Query Translation Methods to Enhance Arabic Information Retrieval

Awni Hammouri

Department of Information Technology Mutah University 61710 Mutah, Karak, Jordan

Summary
Remarkable developments in Information Storage and retrieval technologies have driven the need for information to be retrieved in any language. A rich user experience demands that the Information search isn't language dependent. The predominant language over the web still being English, the goal here is to retrieve information written in a language other than the queried language i;e, Cross Language Information Retrieval (CLIR). CLIR enhances the capacity of search and retrieval technologies resulting in an enhanced user experience. CLIR relies on various translation techniques to retrieve information. Highly malleable languages like Arabic consistently challenge these translational techniques to a great extent. In this study, an overview of cross language information retrieval techniques based on Arabic is presented and its impact on Arabic CLIR schemes is objectively evaluated.

Keywords
Cross-language information retrieval, complex morphology; stemming; stopwords removal; Parallel Corpora-Based Approach;

1. Introduction

Nowadays, there is major shift in the way information is accessed over the internet. Web users; prefer information to be available in their native language. With English still being the dominant language over the internet, there is a growing demand for translation tools. These translation tools retrieve the information stored in a language other than queried language [1]. Although, the translation tools have been existence for decades, many of the proceedings or specialized journals have concentrated on a set of languages beyond English. German, Swedish, Dutch, Spanish, Finnish, Italian and Dutch etc. have found interest in CLEF, whereas Arabic, Chinese and French etc. found solace in TREC.

In this paper, we study the existing approaches in Arabic CLIR and its compelling contribution to Cross Language Information Retrieval systems. The miracle of the Holy Quran is embedded in Arabic language. The history of Islamic Civilization and the Sunnah of the Last Prophet (pbuh) has been right from day one recorded in Arabic. Most of the Universities around the world and especially in the west have heavily relied on Arab books on Medicine and Science. Being one of the six official languages of the United Nations [2] and the mother tongue of billions over the globe, Arabic contains 28 letters, 16 of them have one dot, two or three dots. There are many varying ways of writing and it is written from right to left.

Various adaptations, of the Arabic script have additional letters and some adaptations have ejected few letters [3]. Information Retrieval Systems encounter quite a few problems when dealing with Arabic language because of its richness. Words in Arabic language are arranged in three segments: nouns, verbs and particles with bulk of them extracted from tri, quad, or pent-literal roots.

The swelling demand for Arabic CLIR drove the need for standardization of tools. In 2000, CLEF and TREC introduced standards for evaluation of existing tools. Regrettably, after 2002, no new test collections were composed and the aged assortment has now turned weak. In 2015,[4] have discussed how a contemporary standard Arabic CLIR collection can encourage researchers to extend their work. Various researches have surveyed the CLIR territory.

In 2012, critical approaches in Arabic Machine Translation (MT) were rehashed by Alqudsi et.al [5] and their advantages and disadvantages discussed. The techniques for query or document translation in CLIR were surveyed but as a peculiar case, approaches concerning Arabic language were omitted by Zhou et.al [6]. Precise retrieval issues specific to Arabic CLIR such as Arabic speech, social and web search, Arabic Document Image retrieval etc. were addressed in 2013 [7]. Natural language processing of Semitic one's such as Amharic, Hebrew, Maltese and Arabic were surveyed by Zitouni [8] in 2014.

2. Arabic information retrieval tasks

A. Complex Morphology

A highly improbable two level finite state morphology was proposed by Beesley [9] to circumvent the issues caused by complex morphology in Arabic CLIR. Though occasional, this approach produced umpteen encouraging analyses. Three methods to solve this problem were proposed. A contextual method or corpus statistics in which primarily possible results were achieved by following a deep complete analysis [10]. Few rule based approaches were adopted to eliminate regular suffixes and prefixes from the

Manuscript received November 5, 2017
Manuscript revised November 20, 2017

morphology in the second method with this method being quite similar to light stemming in Arabic [11].
A light stemmer was used to find words which were then classified into classes based on their distributional features. This was the corpus based clustering third method. Despite testing these methods as a relative solution to back Arabic Information Retrieval, an optimal solution hasn't been reached. Recently available highly accurate tools such as MADA, AMIRA and MADAMIRA overcome the drawbacks of existing approaches with nearly 99% accuracy [12].

B. Transliteration

Generally a preprocessing step, Transliteration involves phonetic mapping from one character set. Many Arabic Transliteration schemes were proposed by researchers which survey the standard encoding array of Arabic characters for computers. According to [4] there are no standard transliteration schemes to be pursued in NLP in Arabic IR systems.

