

Arabic open information extraction system using dependency parsing

Sally Mohamed Ali El-Morsy^{1,2}, Mahmoud Hussein¹, Hamdy M. Mousa¹

¹Computer Science Department, Faculty of Computers and Information, Menoufia University, Al Minufya, Egypt

²Higher Institute of Engineering and Technology in Tanta, Tanta, Egypt

Article Info

Article history:

Received Jan 2, 2021

Revised May 27, 2021

Accepted Jun 29, 2021

Keywords:

Arabic

Clause tuples

Dependency parsing

Open information extraction

ABSTRACT

Arabic is a Semitic language and one of the most natural languages distinguished by the richness in morphological enunciation and derivation. This special and complex nature makes extracting information from the Arabic language difficult and always needs improvement. Open information extraction systems (OIE) have been emerged and used in different languages, especially in English. However, it has almost not been used for the Arabic language. Accordingly, this paper aims to introduce an OIE system that extracts the relation tuple from Arabic web text, exploiting Arabic dependency parsing and thinking carefully about all possible text relations. Based on clause types' propositions as extractable relations and constituents' grammatical functions, the identities of corresponding clause types are established. The proposed system named Arabic open information extraction (AOIE) can extract highly scalable Arabic text relations while being domain independent. Implementing the proposed system handles the problem using supervised strategies while the system relies on unsupervised extraction strategies. Also, the system has been implemented in several domains to avoid information extraction in a specific field. The results prove that the system achieves high efficiency in extracting clauses from large amounts of text.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sally Mohamed Ali El-Morsy

Computer Science Department, Faculty of Computers and Information, Menoufia University

Al Minufya, Egypt

Email: smbm222@yahoo.com

1. INTRODUCTION

The advancement in technology promoted the rapid increase in online information volume in recent years. Unstructured knowledge should be organized to take advantage of it and can be used in various fields. Information extraction is employed to organize knowledge from unstructured data. Information extraction (IE) methods extract structured information within relations, objects, entities, events, and many alternative sorts. The volume of structured, unstructured, and semi-structured knowledge and the speed of increasing huge data has led to the need for new techniques to deal with this size and type of information [1]. Several kinds of research treat information extraction among entity extraction and relation extraction in the Arabic language. However, the extracted information has different drawbacks such as uninformative, incoherent, and overly specific relations. These defects appear because of the special nature of the Arabic language with rich morphology. The Arabic language has complicated morphology, meaning each word may consist of one or more prefixes, a stem or root, and one or more suffixes [2].

Furthermore, different challenges face Arabic natural language processing (NLP), especially information extraction. Also, the Arabic language features a lack of capitalization, unlike other languages such as English, in which capital letters are used to recognize name entities. Over the last few years, the researchers

proposed and developed various systems and techniques to overcome obstacles in the Arabic language [1]. Also, the Arabic language faces many challenges due to its inflectional, derivational, and syntactic structures. The grammatical of Arabic sentence has many structures, mainly, nominal sentence composes of subject-verb-object (SVO), and verbal sentence consists of verb-subject-object (VSO) [3]. In addition to the complexity of the Arabic language, different representations, semantic interpretation, and the heterogeneity of data types are often the main problems and the intrinsic properties of the collected raw data massive data. These problems represent the main challenges for extensive data analysis. To overcome these challenges and perform big data analysis, it must be preparing and transforming these raw into a suitable form for analysis. So, the IE process must be efficient enough to handle heterogeneity, dimensionality, and data diversity [4].

Despite recent progress in IE, extracting information from the web presents several challenges for existing systems. There is massive and heterogeneous data on the web, interest relations are not easy to predict, and the number of relations can be huge [5]. One of the most significant challenges for parsers is robustness, the ability to analyze any input. These drawbacks lead to the use the open information extraction to facilitate the discovery of relations in large-scale text and heterogeneous corpora. The relation corpora extracted by open information extraction systems (OIE) systems are valuable resources for downstream tasks like automated knowledge base construction, open question responsive, event schema induction, generating illation rules, or for up OIE systems themselves [6]. Therefore, the Arabic OIE system is designed to overcome the identified challenges for big Arabic data and the limitations of existing IE techniques. Accordingly, this paper proposes a novel framework, called Arabic open information extraction (AOIE), to identify relation tuples in Arabic web text. The grammatical function of its coherent constituent determined the corresponding clause type for every relation. This system used a heterogeneous corpus in the (CoNLL-U) file format by the UDPipe application [7], [8].

The proposed system is expected to improve the information extraction for the Arabic language by providing relation tuples. The proposed system has been evaluated by determining its precision, F-measure, and recall. The results revealed the system's good performance while the precision reaches 91%, recall reaches 84%, and F-measure is 87%. The system has also been applied in several fields: weather, social, sport, health, biomedical, and economical. The overall precision for each field consecutively is 91%, 80%, 81.8%, 91%, and 88%. The evaluation has been done for different sentence complexity levels. These levels are simple, complex, highly complex, and extremely complex. This article is organized as follows, related works are presented, and in section 2, section 3 explains the methodology framework, section 4 is explained the proposed system, section 5 illuminates result and evaluation, and ends with the conclusion and future work.

2. RELATED WORK

Current IE systems focus on analyzing the local context within individual sentences to extract entities and their relationships in a specific field while ignoring the redundant information that can be collectively [9]. In comparison with other languages, we observe a scarcity in efforts related to Arabic-based information extraction, which could be partly imputed to the complexity of Arabic makes it difficult to extract relations automatically [10]. The Arabic name entity recognition (NER) as a base for relation extraction applications has a significant share of this field research. Mesmia *et al.* [9] proposed a system for recognizing the Arabic NER based on two transducers for analysis and synthesis. Darwish and Gao presented simple, effective, and language-independent approaches for improving NER in microblogs for Arabic as an example [11]. Sabty *et al.* [12] proposed a system for extracting Arabic NER dependent on word embedding. A word embedding is a text's representation representing all similar meaning words.

