An Overview of Ontology Learning Process from Arabic Text

Mariam A. Muhammed*[1], Nesrine A. Azim*[2], Mervat H. Gheith**[3]

**Department of Information Systems and Technology, Faculty of Graduate Studies for Statistical Research (FGSSR), Cairo University, Egypt.*
[1]eng.maryamadel@yahoo.com
[2]nesrinealiazim79@hotmail.com
***Department of Computer Science, FGSSR, Cairo University, Egypt.*
[3]mervat_gheith@yahoo.com

Abstract: ***Ontology learning plays an important role in many fields especially in the Semantic Web. The success of the Semantic Web depends on the quality of its ontologies. There is a lot of research work interested in Ontology Learning for Arabic texts. Most of these works focused on three main issues: extracting the terms, extracting the semantic relations, and building the ontology from the Arabic text. In this paper, first, we present the Arabic challenges that were reasons for developing few Arabic Ontology Learning systems. Second, we make a research comparison based on the techniques used and their results. Third, we pointed out the limitations and comments of research works on Arabic Ontology Learning. Finally, we concluded the paper and outlined our future research direction in this area.***

Keywords: ***Arabic Ontology Learning, Term Extraction, Semantic Relation Extraction, Named Entity Recognition.***

1 INTRODUCTION

The World Wide Web is a huge repository of unstructured data. These data are difficult to process by humans and difficult to understand by machines. Tim-Berners Lee invented the Semantic Web to handle the problems related with the huge data and the unstructured format [1]. The ontology allows the semantic web to achieve its aim where the data can be represented in a way that enables machines to understand its meaning. Also, ontology allows data to be shared and reused [2].

Gruber defined the ontology as "a description of the concepts and relationships" [3]. Despite the ontology importance in the semantic web, it is also very important in more fields such as Data Integration and Interoperability, Machine Translation, and Information Retrieval. Concerning Data Integration, the ontology can be used as a semantic reference to several information systems. While Machine Translation task is by finding the exact mapping of concepts across languages. But in the Information Retrieval task, it is used to enrich queries and improve the quality of the results, i.e. meaningful search rather than string-matching search. Due to its importance for the semantic web and other fields, the process of building ontologies is important. The success of the Semantic Web depends on the quality of its underlying ontologies. Ontologies can be built either manually, semi-automatically or automatically. Building the ontology manually needs lots of efforts, time-consuming, and it has error-prone [4]. To overcome the previously mentioned limitations, several researchers tried to build ontologies by using semi-automatic or automatic methods. The process of building the Ontology semi-automatically or automatically is called "Ontology Learning" [5]. Extracting semantic relations is one of the most important phases in the Ontology learning process.

In this paper we overview the research that works on Ontology Learning for Arabic text and their results, we found that the research focused on Extracting the terms, Extracting the semantic relations and building the Ontology automatically.

The outline of this paper is as follows: background about Ontology and Ontology Learning is presented in section 2. The researches on the Term Extraction are reviewed in section 3 and the semantic relation extraction is reviewed in section 4. In section 5 we overviewed approaches and techniques for building the ontology automatically for Arabic text. We explained our comments and limitations of previous work in section 6. Finally, the conclusion and future work of this paper are in section 7.

2 BACKGROUND

In this section, Ontology and Ontology Learning will be discussed.

A. Ontology

Ontology is a structure that describes a set of terms used in a specific domain and their relationships. An ontology consists of individuals, classes and subclasses, attributes, and relations. The individuals are instances of the concepts or objects such as people, animals, and plants as well as abstract individuals such as numbers and words, the ontology may have no individuals. Classes are sets of objects described by a set of attributes. Classes may classify individuals with the help of these attributes. Some examples of classes are Person, car, Thing, etc. Attributes are properties or characteristics that classes can have. For example, (name, age, and height) are properties of a person's class or object. Relationships are links between objects that detect how objects are related to other objects such as (part-of, has-a, is-a, etc.) [6]. An example of the ontology is shown in Fig. 1. This example illustrates the components of the ontology as following:

- "*Person*" is a superclass;
- "*Student*" and "*Professor*" are subclasses of "*Person*" class;
- Also, "*Student*" and "*Professor*" are hyponyms of "*Person*" (their hypernym) where each of them has (is a) hyponym of "*Person*".
- "*E-mail*" and "*Name*" are properties/attributes of the "*Person*" class;
- "*Student_no.*" and "*Research field*" are also properties/attributes of "*Student*" and "*Professor*" classes respectively;
- "*Is-a*" is a relation where (a student *is a* person) and (the professor also *is a* person).
- "*AI*", "*Security*" and "*Big Data*" are individuals of the "*Research field*" class.

