A NLP and Rule-Based Approach to Extract Spatial Entities and Relationships in Arabic Text

Atmane HADJI[1,2,*,†], Mohamed Khireddine Kholladi[3,4,†] and Farid Boumaza[5,6,†]

[1]University of Bejaia, Faculty of exact sciences, Department of Computer Science, 06000 Bejaia, Algeria
[2]LISI Laboratory, Computer Science Department, University Center A. Boussouf Mila, 43000 Mila, Algeria
[3]Department of Mathematics and Computer Science, Faculty of Exact Sciences , University of El-Oued , , HAMMA Lakhdar , El-Oued , Algeria
[4]MISC Laboratory of Abdelhamid Mehri university of Constantine 2, Algeria
[5]Computer Science Department, University of Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj 34030, Algeria
[6]LAPECI Laboratory , University of Oran1, Oran 31000, Algeria

Abstract

Keywords

Information extraction, spatial information, NLP Arabic, rules JAPE

1. Introduction

In recent years, the massive growth of digital information, particularly on the Internet and within Big Data, has highlighted the need to develop efficient information processing systems. This exponential data increase, especially in georeferenced information, has amplified the challenge of information overload, making quick and accurate access to relevant data increasingly crucial, especially in specialized fields like geographic information systems (GIS). In this context, spatial information extraction from raw texts has become a vital area of research, encompassing disciplines such as natural language processing (NLP), information extraction (IE), information retrieval (IR), and GIS [1].

Spatial information extraction, especially in the Arabic language, offers significant advantages across various sectors. It enriches geospatial databases, enhances the accuracy of geographic information systems [2], and optimizes location-based services (LBS) [3]. It also supports decision-making in critical fields such as urban planning, natural resource management, and disaster response. The extraction process involves transforming unstructured textual data into structured information, thereby identifying geospatial entities, relationships, semantic roles, and events for deeper analysis.

However, despite these potential benefits, spatial information extraction from Arabic texts remains a major challenge. Due to the morphological richness of the language and its semantic ambiguities, traditional information extraction methods, whether based on statistical techniques or machine learning, often fall short in addressing these challenges. The Arabic language presents linguistic and grammatical complexities that complicate the identification of georeferenced information, making integration into GIS systems even more challenging. This underscores the importance of developing advanced techniques that can effectively handle these linguistic specifics and overcome the limitations of traditional approaches.

Our approach leverages the complementary strengths of NLP to handle the linguistic intricacies of Arabic while addressing the growing needs of GIS users in the Arab world.

Proceedings of the International IAM'24: International Conference on Informatics and Applied Mathematics, December 04–05, 2024, Guelma, Algeria

[*]Corresponding author.

[†]These authors contributed equally.

a.hadji@centre-univ-mila.dz (A. HADJI); kholladi@univ-eloued.dz (M. K. Kholladi); farid.pgia@gmail.com (F. Boumaza)
0000-0001-6706-6360 (A. HADJI); 0000-0002-3051-1317 (M. K. Kholladi); 0000-0002-9785-420X (F. Boumaza)

In this work, we aimed to address the challenges posed by spatial information extraction in Arabic-language texts, an underexplored field in GIS contexts. To this end, we developed innovative solutions based on NLP techniques and JAPE rules. The objective is to overcome the limitations of traditional approaches to structure geospatial knowledge and facilitate the indexing and extraction of spatial entities and their relationships.

In the first section, we introduce our new approach based on JAPE rules. Section 2 provides a review of related works on information extraction systems in various domains. In Section 3, we present the proposed approach along with the system architecture, detailing the components involved. Section 4 focuses on the application and implementation of our approach. Finally, in Section 5, we discuss the results obtained with our method and perform a comparative evaluation with other approaches.

2. Related works

Rule-based methods have proven their effectiveness in various areas of information extraction, not least thanks to their ability to capture specific relationships by applying defined syntactic and semantic rules. [4] reported a method for extracting and combining spatial and temporal information from Arabic texts that enhances search and exploration capabilities using the GATE (General Architecture for Text Engineering) architecture. [5] Introduced "drNER", a novel rule-based Named Entity Recognition (NER) method designed to extract dietary concepts, and this approach showed significant results for the extraction of evidence-based dietary recommendations.

In the bibliographic domain, [6] applied a rule-based information extraction process to bibliographic data, aiming to establish a database of relevant concepts, refine the retrieved data and automate the local retrieval process. [7] developed a system combining information extraction and ontology creation to facilitate the extraction and visualization of clinical information.

Furthermore, [8] addressed the challenge of automatic information structure extraction from PDF books, proposing an intelligent rule-based approach to accurately extract logical metadata from these documents widely used on the semantic web. [9] presented the VALET (Very Agile Language Extraction Toolkit) framework, a rule-based information extraction system that combines lexical, orthographic, syntactic and corpus-analytic information in a flexible syntax.

[10] proposed an approach integrating automatic natural language processing (ANLP) techniques, rules and gazetteers to extract spatial entities and their relationships from texts, offering a viable solution for enriching GIS with accurate spatial information. [11], demonstrated the effectiveness of a rule-based approach for extracting spatial relationships from annotated corpora, particularly for simple directional relationships. [12] also proposed a system that automatically generates extraction rules from complex Chinese literal features. [13] demonstrated how cross-linguistic alignment based on specific grammatical rules can enrich Open IE datasets for under-represented languages such as Brazilian Portuguese. Finally, [14] illustrated how AIS (Automatic Identification System) data from fishing vessels can be exploited to extract precise spatial information, aimed at improving marine resource management.

3. Proposed JAPE rule-based method

The rule-based method is a classic and widely used approach in the field of information extraction. This method relies on a set of predefined rules that are designed to identify and extract specific information from text or other types of data. These rules are usually expressed in the form of models or patterns that correspond to specific linguistic structures or patterns in the data.

The general architecture of the proposed approach Figure 1 consists of four distinct phases.

