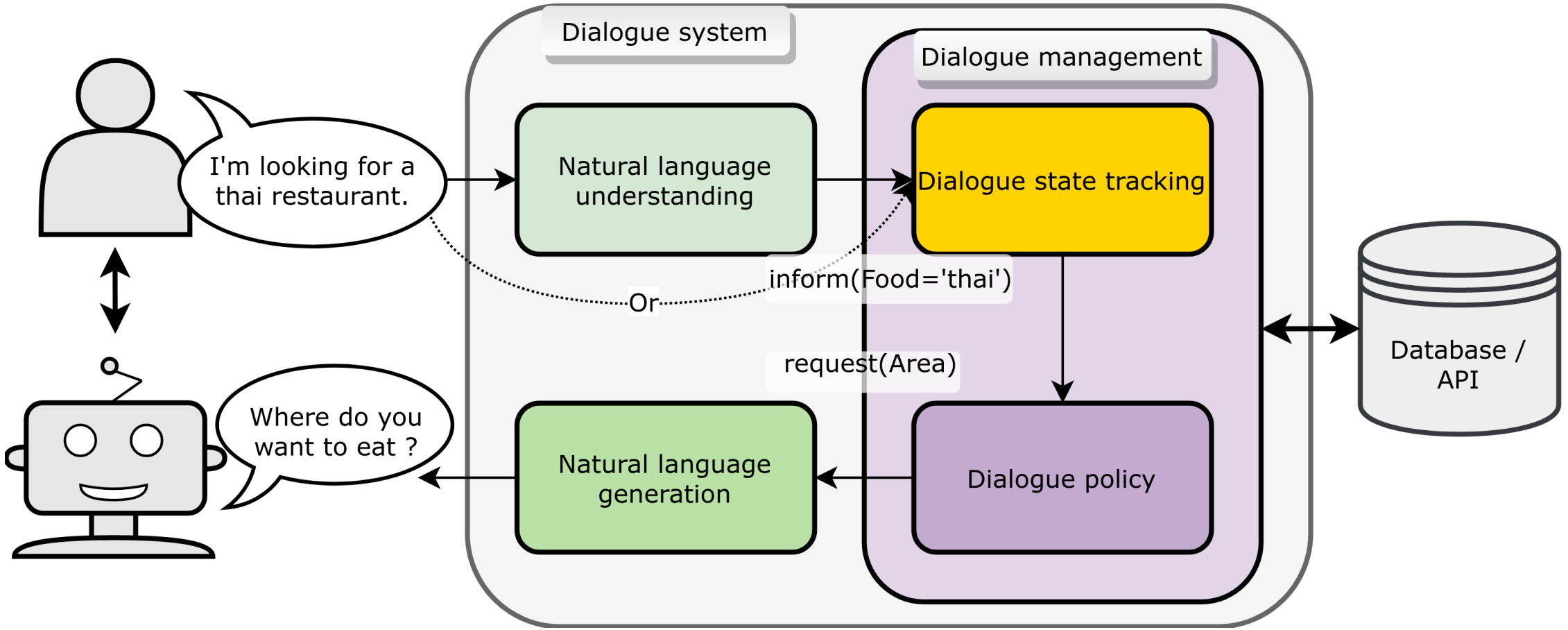
OLISIA: a Cascade System for Spoken Dialogue State Tracking

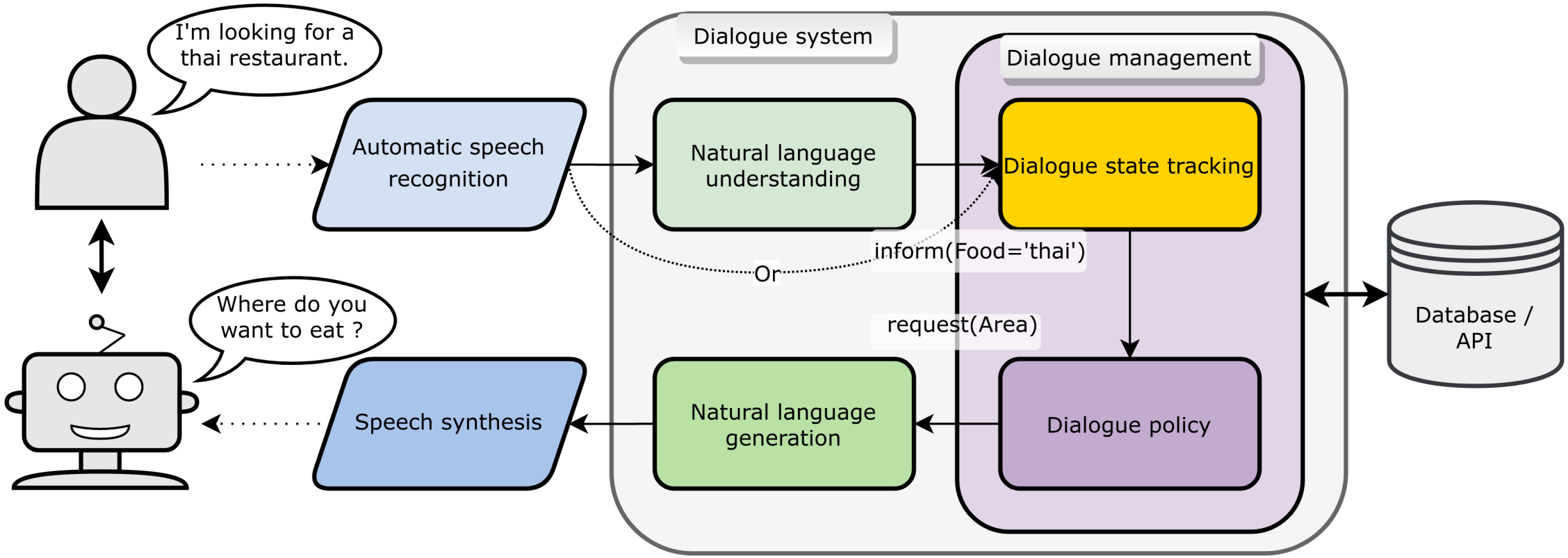
Lucas Druart, Léo Jacqmin
Benoit Favre, Lina M. Rojas-Barahona, Valentin Vielzeuf

