Al-Azhar Bulletin of Science Vol. 31, No. 2, (December) 2020, pp 31-38

AN ONTOLOGY-BASED NAME ENTITY RECOGNITION NER AND NLP SYSTEMS IN ARABIC STORYTELLING

Marwa Elgamal[1*], Mohamed Taha Abu-Kresha[2], Reda Abo Alez[3], Salwa Hamada[4]

[1,3] *Systems and Computer Department, Faculty of Engineering, Al-Azhar University, Cairo, Egypt*
[2] *Mathematics Department, Faculty of Science, (Boys) Al-Azhar University, Cairo, Egypt*
[4] *Electronics Research Institute, Cairo, Egypt*

*Corresponding Author: eng.marwa.m.m@gmail.com

ABSTRACT

Ontology is a descriptive model representing domain knowledge with robust specifications that solve interoperability between humans and machines. In this work, a practical methodology presented for Arabic Storytelling ontology construction for domain ontology extraction from unstructured Arabic story documents. However, the manual construction of ontologies is a time-consuming and challenging process. Still, ontology construction and learning, which extracts ontological knowledge from various data types automatically or semi-automatically, can overcome the bottleneck of knowledge acquisition. This paper intends to investigate the problem of automatically construct and build an Arabic storytelling ontology based on Arabic named entity recognition (NER) from unstructured story text. This paper presents a system designed based on Machine Learning (ML) approach. The system framework is a combination of five main stages: The first stage determines the requirement analysis—second document pre-processing using NLP tasks. The third is Conceptualization. The fourth stage is formal design and construction, and the final step is evaluation.

Keywords: Ontology Construction, Ontology design, Natural Language Processing (NLP), Name Entity Recognition (NER), Semantic Web. Ontology Web Language (OWL).

1. INTRODUCTION

Ontology construction from unstructured textual data is a primary task for building formal ontologies related to a knowledge domain. Ontologies are essential in knowledge representation and linked data that support information retrieval by linking the contextual data that is semantically related by the semantic web. Our aim in this study is to automate the process of extracting the knowledge representation from Arabic story text. Choosing ontology web language (OWL) as a knowledge representation allows us to automatically visualize and analyze the linked data entities' semantics [1]. In this paper, we focus on Arabic OL, to automate the construction and designing of ontology from Arabic stories Datasets. The Arabic Language does not have a perception of capitalization, which makes identifying proper names is very hard, our contribution in this paper has overcome this problem by using Arabic named entity recognition (NER) approach to extract the terms and concepts from the contextual story text that we have from our storytelling corpora [2]. Many ontology tools, like OntoGen [3], Protégé [4], are existing to support the ontologies building, but the automatic ontology construction automatically method still needs research effort studies In the Arabic Language. The advantage of building ontologies from Arabic stories texts concerns in the data semantic linking in the designed storytelling ontology. In this work, a framework for automatic ontology construction will perform and how to extract terms and concepts by named entities recognition approach using natural language processing (NLP) tasks that

can help in Conceptualization and remove the relationships between these concepts.

Aim Of The Study

Information extraction of Arabic concepts based on NER then extracting the semantic relations between them leads to a knowledge representation model. The terms and relations prepared in a hierarchical structure. Ontologies and the semantic web have created a resource of representing domain-specific conceptual knowledge that can exist to facilitate the Arabic storytelling system's semantic capabilities. This ontology used to capture linked experience in ordinarily way and provide an accepted understanding in the field of study. To allow storytelling reasoning, we designed a storytelling ontology to be an essential element in this process as it links all the necessary Arabic story elements of such a case and supports an automated storytelling process. This paper's main contribution is the automation of ontology construction and automatically analysing the semantics of the linked data entities in the storytelling ontology model. The constructed model's design integrated with Protégé, an ontology development environment to build the storytelling model. The paper prepared as follows: section 2, Related work, then Methodology in section 3, and the implementation of the Methodology is discussed in "Implementation of storytelling ontology" in section 4. section 5 Ontology evaluation. The final part is the "Conclusion."

