Salted CX

Dialogue Analysis

And its use in Customer Experience

Martin Váňa, Adam Zíka; 2.11.2023 linkedin.com/in/mvana

About Salted CX

For contact centres where humans and machines work together

- Early-phase SaaS startup (1 year old)
- Experienced team with a successful exit in 2018
- Fully committed to delivering value to customers
- Passionate about analytics, ML/AI, and complex vertical integration

Brief history

Team evolution

- Started in ZOOM International (Eleveo)
- Ytica spinoff: 15 people, successful exit in 2018 (18mo)
- Twilio - Large US company, grew from 750 to ~9K in four years
- Salted CX, 11 people, growing and hiring, mostly engineers now

Contact centres data

To have an idea what we are dealing with

- Call centre agent can spend up to 5h a day on the phone => ~100h / mo
- Contact centre agents can handle up to 10 parallel chat conversations
  - This can be up to 20K tokens
- In many countries CCs are required to collect the data by law.

- Sample: avg. agent on messaging channel
  - ~ 15K tokens
  - ~ 60 conversations
  - ~ 350 turns

What have we learned about CC data

And where have we failed…

- The data collection is costly (Storage!)
  - Adding more costs by processing is hardly an advantage
- Changing hours of audio to piles of transcripts does not solve anything
- Individual results are too fine grained and hardly actionable
  - Limited precision/recall often renders solutions useless
  - Especially if obtained with a large delay, but realtime is expensive

