Spoken Language Understanding Evaluations for Home-Based Basic Math Learning

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

NEURAL INFORMATION PROCESSING SYSTEMS
MATH-AI

INTRODUCTION

- We implement a multimodal task-oriented dialogue system to support play-based learning experiences at home, guiding kids to master basic math concepts.
- This work explores the Spoken Language Understanding (SLU) pipeline of a dialogue system developed for Kid Space, with cascading Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) components evaluated on our home deployment data with kids going through gamified math learning activities.
- We validate the advantages of a multi-task architecture for NLU and experiment with a diverse set of pretrained language representations for Intent Recognition and Entity Extraction tasks in the math learning domain.
- To recognize kids' speech in realistic home environments, we investigate several ASR systems, including Google Cloud and open-source Whisper solutions with varying model sizes.
- We evaluate the SLU pipeline by testing our best-performing NLU models on noisy ASR output to inspect the challenges of understanding children in authentic homes.

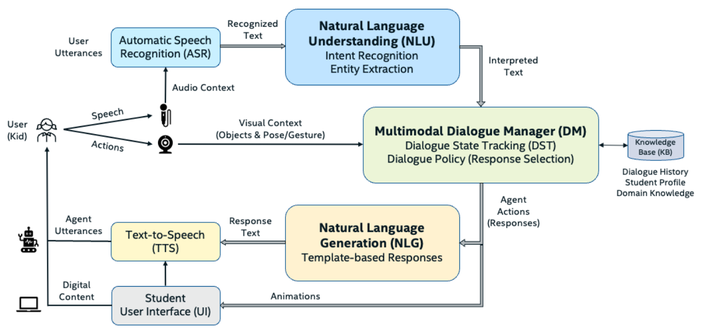

Fig 1: Multimodal Dialogue System Pipeline

METHODS

Datasets

- **POC data**, manually constructed based on UX studies and partially adopted from our previous school data [1], is used to train and cross-validate various NLU models.
- Recent home **deployment data** collected from 12 kids (ages 7-8) experiencing our multimodal math learning system at homes.
- Manually transcribed children's utterances in deployment data [2] are used to test our best NLU models trained on POC data.
- Evaluated multiple ASR engines on audio recordings to compute WER and assess ASR model performances on kids' speech.

Table 1: Kid Space Home POC & Deployment Data Stats

NLU Data Statistics	POC	Deployment
# Intents Types	13	12
Total # Utterances	6,245	733
# Entity Types	3	3
Total # Entities	5,023	497
Min # Utterances per Intent	105	1
Max # Utterances per Intent	1,051	270
Avg # Utterances per Intent	480.4	61.1
Min # Tokens per Utterance	1	1
Max # Tokens per Utterance	40	33
Avg # Tokens per Utterance	4.96	2.88
# Unique Tokens (Vocab Size)	1,992	362
Total # Tokens	30,976	2,113

- We have relatively generic intents (*state-name, affirm, deny, repeat, out-of-scope*) as well as highly domain-specific (*answer-flowers/valid/others, state-color, had-fun-a-lot, end-game*) or math-related intents (*state-number*, *still-counting*).
- Extracted entities are activity-specific (*name*, *color*) and math-related (*number*).

NLU Models

- We investigate several NLU models for Intent Recognition and Entity Extraction tasks by customizing open-source **Rasa** framework [3] as a backbone.
- Baseline approach is inspired by **StarSpace**, a supervised embedding-based model maximizing the similarity between utterances and intents in shared vector space.
  - We enrich this baseline classifier by incorporating **SpaCy** pre-trained embeddings as additional features. **CRF** Entity Extractor is also part of this baseline NLU.
- We explore the advantages of a more recent Dual Intent and Entity Transformer (**DIET**) model [4], a multi-task architecture for joint Intent and Entity Recognition.
  - To observe the net benefits of DIET, we first pass the identical **SpaCy** embed-dings used in our baseline (StarSpace) as dense features to DIET.
  - We adopt DIET with pretrained **BERT**, **RoBERTa**, **DistilBERT** word embeddings, as well as **ConveRT** [5] and **LaBSE** sentence embeddings to inspect the effects of these autoencoding-based language representations on NLU.
  - We also evaluate pretrained embeddings from models using autoregressive training such as **XLNet**, **GPT-2**, and **DialoGPT** on top of DIET.
  - Next, we explore recently-proposed math-language representations pretrained on math corpora, such as **MathBERT**, **Math-aware-BERT**, **Math-aware-RoBERTa**.