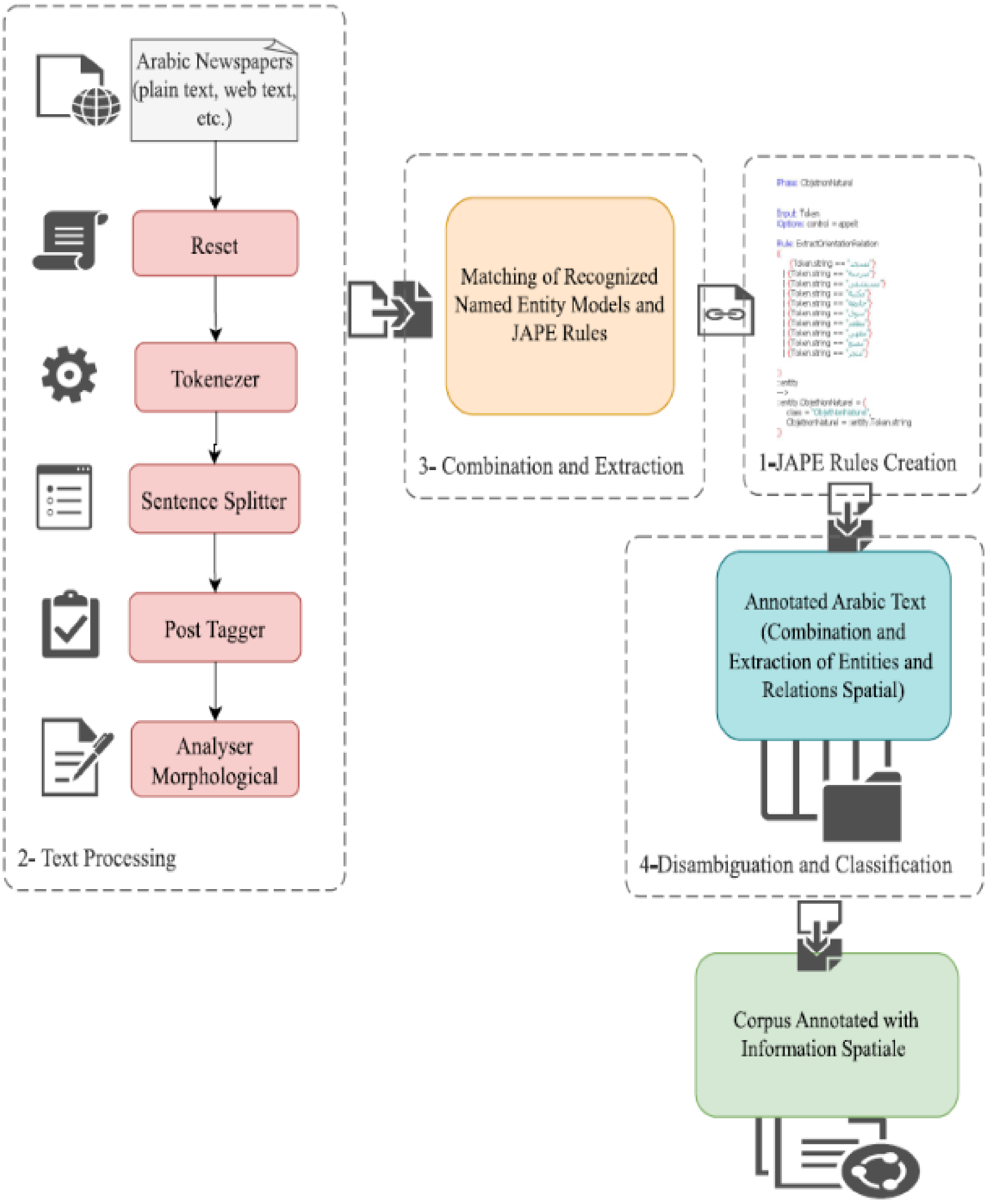

Arabic Newspapers (plain text, web text, etc.)
Reset
Tokenezer
Sentence Splitter
Post Tagger
Analyser Morphological
2- Text Processing
Matching of Recognized Named Entity Models and JAPE Rules
3- Combination and Extraction
1-JAPE Rules Creation
Annotated Arabic Text (Combination and Extraction of Entities and Relations Spatial)
4-Disambiguation and Classification
Corpus Annotated with Information Spatiale

Figure 1: General architecture of proposed approach

3.1. Creation of JAPE rules

In the first phase, concepts related to Arab entities and spatial relationships are identified and collected. These concepts are then used to formulate specific JAPE rules [15]. which are used to annotate and extract relevant spatial information from Arabic texts. JAPE rules are advanced regular expressions developed in Java, enabling the detection of complex patterns in text.

JAPE rules offer significant flexibility in natural language processing, particularly for extracting information from unstructured text. Their main strength lies in the ease of adding or modifying new rules. It is straightforward to integrate new words or expressions into an existing system without disrupting the functionality of previously defined rules. This ability to quickly update the rules based on domain evolution or analysis needs makes JAPE a particularly adaptable and efficient tool for tasks such as entity recognition and contextual information extraction.

3.2. Text processing

The second phase consists of applying natural language processing modules to prepare the raw text. This process includes steps such as normalization, tokenization and annotation of the spatial entities present in the text. These modules are crucial to ensuring that JAPE (Java Annotation Patterns Engine) rules can be applied efficiently and accurately.

3.3. Combination and Extraction

The third phase is based on the application of the JAPE rules created in the first phase. These rules are used to associate text segments with defined classes, subclasses or instances. This phase is essential for automatically extracting structured spatial information from unstructured text, taking into account the linguistic and contextual specificities of the Arabic language.

3.4. Disambiguation and classification

The fourth phase focuses on disambiguation and classification of the extracted spatial entities. This step ensures that each entity and relationship is correctly interpreted in its specific context. JAPE rules are also used here to refine the results, applying disambiguation and classification criteria to improve the accuracy of the extracted data.

4. Application and realization

4.1. Implementation phase

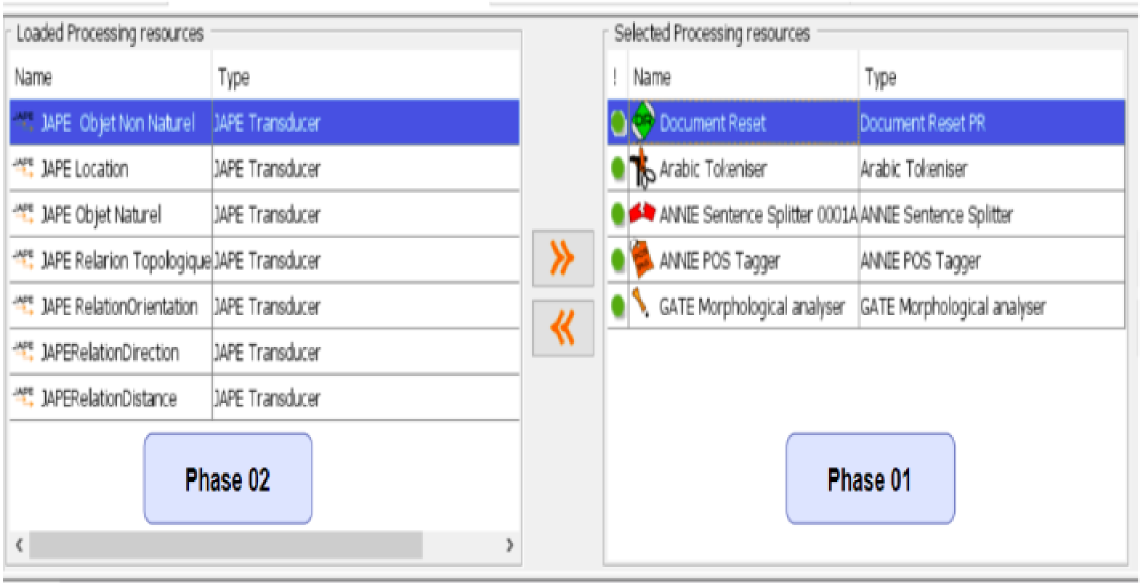
Our JAPE rule-based system architecture consists of two main phases, each playing a crucial role in the extraction of geographic information from natural language text (Figure 2). The first phase uses advanced Natural Language Processing (NLP) techniques to prepare and normalize text data. This preparation includes text cleaning, sentence segmentation and initial annotation of linguistic elements, facilitating better rule application.

