THE 37TH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE
FEBRUARY 7-14, 2023 · WASHINGTON, DC, USA
WALTER E. WASHINGTON CONVENTION CENTER
AAAI-23

MultiSpider: Towards Benchmarking Multilingual Text-to-SQL Semantic Parsing

Longxu Dou[1], Yan Gao[2], Mingyang Pan[1], Dingzirui Wang[1],

Wanxiang Che[1], Dechen Zhan[1], Jian-Guang Lou[2]

[1]Harbin Institute of Technology [2]Microsoft Research Asia

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

SCIR Research Center for Social Computing and Information Retrieval 理解语言 认知社会

Microsoft Research 微软亚洲研究院

Work done during the internship of Microsoft Research Asia

Overview

[Benchmark]

MultiSpider: the largest **text-to-SQL multilingual** dataset covering 7 languages.

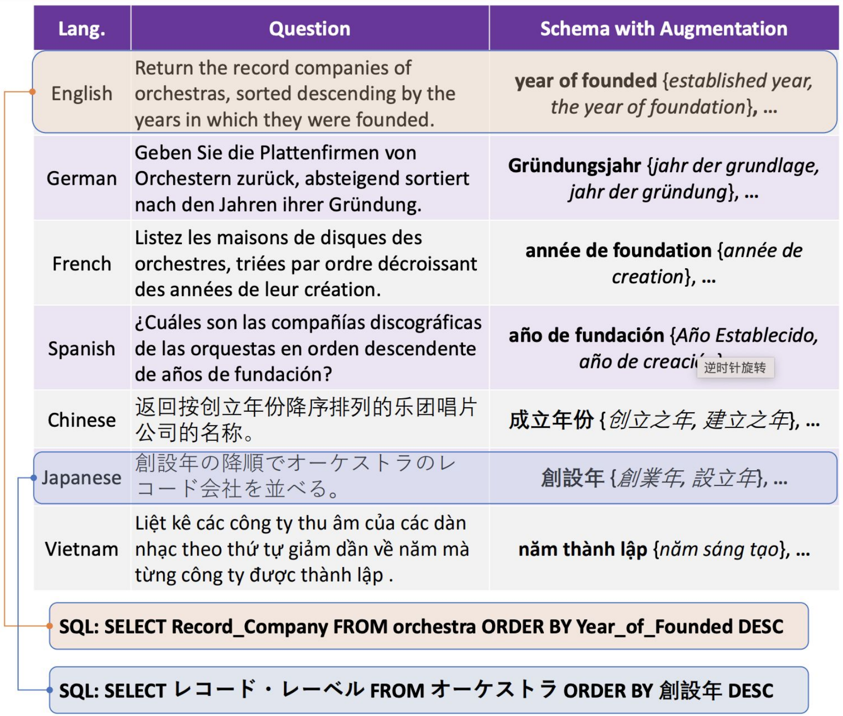
[Analysis&Experiments]

Identify the specific **lexical** challenge and **structural** challenge of MultiSpider.

[Data Augmentation]

SAVe: a data-augmentation method from the **perspective of schema**.

MultiSpider Benchmark



Lang.	Question	Schema with Augmentation
English	Return the record companies of orchestras, sorted descending by the years in which they were founded.	**year of founded** {*established year, the year of foundation*}, …
German	Geben Sie die Plattenfirmen von Orchestern zurück, absteigend sortiert nach den Jahren ihrer Gründung.	**Gründungsjahr** {*jahr der grundlage, jahr der gründung*}, …
French	Listez les maisons de disques des orchestres, triées par ordre décroissant des années de leur création.	**année de foundation** {*année de creation*}, …
Spanish	¿Cuáles son las compañías discográficas de las orquestas en orden descendente de años de fundación?	**año de fundación** {*Año Establecido, año de creaci*逆时针旋转
Chinese	返回按创立年份降序排列的乐团唱片公司的名称。	**成立年份** {*创立之年, 建立之年*}, …
Japanese	創設年の降順でオーケストラのレコード会社を並べる。	**創設年** {*創業年, 設立年*}, …
Vietnam	Liệt kê các công ty thu âm của các dàn nhạc theo thứ tự giảm dần về năm mà từng công ty được thành lập .	**năm thành lập** {*năm sáng tạo*}, …

