Information Retrieval from Unstructured Arabic Legal Data

Imen Bouaziz Mezghanni(✉) and Faiez Gargouri

MIRACL Laboratory, ISIM Sfax, Sfax, Tunisia
imen_bouaziz_miracl@yahoo.com, faiez.gargouri@isimsf.rnu.tn
http://www.miracl.rnu.tn/

Abstract. Given the steady increase of published and stored information in the form of Arabic unstructured texts, current Information Retrieval (IR) systems must be able to suit the nature and requirements of this language for an accurate and efficient search. This paper sheds light on the challenges in Arabic IR (AIR) and proposes an approach for enhancing the process of AIR based on transforming these texts into structured documents in XML format through a document ontology as well as a set of linguistic grammars. The IR system hence is done on the XML documents. The aim of such system is to incorporate the knowledge on the document structure and on specific content elements in computing the relevance of an information element. A query expansion module mainly based on domain ontology as well as user profile is proposed for the enhancement of the search results.

Keywords: Information retrieval · Arabic information retrieval · Unstructured data · Structured data

1 Introduction

Over the past few years, the amount of electronic information is increasing tremendously by the rapid and continual flow of information through independent internet media. Dealing with this perceived explosion information requires Information Retrieval systems (IRS). The main goal of IRS is to retrieve the relevant and only the relevant documents in response to user's queries from mainly unstructured textual data.

Specialized techniques such as Text Mining specifically operating on textual data are becoming inevitable to extract information from such kind of texts. Text Mining techniques are dedicated to discover the implicit structure in the documents. The discovered structure is called Index which contains the most significant terms known as descriptors. Natural Language Processing (NLP) can then be seen as a powerful technology for the vital tasks of IR and knowledge discovery as it allows both facilitating descriptions of document content as well as presenting the user's query usually formulated as a set of natural language keywords.

R. Booth and M.-L. Zhang (Eds.): PRICAI 2016, LNAI 9810, pp. 44–54, 2016.
DOI: 10.1007/978-3-319-42911-3_4

In IRS, index terms play the connecting role between documents and user queries. Usually, the queries fail to match the index terms contained in the relevant documents. Dealing with Arabic legal data doubles the challenge of satisfying the user's need. The challenges arise from the language itself since Arabic is known by its complexity and likewise from the process of IR. Current IRS that access legal databases offer the possibility of a full text search, in which every term can act as a search term, and returns a ranked list of information answers. This answer list can be filtered through a deterministic fulfillment of extra conditions set by the structured information. Thus, on one hand we can search the free texts of the documents and on the other hand the structured information. However, few attempts have yet been made to exploit the structured information in retrieval models. For obtaining more relevant results, our idea is to search on structured documents instead of unstructured ones, we consider in fact that the underlying information should be in a structured form. Furthermore, the integration of a query expansion mechanism can bring up promising answer list and thus improve the retrieval performance.

The rest of the paper is organized as follows, where Sect. 2 addresses the challenges of Arabic language in IR and reviews related works dealing with Arabic IRS, Sect. 3 presents our system architecture for AIR, while Sect. 4 provides conclusions and presents plans for further work.

2 Challenges of Arabic Language in IR

Arabic is the official language of over 420 million people, across 22 countries throughout the Middle East, Europe, and Asia making it the sixth-most spoken language in the world. It is classified into three variants [Ibrahim et al. 2015]:

- Classical Arabic which is the form of the Arabic language used in literary texts and Holy Quran.
- Modern Standard Arabic (MSA) which is the standard and the literary variety of Arabic used in writing and in most formal speech.
- Colloquial Arabic which refers to the many national or regional varieties of spoken Arabic. These differ significantly from MSA and Classical Arabic, as well as from each other and are usually unwritten.

The retrieval of Arabic content, primarily written in MSA, is affected greatly by the properties of Arabic language. We expound these properties in the following section. In the remainder of this article, the given examples are extracted from our corpora. They are given in Arabic along with their English translation and their transliteration using Xerox Morphology System[1].

[1] https://open.xerox.com/Services/arabic-morphology/Consume/Morphological%20Analysis-218.