Background

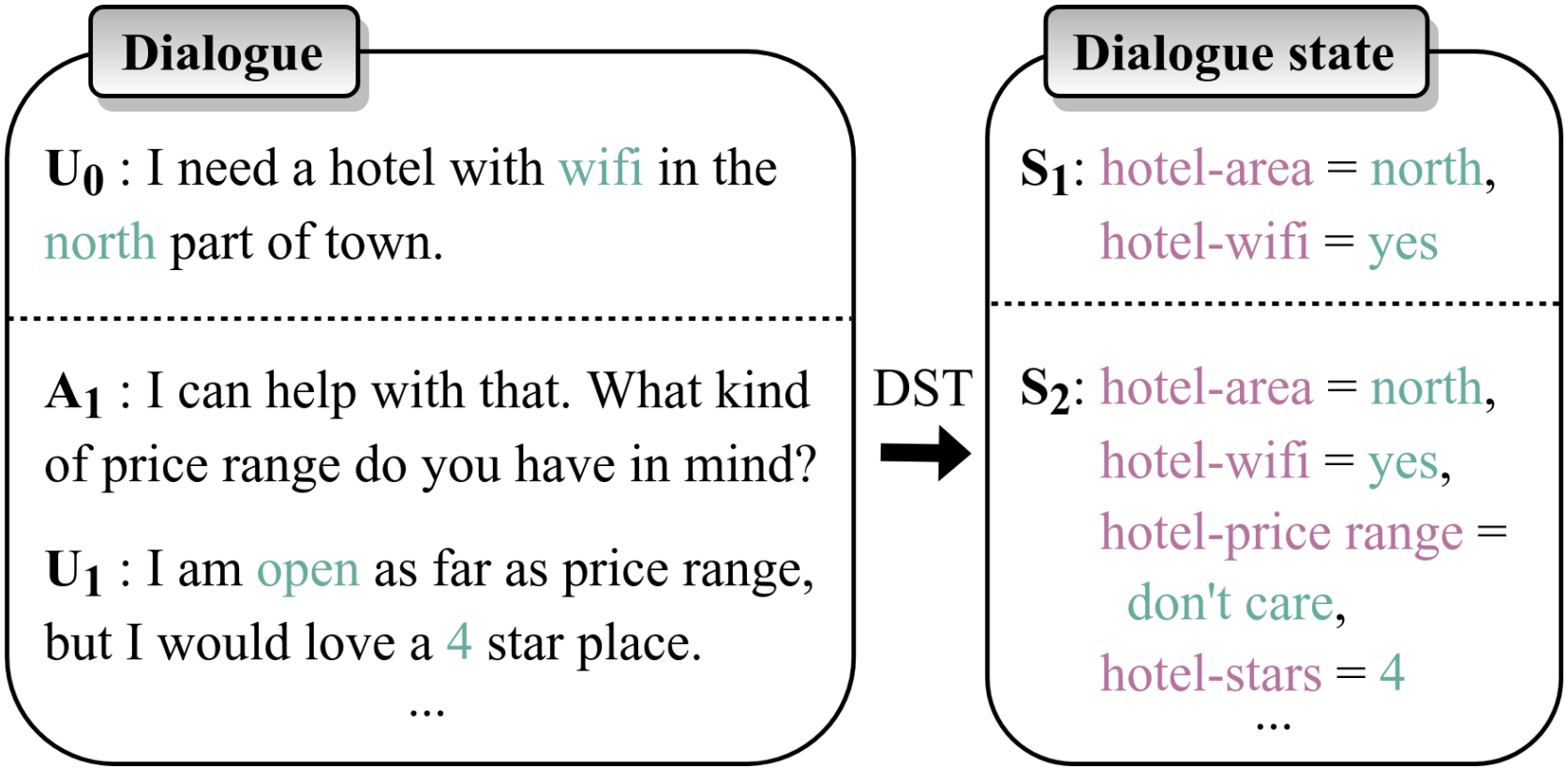
Task-Oriented Dialogue System



Task-Oriented Dialogue System: Spoken Dialogue

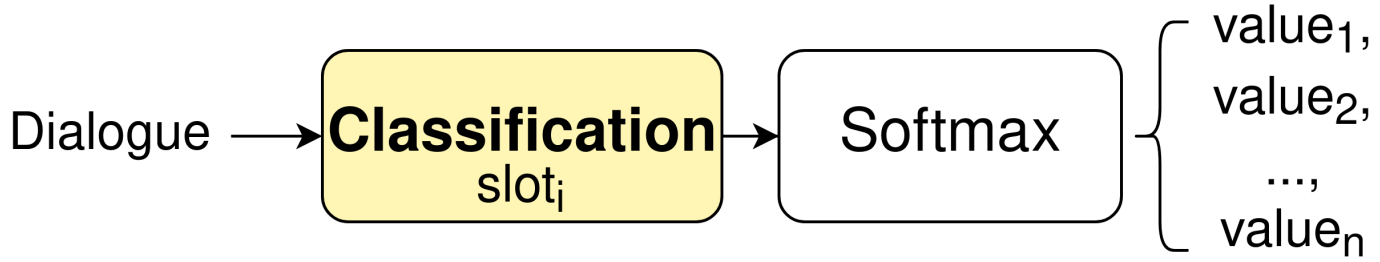


Dialogue State Tracking


Dialogue
U0 : I need a hotel with wifi in the north part of town.
A1 : I can help with that. What kind of price range do you have in mind?
U1 : I am open as far as price range, but I would love a 4 star place.
...
DST
Dialogue state
S1: hotel-area = north, hotel-wifi = yes
