Conversations with Data

Toward more Interactive Natural Language Interfaces

Ahmed Awadallah

https://aka.ms/ahmed

Collaborators: Ahmed Elgohary, Adam Fourney, Saghar Hosseini, Chris Meek, Arpit Narechania, Alex Polozov, Gonzalo Ramos, Matt Richardson, Yu Su, Tao Yu

Microsoft

Microsoft Research

Language is a Universal Interface

AFIPS ‘73

Progress in natural language understanding—An application to lunar geology

by W. A. WOODS

Bolt Beranek and Newman Inc.
Cambridge, Mass.

INTRODUCTION

The advent of comp
(see e.g., Ornstein e
opportunity for acces
different computer f
expectations of a day
rather than an except
take to use a comput
whose languages, forr
him. In this foreseeab
number of different l
scientist would have t
much greater than tha
of his local computing
edgeable assistance is
Natural Language
hereafter refer to as
system to deal with th
tion problems by adap
ordinary natural Eng
adapt to the machine.

English as a query lan

TODS ‘74

SEVEN STEPS TO RENDEZVOUS WITH THE CASUAL USER

by

E. F. Codd
IBM Research Laboratory
San Jose, California

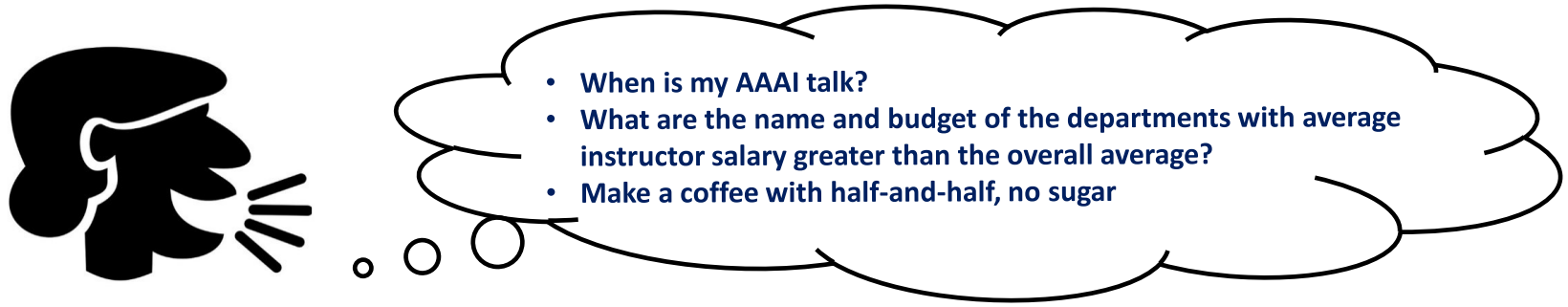
ABSTRACT:
data bas
presently

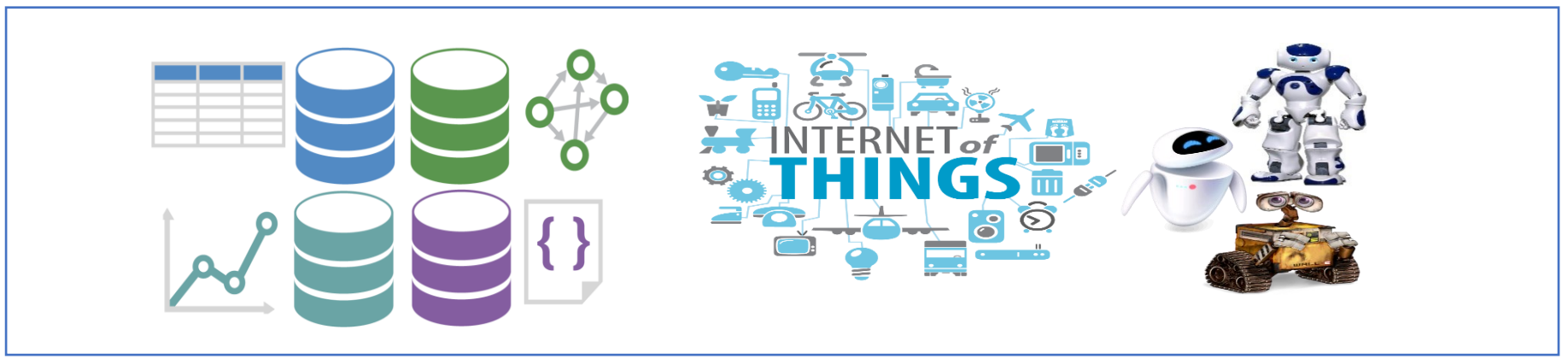
TODS ‘78

Developing a Natural Language Interface to Complex Data

GARY G. HENDRIX, EARL D. SACERDOTI, DANIEL SAGALOWICZ, AND JONATHAN SLOCUM

Natural language interface: One interface for all





Why Now?

- Bigger Opportunity
  - Massiveness and heterogeneity of data and accelerated digitization resulting in increasing need for improved *digital enablement*
- Better Technology
  - Advances in deep learning and program synthesis and availability of compute and benchmarks
- Growing Applications
  - Virtual assistants, language to code, NL search, database QA, etc.

NLIs and Digital Enablement

SQL

SPARQL

{ API }

When is my next meeting with Mike on marketing strategy

Show me the paper Susan sent me last week

Show me all high priority open bugs for Project Florence

Semantic Parsing

NL2SQL

Find all locations whose name contains the word “film”

SELECT Address FROM Locations WHERE Location_Name
LIKE “%film%”;

NL2API

Show me the latest unread messages about AAAI workshop

GET messages? filter=isRead eq false & $search=“AAAI workshop”
& orderby=receivedDateTime desc

Beyond one-shot Semantic Parsing

Show me the latest messages about AAAI workshop that I haven't read

Correction

I want only the messages marked as unread

Follow-up/
decomposition

Were any of them sent by John?

Do you mean John A. or John B.?

Clarification

I meant John B

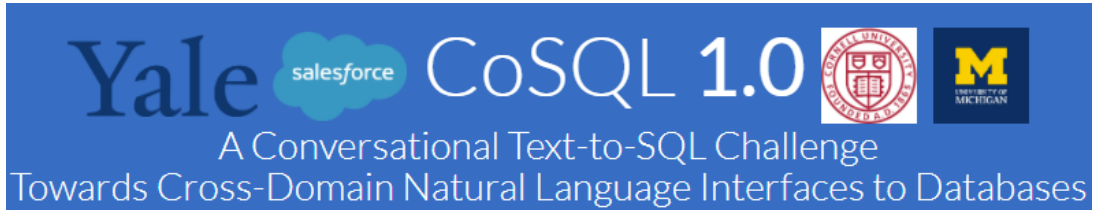
Beyond one-shot Semantic Parsing

Richer Contextual Representations

Richer Models of User Interactions

Richer Contextual Representations

Conversational Semantic Parsing (CSP) is the task of converting a sequence of natural language queries to formal language



MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling

Paweł Budzianowski[1], Tsung-Hsien Wen[2*], Bo-Hsiang Tseng[1],
Iñigo Casanueva[2*], Stefan Ultes[1], Osman Ramadan[1] and Milica Gašić[1]
[1]Department of Engineering, University of Cambridge, UK,
[2]PolyAI, London, UK