2. Related Work

Ontology in computer science concerns in organizing the data into information and relations between them in a set of classes and subclasses. The interpretation rules for knowledge representation and information gathering and axioms that produce new definitions through reasoning engines is an essential component of ontologies. According to W3C (World Wide Web Consortium) "Ontologies define the terms used to describe and represent an area of Knowledge" [5]. Zouaq defend the components of an ontology by the following tuple: O = < C, H, R, A > where O represents ontology, C represents a set of classes (concepts), H represents a set of hierarchical links between the concepts (taxonomic relations), R represents the set of conceptual links (non-taxonomic relations), and A represents the set of rules and axioms [6]. Ontology construction can be defined as an iterative process of creating an ontology from scratch or reusing an existing ontology for enriching or populating [7]. Many implementations of methodologies for efficient ontology automatic and semi-automatic construction aims to create ontologies are existing. Gómez-Pérez [8] define ontology construction as an ordered series of phases that postulate the procedures used in the engineering of an ontology. Lee et al. [9] presented approaches for Chinese text processing for news summarization in Chinees language. Ghneim et al. [10] proposed a framework for Arabic Ontology Learning (ArOntoLearn). The Probabilistic Ontology Model (POM) used in the (Text2Onto) framework to represent the extracted ontology [11]. Their framework acquires new concepts and relations using Lexico-syntactical patterns. Weber and Buitelaar et al. [12] proposed an OL system that uses Wikipedia to create domain ontology, then applies ontology based NER in a given corpus. Then it applies syntactic patterns analysis to recognize new entities. Automated ontology construction systems, such as Text2Onto [11], can extract non-taxonomic relations between concepts using constructed rules and regular expressions. The problems with those tools are their restriction in the efficiency of extracting domain-specific images in the Arabic Language because they identify semantic relations based on Part-Of-Speech (POS) tags only. Most of the researches in Arabic Ontology Learning approaches lack the full extent of Arabic ontology construction.

3. Storytelling Ontology Construction Methodology

3.1 Ontology Requirements

The first phase in ontology construction is specifying the requirement. These requirements are the scope of the storytelling ontology because it helps in ontology domain characterization. The definition of possible scenarios, events, characters, locations, Story types, and story duration time is necessary at this ontology. For ontology construction, the following task performed first; extracting the information from Arabic text based on NER. Second; Specify the domain of the ontology by asking a list of questions that will Specify the knowledge domain and knowledge representation that we want our ontology to answer. Final part; gathering information to produce a list of specialization questions created by a group of domain experts to articulate direct questions that they expect the automatic ontology system to answer once when implemented.

3.2. Conceptualization

Conceptualization is the second critical phase in ontology building and construction that consists of choosing, structuring, and defining the conceptual elements of the ontology knowledge domain model. In this paper, we will use the Arabic NER method [2] to overcome the linguistic difficulties in Arabic Language and to be able to extract the concepts (classes) from name entities like person names, location, time, which are essential in our storytelling ontology. The extracted concepts by NER approach will help in the understanding of the domain knowledge.

3.3. Relation Extraction

The third phase is relation extraction. There are three main approaches for learning ontologies and relations between concepts in the text: Linguistic approach, Statistical approach, and Hybrid approaches. In our paper, we used the Hybrid approach, which combines machine learning with linguistic pre-processing techniques with NLP tasks [13] to extract relations between concepts from the Arabic story text. NLP tasks like data pre-processing with tokenization, POS tagging, word stemming, and data cleaning by stop words removing and punctuation removing are essential tasks in storytelling ontology construction. We set our automatic approach according to the type of knowledge resources which are unstructured dataset in our corpora that is related to Arabic textual stories content. Storytelling ontology construction will be produced from scratch.