S2: hotel-area = north, hotel-wifi = yes, hotel-price range = don't care, hotel-stars = 4
...

Common Approaches

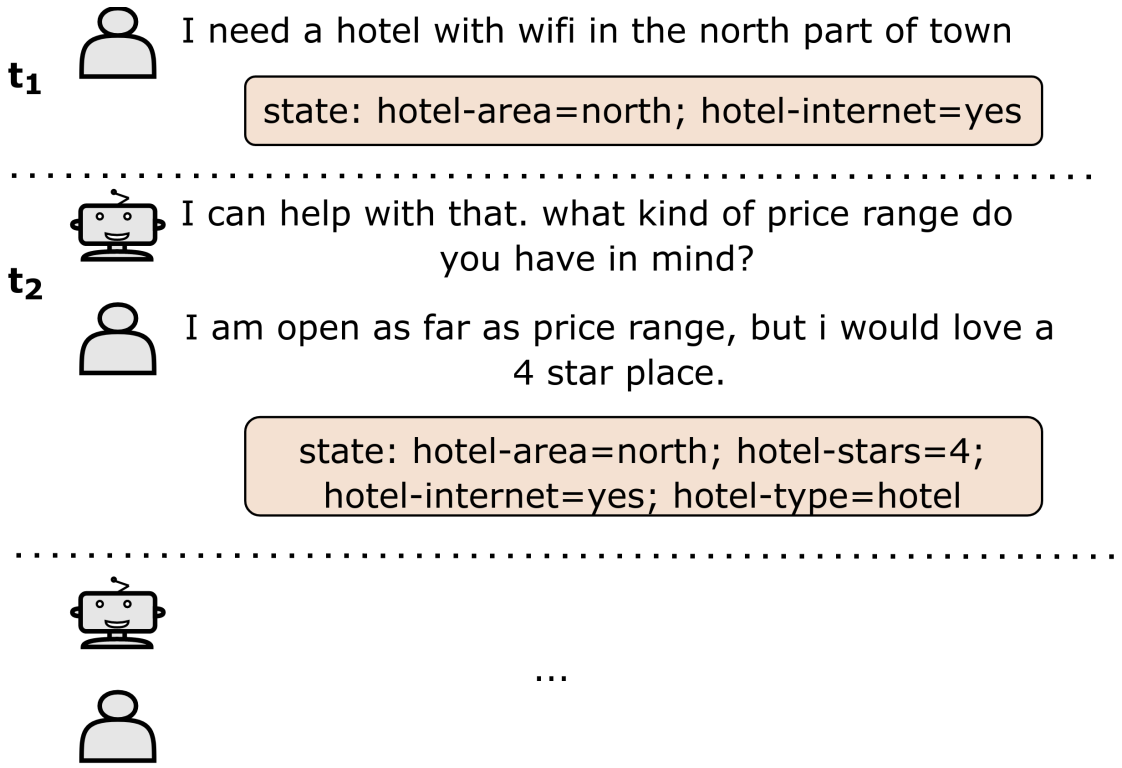
Fixed ontology



Open vocabulary

MultiWOZ Dataset

- No. dialogues: 8438 / 1000 / 1000 (train/dev/test)
- 7 domains:
  train, hotel, restaurant, etc.
- 25 slots
- Avg. turns / dialogues: 13.7



MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling (Budzianowski et al., 2018)