SQL: SELECT Record_Company FROM orchestra ORDER BY Year_of_Founded DESC

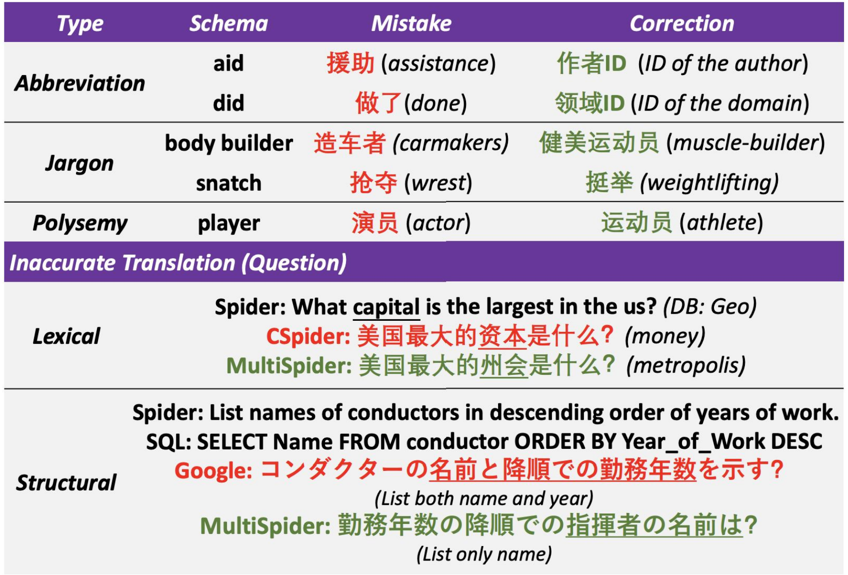
SQL: SELECT レコード・レーベル FROM オーケストラ ORDER BY 創設年 DESC

- Built on the top of **challenging** multi-table cross-database English Spider.

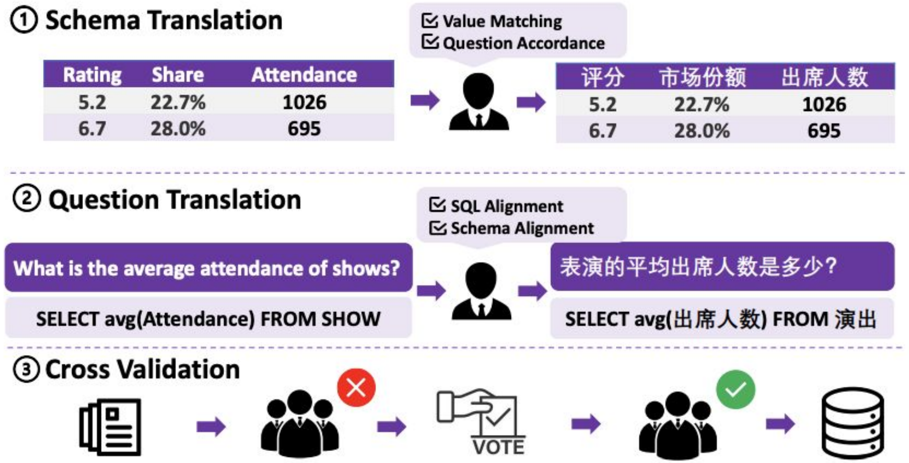
- **Largest** and **high-quality** multilingual text-to-SQL dataset, including **seven** mainstream languages.

- Translate both **question&schema**.

Benchmark Construction



Type	*Schema*	*Mistake*	*Correction*
Abbreviation	**aid**	援助 (*assistance*)	**作者ID** (*ID of the author*)
	did	做了 (*done*)	**领域ID** (*ID of the domain*)
Jargon	**body builder**	造车者 (*carmakers*)	健美运动员 (*muscle-builder*)
	snatch	抢夺 (*wrest*)	挺举 (*weightlifting*)
Polysemy	**player**	演员 (*actor*)	运动员 (*athlete*)
Inaccurate Translation (Question)			
Lexical	**Spider: What capital is the largest in the us?** *(DB: Geo)* **CSpider:** 美国最大的资本是什么? *(money)* **MultiSpider:** 美国最大的州会是什么? *(metropolis)*		
Structural	**Spider: List names of conductors in descending order of years of work.** **SQL: SELECT Name FROM conductor ORDER BY Year_of_Work DESC** **Google:** コンダクターの名前と降順での勤務年数を示す? *(List both name and year)* **MultiSpider:** 勤務年数の降順での指揮者の名前は? *(List only name)*		



① Schema Translation

☑ Value Matching
☑ Question Accordance

Rating	Share	Attendance
5.2	22.7%	1026
6.7	28.0%	695

评分	市场份额	出席人数
5.2	22.7%	1026
6.7	28.0%	695

② Question Translation

☑ SQL Alignment
☑ Schema Alignment

What is the average attendance of shows?

