World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Arabic natural language processing: Data science application in big data environment

Sherif M. Saif *

Department of Computers and Systems, Electronics Research Institute, Cairo, Egypt.

World Journal of Advanced Research and Reviews, 2024, 24(02), 2283–2293

Publication history: Received on 16 October 2024; revised on 22 November 2024; accepted on 25 November 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.2.3602

Abstract

In the era of Big Data and Data Science, Text analysis within, Natural Language Processing (NLP), suffers from the curse of high dimensionality. The use of NLP in applications such as speech processing, semantic webs, and word processing has become a main element in today's Artificial Intelligence and Big Data Applications. A natural language parsing system must incorporate three components of natural language, namely, lexicon, morphology, and syntax. As Arabic is highly derivational, each component requires extensive exploitation of the associated linguistic characteristics. Parsing Arabic sentences still has open challenges due to several reasons including the relatively free word order of Arabic, the length of sentences, and the omission of diacritics (vowels) in written Arabic and the frequency of pro-drop phenomena. This research exploits Visual Prolog to provide a scalable platform for Arabic parser and explains the details of the used lexicon and parser and shows the scalability of the system to address more functions.

Keywords: Arabic NLP; Data Science; Big Data; Prolog; Parser

1. Introduction

Natural Language Processing (NLP) has made significant strides in recent years, with applications ranging from sentiment analysis to machine translation. However, languages like Arabic, with their complex morphological and syntactic structures, pose unique challenges for NLP tasks. The inherent ambiguity in Arabic language, coupled with the lack of standardized resources, further complicates the development of effective NLP systems [1].

To address these challenges, a robust and scalable NLP pipeline is essential. Visual Prolog, a powerful declarative programming language, offers a suitable framework for developing such a pipeline [2]. By leveraging its expressive power and logical reasoning capabilities, Visual Prolog can be used to implement various NLP components, including:

- Tokenization: Breaking down text into individual words or tokens [3].
- Part-of-Speech Tagging: Assigning grammatical categories (e.g., noun, verb, adjective) to words.
- Lemmatization: Reducing words to their base forms (e.g., "runs" to "run") [4].
- Stemming: Extracting the root or stem of a word [5].
- Named Entity Recognition (NER): Identifying and classifying named entities (e.g., persons, organizations, locations) [6].
- Dependency Parsing: Analyzing the grammatical structure of sentences [7].

Visual Prolog's declarative nature and strong pattern-matching capabilities make it well-suited for implementing complex linguistic rules. By formalizing linguistic knowledge as logical rules, Visual Prolog can effectively handle the challenges posed by Arabic language, such as:

* Corresponding author: Sherif M. Saif

- Handling Arabic Morphology: Visual Prolog can be [8] used to define morphological rules and patterns to accurately analyze Arabic words.
- Addressing Word Order Variation: By using sophisticated parsing techniques, Visual Prolog can handle the flexible word order in Arabic sentences.
- Dealing with Diacritics: While diacritics can enhance the clarity of Arabic text, their absence can introduce ambiguity. Visual Prolog can employ statistical techniques and linguistic knowledge to infer missing diacritics.
- Handling Pro-drop Phenomena: Visual Prolog can be used to identify and resolve pro-drop references, where pronouns are omitted.

By addressing these challenges, Visual Prolog can contribute to the development of more accurate and robust Arabic NLP systems.

To make the designs that will be used in Visual Prolog implementation, Augmented Transition Networks (ATN) are a widely used approach for representing the language analysis [9] [10].

This section serves as an introduction to the paper. The remaining sections of the paper are organized as follows: section 2 provides a background about Visual Prolog, section 3 discusses challenges of Arabic Parsing, section 4 presents the proposed system and explains its implementation; section 5 presents the output of the system, and section 6 concludes the research and presents future work.

2. Visual Prolog: Background and Significance

Visual Prolog is a commercial programming language that builds upon the core principles of logic programming, particularly Prolog. It offers a unique blend of logical, functional, and object-oriented programming paradigms, making it a versatile tool for a variety of applications [11].

