A Crowd-Source Based Corpus on Bangla to English Translation

Nafisa Nowshin
Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet, Bangladesh
nafisanowshin107@gmail.com

Zakia Sultana Ritu
Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet, Bangladesh
zakiaritu.cse@gmail.com

Sabir Ismail
Computer Science and Engineering
Stony Brook University
New York, United States
sabir.ismail@stonybrook.edu

Abstract—**In this paper, we present a crowd-source based Bangla to English parallel corpus and evaluate its accuracy. A complete and informative corpus is necessary for any language for its development through automated process. A Bangla to English parallel corpus has importance in various multi-lingual applications and NLP research works. But there is still scarcity of a complete Bangla to English parallel corpus. In this paper we propose a large scale crowd-source method of construction of a Bangla to English parallel corpus through crowd-sourcing. We chose crowd-sourcing method to venture a new approach in corpus construction and evaluate human behavior pattern in doing so. The translations were collected form under graduate students of university to ensure strong language knowledge. A Bangla to English parallel corpus will help in comparing linguistic features of these languages. In this paper we present an initial dataset prepared via crowd-sourcing which will serve as a baseline for further analysis of crowd source based corpus. Our primary dataset is consists of 517 Bangla sentences and for every Bangla sentence, we collected 4 English sentences on an average and 2143 English sentences in total via crowd-sourcing. This data was collected over a period of 2 months and from 62 users. Finally we analyze the dataset and give some conclusive idea about further research.**

Keywords*—*Natural Language Processing(NLP) , machine learning, corpus, crowd-source data, Bangla to English translation.

I. INTRODUCTION

A corpus is basically a collection of written texts or spoken material of a language that is processed to learn about that language's behavior. Construction of a corpus is the fundamental of most language research works. In this sector Bangla is still a little behind than other languages. There are a number of Bangla corpora available right now, like SUMono [1], BdNC01 [2]. But none of these corpora are constructed based on crowd-sourced data. So we wanted to implement crowd source method in constructing a Bangla to English parallel corpus. Crowd-sourced data means, the data or information obtained for a particular task or project by enlisting the services of a large number of people, typically via the Internet. So, what we proposed to do is, by the means of crowd-sourcing, collect the text data of Bangla to English translation and evaluate the data collected to understand the variations and structure of the data.

A parallel corpus is a corpus that contains collection of original texts of a language and their translation in a set of languages. In our case, the Bangla to English parallel corpus has Bangla text data and it's English translation. parallel corpora has many uses in various fields, like comparing linguistic features of two languages, investigating similarities and differences between the source and the target language, helps in translation studies and machine translation related researches. A Bangla to English parallel corpus will help us in all of these sectors and also as it is prepared via crowd-sourcing we also get the information of behavioral pattern of users while translating the sentences. This can help us further in machine translation researches regarding Bangla language and will help determine the behavioral pattern in that case.

Although crowd-sourced data is relatively new in NLP related research sectors, it is gaining popularity fast as training models for machine learning. This process helps get new insights and helps incorporate understanding of human behavior with regard to machine learning. For all these reasons the importance of crowd-sourced data is increasing rapidly. In this paper we have tried to propose a new method of parallel corpus construction applying this popular method.

This paper is arranged as follows, in section 2, we have shed light on the previous works regarding corpus construction and tried to give an overview of the present condition in this sector. In section 3 we have discussed our reasons for choosing this method. Then in section 4 we have discussed in detail the full methodology of our work with examples of our data. We showed an analysis of the data collected in section 5. We conclude in section 6.

II. BACKGROUND STUDY

Corpus construction is one of the most important part of any type of language research work. The strength of the digital presence of a language depends on the availability of that language's proper and complete corpus. So, much attention have been given in corpus construction in NLP sector. Bangla is no different, much work has been done, and many processes has been evaluated in constructing a complete Bangla corpus. We will discuss some of these works below.

