IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 3, 2022, accepted February 21, 2022, date of publication February 28, 2022, date of current version March 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3155521

Recent Developments in Arabic Conversational AI: A Literature Review

AHLAM FUAD AND MAHA AL-YAHYA

Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh 4545, Saudi Arabia

Corresponding author: Ahlam Fuad (aabdulghni@ksu.edu.sa)

This research is supported by a grant from the Researchers Supporting Project No. RSP-2021/286, King Saud University, Riyadh, Saudi Arabia.

ABSTRACT Conversational AI is one of the most active research areas in AI, and it has gained more attention from academia as well as industry. Given recent advancements in several conversational AI systems in addition to the availability of several datasets, the aim of this study is to explore the landscape of Arabic text-based conversational AI systems. In this work, we provide a thorough review of recent Arabic conversational AI systems. We group them into three categories based on their functionality: (1) question-answering (QA) systems, (2) task-oriented dialogue systems (DS), and (3) chatbots. Furthermore, we describe the common datasets used in building and evaluating conversational AI systems in Arabic. Few surveys have targeted the conversational AI field for the Arabic language, and we aim to cover this gap with this study. Our contribution focuses on reviewing and analyzing the literature in the field and highlighting future research directions towards human-like conversational AI systems in Arabic.

INDEX TERMS Conversational AI systems, Arabic language, question answering systems, task-oriented dialogue systems, chatbots.

I. INTRODUCTION

Recently, the idea of interacting with a computer using either text or voice has become more plausible with the emergence of advanced neural deep learning (DL) methods. These methods have improved the performance of digital assistants and chatbots. Such systems aim not only to mimic human conversation but also to answer questions on different topics in different domains. Furthermore, achieving complex tasks such as travel planning with these systems is considered one of the longest-running goals in the field of artificial intelligence (AI). Advances in AI, in addition to the availability of huge amounts of data and vast computing power, have led to an increased interest from both the research and industry communities in conversational AI and have led to a new generation of conversational systems [1].

Conversational AI has received considerable attention in the design of natural user interfaces. The field is growing rapidly and attracting many researchers in the natural language processing (NLP), information retrieval (IR), and machine learning (ML) communities. However, the most considerable achievement is applying DL and reinforcement learning (RL) techniques to conversational AI. Thus, the research community started directing its focus towards improving end-to-end models for conversational AI systems [2], [3].

In the literature, various terms are used to describe conversational AI systems, including chatbots, dialogue systems, virtual digital assistants, personal digital assistants, conversational user interfaces, and conversational agents. Thus, in this survey, conversational AI systems will be reviewed according to three categories based on their functionality: (1) question-answering (QA) systems, (2) task-oriented dialogue systems (DS), and (3) chatbots.

In general, there are slight differences to distinguish between these systems. QA systems aim to provide direct and concise answers to user queries based on knowledge derived from diverse data sources such as web documents and existing datasets. Task-oriented DS aim to accomplish specific tasks for users such as meeting scheduling, flight booking and restaurant reservation. Finally, chatbots aim to converse appropriately and seamlessly with users and may provide useful recommendations to them [1].

Task-oriented DS are developed based on a predefined task-specific schema (a set of user intents, a set of dialogue acts for each intent, and slot-value pairs) and domain-specific

The associate editor coordinating the review of this manuscript and approving it for publication was R. K. Tripathy.

TABLE 1. Main differences between question-answering (QA) systems, task-oriented dialogue systems (DS), and chatbots.

Characteristics	QA Systems	Task-oriented DS	Chatbots
Architecture	A predefined schema	A predefined task-specific schema	Not defined
Task	Defined	Defined	Not defined
Domain	Close/open	Domain-specific	Open domain
Goal	Answer questions	Achieve a certain task	Generate natural responses
Turn	Single / multi	Multi	Multi
Input/Output	The input to the system is a question and the output is an answer	The input to the system is a query and output is not necessarily an answer	The input to the system is a query or a sentence and the output is not necessarily an answer
Method	Retrieval-based	Retrieval-/Generative-based	Retrieval-/Generative-based

components such as natural language understanding (NLU), dialogue management, and language generation. Unlike task-oriented DS, chatbots deal with open-domain knowledge. Furthermore, task-oriented DS and chatbots differ from QA systems in that they are more general. This is because the input to the system is not a question but a dialogue, and the output is not necessarily an answer. The differences between the three categories of systems are illustrated in Table 1.

Recent trends in conversational AI have led to a proliferation of surveys that focus on conversational agents and DL as a new era for these systems, including open-domain conversational agents [3], task-oriented DS [4], QA systems [5], [6], chatbots [7]–[9], and open-domain DS [10].

For Arabic, surveys have tended to study the existing works of conversational agents until 2018, including QA systems and task-oriented DS [11], QA systems [12], [13], and chatbots [14], [15]. All these studies have mainly focused on the challenges of building these systems in Arabic. Although DL approaches have revolutionized NLP for high-resource languages such as English, Arabic is still in its infancy. This work significantly expands the scope of several existing surveys [13]–[15] by going beyond chatbots and QA systems to provide what we believe is the first general survey of text-based conversational AI systems in Arabic. This survey aims to shed light on the recent studies focusing on Arabic conversational AI to provide an overview of the advancement of this field. Thus, the objectives of this study are as follows:

- Explore the landscape of published text-based conversational AI studies developed for the Arabic language, reviewing the progress made and the challenges facing the research community.
- Provide a valuable resource for researchers, students, and software developers, including a unified view of the field's recent studies that explores the techniques they used and the challenges they still face. This may help to gain obvious insights into understanding and creating modern Arabic conversational AI systems, which will be important for creating more Arabic knowledge resources and services available to different users in natural and intuitive ways.

This study is organized into eight sections. Section 2 presents the methodology used to conduct this survey. Section 3 discusses QA systems and tasks and explores the related studies for the Arabic language as well as the approaches to build these systems. Section 4 discusses the task-oriented DS that have been built for the Arabic language, while Section 5 discusses chatbots in Arabic. In Section 6, we explore the existing Arabic conversational AI datasets. We discuss some challenges and open issues for Arabic conversational AI systems in Section 7. Finally, we present our conclusions and discuss future trends in Section 8.

II. METHODOLOGY

In this study, we searched the literature using several steps, including search string selection, databases and search protocol, and data coding and synthesis. For search string selection, we determined the relevant keywords and used several search strings in order to obtain the relevant literature. Since this survey focuses on the Arabic language, we opted for the following search query: "Arabic chatbots" OR "Arabic conversational agents" OR "Arabic question answering systems" OR "Arabic dialogue system" OR "Arabic task-oriented dialogue system." We excluded the papers written in languages other than English.

Databases and search protocol: The search was performed in several computing and science databases, including IEEE Xplore, Science Direct, SpringerLink, Google Scholar, Scopus, and the ACM Digital Library. The search fields were chosen based on the options provided by each database. Considering that the concept of Arabic conversational agents is advancing rapidly and that the available surveys focused on works before 2018 [14], [15], the search range is from 2018 to 2021.

Data coding and synthesis: For each paper included in this study, we provided a brief summary of its main contribution and characteristics. We extracted the problem the papers solved, the domain of their study, the type of study, the approach used, the question types covered in their system, the dataset used, and evaluation metrics. Thus, we categorized the studies based on their functionality into three main categories within the topic of Arabic conversational AI systems: (1) question-answering (QA) systems, (2) task-oriented dialogue systems, and (3) chatbots. In Section 3, we introduce QA systems and then describe the approaches and tasks of these systems. In Section 4, we present the task-oriented DS, and we discuss the chatbots in Section 5.

III. QUESTION-ANSWERING SYSTEMS

Question-answering (QA) systems are a well-known problem in the field of NLP. They aim to retrieve or generate precise answers to given questions in a natural language. Due to the vast amount of information on the internet, it is possible for a user to ask a question and get a specific answer easily and naturally. Nevertheless, searching manually for the related answer in search engines consumes time and effort. This motivated researchers to build automatic QA systems that are more flexible and user friendly. Such systems use a large amount of data, as well as NLP, ML, and IR mechanisms, in order to recognize the text inputs or questions and then generate suitable answers.

Research on English QA systems has witnessed significant progress due to the release of vast datasets such as the Stanford QA Dataset (SQuAD) [16] and WikiQA [17]. However, there has been less literature published for Arabic. Although research in the field of Arabic QA began early in the 1990s, there are insufficient tools and resources to advance the progress of Arabic QA systems.

Recently, Arabic QA systems have gained more popularity with the huge amount of Arabic content on the web. Some of the most common QA datasets for Arabic are Arabic-SQuAD, ARCD [18], DAWQUAS [19], Translated TREC, and CLEF [20]. In addition, some challenges have been released, such as SemEval 2017 [21], which provides a subtask for Arabic QA. Despite high-quality research performed on methods in Arabic QA systems, previous studies have limitations such as the small amounts of data for training and testing and reliance on classical methods [18].

Regarding the domain of QA systems, most research is focused on the medical domain due to the large amount of Arabic language medical data available on the internet [22]–[24]. Other works have addressed open-domain QA systems [18], [25], [26].

Generally, QA systems consist of three main modules: question processing, document processing, and answer processing. Figure 1 illustrates the general architecture of QA systems. Regarding question processing, there are two tasks: question classification (QC) and question semantic similarity (QSS). QC assigns a predefined class to the questions, such as definitional, yes/no, factual, abbreviation, or description questions. Definitional questions are those that answer "who" and "what" questions. Yes/no questions are questions that can be answered with two possible responses: "yes" or "no." Factual questions are questions whose answers are one word or a short phrase, such as questions starting with "who," "where," "what," etc. Abbreviation questions ask about shortened forms of a written word or phrase. Description questions are those whose answers contain definitional information about some terms. Finally, QSS detects semantically similar questions in community QA platforms.

Regarding document processing, document and passage retrieval is responsible for retrieving, analyzing, and ranking documents and then passages, which are more likely to hold the required answer. Regarding answer processing issues, answer extraction is concerned with extracting the answer from relevant documents given both the question and the context. The answer-ranking problem does not generate the answer; instead, it ranks a set of candidate answers based on their relevance to the question. Some studies have focused on building a complete QA system [27], [28]. Others have focused on specific tasks such as QC or QSS in an attempt to improve the quality of the QA systems.

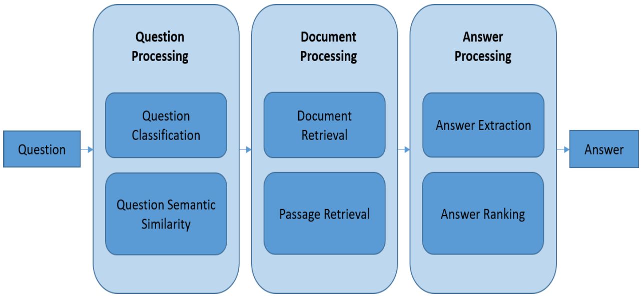

FIGURE 1. General architecture of QA systems.

A. QUESTION-ANSWERING APPROACHES

In this part, we review the recent state of the art in Arabic QA systems. Thus, we classify the studies into three categories according to their approach: semantic and logic-inference systems, rule-based systems, and learning-based systems. We discuss the domain, approach used, dataset, evaluation metrics, and findings of these studies.

