Arabic Semantic Web Applications – A Survey

Aya M. Al-Zoghby
Computer Science Department, Faculty of Computers & Information Science, Mansoura University, Egypt
Email: elzoghby.aya@gmail.com

Ahmed Sharaf Eldin Ahmed
Information Systems Department, Faculty of Computers & Information, Helwan University, Egypt
Email: profase2000@yahoo.com

Taher T. Hamza
Computer science Department, Faculty of Computers & Information Science, Mansoura University, Egypt
Email: taher_hamza@yahoo.com

***Abstract*— Arabic Language is the mother tongue for 23 countries and more than 350 million persons. It is the language of the Holy Quran; therefore, many non-Arabic Islamic countries, like Pakistan, teach Arabic as a second language. Nevertheless, it is observable that the Arabic content on the Web is less than what should be. The evolution of the Semantic Web (SW) added a new dimension to this problem. This paper is an attempt to figure out the problem, its causes, and to open avenues to think about the solutions. The survey presented in this paper concerned with the SW applications regarding the Arabic Language in the domains of Ontology construction and utilization, Arabic WordNet (AWN) exploiting and enrichment, Arabic Named Entities Extraction, Holy Quran and Islamic Knowledge semantic representation, and Arabic Semantic Search Engines. In fact, the study revealed serious deficiencies in dealing semantically with the Arabic Language. That is mainly owing to the rarity of tools that can support the Arabic script. Furthermore, the Arabic resources, if available, are not free. Moreover, there are many technical problems in the semantic dealing with the Arabic context. Therefore, most of the developed applications are not sufficiently proficient. However, due to the significance of the Arabic Language, it is inevitable to overcome these deficiencies in order to put the Arabic Language in the category of the machine-semantically-interpretable languages, rather than just the textually processable ones. This way, we can exploit the power of the Semantic Web features in extracting the essence of the knowledge residing in the Arabic web documents and going beyond dealing with its rigid texts.**

***Index Terms*— Semantic Web (Web 3.0), Arabic Language, Islamic Knowledge, Named Entity Extraction (NEE), Semantic Search Engine (SSE)**

I. INTRODUCTION

The term 'Semantic Web' was coined by Tim Berners-Lee, the inventor of the World Wide Web and the director of the World Wide Web Consortium ("W3C"), which oversees the development of proposed Semantic Web standards. Tim Berners-Lee defined the Semantic Web as "a web of data that can be processed directly and indirectly by machines." He originally expressed the vision of the Semantic Web as follows: "To a computer, the Web is a flat, boring world devoid of meaning. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. The Semantic Web (SW) is not a separate Web but rather an extension of the current one, in which information is given well-defined meaning. Adding semantics to the Web involves allowing documents which have information in machine-readable forms and allowing links to be created with relationship values" [1; 2]. SW, therefore, can enhance the currently existing Web by the interposition of a machine interpretable layer that holds the metadata of the Web document. This metadata will allow computer software to understand what the Web page is about, and thus draw conclusions about it. This revolutionary invention can be utilized by software agents, so the information can easily be found, shared, integrated, and exchanged [3; 4].

An expected contribution of the Semantic Web technology, if it becomes satisfactorily actuated over the World Wide Web, is to extremely enhance the extractability, exchangeability, and hence growth of knowledge on the Web regardless of its nationality, language and expression style. That can be archived due to the utilization of a unified representation of the concepts in a standard machine interpretable language regardless of the native language of the user [5; 6; 7]. It is therefore vital for all natural languages, especially those that have a broad base of utterers, to be supported by the SW tools and applications. Among these natural languages is the Arabic Language. Besides that the Arabic is the language of a significant number of people; it is one of the strongest, richest and most languages able to articulate in the world. The SW, likewise, is the strongest technology that emerged in the field of Web till now. Thus, the integration of these two giants yields a capability that cannot be emulated.

This survey reviews the studies that addressed the Arabic Language and made efforts in order to support it in the environment of the SW. The reviewed studies were grouped thematically in four groups, each of which addressing an essential aspect of developing an Arabic SW application, which are the Arabic-Ontologies development and utilization, the Arabic Named Entities Extraction (ANEE), the Islamic-Knowledge Semantic

doi:10.4304/jetwi.5.1.52-69

Representation, and finally the Arabic Semantic Search Engines (SSEs).

As for the Ontologies, the success of any SW application depends on the design and development of the Ontology; they are the backbone of the SW [8]. This review is concerned with the researches that addressed the Arabic Ontology construction as a target, or those that depended on the Arabic Ontologies in order to accomplish larger systems. Furthermore, the review considered the researches that are based on the Arabic WordNet (AWN) and dealt with it as an Ontology or an alternative to the Ontology. In addition, other researches that aimed to the development and enrichment of the current AWN are included.

Owing to the importance of the NEE tools as one of the pillars upon which many SW applications can be built, we have dedicated a section for discussing the related articles.

As Arabic is the language of the Holy Quran and most of the Islamic Knowledge resources on the web; therefore, exploiting the abilities of the SW technology in the representation of these resources will certainly reflect a qualitative shift in the handling and treatment of such knowledge. The review, thus, contains a specified section for that area.

Although each of the aforementioned issues may constitute a standalone application of the SW; the Semantic Search Engine, on the other hand, is the software that may incorporate all of them as sub-modules. The Arabic Semantic Search Engine is the subject of the last part of this review.

The survey clarifies the inefficiency in handling the Arabic Language semantically, which results in an unqualified Arabic Web content for being a suitable environment of the SW technology implementation. This inefficiency may be ascribed to the sophistication of the Arabic language that makes it complex to the extent that hinders its treatment electronically.

The reviewed articles were selected according to their relevance, currency, and significance. As for the organization methodology, they were organized thematically, then chronologically. Also, the comparable results or methodologies were grouped.

Some abbreviations were used in this research, so we will list them here for convenience: Suggested Upper Merged Ontology (SUMO), Arabic WordNet (AWN), Princeton WordNet (PWN), Semantic Web (SW), Named Entity Recognition (NER), Named Entity Extraction (NEE), Semantic Search Engine (SSE), Google Translation API (GTA), and Word Sense Disambiguation (WSD).

II. Semantic Web (Web 3.0)

The World Wide Web (WWW) dramatically changed the capability of accessing the information electronically. Currently, over 300 million users access around 3 billion Web documents internationally, and continue growing. That constantly increases the difficulty of finding, accessing, and presenting the required information. The information representation in just a human-readable format rather than a machine-interpretable exacerbates the problem. That is because among a heap of information extracted by machine, the user himself must withdraw his exactly desired information or what he is quite looking for in this heap [9]. Therefore, the Web can be at its full power only when its data can be found, shared, processed and understood by means of both human and machine [10; 11]. This vision of the Web can be reached by the new generation of Web, Web 3.0 or SW, which is the hoped solution to narrow that gap between human-readable information form and that machine understandable. It represents the Web content in a more easily machine processable form, and uses smart techniques to get advantages of that representation by adding a semantic layer to the existing Web pages called 'metadata'. Thus, the Web page will contain the formatted information for the human reader presentation, as well as information about their content, metadata, for machine understanding [12]. Since the 'metadata' comprises chunk of the data meaning, the term 'Semantic' hence arose, and 'SW' expression coined. The used technique for adding metadata layer is called 'Semantic Annotation' [13], which is the process of labeling Web pages with the semantics of their contents through mapping the data instances to the corresponding concepts in a predefined Ontology [14; 15].

The Ontology is the corner stone of SW technology. For a certain domain, the Ontology enumerates and gives semantic description(s) of concepts in that domain, defining domain-relevant attributes of concepts and various relationships among them. To this end, the World Wide Web Consortium (W3C) developed formal specifications such as RDF, RDFS, OWL, and SWRL[1] in order to provide an accurate description of the concepts, terms, and relationships within a knowledge domain [16; 17]. In the context of Web, the Ontologies provide a shared understanding of the world, which is required to overcome the terminological variations that may emerge from the different cultures and tongues or the individual expressing ways of the users. This is done by mapping different terminologies that are referring to a particular concept to the ontological representation of that concept. Furthermore, the Ontologies are used to improve the Web search precision by searching for Web pages that hold certain concept instead of search using just keywords and ambiguous terminologies.

Besides the metadata and Ontologies, the SW has two more technologies. The first is the 'Logic and inference' technology, by which automated reasoners can infer conclusions from given knowledge. The other is the 'Agents', which are working independently on behalf of computer programs as they can take tasks to perform, make a certain decision, and provide answers [18; 19]. Hence, the SW will enable the automatic collection and correlation of object's various information parts that are

[1] RDF: Resource Description Framework
RDFS: schema language for declaring basic class and types for describing the terms used in RDF
OWL: The Web Ontology Language, and
SWRL: Semantic Web Rule Language

available at various Web resources. That will certainly keep the time spent for navigating the World Wide Web just to obtain specific information that exists somewhere out there.

Surely, various applications can utilize the significant benefits of the SW. The SW applications are defined as a software program that uses and produces information for the SW such as Semantic Annotation, Semantic Communication, Semantic Search, Semantic Integration, Semantic Personalization, Semantic Proactivity, and Semantic Games [20; 21; 22]. Undoubtedly, there are many challenges faced by the applications that aim to leverage the advantages of the SW. Of these challenges: the Ontology development and advancement, the SW content availability and scalability, the visualization, the SW language standardization, and of course, the multilingualism [23]. As for English or French Languages, there exist large grounded environments that support the development of their SW applications. For the Arabic Language, this is unfortunately not the case, as just few tools were adapted to support it. For instance, there is an obvious deficiency in Arabic editors for OWL language and RDF files, and even for tools that support the Arabic character set; they do not fully support right-to-left script. In addition, there is almost no Arabic Language parser supporting RDF files in semantic editors. Moreover, there is no Arabic metadata definition similar to DCMI[2] in English. Last but not least, there is no open-source for Arabic SW tools and Web services [24].
The main reason for that marked inadequate support of the Arabic Language is the difficulty of the electronic dealing with the sophisticated particularities of such rich and strong language. That will be discussed later in more details.

III. Arabic Language and SW

The Arabic Language, its features, and the challenges faced in the Semantic Web technology implementation are discussed in this section. Moreover, the related developed applications are presented.

A. Arabic Language: Features and Challenges

Arabic Language is the official language of millions of people, and the religious language of all Muslims. It is a Semitic language of 28 alphabets, and it is one of the United Nations official languages [25; 26]. The Arabic Language has some specificity that caused it to be a difficult language and may hamper the development of appropriate SW tools. Among these specificities, its complex morphological, grammatical, and semantic aspects since it is a highly inflectional and derivational language. Therefore, the NLP tools that were designed for English cannot exactly meet the needs of the Arabic Language. Moreover, the Arabic Language lacks the capitalization feature that characterizes English, which directly affects and complicates the extraction of the Arabic Named Entities. Furthermore, the Arabic Language is highly ambiguous for several reasons. One of them is the vowelization feature of the Arabic Language that causes the ambiguity when it absents, which is the usual case. Another is the Polysemous, or multi-meaning words; words that share the same spelling and pronunciation but have different meanings [27; 28]. Another challenging issue is the problem of encoding, since different encodings for the Arabic script exist on the Web [29]. In addition, the Arabic resources such as corpora, gazetteers, and NLP tools are either limited or not available for free. That leads to the wasting of time in the process of collecting, modeling, and developing of such resources. Consequently, SW basic techniques which essentially depend on such resources, such as NER, could be affected [30]. All of these factors will certainly shrink the ability of the development and implementation of Arabic SW applications. Nevertheless, the Arabic Language worth doing the best for overcoming these obstacles in order to bring the success that was reached by the SW in various domains, such as Medicine, e-Commerce, e-Learning and Biology, to those of Arabic. What is hopeful is that the Arabic content on the Web is continuously increasing, so we must exploit that to outspread the success of SW applications to the Arabic Language.

