CCKS-2021

基于深度学习的自然语言语义解析

陈 波

2021-11-04

中文信息处理实验室-让机器理解中文
Chinese Information Processing Laboratory

中国科学院软件研究所 ISCAS
Institute of Software, Chinese Academy of Sciences

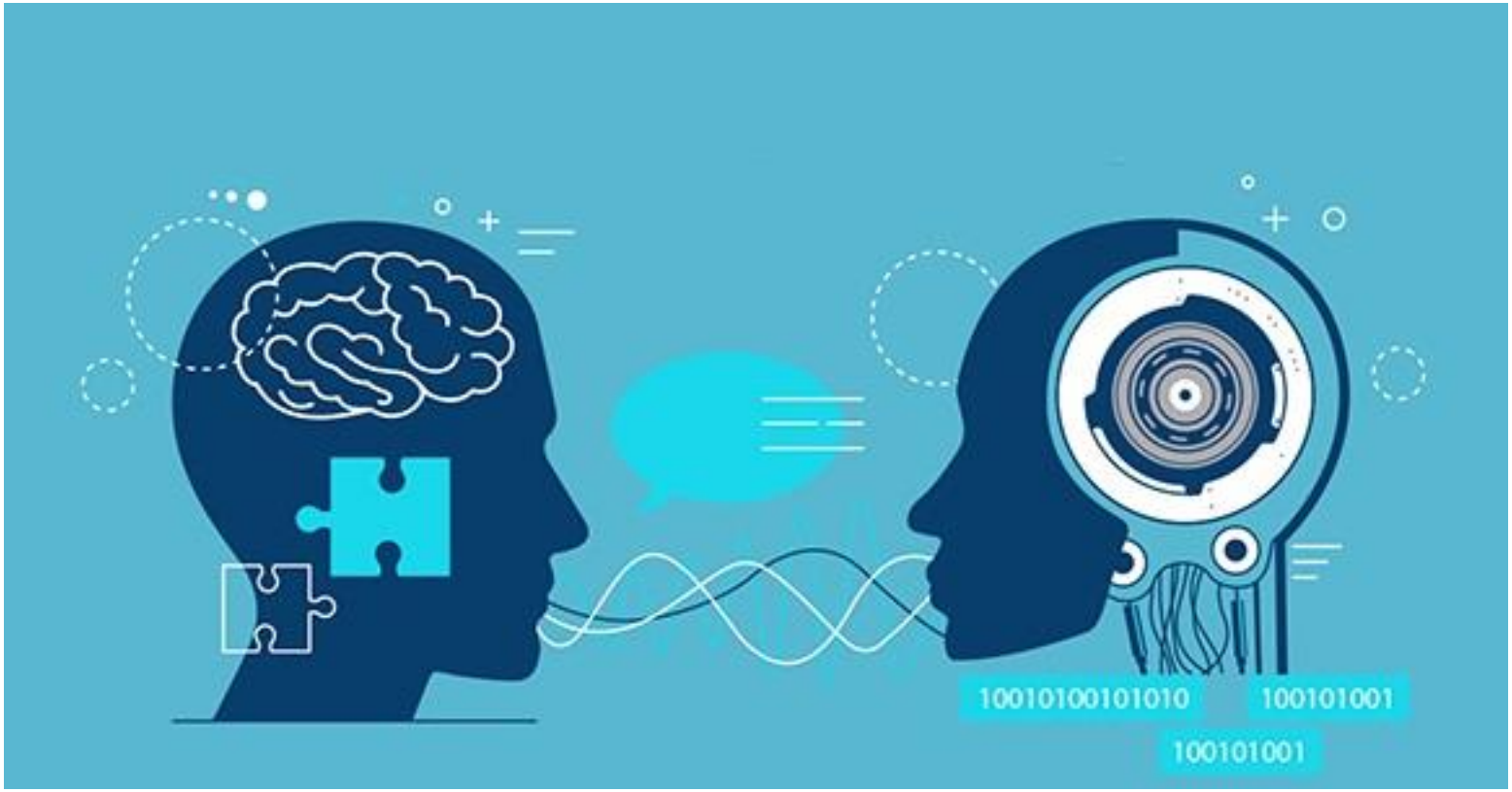
大纲

- 语义解析简介
  - 任务简介
  - 发展历程
- 基于深度学习的语义解析方法
  - Seq2Seq、Seq2Tree、Seq2Action
  - Constrained Decoding
- 基于预训练的语义解析方法
  - 预训练方法在Text-to-SQL任务上的应用
  - PLMs with Constrained Decoding
- 总结与展望

大纲

- 语义解析简介
  - 任务简介
  - 发展历程
- 基于深度学习的语义解析方法
  - Seq2Seq、Seq2Tree、Seq2Action
  - Constrained Decoding
- 基于预训练的语义解析方法
  - 预训练方法在Text-to-SQL任务上的应用
  - PLMs with Constrained Decoding
- 总结与展望

我们憧憬的人机对话



现实中的人机对话


iPad
20:03
82%
“怎么才能上清华大学”
轻按以编辑
好的，下面就是“清华大学”，离你相当远：
北京市海淀区清华大学
北京市海淀区清华大学
清华大学高1楼
华来照相馆
北京清华正大商贸公司
清华图片社
照澜院百货东店
中国工商银行
中国邮政储蓄银行
清华大学高2楼
清华园街道社会保障事务所
清华澜园百货商店
清华大学
中国建设银行
清华澜园超市
北京银行
照澜院商业楼
学明眼镜
中国移动
法律信息
高德地图

自然语言理解（NLU）

- #NLP太难了# （#NLU太难了#）
  - 语言的表达多样性

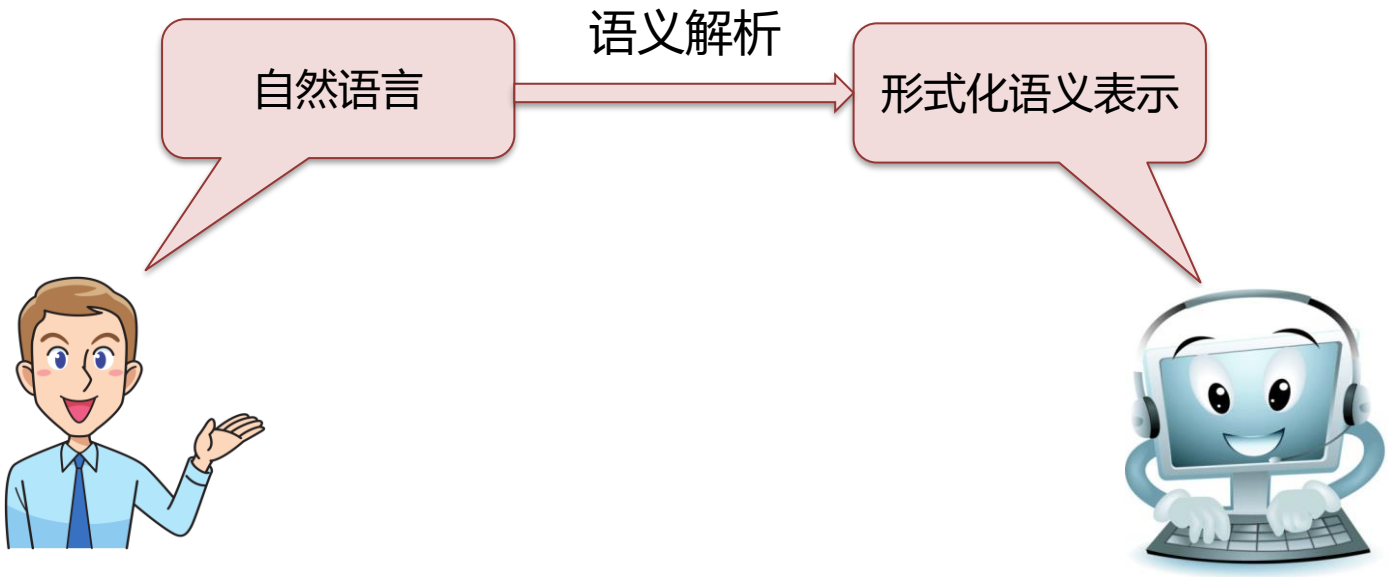
    北京的人口
    住在北京的人有多少
    北京的居民人数
    北京人数量
    ...

  - 语言的歧义性

  - 需要知识

语义解析（Semantic Parsing）

- 语义解析是实现自然语言理解的关键技术之一，目的在于建立自然语言到计算机可以理解的形式化语义表示的映射



语义解析任务定义

- 将自然语言句子转换成计算机可识别的、可计算的、完全的语义表示，如lambda-表达式、SQL、语义图等

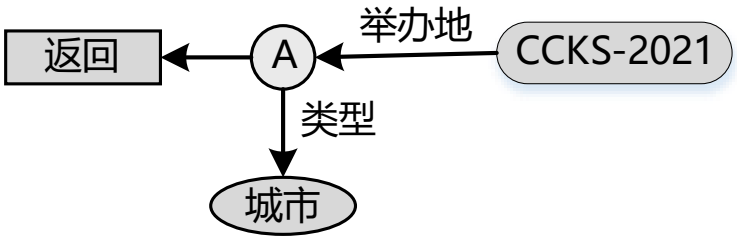
句子：CCKS-2021在哪个城市举办?

语义解析

lambda-表达式：$\lambda x.$城市$(x) \wedge$举办地$(\text{CCKS}-2021，x)$

SQL：SELECT 举办地 FROM 会议 WHERE 会议名=CCKS-2021;

语义图：



John M. Zelle and Raymond J. Mooney. Learning to parse database queries using inductive logic programming. AAAI-1996.

语义解析的任务场景

- 语言到结构化查询语言（language to query）

北京的理工科大学有哪些？ ⟹ $\lambda x.$ 大学(x) ∧ 理工科(x) ∧ 位于(北京)

学校	类型	位置
清华	综合	北京
北大	综合	北京
北航	理工	北京
中科大	理工	合肥
北语	语言	北京
北邮	理工	北京
…	…	…

⟹

北航
北邮
…

Yuk Wah Wong and Raymond J. Mooney.
Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus. ACL-2007.

语义解析的任务场景

- 语言到代码（language to code）

输入NL

Adds a scalar to this vector in place.

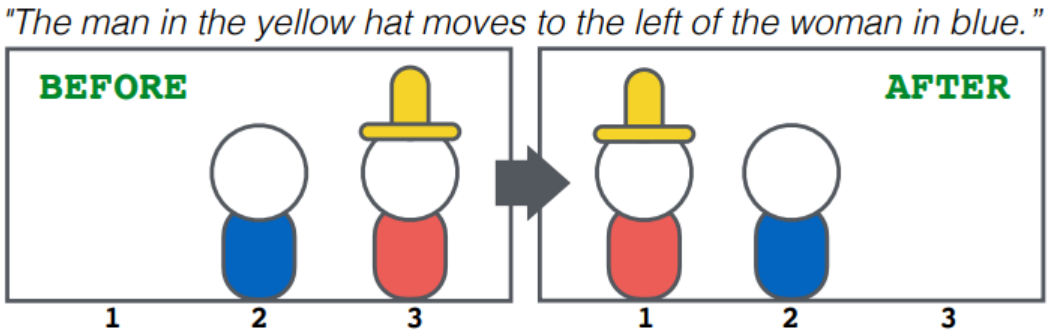
输出代码

```
public void add (final double arg0) {
  for (int loc0 = 0; loc0 < vecElements.length; loc0 ++) {
    vecElements[loc0] += arg0;
    }
}
```

Srinivasan Iyer, Yannis Konstas, Alvin Cheung, Luke Zettlemoyer.
Mapping Language to Code in Programmatic Context. EMNLP-2018.

语义解析的任务场景

- 语言到机器操作指令 （language to instruction）

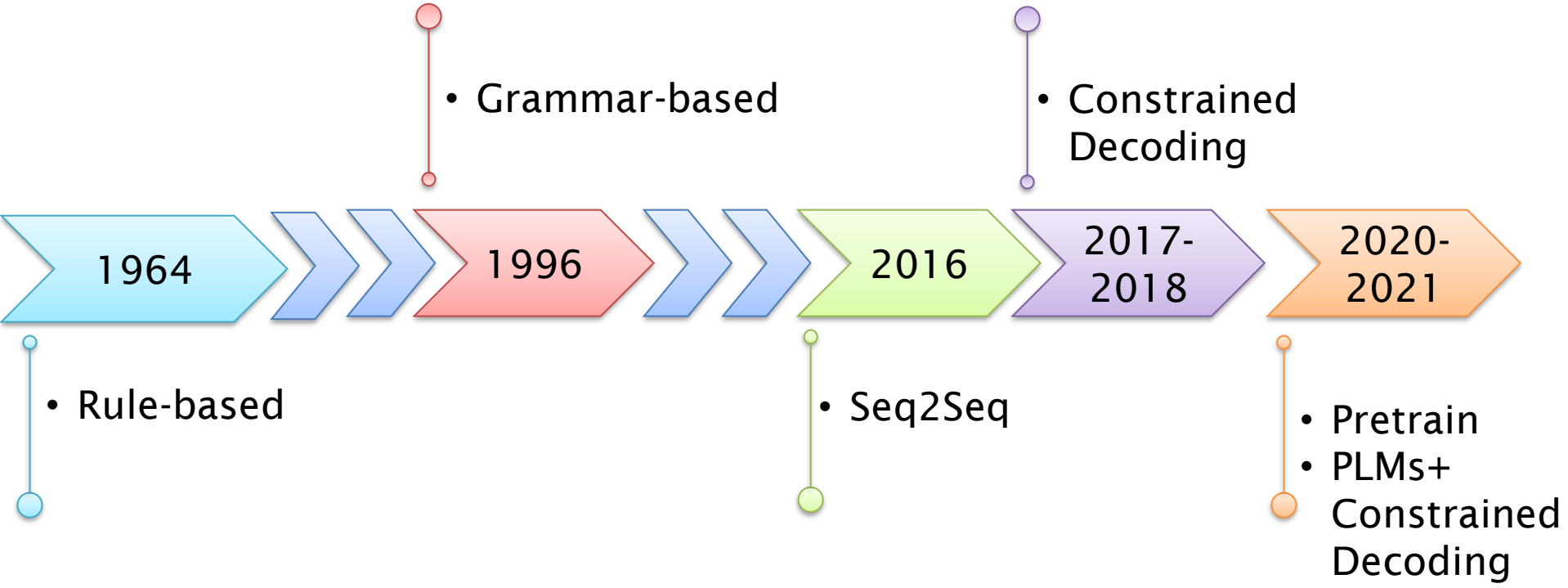


move(hasHat(yellow), leftOf(hasShirt(blue)))

Kelvin Guu, Panupong Pasupat, Evan Zheran Liu, Percy Liang.
From language to programs: bridging reinforcement learning and maximum marginal likelihood. ACL-2017.

语义解析的发展历程

语义解析的发展历程

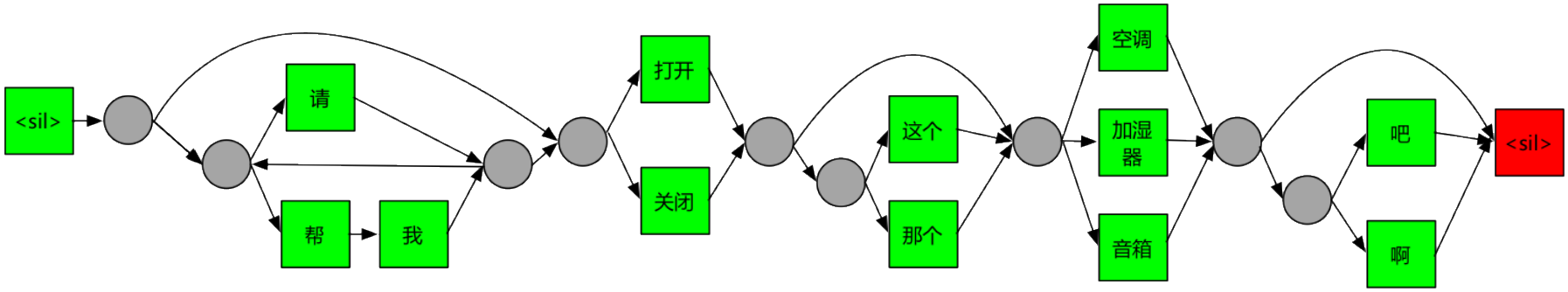

1964
• Rule-based
1996
• Grammar-based
2016
• Seq2Seq
2017-2018
• Constrained Decoding
2020-2021
• Pretrain
• PLMs+ Constrained Decoding

基于规则的语义解析方法

- STUDENT系统 [Bobrow, 1964]：基于规则的线性代数求解
- 规则系统：错误可控可溯源、起步容易/自助构建、增量式模型

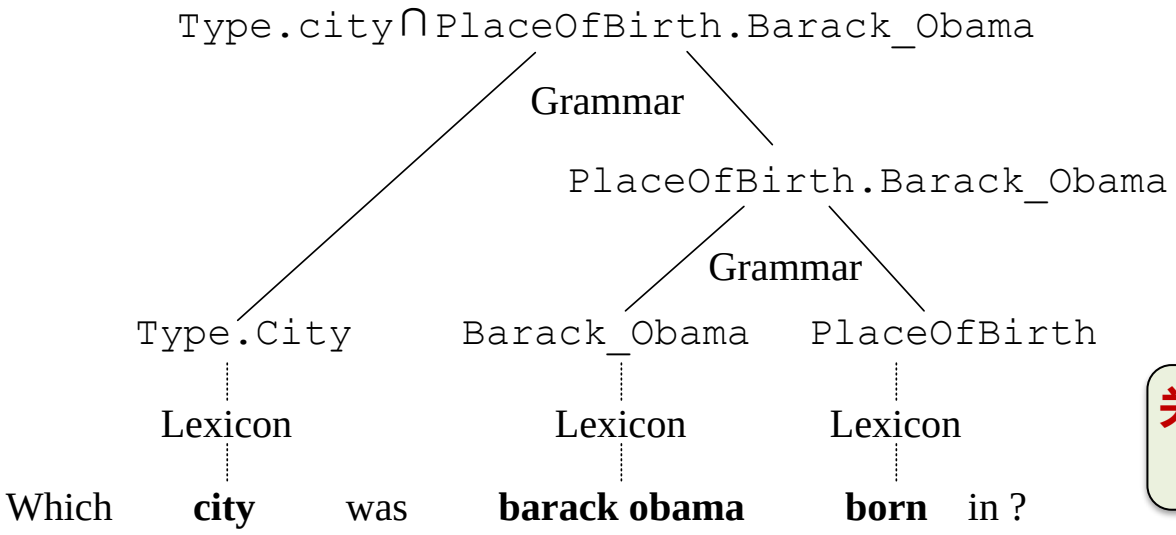
```
public <controlDevice> = <startPolite> <command> <endPolite>;

<command> = <action> <object>;
<action> = (打开|关闭);
<object> = [这个|那个](空调|加湿器|音箱){device};
<startPolite> = (请|帮 我) *;
<endPolite> = [啊|吧];
```



基于组合文法的语义解析方法

- 重要引发点：统计方法的兴起，并逐步替代规则方法
- 代表性方法：基于词典-组合文法的语义解析
  - CCG和DCS
  - 核心组件：词典（lexicon）、组合文法（grammar）、概率模型



关键：词典学习、组合模型

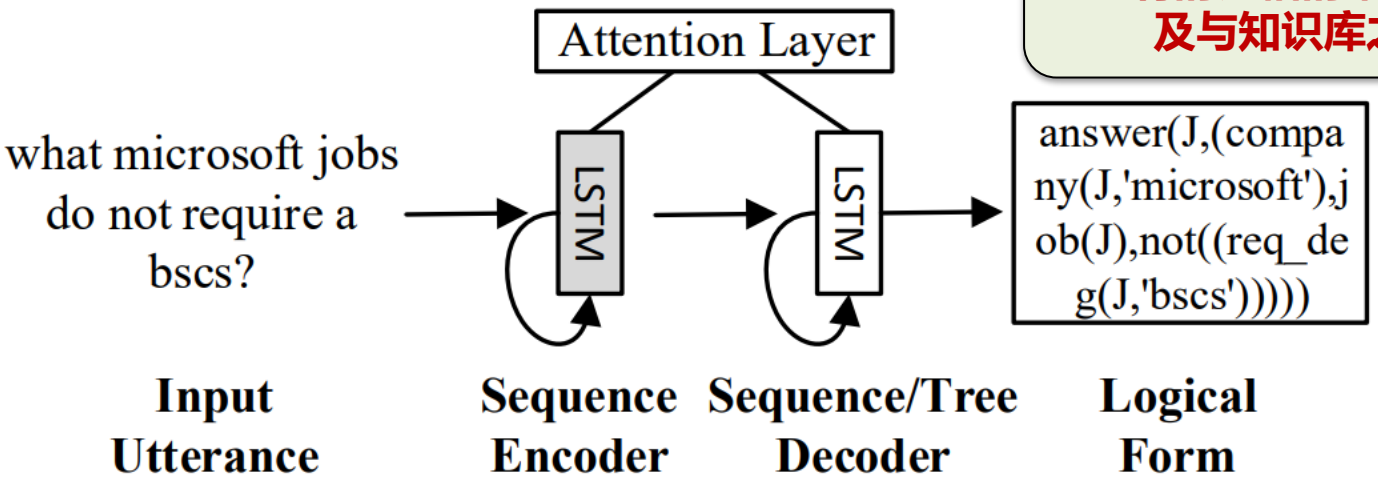
Jonathan Berant, Andrew Chou, Roy Frostig, Percy Liang.
Semantic parsing on Freebase from question-answer pairs. EMNLP-2013.

