NLU for Game-based Learning in Real: Initial Evaluations

Eda Okur, Saurav Sahay, Lama Nachman
Intel Labs
USA
{eda.okur, saurav.sahay, lama.nachman}@intel.com

Abstract

Intelligent systems designed for play-based interactions should be contextually aware of the users and their surroundings. Spoken Dialogue Systems (SDS) are critical for these interactive agents to carry out effective goal-oriented communication with users in real-time. For the real-world (i.e., in-the-wild) deployment of such conversational agents, improving the Natural Language Understanding (NLU) module of the goal-oriented SDS pipeline is crucial, especially with limited task-specific datasets. This study explores the potential benefits of a recently proposed transformer-based multi-task NLU architecture, mainly to perform Intent Recognition on small-size domain-specific educational game datasets. The evaluation datasets were collected from children practicing basic math concepts via play-based interactions in game-based learning settings. We investigate the NLU performances on the initial proof-of-concept game datasets versus the real-world deployment datasets and observe anticipated performance drops in-the-wild. We have shown that compared to the more straightforward baseline approaches, Dual Intent and Entity Transformer (DIET) architecture (Bunk et al., 2020) is robust enough to handle real-world data to a large extent for the Intent Recognition task on these domain-specific in-the-wild game datasets.

Keywords: Spoken Dialogue Systems, Natural Language Understanding, Intent Recognition, Game-based Learning

1. Introduction

Investigating Artificial Intelligence (AI) systems that can help children in their learning process has been a challenging yet exciting area of research (Chassignol et al., 2018; Zhai et al., 2021). Utilizing Natural Language Processing (NLP) for building educational games and applications has gained popularity in the past decade (Lende and Raghuwanshi, 2016; Cahill et al., 2020). Game-based learning systems can offer significant advantages in teaching fundamental math concepts interactively, especially for younger students (Skene et al., 2022). These intelligent systems are often required to handle multimodal understanding of the kids and their surroundings in real-time. Spoken Dialogue Systems (SDS) are vital building blocks for efficient task-oriented communication with children in game-based learning settings. In this study, the application domain is a multimodal dialogue system for younger kids learning basic math concepts through gamified interactions. Such dialogue system technology needs to be constructed and modeled carefully to handle task-oriented game interactions between the children and a virtual character serving as a conversational agent.

Building the Natural Language Understanding (NLU) module of a goal-oriented SDS for game-based interactions usually involves: (i) the definition of intents (and entities if needed); (ii) creation of game-specific and task-relevant datasets; (iii) annotation of the game data with domain-specific intents and entities; (iv) iterative training and evaluation of NLU models; (v) repeating this tedious process for every new or updated game usages. Improving the NLU performances of task-oriented SDS pipelines in low-data regimes is quite challenging. This study primarily explores the potential benefits of a recent transformer-based multi-task architecture proposed for joint Intent and Entity Recognition tasks, especially with limited game datasets. Utilizing that flexible architecture, we focus on increasing the performance of our NLU models trained on small-size task-specific game datasets. The main NLU task we aim to improve is the Intent Recognition from possible user/player utterances during gamified learning interactions. Given an input utterance, the goal of an Intent Recognition model is to predict the user's intent (e.g., what the player wants to accomplish within a game-based interaction).

This work investigates the Intent Recognition model performances on our early proof-of-concept (POC) educational game datasets created to bootstrap the SDS to be deployed later in the real world. We have shown that adopting the recently proposed lightweight Dual Intent and Entity Transformer (DIET) architecture (Bunk et al., 2020) along with the Conversational Representations from Transformers (ConveRT) embeddings (Henderson et al., 2020) is a promising approach for NLU. This method boosts the NLU performance results on our initial small-scale POC game datasets. After the exploratory validation studies were conducted in-the-lab, the final evaluation datasets were collected in-the-wild from students working on fundamental math concepts in a game-based learning space at school. We examine the Intent Recognition performances on these real-world deployment datasets and reveal highly expected performance degradations in-the-wild. Compared to the baseline approaches, we have shown that adopting a DIET classifier with pre-trained ConveRT representations still achieves improved NLU results on our evaluation datasets collected in-the-wild.

2. Related Work

The use of AI technologies to enhance students' learning experiences has gained increasing popularity, especially in the last decade (Chassignol et al., 2018; Aslan et al., 2019; Jia et al., 2020; Baker, 2021; Zhai et al., 2021; Zhang and Aslan, 2021). Intelligent game-based learning systems (Lester et al., 2013; Richey et al., 2021) present significant benefits for practicing math concepts in smart spaces (Pires et al., 2019; Sun et al., 2021), specifically for early childhood education (Skene et al., 2022). Adapting NLP techniques to build various educational applications has been an appealing area of research for quite some time (Meurers, 2012; Blanchard et al., 2015; Lende and Raghuwanshi, 2016; Taghipour and Ng, 2016; Raamadhurai et al., 2019; Cahill et al., 2020; Ghosh et al., 2020). To slightly narrow down on these applications, building conversational agents for the smart education has been widely studied in the community (Graesser et al., 2004; Litman and Silliman, 2004; Kerry et al., 2009; Roos, 2018; Winkler and Söllner, 2018; Palasundram et al., 2019; Winkler et al., 2020). Relatively few number of studies also exist specifically on recognizing goals or intents of players in educational games (Min et al., 2016; Min et al., 2017; Hooshyar et al., 2019).

Since our ultimate goal is to build dialogue systems for interactive educational games, we have outlined the previous studies with applications of AI and NLP for education context until now (e.g., intelligent systems and conversational agents for play-based learning). Next, we will briefly summarize the dialogue system technologies and NLU approaches in a more generic context.