DSTC11 Track 3

Speech-Aware Dialog Systems Technology Challenge

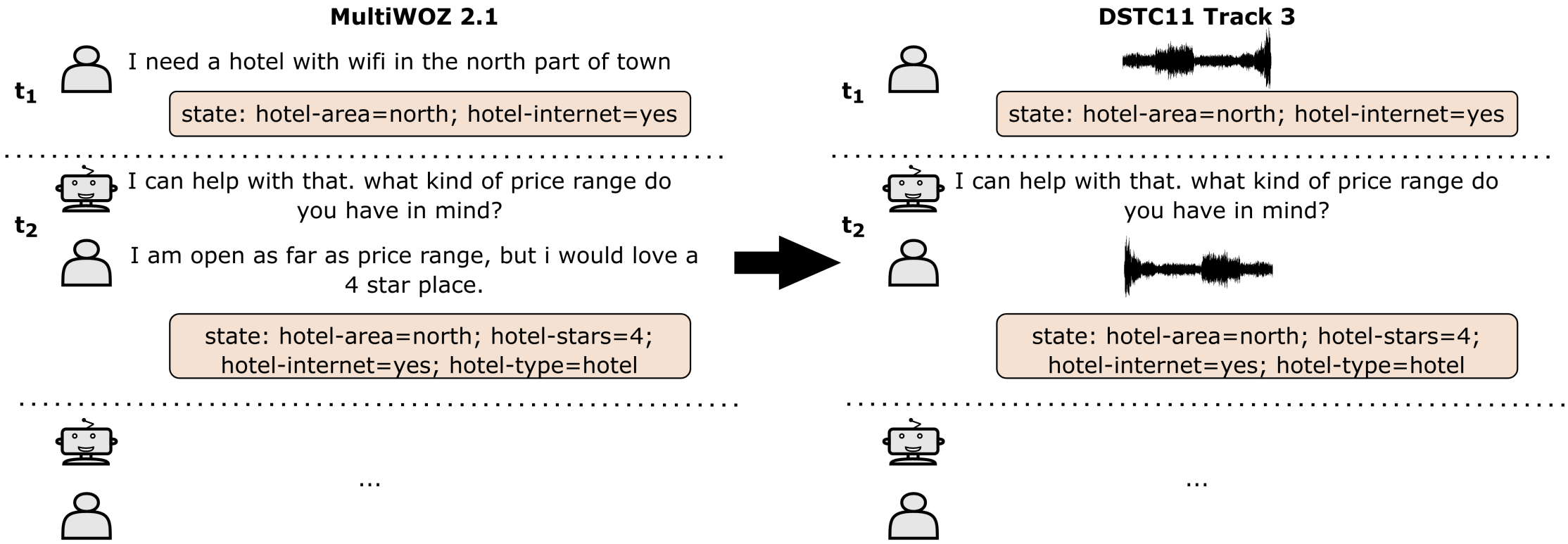
Motivation

- Recent work on task-oriented dialogue systems focused on chat corpora
- Recently, renewed interest in spoken dialogue systems
- Current systems trained on written corpora face robustness issues associated with spoken dialogue, such as ASR errors and disfluencies

Challenge

Released a spoken version of MultiWOZ (2.1)

- Evaluate task-oriented dialog systems end-to-end, from users' spoken utterances to inferred slot values
- Focused on DST since it is more impacted by the switch from written to audio input than response generation
- Modified dev and test sets:
  - New values introduced to encourage generalization for slots:
    `hotel-name, restaurant-name, train-departure, train-destination`
  - Time mentions offset by a constant amount



3 Versions

- **TTS Verbatim**: train, dev & test
- **Human Verbatim**: dev & test
- **Human Paraphrased**: test

Evaluation

Joint Goal Accuracy

$$JGA = \frac{C}{N_t} = \frac{No.\ of\ correct\ dialogue\ state\ predictions}{No.\ of\ turns}$$

–

Slot Error Rate

$$SER = \frac{S+D+I}{N_s} = \frac{No.\ of\ slot\ errors}{No.\ of\ slots\ in\ reference}$$

