

Software Localization Challenges for Bangladesh

Mumit Khan
BRAC University
66 Mohakhali C/A, Dhaka-1212, Bangladesh
mumit@bracuniversity.ac.bd

Introduction

Bangla is the primary language for the 130 million people of Bangladesh, yet there has been scant organized effort until very recently to localize the language for electronic medium. The good news is that we have seen an explosion in the research papers being submitted to conferences that deal with Bangla language computing, especially in forums such as International Conference on Computer and Information Technology (ICCIT), an annual conference held in Bangladesh. The bad news is that there is still no significant synergy among the various research and development teams that are working in this field. However, there is positive movement thanks to various research grants and sponsorship by the government to promote such efforts, and the upcoming year of 2004 does promise a renewed interest in working on the various challenges leading to a comprehensive localization of the Bangla language.

Fonts

The first attempts during the 80s in designing Bangla fonts were all based on bitmap fonts, and were specifically designed for text processors under MS-DOS and later under MS-Windows based operating systems. These fonts have since evolved, and used extensively; however, new efforts are now underway in designing Unicode fonts, which will undoubtedly increase the scope of the use of Bangla language in the electronic medium as various operating systems, especially the open source ones such as Linux and BSD, continue to add and enhance native support for Unicode fonts.

Morphological Analysis

The lack of computational linguistics research on Bangla remains to be a significant challenge in Bangladesh. Researchers in neighboring India however have done noteworthy progress in this area, and we hope to utilize much of that work and contribute in ongoing research. Development of Bangla Morphological parser s one of the highest priority items at BRAC University and elsewhere in Bangladesh, which is essential in creating software frameworks for application level support for spell checking, optical character recognition, text to speech, grammar checker and machine translation. One current project in collaboration with NUCES in Pakistan is focused on developing inflectional and derivational morphology for Bangla.

Lexicon and Dictionary

Building a large and representative lexicon, using morphological rules, requires a significant effort. The various lexicons currently in use do not contain a large number of ever expanding colloquial terms and proper nouns, which hinder spell checkers from being effective tools in everyday work. There is now a proposed project to create a larger and more elaborate lexicon based on the current ones, and using morphological tags to enable end-user applications take full advantage of it. This is however a large task, and will require the cooperation of all concerned. The proposal asks that Bangla Academy, the national body entrusted with language policy matters in Bangladesh, maintain and distribute the lexicon, much like it has done with its Bangla dictionary that is currently the most comprehensive one available and used extensive throughout Bangladesh and the Indian state of West Bengal.

Spell Checker and Optical Character Recognition (OCR)

While “intelligent” spell checkers for many of the western languages are now ubiquitous in word processing software, the state of the art for integrated Bangla spell checker is far from being satisfactory. Two main factors contribute to this problem: (1) lack of morphological parser for Bangla, which requires that the lexicon contain all the variations of a particular word, and (2) lack of integration of Bangla language support in existing word processing software. The problem of finding a “good” distance algorithm is currently being addressed (e.g., the *RecursiveSimulation* algorithm by A.B.A. Abdullah and A. Rahman), and the integration of various local spell-checkers have now established a local talent pool for future work.

Being able to scan documents in Bangla script is now probably the most worked on topic among the various localization efforts. OCR support faces many of the challenges of creating a good spell checker, and additionally faces the issues of pattern recognition and image processing algorithms. Much of the fundamental algorithmic work is of course in common with OCR work done for other languages, but the Bangla specific portions are also quite challenging. Integrating OCR support in existing scanner applications is one topic currently being addressed; there is now work underway to create device drivers for Linux that will allow a TWAIN-supported scanner to scan Bangla documents. However, the support for various typefaces is still very limited, and support for handwritten documents is far from usable at this point. Error detection and correction algorithms for scanned documents is another research topic, and some encouraging work is being performed by the researchers at the Indian Statistical Institute.

Operating Environments (UNIX, Linux, Microsoft Windows, Mac OS X)

The current implementations of Bangla language computing tools are primarily based on the Windows-based operating systems. The focus needs to shift to cover other operating systems as well, especially the open source ones such as Linux and the

others based on BSD, and Mac OS X as well. All the open source operating systems have greatly improved support for Unicode fonts, and there are free and cross-platform word processing software available that can be used to target localized software. The advantages of using open source software cannot be overemphasized, especially when considering the cost of commercial software in the context of a poor nation. One particular problem is the integration of Bangla support in the existing text processors available for these operating systems, many of which are still “in flux” as these are actively being developed, complicating the issue of integration.

Research Collaborations

One particular issue that has been raised recently is the lack of collaboration among the various disparate groups working on the different facets of the entire problem of localization. Conferences such as ICCIT has done a great service of bringing these researchers to the same table, and these international conferences promise to provide a synergistic environment for the researchers to move forward. The last ICCIT conference held in December of 2003 had over 25 papers on the various issues related to localization, and the next one is bound to have even more contributions. Such forums also raise the awareness that is critical to get the university faculty and students to get involved in research projects in this field.

Research Funding and Sponsorship

The lack of consistent and targeted research funding remains to be a significant challenge in developing the skill set required for long-term development in Bangla language computational tools. The current environment is slowly changing, thanks to the intervention by the Government of Bangladesh through its Information and Computer Technology (ICT) initiative, as well as externally sponsored research such as the one funded by Canada’s IDRC in collaboration with NUCES in Pakistan. We hope to build a pool of local talent capable of conducting original research and development in this field, and train the next generation of graduates from local universities.

Conclusions

The current state of localization of Bangla has progressed steadily in the past 5 years or so at a much faster pace than compared to the previous decade. One reason can be attributed to the awareness brought to the fore by various government initiatives to promote IT in Bangladesh, as well as the international conferences that have showcased the work of a small but productive group of researchers. However, many challenges remain in producing Bangla language computing tools, and especially in integrating Bangla support in existing word-processing software, which is essential in making it usable to the larger body of users. Some of the challenges such as font design, input and output device support, linguistic analysis, lexicon and dictionary development are now actively being addressed. The non-technical issues such as sponsored research to develop the local skills are yet to be addressed in an effective manner, which will hopefully change in this new year.