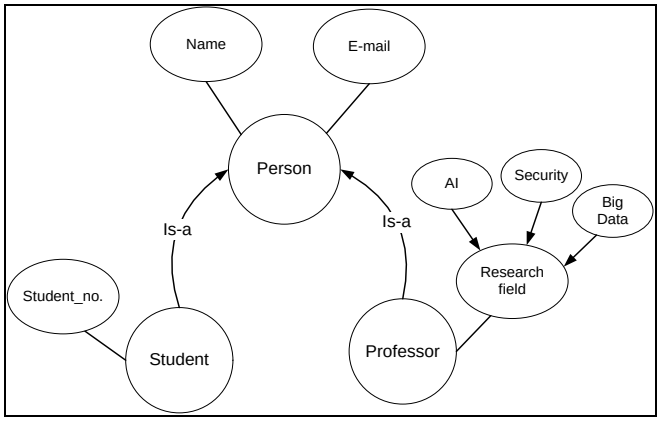

Figure 1: An example of Ontology

Ontologies can be built either manually, semi-automatically or automatically. The main steps of the ontology construction are: (1) build glossary of terms such as concept, instance, relation, and attribute, (2) build the taxonomic relations such as (Is-a hierarchy), (3) build diagrams for non-taxonomic relations, and (4) build concept dictionary [7] as shown in Fig. 2.

Despite the difficulty of building the ontology manually, however, there are a lot of tools that have been developed such as Apollo, OntoStudio, Protégé, Swoop, and TopBraid Composer. (Kapoor & Sharma) in 2010 presented a comparative study between these tools which showed that the "Protégé" is the best tool for building the ontology manually [8].

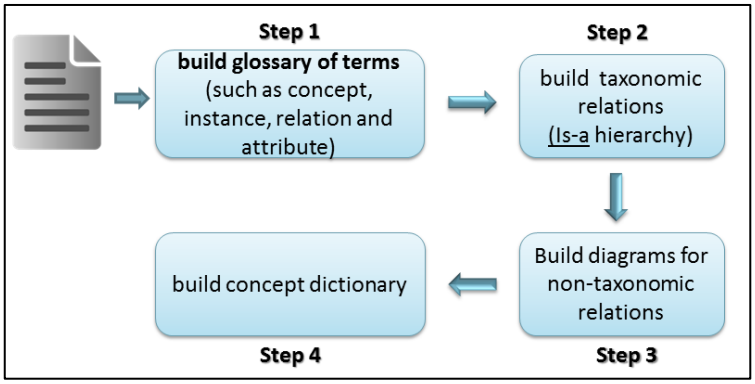

Figure 2: The main steps of the ontology construction process

B. Ontology Learning

Ontology learning (OL) can be defined as the set of techniques and methods from (different fields such as machine learning and natural-language processing [9]) used for building ontology automatically or semi-automatically from a given text corpus in which ontological elements such as concepts and relations are extracted automatically from different resources [4].

Because of ontologies importance in the semantic web, many research developed several ontology learning systems and approaches in different languages such as OntoLearn, Alvis, Text2Onto, and SPRAT [10]. Such systems aren't well for syntactically ambiguous languages such as Arabic compared to English [11]; where several challenges are facing the process of Ontology learning with Arabic text. Such challenges are the absence of capitalization, absence of diacritics, complicated morphology and the lack of resources. Specific terms in Latin languages like English begin with capital letters such as proper names e.g. ("Ahmed", "Mohammed") and abbreviations e.g. ("ACM", "IBM"). But this feature doesn't exist in the Arabic language because the Arabic language can't support the capitalization. Absence of this feature affects the knowledge extraction task [12]. Arabic text also contains the diacritics that "affect the phonetic representation and give a different meaning to the same lexical form" [12]. It also leads to ambiguity because different diacritics represent different meanings. Finding many different patterns for the one Arabic word is one of the characteristics of the Arabic language where each word can consist of one or more prefixes, a stem or root, and one or more suffixes in different combinations, that lead to complicated morphology. The lack of Arabic linguistic resources and tools represents another problem to the process of the semantic relation extraction and Ontology learning [13]. These challenges also affect knowledge extraction from Arabic text. As a result, the Arabic language suffers a lack of Ontologies and semantic web applications [14] compared with other languages such as (English, Chines).

The primary steps in buildings the ontology automatically is extracting both the terms and semantic relations from the text as shown in the following section.

