Semantic Relations Extraction and Ontology Learning from Arabic Texts—A Survey

Aya M. Al-Zoghby, Aya Elshiwi and Ahmed Atwan

Abstract Semantic relations are the building blocks of the Ontologies and any modern knowledge representation system. Extracting semantic relations from the text is one of the most significant and challenging phases in the Ontology learning process. It is essential in all Ontology learning phases starting from building the Ontology from scratch, down to populating and enriching the existing Ontologies. It is challenging, on the other hand, as it requires dealing with natural language text, which represents various challenges especially for syntactically ambiguous languages such as Arabic. In this paper, we present a comprehensive survey of Arabic Semantic Relation Extraction and Arabic Ontology learning research areas. We study Arabic Ontology learning in general while focusing on Arabic Semantic Relation Extraction particularly, as being the most significant, yet challenging task in the Ontology learning process. To the best of our knowledge, this is the first work that addresses the process of Arabic Semantic Relation Extraction from the Ontology learning perspective. We review the conducted researches in both areas. For each research the used technique is illustrated, the limitations and the positive aspects are clarified.

Keywords Knowledge extraction · Arabic semantic relation extraction
Arabic ontology learning · Semantic relation extraction between Arabic NEs
Semantic relation extraction between Arabic ontological concepts

A.M. Al-Zoghby (✉) · A. Elshiwi · A. Atwan
Faculty of Computers and Information Systems, Mansoura University, Mansoura, Egypt
e-mail: aya_el_zoghby@mans.edu.eg

A. Elshiwi
e-mail: Aya.Elshiwi990@gmail.com

A. Atwan
e-mail: Atwan.4@gmail.com

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_11

1 Introduction

Instant growth in the size of the World Wide Web makes it be the world's biggest repository of data. The amount and the unstructured nature of the majority of this data represent two main issues regarding dealing with the WWW for both humans and machines. As for humans, the amount of data is huge to be processed while the unstructured format cannot be understood by machines. As a result, a challenge in the interoperability between humans and machines showed up. To address this challenge, Tim-Berners Lee invented the Semantic Web as an extension of the current web with the vision that is given a well-defined meaning and provides better cooperation between humans and machines [1].

Ontology can be considered as a gateway to achieve the vision of the semantic web. It is used to represent the data in a way that enables machines to understand its meaning, and allow it to be shared and reused. The most commonly used definition of Ontology is "*Formal, explicit specification of a shared conceptualization*" [2]. The definition is explained in [3] as "*conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concept used, and constraints on their use are explicitly defined. Formal refers to the fact that the Ontology should be machine-readable. Shared reflects the notion that an Ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group*".

Ontology is the "backbone" of the semantic web as described by many researchers. Due to the importance and the full reliance of the semantic web upon Ontology, building it automatically is a significant and, at the same time, a very challenging task. On one hand, it requires dealing with the knowledge acquisition bottleneck [4]. On the other hand, it is affected by the heterogeneity, scalability, uncertainty, and the low quality of web data [5].

Ontologies can be built manually, semi-automatically or automatically. The manual construction of the Ontology has limitations as being expensive, time-consuming, and error-prone [6]. Moreover, it requires specialized domain-experts and Ontology-engineers [7]. To overcome these limitations, a new research area called "Ontology Learning" has emerged aiming to automate or semi-automate the process of building the Ontology [8]. Ontology learning involves the extraction of knowledge through two main tasks: the extraction of concepts (that constitute the Ontology), and the extraction of the semantic relationships among them.

Semantic relations are the building blocks of the Ontologies and any innovative knowledge representation system [9]. Extracting semantic relations from the text is one of the most significant and challenging [4] phases in the Ontology learning process. It is essential in all Ontology learning phases starting from building Ontology from scratch down to populating and enriching existing ones. It is challenging, however, as it requires dealing with natural language texts which represent various challenges especially for syntactically ambiguous languages such as Arabic.

Despite the importance of Arabic language as being one of the six most spoken languages in the world [10], and the obvious growth of its content on the web in the past few years, it has little support in semantic relation extraction in particular and Ontology learning in general. Automatic extraction of semantic relations from Arabic text is not well investigated compared to other languages such as English [11, 12]. While, most of the trials to generate Arabic Ontologies are still done manually [13–16]. Moreover, the nature of the Arabic language added extra challenges in knowledge extraction from Arabic text. As a result, the Arabic language suffers a lack of Ontologies and the semantic web applications in general [17, 18].

In this paper, we study the Arabic Ontology learning in general while focusing on Arabic Semantic Relation Extraction particularly, as being the most significant, yet a challenging task in the Ontology learning process.

To the best of our knowledge, this is the first work that addresses the process of Arabic Semantic Relation Extraction from the Ontology Learning perspective. We reviewed the conducted researches in the area. For each research, the used technique is illustrated, the limitations and the positive aspects are clarified.

The rest of this paper is organized as follows: Sect. 2 discusses the Arabic semantic relation extraction and Ontology learning. Section 3 reviews the Arabic semantic relation extraction trials, while the Arabic Ontology learning literature is represented in Sect. 4. Finally, the paper is concluded in Sect. 5.

2 Arabic Semantic Relation Extraction and Ontology Learning

Arabic is the native language of over 20 countries spoken by approximately 400 million native speakers around the world. It is also the 5th most spoken language in the world [19]. The past few years have witnessed an obvious increase in Arabic content on the web. Therefore, the Arabic language requires a further work in building Ontologies that will exploit the web data, lead to achieve the semantic web vision, and to keep up with the importance of the language. However, the Arabic language has challenges that affect the overall process of Ontology learning in general and the semantic relation extraction as a separated stage of Ontology learning, in particular.

Since learning Ontologies from the natural text requires the usage of NLP techniques, the nature of the Arabic language represents a major challenge. As Arabic is a Semitic language that differs from Latin languages in syntax, morphology, and semantics [20], therefor, the NLP algorithms for other languages cannot be applied directly on Arabic [21]. Also, its morphological analysis is complicated due to its agglutinative, derivational, and inflectional characteristics

[22]. In addition, it is written from right to left, missing capitalization, and it has ambiguities related to typography as the Arabic letters shape changes according to their position in the word [23]. Moreover, the availability of Arabic linguistic resources and tools, that are designed according to the specific features of the Arabic language, represents another challenge. Linguistic resources, such as corpora and dictionaries, are either rare or not free [24, 25]. While, according to [26], there are no corpus analysis tools for the Arabic language. Most of the existing tools of morphological analysis and POS have several limitations, such as being unable to provide full parsing or remove all the ambiguities.

All these challenges are standing as a barrier in the way of learning Ontologies from an Arabic text. Recently, researchers are dedicating extensive work to address the Arabic NLP challenges. Some of these that are related to the highly inflectional and derivational Arabic nature, Part of Speech (POS) tagging, and morphological analysis have been resolved to some extent [27]. But, still there is a lack of Arabic Ontologies and, as a result, a lack of its semantic applications.

In this review, we discussed the semantic relation extraction and Ontology learning both together; as they are highly related in the Arabic language. As we found out, the majority of the work dedicated to extract the semantic relationships was for the purpose of building Ontologies. Moreover, the extraction of those relations was a small part of the research except for some that were mainly dedicated to extract the semantic relationships.

3 Arabic Semantic Relation Extraction

In this section, we reviewed all researches considering the extraction of semantic relations from Arabic text.

In our study, we found out that researches in this area fall into two categories. The first contains the studies considering the extraction of the semantic relations between Arabic NEs. That mainly was for the purpose of using this knowledge in several NLP applications such as: question answering, text mining, and automatic document summarization; see [11, 12, 28–32]. The second category, on the other hand, contains researches considering the extraction of semantic relations between Ontological concepts for the purpose of using them in the construction of Arabic Ontologies, as in [11, 12, 14, 15, 24, 33–42].