Microsoft Research Sequential Question Answering (SQA) Dataset

Recent work in semantic parsing for question answering has focused on long and complicated questions, many of which would seem unnatural if asked in a normal conversation between two humans. In an effort to explore a conversational QA setting, we present a more realistic task: answering sequences of simple but inter-related questions. We created SQA by asking crowdsourced workers to decompose 2,022 questions

SParC:
Sequential Text-to-SQL

CoSQL:
Conversational Text-to-SQL

MultiWOZ:
Task-oriented Dialogue

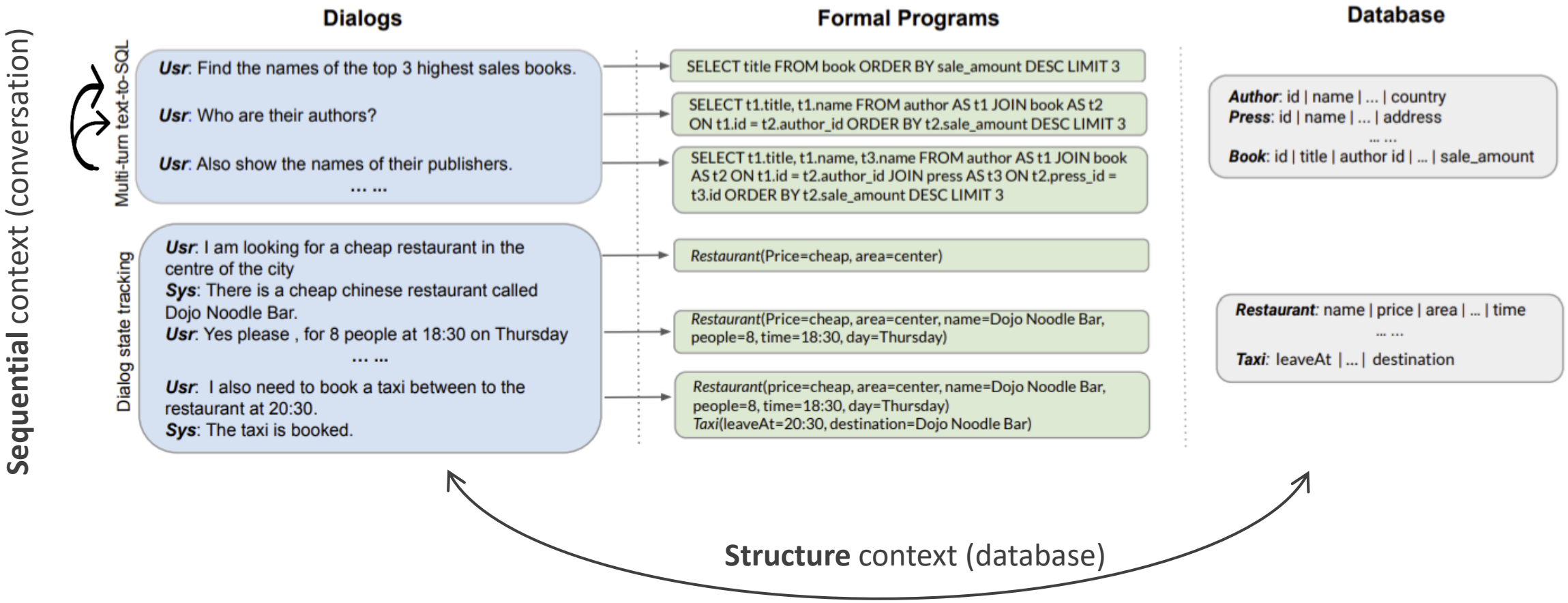
SQA:
Weakly supervised Table QA

Multiple tasks, shared challenges

- Requires lots of annotated data
- Annotation is expensive, hard to collect and not always of good quality
- Learning to represent sequential (conversation) and structure (ontology) contexts is hard

Conversational Semantic Parsing

Challenge: representation context with respect to the sequence of utterances (conversation) and the structure of the underlying ontology (database)



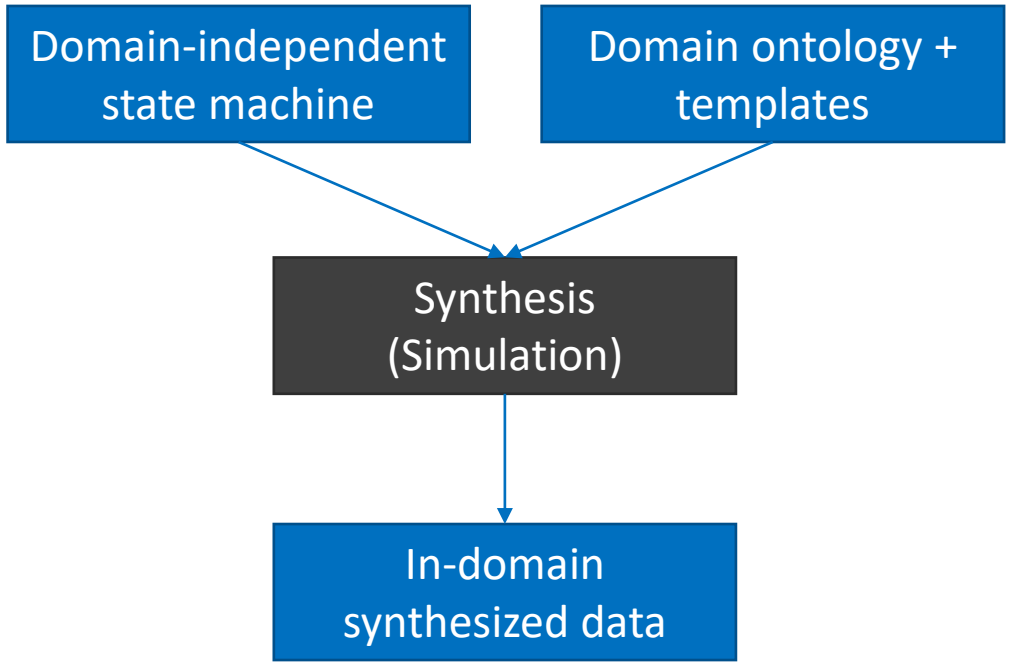
Pre-training for Context Representation

1. Focus on pretraining, not finetuning (exploit shared characteristics).

2. CSP-specific objectives.

3. Pretrain *only* on synthetic data, MLM *only* on natural data.

Synthetic Data Generation



Total of 120k synthetic task-oriented dialogues

S: START
U: Can you help with information regarding a food place? I need to book at 15:45.
S: SEARCHREQUEST restaurant(book_time = "15:45")
A: How about the restaurant with name La Tasca and Italian food?
U: Can you find something which serves seafood?
S: SEARCHREQUEST restaurant(book_time = "15:45", food = "seafood")
A: What date are you looking for?
U: Thursday please.
S: SEARCHREQUEST restaurant(book_time = "15:45", food = "seafood", book_day = "thursday")
A: How about the Copper Kettle? It is a food place with seafood food.
U: What is the price range and the area?
S: SLOTQUESTION restaurant(book_time = "15:45", food = "seafood", book_day = "thursday", price_range =?, area =?)
A: The Copper Kettle is a moderately priced restaurant in the north of the city. Would you like a reservation?
U: No, thanks.
S: CLOSE restaurant(book_time = "15:45", food = "seafood", book_day = "thursday")

