mathematics

MDPI

Article

Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual 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 or 439204463@student.ksu.edu.sa

Abstract: Due to the promising performance of pre-trained language models for task-oriented dialogue systems (DS) in English, some efforts to provide multilingual models for task-oriented DS in low-resource languages have emerged. These efforts still face a long-standing challenge due to the lack of high-quality data for these languages, especially Arabic. To circumvent the cost and time-intensive data collection and annotation, cross-lingual transfer learning can be used when few training data are available in the low-resource target language. Therefore, this study aims to explore the effectiveness of cross-lingual transfer learning in building an end-to-end Arabic task-oriented DS using the mT5 transformer model. We use the Arabic task-oriented dialogue dataset (Arabic-TOD) in the training and testing of the model. We present the cross-lingual transfer learning deployed with three different approaches: mSeq2Seq, Cross-lingual Pre-training (CPT), and Mixed-Language Pre-training (MLT). We obtain good results for our model compared to the literature for Chinese language using the same settings. Furthermore, cross-lingual transfer learning deployed with the MLT approach outperform the other two approaches. Finally, we show that our results can be improved by increasing the training dataset size.

Keywords: cross-lingual transfer learning; task-oriented dialogue systems; Arabic language; mixed-language pre-training; multilingual transformer model; mT5; natural language processing

MSC: 68T50

Citation: Fuad, A.; Al-Yahya, M. Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual Transformer Model mT5. *Mathematics* **2022**, *10*, 746. https://doi.org/10.3390/math10050746

Academic Editors: Cornelia Caragea and Florentina Hristea

Received: 26 January 2022
Accepted: 24 February 2022
Published: 26 February 2022

1. Introduction

Task-oriented dialogue systems (DS) are systems that help users to achieve a vast range of tasks using natural language conversations, such as restaurant reservations, flight bookings, and weather-forecast inquiries [1]. The demand on such systems has increased rapidly, due to the promising performance of English-based conversational DS [2–5]. However, most of these systems are unable to support various low-resource languages because of the lack of high quality annotated data for these languages. Arabic is one of the low-resource languages that is suffering from a lack of task-oriented dialogue datasets, and as such it is lagging behind DS development [6]. One of the common strategies to address this issue is to collect and annotate more data for every language, which is a cost-intensive process. Therefore, in order to address the lack of dialogue datasets, cross-lingual transfer learning is used. In this type of learning, knowledge of high-resource languages (such as English) is transferred into low-resource languages (such as Arabic) with few training data in the low-resource target language [7]. Cross-lingual transfer learning has shown good results for several tasks using pre-trained multilingual models [8,9].

In this work, we aimed to explore the effectiveness of cross-lingual transfer learning in building end-to-end Arabic task-oriented DS using the multilingual pre-trained language model mT5 [10]. These end-to-end systems must handle the tasks of tracking all entities in the dialogue, as well as generating suitable responses. To the best of our knowledge, this is

the first work to examine the cross-lingual transfer ability of mT5 on Arabic task-oriented DS in few-shot scenarios. We aimed to answer two research questions:

- To what extent is cross-lingual transfer learning effective in end-to-end Arabic task-oriented DS using the mT5 model?
- To what extent does the size of the training dataset affect the quality of Arabic task-oriented DS in few-shot scenarios?

The rest of the paper is organized as follows: Related work in the area of task-oriented DS for multilingual models and cross-lingual transfer learning is explored in Section 2. Section 3 delineates the methodology of cross-lingual transfer learning used in this research. In Section 4, we present our experimental setup and the dataset used. The results and findings are discussed in Section 5. Finally, in Section 6, we summarize the research and highlight avenues for future research.

2. Related Works

In task-oriented DS, high-resource languages are those that have many dataset samples, whereas in low-resource languages there are few dataset samples. Therefore, it is important to provide datasets for these languages to advance research on end-to-end task-oriented DS. Cross-lingual transfer learning is a common and effective approach to build end-to-end task-oriented DS in low-resource languages [8,11,12]. Machine-translation and multilingual representations are the most used approaches in cross-lingual research. The authors of [8] introduced a multilingual dataset for task-oriented DS. It contained many annotated utterances in English, but few translated annotated English utterances to Spanish and Thai covering the weather, alarm, and reminder domains. They evaluated their approach using three cross-lingual transfer learning approaches: translating the training data, cross-lingual contextual word representations, and a multilingual machine-translation encoder. Their experiment showed that the latter two approaches outperformed the one in which the training data were translated. However, translating the training data achieves the best results in cases of zero-shot settings where there are no available target-language data. Moreover, they proved that joint training on both high-resource and low-resource target languages improved the performance on the target language.