From our point of view, we consider that the two categories are complementary. This is because the first category can be used for populating the instance level in an existing Ontology, while the second can be used for constructing the Ontology. Both categories will serve the Arabic Ontologies learning process.

Table 1 Summary of extracting Arabic semantic relation between the NEs

Ref.	Technique
[28]	Rule based
[29]	Rule based
[30]	Machine learning
[32]	Machine learning (Enhanced approach)
[31]	Hybrid (Machine learning and Rule based and Manual)
[11]	Hybrid
[12]	Rule based

3.1 Semantic Relation Extraction Between Arabic Named Entities

The extraction of the semantic relations between NEs hasn't been well investigated in Arabic compared to other languages such as English [32].

As mentioned earlier, there are several challenges related to semantic relation extraction between NEs other than the challenges related to the ambiguity and complexity nature of the Arabic language.

Moreover, the Arabic sentences are long, which might cause the existence of more than one NE in the sentence without being semantically related [30, 31]. Also, the position of semantic relation in the Arabic sentences is hard to be determined as it is not fixed; it may occur before first NE, between the NEs, or after the second NE [30].

Furthermore, the Arabic Semantic relation can be noun, verb, or even preposition. That differs from the English language, in which the relation is usually a verb occurs between the NEs pair [30].

In addition, the is a difficulty in determining the implicit relations that can't be directly specified from the text; it can only be understood from the previous context [28]. Moreover, negative, ambiguous, and multiple words relations between NE pairs, represent extra challenges [28, 31].

The previous trials to extract semantic relations between Arabic NEs succeeded to address most of the stated challenges. Table 1 summarizes the trials of Arabic Semantic Relation Extraction between Named Entities; categorized them into: rule based, machine learning and hybrid approaches.

3.1.1 Rule-Based Approach

The articles [28, 29] proposed a rule based approach for the extraction of semantic relations between Arabic NEs. Both of them used the same technique; by first,

identifying a set of linguistic patterns from the training corpus, and then transforming those patterns into rules and transducers using NooJ.[1]

[28] Focused on the extraction of functional relations between the PERS and ORG NEs. A number of journalistic articles and some data from Wikipedia were used as a training corpus. Both NEs and relations were recognized using a collection of manually built dictionaries. For PERS NEs recognition, the *Titles*, *First Names* and *Last Names* dictionaries were used. For ORG NEs recognition, on the other hand, the *Geographical Names*, *Type-Institution* and *Adjective* dictionaries were used. And finally, for relations recognition, the *Functions* dictionary was used. This approach obtained 63%, 78% and 70% in terms of precision, recall and F-measures, respectively.

[29], on the other hand, focused on the extraction of the relations between (PERS-PERS, PERS-LOC, PERS-ORG and ORG-LOC) NEs pairs. For each pair of NEs, several types of relations were identified by the authors. Then, for each type of relations, all the possible syntactic patterns were extracted, and a general pattern for this relation type was built. This approach obtained an F-score of 60%.

The negative aspect of both trials is the manual identification of the relations, which is a tedious time consuming task. This is in addition to the limitations of rule based approach and the necessity to the fully cover of all rules that might satisfy any kind of relations.

3.1.2 Machine Learning Approach

A supervised machine learning method for extracting the semantic relations between Arabic NEs was proposed in [30]. This method was based on the rule mining approach. Its main idea was to extract the position of words surrounding NEs that reflect the semantic relation. In order to apply that idea, two of the previously mentioned challenges of Arabic language were faced. The first was the complex syntax and the length of Arabic sentences, which caused the existence of two NEs in the same sentence without being semantically related. In order to handle this challenge, the authors limited the context by splitting sentences into clauses using Arabic clauses splitter, since each clause is composed of a set of words that contains a subject and a predicate thus, ensuring the existence of a relation between NEs. According to the authors, that solution tackled the problem on average of 80%. The second challenge, on the other hand, was the non-fixed position of the relations in Arabic sentences. In order to handle this challenge, the authors identified the position of words that represent relations in sentences manually. The proposed method consisted of three steps; building training data, automatic rules generation, and selection of the significant rules. In the "first-step", the training dataset was composed by extracting 15 learning features from sentences that contained at least two NEs. Those features were lexical (NEs tag), numerical (number

[1] http://nooj4nlp.net/pages/resources.

of words before, after and between NEs) and semantic (POS of words surrounding NEs). Of these features, 14 were extracted automatically, while the position of the relation was annotated manually by linguistic experts. Each one of the annotated relations was assigned a class that identified its position in the context. In the "second-step", Apriori, tertius, and C4.5 association rules algorithms were applied to automatically extract the highly precision rules. Finally, in the "third-step", the previously generated rules were filtered to select the most significant ones. To evaluate the effectiveness of the proposed method, the authors created an Arabic corpus from electronic and journalistic articles. The method obtained 70%, 53.52% and 60.65% for precision, recall, and f-score, respectively. Authors justified the low recall by the lack of a very efficient Arabic NE recognition tools and the failure of the system to extract the implicit relations between NEs. The main drawback of the system was the need of highly annotated data that was caused from using a supervised machine learning methods. Also, the semantic relations between NEs were extracted fully manually.

In a subsequent enhancement trial to improve the overall coverage of the previous method, authors in [32] used the Genetic Algorithm as a refinement step. The main idea behind using the Genetic Algorithm was to improve the quality of the rules generated from learning algorithms by constructing new rules that are fitter. Crossover and mutation reproduction methods were applied to the selected rules. The results showed that the usage of the Genetic Algorithm increased the precision and recall by 8%.

Another technique to automatically extract relations from the Arabic text was proposed in [12]. The technique was based on the enhancement of Hearst's Algorithm to fit the Arabic language, then integrating it into a four-module framework to extract the relationships. The "preprocessing and feature extraction module" is automatically extracting four different language components; word, POS tag, stem, and phrase. In additionally, three different types of relationships hyponym-hypernym, causality, and hierarchical were manually specified. Using components and relations specified in the previous module, a linguistic expert decided, for each couple of components, the suitable relations. Then, the training set was formed by automatically extracting the matching examples from the text. After that, an enhanced version of Hearst's Algorithm was applied in the "lexical syntactic pattern module". Finally, an evaluation was performed to all the extracted patterns to remove the dirty ones. While, in the "pattern expansion module", the authors expanded the lexical structure of the patterns to include synonyms using Arabic wordNet.[2] Then, enriched the relations with the related concepts. According to the authors, this expansion allowed discovering new relations; however, it caused redundancy as several patterns were covering the same relationship. This problem was handled in the "pattern filtering and aggregation module", where the authors implemented a validation algorithm to filter the redundant patterns that cover the same data instances. To evaluate the efficiency of the framework, it was tested in

[2]http://globalwordnet.org/arabic-wordnet/.

three different datasets: Holy Quran (classical Arabic), Newspapers articles (Modern Standard Arabic), and social blogs (unstructured Arabic text). The overall performance averages of the three datasets in terms of precision, recall and F-measures were 78.57%, 80.71% and 79.54% respectively. According to the authors, when comparing the performance of the proposed technique with the original Hearst's Algorithm, repeated-segments, and co-occurrence based techniques; it achieved the highest performance among all. They, also, studied the effects of different factors such as the type of data in the overall performance of the system. They found that it directly affected the performance; negatively in the classical dataset and positively in Modern Standard Arabic dataset.

