Beyond (fixed) Ontologies

Michael Heck
Dialog Systems and Machine Learning

Düsseldorf, Germany, 2nd July 2021

Introduction

Introduction

Task-oriented Dialog Systems

- Increasing popularity demands robust and scalable solutions for
  - **multi-domain** settings: one system helps accomplish multiple tasks
  - **open-concept** settings: dialogues are typically not bound by pre-defined content

I'm looking for an expensive restaurant in the center of town.

What type of food do you prefer?

Surprise me

“Fitzbillies” serves British food. Would you like more options?

That’s fine. Can you book a table for 5 at 11:30 on Tuesday?

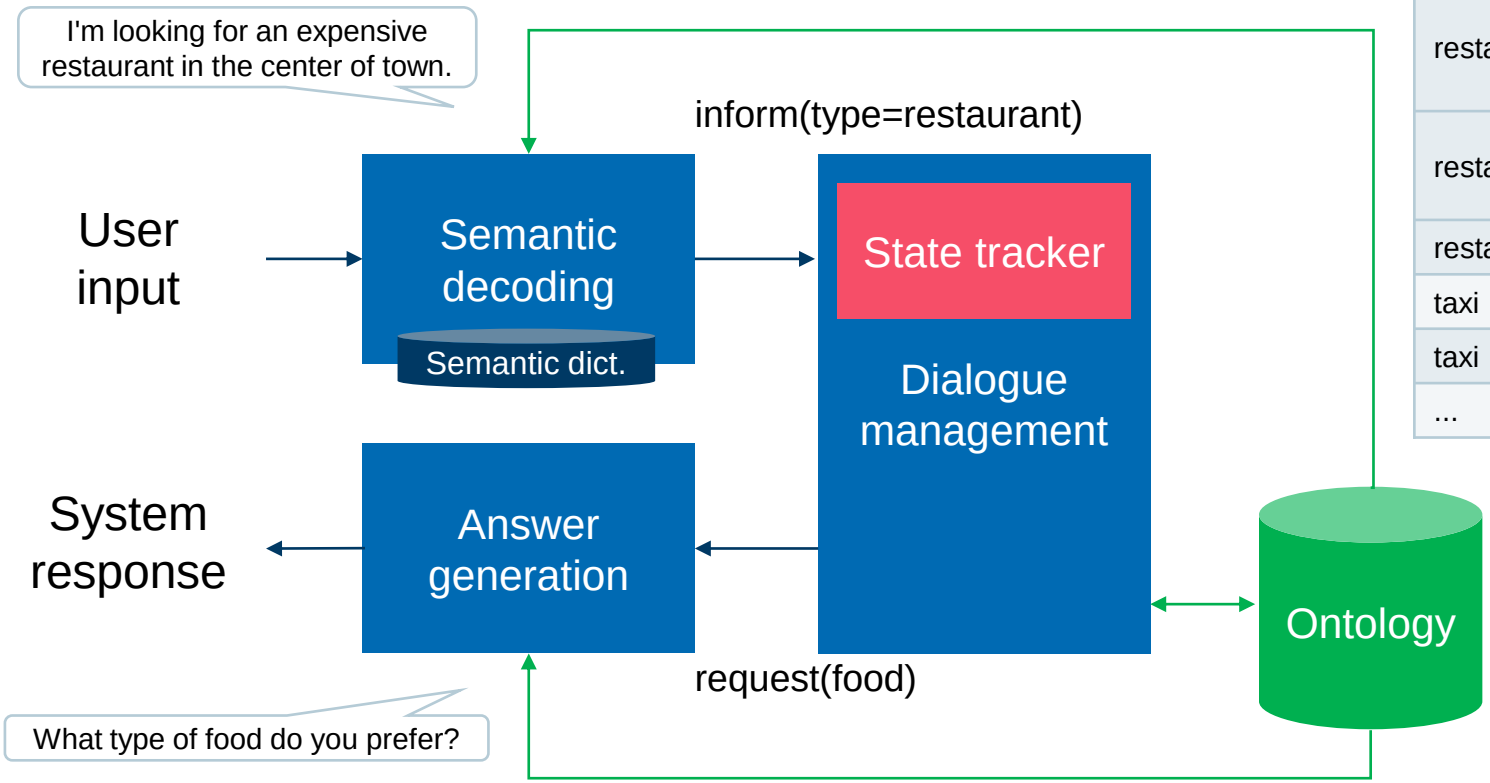
Of course. Is there anything else I can help you with?

I'm also looking for a place to stay in the same area as the restaurant.

domain change

Introduction

Statistical dialogue systems 101



Domains	Slots	Values
restaurant	food	Italian
		Chinese
restaurant	area	centre
		north
restaurant	price	cheap
taxi	depart	station
taxi	arrive	hotel
...	...	...

Introduction

Dialogue State Tracking

- **Dialogue state**: Summary of the conversation till current turn
  - Set of constraints, e.g., **slot-value pairs** from an **ontology**
- Update **dialogue state** at each turn
  - Monitor user‘s goal throughout dialogue
  - Required to determine system actions

Turn	Domain-Slot pair	Value

I'm looking for an expensive restaurant in the center of town.

What type of food do you prefer?

Surprise me

“Fitzbillies” serves British food. Would you like more options?

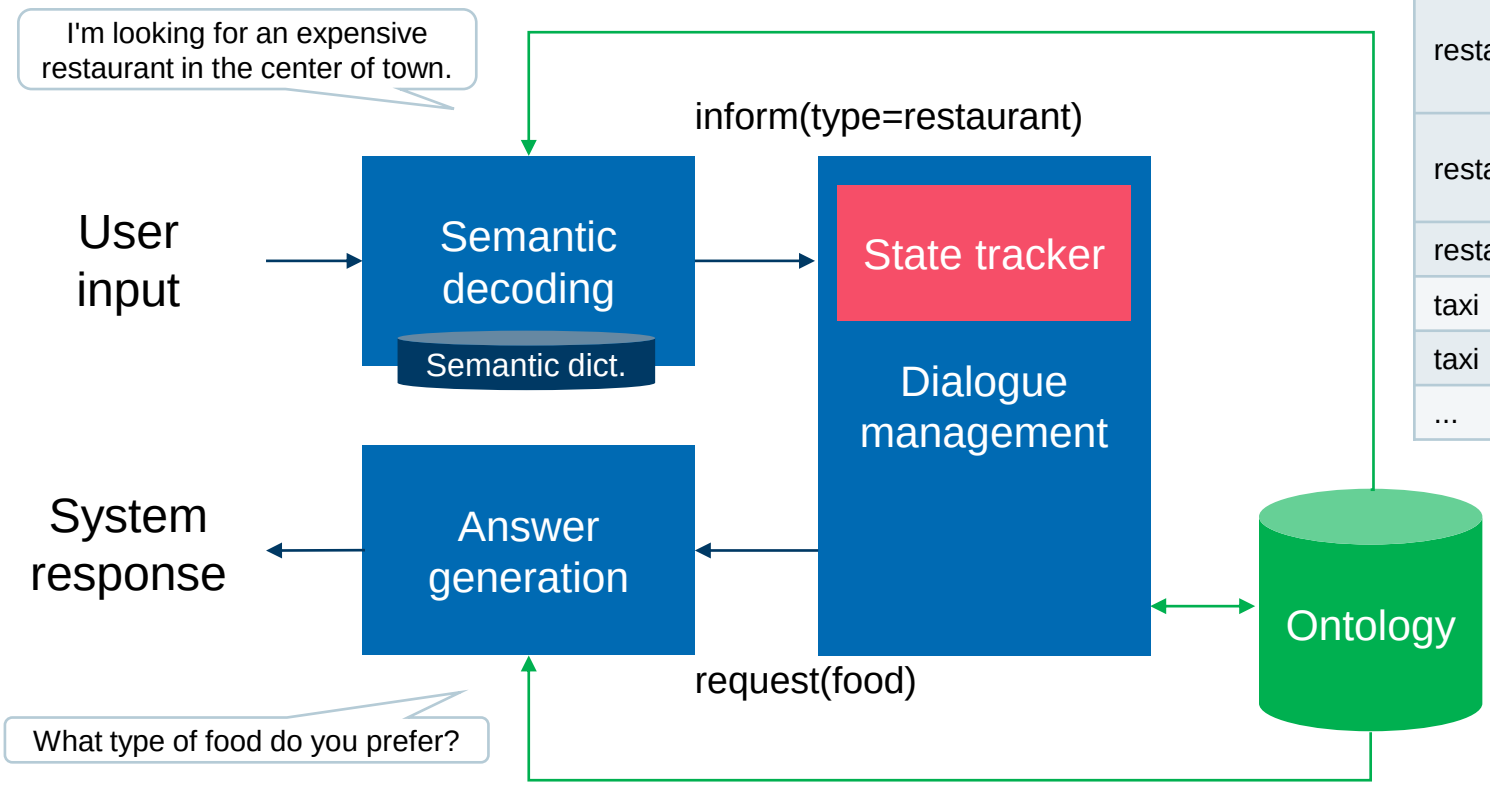
That’s fine. Can you book a table for 5 at 11:30 on Tuesday?

Of course. Is there anything else I can help you with?

I'm also looking for a place to stay in the same area as the restaurant.

Introduction

Statistical dialogue systems 101



Domains	Slots	Values
restaurant	food	Italian
		Chinese
restaurant	area	centre
		north
restaurant	price	cheap
taxi	depart	station
taxi	arrive	hotel
...	...	...

