


Article

# AraConv: Developing an Arabic Task-Oriented Dialogue System Using Multi-Lingual Transformer Model mT5

Ahlam Fuad \*  and Maha Al-Yahya

Department of Information Technology, College of Computer and Information Sciences, King Saud University, P.O. Box 145111, Riyadh 4545, Saudi Arabia; malyahya@ksu.edu.sa

\* Correspondence: aabdulghni@ksu.edu.sa

**Abstract:** Task-oriented dialogue systems (DS) are designed to help users perform daily activities using natural language. Task-oriented DS for English language have demonstrated promising performance outcomes; however, developing such systems to support Arabic remains a challenge. This challenge is mainly due to the lack of Arabic dialogue datasets. This study introduces the first Arabic end-to-end generative model for task-oriented DS (AraConv), which uses the multi-lingual transformer model mT5 with different settings. We also present an Arabic dialogue dataset (Arabic-TOD) and used it to train and test the proposed AraConv model. The results obtained are reasonable compared to those reported in the studies of English and Chinese using the same mono-lingual settings. To avoid problems associated with a small training dataset and to improve the AraConv model's results, we suggest joint-training, in which the model is jointly trained on Arabic dialogue data and data from one or two high-resource languages such as English and Chinese. The findings indicate the AraConv model performed better in the joint-training setting than in the mono-lingual setting. The results obtained from AraConv on the Arabic dialogue dataset provide a baseline for other researchers to build robust end-to-end Arabic task-oriented DS that can engage with complex scenarios.

**Keywords:** task-oriented dialogue systems; Arabic; multi-lingual transformer model; mT5; natural language processing



**Citation:** Fuad, A.; Al-Yahya, M. AraConv: Developing an Arabic Task-Oriented Dialogue System Using Multi-Lingual Transformer Model mT5. *Appl. Sci.* **2022**, *12*, 1881. <https://doi.org/10.3390/app12041881>

**Academic Editors:**  
Arturo Montejo-Ráez and Salud María Jiménez-Zafra

Received: 7 January 2022  
Accepted: 3 February 2022  
Published: 11 February 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Task-oriented dialogue systems (DS) are a type of conversational system designed to help users achieve pre-defined tasks. These systems are designed to help humans perform routine activities, such as make restaurant or hotel reservations, search for attractions, book flights, enquire about the weather forecast, and shop online. Task-oriented DS are considered the core modules of virtual assistants such as Google Assistant, Amazon Alexa, and Apple Siri, which utilize natural language interfaces for various online services [1]. Task-oriented DS allow users to ask questions using natural language and provide answers to those questions in the form of a conversation.

Despite the current progress of state-of-the-art English-based task-oriented DS, it remains a substantial challenge to build systems that can achieve coherent, sustained conversation on diverse topics [2]. Notably, task-oriented DS for Arabic lag behind [3], until now precluding the application of advanced data-intensive deep-learning models for the language [4], especially due to the shortage of Arabic dialogue datasets. Therefore, this study aimed to investigate the effectiveness of the multi-lingual pre-trained language model mT5 [5] for building end-to-end Arabic task-oriented DS. These end-to-end DS must be capable of handling both dialogue state tracking (DST) task and response generation task; in this context, DST is mainly responsible for helping to extract the goals and slot-value pairs from the conversation. As such, this work aimed to answer the following major research questions:

*RQ1: To what extent can mT5, a multi-lingual pre-trained language model, produce satisfactory results for Arabic end-to-end task-oriented DS?*

*RQ2: To what extent can joint-training the mT5 model on Arabic dialogue data and data for one or two high-resource languages (namely, English or English and Chinese) improve the performance of Arabic task-oriented DS?*

To answer these research questions, we conducted several experiments, leading this work to make the following contributions:

- Development of the first Arabic task-oriented dialogue dataset (Arabic-TOD) with 1500 dialogues. By translating the English BiToD dataset [1], we produced a valuable benchmark for further exploring Arabic task-oriented DS. Furthermore, Arabic-TOD is the first code-switching dialogue dataset for Arabic task-oriented DS.
- Introduction of the first Arabic end-to-end generative model, the AraConv model, short for Arabic Conversation, that achieves both DST and response generation tasks together in an end-to-end setting.

The paper comprises five sections. The next section explores related works in the area of task-oriented DS for both English and Arabic. The third section demonstrates the methodology used in this research, including the data collection process and the model architecture. Next, we detail our experiments, discussing the tasks and evaluation metrics, experimental setup, and findings. Finally, the fifth section summarizes our work and the significance of the AraConv before considering possible future research directions.

## 2. Related Works

There are two approaches in applying DS: traditional DS and end-to-end DS. Traditional DS use a pipeline that connects, trains, and evaluates each module separately. End-to-end DS are designed to train all modules as a single unit directly on both knowledge-based information and text transcripts [6]. This section discusses the evaluation of task-oriented DS for the English language before surveying the landscape of Arabic task-oriented DS.

### 2.1. English Task Oriented Dialogue Systems

Given the availability of multi-domain English task-oriented dialogue datasets, work on task-oriented DS in the language has progressed from modularized modeling to generative and end-to-end modeling. Given the fact that the traditional DS design complicates tracking the module responsible for interaction failure [6], some studies have built DS using the end-to-end paradigm [7–15]. However, building powerful task-oriented DS still engenders many challenges due to the system design complexity and the limited availability of human-annotated data. Therefore, the research community has focused on working with the pre-trained language models to reduce human supervision to the extent possible. This approach involves fine-tuning these models and helping to transfer the prior knowledge to improve various NLP tasks, including task-oriented DS. Large pre-trained language models, such as GPT2 and T5, have been used for various NLP tasks, especially language generation tasks. These new approaches model the dialogue pipeline in an end-to-end manner [6].

Given the high costs associated with data collection and annotation, researchers tend to train their models with the least number of samples using transfer learning. Transfer learning represents one of the most successful few-shot learning approaches for task-oriented DS. It refers to pre-training large language models on text or task-related data and then fine-tuning on a few samples. Such systems have proved their success in task-oriented DS such as the work presented in [12–22].

The task-oriented DS literature includes two study categories: studies targeting only DST and studies targeting both DST and response generation. Dialogue state tracking mainly helps to extract the goals (intents) and slot-value pairs from the conversation to maintain the dialogue belief state (BS) and the summary of the dialogue history. The BS contains information about the dialogue from the system perspective [6]. At each user turn

during the conversation, the input to the DST comprises the previous BS, the outputs of the intent classification (the goal), and slot filling information; thus, the DST output is the new /updated BS. For end-to-end dialogue generation, the system indicates the correct required information and generates the appropriate response.

For the first category, studies targeting only DST, some studies focus on handling the DST task to guarantee building a good base for the whole dialogue system [23–29]. Meanwhile, other studies have targeted both DST and response generation in an end-to-end manner [11,12].

Table 1 summarizes the available models for task-oriented DS in English, including datasets and performance measures. Although the models have achieved promising results, they have been designed for English-language task-oriented DS, and, to the best of our knowledge, no research exists concerning Arabic-language task-oriented DS.