B. Arabic Language and SW Applications

This section presents the salient researches that presented Arabic Semantic Web applications in the domains of Ontology construction and utilization, Arabic WordNet (AWN) exploiting and enrichment, Arabic Named Entities Extraction, Holy Quran and Islamic Knowledge semantic representation, and finally Arabic Semantic Search Engines.

1) Arabic Ontologies Applications

As stated before, Ontologies are one of the essential blocks in the SW applications, since they offer a well-defined and standardized form of interoperable, machine understandable repositories. In general, there are two main kinds of Ontologies: 'domain specific Ontology'; which represents the precise meaning(s) of terms as they are interpreted in that domain, and 'upper Ontology'; which represents the general concepts that are related to a wide range of domains [31].

As for Arabic Language, a recent statistic of OntoSelect Ontology library showed that there is a lack of Arabic Ontologies and about 49% of Ontologies were created for the Latin character set [32]. Since each language has its own linguistic environment and cultured conditions, so each language needs its own set of Ontologies. As most of the developed Ontologies are in English, there is an urgent need for developing an Arabic Ontologies to be exploited in the Arabic SW applications. Several researches considered the construction and population of Arabic Ontologies; in fact, the vast majority was of domain-specific category.

In the Legal domain, S.Zaidi et al [33] presented a multilingual Web-based tool for Arabic information retrieval based on an Arabic legal domain Ontology in

[2] DCMI: Dublin Core Metadata Initiative: http://dublincore.org/

order to improve the recall and precision of the search. The tool tries to reduce the amount of noise in the search results using Ontology-based query expansion process. The Ontology was manually constructed by Protégé 2000 using a top-down strategy in which the hierarchal concepts are connected by 'is-a' relationships. For the Ontology population, the United Nations Arabic articles were used as well as some Arabic newspapers. That Ontology was used as the base of subsequent user-query expansion phase in the search system. For instance, if the initial query is " قانون العقوبات", the ontological synonyms of the two legal concepts 'قانون' and 'العقوبات' then will be used to extend the initial query. Moreover, the ontological hyponyms and hyperonyms[3] may also be used in the extension process. Thereby, the original query will be expanded to include the words : " قانون العقوبات, قوانين, عقوبة, العقوبة, يعاقب, نص, نصوص, تشريع, حكم, أحكام, مخالفة, مخالفات, جنحة, جنح, تخريب, اعتداء, اعتداءات, قتل, جريمة, جرائم, قانون_عام_داخلي". That significantly improves the system's results, which will be clarified in more details later at the section of Arabic Semantic Search Engines.

Another domain, the Arabic Agricultural, is conceptualized in the domain Ontology that is presented at [34] . This article proposed a system that automates the process of constructing a taxonomic agriculture Ontology using a domain-specific semi structured Web documents. The system built the Ontology using a set of 180 Arabic extension documents, with 3817 HTML headings and 30 seed concepts that are representing the main concepts of the agricultural production. The Ontology was constructed using two complementary approaches. The first is to utilize the phrases structure that appears in the HTML headings of the used documents. The second is to utilize HTML headings' hierarchical structure for identifying new concepts and their taxonomical relationships, which define the relation between the new concepts themselves, and between them and the seed concepts. The proposed system includes seven modules that act briefly as follows: the 'Heading Extractor' extracts heading from input HTML. A taxonomical Ontology is then extracted in the 'N-gram based Ontology Learner' module using the N-gram phrases in text headings. The 'Ontology Refiner' is a module that filters the extracted Ontology in order to remove noisy or fake concepts. By using the heading structure of the input Web documents, the 'HTML Structure Based Ontology Learner' module learns an Ontology. That Ontology will be extended by discovering new concepts, which have sibling relations with previously learnt concepts in the 'HTML Ontology Refiner' module. Finally, the 'Ontology Merger' module takes the output of 'N-gram based Ontology Learner' and 'HTML Structure Based Ontology Learner' to merge them, then it adjusts the hierarchical structure of the resulted Ontology by discovering the right pattern and level of the concepts existing in the merged Ontologies. The best-obtained result for the lexical evaluation was 72.86 % for precision and 80.52% for recall. For the taxonomic evaluation, the F-measure was 75.66% as 69.08% for precision and 83.62% for recall.

Fatma Zohra et al in [35] and [36] addressed the Arabic Ontology development using Verbs and Roots. In fact, 85% of Arabic words are derived from tri-literal roots, and the verbs are classified by their derivation rules from roots. However, beside the authors did not present an implemented framework to support their hypothesis, the use of the derivational roots as a base for building an Ontology is imprecise since those derived words may hold the same core meaning, but they could be categorized under different classes. For instance, the word 'عِلم' does not belong to the 'Human' class that encapsulates the word 'مُعلّم', although they have the same root.

The research [37] propounded additional model for building and using Arabic domain specific Ontologies. This time, it was for the Locations domain. The research presents an Arabic semantic annotation tool called AraTation for semantically annotating Arabic News content on the Web documents. The system was accomplished on two stages; the first is for the Arabic Information Extraction to identify the Named Entities, and the second is the 'Semantic Annotation' that maps the extracted entities to the corresponding ontological instances. The annotated documents are saved as RDF files to be machine processable and reusable on the Web. Since the annotation process cannot be accomplished without an Ontology for mapping instances with its concepts, the authors of [37] built their own domain Ontology using protégé-OWL editor. In fact, this study suffers from a gap between its Information Extraction (IE) module and the Annotation module. The IE module designed to extract four types of words by means of Person, Organization, Company and Location, while the Ontology is limited to just the Location domain. The research results were evaluated by precision and recall measurement. However, the sample size was limited to the extent that the reliability of results cannot be ensured; it was performed on just 25 documents. The results showed that the average of precision and recall is 67% and 82% respectively. The authors ascribed that rates to the use of the 'part-of' property in the Ontology, and we think that the usage of 'is-a' property to addressing the issue of synonyms and derivations will improve the results as much as possible.

As for the Islamic Knowledge, several researches presented an ontological representation of it. For instance, the research [38] presents a representation of the 'Opposition Terms' field in the Holy Quran, while [39] considered the 'Time Nouns' field. The research [40] on the other hand, depicts a conceptual graph as a wider Ontology for the Holy Quran as a whole. The authors of [41] concern a multilingual Ontology for Islamic portal. We will discuss these articles in detail later at the section of the Islamic Knowledge.

Lilac Al-Safadi et al, in [8], provided an Ontology for representing the knowledge of the *Computer Technology domain* presented at Arabic blogs on the Web. The

[3] Hyponym is a word or phrase whose semantic field is included within that of another word, which is its hypernym (sometimes spelled hyperonym).

research conducted an experimental study on a set of randomly selected Arabic blogs in the domain of computer technology in order to determine the frequent modern Arabic terminologies used in the blogs. The researchers developed a domain Ontology for combining both traditional and modern Arabic terminologies in the computer technology domain. The authors deemed that building an Arabic SSE for blogs relying on just traditional Arabic terminologies is inadequate, and it must utilize the Ontology that has been built for blogs. The developed Ontology consists of 110 computer technology classes and 78 individuals as instances of the predefined classes. Moreover, 48 object relations were defined such as: has-logo:لها شعار , produced by:تنتجه شركة . The Ontology was initially provided as a core, and enabled to be extendable via further reuse and population. The Ontology's relevant terms are gathered by domain users, by sources such as Computer Terms Dictionary, and by translating the content of English Ontology. The Ontology was built using protégé 4.1 editor and tested by protégé 3.44, since protégé 4.1 capable of displaying Arabic text but does not include SPARQL query panel. This deficit is one of the evidences on the weak support of Arabic Language on the SW technologies and tools. In fact, the results of the Ontology testing showed its ability to overcome the semantic gap. We think it is a suitable core for building a SSE for the Arabic blogs.

On the other hand, [42] shows other work that was developed for an Arabic Ontology in the *Computer Technology domain*. It, however, deals with the classic or standard traditional language rather than the blogs modern language. In addition, the presented Ontology is much simpler and does not carry that amount of classes. The presented system will be discussed in more details as an Arabic SSE later.

As for [31], the ongoing project that stated at 2010 in Birzet University, it aims to develop an *Arabic Linguistic or Upper Ontology* rather than a domain Ontology. In fact, many articles that depended on Arabic Ontology in their systems implementation treated the AWN as that Ontology. Nevertheless, Mustafa Jarrar in [31] differentiated them. He indicated that AWN could sometimes be used as an Ontology due to its semantic beside lexical relationships. However, it is neither ideal nor optimal since it lacks formalization and suffers from many fundamental ontological problems. For instance, it, unlike the Ontology, lacks the Hyponymy subtype relation. Moreover, it contains just limited number of words comparing with its English counterpart (PWN). In addition, it is created using the translation methodology, which is ineffective in many cases. That is because different words from different languages may evoke the same concept. In addition, not all concepts are shared cross languages since they are primarily dependent on the culture of that language's users. According to these problems of AWN, the author's aim is to develop an Arabic Ontology that is well founded logically and philosophically. Its top level was defined from the known Top Level Ontologies SUMO and DOLCE. The semantic relations of the Ontology were well defined mathematically. Moreover, the content and structure of the glosses was strictly based on ontological principles. The Ontology is being built through a four-steps approach. The first is to mine the Arabic concepts from dictionaries. This step collects as many glosses/concepts as possible from specialized and general dictionaries. The selected dictionaries should focus on the semantic aspects where multiple meanings are mixed up. In order to conduct this step, a manual mining via Scan and OCR process is firstly done, then the basic cleaning is done automatically. Examples of the used dictionaries are: معجم البلدان, معجم الحاسبات, تعريف مصطلحات القانون الخاص, المعجم الوجيز and others. Step two is to automatically mapping these Arabic concepts with WordNet concepts using a smart algorithm, which takes as input the Arabic gloss and 117k English glosses in WordNet. It outputs the best matches with accuracy of +90%. The next step is to reformulate these glosses according to strict ontological guidelines. Step four then links all concepts with the Arabic Core Ontology. The top levels of the Arabic core Ontology are being built manually basing on DOLCE and SUMO upper level Ontologies. The philosophical and historical aspects of the Arabic concepts terms were taken into consideration. The Core Arabic Ontology (Top 10 levels, ~420 concepts) is being evaluated, and the lower levels are evolving rapidly. Many challenges and future work are considered such as increasing the Ontology size and quality as automatic as possible, and the inclusion of concepts from different Arab countries /communities /eras.

2) Arabic WordNet (AWN)

Concepts are the structural units in the WordNet. They are more than just words, since they consist of compounds, collocations, idiomatic phrases and phrasal verbs. That extends the idea of storing just words to storing its Conceptual Information [43]. The Conceptual Information of the words can be fully detected via both meaning and context. Therefore, linking words to the appropriate senses may help in figuring out that conceptual information. These senses catch the identical meaning of the word and can be linked by means of lexical relationships between the synonyms sets.

The WordNet is that lexical resource that offers a wide range coverage of the general Conceptual Information, which makes it an essential resource for many Information Retrieval tasks that promote the SW functions. WordNet is neither a traditional dictionary nor a thesaurus; it rather combines the features of both. As a thesaurus, the synsets involving all expressing words for a certain concept. As a traditional dictionary, it gives a definition and sample sentences for most of its synsets [27].