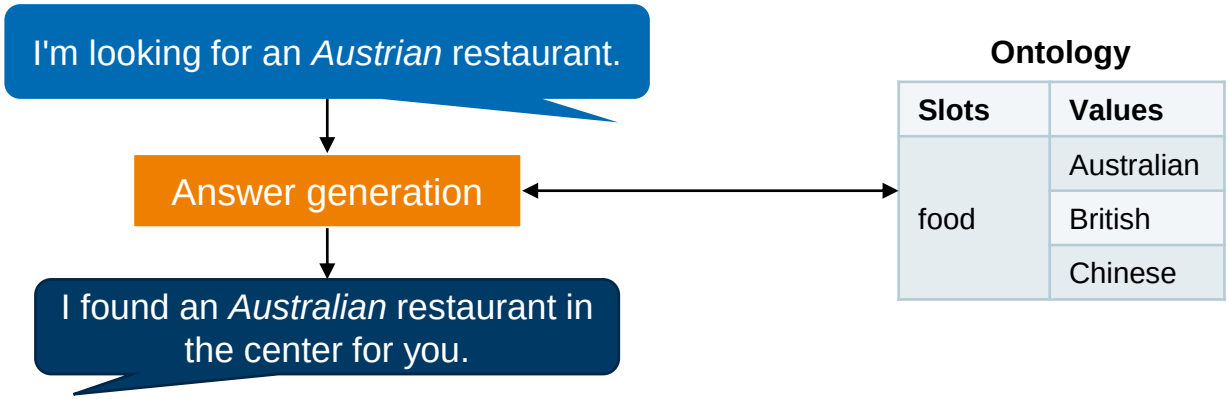
- Discrete representation of concepts limits capacities

Introduction

Limitations caused by fixed ontologies

- Model inflexibility, limited or no scalability due to
  - Closed set of known (i.e., interpretable) concepts
  - Ontology-specific architecture (number of parameters proportional to number of concepts)
  - Methods that do not scale well (e.g., discrete representations, hand crafted features)

- Risk of memorization
  - Model might simply memorize all known concepts and lose ability to generalize



Request can not be handled due to being out-of-ontology

Introduction

Ontology independence

- What does „independence“ mean?
  - Models should not be limited by a fixed ontology
  - Does not mean that there is no underlying ontology, but

    ontology should be **malleable** and/or **interchangeable** during test time

  - I.e., models should be independent from any ontology they have been trained on

- Why is this important? Ontology independence…

… motivates transfer learning

Transfer learning as a means to strenghten models across domains

… motivates data efficiency

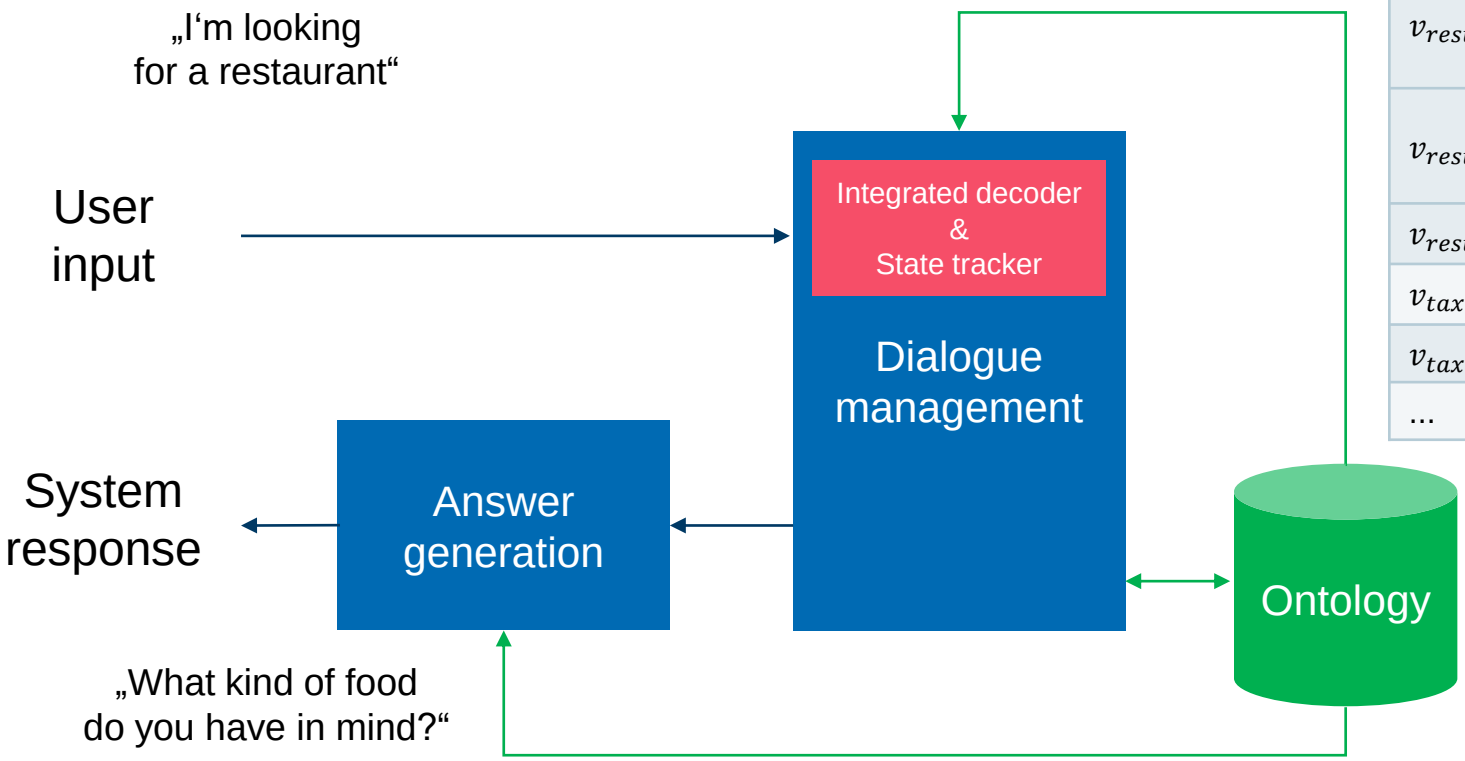
Data efficiency as a means to facilitate ontology independence

… motivates large scale systems

Large scale systems are feasible only with transfer capabilities

Introduction

Continuous representations in dialogue systems



Domains	Slots	Values
$v_{restaurant}$	v_{food}	$v_{Italian}$
		$v_{Chinese}$
$v_{restaurant}$	v_{area}	v_{centre}
		v_{north}
$v_{restaurant}$	v_{price}	v_{cheap}
v_{taxi}	v_{depart}	$v_{station}$
v_{taxi}	v_{arrive}	v_{hotel}
…	…	…

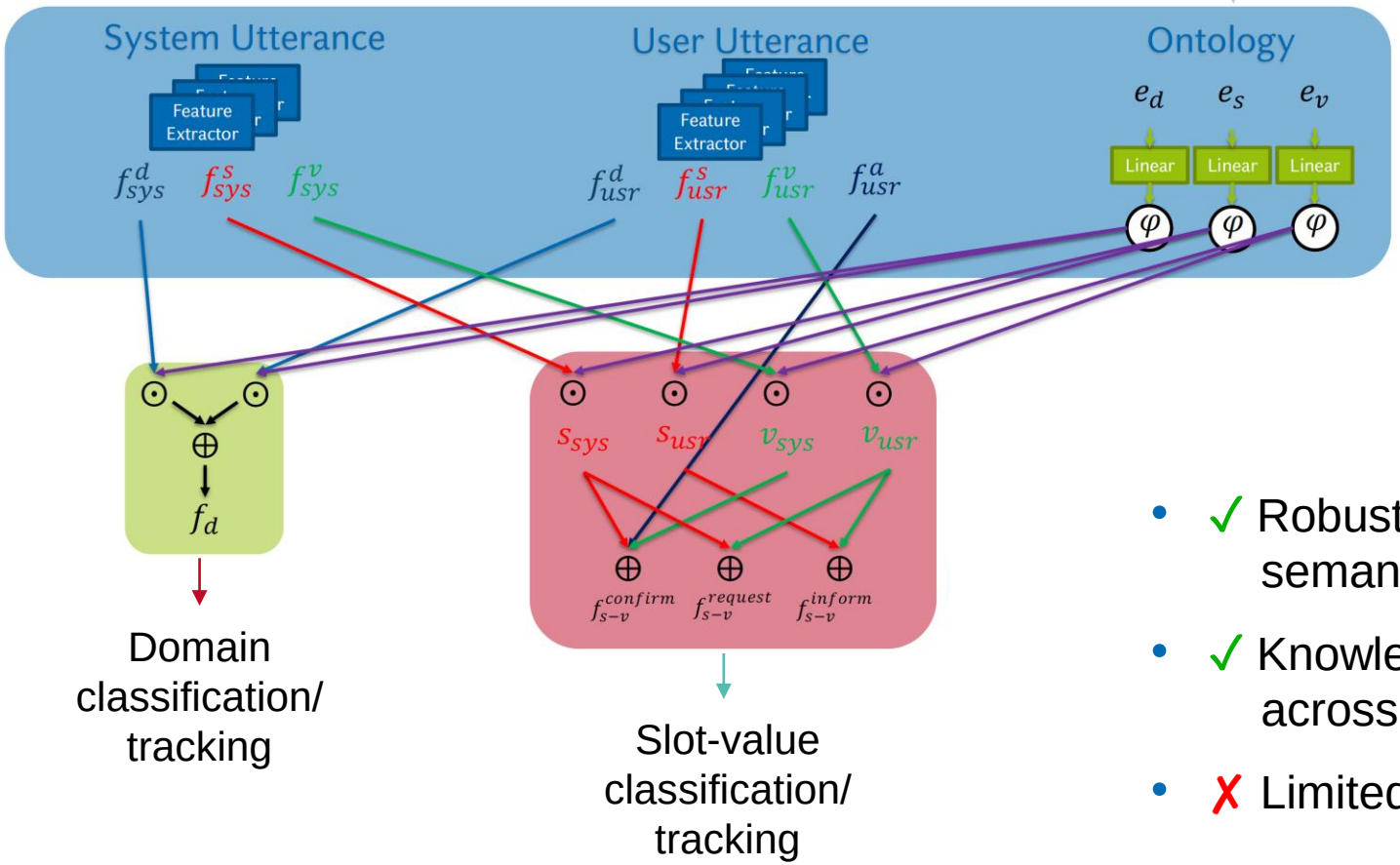
- Vector representations mitigate semantic decoding problem
  - Similarity measures replace exact matching

(Mrksic et al., 2017)

Introduction

Continuous representations in DST

„What food would you like?“ „I‘d like Thai food“



Domains	Slots	Values
restaurant	food	Italian
		Thai
restaurant	area	centre
		north
restaurant	price	cheap
taxi	depart	station
taxi	arrive	hotel
...	...	...