3.1.3 Hybrid Approach

So far, authors in [29, 30, 32] were succeeded to label explicit one-word relations between (PERS-LOC, PERS-PERS, PERS-ORG, ORG-LOC, and LOC-LOC) NEs pairs. The authors successfully addressed some of the previously mentioned challenges such as the long Arabic sentences and the non-fixed position of Arabic relations in sentences. For a further enhancement, the authors presented a hybrid approach in [31]. The approach combined the previously discussed machine learning and rule-based methods with a manual technique to extract the semantic relations between the (PER, ORG, LOC) NEs pairs. It presented several enhancement modules for the purpose of addressing further challenges such as multiple words, negative and other complicated relations.

In the first enhancement module, the overall performance of machine learning method was enhanced by partitioning the dataset into verbal and nominal sentences. According to the authors, the partitioning process proved efficiency as it increased the precision and recall by 6.6% and 3%, respectively, compared to the previous machine learning results. It also proved that, for the Arabic language, the position of the relation between NEs depends on phrase structure. In the second module, the challenges of negative and multiple words relations were addressed. This is done by using handcrafted rules proposed by linguistic experts. As for the negative relations, a set of constraints were added to each rule in order to verify the existence of a negative particle that expressed the negation relation. Taking into account that, the position of the negative particles differs according to the sentence type either it is verb or noun. To handle the multiple words relations, on the other hand, the compound words that expressed a relation between NE pairs were collected into a list. Then, using Nooj,[1] all the syntactic grammars were elaborated. The elaborated grammars represented all the classes that the compound word relations belong to, either before first NE, between the two NEs, or after the second NE. Those grammars were then applied to the corpus to extract further multiple words relations. Finally, in the case of implicit and more complicated relations, the authors relied on manual patterns. To evaluate this approach, the authors compared the results of the system against machine learning method in [30] and rule based method in [29] using the same test corpus. The results showed that the system

achieved the best improvements and obtained 84.8%, 67.6% and 75.2% in terms of the precision, recall and F-score, respectively. We found that this technique is promising as it succeeded to take advantages of the machine learning, pattern-based approaches, and the manually generated rules in order to handle the complicated relations. Moreover, it is the first work to deal with different types of relations and handle the challenges in a good way with satisfactory results. However, to evaluate its performance in general we can say that handling the complicated sentences manually is not the ideal option due to the nature of the Arabic language, which caused several complicated cases that would require extensive time and effort to be dealt with. Also, the manual identification of relations position in sentences is not practical, as the main challenge is to automatically or semi-automatically extract semantic relations from Arabic sentences.

Another hybrid approach was proposed in [11], the approach mixed the statistical calculus and the linguistic knowledge to extract Arabic semantic relations from a vocalized Hadith corpus. The main idea of this approach was to use the statistical measures to calculate the similarity between terms in order to interpret syntactic relations, then exploiting these relations to infer semantic relations. The authors considered three analysis levels; morphological analysis, syntactic analysis and semantic analysis. In the morphological analysis level, AraMorph[3] analyzer was used to analyze the corpus text and extract tokens, their morpho-syntactic category, and the English translation. While in the syntactic analysis level, one type of noun phrases, which is the prepositional phrases, was extracted from the corpus by applying the grammar rules corresponding to this type of phrases. The first component of the noun phrase was considered as the head, and the second component was the expansion. A syntactic network linking the heads and the expansions with the syntactic relations was generated, and every syntactic relation was represented by the preposition. In the semantic analysis level, a dependency graph, which contained the most common syntactic relations (prepositions) in the corpus, was generated. For each preposition, all the correspondence semantic relations were linked to it. This graph was built based on Arabic grammar books, as authors clarified that the Arabic grammar rules were behind the relationship between the syntactic and semantic relations. By that stage, the syntactic relations were exploited to infer the semantic ones. Finally, the statistical measures were used to solve the ambiguity in the extraction of relations between any couple of terms. By first, calculating the similarity between these terms using contingency table based measures. Then, performing an enrichment of signature of this couple by adding the signature of the nearest couple to it depending on the similarity results. To evaluate this approach, the authors performed experiments in drinks, purification, and fasting domains. They presented two results of experimenting two different definitions of contingency table. The first experiment showed that the success rate of semantic relation extraction in the three domains was weak and didn't exceed 65%. They justified that result by a partial failure in the enrichment operation. So, they

[3] http://www.nongnu.org/aramorph/.

performed the same experiment with the second definition of contingency table. The success rate was improved and reached 75%. Further enrichment had been carried out and it improved the success rate as it reached 97 and 100% in the field of purification. When comparing their results to the co-occurrence approach results, according to the authors, their approach was more effective in some cases while, in other cases the two approaches extracted the same relations. They concluded that the two approaches were complementary.

3.2 Semantic Relation Extraction Between Arabic Ontological Concepts

The extraction of semantic relations between Arabic Ontological concepts is affected by the same challenges and limitations caused by the nature of the Arabic language. In this context, we mean by Ontological concepts; concepts from a seed Ontology, concepts that are extracted from text during the trial to build an Ontology, or a given list of concepts to extract semantic relations between them. Extracting semantic relationships between Ontological concepts was performed as a step in the Ontology learning process. So, we discuss the used techniques in details in a subsequent Sect. (4.2).

4 Arabic Ontology Learning

The effort given to Arabic Ontology Learning research area is very little to keep up with the importance of the Arabic language. We discuss all Arabic researches that either build Ontology from scratch, populate, or enrich an existing Ontology. In this section, we review trials to learn Arabic Ontology from natural text. We categorize the Arabic Ontology learning literature, according to [43–45], into upper Ontology and Domain-specific Ontology. Table 2 summarizes the reviewed trials to build Ontology for the Arabic language.

4.1 Upper Ontology

Upper Ontology, also known as (Top level, Foundational or Universal Ontology), is defined as *"an Ontology which describes very general concepts that are the same across all knowledge domains"* [46]. It provides a foundation to guide the development of other Ontologies. Moreover, it facilitates the process of mapping between them as it is easier to map between two Ontologies that are derived from a standard upper ontology [44]. There are many upper level Ontologies for English

Table 2 Summarizes Arabic ontology learning trials

		Ref.	Learning resource	Ontology learning technique	Domain	Relation extraction technique
Domain ontology	General domains	[41], [42]	Structured	Manual	Arabic verbs derivational ontology	Derivation based
		[15]	Unstructured	Manual	Computer domain	Manual
		[14]	Structured	Manual	Computer domain	Manual
		[34]	Unstructured	Manual	Legal domain	Manual
		[39]	Unstructured	Statistical	Arabic linguistics	Hybrid
	Islamic domain	[58], [13]	Semi-structured	Statistical	Agriculture	–
		[59], [18], [60]	Semi-structured	Linguistics	Wikipedia	–
		[24]	Semi-structured	Hybrid (Manual and Translation)	Food, nutrition and health	Manual
		[33]	Structured	Hybrid (Manual and Statistical)	Arabic linguistics domain	Manual
		[35]	Quran	Manual	Quran ontology	Manual
		[63]	Quran	Hybrid (Manual and Statistical)	Stories of prophets	–
		[36], [37]	Quran	Manual	Islamic knowledge	Manual
		[38]	Quran-Quias-Ijmaa	Manual	Solat (Prayer) ontology	Manual
		[40]	Hadith (Sahih AlBukhary)	Hybrid (Manual and Statistical)	Hadith ontology	–
		[16]	Hadith	A proposed system is suggested to adapt the hybrid approach		
Upper ontology	Arabic WordNet ontology	[48], [49], [50]	NA	NA	NA	NA
	Formal Arabic ontology	[47], [43]	NA	NA	NA	NA

such as BFO, Cyc, DOLCE, SUMO and others. The need to use these Ontologies for different languages including Arabic, lead us to a significant question that is: are the upper level Ontology concepts language dependent or independent? We studied the answer from the two perspectives.

