Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 09, No. 03, September 2023.

TRENDS AND CHALLENGES OF ARABIC CHATBOTS: LITERATURE REVIEW

Yassine Saoudi[1] and Mohamed Mohsen Gammoudi[2]

(Received: 29-May-2023, Revised: 27-Jul.-2023 and 19-Aug.-2023, Accepted: 20-Aug.-2023)

ABSTRACT

Conversational systems have recently garnered increased attention due to advancements in Large Language Models (LLMs) and Language Models for Dialogue Applications (LaMDA). However, conversational Artificial Intelligence (AI) research focuses primarily on English. Despite Arabic being one of the most widely used languages on the Internet, only a few studies have concentrated on Arabic conversational dialogue systems thus far. This study presents a comprehensive qualitative analysis of critical research works in this domain to examine the strengths and limitations of existing approaches. The analysis begins with an overview of chatbot history and classification, then explores the language challenges encountered when developing Generative Arabic Conversational AI. Rule-based/Retrieval-based and deep learning-based approaches for Arabic chatbots are also examined. Furthermore, the study investigates the evolution of Generative Conversational AI with the advancements in deep-learning techniques. It also comprehensively reviews various metrics used to assess conversational systems.

KEYWORDS

Chatbot, LLMs, Arabic conversational AI challenges, Generative artificial intelligence, Arabic question answering systems, Taxonomy, Performance evaluation.

1. INTRODUCTION

Conversational agents, commonly known as chatbots, have become integral to our daily lives. They serve as personal assistants on mobile phones, salespeople on e-commerce sites [1] and healthcare assistants [2], engaging in consistent conversations with humans using natural language in text or voice format. While chatbots have been around for decades, recent advancements in artificial intelligence, particularly in human-language processing, have created more efficient, faster and more powerful bots [1], [3]. The field of conversational systems has gained significant attention, driven by the development of Large Language Models (LLMs) and Language Models for Dialogue Applications (LaMDA). However, it is essential to note that conversational systems encompass a broader range of technologies and approaches beyond LLMs and LaMDA.

Chatbots, considered to have specific goals, can be used in many domains, such as in education [4]-[11] and in healthcare [12]-[15]. According to [11], [16]-[18], chatbots reduce the response time to questions; improve customer service; order products online (Alexa from Amazon [19]); research information; guide the user Rahhal [20] (helping tourists at Saudi Arabia) and assist in flight booking [21] (airline ticket booking). Also, solving technical challenges (1+2) is the key to a successful chatbot as: (1) The relevance of the answer: understanding the user's need and not just words and grammar, (2) The structure of the answer: the development of a conversational structure for the user to be comfortable. The dimensions of speed and efficiency appear several times: chatbots must be fast and efficient [22]. They are the most convenient way to deal with consumers in a timely and satisfying way [17].

Based on our extensive reading and research, we highly recommend clearly defining a chatbot: A chatbot is a software application, with or without an avatar, specifically designed to enable conversations using natural language. It utilizes various idioms to serve a specific purpose, aiming to deliver precise information and create a user experience that closely resembles interacting with real individuals regarding efficiency, speed and effectiveness.

Generative Artificial Intelligence (AI) has made remarkable advancements in revolutionizing our lifestyle, work dynamics and interactions with technology. One area that has recently seen significant progress is the development of Large Language Models (LLMs). One prominent example is the

[1] Y. Saoudi is with University of Tunis El Manar, Tunis, Tunisia. Email: yassine.saoudi@fst.utm.tn

[2] M. M. Gammoudi is with ISAM Manouba University, Tunisia. Email: gammoudimomo@gmail.com

Generative Pretrained Transformer (GPT) family, which includes GPT-1 developed by OpenAI in 2018 [23], GPT-2 in 2019 [24], GPT-3 in 2020 [25] and the latest addition, GPT-4 in 2023 [26]. Another notable model is BERT (Bidirectional Encoder Representations from Transformers), introduced by researchers at Google AI Language in 2018 [27]. In addition, LaMDA, a Language Model for Dialogue Applications, was introduced in 2022 [28]. These models have showcased their success in tasks, such as question-answering, text summarization, sentiment analysis and named entity recognition.

This paper presents a comprehensive review of Arabic conversational dialogue system research to identify critical gaps in the existing literature and propose future research directions. Unlike previous surveys, our review encompasses recent studies that address all aspects of the chatbot workflow. To ensure the logical coherence of the paper, we have established a clear roadmap that guides the organization of different sections. Additionally, our review investigates the progression of Arabic chatbot development with advancements in deep-learning techniques. Finally, unlike previous surveys that primarily focused on classification and reviewing existing chatbot systems, we go beyond that by delving into the evolution of deep-learning techniques, the challenges specific to Arabic language and the evaluation metrics for assessing Arabic conversational dialogue systems.

The rest of this paper is organized as follows: Section 2 outlines the methodology employed in conducting this study. Section 3 provides a synthetic background as well as chatbot classification and explores some applications of chatbots. Section 4 examines the challenges encountered in Arabic Conversational AI systems. Subsequently, Section 5 explores rule-based and retrieval-based Arabic chatbots. Section 6 focuses on deep learning-based Arabic Question Answering Systems. Progressing further, Section 7 investigates the evolutionary advancements of deep-learning techniques. The assessment metrics for conversational AI and Question-Answering systems are scrutinized in Section 8. Section 9 draws this survey to a close by offering a discussion of the research findings, while Section 10 presents the conclusions and highlights potential horizons future research.

2. METHODOLOGY

This survey examines various approaches employed in conversational dialogue systems and investigates the achievements and obstacles associated with building conversational AI dialogue systems for Arabic. Our review adheres to the systematic review guidelines outlined by Kitchenham and Xiao in [29]-[30]. Subsequently, we introduce the research questions (RQs) formulated for our systematic review of the problem mentioned above.

RQ1: What is the history of the evolution of the Arabic conversational system and what are the objectives of building conversational chatbots? This question is answered in Section 3. The primary objective of this research question is to illustrate the evolutionary progression of Generative Artificial Intelligence, delving into the development of chatbots and the diverse applications that researchers have explored, particularly in crucial areas, like healthcare or educational support.

RQ2: What approaches are used to perform Generative Arabic conversational AI? This question is answered in Section 7. The aim is to explore state-of-the-art deep-learning techniques.

RQ3: What are the evaluation criteria of the deep-learning techniques used in Arabic conversational AI systems? This question is answered in Section 8. The role of this research question is to specify the measures used to evaluate the deep-learning techniques in the Arabic QA systems.

RQ4: What are the major challenges in building Arabic conversational AI? This question is answered in Section 4 and Section 9. This research question encourages future researchers to explore language-specific techniques by highlighting the significant challenges associated with Arabic conversational AI. The papers reviewed in the various sections were gathered by querying multiple databases, including journal articles and conference proceedings published between 2000 and 2023. The literature collection used the widest publishers, such as IEEE, ACM, Springer, Elsevier, Wiley, Taylor & Francis. Moreover, we searched well-known databases, such as Scopus, Web of Science, DBLP and Google Scholar. Figure 1 shows the percentage of resulting papers per database and Figure 2 illustrates the search results in Scopus from 2000 to 2022 for the keywords TITLE-ABS-KEY ("chatbot*" OR "Generative Artificial Intelligence*” OR ”question answering*” OR ”question answering system*” AND ”Arabic”). The search terms for this review will use various combinations of search terms derived from the research questions. We derive the principal terms from the research questions about Arabic question answering, Arabic Chatbot and deep learning. The search terms consist of the advanced search string construction using identified keywords' search terms using Boolean operators AND/OR:

1) ("BERT" OR "GAN" OR "GPT" OR "Transformers" OR "Seq2Seq" OR "LSTM" OR "Transformers") AND ("question answering" OR "chatbot" OR "conversational agent" OR "Dialogue System”) AND "Arabic".
2) ("Generative Adversarial Networks" OR "Generative Neural Networks" "deep Bidirectional Transformers" OR "DBLSTM" OR “RNN" OR "CNN" OR "DBN" OR "DNN" OR "DANN") AND ("question answering" OR "chatbot" OR "conversational agent" OR "Dialogue System") AND "Arabic".
3) ("deep learning" OR "deep structured learning" OR "hierarchical learning") AND ("chatbot" OR "conversational agent" OR "Dialogue System") AND "Arabic".
4) ("GPT-3" OR "GPT-4" OR "LLMs" OR "LaMDA") AND ("question answering" OR "BARD" OR "conversational agent" OR "ChatGPT") AND "Arabic".

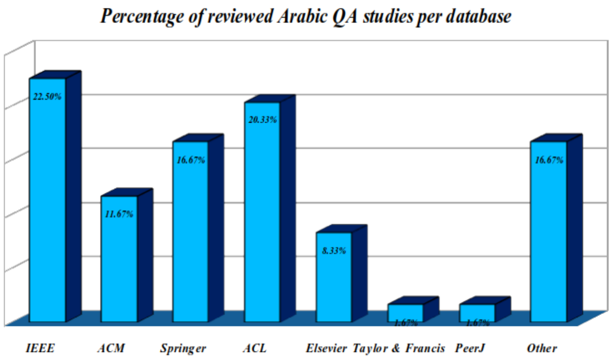

Figure 1. Resulting papers' percentage per database.

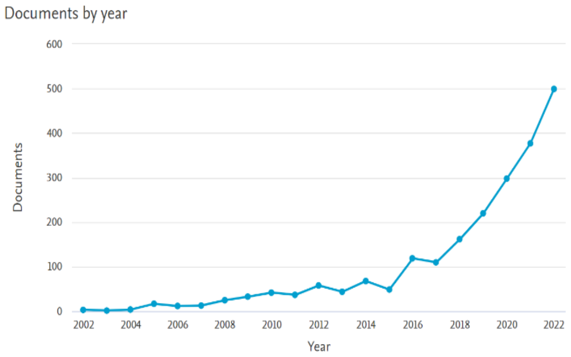

Figure 2. Search results in Scopus, from 2000 to 2022, for the keywords TITLE-ABS-KEY ("chatbot*" OR "Generative Artificial Intelligence*” OR ”question answering*” OR ”question answering system*” AND ”Arabic”).

3. CHATBOTS: HISTORY AND CLASSIFICATION

This section presents a comprehensive overview of chatbots, starting with a brief history of their evolution and highlighting the key milestones that have shaped their development. As we delve into this historical chronology, the transformative power of AI-driven advancements becomes evident. Next, we move on to a detailed classification of chatbots, examining various factors that categorize these intelligent entities. Finally, we showcase the use of chatbots in various sectors, highlighting their application areas.

3.1 Chatbots' History

In 1950, Alan Turing proposed the Turing Test ("Can machines think?") [31] to assess a machine's capacity to exhibit intelligent machine behavior similar to humans. To pass the Turing Test, the machine's responses must be indistinguishable from those of a human during a five-minute test. The origin of chatbots dates back to 1966 with the invention of ELIZA [32]. Eliza is a conversational agent in a very basic Rogerian psychotherapist. It was based on a template-based response mechanism and simple keyword matching. Many chatbots were developed after ELIZA, such as PARRY, an infamous chatbot, created in 1972 by Kenneth Mark Colby, a psychiatrist and computer scientist associated with Stanford's Psychiatry Department. JABBERWACKY in 1988 used contextual pattern-matching technique [33]. In 1990, the Loebner Prize was begun [34] (annual competition for chatbots based on Turing Test). In 1992, the authors of [35] developed Dr. Sbaitso's voice-based chabot.

Later in 1995, Richard Wallace [36] developed ALICE (Artificial Linguistic Internet Computer Entity). This chatbot has become significant, because it led to the development of Artificial Intelligence Markup Language (AIML) [37]. AIML is used to declare pattern-matching rules that link user-submitted words and phrases with topic categories. It is an eXtensible Markup Language (XML)-based language and supports most chatbot platforms and services today. ALICE won the Loebner prize in 2000, 2001 and 2004 [38]. Afterwards, many chatbots were developed based on the ALICE framework [39]. In 2001, SmarterChild chatbot [40] was developed to be compatible with instant messaging applications, such as MSN Messenger and American Online Services (AOI), Instant Messenger (IM) or American Instant Messenger (AIM). In 2005, based on rules written in AIML, Mitsuku (Kuki) [41] was the most widely used stand-alone human-like chatbot. Essential features of Mitsuku are general conversations, which can hold lengthy conversations and multilingual robots that can think logically about a given object. In the Mitsuku chatbot, human curators evaluate incoming data; only the validated data is recorded and used.

Since 2006, new virtual personal assistants have been developed, such as IBM Watson [42], (a rule-based AI chatbot that uses NLP and hierarchical ML methods to generate responses based on the score). Later, many chatbots have been developed, such as Apple Siri [43] (a speech-to-text bot dedicated to Apple products) in 2010, Google Assistant [44] in 2012, Amazon Alexa [19] in 2015, Dialogflow [45] developed by Google in 2016, LUIS [46] developed by Microsoft in 2017 and Amazon Lex [47] developed by Amazon in 2017.