Input encoders produce vector representations

- ✓ Robustness due to semantic representations
- ✓ Knowledge sharing across domains
- ✗ Limited scalability

(Ramadan et al., 2018)

Introduction

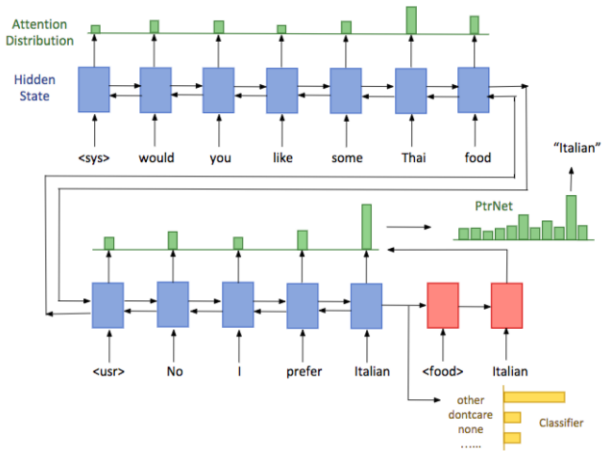
Continuous representations in DST

- Models achieve state-of-the-art performance in fixed-ontology DST evaluations
  - Utilization of **semantic representations** is driving force
    - Leverages semantic similarity of concepts (slots, values, etc.)
    - Representation of previously unseen concepts is possible
    - Tighter integration of DS components utilizes parameter sharing and transfer learning
- Picklist based
  - DS as distribution over all possible slot-values
  - Individual scoring of all slot-value pairs
- Extraction based
  - Find values through span matching or sequence tagging in dialogue context

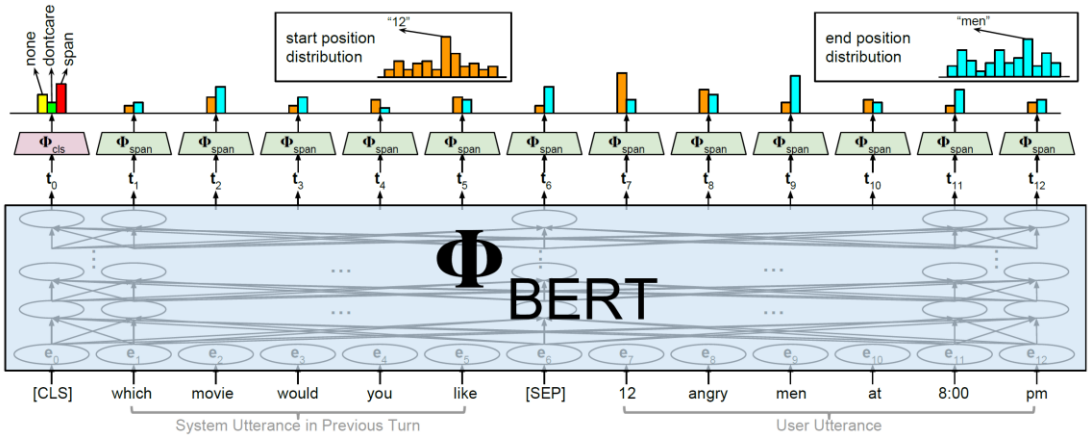
(Henderson et al., 2014) (Wen et al., 2017) (Mrksic et al., 2017) (Ramadan et al., 2018) (Gao et al., 2019) (Chao and Lane, 2019) (Kim et al., 2019)

Introduction

Extraction-based DST



Recurrent NN (PointerNet)



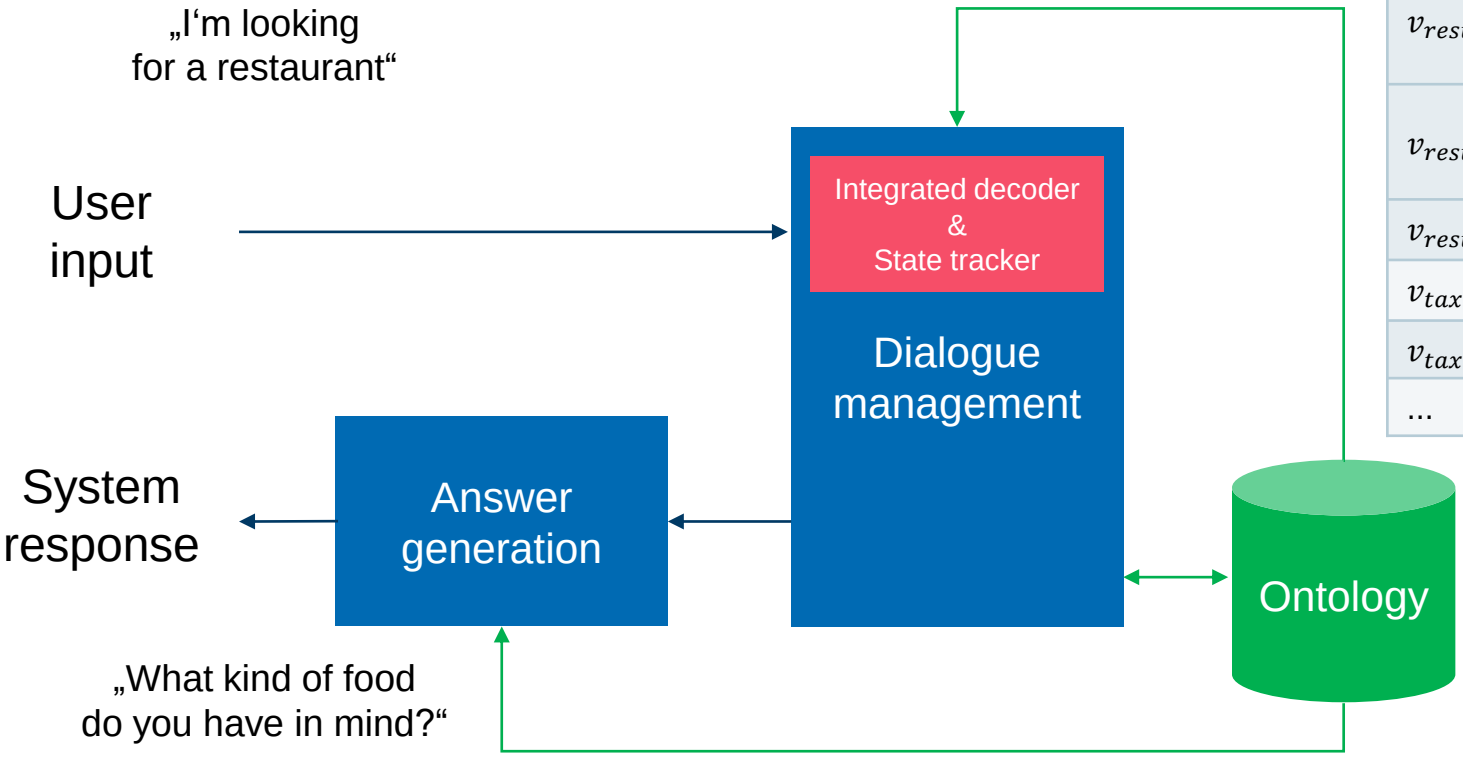
Transformer encoder with classification heads

- PointerNets generates starting and ending index of relevant segment
- Transformer produces contextual representations of input and classfies them
  - **Sentence** representation used to determine presence of value (via „slot gate“)
  - **Token** representations used to determine value span (via start/end position prediction)
- ✗ Limited to extractive values

(Xu and Hu, 2018) (Chao and Lane, 2019)

Introduction

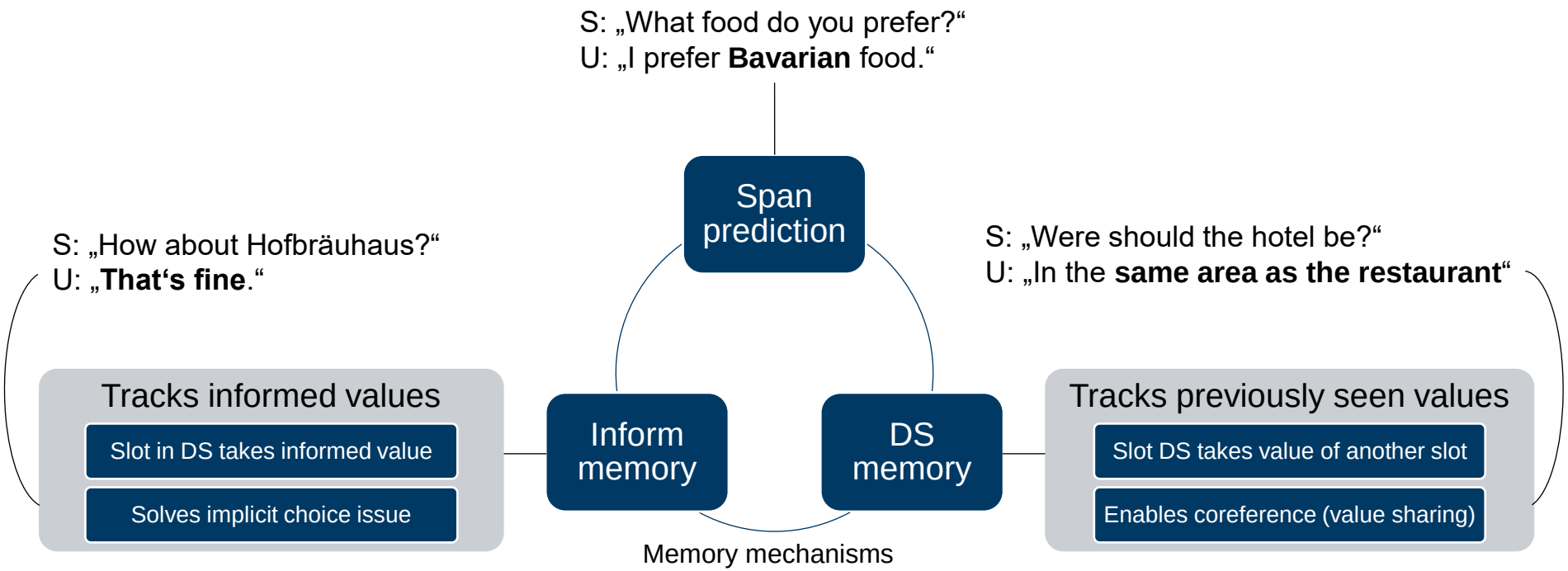
Value independence



Domains	Slots	Values
$v_{restaurant}$	v_{food}	$v_{Italian}$
		$v_{Chinese}$
$v_{restaurant}$	v_{area}	v_{centre}
		v_{north}
$v_{restaurant}$	v_{price}	v_{cheap}
v_{taxi}	v_{depart}	$v_{station}$
v_{taxi}	v_{arrive}	v_{hotel}
…	…	…

