Bengali and Hindi to English CLIR Evaluation

Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar

Abstract

Our participation in CLEF 2007 consisted of two Cross-lingual and one monolingual text retrieval in the Ad-hoc bilingual track. The cross-language task includes the retrieval of English documents in response to queries in two Indian languages, Hindi and Bengali. The Hindi and Bengali queries were first processed using a morphological analyzer (Bengali), a stemmer (Hindi) and a set of 200 Hindi and 273 Bengali stop words. The refined hindi queries were then looked into the Hindi-English bilingual lexicon, 'Shabdanjali' (approx. 26K Hindi words) and all of the corresponding translations were considered for the equivalent English query generation, if a match was found. Rest of the query words were transliterated using the ITRANS scheme. For the Bengali query, we had to depend mostly on the translietrations due to the lack of any effective Bengali-English bilingual lexicon. The final equivalent English query was then fed into the Lucene Search engine for the monolingual retrieval of the English documents. The CLEF evaluations suggested the need for a rich bilingual lexicon, a good Named Entity Recognizer and a better transliterator for CLIR involving Indian languages. The best MAP values for Bengali and Hindi CLIR for our experiment were 7.26 and 4.77 which are 0.20 and 0.13 of our monolingual retrieval, respectively.