SELECT avg(Attendance) FROM SHOW

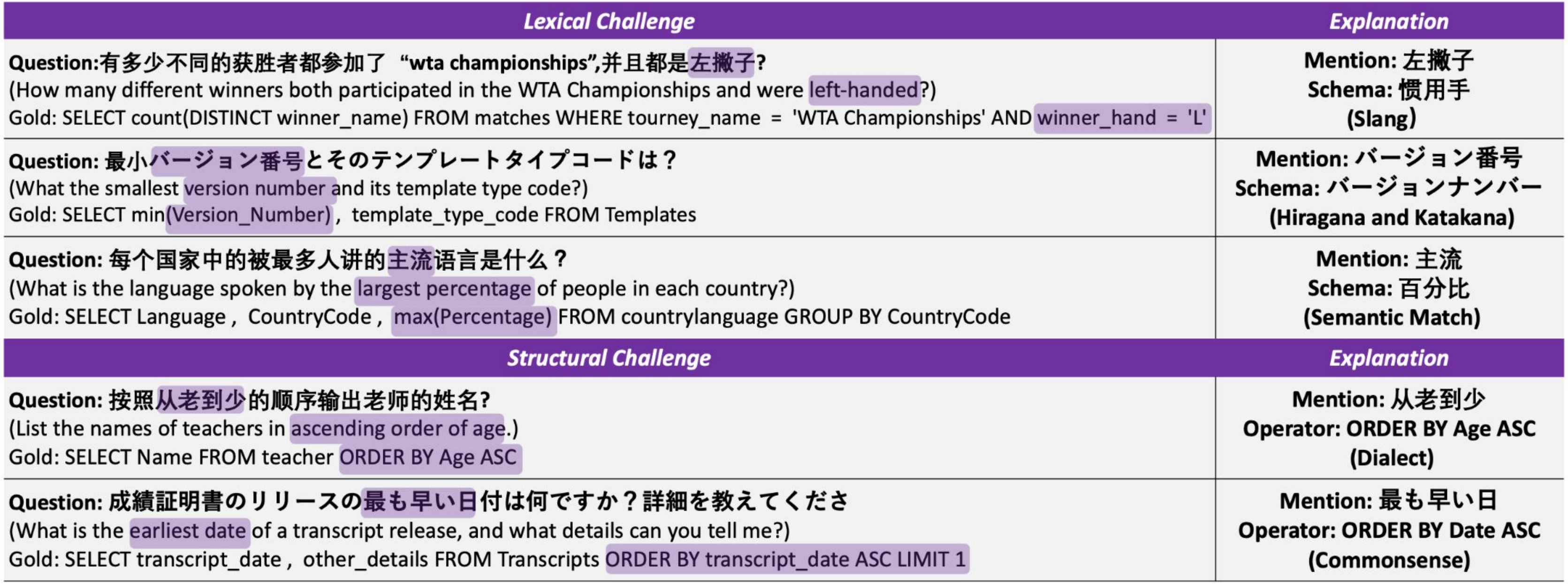
表演的平均出席人数是多少?

SELECT avg(出席人数) FROM 演出

③ Cross Validation

- To ensure the dataset quality, we identify five **typical translation mistakes**.
- Organize the construction pipeline consisting of **multi-round translation**.

