A Review Paper: Bangla Named Entity Recognition

Lamia Hasan Rodoshi
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
BRAC University
Dhaka, Bangladesh
lamia.hasan.rodoshi@g.bracu.ac.bd

Ramisa Fariha Joyee
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
BRAC University
Dhaka, Bangladesh
ramisa.fariha.joyee@g.bracu.ac.bd

Md Humaion Kabir Mehedi
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Md. Farhadul Islam
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
BRAC University
Dhaka, Bangladesh
md.farhadul.islam@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—NER, a component of the Information Extraction in locating named things and their class within a document. This is one of the fundamental tasks in many natural language processing activities. The NER aggregates the supplied entities into many predefined categories, including person, name, organization, date, number, etc. In this essay, we've looked at some works that classified and identified named things in Bengali using a number of methods. DCN BiLSTM, Gated Recurrent Unit (GRU) Long Short-Term Memory (LSTM), HUNER, Support Vector Machine (SVM), Hidden Markov Model (HMM), etc. are a few of them. We also looked at various studies that used hybrid approaches, rule-based approaches, and approaches based on machine learning to identify named entities.**

***Index Terms*—BiLSTM, NER, NLP, DCN**

I. Introduction

Entity recognition is a widely researched technology for numerous uses, including IE, summarization, translation, and POS tagging, machine translation, question answering, information retrieval, and natural language processing (NER). One of the most important tasks in almost all domains of Natural Language Processing is Named Entity Recognition, a subdomain of Information Retrieval [15]. Each of the many languages that are spoken throughout the world has distinctive characteristics that make it challenging to understand them. NER is challenging due to this kind of complexity. A single word in Bengali can be used for a variety of purposes and have multiple meanings. It is easy to distinguish between nouns and other types of speech elements in English since names and nouns are capitalized.However, Bengali does not allow capitalization, which makes NER challenging for Bengali speakers. Therefore, in this article, we evaluated a number of papers to determine which models and classifiers are most frequently used and useful for identifying named entities.

II. Related Works

An essential component of natural language processing(NLP) is named entity recognition (NER). Sadly, Bangla language is still considered as a low resource language in the NLP community. Among the small amount of contributions, we have taken some of the Bangla NER related papers to review and make a comparative study.

The study in [1] proposed Densely Connected Network (DCN) model along with Bi-LSTM for information extraction. In their study, collected dataset from wikipedia and 3 newspapers i.e. Ittefaq, Bangladesh Pratidin and Kaler Kantho consisting of 71 thousand Bangla sentences with proper annotation has been used using IOB tagging format and 4 entity types are annotated i.e. person, location, organisation and object. They have improved the accuracy by using word embedding, character level feature extraction, then Bi-LSTM for sequence labelling and finally, DCN for character feature extraction. On the proposed model DCN-BiLSTM, they attained precision of 68.95, recall of 58.62, and an F1 score of 63.37.

According to another study [8], A Gated Recurrent Unit(GRU) based NER system was recommended to identify four NE (Person, Location, Organization, and Day) from a renowned Bangla Online newspaper "Prothom Alo". On manually annotated data, they used features that are independent of language. They mentioned having little data and using ReLU on Tanh, where it works well but takes a long time to converge. They offer an F1-score of 69 percent for the small dataset. The

NE is tagged by annotators using the IOB format, where I stands for inside, B for beginning, and O for not a NE. GRU is utilized in this case since it is simpler than LSTM. To get around the vanishing gradient issue with simple RNN, they employed a three-layer network and a more versatile unfolded RNN. Their proposed NER system used supervised machine learning to develop the system also they collected raw data and annotated them manually. Then they applied some pre-processing steps and feeded them into GRU model.

In another paper [9], Author suggested HunFlair, a cutting-edge biomedical NER tagger. The HUNER method must be included into the Flair NLP framework by integrating a pre-trained language model in order to generate this HunFlair. On a number of evaluation corpora, it outperforms the state-of-the-art performance. It also trained a cross corpus to prevent bias particular to one corpus. These tools are used to train and assess small gold standard datasets, whereas HUNER tagger is used for large datasets. HUNNER, however, doesn't rely on a previously taught language model. To pretrain, it requires a character-level language model. It accurately recognizes the five key named entity kinds in biomedicine, namely Cell Lines, Chemicals, Diseases, Genes, and Species. HunFlair combines the knowledge by combining character-level LM pretraining and joint training on numerous gold standard corpora, which results in significant improvements over other cutting-edge commercial NER technologies. None of the training corpora for HunFlair include any appreciable textual overlap. Then they contrast HunFlair's tagging accuracy with that of two different rivals. Two types of contemporary research prototypes are available, one being "off-the-shelf" biomedical NER tools. All rivals are outperformed by HunFlair.