Afterwards, specifically with the introduction of transformers' architecture in 2017 by Vaswani et al. [48], many language model based-transformers were developed, such as BERT (Bidirectional Encoder Representations from Transformers), offered by researchers at Google AI Language in 2018 [27] and GPT (Generative Pre-trained Transformer) developed by OpenAI in 2018 [23]. As a result, generative Artificial Intelligence (AI) has made remarkable advancements in revolutionizing our lifestyle, work dynamics and interactions with technology. Recently, one area that has seen significant progress is the development of Large Language Models (LLMs), such as GPT-3 [25], ChatGPT, GPT-4 [26] and LaMDA, a Language Model for Dialogue Applications [28]. Moreover, these models exhibit remarkable accuracy in tasks like text summarization and question-answering. In Section 7, we study this revelational technique in more detail. The evolution of the conversational agent is shown in Figure 3, along with the evolution of relevant techniques and approaches.

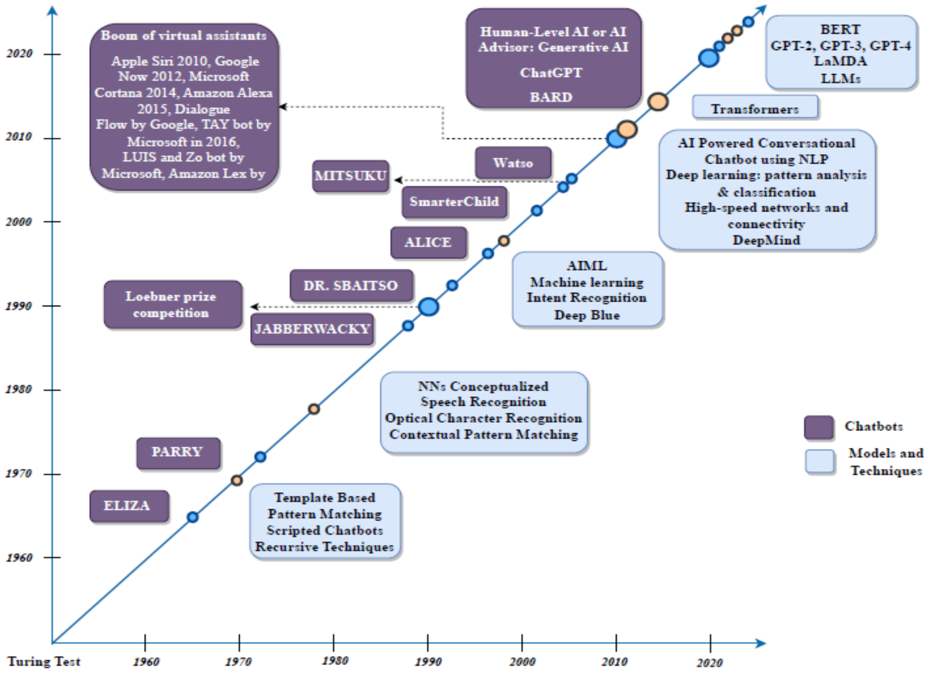
Boom of virtual assistants
Apple Siri 2010, Google Now 2012, Microsoft Cortana 2014, Amazon Alexa 2015, Dialogue Flow by Google, TAY bot by Microsoft in 2016, LUIS and Zo bot by Microsoft, Amazon Lex by
Human-Level AI or AI Advisor: Generative AI
ChatGPT
BARD
BERT
GPT-2, GPT-3, GPT-4
LaMDA
LLMs
Transformers
Watso
MITSUKU
SmarterChild
ALICE
AI Powered Conversational Chatbot using NLP
Deep learning: pattern analysis & classification
High-speed networks and connectivity
DeepMind
DR. SBAITSO
Loebner prize competition
JABBERWACKY
AIML
Machine learning
Intent Recognition
Deep Blue
NNs Conceptualized
Speech Recognition
Optical Character Recognition
Contextual Pattern Matching
PARRY
ELIZA
Template Based
Pattern Matching
Scripted Chatbots
Recursive Techniques
Chatbots
Models and Techniques
2020
2010
2000
1990
1980
1970
1960
Turing Test
1960
1970
1980
1990
2000
2010
2020

Figure 3. The evolution of chatbots: From ELIZA to generative AI chatbots based on AI advancements.

3.2 Chatbots' Classification

In the last few years, the chatbot field has become so dynamic with the emergence of new technologies that more intelligent systems have developed using complex knowledge-based models. Hence, chatbot classification is essential for scientists to compare and evaluate systems, define requirements and select the right tools. Figure 4 illustrates our comprehensive broad classification of chatbots, which is based on the main classification proposed by [33], [49]-[51].

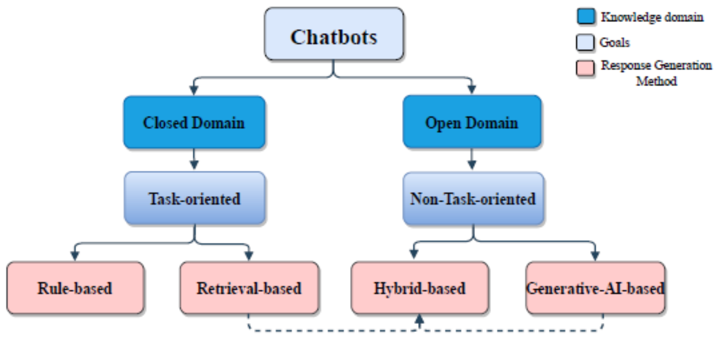

Figure 4. Broad classification of chatbots.

The first superficial level for chatbot classification is based on the following: **Types of questions** [(factoid (when/who/where), confirmation (yes/no), definition, causal (how/why/what), procedural, comparative, opinionated)], **types of knowledge sources** [(structured (RDF graphs, SQL database, CSV, JSON, XML data . . .), unstructured (plain text, e.g. Wikipedia)) and based on **input of user and output of chatbot**, chatbots can be categorized as shown in Table 1.

Table 1. Types of chatbots based user's input and chatbot output.

	Text	**Speech**
Text	Text-to-text bot (TTT): Example: *ELIZA* [32]	Text-to-speech bot (TTS): Example: *Alexa* [52]-[53]
Speech	Speech-to-text bot (STT): Example: *SIRI* [53]-[54]	Speech-to-speech bot (STS): Example: *Cortana* [53]

In Table 2, we see the four main kinds of chatbot classification: the knowledge domain, the response generation method, the goals and the service provided.

Table 2. Taxonomy of chatbot application.

Knowledge domain: Includes the knowledge that a chatbot can access or the amount of data that it is trained upon [49], [55].	Open-domain:A chatbot aiming to establish long-term connections with users who can talk about general topics and respond appropriately. Closed domain:A chatbot operates through information regarding a particular area of interest and aims to provide specific answers concerning only the particular knowledge domain. Generic:A chatbot can answer any user question from whichever domain.
Response-generation method: The distinction is based on the input-processing and response-generation method (the algorithms and the techniques adopted) [56].	Rule-based chatbots: Use a knowledge base organized with conversational patterns, including a list of hand-written responses that correspond to the user's inputs [56]-[57] Three of the most common languages for the implementation of chatbots with the pattern-matching approach are AIML, Rivescript and Chatscript. For these reasons, this model is not robust to spelling and grammatical mistakes in user input. Retrieval-based chatbots: Learn to select responses from the current conversation from a repository with response selection algorithms, heuristics, fairly simple concepts of a rule-based expression match or using a combination of machine-learning classifiers [58]. Also, these chatbots do not produce new responses, but choose one from a pool of predefined responses. Generative chatbots: Generative models are the smartest among the three models in terms of generating answers and can generate more proper responses that could have never appeared in the corpus. These models use machine-learning algorithms and deep-learning techniques. This means that generative chatbots need training with a very large set of data to achieve a good conversation.

Goals: Based on the objectives and the primary goal that the bot aims to achieve, existing dialogue systems are generally divided.	Task-oriented chatbots: are aimed to assist the user with short conversations to complete a particular task, typically used in a closed domain. Non-task oriented chatbots: can simulate a conversation with humans to provide reasonable responses and entertainment. The two major approaches used in non-task-oriented systems are generative and retrieval-based methods. Typically, they focus on conversing with humans on open domains [59].
Service provided: based on the task the chatbot is performing and the amount of intimate interaction that takes place.	Interpersonal chatbots: are usually based on rule-based/ retrieval-based chatbots that offer services like booking servces in restaurants or airlines. Intrapersonal chatbots: exist within the personal domain of the user and understand his/her needs. Inter-agent chatbots: such as Alexa-Cortana integration chatbots to communicate with each other [55].

3.3 Chatbots' Applications

Arabic chatbots have applications in various domains, such as education and healthcare. In the education sector, chatbots can assist students with homework, offer feedback on their work and answer their questions. Examples of these chatbots include the rule-based chatbots [4]-[8] and generative conversational AI [9]-[10]. Moreover, in Section Two, Chapter Four of [11], a review is provided on the various opportunities and challenges associated with educational chatbots. The authors emphasized the advantages of using chatbots in the educational sector, such as their accessibility (24x7 remote access), promotion of self-learning and self-regulation and facilitation of social learning, particularly in creating awareness about societal issues. However, the authors also highlighted specific challenges in implementation. These include issues associated with reliability and accuracy during wide-scale chatbot integration in the learning process, technology limitations in chatbots and insufficient research on various aspects of chatbot technologies. In healthcare, conversational systems like the retrieval-based OlloBot chatbot [12] and AI-based MidoBot chatbot [13]-[15] assist patients in managing their health and answering their medical queries. The potential of conversational dialogue systems to revolutionize human-computer interactions is substantial.

In another study by Abu-Shawar and Atwell [60], a chatbot system called FAQchat is presented. It serves as an interface for Frequently-Asked Questions (FAQ) websites, converting website text into a chatbot-friendly format. The system provides answers using pattern-matching template rules without requiring sophisticated language processing or inference. User trials reveal favorable feedback, with around two-thirds of users preferring FAQ chat over traditional search engines. This demonstrates the practical usability of the chatbot and suggests its potential as a viable alternative for accessing FAQ databases, indicating broader adoption of chatbots in information portal websites.

Artstein et al. [61]-[62] and Traum et al. [63]-[64] introduced New Dimensions in Testimony (NDT). This chatbot application allows users to converse with Holocaust survivor Pinchas Gutter. Developed by the University of Southern California's Institute for Creative Technologies in collaboration with the USC Shoah Foundation, NDT showcases a novel use of AI and natural-language understanding, creating immersive educational experiences. The system goes beyond traditional chatbots by utilizing advanced natural-language processing to generate lifelike virtual avatars of survivors. Based on extensive video interviews, these avatars provide authentic and emotionally impactful storytelling experiences during real-time interactions. While demonstrating the system's effectiveness in simulating conversations with "live" individuals, it has been noted that NDT cannot initiate topics or ask questions in counseling and historical-talk contexts.

In the same context, Abu Ali et al. [65] developed a bilingual (Arabic-English) interactive human avatar dialogue system called TOIA (Time-Offset Interaction Application). The system is inspired by the "new dimensions in testimony demonstration" project by Artstein et al. [61] and simulates face-to-face conversations between humans using digital human avatars recorded in the past. TOIA is a conversational agent based on an actual human being and can be used to preserve and tell stories. The system allows anyone to create an avatar of themselves using a laptop, which facilitates cross-cultural and cross-generational sharing of narratives to wider audiences. TOIA supports monolingual and crosslingual dialogues in Arabic and English, but can be extended to other languages. This system has the potential to bridge the gap in dialectal-speech recognition and overcome the challenges of limited resources, lack of standard orthographic rules and lack of definition in Arabic dialects [61][65].

4. Challenges in Arabic Conversational AI

This section explores the challenges faced in Arabic conversational AI systems. Based on the literature, we outline general problems related to conversational AI (question-answering systems, chatbots, conversational agents and dialogue systems), like context sensitivity, ambiguity and the need for high-quality labeled datasets. Additionally, we explore Arabic-specific challenges, such as the complexity of Arabic morphology, dialectal variations and the phenomena of Arabizi and transliteration. To visually represent these challenges, we present an illustrative overview in Figure 5.

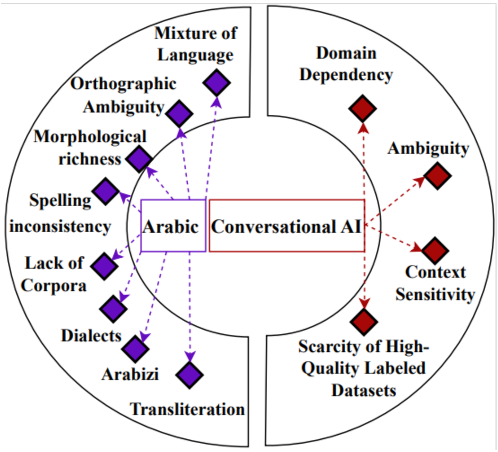

Figure 5. Arabic conversational AI challenges.

4.1 Basic Concepts and General Challenges

Building conversational AI for the Arabic language presents specific challenges due to the complexity and unique characteristics of the language. In this section, we will examine the general problems, which include basic concepts of the Arabic language, ambiguity, data availability and quality, formality and politeness, speech recognition and synthesis and cultural sensitivity.

Basic Concepts: Arabic natural-language processing (NLP) is an increasingly growing field, but it comes with distinctive challenges compared to other languages. Arabic is one of the most widely spoken languages globally, boasting over 421 million speakers as of 2017[1]. Additionally, Arabic encompasses diverse spoken forms [66]-[67]. In his book "On Introduction to Arabic NLP" [68], Habash identifies several challenges that Arabic presents to NLP, including orthographic ambiguity, morphological complexity, dialectal variations and orthographic noise. While some of these challenges may not be exclusive to Arabic, their combination renders Arabic processing notably intricate. Arabic distinctly differs from languages like English in various aspects. For instance, the Arabic alphabet is read and written from right to left and consists of 28 primary characters, 13 of which contain dots, while 15 do not. Furthermore, another specificity differs Arabic from other languages in some punctuation marks, such as the question mark (in Englais ’?’, in Arabic ’؟’), comma (in Englais ’,’, in Arabic ’،’) and semicolon (in Englais ’;’, in Arabic ’؛’).