Campagna et al., ACL'20

Synthetic Data Generation

Turn #	Question-SQL Template	Synthesized Question-SQL
1	"Find the number of TABLE0 with COLUMN0 OP0 VALUE0" SELECT COUNT(*) ORDER BY COLUMN0 OP0 VALUE0	"Find the number of football team with team hometown is not murrieta, california?" SELECT COUNT(*) WHERE TEAM_HOMETOWN != "MURRIETA, CALIFORNIA"
2	"Can you give me their COLUMN1?" TCS: REPLACE(SELECT.COLUMN0), DEL(SELECT.AGG)	"Can you give me their football team player?" SELECT FOOTBALL_TEAM_PLAYER WHERE TEAM_HOMETOWN != "MURRIETA, CALIFORNIA"
3	"How about only show those with AS0 COLUMN2?" TCS: ADD(ORDERBY_AS0.COLUMN2)	"How about only show those with the largest age?" SELECT FOOTBALL_TEAM_PLAYER WHERE TEAM_HOMETOWN != "MURRIETA, CALIFORNIA" ORDER BY AGE DESC LIMIT 1
4	"AS1?" TCS: REPLACE(ORDERBY_AS1.COLUMN2)	"The smallest?" SELECT FOOTBALL_TEAM_PLAYER WHERE TEAM_HOMETOWN != "MURRIETA, CALIFORNIA" ORDER BY AGE AS LIMIT 1

Created a total of 435k text-to-SQL conversations based on 400K tables in WikiTABLES

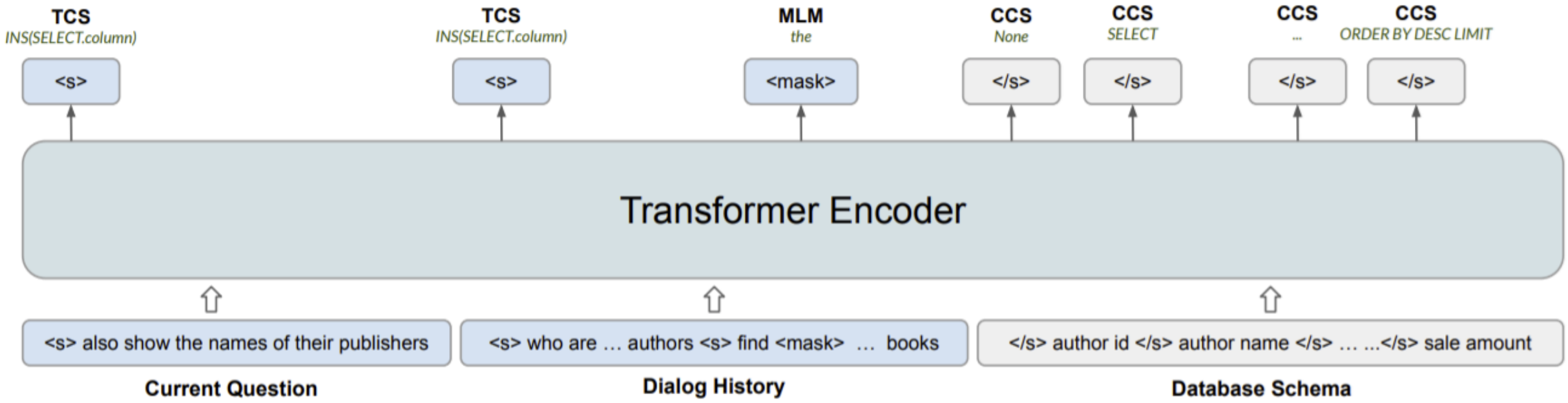
Pre-Training Objectives

Turn Contextual Switch (TCS):

- Aims to represent *diff* between the current and previous SQL

Column Contextual Semantics (CCS):

- *Aims to represent the operation* expected on each schema item



(Yu et. al, 2021)

Significant improvement over all baselines, 3 SOTA results

Sequential Text-to-SQL (SPARC)

Conversational Text-to-SQL (CoSQL)

	SPARC				CoSQL			
	Dev		Test		Dev		Test	
Models	QM	IM	QM	IM	QM	IM	QM	IM
SyntaxSQL (Yu et al., 2018a)	18.5	4.3	20.2	5.2	-	-	14.2	2.2
GAZP + BERT (Zhong et al., 2020)	48.9	29.7	45.9	23.5	42.0	12.3	39.7	12.8
EditSQL + BERT (Zhang et al., 2019b)	47.2	29.5	47.9	25.3	39.9	12.3	40.8	13.7
IGSQL + BERT	50.7	32.5	51.2	29.5	44.1	15.8	42.5	15.0
R^2SQL + BERT	-	-	55.8	30.8	-	-	46.8	17.0
RAT-SQL + BERT (Wang et al., 2019)	56.8	33.4	-	-	48.4	19.1	-	-
+ RoBERTa	58.2	36.7	-	-	50.1	19.3	-	-
+ SCoRe	**62.2**	**42.5**	**62.4**	**38.1**	**52.1**	**22.0**	**51.6**	**21.2**

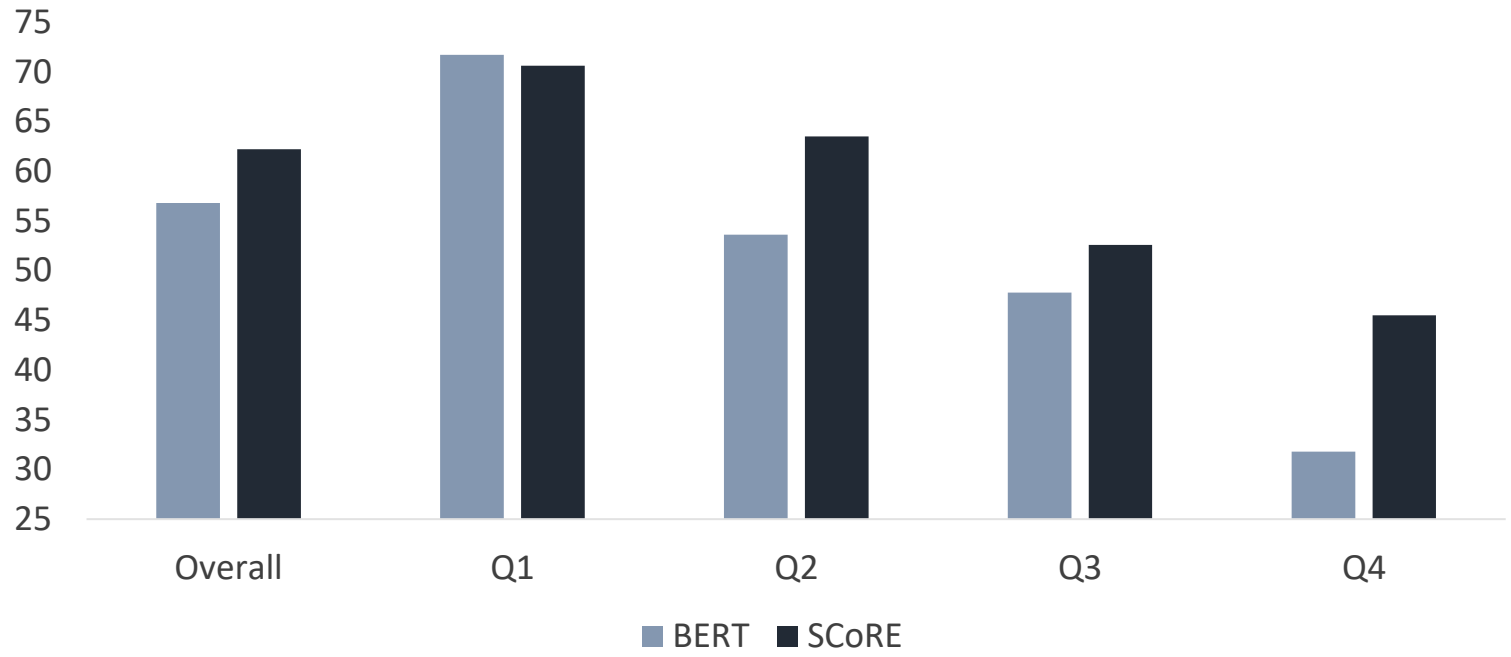
Sequential Question Answering (SQA)