Introduction

TripPy: **Tri**ple co**py** strategy DST

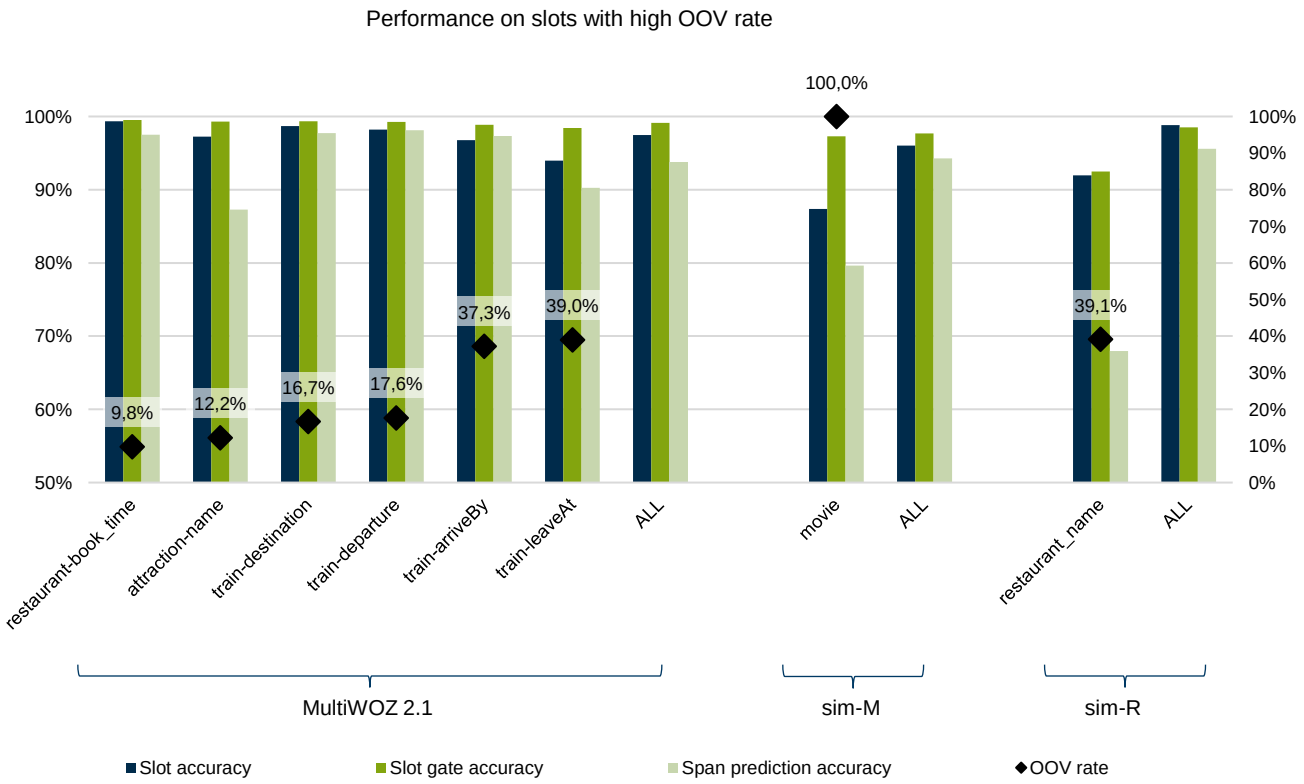


- Dialog context is the only source of information
- Maintains two memories (value candidate lists) **on the fly**

(Heck et al., 2020)

Value Independence of TripPy

- OOV rate has minor effect on performance
  - Good generalization to unseen values



(Heck et al., 2020)

Ontology independence and the schema-guided paradigm

Ontology independence

Shortcomings of recent systems

- Recent systems parse dialogues in terms of **fixed concepts**
  - Lack understanding of the **semantics** of concepts

- Example: “I want to buy tickets for a movie.”
  - Models predict “BuyMovieTickets” based on observed patterns
    - No association with real action of buying movie tickets
    - Similarity to action of buying concert tickets not captured

- Models not robust to changes
  - Need to be retrained as new slots or intents are added
  - Domain-specific parameters unsuitable for zero-shot application

Ontology independence

Challenges of building large-scale systems

- Support of heterogenous services/APIs
  - Might overlap in functionality
- Robustness towards changes in API
  - Robustness towards new slots and intents
  - Generalization to new slot values (with little or no retraining)
- Generalization to new APIs
  - Joint modelling across APIs
  - Zero-shot generalization

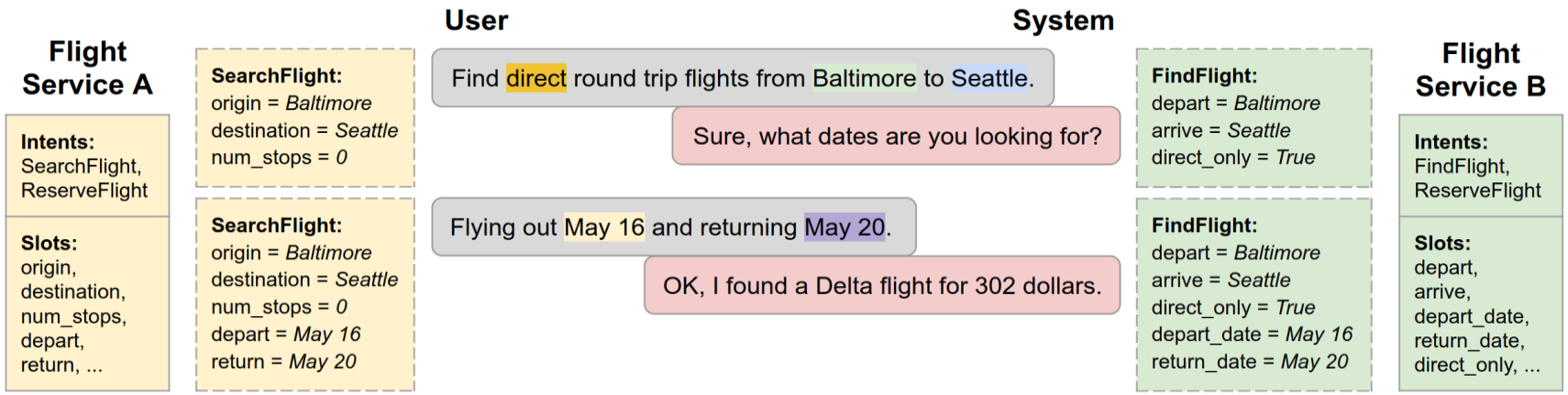
Ontology independence

Schema-guided paradigm

- Many models/evaluations don‘t fully capture reality of scenarios
  - Few domains, one service per domain, static ontologies, fixed concepts
  - Not scalable, transfer not possible due to domain-dependent architectures

vs.

- Many domains, many services (defined by APIs)
  - Mismatch of training and testing conditions
  - Potential sighting of unseen concepts (domains, intents, slots, values)
  - Transfer desirable/necessary



(Rastogi et al. 2020)

Ontology independence

DSTC8: Schema-guided paradigm for dialogue modeling

- Advocates building **unified** dialogue model for all APIs using **semantic conditioning**
  - **A service's schema serves as input to the model**
    - Uses natural language descriptions to obtain **semantic representations** of schema elements
    - Predictions are **conditioned** on semantics of schema
    - Predictions over dynamic sets of intents and slots
  - Facilitates full-scale transfer learning and zero-shot learning for dialogue modeling
    - A model should not contain service specific components

Service

service_name: *"Payment"*
description: *"Digital wallet to make and request payments"*

Slots

name: *"account_type"* categorical: True
description: *"Source of money to make payment"*
possible_values: [*"in-app balance"*, *"debit card"*, *"bank"*]

name: *"amount"* categorical: False
description: *"Amount of money to transfer or request"*

name: *"contact_name"* categorical: False
description: *"Name of contact for transaction"*

Intents

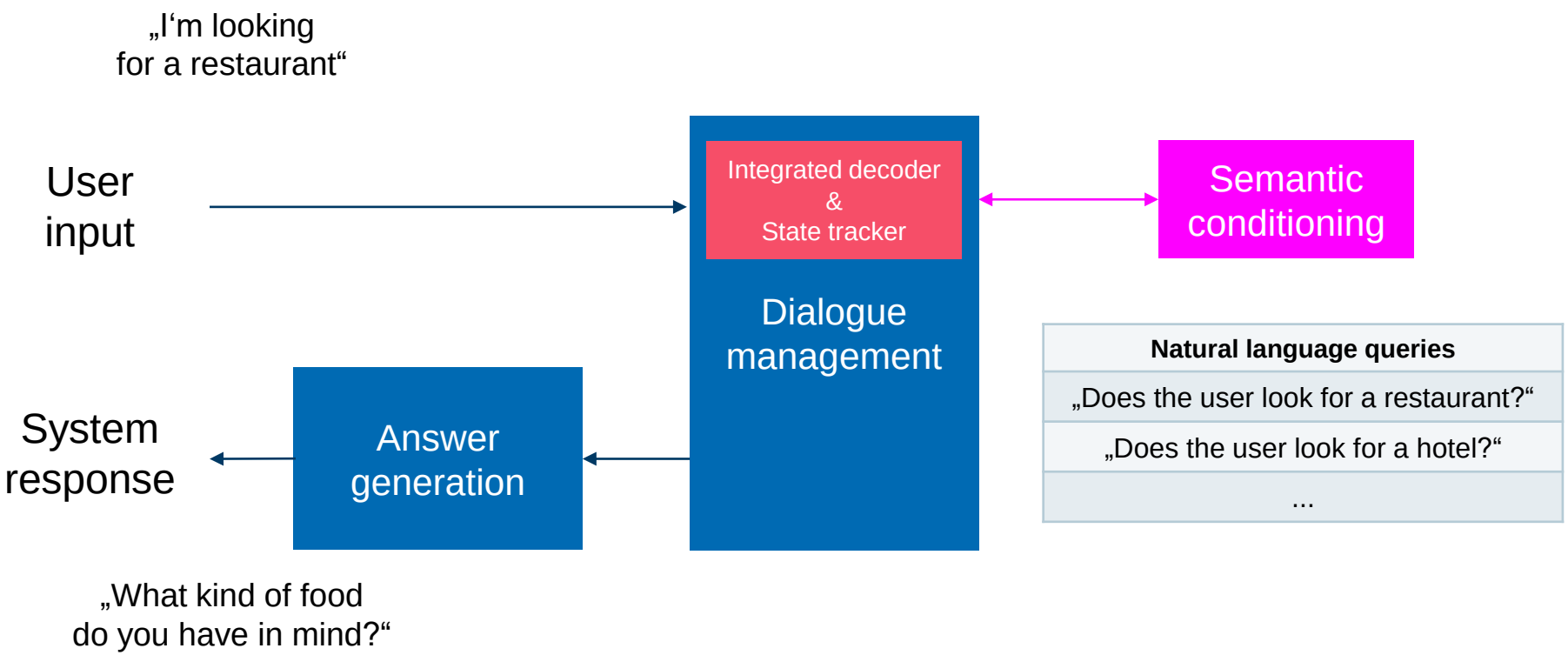
name: *"MakePayment"*
description: *"Send money to your contact"*
required_slots: [*"amount"*, *"contact_name"*]
optional_slots: [*"account_type"* = *"in-app balance"*]

name: *"RequestPayment"*
description: *"Request money from a contact"*
required_slots: [*"amount"*, *"contact_name"*]

(Rastogi et al. 2020)

Ontology independence

Semantic conditioning



Natural language queries
„Does the user look for a restaurant?"
„Does the user look for a hotel?"
...

