Towards Bangla Named Entity Recognition

Conference Paper · December 2018

DOI: 10.1109/ICCITECHN.2018.8631931

3 authors:

Shammur Absar Chowdhury
Qatar Computing Research Institute
46 PUBLICATIONS **219** CITATIONS
SEE PROFILE

Firoj Alam
Qatar Computing Research Institute
60 PUBLICATIONS **464** CITATIONS
SEE PROFILE

Naira Khan
Institute of Education and Research
12 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Towards Bangla Named Entity Recognition

Shammur Absar Chowdhury
University of Trento, Italy
shammur.chowdhury@unitn.it

Firoj Alam
QCRI, Qatar
fialam@hbku.edu.qa

Naira Khan
Dhaka University, Bangladesh
nairakhan@du.ac.bd

***Abstract*—Named Entity Recognition is one of the fundamental problems for Information Extraction and the task is to find the mentioned entities in text. Over the years there has been significant progress in Named Entity Recognition (NER) research for resource-rich languages such as English, Chinese, and Italian. Although, there are a number of studies for Bangla NER, however, most of these studies are conducted almost a decade ago and were focused on a single geographical location (i.e., India). Therefore, in this paper, we present a corpus annotated with seven named entities with a particular focus on Bangladeshi Bangla. It is a part of the development of the Bangla Content Annotation Bank (B-CAB). We also present baseline results, which can be useful for future research. For the baseline results, we employed word-level, POS, gazetteers and contextual features along with Conditional Random Fields (CRFs). Our study also includes the exploration of deep neural networks. Additionally, we investigated another large corpus from a different geographical location (i.e., India) and concluded on the importance of geographic-based NER for a language.**

***Index Terms*—Bangla, Named Entity Recognition, Sequence Labeling, CRF, Neural Network, LSTM**

I. Introduction

The term *Named Entity* was first addressed in the context of the Sixth Message Understanding Conference (MUC-6) [1], to define anything that is associated with a proper name. The MUC-6 Named Entity (NE) recognition task was focused on the automatic identification of names of people, organizations and geographic locations in a text. From then on *Named Entity Recognition* (NER) has become a fundamental step in most Information Extraction tasks. The goal is to identify and categorize proper names that are mentioned in texts. In addition to the named entities, other entity types have also received attention in the research community. It includes temporal expressions (e.g., time, date, duration), unit expressions (e.g., money, percent, speed, rate, age) and bioinformatics (e.g., genes, protein, chemical, disease, DNA, RNA) [2]–[4].

As NER became one of the preliminary tasks in various application scenarios such as Information Extraction, Conversation Summarization, and Topic Detection, a significant amount of effort has gone into developing automatic NER systems using data-driven machine-learning based approaches. Currently, NER systems are available for many languages including English, German, Chinese and Italian with open-source tools (e.g., Stanford NER tagger) as well as commercial applications. A typical NER system takes an unlabeled text as input and labels the entity mentioned with entity types. The current research is also focused on domain-specific entity recognition tasks with noisy social media text [4], [5]. Although research on NER is not new, most of it has been centered on English.

A typical approach of developing an NER system is to use supervised machine learning techniques, which relies heavily on a manually labeled dataset. Such machine learning approaches typically use many hand-crafted features and knowledge sources in which a *word* is associated with a particular Named Entity. The most widely used machine learning algorithms are Hidden Markov Models (HMMs) [6], Conditional Random Fields (CRFs) [7], Maximum Entropy (ME) [8], Maximum Entropy Markov Models (MEMMs) [9], Support Vector Machines (SVMs) [10], [11], and hybrid approaches [12]. Hand-crafted features are mainly orthographic features such as prefix, suffix, the word in context, abbreviation, Part-of-Speech (POS). In addition, a task-specific knowledge source such as gazetteers has also been widely used.

There has been significant progress in the use of machine learning algorithms, i.e. a deep neural network for solving problems in the area of NLP, speech processing and computer vision [13]. The most popular deep learning algorithms include Long Short Term Memory (LSTM) neural networks [14] and Gated Recurrent Unit (GRU) [15] and a combination of LSTM, Convolution Neural Networks (CNN) [16] and CRFs [17]–[19]. Typically, these algorithms are used with distributed words and character representations, also called "word embeddings" and "character embeddings", respectively. The success of both word- and character- level embeddings are reported in many literatures [18]–[21]. It has been reported in the literature that character-level representations can capture word-shape, morphological and orthographic information, which are important for POS tagging, named entity recognition, lemmatization, and syntactic parsing. These word- and character-level embeddings are typically used in order to avoid feature engineering. However, some studies have also used them in combination [22].