	SQA	
Models	QM	IM
Iyyer et al. (2017)	44.7	12.8
Sun et al. (2019a)	45.6	13.2
Müller et al. (2019)	55.1	28.1
Herzig et al. (2020b)	**67.2**	**40.4**
Wang et al. (2019) + RoBERTa	62.8	33.2
with 10% training data	53.3	21.2
Wang et al. (2019) + SCoRe	65.4	38.1
with 10% training data	57.1	26.1

Dialog State Tracking (MultiWOZ 2.1)

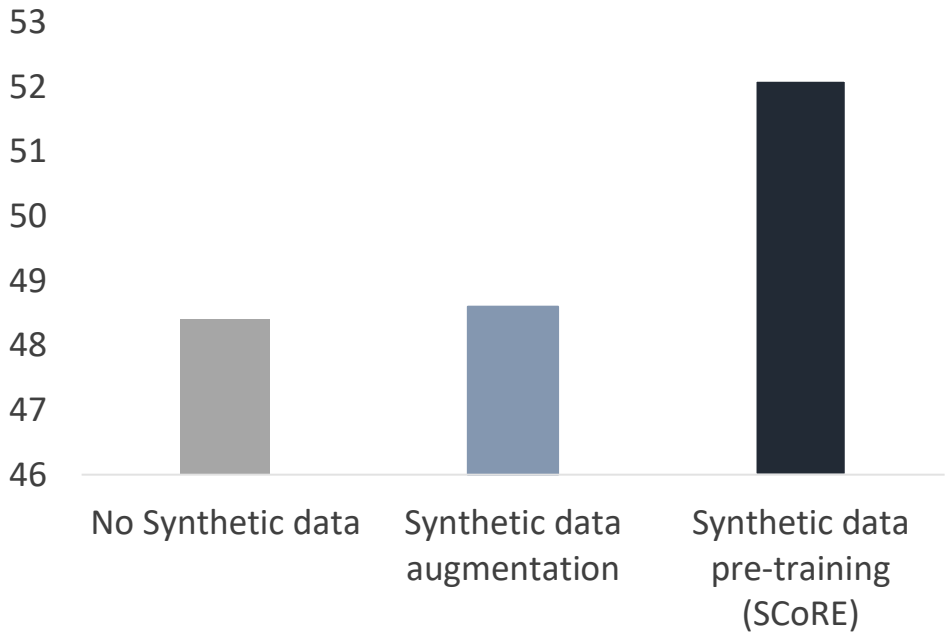
Models	MultiWOZ 2.1
DS-DST	51.21
SOM-DST	52.57
DS-picklist	53.30
TripPy	55.29
SimpleToD	55.72
TripPy (ours)	58.37
+ SCoRe	**60.48**

Accuracy significantly improves on every turn except the first
(in which the task is effectively a single-turn semantic parsing)



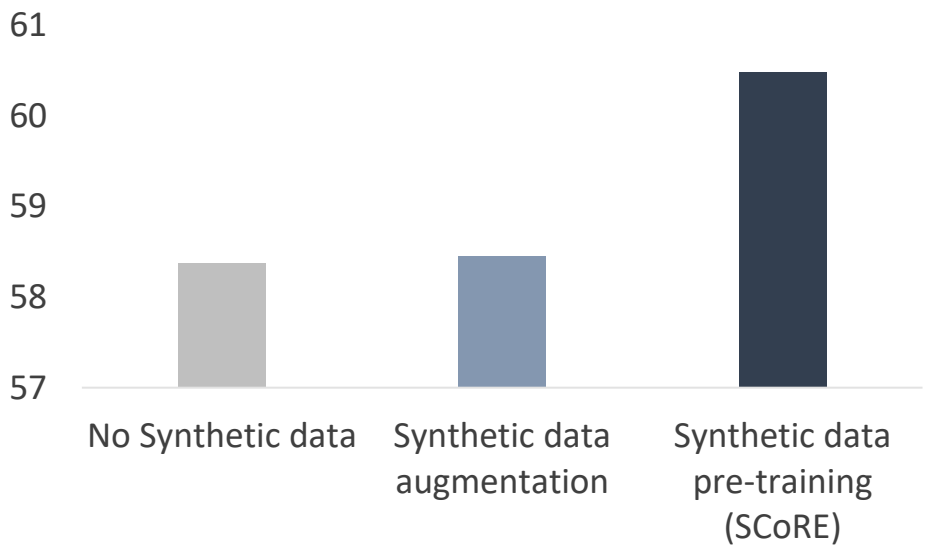
Results for the SQA dataset.
Other datasets exhibit similar behavior.
Comparison with RoBERTa exhibits similar behavior

Pre-training on the synthesized data with the new training objectives is much more effective than using it for data augmentation