978-1-5386-9242-4/18/$31.00 ©2018 IEEE

The process of constructing a Bangla corpus started long ago. Dash and Chaudhuri [3], constructed a small scale Bangla corpus along with 9 other Indian languages called the CIIL corpus. It consists of only 3 million words. Because of the small size of this corpus, it has failed to ensure its representativeness of Bangla language.

Automatic Bangla Corpus Creation was attempted by Sarkar, Pavel and Khan [4]. The process they followed was that they collected all free Bangla documents from the web with the help of a web crawler and collected available offline Bangla text documents. Then they extract all the words in these documents to make a huge repository of text and then converted them to unicode text.

Salam, Yamada and Nishino [5] proposed the first balanced corpus for Bangla language. They built the corpus depending on three independent criteria, time, domain and medium. As their goal was to construct a balanced corpus, they also added necessary additional details to the collected text like sample size, details of the author, topic etc. The source of their data was, literature text data, Bangla academic papers, Bangla text books, newspaper articles, TV and radio news scripts, Bangla technical manuals, Legal documents written in Bangla. We can see that to make the corpus representative and balanced they covered a wide range of text sources.

Mumin, Shoeb, Selim and Iqbal [1], constructed a new Bangla corpus named SUMono. This corpus consists of more than 27 million words. The SUMono corpus was constructed from available online and offline Bangla text data that includes articles from six types of topics. This corpus was constructed following the framework of the American National Corpus (ANC). SUMono corpus includes written texts from writers of various backgrounds, Bangla newspaper articles available online, Bangla text data from various websites etc. Because of the variety of the types of data available in this corpus, its representativeness of Bangla language has been ensured.

They also built a English-Bengali parallel corpus which is known as SUPara [6]. In building this corpus their main focus was to make it a balanced corpus. It contains variety of texts from different domains. They first converted the plain texts to unicode and then they were marked up according to corpus encoding standard. This corpus is open for educational and research purpose.

There also exists specified variations of Bangla corpus. A good example of which is the corpus named “Prothom-Alo”. [7] which is a corpus built solely with news articles published in a popular Bangla newspaper named “Prothom-Alo”, for the year 2005. They first collected the texts from the website of the newspaper. Then the text was extracted and categorized. Then they were converted to Unicode. But as this corpus consists of very specified data, it can not be used in many NLP research works.

Khan, Ferdousi and Sobhan [2] created a new Bangla text corpus named “BdNC01”. Text source for this corpus is, articles collected from web editions of several influential daily newspapers and literary works of old and modern writers. It contains nearly 12 million words. The text data for this corpus was collected over a time of 6 years to avoid time dependencies. After collection and processing of text data, it was added to the repository and statistical computations were done on it for better understanding of Bangla linguistic behavior.

Shamshed and Karim [8]. also proposed a method for Bangla text corpus creation. They proposed to use this corpus for Efficient Information Retrieval system. As they propose to use this corpus for information retrieval, all the text in their corpus are document specified. Their text source was Bangla books and Bangla web data. After collecting and formatting text data, they calculated term frequencies and then applied random walk algorithm on the data. Then they had to assemble the meta data.

Finally, we can say that there is rich literature growing on corpus construction techniques and there is much scope of improving this sector. Most research works discussed in this section has more or less same type of development process. They varied in their text source, their size and their collection of various topics to represent Bangla language. But the most important factor to be noted from this discussion is that none of these works involve crowd-sourced data. As a matter of fact, the process of constructing a text corpus using crowd-sourced data has not been attempted before. So, we are proposing a new process of corpus construction.

III. WHY CROWD-SOURCED CORPUS

There has been various approaches to parallel corpus construction process. They mostly focus on collecting the text document of one language from web pages or written text files and then converting them to unicode. Then they are marked up according to corpus encoding standard for XML and then aligned. But We tried a new approach, first we constructed the Bangla corpus containing simple and small Bangla sentences and then collected crowd-sourced data for the English translation of these sentences. This way we got more than one translated sentence for each Bangla sentence and could compare the output. This process also gives us insight on human behavior in case of translation of one language to another. The process is discussed in detail in the next section.