ASR Models

- We explore 3 main speech recognizers for our math learning application at home:
  - **Rockhopper** ASR is the baseline local approach. Its acoustic models rely on Kaldi generated resources trained on default adult speech data. Its language models fine-tuned with limited in-domain kids' utterances from previous school usages.
  - **Google Cloud** ASR is a commercial solution providing high-quality speech recognition service but requiring connectivity and payment, which cannot be adapted or fine-tuned as Rockhopper.
  - **Whisper** ASR [6] is an open-source adjustable solution that can run locally, achieving new state-of-the-art (SOTA) results. We inspect four configurations of varying model sizes (i.e., **tiny, base**, **small**, and **medium**).

EXPERIMENTAL RESULTS

NLU Model Selection

- We train Intent and Entity Classification models and cross-validate them over the POC dataset to select the best-performing NLU architectures for Kid Space Home.
- Compared to baseline (StarSpace), we gain 2% F1 for intents and 1% F1 for entities with DIET architecture.
- For language representations, BERT family of embeddings achieves higher F1 than the GPT family of embeddings.
- No benefits of employing math-specific representations as all such models achieve worse results than DIET+BERT.
- We select DIET+ConveRT as final model architecture for our NLU tasks at home.

Table 2: NLU Model Selection Results in F1-scores (%) Evaluated on Home POC Data (10-fold CV)

NLU Model	Intent Detection	Entity Extraction
StarSpace+SpaCy	92.83±0.28	97.14±0.21
DIET+SpaCy	94.40±0.08	98.45±0.11
DIET+BERT	97.37±0.26	99.29±0.01
DIET+RoBERTa	95.62±0.21	99.17±0.11
DIET+DistilBERT	97.52±0.23	99.54±0.11
DIET+ConveRT	**98.92±0.28**	99.66±0.02
DIET+LaBSE	98.31±0.21	**99.78±0.03**
DIET+XLNet	95.11±0.22	98.44±0.13
DIET+GPT-2	95.46±0.30	99.07±0.27
DIET+DialoGPT	96.12±0.52	99.00±0.11
DIET+MathBERT-base	94.67±0.25	98.15±0.20
DIET+MathBERT-custom	94.73±0.37	97.54±0.28
DIET+Math-aware-BERT	96.07±0.18	99.00±0.18
DIET+Math-aware-RoBERTa	94.31±0.19	98.81±0.20

NLU Evaluation on Deployment Data

Table 3: NLU Evaluation Results in F1-scores (%) for DIET+ConveRT Models Trained on Home POC Data & Tested on Home Deployment Data

Activity	Intent Detection POC	Intent Detection Deploy	Intent Detection Δ	Entity Extraction POC	Entity Extraction Deploy	Entity Extraction Δ
Intro (Meet & Greet)	99.92	97.46	-2.46	99.32	97.55	-1.77
Warm-up Game	98.91	93.54	-5.37	-	-	-
Training Game	98.48	94.27	-4.21	99.92	99.91	-0.01
Learning Game	99.02	94.37	-4.65	99.95	99.50	-0.45
Closure (Dance)	98.91	98.82	-0.09	-	-	-
All Activities	**98.92**	**94.36**	**-4.56**	**99.66**	**99.42**	**-0.24**

- We evaluate our NLU module on Kid Space Home Deployment data collected at authentic homes over 12 sessions with 12 kids, where each child goes through 5 activities within a session.
- We observe F1% drops (Δ) of 4.6 for intents and 0.3 for entities when our best DIET+ConveRT models tested on home deployment data.
- We witness distributional and utterance-length differences between POC & deployment datasets.
- Real-world data is always noisier than anticipated as these utterances come from younger kids playing math games in dynamic conditions.

ASR Model Evaluation

- Obtained WER before and after standard pre-processing steps (lowercase, punctuation removal) and application-specific filtering (num2word, cleaning).

Table 4: ASR Model Results: Average Word Error Rates (WER) for Child Speech at Kid Space Home Deployment Data

ASR Model	Raw Output	Lowercase (LC)	Remove Punct (RP)	Num2Word (NW)	LC & RP	LC & RP & NW	NW & Clean	LC & RP & NW & Clean
Rockhopper	0.939	0.919	0.924	0.937	0.886	0.884	0.937	0.884
Google Cloud	0.829	0.798	0.775	0.763	0.695	0.602	0.763	0.602
Whisper-tiny	1.055	1.027	1.002	1.027	0.964	0.919	0.983	0.880
Whisper-base	1.042	1.020	0.971	0.985	0.946	0.856	0.622	0.500
Whisper-small	0.834	0.804	0.760	0.756	0.720	0.621	0.537	**0.405**
Whisper-medium	0.905	0.870	0.824	0.814	0.785	0.675	0.522	**0.384**

- Relatively high error rates can be attributed to the characteristics of recordings (incidental voice and phrases), very short utterances (binary yes/no answers, stating numbers) & recognizing kids' speech.
- Still, the comparative results indicate that Whisper ASR solutions perform better on kids, and we can benefit from increasing the model size from tiny to small, while small to medium is close.

