IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 4, August, 2024, pp. 1261-1273
Journal homepage: http://iieta.org/journals/isi

Enhancing Spatial Information Extraction from Arabic Text: A Hybrid Approach with Ontology and Rule-Based

Atmane Hadji[1,2*], Mohammed-Khireddine Kholladi[3,4], Nadezhda Borisova[5]

[1] Department of Computer Science, Faculty of Exact Sciences, University of Bejaia, Bejaia 06000, Algeria
[2] Department of Computer Science, Institute of Mathematics and Computer Science, University Center A. Boussouf Mila, Mila 43000, Algeria
[3] Department of Computer Science, Echahid Hamma Lakhdar University of El Oued, El-Oued 39000, Algeria
[4] MISC Laboratory of Abdelhamid Mehri Constantine 2, University of Constantine, New City Ali Mendjeli 25016, Algeria
[5] Department of Informatics, South-West University "Neofit Rilski", Blagoevgrad 2700, Bulgaria

Corresponding Author Email: a.hadji@centre-univ-mila.dz

https://doi.org/10.18280/isi.290402

Received: 12 January 2024
Revised: 23 July 2024
Accepted: 31 July 2024
Available online: 21 August 2024

Keywords:
automatic information extraction, spatial ontology, jape rules, spatial information, Arabic NLP

ABSTRACT

The abstract presents a new hybrid approach for automatically extracting spatial information from Arabic text documents in geographic information systems. The main objective is to automate and enhance the performance of GIS systems by making certain tasks explicit and improving the resources for Arabic Natural Language Processing (ANLP). The first step of the study involves the construction of a spatial ontology to index, annotate and extract spatial information from Arabic texts. In the subsequent step, JAPE rules are developed and employed to disambiguate and classify different types of spatial information. The evaluation of the proposed system demonstrates promising performance, with a precision rate of 93.8% and a recall rate of 95.2%. Overall, this hybrid approach presents a significant contribution to automating spatial information extraction from Arabic texts, enhancing GIS systems, and improving ANLP resources. The positive experimental results highlight its potential for various practical applications in geographic information retrieval and natural language processing.

1. INTRODUCTION

In recent years, the rapid growth of digital information has highlighted the need for efficient information processing systems, especially for languages that are rich in vocabulary and complex in structure, such as Arabic [1]. Extracting spatial information from raw text has become a crucial research topic in various fields such as Automatic Natural Language Processing (NLP), Information Extraction (IE), Information Retrieval (IR) and Geographic Information Systems (GIS).

Spatial information extraction offers many advantages and can be applied to a variety of domains. It improves the accuracy and relevance of search results, optimises geographic information systems (GIS) [2], enriches geospatial databases, and enhances location-based services (LBS) [3]. It also facilitates decision-making in areas such as urban planning, natural resource management and disaster response. The information extraction process transforms unstructured textual data into structured output by identifying and extracting entities, relationships, semantic roles and events [4].

Due to its complex semantic and morphological features, the Arabic language faces significant challenges in the field of information extraction and retrieval, particularly for spatial information. Existing methods often prove ineffective in meeting these challenges [5, 6]. It is therefore necessary to develop a new approach to improve the accuracy and efficiency of this extraction.

Ontologies appear as an important solution to building a shared and reusable body of knowledge that support human-machine interaction and understanding [7]. They play a crucial role in knowledge representation and contribute to the development of the Semantic Web [8]. However, the exploitation of ontologies for the automatic indexing, retrieval, extraction and annotation of Arabic texts is still little explored. In the context of information extraction and search, the ontology is exploited to: 1) index entities or concepts of a specific domain and their diversities; 2) a hierarchical design to generate retrieval rules by the IE system; 3) the properties of concepts and the relations between these properties that guide the information extraction process; 4) The relationships with the concepts and their properties. The result of the information extraction process should be an updated/improved semantic annotation, model or ontology [9]. In this context, our study aims to answer the question of how to satisfy the needs of users of an Arabic GIS and which techniques to adopt to extract relevant spatial information from Arabic texts. To address these challenges, we propose a hybrid approach combining rule-based and ontology-based methodologies. This approach draws on the strengths of both techniques to deal with the linguistic complexities of Arabic, such as its morphological richness and semantic ambiguities.

The objectives of this study are as follows:

• Develop a hybrid approach that integrates rule-based and ontology-based methodologies for extracting spatial information from Arabic text.

• To construct the Arabic Spatial Toponym Ontology (ASTO) to aid in accurate indexing, annotation, and extraction of spatial data.

• To employ JAPE rules for precise disambiguation and classification of spatial information.

• To evaluate the effectiveness of the proposed approach through rigorous testing and demonstrate its potential to improve GIS application performance.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the hybrid approach and describes the main components of the proposed system. Section 4 discusses the experiments and results. Finally, Section 5 provides the conclusion and perspectives for future work.

2. RELATED WORK

Table 1. Summary of related works

Authors	Input Data	Output Data	Discussion	Year
[10]	Online texts (posts, articles, social media).	Populated Terror Beliefs Ontology (TBO)	-Detection and identification of online posts and personae that espouse radical ideologies.	2024
[11]	Insurance reports describing vehicle damage	Ontology of car damage populated with information extracted from the insurance reports	-The system uses an ontology based on expert knowledge to model car damage and extracts information from insurance reports.	2023
[12]	Text pdf Ontology	Updated Ontology, Graphic results	-OntoHuman is an application that automatically extracts key-value-unit tuples from PDF documents based on ontologies. Allow users to update and enhance the ontologies used. It is designed for use with engineering-related documents and offers an intuitive and collaborative approach to working with ontologies for users.	2022
[13]	Domain Information / Organic Farming	Domain Ontology and information	-This study aims to acquire, store, and provide information on organic farming. To achieve this, Information Extraction (IE) techniques and ontology development are used to create an organic farming extraction system (OBIESOF).	2021
[14]	Medical reports/	Clinical Ontology	-The proposed system involves a combination of IE and creating of an ontology to facilitate the process of extracting and visualizing clinical information.	2020
[15]	Technical documents pdf or html and Generic Ontology	Table extraction	This approach presents an ontology-based method for extracting relevant tables from PDF documents.	2018
[16]	Text and domain ontology	Update ontology	-This approach begins with a domain-specific ontology created by human experts. It then employs techniques to identify and apply heuristics for extracting information from unstructured text, subsequently adding it as structured information to the chosen ontology.	2013
[17]	Message twitter and DBpedia Ontology	Information extraction	-This system integrates Named Entity Recognition (NER) along with a disambiguation module that utilizes syntactical context and a knowledge base such as Freebase	2012
[18]	Text input and ontology predefined/	Information extraction	The main ideas of this approach are the "information extractors," which are components of an IE system that promote the reusability of predefined ontologies, and the "extraction platforms."	2010
[19]	Text and Web pages	Extracts information, expertise base,	-SOBA is a system that performs OBIE from football-related web pages. The purpose of this system is to automatically populate an expertise base that can be used for domain-specific query answering. SOBA can lie in its ability to establish a seamless connection between the ontology, expertise base, and the data extraction component.	2006
[20]	Semantic Web RDFs	Annotated documents Ontology,Information extraction	-KIM framework offers knowledge and information management services, enabling automatic semantic annotation, indexing, and document retrieval. This approach aims to ensure efficiency, facilitate metadata ontology sharing, and promote reusability	2004

Within the semantic-based approach, there are six core techniques: model-driven approach, code-level approach, middleware solution, message interception, rule-based reasoning, and ontology-based solution [21]. In the field of information extraction, various approaches exist in the state of the art, namely: rule-based approaches [22]; learning-based approaches [23]; hybrid approaches that combine these two previous approaches [24]; and ontology-based information extraction (OBIE) [25]. In the following, we will mention some related works in Table 1.

