复杂场景下的对话追踪状态技术研究与实现
中期报告

王德远

2022年7月22日

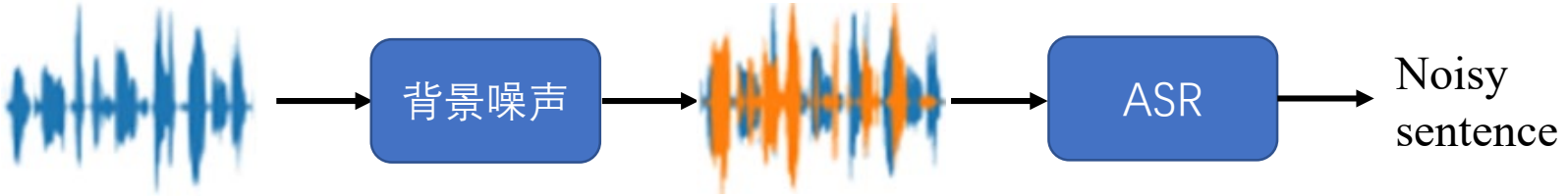
目录

课题背景

研究背景

当前任务型对话系统所基于的对话场景和现实的对话场景有很大的区别，现实场景中当用户通过语音和系统交互时，自身所处的环境可能存在背景噪声（汽车鸣笛、飞机起飞、高斯白噪声、人声等）。

系统将接收到的语音通过 ASR 转换为文字时，出现错误。导致后续对话状态追踪模块无法正确识别用户期望的槽值。



课题背景

课题定义

对话状态跟踪（Dialog State Tracking）是面向任务对话系统的核心组件。DST的目的是提取用户在对话中表达的目标/意图，并将其编码为对话状态的紧凑集合，即一组槽及其对应的值。

今天中午咱们去北邮风味餐厅吃饭吧。

领域：餐厅
餐厅名：北邮风味餐厅

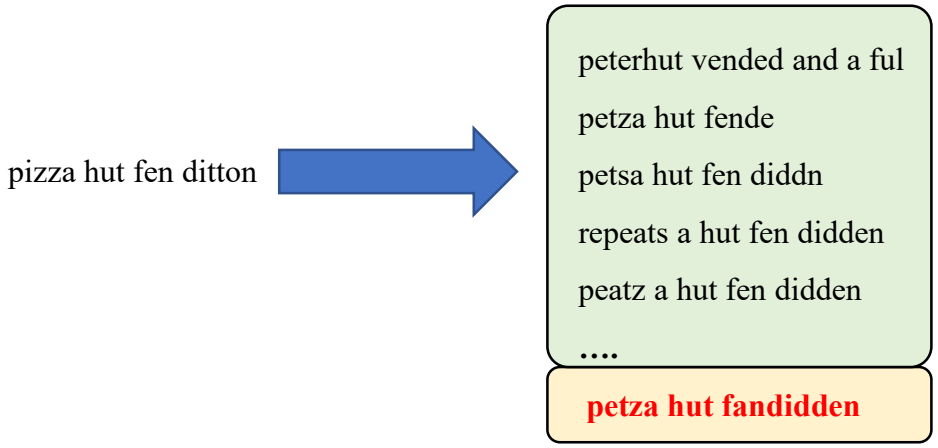
噪声：主要指用户的语音经过ASR转化后的文本错误。例如：“今天中午咱们去贝优风味餐厅吃饭吧。”

课题背景

难点分析1：

同一个实体词错误形式不固定

实体词是嵌在整段连续语音中，其上下文发音不同也会影响ASR对实体词的识别，从而导致同一个实体出现了多种错误形式。



课题背景

难点分析2:

实体词之间的相似带来的挑战

sng0073

User: i would like a taxi from saint john s college to **petza hut fandidden**
Sys: what time do you want to leave and what time do you want to arrive by?
User: i want to leave after seventeen fifteen
Sys: booking completed! your taxi will be [value_car] contact number is [value_phone]
User: thank you for all the help i appreciate it

pizza hut fen ditton

[taxi] destination **pizza hut city centre**
[taxi] departure saint john 's college
[taxi] leave seventeen fifteen

mul2321

User: i need to find information about a certain restaurant can you help with that
Sys: yes i can. what restaurant are you looking for?
User: it is called **maheraji tandori restaurant**

maharajah tandoori restaurant

[restaurant] name **mahal of cambridg**

1、原句中提到的实体是pizza hut fen ditton

2、pizza hut city centre 和 pizza hut fen ditton 都是语料中的实体

课题背景

研究意义

工业：直接对语音识别后的用户语句建模而非语音，即使用户语句中存在噪声，模型也有一定的抗干扰性。从而节省存储成本。

学术：提升对话状态追踪模型的鲁棒性，使得模型可以处理带有噪声的用户自然语言。

目录

研究现状

数据集-DSTC2

餐厅预定领域的多轮对话数据集，其中训练数据有1612组对话，10886个对话轮次，测试集有1117组对话，9160个对话轮次。

Actual input and output	SLU hypotheses and scores	Labels	Example tracker output	Correct?
S: Which part of town? *request(area)*	0.2 inform(food=north_african) 0.1 inform(area=north)	area=north	0.2 food=north_african 0.1 area=north 0.7 ()	✗ ✓ ✗
U: The north uh area *inform(area=north)*		method=byconstraints	0.9 byconstraints 0.1 none	✓
		requested=()	0.0 phone 0.0 address	✓ ✓
S: Which part of town? *request(area)* U: A cheap place in the north *inform(area=north, pricerange=cheap)*	0.8 inform(area=north), inform(pricerange=cheap) 0.1 inform(area=north)	area=north pricerange=cheap	0.7 area=north pricerange=cheap 0.1 area=north food=north_african 0.2 ()	✓ ✗ ✗
		method=byconstraints	0.9 byconstraints 0.1 none	✓
		requested=()	0.0 phone 0.0 address	✓ ✓