基于神经网络的语义解析方法

- 重要引发点：神经网络模型（特别是Seq2Seq模型）在自然语言处理其他任务上的应用及取得的成功

- 将逻辑表达式序列化
- 语义解析转换为词语序列到逻辑表达式序列的翻译过程（Seq2Seq）

关键：如何利用语义表示所具有的严格的结构约束，以及与知识库之间的联系



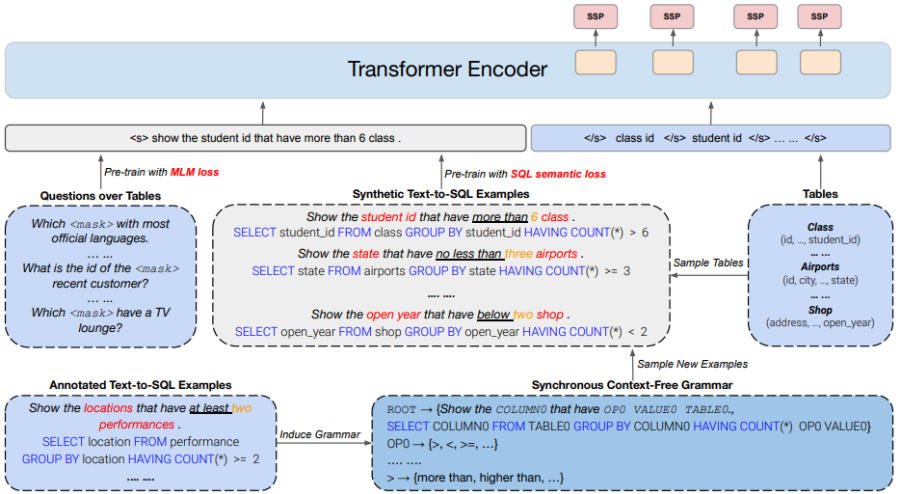
Li Dong, Mirella Lapata. Language to Logical Form with Neural Attention. ACL-2016.

基于预训练的语义解析方法

- 重要引发点：大规模预训练语言模型如EMLo、BERT、GPT等在NLP的多个任务上取得成效
- Pre-training then fine-tuning和prompt成为了NLP的新范式

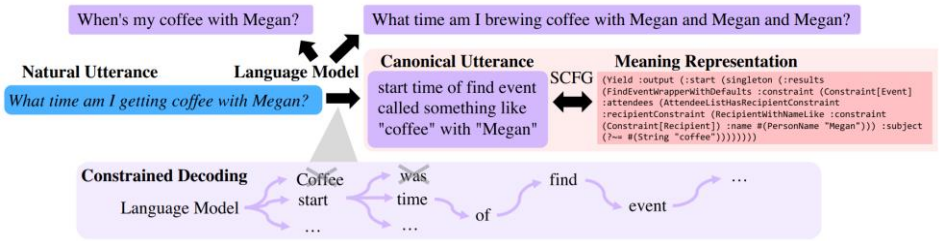
面向语义解析的预训练模型

关键：收集数据、设计自监督学习任务



直接运用通用预训练模型

关键：如何处理生成目标不是自然语言的问题



Tao Yu et al.. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing . ICLR-2021.

Richard Shin et al.. Constrained Language Models Yield Few-Shot Semantic Parsers. EMNLP-2021.

小结：语义解析简介

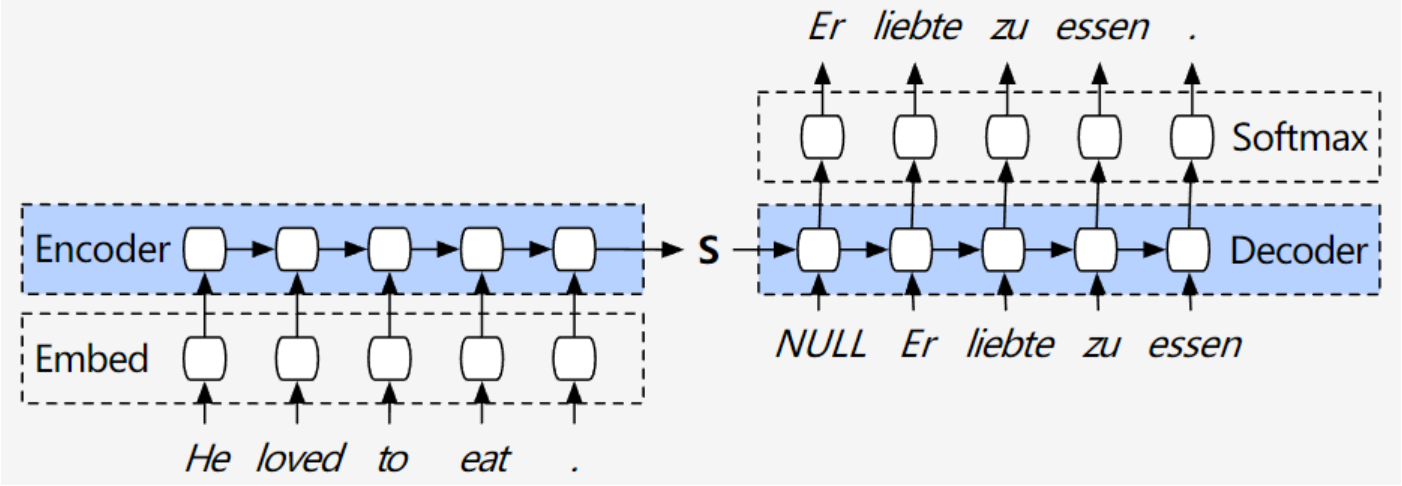
- **语义解析**：从自然语言到计算机语义表示
- **任务设置**：语言到查询、语言到代码、语言到指令
- **发展历程**：
  - 基于规则的方法（60s-90s）
  - 基于组合文法的方法（90s-10s）
  - 基于NN的方法(2016至今)
  - 基于预训练的方法(2020至今)

大纲

- 语义解析简介
  - 任务简介
  - 发展历程
- 基于深度学习的语义解析方法
  - Seq2Seq、Seq2Tree、Seq2Action
  - Constrained decoding
- 基于预训练的语义解析方法
  - 预训练方法在在Text-to-SQL任务上的应用
  - PLMs with Constrained decoding
- 总结与展望

基于神经网络的语义解析兴起前的背景

- 之前的模型都很复杂，模块很多
  - 基于规则方法中的规则设定，基于词典-组合文法方法中的词典学习
- 人工定义特征
- 神经网络模型在NLP其他任务上取得了成功
  - 基于encoder-decoder框架的神经机器翻译



代表性方法

- **Seq2Seq** [Dong & Lapata, 2016; Jia & Liang, 2016], **Seq2Tree** [Dong & Lapata, 2016]

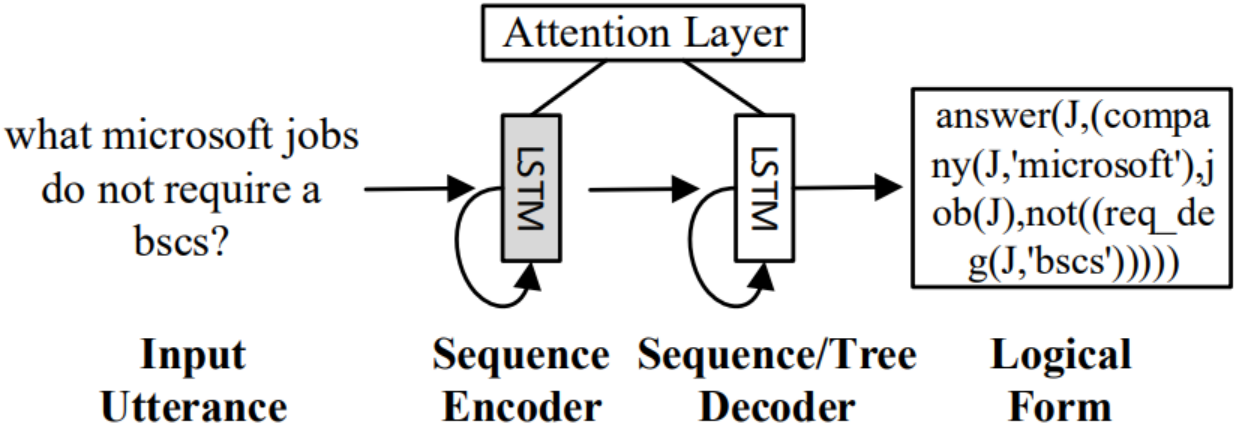
- **Seq2Act** [Chen et al., 2018]

SEQ2SEQ AND SEQ2TREE

Li Dong, Mirella Lapata. Language to Logical Form with Neural Attention. ACL-2016.
Robin Jia and Percy Liang. Data recombination for neural semantic parsing. ACL-2016.

Seq2Seq：把语义解析看做是机器翻译问题

- 将逻辑表达式序列化，看作一系列token，从而将语义解析转化为Seq2Seq问题

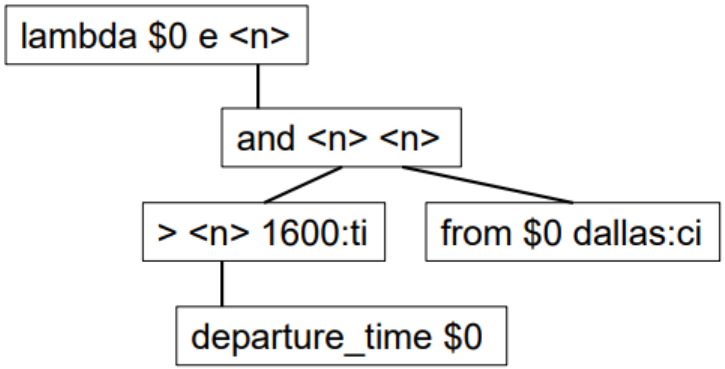


Li Dong, Mirella Lapata. Language to Logical Form with Neural Attention. ACL-2016.
Robin Jia and Percy Liang. Data recombination for neural semantic parsing. ACL-2016.

Seq2Seq中的问题

- 语义解析中的目标语言（语义表示）具有层次结构，而Seq2Seq模型仅把语义表示扁平序列化，从而忽略了层次结构信息

结构化表示



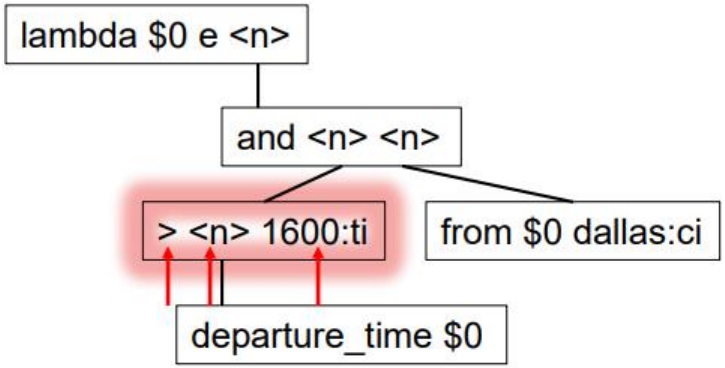
扁平化表示 lambda $0 e (and (> (departure_time $0) 1600:ti) (from $0 dallas:ci))

Seq2Seq中的问题

- 语义解析中的目标语言（语义表示）具有层次结构，而Seq2Seq模型仅把语义表示扁平序列化，从而忽略了层次结构信息

- 导致在解码过程中需要考虑更多的长距离依赖

结构化表示



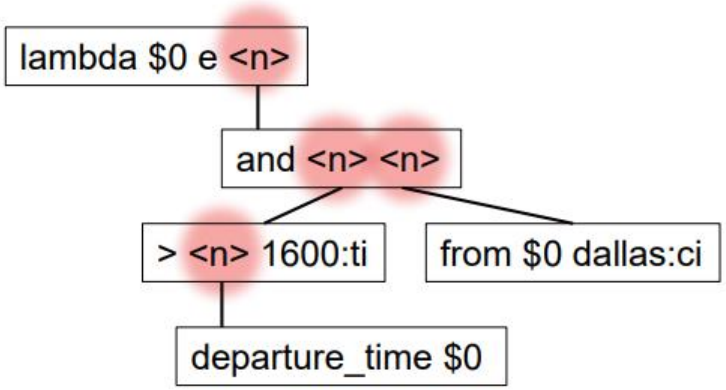
扁平化表示

lambda $0 e (and (> (departure_time $0) 1600:ti) (from $0 dallas:ci))

句子 dallas to san francisco leaving after 4 in the afternoon please

Seq2Tree

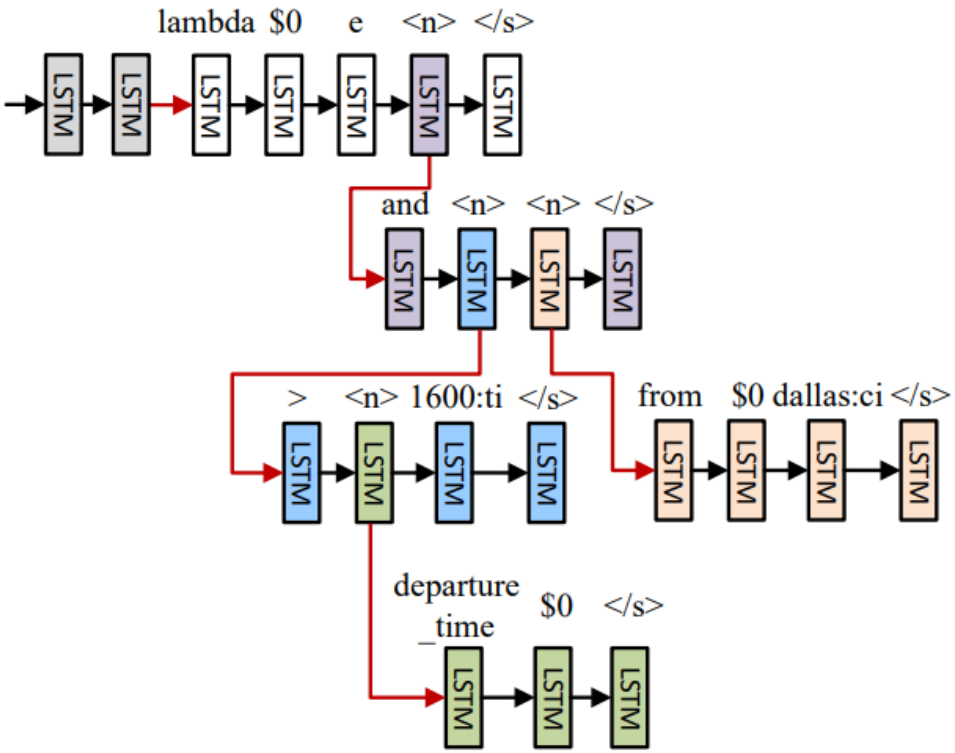
- 层次化的decoder，不生成扁平化的语义表示序列，而是生成层次结构化的语义表示（tree）
- 用<n>来表示树结构中的非终结符



Li Dong, Mirella Lapata. Language to Logical Form with Neural Attention. ACL-2016.

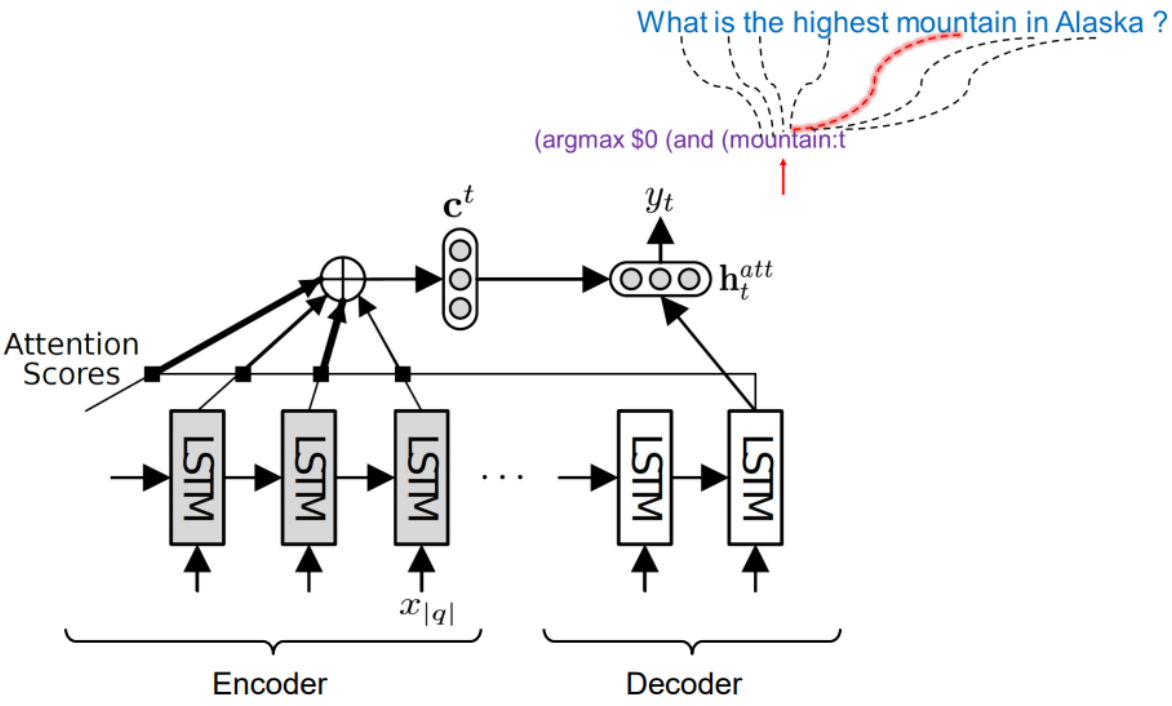
Seq2Tree

- 层次化的decoder，不生成扁平化的语义表示序列，而是生成层次结构化的语义表示（tree）
  - 占位符<n>都存放在一个队列里，直到队列为空才终止。



Seq2Seq/Seq2Tree中的注意力机制

- 注意力机制相当于让模型学习了词语到词语的语义表示之间的软对齐



Seq2Seq和Seq2Tree的性能

JOBS

Method	Accuracy
COCKTAIL (Tang and Mooney, 2001)	79.4
PRECISE (Popescu et al., 2003)	88.0
ZC05 (Zettlemoyer and Collins, 2005)	79.3
DCS+L (Liang et al., 2013)	90.7
TISP (Zhao and Huang, 2015)	85.0
SEQ2SEQ	87.1
– attention	77.9
– argument	70.7
SEQ2TREE	90.0
– attention	83.6

ATIS

Method	Accuracy
ZC07 (Zettlemoyer and Collins, 2007)	84.6
UBL (Kwiatkowski et al., 2010)	71.4
FUBL (Kwiatkowski et al., 2011)	82.8
GUSP-FULL (Poon, 2013)	74.8
GUSP++ (Poon, 2013)	83.5
TISP (Zhao and Huang, 2015)	84.2
SEQ2SEQ	84.2
– attention	75.7
– argument	72.3
SEQ2TREE	84.6
– attention	77.5

GEO

Method	Accuracy
SCISSOR (Ge and Mooney, 2005)	72.3
KRISP (Kate and Mooney, 2006)	71.7
WASP (Wong and Mooney, 2006)	74.8
λ-WASP (Wong and Mooney, 2007)	86.6
LNLZ08 (Lu et al., 2008)	81.8
ZC05 (Zettlemoyer and Collins, 2005)	79.3
ZC07 (Zettlemoyer and Collins, 2007)	86.1
UBL (Kwiatkowski et al., 2010)	87.9
FUBL (Kwiatkowski et al., 2011)	88.6
KCAZ13 (Kwiatkowski et al., 2013)	89.0
DCS+L (Liang et al., 2013)	87.9
TISP (Zhao and Huang, 2015)	88.9
SEQ2SEQ	84.6
– attention	72.9
– argument	68.6
SEQ2TREE	87.1
– attention	76.8

1. 层次结构的decoder使得模型学习到了目标语义表示的结构，这些结构信息能够帮助生成整个语义表示
2. Attention机制让模型学习了词语到词语的语义表示之间的软对齐，起到了类似于词典的效果

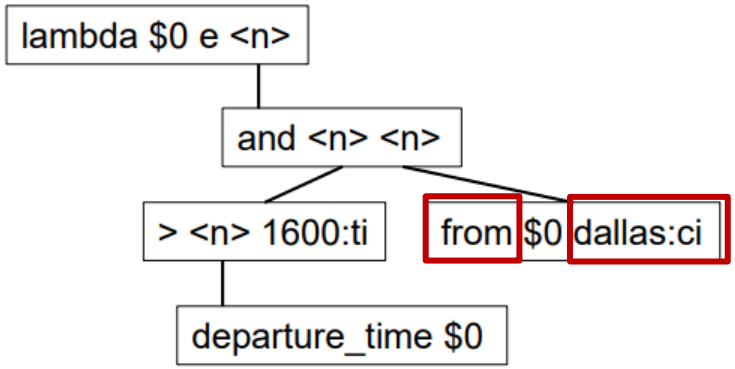
SEQ2ACT

Bo Chen, Le Sun, Xianpei Han.