Visual Prolog aims to simplify the development of complex software systems by providing a visual development environment and a high-level language. It emphasizes declarative programming, where programmers focus on defining the problem rather than specifying the exact steps to solve it. This approach can lead to more concise and elegant solutions, especially for tasks involving knowledge representation and reasoning [12].

Visual Prolog has gained popularity in various domains, including:

- Artificial Intelligence: Developing expert systems, knowledge-based systems, and natural language processing applications.
- Software Engineering: Creating domain-specific languages (DSLs) and generating code.
- Business Applications: Building enterprise applications, such as CRM and ERP systems.
- Education: Teaching programming concepts and logical reasoning.

Key Features of Visual Prolog:

- Declarative Programming: Focuses on specifying what the program should do, rather than how to do it.
- Strong Typing: Enforces type safety, reducing the likelihood of runtime errors.
- Object-Oriented Programming: Supports object-oriented concepts like classes, inheritance, and polymorphism.
- Visual Development Environment: Provides a user-friendly interface for designing and debugging applications.
- Integration with other Technologies: Can be integrated with other programming languages and frameworks.

By combining the power of logic programming with a visual development environment, Visual Prolog offers a unique and effective approach to software development.

3. Arabic Language: Challenges and Status Quo

Arabic Natural Language Processing (ANLP) presents unique challenges due to the language's complex morphology, rich dialectal variations, and the absence of diacritics in written text [13][14].

There are some key challenges in ALP such as:

- Morphological Complexity: Arabic is a morphologically rich language with complex word structures. This poses challenges for tasks like tokenization, stemming, and lemmatization [13].
- Diacritic Ambiguity: The absence of diacritics in written Arabic can lead to ambiguity in word segmentation and part-of-speech tagging [15].
- Dialectal Variation: Arabic has numerous dialects, each with its own unique linguistic features. This diversity makes it difficult to develop general-purpose NLP models [16].
- Lack of High-Quality Corpora: The availability of high-quality annotated Arabic corpora is limited, hindering the development of accurate and robust NLP models [17].

Despite these challenges, significant progress has been made in Arabic NLP in recent years. Researchers and developers have explored various approaches to address these issues:

Data-Driven Approaches [18]:

- Large Language Models: Models like BERT and GPT-3 have been adapted to the Arabic language, achieving impressive results on tasks like text classification, sentiment analysis, and machine translation.
- Pre-trained Language Models: Pre-trained models provide a strong foundation for various NLP tasks, reducing the need for large amounts of training data.

Rule-Based and Statistical Approaches [19]:

- Rule-based systems: Rely on handcrafted rules to analyze and process Arabic text.
- Statistical methods: Utilize statistical techniques to learn patterns from large amounts of data.

Hybrid Approaches [20]:

- Combining Rule-Based and Statistical Methods: Hybrid approaches leverage the strengths of both rule-based and statistical methods to improve accuracy and robustness.

4. Proposed System: Theory, Design, and Implementation

4.1. Theory and background

Arabic is rooted in the Classical or Quranic Arabic. The Arabic language is written from right to left, has 28 letters, that form words generally classified into three main categories: noun, verb, and particle, these categories form two types of sentences: nominal and verbal sentence [21].

Agreement is a major syntactic principle that affects the generation of an Arabic sentence. Agreement in Arabic is full or partial and is determined by word order. An adjective in Arabic usually follows the noun it modifies ("الموصوف") and fully agrees with it with respect to number, gender, case, and definiteness.

The definiteness agreement rules are applied only on the noun stems, if the stem (noun) was defined or undefined or proper name. Adjectives in Arabic usually follow nouns and agree with them in terms of number, gender, and definiteness/indefiniteness.

Let's review the following three combinations that look similar when written in Arabic

- A beautiful cat
- The beautiful cat
- The cat is beautiful

Writing the Arabic sounds in English; the three examples will be:

- Qitta Jamila
- Al-Qitta Al-Jamila
- Al-Qitta Jamila

The sound “Qitta” in Arabic means “cat”. The sound “Jamila” in Arabic means beautiful. The sound “Al” in Arabic means “the”. In Arabic language the following rules apply:

- If an adjective completely agrees with its noun in every aspect, then you have a phrase or a fragmented sentence, as in examples (a) and (b).
- If a noun (subject) is definite and its adjective (predicate) is indefinite, you have a proper complete sentence, as in (c).