Furthermore, for relation extraction, El-Salam *et al.* [13] extracted binary relations between two Arabic-named entities in a specific domain from the web using a semi-supervised technique. Also, Fasha *et al.* proposed the information extraction model for Arabic text that relatively open-text domains. This model contains two-phase. The first phase extracts part-of-speech (POS) tagging relations. Using description logic in the second phase for extracting the implicit knowledge [14]. However, Open IE provides proper data compression, compared to search snippets or reading an original document, while still retaining important information. This can help an end-user in obtaining a summary view of a concept [10]. The previous work in OIE is divided into two generations regarding model considerations. The first generation is the training data-based OIE which generates patterns based on training data represented employing the dependency tree or part of speech (POS) tagged text [15]. This generation has two methods; the first method uses training data and shallow syntax to learn extractors or estimate the confidence of those systems relying on extensive human involvement [10]. The examples for this type are TextRunner [16] and ReVerb [17]. The second method is training data and dependency parsing, which performs POS tagging, syntactic chunking, and dependency parsing and returns a set of relation triples, for example, the OLLIE model [18]. The second generation is Rule-based OIE which relies on hand-crafted patterns from POS-tagged text or rules operating on dependency parse

trees [15]. This generation has two methods; the first method is the rule-based and shallow syntax which extracts relationships based on the simple constraint. Every relational is a verb or a verb followed by a preposition or a verb followed by nouns, adjectives, or adverbs, for example, ReVerb model [15]. The second method is rule-based and dependency parsing which uses hand-crafted heuristics operating on dependency parses, for example, the ClausIE model [19].

Although most of the research in OIE is interested in the English language, several research types focus on other languages, such as [20], which presents the German OIE system (GerIE) depending on hand-crafted rules working on dependency parsed sentences. Furthermore, Jia *et al.* [21] proposed an unsupervised model that extracts open entity relations and solves Chinese linguistic troubles. Also, the dependency semantic normal forms are used to extract entity relation triples. Truong *et al.* proposed a method of OIE for Vietnamese named (vnOIE) using a clause-based approach and generated open relationship and their arguments from Vietnamese. The model formulates Vietnamese dependency parsing considering all possible relationships in a sentence using grammatical clauses [22]. Niklaus *et al.* presented three significant challenges in Open IE systems. The first is automation applying the unsupervised extraction strategies in the open IE systems, automatically detecting possible interest relations with only a single pass over the corpus and automatically generating the relevant training data. The second is corpus heterogeneity which prevents or hinders the progress of the syntactic or dependency parsers. The last is efficiency. Open IE systems are effective if they can scalability and process a large amount of text in various domains [10].

Previous Arabic information extraction research and applications suffer from the high ratio of incoherent output information. Most of them are interested in a specific domain with supervised methods, making the method not used for different purposes. Also, the previous research focuses only on binary relation and name entity recognition while there is a lack of extracting the coherent ratio in Arabic text. OIE practices in different languages also shed light on the most suitable method to use in Arabic following vnOIE, which generates open relations from the Vietnamese language. The system reveals the effectiveness of using dependency parsing in complex language morphology. Accordingly, this paper's proposed system follows training data on heterogeneous corpus depends on the corpus used (CoNLL-U) file format and yields into the more type of relation type. In the next section, more highlights are provided about the proposed system.

3. THE METHODOLOGY FRAMEWORK

Regarding the challenges faced by the Arabic IE and the reviewing of the OIE system developed for different languages, this research proposes an information extraction system called Arabic open information extraction (AOIE). The system extracts relation tuples representing essential clauses or assertions from the text. To formulate effective Arabic OIE system, this research attempts to formulate the research framework as shown in Figure 1. The objective of this system is to yield Arabic clauses to get as coherent information as possible. The complex morphology of the Arabic language is the main obstacle to perform such extraction. Accordingly, some Arabic features like tokenization and part of speech (POS) have been used. The research framework consists of four stages as following.

In the first stage, using the Arabic dependency parsing database developed by Mohamed *et al.* [8]. This corpus contains an index for the sentences and their linguistic meta-data to enable quick mining and search across the corpus. The dependency relation in this corpus has seventeenth morphological annotations and eight features based on identifying the textual structures then recognizing and understanding their grammatical characteristics to perform the dependency relation. The parsing and dependency process is conducted by the universal dependency system and corrected manually [8].

The second stage, build Arabic OIE system; in this stage, the proposed system for Arabic OIE has been built by determining the initial clause types. These clauses are grammatical parts depend on the dependency parsing for examples they are VS ("verb", "subject"), VSO ("verb", "subject t", "object"), VSOA ("verb", "subject", "object", "adjective"), VO ("verb", "object"), VOA ("verb", "object", "adjective"), and VA ("verb", "adjective"). Python programming language has been used to perform the proposed system depending on natural language processing libraries to deal with the Arabic dependency parsing feature included in the (CoNLL-U) format.

The third stage, analyze and evaluate, analyze the results and evaluated them by checking the validity of each clause, and the feedback of this step is considered input for the next stage. Then, improve the system accuracy; this step tries to enhance the system's performance by manually improving the sequences and rules of the system algorithm. For example, in the verbal sentence universal part of speech tagging (UPosTag) = "VERB" and "Head" "0": then the Root = Verb Return (V), and If "Head" = Root" ID" and universal dependency relation (DepRel) = "nsubj": Return (S). Also, in the nominal sentence, if DepRel" value is "nsubj" set it in as the inchoative (I), and if "DepRel" value is "nmod" or "amod" set it as a Root and predicate (E). Finally, the system reaches a satisfying form in several domains to be implemented as approved from the measurement standards including, precision, recall, and F-measure. Therefore, AOIE aims to address three

problems: relying on supervised extraction strategies, Transportability as systems is intended for domain-independent usage, efficiency to scale to large amounts of text readily. Every year, the volume of unstructured data doubles [4], and the extraction of semantic information from such a huge amount of unstructured data become more important.