Nonetheless, the promising performance of pre-trained language models for English-language task-oriented DS has prompted efforts to produce multi-lingual models for task-oriented DS in other languages. Many of these languages are considered low-resource languages due to the absence of high-quality data in the language, and most existing task-oriented DS do not support low-resource languages, creating a gap between the performance of low-resource language systems and high-resource systems. Therefore, providing datasets for low-resource languages is critical to driving the development of efficient end-to-end task-oriented DS for these languages. Several existing studies have built task-oriented DS for low-resource languages using cross-lingual transfer learning [1,30,31]. This involves transferring knowledge from high-resource to low-resource languages, enabling the satisfactory performance of end-to-end task-oriented DS.

**Table 1.** Comparing the performances of the most common English-based task-oriented dialogue systems (DS). Bold numbers indicating the best system according to the column’s metric value.

Model	Dataset	Back-Bone Models	Performance Metrics			
			BLEU	Inform Rate	Success Rate	JGA
DAMD [32]	MultiWOZ 2.1	multi-decoder seq2seq	16.6	76.4	60.4	51.45
Ham [10]	MultiWOZ 2.1	GPT-2	6.01	77.00	69.20	44.03
SimpleToD [11]	MultiWOZ 2.1	GPT-2	15.23	85.00	70.05	<b>56.45</b>
SC-GPT [16]	MultiWOZ	GPT-2	<b>30.76</b>	-	-	-
SOLOIST [12]	MultiWOZ 2.0	GPT-2	16.54	85.50	72.90	-
MARCO [33]	MultiWOZ 2.0	-	20.02	92.30	78.60	-
UBAR [13]	MultiWOZ 2.1	GPT-2	17.0	<b>95.4</b>	<b>80.7</b>	56.20
ToD-BERT [17]	MultiWOZ 2.1	BERT	-	-	-	48.00
MinTL [14]	MultiWOZ 2.0	T5-small	19.11	80.04	72.71	51.24
		T5-base	18.59	82.15	74.44	52.07
		BART-large	17.89	84.88	74.91	52.10
LABES-S2S [20]	MultiWOZ 2.1	A copy-augmented Seq2Seq	18.3	78.1	67.1	51.45
AuGPT [21]	MultiWOZ 2.1	GPT-2	17.2	91.4	72.9	-
GPT-CAN [15]	MultiWOZ 2.0	GPT-2	17.02	93.70	76.70	55.57
HyKnow [22]	MultiWOZ 2.1	multi-stage Seq2Seq	18.0	82.3	69.4	49.2

## 2.2. Arabic Task-Oriented Dialogue Systems

Considering the maturity of research concerning English-based task-oriented DS, we find that task-oriented DS research more broadly remains in its infancy for Arabic. This is due to a lack of fundamental NLP resources and a scarcity of datasets for Arabic task-oriented DS. Most of the research on Arabic task-oriented DS focuses on achieving specific tasks, such as intent classification [34–36] and entity classification [34]. However, there

some attempts to build task-oriented DS have investigated specific domains, including home automation [34], flight bookings [37], education [38–40], hotel reservations [41], and Islamic knowledge enquires [42]. Some Arabic task-oriented DS have been designed to specifically serve the Arabic dialects (e.g., OlloBot [43] and Nabiha [44]). However, this review excludes some of these studies because they are categorized as chatbots rather than task-oriented DS because their system design does not follow a task-oriented DS structure [39,40,42–44].

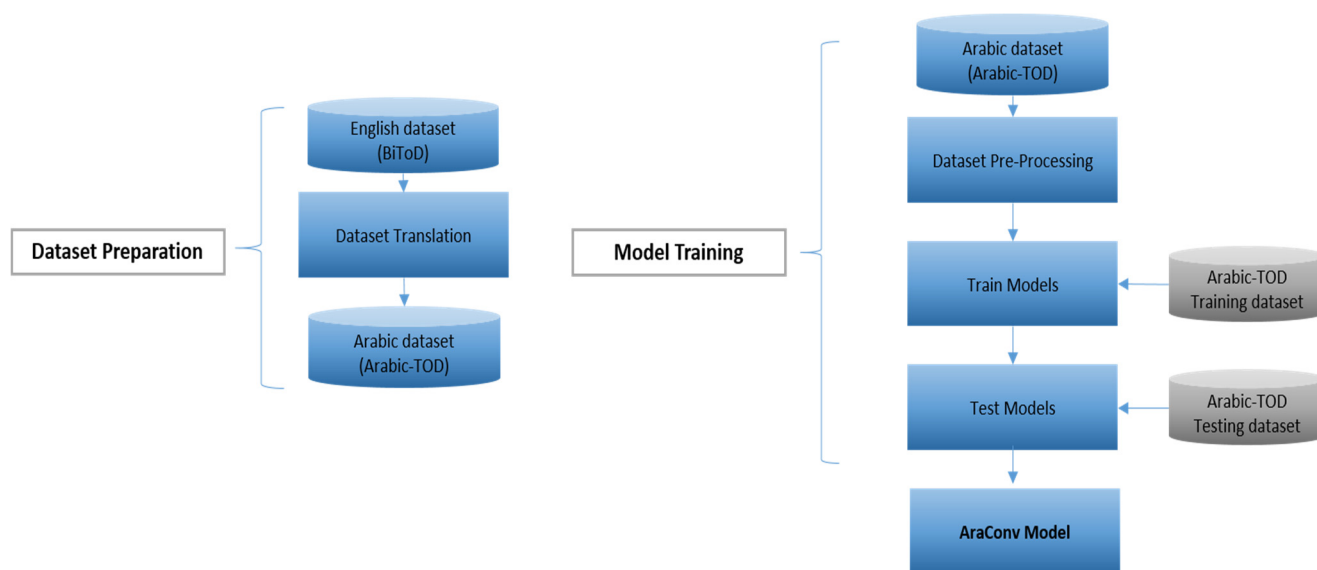
Notably, Bashir et al. [34] used deep learning approaches to build a natural language understanding module for Arabic task-oriented DS for home automation. The module manages of both intent classification and entity extraction tasks. For intent classification, it uses LSTMs and CNNs; for entity extraction, BiLSTM and character-based word embeddings are used. The study used data collected via an online survey and the AQMAR dataset. The data were filtered and labeled according to the Conll-2003 NER format. The findings for the intent classification demonstrated that CNNs performed better than LSTMs (F-score = 94%). For entity extraction, the model obtained comparable results to the named entity recognition benchmarks in English (F-score = 94%).

Meanwhile, Elmadany et al. [35] used a multi-class hierarchical model to solve the dialogue acts classification issue associated with Arabic dialects. They used a manually collected and annotated dataset from multi-genre Egyptian call centers to evaluate their system performance. Using an SVM classifier produced an average F-score of 91.2%, indicating an improvement of 20% compared to the state-of-art approach. Elsewhere, Joukhadar et al. [36] examined different machine learning approaches to recognizing user acts in a text-based DS for the Levantine Arabic dialect. They manually produced 873 sentences for both restaurant orders and flight booking, reporting accuracy of 86% using the SVM model. However, their small dataset was insufficient to build an efficient dialogue system, suggesting an imperative to develop large multi-domain datasets or more efficient techniques.