In [11], Liu et al. proposed an attention-informed Mixed-Language Training (MLT) approach. This is a zero-shot adaptation approach for cross-lingual task-oriented DS. They used a code-switching approach, where code-switching sentences are generated from source-language sentences in English by replacing particular selected source words with their translations in German and Italian. They used task-related parallel word pairs in order to generate code-switching (mixed language) sentences. They obtained better generalization in the target language because of the inter-lingual semantics across languages. Their model achieved a significantly better performance with the zero-shot setting for both cross-lingual dialogue state tracking (DST) and natural language understanding (NLU) tasks than other approaches using a large amount of bilingual data.

Lin et al. [12] proposed a bilingual multidomain dataset for end-to-end task-oriented DS (BiToD). BiToD contains over 7000 multidomain dialogues for both English (EN) and Chinese (ZH), with a large bilingual knowledge base (KB). They trained their models using mT5 and mBART [13]. They evaluated their system in monolingual, bilingual, and cross-lingual settings. In the monolingual setting, they trained and tested the models on either English or Chinese dialogue data. In the bilingual setting, they trained the models on dialogue data in both languages. In the cross-lingual setting, they studied data in a specific language, and thus they used transfer learning to study the transferability of knowledge from a high-resource language to a low-resource language. For end-to-end task evaluation, they used: BLEU; Task Success Rate (TSR), to measure if the system provided the correct entity and answered all the requested information of a specific task; Dialogue Success Rate (DSR), to assess if the system completed all the tasks in the dialogue; and API Call Accuracy (API_{Acc}), to assess if the system generated a correct API call. In addition, they used the joint goal accuracy (JGA) metric to measure the performance of the DST. They suggested the

possibility of improving system performance under low-resource conditions by leveraging the bilingual KB and cross-lingual transfer learning. In fact, BiToD achieved important progress in building a robust multilingual task-oriented DS.

Table 1 summarizes the performance of the most common state-of-the-art models in multilingual task-oriented DS using cross-lingual transfer learning. It is clear that there is a lack of standardized datasets in addition to a lack of fixed standards in using evaluation metrics [14].

Table 1. Performance comparison of the most common state-of-the-art multilingual models in task-oriented dialogue systems (DS) for cross-lingual settings.

Study	Dataset	Dataset Size (Train/Validate/Test)	Metrics	
[8] Cross-lingual settings using cross-lingual pre-trained embeddings	Their own dataset across three domains: weather, alarm and reminder.	English (43k) 30,521/4181/8621 Spanish (8.6k) 3617/1983/3043 Thai (5k) 2156/1235/1692	—	
			Exact match = 75.96, Domain accuracy (acc.) = 99.47, Intent acc. = 97.51, Slot F1 = 83.38 Exact match = 86.12, Domain acc. = 99.33, Intent acc. = 9.87, Slot F1 = 91.51	
[8] Cross-lingual settings using translating the training data		English (43k) 30,521/4181/8621 Spanish (8.6k) 3617/1983/3043 Thai (5k) 2156/1235/1692	—	
			Exact match = 72.49, Domain acc. = 99.65, Intent acc. = 98.47, Slot F1 = 80.60 Exact match = 73.37, Domain acc. = 99.37, Intent acc. = 97.41, Slot F1 = 80.38	
[11] Zero-shot setting using mBERT + transformer	Multilingual WOZ 2.0	English (1k) 600/200/400 German (1k) 600/200/400 Italian (1k) 600/200/400	—	
			Slot acc. = 70.77, JGA. = 34.36, Request acc. = 86.97	
			Slot acc. = 71.45, JGA = 33.35, Request acc. = 84.96	
BiToD [12] cross-lingual	BiToD (cross-lingual) (ZH → EN)	English (EN) (3.6k) 2952/295/442	mT5	TSR = 6.78, DSR = 1.36, API_{Acc} = 17.75, BLEU = 10.35, JGA = 19.86
			mBART	TSR = 1.11, DSR = 0.23, API_{Acc} = 0.60, BLEU = 3.17, JGA = 4.64
			mT5 + CPT	TSR = 44.94, DSR = 24.66, API_{Acc} = 47.60, BLEU = 29.53, JGA = 48.77
			mBART + CPT	TSR = 36.19, DSR = 16.06, API_{Acc} = 41.51, BLEU = 22.50, JGA = 42.84
			mT5 + MLT	TSR = 56.78, DSR = 33.71, API_{Acc} = 56.78, BLEU = 32.43, JGA = 58.3
			mBART + MLT	TSR = 33.62, DSR = 11.99, API_{Acc} = 41.08, BLEU = 20.01, JGA = 55.39
	BiToD (EN → ZH)	Chinese (ZH) (3.5k) 2835/248/460	mT5	TSR = 4.16, DSR = 2.20, API_{Acc} = 6.67, BLEU = 3.30, JGA = 12.63
			mBART	TSR = 0.00, DSR = 0.00, API_{Acc} = 0.00, BLEU = 0.01, JGA = 2.14
			mT5 + CPT	TSR = 43.27, DSR = 23.70, API_{Acc} = 49.70, BLEU = 13.89, JGA = 51.40
			mBART + CPT	TSR = 24.64, DSR = 11.96, API_{Acc} = 29.04, BLEU = 8.29, JGA = 28.57
			mT5 + MLT	TSR = 49.20, DSR = 27.17, API_{Acc} = 50.55, BLEU = 14.44, JGA = 55.05
			mBART + MLT	TSR = 44.71, DSR = 21.96, API_{Acc} = 54.87, BLEU = 14.19, JGA = 60.71