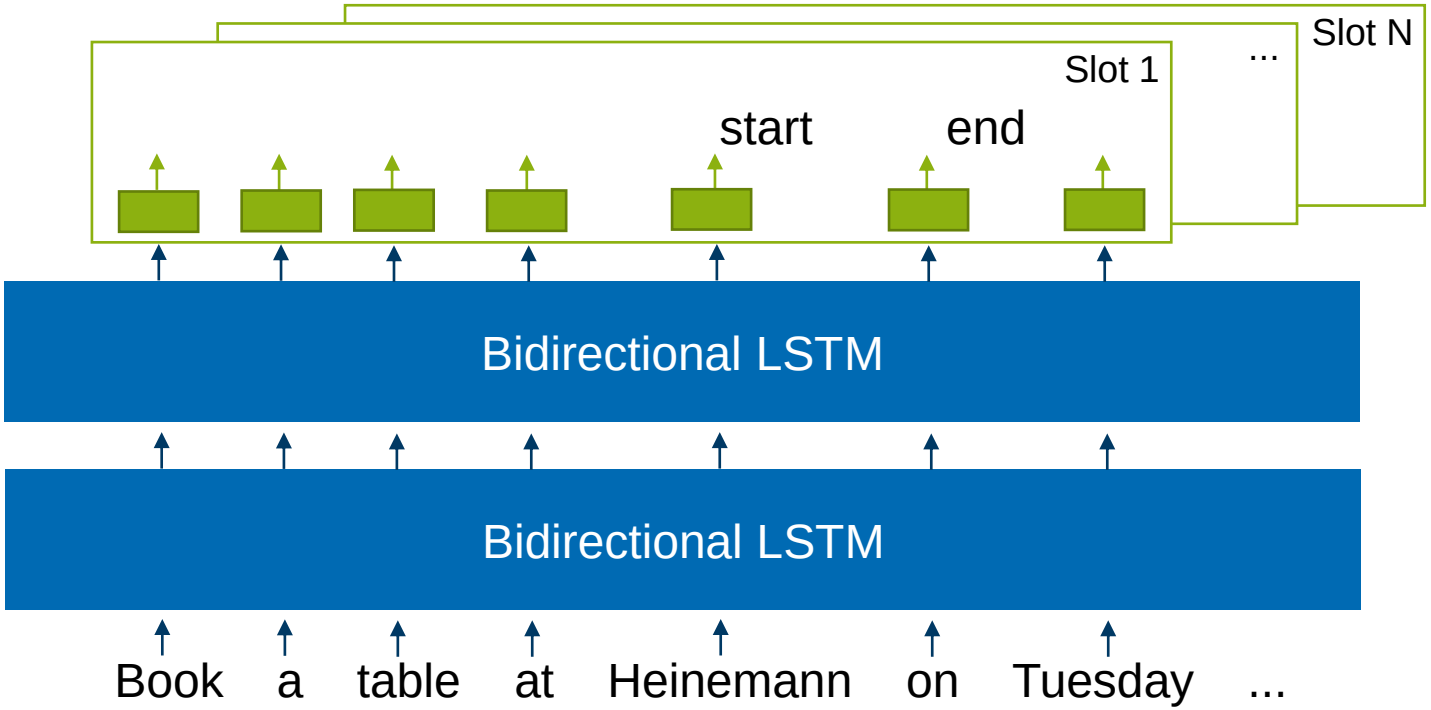
- Conditioning with natural language replaces fixed ontology
  - Measure semantic similarity between input and concepts

(Mrksic et al., 2017)

Ontology independence

Semantic conditioning

- Zero-shot adaptive transfer for DST
  - Infusing semantic slot representations into unified model

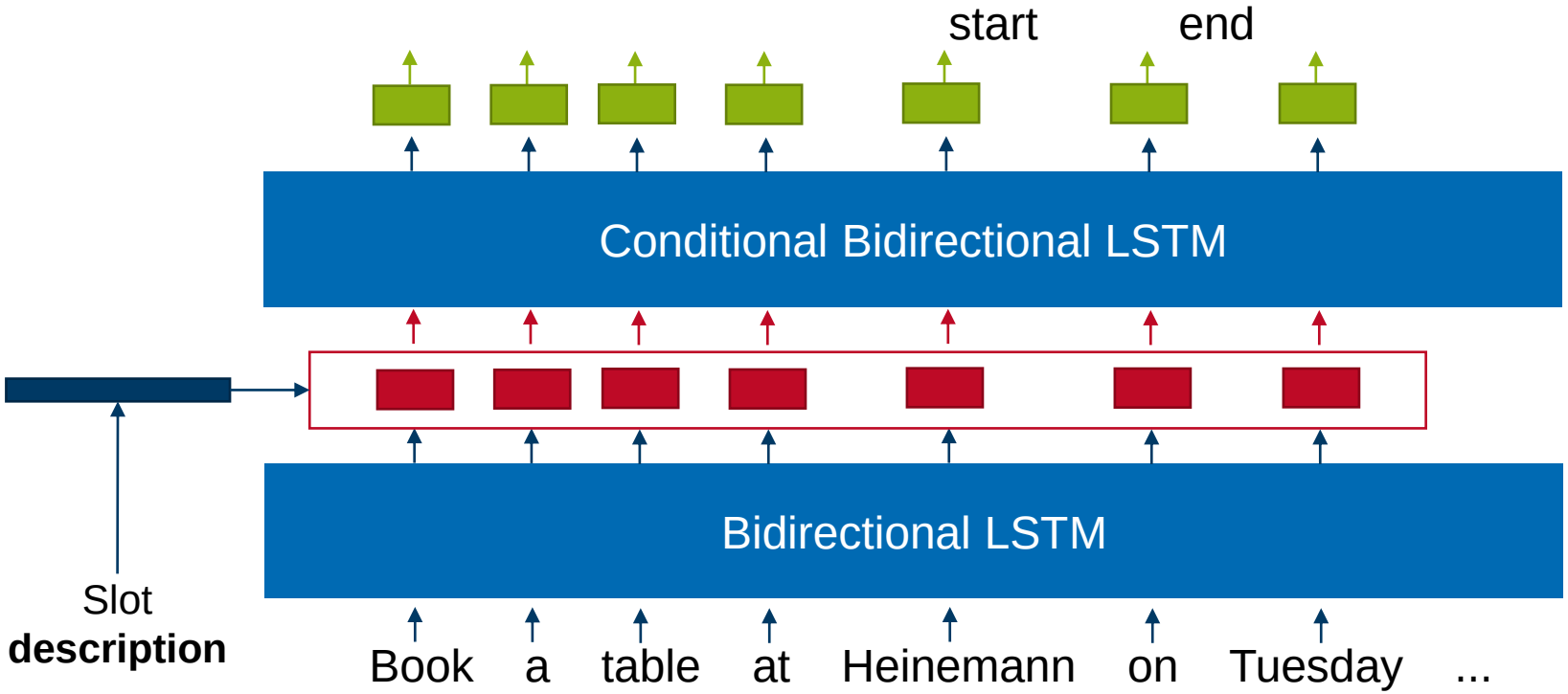


(Bapna et al., 2017) (Shah et al., 2019)

Ontology independence

Semantic conditioning

- Zero-shot adaptive transfer for DST
  - Infusing semantic slot representations into unified model

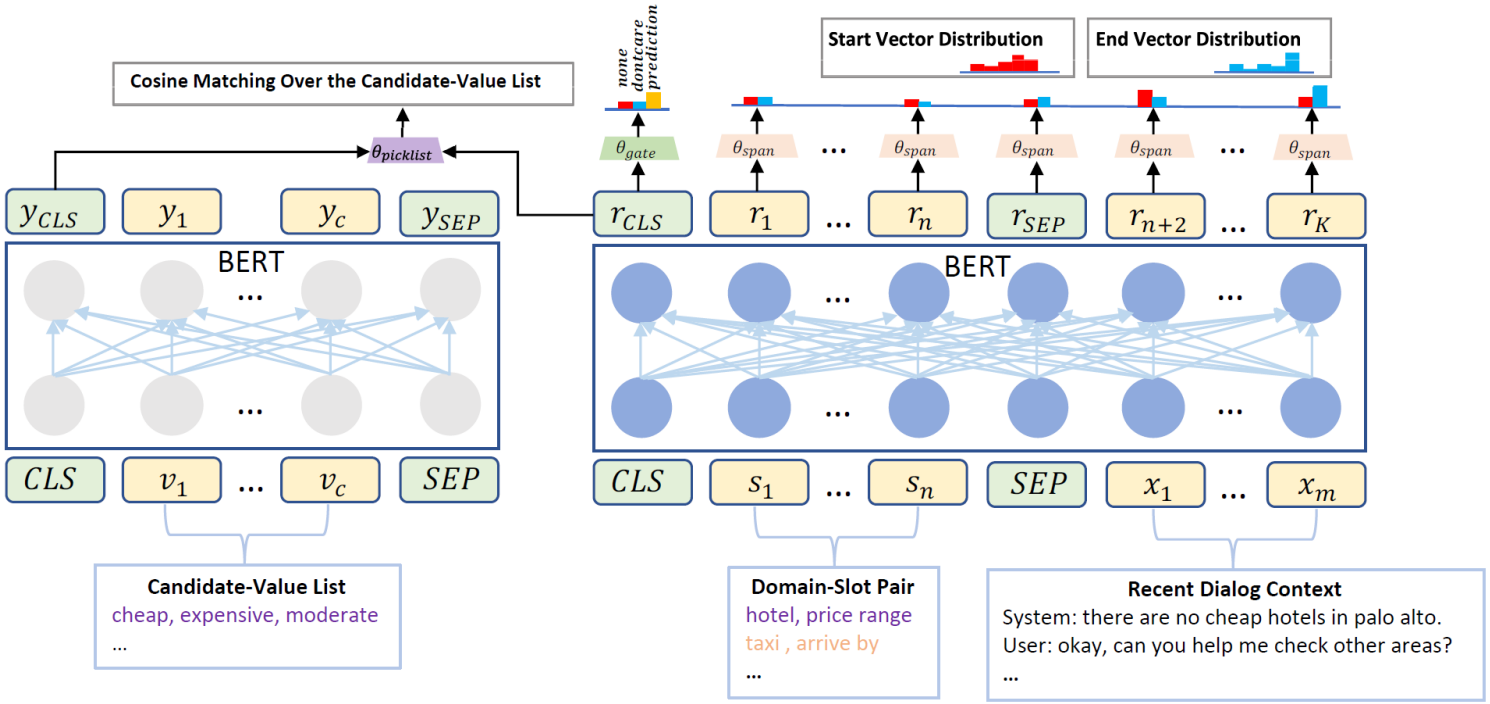


(Bapna et al., 2017) (Lee and Jha, 2018) (Shah et al., 2019)

Ontology independence

Semantic conditioning on the example of „Dual Strategy DST“

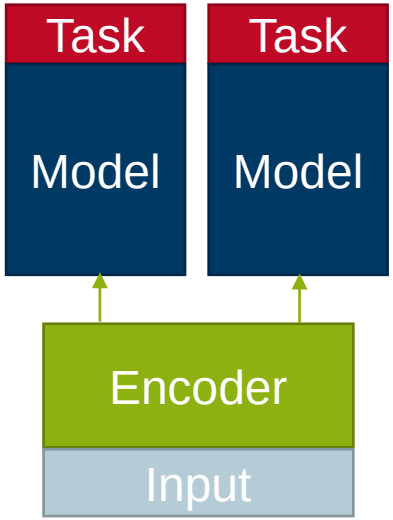
- Adaptation and **transfer learning** for Slot-filling for DST
  - Parameter sharing for domain adaptation and joint training