1) SEMANTIC AND LOGIC-INFERENCE BASED APPROACH

This approach aims to convert Arabic texts into semantic and logical representations to find the textual implication between the question and the passage that contains its answers [28]. Constructing a semantic and logical representation of texts helps in understanding the concepts from texts and the relations between them. However, few studies have focused on the semantic and logic-inference-based approaches to develop Arabic QA systems [28].

W. Bakari *et al.* [28] proposed a novel logic-based approach for Arabic QA systems based on the automatic understanding of Arabic texts in order to convert them into semantic and logical representations. They used conceptual graphs to obtain a logical representation in order to extract a precise answer based on the relationship between the question and text passage. To evaluate their approach, they used a corpus of question-texts (AQA-WebCorp) [29] to answer the factual questions. As a result, they achieved good performance with 74% accuracy. Therefore, their logic-based approach, which is integrated with recognizing textual entailment (RTE), plays a notable role in enhancing the performance of Arabic QA systems. Given that some expressions express the same meaning, RTE helps in determining whether the given answer entails the asked question or not. In another work that used RTE [30], M. Ben-Sghaier *et al.* proposed an Arabic Semantic Logical Textual Entailment Tool, or Ar-SLoTE. Their tool defined the relation of entailment between factual QA couples by extracting both the logical forms of QA pairs and the valuable features from them. Then, they used a decision tree classifier to predict the entailment relationships between the QA pairs. They used 500 pairs of different factual questions with their candidate answers from the AQA-WebCorp dataset. Thus, their experiments showed a relatively good result, achieving 73.33% accuracy. Furthermore, M. Ben-Sghaier *et al.* proposed a semantic-based method for the Arabic RTE [31]. They focused on the sentence level by conducting both a semantic similarity measure and word sense disambiguation process to detect the entailment relationship in the context of a factual Arabic QA system. They used 200 QA pairs derived from the AQA-WebCorp dataset and achieved an accuracy of 70%. In all the previous studies that used RTE, increasing datasets would help to enhance the effectiveness of the QA system.

I. Lahbari *et al.* [32] proposed a passages retrieval method for Arabic QA systems based on both similarity with Arabic Wikipedia documents and a BM25 retrieval model. They started by transforming the Arabic question into a query, then extracted the relevant documents from Arabic Wikipedia based on both the documents titles and Named Entities (NEs) involved in the formulated query. After that, they extracted the candidate passages that contain the correct answers. To evaluate their approach, they translated the TREC and CLEF datasets into Arabic. Hence, their findings indicated a high similarity level between a given question and the candidate passages.

Due to the vast amount of semantic data available on the web, these data can be used in building QA systems over what is called linked open data (LOD). Semantic data is structured data that includes well-defined relationships representing its meanings [25]. Most of the previous research on QAS over LOD is for English, while few studies are available for Arabic. M. Al-Smadi *et al.* tackled the issue of utilizing semantic data on open-domain Arabic QA systems [25]. They suggested a new hybrid approach in order to map the users' questions in Modern Standard Arabic (MSA) to a standard query language for the LOD. This approach is based on extracting the entities from questions and linking them over the web using named-entity recognition and disambiguation, then extracting properties from the extracted named entities via a dependency parsing approach integrated with Wikidata ontology. To evaluate their system, they used a 400-question dataset of three types of questions: definition, factual, and yes/no questions. Therefore, they obtained good results compared to state-of-the-art techniques, achieving 84%, 81.3%, and 82.8% in terms of precision, recall, and F-score, respectively. However, their system is limited to three question types and can only answer questions that contain just one named entity and one property. In their study, they indicated that the lexical gap between Arabic and the LOD terminology requires more investigation. Indeed, after extensive research on using LOD for Automatic QA systems, there is still a limitation in representing Arabic web resources.

2) RULE-BASED APPROACH

In rule-based approaches to QA, some rules are applied to the given question and the passages that contain the answer to identify accurate answers. As outlined by several studies [22]–[24], the authors used the rule-based approach and developed their QA systems using the NooJ linguistic platform.[1] NooJ is a corpus processing system that constructs, tests, and manages formal descriptions in the form of automated dictionaries and grammars in multiple natural languages [23]. E. Bessaies *et al.* [22] focused on yes/no and factual questions. They used a named entity recognizer (NER) component to identify the answers and questions. They evaluated their system on a dynamic corpus composed of Arabic medical journal articles. Therefore, they achieved an acceptable result for named entities identification, with an F-score of 88%. Consequently, the evaluation of question annotation achieved an F-score of 73%. The approach of this study cannot process some types of questions, such as "why" and "how" questions. In addition, I. Ennasri *et al.* proposed a system based on two steps [23]. The first step is the identification of the question type and keywords in order to determine the minimum number of responses. The second is to apply the suitable transducer set to the corpora to extract responses based on generated bilingual dictionaries that contain different Arabic medical terms and their English translations. The authors conducted some experiments to evaluate their system on collected Arabic and English corpora. They obtained an F-score of 72%, although this result was restricted by the structure of some sentences and to only factual questions. Relatedly, S. Dardour *et al.* suggested a new method to resolve Arabic-related ambiguities in the medical domain to support their QA system quality [24]. They collected 350 questions from frequently asked questions (FAQs), discussion forums, and some translated questions from the TREC dataset. In addition, they collected Arabic medical texts for each question from the internet. Thus, they achieved an improvement in the disambiguation process by 28% in terms of the F-score. The system focused on factual as well as complex questions such as "why" and "how" questions. The major limitation of the previous studies is that they are limited to specific terms, as they employed a rule-based approach and used many dictionaries [22]–[24]. To serve the same goal of building a QA system, M. Mtibaa *et al.* tackled the issue of Arabic temporal QA systems. They developed an Arabic temporal resource aimed to provide a relevant answer to the Arabic questions related to time [33]. Although the relevant studies in other languages use grammars, finite state automata, or neural networks to detect the temporal entities, these techniques do not work well for Arabic. This is due to the rich morphology and ambiguity issues of Arabic temporal information. In their work, they used Wikipedia and the internet as sources to collect a set of Arabic temporal information in multiple categories to build their Arabic temporal resource. This temporal resource was useful to unify all the extracted

[1] http://www.nytud.hu/nooj08/nooj.html

temporal sentences under the same writing format in order to facilitate the extraction of the relevant answers. They also collected an Arabic corpus of 500 temporal questions from the TREC corpus and from a list of questions produced in TERQAS Workshop. In spite of their system's encouraging results, it must be improved to cover more domains.

3) LEARNING-BASED APPROACH

While the rule-based approach is still widely used, an alternative approach involving machine learning from data has come to dominate the current QA systems research. The learning-based approach uses machine learning algorithms to help in building QA systems. Recently, huge advances in language models (LMs) have led to promising possibilities for various Arabic NLP tasks. Furthermore, deep neural networks (DNN) have occupied an essential role in developing high-efficiency NLP applications. These models have proven effective for common NLP tasks such as QA systems. Some studies have employed this approach to develop Arabic QA systems.

I. Lahbari *et al.* [27] presented a machine learning (ML) approach using a support vector machine (SVM) in order to classify the questions in a QA system. They adopted an Arabic taxonomy to classify the questions and then transformed the classified questions into queries. This approach can improve the performance of a search engine by retrieving more relevant documents. In their approach, they used hybrid Arabic part-of-speech (POS) tagging to annotate the query tokens and Arabic WordNet for the query expansion. Then, by using Google engine and Arabic Wikipedia as a dataset, they could extract the relevant documents and answers. Given the lack of Arabic resources, they evaluated their approach on the Arabic translation of the text retrieval conference (TREC) and cross-lingual evaluation forum (CLEF) datasets. Thus, they proved that the SVM classifier is more efficient than the other state-of-the-art approaches, achieving an F-score of 90%. However, this study suffered from some translation issues, such as abbreviations that did not exist in Arabic and the loss of the true meaning of the sentence as a natural language. In addition, the study included only a few studied questions types.

Furthermore, M. Al-Shenak *et al.* improved an Arabic QA system [34]. In their system, they used latent semantic analysis (LSA) to model both documents and terms into the same concept space. Then, they used the SVM algorithm to classify the questions into a corresponding class and retrieve the related documents with the selected paragraphs that hold the answer using single-value decomposition (SVD). They evaluated their approach on a dataset of 10,000 Arabic documents in 10 classes. They randomly selected 20 different queries, with two queries for each class. The experiments showed promising results for all the classes, with approximately 98%, 97%, and 98% in terms of precision, recall, and F-score, respectively. However, these measurements may decrease if the number of datasets used for evaluation increases. In another study, H. Mozannar *et al.* aimed to design an open-domain Arabic QA system [18]. Their system is based on two components: a TF-IDF document retriever model and a BERT document reader model. They used hierarchical TF-IDF to select the documents most related to the question, then extracted answers from these documents with the pre-trained bi-directional transformer BERT model. The authors contributed to enriching the Arabic research community by building a corpus for the Arabic QA system; they presented the Arabic Reading Comprehension Dataset (ARCD) composed of 1,395 factual questions extracted from Wikipedia. In addition, they translated about 48,344 questions from the SQuAD dataset into Arabic. Their experiments on the ARCD dataset obtained good results compared to the state of the art; their BERT-based reader and open-domain system achieved F-scores of 61.3% and 27.6%, respectively. Furthermore, one of the most significant findings of this study is that translated data can be effective as a training resource for QA systems. However, increasing the size of ARCD may enhance the end-to-end QA system. In their work, they indicated the possibility of improving the system by using paragraph selection to get the correct answer.

Using a DL approach, Y. El Adlouni *et al.* revealed the effectiveness of applying pairwise approaches combined with DL methods on Arabic community QA systems in the medical domain [35]. They explored both supervised models and unsupervised approaches based on topic modeling. In their approach, they chose a set of potential candidates and then ranked them based on their relevance scores. They evaluated their model on the SemEval 2017 Task 3-D benchmark dataset using three metrics: average precision (mAP) to measure the ranking performance of the model, accuracy, and precision. Thus, their experiments achieved high values using the latent semantic indexing approach (LSA), which outperformed the state-of-the-art approaches. Furthermore, they proved that pairwise approaches outperformed pointwise methods in the ranking process. Pointwise approaches compare each original question to a question–answer pair in order to measure their similarity, while pairwise approaches take a set of pairs of questions in order to find the relationships that may exist between them. Also using DL techniques, A. Almiman *et al.* aimed to address the problem of answer-ranking in Arabic [36]. They worked on a set of candidate question–answer pairs for each unanswered question. Thus, they used both lexical and semantic features in addition to LM-based features derived by the BERT model to measure the similarity between both unanswered and candidate question–answer pairs. They developed a DNN ensemble model to combine these features, measure the relevance probability, and achieve better prediction results. They evaluated their model on the SemEval-2017 dataset using three metrics: mAP; average recall, to measure the accuracy of the model; and mean reciprocal rank, a ranking evaluation method. Their proposed model outperformed the existing ranking models for Arabic QA for the metrics they used. Based on their findings, the system might be improved by using the collective relevance of these similarities instead of using the individual ones as the authors suggested.