IV. METHODOLOGY

A. Data Preparation

In the first step, we focused on preparing the Bangla text data. The Bangla text data in our corpus consists of simple and small Bangla sentences, mostly with only one verb. We have worked with almost the same sentence pattern, the change in the sentence with the change of tense of verb and with the change of person of verb. This way we got many variations of one sentence. The reason behind doing this was to compare the result we get from crowd-sourcing and see their behavioral pattern with small change in sentences. Below is some of the examples of the sentences that is present in our corpus-

- আমি ভাত খাই।
- আমি ভাত খাচ্ছি।

- আমি ভাত খাচ্ছিলাম।
- আমি ভাত খাবো।
- বাবা বাজারে গেছেন।
- বাবা বাজারে যাবেন।
- বাবা বাজারে যাচ্ছেন।
- বাবা কি বাজারে গেছেন?
- কৃষক ক্ষেতে কাজ করতে যাচ্ছে।
- সে কি ঢাকা শহরে বাস করে?
- বৃষ্টি না হলে আমরা বাইরে যাব।
- খেলাধুলা স্বাস্থের জন্যে উপকারী।

For preparing this text corpus we went through some Bangla to English translation books. We prepared the corpus by taking help from school level English grammar books [9]. These books cover Bangla to English translation and grammar structures. As can be seen from the example of the sentences above, we tried to cover assertive, interrogative, negative, conditional and imperative type of sentences. We also tried to focus on the variations of gender, tense, person of the same sentence. The details regarding the Bangla part of the corpus is given in table I.

Table I: DETAILS OF BANGLA PART OF THE CORPUS

Total sentences	517
Total words	2352
Average sentence length	5 words

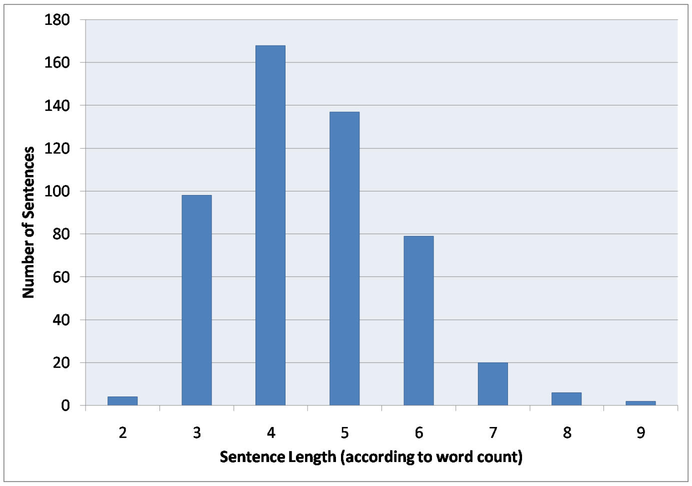
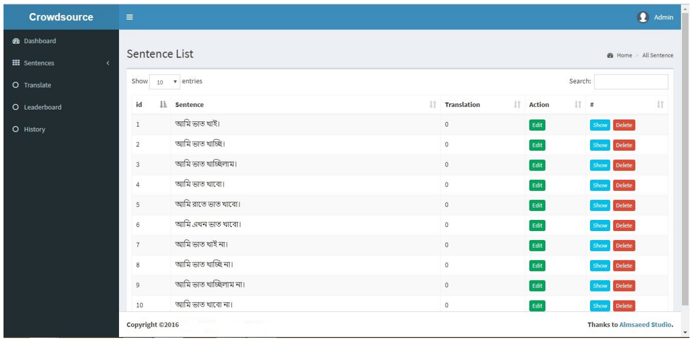


Fig. 1. Statistics of the Bangla Sentences

B. Data Collection

The English translation of our Bangla sentences were collected through crowds-sourcing. for this purpose, we developed a web interface for collecting translations from people. Figure 2 and Figure 3 show some of the screen shots of the interface.