C. Tokenization

In Information retrieval exercises, a neighboring sequence of n items extracted from a given sequence of speech or text is defined as an n-gram. These sequences may be letters, phonemes, base pairs, syllables or words collected from a speech or text corpus. n-grams are from a text or speech corpus. An n-gram in addition to being comprehensive and robust, also provide language autonomy and simplicity. They work relevant to Arabic IR tasks, such as transliteration, root identification, and search of Arabic texts, and have been looked upon by researchers with deep interest.
Conducive performances of n-gram–based retrieval for Arabic IR were exploited in TREC 2002 [13]. n-grams of more than one length were exploited in a hybrid approach employing tokenization within the same term space. Few Arabic symbols or characters which didn't exist in the group of 28 letters were either eliminated or replaced. For instance, HAMZA (أ) , MADDA (آ), and any remaining Arabic letters were eliminated. ALEF MAKSURA (ى) were generalized to YEH (ي) and TEH MARBUTA (ة) to TEH (ت). Multiple length n- grams produced a 10 % improvement over single-length n-grams when measured using mean average precision (MAP).
Bigrams were used to extract roots from Arabic words by using similarity measures such as The Manhattan Distance Measure and Dice's Similarity Measure. For evaluation and testing purposes, the Holy Quran which mainly contains traditional words and Saudi Arabian National Computer Conference proceedings with a corpus of 242 abstracts served as reference. The results of experiments showed The Manhattan Distance Measure being outperformed when an approach consisting of n- grams and Dice's Measure was used [14].
Further, Arabic Documents when indexed using trigrams demonstrated better results compared to a vector space model with the cosine coefficient, Dice's coefficient and TF∗IDF weighting [15]. Based on words and characters, n-grams were exploited as a representation technique. The n-gram approach is quite insufficient when retrieving and indexing legal Arabic documents. This limitation may be compensated by applying a linguistic approach using ontology or a legal thesaurus.

D. Stemming

Despite that various linguistic and light stemmers for the Arabic Language were proposed and tested, in line with the prior state of art tools, they still suffer from many weaknesses vis-à-vis; it may sometime fail to remove the word affixes and as a result fail to select the relevant root. Retrieval effectiveness may also be decreased significantly in an Arabic Language IR system. Root Dictionary needs to be continuously updated to ensure revealed words are aptly stemmed [4].
Several works reveal a very less difference between using roots and stems when assessed for IR effectiveness. A rule based light stemmer was proposed which takes advantage of Larkey stemmer to decide whether certain characters belong to the actual word. The algorithm utilizes a predefined set of rules to find solutions for some ambiguous cases. It breaks down the plural forms into singular ones and groups' words sharing the same sense in a regular form [16].
In a recent approach, a root stemmer was proposed and assessed to enhance Arabic IR effectiveness for Modern Standard Arabic (MAS). This simple Arabic stemmer utilizes the internal morphological structures such as the roots, affixes and patterns of Quranic words to produce Arabic morphological rules. The root of the majority of MSA vocabulary is then enclosed by using a dedicated lexicon [17].

A. Stop-word removal

An effective stopword algorithm plays an important role in many NLP applications such as IRS, question answering systems, spelling normalization and stemming and stem weighting. These words on one hand, while influencing the IR task, tend to appear in a very high frequency thereby decreasing the effect of frequency discrepancies among less common words and as a result the weighing process gets affected. The removal of stopwords also decreases the document length and if not executed accurately, the search effectiveness also drops. A sizable volume of futile processing occurs due to the presence of stopwords by virtue of their occurrence frequency and empty semantic content. The mechanism of removing such stop words by building a list is known as stoplist. These are of two types.

Domain dependent – in which corpus statistics are used to locate those words by specifying few syntactic classes for inclusion or those with the greatest frequency. A hybrid approach which mixes certain syntactic classes with corpus statistics was also proposed to generate stoplists.

3. Arabic clir translation techniques

A. Dictionary-Based Approaches.

A Machine Readable Dictionary was used for Arabic English CLIR[18]. The problem with this approach is the translation ambiguity affiliated to the resources. Three approaches namely every-match(EM) method, first-match(FM) method and two-phase method(TP) were proposed to overcome this. The first approach suffered from ambiguous translations when it was evaluated using simple word by word translation on English Arabic retrieval performance. To reduce the ambiguity, the second approach retains only the first match translation per query term. to loosen the inherent restrictions of the first match method and to overcome the limits of these methods, the two phase method was introduced which uses some but not all of the translations of the Arabic term. The two phase method produced acceptable retrieval effectiveness without using complex resources.