In a proactive step towards using transformer models in Arabic conversational AI, W. Antoun *et al.* pre-trained the BERT model for Arabic, calling the resulting model AraBERT [26]. They aimed to achieve a similar success as BERT did for English. They achieved significant performance on some Arabic NLP tasks, including QA. Consequently, they evaluated their model on ARCD and Arabic-SQuAD datasets using the F-score, exact match score, and sentence match score metrics. They achieved enhancement in the performance on QA tasks with an increase in the sentence match score of 2% compared to multilingual BERT. However, AraBERT needs more improvement in QA tasks to cover Arabic dialects.

To conclude, Table 2 presents a comparison of the existing Arabic QA systems. They are compared with respect to the domain, approach, question type coverage, dataset used, and evaluation metrics.

B. QUESTION-ANSWERING TASKS

In this section, we discuss two issues related to the QA systems studied by the research community: question classification (QC) and question semantic similarity (QSS). These tasks are useful in various NLP applications, such as QA systems, chatbots, and conversational agents. While these issues may appear similar, there is a slight difference between them. In QC, the task is mainly to assign labels to a given question in order to get the correct type of answer based on its class. In QSS, the task is to identify duplicate questions in community QA platforms in order to minimize repeated questions.

1) QUESTION CLASSIFICATION (QC)

This task is concerned with assigning a label to a question. It improves the performance of QA systems by eliminating trivial answer candidates. A few studies have focused on the QC task as part of Arabic-language QA systems [37]. Due to the importance of question taxonomies in the QC task, a wide range of classes are proposed for questions, most of which are not designed for Arabic questions [38]. Indeed, many studies have focused on the "wh-" question types and have designed approaches for limited domains. A few works in the literature have attempted to address this task for the Arabic language and have achieved satisfactory results [37]–[41].

In [37], A. Hamza *et al.* used the embeddings from LMs (ELMo) for the Arabic QC task. To achieve their goal, they suggested building four neural networks to study the behavior of contextual representation. They performed multiclass classification, using seven classes to classify the questions. The authors evaluated their classifiers with two word representations: enriched word2vec and ELMo. Furthermore, they used a large dataset of 3,173 Arabic questions manually collected and annotated on both Arabic taxonomy [39] and an updated Li and Roth taxonomy [42]. Their findings indicated that contextual representation achieves notable results. They obtained 94% for accuracy, macro F-score, and weighted F-score on Arabic taxonomy, and 92%, 83%, and 92% for accuracy, macro F-score, and weighted F-score, respectively, for the updated Li and Roth taxonomy. Compared to the previous studies of Arabic QC, they used a large dataset. Thus, it is worth building an end-to-end Arabic QA system including their model.

Hamza *et al.* proposed an Arabic QC approach based on both ML and words continuous distributed representation [39]. In addition, they proposed a new taxonomy for open-domain Arabic questions extracted from Arabic linguistic rules. They represented the questions with continuous distributed word representation (TF-IDF weighting technique) in order to find the syntactic and semantic relations between words as features used in the classification model. Then, they adopted the SVM classifier to obtain seven types of questions. They evaluated their system using a manually extracted corpus from three datasets: TREC, CLEF, and Moroccan schoolbooks. Their experiments showed comparable results with state-of-the-art techniques by applying the SVM classifier with their suggested Arabic taxonomy, achieving 90% accuracy. As a result, their suggested taxonomy is more appropriate for the Arabic question classification task. Their work suggests that future research should concentrate on building an efficient Arabic QA system using both word embeddings and ML methods. From a different point of view, another study showed that obtaining finer classification will be more efficient using a pre-trained word embedding [40]. In this study, a novel approach was proposed by combining an SVM and a convolutional neural network (CNN) to serve the Arabic QC task using word vectors. The authors used an SVM in order to find the question class. Then, for each class, the CNN model aimed to predict a subclass for the main class. They evaluated their proposed approach on the TALAA-AFAQ corpus [43], achieving promising results for the Arabic QC task, including 82% accuracy for all architecture (including the main and finer classes). Their approach should be applied to an end-to-end QA system to explore its implications and results.

To overcome the limits of the old approaches, which are based on rules in determining the type of question, A. M. Hasan *et al.* improved the QC methodology by leveraging both pattern-based matching and ML [38]. They used a combination of an SVM classifier model and pattern-based approach in the domain of the Arabic Islamic Hadith. They evaluated their approach on a small dataset limited to 200 questions on the Arabic Islamic Hadith derived from Sahih Al-Bukhari. Their experiments demonstrated the efficiency of their proposed approach to QC in Arabic; they obtained 88.39%, 87.66%, and 87.93% in terms of precision, recall, and F-score, respectively. However, their approach is limited to just three types of questions: "who", "where," and "what."

M. Abdel-Latif *et al.* tackled the Arabic CQ task using both lexical and semantic features in order to improve answer retrieval [41]. Their goal was to rank the 30 retrieved question–answer pairs given a new unanswered question based on their relevance to the new one. They explored

TABLE 2. A comparison of recent arabic QA systems based on their approaches (semantic and logic-inference based, rule-based, and learning-based).

Paper	Domain	Approach	Dataset	Question types	Metrics
[28]	Open domain	Semantic and logic-inference based	AQA-WebCorp.	Factual questions	Accuracy= 74%
[30]	Open domain	Semantic and logic-inference based	500 QA pairs extracted from the AQA-WebCorp dataset.	Factual questions	Accuracy=73.33%
[31]	Open domain	Semantic and logic-inference based	200 QA pairs extracted from the AQA-WebCorp dataset.	Factual questions	Accuracy= 70%
[25]	Open domain	semantic and logic-inference based	DBpedia and Wikidata as data sources + 400 questions.	Factual, definitional, and yes/no questions	Recall= 84% Precision= 81.3% F-score= 82.8%
[22]	Medical	Rule-based	Dynamic corpus composed of Arabic medical journal articles.	Factual and yes/no questions	Recall= 72% Precision= 75% F-score= 73%
[23]	Medical	Rule-based	Collected bilingual corpora of Arabic and English.	Factual questions	Recall= 90% Precision= 80% F-score= 72%
[24]	Medical	Rule-based	350 questions collected from FAQs, discussion forums and some translated questions from TREC.	Factual and complex medical-related questions	Recall= 87% Precision= 93% F-score= 89%
[33]	Open domain	Rule-based	500 temporal questions collected from both TREC and TERQAS.	Temporal questions.	Recall= 76% Precision= 70% F-score= 72%
[27]	Open domain	Learning-based (SVM)	Arabic translation of TREC and CLEF datasets.	Factual, definition, abbreviation, description questions	Recall= 89% Precision= 93% F-score= 90%
[34]	Open domain	Learning-based(LSA)	Arabic documents dataset of size 10,000 with 10 classes.	Factual questions	Recall= 97% Precision= 98% F-score= 98%
[18]	Open domain	Learning-based(BERT)	Reading comprehension dataset (ARCD) composed of 1,395 questions + a Arabic-SquAD.	Factual questions	Exact match= 12.8% Sentence match= 29.8% F-score= 27.6%
[35]	Medical	Learning-based(DL)	Semeval 2017 Task 3-D dataset.	All question types	Mean average precision = 61.66% Accuracy = 62.34%
[36]	Medical	Learning-based(DL)	Semeval 2017 Task 3-D dataset.	All question types	Mean average precision = 62.90% Average recall = 86.6% Mean reciprocal rank = 68.86%
[26]	Open domain	Learning-based(AraBERT)	ARCD and Arabic-SQuAD datasets.	Factual questions	AraBERTv0.1/ v1: Exact match = 51.1/ 54.8% F-score = 82.1/82.2% Sentence match = 95.5/95.6%

various sets of features, including the lexical, semantic, and word-embedding features. In addition, they explored multiple ML algorithms, including linear SVM, logistic regression, and random forest, as well as a fully connected DNN approach. The linear SVM achieved the best result compared to the others when using all feature sets; they outperformed the baselines by 62.85% in terms of mAP, with an improvement of 2.3% on the SEMEval2017 dataset.

2) QUESTION SEMANTIC SIMILARITY

One of the most popular applications for QSS is identifying duplicate questions in community QA platforms such as Quora. QSS aims to retrieve all questions that are similar to a given new question posted in the forum. Thus, QSS reduces the number of duplicate questions and saves effort and time. In the case of existing similar questions, users can get an answer to their question quickly. As users may formulate their questions using different words, retrieving the questions will face some difficulties due to word mismatches between questions. In Arabic, it is necessary to take the semantic information into consideration when evaluating the meaning. The Arabic language poses several challenges; it is a low-resource language with various dialects that is rich in morphology, as explained later. Besides these challenges, detecting semantically similar questions is more difficult in Arabic.

Due to the important role of NLP in community QA (cQA), the research community has organized many challenges to encourage research in this field. In the SemEval2017 Task 3 challenge, Task D targeted re-ranking the correct answers for a new question for Arabic [21]. For a given new question, the task was to retrieve all similar questions previously posted in the forum. In the case of existing similar questions, users could get an answer to their question quickly. In this challenge, they provided researchers with a medical dataset to evaluate their work.

A few works in the literature have attempted to address this task for the Arabic language, and acceptable results have been achieved compared to other languages [44]–[48]. Due to the lack of relevant Arabic semantic corpora, rule-based approaches are currently used to achieve this task.

In the work of M. Daoud, a hybrid approach of both rule-based and supervised learning approaches was used to automatically detect the textual and semantic similarity between the questions according to their scope and type [44]. In his approach, the author used 600 Arabic questions selected from the FAQ pages of different United Nations organizations and the website Ejaaba.com.[2] He examined several classifiers via WEKA 3.8, and he obtained the best results using a random forest algorithm with 10-fold cross-validation, achieving 85% precision. Generally, he obtained comparable findings to the state-of-the-art approaches in English. He suggested using a multi-domain Arabic lexicon in order to improve his work.

S. Romeo *et al.* also addressed the QSS issue using an SVM-based model [45]. They used advanced text representations and built a UIMA[3] based processing pipeline that involves a tree-kernel (TK) based ranker. Thus, they extracted useful features, including the lexical and syntactic information of the text, for processing by their model. They evaluated their work on the SemEval[4] 2016 Task 3-D, which consists of an Arabic corpus in the medical domain. Their experiments revealed good results in terms of accuracy, proving that the extracted syntactic information is very useful for such a task.

In their study [46], N. Othman *et al.* suggested a word-embedding-based method to address the QSS task. They calculated the similarity between a given question and previous ones using cosine similarity. Then, they used K-means to cluster the questions into groups. To evaluate their approach, they used a real dataset from Yahoo Answers in English and another translated into Arabic for different topics and categories. They found that the effectiveness of the proposed method for Arabic was less than that of English in terms of mAP and precision at n (P@n). P@n measures the proportion of the top-n retrieved relevant questions. They achieved 46% and 38% in terms of recall for English and Arabic, respectively. The result was worse for Arabic because the suggested approach ignored the language's morphological structure. Unfortunately, their study also neglected lexical ambiguity when representing the words; if these points are addressed, the proposed approach may achieve better performance.

