From Schema to Q&A Agents

Silei Xu

CS294S September 17, 2020

Joint work with Giovanni Campagna, Sina Semnani, Jian Li, and Monica S. Lam

Commercial Assistants

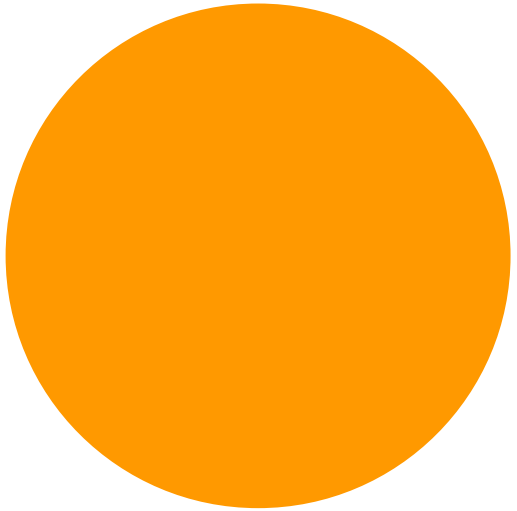
Alexa: Handcode 1 question at a time

get me an upscale restaurants

What are the restaurants around here?

What is the best restaurant?

search for Chinese restaurants



100K Alexa skills
Sep 2019

Commercial Assistants

Alexa: Handcode 1 question at a time

get me an upscale restaurants

What are the restaurants around here?

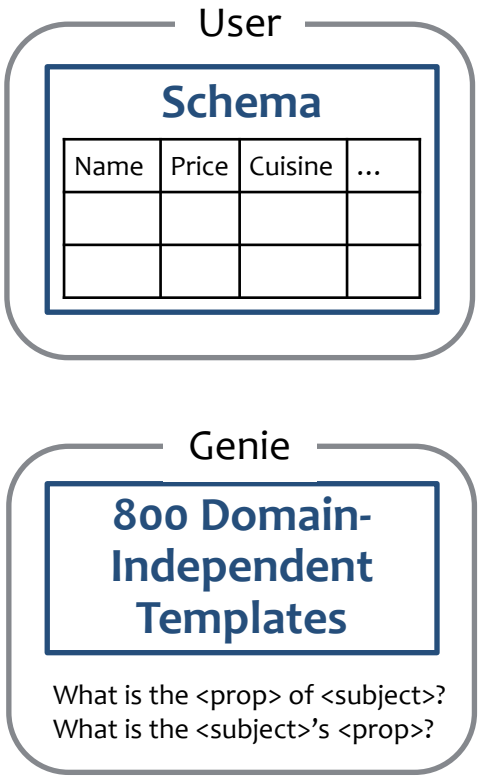
What is the best restaurant?

search for Chinese restaurants

100K Alexa skills
Sep 2019

1.8 billion websites

Genie: Synthesize Question/Code from a Schema



get me an upscale restaurants

What are the restaurants around here?

What is the best restaurant?

search for Chinese restaurants

What is the best restaurant within 10 miles?

Find restaurants that serve Chinese or Japanese food

What is the best non-Chinese restaurant near here?

Show me a cheap restaurant with 5-star review.

Are there any restaurant with at least 4.5 stars?

What is the phone number of Wendy's?

I'm looking for an Italian fine dining restaurant.

Give me the best Italian restaurant.

Find me the best restaurant with 500 or more reviews

Show me some restaurant with less than 10 reviews

Outline

- Representing Questions in ThingTalk
- High-quality Low-cost Training Data Generation by Genie
- Apply Genie on the Web
- AutoQA: Automate Everything!

ThingTalk for Questions

ThingTalk for QA

$$table\ [, filter]^{?}$$

```
now => @QA.restaurant(), geo == new Location(“Stanford”) => notify
```

Show me restaurants in Stanford

ThingTalk for QA

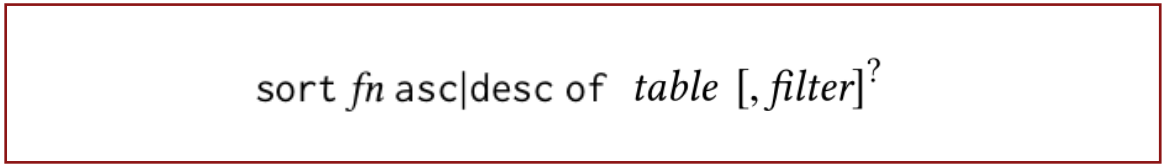
$$table\ [, filter]^?$$

```
now =>  @QA.restaurant(), geo == new Location("Stanford")
                  && servesCuisine =~ "Chinese"  => notify
```

Show me Chinese restaurants in Stanford

ThingTalk for QA

sort *fn* asc|desc of *table* [, *filter*]?

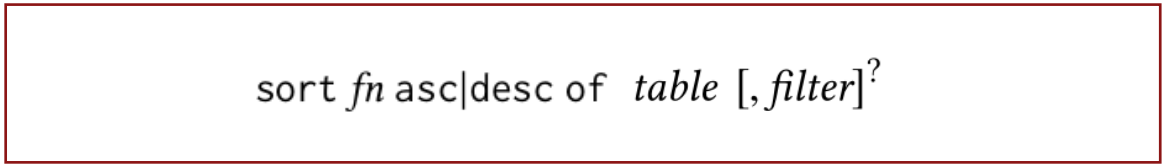


```
now =>  @QA.restaurant(), geo == new Location("Stanford")
                  && servesCuisine =~ "Chinese"  => notify
```

Show me Chinese restaurants in Stanford

ThingTalk for QA

sort *fn* asc|desc of *table* [, *filter*]?



```
now =>  sort aggregateRating.ratingValue desc of (
        @QA.restaurant(), geo == new Location("Stanford")
                    && servesCuisine =~ "Chinese" )  => notify
```

Show me top-rated Chinese restaurants in Stanford

ThingTalk for QA

sort *fn* asc|desc of *table* [, *filter*]? [join *table* [, *filter*]?]*

```
now => sort aggregateRating.ratingValue desc of (
       @QA.restaurant(), geo == new Location("Stanford")
                  && servesCuisine =~ "Chinese" ) => notify
```

Show me top-rated Chinese restaurants in Stanford

ThingTalk for QA

sort *fn* asc|desc of *table* [, *filter*]? [join *table* [, *filter*]?]*

```
now => sort aggregateRating.ratingValue desc of (
       @QA.restaurant(), geo == new Location("Stanford")
                  && servesCuisine =~ "Chinese" )
              join ( @QA.review(), in_array(id, review)
                  && author = "bob" ) => notify
```

Show me top-rated Chinese restaurants in Stanford reviewed by Bob

ThingTalk for QA

sort *fn* asc|desc of *table* [, *filter*]? [join *table* [, *filter*]?]*

[*fn*+ of]? *table* [, *filter*]?

aggregate min|max|sum|avg|count *fn* of *table* [, *filter*]?

