Challenges in Information Retrieval from Unstructured Arabic Data

Hussein Khalil
School of Science and Technology
Nottingham Trent University, NG11 8NS, UK
Hussein.Khalil2007@my.ntu.ac.uk

Taha Osman
School of Science and Technology
Nottingham Trent University, NG11 8NS, UK
Taha.osman@ntu.ac.uk

Abstract—The main issue that currently faces research in the information society is the flood of information; a problem exacerbated by the massive diversity of information on the World Wide Web. It has given researchers access to millions of references, articles, news and services. Regardless of geographic location and language used, much of this information is unstructured data. There is a large body of research on mining unstructured Web data, but little effort for Web pages authored in Arabic. This paper investigates the Semantic Web (SW) support for handling documents that are authored and/or annotated in Arabic, and how to bridge the gap between the SW and Natural Language Processing (NLP). Moreover, to improve the intelligent exploration of unstructured documents in the Arabic domain.

Keywords—NLP; Semantic Web; Text Mining; Information Retrieval; Ontology Engineering; Knowledgebase

I. INTRODUCTION

The success of the Web is fundamentally underpinned by information retrieval technology, in the form of search engines such as Yahoo, Google, etc. In recent developments, Semantic search systems were applied to information retrieval in many applications such as medicine and e-commerce applications [1]. Moreover, such systems employ software agents, which apply ontology concepts to define the meaning of data on a specific domain. An objective of these systems is also to improve search by removing ambiguity associated with a query, and so returning more relevant results. The techniques and technologies allowing machines to read, understand and retrieve the meaning of specific information from the Internet represent the Semantic Web (SW).

This development is an extension of the present Web, whereby defined meaning is given to information, which enables humans and computers to interact meaningfully. The underlying idea requires data on the Internet to be defined and linked to facilitate effective discovery, computerization, integration and reuse by different applications. Such data may include resources that are not exclusively media objects (e.g. Web text, audio, and images), but also other objects, including events, dates, persons, and locations. As such, the SW possesses more than one relation type, other than the hyperlink, between resources, as mentioned [2]. The massively large number of documents available on the Web were created by an equally diverse base of people, and as such represent unstructured data. SW technology works at transforming this unstructured data into structured form, which then represents knowledge as entity sets.

In a marked up domain, entities have specific relationships. Hence, analysis and mapping of unstructured data to a predefined structure must be performed. Natural language Processing (NLP) is a key text analysis technology, combining linguistics and computer science, focusing on the interaction between human (natural) language and computers [3]. NLP tools are commonly applied to the extraction of meaningful information from unstructured contexts. NLP has been used in many research studies in the area. The first step in NLP is to identify and subsequently extract those terms considered to have an important role in the domain of concern. The principal task in extracting and retrieving information, and natural language processing in the domain is called Named Entity Recognition (NER). "A Named Entity Recognition (NER) system is a significant tool in NLP research, since it allows identification of proper nouns in open-domain (i.e. un-structured) text". The NER system mainly performs simple recognition of those instances of linguistic patterns and collates them. This typically takes place without consideration of how this content may be combined or shared with other document sources [4]. The importance of NLP tools lies in that they play a key role in allowing semantic concept tagging of unstructured text, and so realize the SW. However, in the Arabic language domain very little research has been undertaken on the SW processing of unstructured data. Within the Arabic Information Retrieval Domain, the success of the SW requires appropriate supporting tools and applications. This work presents a new framework using NLP and SW...
techniques for information retrieval and knowledge management relating to Arabic data sources [2].

The rest of the paper is organized as follows, where Section II reviews related publications, Section III addresses the Arabic language and the SW challenge, Section IV presents System Architecture and implementation, while Section V provides conclusions and presents plans for further work.

II. LITERATURE REVIEW

In this section, we will discuss various works that were published in the Arabic domain, which have used NLP tools and SW.

A. Arabic Name Entity Recognition (NER)

NER involves identification of proper names in texts, and classification into a set of predefined categories of interest, which is extensively used in NLP. There are no general approaches to extract terms, because they depend on the domain language and its rules. In the Arabic language, domain, for example, there are many approaches to address the extraction of terms.