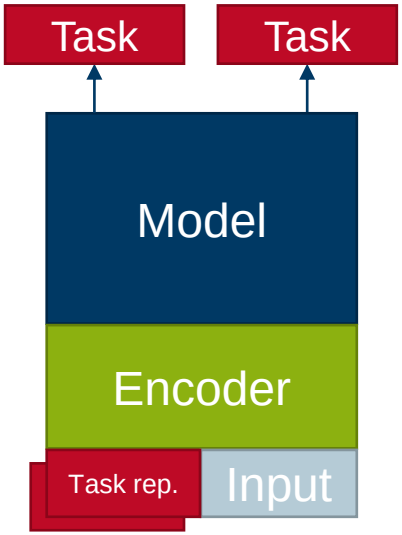
(Zhang et al. 2019)

Ontology independence

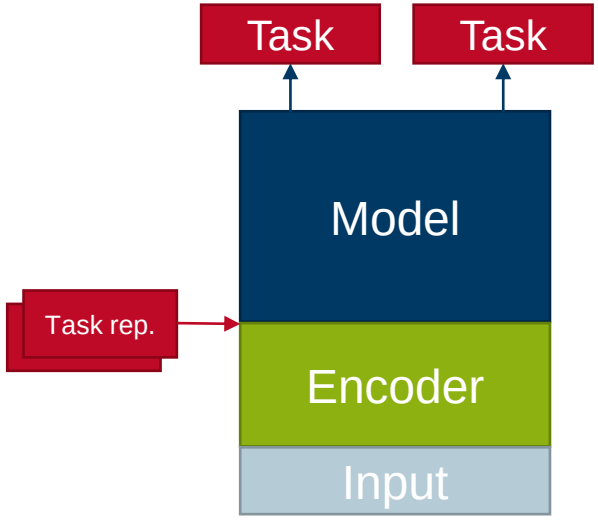
Types of semantic conditioning



None (task specific model architecture)



Semantic conditioning pre-encoding



Semantic conditioning post-encoding

Ontology independence

Heinrich Heine Universität Düsseldorf

Schema-guided paradigm for dialogue modeling

- **Zero-shot learning** by using semantic modeling
- Knowledge sharing by …
  - … relating semantically similar concepts
  - … using single unified model
- Handling of **unseen** services and API changes by using
  - natural language input
  - semantic representations to condition the model

DSTC8: Schema-guided DST

DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

Data

- Benchmark highlighting challenges for large-scale systems

	MultiWOZ	SGD
Domains	7	**16**
Slots	30	**214**
Values	4,510	14,139
Dialogues	8,438	16,142
Avg. turns per dialogue	13.46	20.44

Table: Statistics of training portions of datasets

	train	dev	test
Dialogs	16,142	2,482	4201
Domains	16	16	18
Services	26	17	21
Dialogs w/ unseen APIs	-	**42%**	**70%**
Turns w/ unseen APIs	-	**45%**	**77%**

Table: Split of SGD dataset

Evaluation

- "Joint goal accuracy"
  - Only services that are active in a turn are included in the dialogue state
    - I.e., false positives on inactive services are not counted as error!
  - If ground truth AND prediction == NONE, then don't count
  - Fuzzy value matching

In multi-domain dialogues, false positives in other domains are ignored.

I.e., models are evaluated on single domains (in isolation)

(Rastogi et al., 2020)

DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

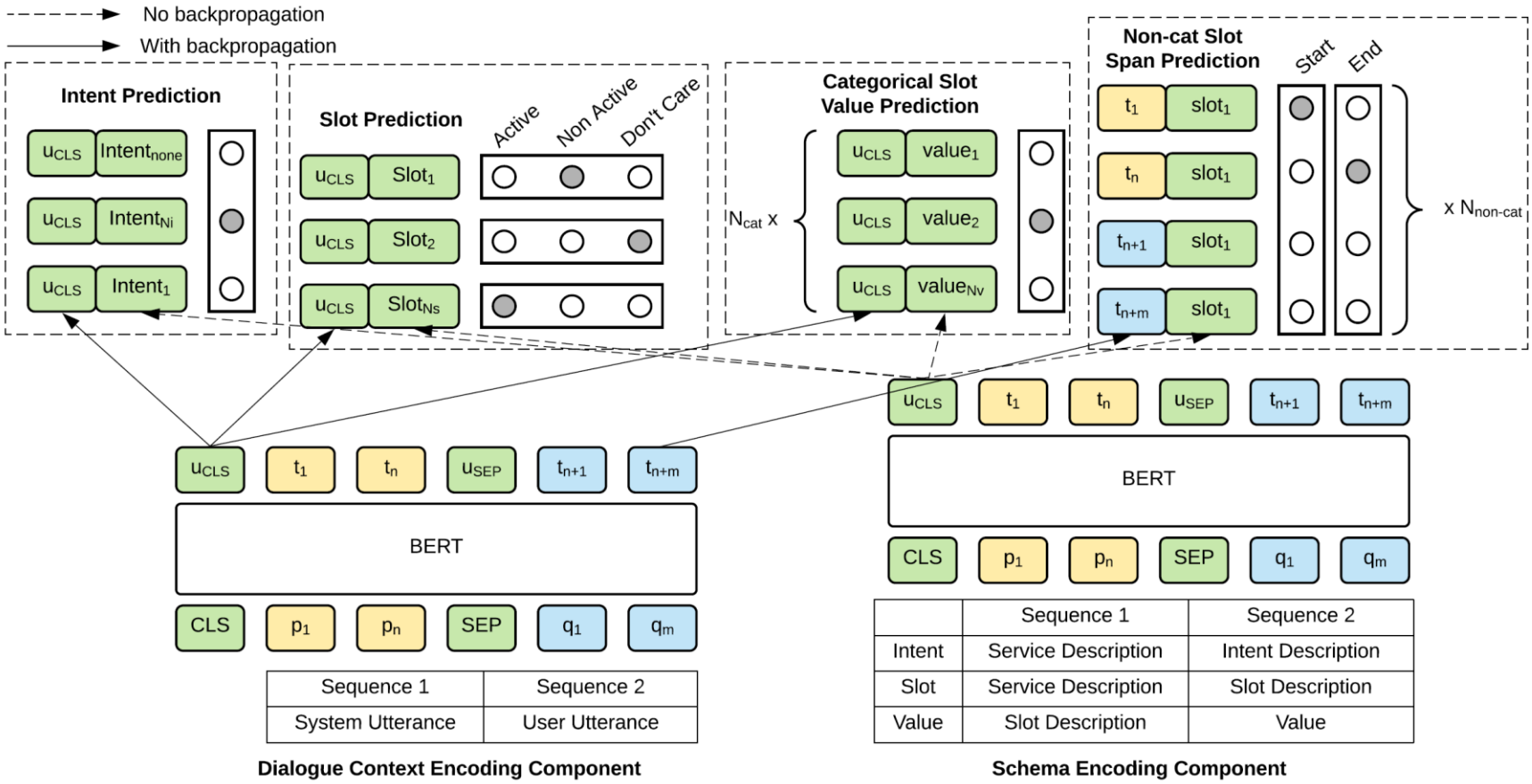
Baseline: Zero-shot dialogue state tracking

- Model is shared among all services and domains
- Uses 2 contextual encoders:
  - **Finetuned** BERT encodes context
  - **Fixed** pre-trained BERT encodes schema element descriptions
    - Intents, slots, categorical slot values
- Schema element-wise classification
  - Concatenate context representation and schema element representation
  - Do for each turn and for each schema element

(Rastogi et al., 2020)

DSTC8: Schema-guided DST

Baseline: Zero-shot dialogue state tracking



(Rastogi et al., 2020)

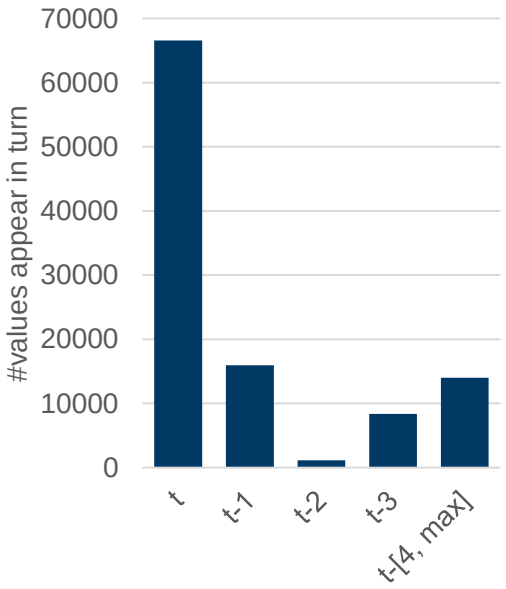
DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

Evaluation results

	All services				Seen services				Unseen services			
	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1
Baseline	0,25	0,56	0,91	0,97	0,41	0,68	0,95	1	**0,2**	0,52	0,89	0,95

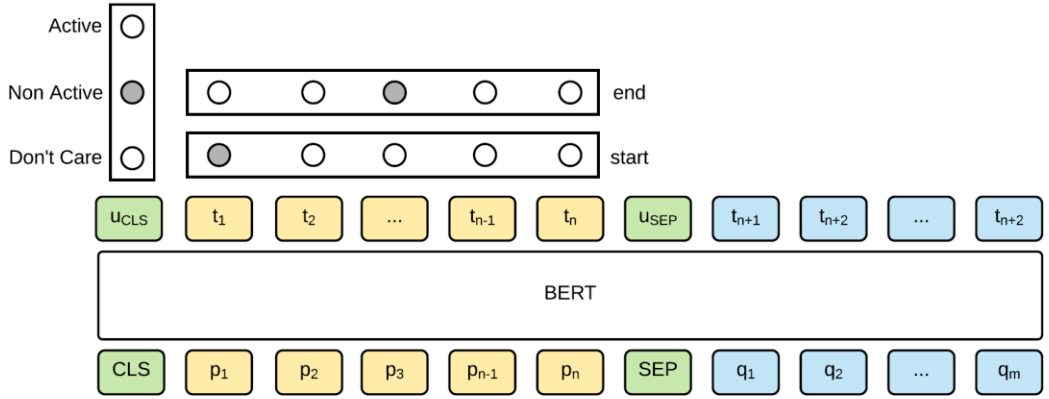
- ✗ Drawbacks
  - No history (only single turn) in context
    - Many slot values appear multiple turns earlier
- Separate models for context and schema
  - Interaction only after encoding
  - No finetuning of schema encoder



DSTC8: Schema-guided DST

Unified span detection framework for DST

- **Single** BERT jointly encodes context and schema elements
  - **Multiple passes** per turn and slot, one for each prediction task
  - Adds (truncated) dialogue history to input
  - Renders all predictions a span prediction problem (for multitask learning effect)



	Sequence 1	Sequence 2
Intent	$turn_{t-k}$ [SEP1] ... $turn_t$	[INTENT] Service Name [SEP2] Intent Name Intent Description
Categorical Slot	$turn_{t-k}$ [SEP1] ... $turn_t$ [CSEP] $value_1$ [CSEP] $value_2$ ... [CSEP] $value_n$	[C_SLOT] Slot Name [SEP2] Slot Name Slot Description
Non Categorical Slot	$turn_{t-k}$ [SEP1] ... $turn_t$	[NC_SLOT] Slot Name [SEP2] Slot Name Slot Description