The successfulness of Princeton WordNet (PWN) opens the way for other promising projects such as the Global WordNet project, which seeks to producing and linking all of the world's languages. As for the Arabic Language, the AWN is a linguistic resource for the

Modern Standard Arabic with a semantic foundation based on the PWN and linked to SUMO [44; 45; 46].

As we indicated before, the author of [31] has a viewpoint that distinguished the AWN from Arabic Ontology, as the AWN is neither an ideal nor alternative Arabic Ontology. The following two sections present the articles that dealt with the AWN in different ways. Some of those tried to exploit AWN to develop new systems and others aimed to enhance and enrich it.

As an attempt to emulate the AWN, AyaSpell-dic project [47] aims to provide a list of Arabic synonyms to be used as a free software. It collects linguistic data from printed dictionary. Currently, it is programming scripts to convert data into OpenOffice thesaurus format and testing thesaurus in OpenOffice on Linux and Windows. It planned to get more data from Arabic dictionary, check data consistency via further tests, and search about suitable applications for thesaurus. However, this application cannot be used as Ontology since its synsets are not mapped to an upper level Ontology and it has not hyponym/hypernym relationships, which can be interpreted as specialization relations between conceptual categories.

The article [48] proposed a framework that can understand Arabic Web content depending on the AWN to be used as a core of other semantic applications such as: SSE, semantic encyclopedia, Arabic QA systems, and semantic Dictionaries. The main goal of the system is to be able to convert any Arabic content into a machine understandable conceptual structure. It needs a WSD module and the AWN as its main components. However, the authors found that the existing AWN does not meet their needs, and decided to customize it. One of the customizations is the usage of a specific stemming-algorithm in order to store the stems. That ascribed to the diversity of the strategies used for storing stems in the original AWN, so the authors need to make their own. Another customization is the integration of the AWN with the PWN to find the interlingual translation. However, we believe that this step is preventable since it is already offered by the AWN, which is essentially built upon the PWN. The third adjustment is the denormalization of the AWN for speeding up the retrieval. The last is to paring the vowelized word with the non-vowels to maximize the findability, and that is a wise idea that may maximize the utilization. The system used a 'Tokenization and Indexing' as a preprocessing module before the WSD goes on. Then the 'Micheal Lesks' algorithm was used as the basis for WSD process. The WSD module faced some problems in dealing with Arabic plurals and conjugations[4]. An additional module then was used to determine the similarity between two conceptual contents, and a hierarchal clustering is performed after using the 'Bisecting K-means' algorithm. An Encyclopedia named Arapedia was developed to evaluate the framework. The results showed that searching for "معدن الذهب" returns contents related to "ذهب", as the 'gold', not those related to "ذهب" that means 'went'.

4 Actually, those problems have already been solved perfectly in the toolkit of RDI : http://www.rdi-eg.com/

3) *Arabic Named Entity Extraction*

The main goal of the SW technology is to annotate the data of the Web documents with predefined Ontologies in order to obtaining a machine readable, understandable, and processable Web. That will allow computers, software agents as well as human beings to work cooperatively together and exchange knowledge and resources [49; 50; 51; 52]. Most of the data that need to be annotated are those of the Named Entities form such as Persons, Locations and Organizations, which need to be extracted via NEE tools [53]. However, the sophisticated characteristics of the Arabic Language may complicate the process of Arabic Named Entities Extraction, since it mainly based on the morphological, syntactic, or even semantic analysis of the Web documents. Nevertheless, considerable attempts have been performed to develop ANEE systems, and they are considered in this section. Some of these attempts developed an ANEE to be used as a tool for improving and enriching the AWN content, as for [54], [55] and [56], while others, on the contrary, used the AWN as the base of the ANEE tool accomplishment as [30]. Moreover, other group as [57], [58], [59], and [60], aimed to build ANEE tools depending on more other techniques such as machine learning. The reviewed researches are organized according to the comparability of methodologies or results. Let's start with the first set.

The main idea of [56] and [61] is to automatically extract the Arabic Named Entities (NEs) from the Arabic Wikipedia, automatically attach them to AWN, and automatically link them to PWN. The overall approach is to extract the candidate instances from PWN and remove the generic types that have not Arabic counterparts or add them manually to the AWN. The WSD process is then executed depending on the 'Extracting Topic Signature' module, which results in a set of English Named Entities with disambiguation information attached. The next process is 'Filtering Candidates', which is the core of the approach. It is based on a local copy of Wikipedia for both Arabic and English loaded into a database. The process uses English NEs to lookup for a corresponding English Wikipedia page. Once it is located, it looks for the occurrence of an 'interwiki-link' to an Arabic page. The title of the corresponding page is returned as the Arabic Named Entity. Since the terms of the original AWN are vowelized, the authors believe that it is more appropriate to perform a further step, which is the vowelization, so the produced English-Arabic NE pairs will also be vowelized in order to harmonize the original version of the AWN. The system shows that 3,854 Arabic words corresponding to 2,589 English synsets were recovered; of which: 3596 (93.3%) were considered to be correct, 67 (1.7%) were wrong, and 191(5%) were not known by the reviewer. In fact, it is excellent work at least in dealing with Arabic Language's problems and difficulties such as the polysemy, WSD, and vowelization;

and trying to solve them in such a comprehensive manner. However, the research is restricted only to considering the Arabic Named Entities that have English counterparts in the English WordNet. Others that do not have corresponding English Named Entities, but have interwiki-links between Arabic and English Wikipedias, are attached as direct hyponyms of the corresponding generic synsets. That means that the Arabic Named Entities must have interwiki-links between Arabic and English Wikipedia to be extractable. Therefore, if the Wikipedia's article is originally constructed in Arabic and have not interlinked to a corresponding English page, then its Arabic Named Entities cannot be extracted. The main reason of that shortcoming is relying on the English Language in the entities extraction process as a consequence of the absence of the capitalization property in Arabic, which is a clue that enables the extraction of NEs easily. As for the weakness that is caused by the limited count of the existing Arabic Wikipedia's documents, comparing to the English ones, it is overlooked by the authors due to the highly growing ratio of the Arabic Wikipedia, which can be exploited to progressively improve the Named Entities coverage in the AWN. However, while the Arabic Wikipedia grows constantly, the translation may be not always kept pace. Therefore, the authors' defense of the problem may itself be a source of another problem.

Other researches that addressed the problem of the AWN enrichment were presented by Lahsen Abouenour et al in the Arabic Q/A system presented at [54] and [55]. In fact, the designed system achieved a mutual usefulness, since the AWN was being enhanced at the same time of the semantic expansion of the user's query. The authors exploited the Yago Ontology in carrying out their research. They studied the effect of the AWN enrichment using Yago Ontology in the context of Arabic Q/A systems. Yago is a large extendable Ontology of high quality, which contains about 3 million entities with 120 million facts. The authors justified the use of Yago that using mere NER system allows the identification of merely NEs, while the use of an Ontology like Yago enables the identification of the semantically related synsets. While [56] used the English WordNet as the ontological base of the AWN enrichment, and passed through the English/Arabic Wikipedia, [54] used the Yago Ontology instead, skipped the step the English Wikipedia translation, and extracted the Arabic entities by translating Yago English entities to the corresponding Arabic. Since Yago itself depends, in its construction, on the English WordNet, therefore, the use of Yago was a wise choice that saves lot of efforts because they started from where others have reached. In other words, it exploited the considerable effort that was exerted in Yago and built upon it. Moreover, the results of Yago were much better than those of [56] as while [56] starts with just 16,873 English NEs, Yago has 3 millions, since the WordNet mainly contains concepts rather than entities or instances. However, the researchers of [54] did not take the full advantages of Yago as it did not translate it to an equivalent Arabic NEs knowledge base. Instead, Yago is invoked just when an Arabic user's query is being processed. Once the query is ready, the Arabic NEs are extracted and translated into English using GTA[5]. The translated entities can then be extracted from Yago accompanied by their associated facts, then they are retranslated again into Arabic and mapped to their related entities in AWN according to synonymy, hypernymy, hyponymy, and SUMO relations. As for results, the system showed that the accuracy of question extending and answering is 23.53% using Yago, while it was 17.49% without. Likewise, 'Mean Reciprocal Rank (MRR)' is improved after using Yago to be 9.59 where it was 7.98 before. Furthermore, the number of the answered questions increased to be 31.37% using Yago, where it was 23.15% without. As we stated above, the relying of both [54] and [56] on the English Language as a first stair in their approaches led them to a trouble; they cannot capture the entities that are originally created in Arabic and have no corresponding English. When we asked the authors of [54] about this issue, the first author said: "Yes, this is right, but in the case of that research, since TREC and CLEF[6] questions were used in the tests and since these questions come from European and American culture, then all NEs exist principally in Yago. If we want to use the system with questions that contain Arabic NEs, we have to extend the AWN relying on Arabic NE resources. Even though, we have a great number of Arabic NEs in Yago". While the answer of the second author was "Yes that is right, not all Arabic entities appear in the English Wikipedia, but the goal of the project was not the full coverage of Arabic Named entities but to extend as much as possible the AWN with available entities and see if this extension has an impact in the query expansion module of a Q/A system". Actually, this impasse can be somewhat overlooked in the case of [54] due to the enormous gains that can be obtained from the exploitation of the powerful and rich resource, Yago, even though, we deem that it is not exploited as it should be.

Mohammad Attia et al demonstrated a system for constructing a repository of an Arabic NEs at [30]. The main difference between this system and the aforementioned ones is its reliance on just Arabic resources, AWN and Arabic Wikipedia, in the NEs extraction process instead of starting by English. The used methodology composed of several steps beginning by the 'Mapping' step, which maps the identified nouns of the AWN that can instantiate NEs to the corresponding categories and hyponym subcategories in the Arabic Wikipedia. Next, the 'NE Identification' step that identifies which of the articles of these categories are NEs to then be extracted and connected to AWN and inserted in the NEs repository. This is done by exploiting Wikipedia's interlingual links to find the correlation between articles in ten different languages to determine the NEs. As noticed before in [56], this method is

5 Google Translation API

6 TREC and CLEF are evaluation campaigns for information retrieval and question answering. There are question datasets from these campaigns annals.

commonly used in the NEs extraction phase due to the lack of something like the English capitalization feature in Arabic. For the Arabic articles that do not have correspondence in any other language, which are about 37%, there are two other heuristics to be used. The first is the 'Keyword Searching' in the article's abstract and then using the 'Geonames' for looking up entities. Subsequently, the 'Post-processing' step looks for further NEs from the Arabic Wikipedia, which are not reachable through AWN. The 'Diacritization', in which the acquired NEs are diacritized, is then taking place. Finally, the 'Vowelization' stage, as that of [56], is performed to uniform the extracted NEs with those that already are in the AWN. However, this step is done here via matching and comparing the Geonames. Whereas the system of [56] extracted 3854 Arabic NEs, about 45000 NEs were extracted by that of [30]. Moreover, its direct start from Arabic resources resolved the problem triggered by the NEs which are originally created in Arabic and have no English equivalents. Nevertheless, the system has some weak points, such as it just deals with the extracted NEs as a lexical repository; there are neither ontological facts nor relationships. As for the restriction of the limited count of Arabic Wikipedia's articles, and hence the abundance of the extracted NEs, it did not get a lot of authors' attention since, as the authors, the process is automated, and so the NEs amount will grow as the growth of Arabic Wikipedia itself. Furthermore, there is an obvious disparity in the number of categories available at [30] and those of Yago knowledge base, which is another factor that affects the amount of the extracted Arabic NEs.