Cultural Sensitivity: Arabic conversational AI systems need to be culturally sensitive and avoid generating responses that may be considered offensive or inappropriate in the Arab world. Understanding cultural norms and values is crucial for building successful conversational AI systems. However, the development and training of Arabic conversational AI systems might suffer from a lack of multicultural representation. If the AI system is primarily designed and trained by individuals from a limited cultural background, this could result in a system biased towards that particular culture and less inclusive of the diverse perspectives and values [69].

Speech Synthesis and Recognition: Compared to English, Arabic voice-based conversational AI

[1] Source: https://www.internetworldstats.com/stats19.htm

systems have less advanced Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) technology. According to [70], the limited availability of Arabic ASR and TTS systems is evident compared to other languages, indicating a relative scarcity of high-quality systems. This lack of resources hinders the development of reliable speech-recognition and synthesis capabilities for Arabic Conversational AI. Moreover, Arabic poses unique challenges in phonetics and pronunciation due to its complex phonetic system, encompassing a wide range of sounds and phonetic variations [68][71]. This complexity creates difficulties for ASR systems to accurately recognize and transcribe Arabic speech, particularly when handling various accents, dialects, limited vocabulary and diacritic marks. These diacritic marks are vital in providing contextual information, making it challenging for ASR systems to achieve precise transcriptions without them.

Ambiguity, Homonymy and Gender Agreement: These aspects pose challenges in the context of Arabic AI systems, particularly in speech recognition [72] and chatbot applications. Arabic nouns, pronouns and verbs exhibit gender agreement, meaning that they vary based on the gender of the speaker and the entities being referred to. This presents difficulties in developing gender-inclusive and culturally sensitive AI systems that respond appropriately to diverse users. Additionally, Arabic words often have multiple meanings depending on the context [70], [73]-[74], leading to ambiguity challenges. Homonymy, where different words share the same pronunciation but have distinct meanings (Table 3), further complicates the accurate interpretation of user input and the provision of appropriate responses.

4.2 Arabic-specific Challenges

In addition to the broader challenges facing conversational AI, there are also specific difficulties related to the Arabic language’s various dialects and morphological structure. Since conversational systems heavily rely on the morphology of the target language, as noted by studies such as [11] and [75], it is essential to consider the unique linguistic characteristics of Arabic, including its dialects, orthography and morphology.

Arabic Varieties: Figure 6 illustrates the three main varieties of Arabic [66]-[68]. The first is classical Arabic, known as Quranic Arabic, used in religious texts and various old Arabic manuscripts. The second variety is Modern Standard Arabic (MSA), the formal means of communication that most Arabic speakers understand. MSA is commonly employed in newspapers as well as in radio television broadcasts. The third type is dialectical or colloquial Arabic, which is utilized in everyday conversations and has recently found its way into TV and radio broadcasts. Arabic dialects can be categorized into five main groups based on geographical distribution (Maghrebi dialect, Egyptian dialect, Levantine dialect, Iraqi dialect and Gulf dialect). However, it is essential to note that this classification is general and relies on the proximity of countries, leading to commonalities in dialectical words and expressions.

Additionally, Farghaly et al. [66] and Ryding et al. [67] proposed a classification of dialectal Arabic into two main categories: western and eastern. The reasons for the existence and growth of many Arabic dialects are geographical, social, political and primarily linguistic conflicts. The advantage of dialect lies in the economy of language, characterized by abandoning common expressions and forms, ignoring, quoting and updating meaning. The nature of life is willing to ignore what should be ignored to quote what is necessary.

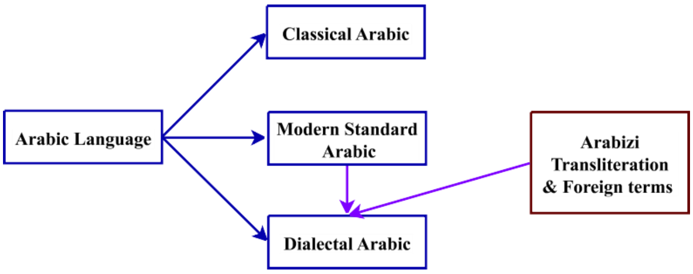

Figure 6. Arabic language varieties.

Arabic Ambiguity and Orthography: The Arabic orthography allows optional diacritical marks to denote short vowels and consonantal doubling. These diacritical marks provide additional information about the pronunciation and structure of words in the written text. However, people commonly omit these marks from the text and do not use upper and lower cases, leading to high ambiguity. Additionally, according to [76], Arabic writers often need to correct spelling, particularly with problematic letters like Alif-Hamza "أ آ" forms and Ta-Marbuta "ة". The problem of orthography is even more challenging for Arabic dialects, as they need standardized orthographies [74][77]. Diacritical marks are used in Arabic to aid pronunciation and clarify the meaning of words. However, proficient speakers can understand the text without these marks, so most Modern Standard Arabic (MSA) texts are written without them, resulting in lexical ambiguity. Diacritical marks include point diacritics, short vowels (fatHah, kasrah and DHammah), linguistic diacritics and decorative marks. This diversity can lead to the problem of having several diacritics on a single base letter. Diacritics are crucial in distinguishing between orthographically similar words, but disambiguated by diacritics. Table 3 illustrates three examples of this.

Table 3. Orthographically similar words disambiguated by diacritical marks.

Triliteral roots	Meaning-1	Meaning-2	Meaning-3
ع، ل، م	عَلِمَ ’to know’	عَلَم ’flag’	عِلْمْ ‘science’
ح، س، ب	حَسَبَ ’to calculate’	حَسِبَ ’to think / assume’	حَسَب ’decent’, ’according to’
ك، ت، ب	كَتَبَ ’to write’	كُتُب ’books’	كَتَّبَ ’to force to write’

Morphological Complexity: Arabic exhibits many inflectional characteristics, in which words undergo various forms based on their context within a sentence, including inflections and derivations. Developing a robust morphological analyzer, a vital component of any natural-language processing system [78] becomes challenging due to this linguistic feature. Habash [68] and Habash-Soudi-Buckwalter [78] highlighted that morphological analysis stands out as one of the most demanding tasks in Arabic NLP due to the language’s rich morphological structure encompassing a considerable number of allomorphs and exceptions.

To address these challenges, researchers are developing new techniques and technologies to enhance the accuracy and effectiveness of Arabic conversational AI systems. For example, Habash-Soudi-Buckwalter [79] introduced an Arabic transliteration (HSBT) scheme. Arabic transliteration involves representing Arabic text using Latin characters, providing a phonetic representation of the script. This enables non-Arabic speakers or those unfamiliar with the Arabic script to read and pronounce Arabic words. Transliteration finds applications in language-learning materials, dictionaries and communication between Arabic and non-Arabic speakers. The HSBT transliteration scheme [79] extends Buckwalter’s original scheme [80], which designates a single, distinct ASCII character for each Arabic letter. HSBT enhances readability by incorporating some non-ASCII characters while maintaining a distinct 1-to-1 mapping between Arabic and Latin characters. For example, the Arabic word السَلَام عَلَيكُم, when transliterated with the HSBT scheme, is written as "As-salāmu alaykum".

An Arabic word manifests various morphological aspects, such as derivation and inflection. Multiple words with distinct meanings originate from the same root in derivation. For instance, from the root "قَلب" (’heart’), we derive "قُلوُب" (’hearts’) and "قَالِب" (’Mold/Template’). On the other hand, the inflection aspect involves varying the same word to denote different grammatical categories expressing the same meaning. For example, the root word "قَرَأَ", meaning ’to read’ in Arabic, appears as "اَقرَأُ" in the present tense, first person singular form and as "قَرَأتُ" in the past tense form.

Arabic’s intricate morphology and unique language features have introduced additional challenges for building Conversational AI, particularly the need for more Arabic question-answering resources. We will now delve into the obstacles linked to the complexity and diversity of the Arabic language, including the lack of corpora, the presence of dialects and Arabizi and the mixture of languages.

Limited Resources and Data Availability: Building a robust and accurate conversational AI system necessitates vast amounts of annotated data for training and evaluating NLP models. The effectiveness of conversational AI is directly linked to the quality and size of the corpus used for training. According to [81]-[82], the availability of Arabic question-answering datasets is limited. Moreover, a comparative analysis of English and Arabic languages in building text-based conversation agents was explored in [73]. The study revealed that constructing conversational AI systems in Arabic is notably more challenging due to the language’s complexity and the scarcity of available resources. Additionally, the authors of [73] pointed out that Arabic speakers often use different linguistic forms based on the conversation's context, sometimes even within the same conversation. This phenomenon, known as diglossia, arises when multiple forms of the same language coexist within the same speech community.

Furthermore, the paper by Antoun et al. [83] highlighted that numerous datasets are derived from translations of other languages, often relying on resources like English-SQuAD, TREC or CLEF. The authors observed that these translated datasets, such as Arabic-SQuAD and ARCD, include text elements in languages other than Arabic, encompassing unknown sub-words and characters. This issue arises from inadequate training samples translated from the English SQuAD dataset. Moreover, Alwaneen et al. [81] pointed out that utilizing translated datasets can introduce language-related complexities. Most Arabic QA datasets are limited, making it impractical to directly compare different systems, as each uses its unique dataset. Furthermore, authors are generally reluctant to share their working code. However, a few exceptions exist, such as the Arabic-squad [82] and TyDi-QA [84] datasets, which incorporate the Arabic language.

Dialectal Variation: Arabic is spoken in many different dialects, each with its distinct vocabulary, grammar and pronunciation. Consequently, constructing a single conversational AI system capable of effectively communicating with users from diverse Arabic-speaking regions presents a considerable challenge [73], [85]. The variations in pronunciation, vocabulary and grammar across these dialects pose significant limitations for NLP systems attempting to process and comprehend all forms of Arabic. Furthermore, Habash's book [68] emphasizes the complexities of dialectal variation in natural-language processing tasks, underscoring the difficulties faced in handling such diversity.

The Arabizi Phenomenon: Arabizi is an informal and non-standardized form of writing that involves using the Latin alphabetic transliteration or writing of Arabic words [68], [86]. It is a term used to describe the practice of phonetically writing Arabic words using a combination of Latin characters and numerals. Arabizi is commonly used in informal digital communication, including social media, chat applications [75] and text messages. This writing style appeals to young Arabic speakers and those who feel more at ease using Latin characters while communicating in Arabic. For example, the number "3" represents the letter "ع" in transliteration, while the number "7" corresponds to the letter "ح". In Arabizi, the Arabic phrase اَلسَلَام عَلَيكُم is written as "as-slm 3lykm".

No Capital Letters and Code Switching in Different Alphabets: Arabic lacks uppercase or lowercase letters, which poses a challenge for factoid question systems that rely on named entity recognition to identify proper names, places and person names [87]. This limitation makes it difficult to differentiate between proper Arabic nouns and other word forms, like adjectives and common nouns [88]. Additionally, Arabic users frequently mix dialectal expressions and Arabizi within their discussions. As both dialects and Arabizi lack standardized rules, they can be written differently, presenting similar linguistic challenges for conversational AI systems.

Furthermore, Arabic speakers are often bilingual or multilingual, particularly in science and medicine. Consequently, users may ask questions that contain more than one language. Pre-processing questions or answers by filtering out Latin letters could result in a loss of meaning, leading the question-answering system to provide inaccurate or empty responses.

4.3 Arabic NLP Tools and Resources

The complexities and characteristics of the Arabic language present a significant challenge for researchers and developers in processing Arabic text. Dealing with ambiguity, diglossia and comprehending the Arabic script necessitates using specialized Arabic NLP tools and resources. These dedicated tools are designed to tackle the unique challenges of Arabic-language processing, ultimately enhancing the accuracy and efficiency of NLP tasks. This sub-section introduces some widely utilized Arabic NLP tools, including MADAMIRA, Farasa and CAMeL, along with the notable online Masader+ catalog of Arabic NLP data and the Special Interest Group on Arabic NLP (SIGARAB). These valuable resources collectively aid in addressing the challenges above and contribute to advancing Arabic-language processing.

MADAMIRA is an Arabic morphological text-analysis tool developed by researchers [89], that combines the two widely-used systems in Arabic pre-processing, MADA and AMIRA [78], [90]-[91], to provide a comprehensive tool for pre-processing Arabic text. This fusion has created the powerful MADAMIRA tool, specifically designed to cater to various linguistic analysis tasks essential for Natural

Language Processing (NLP) in Arabic. Among the key features which MADAMIRA provides are tokenization, lemmatization, part-of-speech tagging, Base Phrase Chunking (BPC) and Named Entity Recognition (NER). Its tokenization capability ensures that input Arabic text is effectively segmented into individual units, preparing it for further analysis. With part-of-speech tagging, MADAMIRA assigns grammatical categories to each token, facilitating a deeper understanding of the sentence's syntactic structure. Additionally, MADAMIRA provides BPC for identifying phrases and their syntactic roles within the sentence and NER capabilities, enabling identifying and extracting named entities, such as names of people, locations, organizations and other relevant entities in the text.