Al-Bataineh *et al.* [47] proposed a new end-to-end prediction model to address QSS. They overcame the morphology and out-of-vocabulary (OOV) issues of Arabic using character-level CNN. Furthermore, they built a Modern Standard Arabic (MSA) Q2Q dataset that is considered one of the largest Arabic Q2Q datasets. They evaluated their approach on their datasets in addition to the generated Dialectic Q2Q dataset using the MADAR parallel corpora [49] for 24 Arabic dialects. Their system outperformed the state-of-the-art approaches, including AraVec, by achieving 93% and 82% in terms of F-score on the MSA and the dialects' benchmarks, respectively.

Another study proposed leveraging the power of DNN to measure the similarity between Arabic question pairs [48]. O. Einea *et al.* used a vector similarity function layer, which is tested on Arabic sentences for the first time in this study, in order to measure sentence similarity. In addition, they explored two famous DDN models: recurrent neural networks (RNN) and CNN. They evaluated their method on the NSURL 2019 Q2Q dataset and the SemEval 2017 QA dataset. Generally, CNN outperformed the state-of-the-art approaches by more than 10% in terms of prediction accuracy; the approach obtained 77% and 58% on both datasets, respectively. Based on their experiments, they proved that the default generated word embeddings achieved better prediction results than those based on pre-trained models. This is due to the inclusion of Arabic semantics as well as the sufficient datasets used in generating a good embedding model. The authors also addressed the topic of imbalanced datasets, suggesting an effective method using three experiments that are built on each other to enhance prediction accuracy. A key strength of this study is that it requires minimal pre-processing to achieve better results. Although the model works well based on medium-sized datasets, the findings will be even better with larger datasets in the future.

To conclude this discussion of Arabic QA systems, Table 3 presents a comparison of the existing systems. They are compared with respect to the domain, type of study, approach, question type coverage, dataset used, and evaluation metrics.

To the best of our knowledge, even with the existing Arabic works, the Arabic QA research field is still limited in terms of question type coverage [22], [38], [44], domain [35], [41], [45], and system performance. Indeed, most Arabic works have resolved the issue of answering factual questions, while few studies have resolved complicated questions. In addition, many works have handled the medical field in Arabic QA.

Furthermore, the recent developments in DL in the NLP area for high-resource languages like English have shown that Arabic is lagging behind. Until now, it has been difficult

[2] https://ejaaba.com/

[3] UIMA is a framework helps to integrate the systems in order to analyze unstructured information whose aim is getting new knowledge relevant to particular context of application.

[4] https://alt.qcri.org/semeval2016/task3/index.php?id=data-and-tools

TABLE 3. A comparison between the recent existing arabic QA systems based on their tasks (QC or QSS). N/A indicates information is not available.

Paper	Domain	Type of study	Approach	Dataset	Question types	Metrics
[37]	Open domain	QC	Learning-based(CNN,RNN)	A dataset of 3,173 Arabic questions collected manually.	All question types	Accuracy= 94% F-score= 94% Weighted F-score= 94%
[39]	Open domain	QC	Learning-based (SVM)	1,302 questions from TREC, CLEF and Moroccan schoolbooks.	All question types	Accuracy= 90% Recall= 90% Precision= 91% F-score= 90%
[40]	Open domain	QC	Learning-based (SVM , CNN)	TALAA-AFAQ corpus.	All question types	Accuracy= 82%
[38]	Islamic (Hadith) domain	QC	Learning-based (SVM) + pattern-based approach	200 questions on Arabic Islamic Hadith derived from Sahih Al-Bukhari.	Three types of questions: “ who”, “where” and “what”	Recall= 87.66% Precision= 88.39% F-score= 87.93%
[41]	Medical domain	QC	Learning-based (SVM, decision trees, and feed-forward networks)	Semeval 2017 Task 3-D dataset.	All question types	Mean average precision = 62.85 %
[44]	Open domain	QSS	Hybrid of learning-based (Random Forests) and rule-based	600 questions from the FAQ pages of various United Nations organizations and ejaaba.com.	All question types	Precision = 84% Recall = 85% F-score= 84%
[45]	Medical	QSS	Learning-based(SVM, tree kernels)	SemEval 2016 Task 3-D.	All question types	NA
[46]	Open domain	QSS	Learning-based(K-means)	Yahoo Answers in Arabic and English.	All question types	Recall for English= 46% Recall for Arabic= 38%.
[47]	Open domain	QSS	Learning-based(deep contextualized word embeddings)	A dataset collected from Tweets, Arabic Wikipedia, and Mawdoo3 articles + the MADAR parallel corpora for 24 Arabic dialects.	Factual questions	F-score on the MSA= 93% F-score on the dialects = 82%
[48]	Open domain	QSS	Learning-based(CNN,RNN)	NSURL 2019 Q2Q dataset + the SemEval 2017 QA dataset.	All question types	Accuracy= 77% and 58% for both datasets, respectively

to apply these new models to Arabic due to the language's unique challenges. However, it is important to find a suitable technique to leverage the power of these new models.

IV. TASK-ORIENTED DIALOGUE SYSTEMS

Recently, task-oriented dialogue systems (DS) are receiving more interest in both the research community and industry. Task-oriented DS are systems built for specific tasks [1]. Such systems have usually been built on top of a domain-structured ontology called the domain knowledge base (KB)[4].

There are several important terms that are frequently used in task-oriented DS: domain, dialogue history, intentions, slots, slot values, dialogue states, and the external KB. Domain represents the current dialogue topics. For instance, the taxi domain is about booking a taxi, and the hotel domain is about hotel reservations. Dialogue history means the context of the whole dialogue in the current conversation (not the whole conversational history between users and the DS). Intentions are the goals behind the user utterance. For instance, the intention of saying “Is it going to be cold tomorrow?” is to check the weather state, and the intention of saying “Book a flight to Riyadh at 8 on Thursday” is to book a flight. Slots represent the predefined variables in the dialogue and can be filled with any slot value. For instance, a location slot can have values such as London and Riyadh. Dialogue states are semantic decoding results such as slot–value pairs. Finally, the external KB is a dynamic and large store of information provided to the DS.

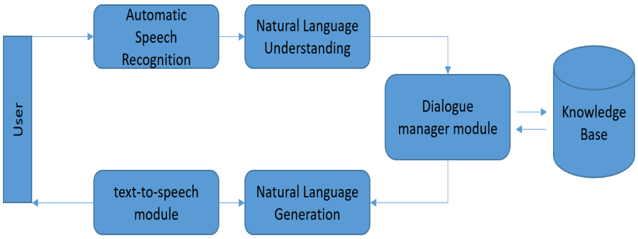
Automatic Speech Recognition
Natural Language Understanding
User
Dialogue manager module
Knowledge Base
text-to-speech module
Natural Language Generation

FIGURE 2. The traditional task-oriented dialogue systems architecture.

The architecture of task-oriented DS is shown in Figure 2, which can be applied to both text-based and speech-based DS. Text-based DS do not have the automatic speech component, which is responsible for converting speech to text in the case of speech-based DS.

Briefly, a natural language understanding module (NLU) analyzes the user's input to determine its meaning and performs several tasks including domain classification, intention detection, and slot-filling. The output of the NLU module is a dialogue act, which represents the function of the sentence or user's utterance, such as a query, question, command, promise, etc. The dialogue manager module is responsible for interacting with the information from KB as an input. It also generates the dialogue actions for the next utterance, which represents the output to the user. In order to obtain the natural language response, the dialogue action will be sent to the natural language generation (NLG) module. Finally, the text-to-speech (TTS) module transforms the text into speech and then answers the user [50].

Two additional important terms occur frequently in task-oriented DS: intent classifier and entity extractor. The intent classifier is responsible for classifying the users' intent in order to direct the DS to the appropriate answer. The entity extractor extracts the main tags from the user commands. It works by giving each word in the sentence a label that identifies its role.

Unlike traditional DS, which uses a pipeline that connects each module, end-to-end neural systems are designed to train directly on both KB information and text transcripts. Recently, due to the success of DS, researchers have started to explore end-to-end solutions for task-oriented DS. However, end-to-end task-oriented DS is still in its infancy, and several issues still need to be addressed [50]. Mainly, traditional DS requires large-scale labeled data to train each dialogue component, which makes the system more stable and interpretable, like most existing commercial systems. While end-to-end systems require less annotation and are thus easier to build, these systems are less controllable.

For Arabic, some task-oriented DS have been built to serve specific domains, such as home automation [51], the medical domain [52], and flight booking [53]. Some are designed to serve specific Arabic dialects, such as OlloBot [52].

To build an intelligent and interactive conversational DS, some platforms have been designed to serve these systems. The Pandorabots[5] platform and the IBM Watson[6] Conversation API are platforms that support Arabic and are easily accessible by multiple users at the same time. The reason for the rarity of such platforms is the complexity and resource scarcity of the Arabic language [52].

In their study [51], A. M. Bashir *et al.* used a DL approach to build an NLU module for an Arabic task-oriented DS for home automation. Their module consists of intent classifier and entity extractor components. The intent classifier was implemented using LSTMs and CNNs, while the entity extractor was implemented using a BiLSTM along with character-based word embeddings. They collected data related to home automation via an online survey and then filtered and labeled the data according to the Conll-2003 NER format. The authors used an NER corpus from the AQMAR dataset, and they employed AraVec [54], a pre-trained Arabic word embedding model, to represent the text data. In their system, they used five intents (greeting, inform, check status, chat, and goodbye) and five entities to be tagged (room, device, actions applied to devices, device speeds, and person names). Their experiments indicated that CNN performance was better than the LSTM, achieving an F-score of 94% for intent classification. Furthermore, the BiLSTM with the char embeddings model achieved comparable results to the named entity recognition benchmarks in English, with an F-score of 94%. As they suggested, it would be interesting to assess the improvements in task-oriented DS by integrating the suggested NLU module with automatic speech recognition (ASR) and NLG modules.

In the medical domain, DS can provide 24/7 support for end users (i.e., patients, physicians, etc.). Such systems can answer simple and repetitive questions using pre-designed answers. One example is OlloBot, a rule-based DS which offers a health tracking system [52]. It supports patients and physicians with care delivery via conversation as a health assistant. A. Fadhil, *et al.* developed their system via the IBM Watson Conversation API to build the dialogue structure and the Telegram Bot Platform[7] to build the chatbot. The conversation slots take the user's input and check its intent, entities, and condition of the conversion to return relevant answers. To test the effectiveness of OlloBot, they conducted a user experiment with 43 Arabic-speaking users using a questionnaire to measure the system's usability. The survey measured ease of use, ease of learning, usefulness, and satisfaction with the application. This contributes additional evidence that DS offers patients an excellent user experience via chatting with the bot in order to get relevant answers to their queries.

Earlier, A. H. Al-Ajmi *et al.* [53] proposed a text-based DS that is capable of handling customers' utterances in flight booking. They used a hybrid rule-based and data-driven approach in building their system. The authors used the Wit.ai framework as the natural language interface for the NLU component, the Telegram Messenger framework as the DS interface for the system's dialogue manager, and Python for the NLG component. They used the Wizard of Oz technique to gather preliminary knowledge about booking scenarios from 28 volunteers. Furthermore, they used crowdsourcing to collect 1,651 training examples, which varied between positive and negative examples to increase the data-driven entity's ability to identify the intended results precisely. They evaluated their system using questionnaires and individual testing to demonstrate the system's effectiveness, ease of use, and ability to self-feed. Despite the small size of the training data, they attempted to use self-feeding to reduce this limitation's effects on DS performance.