According to a recent paper [10], writer focused on the detection bengali named entity. They collected 500+ standards bengali sentences from various renowned Bangladeshi newspapers to detect named entity and classify them in standard category and types. They also tested whether a automatic detection of named entity(NE) is possible or not. In this paper author used Support Vector Machine(SVM) method with simple feature pre processing and Long Short Term Memory(LSTM) method with different word embedding. They got 0.72 to 0.84 F-measure result by using these methods. They also attempted to create a qualitative corpus in Bengali to enhance the resource of Bengali Language to detect NER easily.

In a recent paper [2], the authors have introduced a great resource - a Bengali biomedical named entity annotated corpus which is created from several Bengali health related articles extracted from Bangla Prothom Alo newspaper. As our Bengali language has very low resources of named entity, it is created for biomedical text mining which is in standard IOB format. Writers manually annotated the corpus in four different classes of entity. The table below shows a detailed overview of different Bangla NER related papers.

Paper	Year	Model	Dataset	Training Size	Tagset	Results
Sazzed et el. [2]	2022	Transfer based models	Bangla Newspaper Health Articles	11196 words	4	63.37%
Weber et el. [9]	2021	HunFlair a NER tagger	24 million abstracts of PubMed's articles			
Rashid et el. [10]	2021	SVM and LSTM	Few popular Bengali newspapers	503 target sentences		F1-72% to 84%
Karim et el. [1]	2019	DCN-BiLSTM	Wikipedia and 3 Bangla Newspapers	71,284 tokenized sentences	4	P-68.95% R-58.62% F1-63.37%
Banik et el. [8]	2018	GRU	Bangla newspaper 'Prothom Alo'	420 articles		F1-69 %

III. Issues with Bengali NER

Every language has its own beauty of structure and many challenges. Bengali language has its own structure, grammar etc which are different than other languages. As other languages Bengali language has some unique challenges(Ekbal and Bangapaddhay 2010) [11]. Some are given below-

- Bengali Language lacks capitalization, which makes NER easier to spot in English. A capital term, for instance, makes it clear what kind of word it is.
- Additionally, Bengali language NER is tough to identify when a statement has several meanings. Items like competent POS taggers, gazetteers, or annotated corpora are quite scarce in the Bengali language.
- A sentence in Bengali can have a variety of structures while yet having the same meaning.
- It is similarly problematic to categorize named entity types for multi-word expressions.
- The advancement of technology in NLP is still insufficient.
- Another ambiguity concern in Bengali is the enormous variety of names.

IV. Methods

A. Data Acquisition

For training the proposed model (DCN-BiLSTM) [1], 8,385,616 Bangla sentences from Bangla Wikipedia and 3 Bangla newspaper i.e. Ittefaq, Bangladesh Pratidin and Kaler Knatho were collected using python framework scrapy. Garbage texts and texts of other languages were filtered out and 8,154,503 sentences were finalized. The sententeces were tokenized into words using python library spacy [1].

B. Corpus Annotation

IOB tagging format was implemented and 4 entity tags were used to annotate the dataset i.e. person(PER), location(LOC), organisation(ORG) and object(OBJ). Besides, in another paper, at first, all the annotators got separated into two sets whether the sentences has any NE or not. After making these two sets, they identified each named entity by its type and the

position number of the tokens. After getting approximately 10% data annotated, it was checked that the labeling shows right or wrong. Then all the sentences get annotated by these two sets of annotators with high agreements. Cohen's coefficients for these agreements are fewer than 90, which is a very high number. A number of 0.8 or higher indicates almost complete agreement, whereas a value of 0.6 to 0.8 indicates strong agreement (Artstein and Poesio 2008). A second round of review resulted in the selection of the top annotations. Other systems employed a supervised approach for NER, therefore to annotate the data they had collected, two separate annotators were used. Four categories from the numerous fundamental NE that are usually encountered in online Bangla publications have been taken into account. They are Person (the news story's focus), Location (any specific instance of a place, such as a country, city, or state), Organization (having a unique identity from both domestic and foreign perspectives), and Day (a frequent news articles). It is recommended that annotators tag NEs using the IOB format, where B designates whether a tag represents the start of a NE, I designates whether a tag is inside, and O designates if a tag is outside.