3 TERM EXTRACTION

Term extraction extracts the relevant phrases and terms from the text for a specific domain by applying information extraction (IE) methods to extract terms. A subtask of the information extraction (IE) is Named Entity Recognition (NER) that is used to recognize the proper names in the text such as "Person", "Location", and "Organization". There are three main approaches used in the NER: Rule-Based NER, Machine Learning Based NER, and Hybrid Based NER. The rule-based approach depends on the linguistic rules of the language. Some researchers used this approach for extracting the NER [16]. The machine learning approach depends on the features of the named entity classes in a large training corpus. It overcomes some of the Rule-based challenges. The hybrid-based approach combines the two previously mentioned approaches to overcome some of their drawbacks.

In the following, we review some of the works on these approaches.

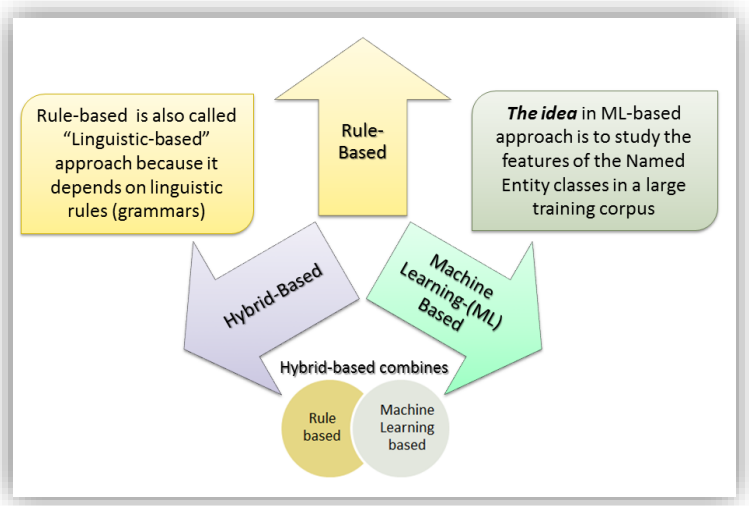

Figure 3 : The main approaches for Named Entity Recognition Task

A. Rule-Based NER

Asharef et al. (2012) presented a rule-based approach to extract Arabic named entities from crime documents [17]. To identify and classify named entities in Arabic crime text, several rules and patterns are applied. The approach includes three modules: (1) pre-processing module that contains these processes: sentence splitting, tokenization, and POS tagging; (2) second module is identifying the named entity and detecting of the tokens boundaries that belong to a named entity; and (3) Final module is the classification of the terms by using set of grammatical rules and patterns and gazetteer. The result showed that this approach is effective and the performance is satisfactory compared with the previous tasks on the crime domain. The drawbacks of this research are needed to increase the tag set and the NEs labels and it is domain-dependent. Another research was by Btoush et al. (2016) who used the rule-based techniques to build a tool for the Named Entity Recognition for the Arabic Language [18]. The Named-Entity detector has applied rules on the text and has given the correct Labels for each word; the labels are for three named entities: Person, Location and Organization. The name entity detector was tested on a file that contains 490 words and successfully tagged 480 words.

The researches that used a rule-based approach play an important role in the process of named entities extraction. But this approach is domain-dependent so, the researches achieved better results in specific domains only; so, if the same approach applied to another domain, the results will be different. Another problem of this approach requires an expensive manual effort and it is time-consuming; and if it is used with languages that have complex morphology such as Arabic, the problem increases.

B. Machine Learning-Based NER

Alsayadi & ElKorany (2016) presented a new model which depends on the machine learning approach [19]. This model aims to combine several linguistic features and to utilize syntactic dependencies to infer semantic relations between three named entities: person, organization and location. The proposed model helped overcome some of the orthographic and morphological problems of the Arabic language. They used the conditional random field as a machine learning classifier for recognizing the named entities. Experimental results show that this approach can achieve good performance where its F-measure was 87.86% for ANERCorp[1] corpus, but they needed to test other ML algorithms to enhance the performance.

There are some researchers who used the neural network algorithms for the named entity recognition task such as [20]. Gridach (2016) proposed a novel neural network architecture [20] by using a combination of bidirectional Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) for Dialectal Arabic

[1] Available for download from http://users.dsic.upv.es/ybenajiba/

and Modern Standard Arabic texts. Experimental results showed that the proposed model achieves state-of-the-art performance on publicly available benchmark for Arabic NER for social media.

Deep learning has performed significantly better than other approaches for different Natural Language Processing tasks including NER. Helwe & Elbassuoni (2017) presented a new approach called "deep co-learning" to detect and classify named entities in any Arabic text [21]. The proposed approach used a small amount of labeled data to overcome the problem of lacking Arabic resources. The approach is based on a semi-supervised learning algorithm known as co-training. They developed a Wikipedia article classifier using an LSTM deep neural network to generate a semi-labeled dataset for the Arabic NER task which outperformed all other compared approaches.