2.1 Arabic Specificities

Unlike Indo-European languages, Arabic includes specific pecularities effecting IR. First of all, Arabic is written in horizontal lines from right to left, and there are no capital letters in the Arabic alphabet making the task of text splitting into sentences difficult since capital letters are considered as cues for text segmentation.

Secondly, it is characterized by orthographic variations (The different typographical forms for one letter) such as ALEF (like ﻷ and آ and أ and ﺇ), YAA with dots or without dots (like ﻲ and ى) and also HAA (like ه and ة). Indeed, the substitution of one of these forms with another alter the meaning of the words. For example (علي/Ely) which indicates a proper noun and (على/ElY/on) which is a preposition.

Besides this, Arabic is a highly agglutinative language with the a rich set of clitics aglutinated to words. Its inflectional and derivational productions introduce a big increase in the number of possible word forms. For instance, prepositions (like ل/li/for), conjunctions (like و/w/and), articles (like ال/Al/the) and pronouns (like ه/h/he), can be agglutinated to nouns, adjectives, particles and verbs which causes several lexical ambiguities.

Moreover, long and complex sentences are frequently used in its discourses. Punctuation marks are narrowly used in such a way that we can simply find an entire long paragraph without any punctuation.

Furthermore, the lack of vowels in current texts and the multiplicity of the vowel forms lead to that a word can have different meanings which make the analysis and the comprehension of Arabic texts more difficult. For example, the word (وهن/whn/weakness) can correspond in English to the noun "Illusion" or to a conjunction (و/wa/and) followed by the pronoun (هن/hun a/they). In Arabic there are principal four categories of words which are noun, proper noun, verbs and prepositions. The absence of short vowels can likewise cause ambiguities within the same category or across different categories. For example, the word (بعد/bEd) corresponds to a preposition with the meaning "After", a Noun with the meaning "Remoteness", a Verb with the meaning "go away".

2.2 Arabic IR

Nowadays, the rising number of Arabic users as well as the amount of digital information available in document repositories has motivated researchers to develop many different Arabic IRS in order to enhance the Arabic documents retrieval process. Indeed, availability of information doesn't mean it is helpful as the user may not always find the needed information.

The classical IRS comprises three main processes, namely Indexing, Query processing and Matching. In indexing phase, documents are indexed using keywords that represent each document in the collection. Indexing mechanism is based on the "inverted index" in which information symbols and all documents containing that symbol are listed. Thus its structure is composed of two elements: the vocabulary defined by the set of all words in the text, and the term

occurrences defining the set of lists comprising the positions of appearance of each word in the text.

Then, the query that is entered by the user is generally pre-processed by the same algorithms used to select the index terms. Additional query processing as query expansion can be performed by using external resources like thesaurus, taxonomies or ontology. A query is reformulated to comply with the information retrieval model as well as to add other keywords or modify the weights of the existent words to achieve better search accuracy.

Finally, in the matching step, the query entered by user will be matched with index. The results of matching between existing index and query will be sorted based on the ranking algorithms which depend on their similarity.

Several information retrieval models exist, of which the most common models include the Boolean model, fuzzy model, and vector space model. Various studies have been elaborated intending the Arabic IR improvement. Among the recent studies, [Maitah et al. 2013] address improving the effectiveness of information retrieval system using adaptive genetic algorithm (AGA) under the vector space model, Extended Boolean model, and Language model in IR.

The main goal of the [Yousef and Khafajeh 2013] research is to design and build an automatic Arabic thesaurus using Local Context Analysis technique to improve the expansion process and so to get more relevance documents for the user's query. The results of this study showed that it improved the retrieval in a remarkable way better than the classical retrieval method. In 2014, the authors developed an algorithm for Arabic word root extraction based on N-gram [Yousef et al. 2014]. Morphological rules aren't used in order to avoid the complexity arising from the morphological richness of the language on one hand and the multiplicity of morphological rules in the Arabic language on the other hand.

[Mahgoub et al. 2014] introduced a query expansion approach using an ontology built from Wikipedia pages in addition to other thesaurus to improve search accuracy. The proposed approach outperformed the traditional keyword based approach in terms of both F-score and NDCG measures.

