



基于槽位共享跨度预测的多领域对话状态追踪模型

中国科学院新疆理化技术研究所
多语种信息技术研究室

艾比布拉·阿塔伍拉

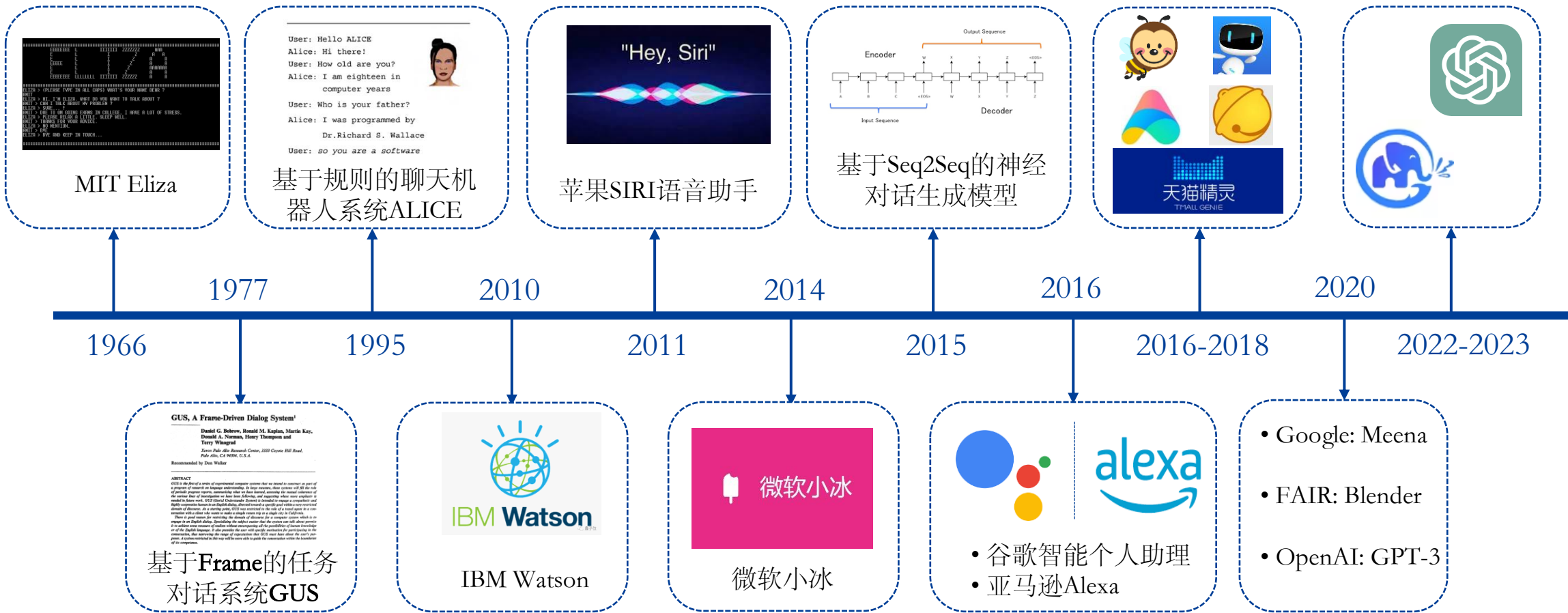
提纲

- 背景介绍
- SSNet模型与方法
- 浅谈未来方向

提纲

- 背景介绍
- SSNet模型与方法
- 浅谈未来方向

对话系统历史



□ 对话系统分类

对话系统是人与机器的一种新型交互方式，作为人工智能领域的重要研究内容，受到学术界和工业界的广泛关注。依据其任务类型不同，对话系统可以分为：智能问答系统、闲聊型对话系统和任务型对话系统。

任务型

有任务目标，系统主动权

通常面向特定的领域，通过与用户的交互，帮助用户便捷处理复杂任务。比如，预定餐厅、订机票等。

典型系统：阿里小蜜，百度小度等。

闲聊型

领域开放，不限定领域

与用户进行无缝且恰当的交谈，并提供有用的建议。

典型系统：微软小冰等。

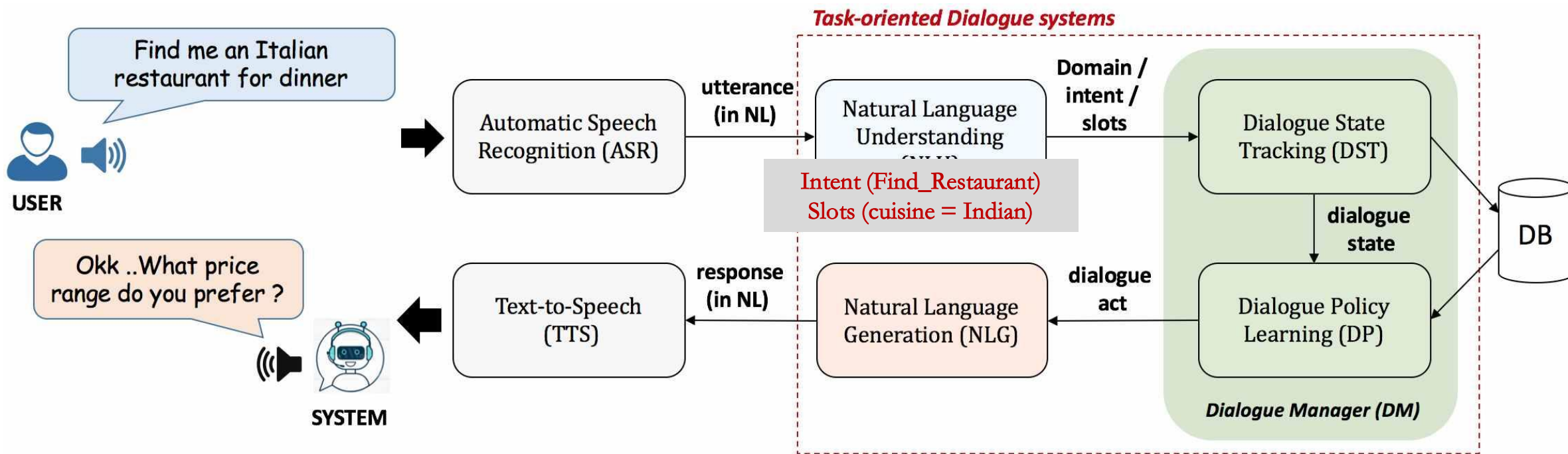
问答型

有任务目标，无参数化需求

根据来自不同数据源的丰富知识，为用户查询提供简明、直接的回答。

典型系统：百度AnyQ。

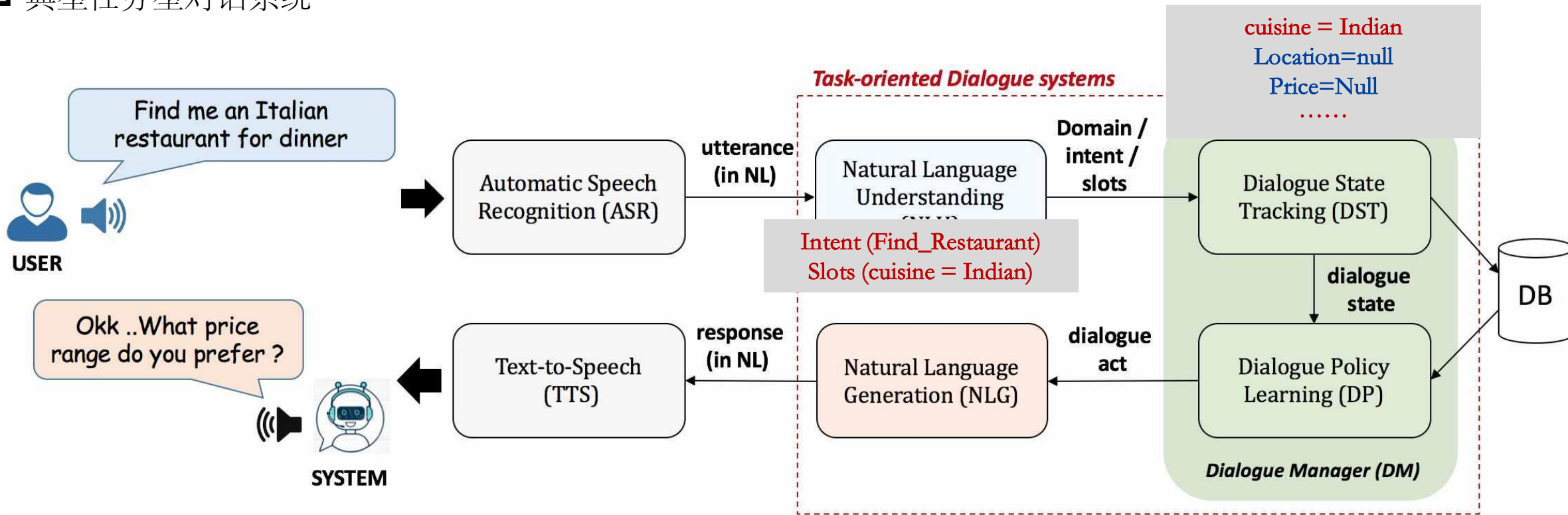
典型任务型对话系统



自然语言理解(NLU): 将输入的自然语言文本转换为结构化的语义框架，通常分为两个子任务：意图识别和语义槽位填充。

输入：用户语句 输出：语义框架 (intents和slots)