CoSQL dataset

Base Model: RAT-SQL + BERT



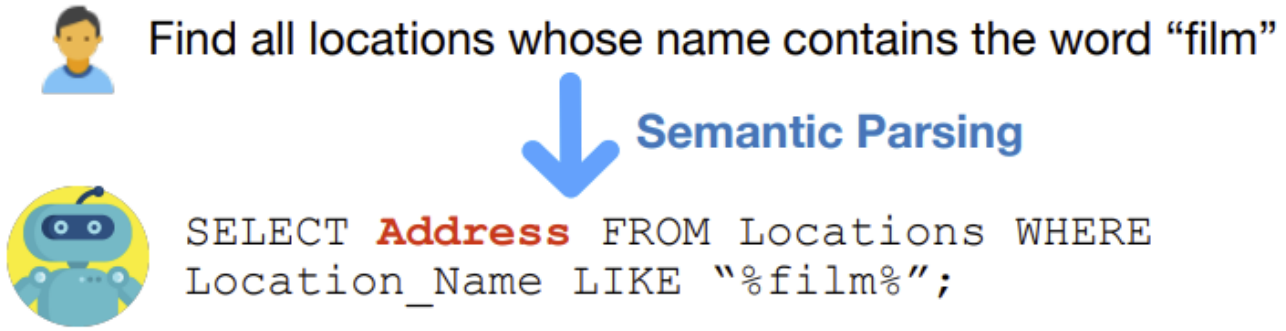
MultiWoZ 2.1 dataset

Base Model: TripPy

Richer Models of User Interactions

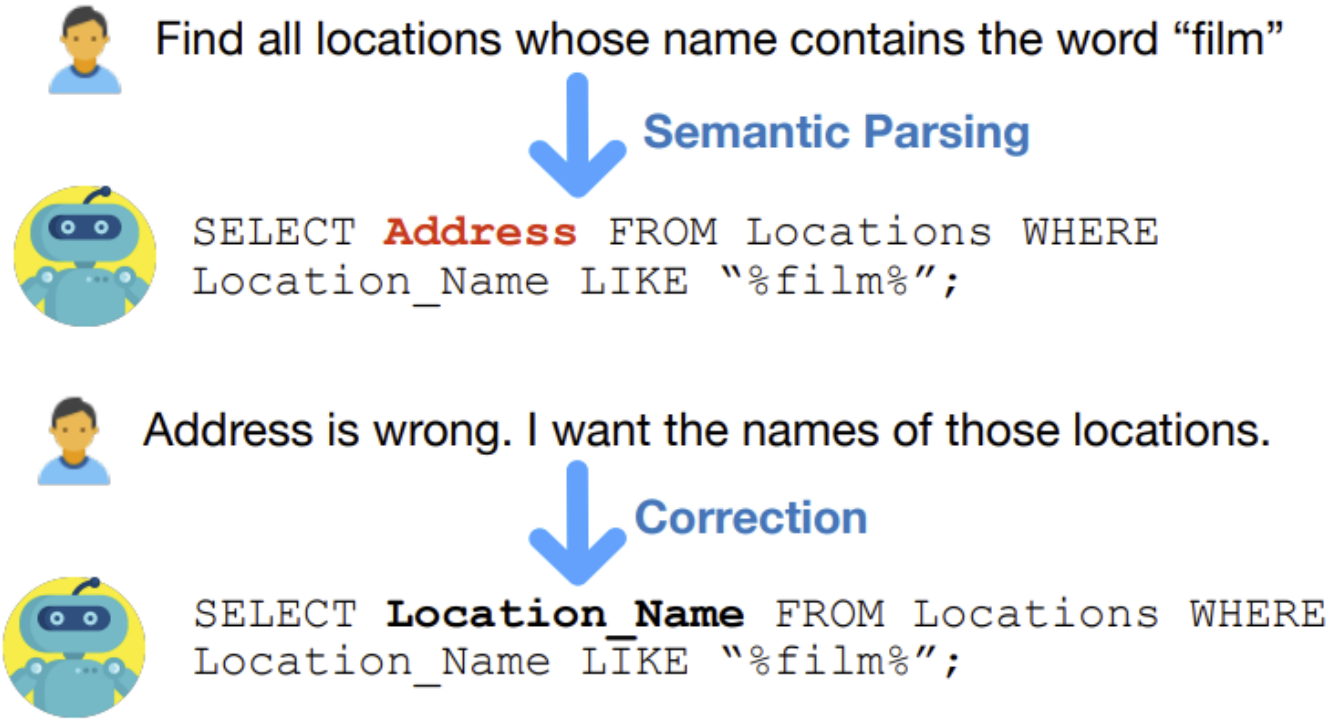
Motivation

- Traditional Semantic Parsing : one-shot translation of an utterance to a corresponding logical form



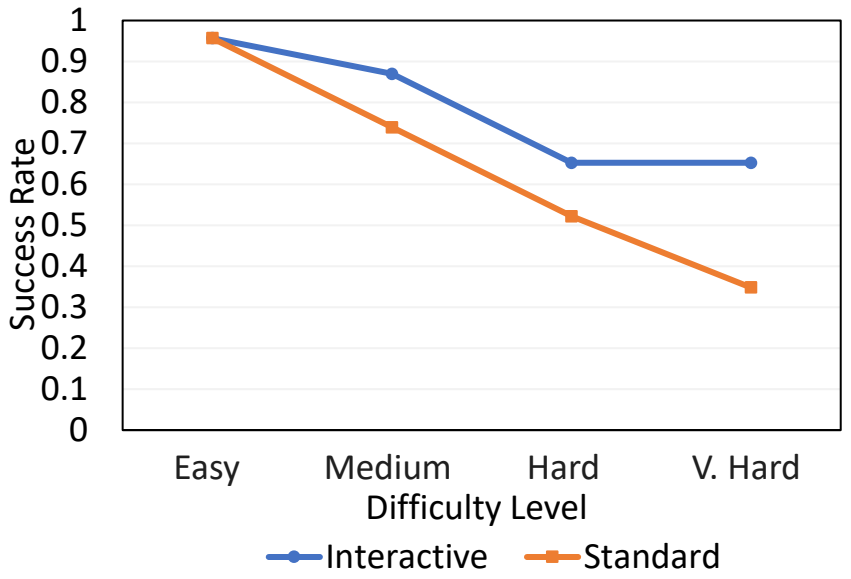
Motivation

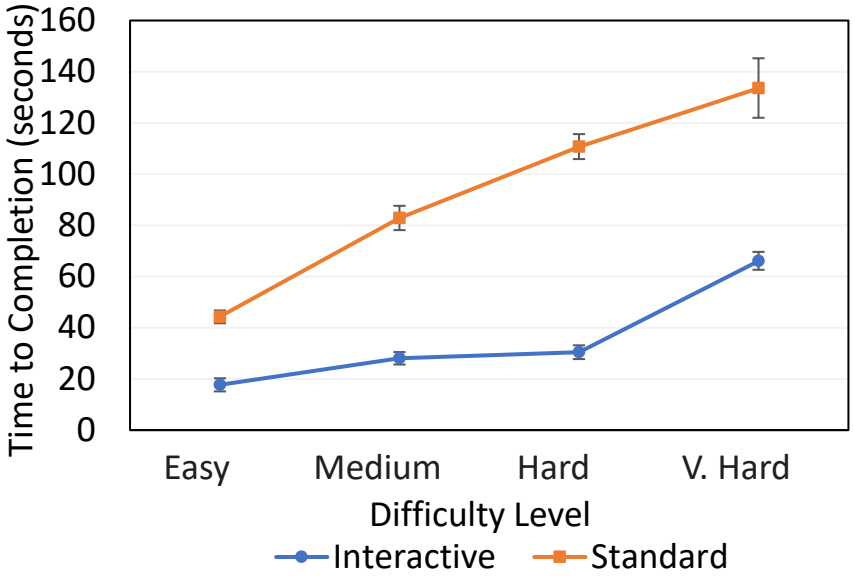
- Interactive Semantic Parsing: humans can further interact with the system by providing free-form natural language feedback to correct the system when it generates an inaccurate interpretation



Motivation

- Many **Semantic Parsing Errors** are minor and can be corrected if humans have a way to continue interacting with the system to correct them





(Su et. al, 2018)

Semantic Parsing Correction with Natural Language Feedback

Utterance: Find all locations whose name contains the word “film”

Initial Parse: `SELECT Address FROM Locations WHERE Location_Name LIKE “%film%”;`

