Automatic Bangla Corpus Creation

Asif Iqbal Sarkar, Dewan Shahriar Hossain Pavel and Mumit Khan
BRAC University, Dhaka, Bangladesh
asif@bracuniversity.net, pavel@bracuniversity.net, mumit@bracuniversity.net

Abstract

This paper addresses the issue of automatic Bangla corpus creation, which will significantly help the processes of Lexicon development, Morphological Analysis, Automatic Parts of Speech Detection and Automatic grammar Extraction and machine translation. The plan is to collect all free Bangla documents on the World Wide Web and offline documents available and extract all the words in them to make a huge repository of text. This body of text or corpus will be used for several purposes of Bangla language processing after it is converted to Unicode text. The conversion process is also one of the associated and equally important research and development issue. Among several procedures our research focuses on a combination of font and language detection and Unicode conversion of retrieved Bangla text as a solution for automatic Bangla corpus creation and the methodology has been described in the paper.

1. Introduction

The development of language resources and its availability is a must for enhancing Language processing capabilities and research in this field. Thus corpus being such an important resource for any language we felt the need to work on it and create a Bangla corpus from existing Bangla documents available online and elsewhere. The plan is to have a system which will collect all Bengali documents from web or from local hard drives and then convert those into Unicode [1] documents in order to create a Bangla corpus and then the corpus will be used for several language processing activities for Bangla language. This huge language resource will be the store house of millions of words, which will be extracted systematically to build a sufficiently tagged Bangla lexicon. The lexicon can be used for several purposes including spell checkers and morphological analysis for Bangla language. So far we have made a little but quite significant progress towards achieving our goal and in continuation of the process we have developed a simple converter, which can convert any Bengali document written in ASCII (TTF) code into equivalent Bengali Document written in Unicode, which would play a vital role in corpus creation. The most significant challenge of this research is the conversion process of the documents available on the web into Unicode. The greatest problem for us is to determine the encodings of all current Bengali e-documents that are mostly edited with TTF. A huge portion of it is in image format, but that would require a sophisticated tool like Bangla OCR, which is beyond the scope of this research. So we are considering only those files which contain edit-able text. Now we have to consider some other related issues to complete the process and our future work will be directed towards coming up with their solutions. The issues include the following:

- Procedure for collection of online and offline Bangla documents.
- Ways for Conversion of the collected documents in to simple Unicode text.
- Storing the text as a corpus.

2. The Value of Language Resources

The value of this effort to create a Bangla corpus is very significant. Developing realistic models of human language that support research and technology development in language related fields requires masses of linguistic data: preferably hundreds of hours of speech, tens of millions of words of text and lexicons of a hundred-thousand words or more, meaning a corpus [2]. Although independent researchers and small research groups now have the desktop capacity to create small- to medium-scale corpora, the collection, annotation and distribution of resources on a larger scale presents not only computational difficulties but also legal and logistical difficulties to challenge most research organizations whether they be educational, commercial or governmental. There has been a significant development in the process of creating a corpus in some European and North American countries in their native language and numerous corporate research groups are routinely engaged in medium- to large-scale corpus creation. But there has been no such recognizable attempt to create Bangla corpus as such. The corpus, which is a published language resource, benefits a broad spectrum of researchers, technology developers and their customers. The presence of community standard resources reduces duplication of effort, distributes production costs and removes a barrier to entry. As research communities mature, published resources are corrected, improved and further annotated. They provide a stable reference point for the comparison of different analytic approaches in language processing. There have been a few incomplete attempts in this area by some contemporary research groups based in India, UK and USA but unfortunately in Bangladesh there has not been any fruitful progress even after realizing the necessity of such a resource. We have realized the need for a Bangla corpus and this paper is just the beginning of our progress towards coming up with a solution.