Dialogue systems are frequently categorized as either task-oriented or open-ended. The task-oriented dialogue systems are designed to fulfill specific tasks and handle goal-oriented conversations. The open-ended systems or chatbots, on the other hand, allow more generic conversations such as chit-chat (Jurafsky and Martin, 2018). With the advancements of deep learning-based language technologies and increased availability of large datasets with computing power in the research community, the dialogue systems trained end-to-end produce promising results for both goal-oriented (Bordes et al., 2017) and open-ended (Dodge et al., 2016) applications. Dialogue Managers (DM) of goal-oriented systems are often sequential decision-making models. The optimal policies can be learned via reinforcement learning from a high number of user interactions (Shah et al., 2016; Dhingra et al., 2017; Liu et al., 2017; Su et al., 2017; Cuayáhuitl, 2017). Unfortunately, building such systems with limited user interactions is extremely challenging. Therefore, supervised learning approaches with modular SDS pipelines are still widely preferred when initial training data is limited, basically to bootstrap the goal-oriented conversational agents for further data collection (Sahay et al., 2019). Statistical and neural network-based dialogue system toolkits and frameworks (Bocklisch et al., 2017; Ultes et al., 2017; Burtsev et al., 2018) are heavily used in the academic and industrial research communities for implicit dialogue context management.

The NLU module within SDS pipeline processes the user utterances as input and often predicts the user intents (along with entities of interest if necessary). LSTM networks (Hochreiter and Schmidhuber, 1997) and Bidirectional LSTMs (Schuster and Paliwal, 1997) have been widely utilized for sequence learning tasks such as Intent Classification and Slot Filling (Mesnil et al., 2015; Hakkani-Tür et al., 2016). Joint training of Intent Recognition and Entity Extraction models have been explored recently (Zhang and Wang, 2016; Liu and Lane, 2016; Goo et al., 2018; Varghese et al., 2020). Several hierarchical multi-task architectures are proposed for these joint NLU approaches (Zhou et al., 2016; Wen et al., 2018; Okur et al., 2019; Vanzo et al., 2019), few of them in multimodal context (Gu et al., 2017; Okur et al., 2020). Vaswani et al. (2017) proposed the Transformer as a novel neural network architecture based entirely on attention mechanisms (Bahdanau et al., 2015). Shortly after, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) became one of the significant breakthroughs in pre-trained language representations, showing strong performance in numerous NLP tasks, including the NLU. Recently, Bunk et al. (2020) introduced the Dual Intent and Entity Transformer (DIET) as a lightweight multi-task architecture that outperforms fine-tuning BERT for predicting intents and entities on a complex multi-domain NLU-Benchmark dataset (Liu et al., 2021). On the efficient representation learning side, Henderson et al. (2020) lately proposed the Conversational Representations from Transformers (ConveRT), which is also a lightweight approach to obtain pre-trained embeddings as sentence representations to be successfully utilized in numerous conversational AI tasks.

3. NLU Models

This section describes the models we examine for the NLU (i.e., Intent Recognition) module within a dialogue system pipeline. We have built our NLU models on top of the Rasa open-source framework (Bocklisch et al., 2017). The former baseline Intent Recognition architecture available in Rasa is based on supervised embeddings provided within the Rasa NLU (Bocklisch et al., 2017), which is an embedding-based text classifier that embeds user utterances and intent labels into the same vector space. This former baseline architecture is inspired by the StarSpace work (Wu et al., 2018), where the supervised embeddings are trained by maximizing the similarity between intents and utterances. The algorithm learns to represent user inputs and intents into a common embedding space and compares them against each other in that vectorial space. It also learns to rank a set of intents given a user utterance and

Figure 1: Learning basic math via game play-based interactions.

provides similarity rankings of these labels. In Sahay et al. (2019), the authors enriched this embedding-based former baseline Rasa Intent Classifier by incorporating additional features and adapting alternative network architectures. To be more precise, they adapted the Transformer network (Vaswani et al., 2017) and incorporated pre-trained BERT embeddings using the `bert-base-uncased` model (Devlin et al., 2019) to improve the Intent Recognition performance. In this work, we employed this improved approach from Sahay et al. (2019) as our baseline NLU model, which we would call TF+BERT in our experiments.

In this study, we explore the potential improvements in Intent Classification performance by adapting the recent DIET architecture (Bunk et al., 2020). DIET is a transformer-based multi-task architecture for joint Intent Recognition and Entity Extraction. It employs a 2-layer transformer shared for both of these NLU tasks. To be more precise, a sequence of entity labels is predicted with a Conditional Random Field (CRF) (Lafferty et al., 2001) tagging layer, which is on top of the transformer output sequence corresponding to the input sentences treated as a sequence of tokens. For the intent labels, the transformer output for the __CLS__ token (i.e., classification token at the end of each sentence) and the intent labels are embedded into the same semantic vector space. The dot-product loss is utilized to maximize the similarity with the target label and minimize similarities with the negative samples. Note that DIET can incorporate pre-trained word and sentence embeddings from language models as dense features, with the flexibility to combine these with token level one-hot encodings and multi-hot encodings of character n-grams as sparse features. These sparse features are passed through a fully-connected layer with shared weights across all sequence steps. The output of this fully-connected layer is concatenated with the dense features from the pre-trained models. This flexible architecture allows us to use any pre-trained embeddings as dense features in DIET, such as GloVe (Pennington et al., 2014), BERT (Devlin et al., 2019), and ConveRT (Henderson et al., 2020).