As mentioned earlier most of the technological advance has been done for resource-rich languages. In comparison, very little attention has been given to under-resourced languages such as Bangla. The reported studies for Bangla NER include Ekbal et al. [23]–[30] and [31]. We discuss these studies in further detail in Section II. The focus of these studies was primarily targeted on the geographical context of India. The text corpus they have collected for annotating and training

978-1-5386-9242-4/18/$31.00

the model is from one of the leading newspapers from India. Another study by Hasanuzzaman et al. [32] used the data collected from a Bangladeshi newspaper. However, in their study the number of annotated entity types is very limited.

In this study, we aimed to widen the scope of Bangla NER to facilitate Information Extraction tasks and the main contributions in this study are as follows:

- We prepared an annotation guideline by following that of Automatic Content Extraction (ACE) [33] and MUC-6 [1]. It currently comprises seven entity types. Their associated subtypes are beyond the scope of this paper.
- We collected Bangladeshi newspaper articles and annotated these with the entity types.
- We prepared gazetteers for Bangla and annotated with *facility*, *organization*, *person* name and *BLACK list*[1] words.
- We present the baseline results, which can be useful for the research community.
- In addition, we investigated another large corpus from a different geographical location (i.e., India) and explored the importance of geography-based NER for a language.

The structure of this paper is as follows. In Section II, we provide a brief overview of the related work on Bangla NER. Followed by Section III, where we discuss the corpus collection, annotation process and how we used it in this study. We present our NER sequence labeling system in Section IV. In Section V, we discuss the results. Finally, we present the conclusions in Section VI.

II. Related Work

The work related to Bangla NER is relatively sparse. The current state-of-the-art for Bangla NER shows that most of the work has been done by Ekbal et al. [23]–[30] and IJCNLP-08 NER Shared Task [34]. Studies by Ekbal et al. comprise the NER corpus development, feature engineering, the use of HMMs, SVMs, ME, CRFs and a combination of classifiers. The reported F1 measure varies from 82% to 91% across a corpus with different number of entity types. The study of Chaudhuri et al. [35] uses a hybrid approach, which includes a dictionary, and rule and n-gram based statistical modeling.

The study in [36] focused on the geographical context of Bangladesh, as they collected and used data from one of the Bangladeshi Newspapers, namely Prothom-Alo [32]. In their study, only three entity types (i.e., tags) are annotated such as *person, location* and *organization*. The reported accuracy of their study is F1 71.99%. Banerjee et al. proposed the Margin Infused Relaxed Algorithm for NER, where they used the IJCNLP-08 NERSSEAL dataset for the experiment [31]. In [37], Ibtehaz et al. reported a partial string matching technique for Bangla NER.

When compared to the studies mentioned above, our study is focused on the development of a new NER corpus with seven different entity types. We also present baseline results using this dataset.

[1]We refereed the stop words and some other tokens as a BLACK list, which can never be an entity.

III. Corpus

Since there is no publicly available NER dataset for Bangladeshi Bangla, therefore, we developed our own corpus, which is a part of the development of the Bangla Content Annotation Bank (B-CAB)[2]. Thus, from here forward it is referred to as B-CAB Named Entity corpus. The text for the corpus has been collected from different popular newspapers in Bangladesh (e.g., Prothom-Alo – https://prothomalo.com). For this study, we used a total of 35 news articles, with topics ranging from politics, sports and entertainment. The extracted texts were then tokenized such that no punctuation, for instance ',' is attached to the word ("প্রাথমিক, মাধ্যমিক স্কুল" $\rightarrow$ "প্রাথমিক , মাধ্যমিক স্কুল") or no $< Cntrl >$ markers among others are present in the text. The cleaned corpus of $\approx 35K$ words contains 2137 sentences and a vocabulary of size $|V| \approx 10K$.

