Bengali Named Entity Recognition: A survey with deep learning benchmark

Md Jamiur Rahman Rifat
Department of CSE
Daffodil Internatinal University
Dhaka, Bangladesh
rifat15-5611@diu.edu.bd

Sheikh Abujar
Department of CSE
Daffodil International University
Dhaka, Bangladesh
sheikh.cse@diu.edu.bd

Sheak Rashed Haider Noori
Department of CSE
Daffodil International University
Dhaka, Bangladesh
drnoori@daffodilvarsity.edu.bd

Syed Akhter Hossain
Department of CSE
Daffodil International University
Dhaka, Bangladesh
aktarhossain@daffodilvarsity.edu.bd

***Abstract*— Sequence labeling is a complex task in natural language processing where the data set used to be biased to a specific class mainly to "not named entity" class. Previously several machine learning approaches were harnessed for Bengali named entity recognition where additional information like Parts Of Speech (POS) tag, suffix value, optimal number of context words were required. This study aspires to give an overview of past methods on Bengali named entity task along with leveraging different neural networks on a new dataset at an easy way. A dataset consisting of 96697 tokens were annotated in the house where 67554 tokens were applied for training and 29143 words were for testing purposes. Then several deep learning methods were exploited where Bidirectional Gated Recurrent Unit (BGRU) came up victorious with f1 score 72.66%. The value may not be promising after comparing with other methods but different studies calculated the precision and recall value differently. Increasing the number of training data could raise the performance metrics along with other forms of word embedding.**

***Keywords*— NER, BLSTM, CNN, BGRU, Deep Learning**

I. INTRODUCTION

Named Entity Recognition (NER) can be regarded as an act of labelling different tokens in a text to some preordained categories. The category could be anything from a drug name to name of a place. It got various applications in Natural Language Processing like Information Extraction, Machine Translation, Question Answering and Summarization etc [1]. In general the common tag set for NER constitutes name of a person, name of company, name of a place. But this can be extended to any entity which will be solely dependent on the annotated corpus. Previously several machine learning approaches were taken for Bengali named entity task Such as Hidden Markov Model (HMM)[2], Conditional Random Fields (CRF) [3], Support Vector Machine (SVM) [4], Maximum Entropy (ME) [5] and Multi Engine Method [1].

It was believed that machine learning approaches were cost efficient [1]. But those methods require explicit feature engineering and selecting the optimal sets of features is so troublesome. For Bangla named entity classification, the multi engine approach incorporated by Ekbal et al [1] displayed the best f-score. But their paper also demonstrated the impact of choosing different named entity feature set. Using context patterns as feature boost up the f-score for multi engine by 2.27%. Like that other dominating feature may exist in the corpus which may need much effort and study to explore Therefore being inspired from the work of Chiu et al [7], deep learning methodologies were introduced to overcome the hurdle of generating superlative feature set. Their work used bidirectional Long Short Term Memory (BLSTM) to capture word level features and Convolutional Neural Network (CNN) for character level features.

Sometimes the success of a NER model depends on the standard of a good annotated corpus. Manual annotation of a big size corpus requires much time and many human annotators. Therefore, this study tried to annotate 96697 tokens in the house. The tokens were collected from a renowned newspaper. Then we have differentiated the classification performance on distinct deep learning models.

II. RELATED WORKS

Though NER is a very necessary topic to NLP researchers but this could draw very little attention to Bengali linguists. A comparative study of different methodologies is summed up in Table I. Among the little contributions the paper published by Asif Ekbal [1-5] is notable. To best of our knowledge, after 2012 -2017 no paper has been published on any renowned journal or conferences about NER in Bangla. In 2017 shamima parvez published a paper on NER [6].

In the paper of Ekbal et al [2] Machine learning over the traditional rule based system was chosen because it is less costlier to maintain. Suffix features and a lexicon have been used here. Moreover a HMM based POS tagger was used to improve the performance. Recall, Precision and F-score was reported as 90.2%, 79.48% and 84.5% respectively.

Ekbal et al [3] annotated the previous corpus of Bengali newspaper with Named Entity tags and obtained better result than the previous experiments with an improvement of 6.2% of F- Score. CRF is very popular to the computational linguistic family for sequence labeling task. Using CRF it was noticed that the Recall value as 93.08%, precision as 87.8%, F score as 90.7%. But those machine learning based models required additional features like part of speech (POS) tag of current token, previous word of current token, prefix value of current token etc which will be totally avoided in our model. In the paper of Hasanuzzaman et al [5] the work used the maximum entropy and marked recall, precision and f score as 88.01%, 82.63%, 85.22% respectively. Later on it was noticed that linguistic features improve the performance of the system.