[Hanandeh and Mabreh 2015] is based on the Genetic Algorithm (GA) to improve the effectiveness such systems. This work uses the Vector Space Model (VSM) and the Extended Boolean Model (EBM) to compute the similarities between queries and documents. Two fitness functions are proposed in this paper: One as fitness function and the other as adaptive mutation. Then comparing each of these functions with a number of ratio mutations that have been introduced to get better results. The experimental results reveal that the proposed cosine function outperformed other fitness models.

A pre-retrieval (offline) method is proposed by [Mohamed 2015] to build a statistical based dictionary for query expansion which is based on a statistical methods (co-occurrence technique and Latent Semantic Analysis (LSA) model) to improve the effectiveness of the search result by retrieving the most relevant documents regardless of their dialect. The evaluation was done using the average recall (Avg-R), average precision (Avg-P) and average F-measure (Avg-F) and

the proposed method proved to be efficient for improving retrieval via expands the query by regional variation's synonyms, with accuracy 83 % in form of Avg-F.

A more recent research of [Atwan et al. 2016] aims to enhance an AIR by improving the processes in a conventional IR framework. An enhanced stopword list is introduced in the pre-processing level and several Arabic stemmers are investigated. In addition, an Arabic WordNet was utilized in the corpus and query expansion levels. A semantic information for the Pseudo Relevance Feedback was adopted. The enhanced Arabic IR framework was built and evaluated and demonstrated an improvement by 49 % in terms of mean average precision, with an increase of 7.3 % in recall compared with the baseline framework.

In summary, the majority of the above reviewed works has tried to handle Arabic challenges for improving the process of AIR. However they reached significant results, we note that they completely ignore document structure. The information is usually searched by means of a full text search, every term in the texts of the documents can function as a search key, missing a great opportunity for a more effective search which motivate us to perform our research work.

3 Proposed Approach

Searching on structured documents can be more fruitful than on unstructured ones. Indeed, the indexing process can be not only by content, but also by their structure which likewise enables users to formulate versatile queries mixing keywords and structural information. Thus, the matching between texts in the corpus with a structured query generates a more precise answer to an information query by returning a structural element (or several elements) instead of the complete document. This way can meet greatly user's need who looks forward to precise answers.

In the legal domain, integrating document structure into a retrieval system for accessing legal documents has a lot of potential especially that legal documents are characterized by diverse structural pecularities:

- **Document's Content:** which describes the content of the document. This property is relevant because it contains and refers to the document's element level (as concepts, facts, people, etc.) and relations that draw on them.
- **Document's logical structure:** which refers to the document's hierarchical outline structure (book, chapter, section, and so on).
- **Document's Metadata:** Metadata are particular information that are not in the original content of the document and that are added to improve comprehension and classification of the document.
- **Document's collections structure:** which refers to references and links existing in and between legal documents.

The whole approach that we follow to take into account these pecularities is articulated around two modules as illustrated in Fig. 1. The former focus on detecting the structure of the documents while the later concerns the process of IR in the output of the first module.