...

Natural Language Programming

What is the top-rated Chinese restaurant in Palo Alto?

```
now =>
sort aggregateRating.ratingValue desc of
( @QA.restaurant(),
geo == new MakeLocation(“Stanford”)
&& servesCuisine =~ “Chinese” )
=> notify;
```

High-quality Low-cost Training Data Generation by Genie

Synthesizing Training Data with Templates

- Templates: Map natural language to database operators

DB Operator	Natural Language	Template	ThingTalk
Selection	restaurants with rating equal to 4 restaurants with rating greater than 4 restaurants with rating less than 4 …	<table> with <property> equal to <value> <table> with <property> greater than <value> <table> with <property> less than <value>	table, property == value table, property >= value table, property <= value …
Projection	rating of restaurant	<property> of <table>	[property] of table
Aggregation	the number of restaurants	the number of <table>	aggregate count of table
…	…	…	…

- Generate natural language and ThingTalk pairs

Discussion

Why this won't work?

Variety in Natural Language

- Fact: "Dr. Smith is Ann's doctor"

Relation	Unknown: Ann	Unknown: Dr. Smith	Part-of-Speech
Doctor	Who has Dr. Smith as a doctor?	Who does Ann have as a doctor?	Noun (has …)
	Who is Dr. Smith a doctor of ?	Who is a doctor of Ann?	Noun (is …)
	Whom does Dr. Smith treat?	Who treats Ann?	Active verb
	Who is treated by Dr. Smith?	By whom is Ann treated?	Passive verb
Patient	Who does Dr. Smith have as a patient?	Who has Ann as a patient?	Noun (has …)
	Who is a patient of Dr. Smith?	Who is Ann a patient of?	Noun (is …)
	Who consults with Dr. Smith?	With whom does Ann consult?	Active verb
	By whom is Dr. Smith consulted?	Who is consulted by Ann?	Passive verb

Previous work: train with paraphrase data based on synthesized sentences

Wang at al. "Building a semantic parser overnight." *ACL* 2015.

Natural Language Annotations

- POS-based annotation for each property

POS	People: worksFor	Restaurants: servesCuisine
Active verb	works for <value>	serves <value> cuisine, offer <value> food
Passive verb	employed by <value>	-
Is-a Noun	an employer of <value>	-
has-a Noun	employee <value>	<value> food, <value> cuisine
Adjective	-	<value>
Prepositional	from <value>	-

Domain-Independent Templates

- A comprehensive set of 800 templates that captures:
  - Different parts of speech

```
now => @QA.restaurant(),
servesCuisine =~ “Chinese” => notify;
```

Show me <table> that <verb>.
Show me <table> with <noun>.
Show me <adjective> <table>.

→

Show me restaurants that serve Chinese cuisine.
Show me restaurants with Chinese food.
Show me Chinese restaurants.