There was an innovative approach of introducing weighted voting techniques in the paper of Ekbal et al [1]. The previously used three models of SVM, CRF and ME were incorporated. Here a new feature was added "Lexical context patterns" obtained from unlabeled 3 million word forms in a semi-automatic way. This model sets a state of the art method for named entity recognition in Bangla. Recall, precision and f score were reported as 93.98%, 90.63% and 92.28% respectively.

In the paper of Chiu et al [7], bidirectional LSTM was used for reading the words from left to right and right to left. Embedding vector plays a vital role here and employed character embedding here also to notice the suffix and prefix relationship. In evaluating their model, two dataset were used. The f score found to be 91.62% for CoNLL-2003 dataset and 86.28% for OntoNotes 5.0 datasetThe paper of Rabby et al. [8] insights the power of deep learning in Bengali domain where he used CNN to recognize Bengali handwritten characters and Abujar et al. [9] implied deep learning for sentence similarity estimation for text summarization.

TABLE I
A COMPARATIVE STUDY ON STATE OF THE ART APPROACHES OF NER IN BANGLA LANGUAGE

Study	Year	Model	Tagset	Training size	Corpus source	Test size/ Methods	Evaluation metrics
Ekbal et al. [2]	2007	HMM	17 as input And 4 as output	150k words (POS annotated not NE annotated)	Bengali newspaper Which contains around 34 million word forms	10 fold Cross validation	R-90.2% P-79.48% F-84.5% (only Bengali language)
Ekbal et al. [3]	2008	CRF	17 with 4 major categories	150k words annotated with NE tags	Bengali Newspaper	10 fold cross validation	R-93.8% P-87.8% F-90.7%
Ekbal et al. [4]	2008	SVM	16 tags	150k words	Bengali Newspaper	10 fold cross validation	R-94.3% P-89.4% F-91.8%
Hasanuzzaman et al. [5]	2009	ME (Maximum Entropy)	4 tags	122467 tokens	IJCNLP-08 NER Shared Task on South and South East Asian Languages (NERSSEAL) [10]	10 fold cross validation	R-88.01% P-82.63% F-85.22%
Ekbal et al. [1]	2009	SVM CRF ME	17 with 4 major tags	272k words	150k of newspaper And 122k of NERSSEAL [10]	35k gold standard	R-93.98% P-90.63% F-92.28%
Chiu et al. [7]	2015	BLSTM CNN	4 tags		CoNLL-2003 dataset		F-91.62%
					OntoNotes 5.0 dataset		F-86.28%
Shamima Parvez [6]	2017	HMM		1 sentence with 21 tagged words		2 sentences	R-1 P-.7 F-.82(chunked person name)
Chowdhury et. al.[12]	2018	LSTM CRF(token+ POS+gazetteers As features)	7 tags		Bangla Content Annotation Bank (B-CAB)		R-.67 P-.78 F1-.72 (Best Model)
Aguilar et. al. [14]	2019	BLSTM CNN CRF					41.86% entity F1-score and a 40.24% surface F1-score.

*Recall, Precision, F score is represented by R, P, F respectively

Chowdhury et. Al [12] found to use different combinations of features with the CRF model. The best result was found when they used token, parts of speech (POS) and gazetteers as features.

Ibtehaz et. al. [13] demonstrated an unorthodox approach of finding NER from unstructured data. Partial string matching based on breadth first search, trie data structure and dynamic programming were harnessed in their approach.

Lastly a recent study of aguilar et. al. [14] showed the use of neural network as feature extractor and CRF as classifier. Their model is so good in addressing noisy data. Evaluation metric recorded was 41.86% entity F1-score and a 40.24% surface F1-score.

III. METHODS

A. Data acquisition

For training the proposed model and annotation, approximately 96697 tokens collected from a renowned newspaper were taken. The news were scrapped from their website. After That using the NLTK tokenizer each words was tokenized including punctuations also. For example, উন্নয়নের বিস্ময় বাংলাদেশ বাংলাদেশের অগ্রগতি উদাহরণ দেওয়ার **মতোই।** this sentence will be turned as like in Table II.

