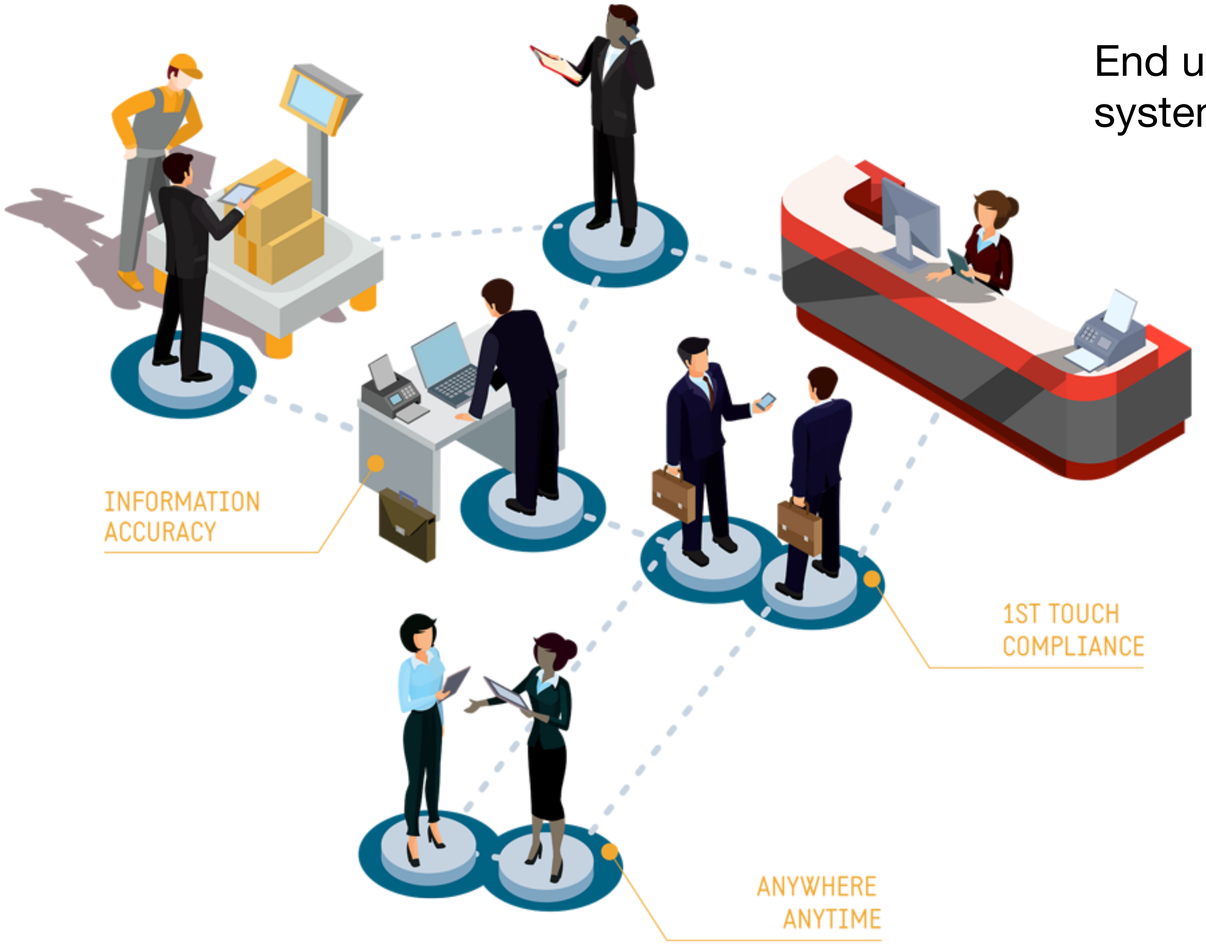
Photon: A Robust Cross-Domain Text-to-SQL System

Jichuan Zeng*, Victoria Lin*, Caiming Xiong, Richard Socher,
Michael R. Lyu, Irwin King, Steven C.H. Hoi

