Stanford CS224v Course

Conversational Virtual Assistants with Deep Learning

Lecture 6

Introduction to Agents for Structured and Unstructured Data

Monica Lam & Shicheng Liu

Why Knowledge Bases?

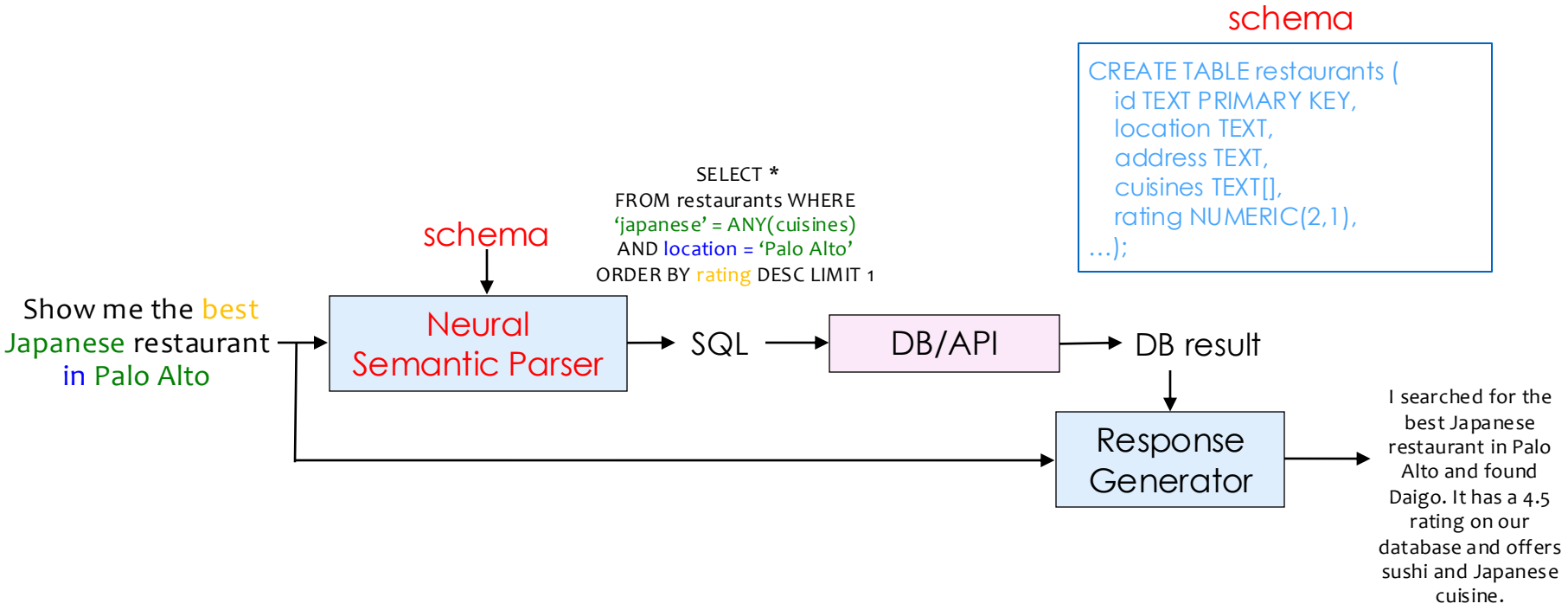
Lots of structured data in the world

- Public databases
  - Schema.org (structured data for the web)
  - Wikidata (knowledge graph with 15 B facts)
- Private databases
  - Every company is a data company [Forbes 2018]
    - Products: Retail (all of ebay), flights, restaurants
      songs, music, books
    - Corporate: People (employees, customers, students, patients)
      Operations (finance, sales, transactions)
    - Personal information: Calendar, emails

Query Languages

- A FEW common query languages:
  - Databases (SQL): Tables with a fixed schema – this lecture
  - Knowledge bases (SPARQL, New4j): – future lecture
    Graphs with nodes (entities), edges (properties)
- Domains are defined by database schemas

Semantic Parsing: NL → Formal Query



Lecture Goals

Introduction to the Problems:

1. Retrieving Data from Databases
2. Retrieving Data from Hybrid Corpora (Databases and Free text)

Focusing only on simple queries in this introduction.

Outline

- **Casual questions are formal queries**
- A basic LLM-based agent that consults a database – Yelp
- Evaluation issues of DB query agents
- Motivation of hybrid corpora
- Techniques for hybrid corpora

Expressiveness of Query Languages

- Domain agnostic
- All queries of any domain in knowledge bases
  are compositions of a FEW relational algebra operations
  - Basic: Selection, Projection, Cartesian product (Join),
    Union, Set Difference
  - Extended: Sort, Aggregate Operators (Sum, Max, Avg, …)
- Expressive, succinct, well-defined

Amazing CS Idea

WHERE clause: filter

```
CREATE TABLE restaurants (
    id TEXT PRIMARY KEY,
    location TEXT,
    address TEXT,
    cuisines TEXT[],
    rating NUMERIC(2,1),
...);
```

SELECT * FROM
table [WHERE *filter*]?

SELECT * FROM restaurants
WHERE location == "Palo Alto"
AND 'chinese' = ANY (cuisines)

Show me Chinese restaurants in Palo Alto

WHERE clause: filter

```
CREATE TABLE restaurants (
    id TEXT PRIMARY KEY,
    location TEXT,
    address TEXT,
    cuisines TEXT[],
    rating NUMERIC(2,1),
...);
```

SELECT * FROM
table [WHERE *filter*]?

SELECT * FROM restaurants
WHERE location = "Palo Alto"
AND "Chinese" = ANY(cuisines)
AND rating >= 4.5

Show me Chinese restaurants in Palo Alto
with at least 4.5 stars

SELECT clause: Projection

```
CREATE TABLE restaurants (
    id TEXT PRIMARY KEY,
    location TEXT,
    address TEXT,
    cuisines TEXT[],
    rating NUMERIC(2,1),
...);
```

SELECT field+ FROM
table [WHERE *filter*]?