Another recent work facilitated the search for Islamic-related information in an interactive way [55]. In this study, F. Bendjamaa *et al.* proposed an Arabic DS based on a pre-existing Quranic ontology. The Quranic ontology is a conceptualization of the Qur'an that contains its chapters and verses as well as each word of the Qur'an with its root and lemma. The system can access Quranic information easily

[5] https://home.pandorabots.com/home.html

[6] https://www.ibm.com/cloud/

[7] https://telegram.org/blog/bot-revolution

via SPARQL queries. However, this system is closer to a QA system than a DS.

Recently, a few works tackled the issue of automatic dialogue act recognition due to the importance of this component for semantic extraction in NLU and DS. Automatic dialogue act recognition has gained some attention in the Arabic research community [56]–[58]. One study [56] classified Arabic dialogue act recognition techniques into rule-based and ML-based techniques based on three criteria: surface features, cue words, and contextual information. In this study, L. Sherkawi *et al.* compared the experimental results for both techniques on a set of 1,500 sentences written in MSA. They found that rule-based systems achieved 98.92% accuracy, while ML-based systems achieved an accuracy of 97.09%, 96.48%, 93.50%, and 93.70% for decision tree, naïve Bayes, neural network, and SVM, respectively.

Joukhadar *et al.* [57] aimed to recognize users' dialogue acts in a text-based DS using several ML approaches with several features and feature selection methods for the Levantine Arabic dialect. They evaluated their work on a set of 873 sentences manually created in two domains: restaurant orders and airplane ticketing. The authors divided the dialogue acts into eight types: Greeting, Goodbye, Confirm, Thanks, Negate, Ask_for_alt, Ask_repeat, and Apology. They achieved the best result using the SVM model, with an accuracy of 86%. In another study, A. Elmadany *et al.* addressed the recognition of automatic dialogue acts in Arabic dialects using a multi-class hierarchical model [58]. The authors evaluated their system's performance using an SVM classifier on a manually collected and annotated dialogue dataset from multi-genre Egyptian call centers. The proposed approach achieved 91.2% in terms of average F-measure scores, which improved the F-measure by nearly 20%. However, larger datasets that consist of the whole context of the sentences and cover different domains are required to build more efficient dialogue systems for predicting dialogue acts.

V. CHATBOTS

Chatbots, one of the most interesting AI applications that uses NLP, imitate conversational skills and human behavior. Such systems have their own knowledge that helps them identify and understand queries or sentences and then generate appropriate responses in an open domain. The term "chatbot" is very popular these days due to its success stories, such as Apple's Siri, Amazon's Alexa, and Google Assistant. Such systems promise to automate customer services and reduce human effort.

Compared to the significant work on chatbots that has been done for English, few chatbots have been developed in Arabic. Whereas 89% of the available chatbots use English as a communication language, only 8% use Arabic [59]. Furthermore, there is little published research compared to the available commercial applications [14]. Thus, in this section, we review the recent chatbot systems in the research community for the Arabic language.

In a recent work, T. Alshareef *et al.* aimed to design an open-domain Arabic chatbot in the Arabic Gulf dialect [60]. The proposed system is based on generative models, using the LSTM-based Sequence-to-Sequence (Seq2Seq) neural network with pre-trained word embeddings. The authors used FastText to capture the semantics of the sentences in order to generate a relevant and correct response. They built their dataset by collecting 5.2k pairs of post replies from Twitter. The model was trained twice, with and without the pre-trained embeddings. They found that the responses of the second model, which included the embeddings, were more related to the input than those of the model without the embeddings. These findings suggest that pre-trained word embedding improved text generation using a small-sized dataset. The authors evaluated their system using the BLEU score [61] and human evaluators; they achieved comparable results with the existing DL models in other languages. Their findings represent a promising step towards building an open-domain conversational system in the Arabic Gulf dialect.

Current research on chatbots focuses on building emotionally intelligent conversational systems, which are called empathetic conversational systems. Such systems are highly desirable because they could increase user satisfaction in several conversational applications [62]. Despite the various works presented on empathetic conversation systems in English [62], [64], a recent study [65] is the first work to investigate building such models for the Arabic language. This study proposed an encoder-decoder model based on a Seq2Seq model with LSTM units combined with Attention. Furthermore, T. Naous *et al.* proposed the first Arabic empathetic conversations dataset (ArabicEmpatheticDialogues) by translating the English EmpatheticDialogues datasets [66]; they obtained 96% accuracy on a sample of the data. The authors evaluated their suggested model using the automated metrics of perplexity (PPL) and BLEU score in addition to the human evaluation metrics of average empathy score and average fluency score. Their proposed model with an embedding dimension of 500 achieved state-of-the-art performance for Arabic, with values of 38.6 and 0.5 for PPL and BLEU, respectively. According to the human evaluation, the proposed model successfully generated responses with an average empathy score of 3.7 and an average fluency score of 3.92. Although the study successfully demonstrated the success of the model in providing empathetic behaviors with emotional responses for user input in Arabic, it has a limitation regarding the size of the dataset. Later, T. Naous *et al.* [67] proposed a transformer-based model initialized using the AraBERT [26] pre-trained weights. By using the ArabicEmpatheticDialogues dataset [65], the authors fine-tuned their proposed empathetic BERT2BERT model to generate Arabic empathetic responses. The findings of this work proved its performance in empathetic response generation compared to the model proposed by T. Naous *et al.* [65]. However, the proposed empathetic BERT2BERT model is limited by its ability to handle neutral chit-chat conversations.

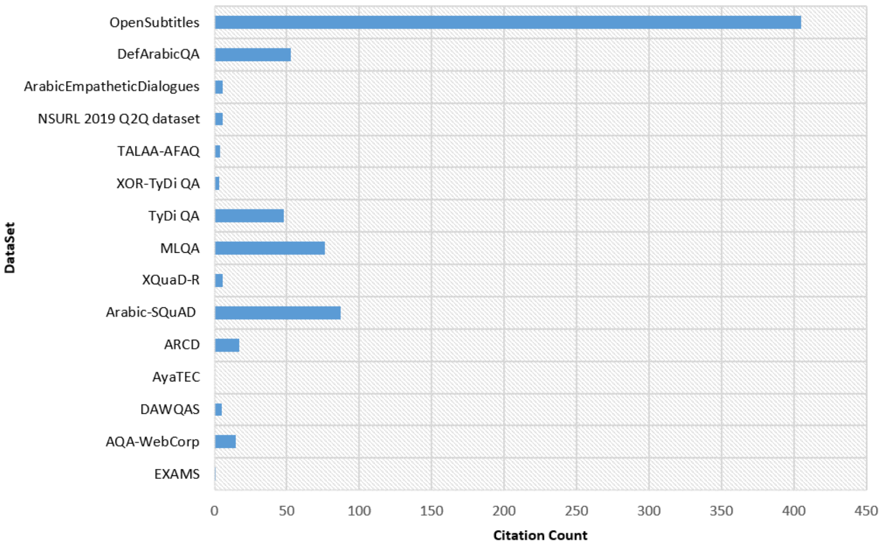

FIGURE 3. Citation count of arabic conversational AI systems datasets in google scholar.

Some chatbots have been designed for specific purposes. For education, Jooka is a bilingual chatbot aimed to improve the university admission process in both Arabic and English [68]. Nabiha is a social chatbot developed to serve the Saudi Arabic dialect [69]. It is intended to serve as an academic counselor for King Saud university students, interacting with students and answering their inquiries on course offerings or any other question related to their academic progress. Nabiha was developed on the Pandorabots platform using a pattern matching approach and the Artificial Intelligence Markup Language (AIML). Labeeb is another chatbot that responds to students' queries related to learning or academic rules [70]. This chatbot uses the Wikipedia API in order to retrieve the first paragraph for the needed queries as XML documents. However, the authors did not clarify the used NLP models. LANA-I is an Arabic chatbot intended to educate children with autism spectrum disorder (ASD) [71]. The system is rule-based and uses pattern matching and text similarity. For Islamic knowledge, SeerahBot is an Arabic chatbot concerned with the Prophet Mohammed's biography [72]. It is based on a retrieval-based approach using 200 questions and answers on the Prophet's biography. SeerahBot used ML to find the QSS then maximize the similarity between both inputs and intents. However, the approach performance is not included in the study.

Most previous Arabic chatbots are rule-based, which encourages using the DL approach to build chatbots. M. Boussakssou *et al.* [73] presented MidoBot, an Arabic chatbot that used a Seq2Seq model to generate new responses from a dataset. MidoBot was built on a manually created dataset from different sources, which helps obtain significant results.

In summary, few recently published works have proposed chatbots, and almost all of them are based on rule-based approaches. Indeed, there is a shortage in published research on Arabic chatbots compared to the available commercial applications. Regarding the commercial chatbots, a few Arabic online chatbots are used in the business domain, such as arabot,[8] labiba,[9] and widebot.[10] Arabot is an intelligent Arabic chatbot built on a state-of-the-art Arabic NLP engine, which has the ability to understand and analyze Arabic conversation accurately and efficiently. Labiba is an intelligent agent that serves customers using their natural language. Finally, widebot can produce human-like conversations in many Arabic dialects. All of these chatbots can interact with customers in a human-like manner in order to provide better marketing and customer experience. While rule-based chatbots are still widely used, specifically in the commercially deployed chatbots, an alternative approach involving machine learning has come to dominate the current Arabic chatbot research [26], [65].

VI. ARABIC CONVERSATIONAL AI SYSTEMS DATASETS

Data collection is an essential step in building efficient conversational AI systems. In this section, we introduce the published Arabic datasets for conversational AI. We focus on the frequently used or recently proposed text datasets, as illustrated in Table 4. Figure 3 shows that there is a variance in the number of citations of the datasets; the datasets designed for multilingual systems have the highest citation counts.

[8] https://arabot.io/
[9] https://www.labiba.ai/
[10] https://widebot.net/
[11] https://github.com/mhardalov/exams-qa
[12] https://github.com/husseinmozannar/SOQAL
[13] https://github.com/deepmind/xquad
[14] https://github.com/google-research-datasets/lareqa
[15] https://github.com/facebookresearch/MLQA
[16] https://github.com/apple/ml-mkqa
[17] https://github.com/google-research-datasets/tydiqa
[18] https://github.com/AkariAsai/XORQA
[19] http://opus.nlpl.eu/OpenSubtitles2018.php

TABLE 4. Common and recent arabic conversational AI datasets. N/A indicates information is not available.