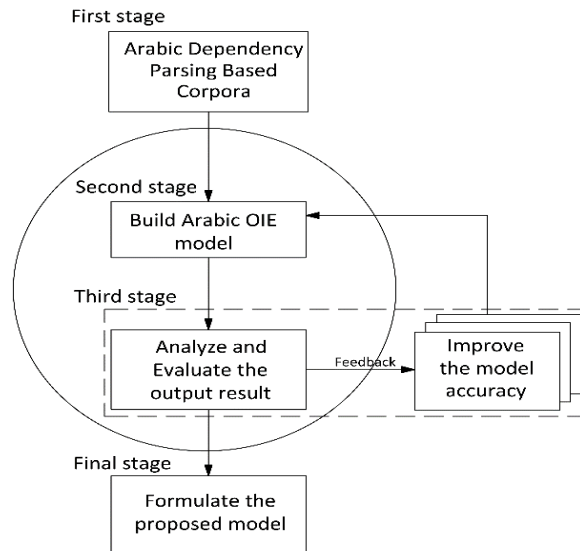


Figure 1. The research framework

3.1. Arabic dependency parsing

In the Arabic language, Arabic grammarians analyze all Arabic words into three main parts-of-speech. This parts-of-speech are distinguished and differentiated into more intricate parts for covering as a whole of the Arabic language. These parts are: i) noun: a noun in Arabic is a name or a word that describes a person, thing, or idea, Verb: it's the most crucial word in the sentence; ii) the Arabic verb classified into perfect, imperfect, and command. Furthermore, the verb can be classified based on gender, number; and iii) particle: the particle includes prepositions, adverbs, and conjunctions [23]. The parts-of-speech (POS) tag is an important feature; these are used in open relation extraction. The Universal dependencies framework has been used to match different types of dependency relations in different languages. There are seventeen dependency relation types provided by parsers trained on Arabic-padt-ud-2.4-data, among the subject and object, relative, adverbial and adnominal clauses, conjunction, auxiliary, and parataxis [24]. Natural language toolkit (NLTK) is an open-source suite of libraries and programs that can be integrated within the Python environment and then used to perform different statistical and rule-based natural language processing tasks POS tagging and parsing [25]. NLTK Applying the top-down as shown in Figure 2 to parse the following Arabic sentences and produce the corresponding dependency-parsing tree for the sentence:

"Coronaviruses are a broad spectrum of viruses that may cause disease in animals and humans"
 "تعتبر فيروسات كورونا ، سلالة واسعة من الفيروسات التي قد تسبب المرض للحيوان والإنسان"

The resulted tree illustrates the parts of speech (POS) tag set, where are some verbs such as; "تكون"; "تعتبر"; and "تسبب"; and some nouns such as "الحيوان", "الإنسان", "المرض", "فيروسات كورونا".

Dependency trees are suitable for various languages [26]. A multilingual parser has been used with a common output tag set for representing the syntactic structure of a sentence. Universal dependency relation has several types, as shown in Table 1. The beginning of the tree is usually the verb that presents the sentence's root connected with nouns, and other adjectives and adverbial connect these nouns. A well-formed dependency tree for an input sentence is simply a tree with the appropriate nodes, their nodes map, one to one, to the tokens due to the morphological analysis and tokenization, and their roots collect the nodes according to the division into sentences or paragraphs. The conversion of these trees was the easiest task as the linguistic representation was already what we needed [27]. Since the (CoNLL-U) has promoted multilingual dependency parsing and provided resources for this, much progress has been made in this area, and the number of freely available dependency parsers has increased.

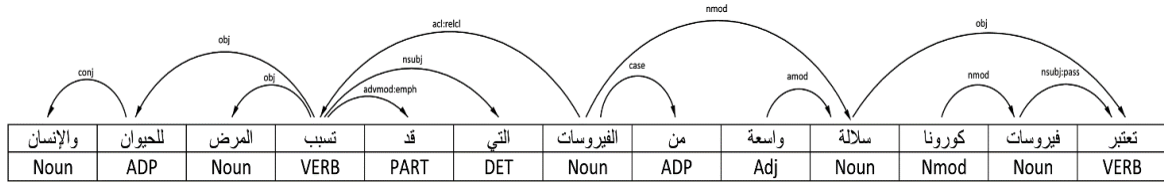


Figure 2. Dependency parsing structure of Arabic sentence

Table 1. Dependency parsing relation

Rel	Definition	Rel	Definition	Rel	Definition
Root	points to the root of the sentence	Parataxis	Used to connect to sentences together	Advmod	adverbial modifier
Nsubj	A nominal subject	Obl	adverbial attaching to a verb, adjective	Aux	Auxiliary
Amod	Adjectival modifier	Case	providing a more uniform analysis of nominal elements	Conj	Conjunct
Obj	Direct object	Nmod	nominal modifier	Xcomp	open clausal complement
Fixed	certain fixed grammaticized	Det	Determiner	Comp	comparison constructions
Cc	coordinating conjunction	Acl	an adverbial clause		

3.2. Arabic dependency corpora

In this study, we conducted experiments on the Arabic dependency parsing based corpora for information extraction, which has been presented in (CoNLL-U) format as dependency parsing (DP) input [8]. This corpus depends on text from the web and includes several fields they are weather, economic, social, sport, health, and biomedical. Table 2 illustrates the sample of the Corpora generated by UDPipe model. UDPipe model conducts (CoNLL-U) format files by performing tokenization, morphological analysis, POS tagging, lemmatization, and dependency parsing for nearly Universal Dependencies 2.5 [28], [29]. UDPipe was developed at Charles University in Prague [28], [30]. The UDPipe output is containing universal POS tags (UPOS), language-specific POS tags (XPOS), a universal subset of morphological features (UFeats), Lemmatization (Lemmas), Universal dependency relation (DepRel), and the Head (root if Head = 0) [28], [31]. The following section introduces the proposed system of Arabic open information extraction.