This can be noted by translating the above three examples:

- “A beautiful cat”: This is a phrase; it is not a complete sentence
- “The beautiful cat”: This is a phrase; it is not a complete sentence
- “The cat is beautiful”: This is a complete proper sentence

There are two main types of sentences in Arabic:

- A nominal statement is a statement that begins with a subject.
- A verbal statement is the one that begins with a verb.

The focus of this research will be the Verbal statements and the Nominal statements that contain verbs.

4.1.1. Nominal statements

In Subject-Verb-Object (SVO) order, the verb agrees with the subject with respect to number and gender.

4.1.2. Verbal statements

The verb in Verb-Subject-Object (VSO) order agrees with the subject in gender only.

The following factors define the agreement parameters:

Number: This means that the subject or the verb refer to

- One person: Singular
- Two persons: Dual
- Three or more persons: Plural

Gender: The word structure vary when referring to:

- Feminine
- Masculine

Person: The person doing the verb may be:

- 1st person: Speaker
  - This means that the verb is referring to the speaker. When I say “I came” this is a 1st type version.
- 2nd person: Listener
  - When the statement says “You came” this is a second type.
- 3rd person: Absent
  - The absent form of the verb such as “He came” or “She came”

The above notations for number, gender and person will be used in the following sections.

4.2. Design of the System

The presented system is composed of two main components, the lexicon and the parser. Figure 1 shows the design of the system.

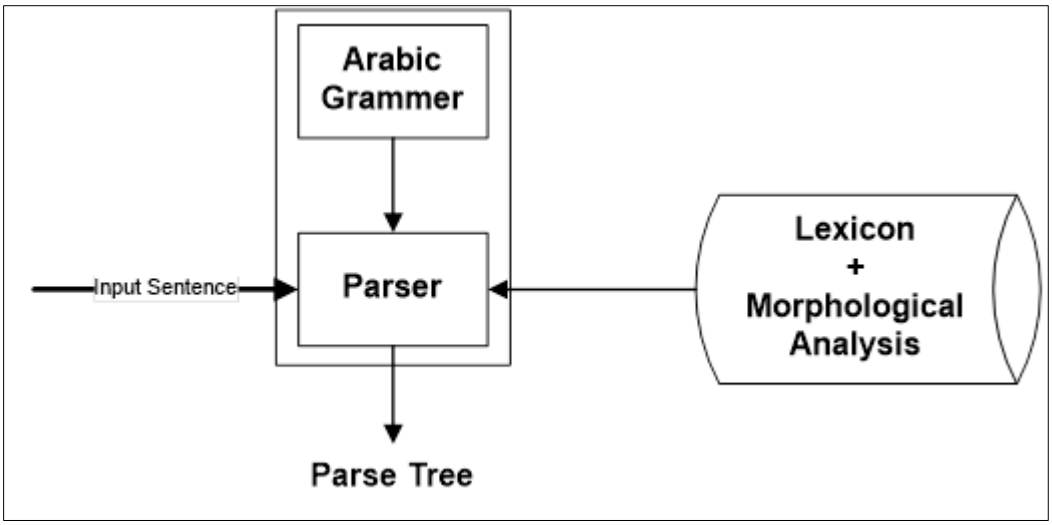

Figure 1 The Architecture of the System

The detailed implementation of the two components is detailed in the following section.

4.3. Implementation Approach

This project is implemented using prolog, PIE Prolog in specific is used since it's an excellent system to start with as it is easy to use and has a standard Windows-style user interface. By comparison with many other Prolog systems, however, PIE Prolog is relatively unsophisticated. It has only a small number of language constructs ('predicates') available and does not have any visual design or debugging tools. Prolog is particularly good at making inferences or deductions from sets of facts and rules. This makes it especially well-suited for the development of complex applications such as the interpretation of human language or diagnostic 'expert systems' such as the presented system.