SLU Pipeline Evaluation

Table 5: SLU Pipeline Evaluation Results in F1-scores (%) for ASR+NLU and VAD-Adjusted ASR+NLU on Home Deployment Data

ASR Model	Intent Detection F1	Intent Detection Adjusted-F1	Entity Extraction F1	Entity Extraction Adjusted-F1
Rockhopper	37.3	15.7	84.4	35.5
Google Cloud	**79.1**	40.3	97.0	49.4
Whisper-tiny	58.1	56.6	94.0	89.1
Whisper-base	64.9	60.3	95.9	91.0
Whisper-small	72.1	68.0	96.5	91.6
Whisper-medium	**76.7**	**73.3**	**98.2**	**93.8**

- For SLU pipeline evaluation, we test our best-performing NLU models (DIET+ConveRT) on noisy ASR output.
- When VAD-adjusted F1-scores are compared, NLU on Whisper ASR performs relatively higher than Google and Rockhopper (aligned with the WER results).
- Increasing the ASR model size from small to medium could be worth the trouble for Whisper.
- When VAD-ASR errors propagate into pipeline, F1 drops from 94.2% with NLU to 73.1% with VAD-ASR+NLU.

Error Analysis

Table 6: NLU Error Analysis: Intent Recognition Error Samples from Home Deployment Data

Sample Kid Utterance	Intent	Prediction
Pepper.	*state-name*	*answer-valid*
Wow, that's a lot of red flowers.	*out-of-scope*	*answer-flowers*
None.	*state-number*	*deny*
Nothing.	*state-number*	*deny*
Yeah. Can we have some carrots?	*affirm*	*out-of-scope*
Okay, Do your magic.	*affirm*	*out-of-scope*
Maybe tomorrow.	*affirm*	*out-of-scope*
He's a bear.	*out-of-scope*	*answer-valid*
I like the idea of a bear	*out-of-scope*	*answer-valid*
Oh, 46? Okay.	*still-counting*	*state-number*
94. Okay.	*still-counting*	*state-number*
Now we have mountains.	*out-of-scope*	*answer-valid*
A pond?	*out-of-scope*	*answer-valid*
Sorry, I didn't understand it. Uh, five tens.	*state-number*	*still-counting*
Ah this is 70, 7.	*state-number*	*still-counting*

Table 7: SLU Pipeline (ASR+NLU): Intent Recognition Error Samples from Home Deployment Data

Human Transcript	ASR Output	ASR Model	Intent	Prediction
Six.	thanks	Rockhopper	*state-number*	*thank*
fifteen	if he	Rockhopper	*state-number*	*out-of-scope*
fifteen	Mickey	Google Cloud	*state-number*	*state-name*
Five.	bye	Google Cloud	*state-number*	*goodbye*
Blue.	Blair.	Whisper-base	*state-color*	*state-name*
twenty	Plenty.	Whisper-base	*state-number*	*had-fun-a-lot*
A lot.	Oh, la.	Whisper-base	*had-fun-a-lot*	*out-of-scope*
A lot.	Oh, wow.	Whisper-small	*had-fun-a-lot*	*out-of-scope*
Two.	you	Whisper-small	*state-number*	*out-of-scope*
Four.	I'm going to see this floor.	Whisper-small	*state-number*	*out-of-scope*
twenty	Swamy?	Whisper-medium	*state-number*	*state-name*
Eight.	E.	Whisper-medium	*state-number*	*out-of-scope*

CONCLUSION

- This study investigates a modular SLU pipeline for kids with cascading ASR and NLU modules, evaluated on our first home deployment data with 12 kids at individual homes.
- For NLU, we examine the advantages of a multi-task architecture & experiment with numerous pretrained language representations for Intent Recognition and Entity Extraction tasks.
- For ASR, we inspect the WER with several solutions that are either low-power and local (Rockhopper), commercial (Google Cloud), or open-source (Whisper) with varying model sizes and conclude that Whisper-medium outperforms the rest on kids' speech at authentic homes.

SELECTED REFERENCES

[1] Okur, E., Sahay, S., Fuentes Alba, R., and Nachman, L. (2022). End-to-end evaluation of a spoken dialogue system for learning basic mathematics. *Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP), EMNLP 2022.*
[2] Okur, E., Sahay, S., Fuentes Alba, R., and Nachman, L. (2022). Inspecting Spoken Language Understanding from Kids for Basic Math Learning at Home. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), ACL 2023.*
[3] Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *Conversational AI Workshop, NIPS 2017.*
[4] Bunk, T., Varshneya, D., Vlasov, V., and Nichol, A. (2020). DIET: lightweight language understanding for dialogue systems. *CoRR, abs/2004.09936.*
[5] Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., and Vulić, I. (2020). ConveRT: Efficient and accurate conversational representations from transformers. *Findings of the Association for Computational Linguistics, EMNLP 2020.*
[6] Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML 2023).*