The 'Machine Translation' and 'Rule Based' are two other approaches used in many NEs Extraction researches. However, they can only work effectively if a vast corpus is used as a source for the NEs. The 'Machine Learning' technique exploits a set of language's features as well as a small set of training data to get accurate results [57]. The abovementioned articles [54] and [30] are examples of the NEs extraction using 'Machine Learning', and we will discuss two others next; [57] and [58], as well as some 'Rule-Based' articles.

As for [57], the authors developed an NER system that is based on a self-training and semi-supervised learning approach. The idea of the research emerged from the authors' recognition of the importance of the presence of other categories besides the common ones: Person, Organization, and Location (POL). As they said: "Non-POL entities that stand out in the history- domain name e.g., important events (wars, famines); cultural movements (romanticism); and political, religious, scientific, and literary texts. Ignoring such domain-critical entities in some sense misses the point". The system identifies only entities mention positions, without associating it to any predefined categories. The system also developed a small corpus of Arabic Wikipedia articles and annotated it to named entities in order to facilitate annotation schemes, which allow annotators to define entity classes more readily. The corpus or dataset was made to provide a testing bed for the evaluation of the new NER models. Since the machine learning is not the subject of this survey, we will not go into the deep details of the system nor its conducting steps. However, in our own perspective, it is a reasonable methodology for determining and categorizing that vast amount of daily produced NEs, since the limitation of predefining a set of categories becomes inefficient. We believe that the efficiency of the Yago-based systems might be in the middle between the two extremes: 'Machine Learning' approach and 'WordNet-based' approach.

The research presented in [58] is another one that exploits the 'Machine Learning' approach in tackling the problem of Arabic NEs recognition. It integrates two techniques, namely, 'Bootstrapping semi-supervised Pattern Recognizer' and 'Condition Random Fields (CRF) classifier' as a supervised learner. The pattern recognizer extracts all expected patterns to let CRF identifying extra Named Entities. The system is supported by the RDI toolkit[7], which facilitates the deal with some of the Arabic Language difficulties. The system identifies ten NE categories, which are Person, Location, Organization, Job, Device, Car, Cell Phone, Currency, Date, and Time. The system consists of three main modules: 'CRF Classifier', the 'Bootstrapping Pattern Recognizer', and the 'Matcher Module'. The 'CRF Classifier' is used for segmenting and labeling the sequential data. To that purpose, it uses 15 features for the word. In fact, the RDI toolkit is used significantly in the specification of the values of these features. For instance, ArabSemantic RDI tool is used in the identification of the word's semantic field, which acts like the synonyms relationships in the WordNet. This is the first variation between [58] and other Arabic NER systems based on the AWN, such as [54], [56], and [30]. Another powerful contribution of RDI in this system is the extraction of the word's lexical features via its 'ArabMorpho-POS' tagger. The 'Bootstrapping Pattern Recognizer' bootstraps an enormous collection of Web pages for finding all occurrences of relation instances in a 150 MB News corpus coupled with their pattern representation. Hence, another variation appears which is the source of the NEs extraction. Here, it is a News corpus rather than the Wikipedia in the others. The last component, 'Matcher Module', is based on the RDI-AraMorpho-POS tagger and RDI-ArabSemantic tool. While NE keywords are frequently occurred within contexts so, using patterns, the identification of NE will depend on some indicators or phrases. That will substitute the most commonly used way in the Arabic NEs identification, which depended on the capitalization property of the English Language, and hence, avoid falling into its shortcomings. The proposed system runs the three modules sequentially till no new NE occurrences extracted. For each NE class, the three modules are run and fed up with the data set collected by ANERCorp and ANERGazet[8], and with a subset of the feature set. The 'CRF Classifier' yields some NE occurrences as the best seeds for the Pattern Recognizer,

7 http://www.rdi-eg.com

8 ANERcorp and ANERgazet are training set corpora and gazetter developed by [64].

which then uses the 'Matching Module' to produce good patterns for boosting 'CRF Classifier'. The results show that the technique may generate good patterns and work cooperatively with CRF using only small size of gazetteers datasets. The system also finds new NE occurrences within contexts easily, without the need of extraction work. These features give preference to use such Machine Learning techniques over the other techniques. The property that makes [57] surpass [58] is its ability to accommodate the increasing number of the developed NEs and hence the classes or categories they belong to, without the restriction of a predefined range. Definitely, significant benefits will be obtained if such system is enhanced with ontological features, in order to acquire a comprehensive semantic Named Entities knowledge base.

According to the implementation of the 'Rule-Based' approach in the ANEE, the articles [59] and its improvement or supplement [62] and [63] are addressed.

About [59], it aimed to extract just Person NEs from an Arabic corpus. Its authors developed a system called 'PERA', in which various Arabic corpora have been analyzed to get the best rules of Person NEs recognition. The recognition system has two components: the 'Gazetteer name list' and the 'Grammar rules'. Person names that are extracted from the available corpora and other resources are then used to build up a lexicon in the form of gazetteers, then the learned patterns and person indicators are used to derive a suitable grammar rules in order to provide high-quality recognition of the Arabic person names. Additional filtration mechanism is then employed for enabling revising capabilities. The system was developed to be incorporated with various language-independent applications. Therefore, the dictionary includes the corresponding English translation of the Arabic names as a metadata. The system results are evaluated by calculating the precision, recall, and F-measure over 46 sets. Their average in people's names recognition was 85.5%, 89%, and 87.5% respectively. The size of the 46 sets was not specified, so the reliability of the results cannot certainly be determined. By the authors, this work has several ongoing extending activities, such as recognition and categorization of other types of Arabic NEs such as Locations and Organizations. They indeed have been accomplished in the subsequent researches [62] and [63], in which a system called 'NERA' was developed. In fact, it is a respectable effort due to the addressing of many challenges of the Arabic Language, such as the complex orthographic system, the ambiguity, and the lack of resources in order to achieve satisfactory results in terms of Precision, Recall, and F-measure. The 'NERA' system is able to recognize 10 categories of Arabic Named Entities instead of just Person names in the previous version, 'PERA'. These categories are Person name, Location, Company, Date, Time, Price, ISBN, Measurement, Phone number, and File name. The system's results are summarized in Table 1, which shows the accumulative recognition accuracy achieved by each category against the reference corpora. Moreover, the authors have more ideas for the results to be improved more and more.

TABLE I.

ACCUMULATED ACCURACY OF THE 10 ARABIC NES [63]

No.	Entity type	Precision (%)	Recall (%)	F-measure (%)
1	Person	86.3	89.2	87.7
2	Location	77.4	96.8	85.9
3	Company	81.45	84.95	83.15
4	Date	91.2	92.3	91.6
5	Time	97.25	94.5	95.4
6	Price	100	99.45	98.6
7	Measurement	97.8	97.3	97.2
8	Phone no.	94.9	87.9	91.3
9	ISBN	94.8	95.8	95.3
10	File name	95.7	97.1	96.4

With respect to [64], it is a research that uses the 'N-gram and Maximum Entropy' method as a base for directing an Arabic Named Entity system called 'ANERsys'. To that end, the authors developed their own training set corpora 'ANERcorp' and gazetteers 'ANERgazet'. The main reason for building these resources is that there are no other free resources on the Web. Therefore, they were used during the research training process, and then released to be used as free resources[9]. ANERcorp consists of 316 articles from various domains, with 150,286 tokens and 32,114 types by 4.67 ratio of tokens to types. The ANERgazet, on the other hand, consists of three sub-gazetteers, which are Location, Person, and Organization. By observation and experiments, the 'Maximum Entropy Classifier' was built in order to specify a list of characteristics about the context in which the Named Entities usually appear. Then it estimates the different weights using the 'General Iterative Scaling (GIS)' algorithm. Finally, it builds a classifier that computes the probabilities for each word to be assigned to the considered class(es). The results of implementing the ANERsys on the ANERcorp testing corpus using the developed ANERgazet are improved from 62.72 % precision without ANERgazet to be 63.21%, and the recall increased to be 49.04% rather than 47.58%. The F-measure also increased to 55.23 from 54.11 without the ANERgazet.

The authors of [64] continued their work and presented the new version called ANERsys 2.0 in [65]. It varies from [64] in that it combines the 'Maximum Entropy' method with the 'POS-tagger' in order to improve the NER process when dealing with longer proper names. That increases the F-measure by 10 points over the first version of the system. It is noteworthy that there is another closely similar attempt to that in [65] in terms of its interest in extracting the Arabic collocation terms; as presented in the article [66], which presents a system that depends on the use of GATE[10]. In order to extract the

9 As we saw, they were used in [58]

10 GATE (General Architecture for Text Engineering): a software toolkit written in Java and widely used worldwide for NLP and IE especially those of Named Entities, http://gate.ac.uk.

Arabic collocation terms, the authors developed new 'Java Annotation Pattern Engine (Jape)' rules and used them on the 'Crescent Quranic Corpus'[11], which is tagged per word, verse, and chapter and contains binary or ternary additional information about morphology and POS. The extracted collocations in [66] are prepositional collocations in forms of Noun-Noun, Adjective-Noun, Verb-Noun, Noun-Proposition-Noun, and so on. The extraction pattern is defined firstly, then it is used to produce the appropriate (Jape) rule, which is passed as a parameter to 'ANNIE' transducer; the main component of the IE in GATE. The Arabic collocations are detected, displayed, and the annotated documents are saved into a data store with the new tags to be exploited in other tasks such as, by the authors, the automatic construction of domain Ontology. The 'AnnotationDiff' tool is used in the results evaluation. It compares the results of system annotation with others manually annotated with Noun-Adjective annotation. The AnnotationDiff calculates the F-measure as 0.66, which is a poor result, but the authors state that they are working on improving it.

The authors of the article [37][12] developed a tool called 'AraTation' for semantically annotating the Arabic News Web documents. The system carried out an 'Information Extraction' module that works as the first stage of the system. It is developed upon a dictionary that contains about 5000 predefined unique Arabic words belong to the News domain and fall in one of the four types: Person, Organization, Location, or Company. The research, unfortunately, has a shortcoming in dealing with the inflection issue of the Arabic words. As an example, the individuality indicator yields five distinct words if it invoked with the five related words: عراقي, عراقيين, العراق, عراق, عراقيون, although all of them are just inflections to the same stem عراق. That of course will affect the overall performance and reduces the rates the precision and recall. Of fairness, the authors were aware of this problem and stated it to be a future concern.

Lastly, a noteworthy work, Arabic Named Entity Extractor (ANEE), is presented in [60]. It is a project that can extract ANEs from both structured and unstructured data using semantic concepts to convey meaning-based results. It can solve linking issues such as finding the 'Arab descent in US cities'. For doing so, the user just selects "US cities" & "Arabic Names" then every mention of "New York City, Washington DC, Los Angeles, and Boston" found with "Mohamed, Ahmed, Usama, and Tarek" will be found. Just select the category and ANEE finds every mention, comprehensively and automatically. Therefore, it is not just a NER system; it is also working as a query answering system. It also finds the alternative names and aliases. For instance, it can recognize that the entity 'Abu Ammar' (أبو عمار) is the same as 'Yaser Arafat' (ياسر عرفات), which is not just great, but also a significant NER task. In addition, it can deal with the prefixes and suffixes issues. The ANEE provides over 25 predefined categories and 100 subcategories. Moreover, ANEE's taxonomy can be customized in order to adding taxonomies or modifying entity concepts. A tool called 'WORDCON' is used to easily Romanize the extracted lists of entities. This is a powerful feature that enables further searching, discovering, and analyzing capabilities. E.g., the Arabic entity "محمد" can be translated into English as "Mohamed", "Mohamad", "Mohamid" or "Mohammad". Another feature of ANEE is its ability of adding relationships between the entities, so the related entities just extracted if they appeared together. In fact, the methodology that is applied in this work is not certainly known. However, something like that may be based on Ontology, which is ensured through a phone conversation with the head of the producing company, as well as from the product's definition poster. As a whole, it seems a powerful application and worth to be tested.

