Educational Dialogue Systems for Visually Impaired Students: Introducing a Task-Oriented User-Agent Corpus

Elisa Di Nuovo – Manuela Sanguinetti – Pier Felice Balestrucci – Luca Anselma – Cristian Bernareggi – Alessandro Mazzei

EU JRC - Ispra (Italy); DMI, University of Cagliari (Italy); DipInfo, University of Turin (Italy); DipInfo, University of Turin (Italy); Polin Lab, University of Turin (Italy); DipInfo, University of Turin (Italy)

1 NOVAGRAPHS FSA CORPUS

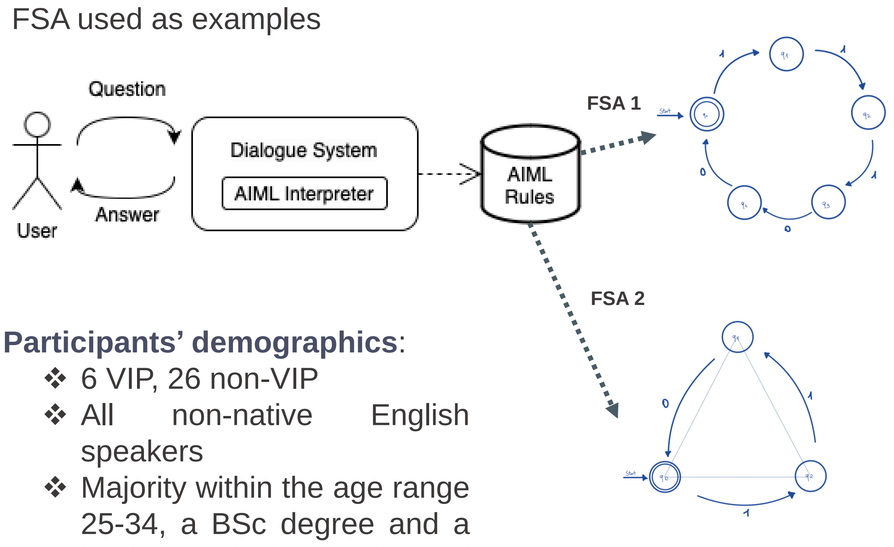
- Dialogues in **English** between users and a **task-oriented conversational agent**
- Interactions revolving around the description of **Finite State Automata (FSA)**
- Part of the **NoVAGraphS** project, aimed at developing tools for an easier access to educational content, in particular for **Visually Impaired People (VIP)** in STEM

In collaboration with: Integr-abile

http://www.integr-abile.unito.it/progetto-novagraphs/

2 DATA COLLECTION

Data obtained from interactions between users and a rule-based DS providing information on one of two FSA used as examples



- Collected in Spring 2023 (Phase 1) + August-September 2023 (Phase 2)
- Users interacted with the system via a web interface compliant with the **Web Content Accessibility Guidelines** (WCAG) 2.1 (VIP used a keyboard and a screen reader)

Participants' demographics:
- 6 VIP, 26 non-VIP
- All non-native English speakers
- Majority within the age range 25-34, a BSc degree and a background knowledge of FSA, male

All participated on a voluntary basis

Corpus metadata:
- Conversation ID
- Turn number
- Participant (User|System)
- Text
- VIP (yes|no)
- No. Tokens
- FSA ID

	VIP	Non-VIP	All
# Dialogues	6	26	32
# Turns	194	512	706
Turns/Dialogue	32.33	19.69	22.06
Tokens/User's turn	3.87	5.91	5.53

Basic statistics of the collected data

3 ANNOTATION DESIGN & RESULTS

LABELS

DIALOGUE ACTS: ISO 24617-2 Standard (Bunt et al., 2017, 2020).

Tagset: 5 dimensions (Task - Ta:, Auto-feedback - Auto:, Turn Management - TuM:, Dialogue Structuring - DS:, Own Communication Management – OCM:) + 6 communicative functions (setQuestion, request, propositionalQuestion, checkQuestion, suggest, answer, autoNegative)

Format: DiAML-TabSW (tabular format, instead of XML); multi-label tagging allowed

ERRORS: Adaptation of error annotation scheme from Sanguinetti et al. (2020). Errors intended as violations of gricean maxims (excl. Quality)
Tagset: Quantity: Lack|Excess of information; Relation: Ignoring question/feedback, Repetition; Topic change, Straight wrong response, Off topic (new). Manner: Indirect response, Non-understandable, Grammatical error. Generic: Non-cooperativity
Format: Tabular; multi-label tagging allowed