ACL 2020 - System Demonstration

Motivation



End users access information system everyday, everywhere…

Live demo: http://naturalsql.com/

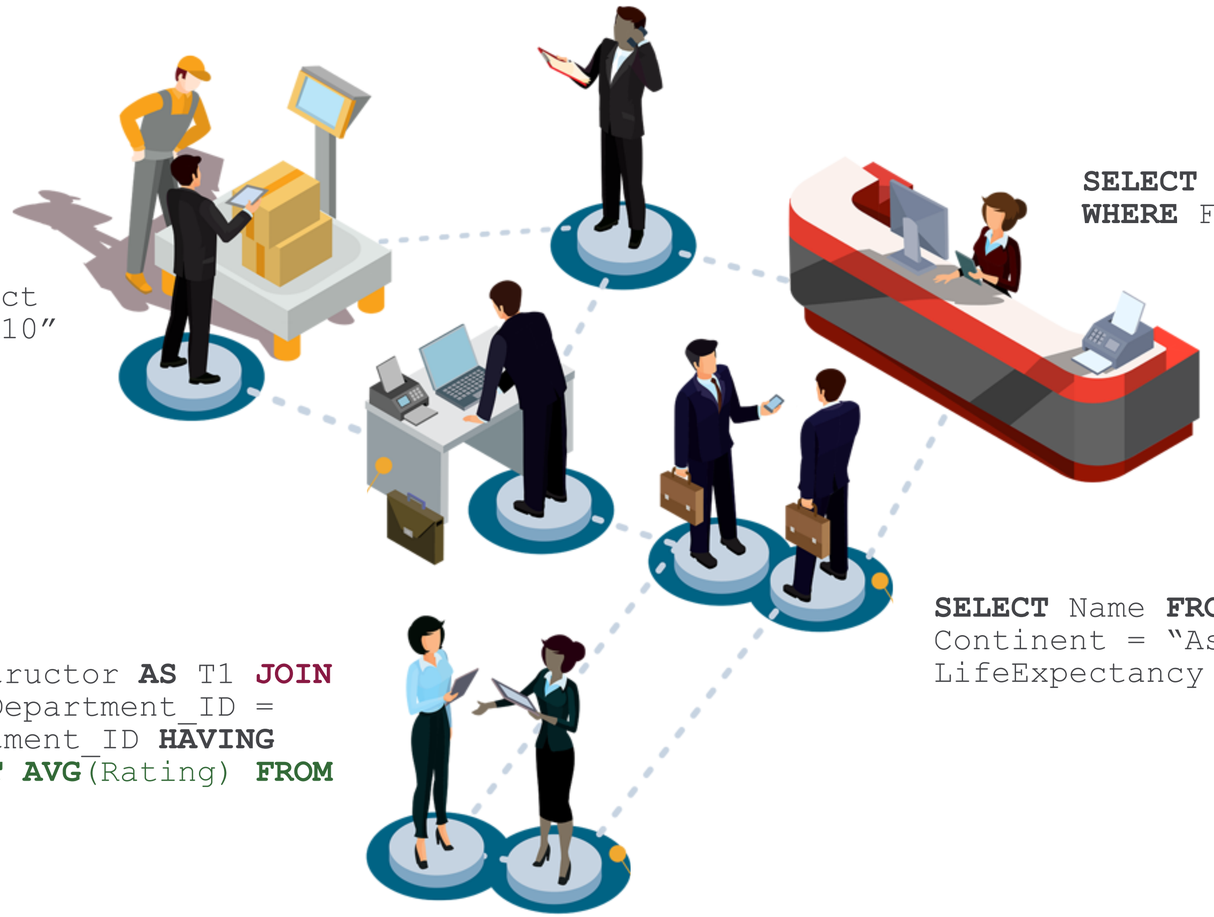
Motivation

Scenario 1: Everyone is a programming master

Live demo: http://naturalsql.com/

Motivation

Scenario 1: Everyone is a programming master



```
SELECT Quantity FROM Product
WHERE Name = "Hoverboard x10"
```

```
SELECT Arriving_Time FROM Flights
WHERE Flight_Number = "CZ327"
```

```
SELECT T2.name FROM Instructor AS T1 JOIN
Department AS T2 ON T1.Department_ID =
T2.ID GROUP BY T1.Department_ID HAVING
AVG(T1.Rating) > (SELECT AVG(Rating) FROM
Instructor)
```

```
SELECT Name FROM Country WHERE
Continent = "Asia" ORDER BY
LifeExpectancy LIMIT 1
```

Live demo: http://naturalsql.com/

Motivation

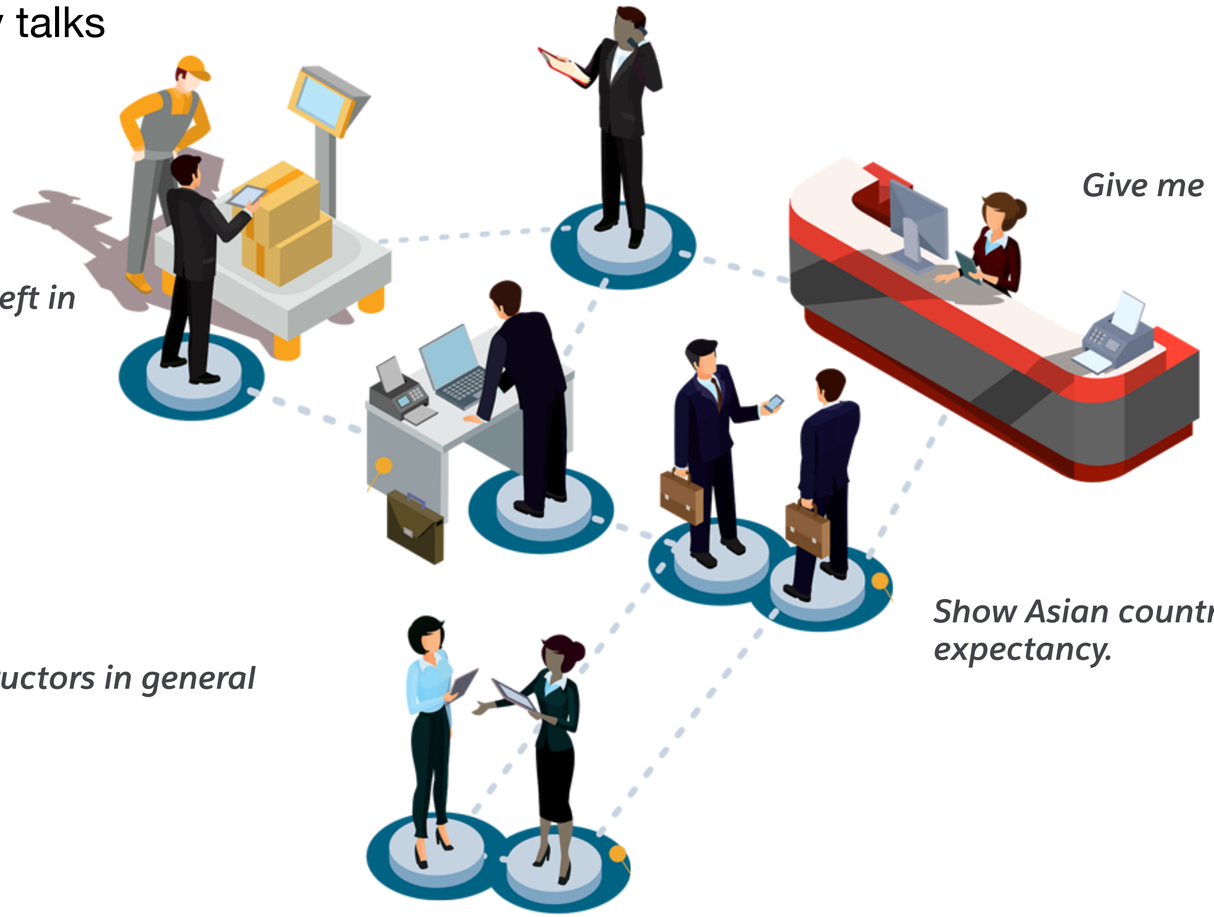
Scenario 2: Everyone simply talks to their information system

Live demo: http://naturalsql.com/

Motivation

Scenario 2: Everyone simply talks to their information system

How many “Hoverboard x10” are left in stock?