Future of contact centres

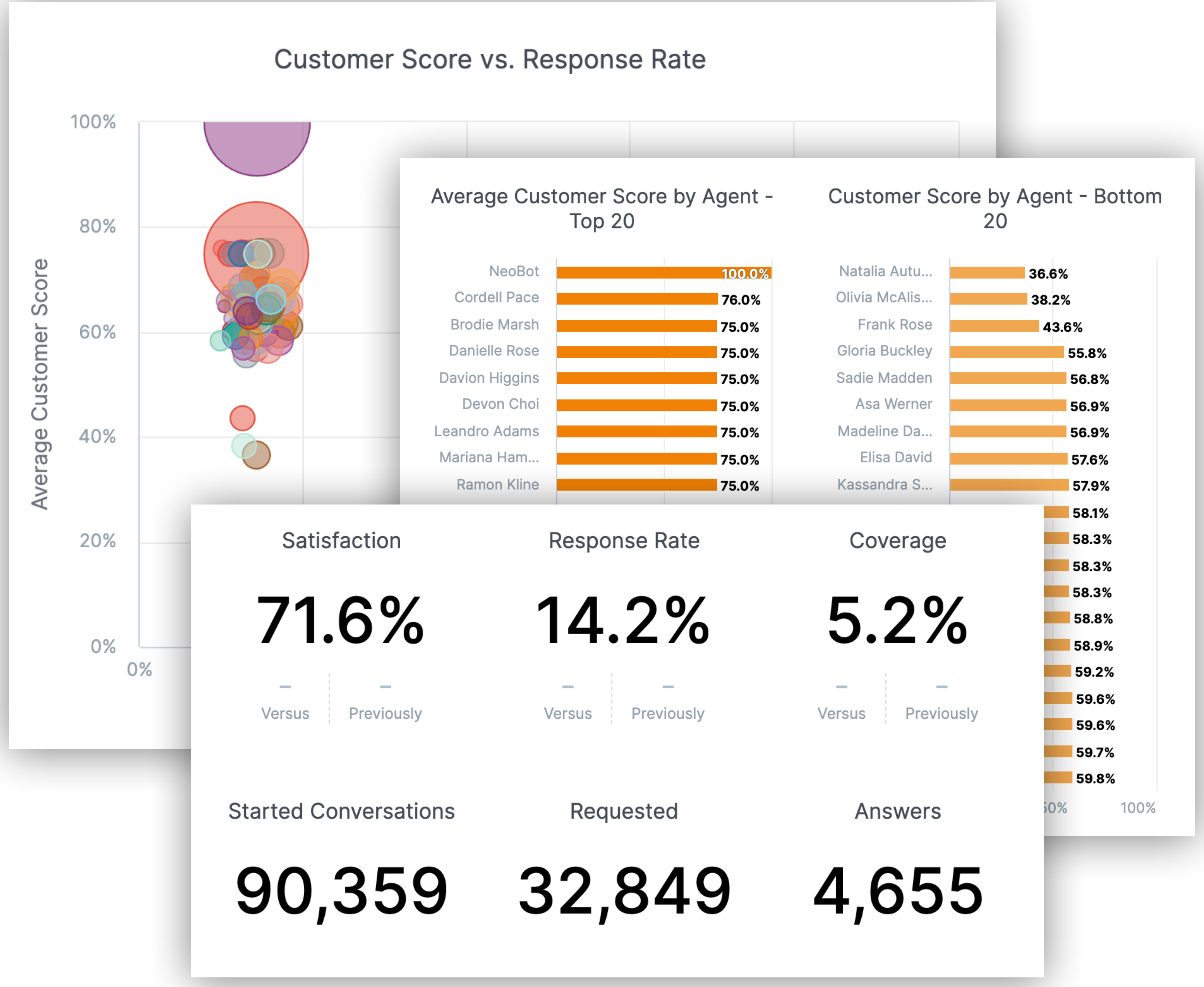
What we dare to guess

- Automation will be the main topic
  - But are we there already?
  - Are the CCs ready for the transition?
  - Is it safe?
  - How will the transition impact the business?
  - If automation reduces price, more data may be collected

Statistical look on CC conversations

How to process the data as a whole

- Reporting
    - Advanced metrics and dashboards
    - Mostly on metadata
    - ML based labelling and categorisation
    - Automated Quality assurance
- Graphical representation of the whole conversation corpora!



CX

Collaboration (JSALT 2023)

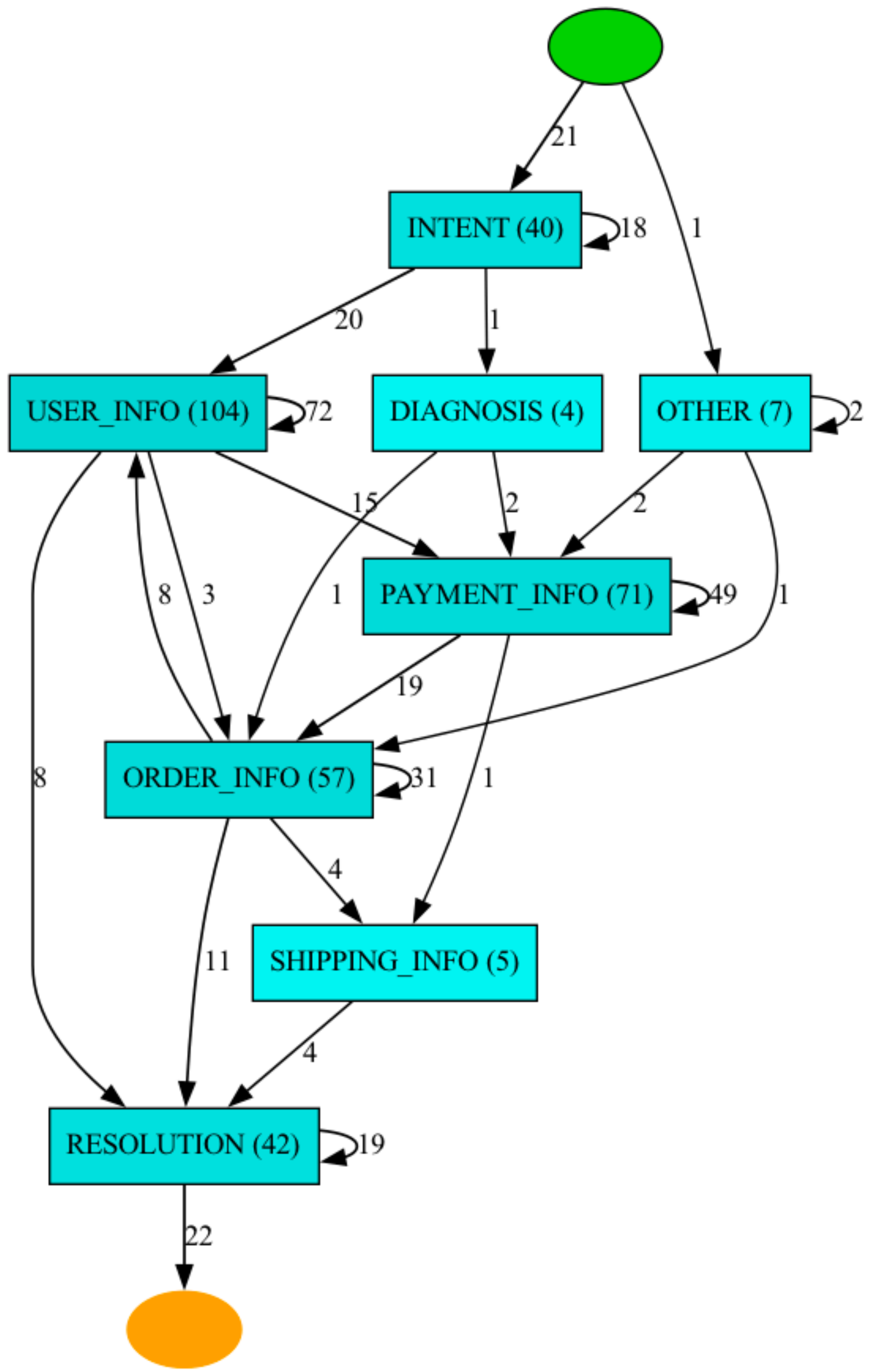
What can a startup gain from open source development