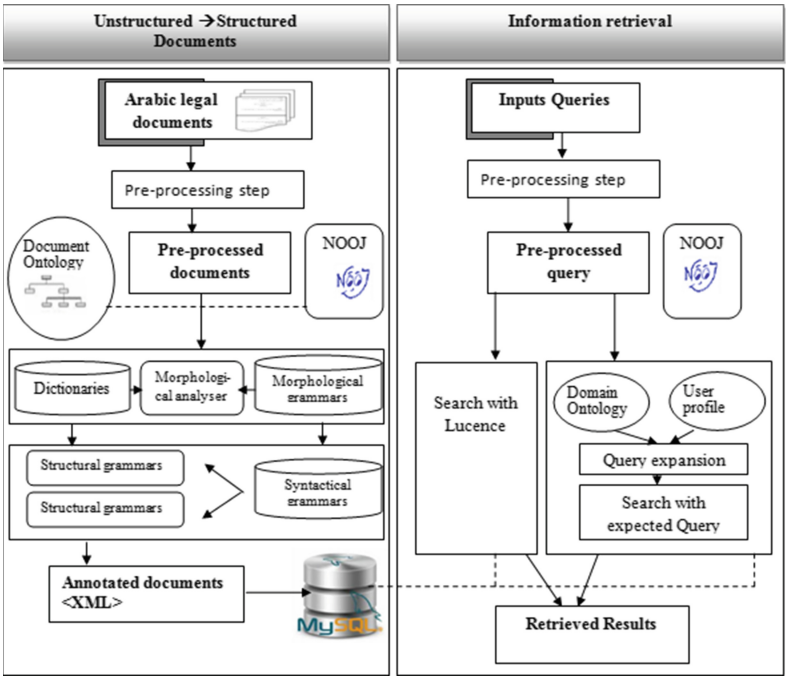

Fig. 1. Approach architecture

3.1 Annotation Module

These particular structures, however are interested in retrieving information, aren't currently explicit in the available legal documents, for this reason before indexation process, an annotation step must be elaborated regrouping five main steps:

– Corpus Selection and Preparation. Faced to the obvious lack of Arabic resources and the unavailability of the existent free corpora, it is primordial to build our own corpus "PenalAr". Dealing with Arabic texts starts in this step, which introduces the beginning of handling Arabic challenges. Firstly, the documents are cleaned from orthographic errors, then segmented to handle the absence of capital letters. Splitting is done using a Clauses splitter, a cascade of finite-state transducers elaborated by [Keskes et al. 2012], which proceed to split extended sentences into a set of clauses. Normalization of documents in a standard format is then applied for easy manipulation as well as orthographic variation resolution. The normalization includes the suppression of special characters if exist, the replacement rules of some Arabic letters:

- Replacement of إ, آ and أ with ا
- Replacement of ة with ه
- Replacement of ي with ى

- Identification of documents Types. We have developed an Arabic Legal Document ontology (ALDO) defining a general legal document class, specifying all types of elements that can be used in the logical structure of a document of this class. Relations between these sub-classes are likewise modeled. The cross-reference between document are also taken into account in the ontology. Identification process is simply carried out by comparing the document title with the ontology concepts. The title is often comprised by the first words in the document.
- Derivation of corresponding structural grammars. Once the document type is identified, the corresponding path in the ontology will be translated by a syntactic grammar in the linguistic NooJ plateform[2] that will be applied to the document for extracting its title and contents table including the books, chapters and sections titles, enumeration lists, and the references... In order to recognize them, we have elaborate grammars represented as directed graphs and apply them to documents of the corpus.
- Application of morphological and syntactic grammars. To handle agglutination issue, a morphological grammar of agglutination is constructed since these forms are not present in the NooJ's dictionary of inflected forms. An excerpt from the used grammars is showed in Fig. 2.
  An Arabic dictionary for legal compound terms is created to tread both simple and compound word to recognize them later. Noting that Arabic NooJ's dictionary generally associates each lexical entry with an inflectional and/or derivational paradigm to describe its syntactic and semantic and inflectional (gender, number, conjugation...), and derivational features. Compound terms are added manually to the dictionary with its semantic attribute referring to the lemma of the term, its category of the compound noun in order to treat each category separately in the syntactical grammar according to NooJ entry format. Each of these features is translated by Inflectional/Derivational grammars based on generic commands to generate the different voweled forms of the dictionary entries and syntactical grammars to extract all related derived and agglutinated forms.
- Projection of grammars on Documents After having applied the grammars to the documents NooJ to produce an annotated text by applying grammar to the documents, the resulting text can be exported as an XML documents, in which XML tags have been inserted.

All the annotated documents are stored in MySQL database to begin the process of indexing and retrieval.

3.2 Indexing and Retrieval

The aim of the XML retrieval system is to incorporate the knowledge on the document structure and on specific content elements in computing the relevance of an information element. This field is fairly fresh. Like any new field, a number

[2] http://www.nooj4nlp.net/.

