Preface

Over 250 million people speak Bangla (or Bengali), an Indo-Iranian language. The overwhelming majority of this population lives in the eastern flank of South Asia that surrounds the Bay of Bengal. They are geographically distributed as follows: over 95% of those living in Bangladesh, and from amongst the Indian states about 26% of those in Andaman and Nicobar Islands, 28% of those in Assam, 67% of those in Tripura, and 85% those in West Bengal. A large Bangla-speaking population is now in diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, United Arab Emirates, United Kingdom, and United States. Although it is the sixth most spoken language in the world, it is not necessarily as highly ranked in terms of the most read, or the most wired, or the most archived, or the most used on the Internet.

Despite significant progress in Information and Communication Technology (ICT) and the availability of a huge, enriched English knowledge database around the globe, the potential ICT benefit continues to elude a large majority of the Bangla-speaking population who are not equipped with either English or the language of their own diaspora. This is complicated by the fact that Bangla is not without its own peculiar nuances and structural issues. It has a relatively large alphabet set that also includes many compound letters, two acceptable forms with varying pronouns and verb conjugations, many regional dialects, and variant spellings for too many of its words. Additionally, there are at least two non-standard dialects of Bangla – Chittagonian and Sylheti, respectively, with 47% and 30% lexical dissimilarity. Bangla Language Processing (BLP) is an evolving computer science discipline that is cognizant of these language realities and seeks to create a robust digital platform for use by a large majority of the Bangla-speaking population who are being bypassed currently by the ICT revolution. The success of BLP is envisioned to have a positive impact for many of the common people and their socio-economic life.

This book had its origins in a major IEEE-sponsored international conference, namely the International Conference in Computer and Information Technology (ICCIT), now in its 16th year, which continues to highlight the latest works in BLP. While there continues to be progress in BLP research, lack of enough archived material outside of that already included in IEEE xPlore continues to force not only the students of computer science and engineering but also the researchers and technologists to fall often in the trap of re-inventing the wheel. We hope that this milestone source book will fill a serious void in both teaching and research, facilitate further BLP research and development, and, consequently, preserve Bangla as a vibrant digitization-ready language for a long time to come.

The 16 chapters of *Technical Challenges and Design Issues in Bangla Language Processing,* selected from 27 initially proposed papers, are authored by 41 researchers from Bangladesh, Canada, India, Ireland, Norway, United Kingdom, and United States. These chapters span seven BLP topical areas – font design, machine translation, character recognition, parsing, speech processing, information retrieval, and

sentiment analysis. Additional acceptable chapters on word processing, spell checking, database management, digital displays, and wireless applications would have made this a much stronger resource book.

In chapter 1, Fiona Ross discusses the key issues that underpin best practices in Bengali digital type design – from a design's conception to its implementation. Aspects of the character set such as dimensions, character fitting, and harmonious multi-script setting are considered from a perspective of non-Latin type design and font development. The chapter elaborates on how past practices in type-making and typesetting has affected current Bengali type forms and how the existing and emerging font technologies can be used effectively to support high-quality cross-platform Open Type Bengali fonts. The next chapter by Hossain, Mahbub-ul-Islam, Azam, Ahamad, and Khan reviews Bangla Braille development and identifies necessary grammatical rules as well as conventions for rule-based Braille translation. A computational model that uses Deterministic Finite Automata (DFA) for machine translation is introduced and studied for its acceptability by the visually impaired community. Architecture for the implementation of machine translation of Bangla to Braille using open source technology is demonstrated and is tested with Bangla Unicode-based text contents, and the generated Braille code is validated after printing in a Braille printer.

Machine Translation (MT) by creating lexical resources allows computers to translate texts from one natural language to another; its importance has become ever more significant with increasing use of the Internet. While the MT for English-Chinese, English-Arabic, and English-French, for example, are already advanced, developing Bangla MT hasn't progressed as much. In chapter three, Ali and Ripon develop a framework for Bangla MT that consists of an EnConverter to convert Bangla native sentences to UNL expressions and a DeConverter, which converts UNL expressions to respective Bangla sentences. In both, the authors consider case structure analysis, Bangla parts of speech, and different forms of verbs along with their prefixes, suffixes, and inflexions. Experimental results confirm that the proposed framework can successfully convert Bangla sentences to UNL expressions, and vice versa. This is followed by chapter four in which Maxim Roy considers the ideas behind Statistical Machine Translation (SMT) systems, which depend otherwise on the availability of bilingual data between language pairs to improve accuracy. The author considers machine-learning approaches such as sentence selection strategies that can improve accuracy without requiring a huge increase in resources. In semi-supervised settings, it is shown that the reversed model approach outperformed all other approaches for Bangla-English SMT and in active learning settings.

Optical Character Recognition (OCR) will play a significant role in the digitization of both printed and handwritten documents and texts. It depends on the performance of classifiers, which in turn relies on the feature extraction methodology used. For Bangla OCR, there are additional challenges to overcome since it is an inflectional language; there are about 300 *basic*, *modified,* and *compound* character shapes in the script, and the characters in a word are often topologically connected. In chapter five, Sarwar, Rahman, Akter, Hossain, Ahmed, and Rahman provide a review of various feature sets and a variability analysis for an optimal feature set by focusing on the specific peculiarities of Bangla such as its different usage as vowel and consonant signs, as well as compound, complex, and connected characters. Islam and Karim then follow up with chapter six, which provides for coverage of optical techniques/system for recognition of Bangla characters. The authors review phase-only filter-based hybrid electro-optical systems but then build upon them to elaborate on the use of joint Fourier transform optical correlators in recognition of Bangla characters.