- Two weeks of lectures, six weeks of work
- In total ~10y of work of mostly senior staff
- Four tracks, four teams
- Track no. 4: *Automatic design of conversational models from observation of human-to-human conversation*

Dialog corpus as a graph

How to compile all dialogs to a single structure

- Representation of the whole corpus
- Described by *nodes* and *arcs*
- "Similar states" collapsed to a single node representing content
- Dialogue as a path in the graph, arcs represents transitions



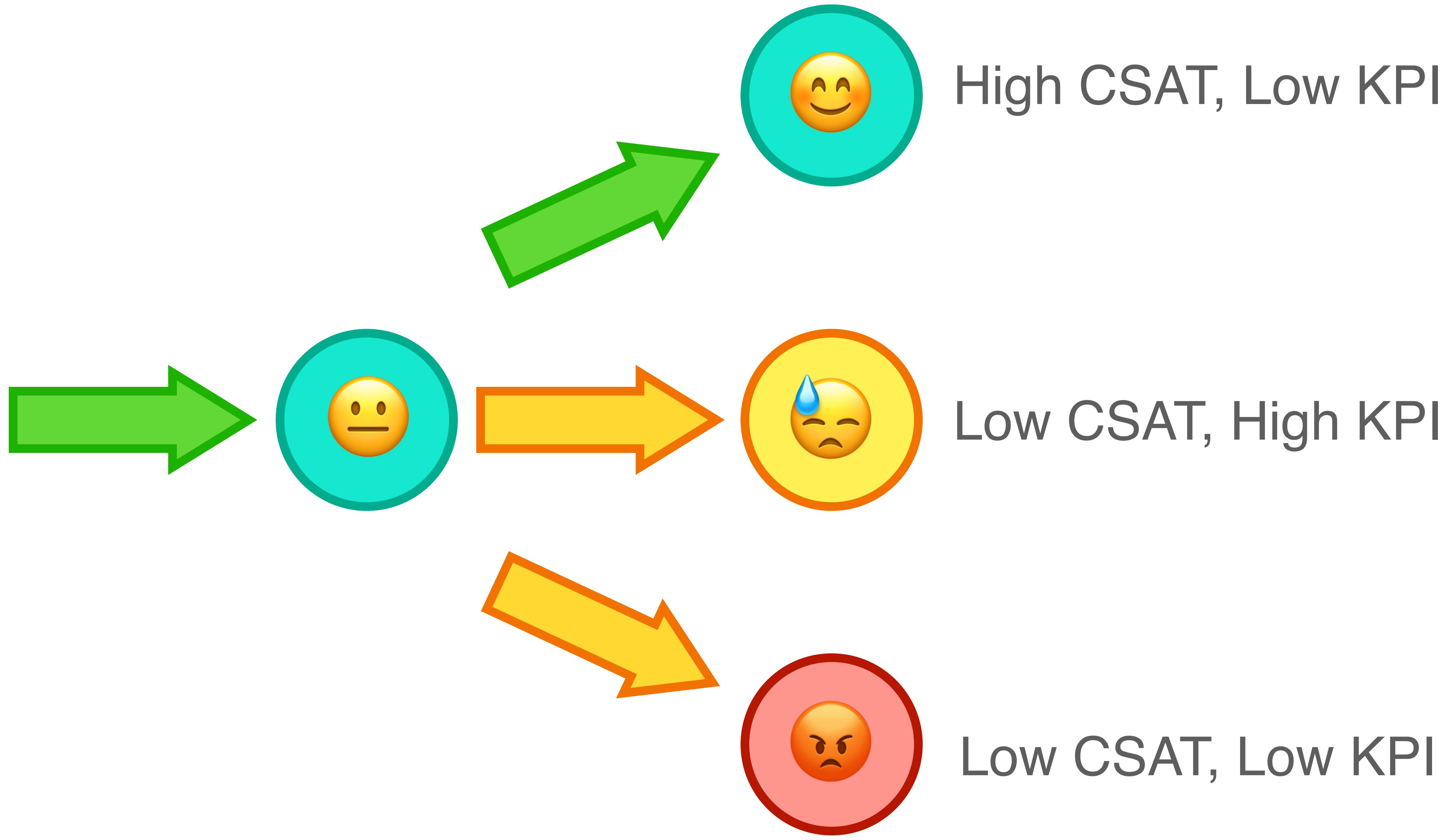
Manon Macary, Pheobe Wong, Allo-Media dataset (unpublished) JSALT 2023