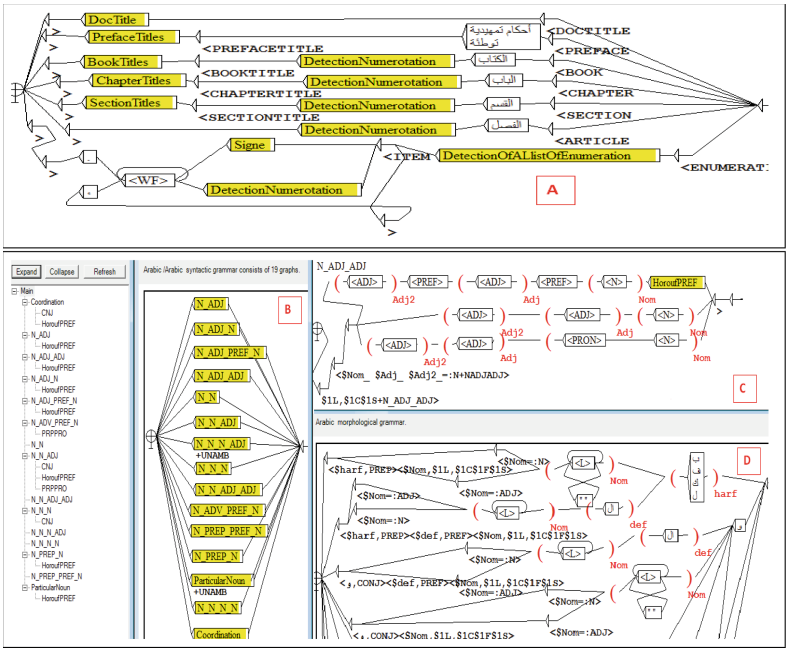

Fig. 2. A. Example of a Structural grammar for legal Code B. C. Graph of Syntactic grammar, D. Graph of morphological grammar

of questions remained unanswered is posed. The most important of these questions, concerns the element to retrieve in response to a query. This is different from traditional IR which considered whole documents to be the only retrievable entity and did not take into account the document's internal structure. The difference here is that the retrieval system must sweep the document at various levels of granularity and compare elements at different levels and by this way it can identify the most specific document element that answers the requested information in the query. Our proposition in this point is to index only elements at the leaf nodes of the XML document. This designates considering each leaf node as a distinct document and indexing it separately.

After the identification step and processing of the query, we propose to the user two alternatives for information retrieval:

– Simple Retrieval without reformulation.
  This alternative is applied when the user wants to send his query directly without enhancement.
  For that, we implement an initial prototype of a search process as illustrated in Fig. 3. We depend on LUCENE, which is free open source information retrieval library released under the Apache Software License with full text indexing and searching capabilities. LUCENE is a high-performance, full-featured text search engine library written entirely in Java. It depends on the Vector Space

Model (VSM) of information retrieval, and the Boolean model to determine how relevant a retrieval element is to a user's query. The search process starts over the annotated documents and the results that match the query will be displayed.

– Retrieval with reformulation.
The quality of responses by IRS depends not only on the quality of the matching step but also of the query made by the user, hence the interest of the reformulation. This alternative is applied when the user wants to enrich his query. The use of ontologies for enrichment (expansion) of the query constitutes a solution among others to solve the problem of semantic variations as they provide resources in the form of semantic relations extending the search field of a query, which improved the search results.
The use of ontologies in IRS may be used either before sending the query, in indexing or in results filtering. In the first case, the query can be enriched by the near concepts of the ontology through the use of different relations such as the generalization/specialization, synonyms. The indexing of documents can be done using the concepts of ontology and not using keywords and Filtering according to a particular domain for users profile.