Ontology-Based Information Extraction (OBIE) is a relatively new research area within information extraction, focusing on using ontologies to process semi-structured or unstructured text and extract specific types of information. In

OBIE, the input information can be a document (PDF, text, HTML, XML), an ontology, or a corpus of documents. These inputs are processed using natural language processing (NLP) algorithms and methods, including tokenization, sentence splitting, and part-of-speech (POS) tagging. The use of ontologies in the information extraction process allows for the detection of synonyms, co-references, and relationships between concepts and properties. The output of such a system can be represented as knowledge in various formats, such as RDF (Resource Description Framework), XML, or semantic annotations [26]. Typically, this knowledge representation is organized as an ontology, contributing to the development of the Semantic Web.

Ontology-based named entity extraction, annotation, and information extraction are successfully applied in various domains, including relevant concept extraction. Several related works are based on OBIE, showcasing its potential and applicability.

3. THE PROPOSED APPROACH

In the scientific community, geographic information is defined as a composition of three concepts: spatial, temporal, and thematic information. The main idea is that the combination of these three types of information allows for describing an event that is occurring or has occurred at a specific location and time. Among these components, this paper focuses on the processing of spatial information.

Currently, the integration of ontologies and rules has garnered significant attention in research related to ontologies and the semantic web [5]. The rise of ontologies is noteworthy as they can represent knowledge across various domains. This knowledge aids in semantic search by enhancing accuracy through understanding the purpose and contextual meaning of terms as they appear within the data space. The proposed approach consists of two main parts: first, the construction of the Arabic Spatial Toponym Ontology (ASTO), which represents spatial entities and relations; and second, an automatic information extraction system that leverages the built ontology along with natural language processing techniques and JAPE rules in GATE [27].

The general architecture of the proposed approach (see Figure 1) comprises four phases. In the first phase (Ontology Creation) (Figure 2), concepts are collected, and classes (representing Arabic spatial entities and relations) (Figure 3) are created for use in the text-matching steps. The second phase (Text Processing) applies various modules to annotate named entities within the text. The third phase (Combination and Extraction) utilizes the ontology to match classes, subclasses, or instances with the text (Figure 4). The fourth phase (Disambiguation and Classification) employs JAPE rules, which consist of algorithms developed and implemented in Java (Figure 5).

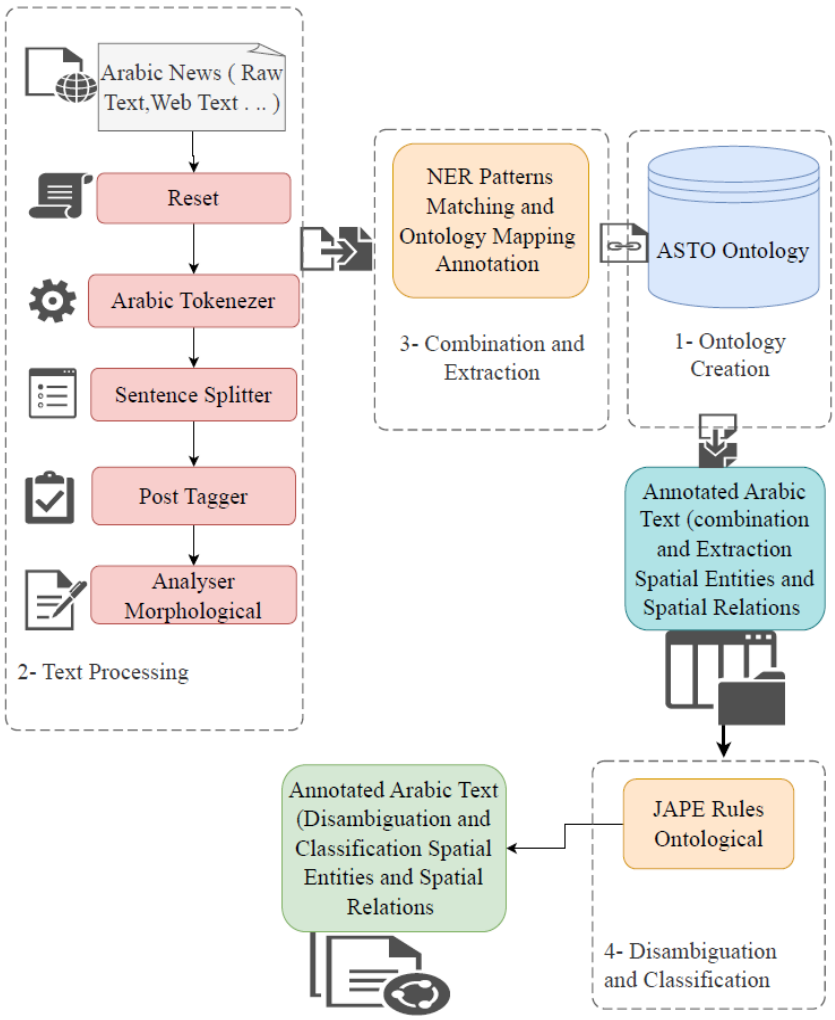

Figure 1. Phases of proposed approach

3.1 Ontology creation (ASTO: Arabic spatial toponym ontology)

Ontologies serve as a crucial solution for developing a set of shared and reusable knowledge that facilitates interaction [28], enabling interoperability and integration of various systems. Automatic extraction of ontological relations from texts is vital for representing documents and their content in a computerized and machine-readable format [29]. Ontology, from a philosophical point of view, is a discipline that deals with the nature and organization of being. With the development of information technology, ontology receives a new definition an Ontology is a formal specification of a shared conceptualization of a domain, it represents concepts using a human-understandable and machine-readable format and it consists of entities, values, relations and axioms [30].

As a conceptual model used for description at the semantic and knowledge levels, ontology is applied across many information processing domains, including knowledge engineering, digital libraries, software reuse, information extraction, and the semantic web [19].

When constructing the ASTO ontology, we follow the methodology outlined by Noy and McGuinness [31], which involves eight stages. The first stage is to define the domain. The next three stages are for constructing the taxonomy and the ontology and providing a formal description. The subsequent four steps include creating classes, conducting the reasoning process, performing a consistency check, and generating the result set (see Figure 2).

The development of an ontology involves iterative refinement: an initial version is constructed, evaluated through applications or expert review, and then refined until a functional ontology is achieved.