Farasa, an Arabic segmenter developed by Abdelali et al. [92] uses SVM-rank to learn feature vectors for each segmentation. The segmentations are then scored with the trained classifier. The tool comes in two varieties: FarasaBase and FarasaLookup. FarasaBase utilizes a classifier to segment words with a lookup list containing concatenated stop-words. On the other hand, FarasaLookup involves a training process where seen segmentations are cached during training and classification is applied to unseen words. Farasa supports several key features and functionalities, including Dialectal Analysis, Diacritization, Text Normalization, POS tagging and Named Entity Recognition (NER). Furthermore, Farasa has demonstrated its performance in Machine Translation (MT), achieving an accuracy rate of 98.94%, outperforming MADAMIRA in this task.

Recently, Obeid et al. [93] introduced **CAMeL** Tools, a comprehensive set of tools and libraries designed explicitly for Arabic natural-language processing tasks. This toolset supports Arabic and Arabic dialect pre-processing, providing many features, such as transliteration orthographic normalization, discretization, dialect identification, morphological modeling, sentiment analysis and named entity recognition. The development of CAMeL Tools was motivated by addressing limitations in previous rule-based systems, such as fragmented tasks spread across different systems and a need for more flexibility. By integrating diverse functionalities into a well-suited toolkit, CAMeL Tools presents a more efficient and flexible approach for tackling Arabic NLP tasks. Additionally, the toolkit includes libraries that facilitate working with diverse text formats like HTML and XML and text classification and clustering capabilities. Experiments conducted on CAMeL Tools demonstrated significant performance improvements compared to some state-of-the-art models and it outperforms MADAMIRA in various processing utilities. Furthermore, Antoun et al. antoun2020arabert utilized CAMeL Tools to fine-tune the pre-trained language model AraBERT, showing superior results compared to the CRF-based system in the named entity recognition task.

Recently, the importance of a corpus as a fundamental component for building accurate question-Answering systems has become increasingly evident. Consequently, researchers have made significant efforts to develop more accessible Arabic language resources. One notable platform in this regard is Masader+, an online catalog of Arabic NLP data, which serves as a comprehensive repository of linguistic resources and datasets for the Arabic language [GitHub[2]]. Developed in 2022 by Alyafeai et al. [94], Masader+ represents an updated version of the original Masader catalog. It offers a wide range of NLP data, including corpora for named entity recognition, question-answering systems and sentiment analysis tasks. The data in the catalog is freely available for researchers and developers working in Arabic NLP. The website provides detailed information about each dataset in the catalog, including data size, format, suitability for specific tasks and data source. Additionally, it offers instructions for data downloading and usage, along with links to related resources and publications. Masader+'s organized and user-friendly interface ensures easy access to high-quality datasets, fostering research, development and evaluation of NLP models and applications tailored to Arabic. As a valuable resource for the latest developments in Arabic NLP, the Masader+ website proves to be beneficial for anyone interested in advancing language processing in the Arabic domain.

The MADAMIRA, Farasa and CAMeL Tools, alongside the online Masader+ catalog of Arabic NLP data, demonstrate remarkable versatility, efficiency and improved performance, making them invaluable resources for researchers and developers in Arabic NLP. With their diverse features and continuous enhancements, these tools play a significant role in advancing Arabic-language processing and enabling the creation of sophisticated and accurate NLP applications for Arabic. Notably, all three tools are open-source and offer command-line interfaces (CLIs) and application programming interfaces (APIs), offering users convenient and flexible integration into their NLP workflows. Additionally, we have the Special Interest

[2] Website available from https://arbml.github.io/masader/

Group on Arabic NLP (SIGARAB), a dedicated professional organization operating under the Association for Computational Linguistics (ACL). SIGARAB strives to promote the growth of Arabic NLP technologies, fostering knowledge exchange among researchers and practitioners. The organization showcases state-of-the-art research in Arabic NLP through regular meetings, conferences, newsletters and proceedings. Its website[3] serves as a valuable resource, providing information on news, events and diverse resources, such as datasets, software and documentation for researchers and developers in the Arabic NLP field.

5. RULE-BASED AND RETRIEVAL-BASED CHATBOTS

In contrast to English and other chatbots, Arabic still needs to be more powerful and efficient. Moreover, Arabic chatbots are comparably scarce due to the challenges coming from the language itself. The challenges originate from the uniqueness of the Arabic representation style, the richness of its morphology, the different meanings of each word and the synonyms provided to express a specific request.

The authors in [4] proposed ArabChat, a rule-based conversational agent developed using the PM approach. In the same context, the authors in [5] proposed a mobile version of ArabChat and the authors of ArabChat [4] provide the "Enhanced ArabChat" [6]. Authors of [8] have simulated Nabiha, a rule-based chatbot based on PM and AIML approaches. Ollobot is an Arabic rule-based chatbot proposed by Fadhil et al. [12] using the AIML method. In [75], the writers stated that they developed a retrieval-based chatbot called BOTTA using several AIML files. Aljameel et al. [7] proposed LANA as an Arabic retrieval-based chatbot based on a combination of Pattern Matching (PM) and Short Text Similarity (STS) to extract the responses. In 2021, an Arabic flight booking dialogue system was proposed using rule-based and data-driven hybrid approaches in [21]. More recently, in 2022, AlHumoud et al. presented Rahhal [20], an Arabic rule-based chatbot for helping tourists visit different Saudi Arabian cities based on the PM approach. The approaches applied in previous works are based essentially on pattern matching. However, these techniques are based on something other than grammatical or linguistic details. The rule-based system is the oldest in chatbot development. Recent advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) enable data-driven approaches to developing chatbots. Table 4 summarizes the pre-mentioned Arabic Chatbot systems.

Table 4. Rule-based and retrieval-based Arabic dialogue system/chatbots.

Chatbot	Application domain	Implementation Technique	Response- generation technique
ArabChat [4]	Closed Domain (assisting the students of ASU)	Pattern Matching	Rule-based
Mobile ArabChat [5]	Closed Domain (assisting the students of ASU)	Pattern Matching	Rule-based
Enhanced ArabChat [6]	Closed Domain (assisting the students of ASU)	Pattern Matching	Rule-based
Botta [75]	Open Domain	AIML	Retrieval-based
Nabiha [8]	Closed Domain (assisting the students of KSU)	AIML and Pattern Matching	Rule-based
LANA [7]	Closed Domain (science for children with autism)	STS and Pattern Matching	Retrieval-based
OlloBot [12]	Closed Domain (Healthcare)	AIML	Retrieval-based
Flight booking [21]	Closed Domain (airline-ticket booking)	Pattern Matching	Rule-based and data-driven
Rahhal [20]	Closed Domain (helping tourists at Saudi Arabia)	Pattern Matching	Rule-based

6. DEEP LEARNING-BASED CHATBOTS

Deep learning is a discipline of machine learning. It is known for learning embedded and abstract representations from raw data with minimal human intervention. Deep-learning models can process large datasets efficiently, saving time by eliminating human intervention and feature engineering [95].

[3] Group available from http://www.sigarab.org/

Deep learning has been recently investigated for the Arabic question-answering system and has shown excellent performance in this area [96]-[97]. These studies use vector-pre-trained word representation datasets for these models, such as Word2vec, Glove and Fasttext, which learn from unlabeled text. In addition, more advanced language representation models, such as BERT [98]-[99] and ELMO [100], have also been used. For Arabic question-answering systems, only a few recent works have explored deep-learning models. Table 5 presents the significant works on deep learning-based Arabic QA systems.

Table 5. Deep learning-based Arabic question-answering systems/chatbots.

Ref.	Year	Domain	Approach	Dataset	Result
[101]	2020	Open Domain	Learning-based (AraBERT)	Arabic-SQuAD and ARCD	AraBERT v1: F1: 82.2% SM: 95.6% EM: 54.8%
[102]	2020	Closed Domain (Medical)	Learning-based (DNN, PCA)	SemEval-2017 CQA-subtask D	MAP: 62.90% MRR: 68.86% AvgRec: 86.6%
[82]	2019	Open Domain	Learning-based (BERT)	ARCD	F1: 27.6% EM: 12.8% SM: 29.8%
[103]	2018	Closed Domain	ML classifier	stack overflow [104]	Random forest Prec: 71.1% Recall: 74.1%
[13]	2022	Closed Domain	DL (seq2seq)	variety of source	N/A
[98]	2021	Closed Domain	BERT, CNN Word2Vec	D1: SemEval-2016 task 3 Subtask B. D2: Quora dataset	BERT/word2vec /CNN+feature D1: Acc:79.57% F1: 71.58% D2: Acc: 88.80% F1 83.18%
[100]	2020	Open Domain	Learning-based (ELMO, CNN, RNN)	Collected manually	Weighted F1: 94% F1: 94%, Acc: 94%
[105]	2021	Open Domain	Learning-based (SVM)	Collected from TREC, CLEF and Moroccan school books	Acc: 90%, Prec: 91% F1: 90%, Recall 90%
[106]	2022	Open Domain	Learning-based (LSTM, CNN, W2V)	Translated from [107]	ASLSTM in Arabic: P@5: 0.37% P@10: 0.28%, MAP: 0.45% Recall: 0.41%
[108]	2020	Open Domain	Semantic and logic-inference-based	AQA-WebCorp [109]	Acc: 0.74%
[110]	2021	Closed Domain	AI-based	-	N/A
[111]	2022	Open Domain	Deep-based DPR, AraELECTRA	491,253 doc: Arabic Wikipedia ARCD [82] TyDiQA GoldP [84]	Single training [84]: EM: 41.8%, F1: 50.1% Single training [82] EM: 15.1%, F1: 35.3% Multi training [84] EM: 43.1%, F1: 51.6% Multi training [82] EM: 15.7%, F1: 36.3%

In fact, Antoun et al. [101] introduced AraBERT, a specialized variant of the BERT (Bidirectional Encoder Representations from Transformers) model designed specifically for Arabic language. The base model consists of 12 attention heads, 12 encoder blocks, 768 hidden layers, a maximum sequence length of 512 and 110M parameters. A large dataset of 70 million sentences, equivalent to 24 GB of text, was used to train the model. Next, the authors conducted evaluations of the pre-trained model on three downstream Arabic tasks: Arabic question-answering, sentiment analysis and named entity recognition. Finally, the authors fine-tuned the pre-trained language model for question answering using two datasets, Arabic-SQuAD and ARCD [82].

Al-Miman et al. [102] proposed a deep neural network model for the answers' ranking problem in Arabic language using the provided Arabic dataset in SemEval-2017 CQA-subtask D. The authors try to find and integrate different types of similar features. The deep model uses the results of the integrated feature set as input. The ranking positions are then generated from the question-answer pairs. The proposed model incorporated features at three levels. The first level utilized Principal Component Analysis (PCA) features, while the second and third levels incorporated similarity features before and after pre-processing, respectively.

The authors of [82] developed a deep-learning system called SOQAL, an open-domain Arabic question-answering system. Their system is based on integrating TF-IDF with a multilingual pre-trained Bidirectional Transformer (mBERT) neural Machine Reading Comprehension (MRC) model to answer open-domain fact queries. TF-IDF is used to group the retrieved documents; i.e., the documents most relevant to the query are selected and responses are extracted from these documents using a multilingual pre-trained Transformer mBERT bidirectional model. A benefit of this research is that the authors first enrich the Arabic research community by creating a corpus that can effectively serve as a training resource for Arabic QA systems. Second, increasing the corpus size can improve the end-to-end QA system. Finally, the authors point out the possibility of improving the system to obtain correct answers using paragraph selection.

Elalfy et al. [103] proposed a hybrid approach using two modules, called content-based and non-content-based modules, to find the best answers on CQA websites. This work used content features, question-answers, answer-answer features and the user-reputation score.

Boussaksso et al. [13] presented an Arabic chatbot called MidoBot, based on the Seq2Seq model, to generate new responses from a dataset. They used AraVec [112] for embedding and the dataset consists of 81,659 lines collected using blogs, plays, Quora Arabic and movie subtitles. Van Tu, N. et al. [98] proposed a model based on the BERT model that integrates various features from other methods.

Hamza, A. et al. [100] proposed a deep model based on distributed word representations (ELMo embeddings) and deep neural (CNN and RNN) models for Arabic question classification (QC). The authors proposed classifying questions into seven categories by finding syntactic and semantic relationships between words and using ELMO to represent the questions. Additionally, they used a dataset of 3,173 Arabic questions, collected and annotated manually, to evaluate their system.

The same authors proposed in [105] a framework based on a machine-learning approach and words' continuously distributed representation for Arabic question classification. First, they proposed a taxonomy of open-domain questions in Arabic, where they represent questions using TF-IDF weighting with n − grams and word representations with a bag of its character n − gram for represented the questions to find the syntactic and semantic relations between words. Then, in the experimentation phase, they used a dataset of 1.302 questions collected from TREC, CLEF and Moroccan school books and they used the SVM algorithm to classify questions.

Othman, N. et al. [106] proposed a deep-learning approach called ASLSTM (Attentive Siamese LSTM-based approach) to tackle similar question-retrieval problems. The ASLSTM method is based on a Siamese architecture with a long short-term memory (LSTM) network, supplemented by an attention mechanism, which enables the model to give extra attention to different words when modeling the questions. To evaluate the proposed approach for the Arabic language, they used Yahoo!Webscope dataset[4], which was translated into Arabic using Google Translate, comprising 1,256,173 questions and 12,512,034 different words. They trained the translation dataset for Arabic word embedding training using Word2Vec (CBOW model); window size = 300 and LSTM training; layer size = 50; embedding layer size = 300.