C. Hybrid Based NER

Abdallah et al. (2012) presented an approach for integrating the rule-based approach with the Machine learning-based approach for Arabic named entity recognition [22]. The authors focused on only three named entities (person name, location, and organization). For training and testing the proposed method, they used two annotated corpora: the ACE 2003[2] Multilingual training set and the ANERcorp corpus. Steps of building their rule-based system: (1) performing the recognition based on a dictionary lookup that containing lists of known named entities, and (2) using a parser, based on a set of grammar rules (represented as regular expressions). Steps of building the proposed integrated approach: (1) using the Stanford POS Tagger to compute some of the general features such as word category and affixation that are defined as machine learning features; (2) Complementing the rule-based features with the other extracted features; and (3) feeding all combining features to a decision tree classifier. The results proved that the proposed hybrid approach is better than the pure rule-based system. Oudah &Shaalan (2017) proposed another hybrid system that integrates both rule-based and machine learning-based NER approaches [23]. Their proposed hybrid NER system is considered state-of-the-art in Arabic NER according to its performance on standard evaluation datasets. They used ACE 2004[3] Newswire standard dataset for extracting new rules for the person, location and organization name recognition. They formulate each new rule based on two feature groups, (1) Gazetteers of each type of named entities and (2) Part-of-Speech tags. The results show that the proposed hybrid approach overcame the drawback of rule-based NER systems and it could improve the performance.

In some cases, the depending only on rule-based features doesn't improve the performance; and the depending only on machine learning-based features doesn't improve performance. But when integrating the features of rule-based with Machine learning classifiers, in this case, the performance can be improved.

TABLE I

SUMMARY OF WORKS ON TERM EXTRACTION FROM ARABIC TEXT

Reference	Technique
Asharef et al. (2012)	Rule-Based
Btoush et al. (2016)	Rule-Based
Alsayadi & ElKorany (2016)	Machine Learning Based
Gridach (2016)	Machine Learning-Based
Helwe & Elbassuoni (2017)	Machine Learning-Based
Abdallah et al. (2012)	Hybrid Based
Oudah &Shaalan (2017)	Hybrid Based

[2] Available to BUID under License from https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

[3] (ACE) 2004 Newswire (NW): the NW data from the Arabic training dataset for the Automatic Content Extraction (ACE) evaluation conducted in 2004, which has been created by Linguistic Data Consortium (LDC)

4 SEMANTIC RELATION EXTRACTION

Extracting the semantic relations from the text is an important step when building the ontology. Several of the research proposed different methods for semantic relation extraction.

Hearst, (1992) is an early researcher who worked on the semantic relation extraction and his proposed algorithm considers the basis for many similar works [24]. He proposed an algorithm that depends on the statistical-approach for the automatic extraction of the lexical relations like (is-a, kind-of or such as) by building lexical patterns of knowledge. This algorithm is considered a low-cost approach for the automatic extraction of semantic lexical relations from text. However, it is inefficient in Arabic text that leads some researchers such as [11] to deal with the Arabic language to overcome such drawbacks.

Al Zamil et al., (2014) proposed a technique that is implemented as an enhanced version of Hearst's algorithm. The steps of this technique are as follows: (1) analyzing Arabic text using lexical-semantic patterns of the Arabic language according to a set of features such as POS tag features; (2) building lexical syntactic patterns of Arabic text by enhancing the algorithm of Hearst; (3) the third phase is to avoid having redundant patterns; and (4) final phase is filtering and aggregation of the pattern. The results of this research show that the technique can enhance extracting ontological relations from Arabic text and they show also that the performance between three different Arabic datasets, Holy Qur'an "Classical Arabic", newspapers "Modern Standard Arabic MSA", and social blogs "unstructured Arabic texts" is not systematic. The Blogs dataset has the lowest performance and the Newspapers dataset for the MSA has the highest performance compared with other datasets. The reason for this is the existence of a few classification errors that affect the performance of the proposed techniques. Table 2 shows the results of a comparison between the original Hearst's algorithm on Arabic texts from different datasets with the Al-Zamil technique on the same datasets which shows that the Al-Zamil technique achieved the highest performance.