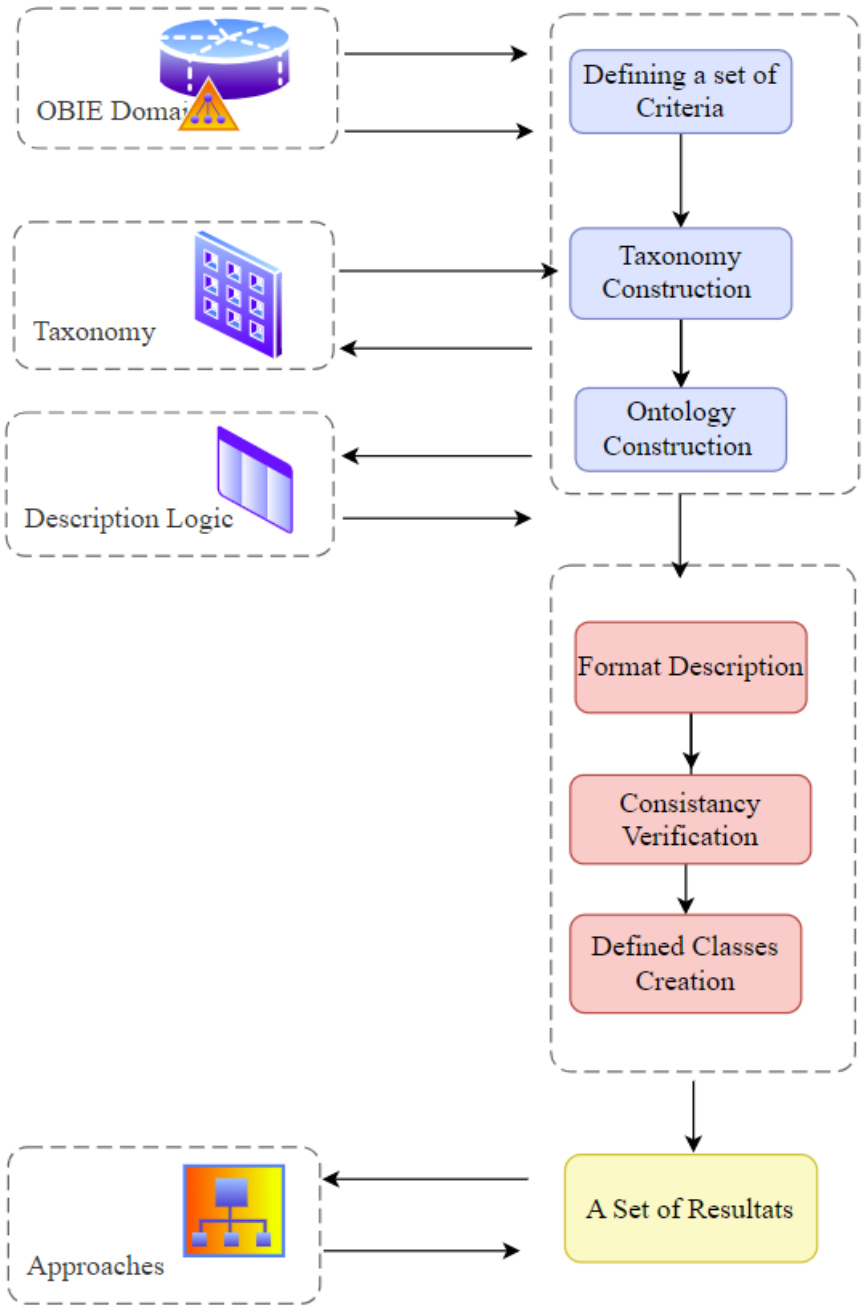

Figure 2. A general procedure of an ontology construction [31]

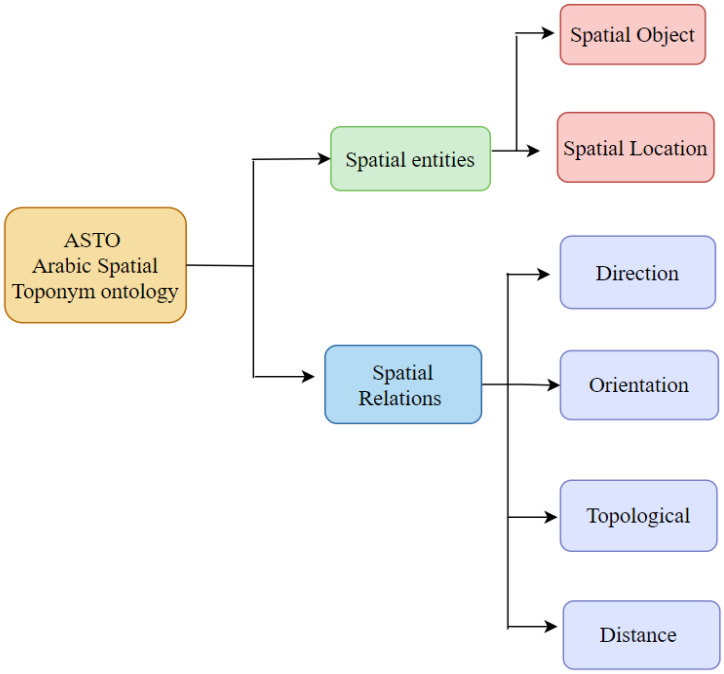

Figure 3. Conceptual Ontology (ASTO) model

```
1. Input: TextRaw T, Ontology  Ont, Sentence S,Word W;
2. Output: Annotated-and-Extracted-Text  CorpusT;
3. Begin
4. Parse T, Read words W from the text;
5.  For each sentence S in T do
6.     For each word  W in S do
7.        For each Class in Onto do
8.           If  Ont.Class == Spatial-Object  Then
9.            While (Ont.Spatial-Object.SubClass <> vide) do
10.                If (W==Ont.Spatial-Object.SubClass.inst) Then
11.                   Annotate (W,Spatial-Object ,CorpusT);
12.                   Extract-in (W,CorpusT , Class-Spatial-Object);
13.                   Onto.Subclass=Onto.Subclass+1;
14.                Endif
15.           End while
16.        Endif
17.         If (Ont.Class == Spatial-Location ) Then
18.           While (Ont.Spatial-location.SubClass <> vide) do
19.                If (W == Onto. spatial-Location.SubClass.inst) Then
20.                   Annotate (W,Spatial-Locaion ,CorpusT) ;
21.                   Extract-in (W,CorpusT, Class-spatial-Location);
22.                   Onto.Subclass = Onto.Subclass+1 ;
23.                   Endif
24.           End while
25.        Endif
26.     EndFor
27.  EndFor
28. EndFor
29. End
```

Figure 4. Pseudo-code for algorithm extraction spatial entities

Initially, as detailed by Konys [25], the process of building an ontology begins with a thorough domain analysis. This involves selecting an appropriate combination of OBIE approaches. Based on this analysis, a final set of characteristics and sub-characteristics is determined, followed by the formation of a class hierarchy. This hierarchy forms the general foundation for constructing taxonomies in existing OBIE systems and tools. During this process, various criteria and sub-criteria are established, and elements such as concepts, relationships, and attributes are integrated from disparate sources into the classification model [32].

Therefore, the taxonomy is based on the construction of the ontology. To implement the ontology, the Web Ontology Language (OWL) is used. OWL provides a formal and structured method for collecting, organizing, and sharing data, along with numerous features for efficient ontology management. The consistency of the created ontology is verified by defining a set of classes and using a reasoning mechanism to validate them and the entire ontology. The accuracy of the results obtained from this process confirms the ontology's overall consistency and coherence. It is important to mention that we have tested this process using various validation queries.

```
// Classification_Spatial_object

Phase: OntoMatching_Spatial_Object
Input: Lookup
Options: control = appelt
Rule: Lookup
(
 {Lookup.class == spatial_object}): Object1
 -->
: Object1.Spatial_Object = {class =: Objct1.Lookup.class,
 inst =: Objct1.Lookup.inst}


// Classification_Spatial_Location

Phase: OntoMatching_Spatial_Location
Input: Lookup
Options: control = appelt
Rule: Lookup
(
 {Lookup.class == spatial_location}): Location1
 -->
: Location1.Spatial_Location = {class =: Location1.Lookup.class,
 inst =: Location1.Lookup.inst}
```

Figure 5. Example of rules JAPE for extraction spatial location and spatial object

3.1.1 Geographic object (spatial entity)

In our model, geographic objects are defined as spatial entities, which can be categorized into multiple geographic types to reflect their physical and natural attributes. Each geographic feature can have multiple names. During the development of the ASTO ontology, several iterations were required for the following reasons: initially, it was unclear whether the collected terms were sufficient to fulfill the ontology's purpose. New terms were added as needed, while unnecessary terms were removed.