Downstream tasks for dialogue graph

What can we learn from this representation

- Outcome prediction
- Best response selection (agent assist, automation)
- Business process optimisation

- Transition based on
  - Probability
  - Outcome
  - Business KPIs



Dialogue representation: nodes & arcs

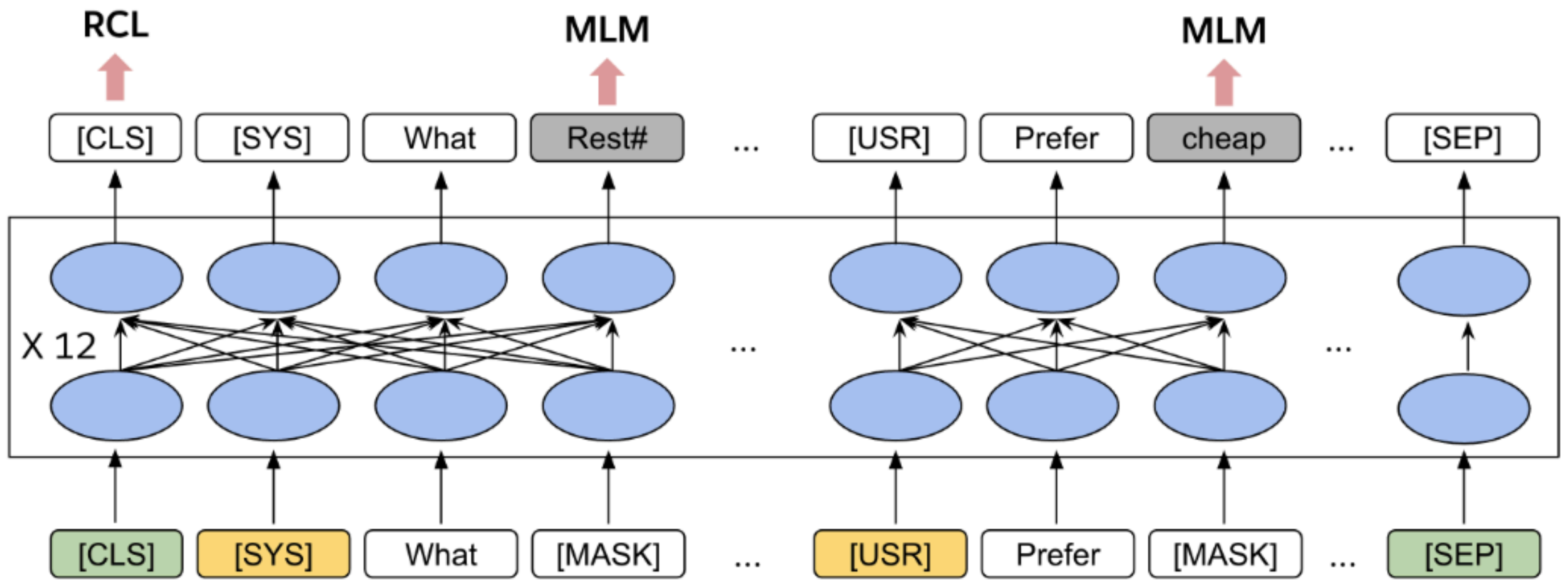
What does similar turns mean?

- Embeddings
  - Encoder —> Sentence BERT, ToD BERT, DSE, …
  - Similarity —> cosine, dot product, …
  - Clustering —> kMeans, Density Based…
- States Dictionary
  - Intent detection —> Classificators, but also generative models
  - CodeBooks