TABLE II

PERFORMANCE COMPARISON BETWEEN AL-ZAMIL TECHNIQUE AND HEARST'S ALGORITHM [11]

Average measurements	Dataset	Precision (%)	Recall (%)	*F*-Measure (%)
Original Hearst's Algorithm (Arabic Text)	Qur'an	46.74	50.32	48.48
	Newspaper	51.23	61.35	55.84
	Blogs	47.43	53.45	50.26
Enhanced Version of Hearst's Algorithm (Arabic Text)	Qur'an	76.28	82.94	79.47
	Newspaper	89.77	84.49	87.05
	Blogs	69.66	74.70	72.09

5 AUTOMATIC ONTOLOGY CONSTRUCTION

Building the ontology automatically can be by using different techniques: NLP techniques like tokenization, stemming and splitting; Machine Learning techniques and Deep learning techniques. There are several researches proposed systems and approaches presented to build the ontology automatically for different languages such as English and Arabic languages. The following are some previous works of such systems for the Arabic language:

Al-Rajebah, & Al-Khalifa (2014) proposed a system that considered each article's title as a concept and extracted its semantic relations from infoboxes and the list of categories contained in each Wikipedia article [25]. The system consisted of three main phases, (1) filtration, (2) extraction and (3) ontology generation. For each article, the infobox was extracted from the text of the article. Each infobox was then parsed to extract (hasFeature), (isRelatedTo) and (hasCategory) relations. The (hasFeature) relation, defined articles features, and their values, the (isRelatedTo) relation identify the related Wikipedia articles, and the (hasCategory) relation extracted article's categories. Then, the final Ontology was generated and written as an OWL file. The evaluation result showed that the average precision of the system equals 65% due to the presence of more duplication in the concepts and relations.

Harrag et al. (2014) proposed a new representation tool for the Quran [26]. A linguistic pattern-based approach was exploited to extract specific concepts from the Quran, while the conceptual relations were found based on association rules technique. The authors proposed the tool and predicted that the proposed Quran ontology will offer a new and powerful representation of Quran knowledge, and the association rules will help to represent the relations between all classes of connected concepts in the Quran ontology.

Al-Arfaj and Al-Salman, (2014) proposed a framework for Arabic ontology construction based on Hadith texts (sayings of Prophet Mohammed) [14]. It consists of the following phases: (1) pre-processing of the corpus, (2) extracting the concepts, (3) extracting the relations between concepts, and finally (4) building the ontology; as it is shown in Fig. 4. The authors discussed the challenges of constructing ontology from Arabic texts and the solutions for each but there aren't results to show.

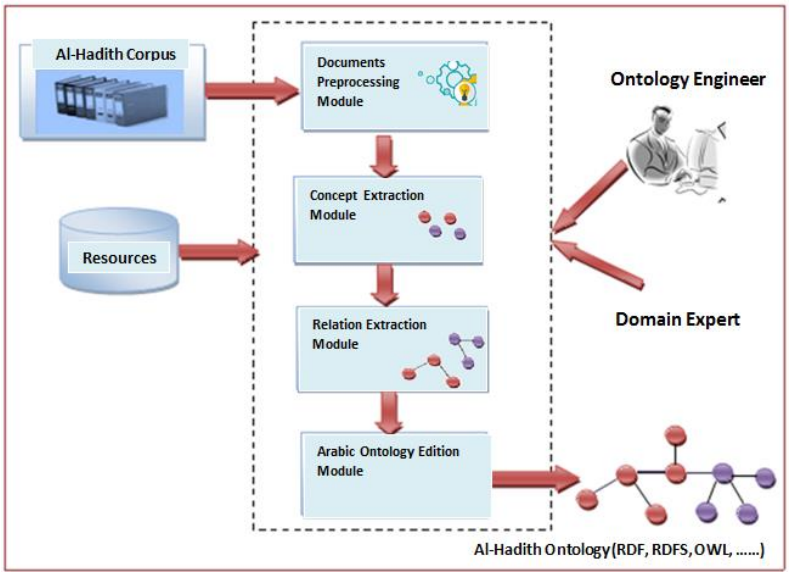

Figure 4: A framework for Ontology construction from Arabic texts [14]

Hawalah, (2018) proposed another approach for building an Arabic ontology from multiple resources (publicly available directory, rich data from the Internet, and Arabic online directory) [13]. The proposed approach consisted of two main phases: (1) building and extracting an Arabic ontology from a publicly available directory, (2) enhancing the Arabic ontology by using richer information from the Internet. The experiments results showed that the classification results of the proposed approach are more accurate than the traditional classification algorithm. But the drawback of this research is using a small number of items in the test dataset that leads to unreliable results.

Albukhitan & Helmy (2013) proposed a new method to development of an Arabic semantic annotation tool. They used information extraction for the domains of food, nutrition, and health [27]. The proposed method aimed to develop Arabic OWL ontologies related to those domains then these ontologies are integrated to produce one ontology. Linguistic patterns are used to determine related relationships between the named entities in the Arabic Web resources. The extracted information was then connected with the corresponding object properties and concepts of the developed ontology to produce the RDF metadata. The results of the proposed method show good precision and recall.