Dataset	Size (*number of samples in the dataset*)	Language	Year	Description	Format	Task
EXAMS[11]	~24K	Multilingual including Arabic	2020	QA for high school examinations in 16 languages, covering 562 questions for Arabic and five subjects.	JSON	QA
AQA-WebCorp [29]	250	Arabic	2016	Question pairs that can be asked in different fields. 25 questions translated from TREC, 25 questions translated from CLEF, 100 questions gathered from forums, and 100 questions from FAQs.	TXT	QA
DAWQAS [19]	3205	Arabic	2018	"Why" QA pairs scraped from public Arabic websites covering eight domains.	CSV, JSON	QA
AyaTEC [74]	207	Arabic	2020	A reusable test collection for verse-based question-answering on the Holy Qur'an. 207 questions with their corresponding 1,762 answers covering 11 different topic categories.	XML, TXT	QA
Arabic Reading Comprehension Dataset (ARCD)[12]	~50K	Arabic	2019	Questions on Wikipedia articles.	JSON	QA, reading comprehension
Arabic-SQuAD (XQuAD)[13]	1190	Multilingual including Arabic	2019	Translation of the SQuAD dataset v1.1 containing ~1.19K Arabic QA pairs.	JSON	QA, reading comprehension
XQuaD-R[14]	N/A	Multilingual including Arabic	2020	Retrieval version of the normal XQuAD dataset. Each question has 11 different languages and 11 parallel correct answers across the languages; contains ~1.19K Arabic QA pairs.	JSON	QA
Multilingual Question Answering(MLQA)[15]	46444	Multilingual including Arabic	2019	Dataset for evaluating the cross-lingual QA performance with ~5K questions in Arabic.	JSON	QA, reading comprehension
Multilingual Knowledge Questions & Answers (MKQA)[16]	260k	Multilingual including Arabic	2020	An open-domain QA dataset aligned across 26 various languages. 10K questions in Arabic.	JSON	QA
TyDi QA[17]	204K	Multilingual including Arabic	2020	~26K Arabic QA pairs.	JSON	QA, reading comprehension
ArabicEmpatheticDialogues [65]	~35K	Arabic	2020	Empathetic conversations in Arabic created by translating the English-language EmpatheticDialogues dataset [66].	CSV	Dialogue
XOR-TyDi QA[18]	40K	Multilingual including Arabic	2020	Open-retrieval multilingual QA dataset that supports cross-lingual answer retrieval; consists of questions in seven languages and answer annotations retrieved from multilingual document collections. ~160K Arabic QA pairs.	JSON	QA
TALAA-AFAQ corpus [43]	~2K	Arabic	2017	A corpus of Arabic factual QA divided into four main classes and 34 finer categories.	NA	QA
SemEvalCQA	N/A	English, Arabic	2016	Dataset for community QA.	XML	QA, reading comprehension
NSURL 2019 Q2Q dataset [75]	12k	Arabic	2019	A balanced number of factual and non-factual questions	TSV	QA
DefArabicQA [76]	50	Arabic	2010	A list of definition questions, a set of 50 files containing snippets collected from Wikipedia, and a set of 50 files containing snippets from the Google engine.	TXT	QA
OpenSubtitles[19]	N/A	Multilingual including Arabic	2016	Multilingual dialogs from movie scripts in 62 languages. ~94K files in Arabic.	XML, XCES	Dialogue

The figure also highlights one of the serious challenges facing research in the field of Arabic conversational AI systems: the absence of sizable Arabic corpora. To the best of our knowledge, no Arabic corpus that contains Arabic dialogue conversations exists.

VII. OPEN CHALLENGES AND ISSUES

Despite the recent advancement of different types of conversational AI systems in English through NLU and knowledge gathering [4], [7], [26], the Arabic language lags behind this progress. Particularly, Arabic conversational AI systems suffer their own issues due to the syntactical richness of Arabic, which affects research in this field. In this section, we highlight these issues.

A. DATA RESOURCES

Unlike English, which has a variety of large datasets, Arabic conversational AI systems have been hindered by the deficiency of massive datasets. Arabic datasets are scarce, limited in size, and not diverse enough to be adaptable to the three discussed types of conversational AI systems [47]. Even with the massive Arabic content on the internet, obtaining large datasets is difficult because the datasets must be annotated. Indeed, annotating such large datasets is costly in terms of both effort and time.

Given the rarity of the Arabic resources (e.g., corpora and taxonomies), most researchers have designed their studies with their own datasets, which need to be annotated and verified manually [27]. For this reason, comparing Arabic conversational AI systems is more difficult. Moreover, these datasets are not big enough to design adaptable systems or more unified Arabic systems or datasets.

B. LANGUAGE-RELATED ISSUES

1) AMBIGUITY

Ambiguity is one of the main challenges of conversational AI systems in Arabic that needs to be resolved. Ambiguity occurs in such systems because a word can be interpreted differently based on the context. It also occurs due to the non-unification of all versions of a letter into one form, especially in the case of a specific domain [24]. In addition, the coverage of the dataset plays a role in increasing the ambiguity, as the complexity of some sentences requires special handling techniques. Some cases of ambiguity can be easily handled, while others require deep linguistic approaches.

Ambiguity negatively influences the performance of conversational AI systems due to confusion in interpreting either the answer, question, or sentence. Furthermore, ambiguity makes the choice of building the taxonomies for conversational AI systems extremely difficult. To reduce this kind of ambiguity in, for example, QA systems, the taxonomies employed should maximize the distance in terms of similarity between the question types [39]. Thus, there is a need for disambiguation solutions to improve Arabic conversational AI systems' performance. A few studies have thoroughly considered the task of disambiguation [24], [25], [31]. Although these studies have obtained interesting results, they have certain limitations in terms of performance, domains, question types, or dataset size. Furthermore, they have only examined factual questions, where the answer is either a single word or a short phrase. Temporal questions are one type of question that has several representations; this causes ambiguity because the temporal entities in the questions may be expressed in various ways [33].

As discussed in the literature, some studies utilized the RTE technique, which is remarkably effective in distinguishing between different expressions that express the same meaning [28], [30], [31]. However, their accuracy ranges between 70% and 73%, suggesting that they must either increase the dataset size or enhance the technique.

2) DIALECTS

Many Arabic internet users use colloquial Arabic instead of MSA to write their speech or even ask their questions. In many cases, this means that dialectal words are recognized as out-of-vocabulary (OOV) words by the Arabic morphological analyzers and other language understanding tools, most of which are designed for MSA. In addition, language variety is another issue that occurs due to the mismatch between question and answer texts in conversational systems. Moreover, users may use a mix between MSA and colloquial Arabic, which increases the challenges of Arabic. Another aspect related to colloquial Arabic is that users neglect writing with punctuation, unlike MSA. Indeed, the proper use of punctuation can be essential for some NLP processing steps, like syntactic parsing and sentence segmentation. Furthermore, some spelling mistakes occur while using dialects, such as swapped letters, missing letters, or split or concatenated words [77].

Very few systems are designed for dialects [69]. However, they suffer from the lack of labeled datasets available for these dialects. Moreover, even if such datasets were available, it would be difficult to train such models [47].

C. MORPHOLOGICAL ISSUES

Arabic has very rich and complex morphological features, including both derivational and inflectional features. Derivational morphology features concern how words are formed, while inflectional morphology features concern how words interact with the syntax. Thus, generating words with different semantic meanings poses challenges to the computational processing of Arabic. These challenges involve many linguistic factors, such as affixes, vowels, root-based systems, diacritical marks, absence of capital letters, free word order, name de-spacing, inconsistent name spelling, dual forms, and grammatical gender variation. The literature provides more information on these challenges [13], [78].

D. DOMAIN-RELATED ISSUES

Obviously, many domains are largely ignored for Arabic when building datasets or conducting research. For example, some researchers face difficulties related to the domain for which their system was built [77]. The medical domain is one example that has many challenges; using English words for medical terms is one of the most common issues. Most of these terms are complex and not understood by Arabic speakers, such as names of diseases and medicines, and especially those that are not used every day.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we reviewed the recent progress towards developing Arabic conversational AI systems. Furthermore, we presented the frequently used corpora for the development and evaluation of Arabic conversational AI. Among Arabic conversational systems, there are many studies that address QA systems, while task-oriented DS and chatbots are still in their infancy. However, QA systems lack both large datasets and diversity in domains.

Despite the recent progress in state-of-the-art conversational AI systems in English, building conversational AI systems that can converse on different topics in addition to achieving coherent, sustained conversations remains extremely challenging [10]. Moreover, Arabic is lagging behind English-language systems. Until now, it has appeared difficult to apply the advanced models to Arabic due to the challenges it faces. However, it is important to find suitable techniques to leverage the power of the new advanced models.

Concerning the future of the Arabic conversational AI research community, there is a need to build more intelligent conversational systems that are capable of dialogue that is consistent, empathetic, personable, and engaging. These systems will combine personalization, self-learning, and generative-based responses in a single solution. Thus, we conclude our paper by discussing some future research avenues, including dialogue-generation models, topic and knowledge bases, empathetic conversational systems, cross-lingual transfer learning, and multilingualism.

A. DIALOGUE-GENERATION MODELS

Recently, there has been increased interest in fully data-driven end-to-end systems that employ user input and generate suitable responses using neural generative models. Seq2seq, transformer, and pre-trained models are the most common choices and have achieved very good performance in generative models [10]. Most generative models adopt the encoder-decoder framework, including BERT and GPT-2. Therefore, there is a need to provide large-scale conversational corpora with verified and cleaned annotations, which will help in the exploration of these models for Arabic. Transformers and pre-trained models can address the data scarcity issue and bring significant advantages by generating contextualized word embeddings. Thus, their low computational cost is suitable for smaller datasets, which are more feasible for Arabic. The ability of recent contextualized embeddings to be trained on unannotated data has made it possible to improve many Arabic NLP tasks. A number of contextualized embedding models have been developed to support Arabic, such as Multilingual BERT [79], AraBERT [26], MarBert [80], GigaBert [81], QARiB,[11] AraGPT2 [82], AraELECTRA [83], ARBERT & MARBERT [84], and AraT5 [85]. Another point that may be exploited to build good Arabic conversational AI systems is achieving controllability for these models. Hybrid models are a good choice to achieve this. These models combine the power of both retrieval and generative models, where retrieved responses can be matched with generated responses to help the retrieval models find a better response.

B. TOPIC AND KNOWLEDGE BASE

Building real-world conversations requires strong and content-rich datasets that include real-world topics and entities; such datasets comprise a system's KB. The KB considerably improves the ability to understand the language in the dialogue context. Consider the slight differences among the types of conversational AI systems previously discussed. For instance, chatbots do not need to deal with annotated dialogue acts nor detect users' intents unlike task-oriented dialogue systems. Due to the Arabic dialects, there is a need to provide large-scale datasets covering different domains.

C. EMPATHETIC CONVERSATIONAL SYSTEMS

Recently, with the advancements in sequence generation models, empathetic conversational models have received significant research interest. In this area, researchers seek to develop conversational models that have the ability to empathize with humans. Emotion is an essential factor to build more effective interactions with empathetic conversation systems. Such systems should be able to realize the user's emotional state and then create emotional responses and conversations. To the best of our knowledge, existing research in this direction is still in its infancy, and only one study sought to design an empathetic conversation system in Arabic [65]. As a result, Arabic lacks such models, as well as the large conversational datasets that support empathetic responses. As a future direction to build empathetic chatbots, using text datasets is insufficient, and speech rhythm and facial expressions may be useful [86], [87].