The second phase focuses on matching JAPE rules to extract specific information. This phase involves the definition and creation of rules, the matching of these rules with the text, disambiguation and the extraction of relevant information. Finally, post-processing is carried out to filter and structure the extracted data, making it ready for further analysis or integration into geospatial databases. Together, these phases ensure accurate and efficient information extraction, tailored to the needs of geographic analysis.

4.2. Application environment

We have chosen to use the GATE environment, a linguistic engineering framework developed by the University of Sheffield and widely adopted since its first release in 1996 for teaching and research. GATE offers a suite of reusable processing resources in JAVA, integrated into an information extraction system called ANNIE (aNearly-New Information Extraction System) [16].

By default, ANNIE is configured for languages other than Arabic. To adapt this tool to our target language, we will use specialized components such as the Arabic tokenizer, sentence splitter, POS tagger and Arabic morphological analyzer. To avoid interference with previous executions, we'll apply the "reset" option to remove all traces of previous processes. Annotations in GATE will be performed



Loaded Processing resources

Name	Type
JAPE Objet Non Naturel	JAPE Transducer
JAPE Location	JAPE Transducer
JAPE Objet Naturel	JAPE Transducer
JAPE Relarion Topologique	JAPE Transducer
JAPE RelationOrientation	JAPE Transducer
JAPERelationDirection	JAPE Transducer
JAPERelationDistance	JAPE Transducer

Selected Processing resources

!	Name	Type
●	Document Reset	Document Reset PR
●	Arabic Tokeniser	Arabic Tokeniser
●	ANNIE Sentence Splitter 0001A	ANNIE Sentence Splitter
●	ANNIE POS Tagger	ANNIE POS Tagger
●	GATE Morphological analyser	GATE Morphological analyser

Figure 2: Phases of our approach

by selecting words in the text and creating new annotation categories, enabling precise extraction of geographical features and other relevant information.

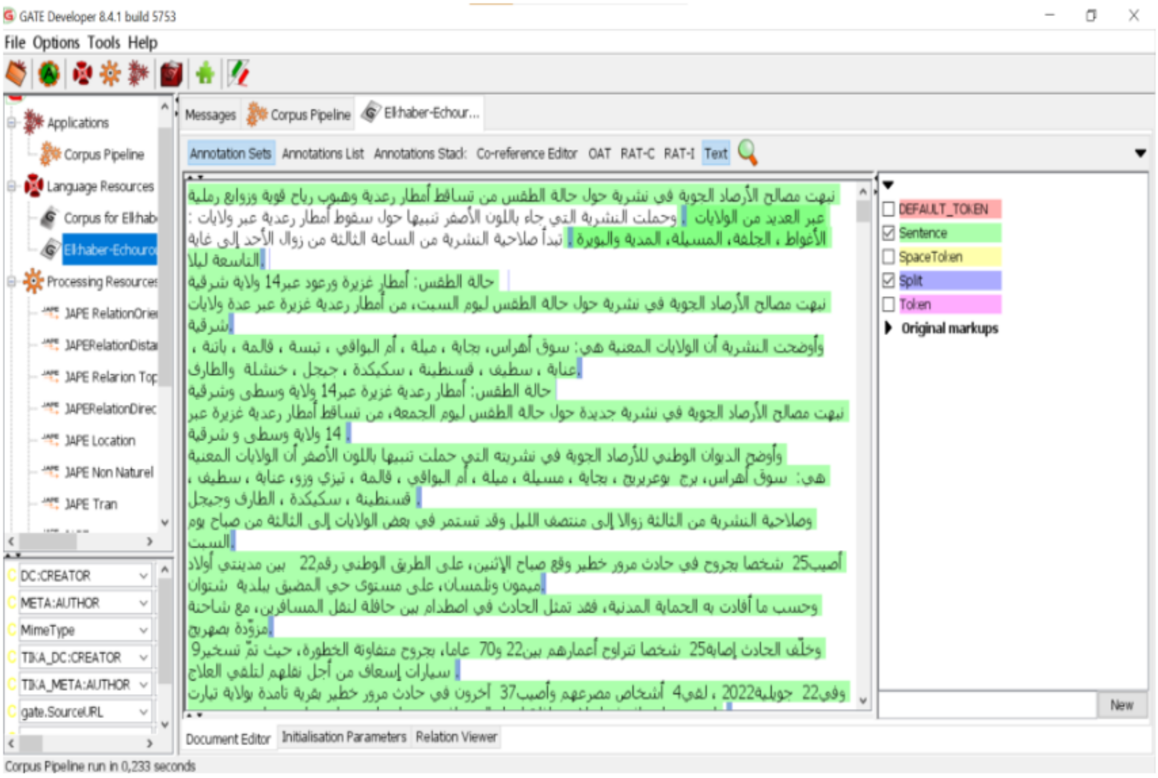
4.3. Phase One: NLP techniques

The first phase of our architecture implements NLP techniques that are essential for processing and understanding natural language text. The aim of this phase is to prepare the textual data in such a way as to facilitate the extraction of relevant information, bearing in mind that we have used the same dataset or corpus discussed in [1].

4.3.1. Linguistic pre-processing

The cleaning of Arabic text is an essential step before applying Natural Language Processing (NLP) techniques. Here are the main steps specific to the cleaning of Arabic texts:

- **Removal of diacritical characters (Tashkeel):** Arabic texts may contain diacritics (harakats) such as Fatha, Damma, Kasra and so on. These diacritics can be removed, as they are often unnecessary for analysis;
- **Removal of special characters and punctuation:** As in other languages, special characters (such as !, @, #, etc.) and punctuation can be removed to simplify the text;
- **Character standardization:** In Arabic, some characters can be written in more than one way. For example, أ, إ, and آ are often normalized to ا. Similarly, ى can be transformed into ي.
- **Removing superfluous spaces:** Arabic texts may contain multiple spaces or spaces before or after punctuation. These spaces need to be normalized to ensure correct analysis.