Albukhitan & Helmy (2016) presented an Ontology Learning framework based on some available NLP tools for Arabic text, utilizing the GATE text analysis system for corpus and annotation management [28]. The main steps in this framework were as follows: (1) Text Preprocessing using NLP techniques, (2) Concept Recognition using Data Mining techniques and algorithms based on the statistical measures, and (3) Relation extraction using the

traditional techniques in addition to some algorithms are built to enhance the accuracy. To evaluate the effectiveness of the proposed framework, a set of 100 Arabic documents were selected and manually annotated by hand. The results showed that the precision was good while the recall was low.
The deep learning techniques were used in solving the ontology learning problem for the Arabic language texts.

Albukhitan et al. (2017) proposed a new system for Arabic ontology learning using deep learning [15]. The researchers used the Continuous Bag of Words (CBOW) and Skip-gram models to examining the performance of implementing deep learning with Arabic ontology learning tasks. The proposed system consists of five steps namely: Data Acquisition, Preprocessing, Deep Learning, Ontology creation, and ontology alignment. The experimental results showed that the proposed system is better than the traditional ontology learning systems.

Table 3 shows a comparison between the work of (Albukhitan & Helmy, 2016) as Traditional Ontology Learning (OL) and (Albukhitan et al. 2017) as Ontology Learning using the Deep Learning approach. It shows that using the Deep Learning approach the performance is enhanced.

TABLE III

COMPARISON BETWEEN THE WORK OF (ALBUKHITAN & HELMY, 2016) AND (ALBUKHITAN ET AL. 2017) [15]

Approach	Task	Count	Extracted	Correct	Precision	Recall
Traditional OL Approach	Concepts	5022	4522	3261	72.1%	64.9%
	Taxonomy Relation	650	551	461	83.7%	70.9%
	Non Taxonomy Relation	180	117	88	75.2%	48.9%
OL using Deep Learning Approach	Concepts	5022	4803	3861	80.39%	76.88%
	Taxonomy Relation	650	591	494	83.59%	76.00%
	Non Taxonomy Relation	180	122	93	76.23%	51.2%

Table 4 shows a summary of some researches on Automatic Ontology Learning from Arabic Text.

TABLE IV

SUMMARY OF SOME WORKS ON AUTOMATIC ONTOLOGY LEARNING FROM ARABIC TEXT

Ref.	**Techniques**	**Datasets**	**Results**		**Description/ Limitations**
AlRajebah, & AlKhalifa (2014)	Linguistic Techniques (NLP tools + pattern matching)	Arabic Wikipedia Articles	Precision = 65 %		More concepts and relations are duplicated.
Harrag et al. (2014)	Linguistic + statistic (Association Rules)	12 Surahs from the Holy Quran related to the stories of prophets in the Qur'an 1407 verses, 16153 words.	No Result to show		They presented a new representation tool for Quran.
Al-Arfaj & Al-Salman, (2014)	Linguistic + statistic + Data Mining techniques	Hadith corpus	No Result to show		They proposed their framework depending on evaluating each step by an expert.
Albukhitan & Helmy (2013)	linguistic techniques (NLP Tools) + Statistical techniques	5000 Web documents about food, nutrition, and health	**More relationships are for the assessment such:** <u>Nutrition - Body Functions:</u> Precision = 69% <u>Food - Body Functions:</u> Precision = 80%		The results of some relationships aren't satisfying.
Albukhitan & Helmy (2016)	NLP + statistical and data mining techniques	100 Arabic documents	**Concepts:** Precision = 84% Recall = 77% **Taxonomic Relations:** Precision = 67% Recall = 46%	**Non Taxonomic Relations:** Precision = 73% Recall = 29%	The results of some relationships are low recall
Albukhitan et al. (2017)	NLP + Deep Learning	About 5 thousand words	**Concepts:** Precision =80% Recall = 77% **Taxonomic Relations:** Precision = 83% Recall = 76%	**Non Taxonomic Relations:** Precision = 76% Recall = 51%	The results of some relationships are low recall

6 COMMENTS AND LIMITATIONS OF RESEARCH WORK ON ARABIC OL

From the previous work on the automatic Ontology construction, we concluded the following:

- There are few efforts for building Arabic ontology automatically.
- Most of them focused on a specific domain.
- Using a small dataset can affect negatively the results.
- Using the Hybrid approaches (Linguistic and Statistical) gives better results.
- Using the deep learning techniques such as Continuous Bag of Words (CBOW), Skip-Grams, and Word embedding can improve the ontology learning process.
- The type of Arabic language like (classical Arabic, Modern Standard Arabic, or Dialect Arabic) has effects on the performance.
- The majority of the work dedicated to extracting the semantic relationships was for Building Ontologies.