Using this website we collected data from people. As seen from the photo of the interface, we gave them a random Bangla sentence from the corpus and they had to add the English translation of the respective sentence. We collected 3 to 5 English translated sentences for each Bangla sentence which results in 4 translated sentences against each Bangla sentence on an average. For the 517 Bangla sentences in our corpus we got a total of 2143 English translated sentences. The details of the English part of the corpus is given in table II.



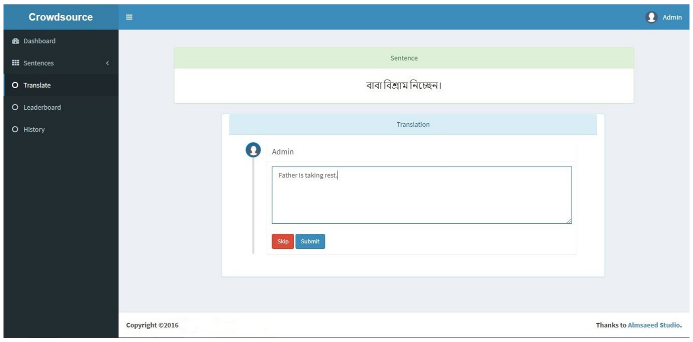
Sentence List

Show 10 entries Search:

id	Sentence	Translation	Action	#
1	আমি ভাত খাই।	0	Edit	Show Delete
2	আমি ভাত খাচ্ছি।	0	Edit	Show Delete
3	আমি ভাত খাচ্ছিলাম।	0	Edit	Show Delete
4	আমি ভাত খাবো।	0	Edit	Show Delete
5	আমি রাতে ভাত খাবো।	0	Edit	Show Delete
6	আমি এখন ভাত খাবো।	0	Edit	Show Delete
7	আমি ভাত খাই না।	0	Edit	Show Delete
8	আমি ভাত খাচ্ছি না।	0	Edit	Show Delete
9	আমি ভাত খাচ্ছিলাম না।	0	Edit	Show Delete
10	আমি ভাত খাবো না।	0	Edit	Show Delete

Copyright ©2016 Thanks to Almsaeed Studio.

Fig. 2. The Sentence List



Sentence

বাবা বিশ্রাম নিচ্ছেন।

Translation

Admin

Father is taking rest.

Copyright ©2016 Thanks to Almsaeed Studio.

Fig. 3. Adding English translation of a sentence

Table II: DETAILS OF ENGLISH PART OF THE CORPUS

Total sentences	2143
Total words	13062
Average sentence length	6 words

In table III we show some of the translated sentences collected through crowd-sourcing and the data we have got through it.

These data were collected from a group of university students where medium of study is English and Bangla is the first language. Our dataset consists of mostly simple sentences, and the user group chosen for collecting the data are well adapt and capable in translating them. There were 62 contributors in total for preparing this dataset. The data was collected over a period of 2 months. Source code of the website that has been used for collecting translations is available in github [10] along with the collected translations.

V. RESULT ANALYSIS

As stated earlier, for each Bangla sentence we got 4 English translated sentences on an average but number of translations received for any sentence varied with sentence