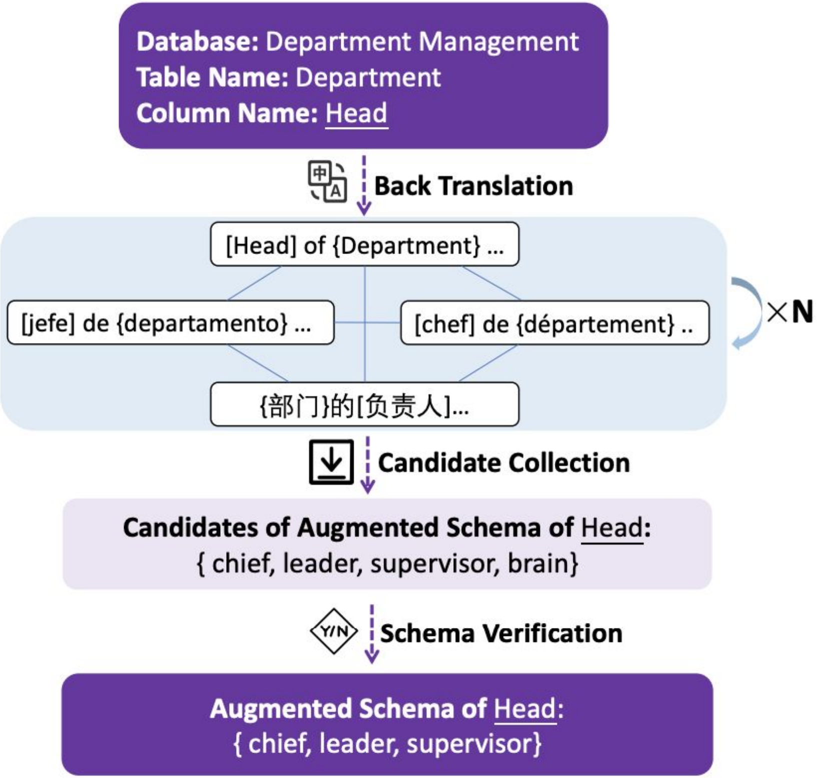
Analysis: More Challenging



Lexical Challenge	*Explanation*
Question:有多少不同的获胜者都参加了 “wta championships”,并且都是左撇子? (How many different winners both participated in the WTA Championships and were left-handed?) Gold: SELECT count(DISTINCT winner_name) FROM matches WHERE tourney_name = 'WTA Championships' AND winner_hand = 'L'	Mention: 左撇子 Schema: 惯用手 (Slang)
Question: 最小バージョン番号とそのテンプレートタイプコードは？ (What the smallest version number and its template type code?) Gold: SELECT min(Version_Number) , template_type_code FROM Templates	Mention: バージョン番号 Schema: バージョンナンバー (Hiragana and Katakana)
Question: 每个国家中的被最多人讲的主流语言是什么？ (What is the language spoken by the largest percentage of people in each country?) Gold: SELECT Language , CountryCode , max(Percentage) FROM countrylanguage GROUP BY CountryCode	Mention: 主流 Schema: 百分比 (Semantic Match)
Structural Challenge	***Explanation***
Question: 按照从老到少的顺序输出老师的姓名? (List the names of teachers in ascending order of age.) Gold: SELECT Name FROM teacher ORDER BY Age ASC	Mention: 从老到少 Operator: ORDER BY Age ASC (Dialect)
Question: 成績証明書のリリースの最も早い日付は何ですか？詳細を教えてくださ (What is the earliest date of a transcript release, and what details can you tell me?) Gold: SELECT transcript_date , other_details FROM Transcripts ORDER BY transcript_date ASC LIMIT 1	Mention: 最も早い日 Operator: ORDER BY Date ASC (Commonsense)

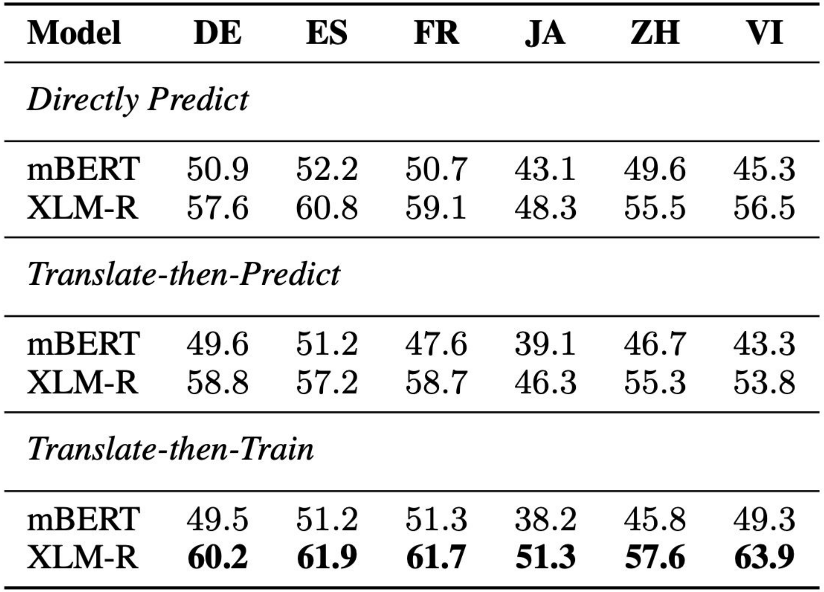
- The **specific language properties** like Hiragana and Katakana (Japanese).
- The **morphologically rich** language (German and French).
- The **dialect and slang sayings** require further commonsense reasoning.