Sequence-to-Action: End-to-End Semantic Graph Generation for Semantic Parsing. ACL-2018.

Seq2Act

- 之前的方法忽略生成的语义表示token之间的联系
  - 如from只能带两个参数
  - 如from的第二个参数只能是城市或者机场



- 用语义图表示语义
  - 与知识库建立紧密联系，充分利用知识库的知识约束

- 利用RNN模型的强表示能力和序列预测能力
  - 端到端

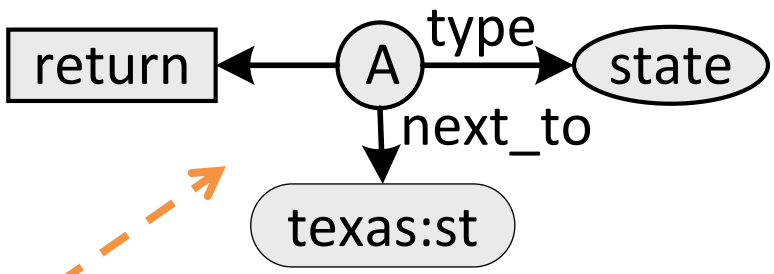
Bo Chen, Le Sun, Xianpei Han.
Sequence-to-Action: End-to-End Semantic Graph Generation for Semantic Parsing. ACL-2018.

Seq2Act: 端到端语义图生成

Which states border Texas?

sentence

Seq2Act: 端到端语义图生成

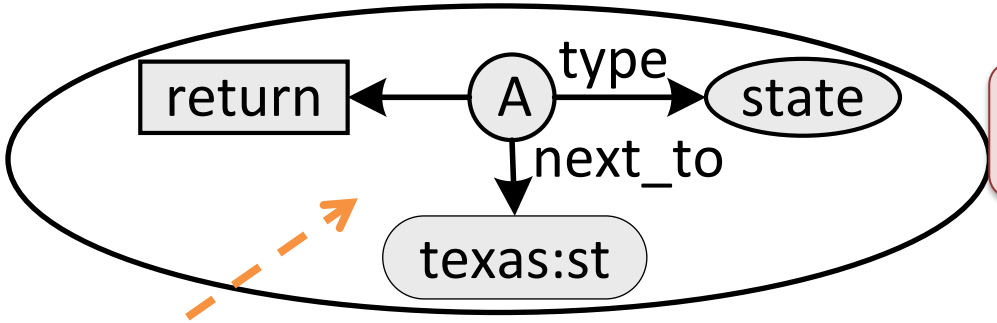


语义图

Which states border Texas?

句子

Seq2Act: 端到端语义图生成

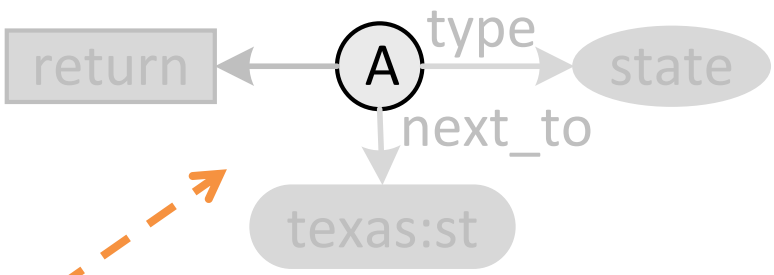


语义图

Which states border Texas?

句子

Seq2Act: 端到端语义图生成



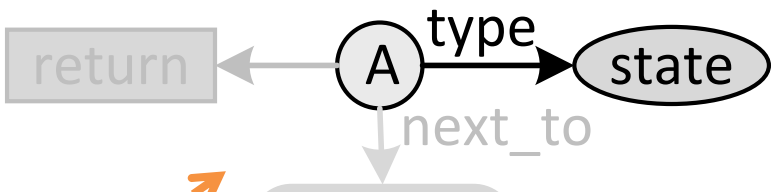
语义图

Action 1: add node A

Which states border Texas?

句子

Seq2Act: 端到端语义图生成



语义图

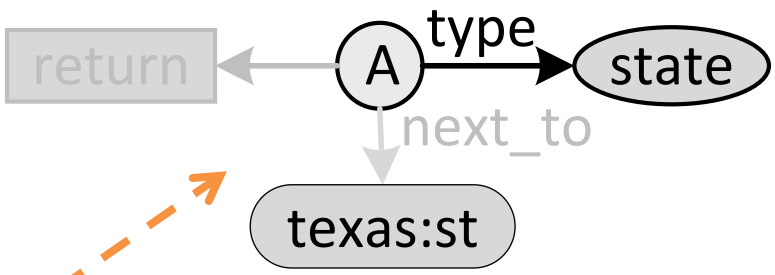
Action 1: add node A

Action 2: add type state

Which states border Texas?

句子

Seq2Act: 端到端语义图生成



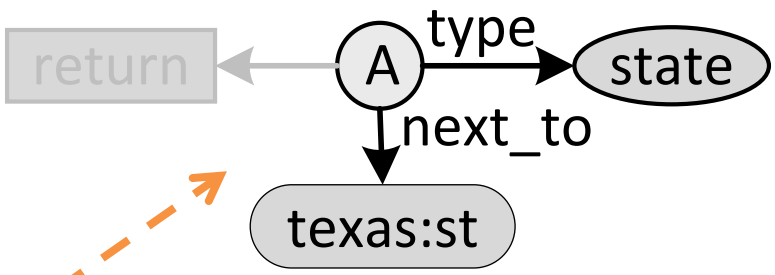
语义图

Action 1: add node A
Action 2: add type state
Action 3: add node texas:st

Which states border Texas?

句子

Seq2Act: 端到端语义图生成



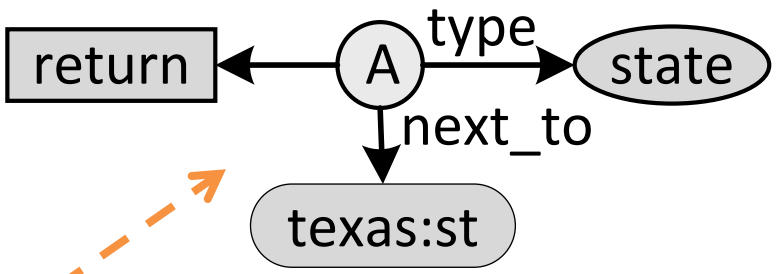
语义图

Action 1: add node A
Action 2: add type state
Action 3: add node texas:st
Action 4: add edge next_to

Which states border Texas?

句子

Seq2Act: 端到端语义图生成



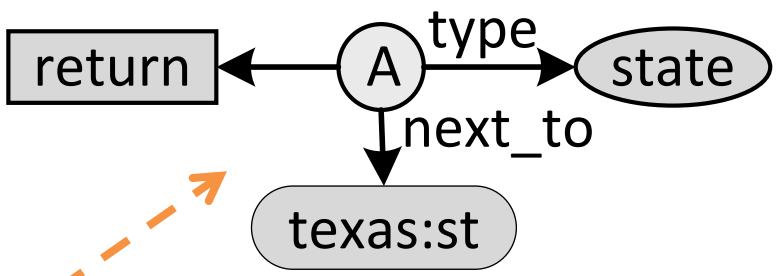
语义图

Action 1: add node A
Action 2: add type state
Action 3: add node texas:st
Action 4: add edge next_to
Action 5: return

Which states border Texas?

句子

Seq2Act: 端到端语义图生成



语义图

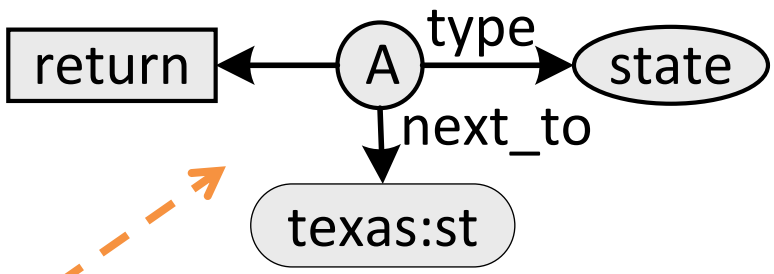
Which states border Texas?

Action 1: add node A
Action 2: add type state
Action 3: add node texas:st
Action 4: add edge next_to
Action 5: return

句子

动作序列

Seq2Act: 端到端语义图生成



语义图

translate

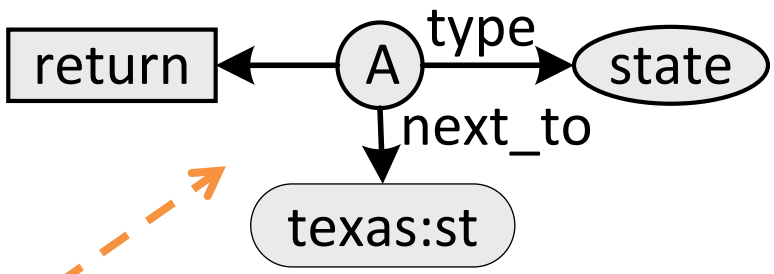
Which states border Texas?

Action 1: add node A
Action 2: add type state
Action 3: add node texas:st
Action 4: add edge next_to
Action 5: return

句子

动作序列

Seq2Act: 端到端语义图生成



语义图

Which states border Texas?

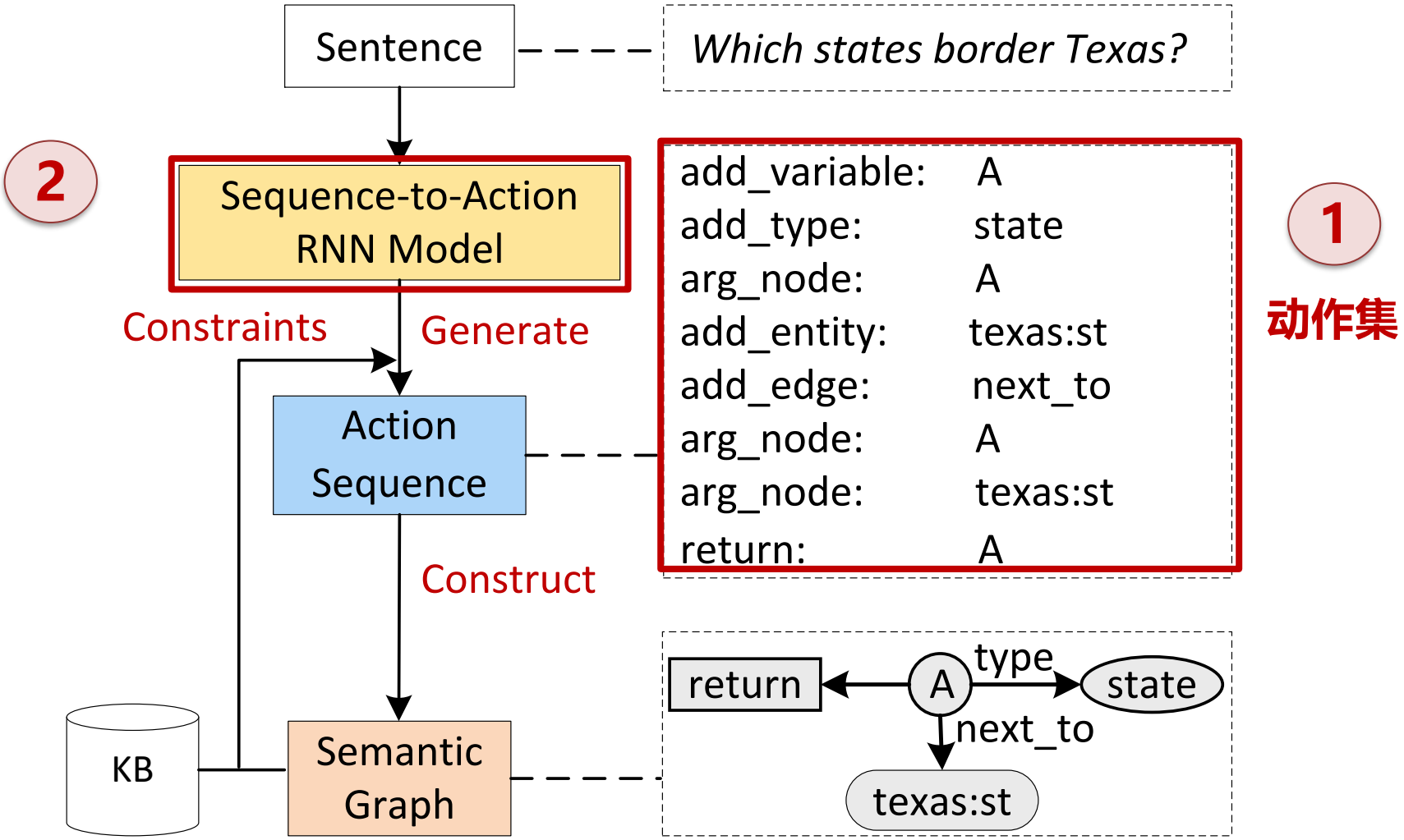
Sequence-to-Action

Action 1: add node A
Action 2: add type state
Action 3: add node texas:st
Action 4: add edge next_to
Action 5: return

句子

动作序列

Seq2Act: 模型框架

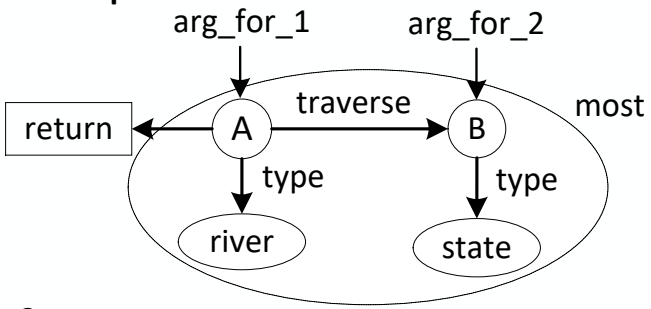


Seq2Act: 动作集

- Add variable node
  - E.g., A
- Add entity node
  - E.g., texas:st
- Add type node
  - E.g., state
- Add edge
  - E.g., next_to
- Operation action
  - E.g., argmax, argmin, count
- Argument action
  - For type node, edge and operation

Sentence: *Which river runs through the most states?*

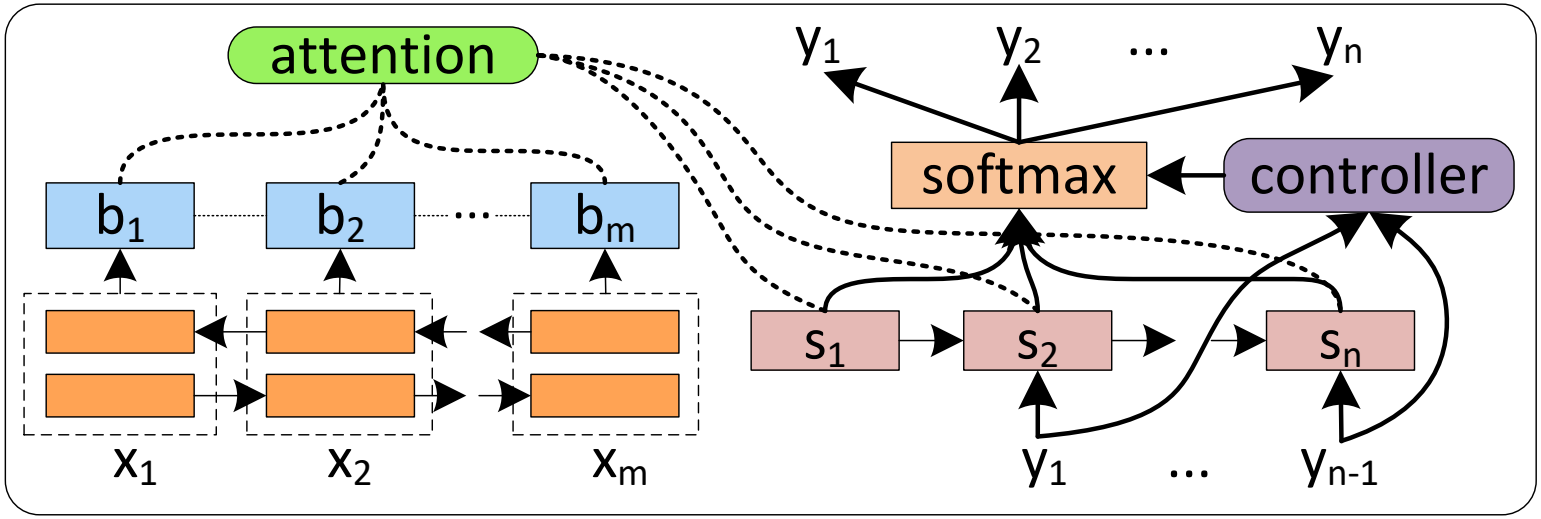
Semantic Graph:



Action Sequence:

Structure	Semantic	Arg
add_operation	most	
add_variable	A	
add_type	river	A
add_variable	B	
add_type	state	B
add_edge	traverse	A, B
end_operation	most	A, B
return	A	

Seq2Act: encoder-decoder model

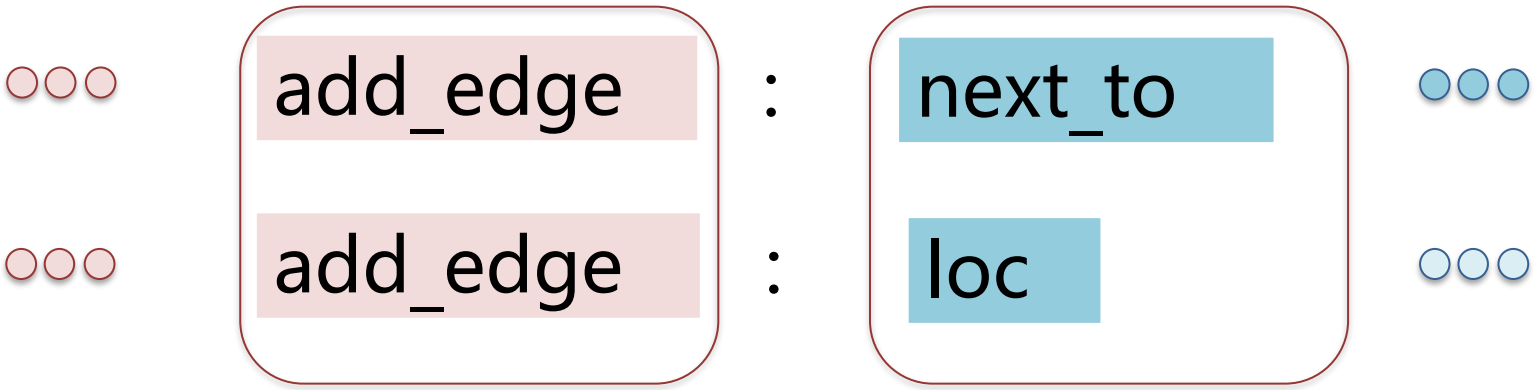


Typical encoder-decoder model (bi-LSTM with attention)

Action embedding

Seq2Act: action embedding

- 包含句法部分和语义部分
  - 分别进行编码，可以一定程度缓解动作的稀疏性（思想与因子化词汇类似）



Φ (add_edge:next to) = [Φ (add_edge); Φ (next_to)]

Seq2Act的性能

- 动作编码更加的紧凑，一定程度解决长距离依赖的问题
- 更容易在解码过程中加入知识库的约束信息，来保证解码时生成符合句法、符合语义的语义表示

	GEO	ATIS
Previous Work		
Zettlemoyer and Collins (2005)	79.3	–
Zettlemoyer and Collins (2007)	86.1	84.6
Kwiatkowksi et al. (2010)	88.9	–
Kwiatkowski et al. (2011)	88.6	82.8
Liang et al. (2011)* (+lexicon)	**91.1**	–
Poon (2013)	–	83.5
Zhao et al. (2015)	88.9	84.2
Rabinovich et al. (2017)	87.1	**85.9**
Seq2Seq Models		
Jia and Liang (2016)	85.0	76.3
Jia and Liang (2016)* (+data)	89.3	83.3
Dong and Lapata (2016): 2Seq	84.6	84.2
Dong and Lapata (2016): 2Tree	87.1	84.6
Our Models		
Seq2Act	87.5	84.6
Seq2Act (+C1)	88.2	85.0
Seq2Act (+C1+C2)	88.9	85.5

	Logical Form	Action Sequence
GEO	28.2	18.2
ATIS	28.4	25.8
OVERNIGHT	46.6	33.3

序列化逻辑表达式的长度与动作序列长度对比

方法小结

- Seq2Seq: 直接把目标语义表示序列化
- Seq2Tree: 考虑到目标语义表示的结构性
- Seq2Act: 用语义图表示语义，用动作序列编码语义图的构建

CONSTRAINED DECODING

Constrained decoder

- 由于目标语言是形式化语言，需符合严格的条件约束，因此对比机器翻译里面的decoder，语义解析中的decoder可以使用严格的约束条件

- 句法条件和语义条件
  - 句法条件：如”(lambda $0 e (and (flight $0) (from $0 “下一个token应该是一个参数（实体或者变量）。
  - 语义条件： “(lambda $0 e (and (flight $0) (from $0 ” 下一个token应该是一个城市或者机场

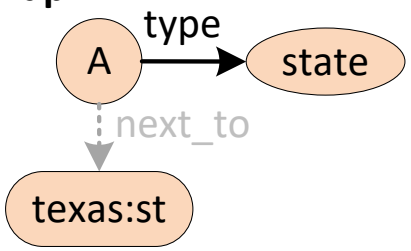
- 避免生成明显错误的token，从而提高最终生成准确语义表示的概率

- 解码过程中的硬性约束，在所有的基于神经网络的语义解析方法中都可以使用

Constrained decoder in Seq2Act

Sentence: *Which states border Texas?*

Partial Semantic Graph:



	Structure	Semantic	Arg	Validity
Generated Actions	add_variable add_type add_entity	A state texas:st	A	
Candidate Next Action	add_type	city	texas:st	✗
	add_edge	loc	A, texas:st	✗
	add_edge	next_to	A, A	✗
	add_edge	next_to	A, texas:st	✓
	⋮	⋮	⋮	⋮

Action 1: violate type conflict

Action 2: violate selectional preference constraint

Action 3: structure constraint

Action 4: YES

Grammar constrained decoder

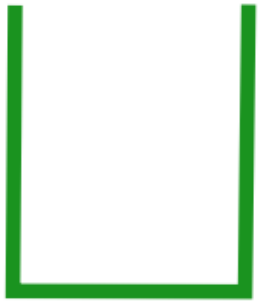
Which athlete was from South Korea after the year 2010?