As shown in Figure 3, the conceptual model of the Arabic Spatial Toponym Ontology consists of various classes and subclasses, with each subclass containing a set of instances and properties (see Table 2).

3.1.2 Spatial relations

In the presented approach, we focus on spatial relationships. Identifying these relationships is crucial for various tasks, such as reducing ambiguity, enhancing geographic information extraction, optimizing navigation systems, managing traffic, spatial reasoning, and responding to queries.

The identification of spatial relationships is a key element in the development of a new spatial information system. In ontology studies, we define relationships between geographic entities considering topological, orientation, distance, and directional relations, as shown in Table 2. We have developed a spatial ontology tailored to the Arabic language to account for the extensive range of terms and meanings, which improves the efficiency of spatial information extraction and adds originality to the work presented in this research.

3.2 Data sources and preprocessing techniques

To build our corpus of Arabic text, we follow rigorous selection criteria to ensure diversity, relevance, and quality. We selected articles from Arabic newspapers based on the diversity of media sources, the relevance of the content to spatial information, and the publication date to ensure the temporal relevance of the data. The newspapers were also chosen from among the most readable, well-known, and widely followed.

Articles were manually collected from four major Arabic information sources (see Figure 6): Echourouk (https://www.echoroukonline.com) News, Elkhaber (https://www.elkhabar.com) News, Ennahar (https://www.ennaharonline.com) News, and Elmihwer (http://elmihwar.dz/ar) News. The final corpus covers various domains such as human disasters (e.g., accidents, fires), natural disasters (e.g., floods), and technological events. Overall, the corpus totals approximately 9,000 tokens, manually collected from the mentioned websites [33].

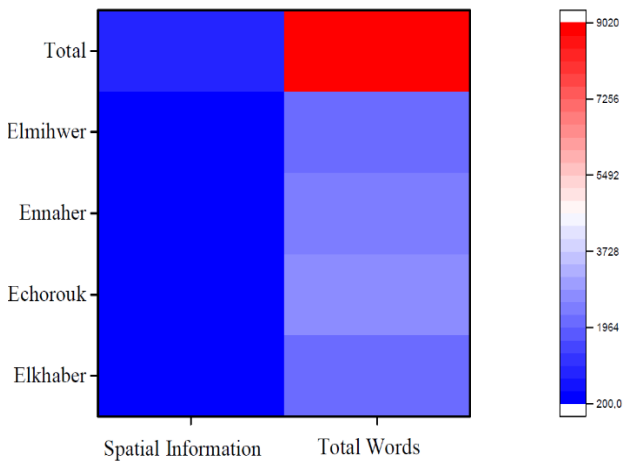

Figure 6. Percentage of spatial information extraction

Table 2. Example of class and instance of ASTO

Class	Subclass	Subclass	Instance
Spatial Entity	Spatial Object	Natural Object	جبل,هضبة,شاطئ,غابة, واد, بحر,صحراء,نهر,
			Mountain, plateau, beach, forest, valley, sea, desert, river,
		Building Object	المدرسة,المستشفى,البناية,العمارة,المسجد,المنزل,الولاية,البلدية,الطريق
			School, hospital, building, architecture, mosque, home, state, municipality, Road….
	Spatial Location	Country	الجزائر,جيجل,ميلة,بجاية,الطاهير,الميلية,فرجيوة
			Algeria, Jijel, Mila, Bejaia, Taher, Milia, Ferdjioua,.....
Spatial Relations	Topological Relations	Inclusion	على ضفة,بعض,جزء,يضع,بين,وسط,داخل,في,
			On the bank, some, part, a few, between, middle, inside, in,...
		Support	على,على مستوى,على محور.....
			on, on a level, on an axis,...
	Direction Relations	Cardinal	شمال,جنوب,شرق,غرب,شمال شرق ...
			North, south, east, west, northeast,....
		Path	نحو, باتجاه, صوب, قصد, عبر, من خلال, حتى...
			Towards, In the direction of, Towards, Intending to, Across, Through, Until ...
	Distance Relations	Quantitative	مسافة,على بعد,تبعد
			distance, at a distance, move away
		Qualitative	قرب,دنو,على قرب,قصيا...
			Near , approach, close up,...
	Orientation Relations	Horizontal	أمام, خلف, قبل, وراء, بعد, يمين,يسار,مقابل.
			Before, behind, before, behind, after, right, left, vs.
		Vertical	فوق, تحت,أعلى,أسفل
			Above ,under ,up ,down.

The preprocessing techniques applied to our corpus of Arabic texts include several crucial steps. We began with data cleaning, removing extra spaces and formatting elements to obtain raw text. Next, the typographical errors were then corrected using both automated and manual tools, followed by text normalization, including removing diacritics, normalizing spaces, and punctuation. Finally, we performed manual tokenization, in which texts were manually divided into meaningful units (sentences or phrases) to ensure accuracy and appropriateness for subsequent analysis.

These cleaning and normalization operations are vital to ensure the texts are properly prepared for thorough and effective analysis.

3.3 Text processing

General Architecture for Text Engineering (GATE) is known for its popularity in text analysis and automatic natural language processing. It has been widely utilized in various projects, including information extraction (IE) in different languages. The main advantage of GATE lies in its extensibility, as it allows the addition of new components or applications to its existing standard components, thereby enabling the expansion of analysis and text processing capabilities. The primary focus of GATE is document annotation, which can be achieved through three methods: fully automatic using applications, manual annotation by users, and semi-automatic annotation through a combination of corpus processing and manual correction or addition of new annotations [34].

GATE offers a range of specialized modules for textual analysis, accessible through a system of plugins. Some of the commonly used modules include tokenizers (segmenters) for breaking text into smaller units, Part of Speech Taggers (morpho-syntactic taggers) for categorizing words, Gazetteers (lexicons) for identifying specific terms, and transducers (JAPE) for pattern-based matching. The named entities extracted by GATE encompass various types such as places, organizations, people, and dates, among others. With GATE, users can create new annotations by loading additional resources like corpora, documents, and texts, as well as incorporating new plug-in. It allows for the combination and parameterization of these resources within the same processing chain [35]. As an open-source Java platform dedicated to text engineering, GATE is well-suited for the development of systems that require natural language processing and information extraction capabilities.

3.3.1 Sentence splitter

In Natural Language Processing (NLP), dividing a text into sentences is a crucial step to extract more meaningful information from it. Each sentence is delimited by identifiers, such as punctuation marks like periods, exclamation marks, or question marks. Initially, a sentence is considered a unit of discourse that is acceptable to the application or NLP system being used. Segmenting a text into sentences allows NLP systems to process and analyze smaller chunks of text at a time, making it easier to understand the context and extract relevant information. By breaking the text into sentences, NLP models can focus on the relationships between words within each sentence, which helps with tasks such as named entity recognition, part-of-speech tagging, sentiment analysis, and other language understanding tasks. Sentence segmentation is an essential preprocessing step in many NLP applications, as it lays the foundation for further analysis and enables the extraction of meaningful insights from a given text.

