

Question Answering on Web Data

Silei Xu

CS294S April 9, 2020

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

Commercial Assistants

Alexa

User hand-codes question/code 1 by 1

get me an upscale restaurants

What are the restaurants around here?

What is the best restaurant?

search for Chinese restaurants

Commercial Assistants

Alexa

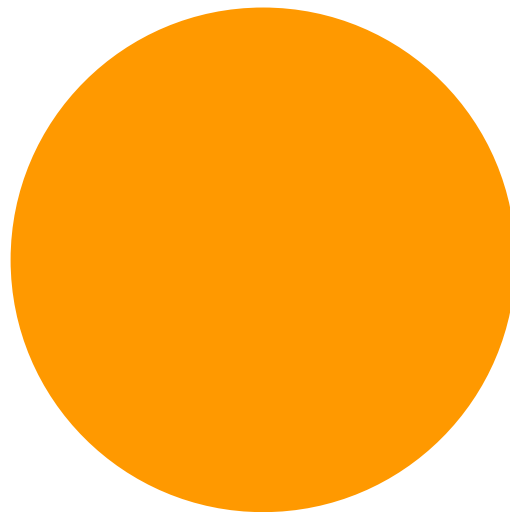
User hand-codes question/code 1 by 1

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

User hand-codes question/code 1 by 1

get me an upscale restaurants

What are the restaurants around here?

What is the best restaurant?

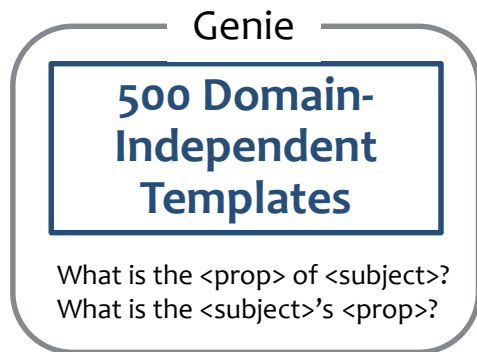
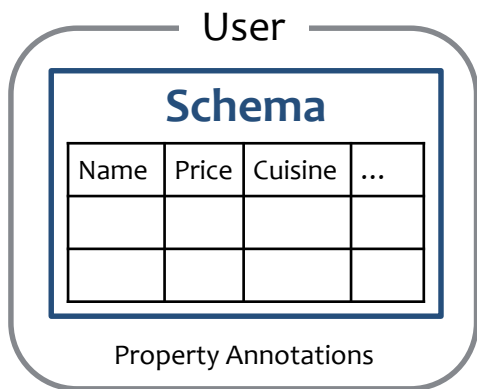
search for Chinese restaurants



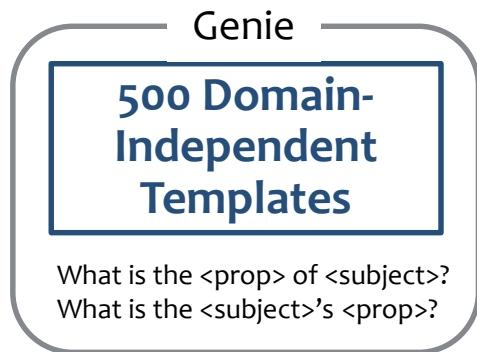
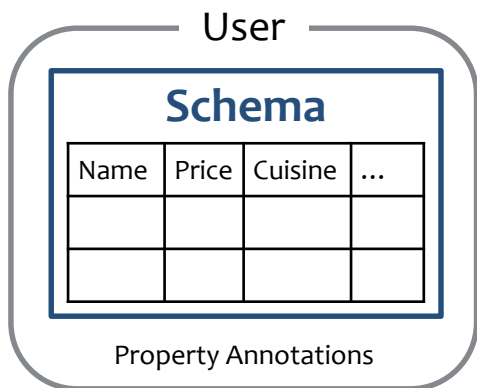
100K Alexa skills
Sep 2019

1.8 billion websites

Genie: Synthesize Question/Code from a Schema



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

The Web Has a Schema!

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 ...

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 ...

```
<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

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

Outline

- Introduction to Schema.org
- Represent Questions in ThingTalk
- LUINet: NL to ThingTalk
 - Training data generation
 - Experimental results
- Work in progress: automate everything!

Introduction to Schema.org

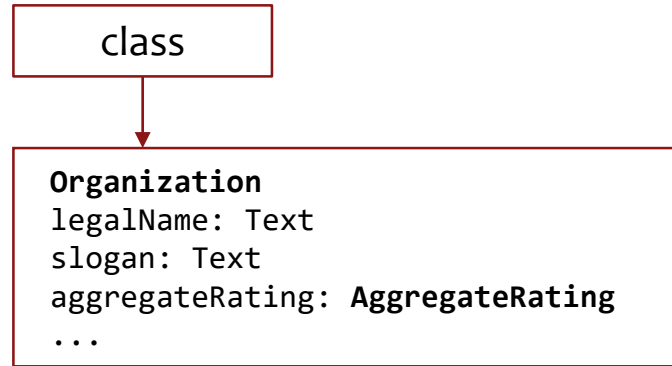


Graph Data Model of Schema.org

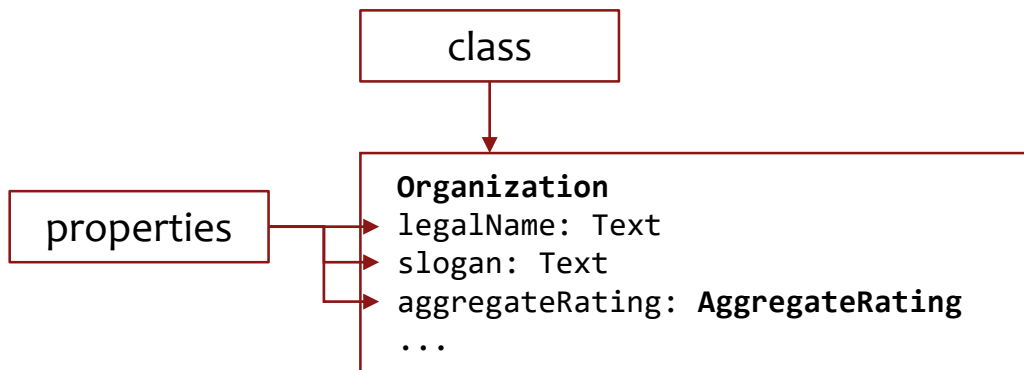
Graph Data Model of Schema.org