C. Word Embedding

To get better accuracy, word embeddings were used instead of normal encoding. Both Word2vec and GloVe models were implemented in training the annotated corpus with 50, 100, 150 and 200 dimensions. In order to learn with neural networks, some people employ two LSTM-based topologies. The different embedding types employed account for the disparity. The first layer in the initial configuration is the embedding layer of Keras. The layer works with input documents that have 50 words each, a vocabulary of 3270 words (unique terms in our dataset), and a vector space of 300 dimensions in which words will be embedded. 50 vectors, each with 300 dimensions, will be the output of the embedding layer, one for each word. They then send the results of the LSTM network (which has 100 hidden nodes and 0.2 dropout) to the Dense output layer. We employ the sigmoid activation function for the dense layer.

D. DCN-BiLSTM Model

2 core features were applied for training the corpus. Firstly, Bi-LSTM was used for labelling the word sequences. To avoid the vanishing gradient problem of Recurrent Neural Network(RNN), Bi-LSTM was used. Secondly, using DCN, character embedding was done to extract character-level characteristics. [1].

E. SVM and LSTM

On next SVM classifiers with various feature combinations and an LSTM classifier are built on a neural network to categorize the phrases into groups indicating whether or not they contained NEs. RBF kernel and SVM classifiers are also used to determine whether a text contained NE or not. The implementation in scikit-learn (Pedregosa et al. 2011) to train

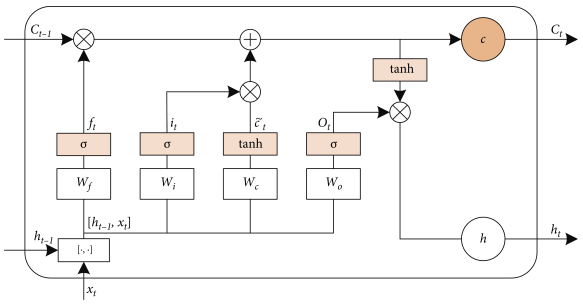

Fig. 1. Pictorial representation of DCN BiLSTM

the classifiers after stratifying the full corpus splits into 80-20 train and test. With the train split, they used 10-fold cross-validation to fine-tune the SVM hyper parameters (C and γ). Following that, the classifier predicted the two labels 1 (at least 1 NE present) and 0 (no NE present).

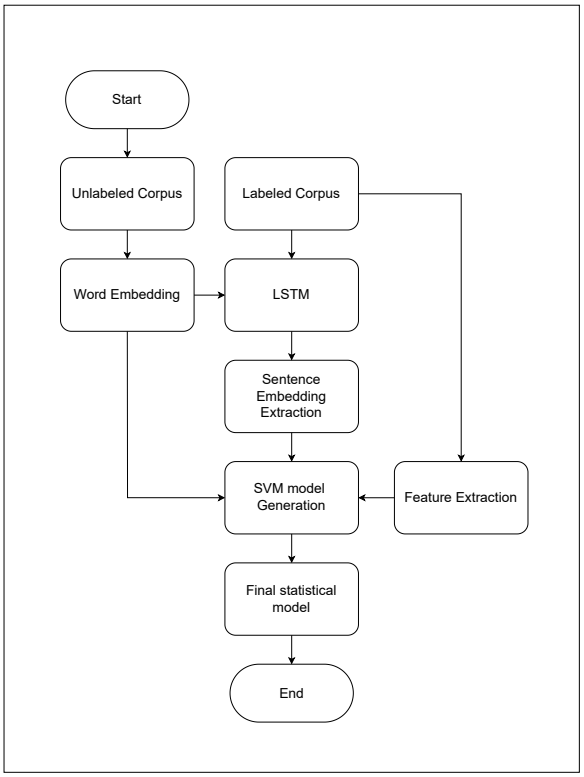

Fig. 2. Pictorial representation of SVM with LSTM

F. HunFlair

Another model, HunFlair [9], performs better than SciSpacy (Neumann et al., 2019), HUNER (Weber et al., 2020), tmChem (Leaman et al., 2015), GNormPlus (Wei et al., 2015), and DNorm (Leaman et al., 2013). HunFlair is simply a rearranged design model of HUNER that is based

on a pre-domain specific character-level language model. [9]

G. GRU & LSTM

Additionally, Recurrent Neural Networks (RNN), a sequential data processor are another tool used in NLP. For detecting Named Entities, the Elman unit, a simple RNN with three layers, often works well. However, the vanishing gradient issue in this simple RNN makes it challenging to learn and tune the earlier layer. [8] To get around this limitation, many neural networks are used, including Long Short Term Memory (LSTM), Gated Recurrent Units (GRUs), and Residual Networks (ResNets). LSTM and GRU are two of them that are often used in NER and other NLP applications. To address a network failure, the Forget gate in the LSTM employs back propagation for an indefinite number of steps. Three gates make up an LSTM: input, forget, and output. Contrarily, GRU only has two gates (Update and Reset). But it also functions much like an LSTM. Similar performance is offered by both for NLP tasks. According to several research studies, GRU outperforms LSTM in terms of application. GRU is also more effective than LSTM.