For the annotation, we prepared an annotation scheme by following the guideline of ACE from the Linguistic Data Consortium[3] [33] and MUC-6 [1]. For the entity annotation, we identified seven entity types and their subtypes. The discussion of subtypes is beyond the scope of this work. The annotated entity types include the following:

- **Person (PER):** Person entities are only defined for humans. A person entity can be a single individual or a group. For example, মাহবুবউল আলম, ইঞ্জিনিয়ার, ডাক্তার, সাংবাদিক, প্রেসিডেন্ট, সভাপতি
- **Location (LOC):** Location entities are defined as geographical entities, which include geographical areas and landmasses, bodies of water, and geological formations. For example, মতিজিল, ঢাকা, ইউরোপ, চট্টগ্রাম, ঢাকা উত্তর, দক্ষিণ ডেন্ডাবর
- **Organization (ORG):** Organization entities are defined by corporations, agencies, and other groups of people. For example, মন্ত্রণালয়, কোর্ট, থানা, বিএনপি, আওয়ামী লিগ, আর্মি, নেভি, গ্রামীণ, বিসিসি, সোনালি ব্যাংক, ঢাকা বিশ্ববিদ্যালয়
- **Facility (FAC):** Facility entities are defined as buildings and other permanent human-made structures. For example, বাংলাদেশ বিমান বন্দর, চামড়া পক্রিয়াকরন কারখানা, হোটেল, স্টেডিয়াম, মিউজিয়াম, জেলখানা, গ্যারেজ, স্টোরেজ, ঘর, বিল্ডিং, রাস্তা, বন্দর, ব্রিজ
- **Time (TIME):** Mentions that represent absolute date and times. It includes duration, days of week, month, year, and time of the day. For example, সকাল, বিকাল, দুপুর, রাত, সময় ১২ টা, ১০/২/২০১৮
- **Units (UNITS):** Units are mentions that include money, number, rate, and age. For example, টাকা, প্রতি ঘণ্টায় ১০ মাইল
- **Misc (MISC):** Any entity that does not fit into the above entities.

[2]https://github.com/Bangla-Language-Processing/Bangla-Content-Annotation-Bank

[3]https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

Instructions have been provided to the annotators to consider the inherent meaning of the text and use contextual knowledge of the text during annotation. An example of an annotated sentence is provided in Figure 1.

For measuring the quality of the annotation, we used four newspaper articles annotated by three annotators. The obtained inter-annotator agreement using Cohen's κ [38], a well known agreement measure, is $(\kappa = 0.72)$.

Once the annotation is complete we then convert the data into the IOB2 format where each token is defined with a certain entity type. It represents whether the token is either inside an entity (defined as "I-"), or at the beginning of an entity (defined as "B-"), or outside the entity (defined as "O").

The following example illustrates the outcome of this conversion process. [দৈনিক$_{B-ORG}$] [ইত্তেফাক$_{I-ORG}$] [ও$_O$] [হাউজিং$_{B-ORG}$] [এন্ড$_{I-ORG}$] [বিল্ডিং$_{I-ORG}$] [রিসার্চ$_{I-ORG}$] [ইনস্টিটিউটের$_{I-ORG}$] [(এইচবিআরআই)$_{I-ORG}$] [যৌথ$_O$] [উদ্যোগে$_O$] [গতকাল$_{B-TIME}$] [রবিবার$_{I-TIME}$] [কাওরানবাজারস্থ$_{B-LOC}$] [ইত্তেফাক$_{B-FAC}$] [কার্যালয়ের$_{I-FAC}$] [মজিদা$_{B-FAC}$] [বেগম$_{I-FAC}$] [মিলনায়তনে$_{I-FAC}$] [এ$_O$] [গোলটেবিল$_O$] [অনুষ্ঠিত$_O$] [হয়$_O$] [।$_O$].

In Figure 2, we present the entity type distribution associated with IOB2 tags. It shows us that more than 50% of the textual content is non-entity mentions tagged as O, which is typical for any named entity corpus. Among the entity types *person* type entities are higher. In Table I, we provide token level statistics for each entity types. On average two to three tokens per entity mentions for each entity type. We also observed that in some cases the number of tokens reached from ten to fifteen due to the fact that title and subtitle are associated with person names. Such entity mentions pose various challenges for the automated recognition system.