According to the issue of the Named Entities Extraction, further detailed discussion is provided in [67], which presents a comprehensive survey on the information extraction approaches applied on Arabic in comparison with Latin languages. Moreover, it gives a detailed explanation of the Information Extraction techniques. It discriminates the NER and Ontology based extraction methods, gives examples for both Latin and Arabic. In addition, it lists the available and needed Arabic/English resources for doing that. Furthermore, it demonstrates the strength, weaknesses, opportunities, and threads of applying NER on Arabic Language.

4) SW and the Islamic Knowledge

As explained earlier, Ontology is the powerful conceptualism used for the SW realization. Therefore, constructing an Ontology for the Islamic knowledge will extend the elegant functionalities and advantages of the SW to the Islamic Knowledge domain, which will be so useful for those who are looking for a trustworthy and comprehensive source of Islamic Knowledge. Since most of the available information resources on Islamic Knowledge were considered with the natural language text processing rather than the semantic manipulation and machine interoperability, the SW technology can provide the semantic meaning that may help in understanding the Islamic Message described in the Holy Quran and Hadith. To the best of our knowledge, the most prominent articles considering the semantic representation of the Islamic Knowledge and Holy Quran were accomplished around 2009 and 2010.

In 2009, [38] is directed. It is one of the rare articles that attempt to represent the Islamic Knowledge using semantic technology. The researchers built a framework for recognizing and identifying *Semantic Opposition* terms using Natural Language Processing and domain Ontologies. At the time of its publication, the work was still in progress and, to our best knowledge, there is no subsequent version that presents a complete system. The authors stated that the Semantic Opposition, which is one of the strong rhetorical techniques of classic Arabic exhibited in the Holy Quran, could be represented as

11 The 'Crescent Quranic Corpus' was developed at Leeds University, The University the Articles [40], [73], and [74] also.

12 The article [37] was mentioned before in section 2.1, and to be discussed here in more details.

Ontologies. The authors' reasoning line used in the argument is logically convincing, and it justified the hypothesis. The proposed framework consists of two components; the 'Domain Ontology', which is limited to handle just the *Women* field, and the 'SemQ tool'. The system's input is a Quranic verse, while output is the list of semantically opposed words in the verse with the degree of opposition. However, it is a just prototype so there are neither results to be evaluated, nor evidences to support the validity of the system. In addition, it has an excessively narrow domain; while the Ontology is limited to the narrow field '*Women*'.

As for the work presented in both [68] and [39] , the authors considered the problem of structuring a computational model that can represent the real meaning of the word transparently. They found that there are two fundamental language's characteristics are not addressed by the NLP community, and they tried to address them in those researches. The first characteristic is the 'definition of word's meaning', which is defined through a collective effort of the language's users. While the second is the 'dynamical behavior of language's words' as some words adapt their meaning, some are obsoleted, and others are originated. Therefore, the research tries to provide a solution based on an ontological model for representing a dynamic and collaboration computational lexicon that can accommodate these two characteristics. The model is applied to the '*Time Nouns*' vocabularies of the Holy Quran. The ontological structure utilizes the word features as semantic units in its semantic representation. Therefore, the words 'woman' & 'girl' can be discriminated by the atomic components (features) gender, and adulthood. While the gender feature will be female for both, the adulthood will be the definition; woman will be an adult, and girl will be a child. This method of meaning representation is implemented over the Holy Quran *'Time Nouns'* scope. The Ontology is developed using UPON [13] ontological engineering approach [69] and OWL standard language representation, in order to enable the resource to be sharable and openly accessible. The Ontology is structured according to the hyponymy/hypernymy relationships between the concepts. The conceptual classification forms the main classes of the Ontology, which is divided into two main categories. The first is the 'top level classes' that represents the major concepts, which are reusable across different semantic fields and different languages. The other category is the 'lexical classes', which represents the actual words within a particular semantic domain. The structure of the ontological lexicon implies that: the deeper going into the hierarchy, the more gained argument by the computational formula, and so, the narrower the meaning. Therefore, using this approach in representing the language's semantics, the meaning of the word can be described by the combination of all features from the top level down until the word itself. For example, in an attempt to answer the question "what is the meaning of the word 'Ghadat – غداة'?" the final formula representing the word's meaning was:

Ghadat = Time + Specific + Tangible + Day +Light+ from dawn till Sunrise + Beginning of the day. Moreover, the Ontology is capable of naming a new phenomenon or concept by viewing the ontological hierarchical classification, while the Ontology is a set of features. The lexicographers, so, can either choose from or add to these features to generate the formula. A set of closed words are then displayed, so the lexicographer can select a target word and add a new sense to represent the new phenomena, or may coin a new word that is based on the meaning of the closely related words. Actually, it is a wise utilization of the ontological capabilities in the representation of the word's meaning transparently, which may open new prospects for the use of Ontologies in the Arabic Language. However, the authors encountered some problems such as the appropriate atomization feature for each concept and the depth level. Another problem is the difference in the meaning description of the same concepts from one person to another.

Saidah et al, on the other hand, attempt to create a framework that aims to develop a method for extracting the Islamic concepts, and then exploiting it in the automatic building of Islamic Knowledge Ontology. The authors published two papers in this regard, [70] and [71]. Of course, the development of Islamic Knowledge Ontology is not that trivial work; at least it needs a group of Ontologies in different fields of the Islamic Knowledge from different Islamic resources. Thus, the authors took just a step towards a hard work. They have presented an approach for the automatic generation of ontological instances from the Holy Quran basing on the combination of Natural Language Processing, Information Extraction, and Text Mining techniques. The system has been developed in order to generate an Ontology that only involves the Quran's verses that holds the 'solat' or prayer phrases. During the pattern extraction process, only the verses containing the keyword solah will be extracted, then the related Quranic verses will be verified with the contents of the 'surahs' for elaborating the hidden meaning, since the obtained verse might be hanging and the super-concept cannot be identified, and that will assist in generating the concept and relation. The verse then will be extracted according to the identified pattern. The whole system is still under construction, and the presented results are just a sample that does not give accurate indicators of performance. However, it is noted that the dealing with Islamic concepts and Quranic verses was in transliteration form rather than absolute Arabic characters. At least, that transliteration could be paired with the corresponding Arabic. That will encourage the Arabic users to use the system.

In Malaysia 2010, Juhana Salim et al treated the issue of Islamic multilingual Ontologies at [41]. Their research aimed to identify a Semantic Retrieval system to retrieve Islamic Knowledge in the three languages: Malay, English, and Arabic. Therefore, they developed a multilingual Islamic Ontology to be used for annotating the documents of the interested domain. To build that

13 Unified Process for ONtology

Ontology, the system used an Islamic Extraction system to extract the content of only reliable Web pages depending on resources and thesaurus such as the 'Library of Congress Subject Heading (LCSH)', the 'Library of Congress Classification (LCC)', and the 'Index Islamicus'. LCSH and LCC are used for expanding the extracted words, after discarding the stop words. All the thesaurus terms are then connected to concepts in the Ontology and given a hierarchical taxonomy using Border and Narrower relationships. They can get further relationships and properties via the BP 77.5 class of LCC. The 'Index Islamicus' is also used for further expansion of the developed domain Ontology. The Ontology, to be multilingual, is translated using 'Kamus Al Irshad empat bahasa'[14] and other Web tools such as Google Translator. The Ontology then is used in the annotation process. Actually, the research did not discuss the results sufficiently as they were presented just as a noisy figure. Therefore, the reliability of the system cannot be evaluated basing on that figure.

Other prominent works, which aimed to apply the SW technologies on various Islamic Knowledge sources for more flexibility and efficiency in the knowledge modeling, storage, publishing, reasoning and retrieval are presented at [72], [40], and [32],

According to [72], the authors presented a semantic-based knowledge representation model for the Holy Quran and its related resources in order to boost Quranic studies and research. The presented framework enables reasoning from a knowledge base of Quran and its related textbooks via Ontologies and semantic reasoning. The importance of this study lies in its attempt to deal with these critical resources comprehensively. In fact, an abundant set of Web sites have been designed to retrieve Islamic Knowledge on the level of keywords, but very few if any, those who provide a machine processable form on the level of information in this field. The system's framework starts by the 'Data Collection' process, which collects the Holy Quran, Ahadith books, and scholarly texts related. They are then standardized through the 'Metadata Generation' process, which parses text to extract data about data, and 'Tag Generation', which extract tags that will be used later in the annotation process. After that, Metadata and Tags are combined for formalizing the data representation in XML format. Subsequently, the 'Knowledge Modeling' process builds Ontology models using Ontology schemas for Quran and related textbooks. The Ontology is then populated by the XML representation documents, and then the populated Ontology is stored in Ontology repository. A contextual association between the Holy Quran and books of Ahadith is then performed in the 'Contextual Modeling' process. Finally, the system processes the given user query in the 'Knowledge Retrieval' process, and the results are passed to the user. In fact, the research faced various challenges due to the complexity and richness of the Islamic Knowledge content; the structural organization and thematic complexity of the Quran are examples. The organic unity and coherence in the structure of the Quran is another challenge. However, from our point of view, the most difficult challenge is the contextual linking between Quran and other related knowledge resources. In our own perspective, the research presented a great and considerable effort, due to the dealing with highly sensitive, critical, affluent, and complicated resources. Such systems may be considered as a source for the new Muslims for learning and understanding their religion. However, the system suffers from the drawback of the use of transliteration rather than the absolute Arabic characters in representing the Arabic words. As with [70] and [71], that will definitely affect and may hamper the use of such great system by a broad sector of the native Arabs. An easy solution is to couple the Arabic words with their transliteration in order to maximize the benefits. In this regard, the Quranic Ontology made in [40; 73] , on the other hand, is great in the sense that it provides the pronunciation both in Arabic and English letters. Nonetheless, it did not handle the related knowledge resources as done in [72] .

The article [74] concentrated on the applications presented at the Website http://corpus.quran.com/. One of these applications is the Ontology of Quranic concepts[15], which defines the concepts in the Quran, and uses the predicate logic to show the relationships between these concepts. The Named Entities, such as the name of historic people and places mentioned in the Quran, are linked to concepts in the Ontology. The reference [40] sheds light upon the other applications carried out in the Leeds University and related to the Arabic Language and Holy Quran.

5) Arabic Semantic Search Engines

Search engine is the most indispensable tool used in navigating the information published on the Web. However, the Web is the biggest global unstructured database, the matter which complicates its machine understandability and processability, and hence the ability to precisely obtain the desired information. Moreover, the query words may sometimes be ambiguous, since different people may use different terminologies for the same concept, synonyms, while the same user, on the other hand, may use the same word for different concepts, polysemous [37; 75]. Thus, most of the search engines encounter the problem of capturing the exact purport of the user's query. Therefore, the main task of the search engines is to correctly interpret the users' needs, process the relevant knowledge from different information sources, and return the accurate and relevant results to each user individually [76; 77; 78].