Give me the arriving time of “CZ327”.

Show Asian countries ordered by life expectancy.

Which departments have instructors in general rated above average?

Live demo: http://naturalsql.com/

Desiderata

Accurately map NL input to executable SQL queries

Live demo: http://naturalsql.com/

Desiderata

Accurately map NL input to executable SQL queries

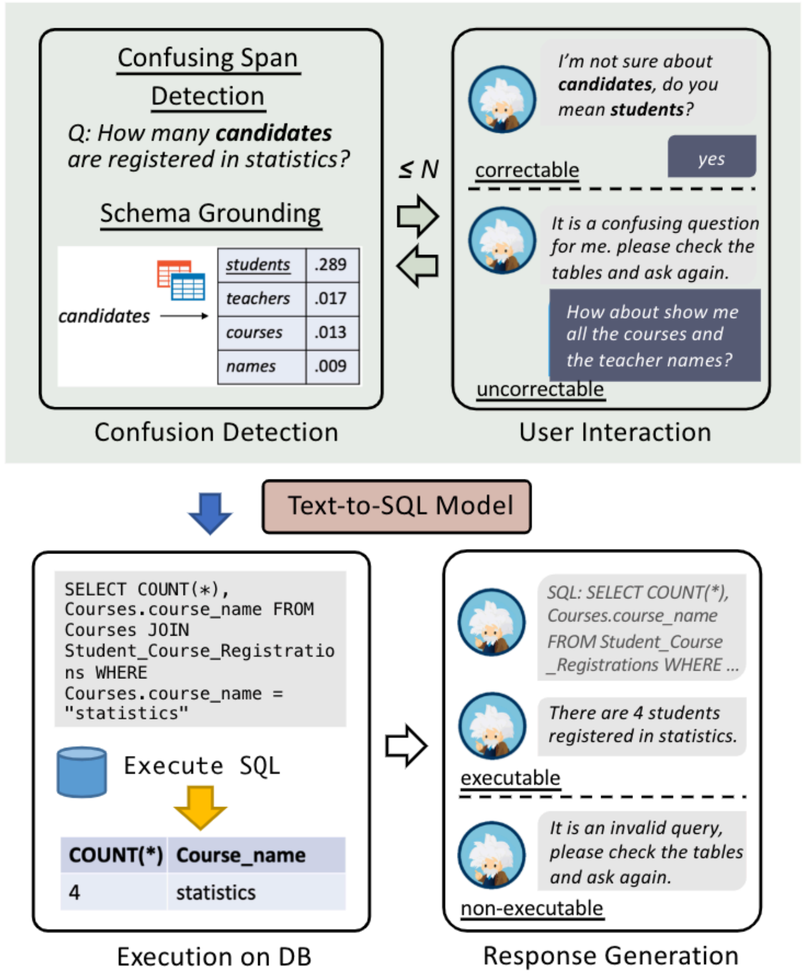
Work across different databases

Robustness - “don’t know” is better than mistakes

Support user interaction

Live demo: http://naturalsql.com/

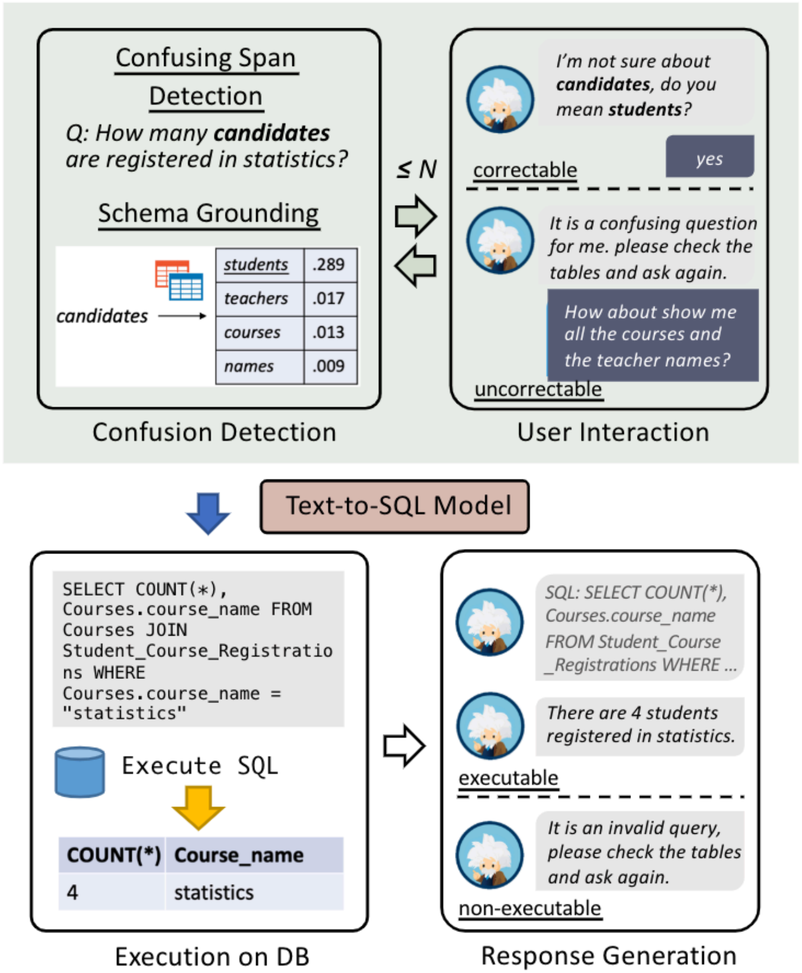
Photon: A Robust Cross-Domain Text-to-SQL System



