Chapter 12
Bengali (Bangla) Information Retrieval

Debasis Ganguly
Dublin City University, Ireland

Johannes Leveling
Dublin City University, Ireland

Gareth J. F. Jones
Dublin City University, Ireland

ABSTRACT

This chapter introduces Bengali Information Retrieval (IR) to students by explaining the fundamental concepts of IR such as indexing, retrieval, and evaluation metrics. This chapter also provides a survey of and comparisons between various Bengali language-specific methodologies, and hence can serve researchers particularly interested in the state-of-the-art developments in Bengali IR. It can also act as a guideline for application developers on how to set up an information retrieval system for the Bengali language. All steps for creating and evaluating an information retrieval system are introduced, including content processing, indexing, retrieval models, and evaluation. Special attention is given to language-specific aspects of Bengali information retrieval. In addition, the chapter discusses cross-lingual information retrieval, where queries are entered in English with an objective to retrieving Bengali documents.

INTRODUCTION

The World Wide Web is growing at an astounding rate, both in terms of the volume of content available and the number of individuals with access to the Web. The majority of professionally authored content is typically still produced in English. However, potentially valuable content is being created in a multitude of other languages. Machine-Translated (MT) versions of this content may be generated for some languages, but the limited availability of high quality MT means that this is not always possible for all language pairs. At the same time, access to digital information is becoming more important and, due to the increase in amount and diversity of data, more difficult.

DOI: 10.4018/978-1-4666-3970-6.ch012

Bangla (or Bengali), one of the more important Indo-Iranian languages, is the sixth-most popular in the world and spoken by a population that now exceeds 250 million, of which more than 193 million are native speakers[1]. Geographical Bangla-speaking population percentages are as follows: Bangladesh (over 95%), and the Indian states of Andaman and Nicobar Islands (26%), Assam (28%), Tripura (67%), and West Bengal (85%). The global total includes those who are now in diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, United Arab Emirates, United Kingdom, and the United States. However, compared to languages such as English, Bengali is a low-resource language, i.e. the range of natural language processing tools and linguistic resources is still small. For example, the English Wikipedia comprises almost 4 million articles while the Bengali Wikipedia has little more than 20,000 articles. Research on language-specific aspects of Bengali information retrieval is still in its infancy.

The process of Information Retrieval (IR) can be broadly defined as satisfying a user's information need by retrieving relevant documents from a collection of documents, where relevant means that a document contains the information necessary to satisfy the user's need. IR encompasses search on collections of text documents, either structured or unstructured, but also search over collections of spoken recordings, music and other audio data, images and video. Most IR approaches still focus on text retrieval or on text annotations of multimedia data. In designing an IR System (IRS), the key issues are to determine methodologies for: (1) document representation; (2) query representation; and (3) a similarity measure for comparing a query with documents. Language-specific adaptations are particularly required for the first two components, namely finding suitable representations for the documents and queries.

Early IR research focused on development of techniques for English (e.g. at TREC, http://trec.nist.gov). More recent work has explored development of effective techniques for European (e.g. at CLEF, http://clef2012.org/) and Asian languages (e.g. at NTCIR, http://research.nii.ac.jp/ntcir/index-en.html). The inception and success of FIRE, the Forum for Information Retrieval and Evaluation (http://www.isical.ac.in/~clia/) has shown the interest in the development and automatic evaluation of IRS for Indian languages and has resulted in the creation of additional language resources for Bengali to aid Natural Language Processing (NLP) and IR.

The rest of this chapter is organized as follows: Section 1 (The Search Process) introduces the search process in general. Section 2 (Content Preprocessing) presents different normalization and conversion methods for transforming a set of files into a document collection. Section 3 (Indexing) introduces content processing and the most widely used index structure, the inverted file. Section 4 (Retrieval Models) presents alternative retrieval models. Section 5 (Enhancing IR effectiveness by Relevance Feedback) describes methods to improve IR effectiveness. Section 6 (Cross-Language IR) presents approaches to cross-language IR. Section 7 (Evaluation) presents evaluation metrics and the most important research results from evaluation initiatives and Section 8 (Bengali IR Benchmarking) reviews the various Bengali IR tasks undertaken till date in open evaluation forums. This is followed by Section 9 (Towards Best Practises), which provides a comparison between various approaches undertaken for Bengali IR. Section 10 (Tools and Resources) provides a list of tools and resources useful for starting Bengali IR experiments. Finally, Section 11 (Conclusion) concludes this chapter.

1. THE SEARCH PROCESS

From a user's point of view, the search process starts with an information need and comprises submitting a search query which attempts to express this need to an IRS. The IRS responds with