For Arabic user-based DS, several studies [37,38,41] have applied either pattern matching, rule-based, or rule-based and data-driven hybrid approaches to task-oriented DS. Nonetheless, it is apparent that most Arabic task-oriented DS use either rule-based or pattern matching approaches, with very few using a hybrid approach. It is understandable that they use these approaches due to the challenges associated with building Arabic task-oriented DS in Arabic [3], among which is the lack of Arabic task-oriented dialogue datasets. Therefore, this study aimed to address this challenge by leveraging the pre-trained language models to build an Arabic task-oriented DS. Multi-lingual language models are among the most popular and common language models, observed to produce good performance on task-oriented DS for many languages. Accordingly, we explored the extent to which mT5 can be useful for building an Arabic task-oriented dialogue system. To the best of our knowledge, this work represents the first attempt at pre-training a large transformer-based language representation model on an Arabic task-oriented dialogue dataset (Arabic-TOD).

### 3. Method

A pre-trained language model is a deep learning model that has been trained on a large amount of data to perform particular NLP tasks [45]. Figure 1 shows a high-level view of the approach adopted. We began with the English BiToD dataset [1], translating the dialogues into Arabic to produce the Arabic-TOD dataset. The dataset was then pre-processed and prepared for the training step. Subsequently, we trained the models on the training Arabic-TOD dataset using different settings. Finally, we used the testing Arabic-TOD dataset to test the models and obtain the results for the AraConv model.



**Figure 1.** High-level view of our approach.

### 3.1. Arabic Task-Oriented DS Dataset

Because Arabic is a low-resource language, no human-annotated Arabic dataset for task-oriented DS has been produced (to the best of our knowledge). To obtain a good-quality dataset, we decided to use an existing dataset, translating a benchmark dataset for task-oriented DS (BiToD [1]) to develop a suitable training dataset for Arabic task-oriented DS.

Translating existing datasets is a practice frequently observed in the literature for low-resource languages, with examples including [46–48]. Recent translation techniques for crowd-sourced annotated datasets have produced reasonable results on training data for different languages, enabling many studies to address the lack of datasets by translating existing datasets for many downstream tasks in NLP. For example, for question answering (QA), the SQuAD dataset has been translated into Arabic [46] and Bengali [47], and for conversation generation, the EmpatheticDialogues dataset has been translated into Arabic [48].

Still, it is imperative for the research community to develop multi-lingual benchmarks to evaluate the cross-lingual transferability of end-to-end systems in general and task-oriented DS in particular [49]. For task-oriented DS, many multi-lingual datasets can be obtained by translating the English datasets. Table 2 presents some of these alongside their corresponding tasks and domains. Translation represents a good choice for low-resource languages to support the reuse of resources and save time spent creating and annotating long dialogues. Additionally, this enables the development of multi-lingual benchmarks for the research community to use.

**Table 2.** Datasets translated from English within the field of task-oriented DS. EN: English, ES: Spanish, DE: German, IT: Italian, TH: Thai, VI: Vietnamese, ZH: Chinese.

Dataset	Task	Language	Domains
Chinese ATIS [50]	Intent classification Slot extraction	ZH	Flight bookings
Multi-lingual WOZ 2.0 [51]	DST	EN, DE, IT	Restaurant bookings
SLU-IT [52]	Intent classification Slot extraction	IT	7 domains (Restaurant, Weather, Music, . . . )
Almawave-SLU [53]	Intent classification Slot extraction	IT	7 domains (Restaurant, Weather, Music, . . . )



**Table 2.** *Cont.*

Dataset	Task	Language	Domains
S. Schuster et al. [30]	Task-oriented DS	ES, TH	3 domains (Weather, Alarm, and Reminder)
Z. Liu et al. [54]	Task-oriented DS	ES, TH	3 domains (Weather, Alarm, and Reminder)
Z. Liu et al. [31]	DST Task-oriented DS	EN, DE, IT ES, TH	Restaurant booking 3 domains (Weather, Alarm, and Reminder)
Vietnamese ATIS [55]	Intent classification Slot extraction	VI	Flight bookings

### 3.2. Structure and Organization of Arabic-TOD Dataset

The Arabic-TOD dataset is based on the BiToD dataset, the first large bilingual task-oriented dialogue dataset created for training and evaluating end-to-end task-oriented DS. It contains annotated English and Chinese dialogues and features a total of 7232 dialogues with 144,798 utterances (3689 dialogues in English and 3543 dialogues in Chinese). The dialogues range between 10 and more than 50 turns with an average length of 19.98 turns. Each turn can be defined as one or more utterances from one speaker [56]. The BiToD dataset includes dialogues in five domains: Hotels, Restaurants, Weather, Attractions, and Metro.

Although there are many other common multi-domain task-oriented dialogue datasets, including MultiWOZ, we chose to translate the BiToD dataset to leverage certain useful features that distinguished it from other datasets [1]. Notably, the BiToD dataset supports mixed-language contexts, also known as code-switching. Some items in the knowledge base (and in daily life) feature mixed-language information, meaning English and Arabic texts appear in the same utterance. For example, there are some restaurant names in English that cannot be translated into Arabic, such as Chom Chom, which maintains the English name even if our conversation is in Arabic (i.e., “*هل يمكنك أن تحجز لي مطعم* Chom Chom”). Another advantageous feature of the BiToD dataset is its use of a deterministic API, which simplifies model evaluations. Deterministic API refers to the ability of the system to recommend the query-matched items on the basis of certain criteria (e.g., user rating). This differs from other API evaluation methods, which randomly return only one or two matched items with the API. Another important aspect of the BiToD dataset is the diversity of user tasks, meaning users might want to book hotels and restaurants within the same dialogue, as they might in a real human-based interactions. As such, we decided to contribute to enriching and augmenting the BiToD dataset by translating the English dialogues into Arabic, producing a multi-lingual dataset enabling the combined use of English, Chinese, and Arabic. Table 3 summarizes the different common multi-domain task-oriented dialogue datasets, indicating the features that we have tried to utilize.

**Table 3.** Summary of the characteristics of different common task-oriented dialogue datasets.

Dataset	Languages	Number of Dialogues	Avg. Turn Length	Number of Domains (Tasks)	Deterministic API	Mixed-Language Context
BiToD	EN, ZH	7232	19.98	5	Yes	Yes
MultiWoZ	EN	8438	13.46	7	No	No
Askmaster	EN	13,215	22.9	6	No	No
MetaLWOZ	EN	37,884	11.4	47	No	No

Table 3. Cont.