Conversational Representations from Transformers (ConveRT) is yet another recent and promising architecture to obtain pre-trained representations that are well-suited for real-world Conversational AI applications, especially for the Intent Classification task. ConveRT is a unique transformer-based dual-encoder network leveraging quantization and subword-level parameterization. In Henderson et al. (2020), the authors show that pre-trained representations from the ConveRT sentence encoder can be transferred to the Intent Classification task with promising results. Both DIET and ConveRT are lightweight architectures with faster and memory/energy-efficient training capabilities than their counterparts. When incorporating the ConveRT embeddings with the DIET classifier, the initial embedding for __CLS__ token is set as the input sentence encoding obtained from the ConveRT model. This way, we can leverage extra contextual information from the complete sentence on top of the word embeddings. For all the above reasons, we adapted the DIET architecture and incorporated pre-trained ConveRT embeddings to potentially improve the Intent Classification performances on our small domain-specific datasets. We would call this approach DIET+ConveRT in our experiments[1].

To investigate the actual benefits of DIET architecture versus the dense features, we also adapted DIET with out-of-the-box pre-trained BERT embeddings using the `bert-base-uncased` model (Devlin et al., 2019), as in our baseline TF+BERT NLU model. When combining these off-the-shelf BERT representations with the DIET classifier, the initial embedding for __CLS__ token is set to the corresponding output embedding of the BERT `[CLS]` token. We would call this approach DIET+BERT in our experiments.

4. Experimental Results

4.1. Datasets

We conduct our experiments on the Kid Space Planting and Watering (PW) games NLU datasets having

---

[1]Please refer to Bunk et al. (2020) for hyper-parameters, hardware specifications, and computational cost details.

Statistics/Dataset	Planting Game	Watering Game
# distinct intents	14	13
total # samples (utterances)	1927	2115
min # samples per intent	22	25
max # samples per intent	555	601
avg # samples per intent	137.6	162.7
# unique words (vocab)	1314	1267
total # words	10141	10469
min # words per sample	1	1
max # words per sample	74	65
avg # words per sample	5.26	4.95

Table 1: KidSpace-PW-POC Dataset Statistics

Statistics/Dataset	Planting Game	Watering Game
# distinct intents	12	11
total # samples (utterances)	2173	2122
min # samples per intent	4	6
max # samples per intent	1005	1005
avg # samples per intent	181.1	192.9
# unique words (vocab)	772	743
total # words	10433	9508
min # words per sample	1	1
max # words per sample	45	44
avg # words per sample	4.80	4.48

Table 2: KidSpace-PW-Deployment Dataset Statistics

utterances from play-based math learning experiences designed for early school-age children (i.e., 5-to-8 years old) (Anderson et al., 2018; Aslan et al., 2022). The use-cases aim to create an interactive smart space for children with traditional gaming motivations such as level achievements and virtually collecting objects. The smart space allows multiple children to interact, which can encourage social development. The intelligent agent should accurately comprehend inputs from children and provide feedback. The AI system needs to be physically grounded to allow children to bring meaningful objects into the play experience, such as physical toys and manipulatives as learning materials. Therefore, the multimodal system would combine various sensing technologies that should interact with children, track each child, and monitor their progress.

The use-cases include a specific flow of interactive games facilitating elementary math learning. The FlowerPot game (i.e., Planting Game in Tables 1 and 2) builds on the math concepts of tens and ones, with the larger flower pots representing 'tens' and smaller pots 'ones'. The virtual character provides the number of flowers the children should plant, and when the children have placed the correct number of large and small pots against the wall, digital flowers appear. In the NumberGrid game (i.e., Watering Game in Tables 1 and 2), math clues (or questions) are presented to children. When the correct number is touched on the number grid (i.e., on the wall), water is virtually poured to water the flowers. The visual, audio, and LiDAR-based recognition technologies enable physically situated interactions. The dialogue system is expected to take multimodal information to incorporate user identity, actions, gestures, audio context, and the objects (i.e., physical manipulatives) in space. For instance, during the FlowerPot game experience, the virtual character asks the children if they are done placing pots, to which they respond 'yes' (or 'no'). The dialogue system needs to use the visual input to have the virtual character respond appropriately to the correct (or incorrect) number of pots being detected.

Figure 1 demonstrates the virtual character (i.e., Oscar

the teddy bear) helping the kids with learning 'tens' and 'ones' concepts along with practicing simple counting, addition, and subtraction operations. The game datasets have a limited number of player utterances, which are manually annotated for intent types defined for each learning game or activity. For the FlowerPot game, we use the Planting Flowers game dataset, and for the NumberGrid game, we use a separate Watering Flowers game dataset. Some of the intents are quite generic across usages and games/activities (e.g., *affirm*, *deny*, *next-step*, *out-of-scope*, *goodbye*), whereas others are highly domain-dependent and game/task-specific (e.g., *intro-meadow*, *answer-flowers*, *answer-water*, *ask-number*, *answer-valid*, *answer-invalid*).

The current learning game activities are designed for two children collaboratively playing with the virtual agent. In addition to kids, an adult user (i.e., the Facilitator) is also present in the space to interact with the agent for game progress and help out the children when needed. Thus, we are dealing with a multiparty conversational system interacting with multiple users (i.e., two kids and one adult) while they progress through several learning games. In this goal-oriented dialogue system, the agent should provide the game instructions (with the Facilitator's help), guide the kids, and understand both the kids' and the Facilitator's utterances and actions to respond to them appropriately.

The NLU models are trained and validated on the initial POC datasets (Sahay et al., 2021) to bootstrap the agents for in-the-wild deployments. These POC game datasets were manually created based on the User Experience (UX) design studies to train the SDS models and then validated with the UX sessions in-the-lab with 5 kids (and one adult) going through play-based learning interactions. Table 1 shows the statistics of these KidSpace-PW-POC NLU datasets. Planting game and Watering game POC datasets have 1927 and 2115 utterances, respectively.