Bilingual Dictionaries were also used for evaluating Arabic English CLIR systems with the sole aim being solving the ambiguous translation problem. Though the simple nature of a bilingual dictionary makes it a good choice for CLIR systems, but query translation performed on bilingual dictionaries reduced the CLIR performance. Varied factors contributed to this cause. It was confirmed that a major challenge associated with CLIR based on bilingual dictionary approaches is the missing coverage of few terms especially when a query containing technical terms is used for which a general dictionary responds cordially. Terminological or idiomatic dictionaries may also be used to help word by word translation methods achieve correct translations of words or phrases which encompass more than a word which regrettably is difficult to find.

B. Parallel Corpora-Based Approaches.

With several deficiencies limiting the use of Dictionary-based methods in Arabic CLIR because of many reasons such a limited coverage, difficulty in choosing a convenient translation from the available set of alternatives given by the dictionary during the query translation, statistical translation models schooled on a corpus of parallel texts collected from the web were beginning to find audience.

A word translation disambiguation method based on parallel corpora and matching schemes was proposed [19] which took advantage of an enormous bilingual corpus and statistical co-occurrence. This was used to identify convenient sense for query translation terms. A suitable translation of every query was found using the cohesion between the query, its possible translation and a similarity score measure. Specific properties of Arabic Language also influenced the correct match. In clear terms, a naïve Bayesian algorithm was exploited for the translation process. This approach fared better for long queries and it was worse for short queries. This was due to the presence of those features which were extracted from a few frequently appearing query terms within the context of dissimilar senses.

An automatic translation process driven Arabic CLIR system called the (Mult) iSearcher, supported by an automatic reduced the user's task by verifying each and every alternative for the query term [19]. Statistical methods helped improve both the disambiguation steps and the translation automatically. A certain level of confidence based on the ranked translation alternatives given by the tool is provided to the user. The extensive free texts available on the web are suited to such tools. Named entity recognition approaches further enhance the performance of this tool. Though these approaches provide acceptable performance they are also tested by many hurdles ranging from domain restrictions to the unavailability of parallel texts with the second one being a more serious problem.

C. Machine Translation -Based Approaches.

The grace of Arabic Language is that the sentences can be communicated in various transformations having identical implications. The three elements subject, verb and object classify an Arabic sentence in four varieties such as VOS, SVO, SOV AND VSO. It is quite difficult for a Machine Translation system to accumulate all CLIR user needs. The application of Machine Translation system degrades the retrieval efficiency (run time performance) and translation quality.

A standard phrase based Arabic to English statistical Machine Translation tool was proposed to recourse the semantic ambiguity of Arabic Language by integrating a local discriminative phrase selection model [20]. This increased the accuracy of phrase selection and a crucial enhancement was found in the full translation task at the syntactic, lexical and semantic levels.

The existing methods for Arabic Machine Translation were also presented by [5]. The authors here conclude that a major portion of Machine Translation systems focus on translating news and official texts. Translation being a compelling requirement on the web and despite the steady growth of statistical machine translation the Machine Translation systems do not meet the user requirements. In order to improve the CLIR effectiveness and taking into account the specifics of Arabic language what is required nowadays is efficient English to Arabic and Arabic to English Machine Translation system.

D. Approaches Combining Arabic Translation Resources.

CLIR effectiveness may be improved by connecting different query specific techniques into a single structured query [21]. Four different CLIR tasks were assessed using this approach with English being the query language. A higher effectiveness and statistically crucial enhancements such as uniform and task specific weighing were found over other combination techniques.

An Arabic – French CLIR query translation method was also proposed by advantaging both parallel corpus and a bilingual dictionary [22]. A suitable query translation for every source term is chosen by mapping two semantic networks. The first semantic network houses all potential senses of each query term whereas a set of sentences extracted from retrieved documents are enclosed in the second semantic network. A significant improvement in the quality of translation was observed.

4. Conclusion and future directions

Arabic CLIR, its related work, together with the preprocessing techniques and specificities of Arabic Language has been reviewed in this paper. Consequently, some current problems are identified and solutions to these and some perspectives are proposed as future work to boost scientific research in this field. Many works on Arabic CLIR not cited here have been published in different research fields and communities.

Compared to other languages, there are a few corpora in the Arabic Language. A dictionary-based CLIR is very helpful as many of the techniques are now well demonstrated and applied. As dictionaries are only one source of translation knowledge, the ambiguity that the Every-Match (EM) method introduces is very high. The performance may also be approximated to half of the monolingual retrieval.

Many senses which may be inappropriate to the query are one of the factors affecting it. A single word may have more than one translation with many senses. Arabic-English CLIR based Machine readable Dictionary is a cost effective option when compared to the various other approaches. The Arabic letters being highly ambiguous in nature are hard to interpret as well. A broad and extensive work is required in the area of spelling normalization and mapping. Lack of high coverage bilingual dictionaries and corpora have also contributed to under developed Arabic translation knowledge.