Generated Actions

START → c
c→(<r,c> r)
<r,c>→(<<c,r>,<r,c>> <c,r>)
<<c,r>,<r,c>>→**reverse**
<c,r>→**athlete**
r→(<r,<r,r>> r r)
<r,<r,r>>→**and**
r→(<c,r> c)
<c,r>→**nation**
c→**south_korea**
r→(<c,r> c)
<c,r>→**year**
c→(<d,c> d)
<d,c>→(<<c,d>,<d,c>> <c,d>)
<<c,d>,<d,c>>→**reverse**
<c,d>→**date**
d→(**>=** d)
d→**2010.mm.dd**

Logical Form

```
((reverse athlete)
 (and (nation south_korea)
      (year ((reverse date)
             (>= 2010-mm-dd)))
```



Non-terminal Stack

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner.
Neural semantic parsing with type constraints for semi-structured tables. EMNLP-2017.

小结：基于神经网络的语义解析方法

- **主流：**基于encoder-decoder的基本框架生成语义表示
- **核心：**选择语义表示，定义生成语义表示的action、grammar或者推导规则。
- **代表性方法：**Seq2Seq、Seq2Tree、Seq2Action等
- **扩展：**Constrained Decoding

- 优点：
  - 模型相对简单、端到端

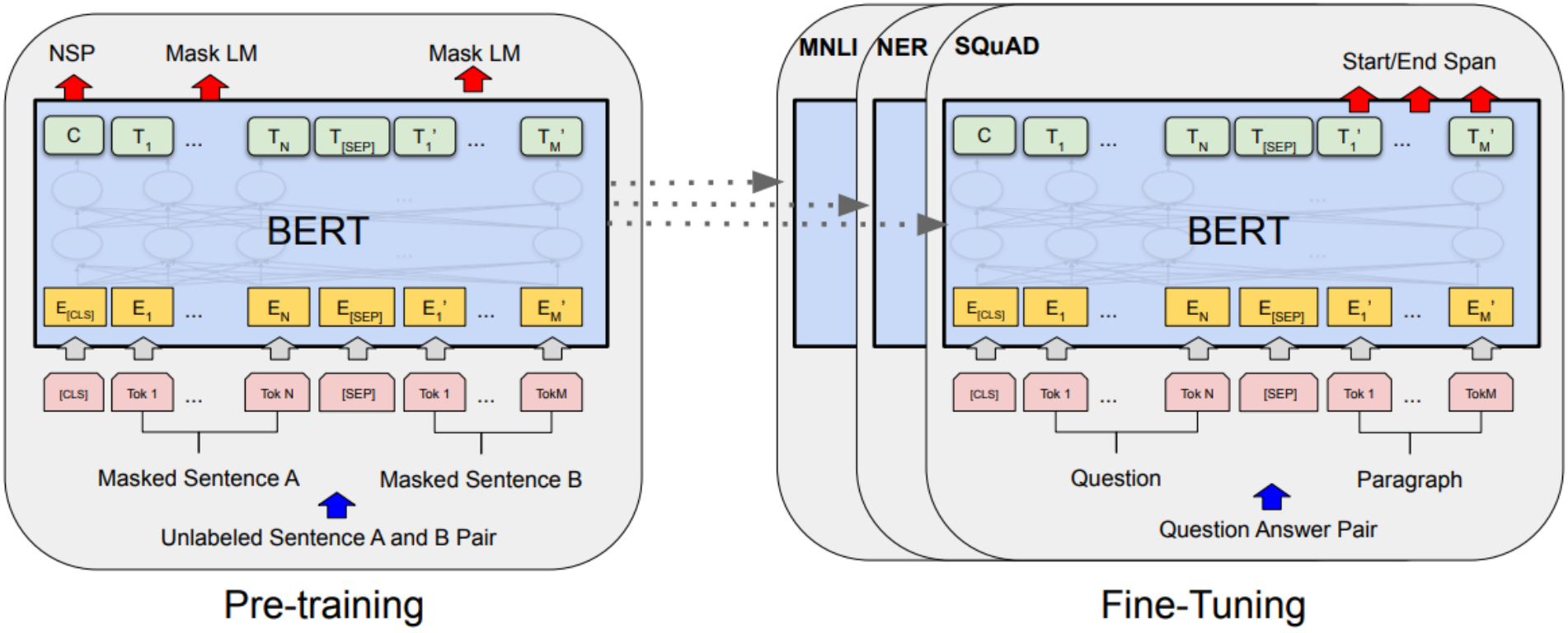
- 缺点：
  - 需要标注数据、可解释性差

大纲

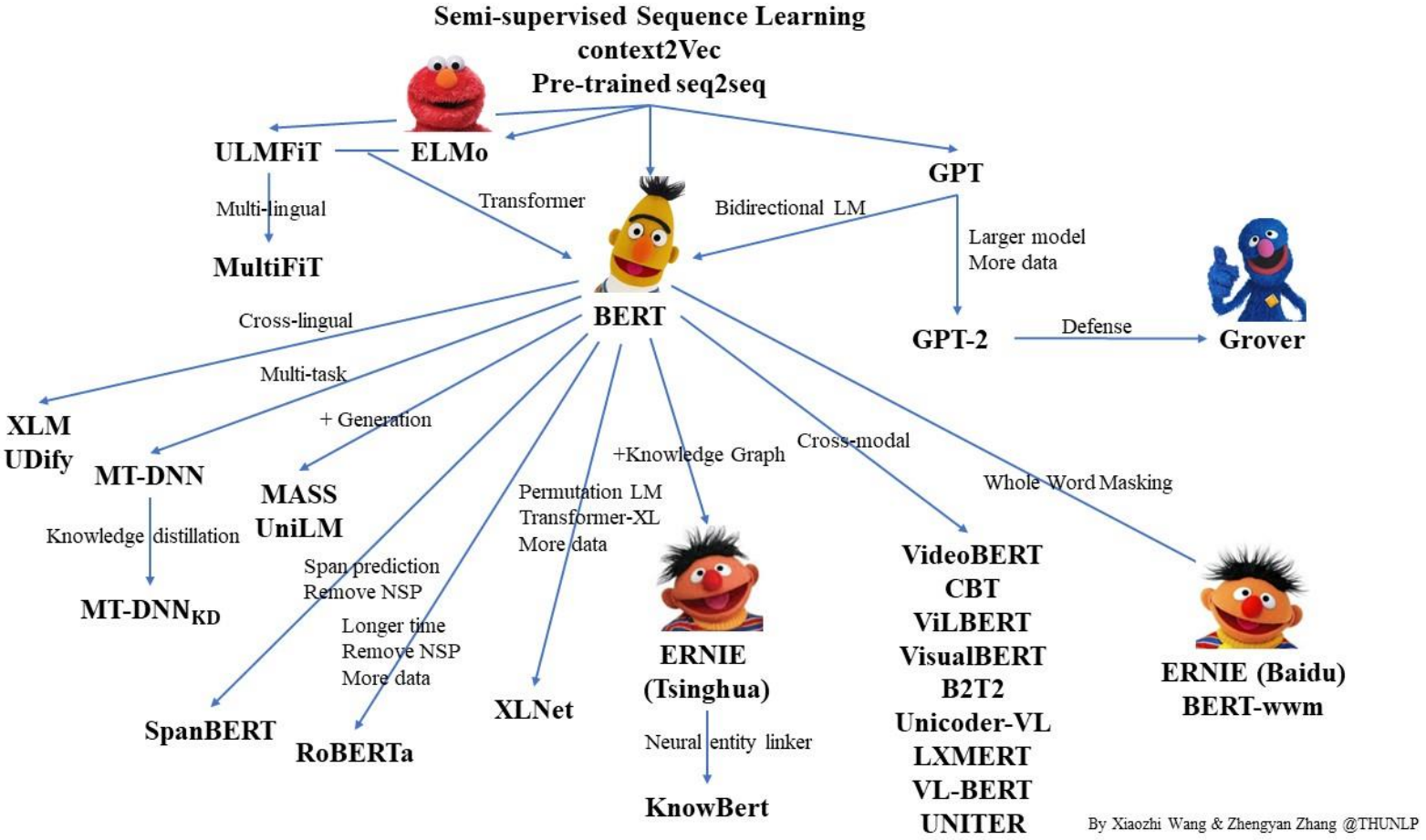
- 语义解析简介
  - 任务简介
  - 发展历程
- 基于深度学习的语义解析方法
  - Seq2Seq、Seq2Tree、Seq2Action
  - Constrained Decoding
- 基于预训练的语义解析方法
  - 预训练方法在Text-to-SQL任务上的应用
  - PLMs with Constrained Decoding
- 总结与展望

预训练语言模型

- 以ELMo，BERT，GPT为代表
- Pre-training then fine-tuning成为了NLP的研究新范式



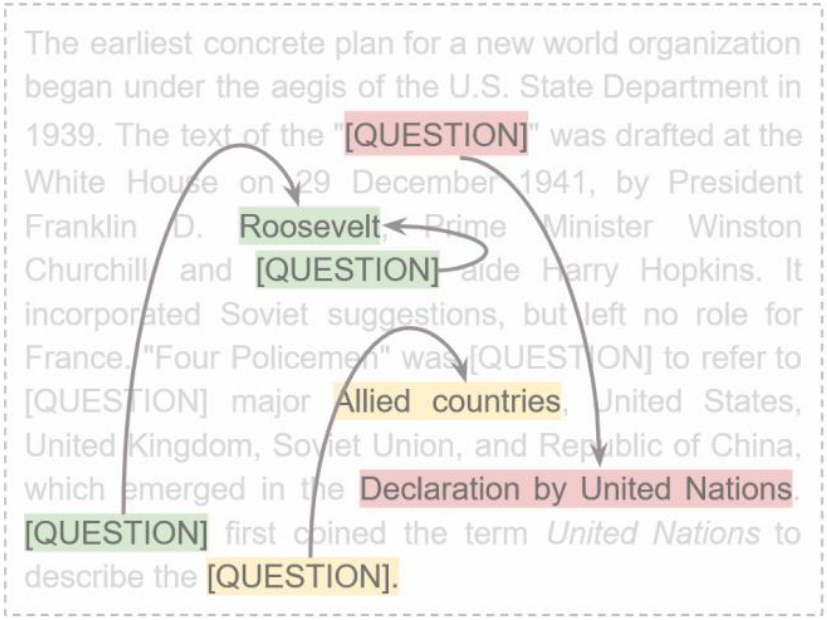
预训练语言模型大家族



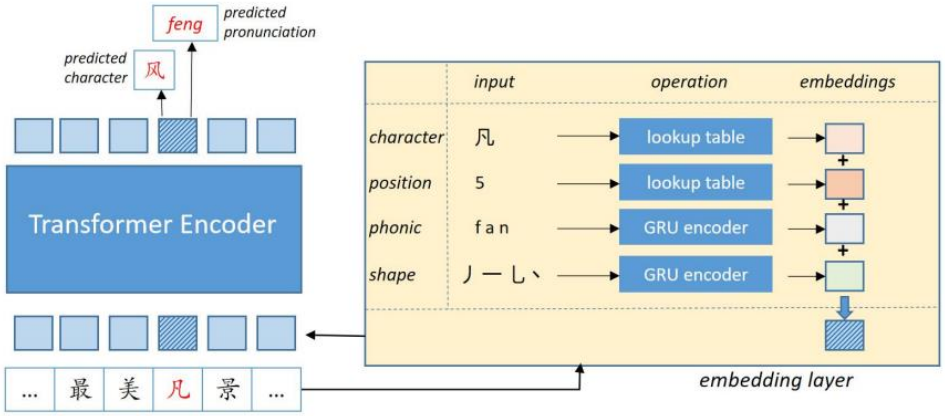
https://github.com/thunlp/PLMpapers

Pre-training for NLP tasks

Splinter: PLM for MRC



PLOME: PLM for Spelling Correction



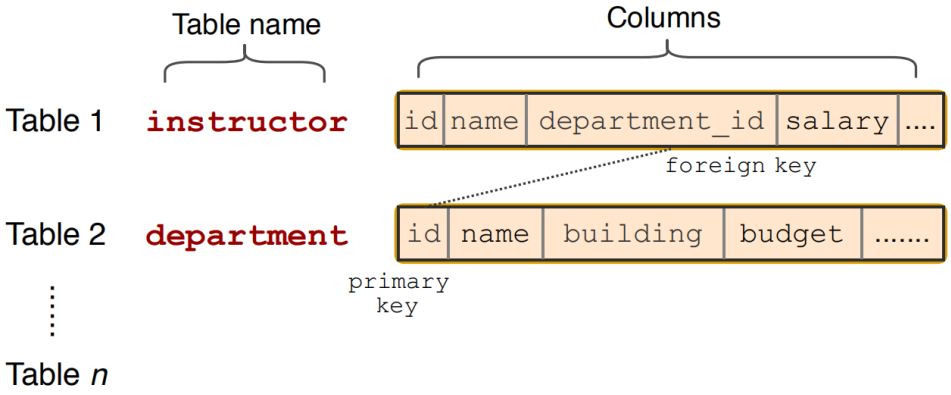
关键：收集数据、设计自监督学习任务

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, Omer Levy. Question answering by pretraining span selection. ACL-2021.

Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, Di Wang. PLOME: Pre-training with Misspelled Knowledge for Chinese Spelling Correction. ACL-2021.

PRE-TRAINING FOR TEXT-TO-SQL

Text-to-SQL



Annotators create:

Complex question What are the name and budget of the departments with average instructor salary greater than the overall average?