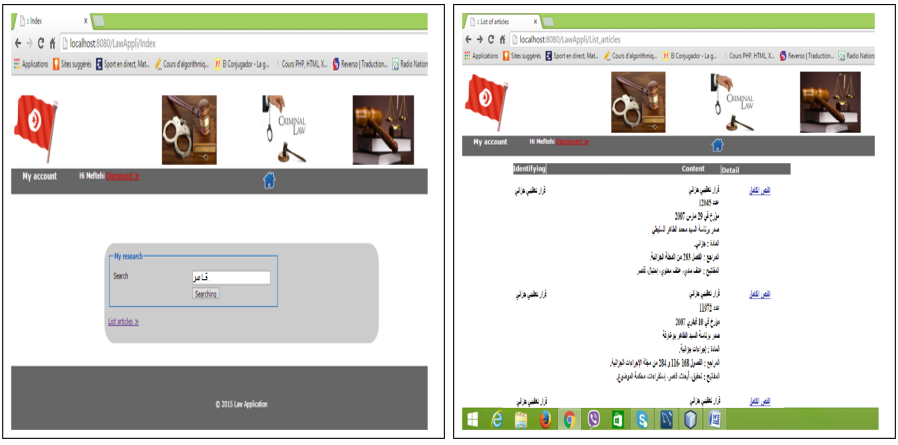

Fig. 3. Initial interface prototype

For very specific queries, a single element may contain the right answer, whereas more general queries could be answered by returning a general element. The models also allow that a group of related elements such as the articles in a section or chapter, contribute to the overall relevance of a single element for example an article.

The retrieval results obtained by the two types alternatives are then saved in various files and various values of recall and precision of the system are calculated for each type of retrieval and query in order to be compared.

3.3 Evaluation

The experiments of the annotation module is done on small corpus consisting of the new-tunisian-constitution's draft released on June 2013 after Tunisia's "Jasmine Revolution" published on the Tunisia OpenGov site[3] and 20 decisions 20 Criminal Law Decisions of the cassation Court available on the web portal site of Justice and Human Rights[4].

We achieved best results through recall and precision which confirm its performance since as regards testing and results on this data, satisfactory ratio are reported for syntactic annotation with 85 % as precision as well as for structural annotation with 73.7 %. The incorrect matches correspond mostly to the syntactic annotation due to sequences containing anaphora. The provided results by the approach have been assessed and validated by two legal experts, a lawyer and a Professor of Law. The former assessed the structural annotation while the latter evaluated the second kind of annotation: the syntactic annotation. Their relevance judgments have achieved an acceptable level of agreement.

4 Conclusion

The present paper has given an overview of how to improve the Arabic legal information retrieval. In our view, it holds great promise in making legal information more transparent and available to more legal professionals. The exploitation of structure requires transforming the data from unstructured to structured in a first stage then applying a retrieval model incorporating the structural information in his ranking algorithm in second stage. The answer list would consist of elements that are highly specific in addition to being relevant to the query and may be found at any level in the document structure. We will be charged of achieving the second stage as perspective.

References

Atwan, J., Mohd, M., Rashaideh, H., Kanaan, G.: Semantically enhanced pseudo relevance feedback for arabic information retrieval. J. Inf. Sci. **42**(2), 246–260 (2016)

Hanandeh, E., Mabreh, K.: Effective information retrieval method based on matching adaptative genetic algorithm. J. Theoret. Appl. Inf. Technol. **81**(3), 446–452 (2015)

Ibrahim, H., Abdou, S., Gheith, M.: Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis. Int. J. Comput. Appl. **118**(11), 26–31 (2015)

Keskes, I., Benamara, F., Belguith, L.H.: Clause-based discourse segmentation of arabic texts. In: Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, pp. 2826–2832 (2012)

Mahgoub, A., Rashwan, M., Raafat, H., Zahran, M., Fayek, M.: Semantic query expansion for arabic information retrieval. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, pp. 87–92 (2014)

[3] http://www.nooj4nlp.net/.
[4] http://ejustice.tn.

Maitah, W., Al-Rababaa, M., Kannan, G.: Improving the effectiveness of information retrieval system using adaptive genetic algorithm. Int. J. Comput. Sci. Inf. Technol. **5**(5), 91–105 (2013)

Mohamed, A.: Design of arabic dialects information retrieval model for solving regional variation problem. Thesis, Sudan University of Science and Technology, Sudan (2015)

Yousef, N., Abu-Errub, A., Odeh, A., Khafajeh, H.: An improved arabic words roots extraction method using n-gram technique. J. Comput. Sci. **10**(4), 716–719 (2014)

Yousef, N., Khafajeh, H.: Evaluation of different query expansion techniques by using different similarity measures in arabic documents. Int. J. Comput. Sci. Inf. Technol. **10**(4), 160–166 (2013)