Table 2. Sample from the Arabic dependency parsing-based corpora for information extraction

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel
# newdoc							
# newpar							
# sent_id = 1							
# text = تعتبر فيروسات كورونا، سلالة واسعة من الفيروسات التي قد تسبب المرض للحيوان والإنسان							
1	تعتبر	اعتَبَر	VERB	VIIP-3FS--	Aspect=Imp Gender=Fem Mood=Ind Number=Sing Person=3 VerbForm=Fin Voice=Pass	0	root
2	فيروسات	فَيْرُوسَات	NOUN	N-----1R	Case=Nom Definite=Cons Number=Plur	1	nsubj
3	كورونا	كُورُونَا	X	U-----	-	2	nmod
4	،	،	PUNCT	G-----	-	3	punct
5	سلالة	سَلَالَة	NOUN	N-----S1I	Case=Nom Definite=Ind Number=Sing	1	obj
6	واسعة	وَاسِعَة	ADJ	A----FS1I	Case=Nom Definite=Ind Gender=Fem Number=Sing	5	amod
7	من	مِنْ	ADP	P-----	AdpType=Prep	8	case
8	الفيروسات	فَيْرُوسَات	NOUN	N-----P2D	Case=Gen Definite=Def Number=Plur	5	nmod
9	التي	الَّتِي	DET	SR----FS2-	Case=Gen Gender=Fem Number=Sing PronType=Rel	11	nsubj
10	قد	قَدْ	PART	F-----	-	11	advmod:emph
11	تسبب	سَبَّبَ	VERB	VIIA-3FS--	Aspect=Imp Gender=Fem Mood=Ind Number=Sing Person=3 VerbForm=Fin Voice=Act	8	acl
12	المرض	مَرَضٌ	NOUN	N-----S4D	Case=Acc Definite=Def Number=Sing	11	obj
13	للحيوان	لِلْحَيَوَانَاتِ	X	X-----	Foreign=Yes	14	Nmod
14	والإنسان	وَالْإِنْسَانَ	X	U-----	-	12	Nmod
15	.	.	PUNCT	G-----	-	1	Punct
# text = جهاز البارومتر هو الجهاز الخاص بقياس الضغط الجو							
1	جهاز	جِهَاز	NOUN	N-----S1R	Case=Nom Definite=Cons Number=Sing	4	nsubj
2	البارومتر	الْبَارُومِتْر	IDAF	U-----	-	1	nmod
3	هو	هُوَ	PRON	SP---MS1-	Case=Nom Gender=Masc Number=Sing Person=3 PronType=Prs	4	nmod
4	الجهاز	جِهَاز	NOUN	N-----S1D	Case=Nom Definite=Def Number=Sing	0	root
5	الخاص	خَاصٌ	ADJ	A----S1D	Case=Nom Definite=Def Gender=Masc Number=Sing	4	amod
6	بقياس	بِقِيَاس	NOUN	N-----S1R	Case=Nom Definite=Cons Number=Sing	4	nmod
7	الضغط	الصَّغْط	NOUN	N-----S2D	Case=Gen Definite=Def Number=Sing	6	nmod
8	الجو	جَوٌّ	NOUN	A----S2D	Case=Gen Definite=Def Gender=Masc Number=Sing	7	amod
9	.	.	PUNCT	G-----	-	4	punct

4. THE PROPOSED SYSTEM

Open information extraction (OIE) is an unsupervised task to extract coherent information task from the text. Its output represents the basic clauses or assertions from the text. Clauses can be defined as coherent and pieces of basic information that are non-over-specified. Arabic clause is a grammatical unit. It is considered part of a sentence that expresses some coherent information [21], [22], [32].

Several stages and steps have been followed to build, test, and improve the proposed system as show in Figure 3, and an Arabic dependency database developed by Mohamed *et al.* [8] has been used as shown in the algorithm of identifying clause type. The used database is an Arabic dependency parsing-based corpora including the different grammatical features such as universal part of speech tagging (UPosTag), Head, and DepRel. In the beginning, the initial clause types have been determined following the Arabic sentence elements. After that, the grammatical rules for identifying each element in the clause have been determined dependent on the initial clause types, and the dependency features these existing in (CoNLL-U) files. Consequently, the system is constructed by employing python programming language using several open-source packages. The source code of the proposed system is uploaded to GitHub [33].

Initially, every sentence in the corpora is separated. The following algorithm of Identifying clause types illustrates the sequence of detecting each clause in the sentence and identifying their types. For the verbal clauses, based on the initial clauses' types and the dependency features these existing in (CoNLL-U) files, the system tries to find the verb (V) in the clause. If the verb is found, then set it as the root. To identify the clause parts, determining the ID of the clause words, then find the related elements such as subject (S), object (O), and adverb (A). Depending on the discovered elements, the clause type is determined as VS, VSO, VO, VOA, VA, SV. If the verb is not found and identified, the clause is recognized as a nominal sentence, and the first word (noun) is set as inchoative (I), and the algorithm tries to find the predictive (E) using the grammatical rules, which is the noun follows the inchoative in the sentence. Consequently, the clause type is determined as (IE).