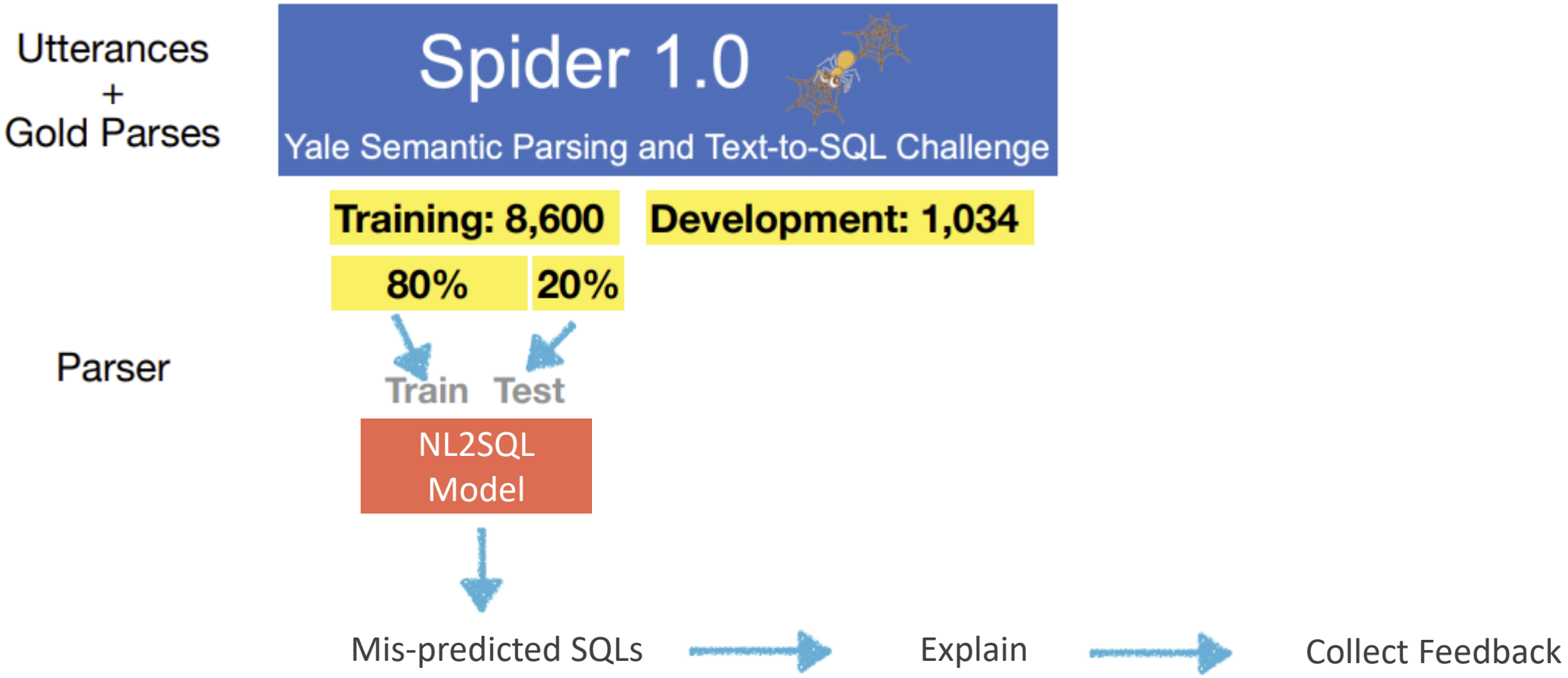
Feedback: Address is wrong. I want the names of those locations.

Schema: Location_ID | Location_Name | Address | Other_Details

Corrected Parse: `SELECT Location_Name FROM Locations WHERE Location_Name LIKE “%film%”;`

(ElGohary et. al, 2020)

SPLASH: A Dataset of NL Correction



SPLASH: Explaining SQL

SQL:

```
SELECT Id, Name FROM Browser GROUP BY Id
ORDER BY COUNT(*) DESC
```

Template: SELECT **$cs0** FROM **$t0** GROUPBY **$c0** ORDERBY **$aggr0** **$c1**

Explanation: Step 1: Find the number of rows of each value of id in browser table.
Step 2: Find id, name of browser table with largest value in the results of step 1.

SPLASH: Explaining SQL

- Types of Feedback

Feedback Type	%	Example
Output		
- Unneeded	4%	No need to return email address
- Missing	13%	I also need the number of different services
- Wrong	36%	Return capacity in place of height
Conditions		
- Unneeded	7%	Return results for all majors
- Missing	10%	Ensure they are FDA approved
- Wrong	19%	Need to filter on open year not register year
Order/Distinct	5%	Only return unique values
Aggregation	6%	I wanted the smallest ones not the largest

SPLASH: Explaining SQL

- Types of Feedback

  - Complete Feedback: 81.5%

    Question: Show the types of schools that have two schools
    Pred. SQL: `SELECT TYPE FROM school GROUP BY TYPE HAVING count(*) >= 2`
    Feedback: You should not use greater than.

  - Partial Feedback: 13.5%

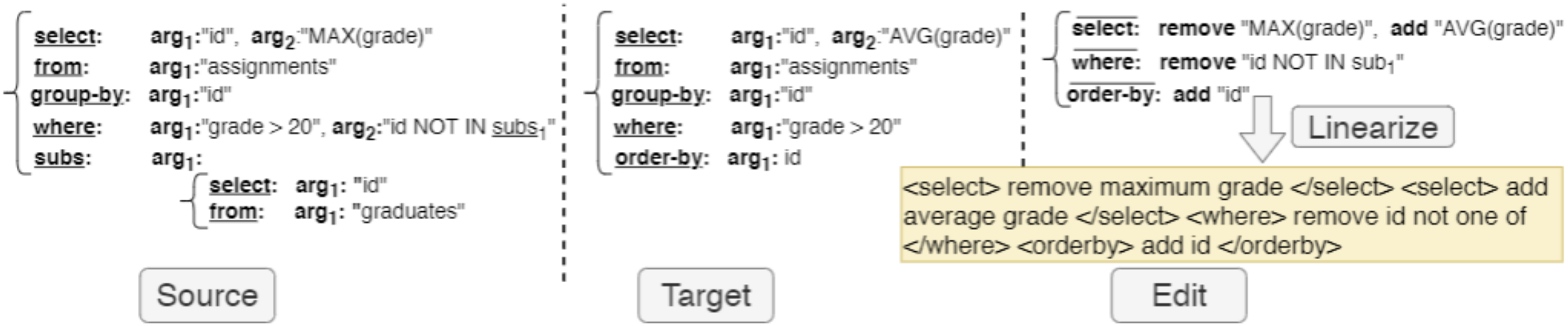
    Question: What are the names of all races held between 2009 and 2011
    Pred. SQL: `SELECT country FROM circuits WHERE lat BETWEEN 2009 AND 2011`
    Feedback: You should use races table.