D. CROSS-LINGUAL TRANSFER LEARNING

Very recently, cross-lingual transfer learning achieved improved results among several languages, including Arabic, with the help of pre-trained multilingual models such as Multi-BERT [81], [88] and AraT5 [85]. Indeed, languages that share specific morpho-syntactic features tend to benefit from transfer learning. Transfer learning helps to transfer knowledge from high-resource languages into low-resource languages. For the MLQA task, the huge number of English datasets can be used for low-resource languages such as Arabic, which has a small number of datasets.

E. MULTILINGUALISM

Multilingualism is one of the main challenges of conversational AI systems, which means the system has the ability to handle conversation with users in many different languages. Nevertheless, recent solutions are based on classical approaches, which are mainly limited to machine translation and manual feature engineering [5]. Additionally, in the last few years, several mutilingual pretrained models have emerged, including mT5 [89], mBART [90], which help in building multilingual conversational systems. However such systems need to multilingual dictionaries and datasets to be trained. For Arabic, such systems can handle different dialects or Arabic

[11] https://github.com/qcri/QARIB

languages with other non-Arabic languages. Thus, these systems need more development with advanced neural approaches and large corpora.

To conclude, our research presented a taxonomy emphasizing text-based conversational AI systems divided into three major categories: QA systems, task-oriented DS, and chatbots. We further attempted to answer various questions regarding the goals and discussed the progress that has been made as well as the future challenges of Arabic conversational AI systems.

REFERENCES

[1] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *Proc. 56th Annu. Meeting Assoc. for Comput. Linguistics, Proc. Conf. Tutorial Abstr. (ACL)*, 2018, pp. 2–7, doi: 10.18653/v1/p18-5002.

[2] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," 2020, *arXiv:2001.09977*.

[3] S. Roller, Y.-L. Boureau, J. Weston, A. Bordes, E. Dinan, A. Fan, D. Gunning, D. Ju, M. Li, S. Poff, P. Ringshia, K. Shuster, E. Michael Smith, A. Szlam, J. Urbanek, and M. Williamson, "Open-domain conversational agents: Current progress, open problems, and future directions," 2020, *arXiv:2006.12442*.

[4] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 2011–2027, Oct. 2020, doi: 10.1007/s11431-020-1692-3.

[5] E. Loginova, S. Varanasi, and G. Neumann, "Towards end-to-end multilingual question answering," *Inf. Syst. Frontiers*, vol. 23, no. 1, pp. 227–241, Feb. 2021, doi: 10.1007/s10796-020-09996-1.

[6] M. A. Calijorne Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 6, pp. 635–646, Jul. 2020, doi: 10.1016/j.jksuci.2018.08.005.

[7] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Mach. Learn. Appl.*, vol. 2, Dec. 2020, Art. no. 100006, doi: 10.1016/j.mlwa.2020.100006.

[8] S. Singh and H. K. Thakur, "Survey of various AI chatbots based on technology used," in *Proc. 8th Int. Conf. Rel., Infocom Technol. Optim. (Trends Future Directions) (ICRITO)*, Jun. 2020, pp. 1074–1079, doi: 10.1109/ICRITO48877.2020.9197943.

[9] S. Fernandes, R. Gawas, P. Alvares, M. Femandes, D. Kale, and S. Aswale, "Survey on various conversational systems," in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (IC-ETITE)*, Feb. 2020, pp. 1–8, doi: 10.1109/ic-ETITE47903.2020.126.

[10] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–32, Jun. 2020.

[11] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, S. R. El-Beltagy, W. El-Hajj, M. Jarrar, and H. Mubarak, "A panoramic survey of natural language processing in the Arab world," *Commun. ACM*, vol. 64, no. 4, pp. 72–81, Apr. 2021.

[12] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh, and B. Gupta, "Deep learning for Arabic NLP: A survey," *J. Comput. Sci.*, vol. 26, pp. 522–531, May 2018, doi: 10.1016/j.jocs.2017.11.011.

[13] H. Samy, E. E. Hassanein, and K. Shaalan, "Arabic question answering: A study on challenges, systems, and techniques," *Int. J. Comput. Appl.*, vol. 181, no. 44, pp. 6–14, Mar. 2019, doi: 10.5120/ijca2019918524.

[14] E. S. Al-Hagbani and M. B. Khan, "Support of existing chatbot development framework for Arabic language: A brief survey," in *Proc. 5th Int. Symp. Data Mining Appl.*, in Advances in Intelligent Systems and Computing, vol. 753. Cham, Switzerland: Springer, 2018, pp. 26–35, doi: 10.1007/978-3-319-78753-4_3.

[15] S. AlHumoud, A. Al, and W. Aldamegh, "Arabic chatbots: A survey," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 8, pp. 535–541, 2018, doi: 10.14569/ijacsa.2018.090867.

[16] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2383–2392, doi: 10.18653/v1/d16-1264.

[17] Y. Yang, W.-T. Yih, and C. Meek, "WikiQA: A challenge dataset for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2013–2018, doi: 10.18653/v1/d15-1237.

[18] H. Mozannar, E. Maamary, K. El Hajal, and H. Hajj, "Neural Arabic question answering," in *Proc. 4th Arabic Natural Lang. Process. Workshop*, 2019, pp. 108–118, doi: 10.18653/v1/w19-4612.

[19] W. S. Ismail and M. N. Homsi, "DAWQAS: A dataset for Arabic why question answering system," *Proc. Comput. Sci.*, vol. 142, pp. 123–131, Jan. 2018, doi: 10.1016/j.procs.2018.10.467.

[20] L. Abouenour, K. Bouzouba, and P. Rosso, "An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering," *Int. J. Inf. Commun. Technol.*, vol. 3, no. 3, pp. 37–51, 2010.

[21] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor, "SemEval-2017 task 3: Community question answering," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, 2017, pp. 27–48, doi: 10.18653/v1/s17-2003.

[22] E. Bessaies, S. Mesfar, and H. B. Ghzela, "Processing medical binary questions in standard Arabic using NooJ," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.*, in Lecture Notes in Computer Science, vol. 10859. Cham, Switzerland: Springer, 2018, pp. 193–204, doi: 10.1007/978-3-319-91947-8_19.

[23] I. Ennasri, S. Dardour, H. Fehri, and K. Haddar, "Question-response system using the NooJ linguistic platform," *Commun. Comput. Inf. Sci.*, vol. 811, pp. 190–199, 2018, doi: 10.1007/978-3-319-73420-0_16.

[24] S. Dardour, H. Fehri, and K. Haddar, "Disambiguation for Arabic question-answering system," in *Proc. Int. Conf. Autom. Process. Natural-Lang. Electron. Texts NooJ*, in Communications in Computer and Information Science, vol. 1153. Cham, Switzerland: Springer, 2020, pp. 101–111, doi: 10.1007/978-3-030-38833-1_9.

[25] M. A. Smadi, I. A. Dalabih, Y. Jararweh, and P. Juola, "Leveraging linked open data to automatically answer Arabic questions," *IEEE Access*, vol. 7, pp. 177122–177136, 2019, doi: 10.1109/ACCESS.2019.2956233.

[26] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools, Shared Task Offensive Lang. Detection*, 2020, pp. 9–15.

[27] I. Lahbari, S. El Alaoui, and K. Zidani, "Toward a new Arabic question answering system," *Int. Arab J. Inf. Technol.*, vol. 15, no. 3A, pp. 610–619, 2018.

[28] W. Bakari and M. Neji, "A novel semantic and logical-based approach integrating RTE technique in the Arabic question-answering," *Int. J. Speech Technol.*, no. 0123456789, 2020, doi: 10.1007/s10772-020-09684-0.

[29] W. Bakari, P. Bellot, and M. Neji, "AQA-WebCorp: Web-based factual questions for Arabic," *Proc. Comput. Sci.*, vol. 96, pp. 275–284, Jan. 2016, doi: 10.1016/j.procs.2016.08.140.

[30] M. Ben-Sghaier, W. Bakari, and M. Neji, "Ar-SLoTE: A recognizing textual entailment tool for Arabic question/answering systems," in *Proc. 7th Int. Conf. ICT Accessibility (ICTA)*, Dec. 2019, pp. 1–6, doi: 10.1109/ICTA49490.2019.9144976.

[31] M. Ben-Sghaier, W. Bakari, and M. Neji, "Recognizing textual entailment for Arabic using semantic similarity and word sense disambiguation," in *Proc. CEUR Workshop*, 2018, vol. 2279, no. 1, pp. 1–10.

[32] I. Lahbari, H. Alami, and S. E. A. O. Khalid Alaoui Zidani, "Towards a passages extraction method for Arabic question answering systems," in *Proc. Int. Conf. Adv. Intell. Syst. Sustain. Develop.*, 2020, vol. 1, no. 1, pp. 230–237, doi: 10.1007/978-3-030-36653-7.

[33] M. Mtibaa, Z. Neji, and M. Ellouze, "A temporal Arabic question answering system based on the construction of a temporal resource," in *Proc. CEUR Workshop*, vol. 2279, 2018, pp. 1–10.

[34] M. Al-Shenak, K. M. O. Nahar, and K. M. H. Halawani, "AQAS: Arabic question answering system based on SVM, SVD, and LSI," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 2, pp. 681–691, 2019.

[35] Y. E. Adlouni, H. Rodríguez, M. Meknassi, S. O. El Alaoui, and N. En-Nahnahi, "A multi-approach to community question answering," *Expert Syst. Appl.*, vol. 137, pp. 432–442, Dec. 2019, doi: 10.1016/j.eswa.2019.07.024.

[36] A. Almiman, N. Osman, and M. Torki, "Deep neural network approach for Arabic community question answering," *Alexandria Eng. J.*, vol. 59, no. 6, pp. 4427–4434, Dec. 2020, doi: 10.1016/j.aej.2020.07.048.

[37] A. Hamza, N. En-Nahnahi, and S. E. A. Ouatik, "Exploring contextual word representation for Arabic question classification," in *Proc. 1st Int. Conf. Innov. Res. Appl. Sci., Eng. Technol. (IRASET)*, Apr. 2020, pp. 1–5, doi: 10.1109/IRASET48871.2020.9092084.

[38] A. M. Hasan and T. H. Rassem, "Combined support vector machine and pattern matching for Arabic Islamic Hadith question classification system," in *Proc. Int. Conf. Reliable Inf. Commun. Technol.*, vol. 843. Cham, Switzerland: Springer, 2019, pp. 278–290, doi: 10.1007/978-3-319-99007-1_27.

[39] A. Hamza, N. En-Nahnahi, K. A. Zidani, and S. E. A. Ouatik, "An Arabic question classification method based on new taxonomy and continuous distributed representation of words," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 2, pp. 218–224, 2021, doi: 10.1016/j.jksuci.2019.01.001.

[40] A. Aouichat, M. S. H. Ameur, and A. Geussoum, "Arabic question classification using support vector machines and convolutional neural networks," in *Natural Language Processing and Information Systems* (Lecture Notes in Computer Science), vol. 10859, M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-91947-8_12.

[41] M. Abdel-Latif, M. Samir, S. Abdel-Aziz, M. Heeba, A. Elmasry, and M. Torki, "A supervised learning approach using the combination of semantic and lexical features for Arabic community question answering," in *Proc. IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2018, pp. 1–7.