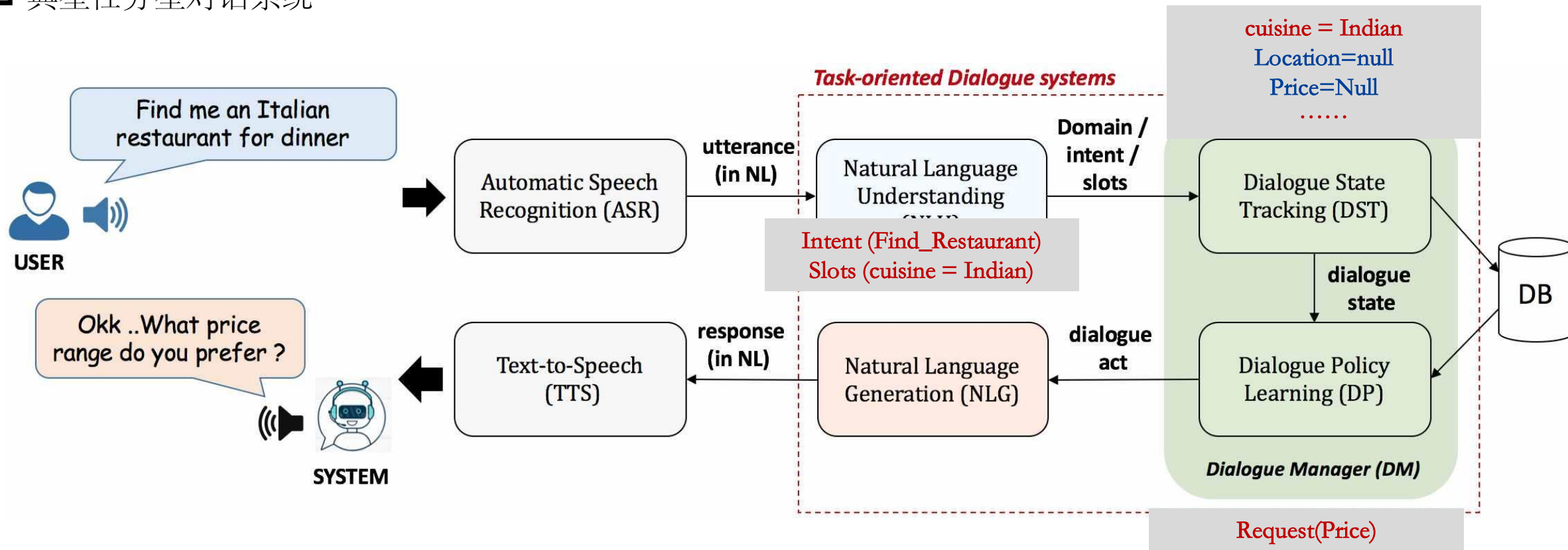
典型任务型对话系统



对话状态追踪(DST): 对话状态追踪任务的目标是在每轮对话回合中根据所有给定的对话上下文预测出当前轮次对话状态。

输入：一段对话/当前对话轮次+前轮对话状态 输出：当前轮次的对话状态

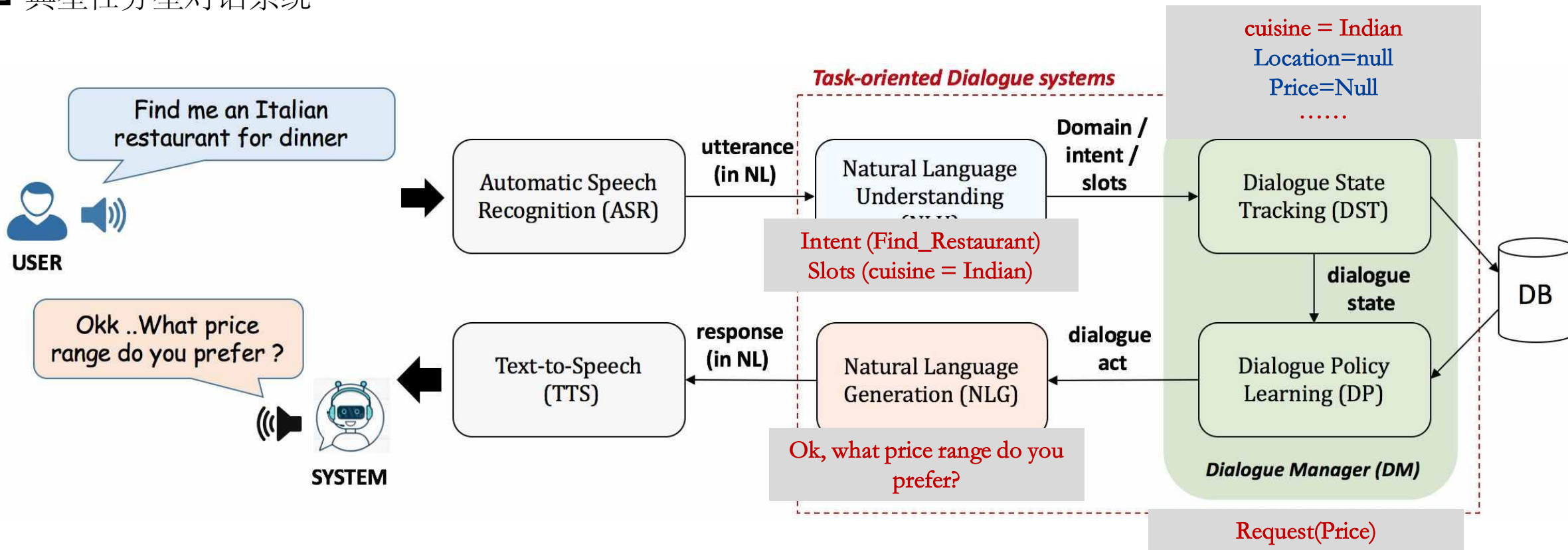
典型任务型对话系统



对话策略学习(DP): 根据当前状态选择下一个操作, 即决定系统下一步应该采取什么系统动作。

输入: 当前对话状态+KB结果 输出: 系统动作

典型任务型对话系统



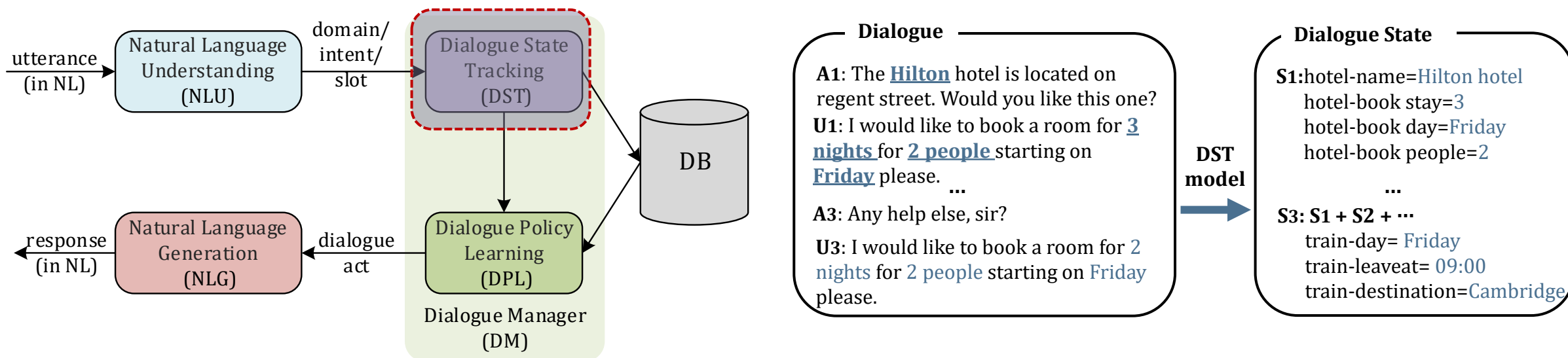
自然语言生成(NLG): 根据系统所决定的对话动作, 生成文本回复, 即将结构化语言转化为自然语言。

输入: 系统动作 输出: 系统恢复

对话状态追踪任务

对话状态是由领域本体指定的一组键值对组成的，键指的是“领域-槽位”，称之为槽位，值指的是对话上下文中需要提取的槽值，称之为槽值。

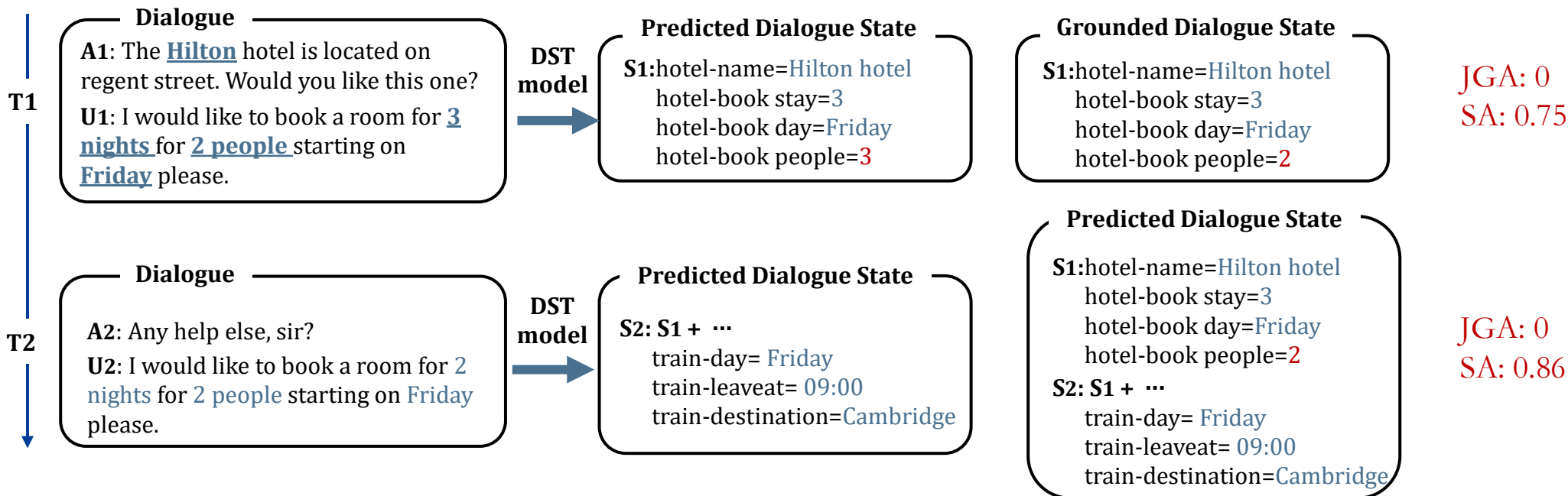
对话状态追踪的任务：目标是在每轮对话回合中根据所有给定的对话上下文预测出当前轮次对话状态。作为中间模块，对话状态追踪模型性能好坏直接影响后续的对话策略选择。



对话状态追踪任务—评价指标

槽位正确率 (Slot Accuracy, SA) :分别将预测的每个槽值对与真实标签进行比较, 正确为1, 否则为0。

联合目标准确率 (Joint Goal Accuracy, JGA) :指的是每个槽位的槽值被正确预测的对话会话轮次的比率, 即在每轮对话中, 只有当所有槽位和对应的槽值都预测正确时, 则该轮JGA为1, 否则为0。