```
Organization  
legalName: Text  
slogan: Text  
aggregateRating: AggregateRating  
...
```

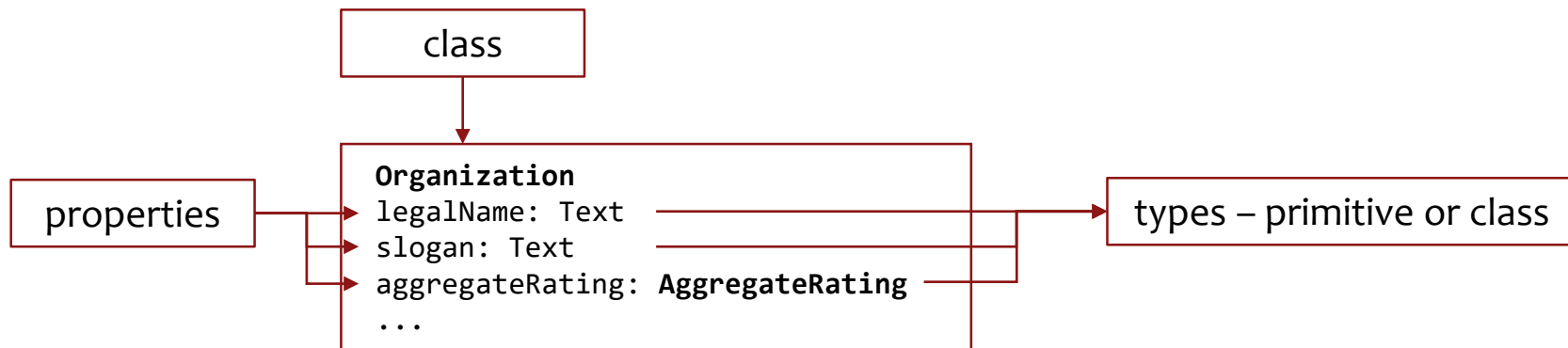
Graph Data Model of Schema.org



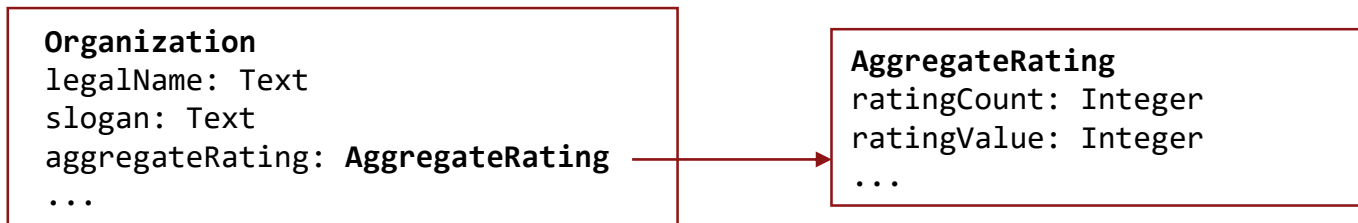
Graph Data Model of Schema.org



Graph Data Model of Schema.org



Graph Data Model of Schema.org



Schema.org Hierarchy

Thing

name: Text
url: URL
...

Organization (Thing)

legalName: Text
slogan: Text
aggregateRating: **AggregateRating**
...

AggregateRating

ratingCount: Integer
ratingValue: Integer
...



Schema.org Hierarchy

Thing

name: Text
url: URL
...

Organization (Thing)

legalName: Text
slogan: Text
aggregateRating: **AggregateRating**
...


AggregateRating

ratingCount: Integer
ratingValue: Integer
...

LocalBusiness (Place, Organization)

openingHours: Text
priceRange: Text
...

Some useful tools

- Google Structured Data Testing Tool  Google Structured Data Testing Tool
 - Show schema.org markups in a web page
- Google Custom Search
 - Search for pages that contain certain schema.org domains

Restrict Pages using Schema.org Types

Restrict pages from the above site list to only those that contain Schema.org types from the list below.

You can add up to ten (10) schema.org types to your Search Engine. Note that when you add a node, all its children automatically get included, so you do not need to add them again. For example, if you add CreativeWork, you do not need to add Book, ImageObject, VideoObject etc. separately.

Restaurant

ThingTalk for Questions



ThingTalk for QA

table [, *filter*]?

ThingTalk for QA

table [, *filter*]?

now =>

```
@QA.restaurant(), geo == makeLocation("Stanford")
```

=> notify

Show me restaurants in Stanford

ThingTalk for QA

table [, *filter*]?

now =>

```
@QA.restaurant(), geo == makeLocation("Stanford")  
&& servesCuisine =~ "Chinese"
```

=> notify

Show me Chinese restaurants in Stanford

ThingTalk for QA

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

now =>

```
@QA.restaurant(), geo == makeLocation("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 == makeLocation("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 == makeLocation("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 == makeLocation("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]? $table$ [, $filter$]?

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*]?

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*]?

...

**LUINet:
NL to ThingTalk**



Natural Language Programming



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

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

Genie Pipeline



Genie Pipeline

Thingpedia Manifest

Schema

Name	Price	Cuisine	...

Natural Language Annotations

cuisine of the restaurant
restaurant's cuisine
cuisine served by the restaurant



Genie Pipeline

Thingpedia Manifest

Schema

Name	Price	Cuisine	...

Natural Language Annotations

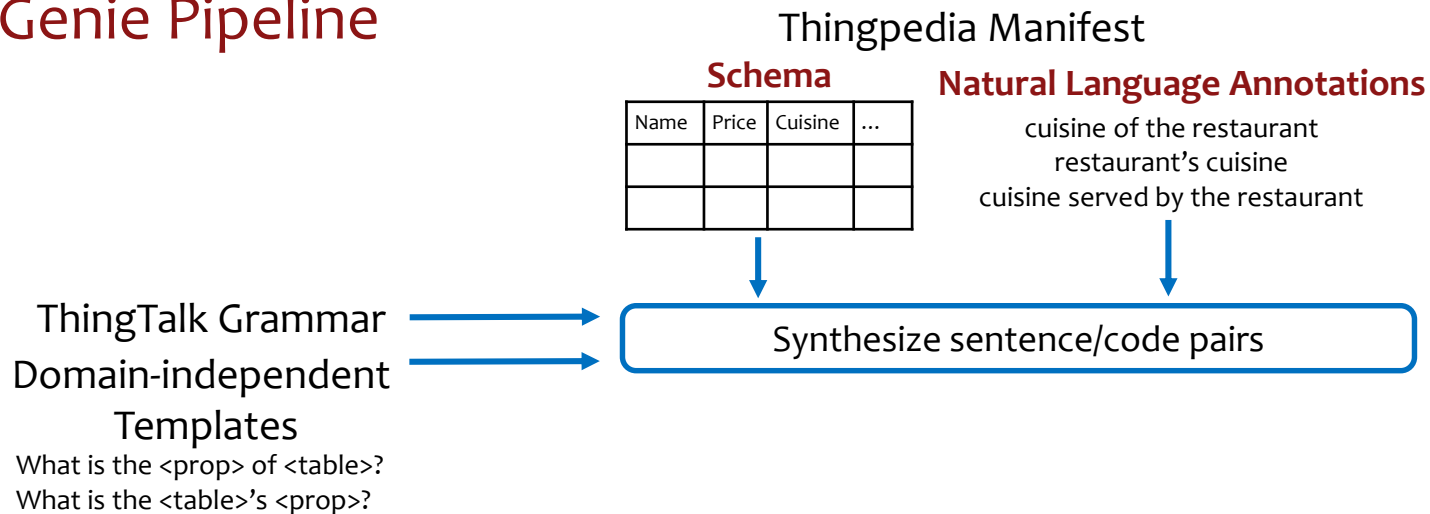
cuisine of the restaurant
restaurant's cuisine
cuisine served by the restaurant

ThingTalk Grammar
Domain-independent
Templates

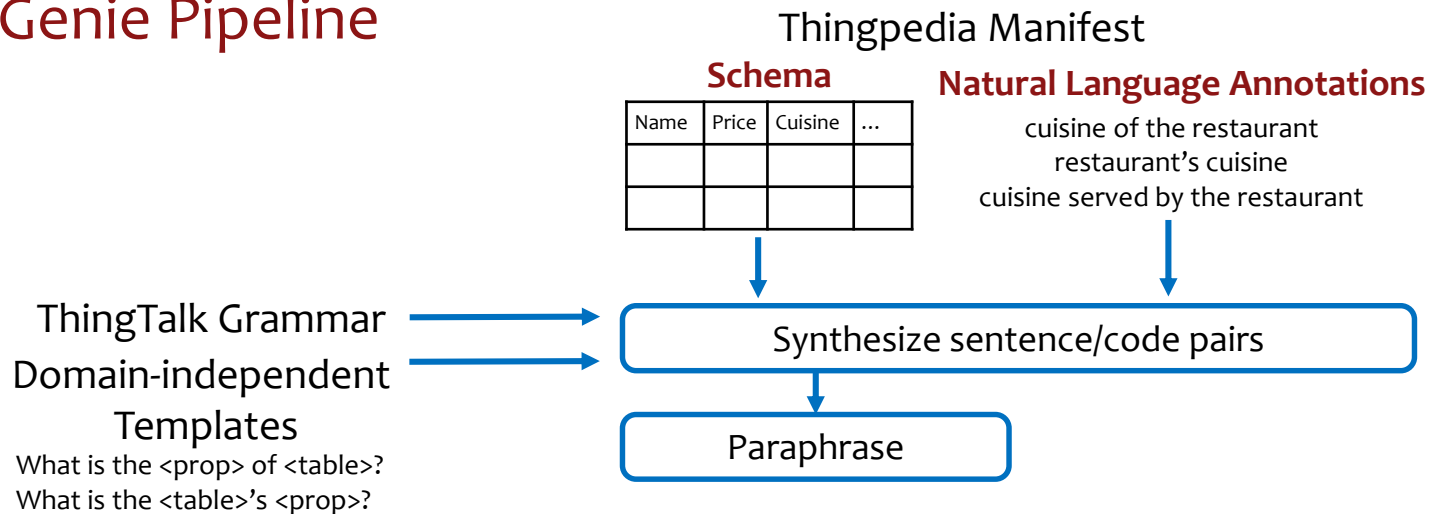
What is the <prop> of <table>?
What is the <table>'s <prop>?