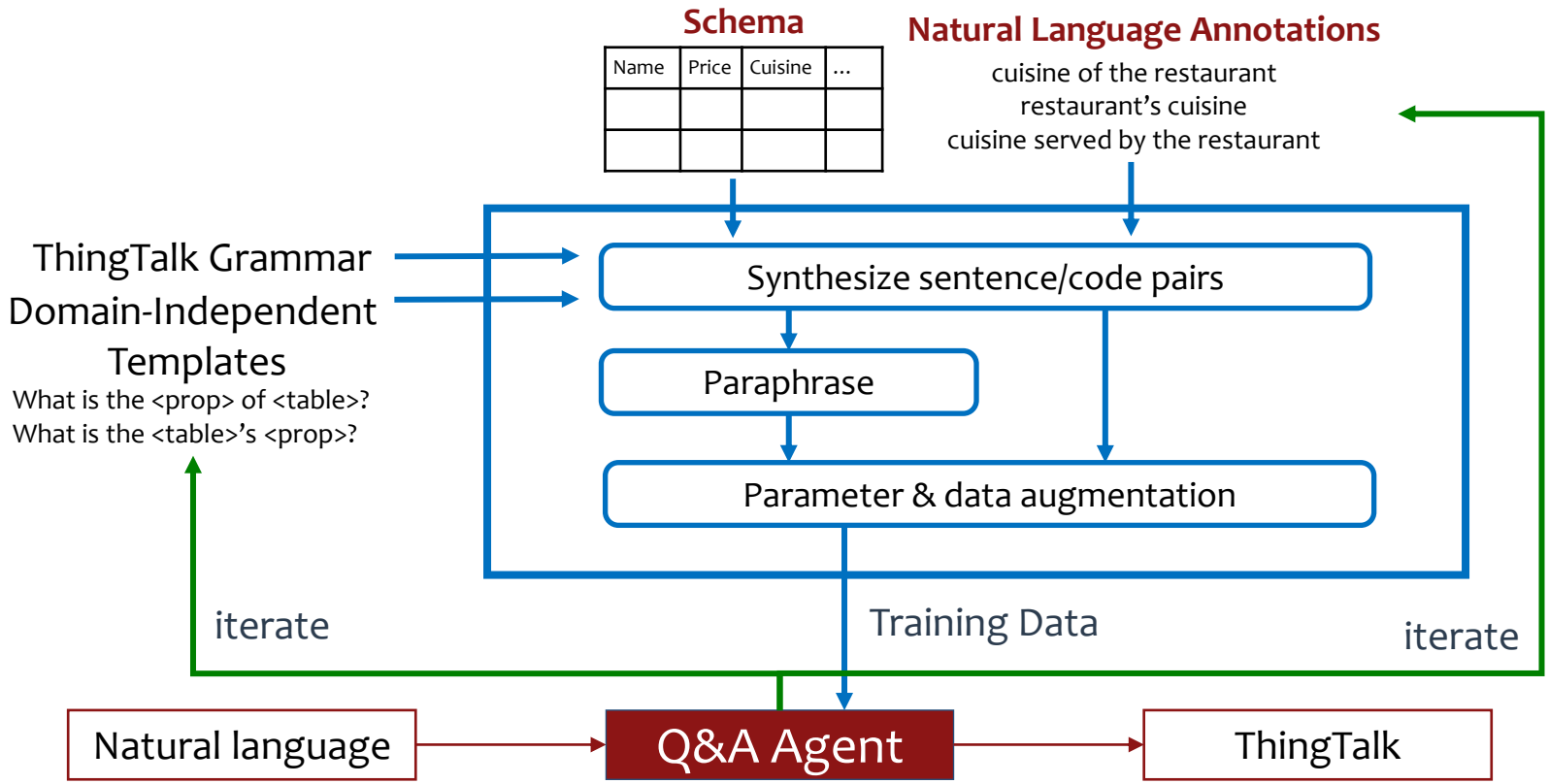
  - Connectives

Show me restaurant that serve Chinese cuisine and with more than 100 reviews.
Show me restaurant with Chinese food and at least 100 reviews.
Show me Chinese restaurant that have more than 100 reviews

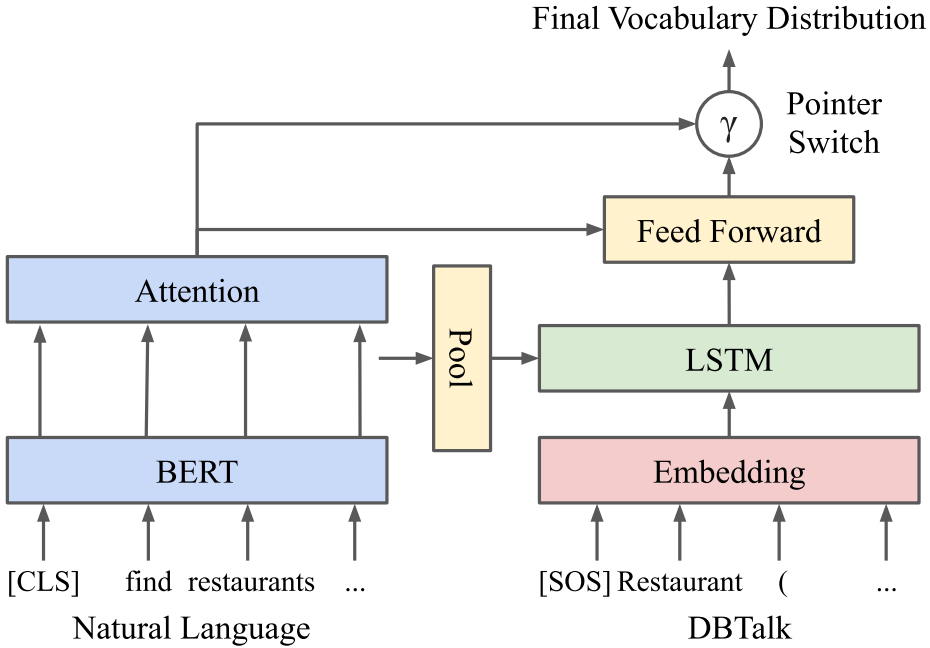
  - Different types

when does the restaurant open?
who owns the restaurant?
how far is the restaurant?

Genie Pipeline



BERT-LSTM Neural Model



Applying Genie to the Web

How do we scale to the web?

- The web has a schema: **Schema.org**
  - Structure data to mark up web pages
  - Mainly used by search engines
  - It covers many domains, including restaurants, hotels, people, recipes, products, news …

40% of the websites use it!