Regarding the application of the language independency point of view for Arabic, using one of the previously mentioned upper Ontologies, requires the integration of some mid-level Ontology that is devoted to the Arabic culture [27]. Also, the translation process should be guaranteed. The application of the language-concept dependency principle, on the other hand, requires building an upper ontology that is specific to the Arabic language. However, there is only one trial to build such Ontology [40, 43], which is still under construction. As there is a challenge to find middle level Ontology for Arabic and the fact that there is no yet standard Arabic Upper Ontology, researchers used the AWN Ontology as an alternative. But, in addition to its practical limitations, mentioned at the following Sect. (4.1.1), it is built based on translation. This presents a big limitation when adopting the language-concept dependency point of view. As, each language has its specific linguistic environment and cultural context [15, 47]. Consequently, this makes it necessary to build the Arabic Ontology that takes into consideration the historical and cultural aspects of the language. In the following two Sects. (4.1.1, 4.1.2) we discuss the two Arabic Upper Ontologies.

4.1.1 Arabic WordNet Ontology

Authors in [48–50] initiated the Arabic WordNet (AWN) project to build a lexical resource for Modern Standard Arabic. AWN was built following the same methodology developed for Princeton WordNet [51] and Euro WordNet [52]. It was constructed by, first, encoding manually the most important concepts to create the core WordNet for Arabic. Then, maximizing the compatibility across other WordNets. A Mapping-based approach was followed to link the Arabic-English corresponding terms. AWN provides a formal semantic framework as it is mapped to SUMO [53] and its associated domain Ontologies. AWN, like all other WordNet projects, was initially constructed as a lexical database, then, it was interpreted and used as a lexical Ontology. Words were collected into sets of synonyms called *synsets* and a number of relations among these *synsets* were recorded. AWN consists of 11,270 synsets and 23,490 Arabic expressions (words & multi words). Interpreting AWN to be used as a lexical ontology is done by formalizing each synset as a concept class. This indeed has led to many practical limitations due to the huge number of concepts (synonym sets) that made it inappropriate for real world applications [54]. Moreover, mapping WordNet to an existing Ontology is a very challenging task [55].

4.1.2 Formal Arabic Ontology

Adopting the language-concept dependency principle point of view, authors in [43, 47] initiated a project to build the first Arabic upper level Ontology. The Ontology was built following the same design as AWN aiming to use it as an alternative. It was constructed following five steps. The first step was, Manual and semi-automatic extraction of Arabic concepts from specialized dictionaries. While, the second step was, reformulation of the concepts manually to strict ontological rules focusing on the intrinsic properties of concepts. Coming to the third step, the generated concepts were mapped automatically with the English WordNet using a smart Algorithm developed by the authors to inherit WordNet semantic relations. Followed by, the fourth step which was, cleaning the inherited semantic relations from WordNet using ONTOClean methodology. Finally, the fifth step was, linking the concepts and the relations with a semantic tree that contains all mother concepts of the Arabic language. This semantic tree was called Arabic Core Ontology; it was built to govern the correctness of the whole Arabic Ontology. It consisted of 10 levels and 400 concepts, and it was built based on DOLCE and SUMO. According to the authors, they have built a logically and semantically well-founded ontology. However, from our point of view, we think that following the same building approach of AWN may lead to the same practical limitations of it.

4.2 Domain Ontology

Domain Ontology, also known as (domain specific Ontology), represents concepts that belong to a specific domain of interest along with relationships interconnecting these concepts [56]. It reduces the conceptual and terminological confusion among users who share electronic documents and various kinds of information that belong to the same domain [40, 57]. Domain Ontology can use upper Ontology as a foundation and extend its concepts, accordingly taking advantage of the semantic richness of the extended concepts and logic that is built into upper Ontology [44]. In this section, we review, to the best of our knowledge, all the conducted trials to build Arabic domain specific Ontology. We categorize domain specific Ontology into two categories, general domains (4.2.1) and Islamic domains (4.2.2).

4.2.1 General Domains

In this section, we review the trials to build Arabic Ontology for general domains such as, computer, legal, linguistics, agriculture, and others. We categorize the trials according to the technique used in the Ontology learning process. In our research, we found out that there are four techniques mainly used in the Arabic Ontology learning process which are manual [14, 15, 34, 41, 42], statistical [13, 39, 58], linguistics [59], and hybrid [24, 33].

Manual Approach

In [41, 42] DEAR-ONTO, a derivational Arabic Meta-Ontology model, was presented. The authors built their hypothesis basing on the fact that Arabic is a "*highly derivational and inflectional language in which morphology plays a significant role*" [22]. The main purpose of that model was to structure the Arabic language into a set of equivalent classes using the derivations and their patterns. To populate the ontology, the authors used a list of selected Arabic verbs and their derivations. The Ontology used verbs as roots, and the derivations formed the equivalent classes. Each equivalent class was represented by a verb and contained all its derived words following derivation and inflection rules of the Arabic language. Then, each equivalent class was modeled as Ontology, and a Meta-Ontology representing the general structures of all those classes was presented. The authors suggested several applications of the model such as, Arabic language development, Arabic language understanding and Arabic morphology analysis.

The work was illustrated as a theoretical stage without practical implementation or evaluation. We found two contradictory opinions regarding that hypothesis. On one hand, authors of [15, 17] criticized it, and proved that, it is imprecise to build Ontology based on the roots. Since, "*85% of Arabic words are derived from trilateral roots"* [49] which definitely lead to, the existence of concepts with different meanings sharing the same root and consequently sharing the same class. Moreover, [15] added that there are words in Arabic that have no roots and the model did not handle those cases. On the other hand, authors of [26] supported the hypothesis, by suggesting using the approach in general domain corpora rather than specific domain. As, in general domain corpora there are more frequently similar terms sharing the same root. Moreover, they clarified that the main weakness in this approach was the over-generation of similarity links.

From our point of view, evaluating that hypothesis basically depends on the applications built upon the proposed Ontology model. As, in case of building applications that take advantage of the derivational nature of the Arabic language and exploit its derivation and inflection rules such as applications that are used to understand Arabic language and its rules, the model is valid. In fact, we found the suggested applications by the authors meet this case and as a result we found the hypothesis valid. On the other hand, in case of building applications that require knowledge to be classified and structured correctly such as information retrieval applications and the applications that require studying specific domain knowledge, the hypothesis is invalid. And the model is affected by the limitations mentioned by Al-Safadi [15] and Al-Zoghby [17].

In [14, 15] a "computer" domain Arabic Ontology was constructed. Both researches built the Ontology to use it as a basis for a semantic search engine. In order to enhance the semantic based search results and to solve the traditional search engines related problems, such as, low query precision.

The Ontology presented by Al-Safadi [15] consisted of 110 classes, 78 instances and 48 relations. It was constructed following three steps. In the first step, the Ontology classes were formed by gathering the most relevant concepts in the

computer technology domain from users. Then, those translating computer domain English Ontology classes into Arabic and using specialized domain dictionaries in addition to, using domain specific articles. While, in the second step, the instances and properties of the predefined classes were defined. Finally, in the third step, the Ontological relationships were manually defined. In this step the concepts were organized into a hierarchy associated via both taxonomic relations such as, (is-a, part-of and type-of) and non-taxonomic relations as, (produced by (تنتجه شركه), has logo (لها شعار) and use (يستعمل). To evaluate the ontology, the authors imposed an Arabic query on it. The Query was tested using protégé 3.4.4 SPARQL Query and the average precision rate of the experiment was 50%.