SELECT address FROM restaurants
WHERE location = "Palo Alto"
AND "Chinese" = ANY(cuisines)
AND rating >= 4.5

Show me the address of Chinese restaurants in Palo Alto with at least 4.5 stars

Subquery

```
CREATE TABLE restaurants (
    id TEXT PRIMARY KEY,
    location TEXT,
    address TEXT,
    cuisines TEXT[],
    rating NUMERIC(2,1),
...);
```

```
CREATE TABLE reviews (
    id TEXT PRIMARY KEY,
    restaurant_id TEXT
    REFERENCES restaurants(id),
    author TEXT,
...);
```

Atom filter:
param op value

Subquery filter:
param op
*(SELECT * FROM …)*

SELECT field+
FROM *table* [WHERE *filter*]?

```
SELECT address FROM restaurants
WHERE location = "Palo Alto"
AND "Chinese" = ANY(cuisines)
AND id IN
(SELECT restaurant_id FROM reviews WHERE author =~ "Bob" )
```

Show me the address of Chinese restaurants in Palo Alto reviewed by Bob

Sorting

```
CREATE TABLE restaurants (
    id TEXT PRIMARY KEY,
    location TEXT,
    address TEXT,
    cuisines TEXT[],
    rating NUMERIC(2,1),
...);
```

```
CREATE TABLE reviews (
    id TEXT PRIMARY KEY,
    restaurant_id TEXT
    REFERENCES restaurants(id),
    author TEXT,
...);
```

SELECT field+ FROM *table* [WHERE *filter*]?
ORDER BY *field* DESC/ASC

```
SELECT address FROM restaurants
WHERE location = "Palo Alto"
AND "Chinese" = ANY(cuisines)
AND id IN
(SELECT restaurant_id FROM reviews WHERE author =~ "Bob" )
ORDER BY rating DESC
```

Show me the address of top-rated Chinese restaurants in Palo Alto reviewed by Bob

Joins

```
CREATE TABLE restaurants (
    id TEXT PRIMARY KEY,
    location TEXT,
    address TEXT,
    cuisines TEXT[],
    rating NUMERIC(2,1),
...);
```

```
CREATE TABLE reviews (
    id TEXT PRIMARY KEY,
    restaurant_id TEXT
    REFERENCES restaurants(id),
    author TEXT,
...);
```

SELECT * FROM
table join *table* [WHERE *filter*]? [on *filter*]?

```
SELECT * FROM restaurants JOIN reviews ON
restaurants.id = reviews.restaurant_id
WHERE location = "Palo Alto"
AND "Chinese" = ANY(cuisines)
ORDER BY rating DESC
```

Show me top-rated Chinese restaurants in Palo Alto
with their reviews

Outline

- Our casual questions are formal queries
- **A basic LLM-based agent that consults a database – Yelp**
- Evaluation issues of DB query agents
- Motivation of hybrid corpora
- Techniques for hybrid corpora

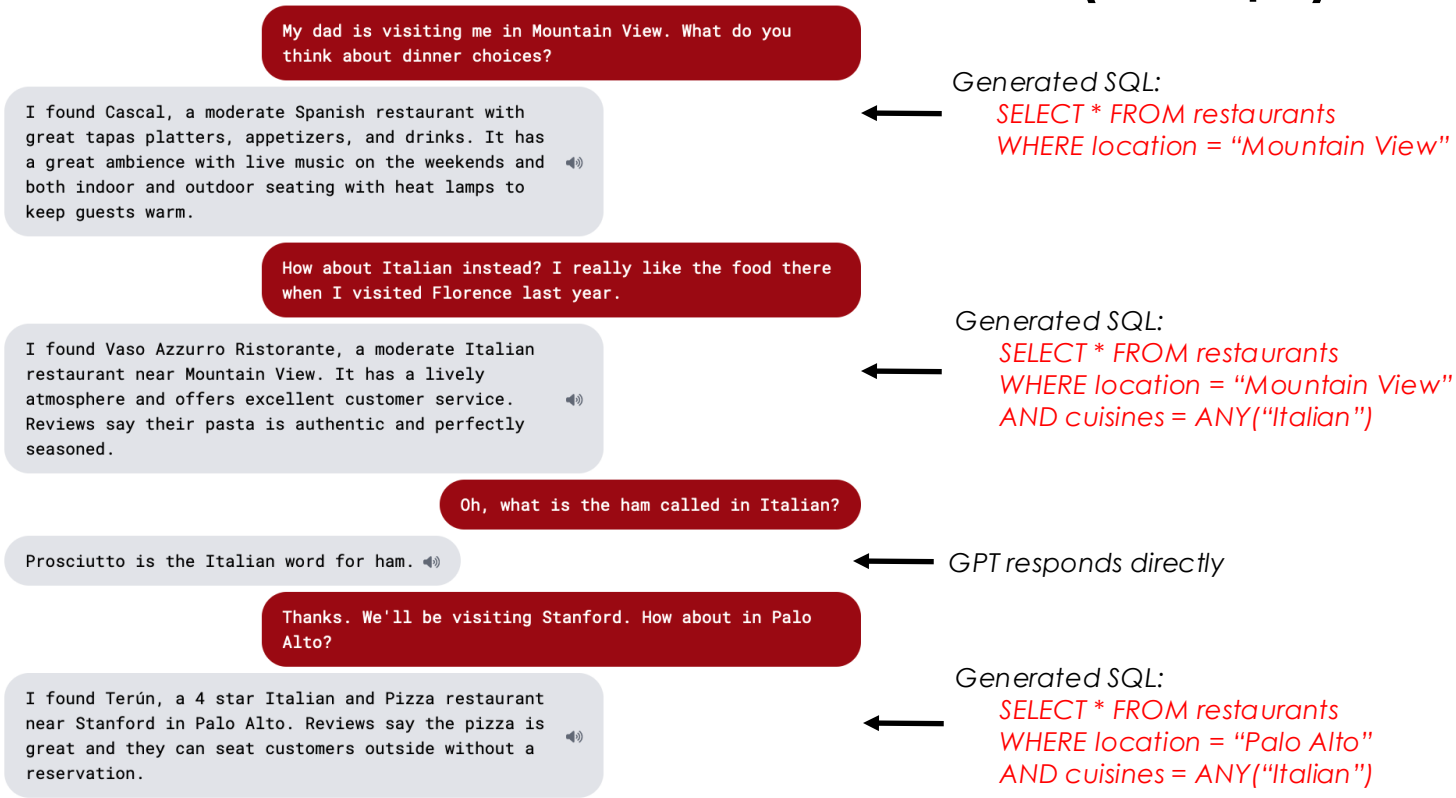
Restaurant Assistant

Schema

```
CREATE TABLE restaurants (
    name            TEXT,
    address         TEXT,
    location        TEXT,
    phone_number    TEXT,
    opening_hours   TEXT,
    cuisines        TEXT[],
    price           ENUM ('cheap', 'moderate', 'expensive', 'luxury'),
    rating          NUMERIC(2,1),
    num_reviews     NUMBER,
    reviews         FREE_TEXT,
    popular_dishes  FREE_TEXT,
);
```

NUMERIC(2,1) means 2-digit precision, 1 digit after decimal

Queries in a Conversation (Yelp)



NL → Formal Database Queries

Natural Language is compositional

A Casual NL Statement Can Correspond to Formal Queries with many clauses

Quiz: For Yelp,

how many possible queries are there?