3. Methodology

The process of automatic corpus building involves several steps, which need to be thoroughly followed in order to create the Bangla corpus. Most of the steps can be implemented in various ways and can be accomplished very easily with existing technologies. The steps for automatic Bangla corpus creation are described below:

3.1. Step 1

First of all we configure a web crawler (robot program) [3] which will search the WWW and return the list of all possible websites which may contain Bangla text. The web crawler program downloads the information into several directories in a specific method. Then a very important job of Language detection is carried out, that is determination of Bangla Language. The language detection process is conducted in several ways depending on the condition of the documents on the web. The procedure is further elaborated below:

- Most of the Bangla text on the web is displayed using HTML and Bangla text could also be in the form of Text files or document files, which can be edited. If the web pages contain text in Unicode then the work becomes easier for us, because our plan is to create a large body of text which contains Unicode, the universally excepted script that can be displayed and processed without hassle in all platforms. So if the web pages contain Unicode text, we simply extract it and add it to our repository of text.
- If a True Type bangle font is used to display the documents, which is text other than Unicode require some processing before they are added to our repository. Unfortunately only a few web pages use Unicode encoding. Most of the Bangla websites either use image, PDF or proprietary True Type Fonts with their own encoding to display text on the web. To view the web pages the user is required to install that TTF in their system. The webpage usually offers the user an option to download the font in their respective systems. The webpage specifies the FONT FACE in their HTML tags, more specifically the FONT tag. Our job is to scan the HTML FONT tag and get the FONT FACE, which is basically the name of the Font.

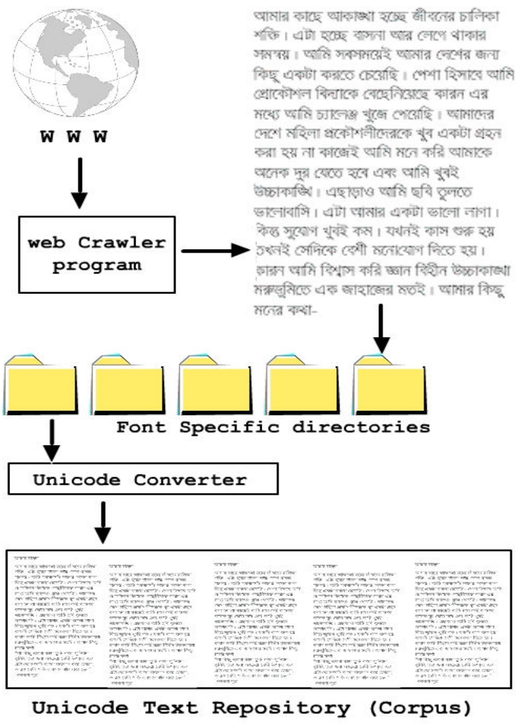
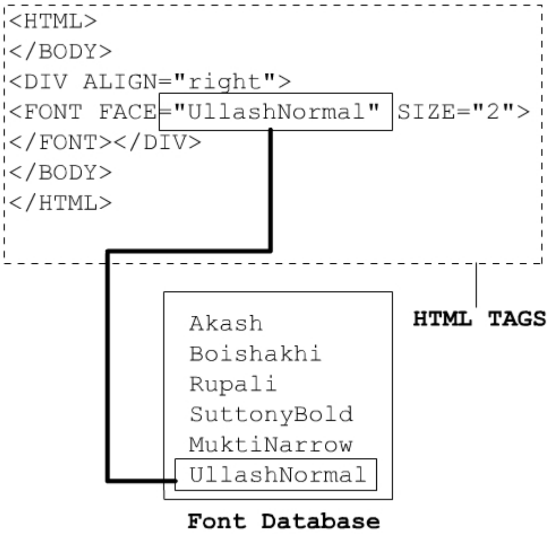

W W W
web Crawler program
আমার কাছে আকাঙ্খা হচ্ছে জীবনের চালিকা
শক্তি। এটা হচ্ছে বাসনা আর লেগে থাকার
সমন্বয়। আমি সবসময়েই আমার দেশের জন্য
কিছু একটা করতে চেয়েছি। পেশা হিসাবে আমি
প্রোকৌশল বিদ্যাকে বেছেনিয়েছে কারন এর
মধ্যে আমি চ্যালেঞ্জ খুজে পেয়েছি। আমাদের
দেশে মহিলা প্রকৌশলীদেরকে খুব একটা গ্রহন
করা হয় না কাজেই আমি মনে করি আমাকে
অনেক দুর যেতে হবে এবং আমি খুবই
উচ্চাকাঙ্খি। এছাড়াও আমি ছবি তুলতে
ভালোবাসি। এটা আমার একটা ভালো লাগা।
কিন্তু সুযোগ খুবই কম। যখনই কাস শুরু হয়
তখনই সেদিকে বেশী মনোযোগ দিতে হয়।
কারন আমি বিশ্বাস করি জ্ঞান বিহীন উচ্চাকাঙ্খা
মরুভূমিতে এক জাহাজের মতই। আমার কিছু
মনের কথা-
Font Specific directories
Unicode Converter
Unicode Text Repository (Corpus)

Figure 1: The automatic Bangla corpus creation process.



```
<HTML>
</BODY>
<DIV ALIGN="right">
<FONT FACE="UllashNormal" SIZE="2">
</FONT></DIV>
</BODY>
</HTML>
```

Figure 2: Font detection from HTML tags.