Complex 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)
```

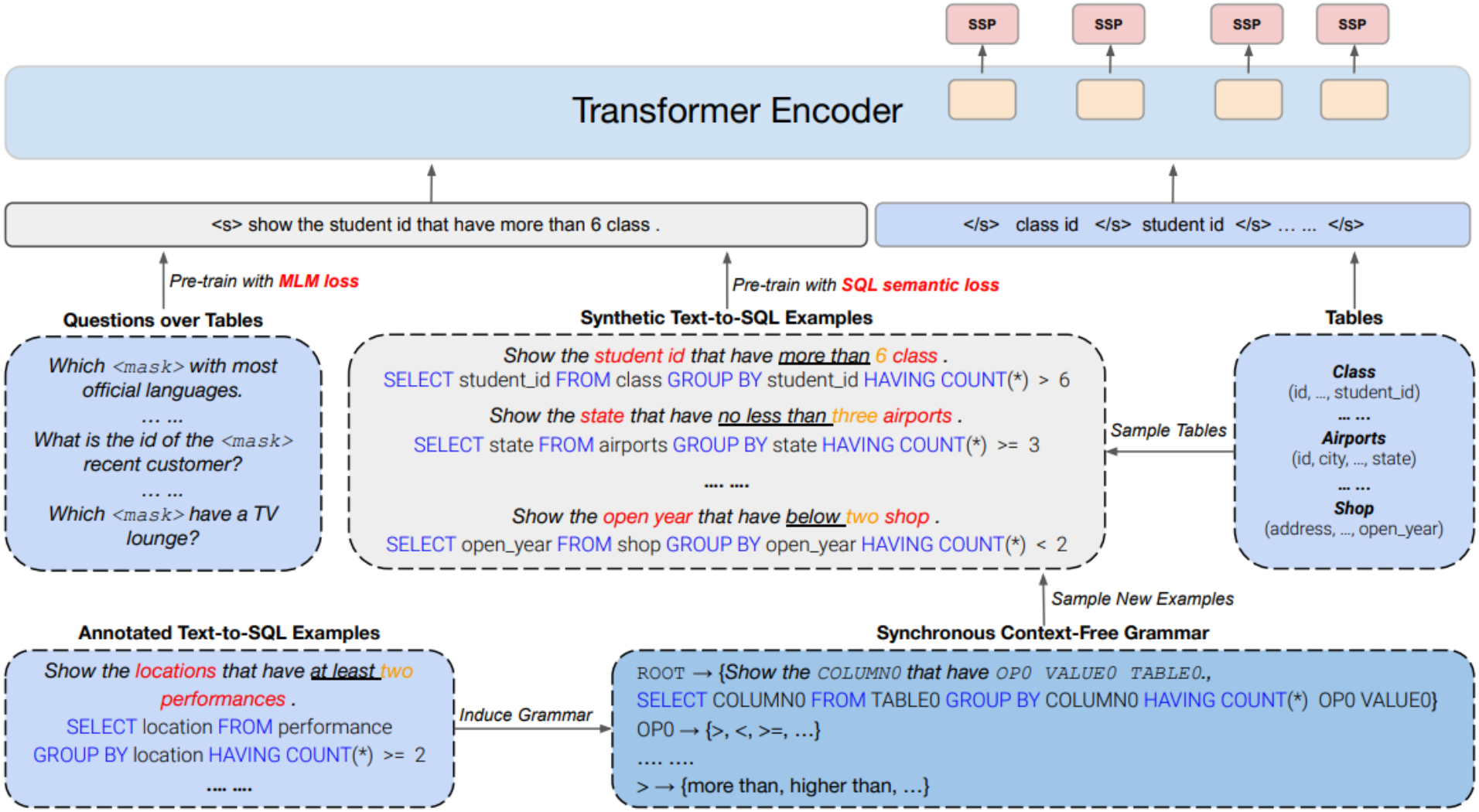
1. **生成目标是SQL，与自然语言大不同**
2. **生成SQL不仅与question有关，还与ontology（tables）相关**

代表性工作

- TABERT [Yin et al., ACL-2020]
- TAPAS [Herzig et al., ACL-2020]
- **GraPPa** [Yu et al., ICLR-2021]
- GAP [Shi et al., AAAI-2021]

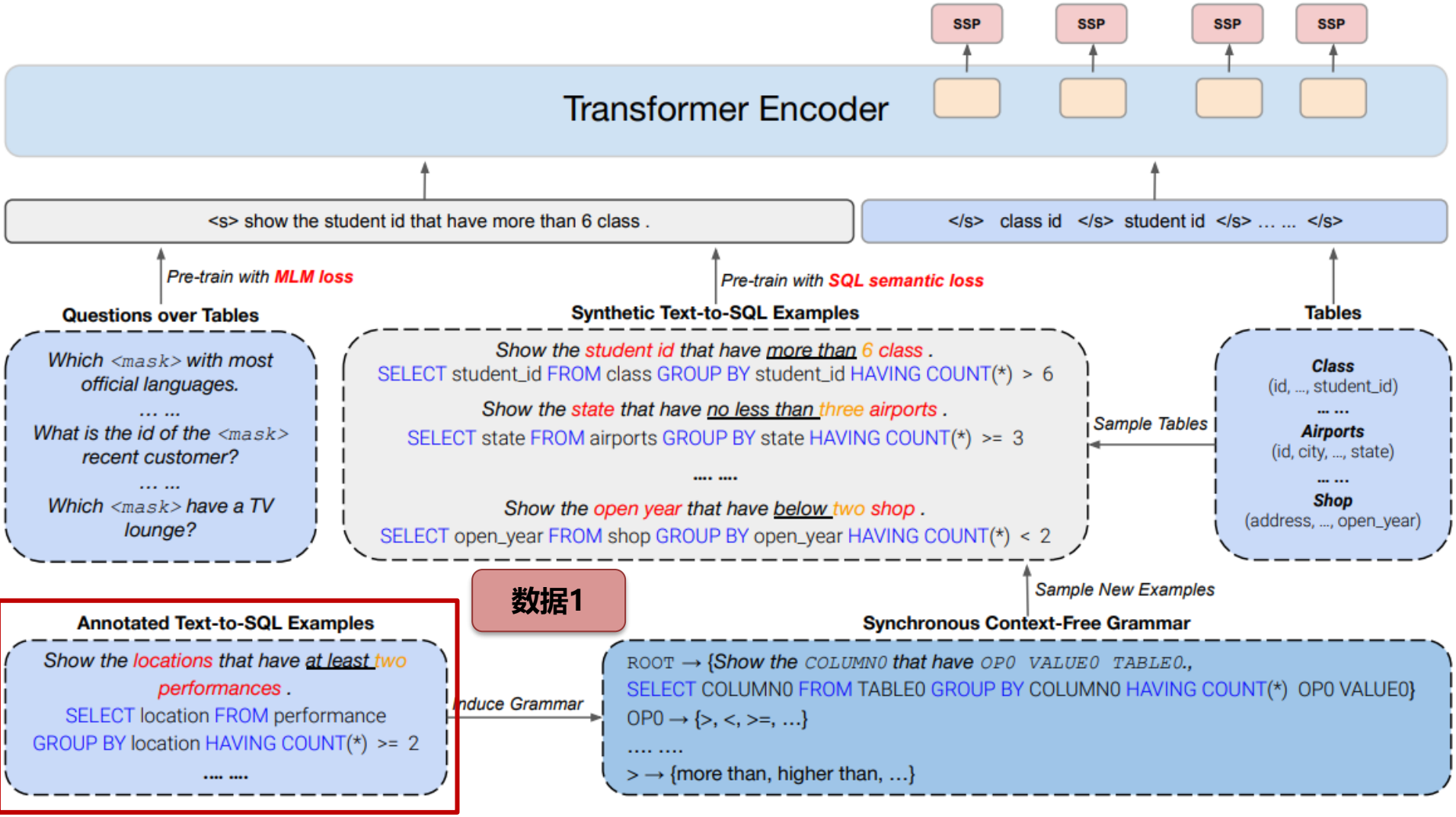
Rank	Model	Dev	Test
1 Sep 1, 2021	S²SQL + ELECTRA (DB content used) *Anonymous*	76.4	**72.1**
1 Jun 1, 2021	LGESQL + ELECTRA (DB content used) *SJTU X-LANCE Lab & AISpeech* (Cao et al., ACL'21) code	75.1	**72.0**
1 Jul 14, 2021	T5-3B+PICARD (DB content used) *Element AI, a ServiceNow company* (Scholak et al., EMNLP'21) code	75.5	**71.9**
4 Nov 19, 2020	DT-Fixup SQL-SP + RoBERTa (DB content used) *Borealis AI* (Xu et al., ACL'21) code	75.0	70.9
5 Nov 19, 2020	RAT-SQL + GraPPa + Adv (DB content used) *Anonymous*	75.5	70.5
6 Nov 19, 2020	SADGA + GAP (DB content used) *Anonymous*	73.1	70.1
7 Dec 25, 2020	RATSQL + GraPPa + GP (DB content used) *OCFT Gamma Big Data Lab* (Zhao et al.,'21)	72.8	69.8
8 Sep 08, 2020	RATSQL + GAP (DB content used) *University of Waterloo & AWS AI Labs* (Shi et al., AAAI'21) code	71.8	69.7
9 Aug 18, 2020	RATSQL + GraPPa (DB content used) *Yale & Salesforce Research* (Yu et al., ICLR'21) code	73.4	69.6
10 Mar 10, 2021	SmBoP + GraPPa (DB content used) *Tel-Aviv University & Allen Institute for AI* (Rubin and Berant, NAACL'21) code	74.7	69.5

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing



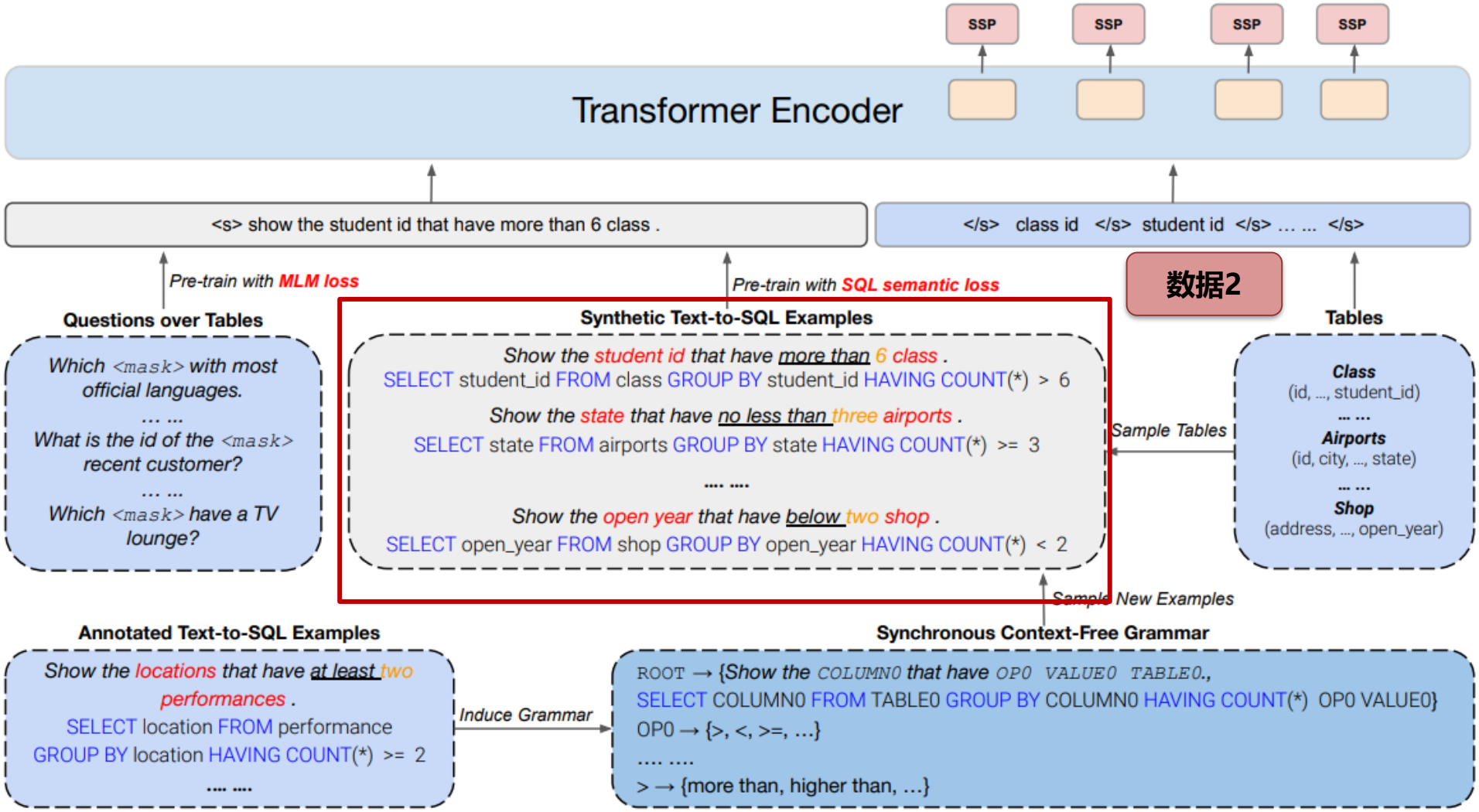
Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher and Caiming Xiong. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. ICLR-2021.

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing



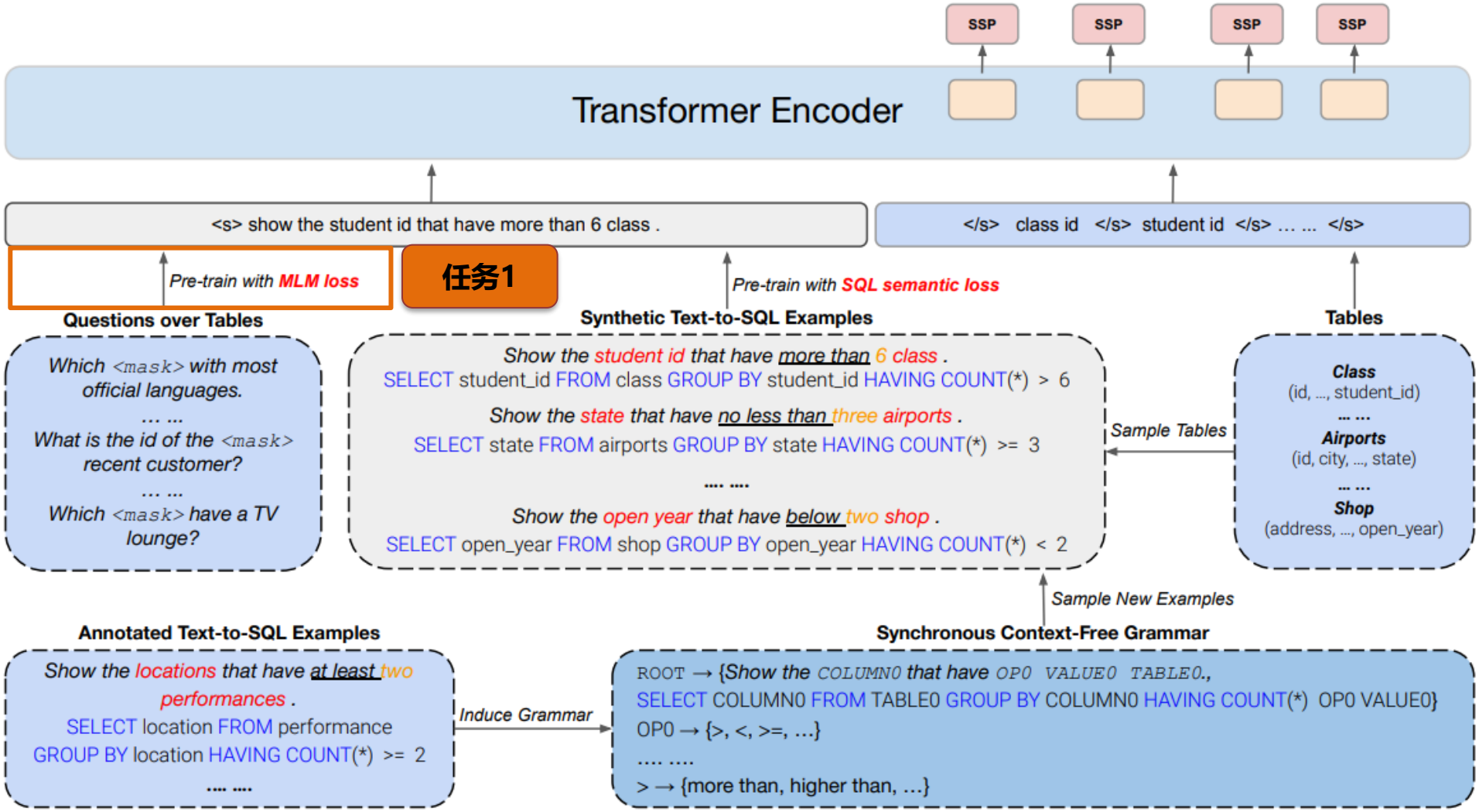
Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher and Caiming Xiong. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. ICLR-2021.

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing



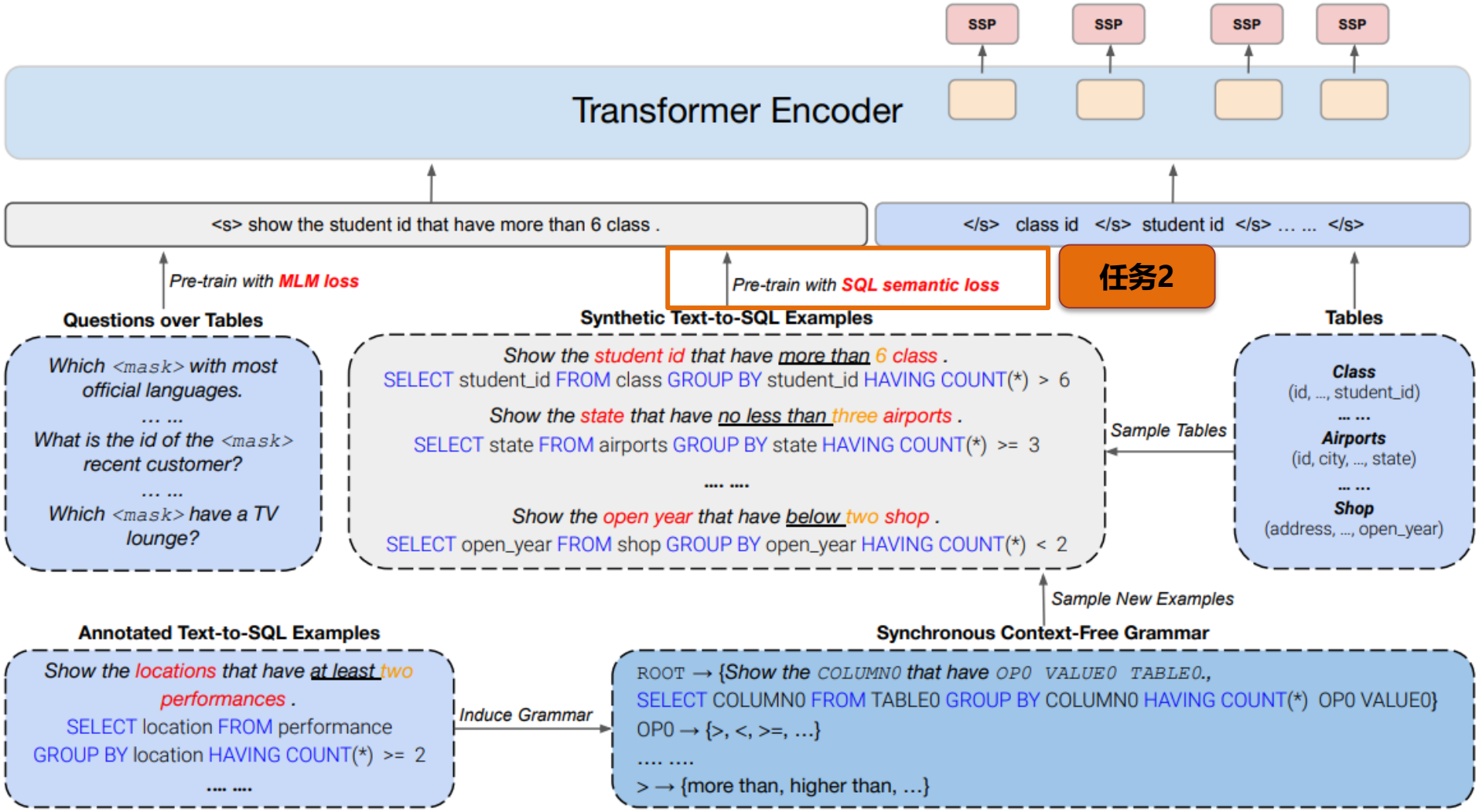
Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher and Caiming Xiong. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. ICLR-2021.

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing



Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher and Caiming Xiong. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. ICLR-2021.

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing



Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher and Caiming Xiong. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. ICLR-2021.

GraPPa：收集数据1

- 只需要natural language utterances和对应的tables（不需要SQL）

	Train Size	# Table	Task
TabFact	92.2K	16K	Table-based fact verification
LogicNLG	28.5K	7.3K	Table-to-text generation
HybridQA	63.2K	13K	Multi-hop question answering
WikiSQL	61.3K	24K	Text-to-SQL generation
WikiTableQuestions	17.6K	2.1K	Question answering
ToTTo	120K	83K	Table-to-text generation
Spider	8.7K	1K	Text-to-SQL generation

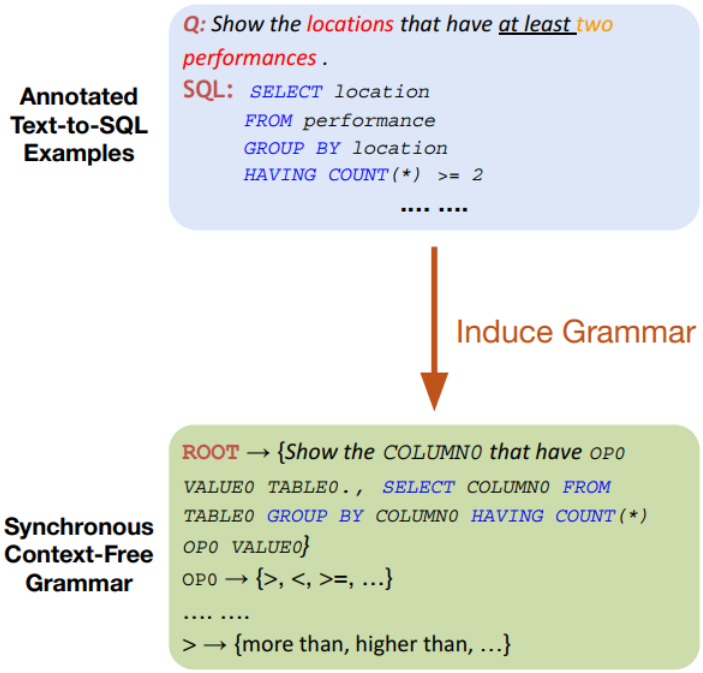
Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

Greece held its last Summer Olympics in which year?

Panupong Pasupat, Percy Liang. Compositional Semantic Parsing on Semi-Structured Tables. ACL-2015.

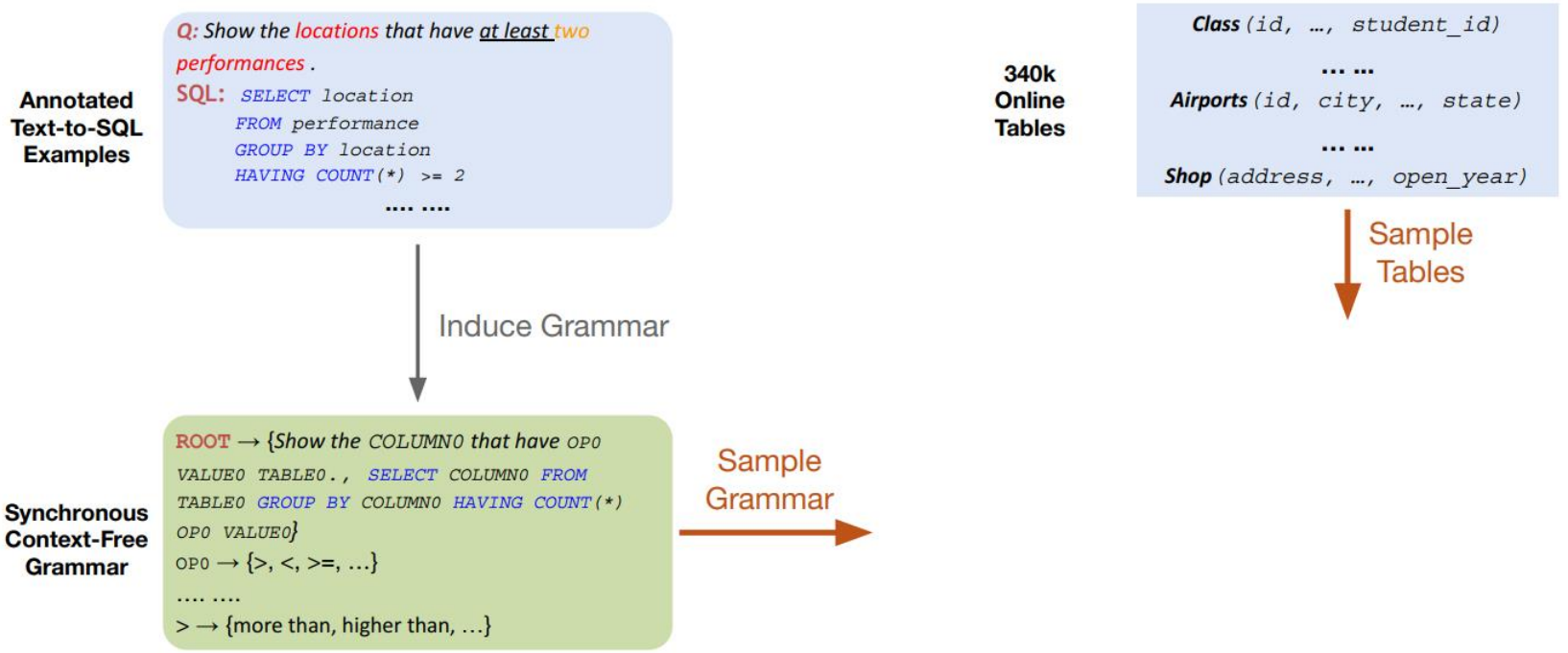
GraPPa：自动生成数据2

- 1. 从已有的标注数据中抽取同步文法



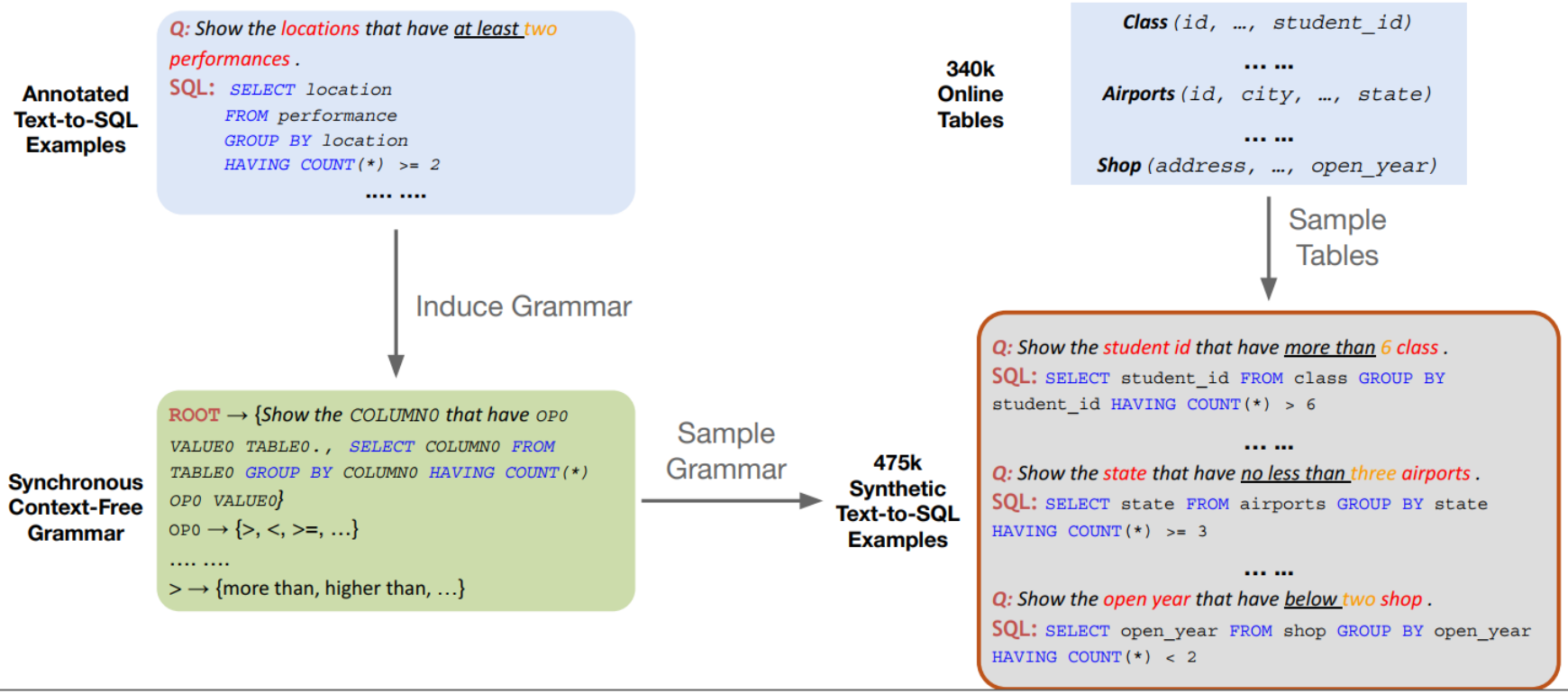
GraPPa：自动生成数据2

- 2. 采样新表格



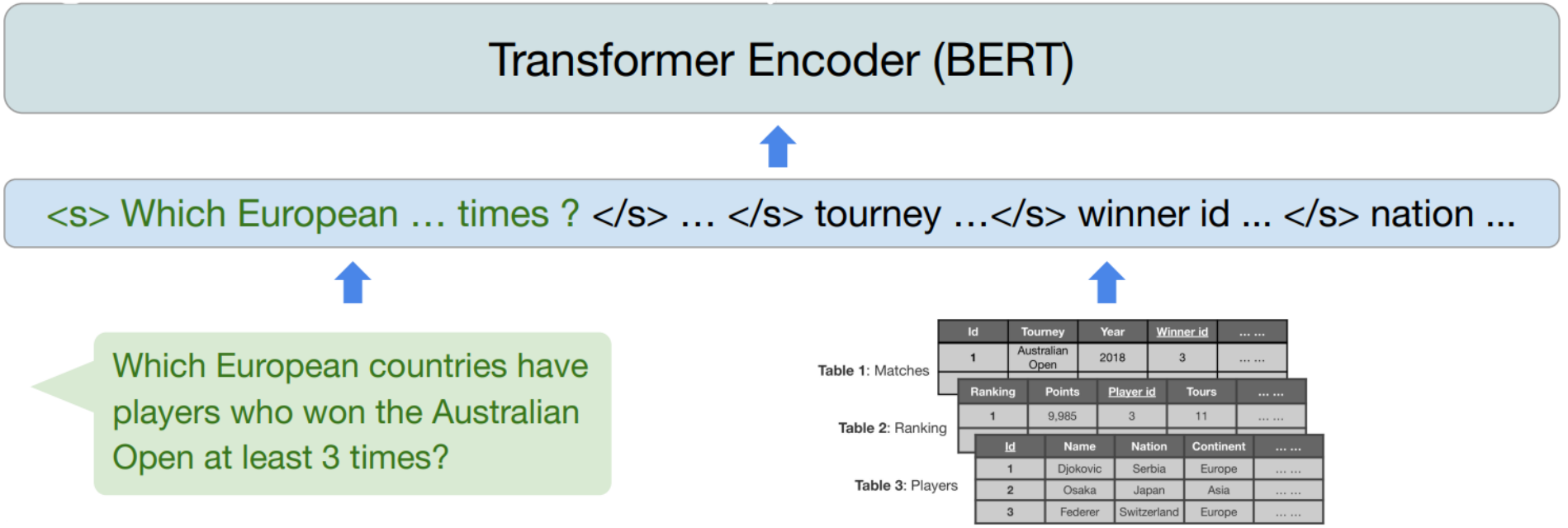
GraPPa：自动生成数据2

- 3. 利用同步文法在新表格的基础上生成新的（query, table, sql）数据



GraPPa：自监督学习任务1

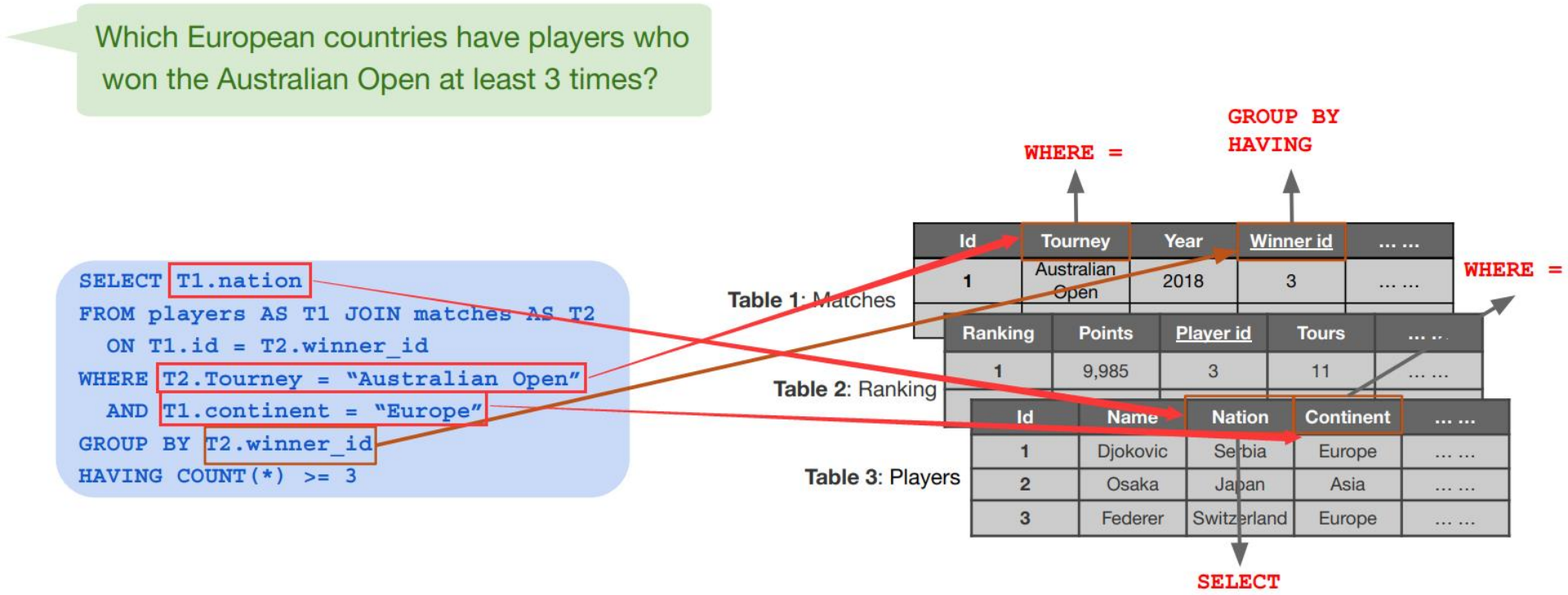
- Masked Language Model (MLM objective)
  - Mask both natural language words and table headers



只在数据1上计算loss

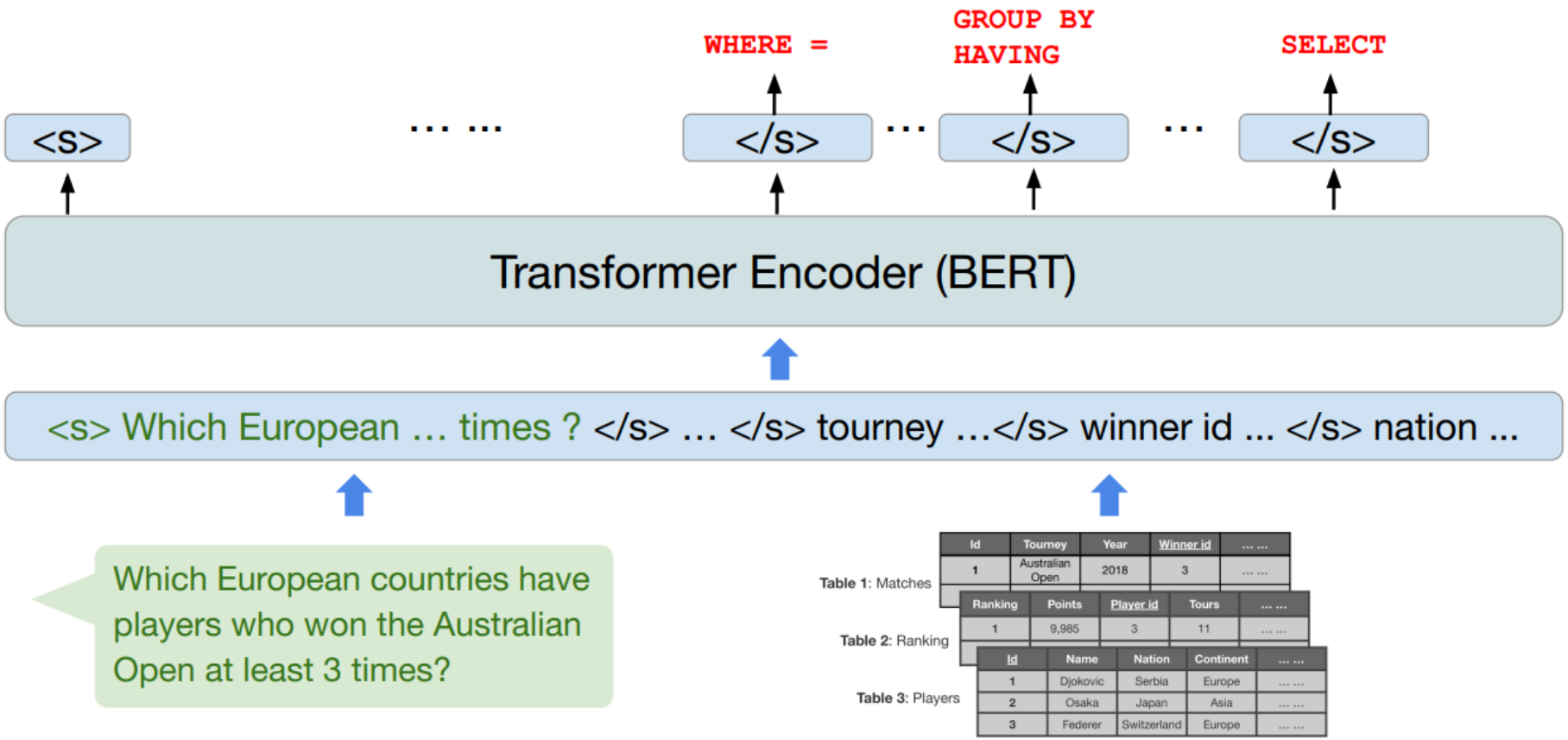
GraPPa：自监督学习任务2

- SQL semantic prediction (SSP objective)
  - 表头的SQL语义标签



GraPPa：自监督学习任务2

- SQL semantic prediction (SSP objective)



只在数据2上计算loss

GraPPa的实验结果

Models	Dev.	Test
Global-GNN (Bogin et al., 2019)	52.7	47.4
EditSQL w. BERT (Zhang et al., 2019b)	57.6	53.4
IRNet w. BERT (Guo et al., 2019)	61.9	54.7
RYANSQL w. BERT (Choi et al., 2020)	70.6	60.6
TranX w. TaBERT (Yin et al., 2020a)	64.5	-
RAT-SQL (Wang et al., 2019)	62.7	57.2
w. BERT-large	69.7	65.6
w. RoBERTa-large	69.57	-
w. GRAPPA (MLM)	71.08	-
w. GRAPPA (SSP)	73.57	67.72
w. GRAPPA (MLM+SSP)	73.43	**69.63**

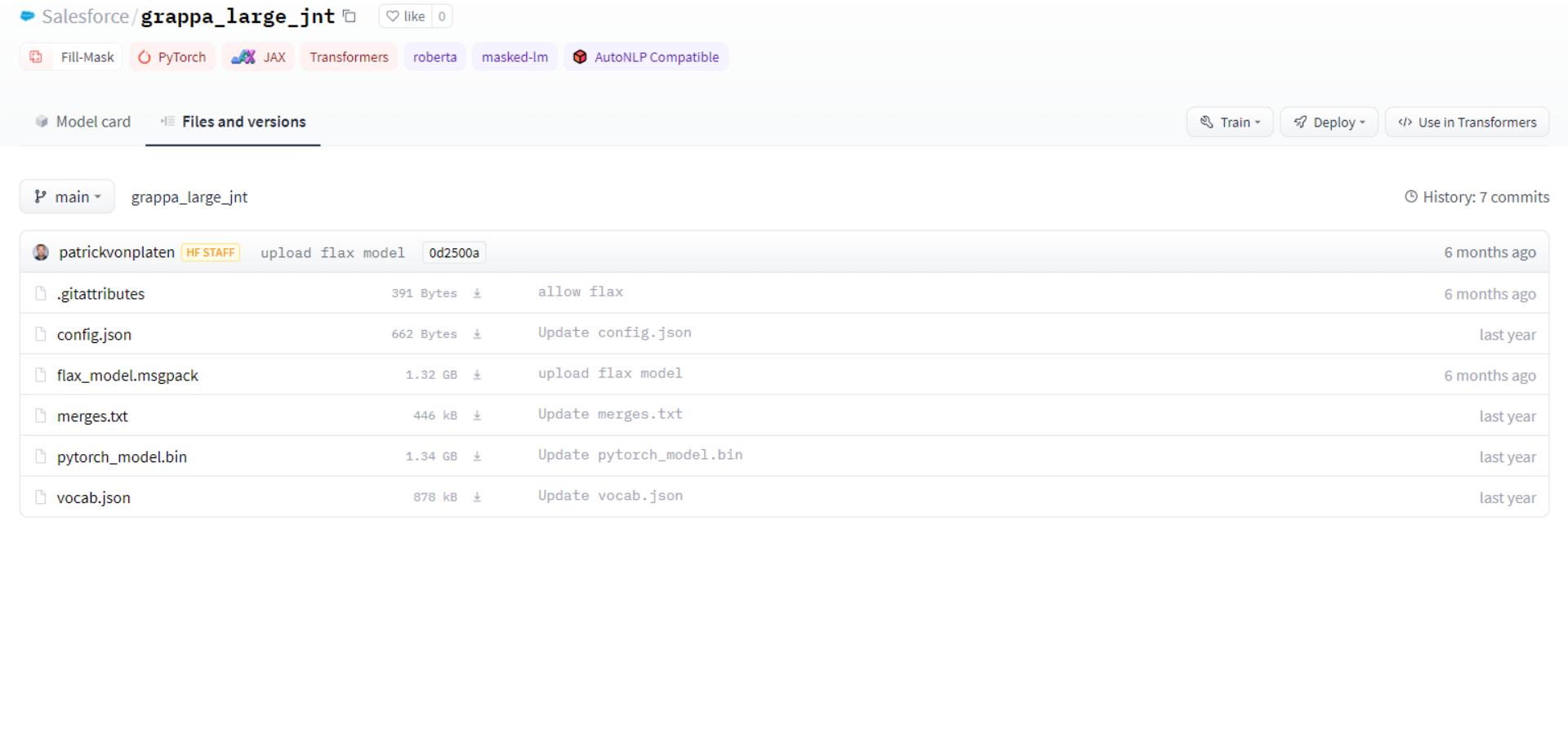
Table 3: Performance on SPIDER. We use RAT-SQL + BERT (Wang et al., 2019) as our base model. We run each model three times by varying random seeds, and the average scores are shown.

Models	Dev.	Test
Pasupat and Liang (2015)	37.0	37.1
Neelakantan et al. (2016)	34.1	34.2
Haug et al. (2017)	-	34.8
Zhang et al. (2017)	40.4	43.7
Liang et al. (2018)	42.3	43.1
Dasigi et al. (2019)	42.1	43.9
Agarwal et al. (2019)	43.2	44.1
Herzig et al. (2020b)	-	48.8
Yin et al. (2020b)	52.2	51.8
Wang et al. (2019)	43.7	44.5
w. RoBERTa-large	50.7(+7.0)	50.9(+6.4)
w. GRAPPA (MLM)	51.5(+7.8)	51.7(+7.2)
w. GRAPPA (SSP)	51.2(+7.5)	51.1(+6.6)
w. GRAPPA (MLM+SSP)	**51.9(+8.2)**	**52.7(+8.2)**
w. RoBERTa-large $\times$10%	37.3	38.1
w. GRAPPA (MLM+SSP) $\times$10%	**40.4(+3.1)**	**42.0(+3.9)**

Table 5: Performance on WIKITABLEQUESTIONS. We use Wang et al. (2019) as a base model. Results trained on 10% of the data are shown at the bottom.

Pre-trained model on Huggingface

- https://huggingface.co/Salesforce/grappa_large_jnt/tree/main



PLMS+CONSTRAINED DECODING

PLMs

- Formulating tasks as text-to-text generation problems:
- Large pre-trained language models (like T5) have shown increasingly impressive performance in a variety of NLP tasks

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

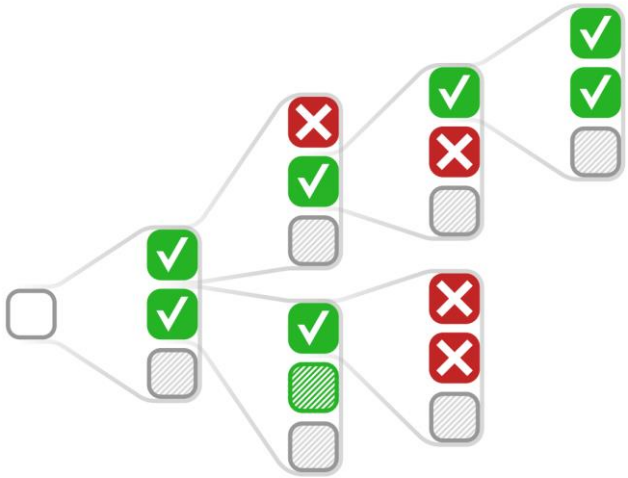
PLMs for Semantic Parsing

Two challenges:

- Meaning representations is grammatical
  - In an unconstrained output space, 10,000s of sub-word tokens can be produced each decoding step.
- Output sequences are much further from the pre-training distribution
  - People living in Beijing
  - ```
    SELECT * FROM Persons WHERE City='Beijing'
    ```
    ```