  - Paraphrase Feedback: 5.0%

    Question: What zip codes have a station with a max temperature greater than or equal to 80 and when did it reach that temperature
    Feedback: Find date , zip code whose max temperature f greater than or equals 80

Learning to Edit with NL Feedback

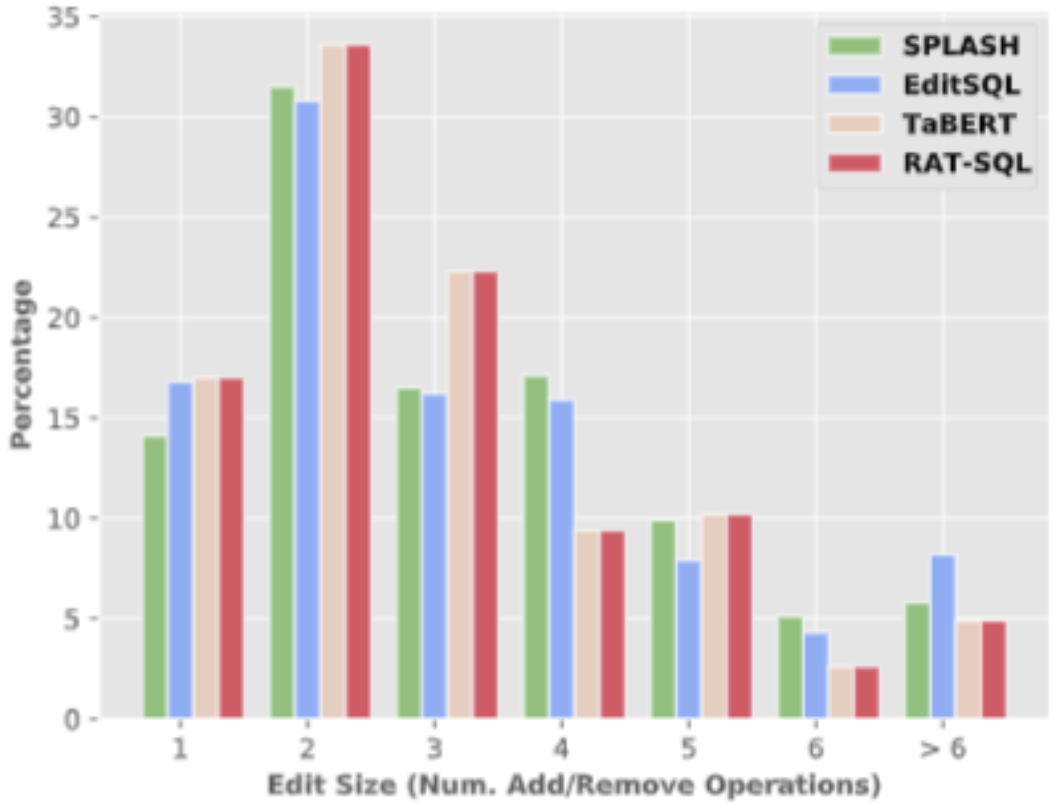
Difference between initial incorrect parse (source) and correct parse (target) is a set of edit operations



(ElGohary et. al, 2021)

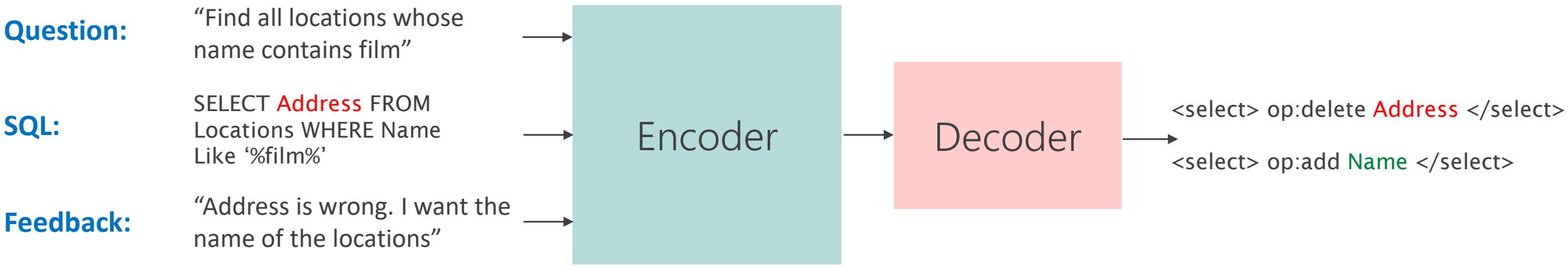
Learning to Edit with NL Feedback

Most corrections involve a small number of edits



Learning to Edit with NL Feedback

Learning to generate edits to correct mistakes based on open-form feedback



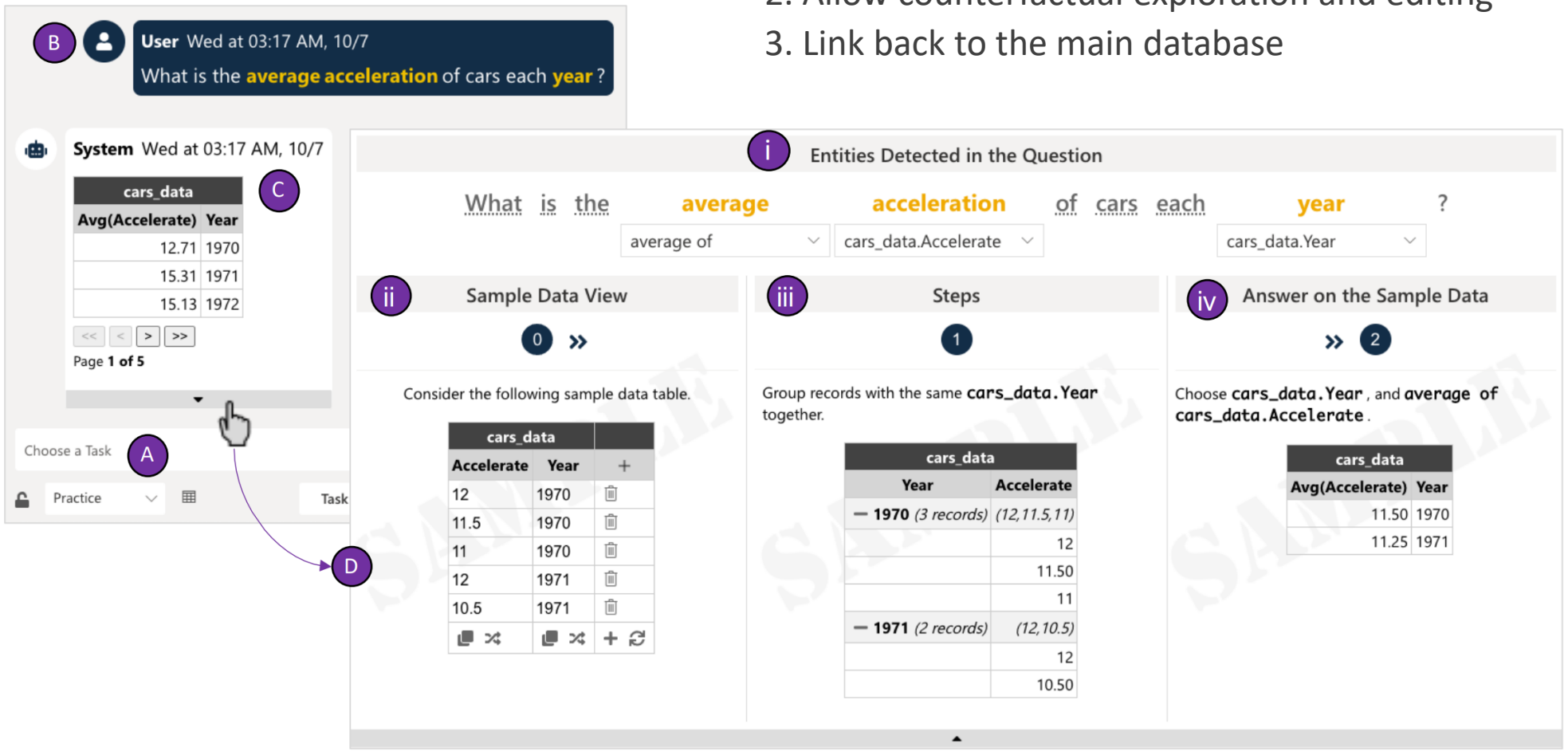
Learning to Edit with NL Feedback

Model	Correction accuracy (%)
Without Feedback:	
- Re-ranking: beam	11.9
- Re-ranking: parser score	11.3
With Feedback:	
- Re-ranking	16.6
- Re-generation (EditSQL)	25.2
With Feedback:	
- Learning to Edit (NLEdit)	41.4