SAVe: Schema Augmentation


Database: Department Management
Table Name: Department
Column Name: Head
Back Translation
[Head] of {Department} ...
[jefe] de {departamento} ...
[chef] de {département} ..
{部门}的[负责人]...
×N
Candidate Collection
Candidates of Augmented Schema of Head:
{ chief, leader, supervisor, brain}
Y/N
Schema Verification
Augmented Schema of Head:
{ chief, leader, supervisor}

- Schema Augmentation-with-Verification
- Back Translation with Machine Translation tools (e.g., Google NMT, M100)
- Schema Verification with Natural Language Inference (e.g., XNLI)

Experiments: Zero-shot Setting



Model	**DE**	**ES**	**FR**	**JA**	**ZH**	**VI**
Directly Predict						
mBERT	50.9	52.2	50.7	43.1	49.6	45.3
XLM-R	57.6	60.8	59.1	48.3	55.5	56.5
Translate-then-Predict						
mBERT	49.6	51.2	47.6	39.1	46.7	43.3
XLM-R	58.8	57.2	58.7	46.3	55.3	53.8
Translate-then-Train						
mBERT	49.5	51.2	51.3	38.2	45.8	49.3
XLM-R	**60.2**	**61.9**	**61.7**	**51.3**	**57.6**	**63.9**

- **Better model** enables better zero-shot transfer (XLM-R > mBERT).

- Directly predict receives better performance about 1.6% (MT creates mistakes) compared with translate-then-predict.

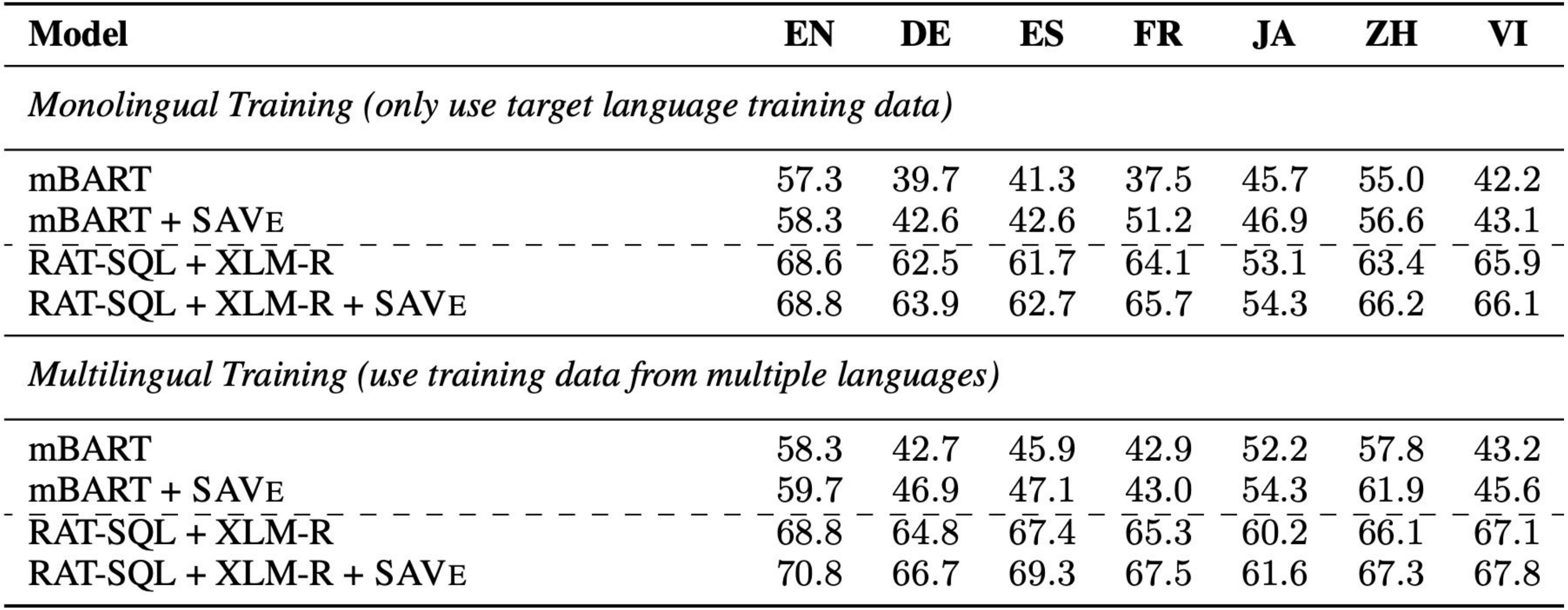
- Strong PLM with Strong MT yields **promising zero-shot** performance.