The symbol (→) means that the model was built using cross-lingual transfer language approach by transfer the knowledge from the source language (on the left of the arrow) to the target language (on the right of the arrow).

In addition to the previous studies, Louvan et al. [15] suggested using a data augmentation approach to resolve the data scarcity in task-oriented DS. Data augmentation aims to produce extra training data, and it proved its success in different NLP tasks for English data [15]. As such, the authors studied its performance for task-oriented DS for non-English data. They evaluated their approach on five languages: Italian, Spanish, Hindi, Turkish, and Thai. They found that data augmentation improved the performance for all languages. Furthermore, data augmentation improved the performance of the mBERT model, especially for a slot-filling task.

To the best of our knowledge, the study of [12] is the most closely related work to ours. The authors leveraged mT5 and mBART to fine-tune a task-oriented dialogue task based on a new bilingual dialogue dataset. Although the performances of the pre-trained language models for task-oriented DS in English has encouraged the emergence of the multilingual models for task-oriented DS in low-resource languages, there is still a gap between the system performance of both low-resource and high-resource languages, due to the lack of high-quality data in the low-resource languages. To the best of our knowledge, no study exists for Arabic language task-oriented DS using multilingual models. Therefore, according to the current landscape of cross-lingual transfer learning in addition to the achieved performance of multilingual language models in task-oriented DS, we aimed to explore how far can mT5 be useful in building an Arabic end-to-end task-oriented DS using cross-lingual settings.

3. Methodology

The workflow of our task-oriented DS is based on a single Seq2Seq model using the pre-trained multilingual model mT5 [10]. D is the dialogue session that represents a sequence of user utterances (U_t) and system utterances (S_t) at turn t, where D = {U_1, S_1, . . . , U_t, S_t}. The interaction between the user and system at turn t creates a dialogue history (H_t) that holds all the previous utterances of both the user and system determined by the context window size (w), where H_t = {U_{t-w}, S_{t-w}, . . . , S_{t-1}; U_t}. Over the turn (t) of a dialogue session, the system tracks the dialogue state (B_t) and knowledge state (K_t), then generates the response (R). We first set the dialogue state and knowledge state to empty strings as B_0 and K_0, respectively. Then, the input at turn t was composed of the current dialogue history (H_t), previous dialogue state (B_{t-1}), and previous knowledge state (K_{t-1}). For each turn, the dialogue state updated from (B_{t-1}) to (B_t) and produced the Levenshtein Belief Spans at turn t (Lev_t), which holds the updated information. Finally, the system was queried with the determined constraint from the dialogue state in order to generate the API name, and then the system updated the knowledge state form (K_{t-1}) to (K_t). Both K_t and the API name were used to generate the response, which was returned to the user.

3.1. Dataset

Arabic-TOD, a multidomain Arabic dataset, was used for training and evaluating end-to-end task-oriented DS. This dataset was created by manually translating the original English BiToD dataset into Arabic. Of 3689 English dialogues, the Arabic-TOD dataset contained 1500 dialogues with 30,000 utterances covering four domains (Hotels, Restaurants, Weather, and Attractions). The dataset comprised 14 intent types and 26 slot types. The Arabic-TOD dataset was preprocessed and prepared for the training step, then divided into 67%, 7%, and 26% for training, validation, and testing, respectively.