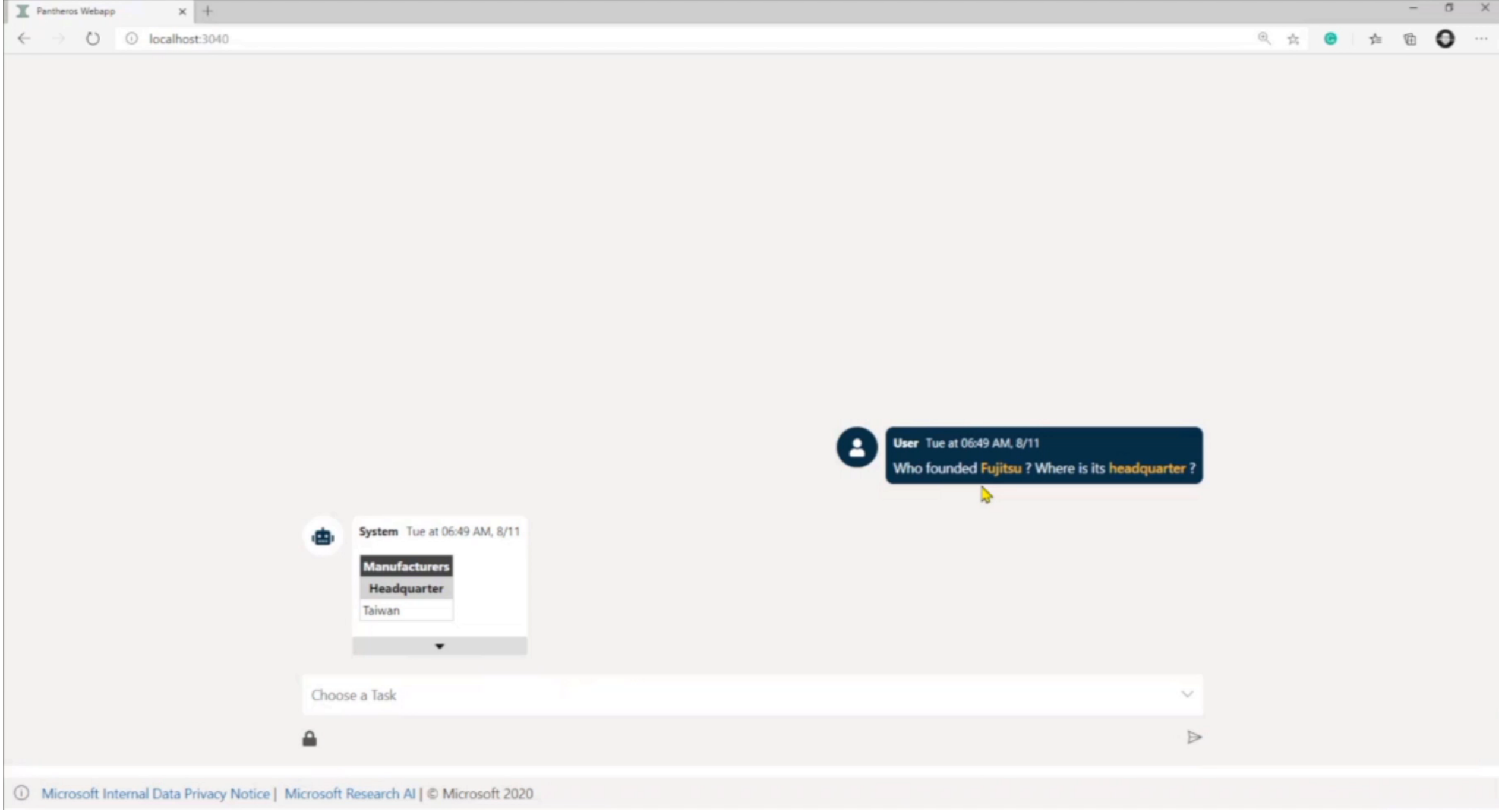
Multi-Modal Interactions as Feedback

"Debug-it-Yourself" Interface:

1. A small-but-relevant example is created
2. Allow counterfactual exploration and editing
3. Link back to the main database



(Narechania et. al, 2021)


Pantheros Webapp
localhost:3040
User Tue at 06:49 AM, 8/11
Who founded Fujitsu ? Where is its headquarter ?
System Tue at 06:49 AM, 8/11
Manufacturers
Headquarter
Taiwan
Choose a Task
Microsoft Internal Data Privacy Notice | Microsoft Research AI | © Microsoft 2020

Take-aways

Richer Contextual Representations

- Importance of leveraging context from interactions and underlying ontology (data)
- Leveraging common challenges across multiple tasks
- Pre-training as a method for contextualization
- Better context representation leads to better few-shot learning abilities

Take-aways

Richer Models of User Interactions

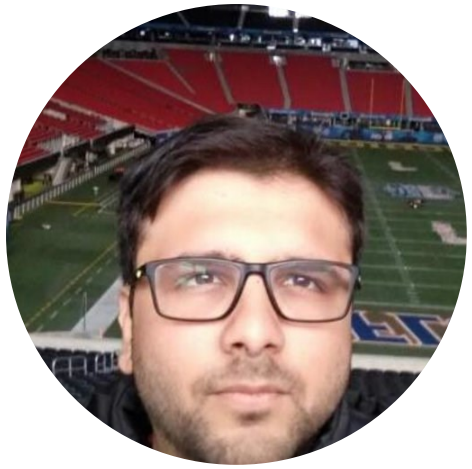
- Toward more collaborative AI systems that can use the user as a teacher
- Richer interaction can lead to better user satisfaction
- Richer models for feedback (binary, natural language , multimodal feedback)
- Interactivity as part of task definition and system evaluation

Take-aways

Many more challenges

- Data Collection and generalization
- Interactivity, continuous learning, and personalization
- Explainability, privacy and trustworthiness
- Evaluation and benchmarks

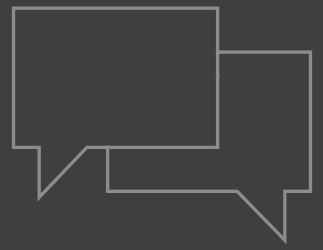
Ahmed Elgohary
University of Maryland



Arpit Narechania
Georgia Tech

Tao Yu
Yale

Collaborators: Ahmed Elgohary, Adam Fourney, Saghar Hosseini, Chris Meek, Arpit Narechania, Alex Polozov, Gonzalo Ramos, Matt Richardson, Yu Su, Tao Yu



Thank you

https://aka.ms/ahmed
https://aka.ms/Conversations-With-Data

Microsoft