Table III: COLLECTED DATA

Bangla sentence	English translation
বাবা বাজারে যাবেন।	• Dad will go to bazar. • Father will go to the market. • father will go to the market • Dad will go to market. • Father will go to office.
বাবা কি আজ বাজারে যাবেন?	• Will father go to market today? • Will father go to bazar today? • Will father go to bazar today?
আমি এখন ভাত খাবো না।	• I won't eat rice now • I will not eat rice now. • I won't eat rice now. • i wont eat rice now
আমি গতকাল ব্যস্ত ছিলাম।	• I was busy yesterday. • I was busy yesterday. • I was busy yesterday. • I was busy yesterday. • I was busy yesterday.
বাচ্চারা মাঠে খেলছে।	• children are playing in the field. • Kids are playing in the field.
রহিম বেড়াতে যাচ্ছে।	• Rahim is going to visit. • Rahim is going outside. • Rahim is going to a tour.
তুমি কি কাজটি শেষ করেছ?	• Are you finished the job ? • Have you done the work? • did you finish the work?
কৃষক কি ক্ষেতে কাজ করছে?	• is farmer working on his farm. • Farmer is working in the field? • Is Farmer working in the field? • Is farmer working in the field.
খেলাধুলা স্বাস্থের জন্যে উপকারী।	• sport is beneficial for health. • The sport is beneficial for health. • Sports are beneficial for health. • Sports is better for health.

Table IV: TIME AND CONTRIBUTOR

Total Contributors	62
Time Required	2 Months

length. Sentences of length 3, 4 and 5 was translated mostly by users. The average number of translations received for any sentence length is shown in the graph in figure 5.

As seen in the previous section, we got a number of translated sentences for each Bangla sentence. The translated sentences has some variations from user to user. We discuss these variations and the reasons behind them in this section.

In case of very simple and small sentences all the translations we got are almost same and correct. for example-

1) আমি ভাত খাই না।
   - I don't eat rice.
   - I don't eat rice.
   - I do not eat rice
   - I don't eat rice.

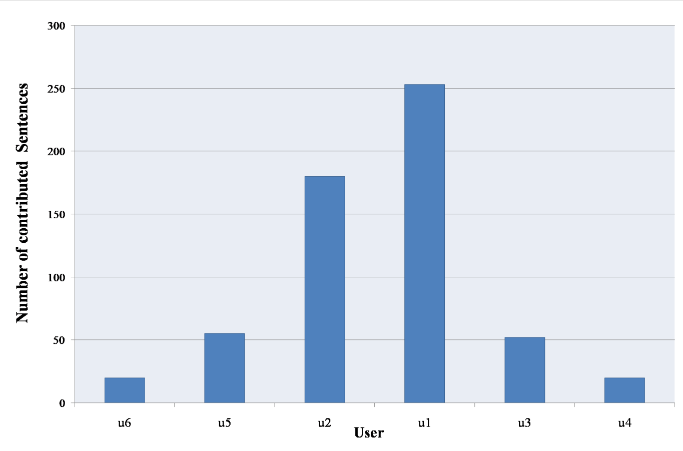


Fig. 4. An overview of user contribution

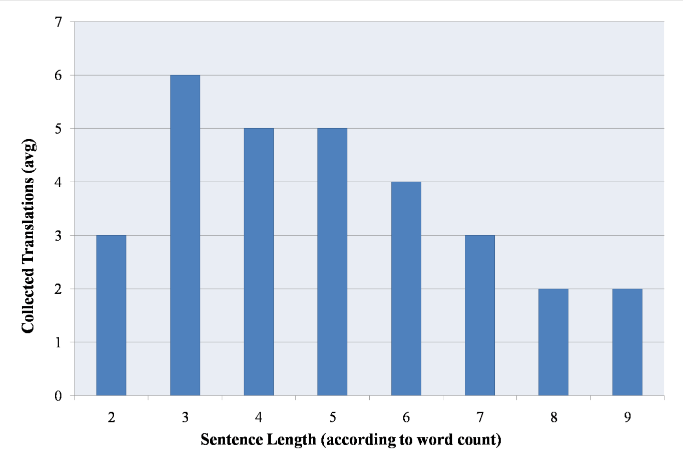


Fig. 5. Statistics of Average translations of Bangla sentence according to length (in words)

As seen in the above example, the sentence is very small and simple and there is not much variation in the way different people translated it. But when the sentence has nouns and pronouns the translation gets more varied. for example-

2) বাবা বাজারে গেছেন।
   - Dad went to the market.
   - Father went to bazar.
   - Father has gone to the market.
   - Father has gone to the market.
   - Father has gone to Market.