Our Participation

System Architecture

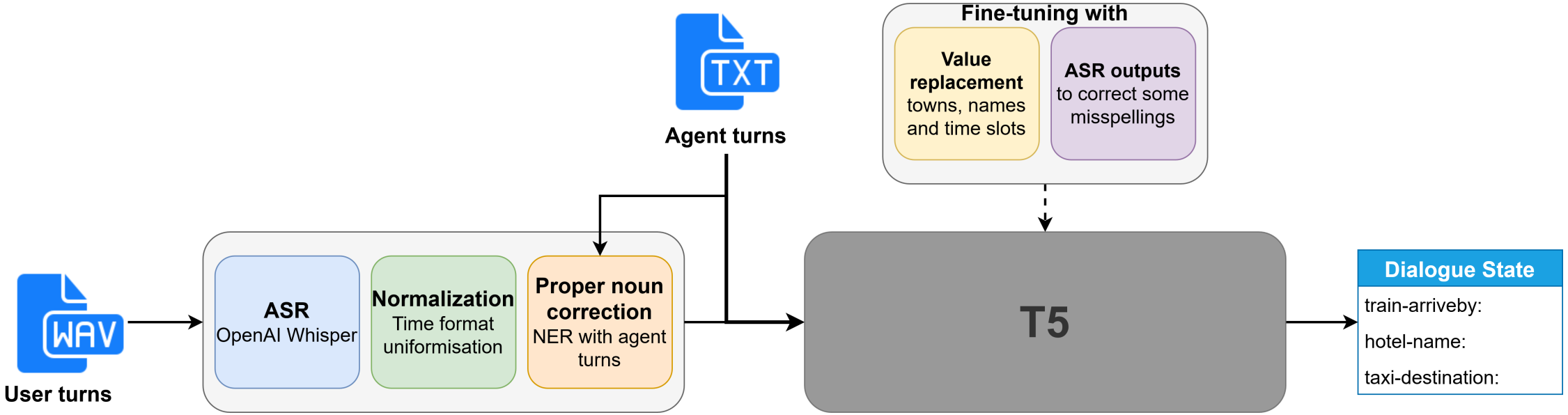
Cascade system

- ASR to transcribe user turns
- DST with text as input

End-to-end system

- Difficulties fusioning speech and text

System Architecture



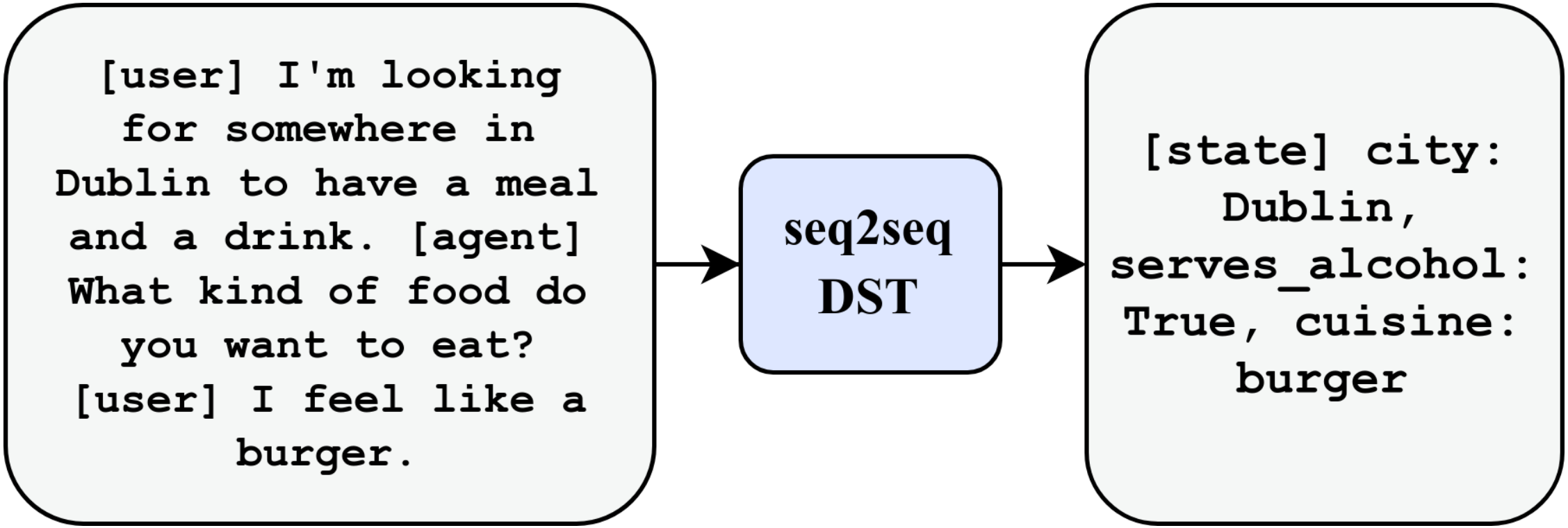
ASR

Whisper Large[1] to transcribe user turns + adaptations:

- Reranking => did not improve much
- Time normalization [hour]:[minutes] [am|pm]
- Value correction
    i. Named Entity Recognition on user and agent turns
    ii. Scoring of each pair of user and agent named entity with Character Error Rate
    iii. If score below tuned threshold
        - Then replace the mispelled user turns' named entities with their match from agent turn
        - Otherwise leave unchanged

[1]Robust Speech Recognition via Large-Scale Weak Supervision (Radford et al., 2022)

DST

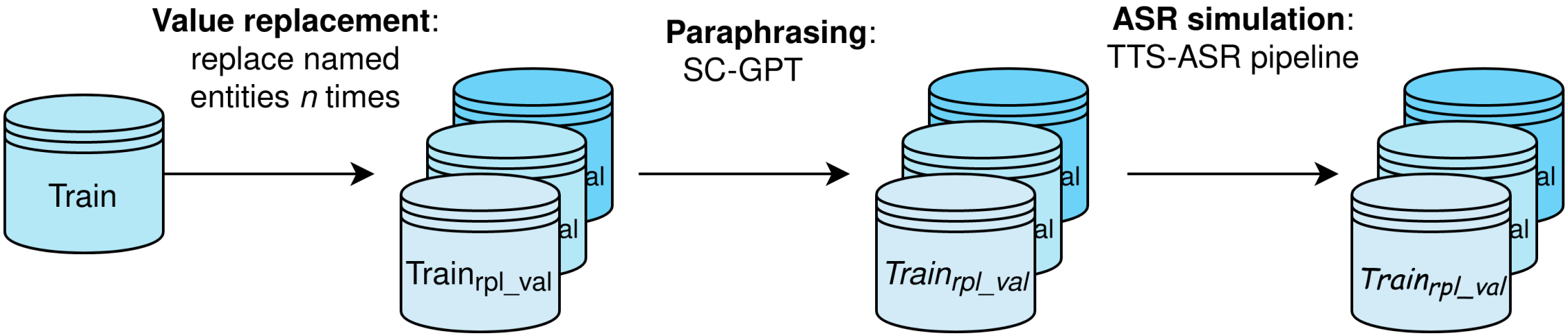

[user] I'm looking for somewhere in Dublin to have a meal and a drink. [agent] What kind of food do you want to eat? [user] I feel like a burger.
seq2seq DST
[state] city: Dublin, serves_alcohol: True, cuisine: burger

DST Data Augmentation

Value replacement

- Create a new ontology to sample from based on data from OpenStreetMap
- For all dialogues
    i. Go through each dialogue state, sample one value from ontology for each distinct value and replace it in the dialogue state
    ii. Track these replacements with a mapping between replaced value and new value
    iii. Based on the obtained mapping, perform a string replacement in the dialogue context

DST Data Augmentation



SC-GPT: Few-shot Natural Language Generation for Task-Oriented Dialog (Peng et al., 2020)