Live demo: http://naturalsql.com/

Photon: A Robust Cross-Domain Text-to-SQL System

A SOTA neural text-to-SQL parser

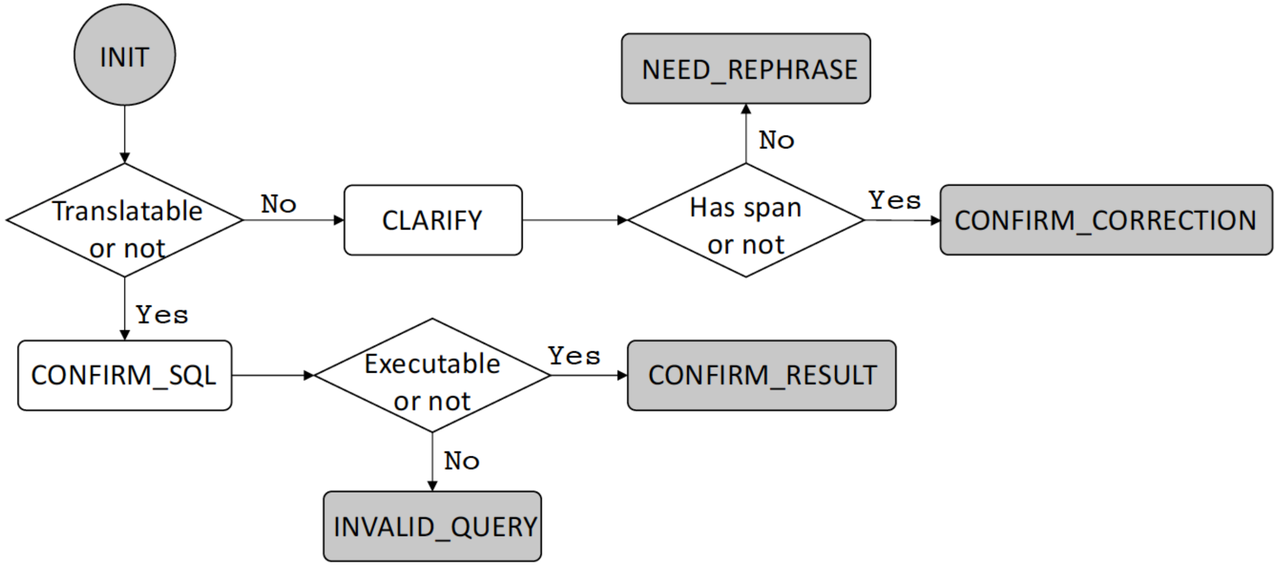


A novel confusion detection approach

Template-based response generation for user interaction

Live demo: http://naturalsql.com/

Interaction Flow



Response Template

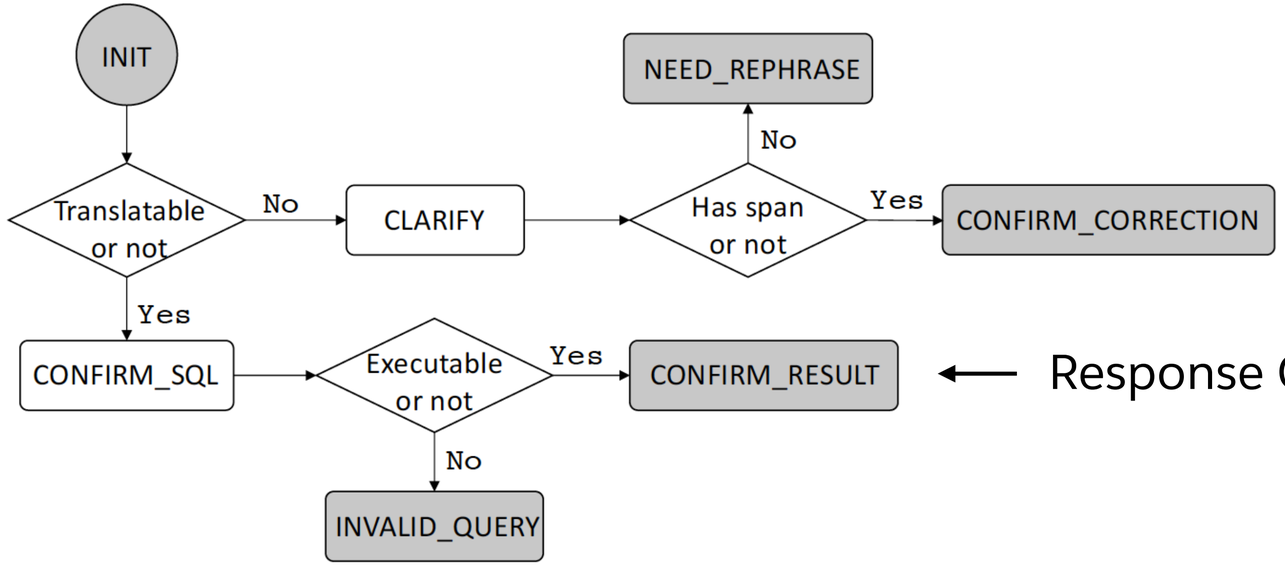
CONFIRM_RESULT	"SQL: {PRED_SQL}. {NL_RESPONSE}"
CONFIRM_CORRECTION	"Sorry, {CONF_TOKENS} is confusing in our scenario, do you mean {CORR_TOKENS}?"
NEED_REPHRASE	"Sorry, it is a confusing question for me, please rephrase your question and ask again."
INVALID_QUERY	"Sorry, it is an invalidate query, please check the table names and associated fields of interest."