While, the Ontology presented in [14] was much simpler. It also was constructed using a set of main concepts in computer domain, such as (computer (الحاسبor العتاد)). Those concepts were associated via inheritance, association, and synonym relations. To evaluate the ontology, the authors performed two queries using a proposed semantic search engine that was built based on the constructed ontology, and "Google", the syntactic search engine. The results of the first query showed that, the semantic search engine returned fewer pages than the syntactic search engine. While, the results of the second query showed that, both search engines returned approximately the same number of pages. According to the authors, the semantic search results were more accurate and specific in both queries. Despite the fact that, the Ontology presented in both trials was initial and it didn't fully cover the computer domain, the trials have a positive aspect in paving the way to build Arabic Ontology based semantic search engine. That provides a better alternative to the keyword based syntactic search engines. While the negative aspects of both trials are, cost, human effort and the time consumed in the manual construction of the Ontology.

In another trial to improve the keyword based search results and to improve Arabic information retrieval in general, [34] presented a simple Ontology to be used as basis of query expansion process in the legal domain. The Ontology was constructed following a top-down strategy, according to the steps mentioned in [7]. Starting with the main concepts in the legal domain, the hierarchy of concepts was constructed. Then, relationships such as, (is-a, and instance-of) were assigned between those concepts. To populate the Ontology, the authors used UN[4] articles in Arabic and a set of selected newspapers articles in the legal domain. In order to improve the precision and recall of the query expansion, the authors associated each concept with its synonyms and derivative set that was selected according to its relevance to the legal domain. To evaluate the efficiency of the generated Ontology in the query expansion process, the authors compared the recall and precision results of an initial query and extended query performed using Arabic engine called Hahooa.[5] The initial query was formulated by a main concept in the legal domain. While the extended query, was formulated by the same concept and all its

[4]http://www.undp.org.

[5]http://www.hahooa.com/nav.php?ver=ar.

synonyms and derivatives after extending it using the generated Ontology. The results showed that, the extended query improved the recall from 115 to 135. while the precision was improved from 2 relevant results out of the first 10 results in case of the initial query, to 7 out of the first 10 results in case of the extended query. Despite the simplicity of the generated Ontology, the results provided by the authors, showed its efficiency in the query expansion process. The main weakness of that trial was the cost of the manual construction of the Ontology and according to the authors, its *"representativeness of the domain"*.

Statistical Approach

Using a statistical approach, authors in [39] constructed Ontology in Arabic Linguistics domain. The Ontology was constructed following a top-down strategy. A seed Ontology was first initialized manually using the general concepts of GOLD. Then, following a three-step process the concepts and relationships between them were extracted and the Ontology was updated. In the first step, a domain corpus was formed and preprocessed. It was formed by 57 documents from books, journal articles, and web documents in Arabic linguistics domain. The documents were selected, prepared manually (by deleting tables, diagrams and graphs), and transformed into plain text. Then a set of preprocessing steps such as normalization, deletion of stop words, and light stemming were performed to prepare corpus for the extraction of ontological elements. While in the second step, the domain concepts were extracted using two techniques; "repeated segments" and "co-occurrence". The "repeated segments" technique considered any term that denotes a concept to consist of four words maximum. The technique extracted all the repeated segments from the corpus after indexing all the words corresponding to their position. The extracted concepts were then filtered to eliminate the unwanted ones using filter of weights and cut filter. The filter of weights filtered concepts according to their total number of occurrences in the corpus compared to a predefined threshold. The cut filter, on the other hand, removed the segments containing certain words, such as verbs, named entities, and numbers. While, the "co-occurrence" technique extracted all the candidate concepts that occur together frequently and at the same time were extracted as repeated segments. Finally, in the third step, the ontological relationships were extracted and the Ontology was updated. The relationships were extracted using two approaches; linguistic markers and hierarchical relations. The linguistic markers approach, used the context between any two candidate concepts to extract elements that identify the relation between them, such as (is-a, and part of). The linguistic markers were organized into categories according to the type of the extracted relation. Hyponym relation category contained (is-a هم) هو هى and meronymy relation category contained (part of تنقسم الى – تتألف من– تتكون من). The hierarchal relations approach, on the other hand, was used as an alternative in case there were no linguistic markers. The approach was responsible for extracting only parent-child or (is-a) relation between two candidate concepts; the first one of them was considered a parent and the second

Table 3 Summarizes the ontology update process using the extracted concepts and relations

		Concept cases		
		Case 1 One concept of the pair was found among the seed ontology concepts and the other one was not	*Case 2* Both concepts of the pair were found among the seed ontology concepts and there was no relation between them	*Case 3* None of pair concepts were found among the seed ontology concepts
Relation extraction	Linguistic Markers Approach	The missed concept of the pair was defined as a new concept and was linked to the other concept with a relation defined by linguistic marker	A New relation between the concept pair was assigned from the linguistic marker	The process does nothing
	Hierarchal Relations Approach	The missed concept was defined as a new son-concept/ (father-concept) and was related to the other concept by a subsumption relation "is-a"	A new relation of subsumption "is-a" was assigned between the concept pair	The process does nothing

was the child. It used rules to ensure the existence of those candidate concepts together more frequently also, to ensure the probability of the occurrence of the first concept (parent) before the second (child) is higher than the reverse. After extracting concept pairs and the relationships between them the seed Ontology was updated as illustrated in Table 3. The authors provided results for only the first two steps; the preprocessing and the concept extraction. There was no illustration of the created ontology or the implementation. Therefore, the generate Ontology can't be evaluated.

[13, 58] Used a semi-automatically Ontology learning system to learn a taxonomical Ontology in agriculture domain. The system used a set of semi-structured HTML web documents and a set of seed concepts as input. The domain concepts and taxonomical relationships between them were extracted using two approaches. The first approach utilized the phrases of the HTML documents headings, while the second approach used the hierarchal structure of those headings to extract the taxonomical Ontology.

In the first approach, the Ontology was generated by searching all the headings phrases after extracting them, to find each concept that was considered children of any seed concept. That was done by extracting all the word sequences or what was called by the authors, the N-gram phrases that had one of the seed concepts as their headwords. The concepts extracted from headings were assigned to their parent concepts and the Ontology was formed. While in the second approach, the Ontology was generated by structuring the concepts in a hierarchy based on the

heading levels or the HTML structure of the documents. The seed concepts were located at the top level of that hierarchy then considering the concepts at the second level as children of the top level and so on. The two Ontologies generated by both approaches were filtered from fake concepts. Then, both were merged to set the right parent and the right level for each concept that was found in both Ontologies. Resulting in, adjusting the hierarchal structure of the final Ontology.

For evaluating the generated Ontology, the authors followed a Golden Standard Evaluation method that consisted of both lexical and taxonomic evaluation. The generated Ontology was compared to a subset of a golden standard Ontology in agriculture domain called AGROVOC. The best F-score results were 76.5% and 75.66% for lexical and taxonomic evaluation, respectively. The main limitation we found in this work was that the authors didn't clarify how the taxonomical relationships between any two concepts were identified. The main focus of both approaches was to extract and structure concepts which can never be true without defining or extracting the right relations between these concepts. As, in the first approach, the existence of two concepts in any N-gram phrase is not an evidence that there is a relation between them. While in the second approach, depending on the structure of the concepts based on the headings levels is not precise according to the nature of the web.

According to [26], the study didn't explain how to recognize the head of each N-gram, nor how to handle N-grams or how to deal with the syntactic ambiguities. However, the positive aspect of this research is the evaluation of the Ontology as it is the first work in Arabic to evaluate the generated Ontology using Ontology standard evaluation technique. This should encourage other Arabic projects to use such techniques.

Linguistic Approach

In [18, 59, 60], the two earlier publications of the same project, the authors proposed a linguistic-based approach to learn Ontology from Arabic Wikipedia. The approach relied on the semantic field theory, in which concepts were defined using their semantic relations. Applying that theory on Wikipedia, the authors considered each article's title as a concept and its semantic relations were extracted from the list of categories and infoboxes, following a bottom up approach. For each article, the infobox was extracted from the articles text. Each infobox was then parsed to extract (hasFeature), (isRelatedTo) and (hasCategory) relations. The (hasFeature) relation, defined articles features and their values, the (isRelatedTo) relation identified the related Wikipedia articles, and the (hasCategory) relation extracted article's categories. Then, the final Ontology was generated and written as an OWL file.