- Directly Predict (Train: English, Test: Target)

- Translate-then-Predict (Train: English, Test: Target to English)

- Translate-then-Train (Train: English to Target, Test: Target)

Experiments: Monolingual & Multilingual Setting

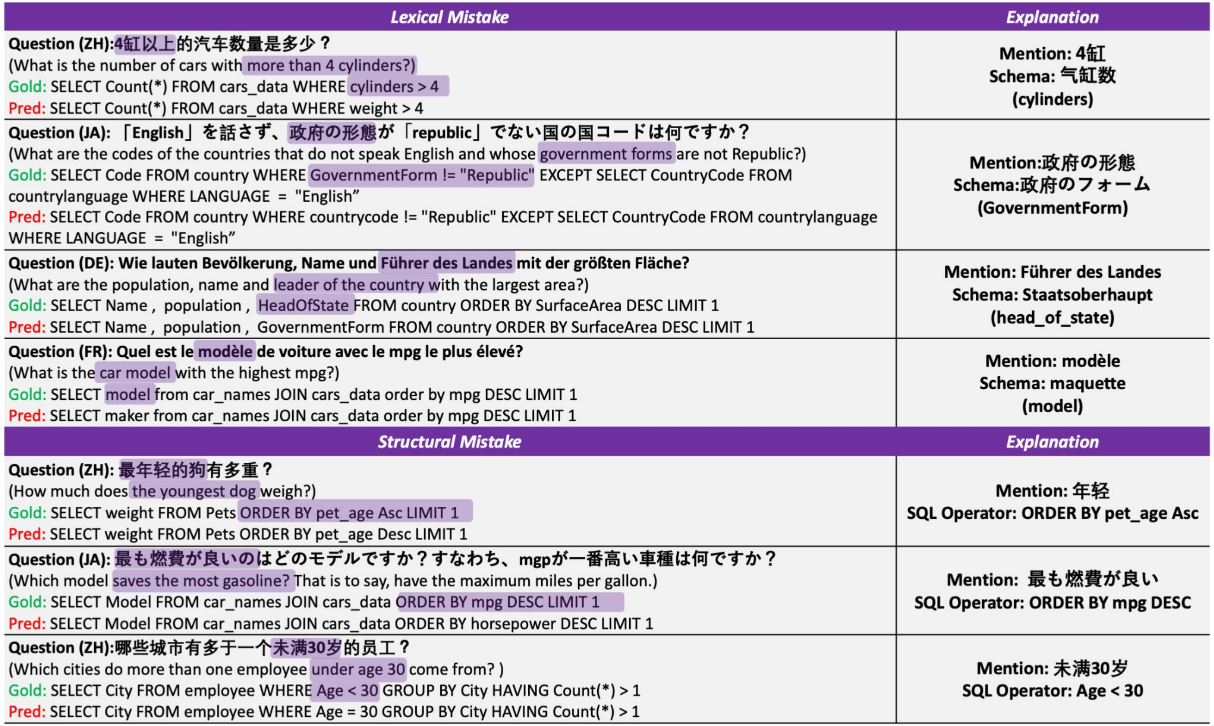


Model	EN	DE	ES	FR	JA	ZH	VI
Monolingual Training (only use target language training data)							
mBART	57.3	39.7	41.3	37.5	45.7	55.0	42.2
mBART + SAVE	58.3	42.6	42.6	51.2	46.9	56.6	43.1
RAT-SQL + XLM-R	68.6	62.5	61.7	64.1	53.1	63.4	65.9
RAT-SQL + XLM-R + SAVE	68.8	63.9	62.7	65.7	54.3	66.2	66.1
Multilingual Training (use training data from multiple languages)							
mBART	58.3	42.7	45.9	42.9	52.2	57.8	43.2
mBART + SAVE	59.7	46.9	47.1	43.0	54.3	61.9	45.6
RAT-SQL + XLM-R	68.8	64.8	67.4	65.3	60.2	66.1	67.1
RAT-SQL + XLM-R + SAVE	70.8	66.7	69.3	67.5	61.6	67.3	67.8

- The performance of **Japanese** is significantly behind other languages.
- The absolute **drop** of accuracy in **non-English** languages is about **6.1%**.
- **SAVe** significantly **improves** the non-English languages **(1.4%-1.9%)**.