In terms of recall/precision, the traditional search engines can be described as: with high-recall, they have low precision, and with low recall, the precision becomes none. That is caused mainly due to the sensitivity of their results to the keywords, and the misinterpretation of the synonymous and Polysemous [79]. Therefore, even if the

14 Melayu - Inggeris - Arab - Urdu ; Inggeris - Melayu - Arab - Urdu

15 corpus.quran.com/ontology.jsp

main relevant pages are retrieved, there are irrelevant documents also retrieved, which affecting the precision parameter. On the other hand, if some or all of the relevant pages are missed, this leads to low or no recall. Consequently, the traditional keyword-based search engine is not appropriate to be used anymore. The alternative is the Semantic Search Engines (SSEs), the new promising generation of search engines. The SSE uses the Ontologies for indexing rather than the usual lexicons used in the traditional search engines. Therefore, the SSE can extract RDF triples that hold the metadata of Web documents, and so provide semantic information about the concepts indicated by the query's keywords and discover the relations between these keywords and concepts to reveal the meaning of the Web contents [24] [80]. By this means, the semantic search engines aim to find pages referring to certain concepts, rather than collecting all pages that just mentioned the query's keywords; which may already be ambiguous too [81; 82]. This manner, the problem of expressing the same semantic concept using different terminologies can be resolved, since all of these terminologies will be recognizable via the use of the ontological representation of that concept. Moreover, the semantic search engines can exploit the generalization/specialization properties of Ontologies hierarchy. Therefore, if it failed to find any relevant documents, it may suggest a more general query, and if too many answers are retrieved, the search engine may suggest some specializations [81; 82]. By means of understanding the meaning of query and its possible dimensions, the returned results will be more relevant, and those missed will be retrieved, which means higher recall with more precision [83].

Unfortunately, however, the Arabic language is still not fully supported through SSEs [8]. Although Swoogle[16], Hakia[17], SenseBot[18], and DeepDyve[19] are among the top SSEs, they have low to no support of Arabic. As Arabic Web sites are increasing constantly, search systems that handle the semantics of Arabic Language come to be essential. A Little work has been done on the Arabic semantic search; the following section presented most of them.

In [33], S.Zaidi et al described a Web-based multilingual Information Retrieval tool, relying on an Arabic legal-domain Ontology, as an attempt to improve both recalling and precision. The tool tries to reduce the amount of noise in the search results using an Ontology-based query expansion, since the query is being expanded through the navigation over the Arabic legal Ontology. In the case that the user wishes to get the results in English or French, the query is translated and expanded through the PWN rather than the Arabic legal Ontology. Note that the AWN had not been accomplished yet. The relevance of the returned documents was calculated on the first ten returned documents. The authors showed that the use of this tool improved the results as that while the recall was 115 and the precision was 2 of the first 10 documents before applying the tool, they are improved to be 1230 for the recall and 7 for precision after the tool application. The authors have a reasonable perspective; they believe that the 'cross-language retrieval' approach is practical in sharing and distributing the information regardless of the used language. However, another opposing perspective, which is presented at [31], states that the interlingual translation is sometimes inconsistent and has conflicts.

At [43], the authors aimed to improve the search results by expanding the Arabic query using synonyms and stems. They implemented a model for IR systems using components such as 'Arabic stemmer' and 'word synonyms structure', for improving the retrieval process of the Arabic search engine. They assumed that the expansion of the search keyword on the base of synonyms could potentially improve the system's recall, since the question's answers may include the synonyms of the basic keywords. The proposed system contains two service sides, namely client and server. The client side acquires the search query, removes stop words, views synonyms-tree, which based on Arabic synonyms database. It shows all the related synonyms located in the same keyword synonyms ring. Next, the search query will be expanded on the bases of the selected synonyms. Lastly, Arabic light stemmer is implemented on that expanded query. As for the server side, there are the 'indexed tokens wildcard cursor', 'wildcard cursor stemmer', and the 'indexed articles database'. Two evaluation approaches are made to evaluate results. The first is based on measuring the performance of the system against related synonymous words and the total number of retrieved relevant words after selecting the synonym(s). As for the other, it focused on comparing the total number of retrieved sense-related words with the result of single query word results. Although the work in the AWN was already standing at the publication time of this research, but it had not been accomplished yet. Nevertheless, this research has the advantage of constructing an Arabic WordNet core that is based entirely on Arabic Language resources, while the AWN, on the other hand, is constructed on the top of the English PWN, which exposes it to criticism of its heterogeneity that comes out of the different languages characteristics such as structure, semantic, and syntax. Nevertheless, the research tries to mimic the idea of the WordNet just in using synonyms and indexed tokens database that could serve as lexicon vocabulary storage and lacks the concepts relationships that characterize the WordNet and the Ontologies in general. Therefore, the research results cannot emulate those that can be obtained if an Arabic Ontology, or even AWN, is used rather than using mere Arabic synonyms DB, since these relationships allow the utilization of the real word's meaning as a concept that has relationships to other concepts with levels of specialization and globalization in the Ontology's hierarchy.

As for [75], it proposes an Arabic search engine, with some semantic level handling. While [43] developed an Arabic synonyms DB, as the AWN had not been yet

16 www.swoogle.com
17 www.hakia.com
18 www.SenseBot.com
19 www.deepdyve.com

available, [75] used a little like approach as [43] although it is published in 2007; i.e. after the producing of the AWN. The researchers developed a terminological dictionary of Arabic words as a handling to the semantic level of the document retrieval process, and it was developed as follows:

Each term xi belonging to the set of all existing terms V in the index of a collection of documents can be represented as:

Dic (xi) = {ti1, ti2, …, tik }, Eq.(1)

where tij is a term that is related or similar to the term xi with a corresponding degree of importance. So, the word "طب" is represented as: "الأمراض، الأوبئة، التشريح، أدوية".

Clearly, the research restricts the semantic functionality to the set of synonyms or related words; although the semantic level may be treated more beneficially if an ontological level is considered rather than just lexical. Therefore, it suffers the same limitations of [43].

Another work, [84], introduces a method for ranking Arabic Web sites using Ontology concepts. In this paper, Arabic Ontology for the electronic commerce domain in Arabic Language has been built, and then used for ranking Arabic documents. The proposed method ranks the documents according to the occurrence frequency of Ontology's concepts in the document, so the document that contains more electronic commerce concepts taking higher rank. The system is implemented by Visual Basic.net, and its performance is compared with three different search engines: AltaVista, Google, and Yahoo. With ranking the first 30 documents, the system's results showed that its ranking methodology is better 4.2 times than AltaVista, 4.5 times than Yahoo, and 2 times than Google. The authors ranked just 30 documents in order to be comparable with the ranking of an expert, and they indicated that the number of the considered documents do not affect the algorithm's performance nor results.

Another Arabic Semantic Search tool that is based on an e-commerce Ontology is SemArab, which is presented at [85] and [86]. This time, the e-commerce Ontology is a tree-like, and it contains just five e-commerce concepts, each with at most 10 branches of related concepts without any interrelationships between the siblings' branches. The relationships are just in terms of ancestor and descendent. SemArab has an easy to use GUI that enables the user to just type his query in a form of keywords, and select its corresponding ontological concepts. The system then searches the Web using a traditional search engine to get the related documents depending on the combination of the user's keyword and its corresponding concept plus all of the related concepts. The first 100 search results are then analyzed and filtered for seeking any occurrences of any of user's query keyword or related concepts. The found results are kept to be ranked after the measurement of the concepts similarity. The results are ranked according to the frequency of the concepts existing in the extracted documents. The research results are summarized in a Table 2, which shows the results of SemArab vs. those of traditional Web search engines.

TABLE 1:

SEMARAB VS. WEB SEARCH ENGINES [85]

Parameter name	SemArab		Google		Bing	
	Relevant	Irrelevant	Relevant	Irrelevant	Relevant	Irrelevant
Ahmad as a person	68 %	32 %	35 %	65 %	32 %	68 %
HP laptop as a sales	90 %	10 %	81 %	19 %	80 %	20 %
AlOthaim as an organization	74 %	26%	33 %	67 %	33 %	67 %
1999 SR as a various	65%	35%	37 %	63 %	29 %	71 %

In fact, the system used the Ontology just to get the function of the word's synonyms rather than to exploit its full power. Moreover, the used methodology, even if it improves the precision factor, will not affect the recall, since it cannot get those missing yet related URIs, which are not fetched ordinarily in the traditional search engines.

Ibrahim Fathy Moawad et al, in [42], showed an Arabic SSE that is based on Arabic Ontology and coupled with a traditional syntactic search engine. The Arabic Ontology provides a semantic interpretation of the user's query, and so improves the search results. Since the construction of an Ontology of the entire Arabic Language is extremely exhausting work and needs to be a standalone project, as that presented in [31], the authors considered just the domain of 'Computer Technology' and used its limited vocabulary in constructing the system's Ontology. Besides building that Ontology, the system's architecture is achieved via conducting other three modules, which are the 'Interactive Semantic Query Analyzer', 'Semantic Ranker', and the 'Interface with Syntactical Search Engine'. As for the 'Interactive Semantic Query Analyzer', it recommends the end user with extra semantic search criteria by means of accessing the Ontology. It then takes the user query as an input and extending it semantically through the matching with the Ontology's concepts, and thus uses it in invoking Google search engine to find the related documents. The search results are ranked in the 'Semantic Ranker' module using the 'concept frequency' technique rather than the 'term frequency' used by the syntactic search engines. The 'Interface with Syntactical Search Engine' module is used for coupling the 'Semantic Ranker' and 'Semantic Query Analyzer' modules with Google search engine. The system is evaluated by two trials using the computer domain terminologies, then the results were compared with those of Google. The returned pages using semantic query were significantly less than those returned by Google. As for the results, the authors just stated that: "this result helps in a great deal with more accurate search results", without any interpretation into measurements such as precision or recall.

Finally, the research [24] describes a method for extracting semantic RDF triples from Arabic Web pages. According to the authors, " the development of the SW technology is to adopt an intelligent algorithm that utilizes the four technologies: NLP, NN, Multi Agent Systems, and Fuzzy Logic Controllers". Thus, they presented a SSE model and referred to it as (543)

semantic model, as an abbreviation to the use of 3 ideas with 4 technologies and 5 meaning theories. The article proposed a detailed overview about some concepts such as RDF triples, OWL, OWL/RDF and Arabic Language characteristics and challenges. However, it did not give any framework nor prototype to implement the suggested (543) model.

A brief outline for more articles related to the Arabic SSEs and beyond can be found in the survey presented at [80]. It was concluded that very few Arabic semantic search systems have been developed and that the appearance of one may dominate the market.

CONCLUSION

In this study, we present a survey that, to the best of our knowledge, considered most of the researches concerning the usage of the Arabic Web content semantically rather than just syntactically. In fact, it is clearly deducible that the Arabic Language is still far away from the environment of the Web's third generation (SW) in terms of both quantity and quality.

Regarding to Ontologies, which are the backbone of the SW applications, they are remarkably rare in Arabic and those that exist are limited to narrow domains, and thus cannot be used in wider fields. Moreover, the only found Upper-Level Ontology for Arabic is still under construction. The study shows that various researches tried to overcome this deficiency using the AWN as an alternative to the Arabic Ontology. However, it has been experimentally proven that this is not appropriate in many cases for many reasons. It is, for example, not as effective as the English WordNet in terms of the count of the available items. Moreover, the reliance on the translation approach in the construction of AWN is not the optimal, due to the inconsistency of some concepts across languages because of the different cultures of their users. Accordingly, it is necessary to develop and improve the AWN to catch up with its English counterpart PWN, in order to achieve the greatest advantage when using it in applications that require the WordNet capabilities. As for those wider applications that require more comprehensive representation of the concepts, meanings and interrelationships, we seriously need more powerful and richer resources for Arabic Upper-Level and Domain Ontologies.