نبهت مصالح الأرصاد الجوية في نشرية حول حالة الطقس من تساقط أمطار رعدية وهبوب رياح قوية وزوابع رملية
عبر العديد من الولايات . وحملت النشرية التي جاء باللون الأصفر تنبيها حول سقوط أمطار رعدية عبر ولايات :
الأغواط ، الجلفة، المسيلة، المدية والبويرة . تبدأ صلاحية النشرية من الساعة الثالثة من زوال الأحد إلى غاية
التاسعة ليلا.
حالة الطقس: أمطار غزيرة ورعود عبر14 ولاية شرقية
نبهت مصالح الأرصاد الجوية في نشرية حول حالة الطقس ليوم السبت، من أمطار رعدية غزيرة عبر عدة ولايات
شرقية.
وأوضحت النشرية أن الولايات المعنية هي: سوق أهراس، بجاية ، ميلة ، أم البواقي ، تبسة ، قالمة ، باتنة ،
عنابة ، سطيف ، قسنطينة ، سكيكدة ، جيجل ، خنشلة والطارف.
حالة الطقس: أمطار رعدية غزيرة عبر14 ولاية وسطى وشرقية
نبهت مصالح الأرصاد الجوية في نشرية جديدة حول حالة الطقس ليوم الجمعة، من تساقط أمطار رعدية غزيرة عبر
14 ولاية وسطى و شرقية .
وأوضح الديوان الوطني للأرصاد الجوية في نشريته التي حملت تنبيها باللون الأصفر أن الولايات المعنية
هي: سوق أهراس، برج بوعريريج ، بجاية ، مسيلة ، ميلة ، أم البواقي ، قالمة ، تيزي وزو، عنابة ، سطيف ،
قسنطينة ، سكيكدة ، الطارف وجيجل .
وصلاحية النشرية من الثالثة زوالا إلى منتصف الليل وقد تستمر في بعض الولايات إلى الثالثة من صباح يوم
السبت.
أصيب25 شخصا بجروح في حادث مرور خطير وقع صباح الإثنين، على الطريق الوطني رقم22 بين مدينتي أولاد
ميمون وتلمسان، على مستوى حي المضيق ببلدية شتوان.
وحسب ما أفادت به الحماية المدنية، فقد تمثل الحادث في اصطدام بين حافلة لنقل المسافرين، مع شاحنة
مزوّدة بصهريج.
وخلّف الحادث إصابة25 شخصا تتراوح أعمارهم بين22 و70 عاما، بجروح متفاوتة الخطورة، حيث تمّ تسخير9
سيارات إسعاف من أجل نقلهم لتلقي العلاج .
وفي22 جويلية2022 ، لقي4 أشخاص مصرعهم وأصيب37 آخرون في حادث مرور خطير بقرية تامدة بولاية تيارت

Figure 3: Execution of Sentence Splitter in GATE

These cleaning steps are crucial to obtaining accurate and relevant results when analyzing Arabic text using Automatic Natural Language Processing (ANLP) techniques. Properly cleaned text reduces noise and errors, enabling analysis algorithms to better understand the complex linguistic structures of Arabic, improve the accuracy of results and ensure better interpretation of textual data. Thus, these cleaning practices are essential for any NLP application, be it named entity recognition, text classification or machine translation.

4.3.2. Application of TALN techniques

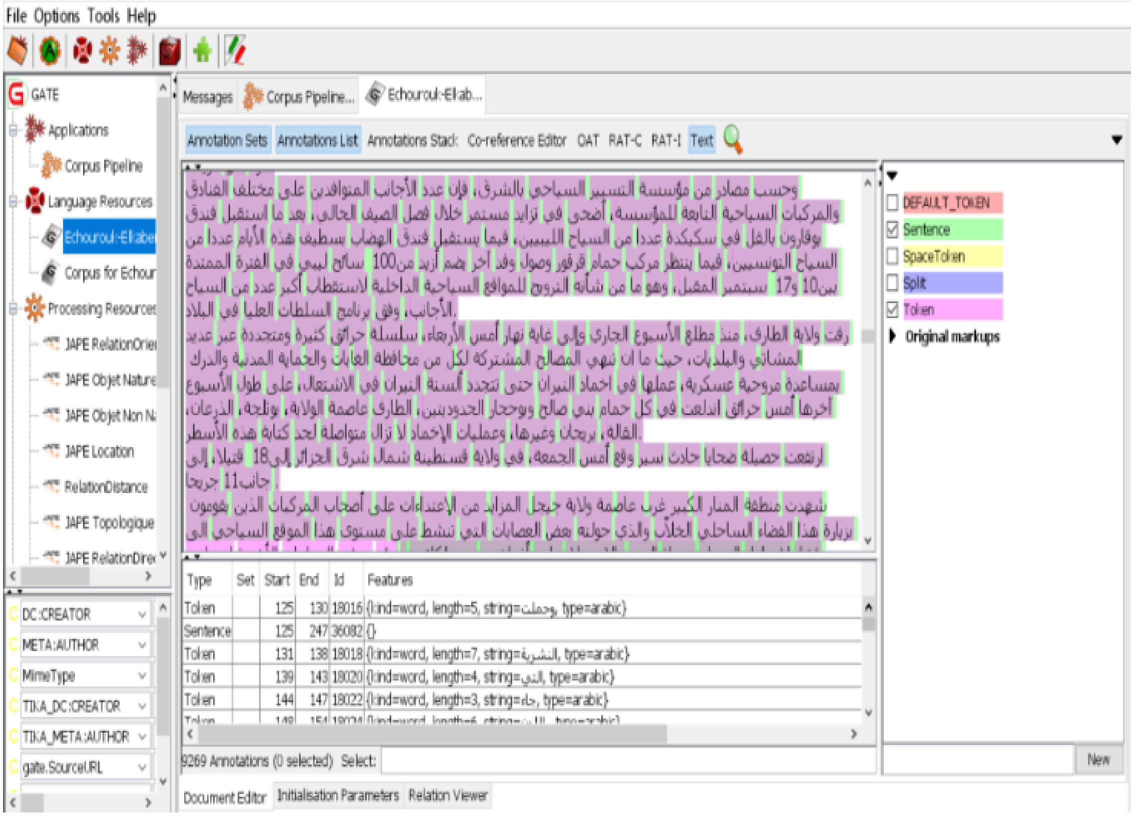
TALN techniques such as Document Reset, Arabic Tokeniser, Sentence Splitter, Post Tagging and Morphological Analyser were explained in detail in the next sections. In this study, we will focus on the practical application of these techniques using the GATE platform [17]. This in-depth exploration is intended to provide a better understanding of GATE and to serve as a practical guide to its use, particularly in the context of Arabic text. Given that online documentation is relatively limited, this chapter plays a vital role in filling this gap and offering clear instructions for taking full advantage of GATE's features.

4.3.3. Sentence Splitter

Figure 3 shows a visualization of GATE during the Sentence Splitter step. This step splits the text into distinct sentences, improving the accuracy of syntactic and grammatical analyses.

4.3.4. Tokenization

Figure 4 shows a screenshot of the GATE platform, illustrating the tokenization process. It shows how GATE segments text into basic units (tokens) for further linguistic analysis.



وحسب مصادر من مؤسسة التسيير السياحي بالشرق، فإن عدد الأجانب المتوافدين على مختلف الفنادق
والمركبات السياحية التابعة للمؤسسة، أضحى في تزايد مستمر خلال فصل الصيف الحالي، بعد ما استقبل فندق
بوقارون بالقل في سكيكدة عددا من السياح الليبيين، فيما يستقبل فندق الهضاب بسطيف هذه الأيام عددا من
السياح التونسيين، فيما ينتظر مركب حمام قرقور وصول وفد آخر يضم أزيد من 100 سائح ليبي في الفترة الممتدة
بين 10 و17 سبتمبر المقبل، وهو ما من شأنه الترويج للمواقع السياحية الداخلية لاستقطاب أكبر عدد من السياح
الأجانب، وفق برنامج السلطات العليا في البلاد.