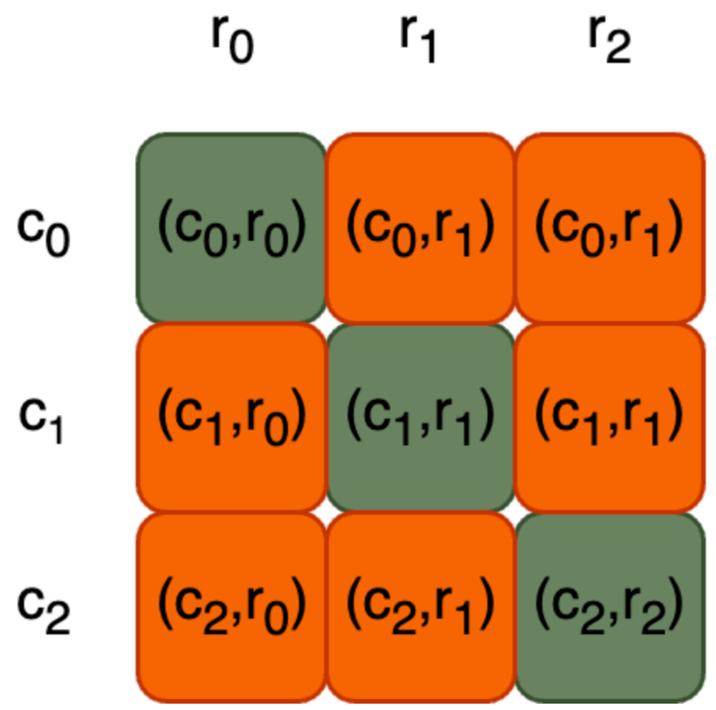
Sentence Embeddings

Improved representation

- TOD-BERT
  - 1.4M utterances from 9 TOD datasets.
  - In-batch negatives via response contrastive loss



$$L_{rcl} = -\sum_{i=1}^{b} \log M_{i,i},$$
$$M = \text{Softmax}(CR^T) \in \mathbb{R}^{b \times b}.$$



Wu, Chien-Sheng, et al. "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue." 2020 ACL

Clusterings

How to reduce the amount of the nodes

- Fixed class count
  - kMeans: Simple, via *scikitlearn*
  - SVRNN: *VAE with Gumbel Softmax* (Vojta Hudeček)
  - Wav2vec: Automated codebook as latent representation (Martin Nykodem)
- Variable class count
  - Density clustering
  - In real world case very hard to configure
  - Mostly ends in one class :(

Dialogue State Tracking (DST)

Ondřej Plátek (JSALT20203)

- Show representation for slots on examples:
  - what was talked about?
  - what the user wants (this time)?
  - what the system attempted to do?

USER:	*I would like a cheap restaurant.*	inform(price=cheap)
SYSTEM:	*Golden plate is cheap in city-center.*	inform(name=Golden plate)
USER:	*What is the cuisine?*	request(cuisine)
SYSTEM:	*They serve chinese food.*	inform(cuisine=chinese)
USER:	*Sounds good. Bye!*	goodbye()
SYSTEM:	*Have a great day.*	goodbye()

Dialogue State Vector

Vector tracking across the dialogue

- Codebook of states derived from intent detection
- Clear understanding of the states
- Combinatoric explosion of state count with intents AND time!
- We should track changes only!

Dialogue	Dialogue State Slot Value
[usr] Can you please help me find a place to go? [sys] I've found 79 places for you to go. Do you have any specific ideas in your mind?	[0, 0, 0] $\rightarrow$ ⓪ ['', '', '']
[usr] I'd like a **sports** place in the **centre** please. [sys] There are no results matching your query. Can I try a different area or type?	[0, 1, 1] $\rightarrow$ ① ['', 'sports', 'centre']
[usr] Okay, are there any **cinemas** in the centre? [sys] We have vue cinema.	[0, 2, 1] $\rightarrow$ ② ['', 'cinemas', 'centre']

Liang Qiu, Chien-Sheng Wu, Wenhao Liu, Caiming Xiong, "Structure Extraction in Task-Oriented Dialogues with Slot Clustering", https://arxiv.org/abs/2203.00073

Comparing clusterings

How to tell you have improved?

	ARI					AMI					SC				
	Taxi	Rest.	Hotel	Attr.	Train	Taxi	Rest.	Hotel	Attr.	Train	Taxi	Rest.	Hotel	Attr.	Train
Random	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	-	-
VRNN	0.05	0.00	0.00	0.00	0.00	0.05	0.02	0.00	0.01	0.06	-	-	-	-	-
BERT-KMeans	0.02	0.01	0.01	0.01	0.01	0.11	0.09	0.02	0.03	0.06	0.11	0.08	0.06	0.13	0.09
TOD-BERT-mlm	0.02	0.01	0.01	0.03	0.02	0.13	0.11	0.03	0.06	0.10	0.12	0.08	0.06	0.17	0.09
TOD-BERT-jnt	0.03	0.02	0.02	0.03	0.03	0.16	0.13	0.06	0.08	0.14	0.09	0.08	0.06	0.13	0.07
BERT-spaCy	0.01	0.06	0.04	0.01	0.01	0.09	0.18	0.12	0.06	0.08	-	-	-	-	-
TOD-BERT-spaCy	0.01	0.03	0.05	0.02	0.01	0.09	0.15	0.12	0.05	0.05	-	-	-	-	-
TOD-BERT-SBD$_{MWOZ}$	**0.15**	0.00	0.00	0.00	0.05	0.17	0.13	0.04	0.06	0.16	**0.39**	**0.34**	**0.27**	**0.44**	**0.34**
TOD-BERT-DET$_{ATIS}$	0.08	0.05	0.09	0.03	0.06	0.26	0.22	0.25	0.15	0.26	-	-	-	-	-
TOD-BERT-DET$_{SNIPS}$	0.06	0.05	0.11	0.03	0.04	0.25	0.23	0.22	0.09	0.22	-	-	-	-	-
TOD-BERT-DET$_{MWOZ}$	**0.15**	**0.22**	**0.24**	**0.33**	**0.24**	**0.39**	**0.48**	**0.44**	**0.44**	**0.44**	-	-	-	-	-

Table 5: Structure extraction results using clustering metrics in the MultiWOZ dataset. SC is omitted for methods that do not encode utterances directly. Results using BERT-Birch and BERT-Agg are reported in Appendix A.