Bakari, W. et al. [108] proposed a system based on logic form and conceptual graph approach for factoid Arabic question-answering systems. Their system is based on an intuitive understanding of Arabic texts to convert them into semantic and logical representations. First, they used conceptual graphs to transform Arabic text to obtain a logical representation and then, they extracted answers based on the relationship between questions and text passages. They evaluated their system through the NArQAS system [113] using the question-text corpus (AQA-WebCorp) [109]. Therefore, their logic-based approach combined with textual participation (RTE) detection significantly improved the performance of Arabic QA systems.

Al-Madi, N.A., et al. [110] developed an intelligent Arabic chatbot system. This social chatbot can assist students from the AlZaytoonah Private University of Jordan and answer their questions regarding their educational progress using the spoken Arabic language.

Alsubhi, K. et al. [111] helped improve the performance of the Arabic open-domain question-answering system. They used deep-learning techniques to build their end-to-end Arabic open-domain question-answering system based on 491,253 documents from Arabic Wikipedia articles. The model consists of two stages; the first is a passage retrieval task, where they retrieved the top 20 passages relevant to the question using ARCD [82] and TyDiQA GoldP [84] datasets using DPR (Dense Paragraph Retrieval) [114]. Then, for the reading comprehension task, they connected DPR to the AraELECTRA [83] passage reader to obtain the first three answers.

AraELECTRA [83] is an Arabic language representation model pre-trained using the Replaced Token Detection (RTD) methodology on a sizeable Arabic text corpus. Dense Passage Retrieval (DPR) [114] is an efficient retrieval method for open-domain question-answering tasks that uses dense representations to compute relevancy. These methods use deep neural networks to embed documents and questions into a shared embedding space. Dense models use transformer-based encoders that are more sensitive to features, such as lexical changes or semantic relationships.

[4] https://https://maktoob.yahoo.com/?p=us/

Due to their complexity, deep learning-based approaches have shown great promise for improving Arabic question-answering systems. Multiple layers in the deep-learning models effectively decompose Arabic text for more accurate understanding and analysis. We find that Arabic language has had a different luck than English language regarding using the deep-learning approach during the development of QA systems [115]. For these reasons, we will give more importance to the consideration of contextualized word embedding, large language models and deep-learning models for building Arabic QA systems. Furthermore, we find that most existing deep learning-based approaches for Arabic QA have focused on Modern Standard Arabic (MSA), with some using datasets derived from translations of other languages, such as English-SQuAD, TREC or CLEF. However, [81][83] noted that Arabic-SQuAD and ARCD datasets may contain text elements in languages other than Arabic, including sub-words and unknown characters. This highlights the importance of developing adequate representations of words for deep-learning models, which can significantly impact their performance. Following the recent works [82][101], where they proposed two pre-trained Arabic language representation models (AraBERT and AraELECTRA) based on the BERT architecture, specifically designed for Arabic language, we believe that these advances demonstrate the potential of deep learning-based approaches to improve QA systems in Arabic and contribute to the growth of this field.

7. DEEP LEARNING TECHNIQUES

Recently, many algorithms based on deep-learning techniques have made significant progress, aiming to solve the related problems of question-answering systems (task-oriented and non-oriented dialogue systems, answer modeling and answers' ranking problems) based on word representation like non-contextual/ contextual Word Embedding and Uni/Bi-directional model by learning feature representations in a high- dimensional distributed fashion and achieving remarkable improvements in these aspects. Figure 7 shows the evolution of deep-learning techniques.

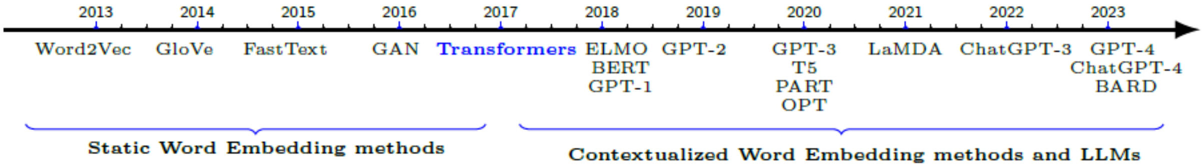

Figure 7. Deep-learning techniques' evolution timeline (2013-2023).

Using distributed representations such, as contextualized or static word embeddings, numeric vectors represent words with similar meanings and contextual usage. Pre-trained embeddings, a successful application of unsupervised learning, can capture semantic and syntactic information, enhancing the performance of various downstream tasks [116]-[117]. A key advantage of pre-trained embeddings is their independence from costly annotation, as they can be derived from extensive unannotated corpora readily available. Consequently, these pre-trained embeddings can be leveraged in downstream tasks that involve limited amounts of labeled data.

7.1 Static Word Embedding

Static Word Embedding methods train the models based on the co-occurrence statistics, such as Word2vec [118], GloVe [119] and FastText [120], which are critical components in many neural language-understanding models.

Word2Vec was introduced by Tomas Mikolov et al. at Google Research [118], helping frame the distributional hypothesis in a predictive approach. Word2Vec has two primary model architectures: the Continuous Bag of Word (CBOW) and the Skip-Gram (SG). These models are algorithmically similar, except that CBOW predicts the target words from the source context words, while Skip-Gram predicts the source context words from the target words. For a set of words in a context window, CBOW sums the vectors representing these words to produce a vector representation of this context. Skip-Gram represents each word and context as d-dimensional vectors to produce similar vector representations of similar words. The CBOW model works better than Skip-Gram on syntactic tasks and much better on the semantic part [121].

Similar to Word2Vec, **GloVe** (Global Vectors for Word Representation) proposed by Jeffrey Pennington [119] is another word vector representation technique. GloVe employs a count-based approach to reduce the dimensionality of the co-occurrence counts' matrix. The main concept behind GloVe is to construct a comprehensive co-occurrence counts' matrix from a given corpus, where each cell in the matrix represents the frequency of a word occurring within a specific context. By normalizing the values for each row of the matrix, we obtain the distribution probability of each context for a given word. Like Word2Vec, word similarity is determined by the similarity between their corresponding vectors. When it comes to dialogue systems, various existing approaches have been utilized. These include retrieval-based methods [122]-[124], as well as generation-based models [125]-[126].

FastText, an open-source project developed by Facebook AI Research Lab, is utilized for constructing scalable solutions in text representation and classification [120]. It serves as an extension to the Word2Vec model and offers an alternative word-embedding approach. Unlike directly learning vectors for individual words, FastText represents each word as an n-gram of characters.

7.2 Contextualized Word Embedding

Unlike traditional word embeddings, such as word2vec or GloVe, contextualized word embeddings consider the surrounding words and sentence structure to generate more nuanced representations. Those approaches have shown effectiveness in various NLP tasks, including sentiment analysis, named entity recognition, machine translation and question-answering. The popular approaches for generating contextualized word embeddings include models like ELMO and transformer-based models like OpenAI's GPT and Google's BERT.

Transformers introduced in 2017 by Vaswani et al. [48] is a type of deep-learning model architecture that has revolutionized various natural-language processing tasks. The transformer has become the unavoidable architecture that forms the basis of the LLMs due to its powerful capabilities. According to [127]-[128], the critical component of a transformer is the self-attention mechanism, also known as scaled dot-product attention. It allows the model to weigh the importance of different words in a sequence when making predictions or generating outputs. In addition, transformers leverage self-attention mechanisms to capture dependencies between words or tokens in parallel. This advantage enables them to effectively model long-range dependencies and capture global context, making them particularly well-suited for tasks involving large text sequences.

LLMs stand of Large Language Models, such as BERT [27] and GPT-3 [25], refer to models trained on a large corpus of text data using unsupervised learning techniques with a massive number of parameters. LLMs are designed to generate coherent and contextually relevant text across various natural-language processing tasks. These models have been widely used for question answering, language translation, text completion and text generation.

LaMDA refers to Language Model for Dialogue Applications [28] as a specialized language model focused on improving dialogue-based applications. LaMDA is explicitly designed for dialogue applications and has the potential to generate responses and improve conversational AI systems and chatbots like Bard. It aims to enhance the flow and coherence of dialogue interactions by addressing challenges, like context understanding, ambiguity resolution and generating more natural and contextually appropriate responses.

ELMo (Embeddings from Language Models) introduced by Matthew Peters et al. [129] is a deep contextualized word-representation model. ELMo word vectors are computed using bidirectional LSTMs (Long Short-Term Memory Networks) to generate word embeddings that capture contextual information.

BERT (Bidirectional Encoder Representations from Transformers), offered by researchers at Google AI Language [27], is a language representation system established with multi-directional language modeling and attention mechanisms. The principal originality of the system is the pre-training approach that captures word and sentence-level representations through Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) training. BERT is pre-trained in diverse languages by employing available unlabeled data. Furthermore, the pre-trained deep bidirectional model has become state-of-the-art in multiple NLP applications like question answering. The concept is to give a general model suitable for diverse applications and a pre-trained architecture that minimizes the necessity of annotated data. For example, for a provided word, its embedding vector is constructed by counting the embeddings of the related word, the sub-words and the position together.

T5 stands of Text-To-Text Transfer Transformer introduced by Raffel et al. in 2019 [130]. It is a transformer-based model with significantly advanced natural-language processing tasks. The T5 model follows a unified text-to-text framework, where a wide range of language tasks, such as translation, summarization and question-answering, can be formulated as a text-to-text conversion problem.

GPT, short for Generative Pre-trained Transformer, is a family of unsupervised transformer-based generative language models developed by OpenAI. The GPT models comprise several large language models (LLMs), including GPT-1, GPT-2, GPT-3/GPT-3.5 and GPT-4. **GPT-1** [23], introduced in 2018, was the initial variant of the Generative Pre-trained Transformer, which included 117 million parameters. In 2019, **GPT-2** [24] was released, boasting 1.5 billion parameters and delivering impressive language-generation capabilities. **GPT-3** [25], released in 2020, took a significant leap forward with staggering 175 billion parameters. The latest is **GPT-4** [26], which was released in 2023 estimated 100 trillion parameters. One notable enhancement in GPT-4 is that it is a multimodal model that can process images and text. GPT models have garnered significant attention and showcased remarkable advancements in natural-language processing, demonstrating their effectiveness across various applications.

8. EVALUATION METHODS

This section summarizes some famous evaluation metrics employed in deep learning-based Arabic QA systems. As indicated by [115], [131]-[132], extractive QA systems commonly utilize span-style datasets employing metrics like F1 score and Exact Match (EM). In contrast, ranking-based QA systems employ metrics such as C@1, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). On the other hand, abstractive QA systems, which rely on text and embedding-based methods, employ metrics like ROUGE, BLEU and several others like Word Mover's Distance (WMD) [133], Sentence Mover's Similarity (SMS), [134], BERTScore, [135] and MoverScore [134]. We classified the evaluation metrics into three main groups: traditional evaluation metrics, ranking-based metrics and textual and embedding-based metrics.

8.1 Traditional Evaluation Metrics

Traditional evaluation metrics, such as Precision, Recall and Accuracy, are often used to develop NLP systems and answer modeling problems. F1-score and Exact Match are commonly employed in QA tasks, particularly in span-style datasets, like the Arabic extractive QA datasets proposed in [82], [84].

Precision (P), Recall (R) and F1-measure (F1): Precision measures the ratio of the number of tokens in the prediction that overlap with the correct answer to the total number of tokens in the prediction (number of correct answers /number of questions answered). Either take: $Correct(q)$ is the set of elements that form the perfect answer for q and $Found(q)$ are those that the QA system returned, then the precision (of the system as regards q) is the fraction of correct responses. The formula of precision is given in Equation 1.

$$Precision = \frac{|Found(q) \cap Correct(q)|}{|Found(q)|} \quad (1)$$

Recall, for each question, there is an expected set of correct answers; these are called the gold standard answers. Recall is the fraction of the correct elements that the system found (number of correct answers/number of questions to be answered). It is computed as shown in Equation 2.

$$Recall = \frac{|Found(q) \cap Correct(q)|}{|Correct(q)|} \quad (2)$$

F1-measure (also called F-score or F1-score) is the harmonic mean of precision and recall. F1 has been generally reserved for evaluating span-based question answering [131]. F1-measure is computed for the words of each predicted answer and each golden answer; i.e., is the average overlap between the words of the predicted and golden answers for a given question. It captures a system's precision and recall capabilities in a single score. It is computed as shown in Equation 3.

$$F1 - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Accuracy (Acc): Let CA denote the number of questions that a system answered correctly and TQ denote the total number of questions in the evaluation dataset. Then, accuracy represents the percentage of questions that a system answered correctly. Accuracy is defined as in Equation 4:

$$Accuracy = \frac{CA}{TQ} \tag{4}$$

Exact Match (EM): The name says it all. Exact match (EM) measures the proportion of predictions that match any reference answers exactly at the token level [131]. Put another way, the prediction is counted as correct only when it matches any reference answers to the given question. If CP denotes the number of correct predictions and TQ denotes the total number of questions in the evaluation dataset. EM is defined as in Equation 5:

$$EM = \frac{CP}{TQ} \tag{5}$$

8.2 Ranking-based Metrics

The performance evaluation of the QA ranking problem is based on the following metrics:

c@1: "A simple measure to assess non-response": This measure is used for question-answering systems that suppose one correct answer for each query. The measure works by evaluating question-answering systems by giving them an option not to answer the question rather than forcing them to provide an incorrect answer [136]. In other words, that differentiates between a wrongly answered question and an unanswered question. Equation 6 defined by [137] represents C@1, where TQ is; total questions, CA is the number of correctly answered questions and UQ is the number of unanswered questions:

$$c@1 = \frac{1}{TQ}(CA + UQ\frac{QA}{TQ}) \tag{6}$$

Mean Average Precision (MAP): It is the mean of the AveP scores for a set of queries. It is defined as in Equation 7.

$$MAP = \frac{1}{TQ}\sum_{i=1}^{TQ} AveP_i \tag{7}$$

Average Precision (AveP): $AveP_i$ is the average precision of the ith question and is computed as shown in Equation 8.

$$AveP = \frac{1}{TQ}\sum_{n=1}^{TQ}\frac{n}{Rank_n} \tag{8}$$

where $rank_i$ is the rank of the nt^{h} correct answer.