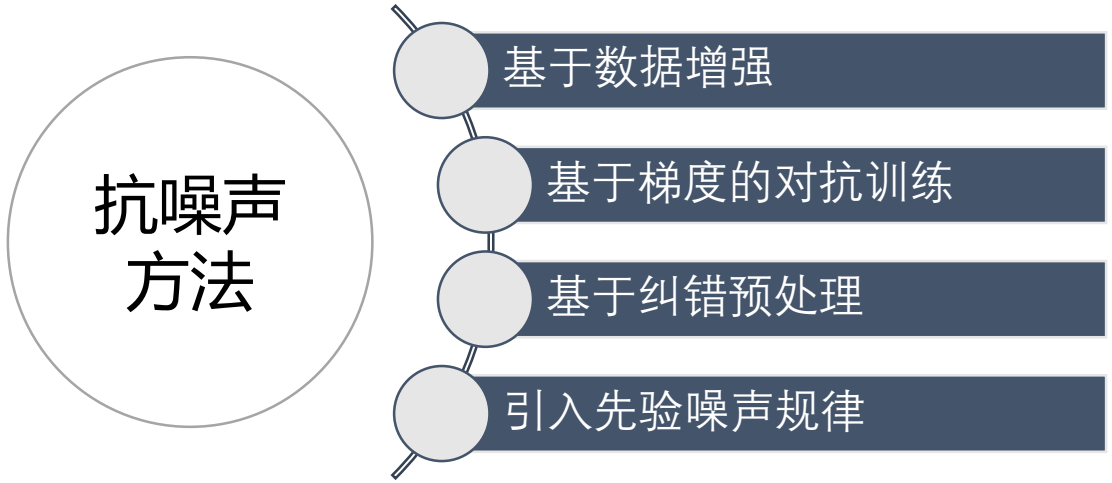
研究现状

数据集-MultiWOZ

大规模多领域多轮对话数据集，训练集包含8438组对话，115424个对话轮次。测试集验证集分别包含1000组对话。

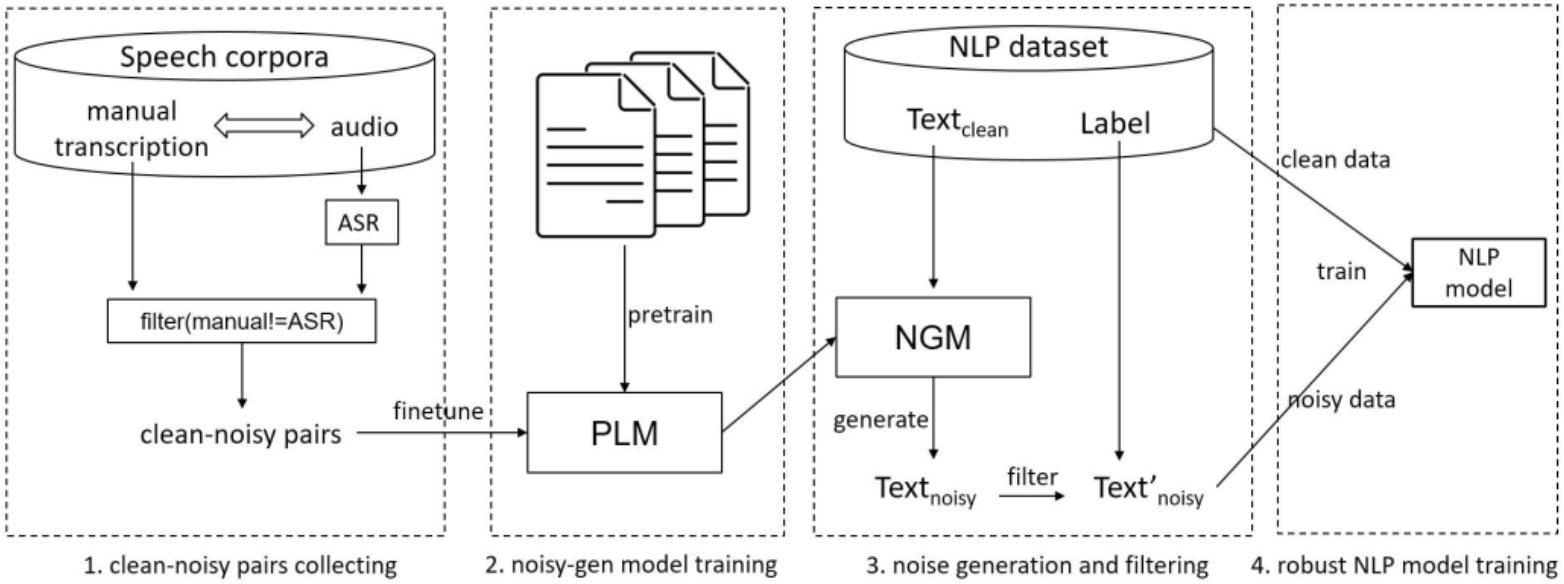
- You are traveling to Cambridge and looking forward to try local restaurants.
- You are looking for a **place to stay**. The hotel should be in the type of **hotel** and should be in the **centre**.
- The hotel should **include free wifi** and should have **a star of 4**.
- Once you find the **hotel** you want to book it for **3 people** and **5 nights** starting from **monday**.
- Make sure you get the **reference number**.
- You are also looking for a **restaurant**. The restaurant should serve **australasian** food and should be in the **moderate** price range.
- The restaurant should be **in the same area as the hotel**.
- If there is no such restaurant, how about one that serves **british** food.
- Once you find the **restaurant** you want to book a table for **the same group of people** at **18:30** on **the same day**.
- Make sure you get the **reference number**