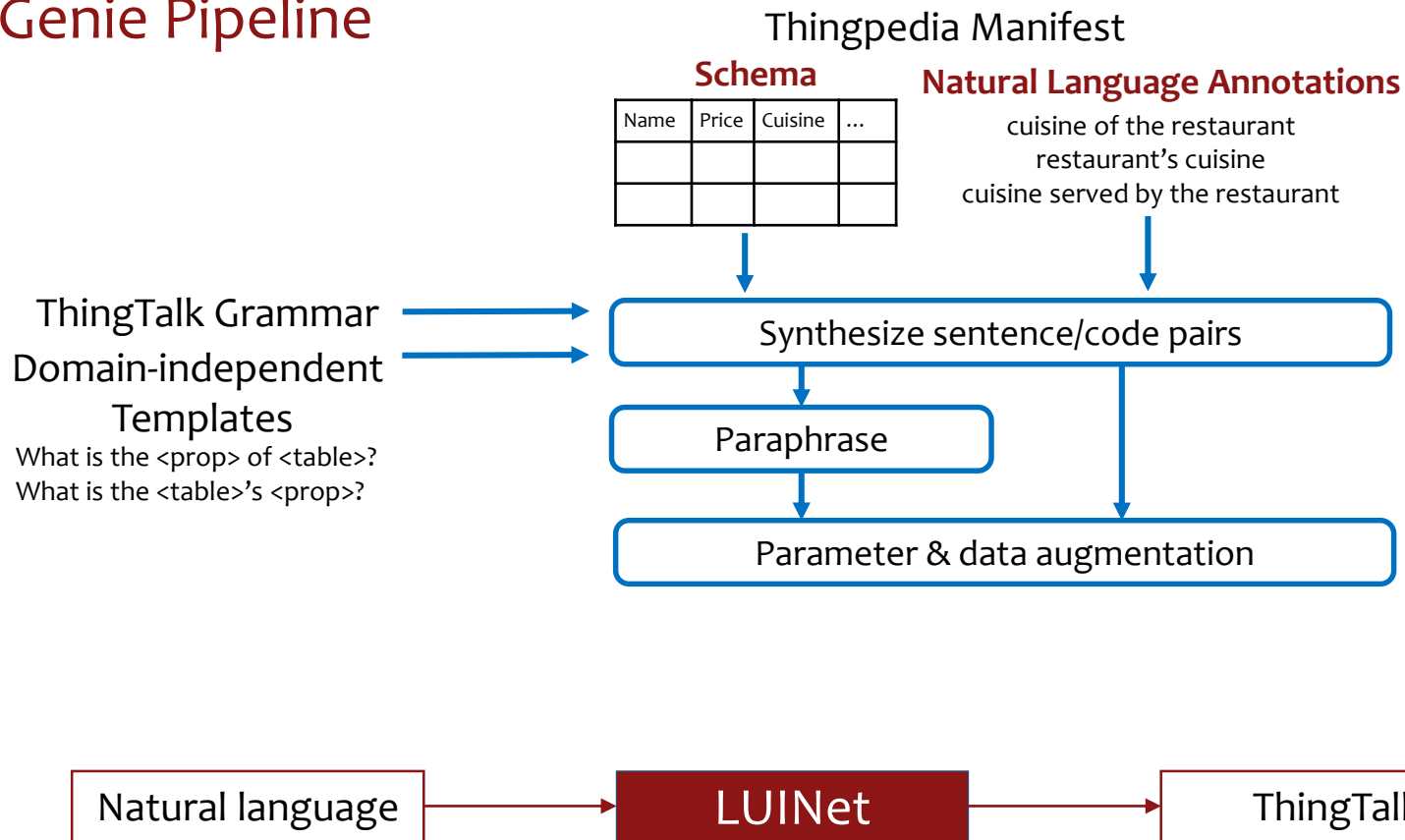
Genie Pipeline



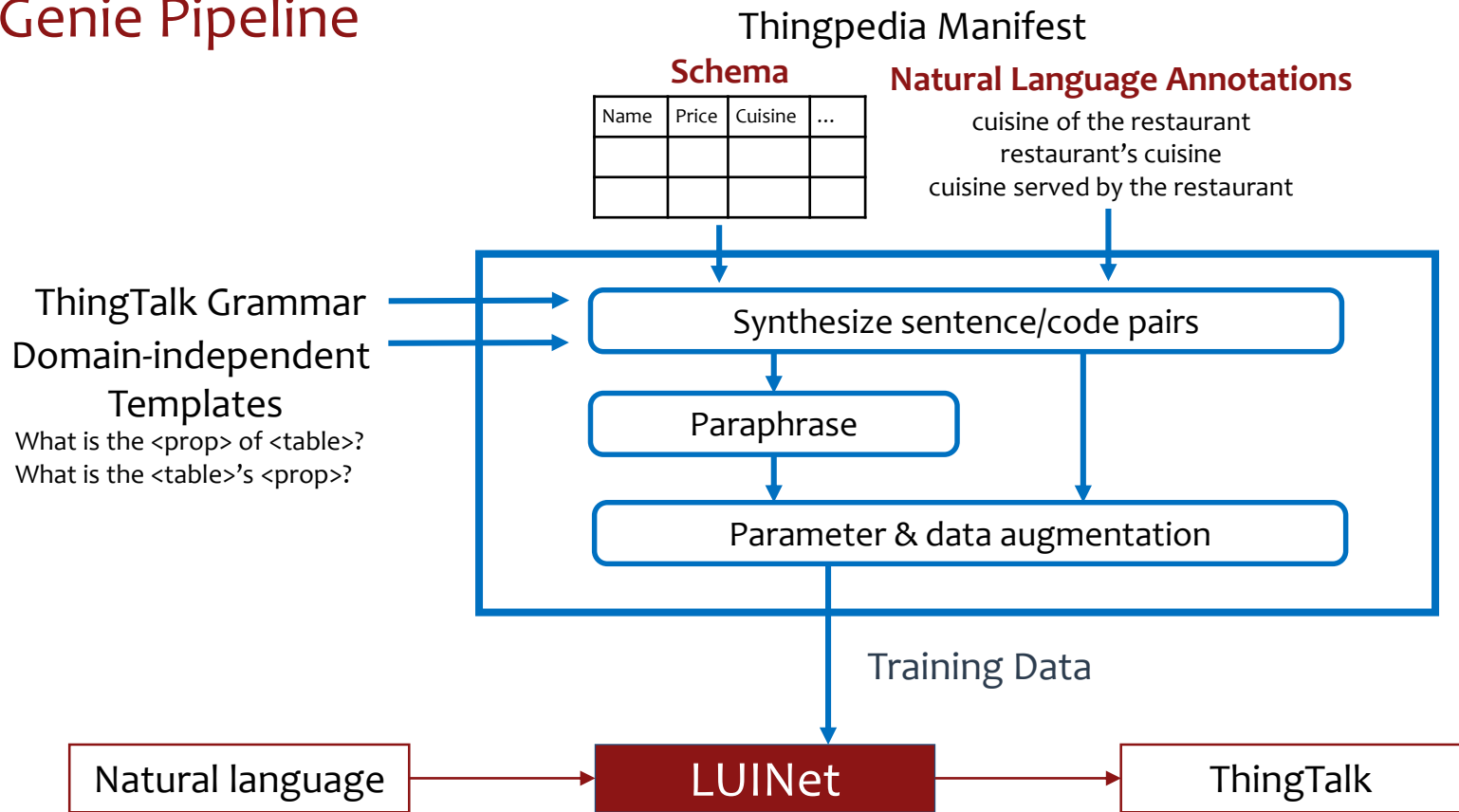
Genie Pipeline



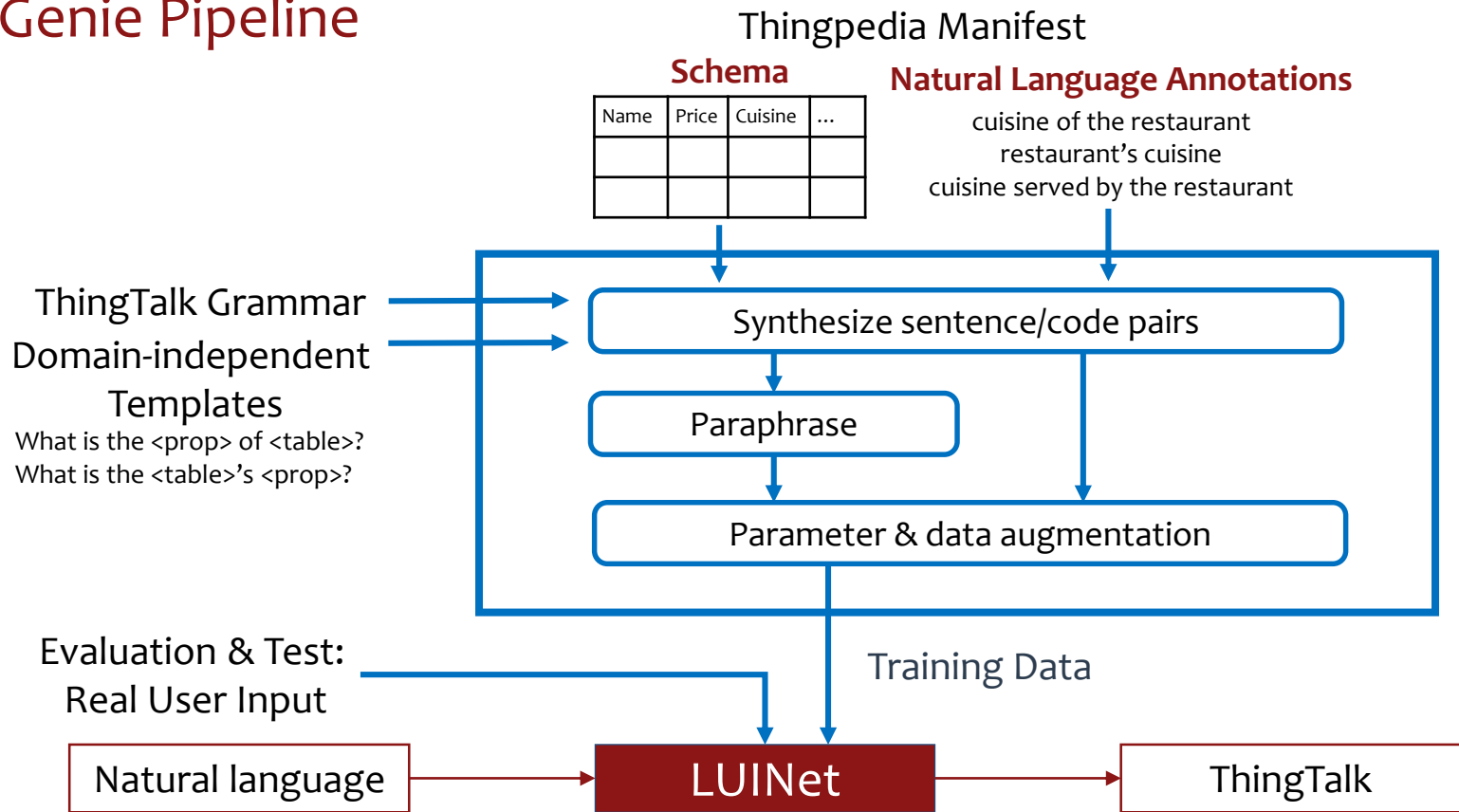
Genie Pipeline



Genie Pipeline



Genie Pipeline



Automatically Turn Schema.org into Thingpedia Manifest

- Tables with hierarchy
 - properties are inherited from parent tables
 - only keep classes & properties with data
 - decide types based on schema.org types and data

```
@org.schema {
  Restaurant extends FoodEstablishment {}
  FoodEstablishment extends LocalBusiness {
    acceptsReservation: Boolean,
    servesCuisine: String, ...
  }
  LocalBusiness extends Place, Organizations {
    priceRange: String,
    openingHours: String, ...
  }
  Organizations extends Thing {
    aggregateRating: {
      ratingCount: Number,
      ratingValue: Number,
    },
    review: Array(Review),
  }
  Thing {
    name: String, ...
  }
}
```

Map properties to natural language

Map properties to natural language

- Long, non-word, property names
 - E.g., ratingValue, servesCuisine
- Variety in natural language usage

Map properties to natural language

- Long, non-word, property names
 - E.g., ratingValue, servesCuisine
- Variety in natural language usage

servesCuisine	ratingValue
Chinese restaurant ✓	4.5 restaurant ✗
Restaurant with Chinese cuisine ✓	Restaurant with 4.5 rating ✓
Restaurant served Chinese cuisine ✗	Restaurant rated 4.5 ✓
Restaurant that serves Chinese cuisine ✓	Restaurant rates 4.5 ✗
Restaurant with Chinese ✗	Restaurant with 4.5 ✗
...	...

NL Annotations by Part-Of-Speech Categories

- “servesCuisine”
 - Noun phrase
 - “cuisine”: e.g., “the cuisine of the restaurant”, “restaurants with Chinese cuisine”
 - Verb phrase
 - “serves # cuisine”, “serves #” : e.g., “restaurant that serves Chinese cuisine”, “what does the restaurant serve”
 - Adjective-phrase value (with no property name)
 - E.g., “Chinese restaurants”