The following predicates are used in the Prolog code:

Table 1 Acronyms that will be used in the Prolog code

Acronym in the code	Meaning
masc	Masculine
femin	Feminine
def	Definite
indef	Indefinite
inf	Infinitive
ppnoun	Prepositional particles
snoun	Demonstrative pronouns
adj	Adjective
intr	Intransitive
tran	Transitive
tr2	Di-transitive

4.4. System Components

4.4.1. Lexicon

The lexicon is used to simulate the output of the morphological analysis and makes use of the agreement rules defined in section 1.

- **Verbs:** A verb has the following features:
  - **Person:** first, second and third person
  - **Number:** singular, dual, plural
  - **Gender:** masculine, feminine
  - **Tense:** past, present
  - **Transitivity:** intransitive / transitive_1_obj / transitive_2_obj
- **Nouns:** A noun has the following features:
  - **Gender:** masculine, feminine
  - **Number:** singular, dual, plural
  - **Humanness:** human, non-human
  - **Definiteness**
- **Adjectives:** An adjective has the following features:
  - **Gender:** masculine, feminine
  - **Number:** singular, dual, plural
  - **Definiteness**
- **Pronouns:** A pronoun has the following features:
  - **Number:** singular, dual, plural
  - **Gender:** masculine, feminine
  - **Person:** first, second and third person

Adverbs, and prepositions have no internal features.

The presented lexicon defines the following particles:

The Personal pronouns: are used to replace nouns. The following is a list of the Singular, dual and plural forms:

- Singular: I, you (masc), you (femin), he, she
- Dual: You (for two), they (for two)
- Plural: We, You (for plural masc), You (for plural femin), They (for plural masc), They (for plural femin)

Demonstrative Pronouns: The use of "this/that & these/those" in Arabic is determined by the number and gender of the noun/adjective they introduce.

- Singulars are: This (masc), and This (femin)
- Dual ar : These (dual masc), and These (dual femin)
- Plurals are: Those

Prepositions:

- To, From, On, In

Adverbs:

- Here, There, Under, Above, At

4.4.2. Parser

There are three types of sentences that our parser expresses in the output form of a parse tree, these sentences are verbal, nominal and equational sentences as will be illustrated bellow.

In order to facilitate the writing of the Arabic grammar rules the Augmented Transition Networks (ATN) [22] to provide us with the different possibilities of the Arabic grammar rules.

Figure 2 illustrates the ATN of the nominal sentence.