研究现状



研究现状

数据增强

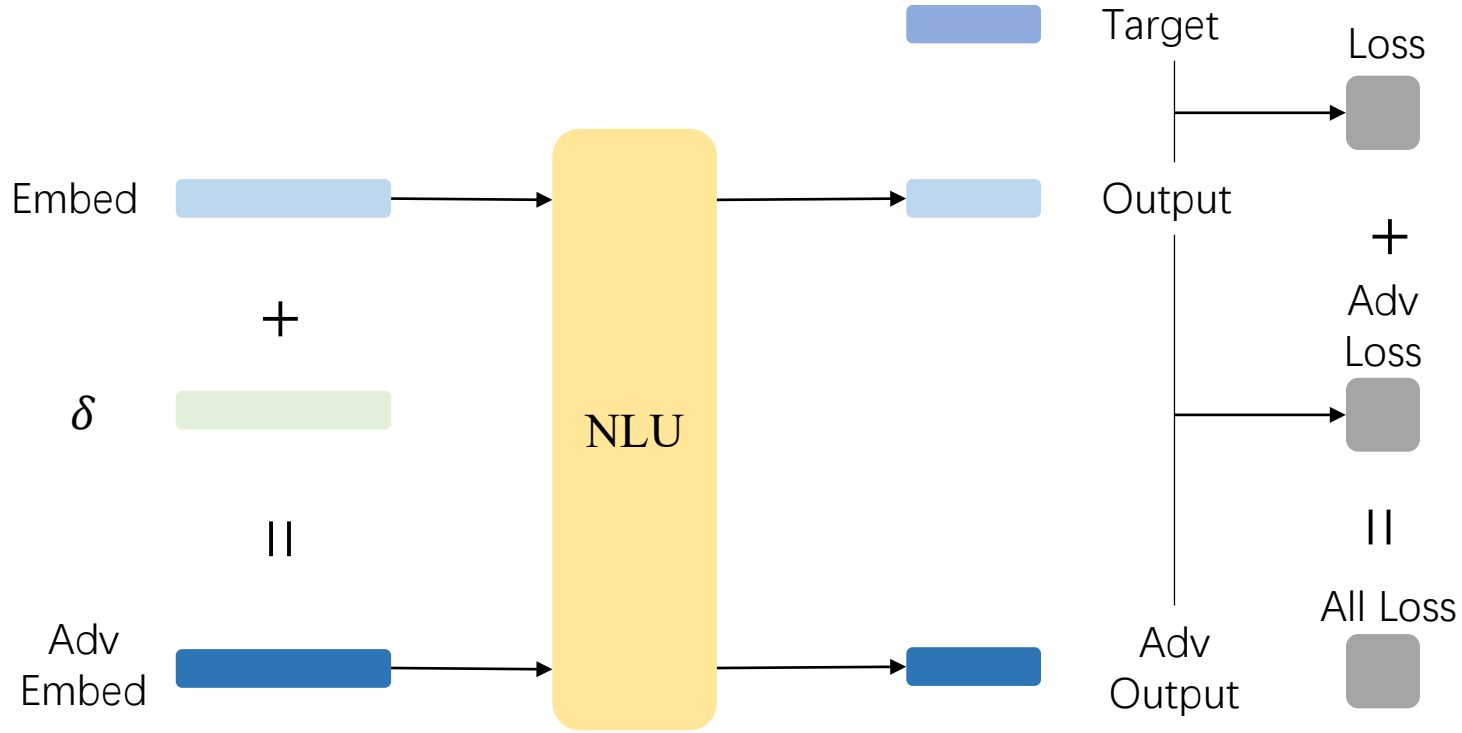


2021.arxiv-An Approach to Improve Robustness of NLP Systems against ASR Errors
2020.findings.acl-A survey of Data Augmentation Approaches
2021.findings-acl-Better Robustness by More Coverage：Adversarial and Mixup Data Augmentation for Robust Finetuning
2021.nips-ASR-GLUE：A New Multi-task Benchmark for ASR-Robust Natural Language Understanding

研究现状

对抗训练

基于梯度对抗训练的方法是针对向量空间，通过最大化对抗损失、最小化模型损失的方式进行训练：



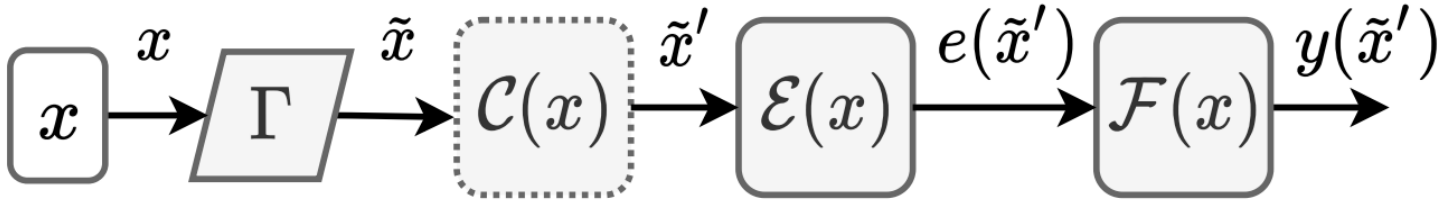
2020.acl-Adversarial Training for Large Neural Language Models
2020.acl-NAT：Noise-Aware Training for Robust Neural Sequence Labeling
2020.aaai-TAVAT：Token-Aware Virtual Adversarial Training for Language Understanding
2021.naacl-Targeted Adversarial Training for Natural Language Understanding

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta} l(f(x+\delta;\theta),y)]$$

研究现状

纠错预处理

将带有噪声的文本经过预先的纠错处理，将纠错后的句子序列输入到下游任务得到结果。

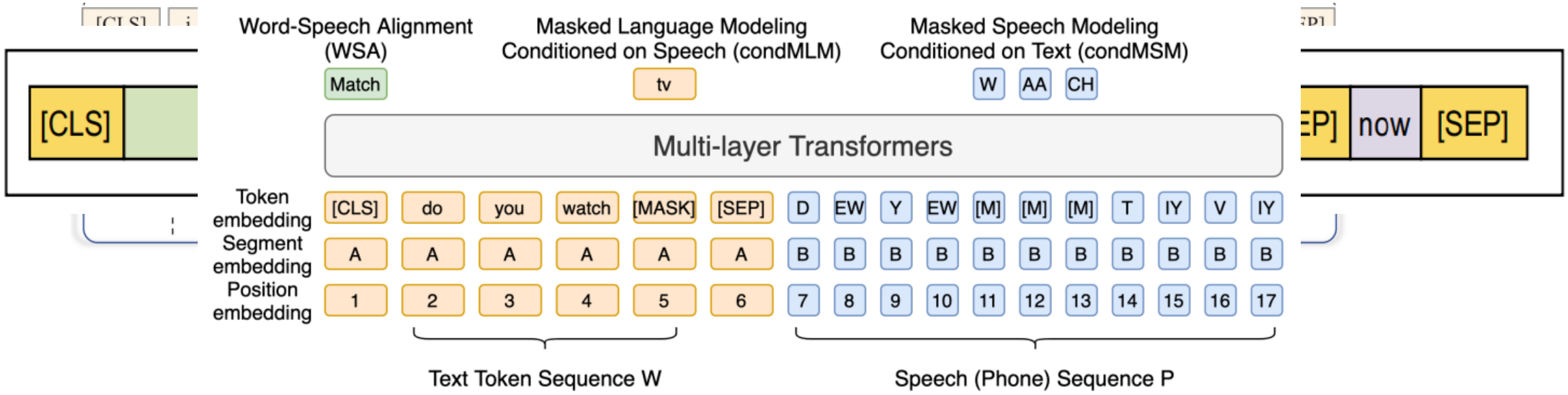


2020.icassp-Joint contextual modeling for asr correction and language understanding
2020.acl-Learning Spoken Language Representations with Neural Lattice Language Modeling
2020.interspeech-ASR Error Correction with Augmented Transformer for Entity Retrieval
2021.findings-acl-Correcting Chinese Spelling Errors with Phonetic Pre-training