Dataset	Languages	Number of Dialogues	Avg. Turn Length	Number of Domains (Tasks)	Deterministic API	Mixed-Language Context
TM-1	EN	13,215	21.99	6	No	No
Schema	EN	22,825	20.3	17	No	No
SGD	EN	16,142	20.44	16	No	No
STAR	EN	5820	21.71	13	No	No
Frames	EN	1369	14.6	3	No	No
Multi-lingual WOZ 2.0	EN, DE, IT	3600	–	1	No	Yes
Arabic-TOD	AR	1500	19.98	4	Yes	Yes

For the translation task, three bilingual speakers of Arabic and English were paid to manually translate the English BiToD dataset into Arabic over 2.5 months, translating the utterances and slot-values in the dataset in the Hotels, Restaurants, Weather, and Attractions categories. We determined the strategy of translation and the used lexicons previously, and we gave them some examples of the target translated dialogues. Of the 3689 English dialogues, 1500 dialogues (30,000 utterances) were translated into Arabic. The translated utterances and slot-values were reviewed to verify the quality of translation and correctness of slot-value pairs on the basis of the English BiToD dataset.

Arabic-TOD dataset contains different lengths of dialogues, some of them with a single task and the others with multiple tasks varying between 2 and 4. For instance, some dialogues include multiple tasks in a single dialogue (e.g., a single dialogue can involve different tasks including enquiring about the weather, finding a restaurant to eat at, and an attraction to visit).

To the best of our knowledge, this Arabic-TOD is the first Arabic dataset supporting a mixed languages context for task-oriented DS that has been annotated following the BiToD dataset's structure [1].

### 3.3. Model Architecture

The AraConv model's generation process is based on a single multi-lingual Seq2Seq (mSeq2Seq) model that uses the pre-trained model mT5 [5], a multilingual variant of T5 [57], which can be formally defined as follows:

Assume the dialogue  $D$  represents a set of user utterances ( $U_t$ ) and system utterances ( $S_t$ ) at turn  $t$ , where  $D = \{U_1, S_1, \dots, U_t, S_t\}$ .

The dialogue history ( $H$ ) holds the previous user and system utterances of turn  $t$ , specified by the context window size ( $w$ ), where  $H_t = \{U_{t-w}, S_{t-w}, \dots, S_{t-1}; U_t\}$ . For turn  $t$ , the dialogue state is represented as  $B_t$ , and the knowledge state is represented as  $K_t$ .

Figure 2 illustrates the proposed workflow for response generation using the mSeq2Seq model based on the BiToD dataset [1].

Initially, we set the dialogue state and knowledge state to empty strings as  $B_0$  and  $K_0$ . Then, we considered the current dialogue history ( $H_t$ ), previous dialogue state ( $B_{t-1}$ ), and previous knowledge state ( $K_{t-1}$ ) as input at turn  $t$ . We added the prompt  $PB = \text{"TrackDialogueState:"}$  to indicate the generation task [57]. Therefore, the mSeq2Seq model produces Levenshtein Belief Spans at turn  $t$  ( $Lev_t$ ), indicating a text span that contains the information for updating the dialogue state from ( $B_{t-1}$ ) to  $B_t$ .  $Lev_t$  can be represented by the following equation:

$$Lev_t = \text{mSeq2Seq}(PB, H_t, B_{t-1}, K_{t-1}) \quad (1)$$

Then, the model generates an output (o/p) based on the new input as the updated dialogue state ( $B_t$ ), and the response generation prompt—referred to as PR = “Response:”—at the current turn  $t$ . If there is a need for an API call, the model will generate an API name according to the following:

$$API = mSeq2Seq(PR, H_t, B_t, K_{t-1}) \tag{2}$$

In this case, the system queries the API with particular constraints in the dialogue state and updates the knowledge state form ( $K_{t-1}$ ) to ( $K_t$ ). The updated knowledge state ( $K_t$ ) and API name (API) are subsequently combined to generate the next turn response. Otherwise, the model generates a textual response (R) that is returned directly to the user:

$$R = mSeq2Seq(PR, H_t, B_t, K_t, API) \tag{3}$$

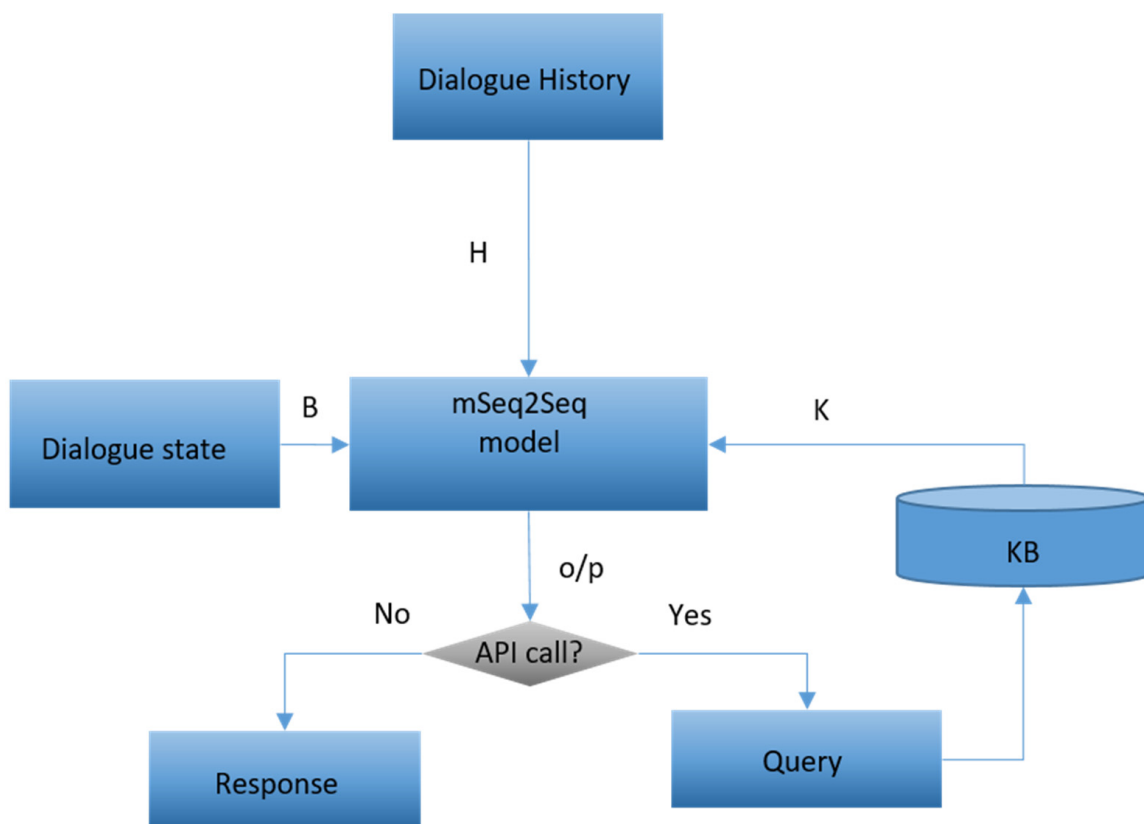


Figure 2. The multi-lingual Seq2Seq model workflow.