PLMs + constrained decoding 代表性工作

- **SSD** [Wu et al., ACL-2021]
- **Constrained-GPT-3** [Shin et al., EMNLP-2021]
- PICARD [Scholak et al., EMNLP-2021]
- Prompt-T5 [Schucher et al., Arxiv-2021]



Shan Wu, Bo Chen, Chunlei Xin, Xianpei Han, Le Sun, Weipeng Zhang, Jiansong Chen, Fan Yang, Xunliang Cai. From Paraphrasing to Semantic Parsing: Unsupervised Semantic Parsing via Synchronous Semantic Decoding. ACL-2021.

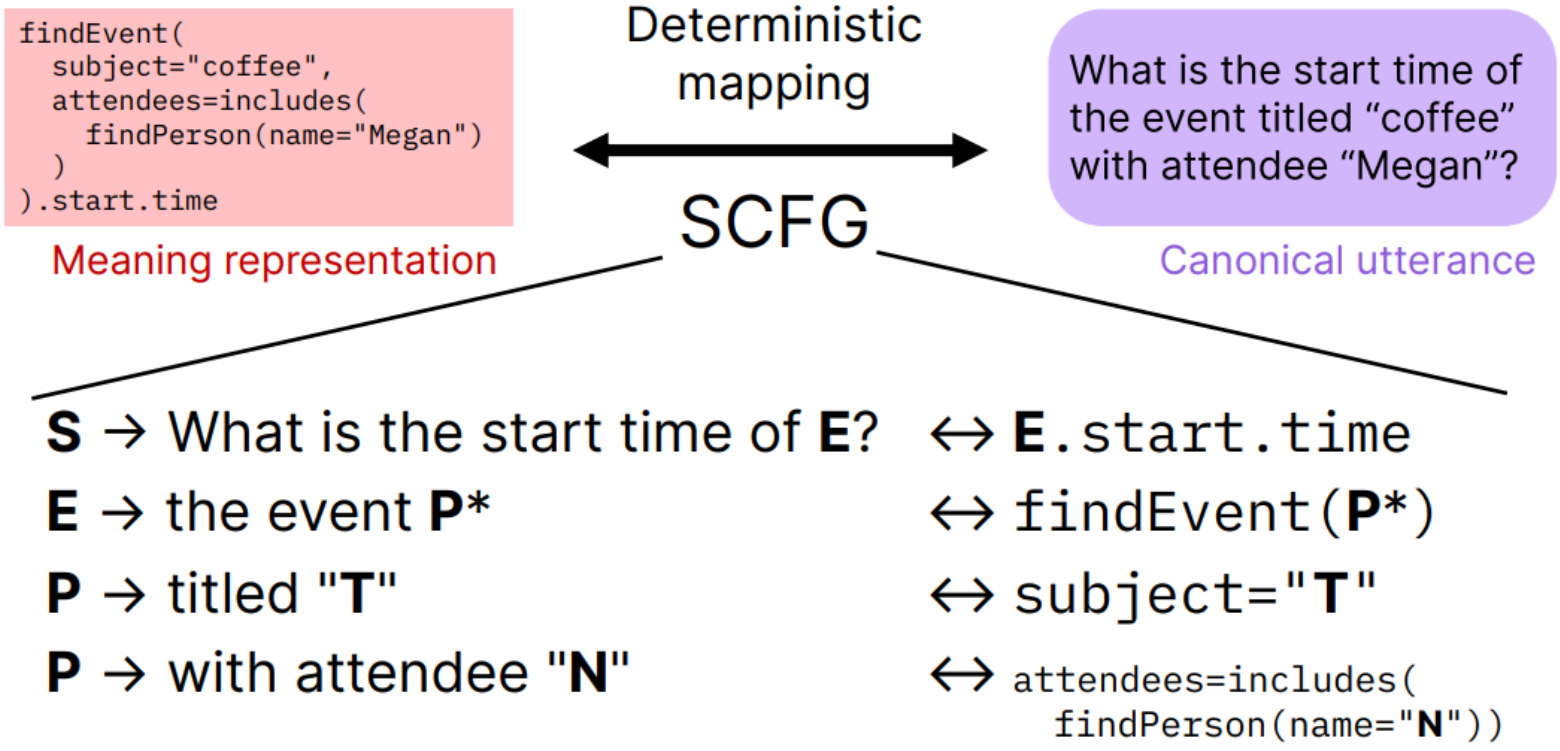
Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, Benjamin Van Durme. Constrained Language Models Yield Few-Shot Semantic Parsers. EMNLP-2021.

Torsten Scholak, Nathan Schucher, Dzmitry Bahdanau. PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models. EMNLP-2021.

Nathan Schucher, Siva Reddy, Harm de Vries. The Power of Prompt Tuning for Low-Resource Semantic Parsing. Arxiv-2021.

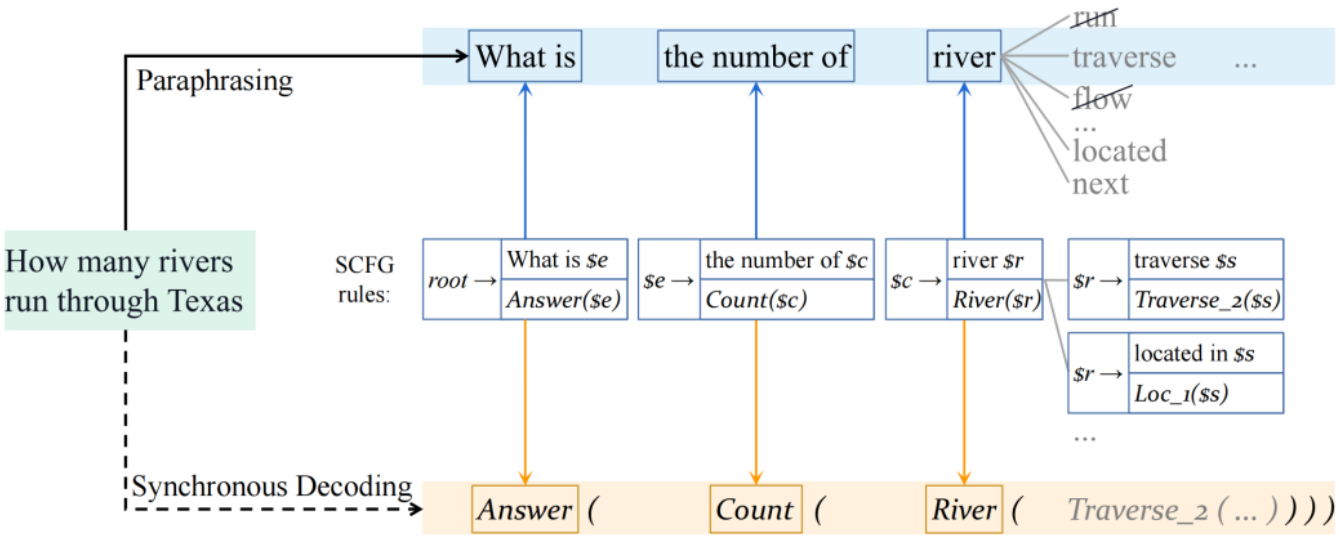
Synchronous Semantic Decoding

- **Canonical Utterance**: pseudo-language representations of logical forms, which have the synchronous structure of logical forms



Controlled Paraphrasing

- The core idea is reformulating semantic parsing into **controlled paraphrasing**.

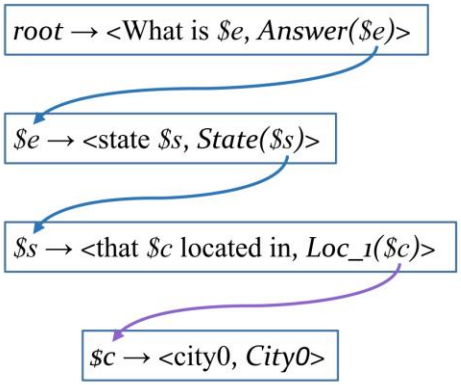


- The sentence is paraphrased to the canonical utterance and semantic parsed to the logical form synchronously.

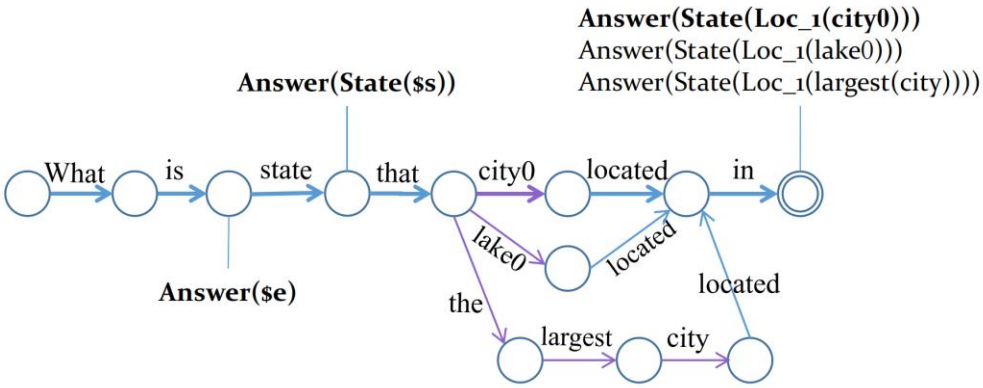
Shan Wu, Bo Chen, Chunlei Xin, Xianpei Han, Le Sun, Weipeng Zhang, Jiansong Chen, Fan Yang, Xunliang Cai. From Paraphrasing to Semantic Parsing: Unsupervised Semantic Parsing via Synchronous Semantic Decoding. ACL-2021.

Inference Algorithms

- How to decoding **valid** canonical utterance by text generation model?



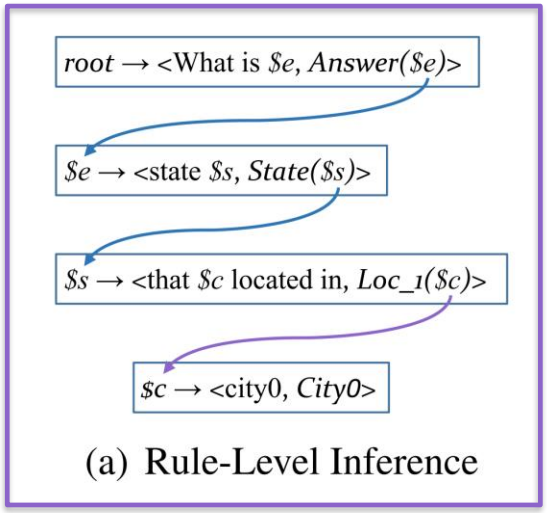
(a) Rule-Level Inference



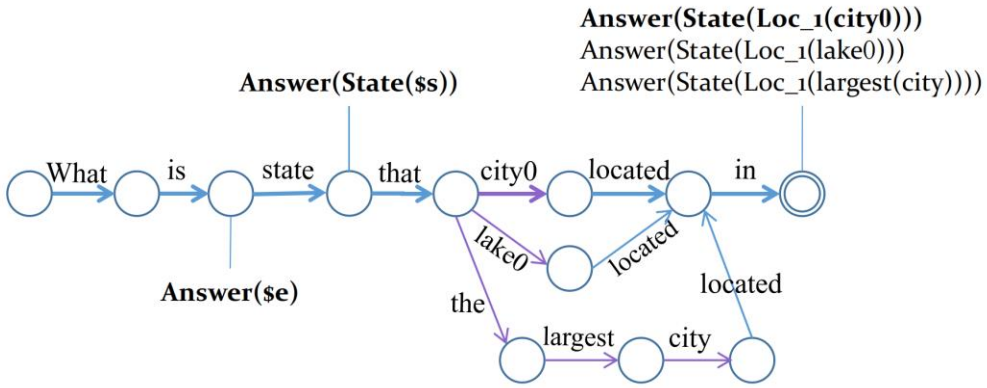
(b) Word-Level Inference

Inference Algorithms

- How to decoding **valid** canonical utterance by text generation model?



(a) Rule-Level Inference

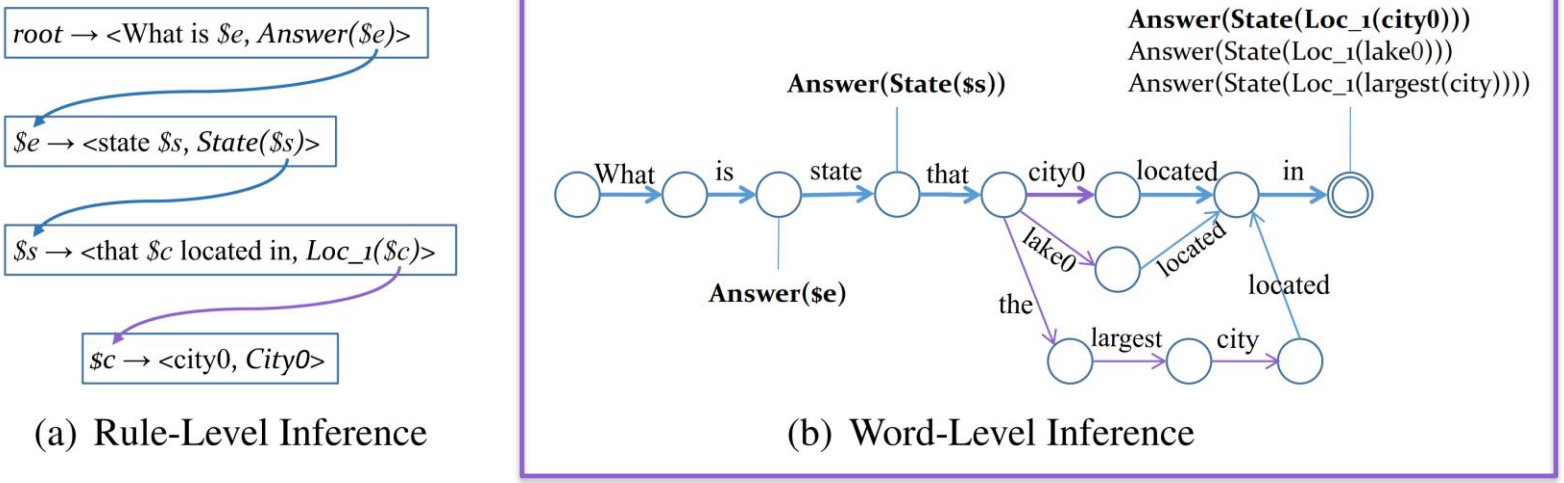


(b) Word-Level Inference

- Rule-level inference: handle non-terminal *$c* by generating the next production rule to expand this rule, until no non-terminal on the left of words, or the generating step reaches the depth of *K*.

Inference Algorithms

- How to decoding **valid** canonical utterance by text generation model?



- Word-level inference: construct the automaton by LR(1) parser. Only the acceptable words in the this state can be generated, and the <EOS> symbol can only be generated when reaching *the final state*.

Language Style Transfer

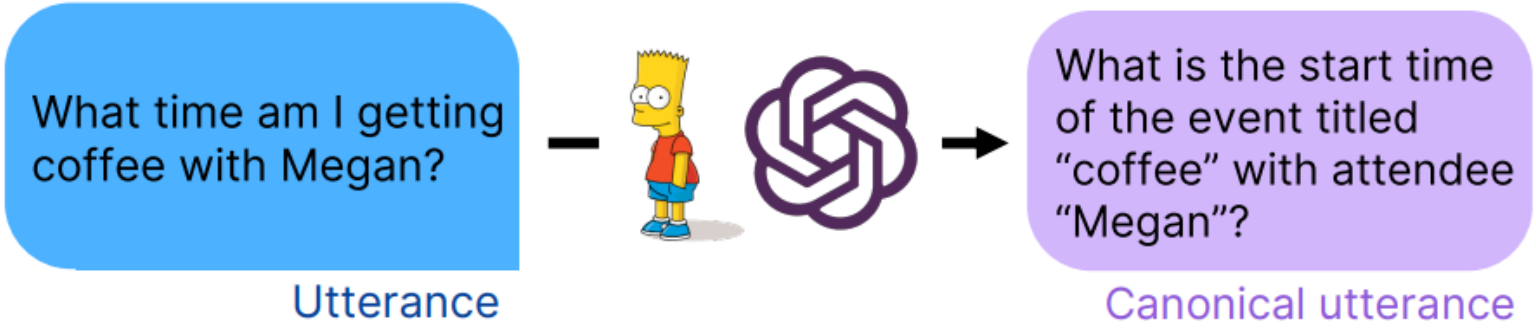
- Style bias:
  - CU: Meetings held in the same place as the weekly standup meeting
  - NL: Meeting whose location is location of weekly standup
- Adaptive Fine-tuning
  - Canonical utterances sampled from SCFG
  - De-stylized paraphrases
  - External paraphrases
- Utterance Reranking
  - Reconstruction score
  - Association score

Experiments

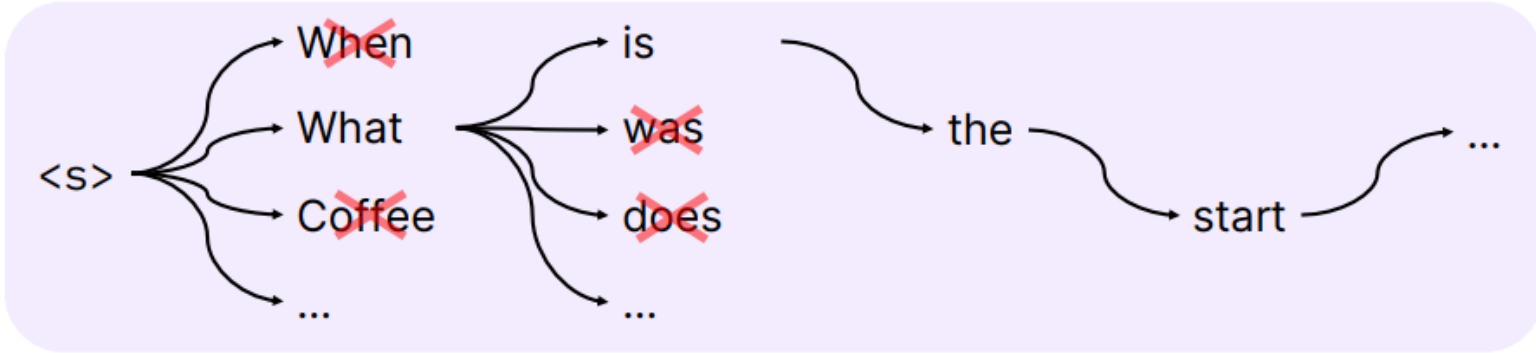
	Bas.	Blo.	Cal.	Hou.	Pub.	Rec.	Res.	Soc.	Avg.
Supervised									
RECOMBINATION (Jia and Liang, 2016)	85.2	58.1	78.0	71.4	76.4	79.6	76.2	81.4	75.8
CROSSDOMAIN (Su and Yan, 2017)	86.2	60.2	79.8	71.4	78.9	84.7	81.6	82.9	78.2
SEQ2ACTION (Chen et al., 2018b)	88.2	61.4	81.5	74.1	80.7	82.9	80.7	82.1	79.0
DUAL (Cao et al., 2019)	87.5	63.7	79.8	73.0	81.4	81.5	81.6	83.0	78.9
TWO-STAGE (Cao et al., 2020)	87.2	65.7	80.4	75.7	80.1	86.1	82.8	82.7	**80.1**
SSD (Word-Level)	86.2	64.9	81.7	72.7	82.3	81.7	81.5	82.7	79.2
SSD (Grammar-Level)	86.2	64.9	81.7	72.7	82.3	81.7	81.5	82.7	79.0
Unsupervised (with nonparallel data)									
TWO-STAGE (Cao et al., 2020)	64.7	53.4	58.3	59.3	60.3	68.1	73.2	48.4	60.7
WMDSAMPLES (Cao et al., 2020)	31.9	29.0	36.1	47.9	34.2	41.0	53.8	35.8	38.7
SSD-SAMPLES (Word-Level)	71.7	58.7	60.1	61.7	57.6	64.3	70.9	46.0	61.4
SSD-SAMPLES (Grammar-Level)	71.3	58.8	60.6	62.2	58.8	65.4	71.1	49.1	**62.2**
Unsupervised									
Cross-domain Zero Shot	-	28.3	53.6	52.4	55.3	60.2	61.7	-	-
GENOVERNIGHT	15.6	27.7	17.3	45.9	46.7	26.3	61.3	9.7	31.3
SYNTH-SEQ2SEQ	16.1	23.6	16.1	30.2	36.6	26.9	43.1	9.2	25.2
SYNTHPARA-SEQ2SEQ	28.4	37.3	33.9	38.1	39.1	41.7	62.7	23.3	38.1
SSD (Word-Level)	68.3	54.9	51.2	55.0	54.7	60.2	65.4	33.6	55.4
SSD (Grammar-Level)	68.8	58.1	56.5	56.1	57.8	59.3	66.9	37.1	**57.6**

Table 1: Overall results on OVERNIGHT.

Constrained GPT-3



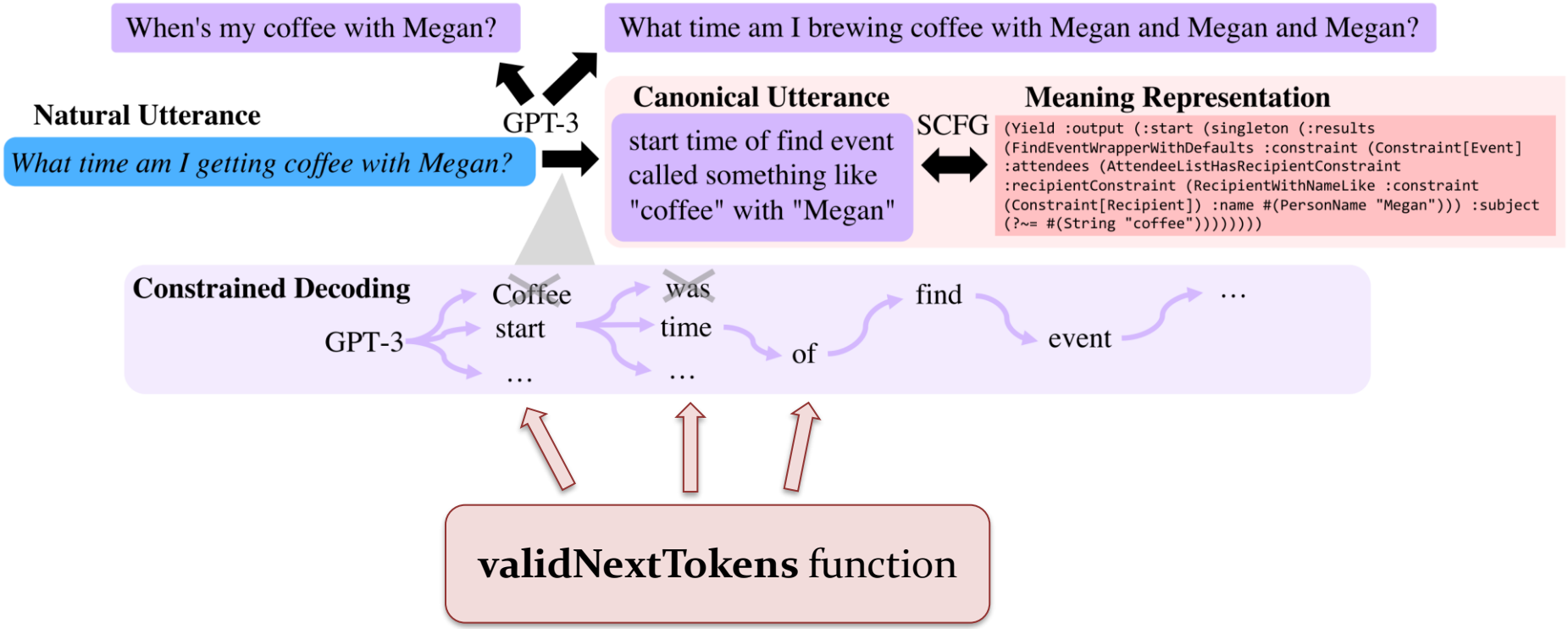
Paraphrasing into a controlled sublanguage



Constrained decoding

Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, Benjamin Van Durme. Constrained Language Models Yield Few-Shot Semantic Parsers. EMNLP-2021.