3.4. Ontology Construction

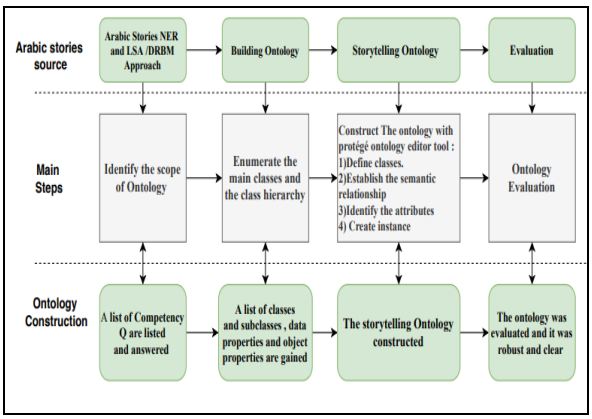

Fig.1. Storytelling Ontology

The fourth stage of the Methodology's objective is integrating all classes into an ontology system with coding all modules by Protégé editor with a standard language. The recent progress in formal ontology is ontology web language OWL. For designing our ontology, we need to: First, Implement the ontology module; we used Protégé development software, an open-source ontology editor used in ontology construction and development. Then, reasoning tasks of taxonomy classification and computing inferred types. Then, Defining the semantic relations between classes and concepts in the ontology

4. Implementation

The objective in this paper is to design and develop an ontology-based on Arabic Named Entity Recognition for the Arabic storytelling system to extract useful textual data from the

Arabic story document. We build our dataset corpora containing 32 Arabic stories with different types (educational, Entertainment, Islamic, and comedian) for children from ages four to twelve. The main task of the designed ontology is to represent the knowledge in the Arabic stories. The knowledge representation must contain the story type, the characters, the location, and the events. To build our system, we need to overcome the difficulties in Arabic linguistic processing, so we used natural language processing NLP tasks to pre-processing stories dataset corpus. We used the Arabic name entity recognition method to extract the concepts from stories and to extract the relations between concepts. We used Python36 with the NLTK package [14] in the pre-processing and NER phases. We build our ontology from the beginning with Protégé. Figure 1 shows the architecture for proposed a case study in the Arabic storytelling Ontology system; it will be described briefly below. The ontology model aims at supporting the automatic Arabic story knowledge representation to review the story type, and represent the stories important parts like characters, location, and events.

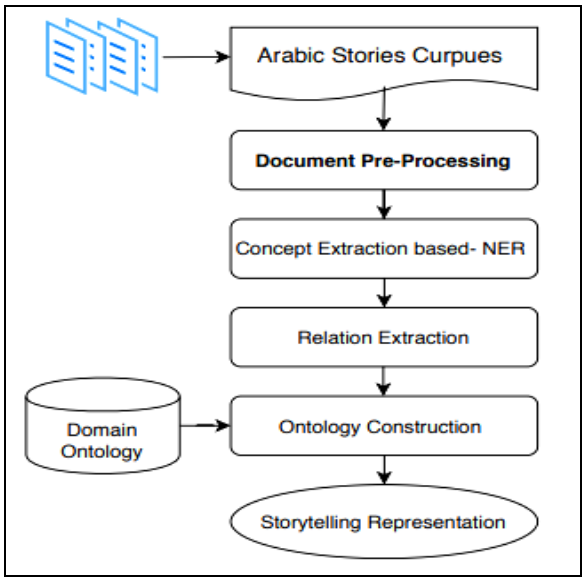

Fig.2. Arabic Storytelling Ontology Architecture.

As a result, defining the textual stories essential components, the next list of specialization questions was defined:

What type of the story?

What are the events and scenarios do exist in the story?

Whom are the characters play different acts in the story?

Where are the location and places that story happened?

When did the story events happen?

We want to extract this information automatically from any written Arabic story, so the ontology construction that we designed will help in representing these pieces of knowledge and data that are semantically related.

In the second phase, we used the natural language processing (NLP) approach in Arabic text pre-processing applications. The basic NLP tasks include data cleaning then tokenization, by dividing each sentence into words (tokens), part of speech (POS) tagging, stop words and punctuation removing, chunking, and stemming. Those tasks for Arabic languages are provided by the NLP package called NLTK in python36.