Reality check

Srikanth Madikeri & Miroslav Hlaváček (JSALT2023)

- Simple Clustering methods are nearly useless
- Even DST produces too many nodes to make the graph actionable
- Testing on MultiWoZ

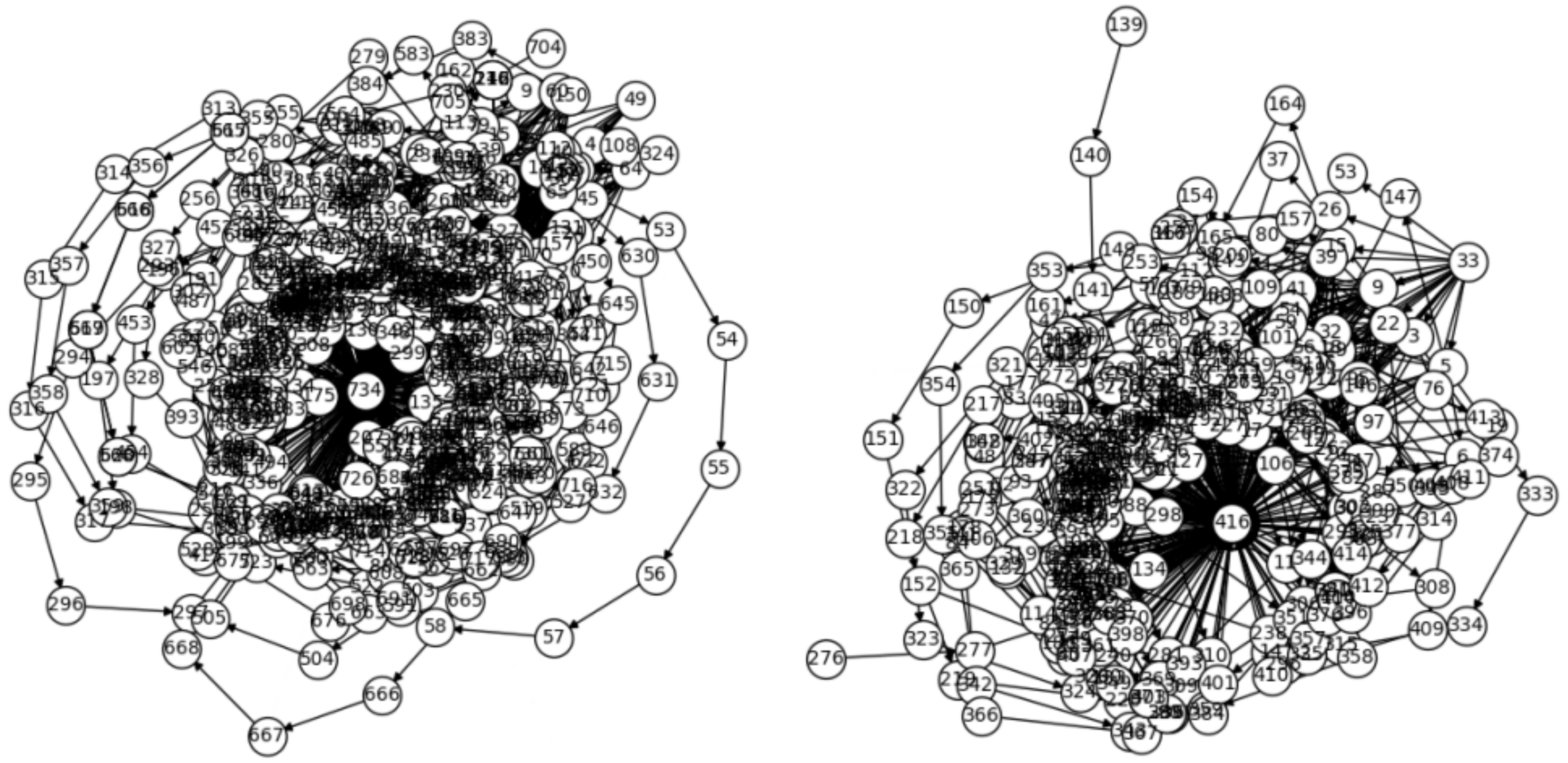


Figure 7: Dialogue structure in the *hotel* domain of the MultiWOZ. The structure on the left is from annotated dialogue states, while the right one is extracted by our approach.

