Research on Bangla Language Processing in Bangladesh: Progress and Challenges

M. S. Islam
Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology
Dhaka-1000, Bangladesh
E-mail: mdsaifulislam@iict.buet.ac.bd

Abstract

Bangla is spoken by about 245 million people of Bangladesh and two states of India. Today, most of the computer based resources and technical journals are in English. Due to the language barrier, the common masses face big obstacle to enjoy the optimum benefits of modern communication and information technology (ICT) as well as huge enriched English knowledge database around the globe. Language processing in mother tongue is the only technology that can be used to remove this barrier. The research work on Bangla language processing (BLP) was started in late 1980s in Bangladesh and it already produced some tangible results. The success of BLP may have a huge impact particularly to learn and use ICT in Bangla for the common people which will enhance their socio-economic life greatly. In this paper, we try to review the various research works that has been carried out in recent years in different areas of BLP in Bangladesh, pointed out the challenges and finally put forward some recommendations to overcome the challenges ahead for successful implementation of BLP.

Introduction

The origin of modern Bangla can be traced back to Vedic Sanskrit (1500 BC – 1000 BC), which is during the middle Indo-Aryan period gave rise to the dialects like *Magadhi* and *Ardha-Magadhi* (600 BC – 200 AD), followed by the *Magadhi-Apabhransa* and finally crystallizing to Bangla (10 century AD). According to the book 'The origin and development of Bangla Language' (Chatterji, 1962), the history of Bangla language can be divided into three distinct periods—formative (950 AD – 1200 AD), middle (1200 AD -1800 AD) and modern (1800 AD onward). The Bangla alphabet is derived from the *Brahmi* alphabet. It is also closely related to the *Devanagari* alphabet, from which it started to diverge in the 11th century AD. The current printed form of Bangla alphabet first appeared in 1778 when Charles Wilkins developed printing in Bangla. A few archaic letters were modernised during the 19th century. Bangla has two literary styles: one is called *Sadhubhasa* (elegant language) and the other *Chaltibhasa* (current language). The former is the traditional literary style based on middle Bangla of the sixteenth century, while the later is a 20th century creation and is based on the speech of educated people in the society. The differences between the two styles are not huge and involve mainly forms of pronouns and verb conjugations.

The people in Bangladesh, West Bengal and Tripura (two states in India) speak and write Bangla as their first language. It is also spoken in Malawi, Nepal, Saudi Arabia, Singapore, Australia, the UAE, UK and USA by Bangla speaking people. About one sixth population of the world is speaking in Bangla. The last decade has been marked by a new phenomenon called globalization and it has a profound impact on different domains of life – social, political and economic. It has also experienced a significant change in the communication dynamics of the world due to the advancement of information and communication technology (ICT). The nature and function of language processing is inevitably affected by these changes. Despite the importance of English language in everyday life especially in ICT field, the mother tongue Bangla may have huge impact for the use of ICT and dissemination of information to the common masses of Bangladesh. Due to the language barrier most of the people cannot get the benefits of Internet and large English database of journals, books etc. The Bangla language processing (BLP) can bridge this gap and Bangla speaking rural people can get maximum benefit out of it. The impact of Bangla computing is not limited to socio-economic aspects only. It will also push the Bangla language identity on the global stage. This paper aims at reviewing the BLP research works that are carried out in Bangladesh in the last few years, identifying its challenges and finally provide some recommendations to address these problems.

Optical Character Recognition

Character is the fundamental attribute for writing and reading a language. Character recognition is a process to classify the input character according to a predefined character class. With increasing the interest of computer applications, the Bangla character recognition of typed or handwritten text bears high importance. Bangla language has 49 letters in its alphabet of which 11 letters are vowels and 38 letters are consonants. There is no capital letter and letters are not connected to the other letter that follows them. The Bangla optical character recognizer (OCR) has infinite applications including fast digitizing of old and rare Bangla books, which would save a lot of time compared to manually typing all the words in those books.

An online Bangla handwritten recognition system (Badruddoza, 2003) was reported that uses neural network for feature selection and extraction, and achieves a recognition rate about 90%. The segmentation of Bangla character plays an important role in recognition because it allows the recognition system to classify the characters more accurately and quickly. Mahmod *et al* (2004) focuses on the segmentation of printed Bangla character. In this scheme, at first the printed document is scanned and digital image is formed and then they have applied two phase filters to identify the segments of each character. A document image can contain every type of character (vowel, consonant and numeric digit). Bangla number extraction and recognition from document image is done by using joint correlation technique (Molla and Talukder, 2007). Center for Research for Bangla Language Processing (CRBLP) of BRAC University recently release the BanglaOCR that the process of converting printed text images to editable unicode text. It also allows users to train data set from any document and observe the recognition performance. BanglaOCR is free and run on Mac, Windows and Linux platforms.