研究现状

引入先验噪声规律

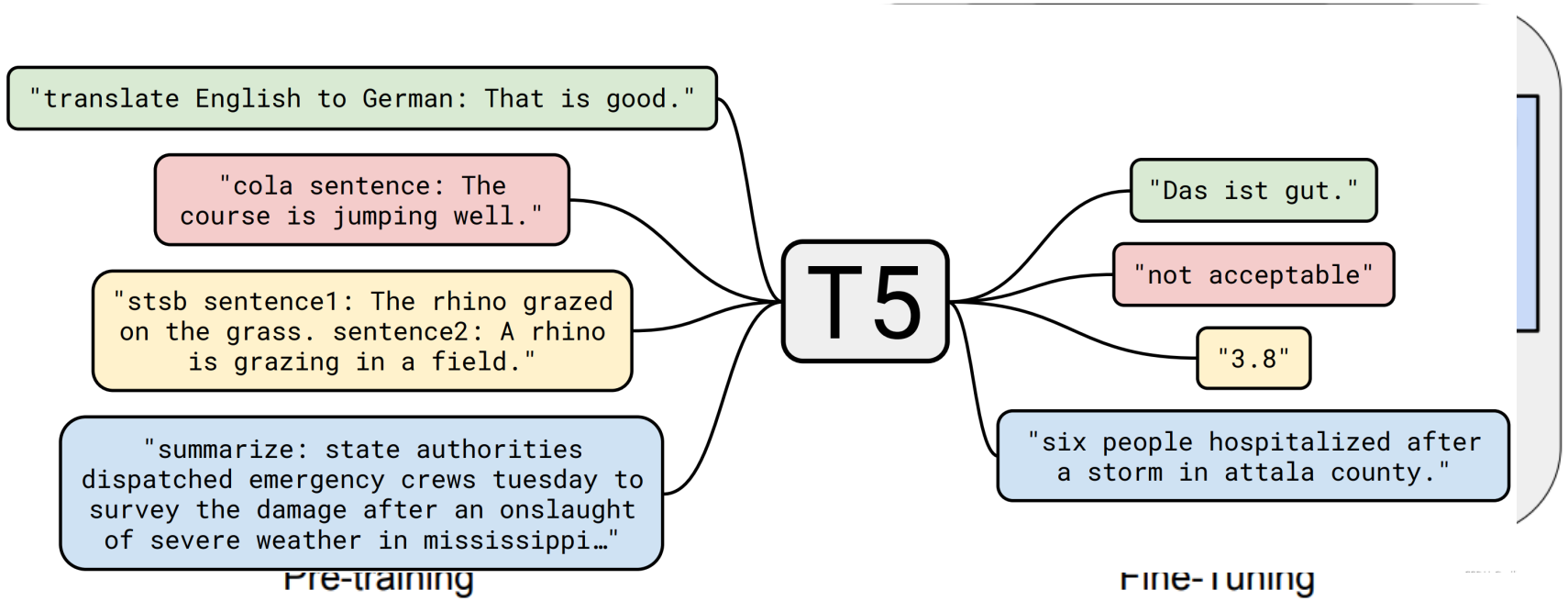
通过引入额外的音素特征、Nbest转录、单词混淆矩阵等来增强句子的表征，将噪声规律引入到句子编码中。



2019.acl-Robust neural machine translation with joint textual and phonetic embedding
2020.acl-Learning Spoken Language Representations with Neural Lattice Language Modeling
2021.findings-acl-Empirical Error Modeling Improves Robustness of Noisy Neural Sequence Labeling
2021.interspeech-Pre-training for Spoken Language Understanding with Joint Textual and Phonetic Representation Learning

研究现状

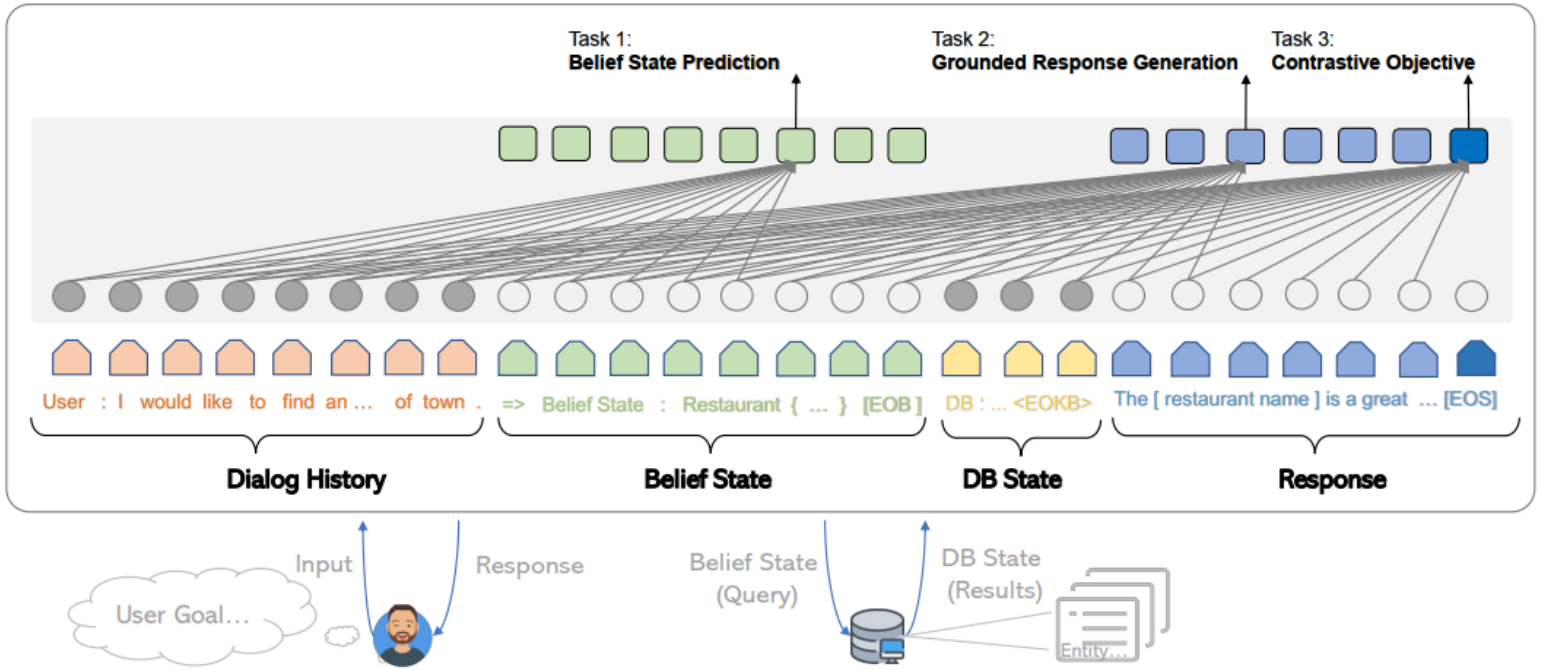
大规模预训练语言模型



目前基于Transformer的预训练语言模型的预训练数据中并没有引入大量的噪声语句，通过直接加载预训练模型提取的语义向量并不能很好适用于噪声文本作为输入的下游任务。