TASK

Turn-level annotation | Pair-wise (spreadsheet) | Max. 3 labels per turn

T17	Is there a transition from q0 to q5	Ta:propositionalQuestion	Ignoring question/feedback
T18	From q0 there is only one transition to q2 marked with 1.	Ta:answer	Indirect response
T19	According to your understanding of Automaton 2 what is its optimal spatial representation	Ta:setQuestion	None
T20	I dont' know. Try asking something like, "Describe the automaton."	AutoF:autoNegative; DS:suggest	Topic change
T21	The automaton accepts a language allowing words made of an odd number of 0s and 1s	Ta:checkQuestion	Ignoring question/feedback
T22	The language accepted by this automaton is made of zero or more words formed by a sequence ofa pair of 1s followed by a 0.	Ta:answer	Indirect response

Label distribution & inter-annotator agreement

Dialogue Acts Users	
Ta:setQuestion	51.54
Ta:request	27.45
Ta:propositionalQuestion	16.25
Ta:checkQuestion	1.96
OCM:selfCorrection	1.40
DS:opening	0.56
TuM:turnAccept	0.56
SOM:initGreeting	0.28
DS	
Ta:answer	48.14
AutoF:autoNegative	24.79
DS:suggest	24.79
Ta:suggest	2.27

Distribution (in %) of single DAs in users' and DS turns

Errors Users	
Repetition	42.75
Ignoring question/feedback	28.24
Grammatical error	13.74
Non-understandable	4.58
Off-topic	3,82
Lack of information	3.05
Non-cooperativity	2.29
Ill-formed	1.53
DS	
Topic change	62.30
Straight wrong response	14.66
Indirect response	10.47
Excess of information	7.33
Lack of information	4.71
Ignoring question/feedback	0.52

Distribution (in %) of single errors in users' and DS turns (calculated over the total of errors)

Cohen's k

	Phase 1	Phase 2
DAs	0.74	0.96
• Communicative Function	0.91	0.96
• Dimension	0.56	0.96
Errors		0.56 (whole corpus)

Inter-Annotator Agreement

4 EXPERIMENTS

Users' DA Annotation with DIET classifier

We addressed the task of recognizing users' DAs structuring this process as an intent detection task with DIET, the default classifier provided with RASA

	BoW	BERT$_{emb}$
Ta:setQuestion	0.9852	0.9963
Ta:request	0.9645	0.9655
Ta:propositionalQuestion	0.9293	0.9583
macro-F1	0.9597	0.9734

Pipelines:
- bag-of-word representation of character n-grams (1 to 4), already set in the default configuration
- BERT pre-trained embeddings

In both configurations:
- 100 epochs, 5-fold cross-validation, 3 runs; selection of the three most frequent users' DAs due to insufficient data instances for training purposes using DIET

Annotating DAs with LLMs

Experiments on LLMs ***zero-shot*** performance in annotating DAs

Models tested:
- ChatGPT 3.5 (via web interface)
- Llama 2
- Tk-Instruct

Prompt building:
All prompts comprised the following points:
- task definition (i.e. annotating DAs in turns)
- context (i.e. the dialogue between a student and a DS programmed to answer about a specific FSA)
- annotation constraints (e.g. one or more DAs can be assigned to each turn, additional co-text)
- presentation of the label set
- DAs decomposition into dimension and communicative function
- propositional questions to guide the selection of the most appropriate dimension and communicative function
- the turn to annotate

Protocol:
10 turns to annotate; task divided into two main steps: 1) dimension; 2) communicative function.
Same prompt run three times and independently on each turn
For each turn the maximum attainable score is computed, considering 1 point per each gold DA (0.5 points per subtask: dimension and communicative function).
Assigned 1 point per complete, correct DA, 0.5 if only one subtask is correct, and 0 if the annotation is incorrect or if the model failed to interpret the prompt correctly.

Prompt	ChatGPT	Llama 2	Tk
1	0.33/1	0.00/1	0.00/1
2	0.33/1	0.00/1	0.50/1
3	1.00/1	0.00/1	0.50/1
4	0.00/1	0.17/1	0.00/1
5	0.50/1	0.00/1	0.00/1
6	0.17/1	0.00/1	0.50/1
7	0.83/2	0.00/2	0.50/2
8	1.00/1	0.00/1	0.50/1
9	1.00/2	0.00/2	0.50/2
10	1.17/3	1.00/3	0.50/3
average	0.63/1.40	0.11/1.40	0.35/1.40
SD	0.41/0.70	0.31/0.70	0.24/0.70
success	45.24%	8.33%	25.00%

DATA AVAILABILITY

Corpus available for research purposes by filling in a form available in the resource repository (see QR code)

SCAN ME

Additional data available along with the corpus:
- Two PNG files with the graphical representation of the automata
- Two HTML files containing the state tables of the automata

References

- Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the iso 24617-2 standard. In Deborah A. Dahl, editor, *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*, pages 109–135. Springer.
- Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The iso standard for dialogue act annotation. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*, pages 549–558. ELRA
- Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalerandi, Dario Mana, and Rossana Simeoni. 2020. Annotating errors and emotions in human-chatbot interactions in Italian. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 148–159, Barcelona, Spain. Association for Computational Linguistics