TABLE II
TOKENIZING THE WORDS

উন্নয়নের
বিস্ময়
বাংলাদেশ
অগ্রগতি
উদাহরণ
দেওয়ার
মতোই
।

Eack token will be saved in a file using newline. For identification of end of a sentence an extra new line will be added. In pre-processing tokens were not stemmed to its root form as the suffix and prefix value can bear additional important information for our model.

B. Annotating the corpus

Before starting the annotation procedure need to define the classes in which the study want to label the tokens.

$$y = f(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where $x \in$ {input word} and $y \in$ {B-PER, I-PER, B-ORG, I-ORG, B-LOC ,I-LOC, TIM, O}.The Table III depicts the label class in details.

The composition of different tags in our corpus is demonstrated in Table IV. It is found that most of the tokens belong to "O" tag which is almost 85.91%. So the data set is highly skewed but we can not make it oversampled or under sampled because this sequencing information is needed to classify the other tokens.

C. Word Embeddings

Word Embeddings is one of the core features that was incorporated in our model. It was observed in [7] that the performance of NER system slightly varies with the choice of embedding methods. Though there lies different publicly available word embedding vectors for English but unfortunately for Bangla it is very rare. In that respect this model used the word embeddings built by Alam et al [11], the dimension of which is 300d and vocab size of the model is 436126.

TABLE III
ANNOTATING SCHEME

Class name	Explanation	Example
B-PER	Marks the beginning of a persons name	***মো*** আজহার রহমান
I-PER	Points the inside of a multiword person name	মো ***আজহার রহমান***
B-ORG	Marks the beginning of an organization's name	***আন্তর্জাতিক*** মুদ্রা তহবিল
I-ORG	Points the inside of a multiword organization's name	আন্তর্জাতিক ***মুদ্রা তহবিল***
B-LOC	Marks the beginning of a location name	***দক্ষিণ***এশিয়াকে
I-LOC	Points the inside of a multiword location name	দক্ষিণ ***এশিয়াকে***
TIM	Marks the time expression	গত, বুধবার
O	Anything other than the above mentioned categories. It is the "not named entity" class.	উদাহরণ, ।

D. Character Embeddings

Induced character embeddings is of 30 dimension and the values were randomly taken with uniform distribution with range [-0.5, 0.5]. the character set contains 101 characters with punctuation marks and special character. The unknown characters were also taken into account and unified to a single embedding. For extracting character level features like prefix and suffix value a CNN layer is used in the model.

TABLE IV
COMPOSITION OF DIFFERENT CLASSES

Tag Name	Composition
B-PER	2.87%
I-PER	2.61%
B-LOC	2.30%
I-LOC	.35%
B-ORG	2.12%
I-ORG	1.39%
TIM	2.41%
O	85.91%

E. BLSTM+CNN based Architecture

The model started with the Input of character. The shape of the character input was (m, 101) where m is the number of instances for our case the number of sentences and As the character list contains 101 characters so the value comes from there. Then the above mentioned character embeddings with a dropout value 0.25 were included. We uniformly used the dropout value as 0.5 but it was seen that [7] the change in the value of dropout affects the value of f score. Then used three time distributed layer for different purposes. The 1st layer is for convolution with the activation function tanh. Then the next layer is for max pooling . The other layer was used to flatten the neural network and at last a dropout was used. This output was concatenated with the output found from the word input vector and its corresponding embeddings. At last a time distributed layer of BLSTM was used. A dropout value 0f 0.5 is also used for BLSTM layer. We have used some variations in our model. Instead of using BLSTM, sometimes we have

used GRU. There lies very subtle differences between LSTM and GRU where LSTM uses three gates meanwhile GRU uses two gates. Reset gate is used to combine new information with previous output whereas update gate determines how much information of previous output should be used for the new computation. Sometimes the dropout value was ignored in our procedure. At last the model was compiled using the "nadam" optimizer. The overall model view is represented in Fig. 1.