The name is then matched with a database containing names of all Bangla fonts (the list of names will be created incrementally). The encoding of all these fonts are also maintained in several files, which are later required for conversion into Unicode. A separate directory is maintained for each font and the text extracted from each page is placed and stored in the directory according to the fonts used to display them. If the font's name does not match with the database they are placed in a miscellaneous directory and those require manual detection and conversion for further processing. The miscellaneous directory normally hosts the text or document files which are downloaded as a result of web crawlers activities.

3.2. Step 2

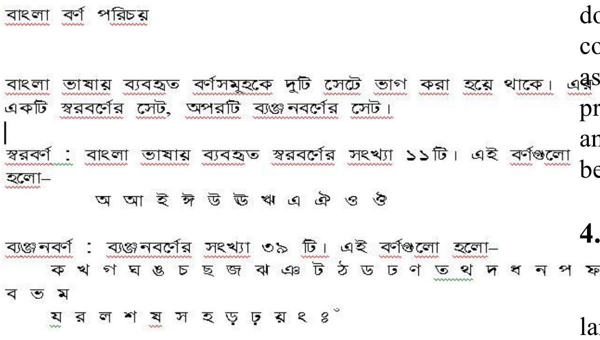
After the text is stored in different directories according to font they are systematically converted to Unicode by applying the converter. But before that, we have to make sure whether the text is in Bangla language. This is the most important step and certainly is a very important research issue that has been studied all over the world. The N-gram statistics [4] method is used to detect Bangla Language for this research project. The method is applied directly on the document and text files received in the miscellaneous directory, because it is not possible to determine their font using the font detection technique mentioned earlier. In the N-gram statistics method, we take the text file stored in directory, open it and randomly choose any 5 (or any number N depending on our choice) consecutive words of a sentence from any line of the text file, convert it to Unicode and match each word with an existing Bangla dictionary (collection of words without tags). If there is a 80% match, that is for the possibility that 4 words out of the 5 are Bangla words, then we take the text to be in Bangla and add it to our corpus (even though the mismatch could occur because of garbage value). If the match is not very significant we consider the text to be non-Bangla.

- The language detection method used here certainly is a research issue that still has drawbacks. For example, the chosen set of words in the text could contain garbage or Unicode points of other scripts while the rest of the document could contain Bangla script. So we would naturally discard a Bangla text by mistakenly considering it as non-Bangla. But the chance of occurrence of such a situation is very rare.

3.3. Step 3

The converter takes in as input the encoding of the fonts that are saved in different files and converts the (Bangla Language detected) text to Unicode according to their individual font and stores the text in a single text file which increases in volume after each conversion. The conversion process has been illustrated in the following images, showing the conversion of TTF text to OTF text for a sample Bangla text.

In this way, the converter can convert any text file written in any TTF to Unicode and thus they can be displayed using some of the Bangla OTF that has been developed so far.



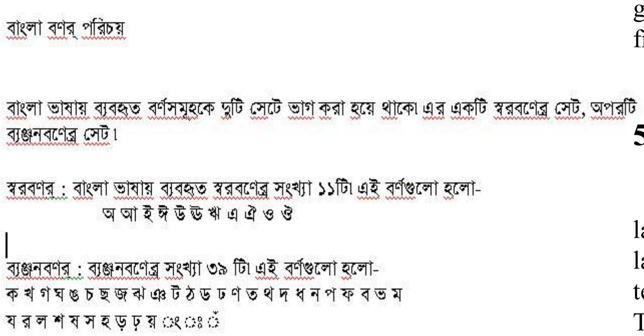
বাংলা বর্ণ পরিচয়

বাংলা ভাষায় ব্যবহৃত বর্ণসমূহকে দুটি সেটে ভাগ করা হয়ে থাকে। এর একটি স্বরবর্ণের সেট, অপরটি ব্যঞ্জনবর্ণের সেট।