研究现状

对话状态追踪



2019.acl-Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems
2020.emnlp-TOD-BERT： Pre-trained Natural Language Understanding for Task-Oriented Dialogue
2020.nips-A Simple Language Model for Task-Oriented Dialogue
2021.acl-SOLOIST：Building Task Bots at Scale with Transfer Learning and Machine Teaching
2022.acl-Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System

研究现状

以往研究的不足：

1、数据增强：数据增强所覆盖的噪声是有限的，无法穷举噪声的所有错误。

2、基于梯度的对抗训练：基于梯度的对抗训练直接将噪声加到连续的向量空间，这类方法与单词级别的噪声不具有一致性。

3、基于纠错：目标任务模型的效果取决于纠错模型的好坏。

4、基于噪声规律先验：需要获得ASR工具更深层的权限，通常很难得到Nbest和单词混淆网络。

5、主流的预训练模型缺乏很好表征噪声文本的能力。

目录

研究内容与目标

研究内容：

1、基于T5的多任务降噪对话状态追踪预训练算法的设计与实现。

2、基于端到端的降噪对话状态追踪算法的设计与实现。

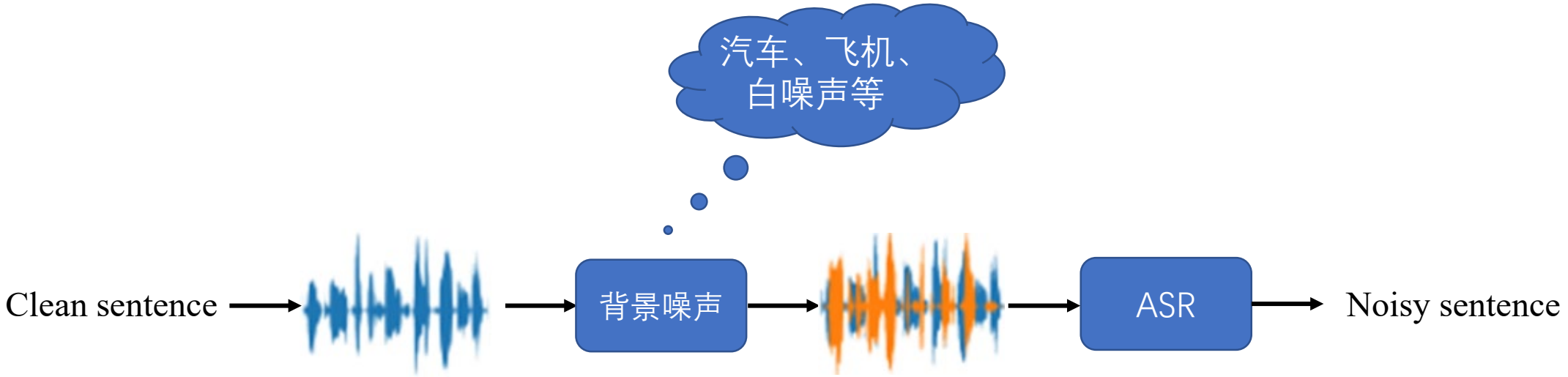
研究目标：

提升对话状态追踪模型的鲁棒性，使得模型可以处理带有噪声的用户自然语言（抗噪声对话状态追踪模型）。

数据集构造

数据集生成

基于MultiWOZ 2.0数据集，对用户语句进行加噪处理，过程如下：



噪声来源：OpenSLR数据集

音频合成：Pydub开源库

Text-to-Speech：pytts开源库

ASR：DeepSpeech2开源库

2017.ICASSP-A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition

数据集构造

数据集标注

Noisy Sentence： i would like a taxi from saint john s college to **petza hut fandidden**

Clean Sentence： i would like a taxi from saint john s college to **pizza hut fen ditton**

Belief State： [taxi] destination **pizza hut city centre** departure saint john 's college

Entity index： [[6, 7, 8, 9], [11, 12, 13, 14]]

1、选择Belief State中的槽值。

2、标注Clean Sentence中槽值实体的出现的位置。

数据集构造

数据集预处理

1、时间格式统一

对于某些阿拉伯数字时间18：15转为语音后可能有多种表达例如：eighteen fifteen、 eighteen a quarter等。将用户语句中的时间以及标注统一为eighteen fifteen形式。

2、量词阿拉伯数字格式统一

例如，3 stars 2 night，统一转换为 three stars two night。

数据集构造

数据集规模

	DSTC2	MultiWOZ 2.0	MultiWOZ-Noise
对话数	1612	8438	16876
轮次	10886	115424	230850
WER	29.99	-	11.4
领域	restaurant	train、hotel、attraction、taxi、restaurant	train、hotel、attraction、taxi、restaurant