Here for the noun word 'বাবা' there can be two English words, 'Father' and 'Dad' which can be used alternatively and both are correct. Same can be said for the word 'বাজারে'. While most people translated it to the English word 'market', one user has treated it as proper noun and translated it to 'bazar'. Similarly synonyms of words can be used alternatively by different users while translating. For example-

3) বাচ্চারা মাঠে ক্রিকেট খেলছে।
   - The kids are playing cricket on the field.
   - Children are playing cricket in the field.

- Children are playing cricket in the playground.
- Kids are playing cricket in the playground.
- Children are playing cricket in the field.

Here for the word 'বাচ্চারা', two synonymous English words ’kids’ and ’children’ has been used alternatively and the same thing happened in case of 'মাঠে', which can be translated to both ’playground’ and ’field’. But the real problem arouses in case of universal truths. Different people translate these types of sentences differently. For example-

4) দুর্ভাগ্যবান তারাই যাদের প্রকৃত বন্ধু নেই।

   - Unlucky are those who don’t have real friend.
   - Those who do not have true friends are unfortunate.
   - Unlucky are those who don’t have real friends.
   - Those are unfortunate who do not have true friends.

   This much variation occurred in this example because universal truth sentences do not usually have a fixed sentence structure. As a result they are perceived differently by different people and the translation gets varied.

So, from the discussion above we can say that the alternative use of nouns, pronouns and synonyms mostly create the variations in the translation process. The sentences containing universal truths also need to be handled differently. So, further work is needed to resolve these issues.

VI. Conclusions

Crowd-sourced data can serve as a promising method of corpus construction in future. It has the advantage of reflecting human behavior while translating from one language to another. This method needs further analysis and more data to construct a complete corpus. Here we worked with an initial dataset to understand this method’s performance and issues regrading the corpus construction. The issues found in analysis of this data needs to be resolved in further works.

References

[1] M. A. Al Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, “Sumono: A representative modern bengali corpus,” SUST Journal of Science and Technology, vol. 21, pp. 78–86, 2014.

[2] S. Khan, A. Ferdousi, and M. A. Sobhan, “Creation and analysis of a new bangla text corpus bdnc01,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 5, 2017.

[3] N. S. Dash, B. B. Chaudhuri, P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja, “Corpus-based empirical analysis of form, function and frequency of characters used in bangla,” in Published in Rayson, P., Wilson, A., McEnery, T., Hardie, A., and Khoja, S.,(eds.) Special issue of the Proceedings of the Corpus Linguistics 2001 Conference, Lancaster: Lancaster University Press. UK, vol. 13, 2001, pp. 144–157.

[4] A. I. Sarkar, D. S. H. Pavel, and M. Khan, “Automatic bangla corpus creation,” BRAC University, Tech. Rep., 2007.

[5] K. M. A. Salam, S. Yamada, and T. Nishino, “Developing the first balanced corpus for bangla language,” in Informatics, Electronics & Vision (ICIEV), 2012 International Conference on. IEEE, 2012, pp. 1081–1084.

[6] M. A. Al Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, “Supara: A balanced english-bengali parallel corpus,” 2012.

[7] K. M. Majumder and Y. Arafat, “Analysis of and observations from a bangla news corpus,” 2006.

[8] J. Shamshed and S. M. Karim, “A novel bangla text corpus building method for efficient information retrieval,” Journal of Convergence Information Technology, vol. 1, no. 1, pp. 36–40, 2010.

[9] Chowdhury and Hossain, Advanced Learner’s Communicative English Grammar & Composition for Class-6 First & 2nd Paper, twenty 1st ed. Advanced Publication, 2016.

[10] “Crowd sourced translator and corpus construction project,” [Online]. Available: https://github.com/ZakiaRitu/Crowdsource_translator/, last accessed 5 November 2018.