Morphological information is integral to parsing, lemmatization, and in applications such as text generation, machine translation, and document retrieval. The next two chapters (7 and 8) focus on parsing that plays a prominent role in computational linguistics. Words consist of individually meaningful

root elements known otherwise as morphemes. Since combining morphemes forms a large number of words, the capacity to produce and understand new words depends often on knowing which morphemes are involved. Morphological analysis of simple and compound Bangla words can be used to make a Universal Natural Language (UNL)-Bangla dictionary for converting the natural Bangla sentences to UNL documents and vice versa. In chapter seven, Al-Mahmud, Sarker, and Hasan review the Context-Free Grammar (CFG) for parsing Bangla language processing. It involves a predictive parser, and the parse table is constructed for recognizing Bangla grammar and overcoming possible syntactical errors when there is no entry for a terminal in the parse table. The top-down CFG parsing suffers often from left recursion issues that are overcome by considering a left-factoring technique. Garain and De, in chapter eight, next consider a constraint-based Dependency Parsing, in general, and a Paninian grammatical model, in particular. The authors attempt to simplify complex and compound sentential structures first, then parse the simple structures so obtained by satisfying the Karaka demands of the verb groups and finally rejoin the parsed structures with the appropriate links and Karaka labels. The trained parser is shown to achieve very high accuracy.

The understanding of human speech by computers is limited in part because it interfaces usually through a keyboard and mouse. Speech recognition as a concept has multiple application possibilities including for user authentication and for use by those who are disabled. The next three chapters focus on varying aspects of speech recognition.

In chapter 9, Kotwal, Hassan, and Huda review Bangla Automatic Speech Recognition (ASR) techniques by evaluating different speech features, such as Mel frequency cepstral coefficients, local features, and phoneme probabilities for different artificial neural networks. The authors have designed three classifiers by male, female, and gender-independent speakers and explored the use of dynamic parameters in their experiments for obtaining higher accuracy in phoneme recognition. Haque follows up in chapter ten and provides an overview of the theory of speech production and analysis and synthesis of Bangla ASR. The author explores nasality, which is a distinctive feature of Bangla vowels, in particular. The chapter provides a review of nasal vowel research, cross-language perception of Bangla vowel nasality and vowel nasality transformation for use in a speech synthesizer. Finally, chapter eleven by Hossain, Rahman, Ahmed, and Sobhan reviews the salient features of Bangla vowels and the sources of acoustic variability in Bangla vowels, and suggests the classification of vowels based on normalized acoustic parameters. The normalization is necessary to remove the effect of non-linguistic factors given that Bangla vowels are spoken differently by different native speakers and by different regions. The authors also consider the study of acoustic features of Bangla dental consonants to identify the spectral differences and parameterize them for the purposes of synthesis.

Ganguly, Leveling, and Jones, in chapter 12, provide an introduction to Bangla information retrieval by reviewing its latest state-of-the-art and identifying the guidelines for application developers on how to set up an information retrieval system, with special attention given to language-specific aspects. The chapter identifies steps for creating and evaluating an information retrieval system including content processing, indexing, retrieval models, and evaluation. They also explore cross-lingual information retrieval, in which queries are entered in English with an objective to retrieve documents in Bangla. Chapter thirteen, by Das, Basu, and Mitra, next explores the specific case of Bangla information retrieval by considering the literary work of Rabindranath Tagore. Tagore's work happens to also provide for richness in both style and language. This work includes a quantitative study of vocabulary size and lexical richness in terms of statistical measures as well as an effective search engine for his works.

The final three chapters of the book are devoted to sentiment analysis, which is turning out to be an important area of natural language processing. The effort to determine whether the expression of a speaker or writer is positive or negative toward a specific subject is becoming relevant given the rapid growth of e-commerce and e-governance. In chapter fourteen, Hasan, Islam, Masur-E-Elahi, and Izhar present a *Sentiment Analyzer* that recognizes Bangla sentiment or opinion about a subject from Bangla text. It relies on specific phrase patterns and their sentiment orientations. Next in chapter fifteen, Amitava and Gambäck explore a particular sentiment analysis to determine if the opinion in question is positive (happy), negative (sad), or neutral (memorable). Finally, in chapter sixteen, Das and Bandyopadhyay consider improving multilingual search engines on the basis of sentiment or emotion and in an effort to build resources for languages other than English. The authors describe the preparation of an emotion corpus and lexicon, termed the Bengali *WordNet Affect*, by considering expansion, translation, and sense disambiguation. Manual annotators develop a Bangla blog corpus for emotion analysis with considerations given to emotional expressions and intensities, emotion holders, and sentential emotion tags.

From a scan of the chapters and topics included in this book, it is clear that there is an effort at exploring research on vital BLP topics. Few of the serious challenges, as we see, remain to be addressed. There still seems to be a lack of detailed morphological analysis of the Bangla language, which is going to impact software framework for the purposes of spell checker, OCR, grammar checker, speech generation, and machine translation. Developing a reliable machine translation system is key to benefitting from the huge English knowledge database of the Internet as well as journals. The need to build a larger and elaborate lexicon as well as a fully Unicode-compatible Bangla operating system in Windows and Linux platforms remain serious issues. Except for newspapers, there exists hardly much of any archived Bangla corpus. Most importantly, there is a lack of coordination and integration not only among the BLP research groups but also between how it is being pursued often differently in Bangladesh and India.

Mohammad A. Karim
Old Dominion University, USA

Mohammad Kaykobad
Bangladesh University of Engineering and Technology, Bangladesh

Manzur Murshed
Monash University, Australia

November 2012