The deployment game datasets were later collected from 12 kids (and two adults), where the Kid Space setup was deployed in a classroom at school (Aslan et al., 2022). Table 2 shows the statistics of KidSpace-

Planting Game Datasets		# Utterances	
Type	Intent	POC	Deployment
Domain	*intro-meadow*	23	7
Specific	*answer-flowers*	110	13
	answer-valid	176	17
	answer-invalid	95	0
	intro-game	134	78
	help-affirm	41	4
	everyone-understand	22	11
	oscar-understand	25	15
	ask-number	34	18
	counting	418	581
Generic	*affirm*	144	370
	deny	125	54
	next-step	25	0
	out-of-scope	555	1005
	Total	1927	2173

Table 3: Intent Class Distributions for Planting Game

Watering Game Datasets		# Utterances	
Type	Intent	POC	Deployment
Domain	*answer-water*	69	9
Specific	*answer-valid*	201	6
	answer-invalid	91	0
	intro-game	102	30
	everyone-understand	44	11
	oscar-understand	25	15
	ask-number	73	21
	counting	476	581
Generic	*affirm*	165	370
	deny	157	54
	next-step	34	0
	out-of-scope	601	1005
	goodbye	77	20
	Total	2115	2122

Table 4: Intent Class Distributions for Watering Game

PW-Deployment NLU datasets, where Planting game and Watering game deployment datasets have 2173 and 2122 utterances, respectively. Note that these deployment datasets are used only for the testing purposes in this study, where we train our NLU models on the POC datasets. For both in-the-lab and in-the-wild datasets, the spoken user utterances are transcribed manually at first. These transcriptions are then manually annotated for the intent types we defined for each game activity. These transcribed and annotated final utterances are analyzed and used in our experiments in this study.

When we compare the POC versus deployment game datasets (in Tables 1-to- 4), we observe above 2.1k sample utterances for each game activity in both cases, except for the Planting POC data with around 1.9k samples. The number of possible user intents we envisioned for the POC was 14 and 13, respectively, for the Planting and Watering games. However, we have not observed any samples for two of the possible intent types for each game in the real-world deployment sessions. These intent types are *next-step* and *answer-invalid*, which were part of our backup intents in case we have technical issues and the users need to skip certain sub-activities (i.e., *next-step*), or in case the users provide highly irrelevant or unexpected answers to our specific questions in the game flow (i.e., *answer-invalid*). The minimum and the maximum number of samples per intent also differ significantly for the POC versus in-the-wild game datasets, which creates a huge difference in class distributions for our test samples (see Tables 3 and 4). Although we expect certain intent types to occur very infrequently in real game-plays (e.g., *help-affirm*), we still have to manually create enough samples ($\geq$20) for each intent type for the model training and validation purposes during the POC. The dominant intent class in both POC and in-the-wild datasets is *out-of-scope* (OOS). That was more or less anticipated as we are dealing with a multi-party conversational game setting here. In these games, the kids are encouraged to talk to each other while collaboratively solving the math puzzles. They can also discuss with or ask for help from the Facilitator. As the agent is in always-listening mode, if the users are not directly addressing Oscar, the system can detect those utterances as OOS (or *counting*, which is the second most frequent intent class, depending on the context). Notice that POC datasets were created with around one-fourth of the utterances as OOS, whereas the deployment datasets have almost half of the utterances tagged as OOS. That was mainly due to a relatively talkative Facilitator at school and some kids' preferences to talk to the Facilitator more often than Oscar in real deployment sessions. We have observed this behavior less often in our in-the-lab UX sessions, as the adult in the room was one of the researchers guiding kids to talk to Oscar instead. Those out-of-distribution and unseen OOS samples create additional challenges for the NLU models when tested on in-the-wild game datasets. We have also observed the vocabulary sizes shrink in-the-wild as we tried to manually curate more variations in the POC datasets to make the NLU models more robust. The average number of tokens per sample (i.e., utterance length) is around 5 in the POC data, yet, we observe slightly shorter utterances in-the-wild that might affect the available contextual information.

4.2. Intent Recognition Results

To evaluate the Intent Recognition performances, the baseline NLU model that we previously explored, TF+BERT, is compared with the DIET+BERT and DIET+ConveRT models that we adapted recently (see section 3). We conduct our evaluations on both the Planting and Watering game datasets. The models are trained and validated on the bootstrap POC datasets and then tested on the school deployment datasets. Table 5 summarizes the Intent Classification performance results on the POC datasets in micro-average

	Planting Game	Watering Game
Model/Dataset		
TF+BERT (Baseline)	90.50±0.25	92.43±0.32
DIET+BERT	94.00±0.38	96.39±0.14
DIET+ConveRT	**95.88±0.42**	**97.69±0.11**
Performance Gain	+5.38	+5.26

Table 5: NLU/Intent Recognition micro-avg F1-scores (%): TF+BERT, DIET+BERT, and DIET+ConveRT models trained and validated on KidSpace-PW-POC datasets (3 runs of 10-fold CV)

	Planting Game	Watering Game
Model/Dataset		
TF+BERT (Baseline)	85.08±0.49	90.06±0.56
DIET+BERT	87.03±0.30	89.63±0.62
DIET+ConveRT	**89.00±0.29**	**90.57±0.86**
Performance Gain	+3.92	+0.51

Table 6: NLU/Intent Recognition micro-avg F1-scores (%): TF+BERT, DIET+BERT, and DIET+ConveRT models trained on KidSpace-PW-POC (3 runs) and tested on KidSpace-PW-Deployment datasets (3 runs)

F1-scores. To test our model extensively on these limited-size POC datasets, we perform a 10-fold cross-validation (CV) by automatically creating multiple train/test splits. We report the average performance results with standard deviations obtained from the 3 runs, where we perform a 10-fold CV over the POC datasets for each run. As one can observe from Table 5, adapting the lightweight DIET architecture (Bunk et al., 2020) with pre-trained ConveRT embeddings (Henderson et al., 2020) significantly improved the Intent Classification performances for the NLU datasets manually created for POC. Specifically, the overall NLU performance gains are higher than 5% F1-scores for both Planting and Watering game datasets. Note that when we keep the dense features (i.e., pre-trained embeddings from BERT language models) constant, we can observe the clear benefits of switching from standard Transformer (TF) architecture to DIET classifier. We gain 3-to-4% F1-scores with DIET architecture, and we improve the Intent Recognition performance by another 1-to-2% with ConveRT embeddings compared to BERT. With these observations, which are consistent across different use-cases (i.e., Planting and Watering games), we updated the NLU component in our multimodal SDS pipeline by replacing the previous TF+BERT model with this promising DIET+ConveRT architecture.