MRR-Mean Reciprocal Rank: The mean reciprocal rank (MRR) is a relative score that calculates each question's average or mean of the reciprocal ranks. The reciprocal rank of a question is the multiplicative inverse of the rank of the first correct answer. It is defined by Equation 9.

$$MRR = \frac{1}{TQ}\sum_{n=1}^{TQ}\frac{1}{Rank_n} \tag{9}$$

8.3 Textual and Embedding-based Metrics

We provide a summary of some textual and embedding-based metrics, such as ROUGE and BLEU. Moreover, it has been observed that many evaluation metrics are primarily designed for English, which can limit their applicability to other languages with unique grammatical structures, such as Arabic. To address this challenge, several studies have proposed alternative metrics tailored to Arabic-text generation, including those based on textual features [138] and embedding-based approaches [139]-[140]. These contributions aim to improve the performance of Arabic-text generation models by better accounting for the linguistic characteristics of the target language.

ROUGE: (Recall-Oriented Under-study for Gisting Evaluation): It consists of a set of measures proposed initially to evaluate automatic text summarization [141]. So far, it is the most popular automatic method for evaluating the content of a summary by comparing the tokens of the candidate and the reference; i.e., it counts the number of over-lapping units, such as n-gram, word-pairs and word-sequences between the system-generated summary to be evaluated and the ideal summaries created by humans. The available variants of ROUGE measures are **ROUGE-N** (N =1,2,3,4), ROUGE-L, ROUGE-W and ROUGE-S. However, in this work, we focus on the most popular ROUGE metrics in question-answering systems, which are ROUGE-N and ROUGE-L, that represent a comparison of texts at different granularities. ROUGE-N measures the ratio of the number of overlaps of unigrams/bigrams/trigrams/four − grams (single tokens) between the generated text and the reference text to the total n-grams in the reference text. It is defined by Equation 10:

$$ROUGE - N = \frac{\sum_{sr \in ref\ erences} \sum_{n-gram \in sr} Count_{match}(n-gram)}{\sum_{sr \in ref\ erences} \sum_{n-gram \in sr} Count(n-gram)} \tag{10}$$

BLEU: (Bilingual Evaluation Understudy), a precision-based metric originally proposed for machine translation [142], computes the $n-gram$ overlap between the reference and the hypothesis. Bleu is generally used for generative question-answering system evaluations. Using the following steps, we calculate the precision of different values of *n*. First, we calculate $Count_{clip}$ for any $n-gram$ as shown in Equation 11. Then, we calculate the modified precision score ($precision_n$), as shown in Equation 12.

$$Count_{clip} = \min(Count, Max\_Ref\_Count) \tag{11}$$

$$precision_n = \frac{\sum_{C \in \{C\ andidates\}} \sum_{garm_n \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{garm_{n'} \in C'} Count_{clip}(n-gram')} \tag{12}$$

We add a brevity penalty to handle too short translations. BP is an exponential decay and is calculated as shown in Equation 13.

$$BrevityPenalty(BP) = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases} \tag{13}$$

With *r* being count of words in a reference translation and *c* count of words in a candidate translation. Finally, BLEU is defined as shown in Equation 14 with *N* no. of $n-grams$ (unigram, bigram, 3-gram, 4-gram); W_n denotes weight for each modified precision.

$$BLEU = BP.\exp(\textstyle\sum_{n=1}^{N} W_n \log P_n) \tag{14}$$

AL-BLEU [139]: Traditional metrics like BLEU may not accurately assess the quality of Arabic translations due to the unique linguistic features of the language. To address this issue, Bouamor et al. introduced AL-BLEU (Arabic Language BLEU), a customized metric specifically designed for evaluating the quality of machine-translation (MT) systems in the context of Arabic-language translation. This metric leverages a human judgment corpus, where bilingual experts assess the quality of translations generated by various MT systems. The judgments are then used to calculate the AL-BLEU score, an adapted version of the original BLEU score. By considering the unique aspects of the Arabic language, AL-BLEU provides a more accurate assessment of translation quality for Arabic texts. The authors likely performed experimental comparisons between AL-BLEU and established metrics, such as BLEU and METEOR, to demonstrate the superiority of AL-BLEU in ranking MT systems. The results indicate that AL-BLEU highly correlates with human judgments and outperforms the other metrics.

Morphologically-enriched Embedding Metrics [140]: In the same context, Guzman et al. proposed an approach to evaluate MT in Arabic by using morphologically-enriched embeddings. The proposed method combines word-level and morpheme-level embeddings to capture the complexities of Arabic grammar and syntax [68]. This approach aims to address the limitations of conventional metrics and provide a more comprehensive assessment of translation quality in the context of Arabic intricate morphological structure, which can facilitate better communication and understanding between Arabic speakers and non-native speakers.

Automated Error Analysis [138]: It is a crucial aspect of natural-language processing (NLP) that helps researchers and developers evaluate the performance of their systems. El Kholy and Habash developed AMEANA, which is open-source and specifically designed to identify morphological errors in the output of MT systems compared to a gold standard reference. AMEANA provides detailed statistical reports on morphological errors and generates an altered version of the production that can be used to assess the impact of these errors using various evaluation metrics.

9. DISCUSSION

The increasing interest in developing an Arabic generative conversational AI system is due to the vital role of the Arab world in the global economy and politics. Consequently, there is a pressing need to address the challenges faced in conversational AI to create valuable resources and effective systems.

Notably, alongside the general challenges encountered in conversational AI, the specific challenges related to Arabic include:

- Limited resources: Arabic is a low-resource language, which means that there is a need for high-quality, annotated datasets for training conversational AI models. This scarcity of resources hinders the development of robust and accurate Arabic-language models [82]-[83], [111].
- Diverse Arabic dialects: Arabic has numerous dialects across different regions, each with its vocabulary, pronunciation and grammar. This diversity challenges building conversational AI systems that understand and generate responses in different Arabic dialects [68], [78].
- The complexity of Arabic script and grammar: Arabic has a rich and complex script with many characters and diacritical marks. The morphology and syntax of the language also add to its complexity. These factors make it challenging to process and analyze Arabic text accurately [68].

Advanced machine-learning algorithms, natural-language understanding models and language-specific libraries and tools are required to overcome these challenges. The development of Arabic AI technologies, such as Farasa [92] and CAMeL [93] Tools, to handle the morphology orthography and ambiguity in Arabic language can facilitate research in building Arabic conversational AI. There is also a need for more information, tips and insight to create chatbots that can converse in Arabic with the accuracy and technology of Arabic natural-language understanding improving daily. However, there are still challenges in creating and maintaining Arabic chatbots, compounded by a shortage in skills. Future research should concentrate on building efficient Arabic conversational systems using both word embeddings and ML methods to overcome the challenges of developing Arabic-language bots, such as the need for more dialectal task-oriented dialogue datasets.

Previous research on Arabic conversational systems has primarily focused on Modern Standard Arabic (MSA) and overlooked some local dialects [8], [75]. MSA is morphologically rich and complex compared to English, but easier to parse than Arabic dialects. Rule-based and hand-crafted feature-based methods form the foundation of NLP techniques for Arabic-language dialogue. However, these methods have limitations due to the complex nature of Arabic morphology orthographic variations, dialectal differences and a high degree of ambiguity. In contrast to English and other chatbots, Arabic chatbots still require further enhancement to become more robust and efficient. Moreover, recently, researchers have exploited the potential of chatGPT, such as Siu et al. [143] who mentioned that GPT models have limited capabilities for low-resource languages such as Arabic. Khoshafah [144], on the other hand, focuses on ChatGPT application in Arabic-English translation and advises users against relying solely on it for translations, recommending the involvement of a professional translator for more accurate results.

Several areas require attention and development for the future of Arabic conversational AI. Firstly, deep-learning techniques have shown promise in achieving high performance for conversational AI systems in English. However, constructing a large dataset for different question types in Arabic is necessary to achieve similar results. Secondly, developing solid tools, such as Farasa [92] and CAMeL Tools [93], to facilitate resolving Arabic morphology orthographic variations, dialectal differences and ambiguity, can improve the accuracy of Arabic conversational AI. Thirdly, improving publicly available pre-trained language models' (PTLMs) scalability, such as AraBERT [101] and ARBERT [145], by augmenting the training data, can enhance the performance of Arabic Conversational AI. Scaling the model's size or the amount of training data improves model capacity for downstream tasks [146]-[147]. Finally, focusing on building Arabic conversational AI on domains, such as criminal law, intellectual property and tax law texts, can enhance the accuracy and relevance of the responses. These developments can bridge the gap in dialectal speech recognition and overcome the challenges of limited resources, lack of standard orthographic rules and lack of definition in Arabic dialects [61][65][68].

10. Conclusion and Future Work

This paper reviews significant works on Arabic conversational systems. We found that these systems have received increasing attention from the NLP research community. We noticed that Arabic conversational agents commonly use rule-based and hand-crafted feature-based methods. At the same time, deep-learning techniques extensively explored in English dialogue systems require further improvement for Arabic. We explore and discuss the challenges of the Arabic language and the proposed solutions in the literature. Three main evaluation methods for question-answering systems and conversational agents are identified: traditional evaluation metrics, ranking-based metrics and textual and embedding-based metrics. We found that deep-learning techniques show great promise in addressing the challenges of Arabic dialogue systems. However, constructing an extensive dataset encompassing various question types specific to the Arabic language is crucial for leveraging the potential of deep learning in Arabic dialogue systems. We highlight the significance of creating tools, such as MADAMIRA, Farasa and CAMeL Tools. These tools address the challenges of Arabic morphology orthographic complexities and ambiguity, ultimately supporting the development of Arabic conversational AI research.

As future work, we plan to create an efficient Arabic conversational system by leveraging contextualized word embedding and deep-learning methods. This approach aims to overcome challenges in creating Arabic-language bots. We also propose exploring Arabic spoken or written dialectical conversational agents, which researchers have not extensively investigated yet. Furthermore, we plan to improve publicly available PTLMs scalability, like ARBERT, by augmenting the training data. As shown by Kaplan et al. [146] and Zhao et al. [147], increasing the model size or the amount of training data can significantly improve the model performance on downstream tasks. Additionally, we aim to develop specialized Arabic question-answering systems tailored to specific domains, such as the judiciary and legal sector. We hope that this survey can provide researchers with a comprehensive overview of the current state of Arabic conversational systems, encompassing advancements in technique, employed approaches and outstanding challenges to facilitate progress and innovation within this field.

REFERENCES

[1] M. Adam, M. Wessel and A. Benlian, "Ai-based Chatbots in Customer Service and their Effects on User Compliance," Electronic Markets, vol. 31, no. 2, pp. 427–445, 2021.

[2] L. T. Car, D. A. Dhinagaran, B. M. Kyaw et al., "Conversational Agents in Health Care: Scoping Review and Conceptual Analysis," Journal of Medical Internet Research, vol. 22, no. 8, p. e17158, 2020.

[3] M. M. Mariani, N. Hashemi and J. Wirtz, "Artificial Intelligence Empowered Conversational Agents: A Systematic Literature Review and Research Agenda," J. of Business Research, vol. 161, p. 113838, 2023.

[4] M. Hijjawi, Z. Bandar, K. Crockett and D. Mclean, "Arabchat: An Arabic Conversational Agent," Proc. of the 6th IEEE Int. Conf. on Computer Science and Information Technology (CSIT), pp. 227–237, 2014.

[5] M. Hijjawi, H. Qattous and O. Alsheiksalem, "Mobile Arabchat: An Arabic Mobile-based Conversational Agent," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 6, no. 10, 2015.

[6] M. Hijjawi, Z. Bandar and K. Crockett, "The Enhanced Arabchat: An Arabic Conversational Agent," Int. J. of Advanced Computer Science and Applications, vol. 7, no. 2, 2016.

[7] S. S. Aljameel, J. D. O’Shea, K. A. Crockett, A. Latham and M. Kaleem, "Development of an Arabic Conversational Intelligent Tutoring System for Education of Children with ASD," Proc. of the 2017 IEEE Int. Conf. on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 24–29, Annecy, France, 2017.

[8] D. Al-Ghadhban and N. Al-Twairesh, "Nabiha: An Arabic Dialect Chatbot," Int. J. of Advanced Computer Science and Applications, vol. 11, no. 3, DOI: 10.14569/IJACSA.2020.0110357, 2020.