To evaluate the approach, the authors performed validation testing of the generated ontology and human judgment, from both crowd and domain experts' evaluation. The Ontology passed the validation against violations of OWL rules successfully. As for the human judgment experiment, the authors published an

online survey to evaluate a sample of the generated ontology extracted for "Saudi Arabia" article. The overall precision of the experiment was 56%. In the experts' validation experiment, on the other hand, two domain experts evaluated Ontology extracted from 24 Wikipedia articles in the geography domain. The individual precision for each expert was 83.82% and 79.41% respectively, while the average precision of the evaluation according to both experts was 82%. The average precision of the approach for the human judges in the two experiments was 65%.

In fact, we found that the system is very promising. It can be improved by increasing the number of concepts and the ontological relations by extracting them from the Wikipedia text not only from the articles or infoboxes. In fact, the authors suggested future enhancements for the system including this point.

Hybrid Approach

Following a hybrid approach of translation and Manual techniques, authors in [24] Presented an early stage integrated Ontology for food, nutrition, and health domains. It was mainly developed to be used in annotating Arabic textual web resources related to the three domains. At first, a simple Ontology for food and nutrition was built by translating food items, food groups, and nutrition names of the USDA[6] database into Arabic. In addition to inheriting the relationships between food items and nutrition values from the USDA.[6] Then, the health domain was added by defining four classes; diseases, part of the body, body biological function, and people status. To integrate the three domains, the authors created object properties between food, nutrition, and health. Object properties consisted of three types of relation: positive, negative and prevent. Prevent relation was used with disease class only. The authors suggested future enhancements including, the integration of the Ontology with international Ontologies, and the expansion of it with additional concepts and relations extracted from web documents related to food, health and nutrition domains.

Merging the manual and statistical techniques, authors in [33] Presented a prototype of linguistic Ontology that was founded on the Arabic Traditional Grammar (ATG). The Ontology was Extracted following two steps. It was first bootstrapped manually by extracting concepts from Arabic linguistic resources and relating them to the concepts in GOLD. Then, it was enriched by implementing an automatic extraction algorithm to extract new concepts from linguistics text. The text was preprocessed by performing segmentation, light stemming, and stop words elimination before applying the extraction algorithm. The algorithm was based on the repeated segments statistical approach. The newly extracted candidate concepts and their relations were proposed to an expert before being inserted in the ontology. In fact, the authors did not provide any clarification of that algorithm or the enrichment step in general. They only provided two conceptual graphs representing

[6] http://www.ars.usda.gov/Services/docs.htm?docid=8964.

the manually constructed Ontology with both the hierarchal and non-hierarchal relations between concepts.

Uncategorized

In this section, we review researches that don't fall in any of the previously mentioned categories.

In an attempt to implement the Arabic domain Ontology construction process, authors in [26] presented a system called ArabOnto. The system takes N corpora representing different domains as an input. The following five steps briefly summarize how the system works. In the first step, it starts by analyzing documents via performing morphological analysis, disambiguation, and POS tagging using MADA [61]. While in the second step, it generates syntactic trees of all Noun Phrases (NPs) using a syntactic parser developed by the authors. Coming to the third step, an algorithm for morpho-syntactic disambiguation and Domain Relevant Term (DRT) extraction is implemented. Followed by the fourth step, in which, conceptual networks that contain DRTs and their syntactic relations are generated. Finally, in the fifth step, a clustering algorithm was applied to group terms in each network. The groups obtained from different networks were managed in order to merge groups that have many common elements. The authors considered that the obtained structure represents the domain ontology, as terms sharing the same hyperonym (i.e., co-hyperonyms) were clustered.

Authors in [27] built Ontology for semantic based question answering system. The main idea behind that work was to integrate the lexical information extracted from Arabic WordNet (AWN) and the semantic, syntactic and lexical information extracted from Arabic VerbNet (AVN). In order to provide a better representation and to add a semantic dimension to the concepts of the generated Ontology, the authors made that combination between AWN and AVN. The Ontology was constructed following a two-phase process; briefly explained as follows. In the first phase, AWN was used to build the concepts hierarchy by transforming each AWN synset into a concept. Then each concept was assigned a lexicon that contained its lexical information and all the words that were members in its synset. Finally, concepts were categorized according to their type into two nodes; nouns and verbs. In the second phase, AVN was used to extract information related to each concept under verbs node. That was done by first, extracting all the frames related to a specific verb from AVN classes. Each verb had three frames; syntactic, semantic and constraints frames. Those frames were then transformed into sub conceptual Graphs and were related to the verb by syntaxOf, semanticOf, and constraintOf relations. Finally, those conceptual graphs were integrated in the Ontology verb nodes as situations of each verb. In fact, the authors presented in details how each frame was transformed into Conceptual Graph; we are not going to discuss it here since our main focus is on how the Ontology was generated. For more details, we refer to the main article. The authors measured the performance of the question answering system using two approaches; surface and semantic similarity based

approaches. The surface similarity based approach measured the similarity between keywords and structure of questions and their candidate passages. While the semantic similarity based approach, on the other hand, used the generated Ontology to build Conceptual Graphs of questions and their candidate passages. Then, it measured the semantic similarity score between those conceptual Graphs. The results showed that the Ontology based semantic approach generally improved the performance of the system. It increased the percentage of the answered questions in general and the correctly answered questions in particular. The percentage of the correctly answered questions increased from 7.39% out of 284 questions in case of surface similarity approach to 16.2% out of 284 questions in case of semantic-based approach. In fact, we found that this research is very promising as, to the best of our knowledge, it is the only research in Arabic that adopted the integration of syntactic and semantic information in order to build semantic based intelligent applications.

4.2.2 Islamic Domain

Islamic Ontology is very important for both Muslims and non-Muslims. As for the non-Muslims, it provides a semantic meaning in order to understand the Islamic Messages as described in Quran and Hadith [37]. As for Muslims, there are 1.7 billion Muslims [62] around the world, most of them do not speak Arabic as their native language. Therefore, it provides them a comprehensive meaning of Quran and Hadith. Arabic plays a significant role in the Islamic scholarship because it is the language of the Holy Quran, and Muslims daily prayers are performed in Arabic [37]. In this section we review trials to build Arabic Islamic domain Ontology. We categorize Islamic Ontology according to the main references that Ontology is built for; to Quran and Hadith.

Quran Ontology

From Al-Quran corpus, authors in [35] built Arabic lexical Ontology called Azhary. The Ontology grouped words into synsets and assigned number of relations between them following the same design as AWN. It contained 26,195 words organized in 13,328 synsets. The Ontology learning system was composed of three modules; word extraction, relation building and Ontology building. The word extraction module, built the seed to start the Ontology. It manually extracted seed words from the Quran corpus. While the relation building module, manually extracted the relations between words from Arabic dictionaries such as (the meaning dictionary (قاموس المعانى), Rich lexicon (معجم الغنى), and Mediator lexicon (المعجم الوسيط). The Ontology contained 7 types of relations; Synonym, antonym, hyponym, holonym, hypernym, meronym and association. The seed words and relationships between them were stored in a table in an excel file. Finally, the Ontology building module, converted the table of words and relations into Ontology. To evaluate the ontology, the authors presented a comparison between Azhary

and AWN. According to the authors, Azhary showed a better response time, contained more words, and recorded more word relations. However, we find this comparison is illogical as both Ontologies are different. AWN is a lexical ontology for the whole Arabic language, while *Azhary* was constructed from Quran as the only source. Also, the manual construction of such Ontology takes a tremendous time, cost and effort. In addition, following the same design approach of AWN may lead to the same practical limitation of AWN as we mentioned earlier in Sect. 4.1 (Upper Ontology Section).