```
<script type="application/ld+json">
{
  @type: "restaurant",
  name: "The French Laundry",
  servesCuisine: “French",
  aggregateRating: {
    @type: "AggregateRating",
    reviewCount: 2527,
    ratingValue: 4.5
  }
  ...
}
```

Schema.org markup on Yelp

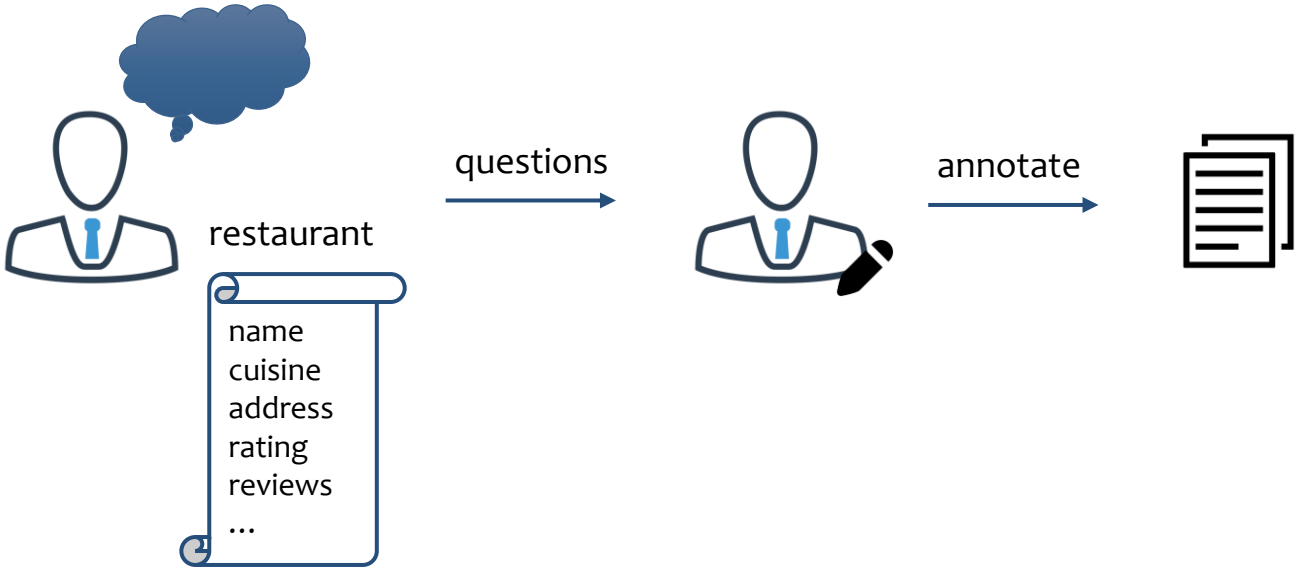
Experiment domains

- 5 domains: restaurant, people, movie, book, and music

	Restaurant	People	Movie	Book	Music	Average
Website	Yelp	LinkedIn	IMDb	Goodreads	Last.fm	-
# of properties	25	13	16	15	19	17.6
# of annotations	122	95	111	96	103	105.4
Synthesized	270,081	270,081	270,081	270,081	270,081	270,081
Paraphrase	6,419	7,108	3,774	3,941	3,626	4,973.6
Total (augmented)	508,101	614,841	405,241	410,141	425,041	472,673