3.2. Evaluation Metrics

For DST evaluation, we used the JGA metric [16] to compare the predicted dialogue states to the ground truth for each dialogue turn. If all the predicted slot values exactly match the ground-truth values, the output of the model is correct.

For the end-to-end generation task, we employed the original evaluation metrics in [12]. The API_{Acc} metric was used to measure if the system generated the correct API call. The TSR metric was used to measure whether the system had found the correct entity and answered all the task's requested information. The DSR metric was used to evaluate if the system completed all the dialogue tasks. Additionally, we computed the BLEU score [17] to measure the fluency of the generated response.

4. Experiments

We investigated the effectiveness of the powerful multilingual model mT5 for few-shot cross-lingual learning for end-to-end Arabic task-oriented DS. In the cross-lingual transfer learning, we transferred the knowledge from a high-resource language (English) to

a low-resource language (Arabic). We used three approaches from the literature, [8,11,12]: mSeq2Seq, cross-lingual pre-training (CPT), and mixed-language pre-training (MLT).

mSeq2Seq approach: In this setting, we utilized the existing pre-trained mSeq2Seq model mT5, and directly fine-tuned these models on the Arabic dialogue data.

Cross-lingual pre-training (CPT) approach: In this setting, we pre-trained the mSeq2Seq model mT5 on English, and then fine-tuned the pre-trained models on the Arabic dialogue data.

Mixed-language pre-training (MLT) approach: In this setting, we used a KB (dictionary) that contained a mixed-lingual context (Arabic and English) for most of the entities. As such, we generated the mixed-language training data by replacing the most task-related keyword entities in English with corresponding keyword entities in Arabic from a parallel dictionary for both input and output sequences. The process of generating the mixed-language context is shown in Figure 1. In this setting, we initially pre-trained the mSeq2Seq model mT5 with the generated mixed language-training dialogue data, then fine-tuned the pre-trained models on the Arabic dialogue data. During the training, our model learned to capture the most task-related keywords in the mixed utterances, which helped the model capture the other unimportant task-related words that have similar semantics, e.g., days of the week—"السبت" (Saturday) and "الثلاثاء" (Tuesday).

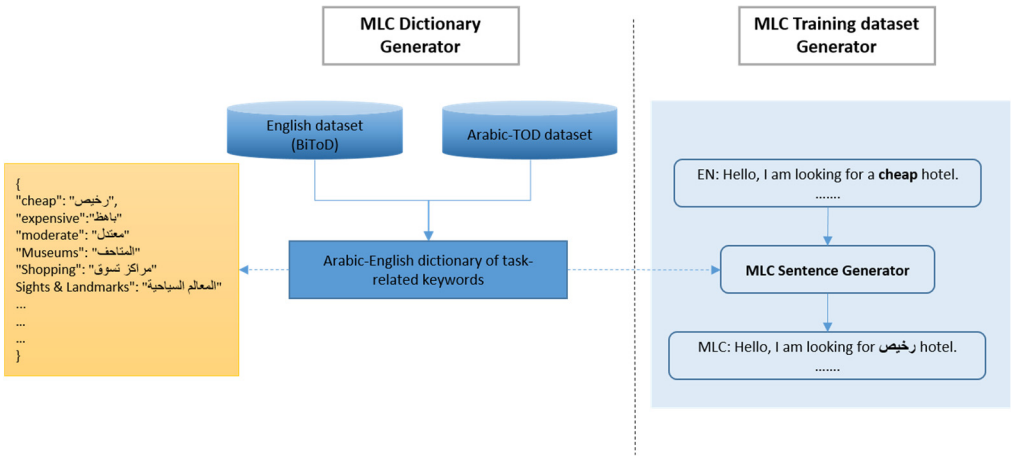

Figure 1. Mixed-language context (MLC) generation process.

In all these three approaches, we used the English BiToD dataset [12]. The Arabic-TOD dataset contained 10% of the English one. We intended to investigate if transferring the pre-trained multilingual language model mT5 to task-oriented DS could handle the paucity of the Arabic dialogue data.