7 CONCLUSION AND FUTURE WORK

After years of development, Ontology Learning Process (OL) for Latin-character languages, such as English, has been refined greatly. Arabic, however, possesses several challenges that make OL more difficult. In This paper, we review these challenges and survey some of the recent research aiming to improve the Arabic Ontology Learning. We found that most of the research works focused on three main issues: (1) extracting the Terms, (2) extracting the semantic relations, and (3) building the ontology automatically. “Extracting the terms” is through three approaches: rule-based approach, machine learning approach, and hybrid approach. The survey showed that the hybrid approach is the best technique for extracting the terms from the text. The works of “extracting the Semantic relations” process showed that using the hybrid techniques such as the NLP techniques with (the statistical techniques or Data Mining techniques) improve the process of semantic relation extraction from the Arabic text. From the review presented in this paper, we concluded that using the Deep Learning techniques will help in improving the Ontology Learning process.
In the future, we will present a new approach for the ontology learning process that will help in solving the problems that are found out in the previous works and trying to improve the performance. The different techniques such as Natural Language Processing, Machine Learning, Deep Learning, and Data mining techniques will be investigated to choose the best among them to be used for each step in developing the approach.

REFERENCES

[1] T. Berners-Lee, J. Hendler and O. Lassila, “The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities”. *Scientif*ic American, 284.5, 2001.

[2] M. J. Somodevilla, D. Vilariño Ayala and I. Pineda, “An Overview on Ontology Learning Tasks”. *Computación y Sistemas*, 22(1), 2018.

[3] T. Gruber, “A Translation Approach to Portable Ontology Specifications”. *Knowledge Acquisition*, 5(2):199-220, 1993.

[4] M. Hazman, S. R. El-Beltagy and A. Rafea, “A survey of ontology learning approaches”. *database*, 7(6), 2011.

[5] A. Barforush, “Ontology learning: revisited”. J. Web Eng.11 (4), 269–289, 2012.

[6] E. Blomqvist, "Fully automatic construction of enterprise ontologies using design patterns: Initial method and first experiences." *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, Berlin, Heidelberg, 2005.

[7] O. Corcho, M. Fernández-López, A. Gómez-Pérez and A. López-Cima, “Building legal ontologies with METHONTOLOGY and WebODE”. *In Law and the semantic web*. Springer, Berlin, Heidelberg. (142-157), 2005.

[8] B. Kapoor and S. Sharma, “A comparative study ontology building tools for semantic web applications”. *International Journal of Web & Semantic Technology* (IJWesT), 1(3), 1-13, 2010.

[9] A. M. Al-Zoghby, A. Elshiwi and A. Atwan, “Semantic Relations Extraction and Ontology Learning from Arabic Texts—A Survey”. *In Intelligent Natural Language Processing: Trends and Applications* (199-225). Springer, Cham, 2018.

[10] T. Gherasim, M. Harzallah, G. Berio and P. Kuntz, “Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application”. *In Advances in Knowledge Discovery and Management* (177-201). Springer, Berlin, Heidelberg, 2013.

[11] Al Zamil, M. G. & Al-Radaideh, Q. (2014). Automatic extraction of ontological relations from Arabic text. Journal of King Saud University-Computer and Information Sciences, 26(4), 462-472.

[12] K. Shaalan, “A survey of Arabic named entity recognition and classification”. *Computational Linguistics*, 40(2), 469-510, 2014.

[13] A. Hawalah, “A Framework for Building an Arabic Multi-disciplinary Ontology from Multiple Resources”. *Cognitive Computation*, 10(1), 156-164, 2018.

[14] A. Al-Arfaj and A. Al-Salman, “Towards ontology construction from Arabic texts-a proposed framework”. *In Computer and Information Technology* (CIT), 2014 IEEE International Conference on (737-742). IEEE, 2014.

[15] S. Albukhitan, T. Helmy and A. Alnazer, “Arabic ontology learning using deep learning”. *In Proceedings of the International Conference on Web Intelligence* (1138-1142). ACM, 2017.

[16] H. Al-Jumaily, P. Martínez, J. L. Martínez-Fernández and E. Van der Goot, “A real time Named Entity Recognition system for Arabic text mining”. *Language Resources and Evaluation*, 46(4), 543-563, 2012.

[17] M. Asharef, N. Omar, M. Albared, Z. Minhui, W. Weiming and Z. Jingjing, “ Arabic Named Entity Recognition In Crime Documents”. *Journal of Theoretical and Applied Information Technology*, 44(1), 1-6, 2012.

[18] M. H. Btoush, A. Alarabeyyat and I. Olab, “Rule based approach for Arabic part of speech tagging and name entity recognition”. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, 7(6), 331-335, 2016.

[19] H. A. Alsayadi and A. ElKorany, "Integrating semantic features for enhancing Arabic named entity recognition". *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, 7(3), 2016.