Live demo: http://naturalsql.com/

Interaction Flow

Confusion Detection →

Text-to-SQL Parsing →



← Response Generation

Response Template

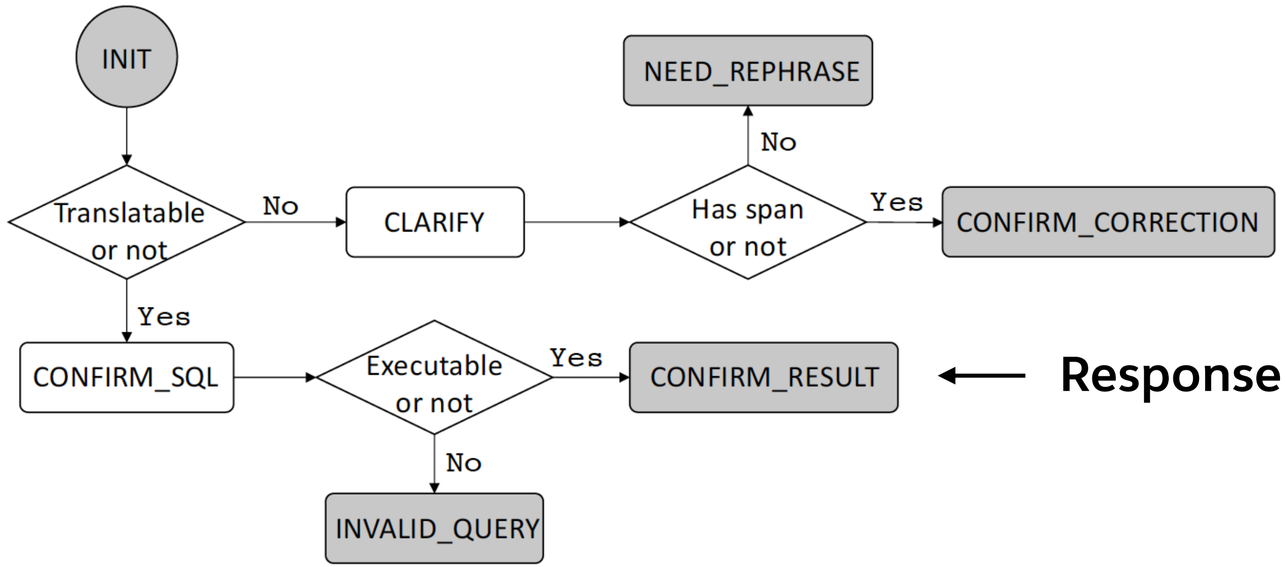
CONFIRM_RESULT	"SQL: {PRED_SQL}. {NL_RESPONSE}"
CONFIRM_CORRECTION	"Sorry, {CONF_TOKENS} is confusing in our scenario, do you mean {CORR_TOKENS}?"
NEED_REPHRASE	"Sorry, it is a confusing question for me, please rephrase your question and ask again."
INVALID_QUERY	"Sorry, it is an invalidate query, please check the table names and associated fields of interest."

Live demo: http://naturalsql.com/

Interaction Flow

Confusion Detection →

Text-to-SQL Parsing →



← **Response Generation**

Response Template

CONFIRM_RESULT	"SQL: {PRED_SQL}. {NL_RESPONSE}"
CONFIRM_CORRECTION	"Sorry, {CONF_TOKENS} is confusing in our scenario, do you mean {CORR_TOKENS}?"
NEED_REPHRASE	"Sorry, it is a confusing question for me, please rephrase your question and ask again."
INVALID_QUERY	"Sorry, it is an invalidate query, please check the table names and associated fields of interest."

Live demo: http://naturalsql.com/

Text-to-SQL Semantic Parsing

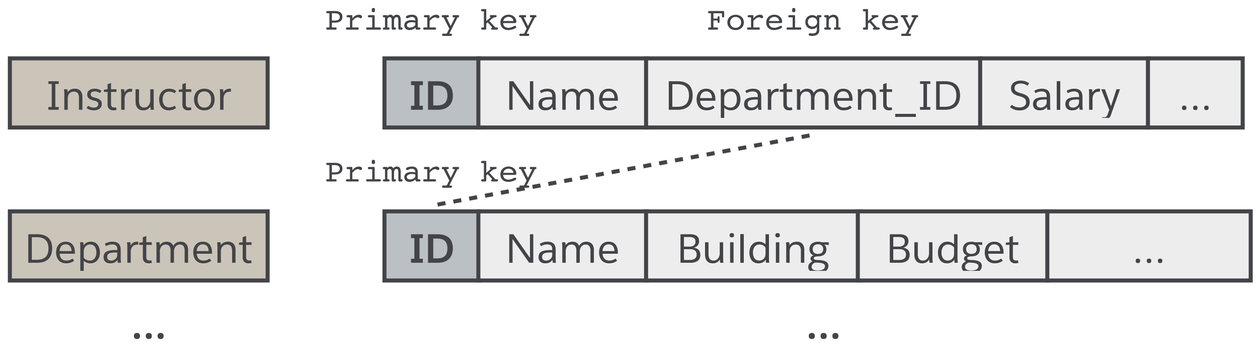
Spider Dataset (Yu et al. 2018)

Expert-annotated, **cross-domain**, **complex** text-to-SQL dataset

	Train	Dev	Test
# DBs	146	20	40
# Examples	8,659	1,034	2,147

Live demo: http://naturalsql.com/

Database



Question What are the name and budget of the departments with average instructor salary above the overall average?