Next, we investigate the NLU model performances on our real-world deployment datasets. The anticipated performance drops occurred when we tested these Intent Recognition models on in-the-wild data, which reflect more realistic game settings from a school deployment. Table 6 summarizes the Intent Classification performance results obtained on the deployment game datasets in micro-average F1-scores. Although the DIET+ConveRT models trained on POC datasets performed very well during the cross-validation (i.e., achieved around 96% and 98% F1-scores for Planting and Watering games, respectively), the performance loss is significantly high (i.e., around 7% F1-score) when tested on in-the-wild datasets. As a result, the same models achieved around 89% and 91% F1-scores when tested on the Planting and Watering deployment sets, respectively. That finding is quite com-

mon and probably not very surprising as the players in-the-wild can often largely deviate from the manual or synthetic data generation inside the labs or data collection through crowd-sourcing for interactive games. We have summarized the game dataset statistics and our preliminary observations regarding the main differences between the POC and deployment sets in section 4.1. We believe such deviations have played a significant role in the observed performance shifts for real-world play-based interactions. More specifically, the sample-class distributions, vocabulary sizes, slightly shorter utterance lengths, frequency of the OOS conversations due to multiparty setup, technical issues during the sessions causing unexpected interactions, etc., would all contribute to these shifts. One should also keep in mind the unprecedented group dynamics for that age group in play-based interactions and the unpredictable nature of kids in game-based learning settings. These factors also play some role in the robustness issues of NLU models developed for such challenging real-world deployments.

Besides the inevitable NLU performance degradations on real-world deployment datasets, Table 6 also compares the baseline TF+BERT models with more recent DIET+BERT and DIET+ConveRT architectures, all trained on POC data and tested on in-the-wild game data. The DIET+ConveRT models still reach the highest Intent Recognition F1-scores on these test sets, but the gap between the baseline and the best-performing models has been narrowed, especially for the Watering game. Compared to the TF+BERT baseline, the performance gain with the DIET+ConveRT approach is +3.92% in Planting and only +0.51% in Watering games when tested in-the-wild. For Planting, the increasing performance trends going from TF+BERT to DIET+BERT and DIET+ConveRT are also distinguishable on the deployment set. However, for Watering, the baseline TF+BERT model performs relatively well when tested on the deployment set, achieving only slightly lower F1-scores than the DIET+ConveRT. Notice that the variances are also relatively high in this case, so we may not observe the significant performance benefits when switching to DIET architecture from baseline TF for Watering game deployment. The

Data	Sample Utterance	Intent	Prediction
Planting Game	oh so like green and blue colors?	*answer-valid*	*answer-flowers*
	thirteen flowers!	*counting*	*answer-flowers*
	so if we had to start at a number what number do you think we should start at?	*counting*	*ask-number*
	or twenty less okay so we're going down	*counting*	*out-of-scope*
	let's add let's add a flower pot what do you think?	*counting*	*intro-game*
	yeah totally! do you wanna plant some next to him?	*affirm*	*intro-game*
	yeah I think that's ninety	*affirm*	*counting*
	no I think it was forty five	*deny*	*counting*
	okay so what do we need to start with?	*out-of-scope*	*help-affirm*
Watering Game	next one?	*ask-number*	*next-step*
	okay so how many more do you think we need?	*counting*	*ask-number*
	we need ten more to water	*counting*	*answer-water*
	to give to have enough water to plant our flowers and make them grow	*intro-game*	*answer-water*
	so when we look at these numbers all of the ones with the two in front, have two tens	*intro-game*	*counting*
	if we get four correct answer	*intro-game*	*counting*
	all right he's gotta go get his watering can that he must have put it away	*out-of-scope*	*intro-game*
	the ground	*out-of-scope*	*answer-valid*
	timber what	*out-of-scope*	*answer-valid*

Table 7: NLU/Intent prediction error samples from Planting and Watering games deployed in-the-wild: DIET+ConveRT model trained on KidSpace-PW-POC datasets and tested on KidSpace-PW-Deployment datasets

possible reason for the baseline model in Watering being already quite robust on real-world data could be the size differences in POC datasets on which the models are trained. To be more precise, the Planting baseline model is trained on 1927 samples and tested on 2173 in-the-wild utterances (see Table 3). Unlikely, the Watering baseline model is trained on 2115 samples and tested on 2122 utterances (see Table 4). In addition, we have one less intent class to predict in total (e.g., 14 vs. 13) and two fewer domain-specific intent types (e.g., 10 vs. 8) in the Watering game compared to the Planting. Having around 10% more data for training, plus having slightly less number of total and domain-specific intents, can explain the relatively higher robustness of the baseline model on Watering deployment data (compared to Planting). On the other hand, due to the consistently significant improvements obtained in all other cases (i.e., Planting-POC, Watering-POC, and Planting-Deployment), DIET+ConveRT still seems a promisingly more robust NLU model for our future use-cases.