数据集构造

数据集展示

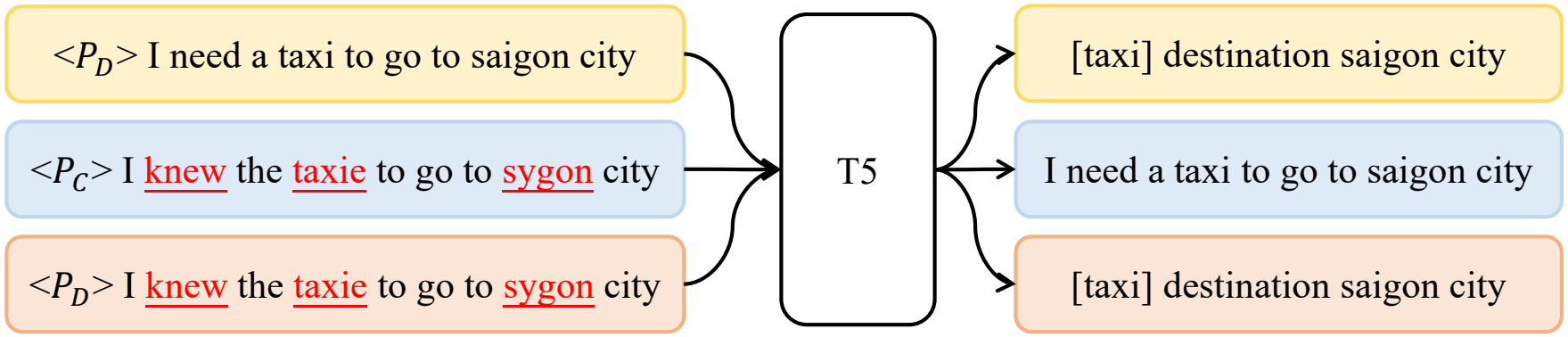
sng0073

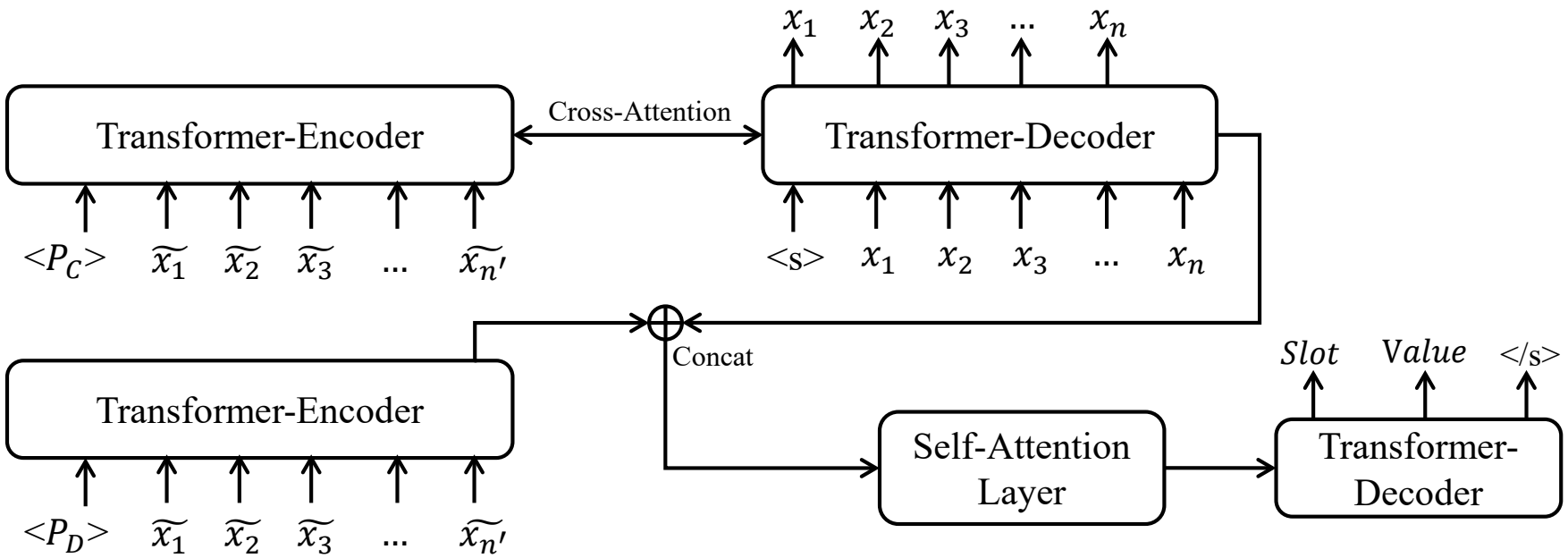
User: i would like a taxi from saint john s college to **petza hut fandidden**
Sys: what time do you want to leave and what time do you want to arrive by?
User: i want to leave after seventeen fifteen
Sys: booking completed! your taxi will be [value_car] contact number is [value_phone]
User: thank you for all the help i appreciate it

pizza hut fen ditton

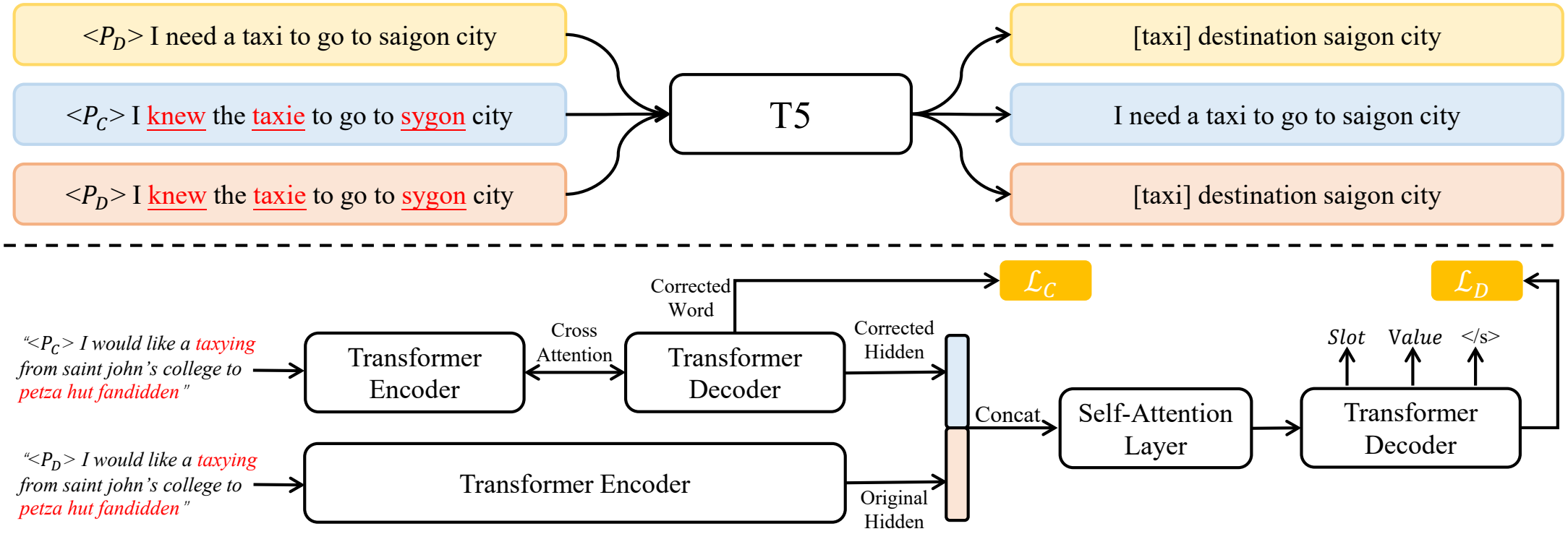
[taxi] destination **pizza hut city centre**
[taxi] departure saint john 's college
[taxi] leave seventeen fifteen

模型


<P_D> I need a taxi to go to saigon city
<P_C> I knew the taxie to go to sygon city
<P_D> I knew the taxie to go to sygon city
T5
[taxi] destination saigon city
I need a taxi to go to saigon city
[taxi] destination saigon city