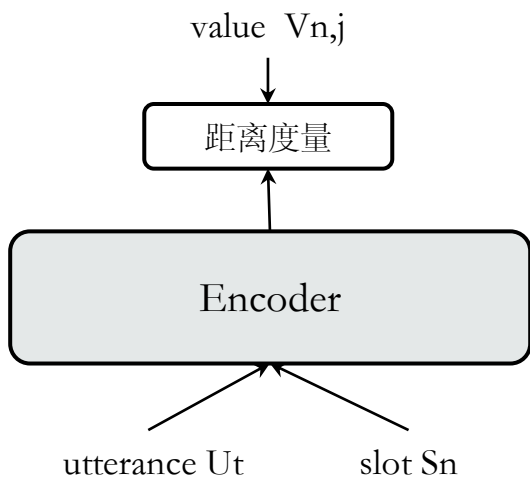
□ 对话状态追踪任务—相关数据集

Multi-domain Wizard-Of-Oz, MultiWOZ : 剑桥大学于2018年提出, 由出租车、餐厅、火车、景点、宾馆等7大领域组成。
数据集收集方法: 人类与人类(H2H)。

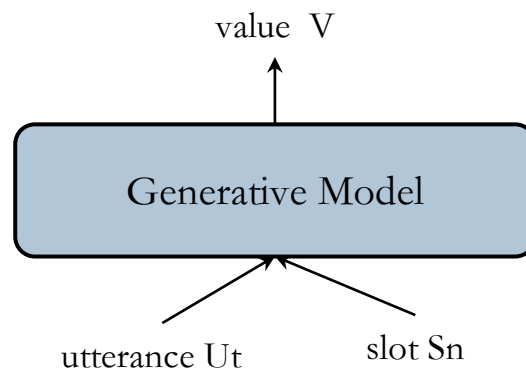
Schema-Guided Dialogue Dataset, SGD: 包含18000条的带注释的任务型对话, 涉及到宾馆、银行等17个领域。对于大多数领域, 数据集包含多个不同的API, 其中许多API 具有重叠的功能, 但接口不同, 这反映了典型的真实场景。数据集收集方法: 人类与机器(H2M)。

数据集	DSTC2	MultiWOZ	SGD
领域数量	1	7	16
对话数量	1612	8438	16142
对话平均轮次	14.5	13.7	20.4
对话状态表示	slot-value	slot-value	slot-value
槽位数量	8	25	214
槽值数量	212	4510	14139

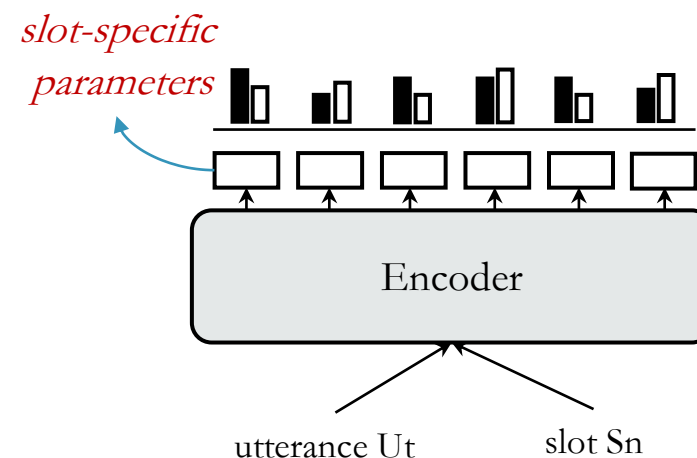
对话状态追踪方法学分类



分类式DST模型



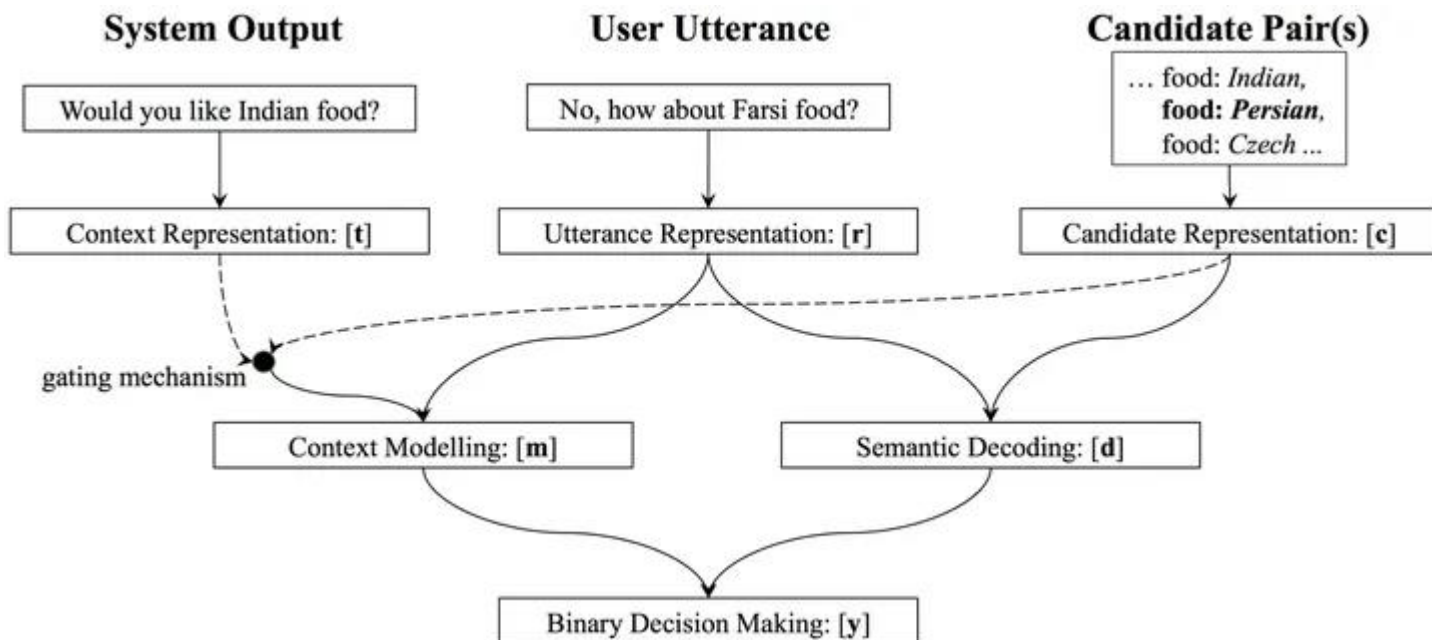
生成式DST模型



抽取式DST模型

典型DST模型—分类式模型

Mrkšić等人提出了NBT模型。该工作的主要动机是克服影响以前的状态跟踪模型的限制。NBT通过推理，对预先训练的单词向量进行建模，学习将其组合为用户话语和对话上下文的分布表示形式。



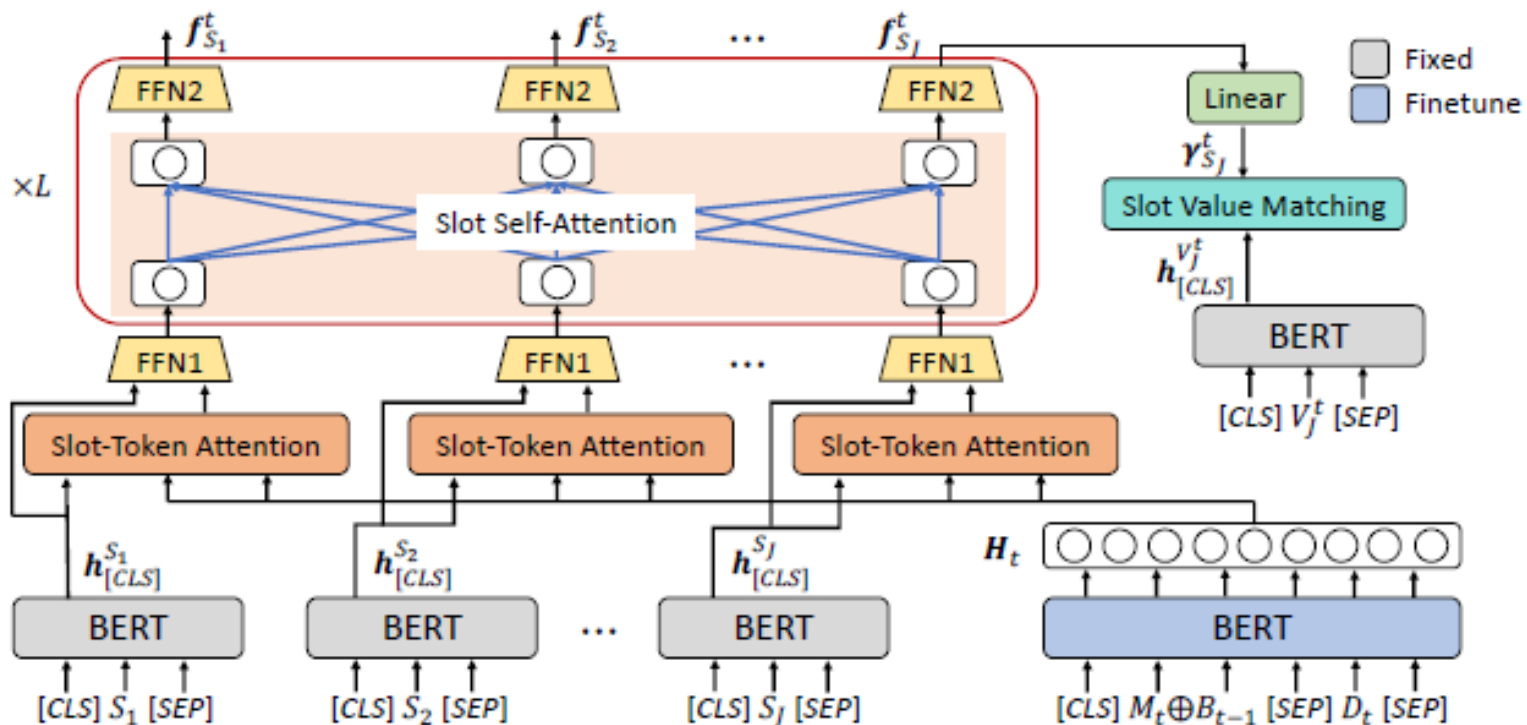
NBT模型框架

贡献:

1. 利用预训练词向量中的语义信息来解决词汇/形态上的歧义
2. 最大化本体值之间共享的参数数量
3. 具有学习领域特定释义和其他变体的灵活性，这使得依靠精确匹配和去词缀化作为一种可靠的策略是不可行的