Evaluation Data Collection

- Evaluating on paraphrase data is misleading!
- Evaluate on a challenging realistic dataset



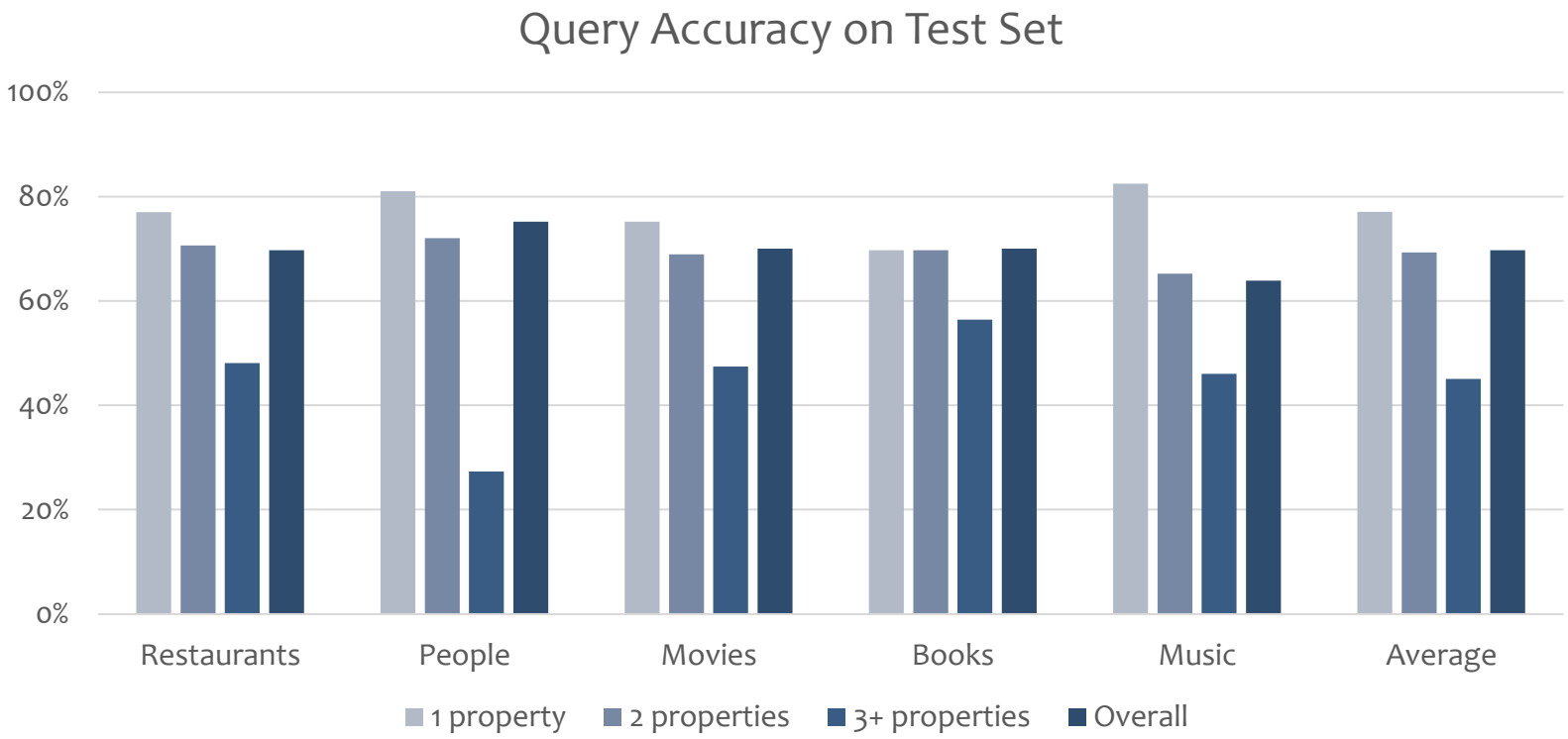
Evaluation Data Collection

- Evaluating on paraphrase data is misleading!
- Evaluate on a challenging realistic dataset