Conceptualization phase concerns in concept recognition from textual data for building our ontology. It can be performed in various ways; we used the Term frequency term-document frequency algorithm (TF-IDF) to extract the terms or Concepts that will be the class's name in storytelling ontology. We used Automatic concept extraction based on named entity recognition (NER) method to extract the Arabic name entities and concepts in the textual Arabic story document (person names, location names, and dates). Arabic names extraction from text is a challenging task because there is no capitalization letter before nouns. To overcome this problem, we extract the concepts with the NER approach.

In this phase, we build our Dataset corpora from Arabic children stories with annotated names with BOI.

Moreover, the NER-based Machine Learning ML methods evaluation gives us high accuracy in our dataset.

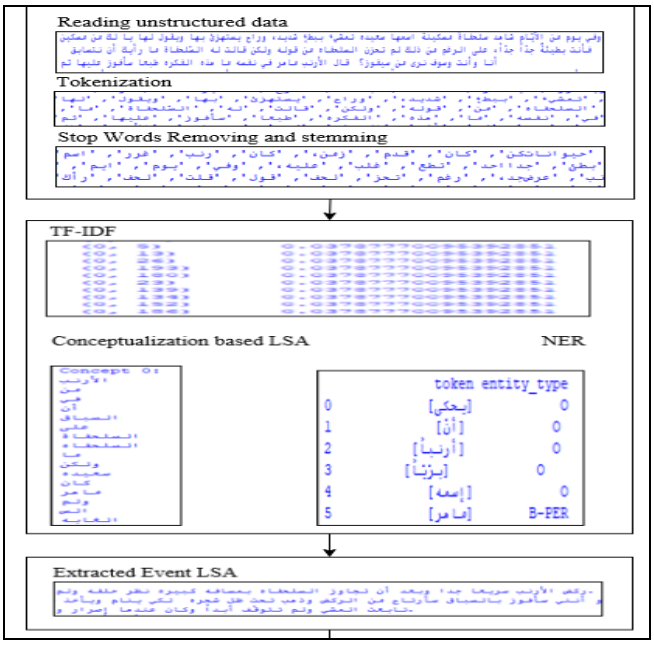

Fig. 3. Arabic NLP storytelling methodology

We divide our dataset into 70% percentage training data and 30% for testing to execute different ML methods to find the best learning evaluate leaning methods with to calculate the accuracy, precious, recall, and F-measure the next table shows the results to classify the Persons beginning name B-PER class entity.

Table [1]: Ontology-based- NER evaluation

Algorithm	Accuracy	Precision	Recall	F-measure
Zero	92%	0.92	0.46	0.96
Naïve Bayesian	93%	0.93	1	0.96
Bayes Net	95%	0.91	0.26	0.41
LibSVM	92%	0.91	0.96	0.96

After extracting the concepts, the relations between them must then be generated to drive the concept hierarchy. Recently, many relation extractions approaches have been proposed that focus on the task of ontology development like learning, extension, and population. These approaches aim to learn taxonomic relations between concepts. In our study, we decide to use clustering methods based on Latent Semantic Analysis (LSA) [15] we aim to organize terms in a hierarchy that can be transformed directly into an ontology, the clustering algorithm with LSA based on the semantic relations and the words which are semantically related. These words will occur closer together in an ontology than those who are less strongly associated. We use the TF-IDF to extract the relations between concepts to help position terms correctly in an ontology and for term disambiguation [16].

4.1.2. *Ontology Construction*

In this phase, the construction method involves the earlier construction of knowledge bases in related domains that we drive and extracted in the previous stages. Validation of the main storytelling ontology concepts decided for Arabic storytelling knowledge representation. First, we introduce the set of concepts (classes) For the domain of knowledge. Second, Ontology Modules identification consists of defining the set of ontologies with classes and subclasses that will conform to the whole ontology system [17]. To implement the ontologies, group similar terms, and concepts that share the same data, which are semantically related. The ontology modules identification formally stated using Description Logics (DL) notation [18]. Story Type ontology defined to represent the different types of any giving story. Story type class consists of subclasses like the educational story, entertainment story, children Islamic story. Figure 3 shows the story type of class hierarchy.