References

[1]. F. C. Gey, N. Kando and C. Peters. 2005. Cross-language information retrieval: The way ahead. Information Processing and Management 41, 3, 415–431.

[2]. Khatib, Ahmed Shafiq,1997," terminological specifications and applications in the Arabic language", cultural fifteenth season of the Arabic Language Academy of Jordan, Amman, Jordan, pp. 177- 213.(Arabic)

[3]. N. Habash. 2010. Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[4]. Bilel Elayeb and Ibrahim Bounhas. 2015. Arabic cross-language information retrieval: A review. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 15, 3, Article 18 (December 2015)

[5]. A. Alqudsi,N. Omar, and K. Shaker. 2012. Arabicmachine translation: A survey. Artificial Intelligence Review42, 4, 549–572.

[6]. D. Zhou, M. Truran, T. J. Brailsford, V. Wade, and H. Ashman. 2012. Translation techniques in cross language information retrieval. ACM Computing Surveys 45, 1, 1–44.

[7]. K. Darwish and W. Magdy. 2013. Arabic information retrieval. Foundations and Trends in Information Retrieval 7, 4, 239–342.

[8]. I. Zitouni (Ed.): 2014. Natural Language Processing of Semitic Languages. Springer-Verlag, Berlin, Germany.

[9]. K. R. Beesley. 1998b. Arabic morphological analysis on the Internet. In Proceedings of the International Conference on Multi-Lingual Computing Arabic.

[10]. K. Darwish. 2002. Building a shallow Arabic morphological analyzer in one day. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. 47–54.

[11]. M. Aljlayl, O. Frieder, and D. A. Grossman. 2002. On Arabic-English cross-language information retrieval: A machine translation approach. In Proceedings of the 2002 International Symposium on Information Technology. IEEE, Los Alamitos, CA, 2–7.

[12]. A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R.M. Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14).

[13]. P. McNamee, C. D. Piatko, and J. Mayfield. 2002. JHU/APL at TREC 2002: Experiments in filtering and Arabic retrieval. In Proceedings of the 11th Text Retrieval Conference (TREC'02). 358–363.

[14]. I. Hmeidi, R. Al-Shalabi, A. T. Al-Taani, H. Najadat, and S. A. Al-Hazaimeh. 2010. A novel approach to the extraction of roots from Arabic words using bigrams. Journal of the American Society for Information Science and Technology 61, 3, 583–591.

[15]. M. Rammel, M. Sanan, and K. Zreik. 2011. Improving Arabic information retrieval system using n-gram method. WSEAS Transactions on Computers 10, 4, 125–133.

[16]. M. Ababneh, R. Al-Shalabi, G. Kanaan, and A. Al-Nobani. 2012. Building an effective rule-based light stemmer for Arabic language to improve search effectiveness. International Arab Journal of Information Technology 9, 4, 368–372.

[17]. M. Algarni, B. Martin, T. Bell, and K. Nehsatian. 2014. Simple Arabic stemmer. In Proceedings of the 23rd ACM

International Conference on Information and Knowledge Management (CIKM'14). ACM, New York, NY, 1803–1806.

[18]. M. Aljlayl and O. Frieder. 2001. Effective Arabic-English cross-language information retrieval via machine readable dictionaries and machine translation. In Proceedings of the 2001 ACM International Conference on Information and Knowledge Management (CIKM'01). ACM, New York, NY, 295–302.

[19]. A. Farag and A. Nurnberger. 2013. Translation ambiguity resolution using interactive contextual information. In Computational Linguistics. Studies in Computational Intelligence, Vol. 458. Springer, 219–240.

[20]. C. Espana-Bonet, J. Gimenez, and L. Marquez. 2009. Discriminative phrase-based models for Arabic machine translation. ACM Transactions on Asian Language Information Processing 8, 4, Article No. 15.

[21]. F. Ture and E. Boschee. 2014. Learning to translate: A query-specific combination approach for cross-lingual information retrieval. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14). 589–599.

[22]. S. Mallat, M. A. Ben Mohamed, E. Hkiri, A. Zouaghi, and M. Zrigui. 2014. Semantic and contextual knowledge representation for lexical disambiguation: Case of Arabic-French query translation. Journal of Computing and Information Technology 22, 3, 191–215.

Awni Hammouri received the B.S. degree in Computer Science from Yarmouk University, Jordan in 1985. M.S. and Ph.D. degrees in Computer Science from Illinois Institute of Technology, Chicago, Illinois, U.S.A in 1989 and 1994 respectively. He is now with Mutah University, Jordan.