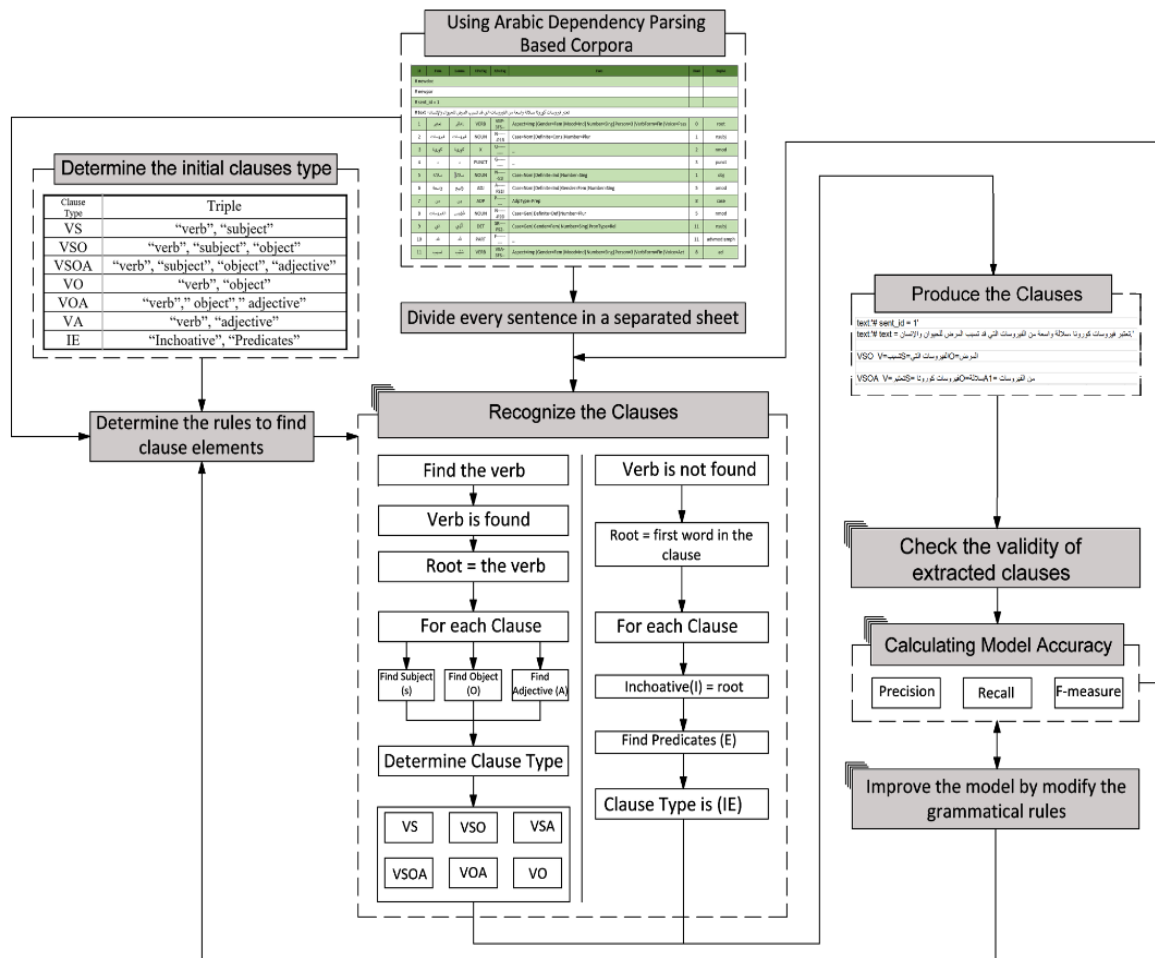


Figure 3. The proposed system building steps

Algorithm: Identifying clause types

```

Input: DP from a sentence
Output: Set of clauses and their types
Do Identifying all clauses in the sentence // recognize each clause in the
      Root, S, V, O, A, I, E = Null // sentence
For every word in clause // S is Subject, V is Verb "Root"
  //Find clause root //O is Object, A is Adjective
  If UPosTag == verb and Head == 0: // verbal clauses
    Root = (V) // (Root_ID)
    //Assign clause parts
    Find (S)
    Find (O)
    Find (A)
  If (S, O and A are found) then
    Clause Type ←VSOA
    Else if (S and O are found) then
      Clause Type ← VSO
    Else if (S and A are found) then
      Clause Type ← VSA
    Else If (S is found) then
      Clause Type ← VS
    Else if (A is found) then
      Clause Type ←VA
    End if // a nominal clause
  Else if UPosTag != verb // I is inchoative, E is the
    Root = (I) predicate
    Find (E)
    Clause Type ← IE
  End if
  Spill clause
  Save clause type and clause tuple
End For
End Do
Save Sorted clause elements and types in excel
sheet

```

After determining the clause type, the clause has been split, and the same processes were repeated for all clauses in the sentence. These processes are repeated for all separated sentences, and the result is sorted in an Excel sheet containing each clause element and its type. After the validity of extracted clauses has been checked manually and the system accuracy has been calculated using the precision, recall, and f-measure. The grammatical extraction rule was modified, and the system was tested to improve the results until the final results were achieved. For example, the sentence shown in Figure 2 has been processed as following sentences:

"تعتبر فيروسات كورونا ، سلالة واسعة من الفيروسات التي قد تسبب المرض للحيوان والإنسان" with the English translation "Coronaviruses are a broad spectrum of viruses that may cause disease in animals and humans"

The (CONLL-U) format file for this sentence has been used in the extraction processes shown in Table 3. First, the system seeks for the verb (V) by finding "UpoTag" value which is "VERB" in this case, if the "Head" of this verb is "0" the verb is considered as the root for this clause which is "تعتبر" in this example then the system finds the linked noun with the root and its "DepRel" value is "nsubj" and consider it as subject (S) which is "فيروسات كورونا". After that, the system looking for the linked words to the root and its "DepRel" value is "obj" or "obl:org" and consider it as the object (O) which is "سلالة", and the system looking for the words with "DepRel" value is "case" or "obl" and consider it as the object (A) which is "من الفيروسات" then the first clause in the sentence is complete and its types set as (VSOA). Finally, the clause sorted in the results as following:

V=تعتبر S= فيروسات كورونا O=سلالة A= من الفيروسات S=Coronavirus V= is a strain A= of virus

For the rest of the sentence, the same processes have been done, and the following is the result for the second clause in the sentence:

V=تسبب S=الفيروسات التي O=المرض S= viruses that may V=cause O=disease

Table 3 illustrates the result of extracted clauses containing clause types and elements for the rest of the paragraph. Another example sentence is: "جهاز البارومتر هو الجهاز الخاص بقياس الضغط الجوي" with English translation "Barometer is a device for measuring atmospheric pressure." The (CONLL-U) format file for this

sentence has been used in the extraction processes shown in Table 2. First, the system finds the first noun of the sentence "UpoSTag" value which is "NOUN" in this case and check if "DepRel" value is "nsubj" then set it in as the inchoative (I) and looking for all linked words which are "جهاز البارومتر", after that check on the following noun and set it as a predicate (E) and looking for the word linked with it which are "بقياس الضغط الجوي" then the clause items as follows:

- I= جهاز البارومتر, E= بقياس الضغط الجوي

- I= thermometer device, E= is a device specially designed to measure the temperature.

For the rest of the text, Table 3 illustrate the result of extracted clauses containing clause types and elements.