To get a complete overview, we consider the applications concerned with the ANEE. The Web explosion raises the need to reuse the included knowledge through the development of applications that can catch knowledge from texts and send it back to the user in the form that meets his requirements. That can only be met by extracting the information held in the Web documents and retrieving its meaning, which can be retrieved if the web document's metadata is annotated to a predefined Ontology. The majority of these information units are in the form of Named Entities. Therefore, they cannot be extracted without effective NEE tools. In the case of Arabic, these tools are limited since they depend mainly on morphological and linguistic tools and other utilities and resources such as Arabic Corpora and Gazetteers, and each has its hindrances. The existing morphological tools are inadequate, and those available for free are insufficient. Moreover, the other resources as Corpora and Gazetteers are not better-off, and their poverty and inefficiency exacerbated the problem.

The Arabic Ontologies, AWN, and ANEE all can be considered the infrastructure for larger applications such as those of Semantic Islamic Knowledge Representation and Retrieval, or those of Arabic SSEs. The performance of these applications is significantly affected by the complications suffered by its infrastructure.

Definitely, the essential reason for all of these difficulties is the nature of the Arabic Language itself, which often leads to the impossibility of dealing with it via machines. Nevertheless, this should not stop more effective efforts for achieving the best possible solutions that enable the Arabic Language and its users to take advantage of the new electronic technologies generally, and the SW particularly. This will not be achieved unless the Arabic electronic infrastructure is built. In a nutshell, we have to construct all of the necessary resources, such as the Arabic Ontologies, Arabic Gazetteers, AWN, as well as the complementary utilities such as Morphological analyzers and NEEs. Likewise, the support of the Arabic script in the SW tools such as OWL/RDF editors must be taken into consideration. If that infrastructure is achieved and is sufficiently strong, the development of wider applications will be better facilitated, and the obtained results will be more reliable and trustworthy.

REFERENCES

[1] Berners-Lee, Tim. "from the Semantic Web to the Web of Data". *http://www.slideshare.net/dpalmisano/from-the-semantic-web-to-the-web-of-data-ten-years-of-linking-up.*. [Cited: March 30, 2012.]

[2] Berners-Lee, Tim. "W3 future directions, Keynote speech". *First International Conference on the World*. http://www.w3.org/Talks/WWW94Tim/. Geneva: May 1994.

[3] Samhaa R. El-Beltagy, Maryam Hazman, and Ahmed Rafea. "Ontology Based Annotation of Text Segments". *SAC '07 Proceedings of the 2007 ACM symposium on Applied computing*. Seoul, Korea : March 11-15, 2007.

[4] Karin Breitman, Marco Antonio Casanova, and Walt Truszkowski. "Semantic Web: Concepts, Technologies and Applications". s.l. : *Springer* London Ltd, October 2010. ISBN 13: 9781849966214. 28.

[5] Noy, Natasha F."What do we need for ontology integration on the Semantic Web (Position Statement)". *Semantic Integration Workshop at the Second International Semantic Web Conference (ISWC-2003)*, Sanibel Island, Vol. 25 No.1. Florida, USA : October 20, 2003,

[6] A. Crapo, X. Wang, J. Lizzi, and R. Larson. "The Semantically Enabled Smart Grid".*Proceedings of the Grid-Interop Forum 2009*. November 17-19, 2009.

[7] Grigoris Antoniou and Frank van Harmelen. "A Semantic Web Primer". *The MIT Press; second edition edition*, London, England : March 21, 2008. ISBN-13: 978-0262012423.

[8] Lilac Al-Safadi, Mai Al-Badrani, and Meshael Al-Junidey.

International Journal of Computer Applications, Vol. 19 No. 4. April 2011.

[9] G. Madhu , A. Govardhan, and T.V. Rajinikanth. "Intelligent Semantic Web Search Engines: A Brief Survey". *International journal of Web & Semantic Technology (IJWesT)*, Vols. 2, No.1. January 2011.

[10] Li Ding, Deborah L. McGuinness, Tim Finin and Anupam Joshi. "Semantic Web Technologies: A Tutorial". *Rochester, NY : presented at Kodak Research Laboratories*. July 18, 2006.

[11] John Davies, Dieter Fensel , and Frank van Harmelen. "Towards the Semantic Web: Ontology-driven Knowledge Management".*Wiley*; 1 edition, ISBN-10: 0470848677. January 21, 2003.

[12] Jorge Cardoso. "On the Move to Semantic Web Services". *World Academy of Science Engineering and Technology WASET*. 2005.

[13] Ian Horrocks."Ontologies and the semantic web". *Communications of the ACM*, Vol. 51 Issue 12. New York. NY, USA : December 2008.

[14] Jiaqiang Dong, Yajun Du, and Mingli Feng. "A Novel Strategy for Constructing User Ontology". *International Journal of Digital Content Technology and its Applications*, Vol. 5 No.5. May 2011.

[15] Thomas B. Passin. "Explorer's Guide to the Semantic Web". *Manning Publications*, March 1, 2004. ISBN-13: 978-1932394207.

[16] Aidan Hogan, Andreas Harth, Juergen Umrich, Sheila Kinsella, Axel Polleres, Stefan Decker. "Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine". *Web Semantics: Science, Services and Agents on the World Wide Web*, Vols. 9, No. 4. 2012.

[17] Dean Allemang & James Hendler. "Semantic Web for the Working Ontologist, 2nd Edition: Effective Modeling in RDFS and OWL". *Morgan Kaufmann*; 1 edition. March 1, 2011. ISBN: 978-0-12-385965-5.

[18] M. Hildebrand. "Search-based user interaction on the semantic web, a survey of existing systems. Technical report". *The Netherlands : Centrum Wiskunde & Informatica*. 2007.

[19] J Davies, Rudi Studer, and Paul Warren. "Semantic Web technologies : trends and research in ontology-based systems". *Hoboken, N.J. : John Wiley & Sons*, 1 edition. July 10, 2006. ISBN-13: 978-0470025963.

[20] Raúl G. Castro, Asunción G. Pérez and Muñoz-García Óscar. "The Semantic Web Framework: A Component-Based Framework for the Development of Semantic Web Applications". *Turin : In DEXA '08: Proceedings of the 2008 19th International Conference on Database and Expert Systems Application*, pp. 185 - 189. 12 September 2008 .

[21] Benjamin Heitmann, Conor Hayes, and Eyal Oren. "Towards a reference architecture for Semantic Web applications". Proceedings of the 1st International Web Science Conference. 2009.

[22] *Jörg Wurzer*. "Application Challenges that may be amenable to Semantic technology solutions". *Riga, Latvia* : The 2011 STI Semantic Summit. July 2011.

[23] Lyndon Nixon, Elena Paslaru, Michal Zaremba, Enrica Dente, Ruben Lara, Walter Binder, Ion Constantinescu, Radu Jurga, Vincent Schickel-Zuber, Vlad Tanasescu, Mark Carman, Loris Penserini, Marco Pistore."State of the Art of Current Semantic Web Services Initiatives". *Knowledge Web Deliverable D1*. July 2004.

[24] Omar Isbaitan, Huda Al-Wahidi. "Arabic model for semantic web 3.0". *International Conference on Intelligent Semantic Web-Services and Applications*. 2011.

[25] Majdi Beseiso, Abdul Rahim Ahmad, and Roslan Ismail. "A Survey of Arabic Language Support in Semantic Web". *International Journal of Computer Applications*, Vols. 9–No.1. November 2010.

[26] Majdi Beseiso, Abdul Rahim Ahmad, and Roslan Ismail. "An Arabic language framework for semantic web". *International Conference on Semantic Technology and Information Retrieval (STAIR).* June 28-29, 2011.

[27] Sabri Elkateb, William Black, Piek Vossen, David Farwell, Adam Pease, & Christiane Fellbaum. "Arabic WordNet and the challenges of Arabic. The Challenge of Arabic for NLP/MT". *London : International conference at the British Computer Society.* October 23, 2006.

[28] Hend S. Al-Khalifa and Areej S. Al-Wabil. "The Arabic Language and the Semantic Web: Challenges and Opportunities". *International Symposium on Computers and the Arabic Language*. Riyadh, Saudi Arabia : November 2007.

[29] Didouh Omar. "Arabic Ontology and Semantic Web". *al-Mu'tamar al-duwali lil-lughah (al-lughat al-'Arabiyah bayn al-inqirad wa al-tatawwur, tahaddiyat wa tawqi'at*). Jakarta, Indonesia : July 22-25, 2010.
الأنطولوجيا العربية و الويب الدلالي : المؤتمر الدولي للغة العربية (اللغة العربية بين الانقراض والتطور -التحديات والتوقعات). جاكرتا, إندونيسيا: 22- 25يوليو .2010

[30] Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini and Josef van Genabith. "An automatically built Named Entity lexicon for Arabic". *LREC European Language Resources Association (2010).* Valletta, Malta : 2010.

[31] Mustafa Jarrar. "Building a Formal Arabic Ontology" (Invited Paper). *In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. Alecso, Arab League. Tunis: April 26-28, 2011.

[32] Soraya Zaidi and M. T. Laskr. "Review of Arabic Textual Terminology Tools for Ontologies Building". *MIC 2009 Management international conference*. Sousse, Tunisia : November 25–28, 2009.

[33] S. Zaidi, M.T. Laskri, and K. Bechkoum. "A Cross-language Information Retrieval Based on an Arabic Ontology in the Legal Domain". *The International Conference On Signal-Image Technology & Internet–Based Systems (SITIS'05)*. Morocco : 2005.

[34] Samhaa R. El-Beltagy, Maryam Hazman, and Ahmed Rafea."Ontology learning from domain specific web documents".*International Journal of Metadata, Semantics and Ontologies*, Vols. 4, No. 1/2, pp. 24 - 33. May 2009.

[35] Fatma Zohra Belkredim and Farid Meziane. "DEAR-ONTO: a derivational Arabic ontology based on verbs". *International Journal of Computer Processing of Languages*, Vols. 21, No.3, pp. :279-291. 2008.

[36] F. Belkridem, and A. El Sebai. "An Ontology Based Formalism for the Arabic Language Using Verbs and Derivatives". *Communications of the IBIMA*, pp. 44–52. 2009.

[37] Layan M. Bin Saleh and Hend S. Al-Khalifa. "AraTation: An Arabic Semantic Annotation Tool". *The 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS2009*). 2009.

[38] Hend S. Al-Khalifa, Maha M. Al-Yahya, Alia Bahanshal, and Iman Al-Odah. "SemQ: A proposed framework for representing semantic opposition in the Holy Quran using Semantic Web technologies". *Current Trends in Information Technology (CTIT).* March 01, 2010.

[39] Maha Al-Yahya, Hend S. Al-Khalifa, Alia Bahanshal, Iman Al-Oud and Nawal Al-Helwa. "An Ontological Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran". *The Arabian Journal for Science and Engineering*, Vols. 35, No. 2C. 2010.

