Current Approaches in Arabic IR: A Survey

Mohammed Mustafa, Hisham AbdAlla, and Hussein Suleman

Department of Computer Science, University of Cape Town,
Private Bag X3, Rondebosch, 7701, South Africa
{mmustafa,hisham,hussein}@cs.uct.ac.za

Abstract. Arabic information retrieval is a popular area of research. This paper presents the current state-of-the-art in Arabic Information Retreival (IR) approaches. Moreover, it provides general guidance for open research areas and future directions.

1 Introduction

Modern Standard Arabic (MSA)is one of the most widely used languages in the world. Previous works in the field of Arabic IR summarize five features of the Arabic language that cause it to be a significant challenge for both information retrieval and search engines: **Orthographic Variations** , its complex **Morphology** ,**Diacritization**, prevalence of **Irregluar / Broken Plural**, and **Synonyms**. The following are some examples for these challanges: Orthographic Variations - either بُورتسُودَان or بُورسُودَان for Port-Sudan city; Morphology - لَنَهدِيَنَّهُم (meaning: we will surely guide them) is a word that consists of different parts; Diacritization - الشَّعر, means hair, while الشِّعر means poem; Broken Plural - the word قَائِد (meaning: leader) changes to قُواد (meaning: leaders) ; and Synonyms - the word أَسَد (meaning:lion) may have different synonyms to lion according to its age.

2 Current Solutions

The above challenges have been solved to different levels. Preprocessing includes removal of non-characters, normalization and removal of stopwords. The non-character removal step [1] includes the removal of punctuation marks, diacritics and kasheeda (the word السُودَان (Sudan) can be written with kasheeda as الـسُودَان). Normalization is used to represent different forms of a letter with a single Unicode representation as in HAMZA (أَ, إِ) and MADDA (آ). For stopwords and thier phrases most exisiting methods use dictionaries or software tools. Tokenization is used intensively in Arabic IR by using different techniques.

A number of studies have been devoted to different approaches of incorporating morphology: stem, root, light stem and no-stemming as well as using non-rule based statistical or n-gram models. Stemming affects all problems mentioned. Kadri and Nie [3] proposed a new stemming technique based on linguistic removal of affixes. Larkey et al[4] presented the best light stemmers (*light10*). Mansour et al [5] proposed a new technique based on Arabic grammatical rules.

For the problem of broken plural, the most used algorithm was proposed by Goweder et al [2]. Other complementary techniques are in regional variation disambigiation [1]. Also, Abdelali et al [1] presented a query expansion mechanism that has the ability to automatically select a corpus related semantically to the query.

3 Current Challenges

Arabic IR has a long road ahead. Major problems of some current Arabic search engines are that queries can only retrieve exact matching documents. Researchers need to devise more effective solutions for Arabic IR. Lack of resources for testbeds, translation and query expansion is one of these challenges. Regional variations, spelling variations and machine translations are contributors to problems in this field. Combinations of resources such as dictionaries, query logs and Web mining techniques is needed for both translation and query expansion. Preprocessing techniques need to be united. Tokenization still has its challenges due to clitics and ambiguity. Linguistic stemming is a very rich area of research. A straightforward algorithm is needed for broken plurals. Automatic diacritization also needs further work. It is expected that the next generations of Arabic search engines will be based on Arabic morphology.

References

1. Abdelali, A. (2006), "Improving Arabic Information Retrieval using Local variations in Modern Standard Arabic", PhD. dissertation, New Mexico Institute of Mining and Technology.
2. Goweder, A., Poesio, M., and De Roeck, A. (2004), Broken plural detection for Arabic information retrieval, in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, pp:566-567.
3. Kadri, Y., and Nie, J. Y. (2006), "Effective stemming for Arabic information retrieval". The challenge of Arabic for NLP/MT Conference, The British Computer Society. London, UK.
4. Larkey, S. L., Ballesteros, L., and Connell, E. M. (2005), Light stemming for Arabic information retrieval. Arabic Computational Morphology: Knowledge-based and Empirical Methods.
5. Mansour, N., Haraty, A. R., Daher, W., and Houri, M. (2008), An auto-indexing method for Arabic text, Information Processing and Management, 44(4), pp:1538-1545.