رفت ولاية الطارف، منذ مطلع الأسبوع الجاري وإلى غاية نهار أمس الأربعاء، سلسلة حرائق كثيرة ومتجددة عبر عديد
المشاتي والبلديات، حيث ما ان تنهي المصالح المشتركة لكل من محافظة الغابات والحماية المدنية والدرك
بمساعدة مروحية عسكرية، عملها في اخماد النيران حتى تتجدد ألسنة النيران في الاشتعال، على طول الأسبوع
أخرها أمس حرائق اندلعت في كل حمام بني صالح وبوحجار الحدوديتين، الطارف عاصمة الولاية، بوثلجة، الذرعان،
القالة، بريحان وغيرها، وعمليات الإخماد لا تزال متواصلة لحد كتابة هذه الأسطر.

ارتفعت حصيلة ضحايا حادث سير وقع أمس الجمعة، في ولاية قسنطينة شمال شرق الجزائر إلى 18 قتيلا، إلى
جانب 11 جريحا.

شهدت منطقة المنار الكبير غرب عاصمة ولاية جيجل المزيد من الاعتداءات على أصحاب المركبات الذين يقومون
بزيارة هذا الفضاء الساحلي الخلاب والذي حولته بعض العصابات التي تنشط على مستوى هذا الموقع السياحي الى

Type	Set	Start	End	Id	Features
Token		125	130	18016	{kind=word, length=5, string=وحملت, type=arabic}
Sentence		125	247	36082	{}
Token		131	138	18018	{kind=word, length=7, string=النشرية, type=arabic}
Token		139	143	18020	{kind=word, length=4, string=التي, type=arabic}
Token		144	147	18022	{kind=word, length=3, string=جاء, type=arabic}
Token		148	154	18024	{kind=word, length=6, string=[illegible], type=arabic}

9269 Annotations (0 selected) Select:

Figure 4: Execution of Tokenization in GATE

4.3.5. Post Tagging and Morphological Analyser

Morpho-syntactic tagging (Post Tagging) and morphological analysis (Morphological Analyser) are essential processes in automatic natural language processing, particularly when integrated into advanced systems such as GATE (General Architecture for Text Engineering) [18]. In this context, these steps are exploited by JAPE (Java Annotation Patterns Engine) rules, which enable sophisticated annotation patterns to be defined for detecting specific linguistic structures within a corpus. When JAPE rules are executed, the annotations generated by Post Tagging and morphological analysis enrich the corpus by adding detailed metadata on grammatical categories and word morphological structure. Although these annotations do not affect the visible display of the text, they play a crucial role in providing invisible but fundamental information for subsequent linguistic analysis and accurate information extraction.

4.4. Second phase: Application of JAPE rules

The second phase focuses on the application of JAPE rules for the extraction of specific spatial information. This phase follows the well-defined steps between rules and spatial information, i.e. the classification of spatial entities and relationships according to the following table

The Table 1 below presents a classification of spatial entities, including natural entities, non-natural entities and entities corresponding to place names or locations.

The Table 2 presents a classification of spatial relationships, detailing the following categories: topological relationships, directional relationships, distance relationships and orientation relationships. These categories help us understand how spatial entities position, orient and relate to each other in a given space. Topological relations describe relationships of contiguity or inclusion, directional relations indicate relative orientations, distance relations measure distances between entities, and orientation

Table 1
Spatial Entities classes

Table 1	Spatial Entities classes	
	Classes	Instances
	Natural Object	جبل، واد، هضبة، شاطئ، غابة، بحر، نهر، صحراء، تلة، بحيرة
Spatial Entities	Building Object	مسجد، مدرسة، مستشفى، مكتبة، جامعة، سوق، مطعم، مقهى، مصنع، محطة، مطار، فندق، جسر، كنيسة، معبد، ملعب، دار
	Location	جيجل، الطاهير، الأمير عبد القادر، الشقفة، القنار نشفي، تاكسنة، العوانة،