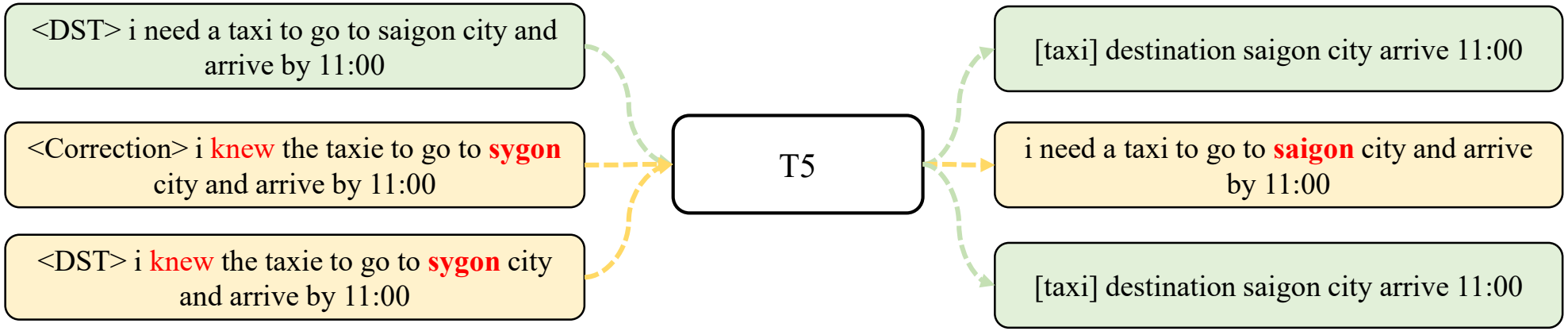

x_1 x_2 x_3 ... x_n
Transformer-Encoder
Cross-Attention
Transformer-Decoder
<P_C> $\widetilde{x_1}$ $\widetilde{x_2}$ $\widetilde{x_3}$ ... $\widetilde{x_n'}$
<s> x_1 x_2 x_3 ... x_n
Concat
Transformer-Encoder
Self-Attention Layer
Transformer-Decoder
Slot Value </s>
<P_D> $\widetilde{x_1}$ $\widetilde{x_2}$ $\widetilde{x_3}$ ... $\widetilde{x_n'}$

模型


<P_D> I need a taxi to go to saigon city
<P_C> I knew the taxie to go to sygon city
<P_D> I knew the taxie to go to sygon city
T5
[taxi] destination saigon city
I need a taxi to go to saigon city
[taxi] destination saigon city
$\mathcal{L}_C$
$\mathcal{L}_D$
Corrected Word
"<P_C> I would like a taxying from saint john's college to petza hut fandidden"
Transformer Encoder
Cross Attention
Transformer Decoder
Corrected Hidden
"<P_D> I would like a taxying from saint john's college to petza hut fandidden"
Transformer Encoder
Original Hidden
Concat
Self-Attention Layer
Slot
Value
</s>
Transformer Decoder

Prompt 多任务预训练

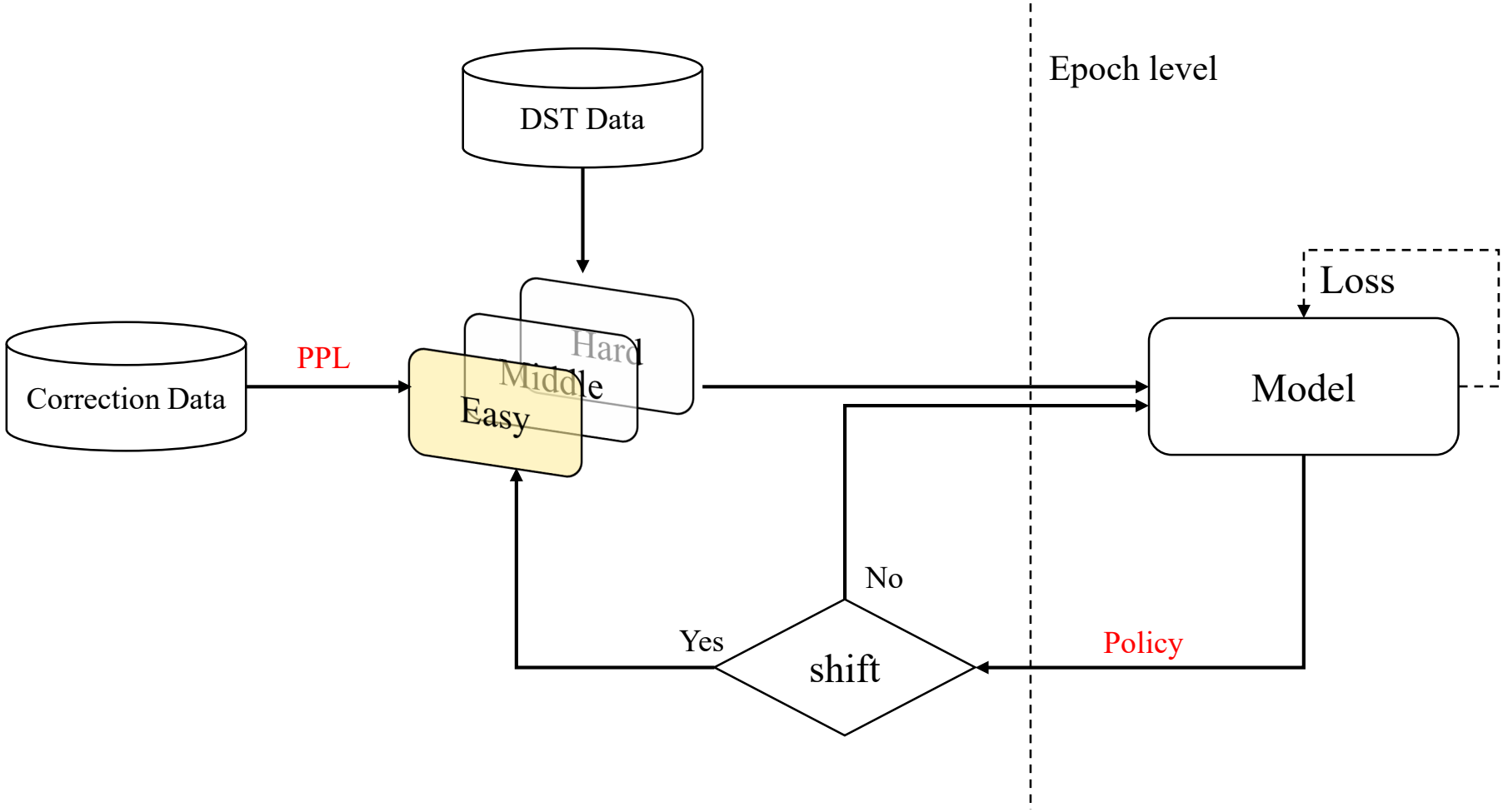
Multi-Task Pre-train



Prompt 模板

<DST> : “translate dialogue to belief state:”
<Correction>: “correct asr errors in text:”

