Arab J Sci Eng (2015) 40:3211–3232
DOI 10.1007/s13369-015-1797-4

RESEARCH ARTICLE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

Enhanced Arabic Document Retrieval Using Optimized Query Paraphrasing

Abeer Al-Dayel[1] · **Mourad Ykhlef**[2]

Received: 4 November 2014 / Accepted: 23 July 2015 / Published online: 8 August 2015

Abstract Query paraphrasing aims to construct a better formulation of user queries in order to enhance retrieval. Formulating search queries remains complicated for a subset of Web users. In a typical situation, a user will not receive satisfactory results from the submitted search query and will subsequently attempt different query paraphrases. The Arabic vocabulary is rich in synonyms and hyponyms. Such richness of synonyms makes automation of the paraphrasing technique crucial for Arabic information retrieval systems in order to facilitate the process of paraphrasing synonyms. In this article, we propose an enhancement for Arabic information retrieval using a query paraphrasing technique. Furthermore, two query paraphrasing optimization techniques are proposed to overcome the time complexity and exhaustive calculation of existing query paraphrasing techniques. One of these techniques uses a genetic algorithm (GA–QP), and the other employs the artificial bee colony algorithm (ABC–QP). The performance of these two algorithms is compared. ABC–QP shows an improvement in Arabic information retrieval performance compared with the genetic algorithm query paraphrasing system.

✉ Abeer Al-Dayel
aabeer@ksu.edu.sa

Mourad Ykhlef
ykhlef@ksu.edu.sa

[1] Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia

[2] Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia

Keywords Arabic language · Arabic information retrieval · Query paraphrasing · Genetic algorithm · Artificial bee colony

1 Introduction

Query paraphrasing is a frequent routine in information retrieval applications; search engine users tend to reformulate the search query until the required information has been retrieved. Reformulating a search query is not a simple task: The user has to test different combinations of query terms in order to identify the best query paraphrase. The query paraphrasing technique aims to assist users by providing the best paraphrase for a search query to accurately retrieve the documents that relate to the specific information required. Query paraphrasing can be defined as the restatement of a query by replacing words with their synonyms and hyponyms and changing the word order, while producing no change in query meaning [1,2].

Paraphrasing has been used in many applications, such as multi-document summarization, text generation, and information retrieval [3]. Research on paraphrasing can be classified into three methods: recognition, generation, and extraction. The recognition method checks whether two sentences are paraphrases. The generation method produces paraphrases of the input sentence. The extraction method extracts the paraphrasing rules (e.g., "X wrote Y" ↔ "Y was authored by X") or similar patterns from corpora. An example is [3], where paraphrases are extracted from a parallel corpus, and similarly [4], presented an extraction procedure for obtaining paraphrases from news articles.

In this article, we focus on generation methods. There are three main approaches to paraphrase generation. The first approach considers paraphrase generation as a machine

translation problem. The second approach considers existing machine translation engines as black boxes and translates each input sentence into a pivot language, and then back into the original language [5]. The third approach uses thesauri to generate lexical paraphrases where the content words are replaced with synonyms [6].

The main difference between query paraphrasing and other query-refinement techniques (relevance feedback, etc.) is that query paraphrasing generates alternative lexical paraphrases from the original query. In addition, in contrast to the selection heuristics used in some query-refinement techniques, query paraphrasing uses corpus-based information in the context of the entire query to calculate the score of a paraphrase and select the best paraphrases [6].

The retrieval process is highly influenced by the query language and document collection. Different linguistic characteristics have an impact on the accuracy of the information retrieval. The Arabic language has been classified as one of the top ten languages on the Internet, with 65.4 million Arabic language users [7]. It is one of the largest members of the Semitic language branch.

In the Arabic language, words are composed from roots that use Arabic schemes, and then the appropriate suffix or prefix is added. For example, from the root K - T - B (ك ت ب), lemmas are formed using Arabic schemes, e.g., (“مكتبة: إسم مكان”, “a library”), and by adding the prefix “فب ” and the suffix “هم ” to the lemma, we obtain the word (“فبمكتبتهم ”, “and in their library”) [8–10].

In our study of Arabic user attitudes toward search engines [11], we found that query paraphrasing is a common search strategy. Approximately 84 % of users tend to paraphrase the search query when they receive no relevant search results. In addition, this study addresses user-paraphrasing skills by asking participants to generate paraphrases for the query “number of people in Riyadh” “عدد سكان مدينة الرياض ”; only 12 % of the users were able to produce all possible paraphrases for the original query. This provides an insight into the need for automatic query paraphrasing in Arabic information retrieval.

The contribution of this article is twofold. First, it proposes a framework for enhancing the retrieval effectiveness of Arabic information retrieval to search for Arabic documents through query paraphrasing techniques. Using query paraphrasing allows the information retrieval system to understand and possibly predict user intent, thereby providing the required assistance at the proper time and thus helping users locate information more effectively and improving their Web search experience. Second, this article presents an enhanced query paraphrasing technique that uses two optimization algorithms, a genetic algorithm and the artificial bee colony algorithm to assess the generation of query paraphrases while reducing complexity. These developments potentially help us approach the development of a truly intelligent information retrieval system. In particular, our computational results reveal that the automation of paraphrasing initial user queries improves Arabic document retrieval performance.

1.1 Related Works

The computational linguistics community has been addressing the topic of paraphrasing over the last decades [12]. Because of the boundless nature of the paraphrasing phenomenon, determining how to characterize paraphrasing within the computational linguistics field has become a challenging issue. In order to answer the questions of where to draw the boundary between paraphrases and non-paraphrases, the work in [12] presents paraphrase linguistic characterization in order to provide Natural Language Processing (NLP) with a more solid base for the development of methods that address paraphrasing. In this work, the authors propose a three-level typology of 24 paraphrase types. Our paraphrasing technique belongs to the lexical-based category, i.e., the same-polarity substitution (or synonymy substitution) type. One of the early works on paraphrasing is [13]. In this study, a paraphrase generation system was developed for implementation in the meaning-text model whose purpose is to establish correspondence between meanings.

Three approaches to lexical paraphrasing have dominated the literature. One approach acquires paraphrases from dictionaries, such as WordNet [1]. In this work, the authors use synonymous paraphrasing of the text based on WordNet synonymy data and Internet statistics of stable word combinations (collocations). The authors of [6] used WordNet and part-of-speech information to propose synonyms for the content words in the queries. Similar to [14], the authors use lexical paraphrasing based on WordNet synonymy data to enhance document retrieval and node identification. In [15], the authors used three information sources to generate lexical paraphrases of Internet queries: WordNet, Webster’s online dictionary, and a combination of the Webster-based thesaurus and WordNet. The second approach collects lexical paraphrases from monolingual or bilingual corpora. In the research described in [16], the authors extracted lexical paraphrases using multiple resources: a monolingual dictionary and corpus, and a bilingual corpus. The authors of [17] used a parallel corpus to identify paraphrases from a corpus of multiple English translations of the same source text. This study [18] used Web as source for generating paraphrases to a given query. The third approach is based on using query logs for extracting paraphrases automatically by acquiring lexical context-specific paraphrases from the Web; this approach is exemplified in [19–21].

The three common methods used in Arabic query refinement are relevance feedback, pseudo-relevance feedback,