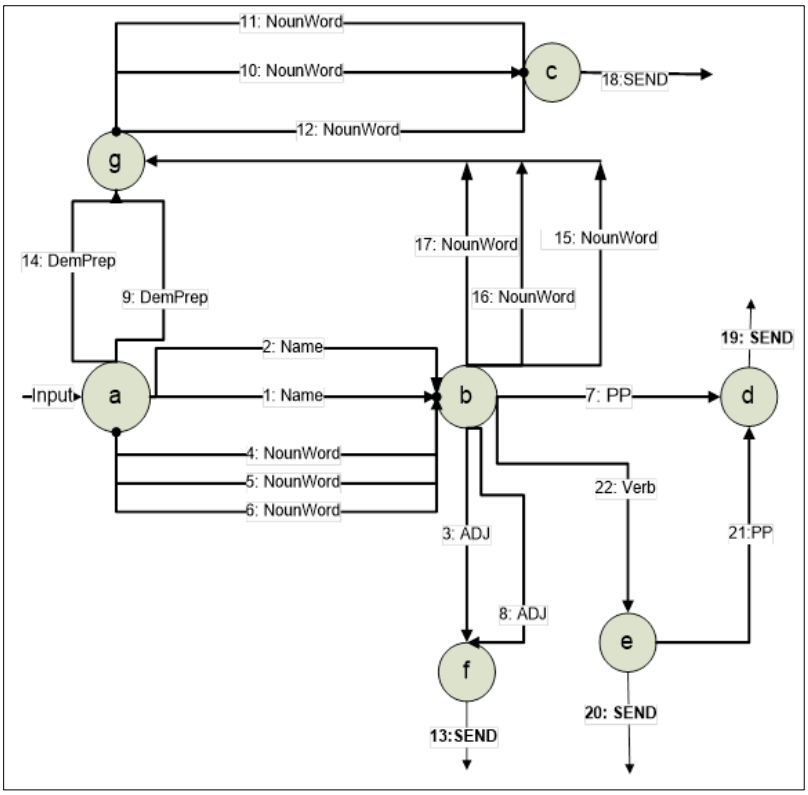

Figure 2 ATN of the Nominal Sentence

Verbal sentence

A verbal sentence begins with a verb. The subject comes after the verb. It is either apparent or implicit. After the subject there are several options, either the acted-upon-noun appears or there can be a semi-sentence which begins by a preposition. Figure 3 illustrates the ATN of the verbal sentence

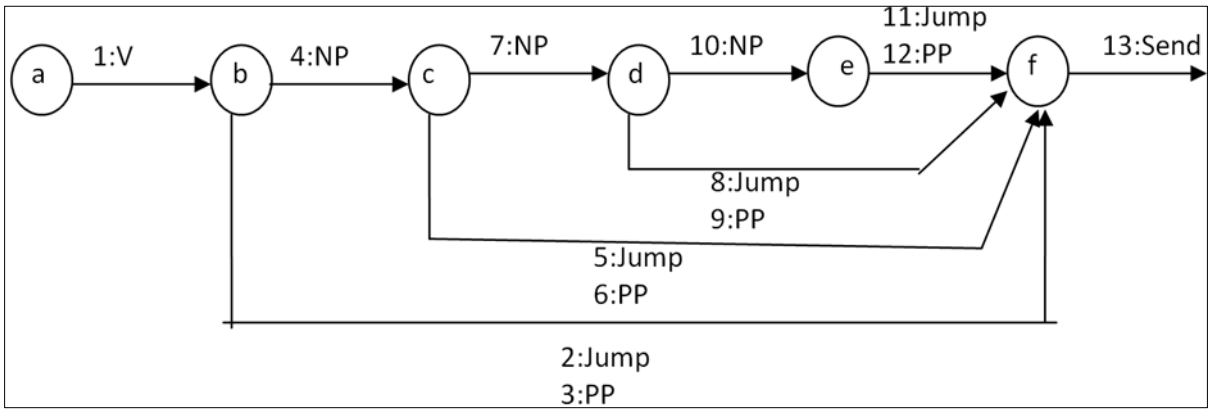

Figure 3 ATN of the Verbal Sentence

There are various types of sentences:

- The verbal sentence is the basic sentence.

Its order is **verb → subject → (object).** In this type of sentence, a verb is marked by the gender of its subject and an object is optional. The presence of the object depends on the type of the verb whether it is transitive or intransitive.

- The Nominal Sentence is where the subject takes an initial position for emphatic purposes, followed by the verb, its order is **subject→verb→(object)**. Consequently, the verb is marked by the number and gender of its subject. The presence of an object depends on the type of the verb whether it is transitive or intransitive.

- The Equational Sentence is made of a subject and a predicate without any expressed verb. The verb "to be" is understood, **subject → predicate**. Both the subject and the predicate have to be in the nominative case.

  NP ➔ Arabic_Subject, Arabic_Predicate

  Arabic_Subject ➔ Noun; Pronoun; Demonstrative pronoun

  Arabic_Predicate ➔ Adjective; Adverb; Preposition phrase; Noun; Verb phrase

Table 2 Grammar Rules

Grammar Rule: First part of the phrase	**Grammar Rule: Secpnd part of the phrase**
Noun (indefinite)	adjective must be (indefinite)
Noun (definite)	adjective (indefinite) ; adjective (definite)
Arabic_Subject(noun,sing,femin,indef)	Arabic_Predicate{adjective,sing,femin,indef}
Arabic_Subject(noun,sing,femin,def)	Arabic_Predicate{adjective,sing,femin,indef}
Arabic_Subject(noun,sing,masc,def)	Arabic_Predicate{adjective,sing,masc,def}
Arabic_Subject(noun,definite)	Arabic_Predicate{preposition phrase}
Arabic_Subject(noun,definite)	Arabic_Predicate{adverb phrase}
Arabic_Subject(Demonstrative Pronouns,sing,femin)	Arabic_Predicate{noun(sing,femin,indef)}

5. Experimentation Output and Results

Test patterns are fed to the system and if the pattern matches an ATN flow it is a valid sentence and it is classified, otherwise it is an invalid sentence.