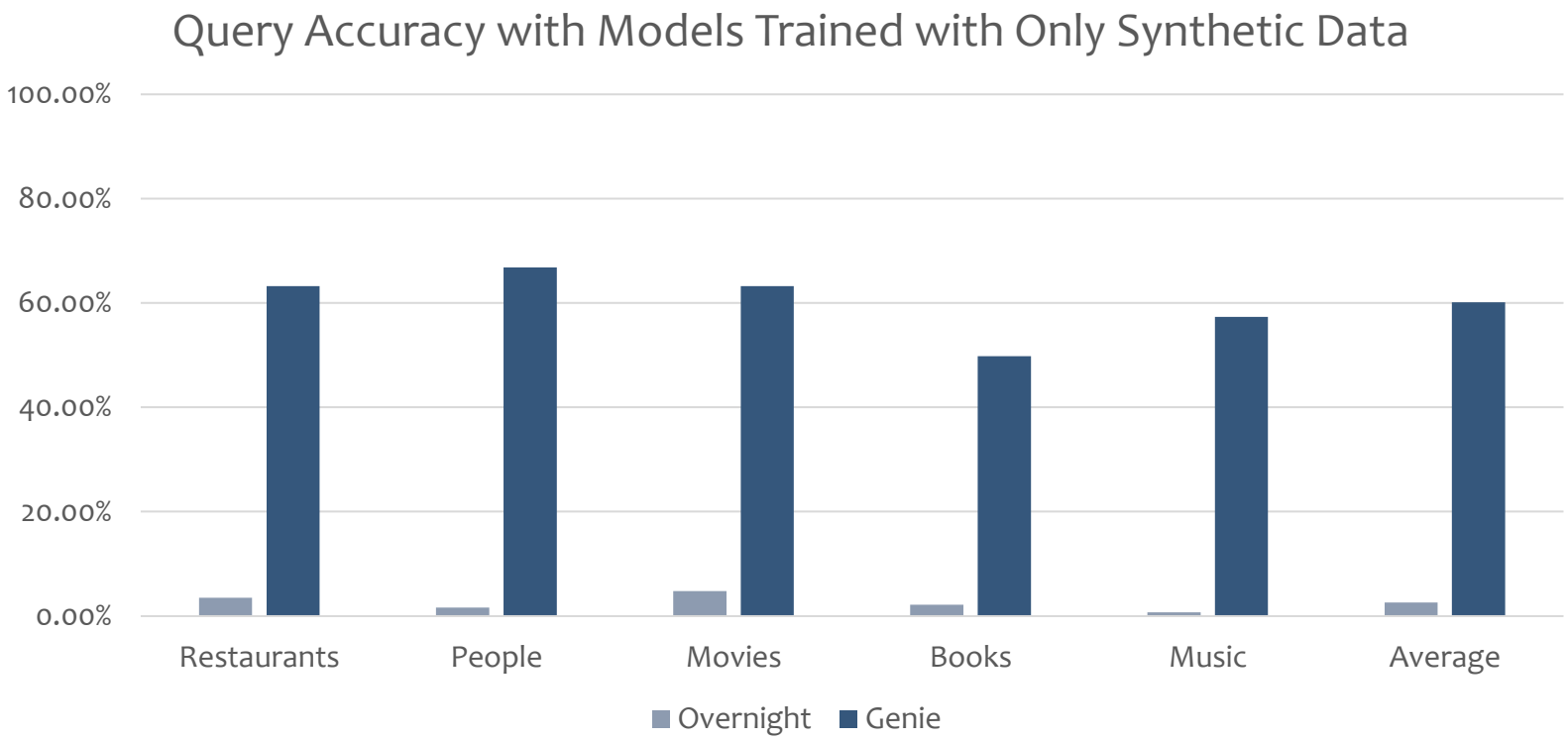
		Restaurant	People	Movie	Book	Music	Average
Dev	1 property	221	127	140	107	62	131.4
	2 properties	219	346	226	222	182	239
	3+ properties	88	26	23	33	82	50.4
	Total	**528**	**499**	**389**	**362**	**326**	**420.8**
Test	1 property	200	232	130	114	44	144
	2 properties	245	257	264	241	181	237.6
	3+ properties	79	11	19	55	63	45.4
	Total	**524**	**500**	**413**	**410**	**288**	**427**

- Over 2/3 of questions have 2+ properties
- Contains unseen values

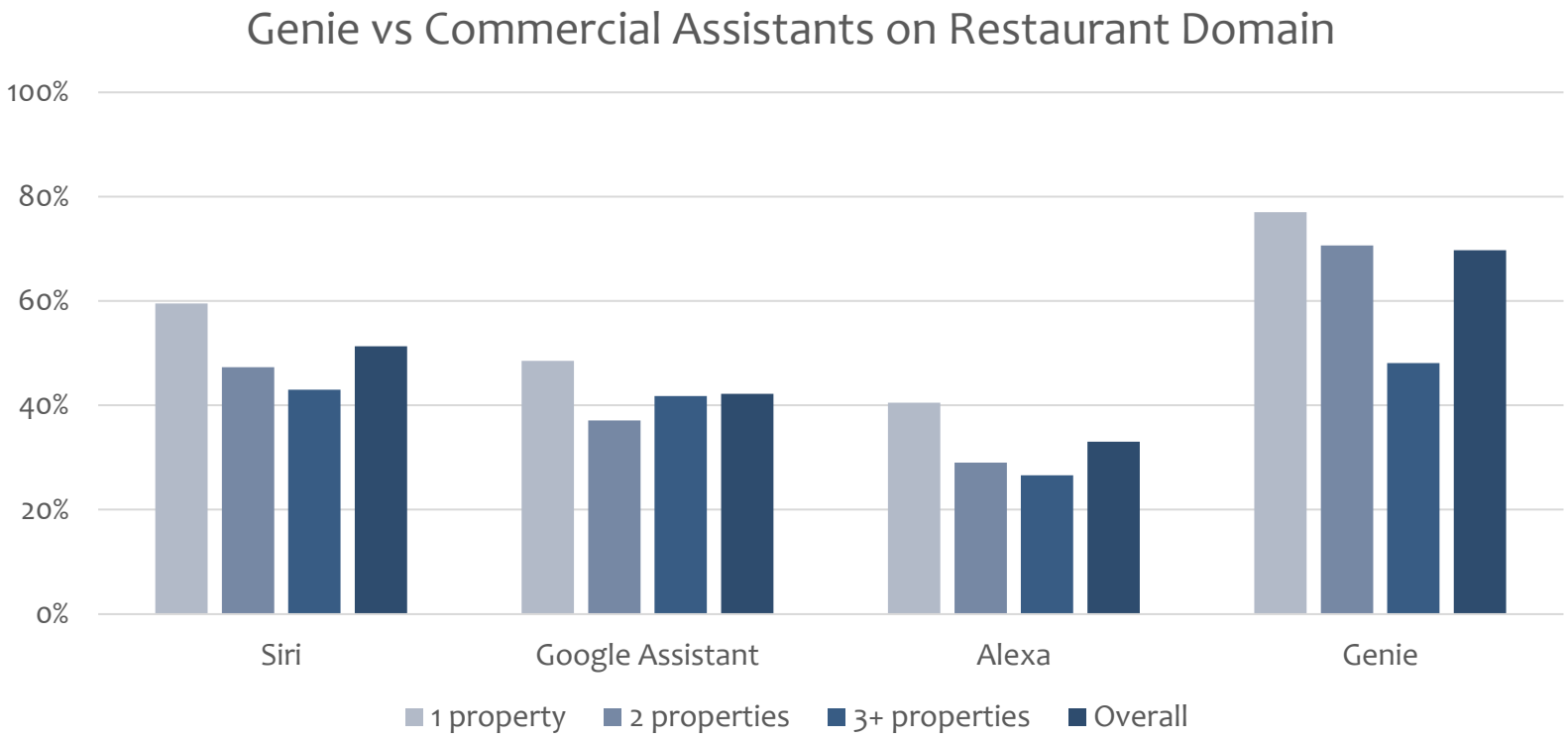
Experimental Results



Experimental Results (Synthetic Only)