SQL

```
SELECT T2.name, T2.budget
FROM Instructor AS T1 JOIN Department AS T2 ON
T1.Department_ID = T2.ID
GROUP BY T1.Department_ID
HAVING AVG(T1.salary) >
     (SELECT AVG(Salary) FROM Instructor)
```

Text-to-SQL Semantic Parsing

Serialize DB schema

T	Instructor	C	ID	C	Name	C	Department_ID	C	Salary	C	...	T	Department	C	ID	C	...

Live demo: http://naturalsql.com/

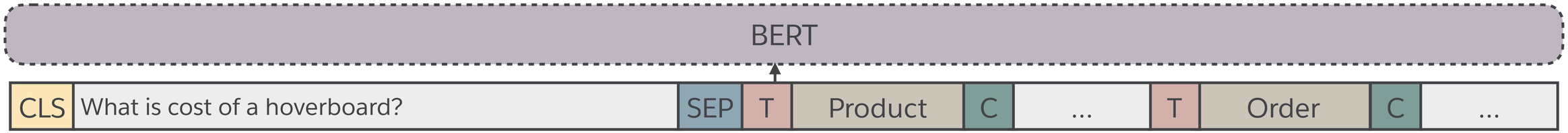
Text-to-SQL Semantic Parsing

Text-table joint encoding

Live demo: http://naturalsql.com/

Text-to-SQL Semantic Parsing

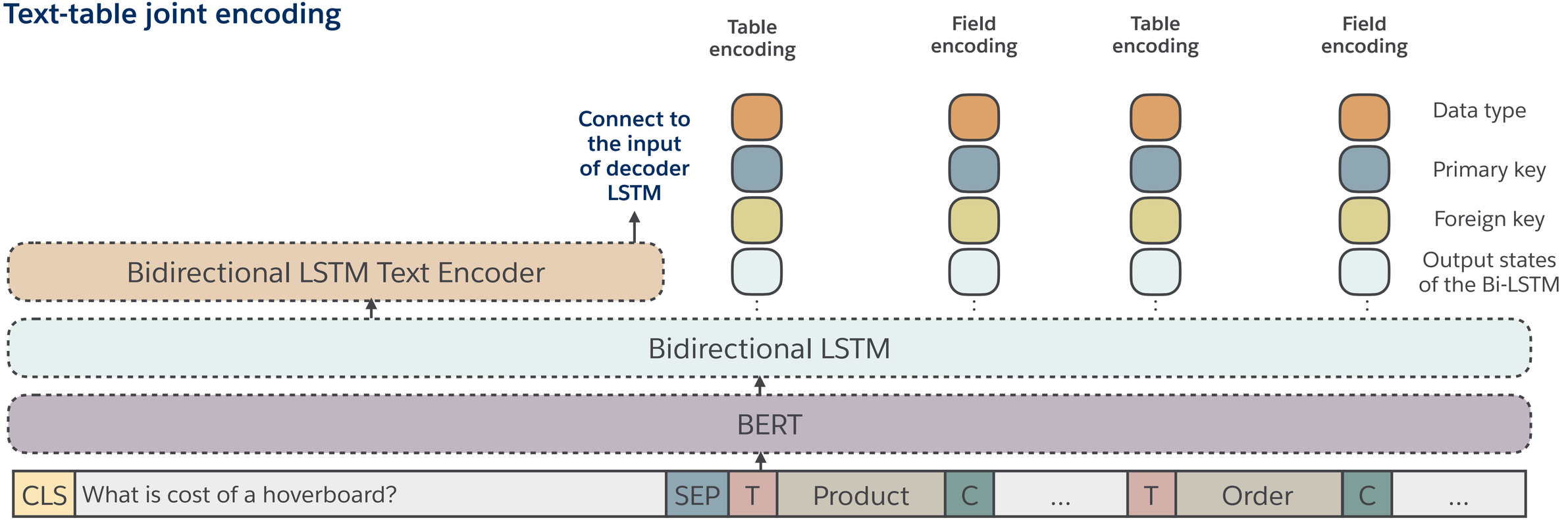
Text-table joint encoding



Live demo: http://naturalsql.com/

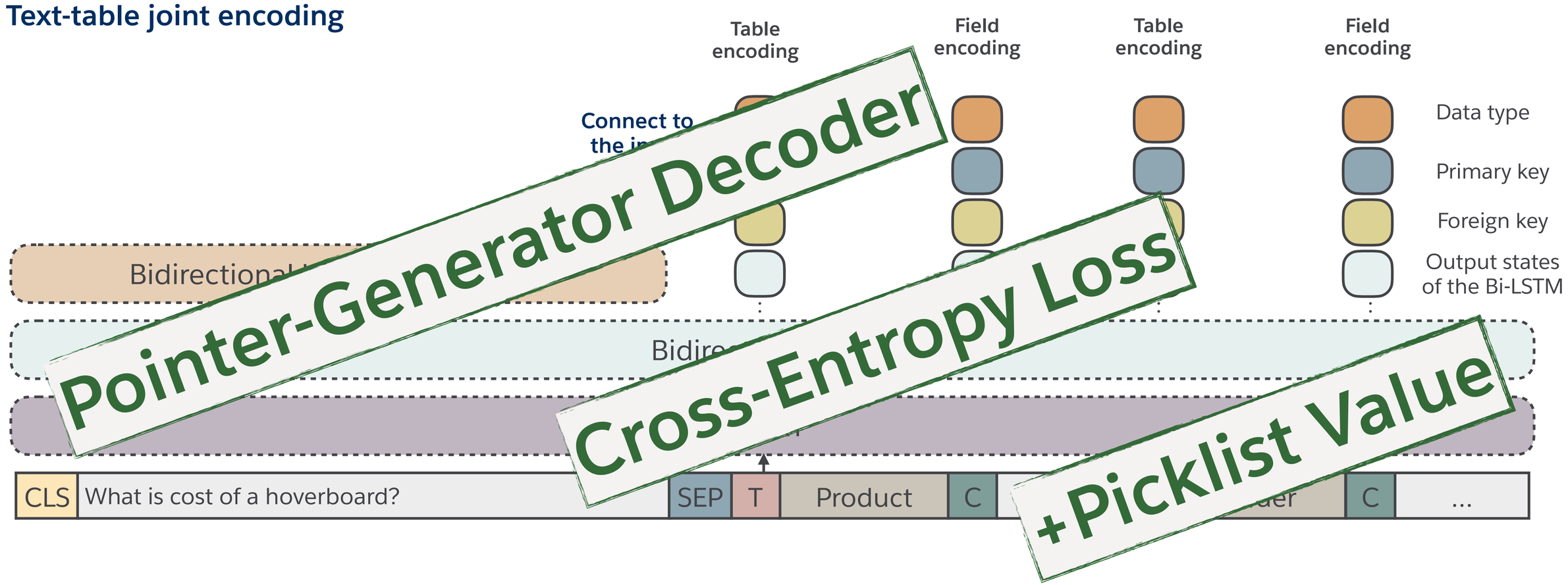
Text-to-SQL Semantic Parsing

Text-table joint encoding



Live demo: http://naturalsql.com/

Text-to-SQL Semantic Parsing



Live demo: http://naturalsql.com/

Text-to-SQL Semantic Parsing Evaluation

Model	EM Acc.
GNN (Bogin et al., 2019a)	40.7
Global-GNN (Bogin et al., 2019b)	52.7
EditSQL + BERT (Zhang et al., 2019)	57.6
GNN+Bertrand-DR† (Kelkar et al., 2020)	57.9
EditSQL+Bertrand-DR† (Kelkar et al., 2020)	58.5
IRNet + BERT (Guo et al., 2019)	61.9
RYANSQL + BERT † (Choi et al., 2020)	66.6
PHOTON	63.2

† denotes unpublished work on arXiv.

Table 3: Experimental results on the Spider Dev set (%). EM Acc. denotes the exact set match accuracy.

Spider leaderboard (May 1st, 2020) https://yale-lily.github.io/spider

Confusion Detection

What is the total?

Show me homes with good schools

How many tourists visited all of the 10 attractions?

Hey, lovely weather

Live demo: http://naturalsql.com/

Confusion Detection

Underspecified

What is the total?

Show me homes with good schools

How many tourists visited all of the 10 attractions?

Hey, lovely weather

Live demo: http://naturalsql.com/

Confusion Detection

What is the total?

Show me homes with good schools

Ambiguous

How many tourists visited all of the 10 attractions?

Hey, lovely weather

Live demo: http://naturalsql.com/

Confusion Detection