Morphology and Morphological analysis

In order to develop computer applications that successfully process natural language data (such as, text and speech), one need good models of the vocabulary and grammar of a particular language as many as possible. According to linguistic theory, words consist of morphemes, which are the smallest individually meaningful elements in a language. Since an immense number of word forms can be constructed by combining a limited set of morphemes, the capability of understanding and producing new word forms depends on knowing which morphemes are involved. Morphology is the branch of grammar which studies the structure or forms of words of a language. It focuses on pattern of word formation within and across languages and attempts to formulate rules that model the knowledge of the speakers of the language. Morphological information is useful for parsing, lemmatization and in several natural language applications: text generation, machine translation, document retrieval, etc.

Modern Bangla morphology is very productive, especially for verbs, with the root verbs takes 168 different forms (Chottapaday, 1989). Bangla lexicon also has a very large number of compound words (words that have more than one root), which can be created from at most any combination nouns, pronouns and adjectives. An effort is made at building a complex morphological parser for Bangla, where it can only handle simple words with a single root (Dasgupta and Khan, 2004). Though the addition of inflectional suffixes in Bangla compound word is fairly complex, the compound word's individual root words may retain their inflectional suffixes. Such a compound word morphological parser is recently developed (Dasgupta, 2006) which can efficiently parse compound words having inflectional suffix and also resolves ambiguities. Ali *et al* (2008) developed some rules to develop the morphological analysis of simple and compound Bangla words that can be used to make universal natural language (UNL)-Bangla dictionary for converting the natural Bangla sentences to UNL documents and vice versa.

Speech Processing and Recognition

Speech can be described as an act of producing voice through the use of the vocal folds and vocal tract to create a linguistic utterance that convey information. Speech is produced by articulator in the vocal tract and muscle. Humans express thoughts, feelings, and ideas through a series of complex movements of vocal tract that alter and shape the sounds. Currently, computers can only understand human speech in a very limited capacity. Because computer interfaces usually consist of a keyboard and mouse, a person has to learn certain skills before they can use a computer. It would be much easier for anyone to use a computer if the computer could understand the person's natural communication method. Speech recognition could also be used to provide access for anybody who has a handicap that prevents use of a keyboard. There is an entire class of people that cannot use a computer at all because they are disabled. Bangla speech recognition could potentially make their lives easier. Computers also need a way to be able to identify who is trying to use them. The most common method of user identification is through the use of passwords. Passwords are not always effective for several reasons. The first reason is that the computer identifies the user purely based on a sequence of characters input by the user. It is easy to see that anyone knowing this sequence can gain access, even if they are not the intended user. Passwords can also be guessed or broken. There are several characteristics to a person's voice that are unique to the individual. Because of this uniqueness, a person's voice could be a very accurate way to authenticate a user. Voice recognition has the benefits of being very user friendly and secure.

Over the years, a significant contribution is already made in Bangla speech processing and recognition. A detailed study on speech production mechanism of Bangla phoneme processing and classification criteria for computer analysis and synthesis of Bangla speech is carried out (Hossain, Rahman and Ahmed, 2005). They also differentiate between vowel and consonant both from the context of linguistics as well as computer processing point of view.

A Bangla speech recognition system is being developed using the linear predictive parameters (LPC) and their derived parameters related to speaker's vocal tract for speaker identification system in which the computer will be able to understand a few simple commands, and identify them accurately (Ali, 2005). In this case, the LPC derived parameters with hidden Markov models has been used in both speech and speaker identification system. On 19 February 2009, CRBLP of BRAC University announced the first official release of its Bangla language processing software package *'Katha'* which converts Bangla text to speech (TTS) (Mahboob, 2009). The TTS run on Linux, Windows and Mac OSX. There is also a web-enabled front-end for the TTS.

Machine Translation

Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. The idea of machine translation may be traced back to 17th century. The idea of using digital computers for translation of natural languages was proposed as early as 1946 by A. D. Booth (British engineer, physicist, born in 1918) and possibly others. Important advances in MT occurs during the 1980s because large-scale access to personal computers and word procesing programs created a conducive environment for machine translation. The beginning of 1990s saw vital developments in machine translation with a radical change in strategy from translation based based on grammatical rules to that based on bodies of texts and examples. To this days as Internet is increasingly multilingual and people need help in finding the information they seek – only the MT can meet up their today's requirements. This situation as led to the creation of online machine translation services such as Altravista, which offer rapid email services, web pages, etc. in the desired language, as well as to the availability of multilingual dictionaries, encyclopedias and free, direct access terminology databases.