Oudah and Mai developed a new system called Named Entity Recognition Arabic (NERA). The purpose of this system is to improve the rules based on Named Entity recognised by means of applying machine learning. The idea of this system is composed of two approaches: a rule-based or machine learning (ML) based approach, and new approaches (a hybrid system). The General Architecture for Text Engineering (GATE) application was used to produce the new system. The NERA system to generate the Named Entity annotation used a rule-based system for Corpora, including ANERcorp corpus, ACE Newswire corpus and ACE Broadcast News corpus. ML classifications, J48, Bagging and END, were used for each dataset, with and without rule-based features (annotations). The Arabic language was used to test the system. Although this experimental system can be seamlessly integrated with rule-based targets of another language, the system has been claimed as the first of its kind to use a hybrid method to improve the rules for extraction of person, location and organization name entity in the Arabic Information Retrieval Domain. This work is capable of recognizing 11 different types of Arabic entity. However, the work focused on extracting the entities without taking into account the relations between them [5]. Harmain, El Khatib and Lakas presents some attempts of mining in Arabic texts that deal with techniques from the research area of Data Mining and NLP. Their system is composed of many stages; the first stage is a pre-processing stage to convert HTML Arabic documents into XML documents. This stage significantly simplifies the analysis process. The processed text is then linguistically analysed from the word level to the text level. The result of this analysis is a semantic network of the entities mentioned in the text and the relationships between them. This semantic network is then used for some specific mining tasks. The author tried to explain the architecture of an Arabic text mining system, and how NLP techniques can play an important role. The system in the research paper might open a new research area in the Arabic text mining application. Nevertheless the research is focused on addressing the HTML document and not other kinds of documents, such as DOC, PDF. Moreover, the paper does not present any evaluation of the work [6]. Discretization is considered one of the major challenges to most Arabic NLP tasks. Many studies worked with NLP, including that of Bahanshal and AL-Khalifa which evaluated the accuracy of some available discretization systems based on fully diacritized text from the Holy Quran and short poems from the period of the advent of Islam. The researchers claimed to enhance the manual evaluation of diacritizer accuracy rate with a computed process using larger text. This paper covered this critical topic in challenges in Arabic language processing. We can use this review as supported in my work [7].

B. Semantic Web (SW) and the Arabic Language

The knowledge represented by an ontology comprises concepts within a domain and the relationships between them; this is considered the backbone of the SW. Alruiy and ‘Meshrif’ presents the basic framework for crime type recognition. It addresses documents that are related to the crime domain in the Arabic language by identifying the types of crime in these documents. Two approaches were used in addressing the document; the first used a gazetteer to perform named entity recognition. The second applied a rule-based system to classify entities not mentioned in the gazetteer. Three types of experiments were performed, where the first and second used the first and second approaches, respectively, and the third used a combination of both approaches. Precision, recall and F-measure were used to evaluate the system. The result of the evaluation was high in the first approach. However, the authors made no mention of any tools used to address the document. In addition, the work dealt with the relation between entities as terms. Furthermore, it
does not explain the knowledge-based (KB), and how it can be populated with mined data [8].

Maynard and Diana aimed to inquire into NLP techniques for ontology population, and how to verify that term recognition is helpful for many tasks of extracting information. Through a combination of rule-based learning and ML, TRUCKS is used to enhance traditional statistical techniques of term recognition [10]. This work tries to explain the relation between term recognition and information extraction, and clarifies the difference between the methods used in each. In this work, the Balanced Distance Metric (BDM) was designed to evaluate ontology-Based Information Extraction, which uses similarity between the key and response instances in an ontology to determine the correctness of the extraction. The approach presents a dependable technique to use the NLP techniques for extraction terms and ontology population [9].

Beseiso, Ahmad, and Ismail, investigated Arabic support in some existing SW technologies and located the ability to stratify the SW with Arabic applications. The chapter presented various studies conducted in the Arabic information extraction area. Different tools, like Protégé, Jena, Sesame, and KOAN, were used in evaluating Arabic support. The evaluation found that Protégé and Jena had Arabic support with Resource Description Framework (RDF), while Protégé had limited Arabic support with OWL and Query. In addition, while Jena had support for Arabic with OWL, it was limited with Query. On the other hand, in terms of Arabic support, results for other tools were not satisfactory. However, the Arabic domain still needs further investigation to catch up with other language domains, like English and French [10].