LLM: Zero-Shot Prompt

You are a semantic parser. Generate a query for a restaurant database with the following signature:

```
CREATE TABLE restaurants (
    name TEXT,
    address TEXT,
    location TEXT,
    phone_number TEXT,
    opening hours TEXT,
    cuisines TEXT[],
    price ENUM ('cheap', 'moderate', 'expensive', 'luxury'),
    rating NUMERIC(2,1),
    num_review NUMBER,
);
```

Casual Assessment

- Show that LLM works surprising well
  - For domains such as restaurants because of its familiarity with the domain
- LLMs already know the basic SQL syntax
- Observed issue: Enumerated values

Issue: Enumerated Types

```
1 CREATE TABLE restaurants ( ...
2     price ENUM ('cheap', 'moderate', '
          expensive', 'luxury'),
3 ...)
```

- Small # values (e.g. 10): included in schema description
- LLMs know how to use these enumerated values

What are some good inexpensive chinese restaurants in Palo Alto?

↓

```
SELECT * FROM restaurants WHERE
'chinese' = ANY (cuisines) AND location = 'Palo Alto'
AND price = 'cheap' AND rating >= 4.0
ORDER BY num_reviews DESC LIMIT 3;
```

Large Enumerated Types

- Example: Yelp cuisines (text) has over 200 values, e.g.
  - "brazilian", "southern", "soul food", "modern european", "champagne bars", "latin american", "health & medical", "singaporean", …
- Problem: Not feasible to put all choices in the schema
- Database search expects an exact match

show me a café

↓

SELECT * FROM restaurants WHERE 'coffee' = ANY (cuisines)

Yelp: no coffee cuisine
Available cuisines:
coffee & tea,
cafe

Large Enumerated Types

show me a café

↓ *Semantic parser*

SELECT * FROM restaurants WHERE 'coffee' = ANY (cuisines)

↓ *Rewrite = operator*

SELECT * FROM restaurants WHERE
'coffee & tea' = ANY (cuisines) OR 'cafe' = ANY (cuisines)

Yelp: no coffee cuisine
Available cuisines:
coffee & tea,
cafe

Solution

- Change '=' operator for a text field into a value classification operation
- classify(x, *V*) finds the closest values of x in *V*,
  returns {} if no close values found
- Given a text field *f*, collect all possible values *V*

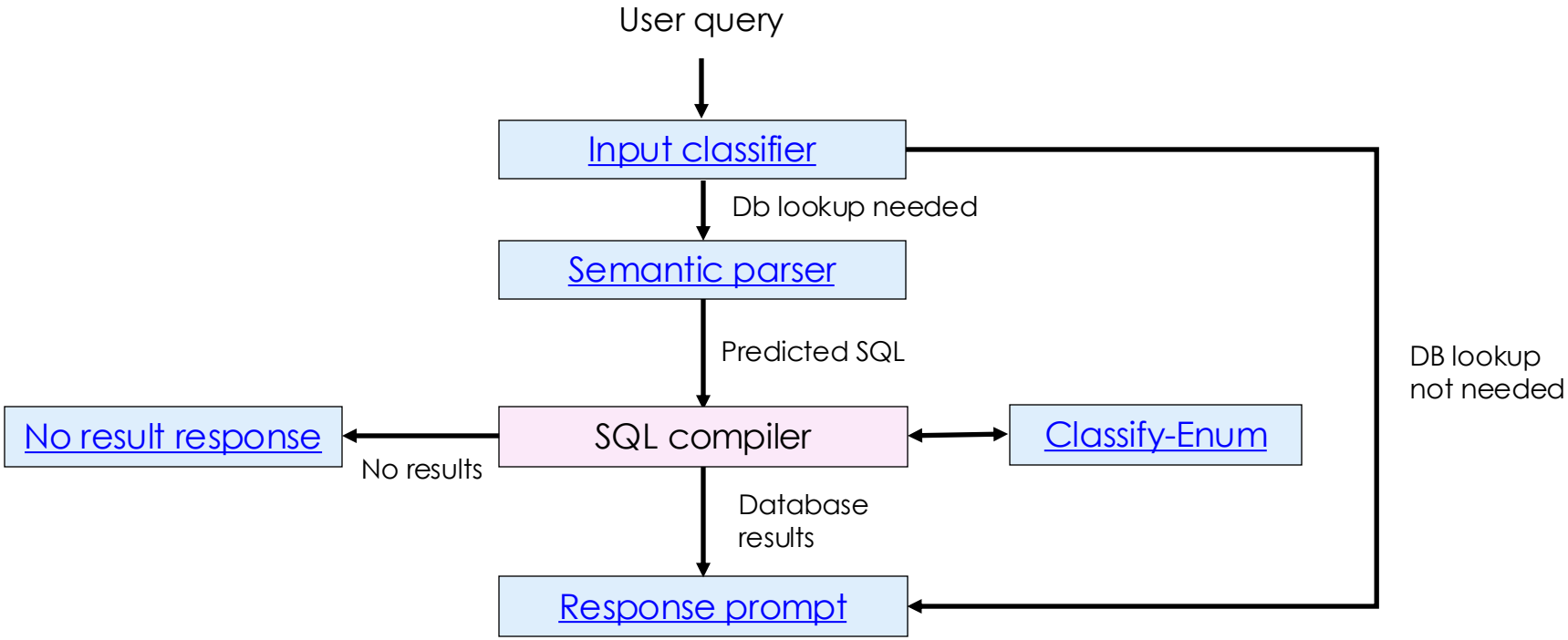
$$x = \text{ANY}\ (f) \Rightarrow c_1 = \text{ANY}\ (f)\ \text{OR}\ c_2 = \text{ANY}\ (f)\ \ldots,$$
$$\text{where } c_i \in \text{classify}\ (x, V)$$

Quiz: How to implement Classify?

The Full Agent Design

Some turns are not queries, let LLMs answer those

Agent Design



Click the links to see the prompts (written in jinja syntax)

Zero-Shot LLM-Based Database Agents are Easy to Build

Do They Perform Well?

Outline

- Our casual questions are formal queries
- A basic LLM-based agent that consults a database – Yelp
- **Evaluation issues of DB query agents**
- Motivation of hybrid corpora
- Techniques for hybrid corpora

Evaluation

- Measure query accuracy of our annotated data
  - Poor accuracy
- Quiz: what should we do next?