Entity types	Avg.	Std.
FACILITY	2.5	1.7
LOCATION	1.6	1.3
MISC	2.0	1.4
ORGANISATION	2.5	1.4
PERSON	2.9	1.9
TIME	2.3	1.3
UNITS	2.5	2.1

Table I: Statistics with the number of token in entity mentions of each entity type.

In order to train the model, we divide the data into training, development and test split. In Table II, we provide the statistics of the dataset with different splits. For the data split, we maintained $\sim$ 70%, $\sim$ 10% and $\sim$ 20% of the tokens, for the training, development and test set, respectively. In Figure 3, we present the distribution of the entity types that are present in the dataset. From the figure, we observe that entity type distribution across dataset is representative for the machine learning experiments.

In addition to the B-CAB dataset, we also experimented with NER dataset that the IJCNLP 2008 NER workshop (NERSSEAL-08 shared task) organizer published. The detail of this corpus can be found in [34]. Due to the considerable difference with the two tagsets, we could not directly use it, therefore, we mapped their tagset to be aligned with our corpus. For example, we mapped their tag "Number" with our tag "Units".

Metric	Train	Dev	Test
# of sentences	1510	200	427
Total	24377	2636	6546
Average	16.14	13.18	15.33
Std. deviation	9.73	7.83	13.45
Max	98	53	85

Table II: Preliminary statistics of the annotated dataset for the training, development and test data split. First row represents total number of sentences in each set. Row 2-5 represents total average, standard deviation, and maximum number of token in each set

IV. Experiments

For the experiment, we conducted baseline studies and designed the models using CRFs, in which we explored features such as Part-of-Speech (POS) labels, gazetteers and word embeddings.

A. Features

For our experiments we used three different kinds of features (i) *word level features* (e.g., token, prefix, suffix), (ii) POS, (iii) list lookup features (e.g., gazetteers), and (iv) word-embedding features.

1) Word level: The *word-level features* include the token, character n-grams with prefixes and suffixes and digit patterns. Character n-grams provide morphological information of the word.

2) POS: In order to develop a POS tagger, we employed distributed representations of the words and characters. We designed the system using LSTMs as the intermediate layers and CRFs as the final layer for designing the model. In addition, we used the pre-trained word embeddings model for the initialization of the network. We used the architecture that has been discussed in [39]. As for the dataset, we used the POS tagged corpus that is publicly available and discussed in [40]. The corpus consists of 30 POS tags [41]. The performance of the POS tagger is F1=85.0%. We used this model to extract POS tag for our NER task.

3) List lookup features: The *list lookup features* include dictionaries, gazetteers, list of peoples names, organization names, word list consists of verbs and stop words. A gazetteer is typically defined as a dictionary that includes a list of place names along with geographical information. In order to match the gazetteers, we used an efficient string matching algorithm (Rabin-Krap) [42].

Our development of gazetteers includes names of countries, major cities, common names among others. Since there is no such resource available, therefore, one of the main contributions of this study is to prepare such a resource for the NER system. For building the gazetteer lists, we crawled different websites and collected the names of different location names, including - districts, divisions, sectors among others. Similarly, we also created a list of organizations (international and in

ORGANISATION দৈনিক ইত্তেফাক ও ORGANISATION হাউজিং এন্ড বিল্ডিং রিসার্চ ইনস্টিটিউটের (এইচবিআরআই) যৌথ উদ্যোগে TIME গতকাল রবিবার LOCATION কাওরানবাজারস্থ FACILITY ইত্তেফাক কার্যালয়ের FACILITY মজিদা বেগম মিলনায়তনে এ গোলটেবিল অনুষ্ঠিত হয়।

Figure 1: An annotation example with different entity types using our deployed collaborative annotation platform.