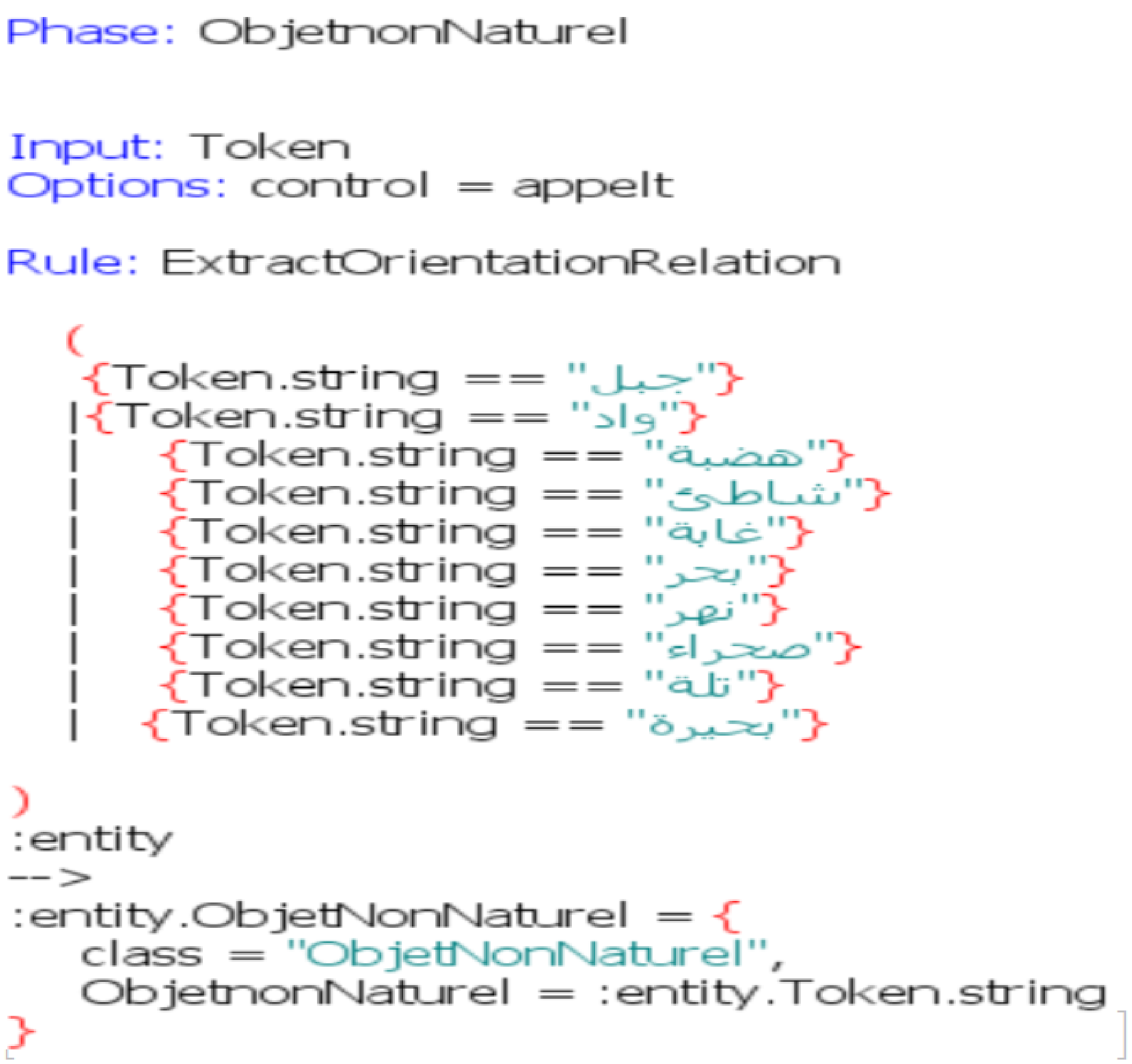
Table 2
Spatial Relations classes

Table 2	Spatial Relations classes	
	Classes	Instances
Spatial Relations	Topological	بعض؛ جزء؛ بضع؛ بين؛ وسط؛ داخل؛ في؛ على؛ علي مستوى؛ على محور؛ على حافة؛ بجانب؛ حول؛ قرب؛ خلف؛ أمام
	Direction	شمال؛ جنوب؛ شرق؛ غرب؛ شمال شرق؛ شمال غرب؛ جنوب شرق؛ جنوب غرب؛ نحو؛ باتجاه؛ صوب؛ قصد؛ عبر؛ من خلال؛ حتى؛ نحو الأعلى؛ نحو الأسفل؛
	Distance	مسافة؛ على بعد؛ تبعد؛ قربٍ؛ دنو؛ علي قرِب؛ قصيا؛ قريب من؛ بعيد عن؛ على مسافة؛ بعيد نسبياً؛ قريب نسبياً؛ على مسافة قصيرة؛ على مسافة طويلة ......
	Topological	بعض؛ جزء؛ بضع؛ بين؛ وسط؛ داخل؛ في؛ على؛ علي مستوى؛ على محور؛ على حافة؛ بجانب؛ حول؛ قرب؛ خلف؛ أمام

relations specify alignments or angles between them.

4.5. Creation of JAPE Rules

A JAPE grammar consists of a set of phases, each containing a series of pattern/action [15]. The phases execute sequentially, forming a cascade of finite-state transducers on the annotations. The left-hand side (LHS) of the rules comprises an annotation pattern description, while the right-hand side (RHS) contains instructions for manipulating the annotations. Matching annotations on the LHS of a rule can be referenced on the RHS using labels attached to pattern elements. Below is an example of a JAPE rule (Figure 5) for extracting named entities from the "Non-Natural Object" class containing specified



```
Phase: ObjetnonNaturel

Input: Token
Options: control = appelt

Rule: ExtractOrientationRelation

  (
   {Token.string == "جبل"}
  |{Token.string == "واد"}
  |    {Token.string == "هضبة"}
  |    {Token.string == "شاطئ"}
  |    {Token.string == "غابة"}
  |    {Token.string == "بحر"}
  |    {Token.string == "نهر"}
  |    {Token.string == "صحراء"}
  |    {Token.string == "تلة"}
  |   {Token.string == "بحيرة"}

)
:entity
-->
:entity.ObjetNonNaturel = {
    class = "ObjetNonNaturel",
    ObjetnonNaturel = :entity.Token.string
}
```

Figure 5: Example of JAPE rule

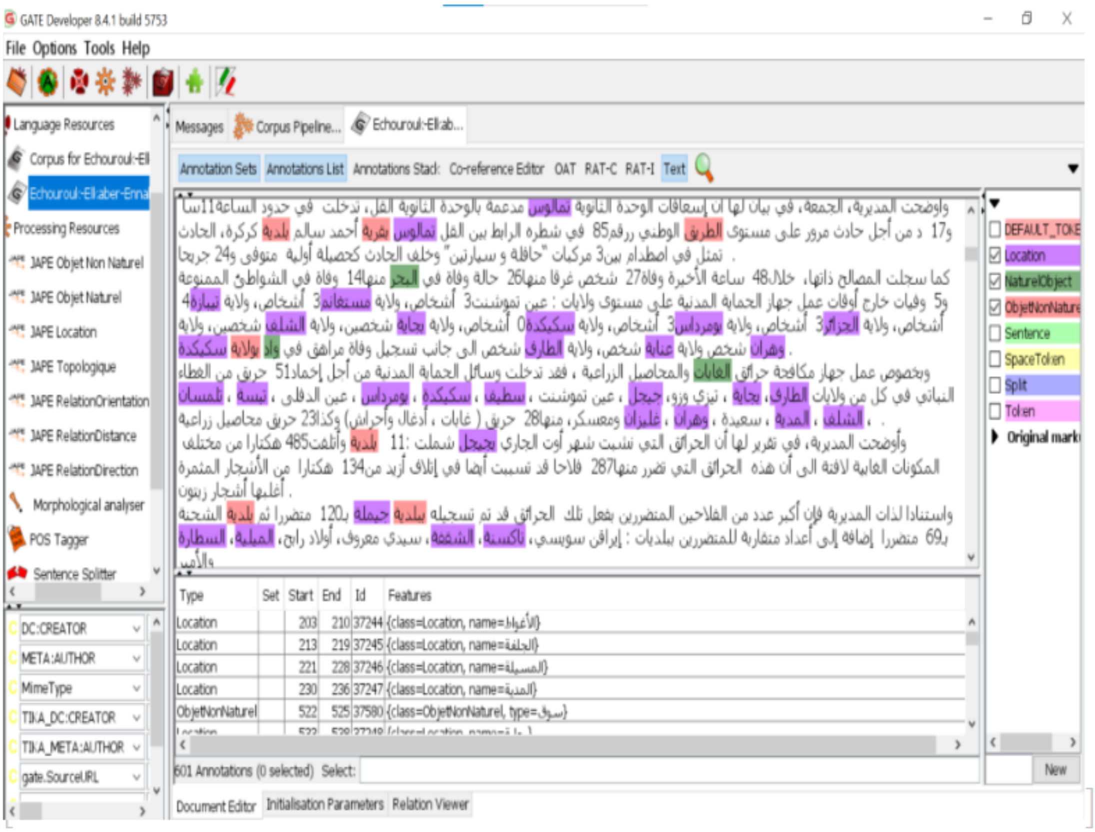
instances: جبل (mountain), هضبة (plateau), شاطئ (beach), غابة (forest), واد (valley), بحر (sea), صحراء (desert), نهر (river).

4.6. Rule Matching

The defined rules (Figure 5) are applied to the text or data to identify segments that match the specified patterns. A rule could be designed to identify geographic entities by searching for phrases containing keywords such as "region," "city," or "country." In our method, the option control = appelt is used to specify that the rules should be executed sequentially. This ensures that each rule is applied in a precise order, thereby maximizing the accuracy of the extraction.