Table 3. Example for analyzing Arabic sentence that illustrated verbal sentences clauses

clause Type	Text	Sentence	Derived clauses
VSOA	من سلالة واسعة من الفيروسات التي قد تسبب المرض للحيوان والإنسان. ومن المعروف أن عدداً من فيروسات كورونا تسبب أمراض تنفسية، تتراوح قوتها من نزلات البرد الشائعة إلى الأمراض الأشد وخامة مثل متلازمة الشرق الأوسط التنفسية (ميرس) والمتلازمة التنفسية الحادة الوخيمة (سارس). و يسبب فيروس كورونا، المكتشف مؤخراً مرض كوفيد-19.	"تعتبر فيروسات كورونا، سلالة واسعة من الفيروسات التي قد تسبب المرض للحيوان والإنسان."	سلالة=O فيروسات كورونا S=تعتبر V= A=من الفيروسات S=Coronavirus, V= is, A= astrain of viruses
VSO	ومن المعروف أن عدداً من فيروسات كورونا تسبب أمراض تنفسية، تتراوح قوتها من نزلات البرد الشائعة إلى الأمراض الأشد وخامة مثل متلازمة الشرق الأوسط التنفسية (ميرس) والمتلازمة التنفسية الحادة الوخيمة (سارس). و يسبب فيروس كورونا، المكتشف مؤخراً مرض كوفيد-19.	"تعتبر فيروسات كورونا، سلالة واسعة من الفيروسات التي قد تسبب المرض للحيوان والإنسان." "Coronaviruses are a broad spectrum of viruses that may cause disease in animals and humans."	V=تسبب S=تسبب O=المرض S= viruses that may, V=cause, O=disease
VSOA	ومن المعروف أن عدداً من فيروسات كورونا تسبب أمراض تنفسية، تتراوح قوتها من نزلات البرد الشائعة إلى الأمراض الأشد وخامة مثل متلازمة الشرق الأوسط التنفسية (ميرس) والمتلازمة التنفسية الحادة الوخيمة (سارس). "It is known that a number of coronaviruses cause respiratory diseases, the strength of which ranges from common colds to more severe diseases such as Middle East Respiratory Syndrome (MERS) and severe acute respiratory syndrome (SARS). Coronavirus, recently discovered, causes Covid-19 disease."	"ومن المعروف أن عدداً من فيروسات كورونا تسبب أمراض تنفسية، تتراوح قوتها من نزلات البرد الشائعة إلى الأمراض الأشد وخامة مثل متلازمة الشرق الأوسط التنفسية (ميرس) والمتلازمة التنفسية الحادة الوخيمة (سارس)." "It is known that a number of coronaviruses cause respiratory diseases, the strength of which ranges from common colds to more severe diseases such as the Middle East Respiratory Syndrome (MERS) and severe acute respiratory syndrome (SARS)."	V=تسبب S=تسبب O=المرض S=عدداً من فيروسات كورونا O=أمراض تنفسية V=caused, S= a number of, O = respiratory diseases V=قوة S=تتراوح O=إلى S= strength, V=ranges, A= to more diseases
VSO	ويُسبب فيروس كورونا، المكتشف مؤخراً مرض كوفيد-19.	"ويُسبب فيروس كورونا، المكتشف مؤخراً مرض كوفيد-19" "Coronavirus, recently discovered, causes Covid-19 disease"	V=يسبب S=فيروس O=مرض كوفيد-19 S=Corona virus, O=discovered, V=causes A=Covid-19 disease
VA	تستخدم عدداً من الأجهزة لرصد نوع المناخ في منطقة ما ومنها . جهاز البارومتر خصيصاً لقياس درجة الحرارة .	"تستخدم عدداً من الاجهزه لرصد نوع المناخ السائد في منطقة ما ومنها . جهاز البارومتر هو الجهاز المعد خصيصاً لقياس درجة الحرارة"	V=تستخدم S=من الاجهزة لرصد نوع المناخ السائد A=من الاجهزة لرصد نوع المناخ السائد V= used, A= from an area
IE	جهاز البارومتر هو الجهاز الخاص بقياس الضغط الجوي	"A number of devices are used to monitor the type of climate prevailing in and from an area. A thermometer is a device specially designed to measure the temperature."	I=جهاز البارومتر E=جهاز المعد خصيصاً لقياس درجة الحرارة I= thermometer device, E= is a device specially designed to measure the temperature.
IE	البارومتر هو الجهاز الخاص بقياس الضغط الجوي	"جهاز البارومتر هو الجهاز الخاص بقياس الضغط الجوي" "Barometer is a device for measuring atmospheric pressure"	I=جهاز البارومتر E=جهاز المعد خصيصاً لقياس الضغط الجوي I= thermometer device, E= atmospheric pressure measuring device

4.1. Evaluation results and discussion

To evaluate our system's extraction, we consider two aspects of each extracted fact: i) coherent or incoherent and ii) minimal or not. A fact is considered coherent, if it keeps the same meaning as in the original sentence. A coherent fact may contain another fact in its arguments. Wrong boundaries, where the relational or argument phrase is either too long or too small. Redundant extraction, as the extraction proposition is already expressed in another extraction. Uninformative extraction, as important information is skipped. Missing extraction (false negative), where an existing relationship is not extracted. The wrong extraction, as there is no meaningful interpretation of the proposition. Finally, measure the quality and efficiency of the system by quantifying the precision, recall, and F-measure using (1):

$$precision = \frac{True\ positives}{True\ positives + false\ positives} \quad (1)$$

Equation (1) presents the precision, which calculates the number of items identified as the number of correctly predicted items. The system's precision has been measured by the ratio between all the extracted facts

and their coherence. We use this ratio to measure the overall precision. Equation (2) presents the recall, which measures how much relevant information the system has extracted. The recall represents a percentage of the total number of correct items for a given topic as the number of correctly predicted items. Equation (3) presents the F-measure to evaluate the overall performance of the system. The above measures have been done for different fields based on several categories of sentences: simple, complex, highly complex, and extremely complex [34]. These categories are determined based on the number of clauses in the sentence. Table 4 shows the results containing each category's ratio and precision in every field [10], [35].

$$Recall = \frac{True\ positives}{True\ positives + false\ negative} \quad (2)$$

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

The complexity of sentence structure affects the number of extracted relations and system precision. However, AOIE achieves high efficiency in determining the relationship between a subject, object, and verbs based on DP analysis, and most of the extracted clause is correct. However, the simple sentence produces one clause and the complex sentence, 2-3 clauses. In highly complex sentences, the numbers of generated clauses are 4-5 clauses, while the extremely complex sentence maybe produce more than five clauses. Table 5 illustrates the system's efficiency applied in different fields of the Arabic text by presenting the overall precision, ratio, and f-measure for each field. Although there is a limitation in Arabic open information research, a binary relation extraction developed by [13] has chosen to compare with AOIE shows a higher accuracy while binary relations model precision ranges from 0.61 to 0.75 while recall ranges from 0.71 to 0.83 and AOIE precision is 0.91 and recall 0.84 as shown in Table 6.