Results

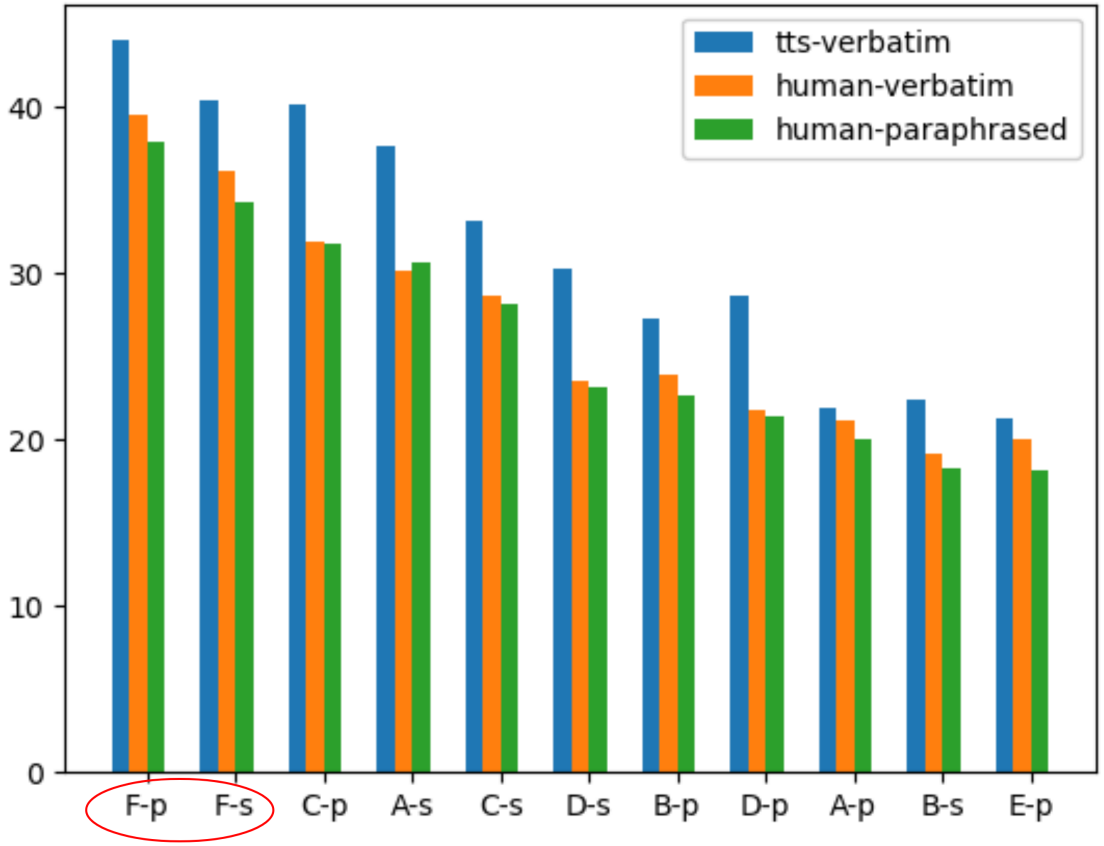
Primary submission (OLISIA$_1$)

Ensemble through majority voting of five T5-large models fine-tuned on different variations of the train set

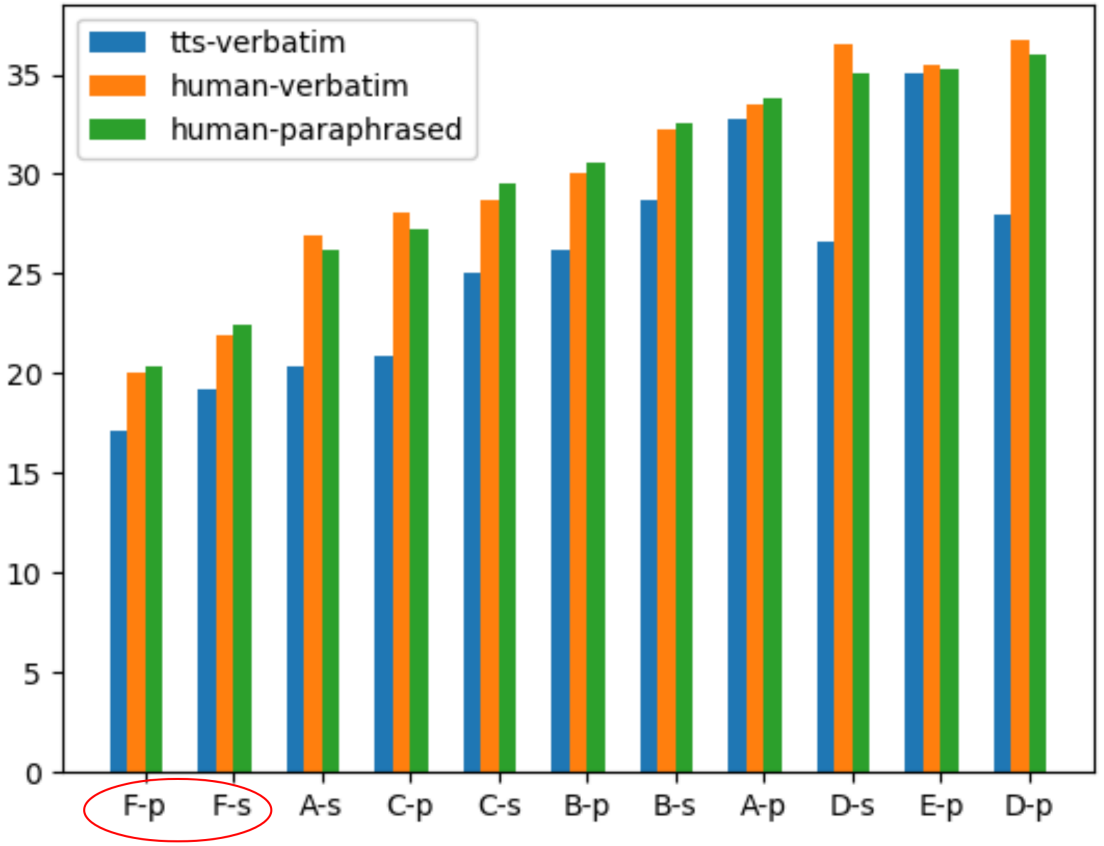
Secondary submission (OLISIA$_2$)