[20] M. Gridach, "Character-aware neural networks for Arabic named entity recognition for social media". *In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)* (23-32), 2016.

[21] C. Helwe and S. Elbassuoni, "Arabic named entity recognition via deep co-learning". *Artificial Intelligence Review*, 52(1), 197-215, 2019.

[22] S. Abdallah, K. Shaalan and M. Shoaib, "Integrating rule-based system with classification for Arabic named entity recognition". *In Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing), Springer- Verlag, Berlin Heidelberg, 311-322, 2012.

[23] M. Oudah and K. Shaalan, "NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic". *Natural Language Engineering*, 23(3), 441-472, 2017.

[24] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora". *In: Proceedings of the 14th conference on Computational linguistics*, vol. 2, (539–545). Association for Computational Linguistics, 1992.

[25] N. I. Al-Rajebah and H. S. Al-Khalifa, "Extracting ontologies from Arabic Wikipedia: A linguistic approach". *Arabian journal for Science and Engineering*, 39(4), 2749-2771, 2014.

[26] F. Harrag, A. Al-Nasser, A. Al-Musnad, R. Al-Shaya and A. S. Al-Salman, "Using association rules for ontology extraction from a Quran corpus". *In Proc. 5th Int. Conf. Arabic Language Process* (1-8), 2014.

[27] S. Albukhitan and T. Helmy, "Automatic ontology-based annotation of food, nutrition and health Arabic web content". *Procedia Computer Science*, 19, 461-469, 2013.

[28] S. Albukhitan and T. Helmy, "Arabic Ontology Learning from Unstructured Text". *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 492-496. DOI: 10.1109/WI.2016.0082, 2016.

BIOGRAPHY

Mariam Muhammad is an Assistant Teacher at the Department of Computer Sciences, Faculty of Statistical Studies and Research (FSSR), Cairo University. She received her B.Sc. in information systems from faculty of computers and information, Helwan University, Egypt. She received her MCs degree in Information Systems from FSSR, Cairo University, Egypt in 2016.

Nesrine Ali Abd el Azim is a lecturer at the Department of Information Systems, (FSSR), Cairo University. She received her B.Sc. in information systems from Faculty of Computers and Information, Cairo University, Egypt. She received her M.Sc. degree in Information Systems from the Faculty of Computers and Information, Cairo University, Egypt. She received her Ph.D. degree in Information Systems, FSSR, Cairo University in 2015.

Mervat Hassan Gheith is a professor at the Department of Computer Sciences, Faculty of Statistical Studies and Research (FSSR), Cairo University. Dr. Mervat was the head of the department. She published many papers in computer sciences fields such as NLP and other artificial intelligence applications.

ARABIC ABSTRACT

استعراض لبعض الدراسات السابقة لعملية بناء الأنطولوجي من النصوص العربية

مريم عادل محمد*[1] , نسرين على عبدالعظيم*[2] , مرفت حسن غيث**[3]

*قسم نظم وتكنولوجيا المعلومات, كلية الدراسات العليا للبحوث الإحصائية, جامعة القاهرة, مصر.

[1]eng.maryamadel@yahoo.com

[2]nesrinealiazim79@hotmail.com

**قسم علوم الحاسب, كلية الدراسات العليا للبحوث الإحصائية, جامعة القاهرة, مصر

[3]mervat_gheith@yahoo.com

ملخص:

لعملية بناء الانطولوجي دور هام فى العديد من المجالات خاصة فى مجال "الويب الدلالي", والذي يعتمد نجاحه على جودة الانطولوجيا المستخدمة فيه. هناك الكثير من الأعمال البحثية المهتمة ببناء الأنطولوجي من النصوص العربية. تعمل هذه الدراسات فى ثلاث اتجاهات أساسية وهي: (1) استخلاص الكلمات والمصطلحات, (2) استخلاص العلاقات ذات المعنى الدلالي, (3) بناء الأنطولوجي من النصوص العربية.

نستعرض فى هذه الورقة خصائص اللغة العربية المميزة عن غيرها من اللغات والتى أدت إلى تطوير عدد قليل من أنظمة بناء الأنطولوجي. كما تم عمل مقارنة بين الأبحاث المقدمة في هذا الموضوع طبقا للأدوات التى استخدمها الباحثون والنتائج التى توصلوا إليها. وبناءا على هذه المقارنة تم استخلاص التحديات التى قابلت الباحثين وتم التعليق عليها واقتراح بعض التعديلات التى قد تساهم في تحسين النتائج.

الكلمات المفتاحية:

عملية بناء الانطولوجي , استخلاص الكلمات والمصطلحات , استخلاص العلاقات ذات المعنى الدلالي.