典型DST模型—分类式模型

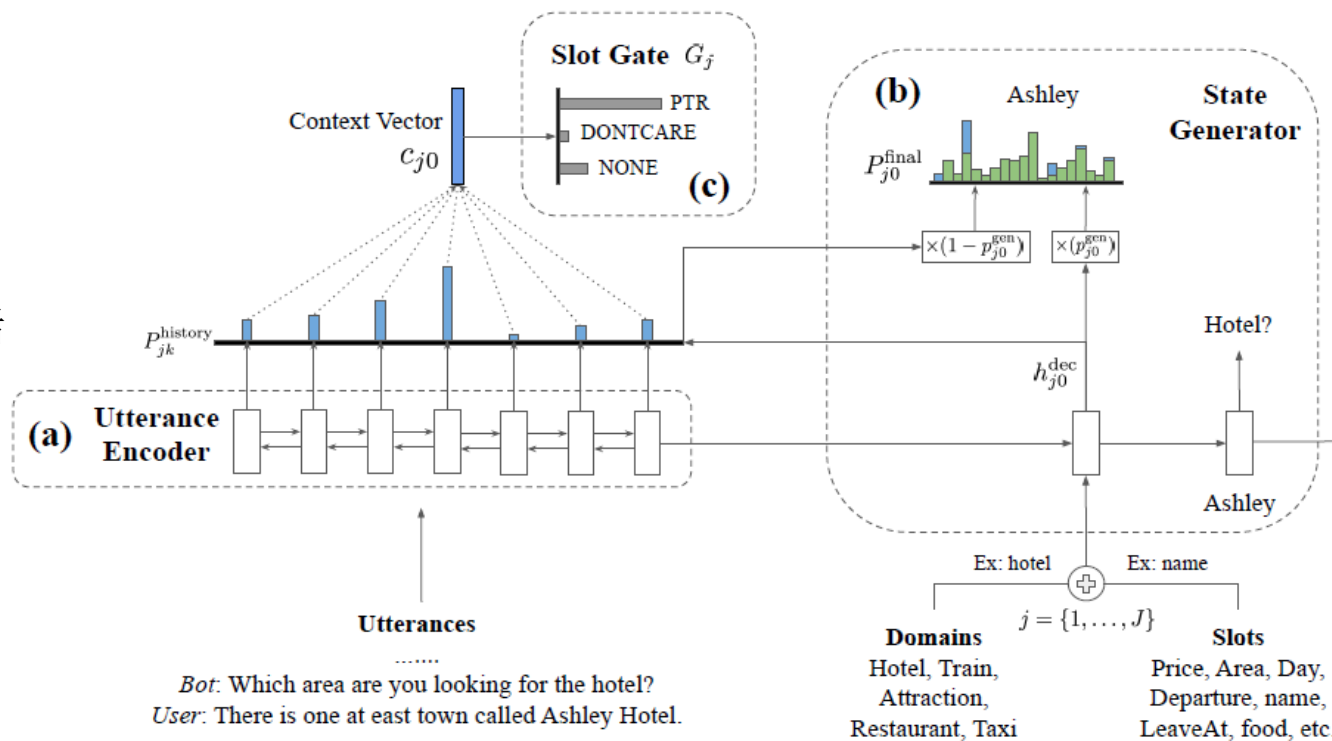
Fanghua等人提出了STAR模型，即通过自注意力机制，建模不同槽位之间的信息交互。



STAR模型框架

典型DST模型—生成式模型

Wu等人提出的TRADE模型结构，是最早的多领域对话状态追踪模型。将copy机制引入到生成式模型当中。

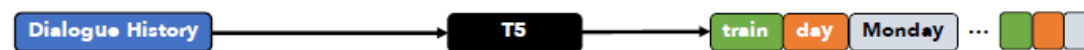


TRADE模型框架

□ 典型DST模型—生成式模型

右边是典型的基于生成式模型的对话状态追踪模型。

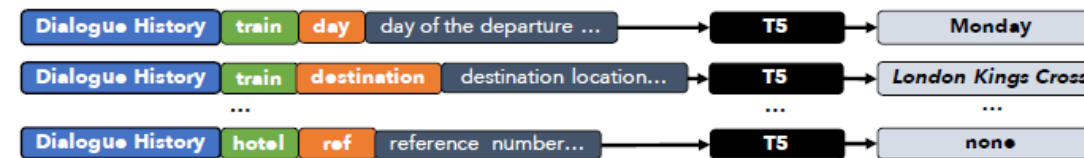
主要区别在于输入输出形式上的不同。



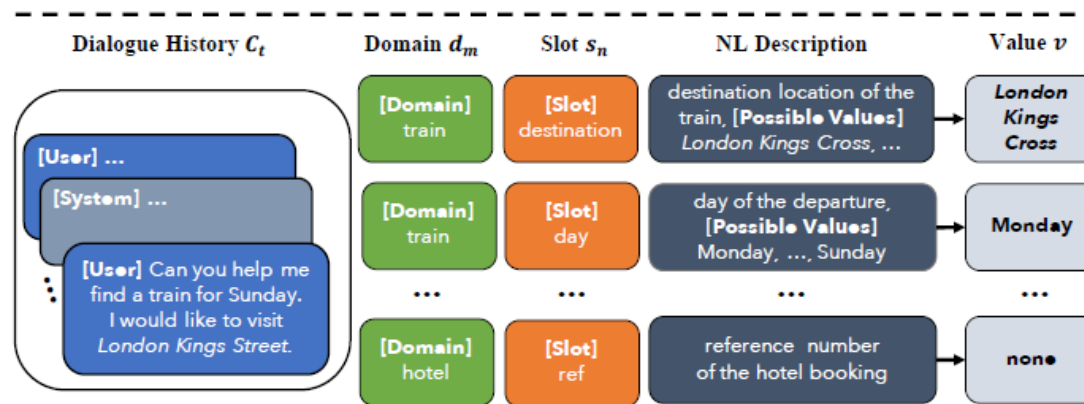
(a) Generation-based DST w/ Sequential Decoding