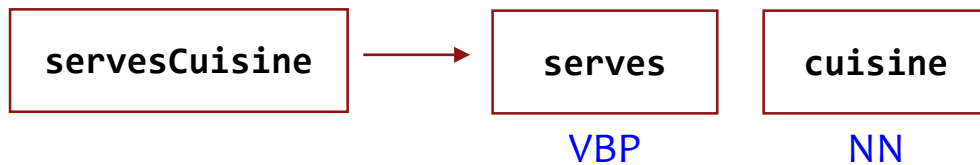
NL Annotation Generation

`servesCuisine`

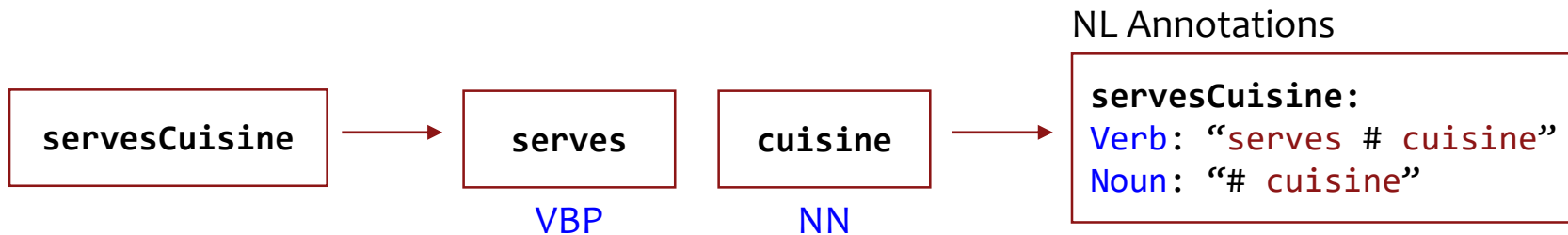
NL Annotation Generation



NL Annotation Generation

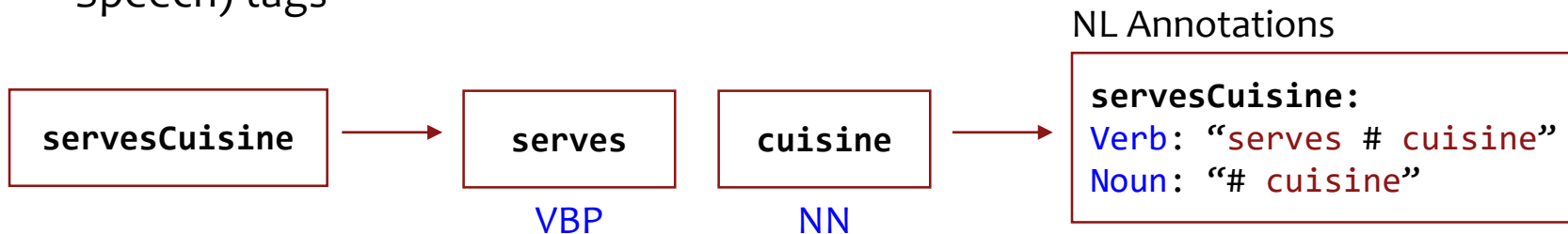


NL Annotation Generation



NL Annotation Generation

- Automatic: Heuristics based on POS (Part-Of-Speech) tags



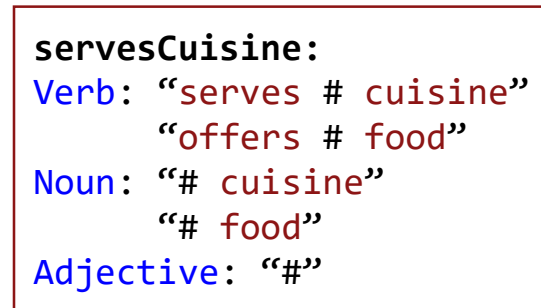
- Manual:
 - Provides additional synonyms, and annotations in different POS categories

NL Annotation Generation

- Automatic: Heuristics based on POS (Part-Of-Speech) tags



- Manual:
 - Provides additional synonyms, and annotations in different POS categories



Domain-independent Templates

servesCuisine:

Verb: “serves # cuisine”

“offers # food”

Noun: “# cuisine”

“# food”

Adjective: “#”

Domain-independent Templates

servesCuisine:

Verb: “serves # cuisine”

“offers # food”

Noun: “# cuisine”

“# food”

Adjective: “#”

Show me <table> that <verb>.

Show me <table> with <noun>.

Show me <adjective> <table>.

Domain-independent Templates

servesCuisine:

Verb: “serves # cuisine”

“offers # food”

Noun: “# cuisine”

“# food”

Adjective: “#”

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



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

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

Domain-independent Templates

```
servesCuisine:  
Verb: "serves # cuisine"  
      "offers # food"  
Noun: "# cuisine"  
      "# food"  
Adjective: "#"
```

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

Show me <table> with <noun:NUMBER>
greater than <value>.

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

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

→ Show me restaurants with rating greater than 4

Domain-independent Templates

servesCuisine:
Verb: “serves # cuisine”
 “offers # food”
Noun: “# cuisine”
 “# food”
Adjective: “#”

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

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

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

Show me <table> with <noun:NUMBER>
greater than <value>.

→ Show me restaurants with rating greater than 4

Show me <table> with <noun:MEASURE(m)>
longer than <value>.

→ Show me surfboard with length longer than 3m

Domain-dependent Templates

Domain-dependent Templates

- Some natural sentences cannot be generated by domain-independent templates:
 - “the top-rated restaurant”, “the best restaurant”
- We allow developers to improve the accuracy by providing domain-dependent templates

Domain-dependent Templates

- Some natural sentences cannot be generated by domain-independent templates:

ThingTalk	Sentence by domain-independent templates
<code>sort aggregateRating.ratingValue desc of @QA.restaurant()</code>	restaurant with the highest rating
	restaurant that have the highest rating
	...

- “the top-rated restaurant”, “the best restaurant”
- We allow developers to improve the accuracy by providing domain-dependent templates

Domain-dependent Templates

- Some natural sentences cannot be generated by domain-independent templates:

ThingTalk	Sentence by domain-independent templates
<code>sort aggregateRating.ratingValue desc of @QA.restaurant()</code>	restaurant with the highest rating
	restaurant that have the highest rating
	...

- “the top-rated restaurant”, “the best restaurant”
- We allow developers to improve the accuracy by providing domain-dependent templates

ThingTalk	Domain-dependent templates
<code>sort aggregateRating.ratingValue desc of @QA.restaurant()</code>	the top-rated restaurant
	the best restaurant
	...

Experiments



Experimental Results

- Domains
 - Restaurants: data from Yelp
 - Person: data from LinkedIn
- Training set

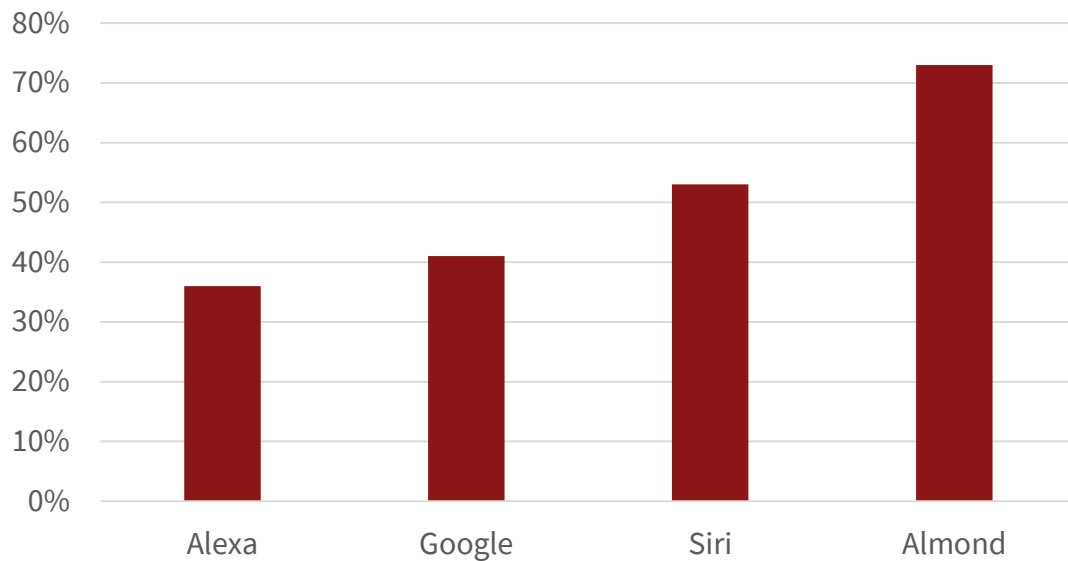
	Restaurant	Person
Synthetic	1,294,278	553,067
Paraphrase	6,288	6,000
Total (augmented)	1,809,109	930,564

- Realistic evaluation set

		Restaurant	Person
Dev	1 property	134	6
	2 properties	47	144
	3+ properties	59	0
	Total	240	160
Test	1 property	96	127
	2 properties	79	106
	3+ properties	40	0
	Total	215	233

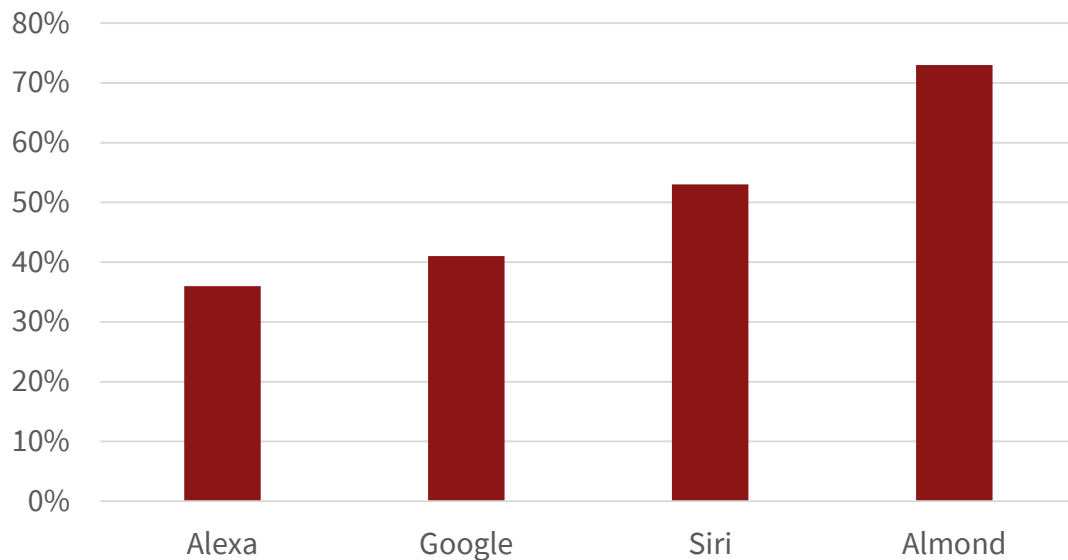
Comparison with Commercial Virtual Assistants