3.3.2 Arabic tokenizer

The tokenizer is an essential component used to divide text into words, numbers, symbols, spaces, and punctuation marks. Each of these units is referred to as a token, representing a distinct syntactic element, which can be a complete word, part of a word, a multi-word phrase, or a punctuation mark. The tokenizer relies on a predefined list of word delimiters, such as whitespace and punctuation marks, and considers the nuances of tokens within words, especially when dealing with word stems and clitics. In this research, complete words, including stems with or without clitics, as well as numbers, are referred to as main tokens. These main tokens are delimited by whitespace or punctuation marks. Additionally, full-form

words can be divided into sub-tokens by separating clitics from stems [35].

3.3.3 POS tagging

POS tagging is an important step in the process before extraction and recognition of named entities. It identifies and adds tags to the tokenized text model, i.e., identifying nouns, verbs, adjectives and other parts of speech for each token. The POS tagger used in this study with GATE is the Hepple tagger [36].

3.3.4 Analysis morphological

Morphology [36], in linguistics, is the study of the internal structure and composition of words and their formation. A morpheme is defined as the shortest meaningful unit of language, which cannot be divided down into shorter parts. The morphological analyzer helps to group words that express similar concepts. The role of the morphological analyzer in Arabic is to identify the morphemes of a word (stem): the affixes (prefix, infix, and suffix) and the root.

3.4 Combination and extraction

Our hybrid spatial OBIE approach is a contribution to the field of information extraction based on the OBIE ontology. Additionally, ontologies allow formal and declarative descriptions of common terms and support automatic or semi-automatic reasoning about shared data in a domain. The main idea is related to the search and retrieval of spatial information.

The process starts with loading the ontology into GATE as an OWL ontology resource and provides facilities for loading corpora from a URL or file to the Pipeline application. The system in this step matches concepts, instances and relations between them of the ontology with an input Arabic text for extraction and annotation of spatial information. Consequently, if the system detects an identical word in the input text that is a class or instance in the ontology, it annotates it. The result obtained at this stage is an annotated corpus containing the set of annotated words (spatial information) with an ambiguity (see Figure 7), we cannot determine the spatial entities, spatial relations, or classes of each entity.

3.5 Disambiguation and classification

JAPE rules [37], consist of a sequence of phases, each consisting of a sequence of model/action rules. Each phase executes sequentially and is a series of state transducers finished as annotations. The rules consist of two sections: the left-hand section (LHS), which describes an annotation model, and the right-hand section (RHS), which consists of annotation manipulation instructions. The corresponding annotations on the left side of a rule can be identified on the right side by means of labels linked to the model elements.

The main functions of the JAPE transducer [37, 38] in this step are to combine and extract the spatial entities, relations and their classes of the ASTO ontology with input corpus. Then, the system searches into the Ontology rules whether the annotated word is contained (as subclass or instance) in the class matched the LHS part of the JAPE rule and if it found, the system classifies the annotated word in this class.

The following algorithm (Figure 4) shows the extraction of the spatial entities (Object and Location):

3.5.1 Detailed algorithm for algorithm extraction spatial entities

Inputs: **T:** Text row to analyze, Ont**:** Ontology containing classes and subclasses, S**:** Sentence extracted from the text, W**:** Word extracted from the sentence.

Output: **CorpusT:** Annotated and classified text corpus

Detailed Explanation of Steps

Initialization:

• The text **T** is parsed to extract individual words W. This is a preprocessing step that prepares the text for further analysis.

Figure 7. Extraction based Ontology without the use of JAPE rules

Sentence Processing:

• Each sentence S in the text is processed individually.

Word Processing: Each word W in the sentence is examined to determine if it matches a class in the ontology **Ont**.

Spatial Object Classes:

• If the current class in the ontology is an object (spatial-Object), the algorithm checks each subclass to see if the word W is an instance of that subclass.

• If a match is found, the word W is annotated and extracted as a spatial object in the output corpus CorpusT. The subclass index is then incremented to continue checking other subclasses.

Spatial Location Classes:

• Similarly, if the current class is a location (spatial-Location), the algorithm checks each class to see if the word W is an instance of that class.

• If a match is found, the word W is annotated and extracted as a spatial location in the output corpus CorpusT. The class index is then incremented to continue checking other classes.

End of Algorithm:

• The algorithm continues until all the words in all the sentences have been processed and extracted according to the ontology classes Ont. We employed the same algorithm for annotating and extracting each spatial relation (topological, distance, orientation, direction).

For example, the spatial entities transducer employs an explicitly defined JAPE rule that is able to match the object entity (Natural/Building) and location entity models which are listed in the following (Figure 5).

We use JAPE rules for disambiguation and classification of spatial entities and spatial relations in our text corpus. Figure 5 illustrates two distinct phases, each aimed at optimizing the annotation and classification of spatial data.

OntoMatching_Spatial_Object Phase: This first phase focuses on the annotation of spatial objects. The applied JAPE rules check if an annotation matches the spatial_object class defined in our ontology. When this match is found, the annotation is enriched with detailed information, including the class and instance of the object. This process ensures not only precise identification of spatial objects but also effective disambiguation by associating each entity with a clear ontological definition.

OntoMatching_Spatial_Location Phase: The second phase follows a similar process for spatial locations. The JAPE rules check for correspondence with the spatial_location class and enrich the relevant annotations accordingly. This treatment improves the accuracy of spatial locations and ensures coherent and informed annotation.

These two phases enable automated identification and enriched classification of spatial entities, based on a well-defined ontology. We used the same JAPE rules for the classification and disambiguation of each spatial relation (topological, distance, orientation, direction).

4. RESULTS AND EVALUATION

In this section, we outline the experiments conducted to validate the system's effectiveness. We introduce a hybrid automatic ontology-based retrieval method that involves the creation of a new spatial ontology containing various Arabic spatial knowledge elements such as concepts, instances, and relations. Additionally, ontological rules derived from the JAPE grammar are employed to extract spatial entities and spatial relations by matching the ontology concepts with the document corpora.

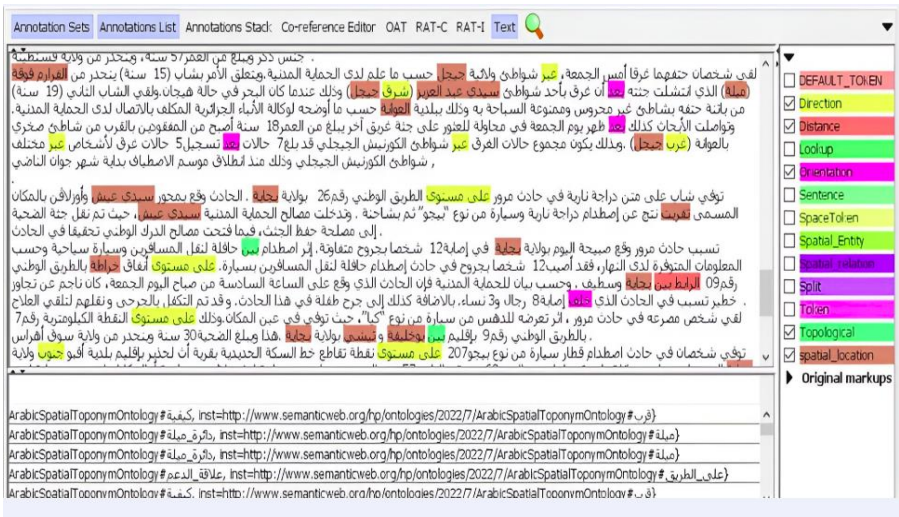

Figure 8. Extraction based Ontology and JAPE rules

Table 3. Number of spatial entity extraction

Newspaper	Spatial Information	Total Words
Elkhaber	216	2072
Echorouk	261	2553
Ennaher	244	2280
Elmihwer	240	2103
Total	961	9008

Table 4. Results of extraction spatial information

Newspaper	Spatial Information	Spatial Entity	Spatial Relations
Elkhaber	216	134	82
Echorouk	261	190	71
Ennaher	244	182	62
Elmihwer	260	165	95
Total	981	671	310

For the initial test, we selected Arabic newspaper texts and compiled a corpus of 9,008 tokens. Within this corpus, 981 words were annotated with spatial information, and 671 spatial entities were identified. In Addition, 310 spatial relations were specifically annotated. Notably, 68.4% of the annotated words with spatial information are spatial entities, and 31.6% of these annotated words actually represent spatial relations. These statistics highlight the richness and diversity of spatial annotations within the corpus, underscoring the complex interconnections between spatial entities and relations. This detailed analysis lays the groundwork for evaluating the effectiveness of the proposed approach in capturing and understanding spatial information in Arabic texts. Details of the corpus annotations are provided in Tables 3 and 4.