#### 4. Experiments

This section first explains the evaluation metrics used to measure the performance of the AraConv model. Next, we describe the experimental setup and detail the experiments performed to test our hypothesis. Finally, we discuss the results of each experiment.

##### 4.1. Evaluation Metrics

This study addresses two main tasks: DST and end-to-end dialogue generation, which includes both DST and response. To evaluate the DST performance of the AraConv model, we used the joint goal accuracy (JGA) metric to compare the predicted dialogue state to the ground truth for each dialogue turn. If all predicted slot values exactly match the ground-truth values, the model’s output is considered correct. To evaluate the performance on the end-to-end generation task by the AraConv model, we used four metrics:



- the BLEU metric to assess the generated response fluency;
- the API call accuracy ( $API_{Acc}$ ) metric to assess if the system generates the correct API call;
- the task success rate (TSR) metric to assess whether the system finds the correct entity and provides all of the requested information for a particular task. TSR can be defined as

$$TSR = \frac{\sum \text{success task}}{\text{total number of tasks}} \quad (4)$$

where the tasks involve searching task and booking task for hotel and restaurant domains, and search task for attraction and weather domains.

- the dialogue success rate (DSR) metric to evaluate whether the system accomplishes all of the dialogue tasks. DSR can be defined as

$$DSR = \frac{\sum \text{success dialogue}}{\text{total number of dialogues}} \quad (5)$$

The evaluation method's main goal is to obtain an automated and repeatable evaluation procedure that enables efficient comparisons of the quality of different dialogue strategies. This involves focusing on the automatic evaluation metrics. However, further measurement of the quality of the generated responses also requires human review. Thus, following the literature [15], we evaluated the AraConv model's performance on end-to-end generation tasks according to two metrics:

- the language understanding score to indicate the extent to which the system understands user inputs; and
- the response appropriateness score to indicate whether the response is appropriate and human-like.

We performed a small-scale human review to measure these scores. The literature indicates two other common metrics used in human evaluation: TSR and DSR [56]. Given the costs and time-intensiveness of human evaluation, we measured these scores automatically (TSR and DSR).

#### 4.2. Experimental Setup

Our framework uses the pre-trained multi-lingual model mT5-small. All of our experiments used the Transformers library [58] and the deep learning framework PyTorch [59]. We trained all of the models using an AdamW optimizer [60] (with an initial learning rate of 0.0005). We set the dialogue context window size ( $w$ ) at 2 and the batch size at 128 in accordance with the approach observed to obtain the best results in the extant literature.

We split our Arabic-TOD dataset into 67%, 7%, and 26% for training, validation, and testing, resulting in 1000, 100, and 400 training, validation, and testing dialogues, respectively. For the mono-lingual setting, we trained the model for 20 epochs; for the bi-lingual and multi-lingual settings, we trained the models for 8 epochs. Training using Google Colab required approximately 22 hours.

#### 4.3. Baseline

As this is the first work to build an Arabic end-to-end generative model for task-oriented DS, there is no directly comparable approach in the previous Arabic studies. Therefore, we experimented with several initial baselines (using the zero-shot setting, that is transferring the model, which is trained to solve task-oriented DS, in English to solve that specific task in Arabic). We trained the mT5 model on English using the English BiToD dataset then tested its performance directly on the Arabic-TOD dataset. This approach is a common practice similar to many downstream tasks such as QA [61,62] or task-oriented DS [63]. The performance of these initial baselines was very low; therefore, we set our baseline using the same concept of zero-shot setting where mT5 model is trained on mixed

language training data by replacing the most task-related keyword entities in English BiToD language with their corresponding in Arabic language from a parallel dictionary.

#### 4.4. Experiments

*RQ1: To what extent can mT5, a multi-lingual pre-trained language model, produce satisfactory results for Arabic end-to-end task-oriented DS?*

This experiment aimed to investigate the performance of an end-to-end Arabic task-oriented dialogue system using an mSeq2Seq model for Arabic. This mono-lingual setting only requires one language to train and test the model. Thus, we trained and tested the proposed mT5 model (AraConv) using the Arabic-TOD dataset. The AraConv model differs from the baseline with the training setting where AraConv trained on Arabic dialogues while the baseline did not (zero-shot learning).

Table 4 shows the results—in terms of BLEU, APIACC, TSR, DSR, and JGA—of the AraConv model in the mono-lingual setting in comparison to the English and Chinese experiments on the BiToD dataset [1]. The observed English results [1] outperformed the AraConv results. This is unsurprising because there are more data for English and Chinese. The model trained and tested on English or Chinese data still performed better than that tested on the Arabic-TOD dataset, which represented only 27% of the BiToD dataset [1]. Where the original mT5 model was trained using multiple languages, the English data represented 5.67% of the whole corpus, and Chinese and Arabic represented 1.67% and 1.66% of the total data, respectively [5], explaining the superior performance for English dialogue. Additionally, Arabic is a language with extensive grammatical case marking [5], which causes lower evaluation metrics compared to English. Meanwhile, despite the comparable sizes of the training data for Arabic and Chinese, the results of the mono-lingual model trained on Chinese BiToD dataset outperformed the AraConv model. This may have been due to the small size of the Arabic-TOD dataset compared to the Chinese BiToD dataset. Nonetheless, the AraConv model achieved a better BLEU value (by approximately 63%) than the Chinese model, meaning that the AraConv model can generate more fluent responses than the Chinese model.

**Table 4.** Mono-lingual experiment dialogue state tracking (DST) and end-to-end dialogue generation results for the AraConv model trained on Arabic-TOD dataset compared to the baseline and the mono-lingual BiToD experiments [1] using English (EN) and Chinese (ZH). Bold numbers indicating the best result according to the column's metric value.

	TSR	DSR	APIAcc	BLEU	JGA
<b>Arabic</b>					
Baseline	3.95	1.16	4.30	3.37	8.21
AraConv	45.07	18.60	48.86	31.05	34.82
<b>Other languages</b>					
EN [1]	<b>69.13</b>	<b>47.51</b>	<b>67.92</b>	<b>38.48</b>	<b>69.19</b>
ZH [1]	53.77	31.09	63.25	19.03	67.35

Still, the AraConv model did not achieve perfect results, potentially due to the complicated nature of the Arabic-TOD dataset, its complex ontology, and its diversity of user goals. Moreover, the DSR result was lower than the TSR result, likely because of the multiple tasks included in the dialogue (2–4 tasks). For instance, some dialogues included multiple tasks in a single dialogue (e.g., a single dialogue can involve the tasks of finding a hotel to stay at, a restaurant to eat at, an attraction to visit, and information about the weather).

*RQ2: To what extent can joint-training the mT5 model on Arabic dialogue data and data for one or two high-resource languages (namely, English or English and Chinese) improve the performance of Arabic task-oriented DS?*

Answering this research question requires performing two experiments to investigate the performance of building an end-to-end Arabic task-oriented dialogue system using an mSeq2Seq model in bi-lingual and multi-lingual settings. Because two languages are used to train and test the model in the bi-lingual setting, we trained the proposed model mT5 on both the Arabic-TOD and English-BiToD datasets [1].