(b) Schema-Based Prompt DST w/ Independent Decoding

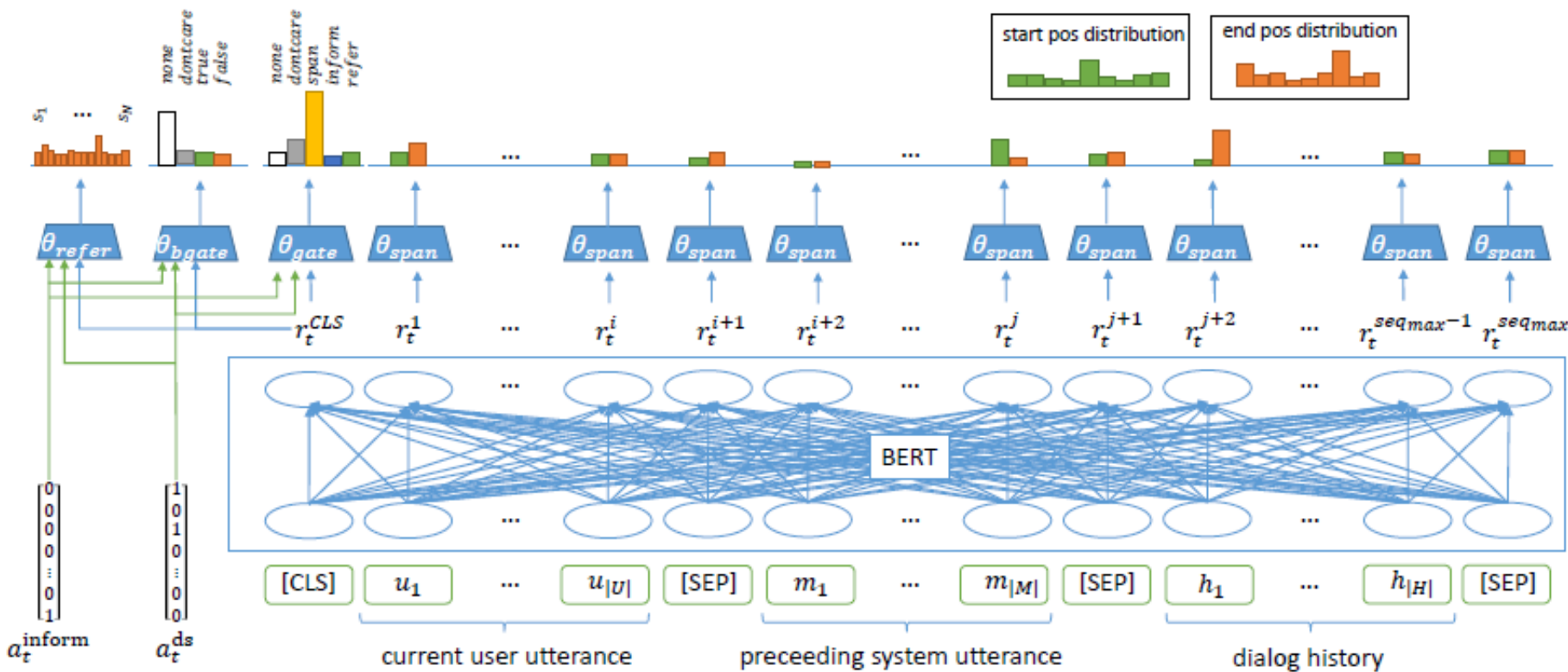


(c) Natural Language Augmented Prompt DST w/ Independent Decoding



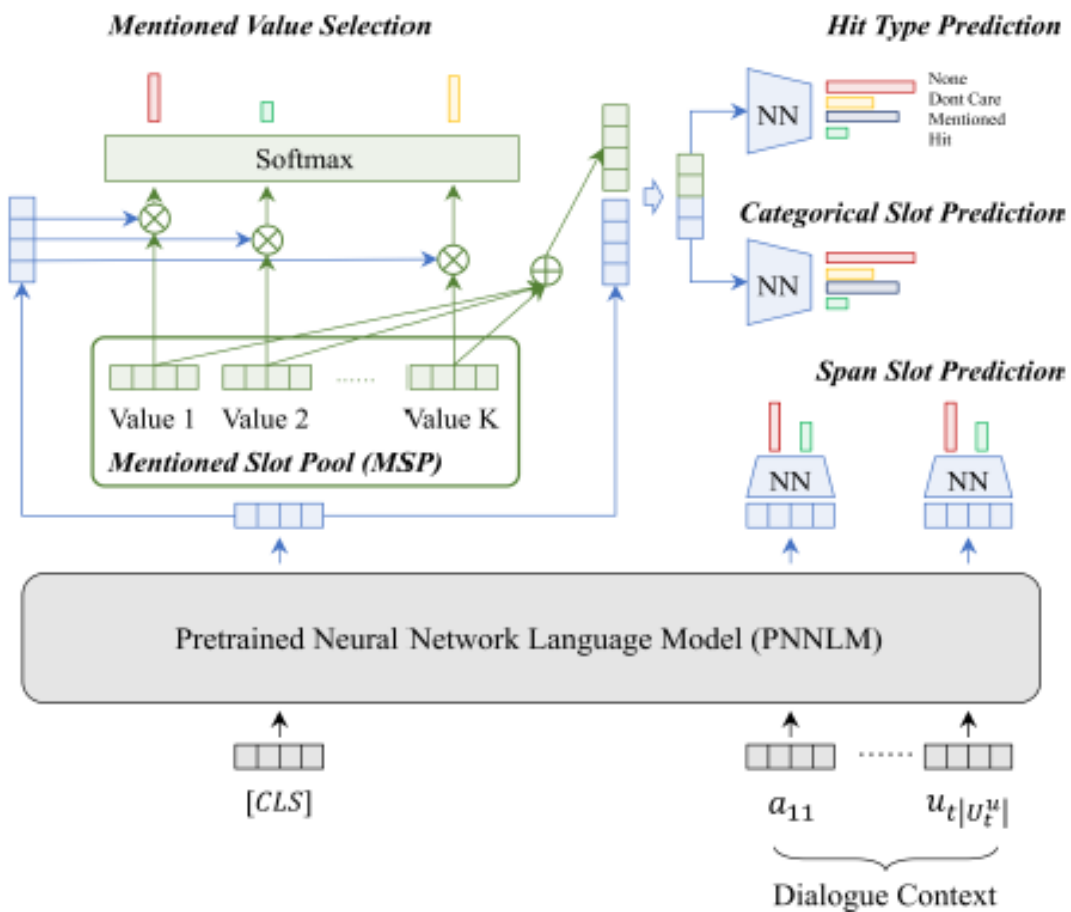
典型DST模型—抽取式模型

Trippy模型进一步优化copy机制，提出三重复制机制。即分别从对话上下文中取、系统提到的槽值中选择以及前期提到的指代槽值中直接拷贝等。



Trippy模型框架

典型DST模型—抽取式模型



MSP-DST模型框架

方法上跟Trippy类似，不同点在于构建了一个系统提到槽位库(MSP)来保存已经提到的槽和对应的槽值，后面出现相同槽位时，符合条件则直接继承。

提纲

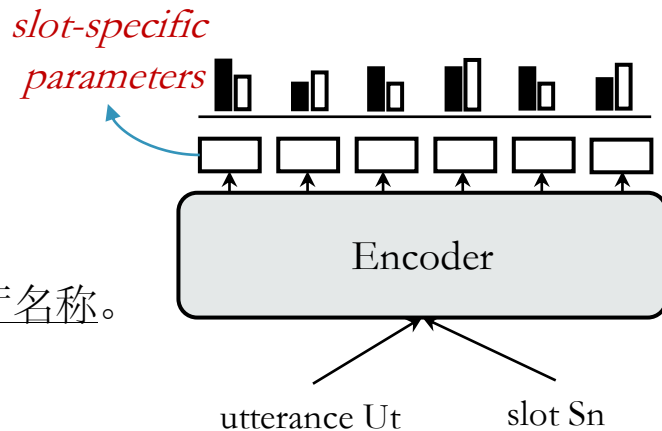
- 背景介绍
- SSNet模型与方法
- 浅谈未来方向

□ 现有方法缺陷

1. 分类式DST模型分类式方法依赖于前期获取的领域本体，无法提取未知槽值。
2. 生成式DST模型由于生成式模型固有特性，生成式方法难以控制生成的字符串。
3. 抽取式DST模型相对稳定，但现有抽取式模型采用槽位独立的槽值提取方法，即针对每一种槽位配置一个槽值提取层。



- (1) 由于提取层数量和槽位数量相关，导致模型可扩展性差。
- (2) 忽略槽位之间的信息交互。这种信息交互包括：
 - A. 不同的槽位可能会共享相同槽值。例如：出租车目的地的槽值可能是餐厅名称。
 - B. 部分槽位有同时出现的可能。例如：出租车目的地和出租车出发地



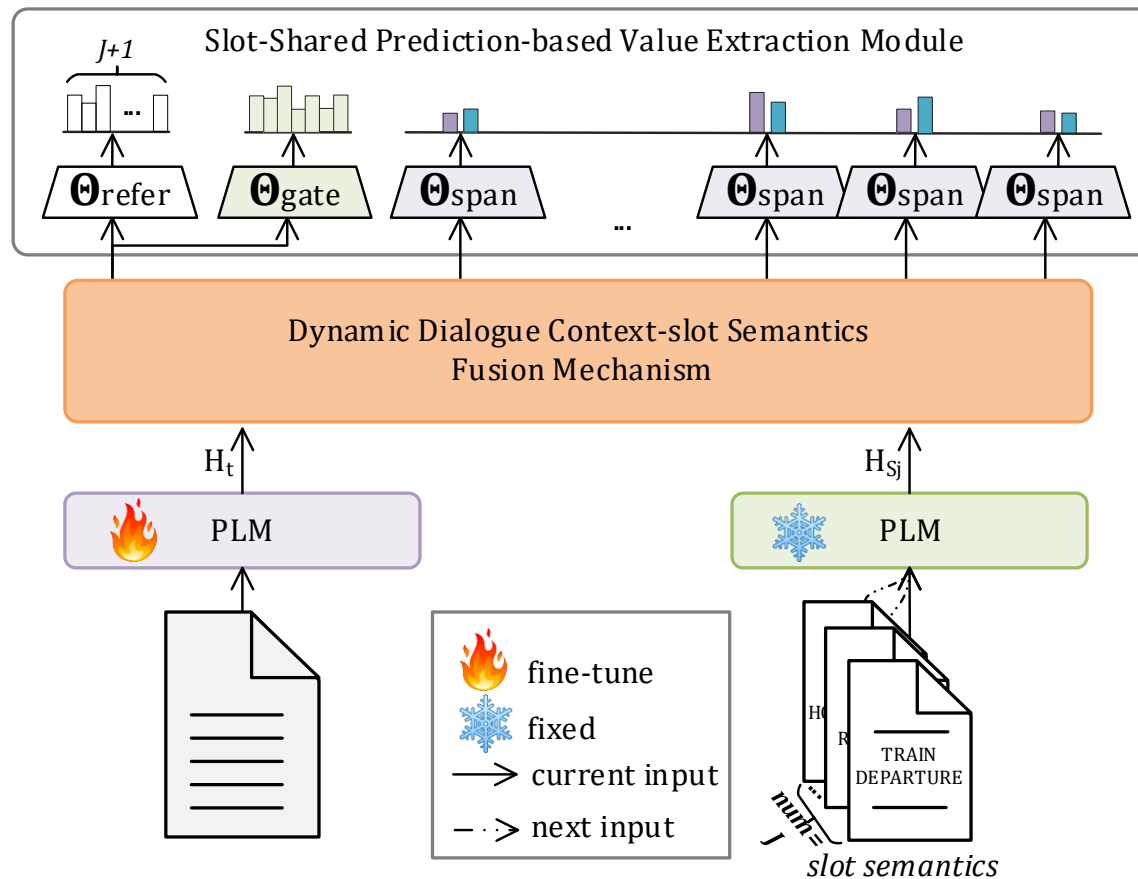
□ 目的与方案

目的：将所有槽位共享同一个槽值提取层，缓解模型扩展性差的问题。

方案：将槽位特征信息作为标识，并将其融合到对话上下文，以区分不同槽位。

SSNet模型

1. 对话上下文与槽位语义编码模块:
2. 对话上下文与槽位语义动态融合模块:
3. 槽位共享槽值提取模块:



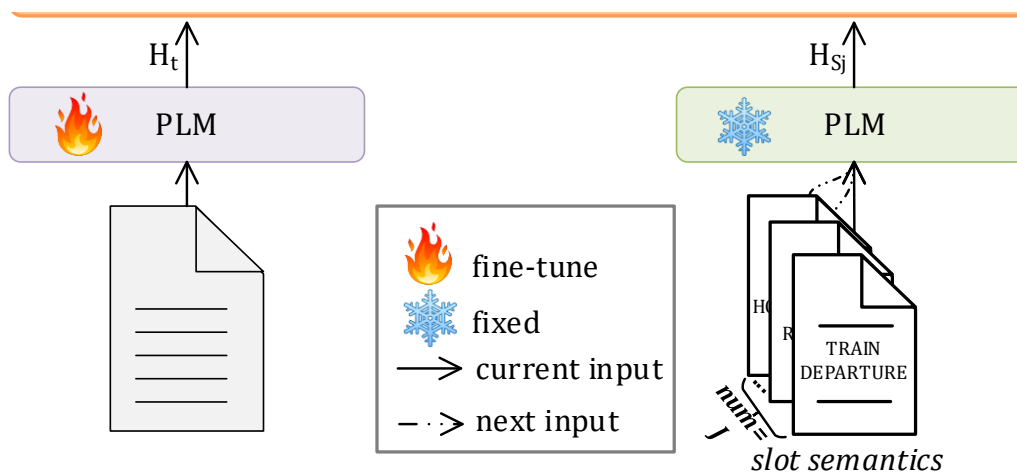
SSNet模型框架

SSNet模型

1. 对话上下文与槽位语义编码模块:

第 t 轮的对话上下文由当前轮次的用户语句 U_t ，上一轮系统回复 R_t 和第1轮至 $t - 1$ 轮的对话历史 $His_t = ((U_1, R_1), (U_2, R_2), \dots, (U_{t-1}, R_{t-1}))$ 组成，表示为 $X_t = [CLS] \oplus U_t \oplus [SEP] \oplus R_t \oplus [SEP] \oplus His_t \oplus [SEP]$ 。通过预训练模型进行编码。

槽语义表示时，将槽名称、槽描述信息和分类属性作为语义信息，引入到槽语义表示当中。具体地，第 j 槽的语义表示为 $SLOT_j = SN_j \oplus SD_j \oplus SC_j$ 。并通过预训练模型进行编码。