Example

What's some great Mexican food around Bernal Heights?

- Gold Target:
  - SELECT FROM restaurants
    WHERE "mexican" = ANY (cuisines) AND location = "bernal heights";
- Predicted Target:
  - SELECT FROM restaurants
    WHERE "mexican" = ANY (cuisines) AND location = "bernal heights"
    AND rating >= 4;

Quiz: Which is correct?

Evaluation example (SQL, BIRD dataset)

Name all cards with 2015 frame style ranking below 100 on EDHRec.

- Gold Target:
  - SELECT id FROM cards WHERE edhrecRank < 100 AND frameVersion = 2015
- Predicted Target:
  - SELECT **name** FROM cards WHERE edhrecRank < 100 AND frameVersion = 2015;

Quiz: which is correct?

https://bird-bench.github.io/

Summary on Simple Structured Data Queries

- Few-shot Chat-GPT parses SQL queries for Yelp
  - Restaurants: well-known domain to ChatGPT
  - Small table: 11 fields
    (incl. 2 Free-text, 1 small, 1 large ENUM)
  - Well-understood field names

Our Discovery Process

- Tried few-shot ChatGPT
  to make it conform to our convention
- Tried to fix the annotations in the benchmark

ChatGPT is better than annotations
by PhD students!

Quiz: What's Next?

How do you find a restaurant on Yelp?

Yelp: DB + Free text

Name TEXT	**Cuisine** TEXT[]	**Rating** NUM(2,1)	**...**	**Popular_dishes** FREE TEXT	**Reviews** FREE TEXT
Hummus Mediterranean Kitchen	mediterranean,halal, salad	4.0	...	Chicken Kebab Plate, Lamb Beef Gyro, Marinated Chicken Gyros, ...	| t l ; d r | This is your place if you're looking for a healthy and filling meal whether it's a quick pick-me-up or casual dining, ...
Penny Roma	italian, venues & event spaces	4.0	...	Cacio E Pepe, Agnolotti Dal Plin, Albacore Tartare, ...	My girlfriend was craving pasta on a Monday night. ... We were not expecting such an intimate and romantic dining experience. The restaurant was candle lit, modern, and perfect for a date night. ...

- Blue columns contain structured data
- Yellow columns are free text
- A lot of rows (thousands)

A lot of info is in free-text

DB + Free-Text Combo Example

I need a family-friendly restaurant
→ "family-friendly" in reviews

I need a family-friendly restaurant in Palo Alto
→ "family-friendly" in reviews; Palo Alto in DB

We Need Hybrid Data

	Structured	Free Text
Restaurants	Cuisines, opening hours, address, rating	Reviews, popular dishes
Products	Product ID, cost, ratings, sizes, physical dimensions, color	Descriptions, reviews
Courses	Class number, time offered, instructor, building, pre-requisites	Descriptions, reviews
Business processes	Receipts, statements	Regulations
Medical	Patient demographics Prescription	Diagnosis history

The Next Problem:

Hybrid Data QA

Combining
Knowledge Bases & Textual QA

Outline

- Our casual questions are formal queries
- A basic LLM-based agent that consults a database – Yelp
- Evaluation issues of DB query agents
- Motivation of hybrid corpora
- **Techniques for hybrid corpora**

Quiz

What is Your Approach to Handle Knowledge Bases & Text?

Overview of Hybrid Corpora Techniques

Discuss 4 major approaches, starting with the simplest

1. **Classify query to decide to retrieve from DB or Text**
2. Retrieve answers from DB and Text separately, choose answer
3. Convert DB to text; use Text Retrieval
4. Retrieve answers from DB and Text separately,
   use LLM to combine the answers
   - HybridQA: dataset requiring composition of DB & Text retrieval

1. Classify the Question: KB or Text

"What do others think?":
Task-Oriented Conversational Modeling with Subjective Knowledge

Chao Zhao [1] **Spandana Gella** [2] **Seokhwan Kim** [2] **Di Jin** [2]
Devamanyu Hazarika [2] **Alexandros Papangelis** [2] **Behnam Hedayatnia** [2]
Mahdi Namazifar [2] **Yang Liu** [2] **Dilek Hakkani-Tur** [2]
zhaochao@cs.unc.edu {sgella,seokhwk,djinamzn}@amazon.com
{dvhaz,papangea,behnam,mahdinam,yangliud,hakkanit}@amazon.com
[1] UNC Chapel Hill [2] Amazon, Alexa

Published 10/3/2023

https://browse.arxiv.org/pdf/2305.12091.pdf

SK-TOD Dataset

Subject-Knowledge Task-Oriented Dataset

"The first dataset, which contains subjective knowledge-seeking dialogue contexts and manually annotated responses grounded in subjective knowledge sources."

- MultiWOZ restaurant and hotel (in Cambridge area)
  - User goal: search and book restaurants and hotels
  - Structure: 13 slots (easier than SQL)
  - Text: 143 entities and 1,430 reviews (8,013 sentences)

Largest TOD Dataset with Subjective Knowledge

	Size	Manual	Dial	TOD	Query	Aspect	Senti	Mul-Knwl	Senti-%
Semeval/MAMS (2016; 2019)	5K/22K	✓	✗	n/a	✗	✓	✓	✗	n/a
Space (2021)	1K	✓	✗	n/a	✗	✓	✓	✓	✗
Yelp/Amazon (2019; 2020)	200/180	✓	✗	n/a	✗	✗	✓	✓	✗
Justify-Rec (2019)	1.3M	✗	✗	n/a	✗	✓	✗	✓	✗
AmazonQA (2016)	309K	✗	✗	n/a	✓	✗	✗	✗	n/a
SubjQA (2020)	10K	✗	✗	n/a	✓	✓	✓	✗	n/a
Holl-E (2018)	9K	✓	✓	✗	✗	✗	✗	✓	✗
Foursquare (2018)	1M	✗	✓	✗	✗	✗	✗	✓	n/a
SK-TOD (Ours)	20K	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between SK-TOD and other benchmarks based on the subjective content. We consider if the dataset is manually annotated, dialogue-based, task-oriented, and query-focused. We also list if it considers aspect and sentiment, multiple knowledge snippets (Mul-Knwl), and the proportion of two-sided sentiments (Senti-%).