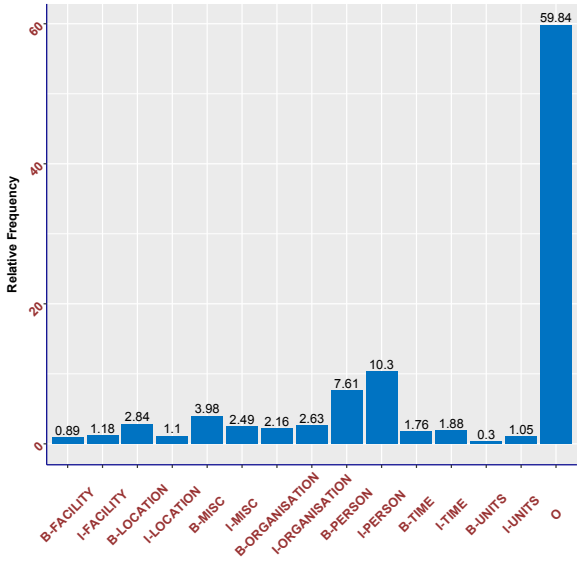

Figure 2: Entity type with IOB2 tag distribution.

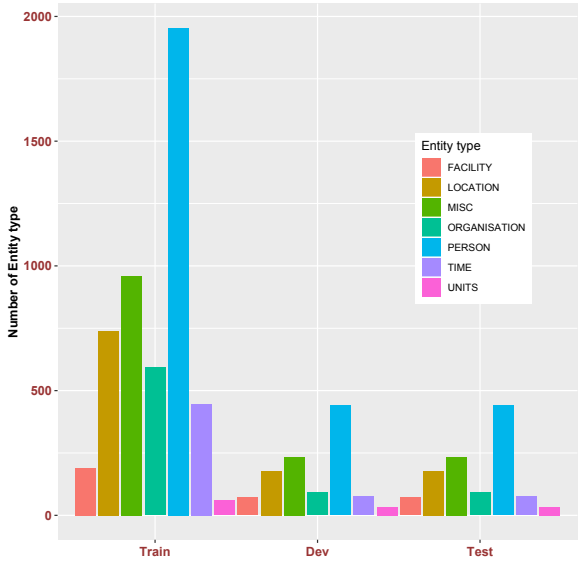

Figure 3: Entity type distribution with training, development and test split.

Bangladesh), person names and proper nouns. We also created a list termed as *black-list*, which includes verbs among others. We collected the entries in *black-list* from the lexicon used in the studies of [43], [44]. These lists have then been cleaned and corrected manually. This is also an ongoing project and falls under B-CAB.

4) Word Embeddings: To use word embedding features we used the publicly available model[4] by Alam et al. [39].

[4]https://github.com/cogniinsight/Word-embedding-model-for-Bangla

B. Classification Algorithms

1) CRFs Model: For designing the sequence classification model, we used CRFs [7] – a popular probabilistic graphical model that exploits the dependency structure. In this study, the contextual dependency (i.e., previous and following context) is being exploited by the CRFs. For a given word sequence $\boldsymbol{W} = \{w_0, w_1, ...w_i, ...w_T\}$ it tries to find the best segmentation along with the output labels $\boldsymbol{SL} = \{l_0, l_1, ...l_i, ...l_N\}$ by modeling conditional distribution $P(\boldsymbol{SL}|\boldsymbol{W})$. Each w_i is basically a vector containing information about the word such as its identity, preceding and succeeding words, gazetteer feature and so on, and each l_j represents its output information for the segmentation and label sequence. The best output sequence is obtained by computing the maximum probability as defined by the Equation 1.

$$\boldsymbol{SL}* = argmax_{SL} P(\boldsymbol{SL}|\boldsymbol{W}) \tag{1}$$

The probability of the output sequence is modeled by the learned parameters and feature functions as shown in Equation 2.

$$P(\boldsymbol{SL}|\boldsymbol{W}, \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{W})} exp\left(\sum_{i=1}^{T}\sum_{j}\lambda_j f_j(l_{i-1}, l_i, w_i)\right) \tag{2}$$

where $f_j(l_{i-1}, l_i, x, i)$ is the feature function and $\boldsymbol{\lambda}$ is the learned parameter. Here, for the sake of simplicity, the feature function considers the current (l_i) and one previous (l_{i-1}) output segment-label, and the current word w_i. The observation dependent normalization $Z(\boldsymbol{W})$ is defined by the equation 3. It ensures the distribution of P sums to 1. The feature function can be designed many different ways. One example is the boolean representation, i.e., presence or absence of a characteristic for a word w_i.

$$Z(\boldsymbol{W}) = \sum_{SL \in \boldsymbol{SL}} \sum_{i=1}^{T}\sum_{j}\lambda_j f_j(l_{i-1}, l_i, w_i) \tag{3}$$

The parameter $\boldsymbol{\lambda}$ is learned using the maximum likelihood estimation from the training data, $D = [(\boldsymbol{W}^1, \boldsymbol{SL}^1), (\boldsymbol{W}^2, \boldsymbol{SL}^2), ..., (\boldsymbol{W}^m, \boldsymbol{SL}^m)]$, that we provide during training. The mathematical details of the parameter estimation can be found in [7], [45].