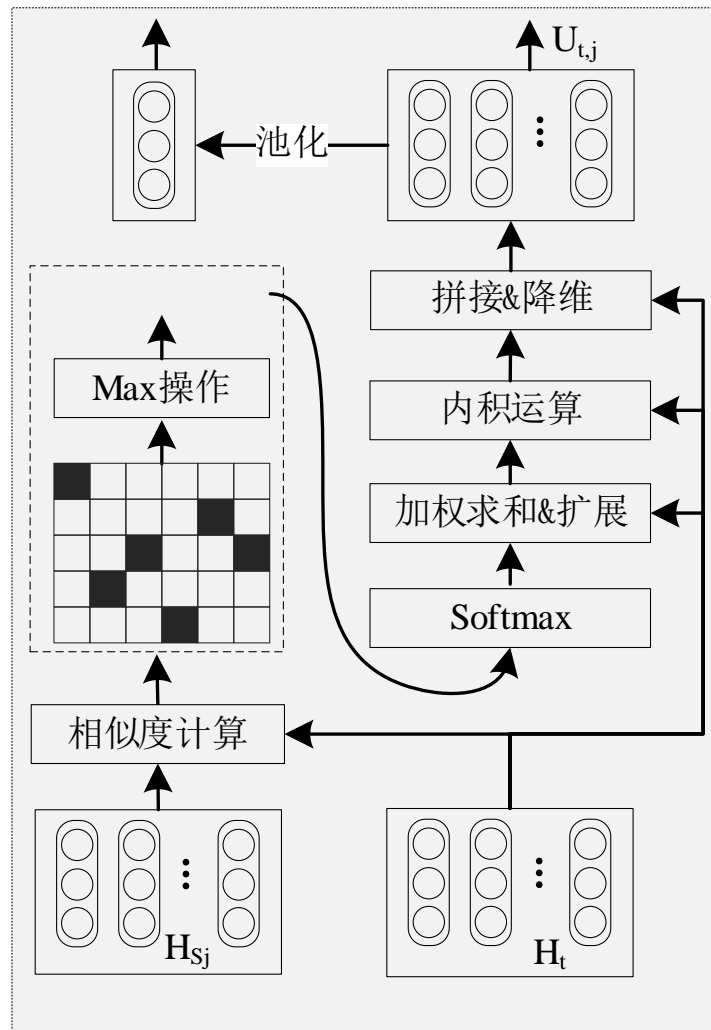


SSNet模型

2. 对话上下文与槽位语义动态融合模块：受BIDAF[1]算法的灵感，提出一种槽语义和对话上下文融合方法。

融合流程：

- 首先引入一个对话上下文与槽位共享的相似度矩阵 $S \in \mathbb{R}^{N \times M}$ ，其中 N 是当前对话上下文长度， M 是槽语义长度，故 $S_{n,m}$ 是第 n 对话Token和第 m 槽语义Token的相似度。
- 然后，通过 max 函数动态地选出对对话上下文词而言，相似度最大槽语义词，进行拼接，并进行 $softmax$ 归一化。
- 注意权重 b 与对话上下文 H_t 加权求和，得到矩阵 \tilde{h} ，并扩展到 N 位。
- 对话上下文和槽语义融合词向量 $U_t^j \in \mathbb{R}^{d \times N}$ 向量由 $U_t^j = Linear_{\Phi}[H_t; H_t \odot \tilde{H}_t]$ 得到。

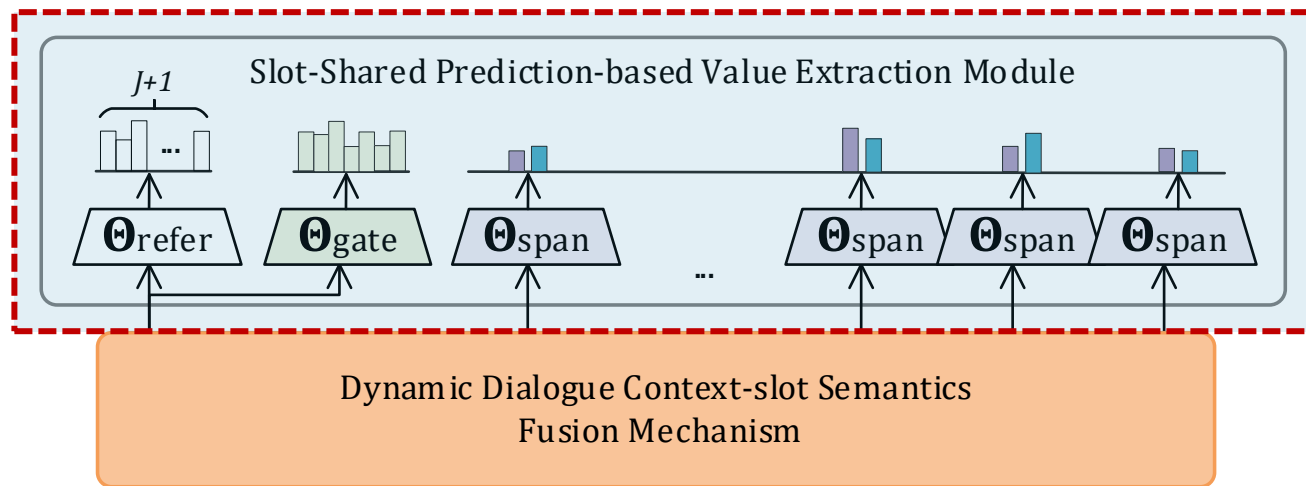


动态融合机制

□ SSNet模型

3. 槽位共享槽值提取层

槽值提取类型分为span、inform、refer、true、false、dontcare和none七种类型，将增强的对话上下文语义向量输入到一个分7类的提取类型分类器，确定当前槽位的槽值提取类型；如果槽值提取类型为inform、true、false、dontcare或者none，则直接确定其槽值；如果槽值提取类型为span，将槽位信息融合的槽位信息融合的对话上下文词向量输入到跨度预测分类器，获取槽值；如果槽值提取类型为refer，将槽位信息融合的对话上下文语义向量指代预测分类器，确定指向的槽位，从而获取槽值。



槽位共享槽值提取层

实验结果

MultiWOZ系列数据集包含涉及7个领域的10438条多领域对话，MultiWOZ 2.2是MultiWOZ 2.1的修订版本，主要修改错误标注和不一致实体。如表1所示，由于数据集中hospital和police领域没有验证集和测试集，本文同其它基线模型一样，在其它restaurant, train, hotel, attraction, taxi 等5个领域中进行相关实验。值得注意的是，本文把数据集提供的槽名称、槽描述和分类属性等信息作为语义表示，没有引入其它的额外信息。

领域	槽位	槽		
		训练集	验证集	测试集
hotel	price range, type, parking, book stay, book day, book people, area, stars, internet, name	3381	416	494
restaurant	food, price range, area, name, book time, book day, book people	3813	438	437
train	destination, day, departure, arrive by, book people, leave at	3103	484	494
attraction	area, name, type	2717	401	395
taxi	leave at, destination, departure, arrive by	1654	207	195
hospital	department, address, phone	287	0	0
police	name, address, phone	245	0	0

MultiWOZ 系列数据集统计

实验结果

在上述提到的MultiWOZ系列数据集上验证模型性能。

Model	Type	MultiWOZ 2.1	MultiWOZ 2.2
TRADE[5]	G	46.00	45.40*
DS-DST[22]	C+S	51.21	51.70*
SOM-DST[6]	G	53.68	53.81*
TripPy[16]	S	55.29	-
STAR[23]	C	56.36	-
AG-DST[7]	G	-	57.26
SDP-DST[9]	G	56.66	57.60
MSP-L[15]	S	57.20	57.70
SSNet	S	59.48	62.10

Joint goal accuracy on MultiWOZ 2.1 and 2.2.

JGA (%)	SA (%)	Per-slot Accuracy (%)			
61.65	77.71	<i>destination</i>	<i>departure</i>	<i>leaveby</i>	<i>arriveat</i>
		66.82	72.45	87.29	84.27

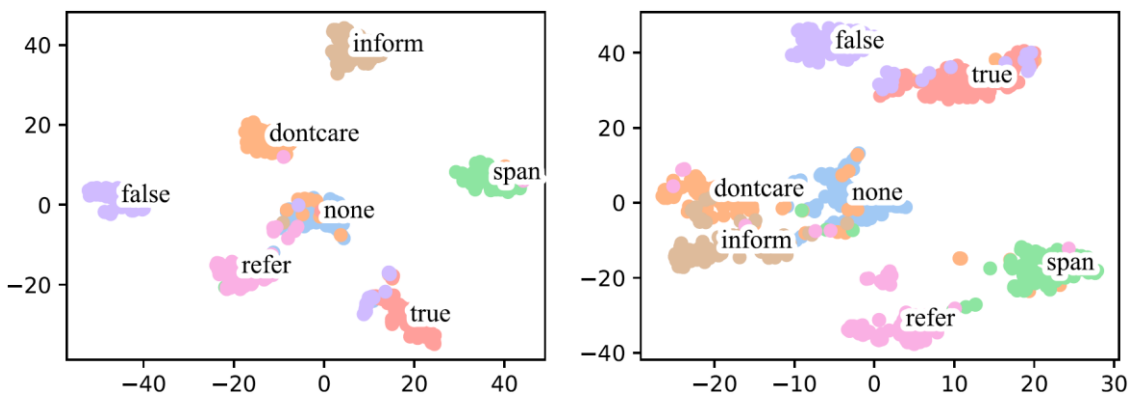
Domain expansion experiment on MultiWOZ 2.2.

Fusion Method	JGA(%)	SA(%)
Tra_ATT	49.77 ↓12.33	96.83 ↓1.11
Mean	60.03 ↓2.07	97.80 ↓0.14
DFM	62.10	97.94

Ablation study on MultiWOZ 2.2 .

实验结果

其它分析实验



t-SNE visualization of different fusion methods. The left is DFM, and the right is Tra ATT.

Class—Number	DFM	Tra_ATT
<i>none</i> — 209897	1.00	0.99
<i>dontcare</i> — 235	0.62	0.02
<i>span</i> — 7397	0.96	0.93
<i>true</i> — 416	0.76	0.70
<i>false</i> — 145	0.85	0.77
<i>inform</i> — 2702	0.93	0.75
<i>refer</i> — 368	0.73	0.58

F1 score of each class using different fusion methods

□ 总结

1. 我们提出一种基于槽位共享跨度预测的多领域对话状态追踪模型SSNet，该模型能够缓解对话状态追踪模型的扩展能力。
2. 此外，提出的动态融合方法能够有效融合对话上下文与不同的槽语义，促使槽位共享槽值提取层区分不同的槽位，以正确提取对应槽值。
3. SSNet在MultiWOZ 2.1和MultiWOZ 2.2任务型对话数据集上分别获得了58.46%和62.10%的联合目标正确率，超过现有性能最好的模型。

提纲

- 背景介绍
- SSNet模型与方法
- 浅谈未来方向

□ 2023—“大模型元年”

Open AI将通用大模型训练的结果通过ChatGPT的应用形式带到大家面前，意味着发展了大半个世纪的人工智能领域正式步入了广泛意义生产力提升的新纪元。



目前大模型在任务型对话上的尝试工作

评估LLM在任务型对话中完成多回合任务和与外部数据库交互的能力。

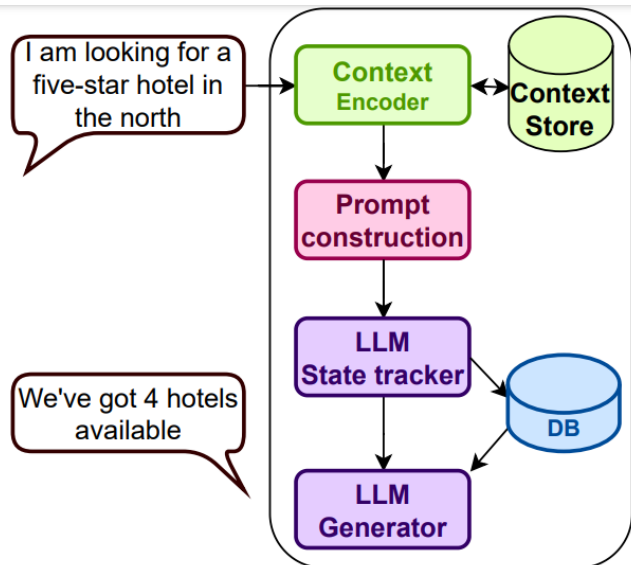


Figure 1: A high level overview of our proposed pipeline. The user input is used to retrieve relevant examples, if available. Then, an initial prompt is constructed and an LLM is asked to provide current dialogue state. Based on that we retrieve database information and construct another prompt. Finally we ask the LLM to provide response.