Liang Qiu, Chien-Sheng Wu, Wenhao Liu, Caiming Xiong, “Structure Extraction in Task-Oriented Dialogues with Slot Clustering”, https://arxiv.org/abs/2203.00073

How about generative model as labeller

Gábor Baranyi (JSALT2023)

- Llama2-7b-chat-hf
- Only little issues with the JSON formatting
- Model was apparently trained on similar data (open sets)

PROMPT:

{dialog}

The model results of the previously given dialog is a list.

Each element of the list is a triplet with the following properties: type, name, value

"type" can be either "intent" or "slot" and "name" and "value" properties are related to the "type".

"name" and "value" are single words in a string, present in the given dialog.

Results only, no explanation. The results are given in valid, standard JSON format.

JSON Results: [{{

Simple example (MultiWoZ)

What are we really dealing with

USER: Can you please help me find a place to go?

SYSTEM: I've found 79 places for you to go. Do you have any specific ideas in mind?

USER: I'd like a sports place in the centre please.

{'type': 'intent', 'name': 'place', 'value': 'Central Station'}

{'type': 'intent', 'name': 'sports', 'value': 'football'}

{'type': 'slot', 'name': 'location', 'value': 'Newcastle'}

{'type': 'intent', 'name': 'go', 'value': 'to'}

{'type': 'slot', 'name': 'sports venue', 'value': "St. James' Park"}

{'type': 'intent', 'name': 'visit', 'value': 'stadium'}

{'type': 'slot', 'name': 'time', 'value': 'evening'}

We need fine-tuning

LoRa, qLoRa, VeRa, and so on…

- Much smaller adapters than the original models
- qLoRa fine tunes well
- Tested on summarisation for open dataset (DialogSum)
- on nVidia A100 one task takes around 3sec, ~1000/h
- On GCP ~ $4/h per instance

Model	Rouge1	Rouge2	RougeL
Human	53.4	26.7	50.8
Top scoring model(s) from DialogSum Challenge 2022	47.6	21.7	45.9
Llama2-7B	13.6	4.2	10.6
Tuned Llama2-7B [25 samples]	13.3	4.0	10.1
Tuned Llama2-7B [50 samples]	12.4	3.5	9.5
Tuned Llama2-7B [100 samples]	10.3	1.9	7.8
Tuned Llama2-7B [200 samples]	22.7	7.2	17.6
Tuned Llama2-7B [400 samples]	32.5	11.1	25.5
Tuned Llama2-7B [all samples]	39.8	15.3	31.8
Tuned Llama2-7B [all samples] - No LORA	41.1	16.3	33.6
Llama2-13B	19.3	6.4	15
Tuned Llama2-13B [100 samples]	26.3	9.4	20.8

Adam Zíka, Salted CX 2023

Llama2/ChatGPT reality

How much it would cost?

- Per conversation analysis ~ 250 tokens + ~ 100 tokens of prompt
  - Can be up to $1 per agent/day
- Per turn analysis ~ 45 tokens + ~ 100 tokens of prompt
  - Can be up to $3 per agent/day
- But we have typically many tasks to be performed!
  - Many prompts to be executed!

How about FST

Finite State Transducers, Miroslav Hlaváček (JSALT2023)