(Shi et al., 2020)

DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

Evaluation results

	All services				Seen services				Unseen services			
	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1
Baseline	0,25	0,56	0,91	0,97	0,41	0,68	0,95	1	**0,2**	0,52	0,89	0,95
Shi	0,54	0,8	0,91	**0,87**	0,53	0,75	0,96	0,85	**0,55**	**0,82**	0,9	0,88

- Observations:
  - ✓ Joint encoding of context and schema improve generalization
  - ✗ Slot-value transfers remain a challenge

DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

Fine-tuning BERT for schema-guided zero-shot DST

- **6** BERT fine-tuned models for prediction

Intent prediction				[CLS]	utterance	[SEP]	intent_1	[SEP]	...	[SEP]	intent_n	[SEP]		
Categorical slot prediction	[CLS]	utterance	[SEP]	slot	[SEP]	null	[SEP]	value_1	[SEP]	...	[SEP]	value_n	[SEP]	
Free-form slot prediction						[CLS]	utterance+null	[SEP]	slot	[SEP]				
Requested slot prediction						[CLS]	utterance+null	[SEP]	slot	[SEP]				
In-domain slot transfer prediction					[CLS]	service	[SEP]	utterance	[SEP]	slot	[SEP]			
Cross-domain slot transfer prediction				[CLS]	utterance	[SEP]	target slot	[SEP]	source slot	[SEP]				

- Adds (truncated) dialogue history to input

(Ruan et al., 2020)

DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

Evaluation results

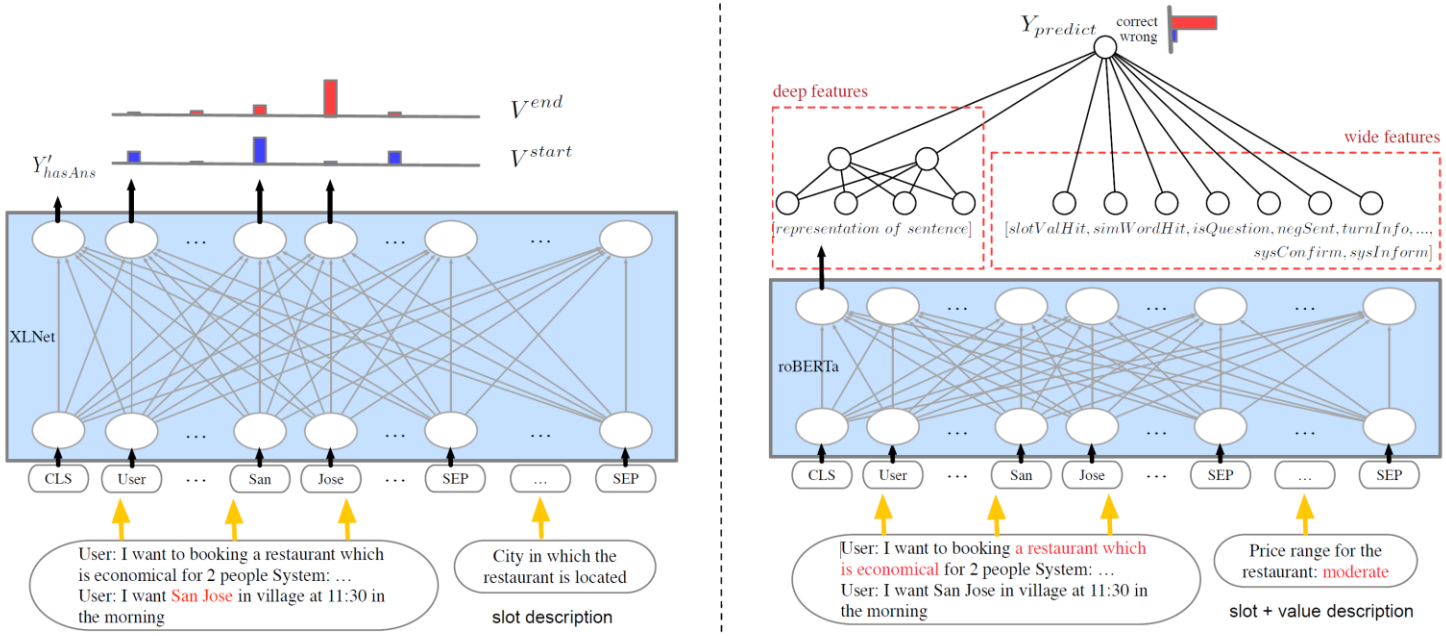
	All services				Seen services				Unseen services			
	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1
Baseline	0,25	0,56	0,91	0,97	0,41	0,68	0,95	1	**0,2**	0,52	0,89	0,95
Shi	0,54	0,8	0,91	**0,87**	0,53	0,75	0,96	0,85	**0,55**	**0,82**	0,9	0,88
Ruan	0,74	0,92	0,92	0,99	0,88	0,96	0,96	**1**	**0,69**	0,91	0,91	0,99

- Observations:
  - ✓Slot transfers significantly improve performance
  - ✗Joint GA drops considerably for unseen services

DSTC8: Schema-guided DST

Reading comprehension and wide & deep DST

- Span prediction for non-categorical slots on untruncated context
- „Wide & deep model“ for categorical slots: Transformer output + hand-crafted features
- Data augmentation via back-translations to vary schema descriptions



a. MRC model for span-based slot and numerical slot

b. Wide & Deep model for boolean and text-based slot

(Ma et al., 2020)

DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

Evaluation results

	All services				Seen services				Unseen services			
	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1	Joint GA	Avg GA	Intent Acc	Req Slot F1
Baseline	0,25	0,56	0,91	0,97	0,41	0,68	0,95	1	0,2	0,52	0,89	0,95
Shi	0,54	0,8	0,91	0,87	0,53	0,75	0,96	0,85	0,55	0,82	0,9	0,88
Ruan	0,74	0,92	0,92	0,99	0,88	0,96	0,96	1	0,69	0,91	0,91	0,99
Ma	**0,87**	**0,97**	**0,95**	0,98	**0,92**	**0,98**	**0,96**	0,99	**0,85**	**0,97**	**0,95**	0,98

- Important details:
  - Hand-crafted features are rule and heuristic based (+10% JGA)
  - Data augmentation by back-translation from Chinese (+6% JGA)
  - Partial delexicalization (phone numbers)
  - Development set used as additional training data

DSTC8: Schema-guided DST

Heinrich Heine Universität Düsseldorf

Analysis

- ✓ What worked?
  - Approach: Reading comprehension + classification
    - Alternative architectures are under-explored
  - Using hand-crafted features and additional data
    - Importance of auxiliary features indicates insufficiencies in semantic encoding
- ? What worked maybe?
  - Compartmentalizing: Specialized models for sub-tasks
    - Best systems employ multiple specialized encoders
    - Unified models are among most robust (benefiting from multitasking)

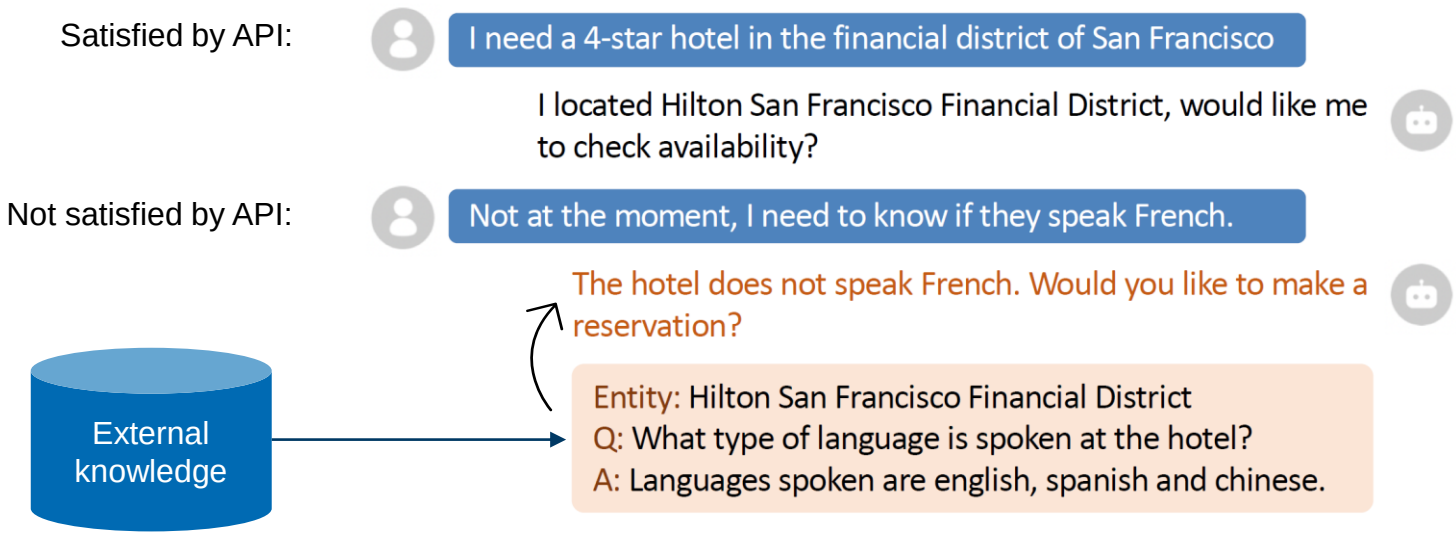
Enabling dialogue beyond ontologies

Enabling dialogue beyond ontologies

Heinrich Heine Universität Düsseldorf

DSTC9: Incorporating access to unstructured knowledge

- Enable systems to go beyond the abilities of their known APIs
- Unstructured knowledge serves as additional knowledge base during inference
- If a user request can not be satisfied by a known service/API/ontology?
  - Query an (unstructured) external knowledge source