What is the total?

Show me homes with good schools

How many tourists visited all of the 10 attractions?

Out-of-scope

Hey, lovely weather

Live demo: http://naturalsql.com/

Confusion Detection

What is the total?

Show me homes with good schools

How many tourists visited all of the 10 attractions?

Hey, lovely weather

Not a query

Live demo: http://naturalsql.com/

Confusion Detection Dataset (UTran-SQL)

Transformation		Original data	Transformed data	Confusing text span
Question	Swap	Q1: How many *conductors* are there? S1: || Conductor_ID || Name || Age || Nationlity|| Year_of_Work ||	Q1: How many **soloists** are there ?	soloists
		Q2: What are the maximum and minimum values of *area codes* ? S2: || Vote_ID || Phone_Number || Area_Code || State || Created ||	Q2: What are the maximum and minimum values of **types**?	types
	Drop	Q1: How many *countries* exist? S1: || CoutryId || CountryName || Continent ||	Q1: How many are there?	WHOLE SENTENCE
		Q2: What is the *official language* spoken in the country whose head of state is Beatrix? S2: || CountryCode || HeadOfState || Captital || Language || IsOfficial || Percentage ||	Q2: What are the people in the country where Beatrix is located?	WHOLE SENTENCE
Schema Drop		Q1: How much *surface area* do the countires in the Carribean cover together? S1: || Name ||Continent || Region || *SurfaceArea* || Population || LifeExpectancy ||	S1: || Name ||Continent || Region || Population || LifeExpectancy ||	surface area
		Q2: Find the name and *age* of the visitor who bought the most tickets at once. S2: ||Customer_ID||Name||Level_of_membership|| *Age* ||	S2: ||Customer_ID||Name||Level_of_membership||	age

Table 5: Examples of question-side and schema-side transformations for generating training data for untranslatable question detection. Let Q denote the question and S denote the schema. For each transformation, we provide two examples, i.e., (Q1, S1) and (Q2, S2). The italic and bold fonts highlight phrases before and after transformations.