Experiment Setup

We set up our experimental framework using the multilingual model mT5-small. We used the PyTorch framework [18] and the Transformers library [19]. We set the optimizer to AdamW [20] with a 0.0005 learning rate. We set the dialogue context window size at 2 and the batch size at 128, based on the best results in the existing literature. We first trained the models on English dialogues for 8 epochs, then fine-tuned the model on Arabic for 10 epochs. All the trainings took about 12 h using Google Colab.

5. Results and Discussion

Table 2 presents our findings compared to the previous experiments described in [12] on English and Chinese with the same settings in the three approaches: mSeq2Seq, CPT, and MLT. The results were calculated in terms of BLEU, API_{ACC}, TSR, DSR, and JGA. We compared our findings with those generated from [12], where in both works English represented the high-resource language used to transfer knowledge to the target language.

In [12], Chinese was considered as the low-resource target language, with 10% of the training dialogue, while in this work, Arabic was the low-resource target language, with 10% of the training dialogue.

Table 2. Cross-lingual experiment results of dialogue state tracking (DST) and end-to-end dialogue generation with 10% of Arabic-TOD dataset (AR) compared to 10% of Chinese dialogues in BiToD (ZH) [12]. Bold numbers indicate the best result according to the column's metric value.

	mSeq2Seq Approach				
	TSR	**DSR**	**API$_{Acc}$**	**BLEU**	**JGA**
AR	**18.63**	**3.72**	**15.26**	**9.55**	**17.67**
ZH [12]	4.16	2.20	6.67	3.30	12.63
	CPT Approach				
AR	42.16	14.18	46.63	23.09	32.71
ZH [12]	**43.27**	**23.70**	**49.70**	13.89	**51.40**
	MLT Approach				
AR	42.16	14.49	46.77	23.98	32.75
ZH [12]	**49.20**	**27.17**	**50.55**	14.44	**55.05**

Our model outperformed the Chinese model using the cross-lingual model deployed by the mSeq2Seq approach, while the Chinese model outperformed ours on the other two approaches, except for BLEU. However, these poor results can be improved by increasing the Arabic-TOD dataset size under the few-shot learning scenarios. Due to the smallness of the Arabic-TOD dataset, we tried to re-implement the experiment utilizing all the data of the Arabic-TOD dataset, which comprised 27% of the size of the Chinese dataset in [12]. In so doing, we improved the results of the Arabic models, as shown in Table 3. We found a high improvement from the first approach, which still outperformed the Chinese model. Furthermore, Arabic obtained better results in terms of TSR, API$_{Acc}$, and BLEU than Chinese using the cross-lingual model deployed by the CPT approach. The cross-lingual model deployed by the MLT approach outperformed the Chinese model in terms of API$_{Acc}$ and BLEU.

Table 3. Three approaches in cross-lingual settings results of DST and end-to-end dialogue generation with all the Arabic-TOD dataset. Numbers in bold font indicate superior corresponding values to the Chinese model that were mentioned in Table 2.

Approach \ **Evaluation Metrics**	**TSR**	**DSR**	**API$_{Acc}$**	**BLEU**	**JGA**
AR (mseq2seq)	**42.88**	**13.95**	**48.68**	**29.28**	**35.74**
AR (CPT)	**47.18**	18.14	**52.10**	**31.16**	36.32
AR (MLT)	48.10	18.84	**52.58**	**31.74**	37.17

Overall, our findings indicate the excellent transferability of the cross-lingual multilingual language model mT5. Moreover, the MLT approach improved the performance of few-shot cross-lingual learning, which indicates that bilingual KB can facilitate the cross-lingual knowledge transfer in low-resource scenarios, such as in Arabic. In addition, the JGA values were relatively small for the Arabic models, due to the difficulty and multiplicity of tasks existing in Arabic-TOD datasets. As we mentioned earlier, we only translated the task-related keywords in the dialogues and did not translate names,

locations, and addresses, which in return made parsing the Arabic utterances easier in a cross-lingual setting.

In summary, the cross-lingual setting is an effective approach for building an Arabic end-to-end task-oriented DS, in cases in which there is a scarcity of training data. Our results can be considered a baseline for the future of Arabic conversational systems.