2) DNN Model: To explore the power of neural networks and to check whether neural networks perform well we designed the sequence to sequence model using the bidirectional LSTMs - CRFs [39]. In this architecture, we exploited word- and character- level embeddings along with POS tags and gazetteers as features. For training the model, we used bidirectional LSTMs followed by a CRFs layer.

C. Evaluation

For the NER system, the evaluation metrics measure the ability of the model to find the boundaries of name-entity and their correct label. Therefore, it includes the performance of both segmentation and labeling of the entity. For such tasks, most evaluation techniques measure the exact match – on both boundary and segment label. However, in some cases, such as NER or semantic annotations [46], [47], the exact boundary detection is not so important as long as the major part of the name has been identified. For example, the segment, [যিনি$_{B-PER}$] [মৃত$_{I-PER}$] [মানুষের$_{I-PER}$] [ভুয়া$_{O}$] [সার্টিফিকেট$_{O}$] can be recognized as – [যিনি$_{B-PER}$] [মৃত$_{O}$] [মানুষের$_{B-PER}$] [ভুয়া$_{O}$] [সার্টিফিকেট$_{O}$].

The above two examples are almost the same except with the minor differences in the adjective 'মৃত', which is not an entity but a modifier. Therefore, another matching protocol can be used (as adopted in [1]), which is based on loose matching conditions. It allows credits to any correct entity mention detected regardless of its boundary as long as there is an overlap with the gold annotation.

Therefore, for this study, we used two matching protocols:

- **Exact** matching for both boundary and entity mention (widely used in CONLL shared tasks)
- **Partial** matching for the boundary for which the detected entity mention is correct only.

Moreover, for each protocol, we calculated the precision, recall and its corresponding F-measure as our evaluation measure.

V. Results and Discussion

In Table III, we present the performance of our system on the test set, which also includes baseline results. We obtained the baseline results using the token as a feature and CRFs to train the model. As can be seen in the table, there is a significant improvement in performance (exact match - 19%; partial match - 16%) in terms of F-measure, for both matching protocols, while using the POS and gazetteers along with word-level features and CRFs architecture with our training dataset.

We also used bidirectional LSTMs-CRFs architecture to explore how a neural model performs. We observed that the result of our DNN architecture is significantly better than the baseline, however, it could not outperform our CRFs-based model with additional features, $E2\ (O)$. This pattern has also been seen in literature [22], the deep neural network could not outperform CRFs based model.

In order to understand whether data collected in a different geographic location could be useful, we combined IJCNLP-08 NER shared task dataset with our training set. As mentioned earlier, we mapped the tags of this dataset with the B-CAB dataset to be compatible. Our experimental results show that it reduces the performance significantly. To get more insights we have done an analysis to understand the differences. In Table IV, we report top entity mentions of type location of the two datasets. It is clearly visible that two datasets represent the location of a different geographic location. Another phenomena that we observed that in many cases the spelling is different in two datasets. For example, the word বাঙলার is present in IJCNLP, whereas it is বাংলার in B-CAB corpus. Another example is রঙপুর *vs* রংপুর.

EXP	Exact			Partial		
	P	R	F1	P	R	F1
E1 (B-CAB)	0.48	0.34	0.40	0.66	0.48	0.56
E2 (B-CAB)	0.65	0.53	**0.58**	0.78	0.67	**0.72**
E3 (B-CAB)	0.56	0.56	0.56	0.67	0.66	0.67
E2 (B-CAB+IJCNLP 08)	0.69	0.39	0.50	0.82	0.51	0.63

Table III: Results on B-CAB test set. E1 represents the baseline results with CRFs architecture and token (T) features, E2 presents the results from CRFs model with T, POS (P) and gazetters (G) as features. E3 presents results using DNN architecture and word embedding (W), P and G as feature. (.) represents the training data used for training the model.