[36–38] started another project to build an Islamic Knowledge Ontology. In [36, 37], as a first stage in the project, authors attempt to build Islamic Ontology from Quran text. However, this trial faced some challenges regarding the used approach to extract ontological elements from text. Additionally, the structure of the Quran, which requires another source of Islamic knowledge in order to build a more complete ontology. In [38], as a forward step in the way of Islamic Ontology construction, authors enhanced the Ontology learning approach and also used another knowledge source to build the Ontology. In this trial, Prayers (Solat صلاه) Ontology was presented. It was constructed semi-automatically from Quran as the primary resource, and from Qiyas-analogy and Ijma-consensus as a secondary resource. The Ontology focused on two types of Solat; Obligatory and Sunnah. It was constructed by using the Quran indexes as upper layer TBox for the Ontology. Listing all the important Solat terms and the different types of Solat such as obligatory (Fardhu فرض) and Sunnah (سُنه) to develop the hierarchal taxonomy. In the formation of this taxonomy, the few top-level concepts were associated to the middle level, then all the other classes that can be expanded from Solat were generated. The generated ontology had 48 concepts, 51 relationship properties, and 282 instances. Authors provided a visualization of the generated ontology and we find that it completely covers the two types of solat. We think the approach has proved its efficiency when applied to this small area. The only negative aspect of the research is the cost, time, and effort it requires due to human intervention.

In [63], an attempt to construct Ontology for the holy Quran Chapters (*Surah*) related to stories of the prophets was presented. Al-Quran corpus was used as the knowledge source. The authors followed the same approach as [40] that we illustrated in details in the following section (section "Hadith Ontology").

Hadith Ontology

Authors in [16] suggested a framework for semi-automatic Ontology construction from the Hadith corpus. The presented framework consisted of four modules; documents preprocessing, concept extraction, relation extraction and Ontology edition. It was based on NLP, statistical, and data mining techniques to extract concepts and semantic relations. However, the implementation of this system is still under progress.

Using association rules algorithm, authors in [40] presented a trial to build Ontology for prophetic traditions (Hadith حديث). The trial was considered a first step

in the way of building a fully functional Hadith Ontology. The Ontology was constructed from Sahih Al-Bukhari[7] (صحيح البخارى) book as the only knowledge source from the entire hadith collection. The Ontology consisted of two parts; Hadith Metadata Ontology, and Hadith Semantic Ontology. The Hadith Metadata Ontology was built from the structural taxonomy of Sahih Al-bukhary book. By, creating a sub-class-of relationship between a concept and a sub concept based on the structure of the book. For example, each chapter name was considered a concept and all the sections names related to that chapter were considered sub concepts. Hadith semantic Ontology, on the other hand, was built from concepts and semantic relationships extracted from the texts of Hadiths. Concepts were extracted according to the following steps. First, key phrases were extracted using KP-miner[8] after applying preprocessing and tagging operation to the text. Then, key phrases were stemmed to transform all word derivatives to their roots in order to extract the different forms of the same root. Finally, the frequencies of all words that shared the same root were calculated and the words with higher frequencies were defined as concepts.

Relationships were extracted according to the following steps. First, the authors specified certain types of relations such as "kind-of", "Part-of", and "Synonym-of" and assigned tags to the words in the text that represent any type of those relations. Then, Apriori algorithm was applied to extract all the association rules between concepts and the predefined relationships. After that, the rules were selected based on the satisfaction of a condition that; in any rule the higher concept must occur after the relationship. To clarify, when C1 is part of C2; C2 must be the higher concept in the pair. Finally, the confidence of the selected rules was calculated and the rules with higher confidence were extracted. In fact, the authors didn't clarify practically how the higher concepts were identified in case that those concepts weren't included in the first part of the Ontology (Hadith metadata Ontology). The authors presented an illustrative example to extract part of the relationship between concepts from four Hadiths. The authors mentioned that they used OWL to represent the Ontology; they did not provide any representation not even visually of the Ontology.

5 Conclusion

In this review, we studied two highly related topics in Arabic; Arabic Semantic Relation Extraction and Arabic Ontology Learning. Arabic Semantic Relation Extraction is one of the most significant however, least tackled tasks in the Arabic Ontology learning process. As we proceeded in our study, we noticed a gap in this area as we found that most researches considering the extraction of semantic

[7] https://fr.wikipedia.org/wiki/Sahih_al-Bukhari.

[8] http://www.claes.sci.eg/coe_wm/kpminer/.

relationships fall into two categories. The first category, considers the extraction of semantic relations between pairs of NEs for the purpose of using them in several NLP applications. The second category, on the other hand, considers the extraction of semantic relations between Ontological concepts for the purpose of building Arabic Ontologies. We suggest that both categories should be integrated as the first category can be used in the enrichment and population of Ontologies in instance level, while, the second category can be used to build Arabic Ontologies. This integration will certainly enhance the Arabic Ontology learning. Regarding the extraction of Semantic relations between NEs, we found that it is very challenging to fully automate this process due to the nature of Arabic Language and the many odd cases that can only be handled manually. Additionally, research in this area is very little and most of the work was done by the same group of authors. The effort of the authors is well appreciated, but still this area of research in Arabic needs more work in order to be able to compare techniques conducted by other researchers and conducted from different points of view.

Semantic Relation Extraction between Ontological concepts, on the other hand, was discussed as a part of the whole Ontology learning process. We reached a conclusion that, most works extracted it manually or used a set of predefined relations. The main focus was to build the full Ontology neglecting the semantic relation extraction phase.

Arabic Ontology learning is a very critical topic based on which the future of the Arabic semantic web will be determined. Very little trials were directed toward building Arabic Ontologies and these trials are too little to keep up with the significance of the Arabic language. Arabic Ontology learning is facing several obstacles starting from the nature of the Arabic language that makes it very challenging to deal with the Arabic text and the lack of the Arabic linguistic resources and tools. One more obstacle facing the Arabic Ontology construction is the fact that there is no standard upper-level Ontology to work as a foundation to build other Ontologies which causes lack of coherence among the Arabic Ontologies. Regarding the reviewed trials to build domain specific Ontologies, most of these trials are constructed either fully manually or partially manually. Which lead to the limitations of manual techniques represented in time, cost and the simplicity of the generated Ontology. The study also showed an obvious lack in the Islamic Ontologies and the majority of the trials are directed to build a simple domain Ontology. Ontology evaluation and validation approximately not tackled in Arabic researches despite its significance in reflecting the performance of applications. Only few researches performed evaluation of their generated Ontologies. It is necessary for the future development to work on overcoming these limitations and increasing the research work in Arabic Ontology Learning. This subsequently will lead to enhancing the Arabic semantic web.