Binary Classifier (IR vs Semantic Parsing)

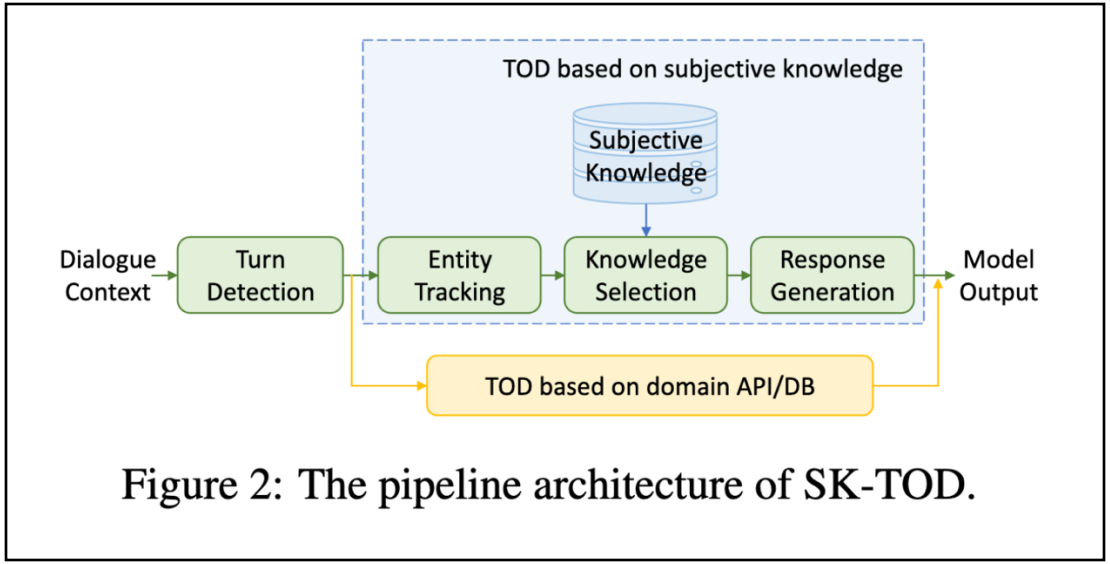


Figure 2: The pipeline architecture of SK-TOD.

Turn detection: binary classification

INFORMATION RETRIEVAL (IR)

Entity Tracking
Fuzzy n-gram matching between dialogue &entities. (A few entities in MultiWOZ)

Knowledge Selection
Calculate relevant knowledge score between dialogue context & knowledge snippets

Response Generation
Use pretrained GPT-2/Bart

2. Choose Later: Stanford Chirpy Cardinal

Neural Generation Meets Real People: Building a Social, Informative Open-Domain Dialogue Agent

Ethan A. Chi*, Ashwin Paranjape*, Abigail See*, Caleb Chiam*, Trenton Chang, Kathleen Kenealy, Swee Kiat Lim, Amelia Hardy, Chetanya Rastogi, Haojun Li, Alexander Iyabor, Yutong He, Hari Sowrirajan, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, Jillian Tang, Avanika Narayan, Giovanni Campagna, and Christopher D. Manning

- Alexa Social Bot Prize Runner up, 2021
- Different modules to handle different kinds of questions
- Run generators in parallel
- Choose at the end

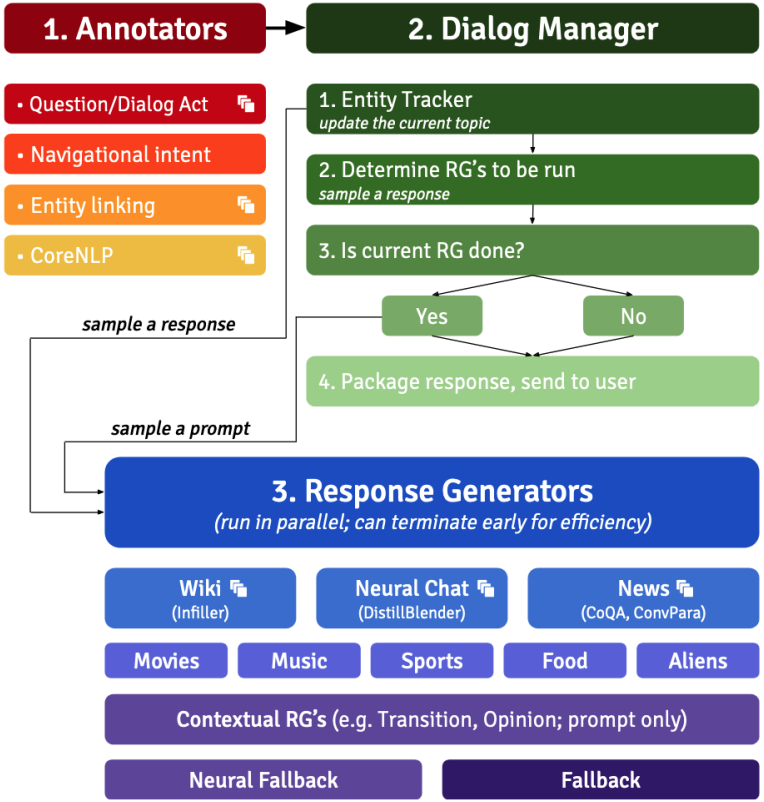


Figure 4: Overall system design.

Proceedings of the SIGdial 2022 Conference, pages 376–395

Quiz

What is the Weakness of Choosing Just the KB or Text?

What if We Need Both to Answer the Questions?

Hey! Can you recommend me an Italian restaurant with a romantic atmosphere?

Name TEXT	Cuisine TEXT[]	Rating NUM(2,1)	...	Popular_dishes FREE TEXT	Reviews FREE TEXT
Hummus Mediterranean Kitchen	mediterranean,halal, salad	4.0	...	Chicken Kebab Plate, Lamb Beef Gyro, Marinated Chicken Gyros, ...	| t l ; d r | This is your place if you're looking for a healthy and filling meal whether it's a quick pick-me-up or casual dining, ...
Penny Roma	italian, venues & event spaces	4.0	...	Cacio E Pepe, Agnolotti Dal Plin, Albacore Tartare, ...	My girlfriend was craving pasta on a Monday night. ... We were not expecting such an intimate and romantic dining experience. The restaurant was candle lit, modern, and perfect for a date night. ...