Our corpus		IJCNLP corpus	
বাংলাদেশ	21	বাঁকুড়ার	11
ঢাকা	20	পাঁচমুড়ার	10
পাকিস্তানের	17	কৃষ্ণনগরের	6
জেরুজালেম	13	থ্যাকারে ম্যানসনের	6
সারাদেশে	13	উত্তরবঙ্গের	5
বাংলাদেশের	13	কাশীতে	4
টাঙ্গাইল	10	ভারতের	4
বরিশাল	10	পশ্চিমবঙ্গে	4
রংপুর	10	বাঙলার	3

Table IV: Location type entity mentions in two corpora.

Simultaneously, to better understand our results we have done some error analysis. We realized that in many cases entity mentions are confusing with one another. For example, ঢাকা is a location mention, whereas it also represents as O tag when appears as a verb.

VI. Conclusions

In this paper, we present a Bangla Named Entity Corpus, named as B-CAB Named Entity Corpus, consists of seven entity types. For which, we developed an annotation guideline for the annotation. We then annotated newspaper articles collected from different newspapers. Using the annotated dataset we have done our preliminary experiments and present the baseline results. We have also done an analysis by using IJCNLP 08 named entity corpus to understand whether that could be useful. We realized that as the data collected from different geographical location there are considerable differences in the dataset, which is limiting us to use them. In order to develop a general NER, we believe it is necessary to develop a corpus that covers a wide variety, geographical location, and domains. This is what we aim to address in the future. In addition, we aim to enrich the B-CAB corpus with more annotated data and explore with deep neural networks.

Acknowledgment

The research leading to these results has received funding from Cognitive Insight Limited – https://cogniinsight.com/.

References

[1] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history." in *COLING*, vol. 96, 1996, pp. 466–471.

[2] F. Alam, B. Magnini, and R. Zanoli, *Comparing Named Entity Recognition on Transcriptions and Written Texts, Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*. Springer, 2015, vol. 589.

[3] F. Alam, A. Corazza, A. Lavelli, and R. Zanoli, "A knowledge-poor approach to biocreative v dner and cid tasks," in *In Proc. of the Fifth BioCreative Challenge Evaluation Workshop*, 2015, pp. 274–279.

[4] M. Eckert, L. Clark, H. Lind, J. Kessler, and N. Nicolov, "Structural sentiment and entity annotation guidelines," *JD Power and Associates.*, 2008.

[5] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in twitter data with crowdsourcing," in *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. ACL, 2010, pp. 80–88.

[6] T. Brants, "Tnt: a statistical part-of-speech tagger," in *Proc. of the sixth conference on Applied natural language processing*. ACL, 2000, pp. 224–231.

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. of 8th ICML*, vol. 1, 2001, pp. 282–289.

[8] A. Ratnaparkhi *et al.*, "A maximum entropy model for part-of-speech tagging," in *Proc. of EMNLP*, vol. 1. Philadelphia, USA, 1996, pp. 133–142.

[9] A. McCallum, D. Freitag, and F. C. Pereira, "Maximum entropy markov models for information extraction and segmentation." in *Proc. of ICML*, vol. 17, 2000, pp. 591–598.

[10] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," in *Proc. of NACL*. ACL, 2001, pp. 1–8.

[11] H. Yamada and Y. Matsumoto, "Statistical dependency analysis with support vector machines," in *Proc. of IWPT*, vol. 3, 2003, pp. 195–206.

[12] F. Alam, "Evalita 2011: Named entity recognition on transcription using cascaded classifiers," in *Working Notes of EVALITA 2011*, 2012.

[13] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *JMLR*, vol. 11, no. Feb, pp. 625–660, 2010.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[17] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[18] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *arXiv preprint arXiv:1603.01354*, 2016.

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[20] C. N. dos Santos and B. Zadrozny, "Learning character-level representations for part-of-speech tagging." in *ICML*, 2014, pp. 1818–1826.

[21] W. Ling, T. Luís, L. Marujo, R. F. Astudillo, S. Amir, C. Dyer, A. W. Black, and I. Trancoso, "Finding function in form: Compositional character models for open vocabulary word representation," *arXiv preprint arXiv:1508.02096*, 2015.