Table 4. Experiment results in different sentence structure cases

Field	Category of sentence	Ratio	Precision	Field	Category of sentence	Ratio	Precision
Sport & Health	Simple	45%	80%	Economic	Simple	20%	94%
	Complex	40%	80%		Complex	25%	79%
	High Complex	10%	82%		High Complex	30%	85%
	Extremely Complex	5%	90%		Extremely Complex	25%	90%
weather	Simple	50%	95%	Biomedical	Simple	50%	95%
	Complex	30%	84%		Complex	30%	84%
	High Complex	10%	82%		High Complex	10%	82%
	Extremely Complex	10%	80%		Extremely Complex	10%	80%
Economic & Social	Simple	65%	89%	NEWS	Simple	40%	90%
	Complex	10%	90%		Complex	20%	50%
	High Complex	15%	85%		High Complex	25%	72%
	Extremely Complex	10%	88%		Extremely Complex	15%	42%

Table 5. Experiment results for different field

Field	Precision	Recall	F-Measure	Field	Precision	Recall	F-Measure
Weather	91%	84%	87%	Biomedical	84%	69%	75%
Economic & Social	81%	68%	73%	Economic	88%	79%	83%
Sport & Health	81.8%	64%	71%	News	67%	50%	57%

Table 6. Comparison between the previous model and the proposed system

Comparison item	Binary relations model	AOIE
PRECISION	75%	91%
RECALL	83%	84%
F-MEASURE	76%	87%

This research is the first comprehensive of the relation extraction system in the Arabic language to the best of our knowledge. AOIE is the first Arabic open information extraction system based on the Arabic language's grammatical clauses, highly scalable in terms of clause extraction, and domain-independent. The system exploits DP analysis to extract relation tuples based on grammatical clauses quickly. This system achieves a precision for performance from 71% to 91%, the recall from 83% to 84, and the F-measure from 76% to 87%. However, some unexpected incorrect extractions could result from the output of Arabic sparing. Unlike English DP, the DP in Arabic may not detect the details of auxiliary adverbs, which could result in incorrect extractions caused by wrongly labeled main verbs. The problem was due to AOIE using heuristic rules to find significant verbs in a sentence based on DP. Secondly, Arabic DP has a limitation on distinguishing

between essential adverbs and verbs. Essential verbs are required, while adverbs may or may not appear in extracted clauses.

In some cases, the AOIE system failed to determine clauses such as VS, VSO, VO, VOA, VA, where components S are subject, V are verb, O are objects, and A is an adverb. Accordingly, the proposed system's implementation addresses the previous problems mentioned in the literature while the system relies on unsupervised extraction strategies and is implemented in several domains. The results also prove that the system achieves high efficiency in extracting clauses from large amounts of text.

5. CONCLUSION

This system presents a solution for the Arabic resources shortage problem where language has a limited number of available resources to be used in NER systems. This paper presents the first attempt to implement the Arabic open information extraction systems. The proposed system takes advantage of the grammatical clause-based approach. By using grammar rules, the proposed system extracts all possible clauses in a sentence. The proposed system identifies the corresponding clause type based on propositions as extractable relations and constituents' grammatical functions. In the experiments, the system has been evaluated using several factors such as grammatical structures of sentences and the number of verbs existing in a sentence. Also, the proposed system addresses the problem of using supervised strategies while the system relies on unsupervised extraction strategies. Then, the system has been implemented in several domains to avoid information extraction in a specific field. The results prove that the system achieves high efficiency in extracting clauses from large amounts of text. The proposed system's output is a set of new relations that contains the most important part of the sentence. The results show that the proposed system delivers promising results. AOIE could be applied to Arabic answering systems or integrated into higher Arabic NLP tasks such as text similarity or text summarization. The complex nature of the Arabic language while the nominal sentence not has a specific form and not contain a verb which makes it difficult to extract its parts. In this regard, future research should interest in solving such problems.




REFERENCES

- [1] S. Badreddine and H. Mouloud, "Urban seismic vulnerability assessment: application to the city of Constantine," (in French) *ACM Transactions Asian Lang. Inf. Process.*, vol. 8, pp. 1-10, 2013.
- [2] M. Mahmoud, A. Shquier, and K. M. Al-howiti, "Fully automated arabic to English machine translation system : Transfer-based approach of AE-TBMT," *International Journal of Information and Communication Technology*, vol. 10, no. 4, pp. 376–391, 2017, doi: 10.1504/IJICT.2017.084341.
- [3] W. Etaiwi, A. Awajan, and D. Suleiman, "Statistical Arabic name entity recognition approaches: A survey," *Procedia Computer Science*, vol. 113, pp. 57–64, 2017, doi: 10.1016/j.procs.2017.
- [4] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, vol. 6, no. 1, pp. 1-38, 2019, doi: 10.1186/s40537-019-0254-8.
- [5] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the Web," *Proceedings of the 20th international joint conference on Artificial Intelligence*, 2007, pp. 2670–676.
- [6] K. Gashteovski, S. Wanner, S. Hertling, S. Broscheit, and R. Gemulla, "OPIEC: an open information extraction corpus," *arXiv Prepr. arXiv*, 2019.
- [7] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: an overview," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 5, pp. 497-507, 2019, doi: 10.1016/j.jksuci.2019.02.006.
- [8] S. Mohamed, M. Hussien, and H. M. Mousa, "ADPBC: Arabic dependency parsing based corpora for information extraction," *International Journal Information Technology and Computer Science.*, pp. 54-61, 2021, doi: 10.5815/ijitcs.2021.01.04.
- [9] F. Ben Mesmia, K. Haddar, N. Friburger, and D. Maurel, "CasANER: Arabic named entity recognition tool," *Studies in Computational Intelligence.*, vol. 740, pp. 173–198, 2018, doi: 10.1007/978-3-319-67056-0_10.
- [10] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A survey on open information extraction," *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August 20-26, 2018, pp. 3866–3878.
- [11] K. Darwish and W. Gao, "Simple effective microblog named entity recognition: Arabic as an example," in *Lrec*, pp. 2513–2517, 2014.
- [12] C. Sabty, M. Elmahdy, and S. Abdennadher, "Arabic named entity recognition using word representations," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 14, no. 8, pp. 956–965, 2016.
- [13] S. M. A. El-salam, E. M. F. El Houby, and E. Division, "Extracting Arabic relations from the web," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 8, no. 1, pp. 85–102, 2016, doi: 10.5121/ijcsit.2016.8107
- [14] M. Fasha, N. Obeid, and B. Hammo, "A proposed model for extracting information from Arabic-based controlled text domains," *arXiv Prepr. arXiv*, 2017.
- [15] D. S. Batista, "Large-scale semantic relationship extraction for information discovery," Ph.D. Thesis, Instituto Superior Técnico, Lisbon, Portugal, 2016.
- [16] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1535-1545.
- [17] L. Qiu and Y. Zhang, "ZORE: A Syntax-based System for Chinese Open Relation Extraction," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1870-1880.
- [18] D. Vo and E. Bagheri, "Open information extraction," *Encyclopedia with Semantic Computing and Robotic Intelligence*, vol. 1, no. 1, 2016, doi: 10.1142/S2425038416300032.
- [19] P. Gamallo, "An overview of open information extraction," *3rd Symposium on Languages, Applications and Technologies (SLATE'14)*, pp. 13–16, 2014, doi: 10.4230/OASfcs.SLATE.2014.13.