[42] X. Li and D. Roth, "Learning question classifiers," in *Proc. 19th Int. Conf. Comput. Linguistics*, 2002, pp. 1–7, doi: 10.3115/1072228.1072378.

[43] A. Aouichat and A. Guessoum, "Building TALAA-AFAQ, a corpus of Arabic FActoid question-answers for a question answering system," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2017, pp. 380–386, doi: 10.1007/978-3-319-59569-6.

[44] M. Daoud, "Novel approach towards Arabic question similarity detection," in *Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2019, pp. 1–6, doi: 10.1109/ICTCS.2019.8923102.

[45] S. Romeo, G. Da San Martino, Y. Belinkov, A. Barrón-Cedeño, M. Eldesouki, K. Darwish, H. Mubarak, J. Glass, and A. Moschitti, "Language processing and learning models for community question answering in Arabic," *Inf. Process. Manage.*, vol. 56, no. 2, pp. 274–290, Mar. 2019, doi: 10.1016/j.ipm.2017.07.003.

[46] N. Othman, R. Faiz, and K. Smaïli, "Enhancing question retrieval in community question answering using word embeddings," *Proc. Comput. Sci.*, vol. 159, pp. 485–494, Jan. 2019, doi: 10.1016/j.procs.2019.09.203.

[47] H. Al-Bataineh, W. Farhan, A. Mustafa, H. Seelawi, and H. T. Al-Natsheh, "Deep contextualized pairwise semantic similarity for Arabic language questions," in *Proc. IEEE 31st Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2019, pp. 1–14.

[48] O. Einea and A. Elnagar, "Predicting semantic textual similarity of Arabic question pairs using deep learning," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–4, doi: 10.1109/AICCSA47632.2019.9035362.

[49] H. Bouamor, N. Habash, M. Salameh, and W. Zaghouani, "The MADAR Arabic dialect corpus and lexicon," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2019, pp. 3387–3396.

[50] M. McTear, *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*, vol. 13, no. 3. San Rafael, CA, USA: Morgan & Claypool.

[51] A. M. Bashir, A. Hassan, B. Rosman, D. Duma, and M. Ahmed, "Implementation of a neural natural language understanding component for Arabic dialogue systems," *Proc. Comput. Sci.*, vol. 142, pp. 222–229, Jan. 2018, doi: 10.1016/j.procs.2018.10.479.

[52] A. Fadhil and A. AbuRa'ed, "OlloBot–towards a text-based Arabic health conversational agent: Evaluation and results," in *Proc. Natural Lang. Process. Deep Learn. World*, Oct. 2019, pp. 295–303, doi: 10.26615/978-954-452-056-4_034.

[53] A.-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, doi: 10.1109/ACCESS.2021.3049732.

[54] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic word embedding models for use in Arabic NLP," *Proc. Comput. Sci.*, vol. 117, pp. 256–265, Jan. 2017, doi: 10.1016/j.procs.2017.10.117.

[55] F. Bendjamaa and T. Nora, "A dialogue-system using a Qur'anic ontology," in *Proc. 2nd Int. Conf. Embedded Distrib. Syst. (EDiS)*, Nov. 2020, pp. 167–171.

[56] L. Sherkawi, N. Ghneim, and O. A. Dakkak, "Arabic speech act recognition techniques," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 3, pp. 1–12, May 2018, doi: 10.1145/3170576.

[57] A. Joukhadar, H. Saghergy, L. Kweider, and N. Ghneim, "Arabic dialogue act recognition for textual chatbot systems," *Proc. 1st Int. Workshop NLP Solutions Under Resourced Lang. (NSURL) Co-Located ICNLSP-Short Papers*, 2019, pp. 43–49.

[58] A. R. A. Elmadany, S. M. Abdou, and M. Gheith, "Improving dialogue act classification for spontaneous Arabic speech and instant messages at utterance level," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2019, pp. 128–134.

[59] P. Smutny and P. Schreiberova, "Chatbots for learning: A review of educational chatbots for the Facebook messenger," *Comput. Educ.*, vol. 151, Jul. 2020, Art. no. 103862, doi: 10.1016/j.compedu.2020.103862.

[60] T. Alshareef and M. A. Siddiqui, "A Seq2Seq neural network based conversational agent for Gulf Arabic dialect," in *Proc. 21st Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2020, pp. 1–7, doi: 10.1109/ACIT50332.2020.9300059.

[61] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.

[62] Ö. N. Yalcin and S. DiPaola, "A computational model of empathy for interactive agents," *Biologically Inspired Cognit. Archit.*, vol. 26, pp. 20–25, Oct. 2018, doi: 10.1016/j.bica.2018.07.010.

[63] N. Asghar, I. Kobyzev, J. Hoey, P. Poupart, and M. Bilal Sheikh, "Generating emotionally aligned responses in dialogues using affect control theory," 2020, *arXiv:2003.03645*.

[64] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "EMMA: An emotion-aware wellbeing chatbot," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 15–21, doi: 10.1109/ACII.2019.8925455.

[65] T. Naous, C. Hokayem, and H. Hajj, "Empathy-driven Arabic conversational chatbot," in *Proc. 5th Arabic Natural Lang. Process. Workshop*, 2020, pp. 58–68.

[66] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5370–5381, doi: 10.18653/v1/p19-1534.

[67] T. Naous, W. Antoun, R. A. Mahmoud, and H. Hajj, "Empathetic BERT2BERT conversational model: Learning Arabic language generation with little data," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Kyiv, Ukraine, 2021, pp. 164–172.

[68] S. A. Walid El Hefny, Y. Mansy, and M. Abdallah, "Jooka: A bilingual chatbot for University admission," in *Proc. WorldCIST*, no. 3, 2021, pp. 671–681, doi: 10.1007/978-3-030-72660-7.

[69] D. Al-Ghadhban and N. Al-Twairesh, "Nabiha: An Arabic dialect chatbot," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 452–459, 2020, doi: 10.14569/ijacsa.2020.0110357.

[70] Y. Almurtadha, "LABEEB: Intelligent conversational agent approach to enhance course teaching and allied learning outcomes attainment," *J. Appl. Comput. Sci. Math.*, vol. 13, no. 1, pp. 9–12, 2019, doi: 10.4316/jacsm.201901001.

[71] S. Aljameel *et al.*, "LANA-I: An Arabic conversational intelligent tutoring system for children with ASD," in *Proc. Comput. Conf. Intell. Comput.*, vol. 997. Cham, Switzerland: Springer, 2019, pp. 498–516, doi: 10.1007/978-3-030-22871-2_34.

[72] M. Z. Khan and S. M. Yassin, "SeerahBot: An Arabic chatbot about prophet's biography," *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 9, no. 2, pp. 89–97, 2021, doi: 10.21276/ijircst.2021.9.2.13.

[73] M. Boussakssou, H. Ezzikouri, and M. Erritali, "Chatbot in Arabic language using seq to seq model," *Multimedia Tools Appl.*, vol. 81, pp. 2859–2871, 2021, doi: 10.1007/s11042-021-11709-y.

[74] R. Malhas and T. Elsayed, "*AyaTEC* building a reusable verse-based test collection for Arabic question answering on the holy Qur'an," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 6, pp. 1–21, Nov. 2020, doi: 10.1145/3400396.

[75] H. Seelawi, A. Mustafa, H. Al-Bataineh, W. Farhan, and H. T. Al-Natsheh, "NSURL-2019 shared task 8: Semantic question similarity in Arabic," 2019, *arXiv:1909.09691*.

[76] O. Trigui, L. H. Belguith, and P. Rosso, "DefArabicQA: Arabic definition question answering system," in *Proc. Workshop Lang. Resour. Hum. Lang. Technol. Semitic Lang., 7th LREC*, 2010, pp. 40–44.

[77] P. Nakov, L. Màrquez, A. Moschitti, and H. Mubarak, "Arabic community question answering," *Natural Lang. Eng.*, vol. 25, no. 1, pp. 5–41, Jan. 2019, doi: 10.1017/S1351324918000426.

[78] N. Y. Habash, *Introduction to Arabic Natural Language Processing*, vol. 3, no. 1. San Rafael, CA, USA: Morgan & Claypool, 2010.

[79] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL HLT)*, vol. 1, 2019, pp. 4171–4186.

[80] M. Abdul-Mageed, C. Zhang, A. Elmadany, and L. Ungar, ''Toward micro-dialect identification in diaglossic and code-switched environments,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 5855–5876, doi: 10.18653/v1/2020.emnlp-main.472.

[81] W. Lan, Y. Chen, W. Xu, and A. Ritter, ''An empirical study of pre-trained transformers for Arabic information extraction,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 4727–4734.

[82] W. Antoun, F. Baly, and H. Hajj, ''AraGPT2: Pre-trained transformer for Arabic language generation,'' in *Proc. 6th Arabic Natural Lang. Process. Workshop*, 2021, pp. 196–207.

[83] W. Antoun, F. Baly, and H. Hajj, ''ARAELECTRA: Pre-training text discriminators for Arabic language understanding,'' in *Proc. 6th Arabic Natural Lang. Process. Workshop*. Kyiv, Ukraine: Association for Computational Linguistics, 2021, pp. 191–195.

[84] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, ''ARBERT & MARBERT: Deep bidirectional transformers for Arabic,'' in *Proc. 59th Annu. Meeting Assoc. for Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2021, pp. 7088–7105.

[85] E. Moatez Billah Nagoudi, A. Elmadany, and M. Abdul-Mageed, ''AraT5: Text-to-text transformers for Arabic language understanding and generation,'' 2021, *arXiv:2109.12068*.

[86] A. Saha, M. M. Khapra, and K. Sankaranarayanan, ''Towards building large scale multimodal domain-aware conversation systems,'' in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 696–704.

[87] C. Cui, W. Wang, X. Song, M. Huang, X.-S. Xu, and L. Nie, ''User attention-guided multimodal dialog systems,'' in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 445–454, doi: 10.1145/3331184.3331226.

[88] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, ''Zero-shot cross-lingual transfer with meta learning,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.* 2020, pp. 4547–4562, doi: 10.18653/v1/2020.emnlp-main.368.

[89] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, ''mT5: A massively multilingual pre-trained text-to-text transformer,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2020, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41.

[90] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, ''Multilingual denoising pre-training for neural machine translation,'' *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020, doi: 10.1162/tacl_a_00343.

AHLAM FUAD received the B.S. degree in information technology from the College of Engineering and Information Technology, Taiz University, Taiz, Yemen, in 2011. She is currently pursuing the M.S. degree with the Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Previously, she worked as a Teacher Assistant with the Information Technology Department, College of Engineering and Information Technology, Taiz University. Her current research interests include machine learning, dialogue systems, Arabic NLP, and data science.

MAHA AL-YAHYA received the M.Sc. degree in computer science from Bristol University, U.K., and the Ph.D. degree in computer science from the University of Nottingham, U.K. She is currently an Associate Professor with the Information Technology Department, College of Computer Information Sciences, King Saud University, Riyadh, Saudi Arabia, where she is a member of the IWAN Research Group. Her research interests include Arabic computing and Arabic NLP.