In the experiments described in [1], the models were trained on almost the same number of English and Chinese dialogues (2952 and 2835). However, our Arabic-TOD dataset includes only 27% of the data included in the BiToD datasets. Accordingly, we investigated two cases:

- Non-equivalent (NQ): The size of the Arabic-TOD dataset is not equal to the English BiToD dataset. We trained the model with 1000 Arabic dialogues and 2952 English dialogues.
- Equivalent (Q): The size of the Arabic-TOD dataset and the English BiToD dataset are equal (1000 dialogues for training).

Because three languages were used to train and test the model for the multi-lingual experiment, the mT5 model was trained on the Arabic-TOD, the English BiToD, and the Chinese BiToD datasets [1]. As in the previous experiment, we investigated two cases:

- Non-equivalent (NQ): The size of the Arabic-TOD dataset is not equal to the English or Chinese BiToD dataset. We trained the model with 1000 Arabic dialogues, 2952 English dialogues, and 2835 Chinese dialogues.
- Equivalent (Q): The size of the Arabic-TOD dataset, the English BiToD dataset, and the Chinese BiToD are equal (1000 dialogues for training).

For the bi-lingual setting, Table 5 compares the AraConv results—in terms of BLEU, APIACC, TSR, DSR, and JGA—to the experiments reported in [1] regarding English and Chinese dialogues with the same settings. We observed that the non-equipollent bi-lingual AraConv model (AraConv<sub>Bi-NQ</sub>) outperformed the equipollent bi-lingual AraConv model (AraConv<sub>Bi-Q</sub>), demonstrating the impact of training dialogue dataset size on the final model given that the AraConv<sub>Bi-NQ</sub> model is trained on more data. Therefore, using more English data in training with Arabic helps to improve the result because of the semantics of the conversation, which is almost similar to Arabic, especially for the task-related words. However the model in [1], which was trained on both English and Chinese data and then tested on English, outperformed all models, assuming the dialogues in the two datasets were almost the same. As discussed, the distinguished performance of the English model could have been due to the amount of English data used to train the mT5 model. Nonetheless, we observed that the AraConv model performed better according to the BLEU metric than the Chinese model, despite training on the same English dataset (as a second dataset for joint-training), confirming the greater fluency of the AraConv model.

For the multi-lingual setting, Table 6 presents AraConv results calculated in terms of BLEU, APIACC, TSR, DSR, and JGA. Our findings emphasize the previous results of AraConv in the bi-lingual experiment, which saw the non-equipollent multi-lingual AraConv model (AraConv<sub>M-NQ</sub>) perform better than the equipollent multi-lingual AraConv model (AraConv<sub>M-Q</sub>). Accordingly, we recognize that joint-training on multiple languages including the target language (in this case, Arabic) improves the results in experiments on the target language, which aligns with the results reported in [30].

**Table 5.** Bi-lingual experiment DST and end-to-end dialogue generation results for the AraConv model trained on the Arabic-TOD dataset compared to the bi-lingual BiToD experiments [1] using English and Chinese BiToD datasets. The bold letters refer to the target language in the corresponding experiments (used to test the model). Bold numbers indicating the best result according to the column’s metric value.

	TSR	DSR	APIAcc	BLEU	JGA
<b>Arabic</b>					
AraConvBi-NQ ( <b>AR</b> , EN)	45.57	21.90	56.23	30.41	37.35
AraConvBi-Q ( <b>AR</b> , EN)	44.62	16.98	46.32	27.36	35.58
<b>Other languages</b>					
ZH, EN [1]	<b>71.18</b>	<b>51.13</b>	<b>71.87</b>	<b>40.71</b>	<b>72.16</b>
ZH, EN [1]	57.24	34.78	65.54	22.45	68.70

**Table 6.** Multi-lingual experiment DST and end-to-end dialogue generation results for the AraConv model on the Arabic-TOD dataset. The bold letters refer to the target language. Bold numbers indicating the best result according to the column’s metric value.

	TSR	DSR	APIAcc	BLEU	JGA
AraConvM-NQ ( <b>AR</b> , EN, ZH)	<b>51.27</b>	<b>20.00</b>	<b>55.44</b>	<b>32.58</b>	<b>37.68</b>
AraConvM-Q ( <b>AR</b> , EN, ZH)	47.17	16.98	53.07	31.05	36.13

Generally, for bi-lingual and multi-lingual experiments, the trained models can simultaneously handle dialogues in multiple languages (whether English, Chinese, or Arabic) without using any of the language identifiers supplied during testing.

For the human review, we aimed to rate dialogue or utterances on the basis of certain metrics identified in the literature [56]. Five expert researchers (who are independent from this paper author) were chosen for this task. We randomly selected 20 complete dialogue sessions from the generated dialogues of AraConv model. The researchers were asked to rate these dialogues by providing language understanding and response appropriateness scores. Their scores ranged from 0 (extremely bad) to 5 (extremely good), depending on the system’s response. Subsequently, we evaluated the reliability of their rating using Fleiss’ Kappa [64]. The overall Fleiss’ kappa values for the language understanding and appropriateness scores were 0.253 and 0.229, respectively, indicating “fair agreement”.

## 5. Conclusions and Future Work

To the best of our knowledge, this work represents to the first attempt to build an end-to-end Arabic task-oriented dialogue system (AraConv) using a pre-trained transformer-based multi-lingual language model. We utilized the highly regarded multi-lingual model mT5 to build an end-to-end Arabic task-oriented dialogue system with different settings and presented an Arabic-TOD dataset based on translating 27% of the BiToD dataset’s English dialogue data into Arabic. The Arabic-TOD dataset is considered the first dialogue dataset for the Arabic task-oriented DS that supports code-switching. Although using the Arabic-TOD dataset to train and test the model in a mono-lingual setting demonstrates a reasonable performance for the AraConv model compared to the results observed for the English and Chinese BiToD datasets in the same settings, the performance is undermined by the small size of the Arabic TOD dataset. To overcome this problem, we considered joint-training the model on Arabic dialogue data and one or two high-resource languages (English or both English and Chinese). The findings reveal that the AraConv model in the multi-lingual setting outperformed the AraConv model in the mono-lingual setting, with multi-lingual training with English, Chinese, and Arabic observed to be better than bi-lingual training with only English and Arabic data. Thus, the AraConv model can be

considered a good baseline for building robust end-to-end Arabic task-oriented DS that can engage with complex scenarios.

The main limitation of this work is the small size of the Arabic-TOD dataset. A related limitation concerns the Arabic-TOD dataset using non-Arabic entities, with the dataset code-switching due to entities in the original BiToD dataset. However, we leveraged this property to align the model with the routine usage of such entities in conversation. In the future, we aim to extend the Arabic-TOD dataset to equal the BiToD dataset in terms of the number of dialogues. Additionally, we plan to examine cross-lingual models, especially involving the Arabic-TOD dataset. Furthermore, we plan to develop Arabic task-oriented DS using other multilingual language models (e.g., mBART [65]). Another possible venue for future work is using a pre-trained Arabic model for Arabic task-oriented DS such as AraT5 [66], which was yet to be deployed at the time of working on this paper.