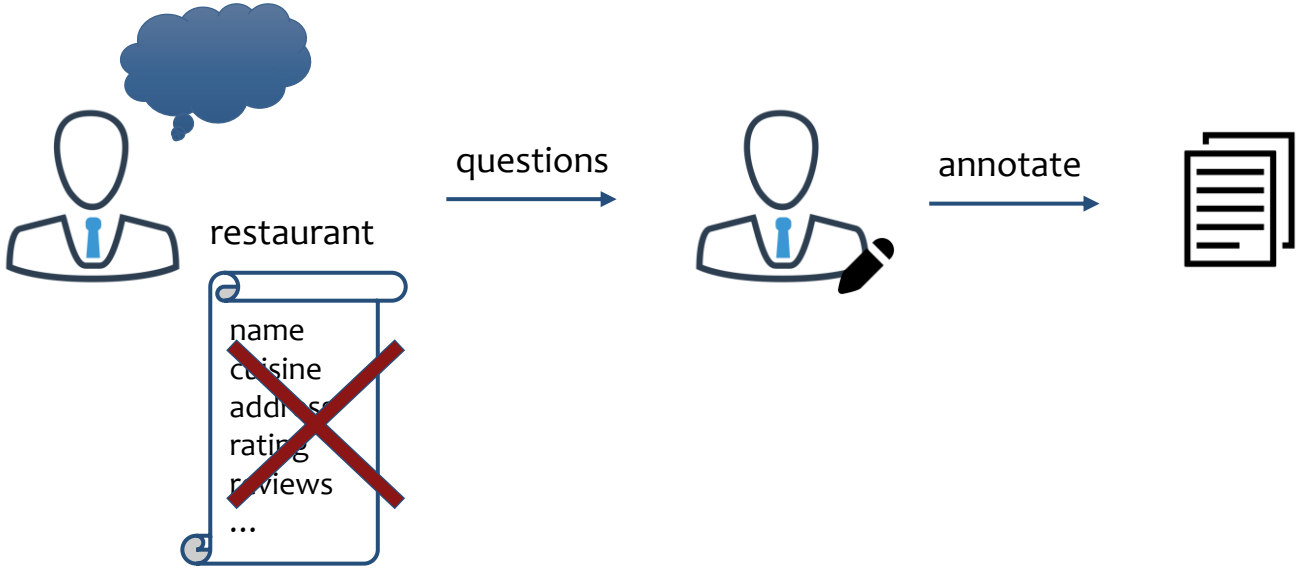
Comparison with Commercial Assistants



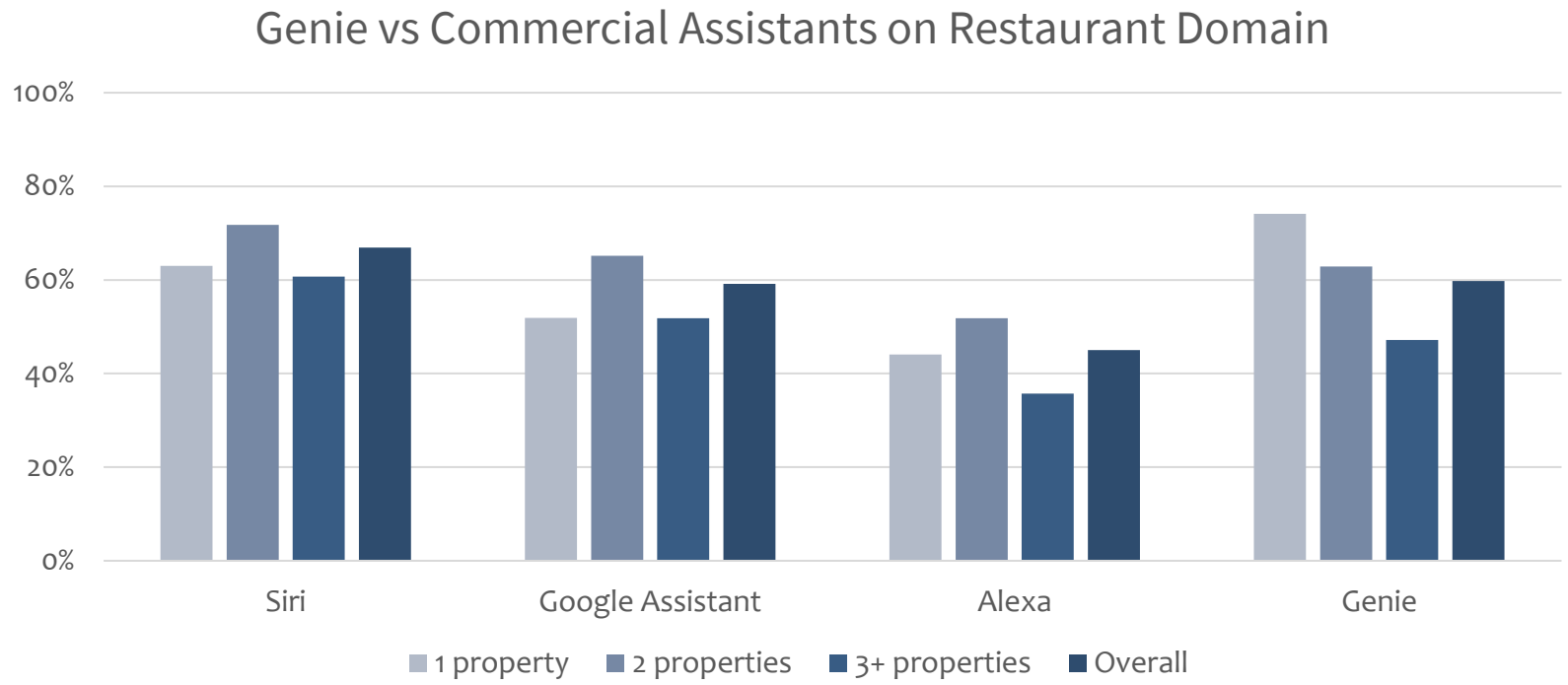
Example Questions

	Siri	Google	Alexa	Genie
Show restaurants near Stanford rated higher than 4.5	✗	✗	✗	✓
Show me restaurants rated at least 4 stars with at least 100 reviews	✗	✗	✗	✓
What is the highest rated Chinese restaurants in Hawaii?	✓	✗	✓	✓
How far is the closest 4 star and above restaurant?	✗	✗	✗	✓
Find a W3C employee that went to Oxford	✗	✗	✗	✓
Who worked for both Google and Amazon?	✗	✗	✗	✓
Who graduated from Stanford and won a Nobel prize?	✗	✓	✗	✓
Who worked for at least 3 companies?	✗	✗	✗	✓
Show me hotels with checkout time later than 12PM	✗	✗	✗	✓
Which hotel has a swimming pool in this area?	✗	✓	✗	✓

Evaluate on Common Questions



Comparison with Commercial Assistants on Common Questions



Discussions

Why do commercial assistants do a poor job on the first task but do a much better job in the second?

Discussions

- Why do commercial assistants do a better job in the second experiment?
  - they are tuned for common questions
  - they do a great job on recognizing common named entities
  - they can answer question correctly even with limited understanding of the question

- Why do commercial assistant do a poor job in the first experiment?
  - they are not tuned for complex long-tail questions
  - they don't even include some of the less-common properties (e.g., review count)
  - they do a poor job on numeric comparison

Error Analysis

- 50% of the errors are due to named entity recognition
  - work in progress (potential class project)
- 14% of the error can potentially be solvable with new templates
  - E.g., two fields with the same value: “movies produced and directed by Steven Spielberg”
- If we fix these two, we can get close to 90%!
- Others: typos, joins operators

Can We Do Better?

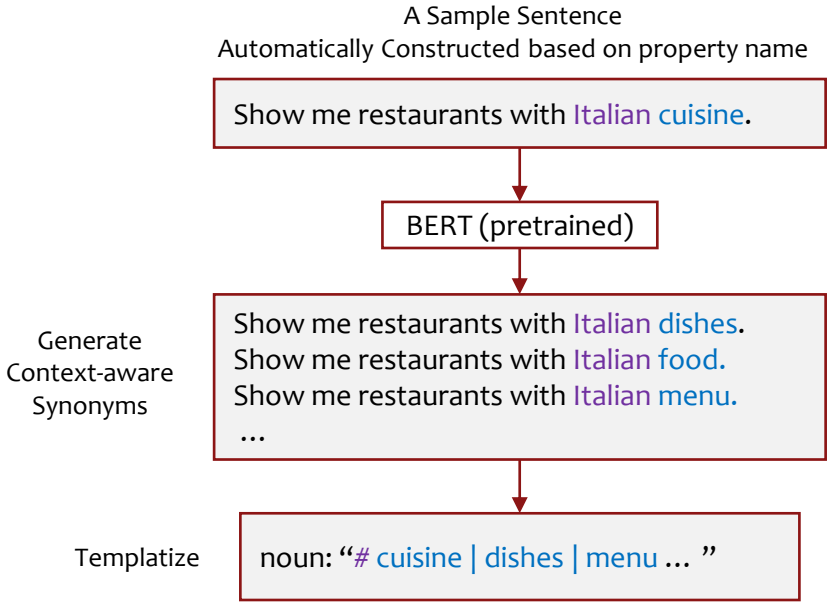
Manual Steps in Genie Pipeline

- Natural language annotations
  - We ask developers to provide natural language annotations, and it takes a few iterations to get a good quality set of annotations
- Paraphrase
  - We ask crowd workers to manually paraphrase synthetic sentences
  - We can only do this for a small sample of synthetic because of cost

- Can we replace them with something **automatic**?