[9] D. Baidoo-Anu and L. O. Ansah, "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of Chatgpt in Promoting Teaching and Learning," SSRN 4337484, [Online], Available: https://dx.doi.org/10.2139/ssrn.4337484, 2023.

[10] E. Kasthuri and S. Balaji, "A Chatbot for Changing Lifestyle in Education," Proc. of the 2021 3rd IEEE Int. Conf. on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 1317–1322, Tirunelveli, India, 2021.

[11] M. A. Kuhail, B. Abu Shawar and R. Hammad, Trends, Applications and Challenges of Chatbot Technology, Advances in Web Technologies and Engineering, IGI Global, ISBN10: 1668462346, 2023.

[12] A. D Fadhil et al., "Ollobot-towards a Text-based Arabic Health Conversational Agent: Evaluation and Results," Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2019), pp. 295–303, Varna, Bulgaria, 2019.

[13] M. Boussakssou, H. Ezzikouri and M. Erritali, "Chatbot in Arabic Language Using Seq to Seq Model," Multimedia Tools and Applications, vol. 81, no. 2, pp. 2859–2871, 2022.

[14] L. Xu, L. Sanders, K. Li et al., "Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review," JMIR Cancer, vol. 7, no. 4, p. e27850, 2021.

[15] L. Athota, V. K. Shukla, N. Pandey and A. Rana, "Chatbot for Healthcare System Using Artificial Intelligence," Proc. of the 2020 IEEE 8th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pp. 619–622, Noida, India, 2020.

[16] A. Stöckl, "Classification of Chatbot Inputs," [Online], Available: https://www.researchgate.net/publication/ 318661551_Classification_of_Chatbot_Inputs, 2017.

[17] N. K. Manaswi and S. John, Deep Learning with Applications Using Python, ISBN-10: 1484240510, Springer, 2018.

[18] N. M. Radziwill and M. C. Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents," arXiv preprint, arXiv: 1704.04579, 2017.

[19] Amazon Lab126, "Amazon Alexa," [Online], Available: https://en.wikipedia.org/wiki/Amazon_Alexa/, 2013.

[20] S. Alhumoud et al., "Rahhal: A Tourist Arabic Chatbot," Proc. of the 2022 2nd IEEE Int. Conf. of Smart Systems and Emerging Technologies (SMARTTECH), pp. 66–73, 2022.

[21] Al-H. Al-Ajmi and N. Al-Twairesh, "Building an Arabic Flight Booking Dialogue System Using a Hybrid Rule-based and Data Driven Approach," IEEE Access, vol. 9, pp. 7043–7053, 2021.

[22] A. M. Rahman, A. Al Mamun and A. Islam, "Programming Challenges of Chatbot: Current and Future Prospective," Proc. of the 2017 IEEE Region 10 Humanitarian Technology Conf. (R10-HTC), pp. 75–78, Dhaka, Bangladesh, 2017.

[23] A. Radford, K. Narasimhan, T. Salimans et al., "Improving Language Understanding by Generative Pre-training," pp. 1-12, [Online], Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language Models Are Unsupervised Multitask Learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[25] T. Brown, B. Mann, N. Ryder et al., "Language Models are Few-shot Learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[26] OpenAI, "Gpt-4 Technical Report," [Online], Available: https://cdn.openai.com/papers/gpt-4.pdf, 2023.

[27] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint, arXiv: 1810.04805, 2018.

[28] Romal Thoppilan et al., "LaMDA: Language Models for Dialog Applications," arXiv preprint, arXiv: 2201.08239, 2022.

[29] P. Brereton, B. A Kitchenham, D. Budgen, M. Turner and M. Khalil, "Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain," Journal of Systems and Software, vol. 80, no. 4, pp. 571–583, 2007.

[30] Y. Xiao and M. Watson, "Guidance on Conducting a Systematic Literature Review," Journal of Planning Education and Research, vol. 39, no. 1, pp. 93–112, 2019.

[31] A. M. Turing, Computing Machinery and Intelligence, in Book: Parsing the Turing Test, pp. 23–65, Springer, 2009.

[32] J. Weizenbaum et al., "Eliza: A Computer Program for the Study of Natural Language Communication between Man and Machine," Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.

[33] E. Adamopoulou and L. Moussiades, "Chatbots: History, Technology and Applications," Machine Learning with Applications, vol. 2, p. 100006, 2020.

[34] Loebner Prize, [Online], Available: https://en.wikipedia.org/wiki/Loebner_Prize.

[35] Ina, [Online], Available: https://onlim.com/en/the-history-of-chatbots/, 2021.

[36] R. Wallace, Artificial Linguistic Internet Computer Entity (Alice), [Online], Available: https://www.chatbots.org/chatbot/a.l.i.c.e/, 1995.

[37] B. Abu Shawar and E. Atwell, "Chatbots: Are They Really Useful?" J. for Language Technology and Computational Linguistics, vol. 22, no. 1, pp. 29–49, 2007.

[38] A. M. Ezzeldin and M. Shaheen, "A Survey of Arabic Question Answering: Challenges, Tasks, Approaches, Tools and Future Trends," Proc. of the 13th Int. Arab Conf. on Information Technology (ACIT 2012), pp. 1–8, 2012.

[39] Y. Wu, G. Wang, W. Li and Z. Li, "Automatic Chatbot Knowledge Acquisition from Online Forum *via* Rough Set and Ensemble Learning," Proc. of the 2008 IEEE IFIP Int. Conf. on Network and Parallel Computing, pp. 242–246, Shanghai, China, 2008.

[40] G. Molnár and Z. Szüts, "The Role of Chatbots in Formal Education," Proc. of the 2018 IEEE 16th Int. Symposium on Intelligent Systems and Informatics (SISY), pp. 000197–000202, Subotica, Serbia, 2018.

[41] S. Worswick, Mitsuku Chatbot, [Online], Available: https://www.pandorabots.com/mitsuku/, 2005.

[42] David Ferrucci, IBM Watson, [Online], Available: https://www.ibm.com/watson, 2006.

[43] T. G. D. Kittlaus and UCLA Alumnus Adam Cheyer, Apple Siri, [Online], Available: https://www.apple.com/siri/, 2010.

[44] Google, Google Assistant, [Online], Available: https://assistant.google.com/, 2012.

[45] Google Cloud, Dialogflow, [Online], Available: https://cloud.google.com/dialogflow, 2016.

[46] Microsoft Azure, Language Understanding (LUIS), [Online], Available: https://www.luis.ai/, 2017.

[47] Amazon, Amazon Lex- Conversational AI for Chatbots, [Online], Available: https://aws.amazon.com/ lex/?nc=sn&loc=0, 2017.

[48] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention Is All You Need," Advances in Neural Information

Processing Systems, vol. 30, arXiv:1706.03762, 2017.

[49] S. Hussain, O. A. Sianaki and N. Ababneh, "A Survey on Conversational Agents/Chatbots Classification and Design Techniques, Proc. of the Workshops of the Int. Conf. on Advanced Information Networking and Applications, pp. 946–956, Springer, 2019.

[50] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," Proc. of the IFIP Int. Conf. on Artificial Intelligence Applications and Innovations, pp. 373–383, Springer, 2020.

[51] E. H. Almansor and F.h K. Hussain, "Survey on Intelligent Chatbots: State-of-the-art and Future Research Directions," Proc. of Conf. on Complex, Intelligent and Software Intensive Systems, pp. 534–543, Springer, 2019.

[52] R. Dale, "The Return of the Chatbots," Natural Language Engineering, vol. 22, no. 5, pp. 811–817, 2016.

[53] M. B. Hoy, "Alexa, Siri, Cortana and More: An Introduction to Voice Assistants," Medical Reference Services Quarterly, vol. 37, no. 1, pp. 81–88, 2018.

[54] A. L. Guzman, "Making AI Safe for Humans: A Conversation with Siri," Chapter in Book: Socialbots and Their Friends, 1st Edn., pp. 85–101, Routledge, 2016.

[55] K. Nimavat and T. Champaneria, "Chatbots: An Overview of Types, Architecture, Tools and Future Possibilities," Int. J. Sci. Res. Dev. (IJSRD), vol. 5, no. 7, pp. 1019–1024, 2017.

[56] H. T. Hien, P.-N. Cuong, L. N. H. Nam, H. L. T. K. Nhung and L. D. Thang, "Intelligent Assistants in Higher-education Environments: The FIT-EBoT, a Chatbot for Administrative and Learning Support," Proc. of the 9th Int. Symposium on Inform. and Comm. Techn. (SoICT '18), pp. 69–76, 2018.

[57] K. Ramesh, S. Ravishankaran, A. Joshi and K. Chandrasekaran, "A Survey of Design Techniques for Conversational Agents," Proc. of the Int. Conf. on Information, Communication and Computing Technology, pp. 336–350, Springer, 2017.

[58] Z. Ji, Z. Lu and H. Li, "An Information Retrieval Approach to Short Text Conversation," arXiv preprint, arXiv: 1408.6988, 2014.

[59] R. Artstein, S. Gandhe, J. Gerten, A. Leuski and D. Traum, "Semi-formal Evaluation of Conversational Characters," in Book: Languages: From Formal to Natural, vol. 5533, pp. 22–35, Springer, 2009.

[60] E. Atwell, "A Chatbot As a Question Answering Tool," DOI: 10.17758/ur.u0915120, 2015.

[61] R. Artstein, A. Gainer, K. Georgila et al., "New Dimensions in Testimony Demonstration," Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 32–36, San Diego, California, USA, 2016.

[62] R. Artstein, A. Leuski, H. Maio et al., "How Many Utterances Are Needed to Support Time-offset Interaction?" Proc. of the 28th Int. Florida Artificial Intelligence Research Society Conference (FLAIRS 2015), pp. 144-149, 2015.

[63] D. Traum, K. Georgila, R. Artstein and A. Leuski, "Evaluating Spoken Dialogue Processing for Time-offset Interaction," Proc. of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 199–208, Prague, Czech Republic, 2015.

[64] D. Traum et al., "New Dimensions in Testimony: Digitally Preserving a Holocaust Survivors' Interactive Storytelling," Proc. of the 8th Int. Conf. on Interactive Digital Storytelling (ICIDS 2015), Copenhagen, Denmark, Proceedings 8, pp. 269–281, Springer, 2015.

[65] D. Abu Ali et al., "A Bilingual Interactive Human Avatar Dialogue System," Proc. of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pp. 241–244, Melbourne, Australia, 2018.

[66] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, no. 4, pp. 1–22, 2009.

[67] K. C. Ryding, A Reference Grammar of Modern Standard Arabic, ISBN: 0521777712, Cambridge University Press, 2005.

[68] N. Y. Habash, Introduction to Arabic Natural Language Processing, ISBN: 1598297953, Springer Nature, 2022.

[69] C. Zhai, "A Systematic Review on Artificial Intelligence Dialogue Systems for Enhancing English As Foreign Language Students' Interactional Competence in the University," Computers and Education: Artificial Intelligence, vol. 4, p. 100134, 2023.

[70] K. Chemnad and A. Othman, "Advancements in Arabic Text-to-speech Systems: A 22-year Literature Review," IEEE Access, vol. 11, pp. 30929 – 30954, 2023.

[71] K. Darwish, N. Habash, M. Abbas et al., "A Panoramic Survey of Natural Language Processing in the Arab World," Communications of the ACM, vol. 64, no. 4, pp. 72–81, 2021.

[72] Amira Dhouib et al., "Arabic Automatic Speech Recognition: A Systematic Literature Review," Applied Sciences, vol. 12, no. 17, p. 8898, 2022.

[73] M. Hijjawi and Y. Elsheikh, "Arabic Language Challenges in Text Based Conversational Agents Compared to the English Language," IJCSIT, vol. 7, no. 5, pp. 1–13, 2015.

[74] N. Habash, M. T. Diab and O. Rambow, "Conventional Orthography for Dialectal Arabic," Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC'12), pp. 711–718, Istanbul, Turkey, 2012.

[75] D. Abu Ali and N. Habash, "Botta: An Arabic Dialect Chabot," Proc. of COLING 2016, the 26th Int. Conf. on Computational Linguistics: System Demonstrations, pp. 208–212, Osaka, Japan, 2016.

[76] W. Zaghouani et al., "Large Scale Arabic Error Annotation: Guidelines and Framework," Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14), pp. 2362-2369, Reykjavik, Iceland, 2014.

[77] R. Eskander, N. Habash, O. Rambow and N. Tomeh, "Processing Spontaneous Orthography," Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 585–595, Atlanta, USA, 2013.

[78] N. Habash, R. Roth, O. Rambow, R. Eskander and N. Tomeh, "Morphological Analysis and Disambiguation for Dialectal Arabic," Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Lang. Techn., pp. 426–432, Atlanta, USA, 2013.

[79] N. Habash, A. Soudi and T. Buckwalter, "On Arabic Transliteration," in Chapter: Arabic Computational Morphology: Knowledge-based and Empirical Methods, Part of the Text, Speech and Language Technology Book Series, vol. 38, pp. 15–22, 2007.

[80] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0," Catalog no. LDC2002L49, Linguistic Data Consortium, University of Pennsylvania, 2002.

[81] T. H. Alwaneen et al., "Arabic Question Answering System: A Survey," Artificial Intelligence Review, vol. 55, no. 1, pp. 207–253, 2022.