5.1. Testing Patterns

Patterns similar to the following are used to test the system:

- The cat is beautiful.
- The tree is beautiful.
- This is a cat.
- Ahmed is in the park.

5.2. Identified Patterns: Nominal Sentences

For Nominal Sentences, figure 4 show examples of the output:

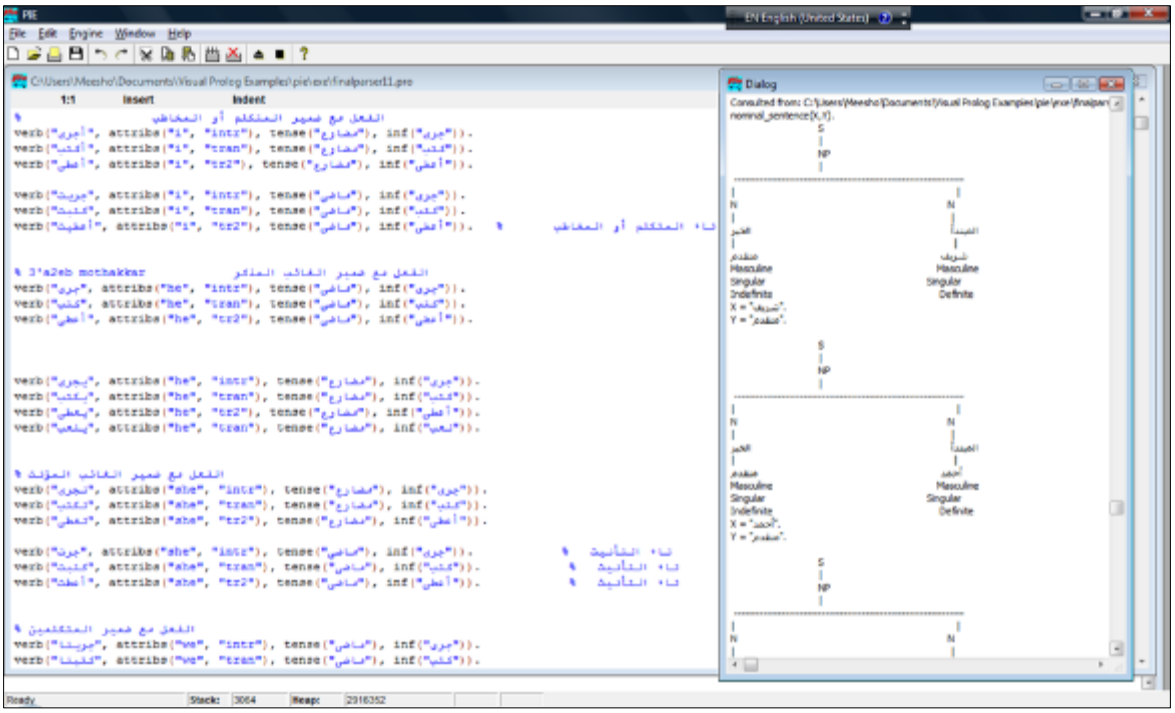

Figure 4 System Output for Nominal Sentences

5.3. Identified Patterns: Verbal Sentences

For Verbal Sentences, figure 5 shows examples of the output:

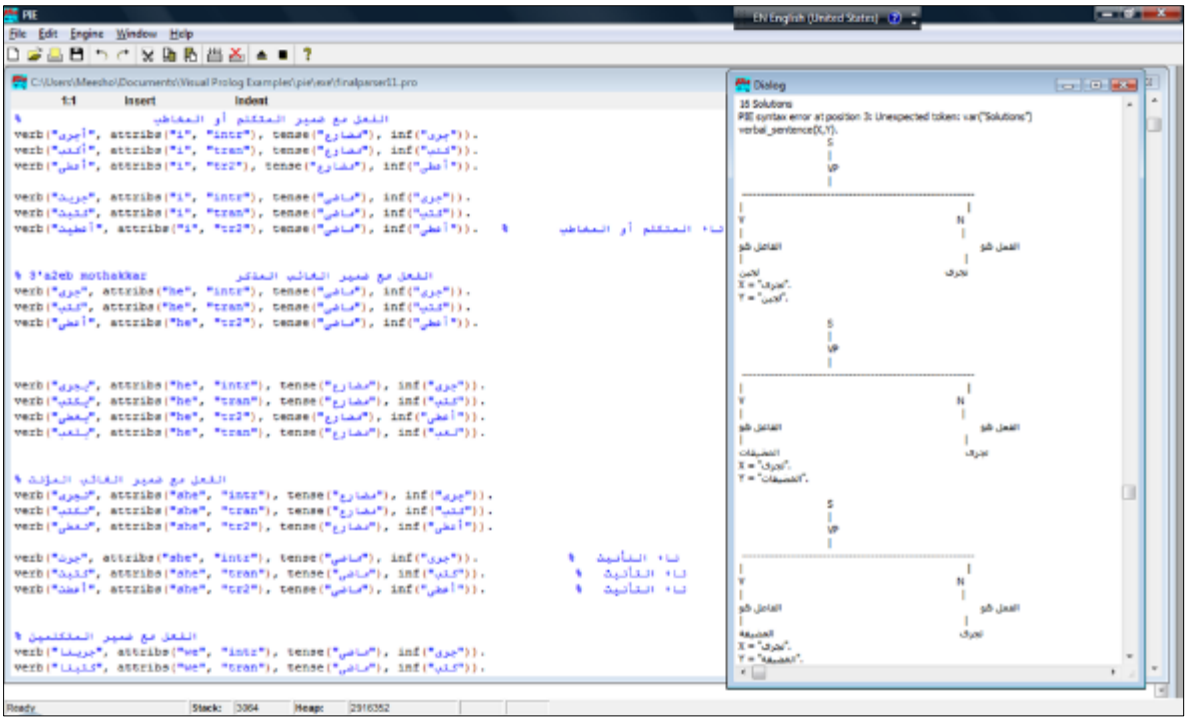

Figure 5 System Output for Verbal Sentences

6. Conclusion

This research has demonstrated the effectiveness of Visual Prolog as a powerful tool for implementing a natural language parser, specifically for Arabic. By leveraging its declarative programming paradigm and strong pattern-

matching capabilities, this work have successfully addressed the complex linguistic challenges posed by Arabic, including its rich morphology and flexible word order.

The implementation of the lexicon and parser components has been detailed, showcasing the use of Visual Prolog to represent and process Arabic linguistic structures. The system has been shown to accurately handle various sentence constructions, including nominal sentences with embedded verbs and verbal sentences with different verb types.

While the presented system has achieved promising results, there are still opportunities for further improvement. Future work could focus on enhancing the system's accuracy by incorporating more sophisticated linguistic knowledge and machine learning techniques. Additionally, expanding the system to handle a wider range of Arabic dialects and domains would be a valuable contribution.

By continuing to refine and extend this research, more advanced Arabic NLP systems can be developed, that can better understand and process natural language, ultimately leading to more sophisticated applications and services.

Future Directions

Future research directions in Arabic NLP include:

- Improving the quality and quantity of Arabic language resources.
- Developing more advanced language models and techniques.
- Addressing the challenges of low-resource dialects.
- Exploring the potential of multilingual NLP models.

By leveraging the latest advancements in NLP, researchers can develop more sophisticated and accurate Arabic NLP systems, enabling a wide range of applications, from information retrieval to machine translation.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Alrayzah, Basmah, Ammar A. Alrayes, Mohammad W. Alrayes, et al. Challenges in Arabic NLP: Morphological Richness, Orthographic Ambiguity, and Dialectal Variations. *PeerJ Computer Science*, 2023. DOI: 10.7717/peerj-cs.1633.