H. Rule-based Approach

The rule-based approach employs a fixed state pattern matching technique that is manually designed. In general, it works like a regular expression matcher and tries to match against a list of words.

I. Machine Learning-based Approach(ML)

Machine learning is more widely used than rule-based approaches since it is less expensive to train and adapt. Numerous machine learning (ML) techniques are being used, including the Hidden Markov Model (HMM) [12], Conditional Random Fields (CRFs) [13], and Maximum Entropy [14]. Due to the fact that machine learning (ML) systems may be maintained in a variety of languages and are less expensive to maintain than rule-based systems, these techniques are widely popular.

J. Hybrid

The two approaches are combined to create hybrid models which are rule-based and machine learning technique. The output of rule-based approach is used as features and it is used with other language independent features in the ML based approach.

V. Result

Any combination of features outperforms the majority baseline in the SVM system (F-measure: 0.45). When the first word is used as a feature, the F-measure rises to 0.56, which is When the length range is included as a feature, the rise is furthered. The sentence's bag-of-words representations result in an F-measure of 0.68. It is advantageous to include the tfidf representations in the feature set because they produce 0.72 F-measure. Using a corpus of 10k words, a system was created that combines ML and Rule-based approaches for NER which produce 71.59% F1-score. The setup that employs the embedding layer provided by Keras provides a substantially higher F-measure of 0.84 compared to the NN setup that uses weights from a 300 dimensional pre-trained word vector for Bengali words. Besides, We have taken into account the F1- score (or f-measure), which is the harmonic mean of precision and recall, from a variety of evaluation measures.

$$F1score = 2 \times \frac{P \times R}{P + R} \tag{1}$$

Here R refers for recall while P stands for precision.

TABLE I
Outcomes of Different Models

Model	Precision	Recall	F1-Score
SVM (TFIDF)	72.12	72.09	72.10
DCN-BiLSTM	68.95	58.62	63.37
Rule based Approach	70.32	68.39	68.98
ML approach (HMM)	72.92	70.67	71.59
LSTM (Keras)	83.56	84.63	84.09

VI. Conclusion

To enhance our knowledge about Bengali NER tasks and Bengali NLP community, we presented a survey study of five Bengali NER tasks with literature review. We reviewed 30 papers covering years of research along with proposed approaches. We also discussed about the proposed models' advantages and disadvantages thoroughly. Our comparison study can help future researchers to use the said models in a best fitting way for various Bangla NLP tasks and help future researchers to contribute in enriching Bengali NER research.

References

[1] Karim, R., Islam, M., Simanto, S., Chowdhury, S., Roy, K., Neon, A., Hasan, M., Firoze, A. & Rahman, M. A step towards information extraction: Named entity recognition in Bangla using deep learning. *Journal Of Intelligent Fuzzy Systems*. pp. 1-13 (2019,7)

[2] Sazzed, S. BanglaBioMed: A Biomedical Named-Entity Annotated Corpus for Bangla (Bengali). *Proceedings Of The 21st Workshop On Biomedical Language Processing*. pp. 323-329 (2022,5), https://aclanthology.org/2022.bionlp-1.31

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
[8] Banik, N., Rahman, M. H. H. (2018, December). Gru based named entity recognition system for bangla online newspapers. In 2018 International Conference on Innovation in Engineering and Technology (ICIET) (pp. 1-6). IEEE.
[9] Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., Akbik, A. (2021). HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics, 37(17), 2792-2794.
[10] Rashid, F., Hamid, F. (2021, April). Detecting the Presence of Named Entities in Bengali: Corpus and Experiments. In The International FLAIRS Conference Proceedings (Vol. 34).
[11] Ekbal, A., and Bandyopadhyay, S. 2010. Named entity recognition using support vector machine: A language independent approach. International Journal of Electrical and Computer Engineering 4(3):589 – 604.
[12] D. M. Bike, R. L. Schwartz and R. M. Weischedel, ”An Algorithm that Learns What’s in a Name”, 1999.
[13] S. Song, N. Zhang and H. Huang, ”Named entity recognition based on conditional random fields”, Cluster Computing, 2017.
[14] 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.
[15] Asif Ekbal and Sivaji Bandyopadhyay. 2009. A conditional random field approach for named entity recognition in Bengali and Hindi. Linguistic Issues in Language Technology 2, 1 (2009), 1–44.