The purpose of these experiments was to demonstrate the system's capability to effectively extract spatial information from Arabic texts using the hybrid approach, thus showcasing its potential in information retrieval and natural language processing tasks related to geographic information systems.

As shown in (Figure 8), the annotated words are, for the most part, rich in detailed content of spatial entities or relationships.

4.1 Analysis

This research is a significant contribution to enriching the Arabic language in the domains of digitization, media, information technology, and communication. It aims to enhance the Arabic language's capabilities in automatic natural language processing, particularly in geographic information systems, which have become increasingly essential and widely used across various domains. The majority of events or information nowadays are related to geographical or spatial information, highlighting the significance of this research.

Through the initial experiment involving the processing of our corpus, spatial information was found to constitute (see Figure 9) approximately 10% of the total information. Within the spatial entities, place names and other spatial objects were distributed nearly equally at 33% and 36%, respectively.

Regarding spatial relationships, they were distributed as follows: 15% for direction relations, 9% for orientation, 5% for topology, and 3% for distance relations. These findings shed light on the prevalence and importance of spatial information in Arabic texts and emphasize the potential for further advancements in the field of automatic natural language processing, especially in the context of geographic information systems**.**

To our experience from the set of processed corpora in Arabic text, the distribution or dispersion of entities and spatial relations is represented in the following Table 5 and in Figure 9 and Figure 10.

Table 5. Results of distribution the spatial information

News Paper	Spatial Entity		Spatial Relations			
	LOC	OBJ	DIR	ORI	TOP	DIS
El khaber	56	78	43	15	16	08
Echorouk	104	86	34	10	25	02
Ennaher	78	104	28	15	14	05
Elmihwer	83	82	40	13	33	09
Total	321	350	145	53	88	25

To evaluate and compare our proposed hybrid approach, we will use the metrics of precision, recall, and F-measure. Precision indicates the correctness of the extraction, while recall indicates completeness. The F-measure provides the harmonic means between precision and recall [39].

As stated by Maynard and Peters [40], precision is defined as the ratio of valid annotations to the total number of identified annotations. This can be formally expressed as follows:

$$\text{Precision} = \frac{(\text{True Positives})}{(\text{True Positives} + \text{False Positives})} \quad (1)$$

Recall is defined as the ratio of valid annotations relative to the total number of annotations. Formally, it is expressed as:

$$\text{Recall} = \frac{(\text{True Positives})}{(\text{True Positives} + \text{False Negatives})} \quad (2)$$

The F-measure is calculated as the harmonic mean of precision and recall. It is formally expressed as:

$$\text{F} - \text{mesure} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \quad (3)$$

The results and evaluation of spatial information are presented in Tables 6 and 7. Table 6 provides a detailed analysis of performance in terms of correct, incorrect, and missing results. In contrast, Table 7 offers an evaluation of different methods using precision, recall, and F-measure metrics.

Figure 11 presents the results and evaluation of precision, recall, and F-measure scores obtained for each newspaper, namely Ennahar, Echourouk, El Khabar, and El Mihwar.

The results obtained in this study are highly satisfactory, as demonstrated by the high Precision and Recall rates (refer to Table 6 and Table 7). However, it is worth mentioning that some errors in Precision are mainly attributed to wrong annotations associated with specific words like " بين, تحت, على, في, حول" in English "between, under, on, in, around"(see the example in Table 8). These words have multiple meanings, sometimes indicating spatial relationships and other times depending on the sentence context. This ambiguity poses challenges for the extraction system.

On the other hand, the extraction of spatial entities did not face such ambiguity, and the results were accurate. It is important to consider that the quality of spatial relation extraction can be influenced by errors in semantic analysis and syntactic parsing.

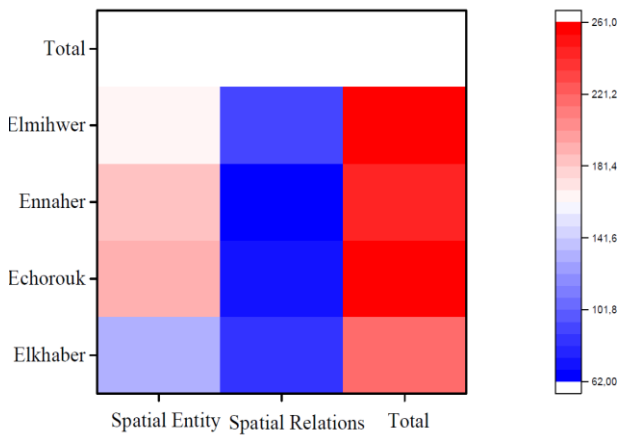

Figure 9. Panoramic information spatial in Arabic

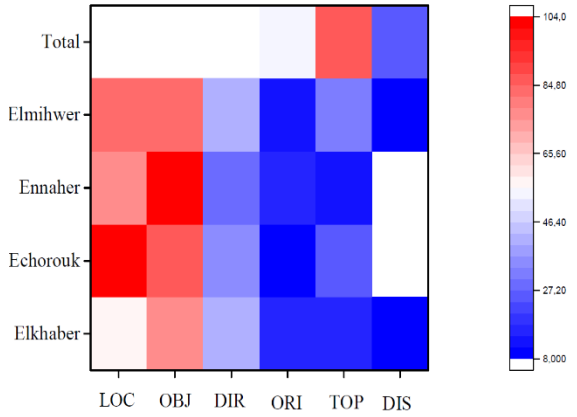

Figure 10. Distribution of spatial information by category

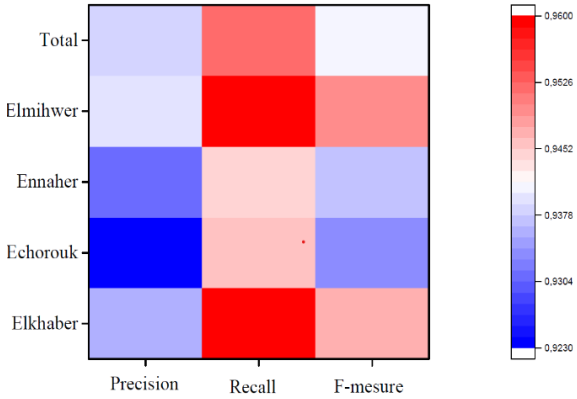

Figure 11. Evaluation of extraction Arabic spatial information

Table 6. Results and evaluation of spatial information

News Paper	Correct	Incorrect	Missing
EnnaherNews	220	15	9
EchoroukNews	229	19	13
ElkhaberNews	191	14	11
ElmihwerNews	235	15	10
Total	875	63	43

Table 7. Results and evaluation of spatial information

News Paper	Precision	Recall	F-mesure
EnnaherNews	0.936	0.960	0.947
EchoroukNews	0.923	0.946	0.934
ElkhaberNews	0.931	0.945	0.937
ElmihwerNews	0.940	0.959	0.949
Total	**0.938**	**0.952**	**0.941**

Table 8. Error syntaxic and semantic

Semantic Error	Related to time.	تتراوح أعمارهم بين 20 و 35 سنة Their ages range between 20 and 35 years
	Relation spatial	وقع الحادث بين ولايتي بجاية وجيجل The accident occurred between the states of Bejaia and Jijel
Syntax Error	Verb	خلف الحادث أربع ضحايا The accident left four victims
	Adverb (Relation spatial)	يقع المستشفى خلف مسجد المدينة The hospital is located behind the city mosque

4.2 Comparison and discussion

The following Table (Table 9) presents the results obtained by other authors for comparison, alongside our own results. This comparison highlights the differences and improvements achieved by our approach.