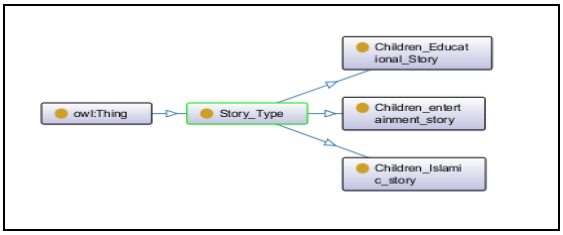

Fig.4. Story types of the class hierarchy

Person class is an ontology that defines the different persons relates to the story. The subclasses, like story characters, actors, writers, and editors, each person has a data property like names, type, and age—location class, which defines the locations and places where the events and scenarios have been placed. The person's names and location are extracted in the concept recognition phase by the NER approach. The Arabic NER extracts the Arabic Persons' names and Location names.

Date class is the ontology that represents the times and dates that are described in the story. This information is essential in our storytelling ontology because it means whether the story was historical or recent. Moreover, this information is necessary in our case study.

The final class is designed for the story's events. The story events and scenario extracted by the clustering algorithm will be defined, and the data are semantically related to previous classes, and this will help in linking data via semantic web to represent the needed knowledge automatically from a given story. For the Arabic storytelling ontology construction, the previous ontologies modules are integrated to build the overall storytelling ontology. The constructed ontology contains the classes, the subclasses, and the object properties and data properties. Each class has its individuals and instance and axioms.

The next figure shows the designed ontology.

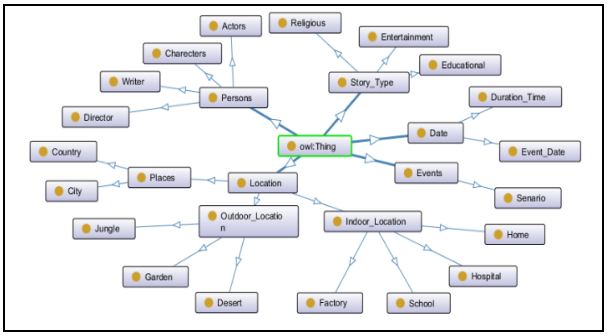

Fig .5. Arabic storytelling ontology

4.1.3. Ontology Learning:

In ontology learning to information extraction like concepts and instance for the storytelling ontology, constructed ontology will help discover such relationships. After building our ontology, an OWL file can be generated and uploaded on the web to automate the information extraction and knowledge representation from ontology. The procedure of instance, learning, and classification are represented as follows. First, we build our ontology based on the linguistic and statistical methods; then, we add the object properties, data properties, individuals, and concepts hierarchy. Second, The OWL file populated on a web page the classes and subclasses are defined in the web page as a URL file. Third, the ontology concepts that are semantically related are linked via the semantic web. Finally, the quiring method is used to automate the information extraction from ontology from the semantic web because all the linked data in the designed ontology are semantically related. The next figure shows an example of the linked data in our developed Storytelling ontology.

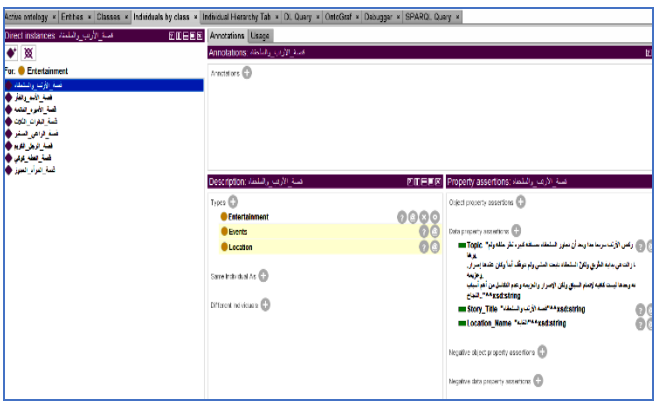

Fig.6. Storytelling Learning Ontology

To generate the story knowledge representation from constructed ontology, we defined different classes like *story title, Story locations, Events, Date*, and *Persons*. Each class contains subclasses, like in *Person class*, the subclasses are *Actors, Characters, and writers*. To define an instance like *Entertainment* in *story type* subclass, we describe the object properties, which is the relations between the concepts.