Zaidi, Soraya, M. T. Laskri, and A. Abde used a rule-based approach to extract Arabic collocation by using the GATE tool to create a KB domain. JAPE rules were used to write the new rules for collocation. The Quranic corpus was used to extract the collocations. The output of this stage is then used to create domain knowledge. The precision, recall and f-measure were used to evaluate the system, returning 0.5, 1, and 0.66, respectively. However, we observed different values between recall and precision. This is due to the use of JAPE rules without taking gazetteers into account. We can conclude that working in the Arabic language domain still needs further investigation[11].

In summary, we note that most of the previous studies focussed on how to extract Arabic terms, but without looking at how to find the relation among these terms.

These relations are considered very important to create the Knowledge-based, which is considered the backbone of the SW. Moreover, the construction of Arabic ontologies is rare in these studies. Therefore, our work is focused on how the SW can support handling documents that are authored and/or annotated in Arabic. In addition, how to create a system able to populate a KB from Arabic unstructured data. It also addresses how to bridge the gap between the SW and NLP, and so improve the intelligent exploration of unstructured documents in the Arabic domain.

III. ARABIC LANGUAGE AND THE SEMANTIC WEB (SW) CHALLENGE:

Arabic is the native language of nearly 500 million people located in 23 different countries, and is the largest member of the Semitic language family [12]. It is also the language of the holy Quran. With over 200,000 Arabic websites on the Internet [13], most published Arabic information is in unstructured (raw) format. Compared to Latin language processing, the Arabic text poses many challenges that have influenced the development of language processing tools, such as short vowels, absence of capital letters, complex morphology; moreover, it is orthographic with diacritics, and is highly inflectional and derivational etc. According to [14], “Arabic is a highly inflected language which has a rich and complex morphological system. The Arabic word takes more than one word form to represent it. That includes root, prefixes, suffixes and clitics. Arabic diacritization, defined as the full representation of short vowels, is considered one of the major challenges to most Arabic NLP tasks. However, modern standard Arabic is usually used nowadays in newspapers, books, and web without diacritization. Therefore, our work will take diacritization into account. There are a variety of tools to extract terms in languages, such as English and French, but these tools are not fully suitable for the Arabic language; which poses a different set of challenges in language processing; for example, if diacritization is not applied, then the phrase "كتاب الولد في المدرسة" may take the meaning: the books of a boy are in the school, or the boy wrote in the school. Therefore diacritization can improve clarifying the context of a sentence or paragraph, [15] but can also introduce diacritization challenges in terms of associating distinct meaning to the same word, such as in the example below Table 1.
TABLE 1: DIACRITIZATION IN THE ARABIC LANGUAGE

<table>
<thead>
<tr>
<th>Word 1 meaning</th>
<th>Word 2 meaning</th>
<th>Word 2 meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>علم</td>
<td>علم</td>
<td>علم</td>
</tr>
<tr>
<td>Science (noun)</td>
<td>Flag (noun)</td>
<td>Knew (verb)</td>
</tr>
</tbody>
</table>

IV. SYSTEM ARCHITECTURE AND IMPLEMENTATION

In the previous sections, we discussed different approaches, and how they deal with Arabic SW challenges.

In this work, we present a framework to extract terms from unstructured text, in order to populate the Knowledgebase (ontology) and improve the information retrieval results. Figure 1 shows the architecture of our system, which has three components; the first involves NLP techniques to extract Arabic Terms and ontology population. The second component is to construct the KB. Finally, the last component uses the result of the first stage to extract information with the implicit knowledge to enhance the Knowledgebase and update gazetteers.