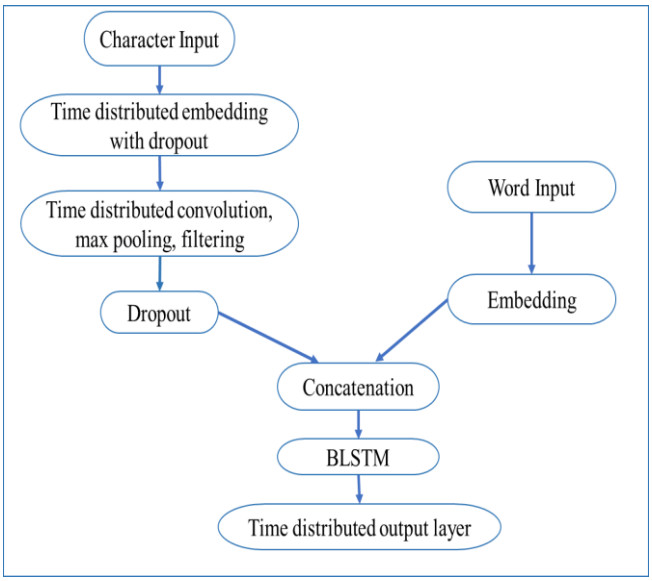

Fig. 1. BLSTM+CNN based Model

F. BLSTM+CNN+CRF based Architecture

This model is pretty much similar with the model in Fig. 1. Here we have used a CRF layer as output layer. CRF is so good at sequence labelling. It takes the context into account for setting a class of a token. Keras implementation of CRF is used. For compiling the model "crf.loss_function" was used. The dropout value was totally avoided in the model. A variation was made by replacing GRU with BLSTM. The model is illustrated in Fig 2.

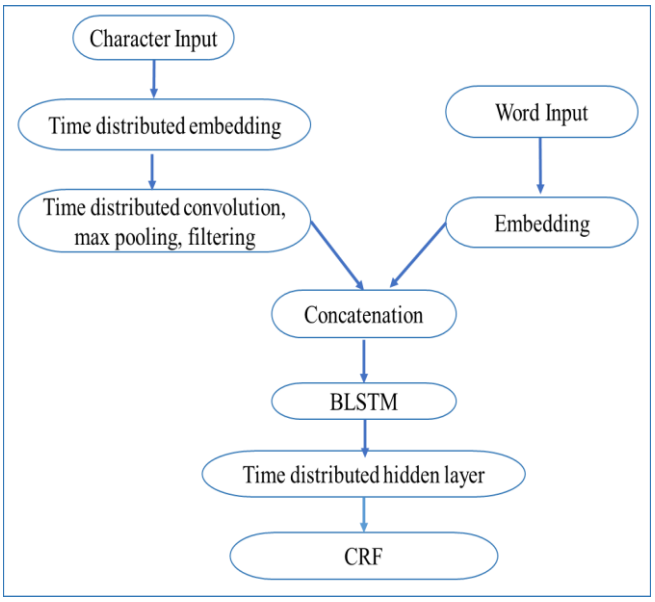

Fig. 2. BLSTM+CNN+CRF Based Model

IV Result

The performance was evaluated on the basis of precision, recall and f1 score. Ekbal et al [2] calculated the precision value after dividing the number of correctly tagged named entities by the total number of named entities and recall value by diving the number of predicted named entities by the total number of named entities. But exact matching was employed in our model. Though "B-PER" and "I-PER" both indicate a person's name, but we have considered them as two distinct class.

TABLE V
OUTCOMES OF DIFFERENT MODELS WHERE BESTS ARE MARKED BOLD

Model	No of epochs	Precision	Recall	F1
BLSTM+CNN-dropout	35	77.00	68.22	71.92
BLSTM+CNN+dropout	35	**85.47**	65.27	71.90
BLSTM-dropout	35	80.35	60.82	68.00
BGRU+CNN	35	73.32	**72.27**	**72.66**
BLSTM+CNN+CRF	35	79.71	65.41	70.79
BGRU+CNN+CRF	35	81.63	63.80	70.44

From table 4, it is observed that, the F1 score is almost similar where the value of recall has fluctuated more. Though BGRU+CNN model gives us the best recall and F1 score, but it gives us the lowest precision value. BLSTM+CNN+dropout obtained highest precision value. The lowest f1 score of Only BLSTM model gives insight that CNN was so powerful in finding character level information. And Fig. 3 describes the confusion matrix of our best model.