- [20] A. Bassa, "GerIE: open information extraction for German texts," M.S. thesis, Knowledge Technologies Institute, Graz Univ. Technol., Graz, Austria, 2016.
- [21] S. Jia, M. Li, and Y. Xiang, "Chinese open relation extraction and knowledge base establishment," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 17, no. 3, pp. 1–22, 2018, doi: 10.1145/3162077.
- [22] D. Truong, D. T. Vo, and U. T. Nguyen, "Vietnamese open information extraction," *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 2017, pp. 135–142, doi: 10.1145/3155133.3155171.
- [23] S. Khoja, "APT : Arabic part-of-speech tagger," *Proceedings of the Student Workshop at NAACL*, 2001, pp. 20–25.
- [24] O. Lyashevskaya and I. Panteleeva, "Automatic dependency parsing of a learner english corpus realec," *Higher School of Economics Research Paper No. WP BRP 62/LNG/2017*, 2018.
- [25] M. Shatnawi and B. Belkhouche, "Parse trees of Arabic sentences using the natural language toolkit," *International Journal of Speech Technology*, 2015.
- [26] M. Dragoni, M. Federici, and A. Rexha, "An unsupervised aspect extraction strategy for monitoring real-time reviews stream," *Inf. Process. Manag.*, vol. 56, no. 3, pp. 1103–1118, 2019, doi: 10.1016/j.ipm.2018.04.010.
- [27] S. Buchholz and E. Marsi, "CoNLL-X shared task on multilingual dependency parsing," *Proc. Tenth Conf. Comput. Nat. Lang. Learn.*, 2006, pp. 149–164.
- [28] M. Straka, J. Hajič, and J. Straková, "UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing," *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4290–4297.
- [29] D. Taji, N. Habash, D. Zeman, F. Albogamy, and A. Ramsay, "Universal dependencies for Arabic," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2017.
- [30] M. Diab, N. Habash, O. Rambow, and R. Roth, "LDC Arabic treebanks and associated corpora: data divisions manual," Center for Computational Learning Systems - CCLS, Columbia University, New York, NY, USA, No. CCLS-13-02, 2013.
- [31] D. Kondratyuk and M. Straka, "75 languages, 1 model: parsing universal dependencies universally," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019, pp. 2779–2795.
- [32] Y. Zhang, Y. Chen, Q. Li, J. Han, and X. Wang, "Open information extraction with meta-pattern discovery in biomedical literature," *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018, pp. 291–300, doi: 10.1145/3233547.3233594.
- [33] S. El-Morsy, "Implementation-of-Arabic-open-information-extraction" Github.com. <https://github.com/salsama/Implementation-of-arabic-open-information-extraction> (accessed Dec. 1, 2020).
- [34] A. R. Mohamed, "Syntactic treebank built and employed within the framework of artificial intelligence techniques," *King Abdullah Bin Abdul Aziz Int. Cent. Arab. Lang.*, 2017.
- [35] Q. Zhu, X. Ren, J. Shang, Y. Zhang, A. El-Kishky, and J. Han, "Integrating local context and global cohesiveness for open information extraction," *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 42–50, doi: 10.1145/3289600.3291030.

BIOGRAPHIES OF AUTHORS






Sally Mohamed Ali El-Morsy    received BSc and MSc in computer and automatic control engineering from Tanta 2004 and 2014, respectively. Her main research interest includes Data Mining, Machine Learning and Embedding Network. She can be contacted at email: smbm222@yahoo.com.



Mahmoud Hussein    received his BSc. and MSc. in Computer Science from Menoufia University, Faculty of Computers and Information in 2006 and 2009 respectively and received his PhD in Software Engineering from Swinburne University of Technology, Faculty of Information and Communications Technology in 2013. His research interest includes Software Engineering, Data Mining, Machine Learning, Data Privacy, and Security. He can be contacted at email: mahmoud.hussein@ci.menofia.edu.eg.



Hamdy M. Mousa    received the B.S. and M.S. in Electronic Engineering and Automatic control and measurements from Menoufia University, Faculty of Electronic Engineering in 1991 and 2002, respectively and received his Ph.D. in Automatic control and measurements Engineering (Artificial intelligent) from Menoufia University, Faculty of Electronic Engineering in 2007. His research interest includes intelligent systems, Natural Language Processing, privacy, security, embedded systems, GSP applications, intelligent agent, Bioinformatics, Robotics. He can be contacted at email: hamdimmm@hotmail.com.