Confusion Detection Dataset (UTran-SQL)

Transformation		Original data	Transformed data	Confusing text span
Question	Swap	Q1: How many *conductors* are there? S1: || Conductor_ID || N…	Q1: How many **soloists** are there ? …ge || Nationlity|| Year_of_Work ||	soloists
		Q2: What are the maximum a… values of *area code*…	Q2: What are the maximum and minimum values **types**? …ber || Area_Code || State || Created ||	types
		Q1: … S1: || CoutryId || CountryName || Continent ||	Q1: How many are there?	WHOLE SENTE…E
		…*official language* spoken in the …y whose head of state is Beatrix? S2: || CountryCode || HeadOfState || Captital || Language || IsOfficial || Pere…	Q2: What are the people in the country where Beatrix is located?	
Schema Drop		Q1: How much *surface area* do the countires in the Carribean cover to… S1: || Name ||Continent || Region || *SurfaceArea* || Population || LifeExpectancy ||	S1: || Nam… Lif…	surface area
		Q2: Find the name and *age* of the visitor… S2: ||Customer_ID||Name||Level_of_mem… *Age* ||	…ID||Name||Level_of_membership||	age

Back-Translation

Adversarial Filtering

Table 5: Examples of question-side and schema-side transformations for generating training data for untranslatable question detection. Let Q denote the question and S denote the schema. For each transformation, we provide two examples, i.e., (Q1, S1) and (Q2, S2). The italic and bold fonts highlight phrases before and after transformations.

UTran-SQL Data Statistics

	Spider		$\text{Spider}_{\text{UTran}}$	
	Train	Dev	Train	Dev
# Q	8,659	1,034	13,392	1,631
# UTran Q	0	0	4,733	597
# Schema	146	20	918	112

Table 1: Data split of Spider and $\text{Spider}_{\text{UTran}}$. Q represents the all the questions, UTran Q represents the untranslatable questions.

Live demo: http://naturalsql.com/

UTran-SQL Data Statistics

	Spider		$\text{Spider}_{\text{UTran}}$	
	Train	Dev	Train	Dev
# Q	8,659	1,034	13,392	1,631
# UTran Q	0	0	4,733	597
# Schema	146	20	918	

Table 1: Data split of Spider a… …ep-resents the all the que… …presents the untranslatable question…

Strong Supervision

Live demo: http://naturalsql.com/

Confusion Detection Model

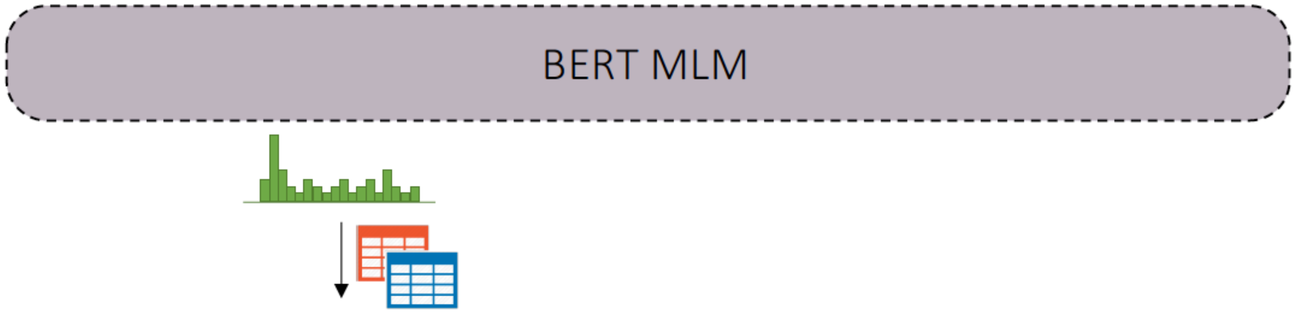
- **Translatability prediction:** binary classification based on `[CLS]` representation of the BERT text-table encoder
- **Confusion span detection:** predicting the start and end token indices

Live demo: http://naturalsql.com/

Question Rephrasing Model

Original input: How many **candidates** are registered in statistics ?

Processed input: How many [MASK] are registered in statistics ? [TABLE NAMES]



Table&Column Names

students	.289
teachers	.017
courses	.013
names	.009
student details	.008

...

System: ***candidates*** *is confusing here, do you mean* ***students****?*

Live demo: http://naturalsql.com/

Limitations

- We assume only one confusion span per sentence
- We assume the confusion span is a column mention
- The transformation rules can introduce errors
- Alternatives for confusion detection in text-to-SQL are worth exploring
  - Yao et al. 2019
  - Yao et al. 2020 (concurrent)
- Limited set of user actions are considered

Live demo: http://naturalsql.com/

Additional Demo Features

- Upload your own DBs for testing
- Effective DB schema visualization and data browsing
- Rate your experience and provide feedback

Live demo: http://naturalsql.com/

Related Work

ATIS Corpus collection

Hemphill et al. 1990
Dahl et al. 1994

Learning logical-form based semantic parsers for NLIDBs

Zelle and Mooney 1996
Popescu et al. 2003
Zettlemoyer and Collins 2005

Neural networks widely adopted in NLP

Sutskever et al. 2014
Bahdanau et al. 2015

Seq2Seq-style neural semantic parsing

Dong and Lapata 2016

WikiSQL: a large-scale, cross-domain text-to-SQL corpora

Zhong et al. 2017

TypeSQL, column attention, sketch-based, execution guided, RL, meta-learning

Xu et al. 2017
Dong and Lapata 2018
Wang et al. 2018
Yu et al. 2018a

Yu et al. 2018b

Spider: expert-annotated, large-scale, cross-domain, complex

LM pre-training: BERT

Devlin et al. 2018

Table-Aware BERT Encoder, surpassed human-performance on WikiSQL

Hwang et al. 2019

Syntax-guided, GNN, schema linking, SemQL

Yu et al. 2018c
Bogin et al. 2019
Shin et al. 2019
Guo et al. 2019
Wang et al. 2020
...

Live demo: http://naturalsql.com/

Related Work

- Most state-of-the-art cross-domain, complex text-to-SQL semantic parsers are not well packaged for user test and interaction

- Most existing NLIDB systems are DB-specific or non-interactive

Live demo: http://naturalsql.com/

Live Demo: http://naturalsql.com/

Join us at the Q&A sessions

Tuesday July 7, 2020 UTC+0 17:00-17:45
Tuesday July 7, 2020 UTC+0 20:00-20:45