Query Translation and Expansion for Searching Normal and OCR-Degraded Arabic Text

Tarek Elghazaly and Aly Fahmy

Faculty of Computers and Information, Cairo University, Giza, Egypt,
{t.elghazaly, a.fahmy}@fci-cu.edu.eg

Abstract. This paper provides a novel model for English/Arabic Query Translation to search Arabic text, and then expands the Arabic query to handle Arabic OCR-Degraded Text. This includes detection and translation of word collocations, translating single words, transliterating names, and disambiguating translation and transliteration through different approaches. It also expands the query with the expected OCR-Errors that are generated from the Arabic OCR-Errors simulation model which proposed inside the paper. The query translation and expansion model has been supported by different libraries proposed in the paper like a Word Collocations Dictionary, Single Words Dictionaries, a Modern Arabic corpus, and other tools. The model gives high accuracy in translating the Queries from English to Arabic solving the translation and transliteration ambiguities and with orthographic query expansion; it gives high degree of accuracy in handling OCR errors.

Keywords: Query Translation, Orthographic Query Expansion, Cross Language Information Retrieval, Arabic OCR-Degraded Text, Arabic Corpus.

1 Introduction

The importance of Cross Language Information Retrieval (CLIR) appears clearly when we consider a case like the Library of Congress [1] which has more than 134 million items and approximately half of the library's book and serial collections are in 460 languages other than English. When people like to retrieve the whole set of documents that represent some interest, they have to repeat search process in each language. Furthermore, as a big number of books and documents are available only in print especially the Arabic ones, they are not 'full text' searchable and they need applying the Arabic OCR process whose accuracy is far from perfect [2]. The goal of this paper is to provide a solid English/Arabic query translation and expansion model to search both normal and OCR-Degraded Arabic Text.

The outline of this paper is as follows: The previous work is reviewed in Section 2. The proposed work is presented in the next sections. Arabic words formalization, normalization and stemming are presented in Section 3. Corpus and Dictionaries are presented in Section 4 and 5. In Section 6 & 7 the work done for CLIR through Query Translation and expansion respectively is detailed, followed by the experimental results and the conclusions in Sections 8 & 9.