The MT for English-Chinese, English-Arabic, English-French, and many other pairs of languages are relatively quite advanced. On the other hand, very little work has been done in developing Bangla MT in the field of automatic translation, parsing and syntax analysis to develop software for translating English-Bangla (E2B) or Bangla-English (B2E) vice-versa in Bangladesh. Some work has been done; it's for the Bangla of West Bengal in India. Here, we are citing some Bangla MT research work in Bangladesh. A significant part of the development of any MT system is the creation of lexical resources that the system will use. For accurate and efficient transformation from one language to another the necessity for a MT dictionary is obvious for specific domain. An attempt is made to develop MT Bangla dictionaries that address the organization, contents and details of the information (Ali and Ali, 2002). Saha (2005) developed low cost English to Bangla (E2B)-ANUBAD that translates English text into Bangla text with disambiguation. It uses both the rule-based and transformation-based MT schemes along with three level of parsing. An effort is made to develop a statistical Bangla to English translation engine using only simple Bangla sentences that contains a subject, an object and a verb (Uddin, Ashraf *et al*, 2004).

State of Bangla Computing and Challenges

We already mentioned that Bangla is the first language of Bangladesh. However, organized efforts in software and computer based content and software system localization in Bangla are not very visible in the country. It is obvious that before any content can be generated or any application developed, some basic standards for encoding the language must be developed. The first attempt to use Bangla in computing was made in the early 1980s with Bangla font development in the Windows environment. These efforts were led by commercial vendors. It was in 1986 when Bangla language first entered the computer system through '*Shahid Lipi'*. It was a breakthrough. The introduction of *'Bijoy*' Bangla software also added a new dimension to the Bangla computing initiative. The main problem at that period was the compatibility issue of Bangla language in different platform. Bangla was not usable as a general language on every system as there was no unique way to represent Bangla. In the late 1990s unicode shed new light on the issue and the process of Bangla computing began to take a new shape in the country.

The open source movement has some impact on Bangla in computing. In 1998, J. Ahmed (*wiki.mozilla.org/L1on:Teams:bn-BD*), a software developer in Bangladesh, first solved the Bangla issue in computing and started a process of Bangla version of Linux. In the late 1990s, a voluntary group named Ankur (*www.ankurbangla.org*) started Bangla open source software like Linux, OpenOffice.org, Gaim, etc. Another voluntary organization, Ekushey (*ekushey.org*), started developing open source unicode fonts and a Bangla input system (*i.e.* determining how Bangla fonts can be arranged using the existing keyboard). In 2004, the Bangladesh Computer Council (BCC) took the initiative from the government side and came up with a national Bangla keyboard mapping and a collation sequence. Around this time, CRBLP at BRAC University started conducting research projects that dealt with Bangla language processing. Presently, the research team of CRBLP is working on Bangla information retrieval (*i.e.*, Bangla spell-checking and a Bangla search engine etc.), morphological analysis, developing a digital lexicon and an online dictionary, optical character recognition and speech processing. In 2005, the Bangladesh Open Source Network (BdOSN: *bdosn.org*) was formed with local open source volunteers. BdOSN took Bangla in computing as one of its main issues. As a result, open source in Bangla has started to thrive.

Though there is a positive movement on BLP research to use of Bangla for the benefit of common masses to use and learn ICT, but it faces the various following challenges leading to comprehensive use of Bangla in ICT:

i) There is a lack of detailed morphological analysis of Bangla language which is very essential to develop software framework for application level support like spell checker, OCR, TTS, grammar checker, MT etc, remains to be a significant challenge.

ii) Still we do not have a large and representative lexicon of Bangla language and the various lexicons currently in use do not contain a large number of ever expanding colloquial terms and proper nouns. Now, there is a need to build a larger and elaborate lexicon.

iii) Most of the current Bangla language computing tools are primarily based on Microsoft Windows operating system. As all the open source operating greatly improved and support unicode fonts, focus need to be shifted to cover these free platforms fully.

iv) There is a lack of coordination and integration among the various research groups working on the different areas of BLP. Sometimes same or almost similar work is carried out simultaneously that bears little value or no value at all.

v) There is a lack of consistent and targeted research funding to develop the skill set required for long term development of BLP. Consistent and sponsored research funding is outmost necessary to carry out the on going research as well as to build future human resources in the field of BLP an effective manner.