- Is it common? Yes!
  - Real-user queries about restaurants: 55/100 need a combo
- Separate modules cannot answer these questions

Overview of Hybrid Corpora Techniques

Discuss 4 major approaches, starting with the simplest

1. Classify query to decide to retrieve from DB or Text
2. Retrieve answers from DB and Text separately, choose answer
3. **Convert DB to text; use Text Retrieval**
4. Retrieve answers from DB and Text separately,
   use LLM to combine the answers
   - HybridQA: dataset requiring composition of DB & Text retrieval

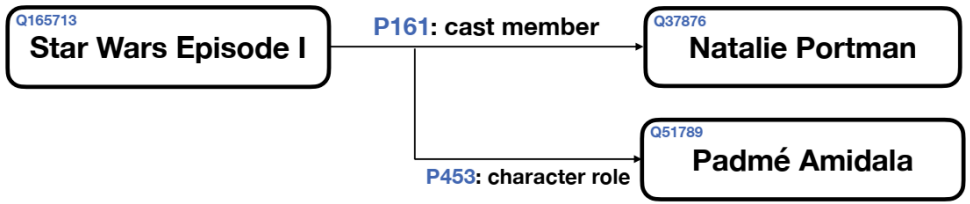
3. Converting All to Text

- Unified Knowledge Question Answering (UniK-QA)
- Converts everything to text
  - Text
  - Semi-structured data e.g. Wikipedia lists, tables, info boxes
  - Knowledge base
- Then uses the usual pipeline designed for textual QA

Oguz et al. "UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering." arXiv preprint arXiv:2012.14610 (2021).

Wikipedia Tables / Wikidata

- Wikipedia tables are linearized simply by concatenating cell values with special tokens to separate rows
- Wikidata:



Oguz et al. "UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering." Findings of NAACL (2022).

Quiz

What is the Weakness of Converting Databases to Text?

Unified Knowledge Question Answering (UniK-QA)

- Retrieve and read QA pipeline with SOTA components
  - Neural retriever; T5 reader
- Exact match scores:

Knowledge Source	Natural Questions	TriviaQA	WebQuestions
Wikipedia Text	49.0	64.0	50.6
Wikipedia Tables	36.0	34.5	41.0
Wikidata	27.9	35.4	55.6
Text + Tables	**54.1**	**65.1**	50.2
Text + Tables + Wikidata	54.0	64.1	**57.8**

Quiz: What do you observe?

Testing with Yelp

linearize

Name, Hummus Mediterranean Kitchen, cuisine, mediterranean, ..., reviews, | t l ; d r | This is your place

Name, Penny Roma, cuisine, italian, ... , reviews, My girlfriend was craving pasta on a Monday night ...

Embedding model

0.076	0.865	...	-0.732
-0.746	-0.301	...	0.495

Cos Sim

0.376	-0.265	...	-0.922

Linearization (Vector DB)

Hey! Can you recommend me an **Italian** restaurant with a **romantic atmosphere**?

I found Hummus Mediterranean Kitchen. It is a 4 star **Mediterranean** restaurant in San Francisco with **a clean and welcoming atmosphere**.

wrong cuisine; no mention of romantic atmosphere in review

Problems with Linearization

- How to combine different clauses together?
  (Algebraic operators handle them naturally)
  - e.g. multiple clauses: "Are there any restaurants that have **a great view of the Golden Gate bridge** in **San Francisco**?"

- Hard to do column computations
  (rank, max, min, average, count) on flattened data
  - e.g. "what is the **top-rated** Chinese restaurant?"

Overview of Hybrid Corpora Techniques

Discuss 4 major approaches, starting with the simplest

1. Classify query to decide to retrieve from DB or Text
2. Retrieve answers from DB and Text separately, choose answer
3. Convert DB to text; use Text Retrieval
4. **Retrieve answers from DB and Text separately, use LLM to combine the answers**
   - HybridQA: dataset requiring composition of DB & Text retrieval

HybridQA Dataset

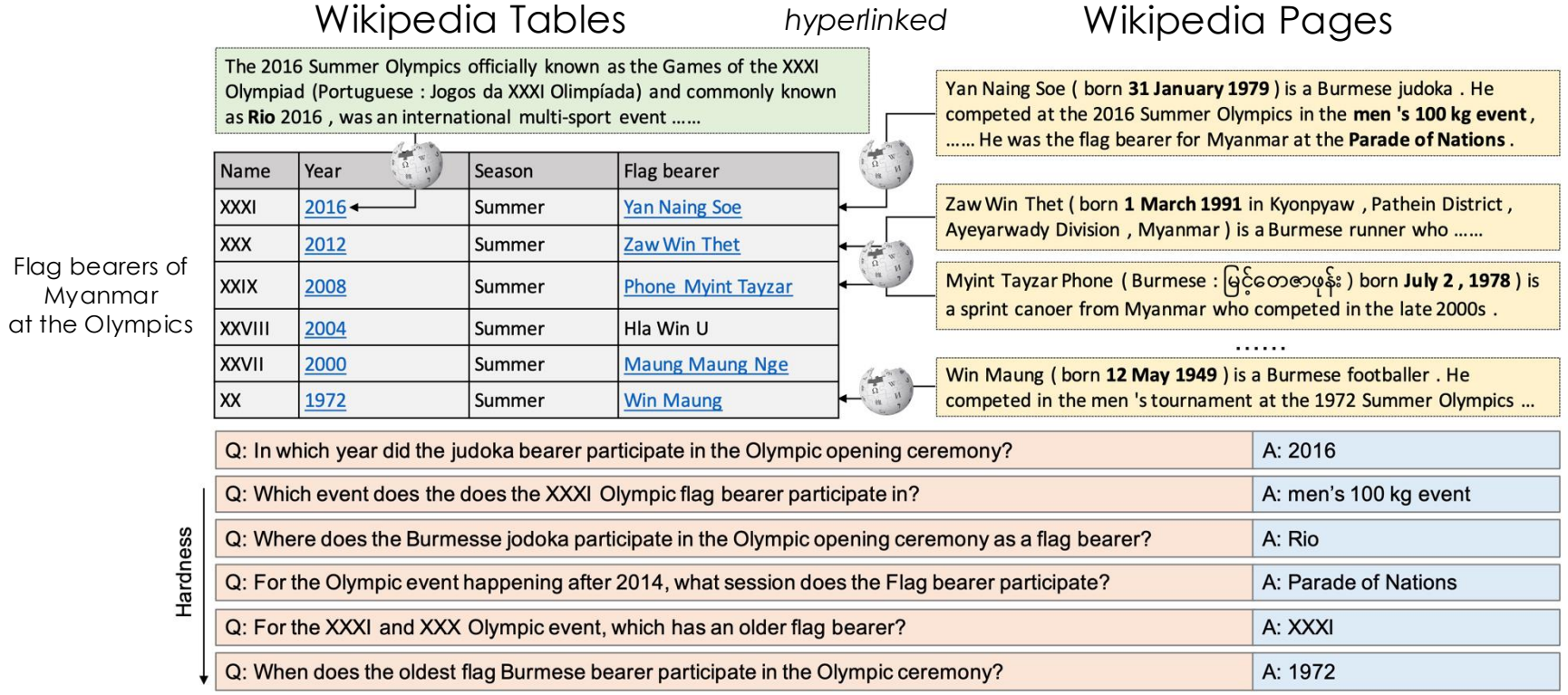
HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data