4.7. Information Extraction

When matching segments are identified, relevant information is extracted (Figure 6). This may include capturing specific words or phrases or identifying relationships between different entities. For example, the rule "Extract Natural Object" in our JAPE script checks for text matching one of the specified



واوضحت المديرية، الجمعة، في بيان لها أن إسعافات الوحدة الثانوية تمالوس مدعمة بالوحدة الثانوية القل، تدخلت في حدود الساعة11سا
و17 د من أجل حادث مرور على مستوى الطريق الوطني ررقم85 في شطره الرابط بين القل تمالوس بقرية أحمد سالم بلدية كركرة، الحادث
تمثل في اصطدام بين3 مركبات "حافلة و سيارتين" وخلف الحادث كحصيلة أولية متوفى و24 جريحا .
كما سجلت المصالح ذاتها، خلال48 ساعة الأخيرة وفاة27 شخص غرقا منها26 حالة وفاة في البحر منها14 وفاة في الشواطئ الممنوعة
و5 وفيات خارج أوقات عمل جهاز الحماية المدنية على مستوى ولايات : عين تموشنت3 أشخاص، ولاية مستغانم3 أشخاص، ولاية تيبازة4
أشخاص، ولاية الجزائر3 أشخاص، ولاية بومرداس3 أشخاص، ولاية سكيكدة0 أشخاص، ولاية بجاية شخصين، ولاية الشلف شخصين، ولاية
وهران شخص ولاية عنابة شخص، ولاية الطارف شخص الى جانب تسجيل وفاة مراهق في واد بولاية سكيكدة .
وبخصوص عمل جهاز مكافحة حرائق الغابات والمحاصيل الزراعية ، فقد تدخلت وسائل الحماية المدنية من أجل إخماد51 حريق من الغطاء
النباتي في كل من ولايات الطارف، بجاية ، تيزي وزو، جيجل ، عين تموشنت ، سطيف ، سكيكدة ، بومرداس ، عين الدفلى ، تبسة ، تلمسان
، الشلف ، المدية ، سعيدة ، وهران ، غليزان ومعسكر، منها28 حريق (غابات ، أدغال وأحراش) وكذا23 حريق محاصيل زراعية
وأوضحت المديرية، في تقرير لها أن الحرائق التي نشبت شهر أوت الجاري شملت :11 بلدية وأتلفت485 هكتارا من مختلف
المكونات الغابية لافتة الى أن هذه الحرائق التي تضرر منها287 فلاحا قد تسببت أيضا في إتلاف أزيد من134 هكتارا من الأشجار المثمرة
أغلبها أشجار زيتون .
واستنادا لذات المديرية فإن أكبر عدد من الفلاحين المتضررين بفعل تلك الحرائق قد تم تسجيله ببلدية جيملة بـ120 متضررا ثم بلدية الشحنة
بـ69 متضررا إضافة إلى أعداد متقاربة للمتضررين ببلديات : إيراقن سويسي، تاكسنة، الشقفة، سيدي معروف، أولاد رابح، الميلية، السطارة
والأمير

Type	Set	Start	End	Id	Features
Location		203	210	37244	{class=Location, name=الأغواط}
Location		213	219	37245	{class=Location, name=الجلفة}
Location		221	228	37246	{class=Location, name=المسيلة}
Location		230	236	37247	{class=Location, name=المدية}
ObjetNonNaturel		522	525	37580	{class=ObjetNonNaturel, type=سوق}
Location		522	528	37248	{class=Location, name=[illegible]}

601 Annotations (0 selected) Select:

Figure 6: Extraction of spatial information based JAPE Rules

instances in a list of natural objects, such as غابة (forest), واد (valley), or بحر (sea). When a token matches a word in the text, a named entity "Natural Object" is created, and specific attributes, such as class = "Natural Object" and type (containing the character string of the identified entity), are associated with this entity.

5. Results and Evaluation

In this section, we examine the results of experiments aimed at evaluating the effectiveness of our spatial information extraction system, which uses JAPE rules. This method is based on applying these rules to automatically extract spatial entities and relationships from a corpus of Arabic texts.

To evaluate and compare the methods we studied, we will use metrics: Precision, Recall, and F-scale. Precision refers to the correctness of the retrieval, while recall refers to the completeness of the retrieval. The F-measure provides the harmonic mean between precision and recall [17].

According to [18]:

- Precision is the fraction of the valid annotations over the total number of identified annotations. It is formally defined as:

$$Precision = (Correct)/(Correct + Spurious) \tag{1}$$

- Recall is the fraction of the valid annotations over the total amount of annotations. It is formally defined as:

$$Recall = Correct/(Correct + Missing) \tag{2}$$

Table 3
Distribution of annotations spatial entities and relations

Table 3
Distribution of annotations spatial entities and relations

NewsPaper	Spatial Entity	Spatial Relation	Total Words
Total	611	197	9008

Table 4
Results of Distribution of spatial entities and relationships

Spatial Entity			Spatial Relation			
Location	Natural	Building	Direction	Orientation	Topological	Distance
301	45	265	88	58	39	12

Table 5
Evaluation Annotation Metrics

Source	Correct	Incorrect	Missing
Newspaper	635	64	109

- F-measure is defined as the harmonic mean of two factors, precision and recall. It is formally as:

$$F - mesure = (2 * Precision * Recall)/((Precision + Recall)) \tag{3}$$

The following Tables present the evaluation results for the extraction of information related to natural disasters. The corpus used for this evaluation is the same as described in the study (Hadji et al., 2024), comprising a total of 9,008 words extracted from four different Algerian newspapers. This corpus was annotated to identify 908 spatial information elements, of which 611 are spatial entities and 197 are spatial relationships. The results obtained (Table 3) show the distribution of annotations across the different newspapers and indicate that spatial entities make up the majority of annotations, representing 68.4% of the total, while spatial relationships account for 31.6%.

The results (Table 4) show that the most frequently annotated spatial entities are non-natural objects (265) and places (301), while natural objects are less represented. This may reflect a focus on entities deemed more relevant in the contexts of the newspapers studied. Regarding spatial relationships, directional relationships are the most commonly annotated (88), followed by orientation relationships (58). Topological and distance relationships are much less frequent, which could indicate that they are considered less important or less complex in the analyzed corpora.

The following Table 5 shows the number of correct, incorrect, and missing annotations for Algerian newspapers. This data provides an assessment of the accuracy of spatial annotations performed on press articles, offering an overview of the quality of the results obtained during the extraction process.

6. Analysis and Discussion

In evaluating the effectiveness of our proposed JAPE rule-based approach for extracting spatial information from Arabic texts, we compared our results with those of various other methodologies, including rule-based and hybrid approaches. The following table summarizes the precision, recall, and F-measure of each method.

6.1. Analysis

- **Precision:** Our approach demonstrates a high precision of 0.90, indicating that 90% of the spatial information extracted is relevant and correct. This is a significant advantage, especially

Table 6
Performance Evaluation Metrics

	Precision	Recall	F-measure
Our Approach	0.90	0.85	0.87
[4] Based Rule	0.80	0.91	0.85
[19] Based Rule	0.85	0.88	0.85
[1] Hybrid	0.93	0.95	0.94

in applications where accuracy is paramount, such as in disaster response scenarios or when processing sensitive geographic data. In comparison, the rule-based methods [4] and [19] show lower precision values of 0.80 and 0.85, respectively. This discrepancy suggests that while these methods can recall a broader range of information, they may also include a higher number of false positives.