We define the data properties extracted from NER for persons names and location names.

Then we add the Event in data properties that define the relations between instances. The Event is extracted by clustering algorithm based LSA and Deep learning based RBM; then, it added in the ontology to describe the Axiom. After reasoning our designed ontology for Arabic storytelling, the ontology was Coherence and built correctly—an example represented in figure five.

5. Ontology Evaluation

Ontology Evaluation states the correct construction of the content in the ontology, certifying that its descriptions of the ontology are correctly implemented according to the requirements and competency questions. Gómez-Pérez (1994). Two critical issues are used for evaluation: the clarity of the ontology and the quality requirements.

Clarity, the ontology concepts were implemented using formal axioms; in storytelling, ontology was clear and defined by story type, Persons, Locations, Date, and events.

Coherence, Coherence was verified by performing the reasoning tasks of reliability checking. In our case study, the ontology was Coherence. The constructed ontology can be exported as an OWL/XML file, and the linked data in the story ontology can be extracted automatically by SPARQL queries in the semantic web.

6. Conclusion and future work

In this paper, the development of a practical knowledge graph and information modelling-based ontology for Arabic stories is presented. The dataset contains a corpus of Arabic children's stories containing stories in different types like educational and Entertainment stories. The concepts and the semantic relations in Arabic stories are extracted with NLP applications. These extracted information like Characters, Locations, Date, events and Scenario classes are defined and modelled in the ontology and knowledge graph. The Methodology in this paper provides a solution in ontology construction in Arabic storytelling domain knowledge to improve the automatic storytelling representation based on the designed ontology. To overcome the difficulties in Conceptualization in the data written in the Arabic Language, we developed a system with NER approach to extract the concepts from Arabic stories documents. We execute different ML methods for learning evaluation, and we found that BayesNet gave the best learning accuracy in names recognition, the accuracy

was 95%, precious 91%, and F-measure 41%. To extract the relations between these concepts, we used the clustering algorithm based in LSA to extract the data that are semantically related in the Arabic text. The designed ontology module is robust, coherent, and precise.

We build the OWL file published in the semantics web W3C. In future work; we aim to enrich the storytelling ontology with more classes like animated cartoon class that links the stories text with different multimedia objects like video, audio, and AR.

REFERENCES

[1] Ontology Web Language (OWL).

https://en.wikipedia.org/wiki/Web_Ontology_Language

[2] Khaled Shaalan, “A Survey of Arabic Named Entity Recognition and Classification," School of Informatics, University of Edinburgh, UK, (2012).

[3] F. Blaž, G. Marko, M. Dunja, "OntoGen: Semi-automatic Ontology Editor," Book (2007).

[4] Protége, Stanford, CA, USA (2020).

[5] Mishra S, Jain S, "A study of various approaches and tools on ontology". In: 2015 IEEE international conference on computational intelligence and communication technology (CICT), pp 57–61, (2015).

[6] Zouaq A, "An overview of shallow and deep natural language processing for ontology learning", Inf System 36:1064– 1081, (2011).

[7] Fatima N., Huah Y. "Automatic ontology construction from text: a review from shallow to deep learning trend", Artificial Intelligence Review, 53:3901–3928, (2020).

[8] Fernández M., Gómez-Pérez A.; Juristo N., "Methodologies: From Ontological Art Towards Ontological Engineering. Symposium on Ontological Engineering of AAAI. Stanford (1997).

[9] Chang-Lee, Mei Wang, et al., "Automated ontology construction for unstructured text documents," Science direct journal, Data & Knowledge Engineering 547–566 (2007).