4.3. Error Analysis and Discussion

In this subsection, we aim to investigate further the differences between the POC and the real-world deployment datasets for NLU in our game-based learning activities. When best-performing DIET+ConveRT models were tested in-the-wild, we discovered overall F1-score performance drops of around 7% for Intent Recognition, consistently for both game activities (i.e., Planting and Watering). When we analyze the intent-wise results, we identify some generalization issues between the POC to in-the-wild datasets, especially with the highly domain-specific intents.

For the Planting Flowers game, the top 5 intent classes with highest performance drops (≥20%) are *answer-valid*, *help-affirm*, *ask-number*, *answer-flowers*, and

intro-game. Among these, *help-affirm* had quite low test samples (only 4 utterances observed during deployments), which could explain the high variance in the detection performance. Regarding these top 5 erroneous intent classes, we realize that these are highly domain-dependent and activity-specific intent types, where we expect vastly specific answers from the kids based on the game flow design. To illustrate, during this Planting game, kids are helping Oscar to make the meadow look nicer. At the beginning of their interactions, the virtual character is asking "*Let's see, what could we add to the meadow... What kind of plants have pretty colors and smell nice?*" (or something along those lines as we use variations in response templates). We expect the kids to answer with "*flowers*" or its variations at this point in the game, where these short utterances should be classified as *answer-flowers* intent. However, kids can also answer with other plants (or animals, etc.) that belong to the meadow, like "*trees*", "*bushes*", "*butterflies*", "*birds*", etc. These viable but incorrect answers would ideally be classified as *answer-valid* intent. As you can see, these are extremely task-specific intents, and numerous things can go wrong in-the-wild for these, which may be beyond our assumptions. The *intro-game* intent is also highly game-specific as it is designed to cover the possible utterances from the Facilitator while s/he is introducing the game and explaining the rules (e.g., how to use the big and small pots for 'tens' and 'ones' for this Planting Flowers game). For more generic intent types that can be shared across other activities (e.g., *affirm*, *deny*, *out-of-scope*), we observed relatively less performance degradation in-the-wild using DIET+ConveRT.

For the Watering game activity, the top 4 intents with highest performance degradations (≥20%) are *answer-valid*, *ask-number*, *intro-game*, and *answer-water*. This time, *answer-valid* had very few test samples (only

34

6 utterances observed during the Watering game at school sessions), which might again explain the high variance in its performance. All these four intent types are also highly task-specific, and we anticipate more vulnerability for deviations in-the-wild for them, in contrast to the generic intent classes (e.g., *affirm*, *deny*, *out-of-scope*, *goodbye*). During the Watering game, this time, Oscar is asking "*What do you think we need to help the flowers bloom?*". We expect the kids to answer with "*water*" or its variations, where such utterances should be recognized as *answer-water* intent. Once again, kids can say other viable answers that could help the flowers grow/bloom, such as "*sunlight*", "*soil*", "*bees*", etc., which should be classified as *answer-valid* intent. Similarly, the *intro-game* intent is extremely domain/game-specific and aims to detect Facilitator utterances while s/he is introducing/explaining the game rules (e.g., how to use the number grid projected on the wall for touch-based interactions in this Watering game). Note that these valid answers or game introductions differ substantially based on which game we are playing, and we need to train separate NLU models for each game using these game domain-specific samples.

Table 7 depicts some of the user utterances collected in-the-wild as concrete examples from both deployment datasets. The ground truth intent labels and the predicted intent classes are compared, emphasizing the errors made on some of the most problematic game-specific intents. Here we use our best-performing DIET+ConveRT models for the Intent Classification task. These prediction errors are expected to occur in real-world deployments for various reasons. Some of these user utterances could have multiple intents (e.g., "*yeah totally! do you wanna plant some next to him?*" starts with *affirm*, then the Facilitator continues guiding the kids during *intro-game*). Others could fail due to subtle semantic differences between these classes (e.g., "*if we get four correct answer*" is used by the Facilitator while explaining the NumberGrid game rules but can easily be mixed with *counting* too). There exist some utterances where we see "*flowers*" or "*water*" while *counting* with numbers (e.g., "*thirteen flowers!*", "*we need ten more to water*"), which are confusing for the models trained on much cleaner datasets. Note that the majority of these classification errors occur for the user utterances during multiparty conversations, i.e., the users are talking to each other instead of Oscar, the virtual game character, but the SDS fails to recognize that (e.g., "*okay so what do we need to start with?*"). These sample utterances also depict several cases where our highly vocal adult Facilitator at school is talking to the kids to introduce the games, explain the rules, guide them to count loudly, and help them find the correct answers in the game flow. It is highly challenging to predict those nearly open-ended conversations and include all possibilities in the POC training datasets to make the NLU models more robust for

real-world deployments. However, we are working towards clustering-based semi-supervised intent discovery and human-in-the-loop (HITL) bulk labeling approaches (Sahay et al., 2021; Shen et al., 2021) for cleaner design and separation of intent classes on in-the-wild datasets. We also plan to continue our data augmentation with paraphrase generation efforts to increase the limited POC samples and add more variations during training to make the NLU models more robust in future deployments (Okur et al., 2022).

5. Conclusion

Dialogue systems are vital building blocks to carry out efficient task-oriented communication with children for game play-based learning settings. This study investigates a small step towards improving contextually aware multimodal agents that need to understand and track children's activities and interactions during educational games, support them in performing learning tasks and provide insights to teachers and parents to help personalize the learning experiences. We focus on building task-specific dialogue systems for younger kids learning basic math concepts via gamified interactions. We aim to improve the NLU module of the goal-oriented SDS pipeline with domain-specific game datasets having limited user/player utterances.

In this exploration, we experimented with a flexible and lightweight transformer-based multi-task architecture called DIET (Bunk et al., 2020) to improve the NLU performances on our task-specific game datasets with limited sizes. These domain-specific datasets are manually created for the POC first and then tested on in-the-wild deployment data. Based on the results obtained on POC game datasets, using the DIET classifier with pre-trained ConveRT embeddings has shown to be a promising approach yielding remarkably higher F1-scores for Intent Classification. The NLU results on the real-world deployment game datasets also support these preliminary findings but to a lesser extent.