Answer Accuracy on Restaurant Queries



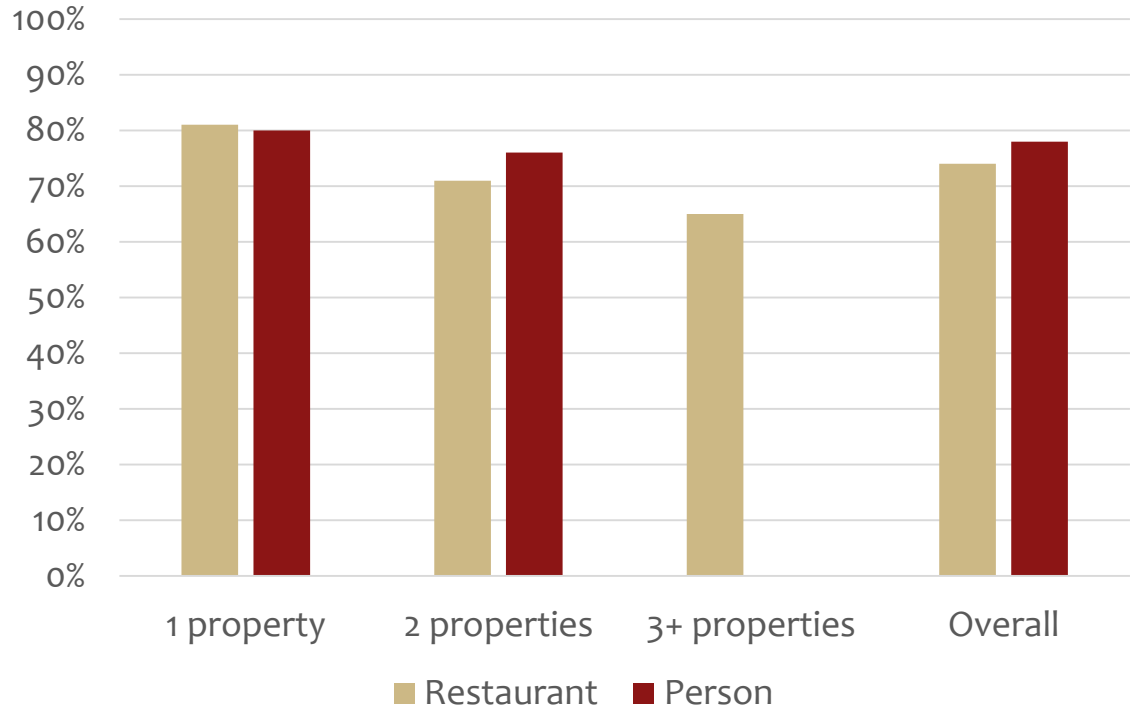
Comparison with Commercial Virtual Assistants

Answer Accuracy on Restaurant Queries



Trained with no real data!

Experimental Results



Can We Do Better?



Manual Effort in Genie Pipeline

Manual Effort in Genie Pipeline

- Natural language annotations
 - The heuristics based on part-of-speech doesn't provide good variety, and sometimes unnatural
- Paraphrase
 - We ask crowdworkers 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

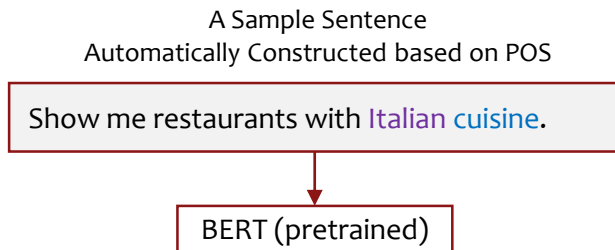
- Generate context-aware synonyms by a language model

A Sample Sentence
Automatically Constructed based on POS

Show me restaurants with [Italian cuisine](#).

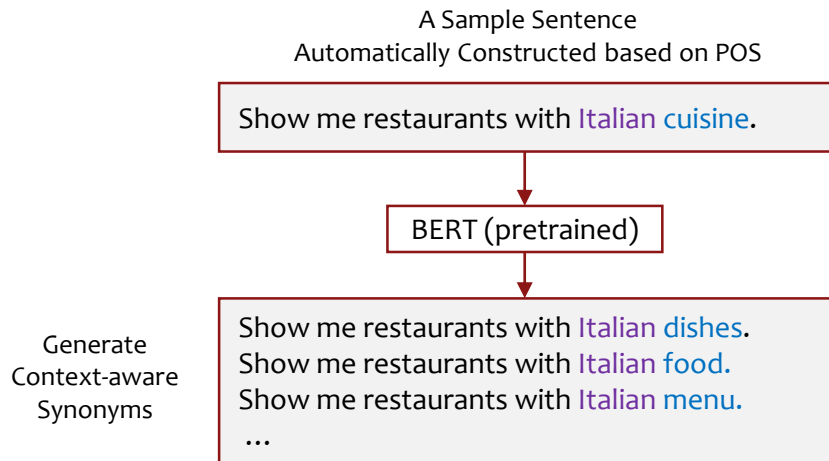
Automatic NL Annotation Generation

- Generate context-aware synonyms by a language model



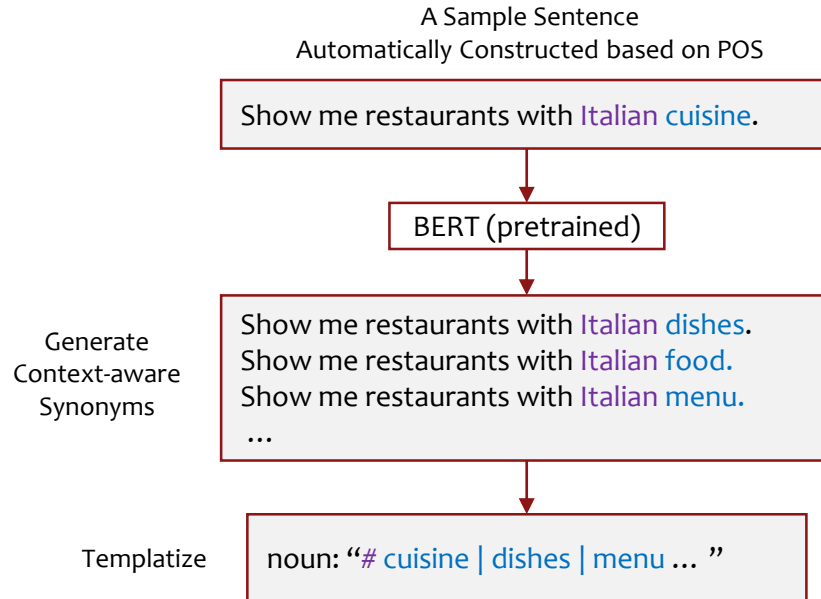
Automatic NL Annotation Generation

- Generate context-aware synonyms by a language model



Automatic NL Annotation Generation

- Generate context-aware synonyms by a language model



Automatic NL Annotation Generation (cont.)

- Predict adjective properties by a language model

Automatic NL Annotation Generation (cont.)