Conclusion and Recommendations

This paper is an attempt to survey the research work on BLP in Bangladesh and the overall status of use of Bangla language in computing. We have particularly focused only few important areas of language processing to show what we have achieved and where we are lagged behind. Though BLP research in Bangladesh started on late 1980s – about two and half decades of slow research and development, we found that efforts are now showing the hopeful results. In comparison with the language processing efforts in Europe, America and Japan, which are at more than four decade old, it would seem that we have progressed a little on BLP in Bangladesh has long way to go to reap the full benefit of it. However this can advantage, because Bangladeshi researcher can learn from the experience of their global counterpart.

These days, we are talking about widespread use of ICT and digital Bangladesh; this only can be achieved if the computational tools are in Bangla. Without the development of Bangla processing computation tools, the fruits of ICT can not reached to the common masses. To expedite the basic research on BLP and Bangla language tools development, we need to address the following issues:

i) Developing a large and representative Bangla corpus which will be helpful for spelling and grammar checking, speech reconstruction, speech generation, topic detection, message understanding and many other related topics.

ii) Developing a fully unicode compatible Bangla operating system in Widows and Linux platforms.

iii) Developing a local language dialect database and customized dialog system for easy access for all people to the necessary information according to their need.

iv) Developing a good English-Bangla (E2B) and Bangla-English (B2E) machine translation system which will immensely help the students, researchers and other people to use the huge English knowledge database of the Internet as well as journals.

v) Developing mechanism of consistent and long term research funding for BLP with targeted goals within specific timeframe.

vi) Attracting the fresh and potential young researchers with attractive salary package, imparting training with appropriate skill and offering scholarship for higher degrees to create a pool of competent human resources to carry out research work on BLP.

References

Chatterji, S. K. (1962). *The Origin and Development of the Bengali Language*. Published by Rupa & Co., New Delhi, India.

Badruddoza, M (2003). Recognition of Bangla hand written letters using self-organizing map (SOM). *Proceedings of 6th International Conference on Computer and Information Technology (ICCIT*), 357-360.

Mahmud, S. M. M, Shahrier, N., Hossain, A. S. M., Chowdhury, M. T. M. & Sattar, M. A. (2004). An efficient segmentation scheme for the recognition of printed Bangla characters.

Proceedings of 7th International Conference on Computer and Information Technology (ICCIT), 779-781.

Molla, M. K. I., & Talukder K. M. (2007). Bangla number extraction and recognition from document image. *Proceedings of 10th International Conference on Computer and Information Technology (ICCIT*), 512-517.

Chottapadday, S. K. (1998). *Vasha Prokash Bangla Bakaran*, Annada Publishers, Kolkata, India.

Dasgupta, S., & Khan, M.(2004). Morphological parsing of Bangla words using PC-KIMMO. *Proceedings of 7th International Conference on Computer and Information Technology (ICCIT*), 264-267.

Dasgupta, S. (2006). Morphological analysis of inflectional compound cords in Bangla. *Proceedings of 9th International Conference on Computer and Information Technology (ICCIT*), 345-348

Ali, M. N. Y., Al-Mamun, S. M. A., Das, J. K., & Nurunnabi, A. B. (2008). Morphological Analysis of Bangla Words for Universal Networking. *Proceeding Third International Conference on Digital Information Management, (ICDIM 2008*), 532-537

S. A. Hossain, S. A., Rahman, M. L., & Ahmed, F. (2005). A review on Bangla phoneme production and perception for computational approaches. *World Scientific and Engineering Academy and Society (WSEAS*), Sofia, Bulgaria*,* 346 - 354

Ali, M. E. (2005). *An Approach to Implementation of Bangla Speech Recognition using Hidden Markov Model.* M. Sc. Engg. Thesis dissertation, Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka.

Mahboob, M. (2009). *TechSpotlight: The power of Bangla*, The Daily Star, Dhaka, Bangladesh.

Ali, M., & Ali, M. M. (2002). Development of machine translation dictionaries for Bangla Language. *Proceedings of 7th International Conference on Computer and Information Technology (ICCIT*), 272-276.

Saha, G. K. (2005). The E2B machine translation: A new approach to HLT. Ubiquity archive, *Association of Computing Machineries (ACM),* 6(32), New York.

Uddin, M. G., Ashraf, H., Kamal. A. H. M, & Ali, M. M. (2004). New parameters for Bangla to English statistical machine translation. *Proceedings of 3rd International Conference on Electrical & Computer Engineering (ICECE 2004),* 545-548