Discussion

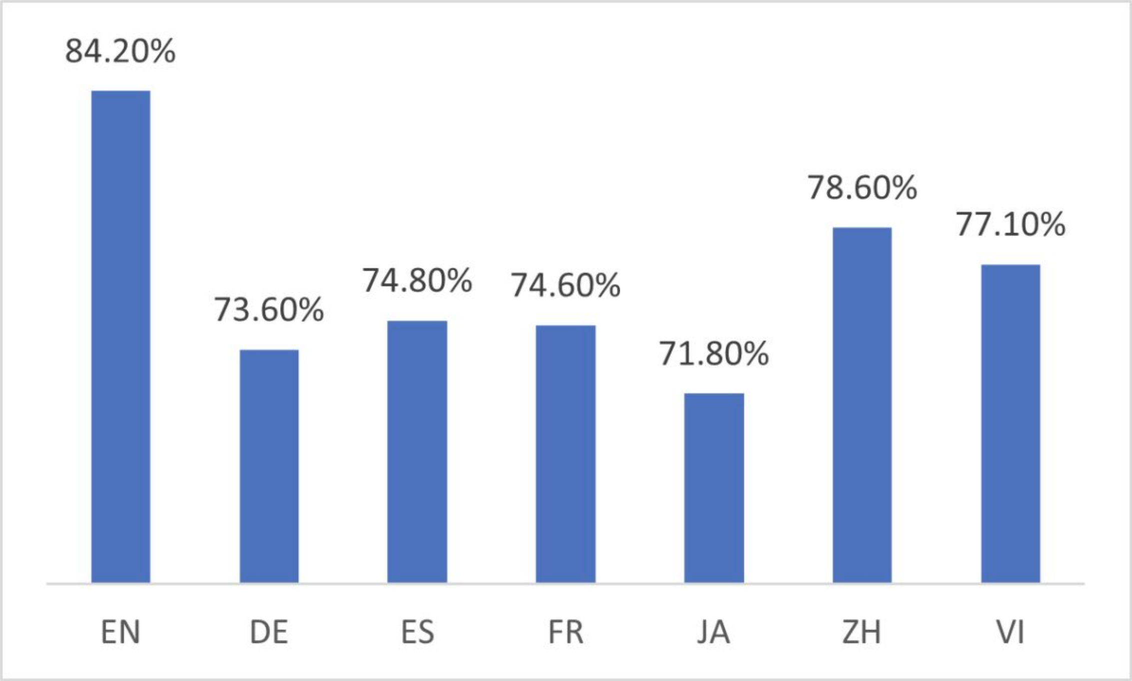
- What causes the performance drop in non-English languages?
  - **Specific language properties** and the **dialect sayings** lead to the performance drop in non-English languages.



Lexical Mistake	*Explanation*
Question (ZH):4缸以上的汽车数量是多少？ (What is the number of cars with more than 4 cylinders?) Gold: SELECT Count(*) FROM cars_data WHERE cylinders > 4 Pred: SELECT Count(*) FROM cars_data WHERE weight > 4	**Mention: 4缸** **Schema: 气缸数** **(cylinders)**
Question (JA): 「English」を話さず、政府の形態が「republic」でない国の国コードは何ですか？ (What are the codes of the countries that do not speak English and whose government forms are not Republic?) Gold: SELECT Code FROM country WHERE GovernmentForm != "Republic" EXCEPT SELECT CountryCode FROM countrylanguage WHERE LANGUAGE = "English" Pred: SELECT Code FROM country WHERE countrycode != "Republic" EXCEPT SELECT CountryCode FROM countrylanguage WHERE LANGUAGE = "English"	**Mention:政府の形態** **Schema:政府のフォーム** **(GovernmentForm)**
Question (DE): Wie lauten Bevölkerung, Name und Führer des Landes mit der größten Fläche? (What are the population, name and leader of the country with the largest area?) Gold: SELECT Name , population , HeadOfState FROM country ORDER BY SurfaceArea DESC LIMIT 1 Pred: SELECT Name , population , GovernmentForm FROM country ORDER BY SurfaceArea DESC LIMIT 1	**Mention: Führer des Landes** **Schema: Staatsoberhaupt** **(head_of_state)**
Question (FR): Quel est le modèle de voiture avec le mpg le plus élevé? (What is the car model with the highest mpg?) Gold: SELECT model from car_names JOIN cars_data order by mpg DESC LIMIT 1 Pred: SELECT maker from car_names JOIN cars_data order by mpg DESC LIMIT 1	**Mention: modèle** **Schema: maquette** **(model)**
Structural Mistake	***Explanation***
Question (ZH): 最年轻的狗有多重？ (How much does the youngest dog weigh?) Gold: SELECT weight FROM Pets ORDER BY pet_age Asc LIMIT 1 Pred: SELECT weight FROM Pets ORDER BY pet_age Desc LIMIT 1	**Mention: 年轻** **SQL Operator: ORDER BY pet_age Asc**
Question (JA): 最も燃費が良いのはどのモデルですか？すなわち、mgpが一番高い車種は何ですか？ (Which model saves the most gasoline? That is to say, have the maximum miles per gallon.) Gold: SELECT Model FROM car_names JOIN cars_data ORDER BY mpg DESC LIMIT 1 Pred: SELECT Model FROM car_names JOIN cars_data ORDER BY horsepower DESC LIMIT 1	**Mention: 最も燃費が良い** **SQL Operator: ORDER BY mpg DESC**
Question (ZH):哪些城市有多于一个未满30岁的员工？ (Which cities do more than one employee under age 30 come from?) Gold: SELECT City FROM employee WHERE Age < 30 GROUP BY City HAVING Count(*) > 1 Pred: SELECT City FROM employee WHERE Age = 30 GROUP BY City HAVING Count(*) > 1	**Mention: 未满30岁** **SQL Operator: Age < 30**