Impact of Arabic-TOD Dataset Size on Arabic Task-Oriented DS Performance

We also investigated how our cross-lingual model deployed with MLT performed with different training dataset sizes in few-shot learning settings. We conducted further experiments with varying training data percentages on the Arabic-TOD dataset, ranging from 5% (50 examples) to 100% (1000 examples). We observed improvements when increasing the dataset size for cross-lingual training, as shown in Table 3. In this experiment, we focused on the cross-lingual model with the MLT approach, due to its previous performance. We fine-tuned the pre-trained models with the MLT approach on the few-shot Arabic dialogue data. To conclude, our results can be improved with dataset increases, as shown in Table 4. Although the dialogue dataset was small, it was sufficient to study the effectiveness of the cross-lingual model using the multilingual language model mT5 for Arabic end-to-end task-oriented DS.

Table 4. Few-shot learning results of cross-lingual model deployed with MTL approach on the Arabic-TOD dataset using training dataset of different sizes. Bold numbers indicate the best result according to the column's metric value.

Dataset Size \ Evaluation Metrics	TSR	DSR	API_{Acc}	BLEU	JGA
5%	30.09	10.23	33.07	20.26	24.85
10%	34.90	11.86	37.89	20.87	28.26
20%	40.73	14.42	44.47	23.84	32.05
50%	42.16	14.88	48.51	24.94	34.03
100%	**48.10**	**18.84**	**52.58**	**31.74**	**37.17**

6. Conclusions and Future Work

In this work, we studied the effectiveness of cross-lingual transfer learning using the multilingual language model mT5 for Arabic end-to-end task-oriented DS. We used the Arabic-TOD dataset in training and testing the model. To address the problem of the small Arabic dialogue dataset, we presented cross-lingual transfer learning using three approaches. We obtained good results for our model compared to those in the literature for Chinese with the same settings. Therefore, cross-lingual transfer learning can improve the system performance of Arabic in cases of small datasets. Furthermore, we explored the impact of Arabic training dialogue data size on cross-lingual learning in few-shot scenarios and found improvements when increasing the training dataset size. Finally, the results obtained from our proposed model on the Arabic-TOD dataset can be considered a baseline for future researchers to build robust end-to-end Arabic task-oriented DS that tackle complex scenarios. In addition to the training dataset size, there are other factors that may influence the model's performance, such as the number of tasks and the number of turns (length) in the dialogue. Further experimentation to validate the influence of these factors will be addressed in future work.

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.

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. McTear, M. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*; Morgan & Claypool Publishers LLC: San Rafael, CA, USA, 2020; Volume 13.
2. Wu, C.S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; Fung, P. Transferable multi-domain state generator for task-oriented dialogue systems. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 808–819. [CrossRef]
3. 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. [CrossRef]
4. Yang, Y.; Li, Y.; Quan, X. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. *arXiv* **2020**, arXiv:2012.03539.
5. 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.
6. AlHagbani, E.S.; Khan, M.B. Challenges facing the development of the Arabic chatbot. In Proceedings of the First International Workshop on Pattern Recognition 2016, Tokyo, Japan, 11–13 May 2016; Volume 10011, p. 7. [CrossRef]
7. 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, 3–7 November 2019; pp. 1297–1303. [CrossRef]
8. Schuster, S.; Shah, R.; Gupta, S.; Lewis, M. Cross-lingual transfer learning for multilingual task oriented dialog. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 3795–3805. [CrossRef]
9. Zhou, X.; Dong, D.; Wu, H.; Zhao, S.; Yu, D.; Tian, H.; Liu, X.; Yan, R. Multi-view response selection for human-computer conversation. In Proceedings of the EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 372–381.
10. 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, 6–11 June 2020; pp. 483–498. [CrossRef]
11. 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. In Proceedings of the AAAI 2020—34th Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 8433–8440. [CrossRef]
12. 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.
13. 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]
14. Sitaram, S.; Chandu, K.R.; Rallabandi, S.K.; Black, A.W. A Survey of Code-switched Speech and Language Processing. *arXiv* **2019**, arXiv:1904.00784.
15. Louvan, S.; Magnini, B. Simple data augmentation for multilingual NLU in task oriented dialogue systems. In Proceedings of the Seventh Italian Conference on Computational Linguistics CLIC-IT 2020, Bologna, Italy, 30 November–2 December 2020; Volume 2769. [CrossRef]
16. Henderson, M.; Thomson, B.; Williams, J. The second dialog state tracking challenge. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Philadelphia, PA, USA, 18–20 June 2014; pp. 263–272. [CrossRef]
17. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2001; pp. 311–318. [CrossRef]
18. PyTorch. Available online: https://pytorch.org/ (accessed on 17 November 2021).

19. 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).
20. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2017.