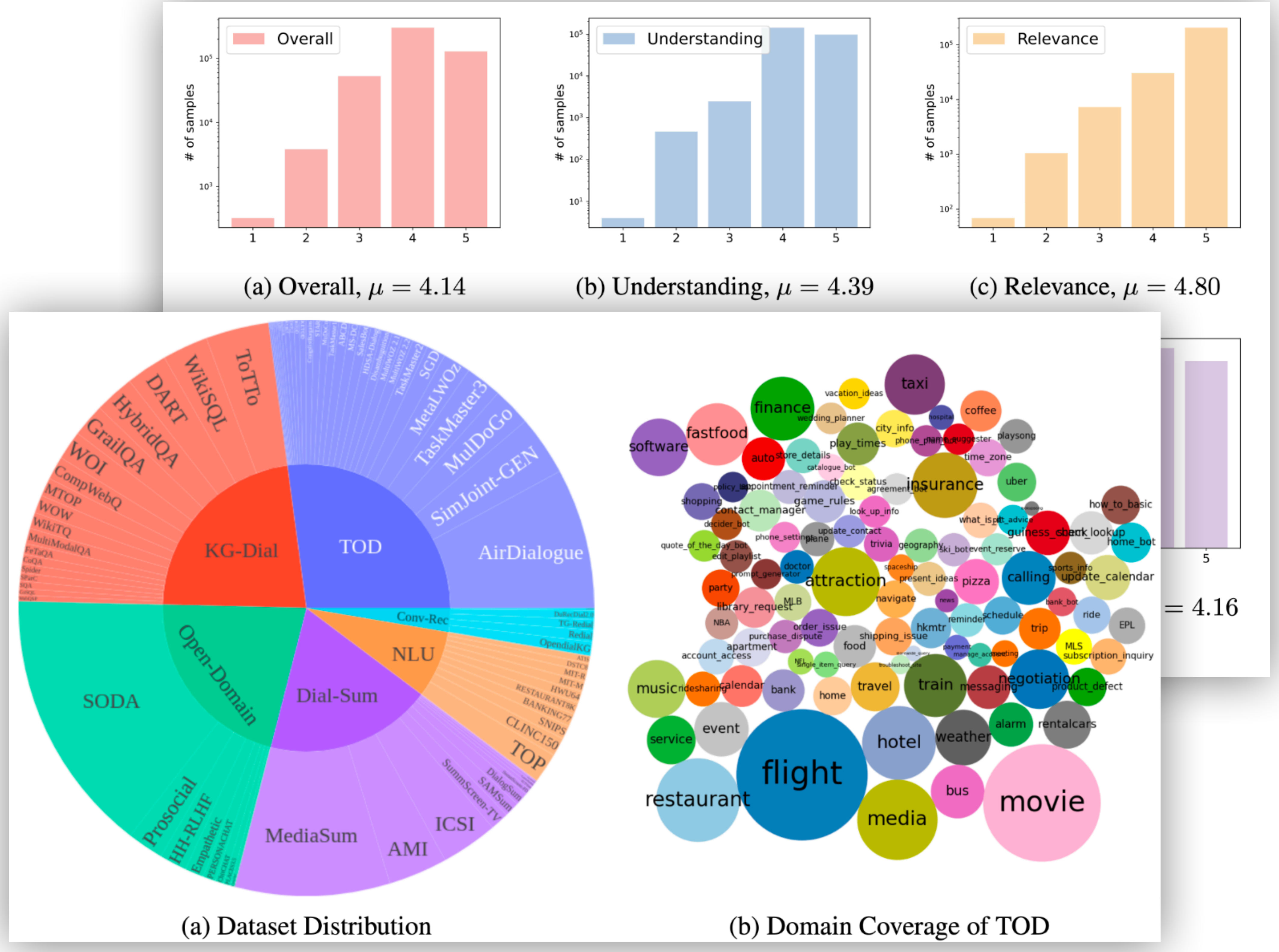
- Natural representation for the graph structure
  - Encode user/system turns on the arcs
  - Quantise turns to a small set of classes
  - Reduce graph via composition, determinization, minimization, and pruning
- Open source tooling
  - OpenFST
  - Julia

Miroslav Hlaváček, Allo-Media dataset (unpublished) JSALT 2023

Open dialogue datasets

What are we using for generic models

- Many on *huggingface* but way too fractured
- Format varies heavily across the sets
- DialogStudio (Salesforce)
  - One place for many datasets
  - Unfortunately not normalised!
- **Any open dataset must be considered as training for most LLMs!**



Jianguo Zhang et al., DialogStudio: Towards Richest and Most Diverse Unified Dataset Collection for Conversational AI, arXiv:2307.10172 [cs.CL]

LLM as data generator

And how that can help us (Adam Zíka)

- We can use very large models as we do not generate often
- We can use generator prompt inputs as labels (intents, emotion, …)
- We do not intend to use data for training, but for testing
- Very helpful for showcasing your product

Natalia Autumn

Hi there! Welcome to Demo Adventures. How can I help you today?

Customer +2m

Yeah, I want to book a trip to Hungary. I saw this really cool ad online and I really want to go.

Natalia Autumn +6m

Oh, unfortunately, we don't have any trips to Hungary available. That offer has expired.

Customer +5m

What? Why didn't you tell me that earlier? I've been looking forward to this trip for weeks!

Natalia Autumn

I apologize, but we only offer trips to countries where our packages are available. Hungary is not one of those countries.

Customer +5m

Can't you just make an exception? I really want to go to Hungary.

Natalia Autumn +1m

I understand, but I can't make exceptions to our policies. It's not fair to other customers if we start making exceptions for certain people.

Conclusion

What have we learned so far

- Learning dialog structure is quite a hard task
  - Simple clustering methods do not work even for easiest cases
  - Dialog state tracking is too fine-grained to work
  - Generative models may show better results but large models are expensive

- Results for open datasets are skewed as the data was used to pre-train large models
- Brute force approach does not scale well with task count!

Thank you

We are hiring!