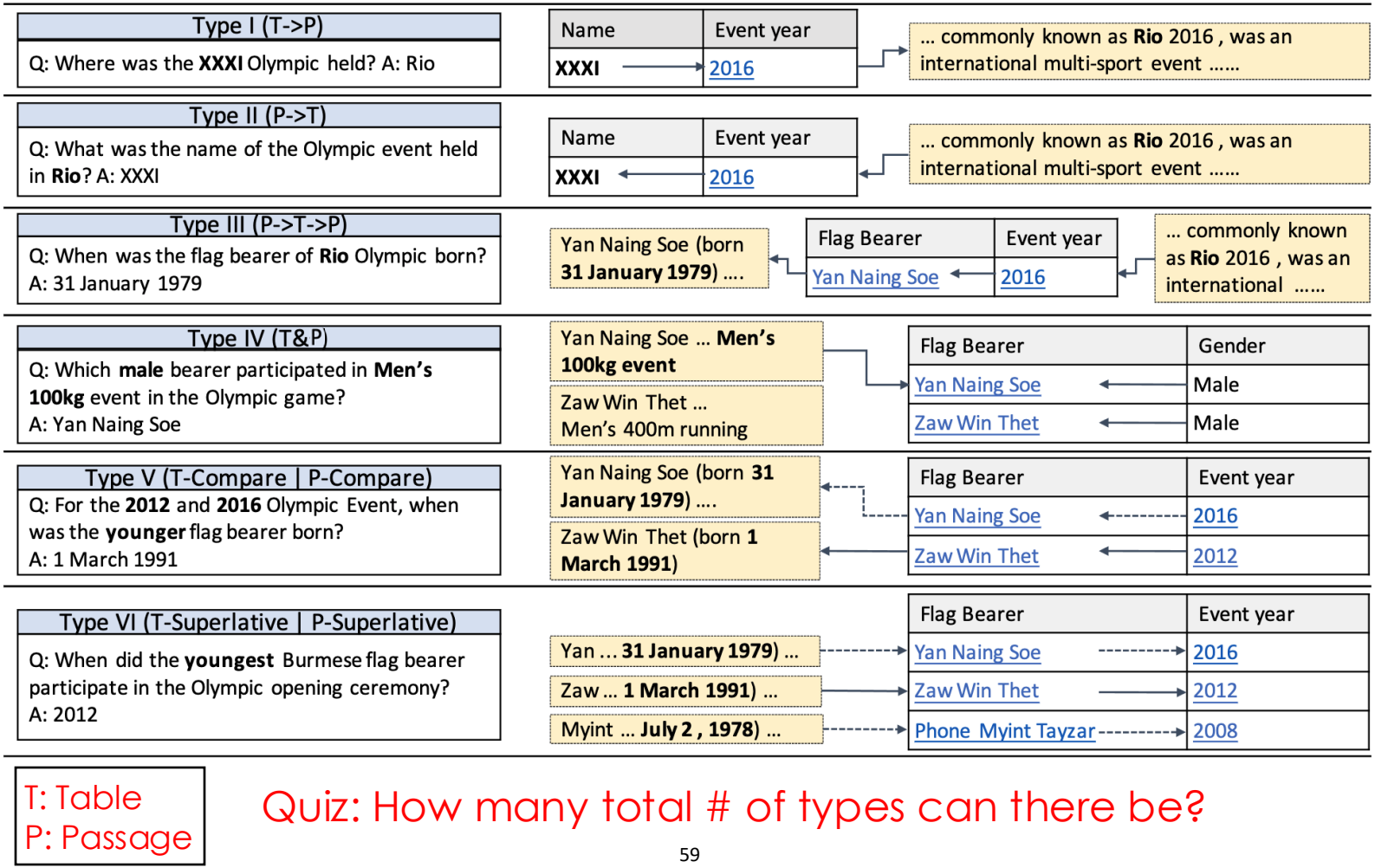
Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, William Wang
University of California, Santa Barbara, CA, USA
{wenhuchen, hwzha, xwhan}@cs.ucsb.edu,
{zhiyuchen, hongwang600, william}@cs.ucsb.edu

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1026–1036
November 16 - 20, 2020.

Earlier than SK-TOD. Note: this is not conversational

HybridQA Dataset




Type I (T->P)
Q: Where was the XXXI Olympic held? A: Rio
Name
Event year
XXXI
2016
... commonly known as Rio 2016 , was an international multi-sport event .....
Type II (P->T)
Q: What was the name of the Olympic event held in Rio? A: XXXI
Name
Event year
XXXI
2016
... commonly known as Rio 2016 , was an international multi-sport event .....
Type III (P->T->P)
Q: When was the flag bearer of Rio Olympic born? A: 31 January 1979
Yan Naing Soe (born 31 January 1979) ....
Flag Bearer
Event year
Yan Naing Soe
2016
... commonly known as Rio 2016 , was an international .....
Type IV (T&P)
Q: Which male bearer participated in Men's 100kg event in the Olympic game? A: Yan Naing Soe
Yan Naing Soe ... Men's 100kg event
Zaw Win Thet ... Men's 400m running
Flag Bearer
Gender
Yan Naing Soe
Male
Zaw Win Thet
Male
Type V (T-Compare | P-Compare)
Q: For the 2012 and 2016 Olympic Event, when was the younger flag bearer born? A: 1 March 1991
Yan Naing Soe (born 31 January 1979) ....
Zaw Win Thet (born 1 March 1991)
Flag Bearer
Event year
Yan Naing Soe
2016
Zaw Win Thet
2012
Type VI (T-Superlative | P-Superlative)
Q: When did the youngest Burmese flag bearer participate in the Olympic opening ceremony? A: 2012
Yan ... 31 January 1979) ...
Zaw ... 1 March 1991) ...
Myint ... July 2 , 1978) ...
Flag Bearer
Event year
Yan Naing Soe
2016
Zaw Win Thet
2012
Phone Myint Tayzar
2008
T: Table
P: Passage
Quiz: How many total # of types can there be?