Constrained GPT-3



Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, Benjamin Van Durme. Constrained Language Models Yield Few-Shot Semantic Parsers. EMNLP-2021.

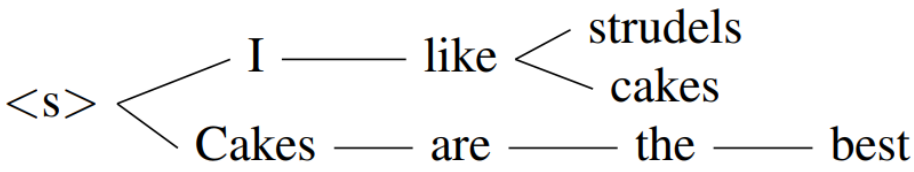
Case 1: Overnight

which january 2nd meetings is alice attenting [sic]

```
(call listValue  (call filter
  (call filter (call getProperty
      (call singleton en.meeting) (string !type))
    (string date) (string =) (date 2015 1 2))
  (string attendee) (string =) en.person.alice)
```

meeting whose date is jan 2 and whose attendee is alice

- **validNextTokens** function
  - build a large **trie** that contains all of the canonical utterance strings



Case 1: Overnight

Model	Train n	Basketball	Blocks	Calendar	Housing	Publications	Recipes	Restaurants	Social
GPT-3 Constrained Canonical	200	0.859	**0.634**	**0.792**	**0.741**	**0.776**	0.792	**0.840**	**0.687**
BART^f Constrained Canonical	200	**0.864**	0.554	0.780	0.672	0.758	**0.801**	0.801	0.666
GPT-2^f Constrained Canonical	200	0.836	0.540	0.766	0.666	0.715	0.764	0.768	0.623
Cao et al. (2019)	200	0.772	0.429	0.613	0.550	0.696	0.671	0.639	0.566
Cao et al. (2019)	640–3535	**0.880**	**0.652**	**0.807**	**0.767**	**0.807**	**0.824**	**0.840**	**0.838**
BERT-LSTM (Xu et al., 2020)	640–3535	0.875	0.624	0.798	0.704	0.764	0.759	0.828	0.819
AutoQA (Xu et al., 2020)	>400,000†	0.739	0.549	0.726	0.709	0.745	0.681	0.786	0.615

Model	Train n	Basketball	Blocks	Calendar	Housing	Publications	Recipes	Restaurants	Social
GPT-3 Constrained Canonical	200	**0.80***	**0.62***	**0.82***	**0.71***	0.79*	**0.84***	**0.89***	**0.72***
GPT-3 Constrained Meaning	200	0.68*	0.53*	0.68*	0.58*	0.63*	0.75*	0.78*	0.63*
GPT-3 Unconstrained Canonical	200	0.76*	0.46*	0.68*	0.56*	0.58*	0.74*	0.74*	0.55*
GPT-3 Unconstrained Meaning	200	0.56*	0.39*	0.50*	0.42*	0.46*	0.66*	0.58*	0.48*
GPT-3 Constrained Canonical	20	**0.80***	0.55*	0.67*	0.68*	**0.81***	0.60*	0.76*	0.67*
BART^f Constrained Canonical	200	**0.864**	**0.554**	**0.780**	**0.672**	**0.758**	**0.801**	**0.801**	**0.666**
BART^f Constrained Meaning	200	0.834	0.499	0.750	0.619	0.739	0.796	0.774	0.620
BART^f Unconstrained Canonical	200	0.852	0.539	0.726	0.656	0.714	0.773	0.756	0.585
BART^f Unconstrained Meaning	200	0.813	0.476	0.732	0.566	0.696	0.778	0.720	0.536
GPT-2^f Constrained Canonical	200	0.836	0.540	0.766	0.666	0.715	0.764	0.768	0.623
GPT-2^f Constrained Meaning	200	0.760	0.479	0.736	0.571	0.645	0.699	0.660	0.606

Case 2: Break

What color are a majority of the objects?

```
1. objects
2. colors of #1
3. number of #1 for each #2
4. #2 where #3 is highest
```

(colors of (objects)) where (number of (objects for each (colors of (objects))) is highest)

- **validNextTokens** function
  - words or their inflections that appear in the questions,
  - the predefined set of function words
  - opening and closing parentheses.

Case 2: Break

Model	Train n	nem
Coleman & Reneau	44,321	0.42
Wolfson et al. (2020)	44,321	0.29
Arad & Sapir	44,321	0.16
GPT-3 Constrained Canonical	1,000	0.32*
GPT-3 Constrained Canonical	100	0.24*
GPT-3 Constrained Canonical	25	0.20*
GPT-3 Constrained Canonical	200	0.31*
GPT-3 Constrained Meaning	200	0.24*
GPT-3 Unconstrained Canonical	200	0.20*
GPT-3 Unconstrained Meaning	200	0.17*
GPT-3 Constrained Canonical	200	0.24
BART^f Constrained Canonical	200	0.22
BART^f Constrained Meaning	200	0.22
BART^f Unconstrained Canonical	200	0.18
BART^f Unconstrained Meaning	200	0.19

Case 3: SMCalFlow

What did I set as my response status for the team meeting?

```
(Yield :output
  (:responseStatus (singleton (:results
    (FindEventWrapperWithDefaults
      :constraint (Constraint[Event]
        :subject (?~= #(String “team meeting”)))))))))
```

my response status of find event called something like "team meeting"

- **validNextTokens** function
  - more SCFG-friendly MR
  - then use Earley parser

Case 3: SMCaFlow

Model	Train n	Accuracy
Semantic Machines et al. (2020)	133,821	0.73
GPT-3 Constrained Canonical	300	0.44*
GPT-3 Constrained Meaning	300	0.30*
GPT-3 Unconstrained Canonical	300	0.26*
GPT-3 Unconstrained Meaning	300	0.21*
GPT-3 Constrained Canonical	300	0.38
BART^f Constrained Canonical	300	0.47
BART^f Constrained Meaning	300	0.32
BART^f Unconstrained Canonical	300	0.40
BART^f Unconstrained Meaning	300	0.32

小结：基于预训练的语义解析方法

面向语义解析的预训练模型

- **核心**
  - 收集数据
  - 定义预训练任务
- 优点：
  - 如BERT一般可用于任何semantic parser
- 缺点：
  - 受限于数据，目前只在text-to-sql和code generation等情境中应用

直接利用已有的预训练模型

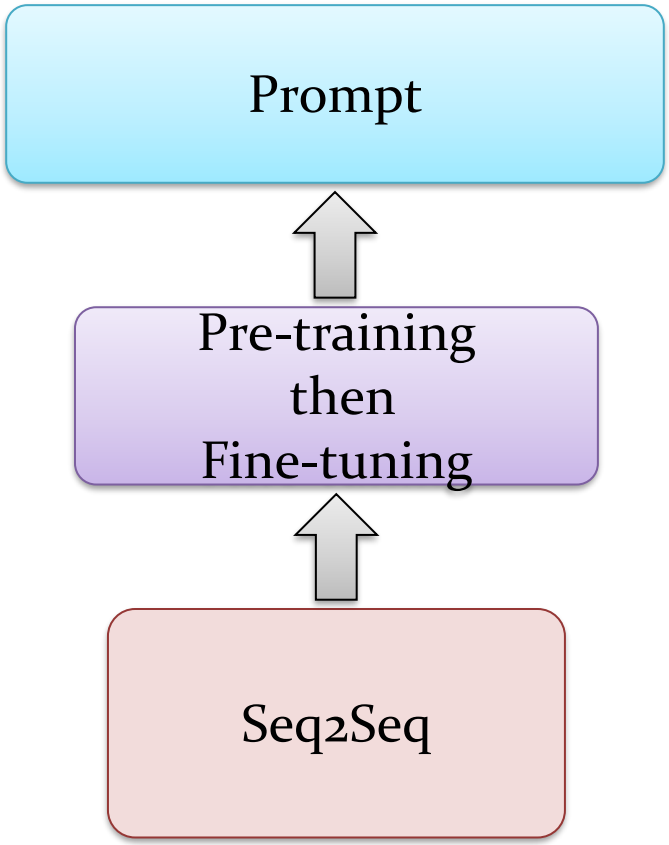
- **主流**
  - Controlled generation (paraphrasing + constrained decoding)
- 优点：
  - 不需要重新训练大模型
- 缺点：
  - 需要同步文法、需要人工定义约束条件

大纲

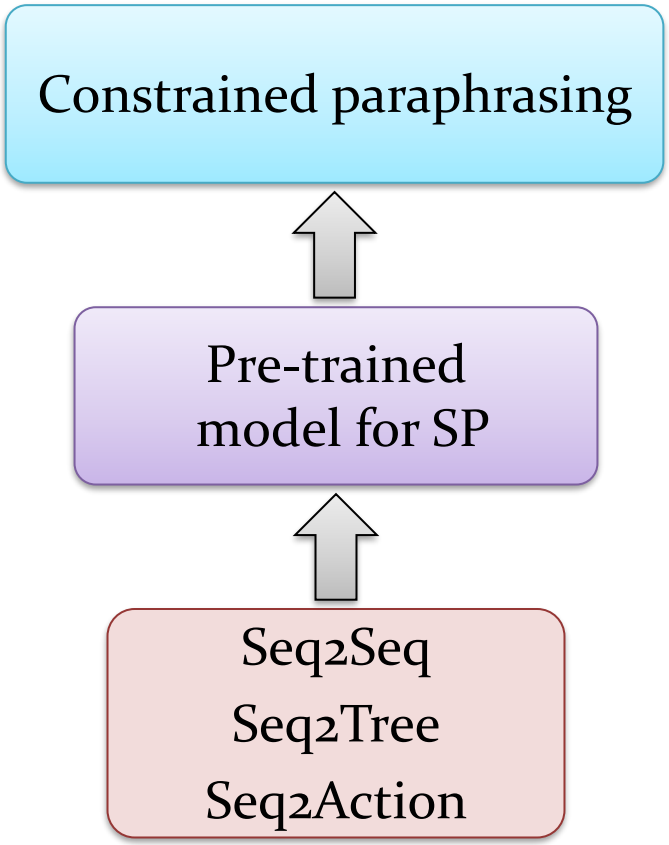
- 语义解析简介
  - 任务简介
  - 发展历程
- 基于深度学习的语义解析方法
  - Seq2Seq、Seq2Tree、Seq2Action
  - Constrained decoding
- 基于预训练的语义解析方法
  - 预训练方法在Text-to-SQL任务上的应用
  - PLMs with Constrained Decoding
- 总结与展望

总结

- 语义解析紧跟NLP大流



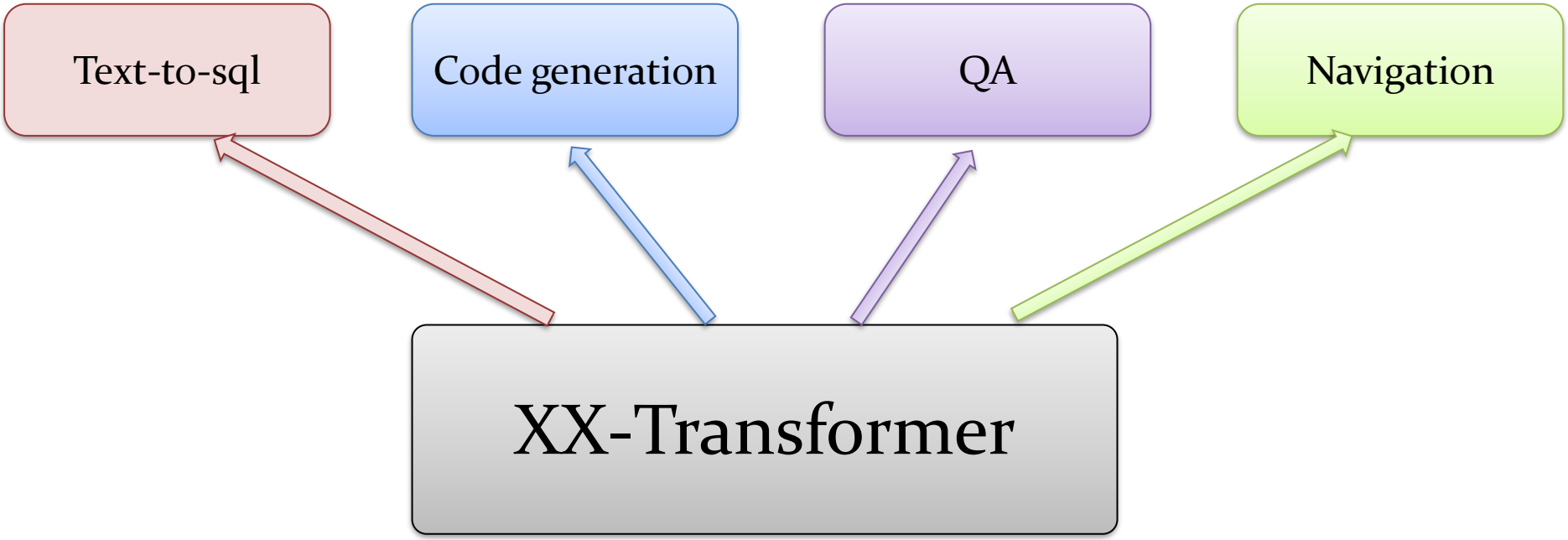
NLP发展之势



语义解析发展之势

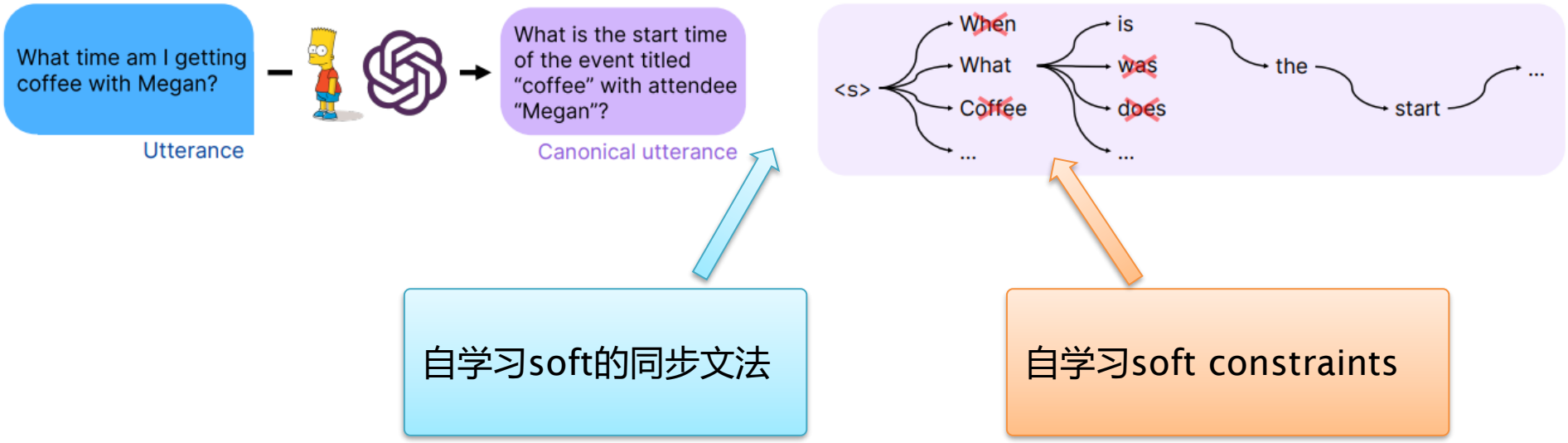
展望1：General Pre-trained Model for NLU

- Pre-trained model不局限于text-to-sql和code generation等情境，而是面向并同时面向更general的SP情境，如open-domain QA，language to instruction等。



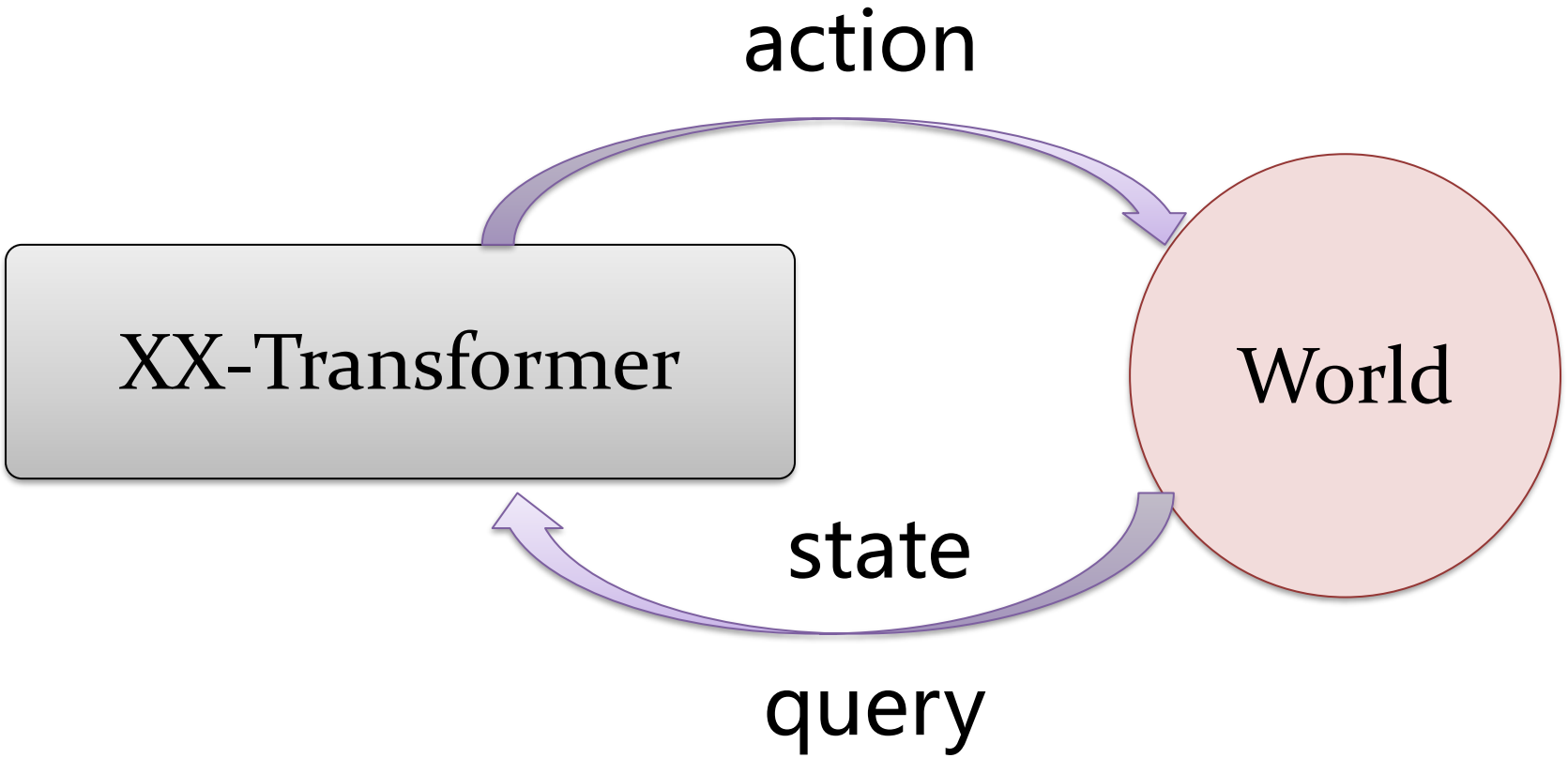
展望2：自学习的controlled generation

- PLMs + Constrained decoding已展示了威力，但constraints仍然需要人来参与，且需要人工定义同步文法。



展望3：状态感知的pre-trained model for NLU

- 目前的模型与世界没有交互，如何预训练一个面向NLU的能与世界进行交互的大模型



语义解析资源

- 

https://github.com/casnlu/Semantic-Parsing
  - 数据集
  - 平台
  - 博士论文
  - Tutorials
  - 顶会论文

敬请大家批评指正！

陈 波

实验室链接：http://www.icip.org.cn