Using the best performing DIET+ConveRT approach, we observed significant performance drops when the NLU models were tested on in-the-wild game datasets compared to the initial POC datasets. That finding was foreseeable as the player utterances in real-world deployments may usually diverge from the samples within the POC data manually generated for bootstrapping purposes. We investigated these game datasets and shared our exploratory insights for the deviations between POC and in-the-wild datasets. Our preliminary observations suggest that the highest performance shifts occur for the more domain-specific intents in each educational game set. We are working towards making the NLU models and eventually the SDS pipeline more robust for such deviations in-the-wild by empowering the interactive intent labeling with HITL learning techniques and the data augmentation with paraphrasing.

6. Acknowledgements

7. Bibliographical References

Anderson, G. J., Panneer, S., Shi, M., Marshall, C. S., Agrawal, A., Chierichetti, R., Raffa, G., Sherry, J., Loi, D., and Durham, L. M. (2018). Kid space: Interactive learning in a smart environment. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA. Association for Computing Machinery.

Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., and Arslan Esme, A. (2019). Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Aslan, S., Agrawal, A., Alyuz, N., Chierichetti, R., Durham, L. M., Manuvinakurike, R., Okur, E., Sahay, S., Sharma, S., Sherry, J., et al. (2022). Exploring kid space in the wild: a preliminary study of multimodal and immersive collaborative play-based learning experiences. *Educational Technology Research and Development*, pages 1–26.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR 2015)*.

Baker, R. S. (2021). Artificial intelligence in education: Bringing it all together. *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, pages 43–51.

Blanchard, N., Brady, M., Olney, A. M., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S., and D'Mello, S. (2015). A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In Cristina Conati, et al., editors, *Artificial Intelligence in Education*, pages 23–33, Cham. Springer International Publishing.

Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. In *Conversational AI Workshop, NIPS 2017*.

Bordes, A., Boureau, Y.-L., and Weston, J. (2017). Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations (ICLR 2017)*.

Bunk, T., Varshneya, D., Vlasov, V., and Nichol, A. (2020). DIET: lightweight language understanding for dialogue systems. *CoRR*, abs/2004.09936.

Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., and Zaynutdinov, M. (2018). Deep-Pavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia, July. Association for Computational Linguistics.

Cahill, A., Fife, J. H., Riordan, B., Vajpayee, A., and Galochkin, D. (2020). Context-based automated scoring of complex mathematical responses. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 186–192, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Chassignol, M., Khoroshavin, A., Klimova, A., and Bilyatdinova, A. (2018). Artificial intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136:16–24. 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece.

Cuayáhuitl, H. (2017). Simpleds: A simple deep reinforcement learning dialogue system. In *Dialogues with Social Robots*, pages 109–118. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dhingra, B., Li, L., Li, X., Gao, J., Chen, Y.-N., Ahmed, F., and Deng, L. (2017). Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–495, Vancouver, Canada, July. Association for Computational Linguistics.

Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A. H., Szlam, A., and Weston, J. (2016). Evaluating prerequisite qualities for learning end-to-end dialog systems. In *International Conference on Learning Representations (ICLR 2016)*.

Ghosh, D., Beigman Klebanov, B., and Song, Y.

(2020). An exploratory study of argumentative writing by young students: A transformer-based approach. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 145–150, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Goo, C.-W., Gao, G., Hsu, Y.-K., Huo, C.-L., Chen, T.-C., Hsu, K.-W., and Chen, Y.-N. (2018). Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757, New Orleans, Louisiana, June. Association for Computational Linguistics.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., and Louwerse, M. M. (2004). Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2):180–192.

Gu, Y., Li, X., Chen, S., Zhang, J., and Marsic, I. (2017). Speech intention classification with multimodal deep learning. In *Canadian conference on artificial intelligence*, pages 260–271. Springer.

Hakkani-Tür, D., Tur, G., Celikyilmaz, A., Chen, Y.-N., Gao, J., Deng, L., and Wang, Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. *Interspeech 2016*, pages 715–719.

Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., and Vulić, I. (2020). ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online, November. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hooshyar, D., Yousefi, M., and Lim, H. (2019). A systematic review of data-driven approaches in player modeling of educational games. *Artificial Intelligence Review*, 52(3):1997–2017.

Jia, J., He, Y., and Le, H. (2020). A multimodal human-computer interaction system and its application in smart learning environments. In Simon K. S. Cheung, et al., editors, *Blended Learning. Education in a Smart Learning Environment*, pages 3–14, Cham. Springer International Publishing.

Jurafsky, D. and Martin, J. H. (2018). *Ch 24: Dialog Systems and Chatbots. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 3rd (draft) edition.

Kerry, A., Ellis, R., and Bull, S. (2009). Conversational agents in e-learning. In Tony Allen, et al.,

editors, *Applications and Innovations in Intelligent Systems XVI*, pages 169–182, London. Springer London.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, ICML, pages 282–289.

Lende, S. P. and Raghuwanshi, M. (2016). Question answering system on education acts using nlp techniques. In *2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave)*, pages 1–6. IEEE.

Lester, J. C., Ha, E. Y., Lee, S. Y., Mott, B. W., Rowe, J. P., and Sabourin, J. L. (2013). Serious games get smart: Intelligent game-based learning environments. *AI Magazine*, 34(4):31–45, Dec.

Litman, D. J. and Silliman, S. (2004). Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, page 5–8, USA. Association for Computational Linguistics.

Liu, B. and Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016*, pages 685–689.

Liu, B., Tur, G., Hakkani-Tur, D., Shah, P., and Heck, L. (2017). End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. In *Conversational AI Workshop, NIPS 2017*.