[10] Ghneim, et al., "Building a Framework for Arabic Ontology Learning", The International Business Information Management Conference (13th IBIMA), (2009).

[11] P. Cimiano, J. Volker, "Text2onto – a framework for ontology learning and data-driven change discovery," 2nd European Semantic Web Conference (ESWC'05), (2005).

[12]N. Weber, P. Buitelaar, "Web-based Ontology Learning with ISOLDE", Language Technology Lab Saarbrücken, Germany, (2006).

[13] V. Gurusamy, et al., " Preprocessing Techniques for Text Mining," Conference paper, RTRICS, (2014).

[14] NLTK, Natural Language Toolkit.

https://www.nltk.org/

[15] K. Al-Sabahi1, Z. Zhang1 et al., "An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization," Arabian Journal for Science and Engineering, 43:8079–8094, (2018).

[16] Y. Yang, et. al, "Text Features Extraction based on TF-IDF Associating Semantic", IEEE 4th International Conference on Computer and Communications (ICCC), (2018).

[17] T.R.Gruber, "A Translation Approach to Portable Ontology Specifications," Knowledge Systems Laboratory, Technical Report KSL 92-71,(1993).

[18] J. Lamy, "The great table of Description Logics and formal ontology notations," the university of Paris, (2018).

بناء نظام اللأنطلوجيا الدلالية القائم علي التعرف علي كيان الاسم وأنظمة معالجة اللغة الطبيعية في رواية القصص العربية

مروه الجمل [1], رضا أبو العز [2] , محمد طه أبو كريشة [3], سلوى حماده [4]

[1,2] قسم هندسة النظم والحاسبات (بنين), كلية الهندسة , جامعة الأزهر
[3] قسم الرياضيات ,كلية العلوم (بنين) ,جامعة الأزهر
[4] معهد بحوث الإلكترونيات , الدقي , القاهرة

الملخص العربي

علم الوجود (الأنطولوجيا) في علوم الكمبيوتر هو نموذج وصفي يمثل نمذجة البيانات والمفاهيم المستخرجة من النصوص و ربطها بعلاقات ممثلة في شبكة دلالية بمواصفات قوية تحل إمكانية االتفاعل بين البشر والآلات لإستنباط البيانات. في هذه الدراسة ، تم تقديم منهجية عملية لبناء أنطولوجيا سرد القصص لاستخراج البيانات اليا من وثائق القصص المكتوبة باللغة العربية. إن بناء ونمذجة البيانات بالأنطولوجيا لإستخراج المعرفة اوالعلاقات الدلالية من أنواع البيانات المختلفة أوتوماتيكيا يمكن أن يتغلب على الصعوبة في اكتساب المعرفة . تهدف هذه الدراسة إلى التحقيق في مشكلة الإنشاء والبناء تلقائيًا لأنطولوجيا سرد القصص استنادًا إلى البيانات المستخرجة بطريقة التعرف على الكيانات والمفاهيم المكتوبة باللغة العربية (NER) وإستنباط العلاقات الدلايلة بين المفاهيم من نص القصص غير المنظم. كما تقدم هذه الدراسة نظامًا مصممًا على أساس نهج التعلم الآلي والذكاء الإصطناعي. يتكون إطار عمل النظام من خمس مراحل رئيسية: المرحلة الأولى تحديد نطاق الانطولوجيا و تحليل المتطلبات .المرحلة الثانية معالجة النصوص العربية باستخدام مهام معالجة اللغات الطبيعية. ثالثا تحديد المفاهيم والعلاقات الدلالية وتعريف خصائص الموضوعات والبيانات المراد عمل نموذج لها. المرحلة الرابعة هي التصميم و والبناء بإستخدام محرر protégé، والمرحلة النهائية هي التقييم للنظام المقترح. وقد أثبتت الحالة الدراسية أن النظام المقترح لنمذجة القصص العربية واضح و فعال ومنطقي لإستنباط البيانات المختلفة من القصص العربية.