HybridQA

A Good Dataset that Illustrates the Naturalness of Hybrid Data Queries

Need to Support Full Compositionality Though

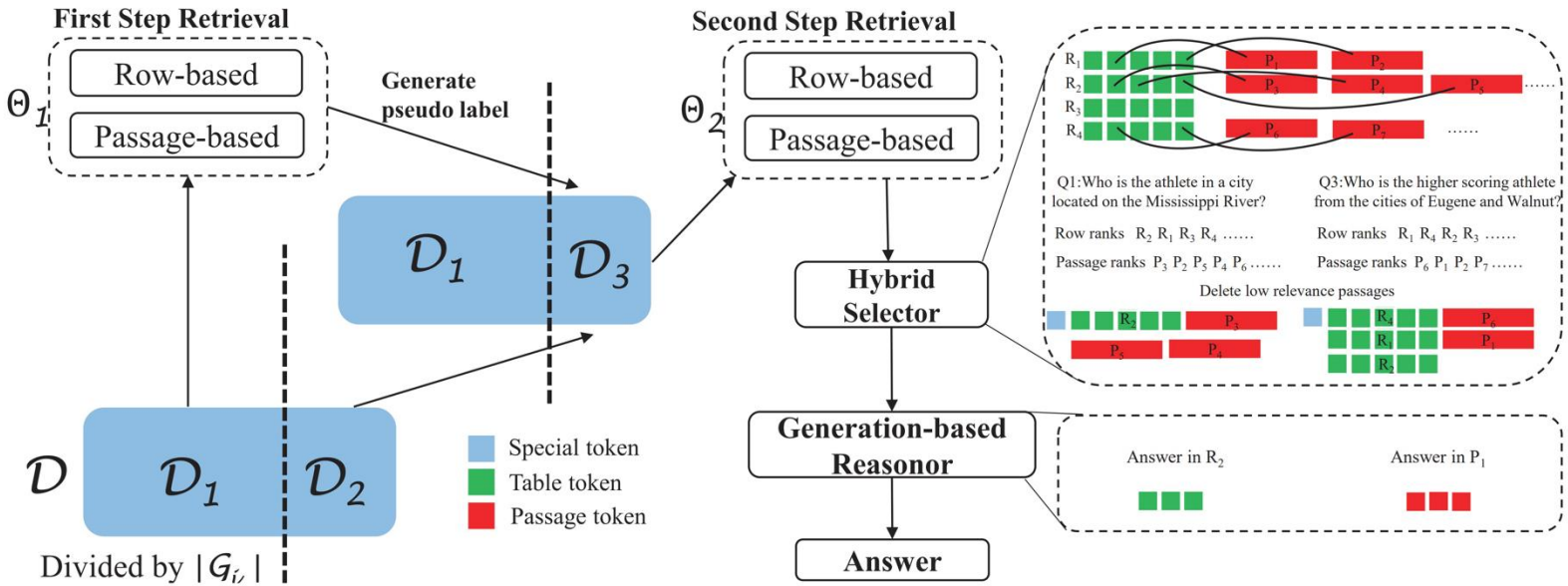
SOTA Model in 2023

S^{3}HQA: A Three-Stage Approach for Multi-hop Text-Table Hybrid Question Answering

Fangyu Lei[1,2], **Xiang Li**[1,2], **Yifan Wei**[1,2],
Shizhu He[1,2], **Yiming Huang**[1,2], **Jun Zhao**[1,2], **Kang Liu**[1,2]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{leifangyu2022, lixiang2022, weiyifan2021, huangyiming2023}@ia.ac.cn
{shizhu.he, jzhao, kliu}@nlpr.ia.ac.cn

https://browse.arxiv.org/pdf/2305.11725.pdf

Hybrid Retrievers



Only uses LLM in the "generation-based reasoner" step

https://github.com/wenhuchen/HybridQA#recent-papers
SOTA model: https://arxiv.org/abs/2305.11725

3 Stages

1. Retrieve information from Tables and Free Text
   - Extra complexity to recover from annotation errors
2. Hybrid selector
   - Sort and filter based on relevance
3. LLM-Based Reasoner
   - Combines the filtered information from DV and Free Text
     - Question types: Count, compare
     - Use lexical analysis to identify the question type
     - Chain-of-thought prompting

SOTA Result

	Table				Passage				Total			
	Dev		Test		Dev		Test		Dev		Test	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Unsupervised-QG (Pan et al., 2021)	-	-	-	-	-	-	-	-	25.7	30.5	-	-
HYBRIDER (Chen et al., 2020b)	54.3	61.4	56.2	63.3	39.1	45.7	37.5	44.4	44.0	50.7	43.8	50.6
DocHopper (Sun et al., 2021)	-	-	-	-	-	-	-	-	47.7	55.0	46.3	53.3
MuGER[2] (Wang et al., 2022b)	60.9	69.2	58.7	66.6	56.9	68.9	57.1	68.6	57.1	67.3	56.3	66.2
POINTR (Eisenschlos et al., 2021)	68.6	74.2	66.9	72.3	62.8	71.9	62.8	71.9	63.4	71.0	62.8	70.2
DEHG (Feng et al., 2022)	-	-	-	-	-	-	-	-	65.2	**76.3**	63.9	75.5
MITQA (Kumar et al., 2021)	68.1	73.3	68.5	74.4	66.7	75.6	64.3	73.3	65.5	72.7	64.3	71.9
MAFiD (Lee et al., 2023)	69.4	75.2	68.5	74.9	66.5	75.5	65.7	75.3	66.2	74.1	65.4	73.6
S^{3}HQA	**70.3**	**75.3**	**70.6**	**76.3**	**69.9**	**78.2**	**68.7**	**77.8**	**68.4**	75.3	**67.9**	**75.5**
Human	-	-	-	-	-	-	-	-	-	-	88.2	93.5

Table 1: Performance of our model and related work on the HybridQA dataset.

Quiz: What do you observe?

Problem

- Data are retrieved in just one hop
  - Cannot solve cascading multi-hop questions, where the data to retrieve depends on the 1st hop answer
- Imprecision with the lexical analysis
- Lacks completeness and compositionality
  - Just one “count” or one “compare” operation

Conclusion

- **LLMs can handle simple database queries**
- **Many questions require composition of information retrieved from DB and Text**
- **Structure OR Text: is inadequate**
  1. Binary classifier up front (SK-TOD, 2023)
  2. Pick afterwards (Stanford Chirpy Cardinal, 2021)
- **Different approaches to combine structures and free-text**
  3. Structures → Text: Linearization (one hop)
  4. Hybrid: Retrieve from both and combine (one hop each)

We need better techniques for complex db queries and hybrid QA