Subsequently, we will compare these results (see Figure 12) to highlight the strengths and limitations of our method relative to existing approaches.

Table 9. Our approach evaluation

	Precision	Recall	F-mesure
Our Approach	0.93	0.95	0.94
[5]	0.80	0.91	0.85
[6]	0.88	0.76	0.82
[41]	0.64	0.80	0.71
[42]	0.97	0.96	0.97
[43]	0.62	0.59	0.61
[44]	0.83	0.86	0.84

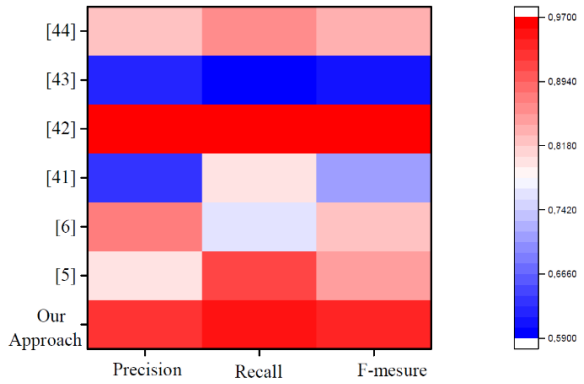

Figure 12. Evaluation of our system with other systems

In the study of Feriel and Kholladi [5], the author presents: A method based on a corpus and rules for extracting spatio-temporal information in Arabic, which achieves good results but is still inferior to our approach, highlighting the effectiveness of our spatial ontology.

-According to the study by Vasilopoulos et al. [6]: "Automatic text extraction from Arabic newspapers" also focuses on Arabic but shows lower recall and F-measure performance, suggesting that the integration of JAPE rules in our approach helps better manage linguistic ambiguities.

-As is shown by Haris et al. [41]: This English-based method, which relies on co-occurrence for extracting spatial information from travel narratives, shows inferior results, underscoring the robustness of our hybrid method.

-Reported by Li et al. [42]: Although very effective for Chinese texts, this method is not directly comparable due to linguistic differences.

-From the study by Shin et al. [43]: "BERT-based spatial information extraction" shows acceptable results, even though BERT-based models are generally powerful.

-In the study of Zenasni et al. [44], it is demonstrated: This method for extracting spatial information from short messages in English achieves good results but remains inferior to our approach.

Our hybrid approach demonstrates robust and balanced performance in extracting spatial information from Arabic texts. The combination of a spatial ontology and rules

effectively addresses the linguistic and semantic complexities of Arabic, offering a promising solution for applications in GIS and natural language processing.

5. CONCLUSIONS

This study introduces an innovative hybrid approach for the automatic extraction of spatial information from Arabic text documents, combining ontology and JAPE rules to enhance Geographic Information Systems (GIS) and Arabic Natural Language Processing (ANLP) resources. The development of the Arabic Spatial Toponym Ontology (ASTO) represents a significant advancement by structuring and formalizing spatial entities and relationships specific to Arabic, thus facilitating the indexing, annotation, and extraction of spatial data. The integration of JAPE rules has enabled effective disambiguation and classification, significantly improving the management of ambiguities and linguistic variations. The experimental results confirmed the effectiveness of this hybrid method, demonstrating its potential to optimize GIS systems and enhance the information retrival in Arabic.

The primary contributions of our work include the creation of ASTO, which structures Arabic spatial information, and the use of JAPE rules to enhance the accuracy of data extraction and classification. These contributions not only optimize GIS systems but also strengthen ANLP resources. Future work should focus on extending ASTO to cover additional geographic and semantic aspects, adapting the methodologies to other languages with complex structures, and exploring applications in complementary fields, such as social science text analysis and crisis management. Furthermore, ongoing refinement of the JAPE rules and JAPE algorithms is crucial to handle more complex cases and subtle linguistic variations, thus strengthening the robustness and reliability of information extraction systems. In summary, this research makes a significant contribution to the advancement of spatial information extraction technologies in Arabic and opens promising avenues for the future development of ANLP and GIS.

REFERENCES

[1] Alrayzah, A., Alsolami, F., Saleh, M. (2024). AraFast: Developing and evaluating a comprehensive modern standard arabic corpus for enhanced natural language processing. Applied Sciences, 14(12): 5294. https://doi.org/10.3390/app14125294

[2] Aguilar, A.J., Pinos-Navarrete, A., Domingo Jaramillo, C., de la Hoz-Torres, M.L. (2024). Geographic information systems and web GIS in higher education: a collaborative tool for the analysis of accessibility in the urban and built environment. In Teaching Innovation in Architecture and Building Engineering: Challenges of the 21st Century, pp. 401-415. https://doi.org/10.1007/978-3-031-59644-5_23

[3] Reddy, K.R., Sharma, V.K., Anusha, M., Jhade, S., Dhanasekaran, B. (2024). Progressive collaborative method for protecting users privacy in location-based services. In MATEC Web of Conferences, 392: 01089. http://doi:org/10.1051/matecconf/202439201089

[4] Hkiri, E., Mallat, S., Zrigui, M. (2016). Events automatic extraction from Arabic texts. International Journal of Information Retrieval Research (IJIRR), 6(1):36-51. https://doi.org/10.4018/IJIRR.2016010103

[5] Feriel, A., Kholladi, M.K. (2015). Automatic extraction of spatio-temporal information from Arabic text documents. International Journal of Computer Science & Information Technology (IJCSIT), 7(5): 97-107. https://doi.org/10.5121/ijcsit.2015.7507

[6] Vasilopoulos, N., Wasfi, Y., Kavallieratou, E. (2018). Automatic text extraction from Arabic newspapers. In Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, pp. 505-510. https://doi.org/10.1007/978-3-319-93000-8_57

[7] Hadji, A., Kholladi, M.K. (2012). Traitement et fusion de données dans le cadre de l'interopérabilité sémantique des systèmes d'information géographiques. In Proceedings Conference CTIC Adrar, Algeria, pp. 1-6.