[22] D. Bonadiman, A. Severyn, and A. Moschitti, "Deep neural networks for named entity recognition in italian," *CLiC it*, p. 51, 2015.

[23] A. Ekbal and S. Bandyopadhyay, "Named entity recognition using support vector machine: A language independent approach," *International Journal of Electrical, Computer, and Systems Engineering*, vol. 4, no. 2, pp. 155–170, 2010.

[24] A. Ekbal and S. Bandyopadhyay, "Bengali named entity recognition using classifier combination," in *ICAPR*. IEEE, 2009, pp. 259–262.

[25] A. Ekbal and S. Bandyopadhyay, "Named entity recognition in bengali: A multi-engine approach," *Northern European Journal of Language Technology*, vol. 1, no. 2, pp. 26–58, 2009.

[26] A. Ekbal and S. Bandyopadhyay, "A web-based bengali news corpus for named entity recognition," *Language Resources and Evaluation*, vol. 42, no. 2, pp. 173–182, 2008.

[27] A. Ekbal and S. Bandyopadhyay, "Development of bengali named entity tagged corpus and its use in ner systems," in *Proc. of the 6th Workshop on Asian Language Resources*, 2008.

[28] A. Ekbal and S. Bandyopadhyay, "Bengali named entity recognition using support vector machine," in *Proc. of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[29] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Named entity recognition in bengali: A conditional random field approach," in *Proc. of the 3rd Joint Conference on NLP*, 2008.

[30] A. Ekbal and S. Bandyopadhyay, "A hidden markov model based named entity recognition system: Bengali and hindi as case studies," in *International Conference on PRML*. Springer, 2007, pp. 545–552.

[31] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, "Bengali named entity recognition using margin infused relaxed algorithm," in *International Conference on TSD*. Springer, 2014, pp. 125–132.

[32] 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*, vol. 1, no. 1, p. 408, 2009.

[33] L. D. Consortium *et al.*, "Ace (automatic content extraction) english annotation guidelines for entities," *Version*, vol. 5, no. 6, pp. 2005–08, 2005.

[34] A. K. Singh, "Named entity recognition for south and south east asian languages: taking stock," in *Proc. of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[35] B. B. Chaudhuri and S. Bhattacharya, "An experiment on automatic detection of named entities in bangla," in *Proc. of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[36] K. S. Hasan, V. Ng *et al.*, "Learning-based named entity recognition for morphologically-rich, resource-scarce languages," in *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 354–362.

[37] 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*, 2018.

[38] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[39] F. Alam, S. A. Chowdhury, and S. R. H. Noori, "Bidirectional lstms—crfs networks for bangla pos tagging," in *Proc. of 19th ICCIT*. IEEE, 2016, pp. 377–382.

[40] M. C. Kalika Bali and P. Biswas, "Indian language part-of-speech tagset: Bengali ldc2010t16," Philadelphia: Linguistic Data Consortium, Tech. Rep., 2010.

[41] S. Baskaran, K. Bali, T. Bhattacharya, P. Bhattacharyya, G. N. Jha *et al.*, "A common parts-of-speech tagset framework for indian languages," in *Proc. of LREC 2008*. Citeseer, 2008.

[42] R. M. Karp and M. O. Rabin, "Efficient randomized pattern-matching algorithms," *IBM journal of research and development*, vol. 31, no. 2, pp. 249–260, 1987.

[43] F. Alam, S. M. Habib, and M. Khan, "Bangla text to speech using festival," in *Conference on Human Language Technology for Development*, 2011, pp. 154–161.

[44] S. A. Chowdhury, "Implementation of speech recognition system for bangla," Ph.D. dissertation, BRAC University, 2010.

[45] C. Sutton, A. McCallum *et al.*, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2012.

[46] S. A. Chowdhury, A. Ghosh, E. A. Stepanov, A. O. Bayer, G. Riccardi, and I. Klasinas, "Cross-language transfer of semantic annotation via targeted crowdsourcing," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[47] E. A. Stepanov, S. A. Chowdhury, A. O. Bayer, A. Ghosh, I. Klasinas, M. Calvo, E. Sanchis, and G. Riccardi, "Cross-language transfer of semantic annotation via targeted crowdsourcing: task design and evaluation," *Language Resources and Evaluation*, vol. 52, no. 1, pp. 341–364, 2018.