**Author Contributions:** Conceptualization, A.F. and M.A.-Y.; methodology, A.F.; software, A.F.; validation, A.F.; formal analysis, A.F. and M.A.-Y.; investigation, A.F. and M.A.-Y.; resources, A.F. and M.A.-Y.; data curation, A.F.; writing (original draft preparation), A.F.; writing (review and editing), A.F. and M.A.-Y.; visualization, A.F. and M.A.-Y.; supervision, M.A.-Y.; project administration, M.A.-Y.; funding acquisition, M.A.-Y. All authors have read and agreed to the published version of the manuscript.

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

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Upon request.

**Acknowledgments:** The authors extend their appreciation to the Researchers Supporting Project number RSP-2021/286, King Saud University, Riyadh, Saudi Arabia.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Lin, Z.; Madotto, A.; Winata, G.I.; Xu, P.; Jiang, F.; Hu, Y.; Shi, C.; Fung, P. BiToD: A Bilingual Multi-Domain Dataset For Task-Oriented Dialogue Modeling. *arXiv* **2021**, arXiv:2106.02787.
2. Huang, M.; Zhu, X.; Gao, J. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.* **2019**, *38*, 1–32. [[CrossRef](#)]
3. AlHagbani, E.S.; Khan, M.B. Challenges facing the development of the Arabic chatbot. *First Int. Work. Pattern Recognit. Int. Soc. Opt. Photonics* **2016**, *10011*, 7. [[CrossRef](#)]
4. Darwish, K.; Habash, N.; Abbas, M.; Al-Khalifa, H.; Al-Natsheh, H.T.; Bouamor, H.; Bouzoubaa, K.; Cavalli-Sforza, V.; El-Beltagy, S.R.; El-Hajj, W.; et al. A Panoramic Survey of Natural Language Processing in the Arab World. *Commun. ACM* **2021**, *64*, 72–81. [[CrossRef](#)]
5. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou', R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, online, 15–20 June 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 483–498. [[CrossRef](#)]
6. McTear, M. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*; Morgan & Claypool Publishers LLC: San Rafael, CA, USA, 2020; Volume 13.
7. Qin, L.; Xu, X.; Che, W.; Zhang, Y.; Liu, T. Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, online, 5–10 July 2020; pp. 6344–6354. [[CrossRef](#)]
8. Lei, W.; Jin, X.; Ren, Z.; He, X.; Kan, M.Y.; Yin, D. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 1437–1447. [[CrossRef](#)]
9. Budzianowski, P.; Vulić, I. Hello, It's GPT-2-How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019), Hong Kong, China, 4 November 2019; pp. 15–22. [[CrossRef](#)]



10. Ham, D.; Lee, J.-G.; Jang, Y.; Kim, K.-E. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; Volume 2, pp. 583–592. [[CrossRef](#)]
11. Hosseini-Asl, E.; McCann, B.; Wu, C.S.; Yavuz, S.; Socher, R. A simple language model for task-oriented dialogue. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 20179–20191.
12. Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; Gao, J. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 807–824. [[CrossRef](#)]
13. Yang, Y.; Li, Y.; Quan, X. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. *arXiv* **2020**, arXiv:2012.03539.
14. Lin, Z.; Madotto, A.; Winata, G.I.; Fung, P. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 3391–3405. [[CrossRef](#)]
15. Wang, W.; Zhang, Z.; Guo, J.; Dai, Y.; Chen, B.; Luo, W. Task-Oriented Dialogue System as Natural Language Generation. *arXiv* **2021**, arXiv:2108.13679.
16. Peng, B.; Zhu, C.; Li, C.; Li, X.; Li, J.; Zeng, M.; Gao, J. Few-shot Natural Language Generation for Task-Oriented Dialog. *arXiv* **2020**, arXiv:2002.12328.
17. Wu, C.-S.; Hoi, S.C.H.; Socher, R.; Xiong, C. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 917–929. [[CrossRef](#)]
18. Madotto, A.; Liu, Z.; Lin, Z.; Fung, P. Language Models as Few-Shot Learner for Task-Oriented Dialogue Systems. *arXiv* **2020**, arXiv:2008.06239.
19. Campagna, G.; Foryciarz, A.; Moradshahi, M.; Lam, M. Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking. *arXiv* **2020**, arXiv:2005.00891.
20. Zhang, Y.; Ou, Z.; Hu, M.; Feng, J. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 9207–9219. [[CrossRef](#)]
21. Kulhánek, J.; Hudeček, V.; Nekvinda, T.; Dušek, O. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation. *arXiv* **2021**, arXiv:2102.05126.
22. Gao, S.; Takanobu, R.; Peng, W.; Liu, Q.; Huang, M. HyKnow: End-to-End Task-Oriented Dialog Modeling with Hybrid Knowledge Management. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1591–1602. [[CrossRef](#)]
23. Lee, H.; Lee, J.; Kim, T.Y. SUMBT: Slot-utterance matching for universal and scalable belief tracking. *arXiv* **2019**, arXiv:1907.07421.
24. Chao, G.L.; Lane, I. BERT-DST: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *Proc. Annu. Conf. Int. Speech Commun. Assoc. Interspeech* **2019**, *2019*, 1468–1472. [[CrossRef](#)]
25. Kim, S.; Yang, S.; Kim, G.; Lee, S.-W. Efficient Dialogue State Tracking by Selectively Overwriting Memory. *arXiv* **2020**, arXiv:1911.03906. [[CrossRef](#)]
26. Kumar, A.; Ku, P.; Goyal, A.; Metallinou, A.; Hakkani-Tur, D. MA-DST: Multi-Attention-Based Scalable Dialog State Tracking. *Proc. Conf. AAAI Artif. Intell.* **2020**, *34*, 8107–8114. [[CrossRef](#)]
27. Heck, M.; van Niekerk, C.; Lubis, N.; Geishauser, C.; Lin, H.-C.; Moresi, M.; Gašić, M. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. *arXiv* **2020**, arXiv:2005.02877.
28. Li, S.; Yavuz, S.; Hashimoto, K.; Li, J.; Niu, T.; Rajani, N.; Yan, X.; Zhou, Y.; Xiong, C. CoCo: Controllable Counterfactuals for Evaluating Dialogue State Trackers. *arXiv* **2020**, arXiv:2010.12850.
29. Wang, D.; Lin, C.; Liu, Q.; Wong, K.-F. Fast and Scalable Dialogue State Tracking with Explicit Modular Decomposition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 6–11 June 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 289–295. [[CrossRef](#)]
30. Schuster, S.; Shah, R.; Gupta, S.; Lewis, M. Cross-lingual transfer learning for multilingual task oriented dialog. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 3795–3805. [[CrossRef](#)]
31. Liu, Z.; Winata, G.I.; Lin, Z.; Xu, P.; Fung, P. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8433–8440. [[CrossRef](#)]
32. Zhang, Y.; Ou, Z.; Yu, Z. Task-oriented dialog systems that consider multiple appropriate responses under the same context. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 9604–9611. [[CrossRef](#)]
33. Wang, K.; Tian, J.; Wang, R.; Quan, X.; Yu, J. Multi-Domain Dialogue Acts and Response Co-Generation. *arXiv* **2020**, arXiv:2004.12363.
34. Bashir, A.M.; Hassan, A.; Rosman, B.; Duma, D.; Ahmed, M. Implementation of A Neural Natural Language Understanding Component for Arabic Dialogue Systems. *Procedia Comput. Sci.* **2018**, *142*, 222–229. [[CrossRef](#)]