[8] Wimalasuriya, D.C., Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36(3): 306-323. https://doi.org/10.1177/0165551509360123

[9] Shah, R., Jain, S. (2014). Ontology-based information extraction: An overview and a study of different approaches. International Journal of Computer Applications, 87(4): 6-8. https://doi.org/10.5120/15194-3574

[10] Etudo, U., Yoon, V.Y. (2024). Ontology-based information extraction for labeling radical online content using distant supervision. Information Systems Research, 35(1): 203-225. https://doi.org/10.1287/isre.2023.1223

[11] Ahaggach, H., Abrouk, L., Lebon, E. (2023). Information extraction and ontology population using car insurance reports. In International Conference on Information Technology-New Generations, pp. 405-411. https://doi.org/10.1007/978-3-031-28332-1_46

[12] Opasjumruskit, K., Böning, S., Schindler, S., Peters, D. (2022). OntoHuman: Ontology-based information extraction tools with human-in-the-loop interaction. In International Conference on Cooperative Design, Visualization and Engineering, pp. 68-74. https://doi.org/10.1007/978-3-031-16538-2_7

[13] Abayomi-Alli, A.A., Misra, S., Akala, M.O., Ikotun, A. M., Ojokoh, B.A. (2021). An ontology-based information extraction system for organic farming. International Journal on Semantic Web and Information Systems (IJSWIS), 17(2): 79-99. https://doi.org/10.4018/IJSWIS.2021040105

[14] Jusoh, S., Awajan, A., Obeid, N. (2020). The use of ontology in clinical information extraction. In Journal of Physics: Conference Series, 1529(5): 052083. https://doi.org/10.1088/1742-6596/1529/5/052083

[15] Rizvi, S.T.R., Mercier, D., Agne, S., Erkel, S., Dengel, A., Ahmed, S. (2018). Ontology-based information extraction from technical documents. In Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018), Funchal, Madeira, Portugal, pp. 493-500. https://doi.org/10.5220/0006596604930500

[16] Anantharangachar, R., Ramani, S., Rajagopalan, S. (2013). Ontology guided information extraction from unstructured text. International Journal of Web & Semantic Technology (IJWesT), 4(1): 19-36. https://doi.org/10.5121/ijwest.2013.4102

[17] Nebhi, K. (2012). Ontology-based information

extraction from twitter. In Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, Mumbai, pp. 17-22.
[18] Wimalasuriya, D. C., Dou, D. (2010). Components for information extraction: Ontology-based information extractors and generic platforms. In Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto ON Canada, pp. 9-18. https://doi.org/10.1145/1871437.1871444
[19] Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M. (2006). Ontology-based information extraction with SOBA. In Proceedings of the international conference on language resources and evaluation (LREC).
[20] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A. (2004). KIM–A semantic platform for information extraction and retrieval. Natural language engineering, 10(3-4): 375-392. https://doi.org/10.1017/S135132490400347X
[21] Andročec, D., Vrček, N. (2016). Ontologies for platform as service APIs interoperability. Cybernetics and Information Technologies, 16(4): 29-44. https://doi.org/10.1515/cait-2016-0065
[22] Wu, L.T., Lin, J.R., Leng, S., Li, J.L., Hu, Z.Z. (2022). Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. Automation in Construction, 135: 104108. https://doi.org/10.1016/j.autcon.2021.104108
[23] Bao, X., Xie, S., Zhang, K., Song, K., Yang, Y. (2019). Machine learning based information extraction for diabetic nephropathy in clinical text documents. In 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, pp. 1438-1442. https://doi.org/10.1109/ICSAI48974.2019.9010211
[24] Silva, E.F., Barros, F.A., Prudencio, R.B. (2006). A hybrid machine learning approach for information extraction. In 2006 Sixth International Conference on Hybrid Intelligent Systems (HIS'06), Rio de Janeiro, Brazil, pp. 44-44. https://doi.org/10.1109/HIS12093.2006
[25] Konys, A. (2018). Towards knowledge handling in ontology-based information extraction systems. Procedia Computer Science, 126: 2208-2218. https://doi.org/10.1016/j.procs.2018.07.228
[26] Biniam, P. (2020). Ontology-based information extraction from legacy surveillance reports of infectious diseases in animals and humans.Digitala Vetenskapliga Arkivet.
[27] GATE: http://gate.ac.uk/, accessed July. 23, 2024.
[28] Jones, C.B., Alani, H., Tudhope, D. (2001). Geographical information retrieval with ontologies of place. In Spatial Information Theory: Foundations of Geographic Information Science International Conference, COSIT 2001 Morro Bay, CA, USA, pp. 322-335. https://doi.org/10.1007/3-540-45424-1_22
[29] Ping, D., Yong, L. (2009). Building place name ontology to assist in geographic information retrieval. In 2009 International Forum on Computer Science-Technology and Applications, Chongqing, China, pp. 306-309. https://doi.org/10.1109/IFCSTA.2009.80
[30] Guarino, N., Oberle, D., Staab, S. (2009). What is an ontology? Handbook on Ontologies, pp. 1-17. https://doi.org/10.1007/978-3-540-92673-3_0
[31] Noy, N., McGuiness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. Stanford, CA, Stanford Knowledge Systems Laboratory, USA.
[32] Abu-Errub, A., Odeh, A., Shambour, Q., Hassan, O.A.H. (2014). Arabic roots extraction using morphological analysis. International Journal of Computer Science Issues (IJCSI), 11(2): 128-134.
[33] Aljabari, A., Duaibes, L., Jarrar, M., Khalilia, M. (2024). Event-arguments extraction corpus and modeling using BERT for Arabic. Computation and Language, arXiv:2407.21153. https://doi.org/10.48550/arXiv.2407.21153
[34] Al-Laith, A., Shahbaz, M., Alaskar, H.F., Rehmat, A. (2021). Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus. Applied Sciences, 11(5): 2434. https://doi.org/10.3390/app11052434
[35] Attia, M. (2007). Arabic tokenization system. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic, pp. 65-72. https://doi.org/10.3115/1654576.1654588
[36] Borisova, N. (2014). An approach for ontology based information extraction. Information Technologies and Control, 12(1): 15-20. https://doi.org/10.1515/itc-2015-0007
[37] GATE Team. (2024). Chapter 8, JAPE: Regular Expressions over Annotations, GATE: A General Architecture for Text Engineering. https://gate.ac.uk/sale/tao/splitch8.html, accessed on August 13, 2024.
[38] Thakker, D., Osman, T., Lakin, P. (2009). Gate jape grammar tutorial. Nottingham Trent University, UK, Phil Lakin, UK, Version, 1.
[39] Gutierrez, F., Dou, D., Fickas, S., Wimalasuriya, D., Zong, H. (2016). A hybrid ontology-based information extraction system. Journal of Information Science, 42(6): 798-820. https://doi.org/10.1177/0165551515610989
[40] Maynard, D., Peters, W., Li, Y. (2006). Metrics for evaluation of ontology-based information extraction. In Proceeding15th International Conference on World Wide Web in Workshop on Evaluation of Ontologies for the Web, EON@ WWW, Edinburgh, UK.
[41] Haris, E., Gan, K.H., Tan, T.P. (2020). Spatial information extraction from travel narratives: Analysing the notion of co-occurrence indicating closeness of tourist places. Journal of Information Science, 46(5): 581-599. https://doi.org/10.1177/0165551519837188
[42] Li, X., Zhang, W., Wang, Y., Tan, Y., Xia, J. (2023). Spatio-temporal information extraction and geoparsing for public Chinese resumes. ISPRS International Journal of Geo-Information, 12(9): 377. https://doi.org/10.3390/ijgi12090377
[43] Shin, H.J., Park, J.Y., Yuk, D.B., Lee, J.S. (2020). BERT-based spatial information extraction. In Proceedings of the Third International Workshop on Spatial Language Understanding, pp. 10-17. https://doi.org/10.18653/v1/2020.splu-1.2
[44] Zenasni, S., Kergosien, E., Roche, M., Teisseire, M. (2018). Spatial information extraction from short messages. Expert Systems with Applications, 95: 351-367. https://doi.org/10.1016/j.eswa.2017.11.025