References

1. Berners-Lee, T.: The semantic web. a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Sci. Am. **284**(5), 1–5 (2001)
2. Gruber, T.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum. Comput. Stud. **43**(5), 907–928 (1995)
3. Studer, R.: Knowledge engineering: principles and methods. Data Knowl. Eng. **25**(1), 161–197 (1998)
4. Maedche, A.: Ontology learning for the semantic web. Springer Science & Business Media (2002)
5. LEHMANN, J.: An Introduction to Ontology Learning
6. Hazman, M.: A survey of ontology learning approaches. Database **7**(6) (2011)
7. Noy, N.: Ontology development 101: a guide to creating your first ontology (2001)
8. Barforush, A.: Ontology learning: revisted. J. Web Eng. **11**(4), 269–289 (2012)
9. Auger, A.: Pattern-based approaches to semantic relation extraction: a state-of-the-art. Terminology, 1–19 (2008)
10. Wikipedia. Arabic language—wikipedia, the free encyclopedia (2015) [Online; accessed November-2015]
11. Lahbib, W.: A hybrid approach for Arabic semantic relation extraction, pp. 315–320 (2013)
12. Al Zamil, M.: Automatic extraction of ontological relations from Arabic text. J. King Saud Univ. Comput. Inf. Sci. 462–472 (2014)
13. Hazman, M.: Ontology learning from domain specific web documents. Int. J. Metadata Semant. Ontol. **4**(1), 24–33 (2009)
14. Moawad, I.: Ontology-based architecture for an Arabic semantic search engine. In: The Tenth Conference on Language Engineering Organized by Egyptian Society of Language Engineering (ESOLEC'2010) (2010)
15. Al-Safadi, L.: Developing ontology for Arabic blogs retrieval. Int. J. Comput. Appl. **19**(4), 40–45 (2011)
16. Al Arfaj, A.: Towards ontology construction from Arabic texts-a proposed framework. In: IEEE International Conference on Computer and Information Technology (CIT) (2014)
17. Al-Zoghby, A.: Arabic semantic web applications–a survey. J. Em. Technol. Web Intel. **5**(1), 52–69 (2013)
18. Al-Rajebah, N.: Extracting ontologies from Arabic Wikipedia: a linguistic approach. Arabian J. Sci. Eng. **39**(4), 2749–2771 (2014)
19. Wikipedia. List_of_languages_by_total_number_of_speakers, the free encyclopedia (2015) [Online; accessed November-2015]
20. Elkateb, S.: Arabic WordNet and the challenges of Arabic. In: Proceedings of Arabic NLP/MT Conference, London, UK (2006)
21. Al-Khalifa, H.: The Arabic language and the semantic web: challenges and opportunities. In: The 1st International Symposium on Computer and Arabic Language (2007)
22. Attia, M.: Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation. Dissertation. University of Manchester (2008)
23. Farghaly, A.: Arabic natural language processing: challenges and solutions. ACM Trans. Asian Lang. Inf. Process. (TALIP) **8**(4), 14 (2009)
24. Albukhitan, S.: Automatic ontology-based annotation of food, nutrition and health Arabic web content. Procedia Comput. Sci. **19**, 461–469 (2013)
25. Soudani, N.: Toward an Arabic ontology for Arabic word sense disambiguation based on normalized dictionaries. In: On the Move to Meaningful Internet Systems: OTM 2014 Workshops. Springer, Berlin, Heidelberg (2014)
26. Bounhas, I.: ArabOnto: experimenting a new distributional approach for building Arabic ontological resources. Int. J. Metadata Semant. Ontol. **6**(2), 8195 (2011)
27. Abouenour, L.: Construction of an ontology for intelligent Arabic QA systems leveraging the conceptual graphs representation. J. Intel. Fuzzy Syst. **27**(6), 2869–2881 (2014)

28. Hamadou, A.: Multilingual extraction of functional relations between Arabic named entities using NooJ platform. In: Nooj 2010 International Conference and Workshop (2010)
29. Ines Boujelben, S.: Rules based approach for semantic relation extraction between Arabic named entities. In: NooJ (2012)
30. Boujelben, I.: Enhancing machine learning results for semantic relation extraction. In: Natural Language Processing and Information Systems, 337–342 (2013)
31. Boujelben, I.: A hybrid method for extracting relations between Arabic named entities. J. King Saud Univ. Comput. Inf. Sci. 425–440 (2014)
32. Boujelben Ines, B.: Genetic algorithm for extracting relations between named entities. In: LTC, pp. 484–488 (2013)
33. Aliane, H.: Al-Khalil: the Arabic linguistic ontology project. In: LREC (2010)
34. Zaidi, S.: A cross-language information retrieval based on an Arabic ontology in the legal domain. In: Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05) (2005)
35. Ishkewy, H.: Azhary: an Arabic lexical ontology (2014). arXiv:1411.1999
36. Saad, S.: Islamic knowledge ontology creation. In: International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009, IEEE (2009)
37. Saad, S.: Towards context-sensitive domain of islamic knowledge ontology extraction. Int. J. Infon. (IJI) **3**(1), 197–206 (2010)
38. Saad, S.: A process for building domain ontology: an experience in developing Solat ontology. In: International Conference on Electrical Engineering and Informatics (ICEEI), 2011. IEEE (2011)
39. Mazari, A.: Automatic construction of ontology from Arabic texts. In: ICWIT (2012)
40. Harrag, F.: Ontology extraction approach for prophetic narration (Hadith) using association rules. Int. J. Islamic Appl. Comput. Sci. Technol. **1**(2), 48–57 (2013)
41. Belkredim, F.: DEAR-ONTO: a derivational Arabic ontology based on verbs. Int. J. Comput. Process. Lang. **21**(03), 279–291 (2008)
42. Belkredim, F.: An ontology based formalism for the Arabic language using verbs and their derivatives. Commun. IBIMA **11**, 44–52 (2009)
43. Jarrar, M.: Building a formal Arabic ontology (invited paper). In: Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League, Tunis (2011)
44. Semy, S.: Toward the use of an upper ontology for US government and US military domains: an evaluation. In: No. MTR-04B0000063. MITRE CORP BEDFORD, MA (2004)
45. Mizoguchi, R.: Part 1: introduction to ontological engineering. New Gener. Comput. **21**(4), 365–384 (2003)
46. Wikipedia. Upper ontology—wikipedia, November (2015)
47. Jarrar, M.: The Arabic ontology. In: Lecture Notes, Knowledge Engineering Course (SCOM7348), Birzeit University, Palestine (2010)
48. Black, W.: Introducing the Arabic WordNet project. In: Proceedings of the Third International WordNet Conference (2006)
49. Elkateb, S.: Building a WordNet for Arabic. In: Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006) (2006)
50. Rodríguez, H.: Arabic wordnet: current state and future extensions. In: Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary (2008)
51. Fellbaum (ed.): WordNet—An Electronic Lexical Database. The MIT Press, Cambridge (1998)
52. Vossen, P.: EuroWordNet, A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, The Netherlands (1999)
53. Pease, A.: The suggested upper merged ontology: a large ontology for the semantic web and its applications. In: Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, vol. 28 (2002)
54. Wikipedia. Lexical Ontology—wikipedia, the free encyclopedia (2015) [Online; accessed November-2015]

55. Wikipedia. Arabic WordNet Ontology—wikipedia, the free encyclopedia 2015 [Online; accessed November-2015]
56. Wikipedia. Domain Ontology—wikipedia, the free encyclopedia 2015 [Online; accessed November-2015]
57. Navigli, R.: Learning domain ontologies from document warehouses and dedicated web sites. Computational Linguistics, 151–179 (2004)
58. Hazman, M.: Ontology learning from textual web documents. In: 6th International Conference on Informatics and Systems, NLP (2008)
59. Al-Rajebah, N.: Building ontological models from Arabic Wikipedia: a proposed hybrid approach. In: Proceedings of the 12th International Conference on Information Integration and Web-based Applications and Services. ACM (2010)
60. Al-Rajebah, N.: Exploiting Arabic Wikipedia for automatic ontology generation: a proposed approach. In: 2011 International Conference on Semantic Technology and Information Retrieval (STAIR), IEEE (2011)
61. Habash, N., Rambow, O., Roth, R.: MADA + TOKAN: a toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In: Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), vol. 41, p. 62. Cairo, Egypt, April (2009)
62. Wikipedia. Islam_by_country–wikipedia, the free encyclopedia (2015) [Online; accessed November-2015]
63. Harrag, F.: Using association rules for ontology extraction from a Quran corpus