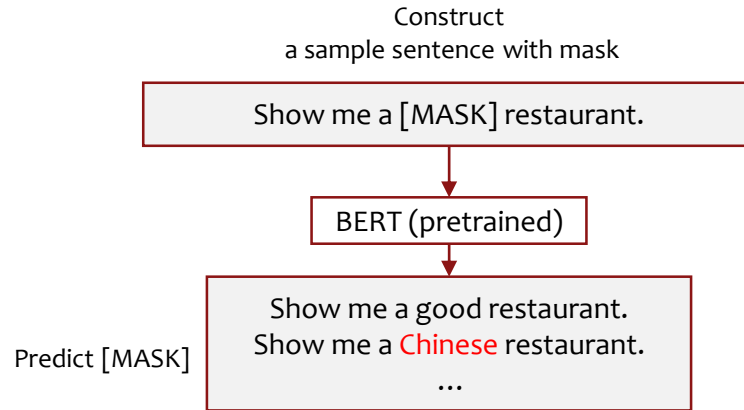
- Predict adjective properties by a language model

Construct
a sample sentence with mask

Show me a [MASK] restaurant.

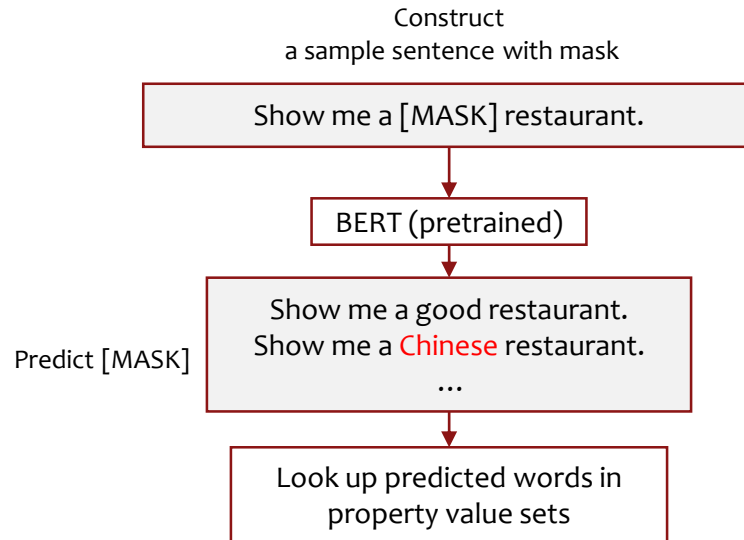
Automatic NL Annotation Generation (cont.)

- Predict adjective properties by a language model



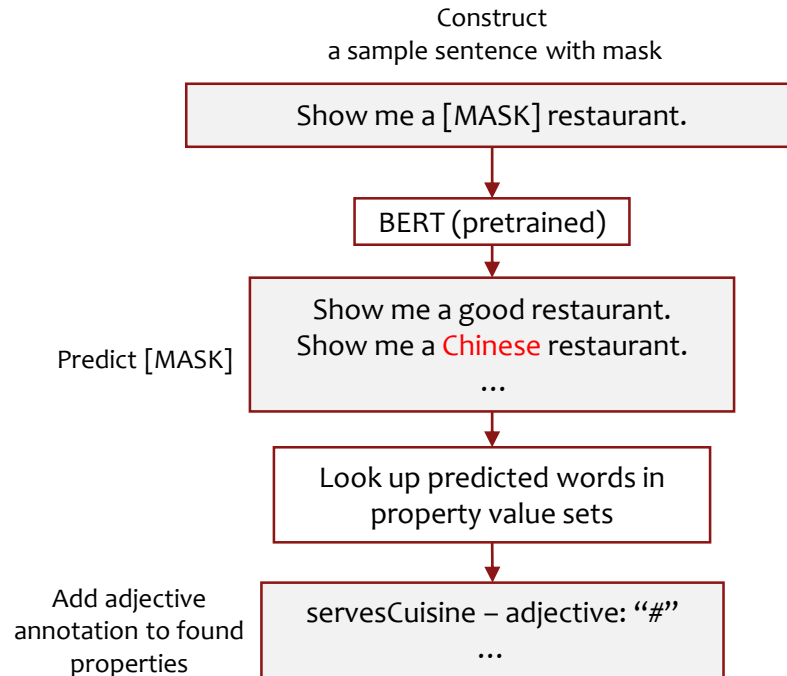
Automatic NL Annotation Generation (cont.)

- Predict adjective properties by a language model

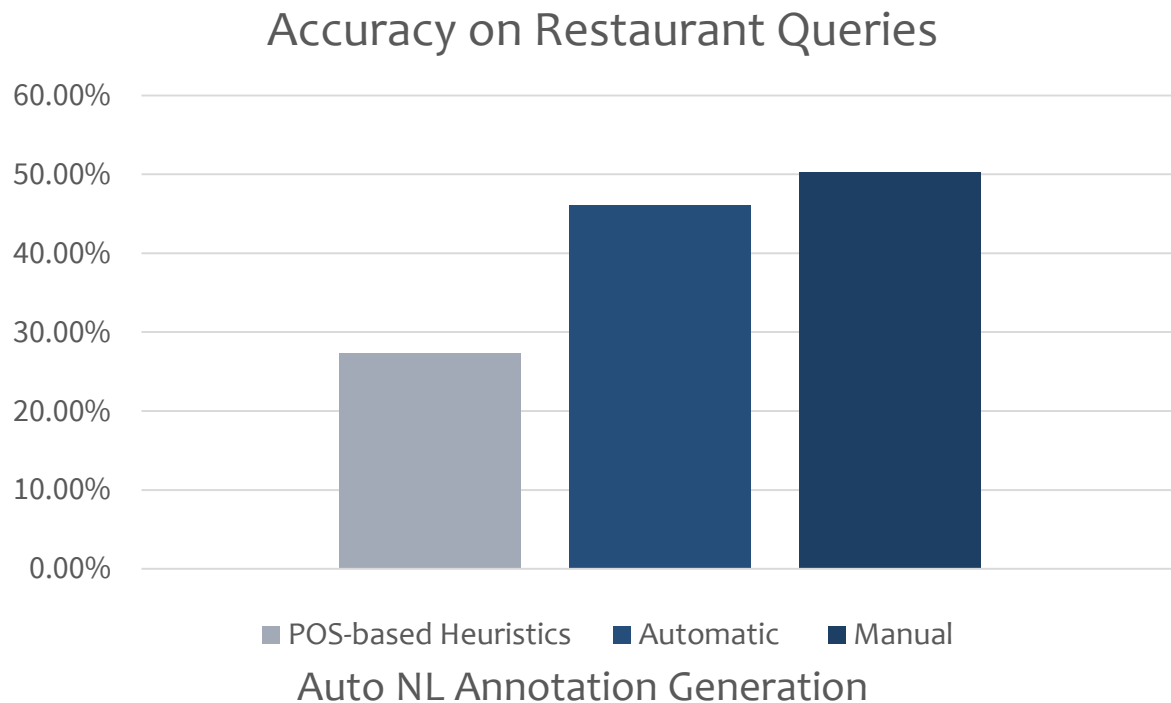


Automatic NL Annotation Generation (cont.)