- **Recall:** In terms of recall, our approach achieves a score of 0.85, which indicates a solid ability to capture a significant proportion of the actual relevant spatial information present in the texts. Although this recall rate is lower than that of [4] (0.91), it remains competitive, especially when considering that higher recall often comes at the cost of lower precision. The approaches based on rules [19] and the hybrid method exhibit similar performance levels in recall, with scores of 0.88 and 0.95, respectively. This suggests that while our method may miss some relevant entities compared to the others, it does so while maintaining a high level of accuracy.
- **F-measure:** The F-measure, which balances precision and recall, is another critical metric for assessing the overall performance of the approaches. Our method achieves an F-measure of 0.87, which reflects a strong performance overall. The hybrid approach [1] leads in this area with an impressive F-measure of 0.94, underscoring the effectiveness of combining different techniques to leverage their respective strengths. While our approach does not outperform this hybrid model, it still outshines the purely rule-based approaches, [4] and [19], which both yield lower F-measure scores of 0.85.

6.2. Discussion

The analysis indicates that while our JAPE rule-based method excels in precision, making it a robust option for applications that require accuracy, it falls slightly behind in recall compared to some other methods. This presents a critical trade-off in the context of information extraction: achieving a high precision often limits the breadth of recall. The hybrid method, although potentially more complex to develop and implement, demonstrates the highest overall effectiveness, suggesting that a combined strategy could yield the best results in future applications.

Moving forward, our findings advocate for a nuanced approach to spatial information extraction that considers the specific requirements of each task. For instance, in scenarios where precision is paramount, our JAPE method stands out as an ideal choice. Conversely, for applications requiring extensive coverage of information, exploring hybrid methodologies could enhance performance significantly. Further research could involve developing a hybrid model that integrates the best features of our JAPE approach with the comprehensive capabilities of hybrid and machine learning methods, aiming to improve both precision and recall without sacrificing efficiency.

7. Conclusion

This research introduced an innovative approach for extracting spatial information from Arabic texts within Geographic Information Systems (GIS), utilizing JAPE rule-based techniques. This methodological choice proved effective for annotating and identifying spatial entities, such as natural objects, artificial objects, and locations, as well as spatial relations, including distance, topology, orientation, and directional relationships. The use of JAPE rules presents several advantages: it simplifies the creation of specific linguistic patterns, making it a swift and suitable solution for systems with focused objectives where ambiguities are minimal. Thus, for targeted applications and well-defined contexts, the JAPE approach ensures reliable and systematic extraction of spatial information.

However, our study also highlighted the limitations of this approach in addressing the linguistic nuances of the Arabic language, which often require labor-intensive manual adjustments and advanced linguistic expertise. In comparison, ontology-based and machine learning methods, though promising in terms of generalization and adaptability, demand significant resources to build comprehensive ontologies and annotated datasets, making them less accessible for applications requiring rapid deployment.

In conclusion, our work underscores the relevance of the JAPE rule-based approach for extraction systems where simplicity and quick implementation are paramount. For future applications, it would be beneficial to explore the hybridization of this method with machine learning and deep learning techniques, aiming to combine their precision with the adaptability and contextualization capacities that these approaches offer. Such a combination could lead to more robust and versatile spatial information extraction systems, tailored to the diverse challenges presented by Arabic texts and GIS contexts.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Hadji, M.-K. Kholladi, N. Borisova, Enhancing spatial information extraction from arabic text: A hybrid approach with ontology and rule-based, Ingenierie des Systemes d'Information 29 (2024) 1261.

[2] A. J. Aguilar, A. Pinos-Navarrete, C. Domingo Jaramillo, M. L. de la Hoz-Torres, 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, Springer, 2024, pp. 401–415.

[3] K. R. Reddy, V. Sharma, M. Anusha, S. Jhade, B. Dhanasekaran, Progressive collaborative method for protecting users privacy in location-based services, in: MATEC Web of Conferences, volume 392, EDP Sciences, 2024, p. 01089.

[4] A. Feriel, M. Kholladi, Automatic extraction of spatio-temporal information from arabic text documents, Int. J. Comput. Sci. Inf. Technol 7 (2015) 97–107.

[5] T. Eftimov, B. Koroušić Seljak, P. Korošec, A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations, PloS one 12 (2017) e0179488.

[6] V. Makhija, S. Ahuja, Rule based text extraction from a bibliographic database., DESIDOC Journal of Library & Information Technology 38 (2018).

[7] S. Jusoh, A. Awajan, N. Obeid, The use of ontology in clinical information extraction, in: Journal of Physics: Conference Series, volume 1529, IOP Publishing, 2020, p. 052083.

[8] A. Alamoudi, A. Alomari, S. Alwarthan, et al., A rule-based information extraction approach for extracting metadata from pdf books, ICIC Express Letters, Part B: Applications 12 (2021) 121–132.

[9] D. Freitag, J. Cadigan, R. Sasseen, P. Kalmar, Valet: Rule-based information extraction for rapid deployment, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 524–533.

[10] N. Hassini, K. Mahmoudi, S. Faiz, A hybrid approach for spatial information extraction from natural language text, in: 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), IEEE, 2023, pp. 1–8.

[11] Q. Qiu, Z. Xie, K. Ma, Z. Chen, L. Tao, Spatially oriented convolutional neural network for spatial relation extraction from natural language texts, Transactions in GIS 26 (2022) 839–866.

[12] Y. Liao, J. Hua, L. Luo, W. Ping, X. Lu, Y. Zhong, Aprcoie: An open information extraction system for chinese, SoftwareX 26 (2024) 101649.

[13] X. Zhao, A. Rios, Utsa-nlp at chemotimelines 2024: Evaluating instruction-tuned language models for temporal relation extraction, in: Proceedings of the 6th Clinical Natural Language Processing Workshop, 2024, pp. 604–615.

[14] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, ACM Computing Surveys 56 (2023) 1–62.

[15] D. Thakker, T. Osman, P. Lakin, Gate jape grammar tutorial, Nottingham Trent University, UK, Phil Lakin, UK, Version 1 (2009).

[16] Gate.ac.uk, JAPE: Regular Expressions over Annotations, https://gate.ac.uk/sale/tao/splitch8.html, 2024. Last accessed July 23, 2024.

[17] A. Hadji, M. K. Kholladi, Advanced nlp methods for disaster information extraction: Analyzing jape rules, ontologies, and machine learning approaches, in: Proceedings of the 3rd International Conference on Computer Science's Complex System and their Application (CCSA'2024), Computer Science Book Series, Springer Nature, 2024. In press.

[18] A. Hadji, M.-K. Kholladi, Automatic opinion extraction from football-related social media: A gazetteer and rule-based approach, NCAIA'2023 (2023) 61.

[19] S. Panda, A. Pradhan, V. Behera, A. Mohanty, A rule-based information extraction system, International Journal of Innovative Technology and Exploring Engineering 8 (2019) 1613–1617.