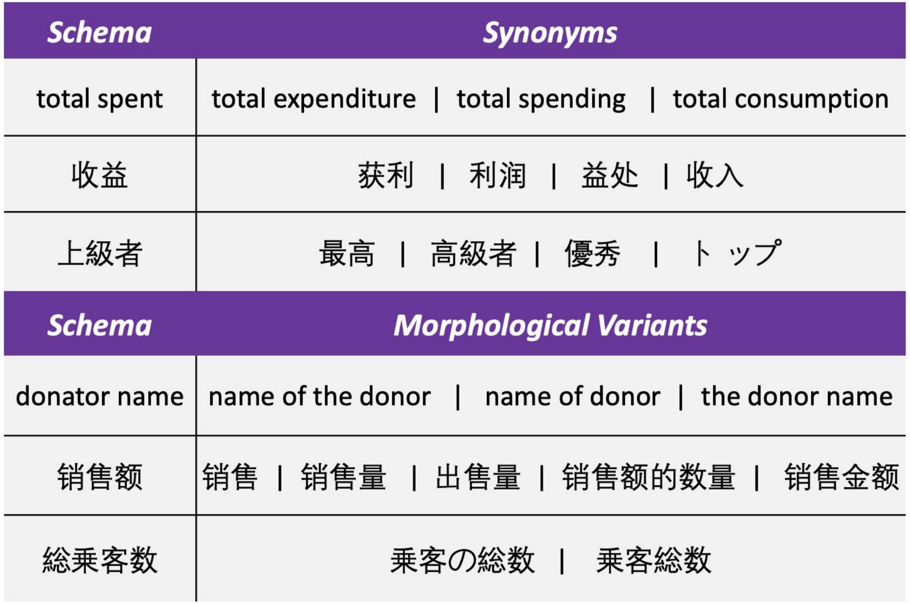
Discussion

- What causes the performance drop in non-English languages?
  - **Schema-linking** becomes more **challenging** in non-English languages.



Discussion

- How schema augmentation SAVE improves the model?
  - **Synonyms** that semantically identical with the original schema but with different lemmas.
  - **Morphological variants** that change the forms of schema syntactically.



Schema	*Synonyms*
total spent	total expenditure | total spending | total consumption
收益	获利 | 利润 | 益处 | 收入
上級者	最高 | 高級者 | 優秀 | ト ップ

Schema	*Morphological Variants*
donator name	name of the donor | name of donor | the donor name
销售额	销售 | 销售量 | 出售量 | 销售额的数量 | 销售金额
総乗客数	乗客の総数 | 乗客総数

Future Work

(1) Developing a multilingual text-to-SQL system and apply it in the real **globalization scenario**.

(2) Leveraging better pretrained model and advancing **architecture design** to address the lexical challenge and structural challenge in multilingual settings.

(3) **Expanding SAVe** to other table-related task (e.g., TabFact) and further improve the schema verification accuracy.

THE 37TH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE
FEBRUARY 7-14, 2023 · WASHINGTON, DC, USA
WALTER E. WASHINGTON CONVENTION CENTER
AAAI-23

Thanks!

Longxu Dou[1], Yan Gao[2], Mingyang Pan[1], Dingzirui Wang[1], Wanxiang Che[1], Dechen Zhan[1], Jian-Guang Lou[2]

[1]Harbin Institute of Technology [2]Microsoft Research Asia

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

SCIR Research Center for Social Computing and Information Retrieval 理解语言 认知社会

Microsoft Research 微软亚洲研究院

Paper/Slides/Code in https://longxudou.github.io/