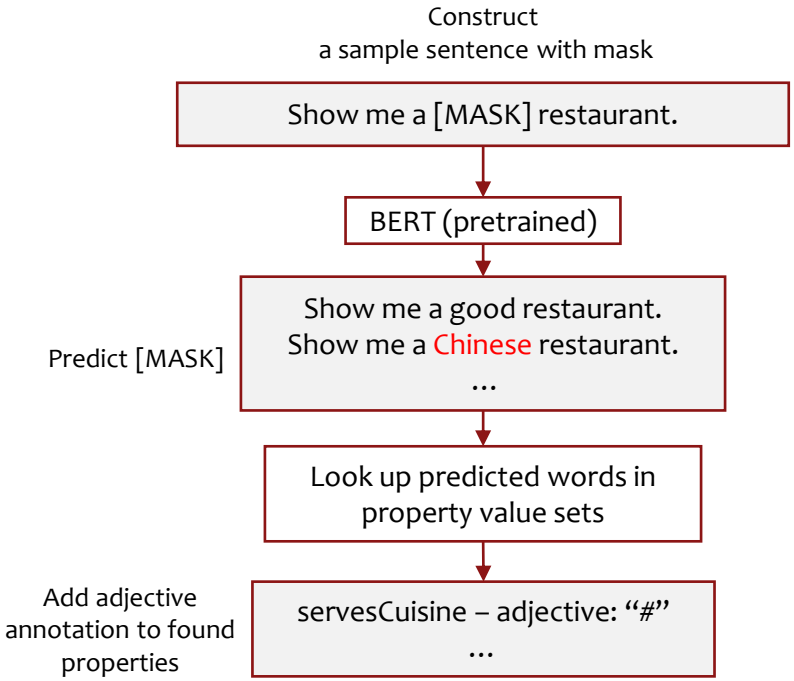
Automatic NL Annotation Generation

- Generate context-aware synonyms by a language model

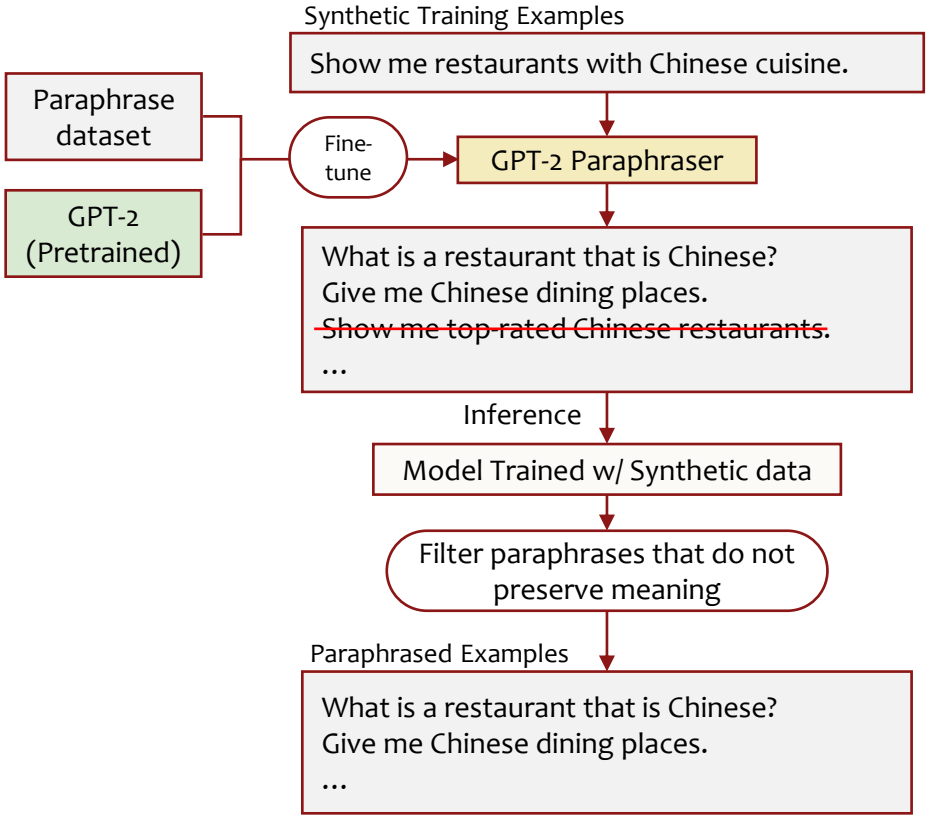


Automatic NL Annotation Generation (cont.)

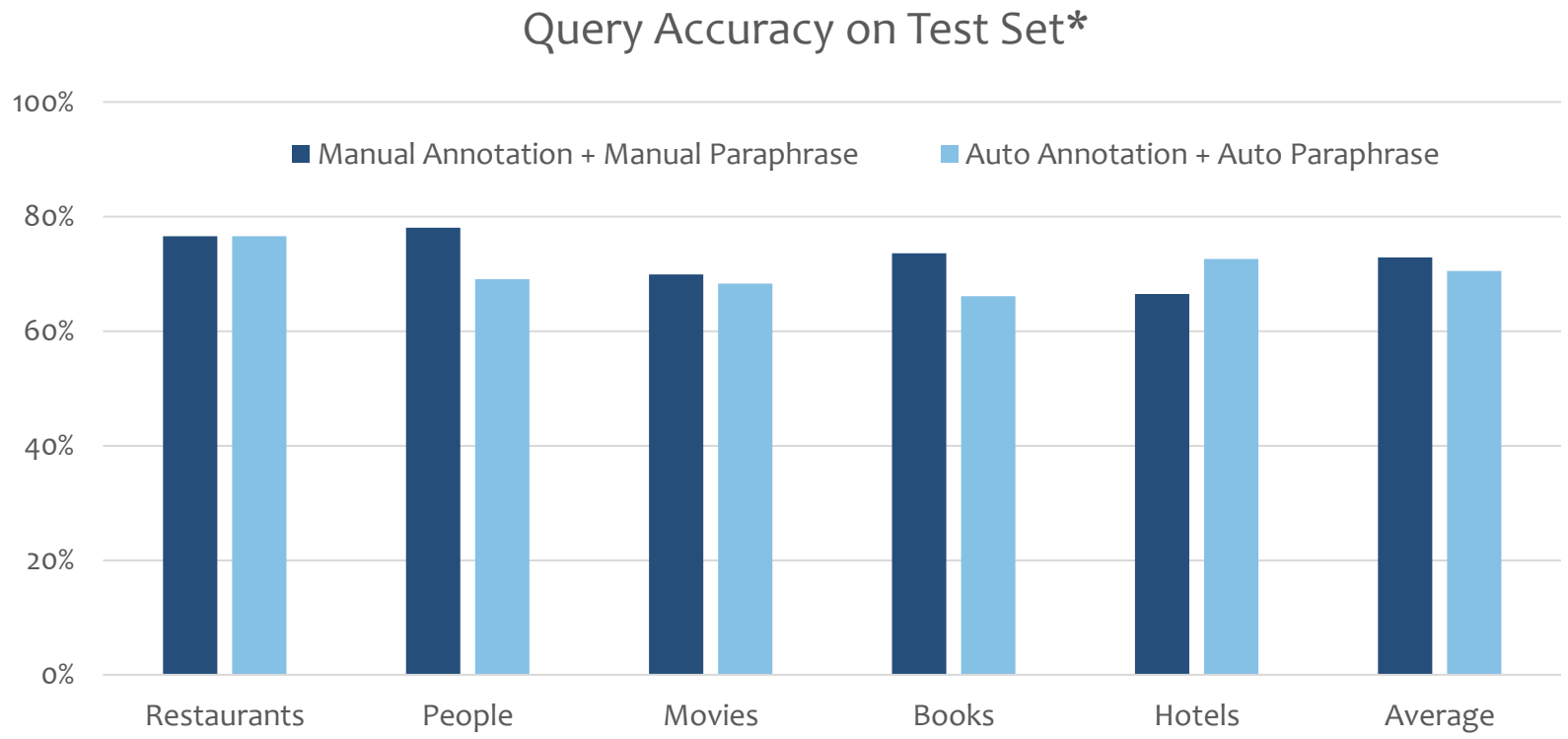
- Predict adjective qualifiers by a language model



Automatic Paraphrasing



Experimental Result



* evaluated on an older version of the dataset with fewer properties per domain

Thank you!