该工作结论:

1. 大模型在对话状态追踪方面表现不佳，即使提供了context example。但是，随着指令调优技术和进步，有可能进行改进。
2. 在对话成功率方面，模型能够很好地与用户进行交互，提供有用的信息，满足用户的需求。

- 目前大模型在任务型对话上的尝试工作在LLM中注入领域知识。

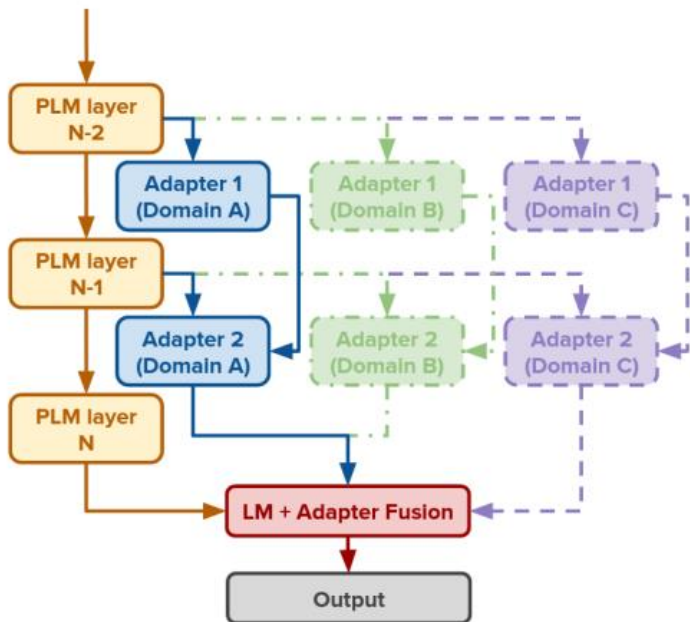


Figure 1: A high-level representation of the KB-adapter architecture (decoder only, for clarity). Adapter states are fused with the hidden states of the PLM to produce a knowledge-informed predictive distribution. Dashed elements are used only if multiple adapters are active.

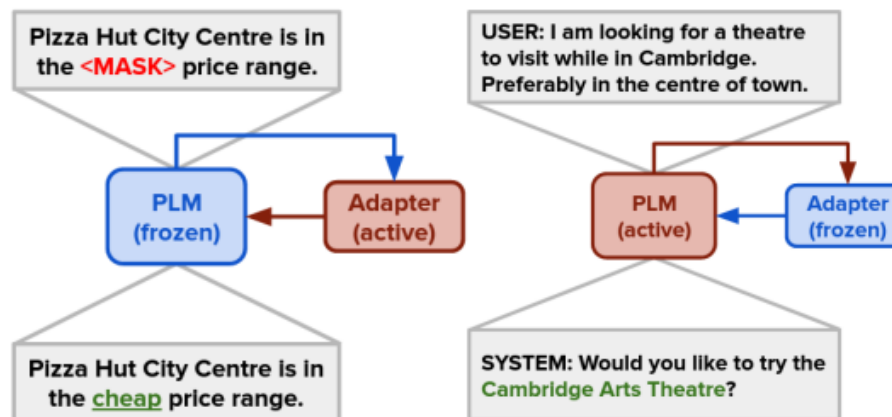


Figure 3: On the *left*, a schematic representation of the **memorization** stage, where the adapter is trained to memorize KB contents by reconstructing corrupted statements derived from KB facts. On the *right*, a representation of the **utilization** stage, where the adapter-augmented PLM is fine-tuned on a downstream TOD task and learns how to utilize adapter knowledge.

目前大模型在任务型对话上的尝试工作

测试LLM在零样本对话状态追踪任务上的性能。

Models	attr.	hotel	rest.	taxi	train	avg.
TRADE (2019; 2020)	22.8	19.5	16.4	59.2	22.9	28.16
TripPy-R (2022)	27.1	18.3	15.3	61.5	23.7	29.18
TransferQA (2021)	31.3	22.7	26.3	61.9	36.7	35.78
Li et al. (2021)	42.4	24.9	27.7	60.3	41.1	39.28
D3ST (2022)	56.4	21.8	38.2	78.4	38.7	46.70
Campagna et al. (2020)	52.8	36.3	45.3	62.6	46.7	48.74
ChatGPT	52.7	42.0	55.8	70.9	60.8	56.44
IC-DST ⁵ (2022)	60.0	46.7	57.3	71.4	49.4	56.96

Table 1: Performance comparison between state-of-the-art methods for zero-shot DST in per-domain JGA.

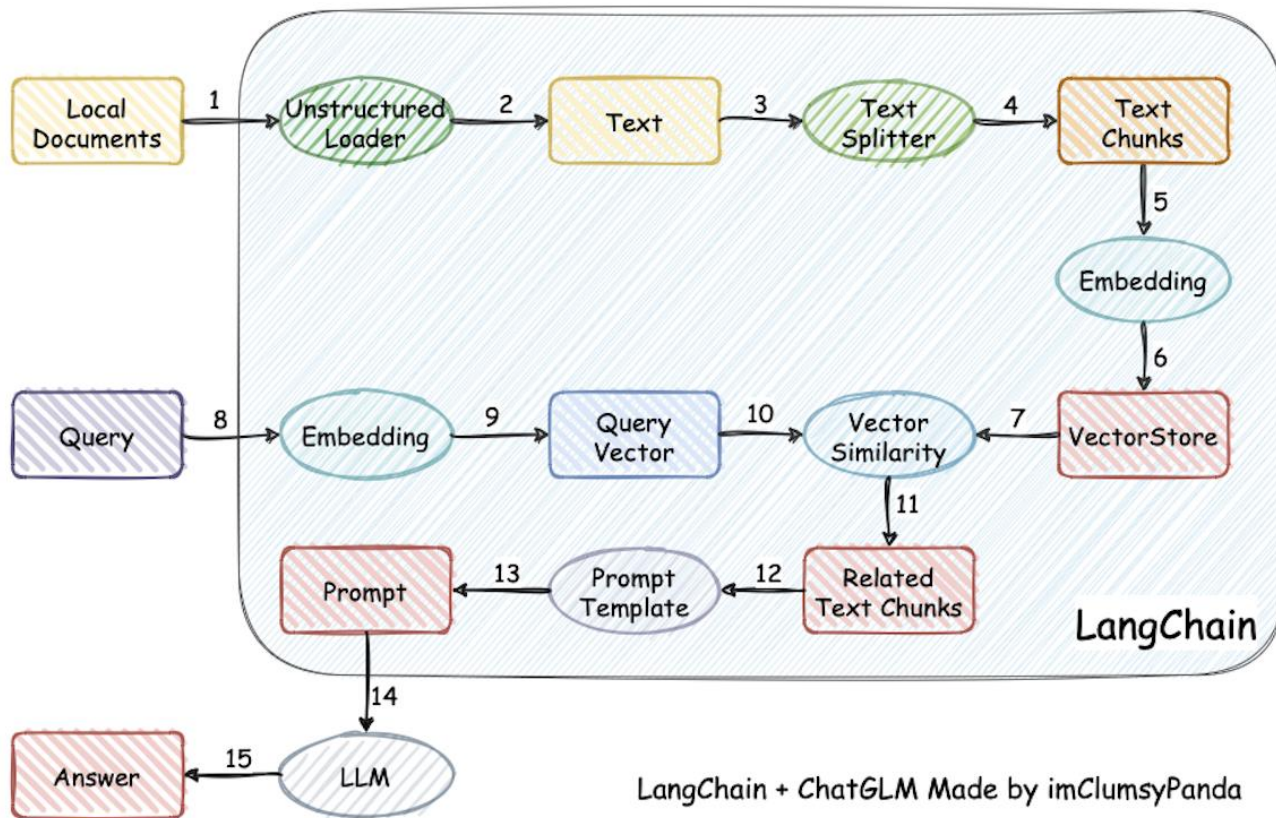
a) PMUL4050	system: "I'd recommend the Autumn House. Would you like to make a booking?" user: "Yes please. I need the reservation to be for 8 people and 2 nights starting on Tuesday." Prediction: ... hotel-name: none Label: ..., hotel-name: autumn house
b) PMUL0117	user: "Yes I also need a taxi that will get me to the restaurant by the booked time please." Prediction: taxi-destination: hotel, taxi-departure: restaurant Label: taxi-destination: the gonville hotel, taxi-departure: la mimosa
c) SNG01873	user: "I need to be picked up from pizza hut city centre after 04:30" Prediction: ..., hotel-name: dontcare, ..., attraction-type: dontcare, ... Label: ...
d) PMUL0599	user: "[...] Can you just help me find a high-end Mexican restaurant?" Prediction: ..., restaurant-pricerange: high-end Label: ..., restaurant-pricerange: expensive
e) MUL2051	user: "Can I get address and postcode for the hotel?" Prediction: hotel-address: ?, hotel-postcode: ? Label: - system: "The address is 74 chesterton road, the postal code is cb41er, can I assist with anything else?" user: "That is all for now, goodbye." Prediction: hotel-address: 74 chesterton road, hotel-postcode: cb41er Label: -
f) MUL0524	user: "I'm going to Cambridge on saturday and want to arrive by 14:15 please." Prediction: ..., train-day: Saturday Label: ..., train-day: saturday
g) PMUL4246	user: "i need a place to go and should be a museum" Prediction: attraction-type: museum Label: attraction-type: museum system: "Okay! There are several museums in Cambridge. What part of town would you like to visit?" user: "How about ones in the centre, what's available?" Prediction: attraction-type: museum, attraction-area: centre Label: attraction-area: centre

Table 2: Examples of reoccurring errors as listed in Section 4.3.