- Predict adjective properties by a language model



Preliminary Experimental Result



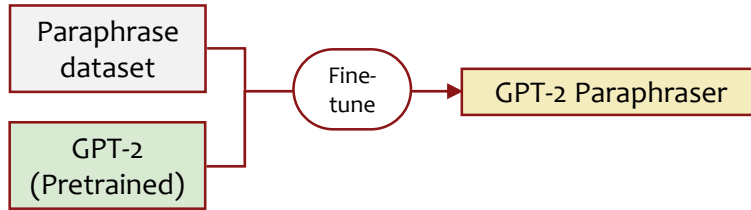
Automatic Paraphrasing

Automatic Paraphrasing

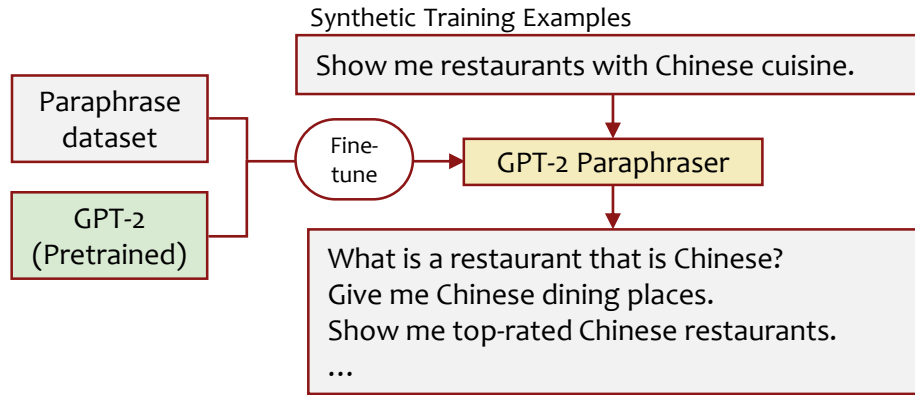
Paraphrase
dataset

GPT-2
(Pretrained)

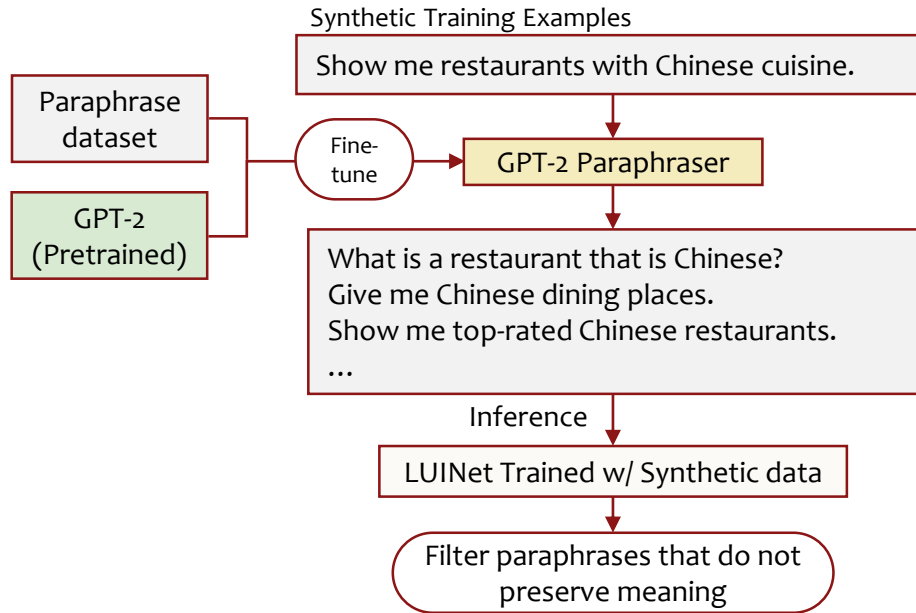
Automatic Paraphrasing



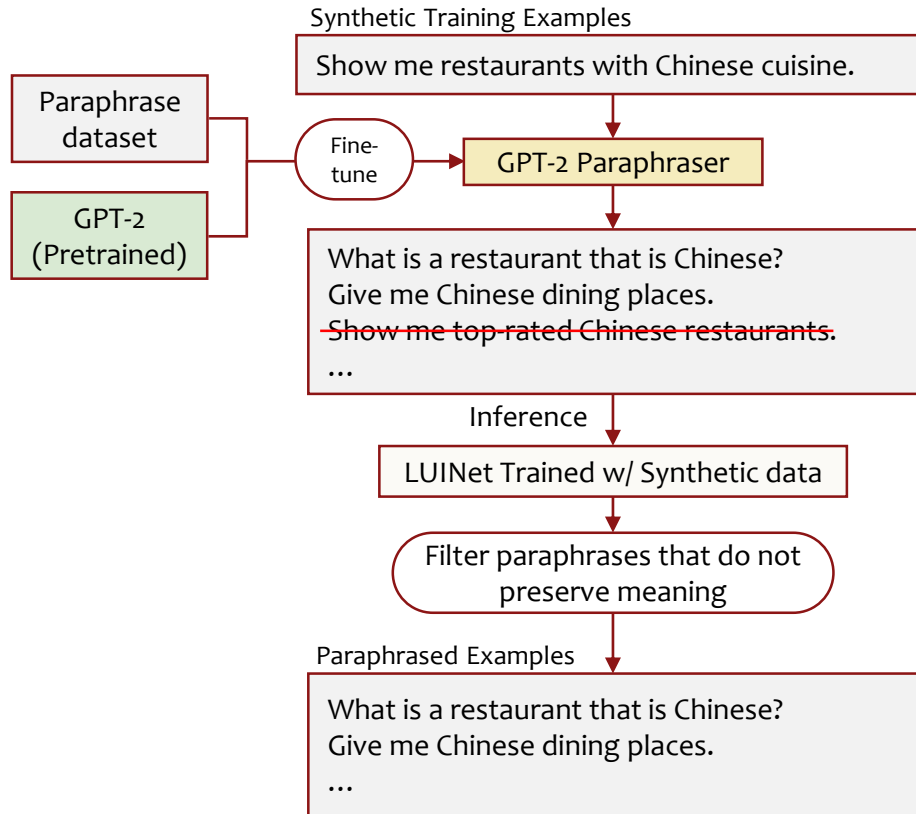
Automatic Paraphrasing



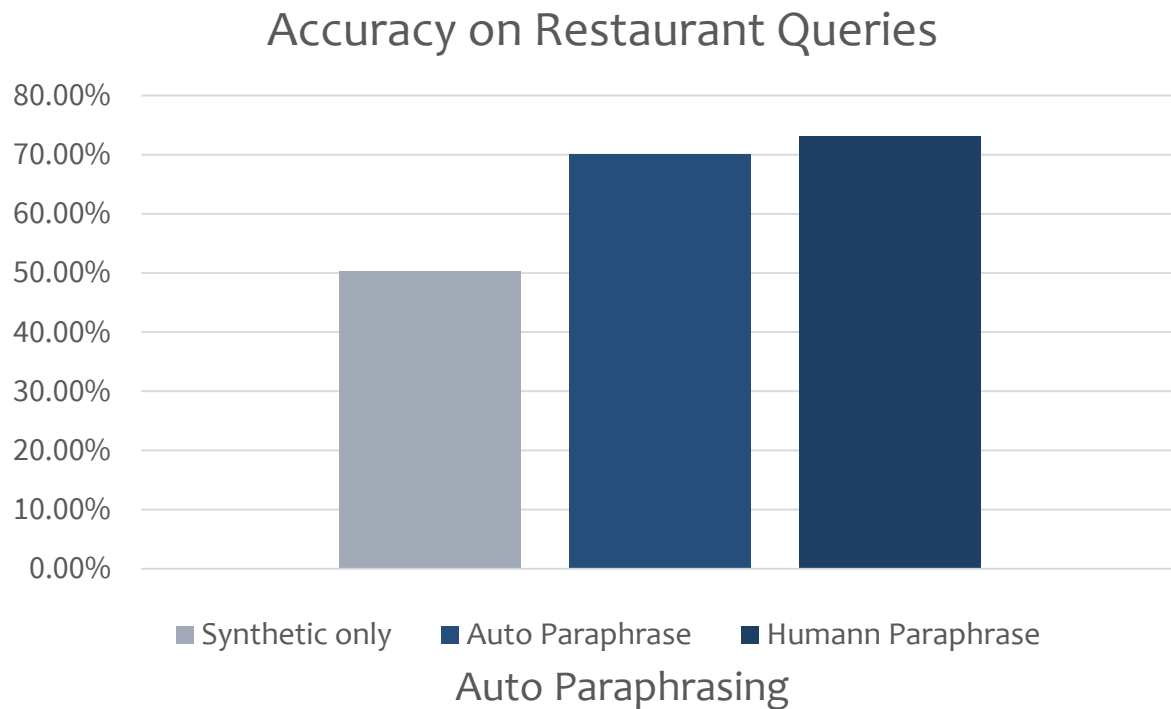
Automatic Paraphrasing



Automatic Paraphrasing



Preliminary Experimental Result



Thank you!

Hope you will enjoy your homework 😊