স্বরবর্ণ : বাংলা ভাষায় ব্যবহৃত স্বরবর্ণের সংখ্যা ১১টি। এই বর্ণগুলো হলো–
অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

ব্যঞ্জনবর্ণ : ব্যঞ্জনবর্ণের সংখ্যা ৩৯ টি। এই বর্ণগুলো হলো–
ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ
ব ভ ম
য র ল শ ষ স হ ড় ঢ় য় ং ঃ ঁ

Figure 3: TTF text before conversion.



বাংলা বণর্ পরিচয়

বাংলা ভাষায় ব্যবহৃত বর্ণসমূহকে দুটি সেটে ভাগ করা হয়ে থাকে। এর একটি স্বরবণের্র সেট, অপরটি ব্যঞ্জনবণের্র সেট।

স্বরবণর্ : বাংলা ভাষায় ব্যবহৃত স্বরবণের্র সংখ্যা ১১টি। এই বর্ণগুলো হলো-
অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

ব্যঞ্জনবণর্ : ব্যঞ্জনবণের্র সংখ্যা ৩৯ টি। এই বর্ণগুলো হলো-
ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ভ ম
য র ল শ ষ স হ ড় ঢ় য় ◌ং◌ঃ◌ঁ

Figure 4: OTF text after conversion to Unicode.

3.4. Step 4

There are ways to quicken the task by taking help of the existing searching engines to do the search work for us. For instance, we can use the popular search engine on the web namely “Google” and we’ll set our searching key words like “Bangla, Bengali” etc. We are anticipating that this kind of search must narrow down the task of the web crawler. Then the web crawler will retrieve all the documents in the list of web links generated by “Google” and down load all those in our local drives. Then the language detection and conversion process will start accordingly. Employing the search engine readily available will greatly enhance the speed and accuracy of the text retrieval process.

3.5. Step 5

After detection, conversion of the text acquired from online documents they are all added to a single repository that we call the corpus. We can also employ the same method for offline Bangla documents and in this way we can gather a huge collection of text from different offline sources, such as book publishers, Newspaper archives and other print media. If the steps are properly followed we anticipate a significant number of Bangla words will be collected for the corpus.

4. Future Work

The corpus creation is the beginning of a series of language processing work that will be conducted using the corpus as the language resource. To implement a complete Bangla tagged lexicon [5] the corpus will be required. Beside this automatic Bangla grammar extraction, ma-chine translation and frequency analysis of words can also be done from it.

5. Conclusion

Corpus is considered as basic resource for language analysis and research for many foreign languages. This reflects both ideological and technological change in the area of language research. This change is probably caused due to the introduction of computer and corpus in linguistic research which, as a result, have paved out many new applications of language (and linguistics) in the fields of communication and information exchange. The use of corpus in Bangla language for various technological developments as well as for various linguistic studies in Bangla language can open up many new avenues for us. This corpus can be useful for producing many sophisticated automatic tools and systems, besides being good resources for language description and theory making.

6. Acknowledgment

This work has been partially supported through PAN Localization Project (www.PANL10n.net) grant from the International Development Research Center, Ottawa, Canada, administered through Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan.

The research team would also like to thank Mr. M. Saad Abdullah for his valuable contribution in idea formulation for the text acquisition and Language detection.

Bengali

7. References

[1] www.unicode.org

[2] C. Cieri and M. Liberman, *Issues in Corpus Creation and Distribution: The Evolution of the Linguistic Data Consortium*, University of Pennsylvania and Linguistic Data Consortium Philadelphia, Pennsylvania, USA.

[3] www.w3c.org/robot

[4] I. Suzuki, Y. Mikami, A. Ohsato, “A Language and Character Set Determination Method Based on N-gram Statistics”, *ACM Transactions on Asian Language Information Processing, Vol. 1, No. 3, September 2002, Pages 269-278.*, Nagaoka University of Technology and Yoshihide Chubachi Numeric & Co., Ltd.

[5] J. Hasan. “Automatic dictionary construction from large collections of text”. Master's thesis, School of Computer Science and Information Technology, RMIT University, 2001.