Single T5-large model fine-tuned on best train set

Joint Goal Accuracy



Slot Error Rate



JGA	Dev		Test		
	TTS-v	**Human-v**	**TTS-v**	**Human-v**	**Human-p**
Challenge baseline	26.3	22.6	_	_	_
OLISIA$_2$	44.1	40.3	40.4	36	34.3
OLISIA$_2$ (text oracle)	55		51.1		

ASR Cascade Adaptations

Metrics: JGA↑ / WER↓

Test set

	TTS-Verbatim	Human-Verbatim	Human-paraphrased
Whisper raw outputs	37.9 / 4.92	33.8 / 8.40	32.0 / _
+ Time normalization	40.0 / 4.49	35.6 / 7.89	33.5 / _
+ Noun correction	40.3 / 4.36	36.1 / 7.71	34.3 / _

DST Data Augmentation

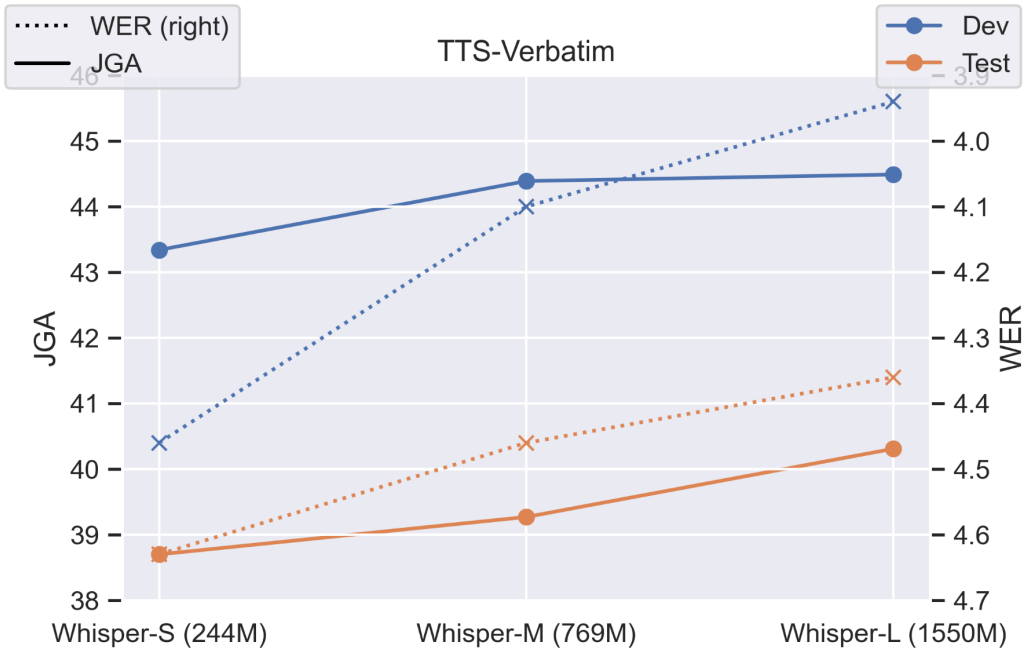
Metric: JGA

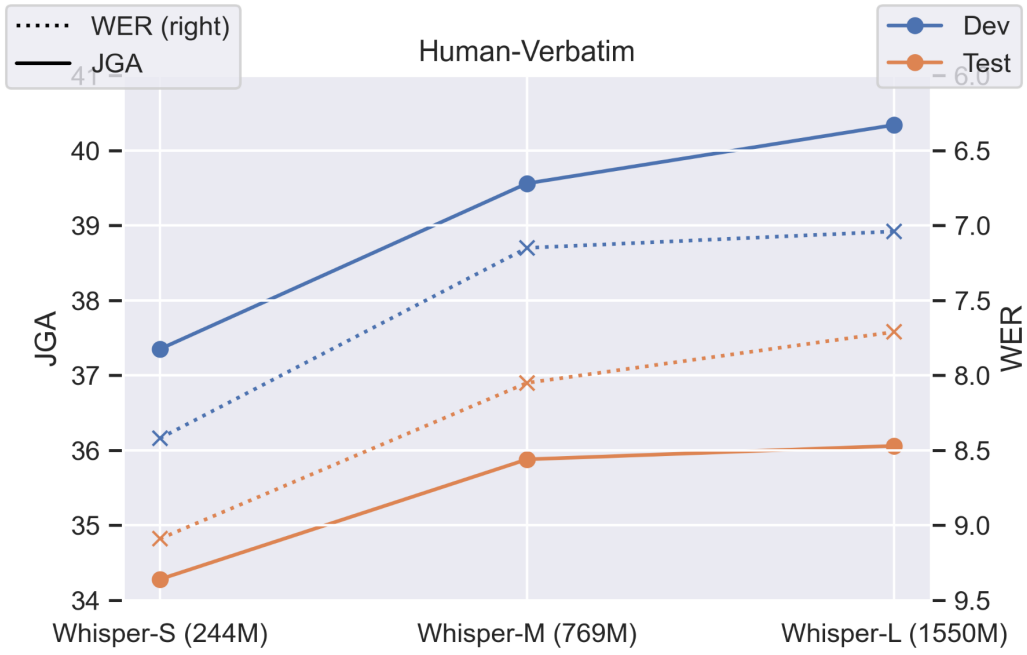
Test set

	TTS-Verbatim	Human-Verbatim	Human-paraphrased
Default train set	32.2	28.3	26.8
+ Value replacement	40.2	35.5	33.2
+ Speech simulation	40.3	36.0	34.3
+ Paraphrasing	37.3	33.9	31.5