研究意义：幻觉（hallucinations）可能是一个问题，但它们也为实现零样本提供了机会。我们观察到许多插槽幻觉是合理的，并且指向了对对话有意义的元素。

目前大模型在任务型对话上的尝试工作

2. LangChain+LLM: 一种利用 `langchain` 思想实现的基于本地知识库的问答应用，目标期望建立一套对中文场景与开源。



□ 大模型在任务型对话上的应用有哪些优势?

1. 具备初步的系统主动性

A 你好，可以给我订一家酒店嘛？



当然可以！请问您要前往哪个城市，并告诉我您的入住和离店日期以及预算范围，我将帮您找到适合的酒店选项。

2. 良好的语言理解能力：指代等

A 它里边有wifi嘛？



一般来说，大多数酒店都提供免费的Wi-Fi服务，但具体情况可能因酒店而异。我无法确切知道该青年旅舍是否提供免费的Wi-Fi服务。您可以在预订前查看酒店的官方网站或联系酒店前台，向他们确认是否提供Wi-Fi服务以及相关细节。

大模型在任务型对话上的应用有哪些优势?

3. 支持多轮对话，能够很好地与用户进行交互。

A 你好，你可以给我推荐一家餐厅吗？

当然可以！请告诉我您所在的城市或地区，以便我能够为您提供附近的推荐餐厅。

A 我在新疆乌鲁木齐。

在乌鲁木齐，有一家非常受欢迎的餐厅叫做「贝聿铭西餐厅」(Bey's Dining)。这是一家以西式菜肴为主的高级餐厅，提供精致的用餐体验和优质的食物。餐厅的装修非常精美，氛围温馨舒适，适合特殊场合或浪漫的晚餐。

另外，如果您想品尝当地的新疆菜，我推荐「香满楼」(Fragrant Pavilion) 餐厅。这家餐厅以正宗的新疆美食而闻名，提供烤全羊、手抓饭、大盘鸡等传统菜肴。它的装修风格也充满了新疆的特色，能够为您带来真正的新疆餐饮体验。

无论您是想享受西式美食还是尝试当地特色菜肴，这些餐厅都是乌鲁木齐的热门选择。希望您在这里用餐愉快！

A 贝聿铭西餐厅是在乌鲁木齐嘛？

A 它里边有wifi嘛？



一般来说，大多数酒店都提供免费的Wi-Fi服务，但具体情况可能因酒店而异。我无法确切知道该青年旅舍是否提供免费的Wi-Fi服务。您可以在预订前查看酒店的官方网站或联系酒店前台，向他们确认是否提供Wi-Fi服务以及相关细节。

大部分酒店在设施中会明确标明是否提供Wi-Fi服务，有些酒店甚至在房间描述中提供更详细的Wi-Fi信息，如网络覆盖范围、连接方式等。这些信息可以帮助您做出更好的决策。

如果Wi-Fi对您的住宿非常重要，您也可以选择在预订时选择高级酒店或商务酒店，它们通常提供更稳定和高速的Wi-Fi服务。

希望这些信息能对您有所帮助！如有其他问题，请随时提问。

A 简短回答 它有wifi还是没有wifi



很抱歉，我无法直接提供有关该青年旅舍是否提供Wi-Fi的准确信息。建议您在预订前查看酒店的官方网站或直接联系酒店前台以获取确切的Wi-Fi信息。

□ 大模型在任务型对话上的应用有哪些不足？

1. 由于缺乏专业领域知识，无法实时获取信息，会产生与事实违背的回复，然而任务型对话基本要求是精确和可控。

A 我在新疆乌鲁木齐，今天入住，价格200元左右。

 好的，我将为您找到适合您需求的酒店选项。请稍等片刻。

在乌鲁木齐，以下是一家价格适中的酒店推荐给您：

1. **乌鲁木齐市中心青年旅舍** 位于乌鲁木齐市中心，价格实惠，提供干净舒适的住宿环境。房间价格大约在200元左右。**地址：乌鲁木齐市天山区人民路123号。**

请注意，酒店价格可能会随时变动，建议您在预订前再次确认价格和可用性。您可以使用在线预订平台如Booking.com、Ctrip等来预订酒店，或者联系酒店前台直接预订。

祝您在乌鲁木齐度过愉快的旅行！如有其他问题，我随时为您服务。



□ 大模型在任务型对话上的应用有哪些不足？

2. 一部分研究把大模型可以看作是知识库。按这个角度讲，它可以是一个静态的，通用的知识库，缺乏实时性和私人性。

静态性

缺乏实时性，知识更新成本高（在新的数据上重新训练）

研究点一：语言模型与数据库交互。

研究点二：新知识（领域、任务、结构化知识等）的融入。

通用

缺乏用户个性化和对话系统同理心。

Empathy in Dialogue Systems

Human: I messed up my project today.

System: What is your project?

Human: I think I might get fired...

System: Then you need to find another job!

Human: 😞

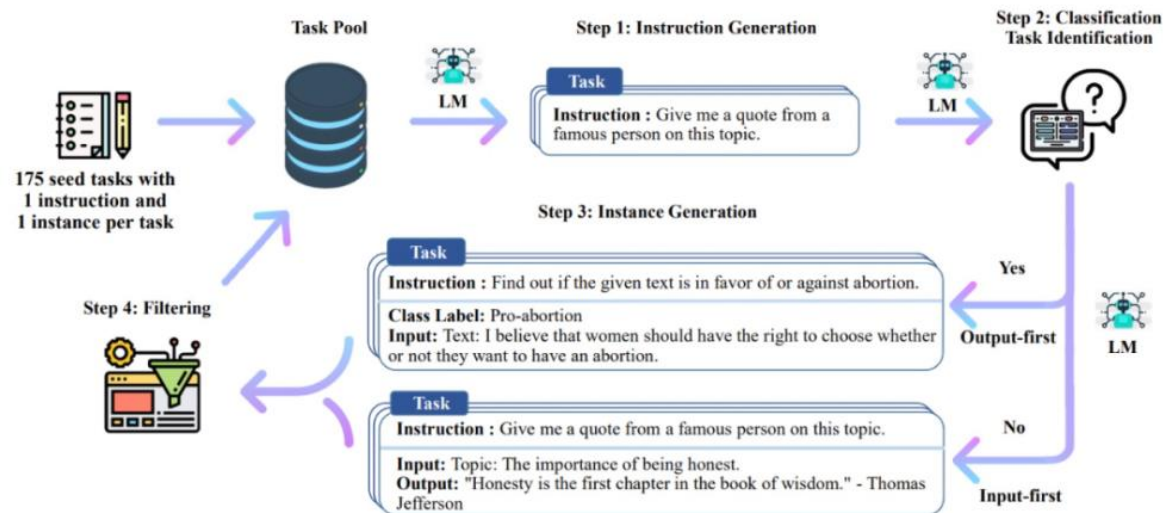


大模型在任务型对话上的应用有哪些不足？

3. 高度黑盒：尽管包含相当多的知识与常识，目前的指令提示并不能够稳定挖掘其中知识。

```
36 select A.姓名,C.时间,
37 case when B.部门代码 = C.变动前部门 then A.部门 end '变动前部门',
38 case when B.部门代码 = C.变动后部门 then A.部门 end '变动后部门'
39 from A,B,C
40 where A.姓名 = C.姓名
41       and A.部门=B.部门名称
42       and C.时间 = '2015-01-01'
```

姓名	时间	变动前部门	变动后部门
1 赵	2015-01-01	测试一部	NULL
2 钱	2015-01-01	NULL	测试三部
3 孙	2015-01-01	测试二部	NULL



数据库查询 VS 大模型指令提示

□ 大模型==任务型对话系统???

- 1.大模型具备任务型对话系统具备的系统主动、强大的语言理解能力、支持多轮等优势。
2. 目前由于指令调优技术不够成熟且带有不确定性，缺乏准确挖掘大模型知识的有效工具。现阶段还需要通过引入langChain、传统数据库等阶段知识更新问题。
3. 按任务型对话系统角度来讲，未来，连续学习与终身学习可能是解决大模型知识更新问题的未来方向。研究更加有效的大模型知识挖掘（激发）工具也是必不可少的。

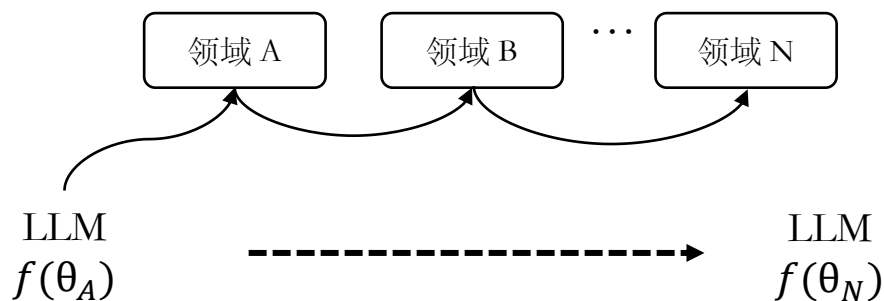


图 Continual Learning

大模型生态下的机遇与挑战

大模型“画龙”，小数据“点睛”

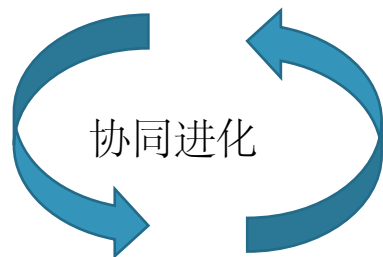
大规模预训练数据为“基座”，小场景数据微调。保留模型通用视觉、语言、动作理解能力，同时提升对特定任务的适应性。
大模型为教师，小模型蒸馏获取小场景任务能力。

小场景交互性与自学习性

交互式学习给予模型反馈，帮助模型理解特定场景需求（e.g., 人类反馈学习）大模型为教师，小模型蒸馏获取小场景任务能力。
自学习自动更新模型知识，适应特定任务新场景（e.g., 增量学习）

小场景

样本有限，标注精细，结构化



大模型

数据量大，有噪音，多种模态并存





报告完毕！
请各位老师同学们批评指正

中国科学院新疆理化技术研究所
多语种信息技术研究室

艾比布拉·阿塔伍拉