Liu, X., Eshghi, A., Swietojanski, P., and Rieser, V., (2021). *Benchmarking Natural Language Understanding Services for Building Conversational Agents*, pages 165–183. Springer Singapore, Singapore.

Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., and Zweig, G. (2015). Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539, March.

Meurers, D. (2012). Natural language processing and language learning. *Encyclopedia of Applied Linguistics*, pages 4193–4205.

Min, W., Mott, B. W., Rowe, J. P., Liu, B., and Lester, J. C. (2016). Player goal recognition in open-world digital games with long short-term memory networks. In *IJCAI*, pages 2590–2596.

Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K. E., and Lester, J. (2017). Multimodal goal recognition in open-world digital games. In *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence And Interactive Digital Entertainment*, volume 13, pages 80–86.

Okur, E., Kumar, S. H., Sahay, S., Esme, A. A., and Nachman, L. (2019). Natural language interactions in autonomous vehicles: Intent detection and slot filling from passenger utterances. *20th International*

Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019), April.

Okur, E., H Kumar, S., Sahay, S., and Nachman, L. (2020). Audio-visual understanding of passenger intents for in-cabin conversational agents. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML), ACL 2020*, pages 55–59, Seattle, USA, July. Association for Computational Linguistics.

Okur, E., Sahay, S., and Nachman, L. (2022). Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*.

Palasundram, K., Sharef, N. M., Nasharuddin, N., Kasmiran, K., and Azman, A. (2019). Sequence to sequence model performance for education chatbot. *International Journal of Emerging Technologies in Learning (iJET)*, 14(24):56–68, December.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Pires, A. C., González Perilli, F., Bakała, E., Fleisher, B., Sansone, G., and Marichal, S. (2019). Building blocks of mathematical learning: Virtual and tangible manipulatives lead to different strategies in number composition. *Frontiers in Education*, 4.

Raamadhurai, S., Baker, R., and Poduval, V. (2019). Curio SmartChat : A system for natural language question answering for self-paced k-12 learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 336–342, Florence, Italy, August. Association for Computational Linguistics.

Richey, J. E., Zhang, J., Das, R., Andres-Bray, J. M., Scruggs, R., Mogessie, M., Baker, R. S., and McLaren, B. M. (2021). Gaming and confrustion explain learning advantages for a math digital learning game. In Ido Roll, et al., editors, *Artificial Intelligence in Education*, pages 342–355, Cham. Springer International Publishing.

Roos, S. (2018). Chatbots in education: A passing trend or a valuable pedagogical tool? Master's thesis, Uppsala University, Department of Informatics and Media.

Sahay, S., Kumar, S. H., Okur, E., Syed, H., and Nachman, L. (2019). Modeling intent, dialog policies and response adaptation for goal-oriented interactions. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*, London, United Kingdom, September. SEMDIAL.

Sahay, S., Okur, E., Hakim, N., and Nachman, L. (2021). Semi-supervised interactive intent labeling. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances,*

NAACL 2021, pages 31–40, Online, June. Association for Computational Linguistics.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, November.

Shah, P., Hakkani-Tur, D., and Heck, L. (2016). Interactive reinforcement learning for task-oriented dialogue management. In *Deep Learning for Action and Interaction Workshop, NIPS 2016*.

Shen, X., Sun, Y., Zhang, Y., and Najmabadi, M. (2021). Semi-supervised intent discovery with contrastive learning. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 120–129, Online, November. Association for Computational Linguistics.

Skene, K., O'Farrelly, C. M., Byrne, E. M., Kirby, N., Stevens, E. C., and Ramchandani, P. G. (2022). Can guidance during play enhance children's learning and development in educational contexts? a systematic review and meta-analysis. *Child Development*.

Su, P.-H., Budzianowski, P., Ultes, S., Gasic, M., and Young, S. (2017). Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 147–157, Saarbrücken, Germany, August. Association for Computational Linguistics.

Sun, Y., Play, T., Nambiar, R., and Vidyasagaran, V. (2021). Gamifying math education using object detection. In *Workshop on Math AI for Education (MATHAI4ED), 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Taghipour, K. and Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, November. Association for Computational Linguistics.

Ultes, S., Rojas Barahona, L. M., Su, P.-H., Vandyke, D., Kim, D., Casanueva, I., Budzianowski, P., Mrkšić, N., Wen, T.-H., Gasic, M., and Young, S. (2017). PyDial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada, July. Association for Computational Linguistics.

Vanzo, A., Bastianelli, E., and Lemon, O. (2019). Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 254–263, Stockholm, Sweden, September. Association for Computational Linguistics.

Varghese, A. S., Sarang, S., Yadav, V., Karotra, B., and Gandhi, N. (2020). Bidirectional lstm joint model for intent classification and named entity recognition in natural language understanding. *International Journal of Hybrid Intelligent Systems*, 16(1):13–23.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Wen, L., Wang, X., Dong, Z., and Chen, H. (2018). Jointly modeling intent identification and slot filling with contextual and hierarchical information. In Xuanjing Huang, et al., editors, *Natural Language Processing and Chinese Computing*, pages 3–15, Cham. Springer International Publishing.

Winkler, R. and Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Annual Meeting (AOM)*.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., and Leimeister, J. M., (2020). *Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent*, page 1–14. Association for Computing Machinery, New York, NY, USA.

Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., and Weston, J. (2018). Starspace: Embed all the things! In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., and Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, 2021.

Zhang, K. and Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, 2:100025.

Zhang, X. and Wang, H. (2016). A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2993–2999. AAAI Press.

Zhou, Q., Wen, L., Wang, X., Ma, L., and Wang, Y. (2016). A hierarchical lstm model for joint tasks. In Maosong Sun, et al., editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 324–335, Cham. Springer International Publishing.