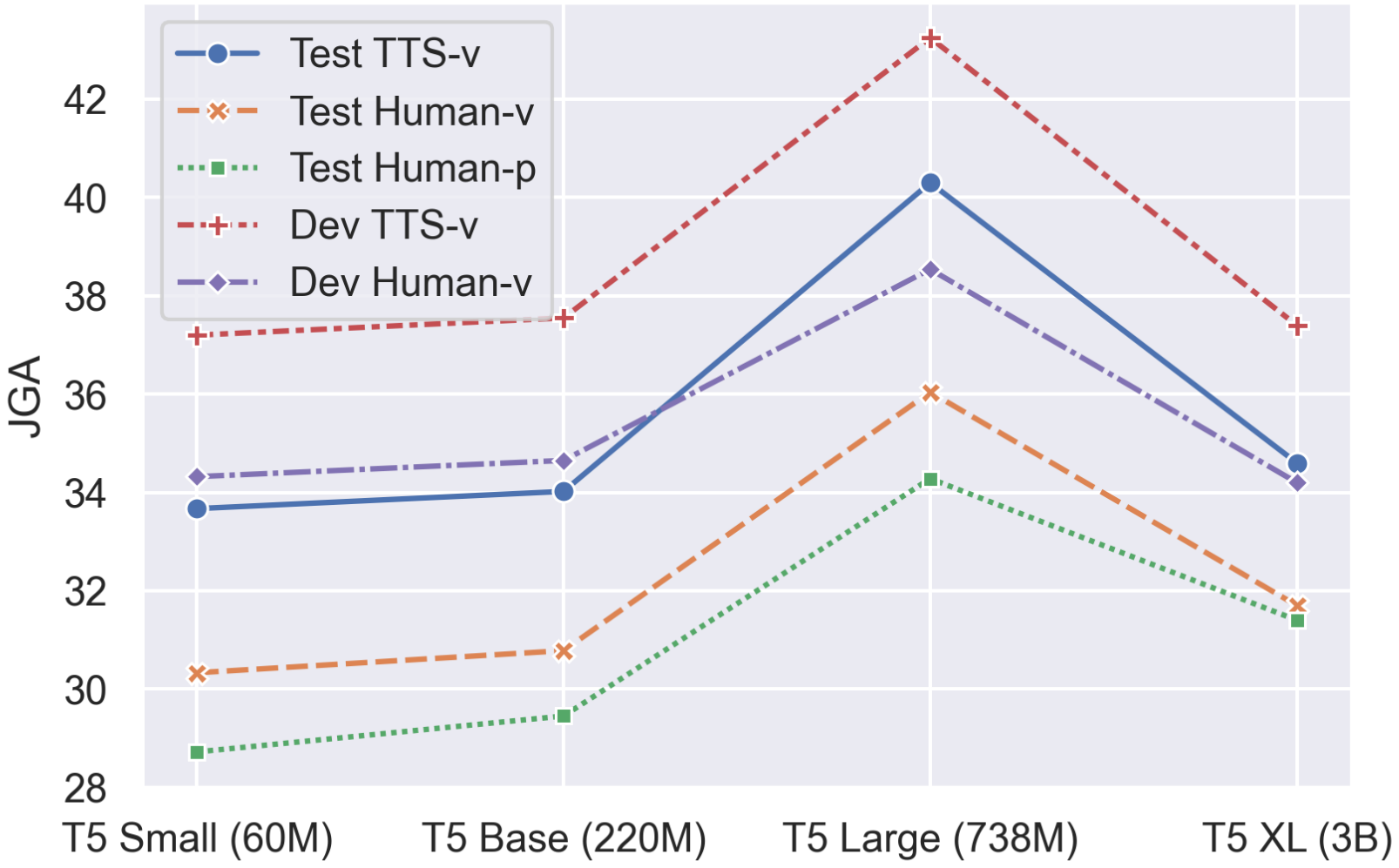
Model Size

ASR





DST


Test TTS-v
Test Human-v
Test Human-p
Dev TTS-v
Dev Human-v
JGA
42
40
38
36
34
32
30
28
T5 Small (60M)
T5 Base (220M)
T5 Large (738M)
T5 XL (3B)

Ensembling strategy

Metric: JGA

Dev set

	TTS-verbatim	Human-verbatim
1 model (no ensemble)	44.1	40.2
5 models - same train set	44.4	40.4
5 models - different train sets	47.8	43.5
9 models - different train sets	48.5	43.9

Contribution

We show

- the impact of scaling up the model size (both for ASR and DST)
- the relevance of different data augmentation techniques for DST
- the need for post-processing the ASR output in a cascade system

What's next?

- Multimodal model: encoding both speech and text
- Comparison between cascade and end-to-end systems for spoken DST
- Encoding n-best ASR hypotheses with Transformer

Thanks!