[2] Hollander, Judd E. *A Guide to Artificial Intelligence with Visual Prolog*. Gabon: Salon Numérique Gabon, n.d. Accessed November 19, 2024. https://salonnumeriquegabon.campusfrance.org, Salon Numérique Gabon

[3] Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 2020.

[4] Straka, Milan. Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing, and NER. *Proceedings of Text, Speech, and Dialogue (TSD)*, 2020.

[5] Kondratyuk, Daniel, and Milan Straka. Towards JointUD: Part-of-Speech Tagging and Lemmatization Using Recurrent Neural Networks. *Proceedings of the CoNLL 2018 Shared Task*, 2018.

[6] Jivani, Anjali G. A Comparative Study of Stemming Algorithms. *International Journal of Computer Technology and Applications* 2, no. 6 (2011): 1930–1938.

[7] Straka, Milan. Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing, and NER. *Proceedings of Text, Speech, and Dialogue (TSD)*, 2020.

[8] Eisenstadt, Marc, and Ian Brayshaw. Visual Representations for Prolog Programming. *Proceedings of the Psychology of Programming Interest Group (PPIG)*, 1987. Accessed November 19, 2024. http://www.ppig.org

[9] Fifty Years of Prolog and Beyond. *Association for Logic Programming*. Accessed November 19, 2024. https://alplogic.org, Cambridge University Press & Assessment

[10] Colmerauer, Alain, and Philippe Roussel. The Birth of Prolog. *Communications of the ACM* 28, no. 4 (1986): 319–322. Accessed November 19, 2024. https://doi.org/10.1145/6327.

[11] International Journal of Computer Applications. Using Augmented Transition Network for Morphological Processing of Arabic. *International Journal of Computer Applications* 25, no. 10 (July 2011): 21–28. Available at: IJCA Online

[12] Woods, William A. Transition Network Grammars for Natural Language Analysis. *Communications of the ACM* 13, no. 10 (1970): 591–606.

[13] Darwish, Kareem, Nizar Habash, Mourad Abbas, et al. A Panoramic Survey of Natural Language Processing in the Arab World. arXiv preprint arXiv:2011.12631 (2020). This comprehensive survey examines the development, challenges, and future directions of Arabic NLP.

[14] Al-Sarayreh, Sallam, Azza Mohamed, and Khaled Shaalan. Challenges and Solutions for Arabic Natural Language Processing in Social Media. In Business Intelligence and Information Technology, 293–302. Springer, 2023.

[15] Mhamdi, Abderrahim, Marwan Tuffaha, Abdullah Rashwan, and Sherif Abdou. Exploring Large Language Models for Arabic AI: Benchmarks and Innovations. *arXiv preprint*, 2023.

[16] Guellil, Imane, Omar Nouali, and Abdelaziz Mourad. Arabic Dialect Processing: Advancements and Opportunities. *Natural Language Engineering*, 29 (2023): 315-330.

[17] Alyafeai, Zaid, Mohammed AlShaibani, and Ismail A. Ahmad. Arabic NLP: Current Trends, Challenges, and Future Prospects. *Journal of Computational Linguistics* 49, no. 1 (2023): 123-145.

[18] Koubaa, Anis, Adel Ammar, Lahouari Ghouti, et al. ArabianGPT: Native Arabic GPT-based Large Language Model. arXiv preprint arXiv:2402.15313 (2024).

[19] Alhaj, Mohammed. NLP Challenges in Arabic Language. Medium, May 2023.

[20] Alhafni, Bashar, Sarah Al-Towaity, Ziyad Fawzy, et al. Exploiting Dialect Identification in Automatic Dialectal Text Normalization. arXiv preprint arXiv:2407.03020 (2024).

[21] Zaki, Asma, Amira Ben Hamadou, and Ahmed Serhrouchni. Morphological Challenges in Arabic Text Mining. *International Journal of Advanced Computer Science and Applications* 14, no. 2 (2023): 44-51.

[22] Proceedings of the National Conference on Recent Innovations in Emerging Computer Technologies. Natural Language Processing Algorithms and ATNs. Presented at NCRIECT 2023, Raipur: Kalinga University. ISBN 979-8-3938-8807-7, May 2023.