35. Elmadany, A.R.A.; Abdou, S.M.; Gheith, M. Improving dialogue act classification for spontaneous Arabic speech and instant messages at utterance level. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 7–12 May 2018; pp. 128–134.
36. Joukhadar, A.; Saghergy, H.; Kweider, L.; Ghneim, N. Arabic Dialogue Act Recognition for Textual Chatbot Systems. In Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019), Trento, Italy, 11–12 September 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 43–49.
37. Al-Ajmi, A.H.; Al-Twairsh, N. Building an Arabic Flight Booking Dialogue System Using a Hybrid Rule-Based and Data Driven Approach. *IEEE Access* **2021**, *9*, 7043–7053. [[CrossRef](#)]
38. Hijjawi, M.; Bandar, Z.; Crockett, K.; McLean, D. ArabChat: An arabic conversational agent. In Proceedings of the 2014 6th International Conference on Computer Science and Information Technology, CSIT 2014-Proceedings, Amman, Jordan, 26 March 2014; pp. 227–237. [[CrossRef](#)]
39. Almutadha, Y. LABEEB: Intelligent Conversational Agent Approach to Enhance Course Teaching and Allied Learning Outcomes attainment. *J. Appl. Comput. Sci. Math.* **2019**, *13*, 9–12. [[CrossRef](#)]
40. Aljameel, S.; O’shea, J.; Crockett, K.; Latham, A.; Kaleem, M. LANA-I: An Arabic Conversational Intelligent Tutoring System for Children with ASD. *Adv. Intell. Syst. Comput.* **2019**, *997*, 498–516. [[CrossRef](#)]
41. Moubaidin, A.; Shalbak, O.; Hammo, B.; Obeid, N. Arabic dialogue system for hotel reservation based on natural language processing techniques. *Comput. Syst.* **2015**, *19*, 119–134. [[CrossRef](#)]
42. Bendjamaa, F.; Nora, T. A Dialogue-System Using a Qur’anic Ontology. In Proceedings of the 2020 Second International Conference on Embedded & Distributed Systems (EDiS), Oran, Algeria, 3 November 2020; pp. 167–171.
43. Fadhil, A.; AbuRa’Ed, A. Ollobot-Towards a text-based Arabic health conversational agent: Evaluation and results. In Proceedings of the Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria, 2–4 September 2019; pp. 295–303. [[CrossRef](#)]
44. Al-Ghadhban, N.; Al-Twairsh, D. Nabiha: An Arabic dialect chatbot. *Int. J. Adv. Comput. Sci. Appl. Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 452–459. [[CrossRef](#)]
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5999–6009.
46. Mozannar, H.; Maamary, E.; el Hajal, K.; Hajj, H. Neural Arabic Question Answering. *arXiv* **2019**, arXiv:1906.05394.
47. Mayeesha, T.T.; Sarwar, A.M.; Rahman, R.M. Deep learning based question answering system in Bengali. *J. Inf. Telecommun.* **2020**, *5*, 145–178. [[CrossRef](#)]
48. Naous, T.; Hokayem, C.; Hajj, H. Empathy-driven Arabic Conversational Chatbot. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, Barcelona, Spain, 8 December 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 58–68.
49. Razumovskaia, E.; Glavaš, G.; Majewska, O.; Ponti, E.M.; Korhonen, A.; Vulić, I. Crossing the Conversational Chasm: A Primer on Natural Language Processing for Multilingual Task-Oriented Dialogue Systems. *arXiv* **2021**, arXiv:2104.08570.
50. He, X.; Deng, L.; Hakkani-Tur, D.; Tur, G. Multi-style adaptive training for robust cross-lingual spoken language understanding. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8342–8346.
51. Mrkšić, N.; Vulić, I.; Séaghdha, D.; Leviant, I.; Reichart, R.; Gašić, M.; Korhonen, A.; Young, S. Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 309–324. [[CrossRef](#)]
52. Castellucci, G.; Bellomaria, V.; Favalli, A.; Romagnoli, R. Multi-lingual Intent Detection and Slot Filling in a Joint BERT-based Model. *arXiv* **2019**, arXiv:1907.02884.
53. Bellomaria, V.; Castellucci, G.; Favalli, A.; Romagnoli, R. Almwave-SLU: A new dataset for SLU in Italian. *arXiv* **2019**, arXiv:1907.07526.
54. Liu, Z.; Shin, J.; Xu, Y.; Winata, G.I.; Xu, P.; Madotto, A.; Fung, P. Zero-shot cross-lingual dialogue systems with transferable latent variables. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 11 November 2019; pp. 1297–1303. [[CrossRef](#)]
55. Dao, M.H.; Truong, T.H.; Nguyen, D.Q. Intent Detection and Slot Filling for Vietnamese. *arXiv* **2021**, arXiv:2104.02021.
56. Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; Cieliebak, M. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.* **2020**, *54*, 755–810. [[CrossRef](#)]
57. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
58. Huggingface/Transformers? Transformers: State-of-the-Art Natural Language Processing for Pytorch, TensorFlow, and JAX. Available online: <https://github.com/huggingface/transformers> (accessed on 17 November 2021).
59. PyTorch. Available online: <https://pytorch.org/> (accessed on 17 November 2021).
60. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
61. Siblini, W.; Pasqual, C.; Lavielle, A.; Challal, M.; Cauchois, C. Multilingual Question Answering from Formatted Text applied to Conversational Agents. *arXiv* **2019**, arXiv:1910.04659.

62. Hsu, T.Y.; Liu, C.L.; Lee, H.Y. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5933–5940. [[CrossRef](#)]
63. Upadhyay, S.; Faruqui, M.; Tür, G.; Dilek, H.-T.; Heck, L. (Almost) Zero-shot cross-lingual spoken language understanding. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6034–6038.
64. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378–382. [[CrossRef](#)]
65. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual Denoising Pre-training for Neural Machine Translation. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 726–742. [[CrossRef](#)]
66. Nagoudi, E.M.B.; Elmadany, A.; Abdul-Mageed, M. AraT5: Text-to-Text Transformers for Arabic Language Understanding and Generation. *arXiv* **2021**, arXiv:2109.12068.