<table>
<thead>
<tr>
<th>Figure 1: Architecture of our system</th>
</tr>
</thead>
<tbody>
<tr>
<td>In The first component, NLP techniques are deployed to identify known entities and patterns within the Arabic corpus to aid tagging them with semantic concepts. The second stage comprises building the KB ontology using the key concepts and the relationships of the analysed domain, and then populate the KB ontology with the entities tagged at the NLP stage. The final stage introduces intelligent inference to the Knowledgebase by exploring the engineered object relations between the concepts, and also independent roles that can further extract implicit inferred knowledge. Initially, the documents are collected from different Arabic resources related to the economic domain, such as Aljazeera News, Arab news, BCC News, etc. The NLP stage aims to extract structured information from unstructured textual data. In this stage, each document is preprocessed. The process of text classification is implemented using the GATE Open Source toolkit. The linguistic component uses the infrastructure and the following resources from GATE, namely tokenize, sentence splitter, part-of-speech tagger, morphological analyser and VP chunker. On the top of these resources, which produce syntactic annotations for the input text, the linguistic component uses a Gazetteer to recognize the terms in the text. It looks for matches between these lists with each word in the text, each gazetteer list presents a set of names, such as an organization, stop words, economic terms, resources, locations act. Gazetteer data is collected from different resources, such as Maknaz an Arabic database of organizations, and Internet resources [16]. Grammar is used to identify linguistic triples. This grammar was implemented in JAPE [17], which allows the definition of patterns to recognize regular expressions using the annotations provided by GATE. We use JAPE to manipulate annotations on the document, and to further identify patterns. In addition the JAPE rules are used to classify entities not mentioned in Gazetteer, then using Wikipedia SPARQL to increase the confidence of the pattern recognition process. In addition, enriching the Knowledgebase and gazetteers with the similar entities. Indeed, entities and facts from the document, which were not recognized in the past. Figure 2 shows Arabic Terms extraction. This approach will hopefully help overcome the challenges of Arabic language processing, such as spelling and the typographic forms problem. The major stage in our framework is event extraction. It represents the relation among entity like (Export, Import). The extract name entity and event extraction are used to create a semantic graph. The second stage is constructing the Knowledgebase. The success of the Semantic Web application depends on the Design and development of well-structured ontologies figure 4 shows our Ontology. The Knowledge has been represented by an ontology as</td>
</tr>
</tbody>
</table>
concepts within a domain and the relationships between those concepts. In this work, we construct the Arabic ontology in the economic domain to serve the semantic search and information retrieval of Arabic on the Web. The event and. Entities are used to define the domain by mapping rules. The Protégé tool is used to build ontologies, while the Jena Framework is used to extract and write from/to the RDF graph. When the knowledge representation in the Knowledgebase is completed, we will use the reasoning mechanism to compute the inferred ontology class hierarchy.

The final stage used the output of the second stage to infer new facts from these semantic annotated. On the other hand, some words or phrases in the documents might not be clearly associated with classified entities. If the NLP process is not certain about the recognised entities, then public dataset such as DBpedia (the semantic version of Wikipedia) [18] can be queried to resolve information ambiguity. For instance, in the scenario exemplified in figure 3 above, JAPE grammar rules were engineered to detect a pattern, where the NLP process can predict the class of a specific entity (ALhariga – الحريقة) from the linguistic context of the analysed text. The rule predicts that (ALhariga – الحريقة) is a city in Libya, a fact that is verified by querying the public dataset DBpedia about Libyan cities. Moreover, the query also; returns additional knowledge that the city is an oil exporting port; and finally, all these new facts are used both to enrich the Semantic knowledgebase, and also to update the NLP gazetteers in order to automate future recognition of the new facts.
V. CONCLUSION

This paper investigated the challenges an Arabic information retrieval as applicable to both NLP and Semantic Web technologies by reviewing published work and experimenting with current technologies it was established that Arabic information retrieval is tagging behind the Latin language in terms of maturity of the processing technologies to address the short comings , we proposed a hybrid NLP-Semantic framework that uses fundamental NLP techniques and proprietary grammar rules to semantically tag Arabic text for a specific domain (Economy was our use case), which will be followed by Knowledge inference in an appropriately modelled semantic knowledgebase to discover new facts in the mined text. We also demonstrate how a public semantic dataset from the linked Data cloud can be used to resolve any ambiguity in the mind data. The next stage of our research will focus on improving the intelligence of the semantic knowledgebase, deploying sophisticated entity relationships (transitive, functional, etc.) as well as engineerly SWRL [19] that can future aid the classification of Semantic knowledge to infer new facts about the target domain.

REFERENCES

[16] http://maknaz-so0oso0o.blogspot.co.uk/.