[82] H. Mozannar, K. El Hajal, E. Maamary and H. Hajj, "Neural Arabic Question Answering," arXiv preprint, arXiv: 1906.05394, 2019.

[83] W. Antoun, F. Baly and H. Hajj, "Araelectra: Pre-training Text Discriminators for Arabic Language Understanding," arXiv preprint, arXiv: 2012.15516, 2020.

[84] Jonathan H Clark et al., "TyDi QA: A Benchmark for Information-seeking Question Answering in Typologically Diverse Languages," Transactions of the Association for Computational Linguistics, vol. 8, pp. 454–470, 2020.

[85] B. Abu Shawar, "A Chatbot As a Natural Web Interface to Arabic Web QA," Int. J. of Emerging Technologies in Learning (iJET), vol. 6, no. 1, pp. 37–43, 2011.

[86] M. A. Yaghan, ""Arabizi": A Contemporary Style of Arabic Slang," Design Issues, vol. 24, no. 2, pp. 39– 52, 2008.

[87] K. Shaalan, "A Survey of Arabic Named Entity Recognition and Classification," Computational Linguistics, vol. 40, no. 2, pp. 469–510, 2014.

[88] W. Bakari, P. Bellot and M. Neji, "Researches and Reviews in Arabic Question Answering: Principal Approaches and Systems with Classification," Proc. of the Int. Arab Conf. on Information Technology (ACIT'2016), pp. 1–9, 2016.

[89] Arfath Pasha et al., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14), pp. 1094–1101, Reykjavik, Iceland, 2014.

[90] N. Habash and O. Rambow, "Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop," Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 573–580, Ann Arbor, Michigan, USA, 2005.

[91] N. Habash, O. Rambow and R. Roth, "MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization," Proc. of the 2nd Int. Conf. on Arabic Language Resources and Tools (MEDAR), vol. 41, p. 62, Cairo, Egypt, 2009.

[92] A. Abdelali, K. Darwish, N. Durrani and H. Mubarak, "Farasa: A Fast and Furious Segmenter for Arabic," Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 11–16, San Diego, California, USA, 2016.

[93] O. Obeid et al., "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing," Proc. of the 12th Language Resources and Evaluation Conference, pp. 7022–7032, Marseille, France, 2020.

[94] Zaid Alyafeai et al., "Masader: Metadata Sourcing for Arabic Text and Speech Data Resources," arXiv preprint, arXiv: 2110.06744, 2021.

[95] N. Al-Twairesh, H. Al-Khalifa, A. Alsalman and Y. Al-Ohali, "Sentiment Analysis of Arabic Tweets: Feature Engineering and a Hybrid Approach," arXiv preprint, arXiv: 1805.08533, 2018.

[96] P. Liu, X. Qiu, J. Chen and X.-J. Huang, "Deep Fusion LSTMs for Text Semantic Matching," Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 1034–1043, Berlin, Germany, 2016.

[97] J. Mueller and A. Thyagarajan, "Siamese Recurrent Architectures for Learning Sentence Similarity," Proc. of the AAAI Conf. on Artificial Intelligence, vol. 30, no. 1, DOI: 10.1609/aaai.v30i1.10350, 2016.

[98] N. Van Tu et al., "A Deep Learning Model of Multiple Knowledge Sources Integration for Community Question Answering," VNU J. of Science: Computer Sci. and Comm. Eng., vol. 37, no. 1, 2021.

[99] J. D. Ming-Wei C. Kenton and L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. of NAACL-HLT, pp. 4171– 4186, Minneapolis, USA, 2019.

[100] A. Hamza, N. En-Nahnahi and S. El Alaoui Ouatik, "Exploring Contextual Word Representation for Arabic Question Classification," Proc. of the 2020 1st IEEE Int. Conf. on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1–5, Meknes, Morocco, 2020.

[101] W. Antoun, F. Baly and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," arXiv preprint, arXiv: 2003.00104, 2020.

[102] A. Almiman, N. Osman and M. Torki, "Deep Neural Network Approach for Arabic Community Question Answering," Alexandria Engineering Journal, vol. 59, no. 6, pp. 4427–4434, 2020.

[103] D. Elalfy, W. Gad and R. Ismail, "A Hybrid Model to Predict Best Answers in Question Answering Communities," Egyptian Informatics Journal, vol. 19, no. 1, pp. 21–31, 2018.

[104] D. Elalfy, W. Gad and R. Ismail, "Predicting Best Answer in Community Questions Based on Content and Sentiment Analysis," Proc. of the 2015 IEEE 7th Int. Conf. on Intelligent Computing and Information Systems (ICICIS), pp. 585–590, Cairo, Egypt, 2015.

[105] A. Hamza et al., "An Arabic Question Classification Method Based on New Taxonomy and Continuous Distributed Representation of Words," Journal of King Saud University-Computer and Information Sciences, vol. 33, no. 2, pp. 218–224, 2021.

[106] N. Othman, R. Faiz and K. Smaïli, "Learning English and Arabic Question Similarity with Siamese Neural Networks in Community Question Answering Services," Data & Knowledge Engineering, vol. 138, p. 101962, 2022.

[107] W.-N. Zhang et al., "Capturing the Semantics of Key Phrases Using Multiple Languages for Question Retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 4, pp. 888–900, 2015.

[108] W. Bakari and M. Neji, "A Novel Semantic and Logical-based Approach Integrating RTE Technique in the Arabic Question–answering," International Journal of Speech Technology, vol. 25, pp. 1–17, 2020.

[109] W. Bakari, P. Bellot and M. Neji, "AQA-WebCorp: Web-based Factual Questions for Arabic," Procedia Computer Science, vol. 96, pp. 275–284, 2016.

[110] N. A. Al-Madi et al., "An Intelligent Arabic Chatbot System Proposed Framework," Proc. of the 2021 IEEE Int. Conf.on Information Technology (ICIT), pp. 592–597, Amman, Jordan, 2021.

[111] K. Alsubhi, A. Jamal and A. Alhothali, "Deep Learning-based Approach for Arabic Open Domain Question Answering," PeerJ Computer Science, vol. 8, p. e952, 2022.

[112] A. Soliman, K Eissa and S. R El-Beltagy, "Aravec: A Set of Arabic Word Embedding Models for Use in Arabic NLP," Procedia Computer Science, vol. 117, pp. 256–265, 2017.

[113] M. Ben-Sghaier, W. Bakari and M. Neji, "An Arabic Question-answering System Combining a Semantic and Logical Representation of Texts," Proc. of the Int. Conf. on Intelligent Systems Design and Applications, pp. 735–744, Springer, 2017.

[114] V. Karpukhin et al., "Dense Passage Retrieval for Open-domain Question Answering," arXiv preprint, arXiv: 2004.04906, 2020.

[115] H. Abdel-Nabi, A. Awajan and M. Z. Ali, "Deep Learning-based Question Answering: A Survey," Knowledge and Information Systems, vol. 65, no. 4, pp. 1399–1485, 2023.

[116] S. T. Chung and R. L. Morris, "Isolation and Characterization of Plasmid Deoxyribonucleic Acid from Streptomyces Fradiae," Paper presented at the 3rd Int. Symposium on the Genetics of Industrial Microorganisms, University of Wisconsin, Madison, 4–9 June 1978.

[117] Z. Hao, A. AghaKouchak, N. Nakhjiri and A. Farahmand, "Global Integrated Drought Monitoring and Prediction System," Scientific Data, vol. 1, p. 853801, 2014.

[118] S. A. Babichev, J. Ries and A. I. Lvovsky, "Quantum Scissors: Teleportation of Single-mode Optical States by Means of a Nonlocal Single Photon," arXiv preprint, arXiv: 0208066v1, 2002.

[119] M. Beneke, G. Buchalla and I. Dunietz, "Mixing Induced CP Asymmetries in Inclusive B Decays," Physics Letters, B393, pp. 132–142, 1997.

[120] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Subword Information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135– 146, 2017.

[121] B. Stahl, "DeepSIP: Deep Learning of Supernova Ia Parameters," Astrophysics Source Code Library, Bibcode: 2020ascl.soft06023S, 2020.

[122] R. Yan, Y. Song and H. Wu, "Learning to Respond with Deep Neural Networks for Retrieval-based Human-computer Conversation System," Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '16), pp. 55–64, 2016.

[123] Y. Wu, W. Wu, C. Xing, M. Zhou and Z. Li, "Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots," arXiv preprint, arXiv: 1612.01627, 2016.

[124] M. Wang, Z. Lu, Hang Li and Q. Liu, "Syntax-based Deep Matching of Short Texts," Proc. of the 24th Int. Joint Conf. on Artificial Intelligence (IJCAI 2015), pp. 1354-1361, 2015.

[125] I. Serban, A. Sordoni, Y. Bengio, A. Courville and J. Pineau, "Building End-to-end Dialogue Systems Using Generative Hierarchical Neural Network Models," Proc. of the AAAI Conf. on Artificial Intelligence, vol. 30, pp. 3776–3783, 2016.

[126] W. Zhang et al., "Context-sensitive Generation of Open-domain Conversational Responses," Proc. of the 27th Int. Conf. on Computational Linguistics, pp. 2437–2447, Santa Fe, New Mexico, USA, 2018.

[127] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai and X. Huang, "Pre-trained Models for Natural Language Processing: A Survey," Science China Technological Sciences, vol. 63, no. 10, pp. 1872–1897, 2020.

[128] K. S. Kalyan, A. Rajasekharan and S. Sangeetha, "AMMUS: A Survey of Transformer-based Pre-trained

Models in Natural Language Processing," arXiv preprint, arXiv: 2108.05542, 2021.

[129] M. E. Peters et al., "Deep Contextualized Word Representations," Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers), pp. 2227–2237, New Orleans, USA, 2018.

[130] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer," The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020.

[131] P. Rajpurkar, J. Zhang, K. Lopyrev and P Liang, "Squad: 100,000+ Questions for Machine Comprehension of Text," arXiv preprint, arXiv: 1606.05250, 2016.

[132] A. Celikyilmaz, E. Clark and J. Gao, "Evaluation of Text Generation: A Survey," arXiv preprint, arXiv: 2006.14799, 2020.

[133] M. Kusner, Y. Sun, N. Kolkin and K. Weinberger, "From Word Embeddings to Document Distances," Proc. of the Int. Conf. on Machine Learning, PMLR, vol. 37, pp. 957-966, 2015.

[134] E. Clark, A. Celikyilmaz and N. A. Smith, "Sentence Mover's Similarity: Automatic Evaluation for Multi-sentence Texts," Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2748–2760, Florence, Italy, 2019.

[135] W. Zhao et al., "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance," arXiv preprint, arXiv: 1909.02622, 2019.

[136] P. Forner et al., "Evaluating Multilingual Question Answering Systems at CLEF," Proc. of the 7th Int. Conf. on Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010.

[137] A. Peñas and A. Rodrigo, "A Simple Measure to Assess Non-response," Proc. of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 1415–1424, Portland Oregon, 2011.

[138] A. El Kholy and N. Habash, "Automatic Error Analysis for Morphologically Rich Languages," Proc. of Machine Translation Summit XIII: Papers, pp. 225-232, 2011.

[139] H. Bouamor, H. Alshikhabobakr, B. Mohit and K. Oflazer, "A Human Judgement Corpus and a Metric for Arabic MT Evaluation," Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 207–213, Doha, Qatar, 2014.

[140] F. Guzmán, H. Bouamor, R. Baly and N. Habash, "Machine Translation Evaluation for Arabic Using Morphologically-enriched Embeddings," Proc. of COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, pp. 1398–1408, Osaka, Japan, 2016.

[141] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Proc. of Text Summarization Branches Out, pp. 74–81, Association for Computational Linguistics, Barcelona, Spain, 2004.

[142] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, Philadelphia, USA, 2002.

[143] S. C. Siu, "ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation," SSRN 4448091, DOI: 10.2139/ssrn.4448091, 2023.

[144] F. Khoshafah, "ChatGPT for Arabic-English Translation: Evaluating the Accuracy," Research Square, DOI: 10.21203/rs.3.rs-2814154/v1, 2023.

[145] M. Abdul-Mageed, A. Elmadany and M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," arXiv preprint, arXiv: 2101.01785, 2020.

[146] Jared Kaplan et al., "Scaling Laws for Neural Language Models," arXiv: 2001.08361, 2020.

[147] Wayne Xin Zhao et al., "A Survey of Large Language Models," arXiv preprint, arXiv: 2303.18223, 2023.

ملخص البحث:

تقدّم هذه الدّراسة عرضاً شاملاً للأعمال البحثية في هذا المجال؛ من أجْل بيان مزايا الأنظمة المتعلّقة بالمحادثة باللّغة العربية، الى جانب التّحدّيات التي تعترضها. يبدأ التّحليل بنظرةٍ على تاريخ أنظمة المحادثة وتصنيفها، ثمّ يتمّ الانتقال الى الصّعوبات والتّحدّيات التي تواجه تصميم أنظمة المحادثة باللّغة العربية باستخدام تقنيات الذّكاء الاصطناعي. كذلك تتطرّق الدّراسة الى ما حصل من تطوّر في أنظمة المحادثة باللّغة العربية بفضل التّقدُّم الّذي حدث في تقنيات التّعلُّم العميق. بالإضافة الى ما سبق، تستعرض الدّراسة مؤشّرات التّقييم الّتي تُستخدم في الحُكم على أنظمة المحادثة باللّغة العربية والمفاضلة بينها.