[40] Abdul-Baquee Sharaf, Eric Atwell, Kais Dukes, Majdi Sawalha, Amal Al-Saif, Serge Sharoff , Katja Markert, Latifa Al-Sulaiti, Bayan Abu Shawar, Nora Abbas and Andy Roberts. "Arabic and Quranic Computational Linguistics Projects at the University of Leeds". *Proceedings of the workshop of Increasing Arabic Contents on the Web, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO).* Damasus, Syria : October 16-19, 2010.

[41] Juhana Salim, Siti Farhana Mohamad Hashim, and Akmal Aris ."A framework for building multilingual ontologies for Islamic portal". *Information Technology (ITSim), 2010 International Symposium*, Vol. 3, pp. 1302 - 1307. Kuala Lumpur : June 15-17, 2010. ISBN: 978-1-4244-6715-0.

[42] Ibrahim Fathy Moawad, Mohammad Abdeen, and Mostafa Mahmoud Aref. "Ontology-based Architecture for an Arabic Semantic Search Engine".*The Tenth Conference. On Language Engineering Organized by Egyptian Society of Language Engineering (ESOLEC'2010).* Cairo, Egypt : December 15-16, 2010.

[43] Hayder K. Al Ameed, Shaikha O. Al Ketbi, Amna A. Al Kaabi, Khadija S. Al Shebli,Naila F. Al Shamsi, Noura H. Al Nuaimi, Shaikha S. Al Muhairi. "Arabic Search Engines Improvement: A New Approach using Search Key Expansion Derived from Arabic Synonyms Structure". *Proceedings of the IEEE International Conference on Computer Systems and Applications (AICCSA '06)*, 2006.

[44] Sabri Elkateb , William Black , Horacio Rodríguez , Musa Alkhalifa , Piek Vossen , Adam Pease , and Christiane Fellbaum. "Introducing the Arabic WordNet Project". *Proceedings of the third International WordNet Conference (GWC-06).* South Jeju Island, Korea : January 22-26, 2006.

[45] Christiane Fellbaum, Musa Alkhalifa, William J. Black, Sabri Elkateb, Adam Pease, Horacio Rodríguez, and Piek Vossen. "Building a WordNet for Arabic". *Proceedings of the the 5th Conference on Language Resources and Evaluation LREC2006.* May 2006.

[46] Karim Bouzoubaa. "ArabicWordnet Use and Enrichment". [pdf Document] *Mohammadia School of Engineers*. Rabat, Morocco : Oct 17, 2010.

[47] Taha Zerrouki. "AyaSpell-Dic. Arabic thesaurus project". http://groups.google.com/group/ayaspell-dic/msg/4bf02344837b16af. [Cited: March 4, 2012.]

[48] Bassel AlKhatib, Mouhamad Kawas, Wajdi Bshara, and Mhd. Talal Kallas. "Ontology-Based Semantic Context Framework (OBSC) Framework for Arabic Web Contents". *International Symposium on Web Services - Zayed University*. Dubai : April 9-10, 2008.

[49] Jim Hendler.″ "Why the Semantic Web will never work" (note the quote marks) ". *8th Extended Semantic Web Conference - ESWC 2011.* Heraklion, Greece : June 2011.

[50] Aldo Gangemi, Sean Bechhofer, Asunción Gómez-Pérez, and Jim Hendler. "Introduction to the Introduction to the Semantic Web". *7th International Semantic Web Conference (ISWC 2008)*. Karlsruhe, Germany : October 2008.

[51] Rudi Studer. "Ontologies and Linked Data". *2011 STI Semantic Summit.* Riga, Lativa : 2011.

[52] Barry Norton. "Semantic Technologies: Origins, Linked Data and Beyond". *AI Bootcamp 2011 @ Ghana-India Kofi Annan Centre of Excellence in ICT (AITI-KACE)*. Accra, Ghana : March 2011.

[53] Mohit Behrang, Kemal Oflazer, and Noah Smith. "Named entity recognition from Arabic Wikipedia". *Qatar Foundation (QF) First Annual Research Forum*, p140. 2010.

[54] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. "Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet". *Workshop on LR & HLT for Semitic Languages, 7th Int. Conf. on Language Resources and Evaluation, LREC-2010*. Malta : May 2010.

[55] Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. "On the Extension of Arabic Wordnet Named Entities and Its Impact on Question / Answering". *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, KEOD 2010.* Valencia, Spain : October 2010.

[56] Musa Alkhalifa and Horacio Rodríguez. ″Automatically Extending Named Entities coverage of Arabic WordNet using Wikipedia". *International Journal on Information and Communication Technologies*, Vols. 3, No.3. June 2010.

[57] Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. "Recall-oriented learning for named entity recognition in Wikipedia". *Technical Report, Carnegie Mellon University CMU-LTI-11-012*. Pittsburgh, Pennsylvania : August 2011.

[58] Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy and Aly Fahmy. "Integrated Machine Learning Techniques for Arabic Named Entity Recognition". *International Journal of Computer Science Issues IJCSI*, Vols. 7, Issue 4, No 3. July 2010.

[59] Khaled Shaalan and Hafsa Raza. "Person Name Entity Recognition for Arabic". *Semitic '07 Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pp. 17-24. Prague, Czech Republic : June 2007.

[60] Taghride Anbar. "ANEE Product : COLTEC Computer Language Technology". http://www.coltec.net/default.aspx?tabid=221. [Cited: March 4, 2012.]

[61] Musa Alkhalifa and Horacio Rodríguez."Automatically Extending NE coverage of Arabic WordNet using Wikipedia". *3rd International Conference on Arabic Language Processing (CITALA'09)*. Rabat, Morocco : May 4-5, 2009.

[62] Khaled Shaalan, and Hafsa Raza."Arabic Named Entity Recognition from Diverse Text Types". *GoTAL '08 Proceedings of the 6th international conference on Advances in Natural Language Processing*, ISBN: 978-3-540-85286-5. Berlin : 2008.

[63] Khaled Shaalan, and Hafsa Raza."NERA: Named Entity Recognition for Arabic". *The Journal of the American Society for Information Science and Technology (JASIST)*.

NJ, USA : July 2009.

[64] Yassine Benajiba, Paolo Rosso, and José Benedíruiz. "ANERsys: An Arabic Named Entity Recognition. System Based on Maximum Entropy". *Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing,* pp. 143-153. 2007.

[65] Yassine Benajiba, and Paolo Rosso. "ANERsys 2.0 : Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information". *Proceedings of the 3rd Indian International Conference on Artificial Intelligence IICAI,* pp. 1814-1823. Pune, India : December 17-19, 2007.

[66] Soraya Zaidi, M. Laskri, and A. Abdelali. "Arabic collocations extraction using Gate". *Machine and Web .Intelligence (ICMWI), 2010 International Conference*, pp. 473 - 475. Algiers : 3-5 Oct. 2010 .

[67] Samir AbdelRahman, Maryam Hazman, Marwa Magdy, and Aly Fahmy. "Information Extraction". *ARABIC LANGUAGE TECHNOLOGY CENTER (ALTEC)* : The Pre-SWOT Analysis. unpublished

[68] Maha Al-Yahya, Hend S. Al-Khalifa, Alia Bahanshal, Iman Al-Oud and Nawal Al-Helwa. "An Ontological Model for Representing Computational Lexicons: A Componential Based Approach". I*n Proceedings of The 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE'10*. Beijing, China : Aug.21-23, 2010.

[69] De Nicola, A., Missikoff, M., Navigli, R."A software engineering approach to ontology building". *Information Systems, Elsevier,* Vol. 34, No. 2, pp. 258-275. 2009.

[70] S. Saad, N. Salim, Z. Ismail and H. Zainal. "A framework for Islamic knowledge via ontology representation". *Proceeding of International conference on information retrieval and knowledge management, CAMP'10*. Shahalam, Malaysia : 2010.

[71] S. Saad, N. Salim, and H. Zaina. "Islamic Knowledge Ontology Creation". *International Conference for Internet Technology and Secured Transactions ICITST*. London : 2009.

[72] Sumayya Baqai, Amna Basharat, Hira Khalid, Amna Hassan, and Shehneela Zafar. "Leveraging semantic web technologies for standardized knowledge modeling and retrieval from the Holy Qur'an and religious texts". *FIT '09 Proceedings of the 7th International Conference on Frontiers of Information Technology*. 2009.

[73] Kais Dukes, Eric Atwell and Nizar Habash. "Supervised Collaboration for Syntactic Annotation of Quranic Arabic". *Language Resources and Evaluation Journal (LREJ). Special Issue on Collaboratively Constructed Language Resources*. November 2011.

[74] Eric Atwell, Claire Brierley, Kais Dukes, Majdi Sawalha, and Abdul-Baquee Sharaf."A An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet". *Proceedings of NITS 3rd National Information Technology Symposium*. 2011.

[75] Naima Tazit, El Houssine Bouyakhf, Souad Sabri, Abdellah Yousfi, Karim Bouzouba."Semantic internet search engine with focus on Arabic language". *The 1st International Sysmposium on Computers and Arabic Language & Exhibition 2007 © KACST & SCS.* 2007.

[76] Jorge Cardoso. "Semantic Web services: theory, tools, and applications*". IGI Global*, ISBN-13: 978-1599040455. Mar 30, 2007.

[77] Martin Hepp, Pieter De Leenheer, and Aldo de Moor. "Ontology management: semantic web, semantic web services, and business applications". *Springer*, ISBN: 978-0-387-698899-1. New York ; [London] : 2008.

[78] Vipul Kashyap, Christoph Bussler, and Matthew Moran. "The Semantic Web: Semantics for Data and Services on the Web (Data-Centric Systems and Applications)". *Springer* , ISBN-13: 978-3540764519.15 . Aug 2008.

[79] Soumyarashmi Panigrahi and Sitanath Biswas. "Next Generation Semantic Web and Its Application". *IJCSI International Journal of Computer Science Issues*, Vols. 8, Issue 2. March 2011,

[80] Samhaa R. El-Beltagy. *Technology : Semantic Search.* s.l. : ARABIC LANGUAGE TECHNOLOGY CENTER (ALTEC) : The Pre-SWOT Analysis, Feb 2010.

[81] Meena Unni , K. Baskaran. *"*OVERVIEW OF APPROACHES TO SEMANTIC WEB SEARCH". *International Journal of Computer Science and Communication*, Vols. 2, No. 2, pp. 345-349. July-December 2011.

[82] Walter Renteria-Agualimpia, Francisco J. López-Pellicer,Pedro R. Muro-Medrano, Javier Nogueras-Iso, and F.Javier Zarazaga-Soria1. "Exploring the Advances in Semantic Search". *International Symposium on Distributed Computing and Artificial Intelligence*. 2010.

[83] Junaidah Mohamed Kassim and Mahathir Rahmany. "Introduction to Semantic Search Engine". *International Conference on Electrical Engineering and Informatics ICEEI '09*. Selangor :2009.

[84] Zakaryia Qawaqneh, Eyas El-Qawasmeh, and Ahmad Kayed. "New Method for Ranking Arabic Web Sites Using Ontology Concepts". *2nd International Conference* on *Digital Information Management* (*ICDIM '07*), pp. 649-656. Lyon, France : October 2007.

[85] Majdi Beseiso, Abdul Rahim Ahmad , and Jamilin Jais. "Semantic Arabic Search Tool". *Semantic Technology and Knowledge Engineering - STAKE* . Kuching, Sarawak : July 2010.

[86] Majdi Beseiso, Abdul Rahim Ahmad , Jamilin Jais. "A New Architecture for Semantic Arabic Search Tool". *International Conference on Islamic Arts & Architecture (ICIAA)*. Bangalore, India : July 2010.