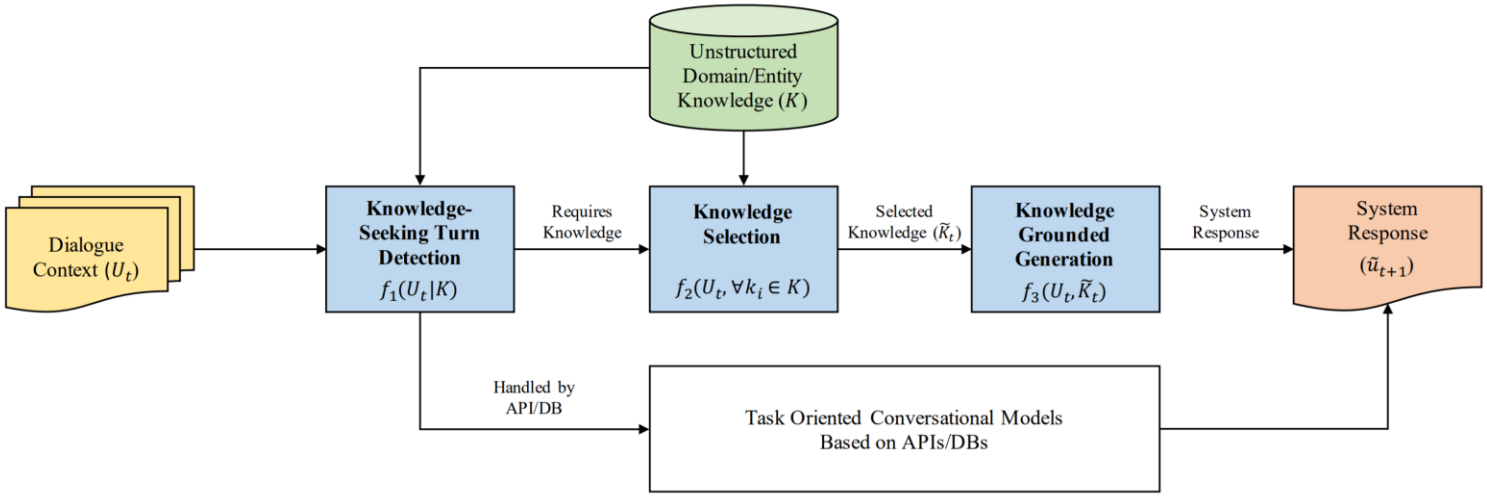


(Kim et al. 2020) (He et al. 2021)

Enabling dialogue beyond ontologies

DSTC9: Incorporating access to unstructured knowledge

- Knowledge seeking turn detection:
  - Does the current request need to be satisfied with external knowledge?
- Knowledge selection:
  - Which external knowledge satisfies the current request best?
- Knowledge grounded generation:
  - Generate a suitable response to the request



(Kim et al. 2020)

Enabling dialogue beyond ontologies

Heinrich Heine Universität Düsseldorf

DSTC9: Incorporating access to unstructured knowledge

- Challenge data:
  - Generalization to new domains required

	# Samples	# Knowledge Snippets	Domains	Locales
Train	71,348	2,900	hotel, restaurant, taxi, train	Cambridge
Valid	9,663	2,900	hotel, restaurant, taxi, train	Cambridge
Test	4,181	12,039	hotel, restaurant, taxi, train, attraction	Cambridge, San Francisco

Knowledge snippets in FAQ format
Domain (e.g. „hotel“)
Entity (e.g. „Hilton San Francisco)
Question (e.g. „are pets allowed?“)
Answer (e.g. „pets are not allowed“)

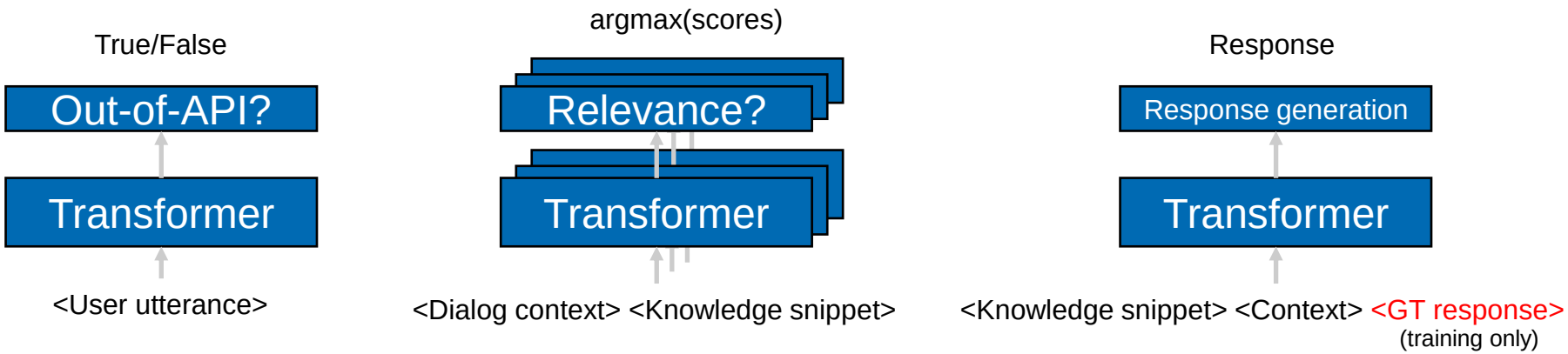
- Challenge outcomes:
  - Pipelined approaches are the vast majority
  - Little divergence from the baseline
  - Major improvements primarily due to extensive use of ensembling and additional data

(Kim et al. 2020) (He et al. 2021)

Enabling dialogue beyond ontologies

DSTC9: Baseline

- Knowledge seeking turn detection
  - Simple binary classification of current user turn, using a transformer encoder
  - Data augmentation: Treat external knowledge as user requests
- Knowledge selection
  - Dialog context and candidate knowledge snippet as joint input to transformer encoder
  - Compute relevance score for knowledge candidates and pick most probable fit
- Response generation
  - Generative transformer model produces response given context and knowledge candidate

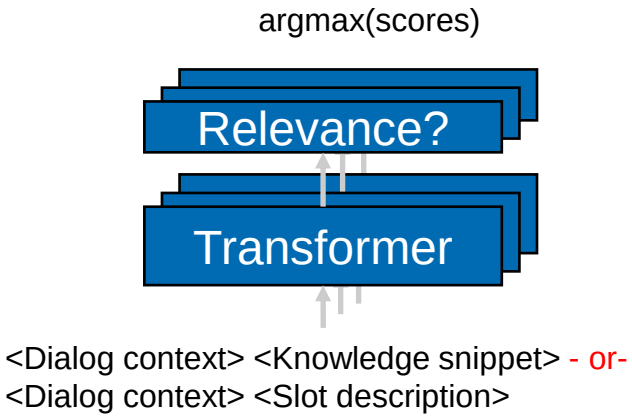


(Kim et al. 2020) (He et al. 2021)

Enabling dialogue beyond ontologies

DSTC9: Winning team

- Knowledge seeking turn detection
  - Adopts schema guided paradigm: slots and knowledge snippets are competing elements
  - Training uses negative sampling
  - Massively larger transformer (32 layers vs. 12 layers)
  - If a knowledge snippet is more likely to explain the context than any slot description, continue with knowledge selection



(He et al. 2021)

Enabling dialogue beyond ontologies

Heinrich Heine Universität Düsseldorf

DSTC9: Winning team

- Knowledge selection
  - Same approach as baseline, but
    - Massively larger transformer (32 layers vs. 12 layers)
    - Massive ensembling (7 large transformer models)
  - Multi-level negative sampling, i.e., negative sampling from set of different granularity
    - **Random**: negative sample from full set of knowledge snippets
    - **In-domain**: negative sample from same domain
    - **In-entity**: negative sample belonging to same entity
    - **Cross-entity**: negative sample belonging to previously encountered entity
    - -> Sample 1 negative sample from each level -> pos/neg ratio = 1/4

- Response generation
  - Similar to baseline
  - Different transformer type (32 layers, 32 attention heads, embedding dim. 2048)
  - Different training data (684M samples from Reddit)

(He et al. 2021)

Enabling dialogue beyond ontologies

Heinrich Heine Universität Düsseldorf

DSTC9: Results

	Task 1	Task 2		Task 3		
	F1	**R@1**	**R@5**	**BLEU-1**	**BLEU-4**	**Human score**
Baseline	0.9455	0.6201	0.8772	0.3031	0.0655	3.8271
Team 19	0.9886	0.9235	0.9814	0.3803	0.1081	4.3920

Achievements & Limitations

- ✓ First large-scale investigation into incorporating external knowledge in dialog flow
- ✓ Promising results with data augmentation (w/o need for massive ensembles)

- ✗ In-API and out-of-API handling are entirely separate processes
  - No inter-communication
  - No tracking (dialogue state becomes invalid after out-of-API request)

(Gunasekara et al. 2020) (Kim et al. 2020) (He et al. 2021)

Conclusion & Outlook

Conclusion & Outlook

- ✓ Ontology independence relies on semantic representations and semantic conditioning
  - ✗ Ways of semantic conditioning are underexplored
- ✓ Zero-shot abilities of models raised by data augmentation and smart negative sampling
  - ✗ Gap between performance on known and unknown domains still large
- DSTC8: Ontology independence
  - ✓ Huge step from baseline to winning approach (25% JGA to 87% JGA)
  - ✗ Zero-shot abilities are still limited
  - ✗ Expressiveness of evaluation is limited
- DSTC9: Handling requests beyond ontologies
  - ✓ Out-of-API requests can be detected and handled with very high accuracy
  - ✗ No proper integration yet, i.e. real dialogue systems with out-of-API handling

Thank you!

Heinrich Heine
Universität Düsseldorf

References

- Mrksic et al., 2017, Neural Belief Tracker - Data-Driven Dialogue State Tracking
- Henderson et al., 2014, Word-based dialog state tracking with recurrent neural networks
- Wen et al., 2017, A network-based end-to-end trainable task-oriented dialogue system
- Ramadan et al., 2018, Large-Scale Multi-Domain Belief Tracking with Knowledge Sharing
- Gao et al., 2019, Dialog state tracking: A neural reading comprehension approach
- Chao and Lane, 2019, BERT-DST: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer
- Kim et al., 2019, Efficient dialogue state tracking by selectively overwriting memory
- Zhang et al., 2019, Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking
- Bapna et al., 2017, Towards Zero-Shot Frame Semantic Parsing for Domain Scaling
- Shah et al., 2019, Robust Zero-Shot Cross-Domain Slot Filling with Example Values
- Rastogi et al., 2017, Scalable Multi-Domain Dialogue State Tracking
- Rastogi et al., 2020, Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset
- Rastogi et al., 2020, Schema-Guided Dialogue State Tracking Task at DSTC8
- Shi et al., 2020, A BERT-based Unified Span Detection Framework for Schema-Guided Dialogue State Tracking
- Gulyaev et al., 2020, Goal-Oriented Multi-Task BERT-Based Dialogue State Tracker
- Ruan et al., 2020, Fine-Tuning BERT for Schema-Guided Zero-Shot Dialogue State Tracking
- Ma et al., 2020, An End-to-End Dialogue State Tracking System with Machine Reading Comprehension andWide & Deep Classification
- Gunasekara et al., 2020, Overview of the Ninth Dialog System Technology Challenge: DSTC9
- Heck et al., 2020, TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking
- Kim et al., 2019, The Eighth Dialog System Technology Challenge
- Rastogi et al., 2020, Schema-Guided Dialogue State Tracking Task at DSTC8
- Zhang et al., 2019, Find or Classify? Dual Strategy for Slot-Value Predictions on Multi-Domain Dialog State Tracking
- Xu and Hu, 2018, An End-to-end Approach for Handling Unknown Slot Values in Dialogue State Tracking
- He et al., 2021, Learning to Select External Knowledge with Multi-Scale Negative Sampling