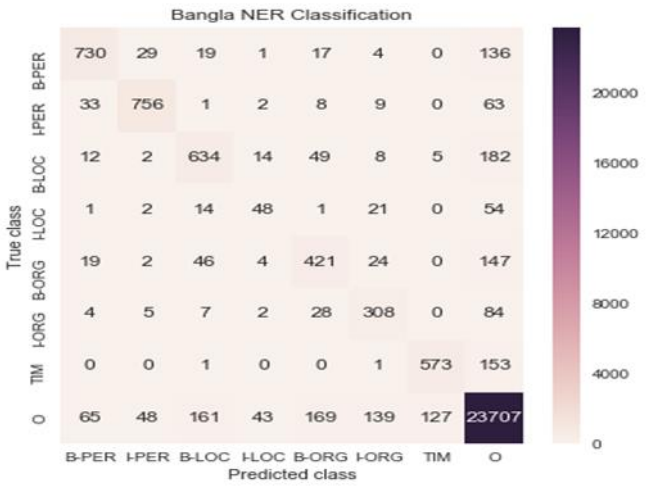

Fig. 3. Confusion Matrix of BGRU+CNN

Fig. 3 delineates that most of the wrong prediction occurs with the "O" class. If we consider Beginning-Inside (B-I) format as a single class then there will be increasimg in the accuracy value.

V. Discussion And Future Work

As the data set was highly skewed to "O" class, so from the figure it is found that most of the false class belong to "O"

tag. Accumulating more training data set can significantly improve our system. There lies a provision to include some additional features in our model. For English chiu et al. [7] used extra casing features, lexicon value. As Bangla does not contain any casing properties so need not to worry about it. But other features like lexicon value, part of speech value can be used in our model. Also further experiments can be run to judge the effect of different word embeddings method for our case. A lexicon can be built in an unsupervised way to label unknown words perfectly.

VI. Conclusion

This paper introduces an easier way of Bengali NER task with descent performance metrics which discourages the mundane system of using several handcrafted features. Convolutional and Recurrent layers are so powerful in deep learning models, so we have applied them to get the best output. There subtle differences in their f1 value proves our assumption to be right. One of the limitations of our work is that we could not handle the skewed class "O" in our work, for that reason we are getting so many errors for this particular class. Also the use of some important features like POS and gazetteers were absent in our model.

Acknowledgment

We would like to thank DIU-NLP and Machine Learning Research LAB for their tremendous support..

References

[1] A. Ekbal and S. Bandyopadhyay, "Named Entity Recognition in Bengali: A Multi-Engine Approach", *Northern European Journal of Language Technology*, vol. 1, pp. 26-58, 2010.

[2] A. Ekbal and S. Bandyopadhyay, "A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies," in *Pattern Recognition and Machine Intelligence*, 2007.

[3] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Named entity recognition in Bengali: A conditional random field approach," *Proc. IJCNLP*, 2008.

[4] A. Ekbal and S. Bandyopadhyay, "Bengali Named Entity Recognition using Support Vector Machine," in *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*, 2008.

[5] M. Hasanuzzaman, A. Ekbal, and S. Bandyopadhyay. "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi." *International Journal of Recent Trends in Engineering* 1.1, pp. 408-412, 2009.

[6] S. Parvez, "Named Entity Recognition from Bengali Newspaper Data," *Int. J. Nat. Lang. Comput.*, vol. 6, no. 3, pp. 47–56, 2017.

[7] J. Chiu and E. Nichols. "Named entity recognition with bidirectional LSTM-CNNs." *Transactions of the Association for Computational Linguistics*, *4*, pp.357-370, 2016.

[8] A. Rabby, S. Haque, S. Islam, S. Abujar and S. Hossain, "BornoNet: Bangla Handwritten Characters Recognition Using Convolutional Neural Network", *Procedia Computer Science*, vol. 143, pp. 528-535, 2018.

[9] S. Abujar, M. Hasan, and S. Hossain. "Sentence Similarity Estimation for Text Summarization Using Deep Learning." In *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology*, pp. 155-164. Springer, Singapore, 2019.

[10] http://ltrc.iiit.ac.in/ner-ssea-08. Accessed on 10 January 2019

[11] F. Alam, S. A. Chowdhury, and S. R. H. Noori, "Bidirectional LSTMs - CRFs networks for bangla POS tagging," in *19th International Conference on Computer and Information Technology, ICCIT 2016*.

[12] S. Chowdhury, F. Alam and N. Khan, "Towards Bangla Named Entity Recognition", *2018 21st International Conference of Computer and Information Technology (ICCIT)*, 2018.

[13] N. Ibtehaz and A. Satter, "A Partial String Matching Approach for Named Entity Recognition in Unstructured Bengali Data", *International Journal of Modern Education and Computer Science*, vol. 10, no. 1, pp. 36-45, 2018.

[14] G. Aguilar, S. Maharjan, and T. Solorio, "A multi-task approach for named entity recognition in social media data." *arXiv preprint arXiv:1906.04135,* 2019.