pattern wise entities. But in the newspaper, these kinds of pattern wise entities were not available. As a result few extra sentences were added which had patterned entities. It helps to eliminate the lacking of those entities

VI. Conclusion And Future Work

In this paper, a system was developed which merged ML approach and Rule based approach for NER by using corpus containing 10k words. It has been observed that both of the ways have given close F-score but in the second mentioned way, a rather better result was obtained. To analyze the performance of our developed system, corpus containing more vocabularies can be used. For machine learning Hidden Markov Model (HMM) was used where other methods like CRF, Support Vector Machines (SVMs) etc. can also be applied where these can be merged with Rule base approach.

TABLE III. 10 FOLD CROSS VALIDATION RESULT FOR SECOND WAY

Test Set No.	Precision	Recall	F- Score (%)
1	75.48	69.97	72.62
2	68.03	65.97	66.98
3	85.86	73.55	79.23
4	81.01	74.81	77.79
5	44.42	44.80	44.61
6	74.20	69.40	71.72
7	68.75	81.47	74.57
8	87.91	77.38	82.31
9	68.03	77.02	72.25
10	75.55	72.35	73.91
Average	72.92	70.67	71.59

REFERENCES

[1] S. Song, N. Zhang and H. Huang, "Named entity recognition based on conditional random fields", Cluster Computing, 2017.

[2] D. M. Bike, R. L. Schwartz and R. M. Weischedel, "An Algorithm that Learns What's in a Name", 1999.

[3] S. Saha, P. Mitra and S. Sarkar, "A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition", Knowledge-Based Systems, vol. 27, pp. 322-332, 2012.

[4] M. Hjouj, A. Alarabeyyat and I. Olab, "Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition", International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016.

[5] J. Islam, M. Mubassira, M. R. Islam and A. K. Das, "A Speech Recognition System for Bengali Language using Recurrent Neural Network," 2019 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019.

[6] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, " An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019.

[7] A. K. Das, A. Ashrafi and M. Ahmmad, "Joint Cognition of Both Human and Machine for Predicting Criminal Punishment in Judicial System," 2019 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019.

[8] S. Sadat, "Named Entity Detection in Bangla Text Using Knowledge Base and Elimination Method", 2017, pp. 31, 35-46.

[9] A. K. Das, T. Adhikary, M. A. Razzaque and C. S. Hong, "An intelligent approach for virtual machine and QoS provisioning in cloud computing," The International Conference on Information Networking 2013 (ICOIN), Bangkok, 2013, pp. 462-467.

[10] A. Tashnim, S. Nowshin, F. Akter and A. K. Das, "Interactive interface design for learning numeracy and calculation for children with autism," 2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE), Phuket, 2017, pp. 1-6.

[11] F. T. Zohora, M. R. R. Khan, M. F. R. Bhuiyan and A. K. Das, "Enhancing the capabilities of IoT based fog and cloud infrastructures for time sensitive events," 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, 2017, pp.224-230.

[12] Y. Li, R. Krishnamurthy, S. Raghavan and S. Vaithyanathan, "Regular Expression Learning for Information Extraction", in *Empirical Methods in Natural Language Processing*, 2008.

[13] A. K. Das, T. Adhikary, M. A. Razzaque, M. Alrubaian, M. M. Hassan, Z. Uddin, and B. Song, "Big media healthcare data processing in cloud: a collaborative resource management perspective," Cluster Computing, Volume 20, Issue 2, pp 1599-1614, June 2017.

[14] M. R. Ullah, M. A. R. Bhuiyan and A. K. Das, "IHEMHA: Interactive healthcare system design with emotion computing and medical history analysis," 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT), Himeji, 2017, pp.1-8.

[15] D. Jurafsky and J. H. Martin, "Speech and Language Processing", 2009, pp. 151-172.

[16] D. Chopra and S. Morwal, "Named Entity Recognition in English Using Hidden Markov Model", International Journal on Computational Science & Applications, vol. 3, no. 1, pp. 1-6, 2013.

[17] M. A. A. Mamun, J. A. Puspo and A. K. Das, "An intelligent smartphone based approach using IoT for ensuring safe driving," 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, 2017, pp. 217-223.

[18] M. Akter, F. T. Zohra and A. K. Das, "Q-MAC: QoS and mobility aware optimal resource allocation for dynamic application offloading in mobile cloud computing," 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, 2017, pp. 803-808.

[19] A. Jayaweera and N. Dias, "Hidden Markov Model Based Part of Speech Tagger for Sinhala Language", International Journal on Natural Language Computing, vol. 3, no. 3, pp. 9-23, 2014.

[20] T. Adhikary, A. K. Das, M. A. Razzaque, A. Almogren, M. Alrubaian, and M. M. Hassan, "Quality of Service Aware Reliable Task Scheduling in Vehicular Cloud Computing," Mobile Networks and Applications, Volume 21, Issue 3, pp 482-493, June 2016.