引入 Curriculum Learning



Curriculum Learning策略分析

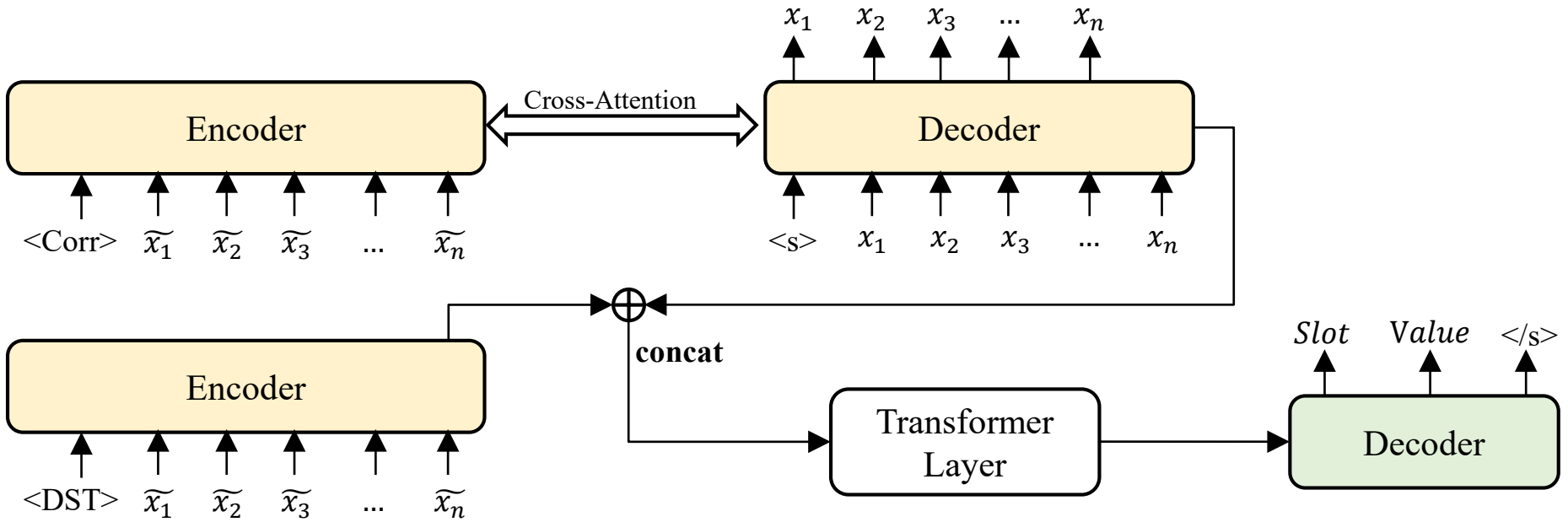
Setting	Joint_acc	WER
Easy_Dataset_Only	45.73	7.64
Middle_Dataset_Only	45.79	6.24
Hard_Dataset_Only	46.25	4.6
Easy_to_Hard	46.30	5.11
轮询	**46.79**	4.4

对比结论：

1. 轮询方式效果最好， Easy_to_Hard和只用Hard的效果相当。

降噪DST

Denoising DST

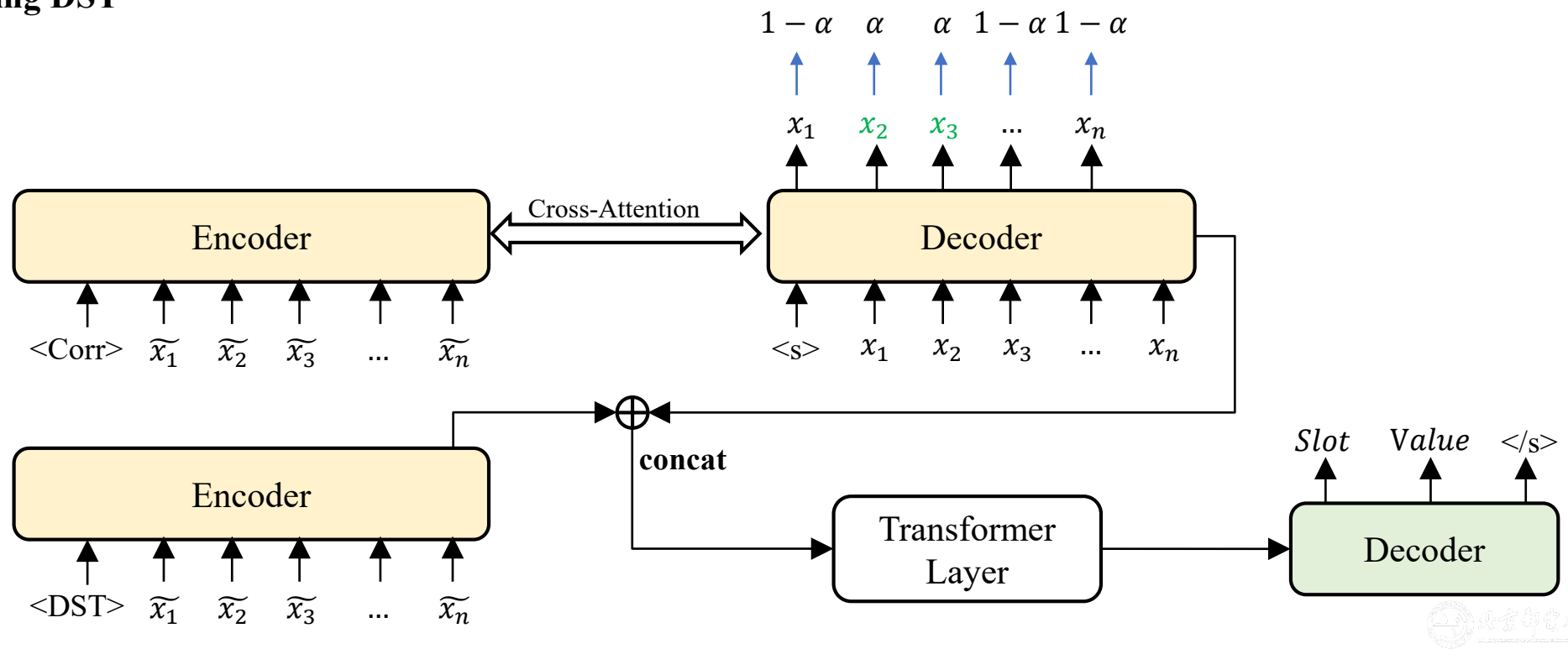


1、纠错的效果会影响后续DST，把纠错后的语义向量传给DST而不是用纠错后的token。

2、防止纠错把正确的纠错，两个向量拼接在一定程度上增强了鲁棒性。

降噪DST

Denoising DST



$$L_{corr} = \alpha \frac{1}{N_e} \sum_{i}^{N_e} \log p(y_i) + (1-\alpha) \frac{1}{N_w} \sum_{j}^{N_w} \log p(y_j)$$

$$L_{\mathrm{DST}} = \frac{1}{N_{\mathrm{D}}} \sum_{i}^{N_{\mathrm{D}}} \log p(y_i)$$

评价指标

词错率（Word Error Rate, WER）

$$WER = \frac{S + D + I}{N}$$

联合目标准确率（Joint Goal Accuracy）

$$\text{Joint}\_GA = \frac{\text{N}_{All\_right}}{N}$$

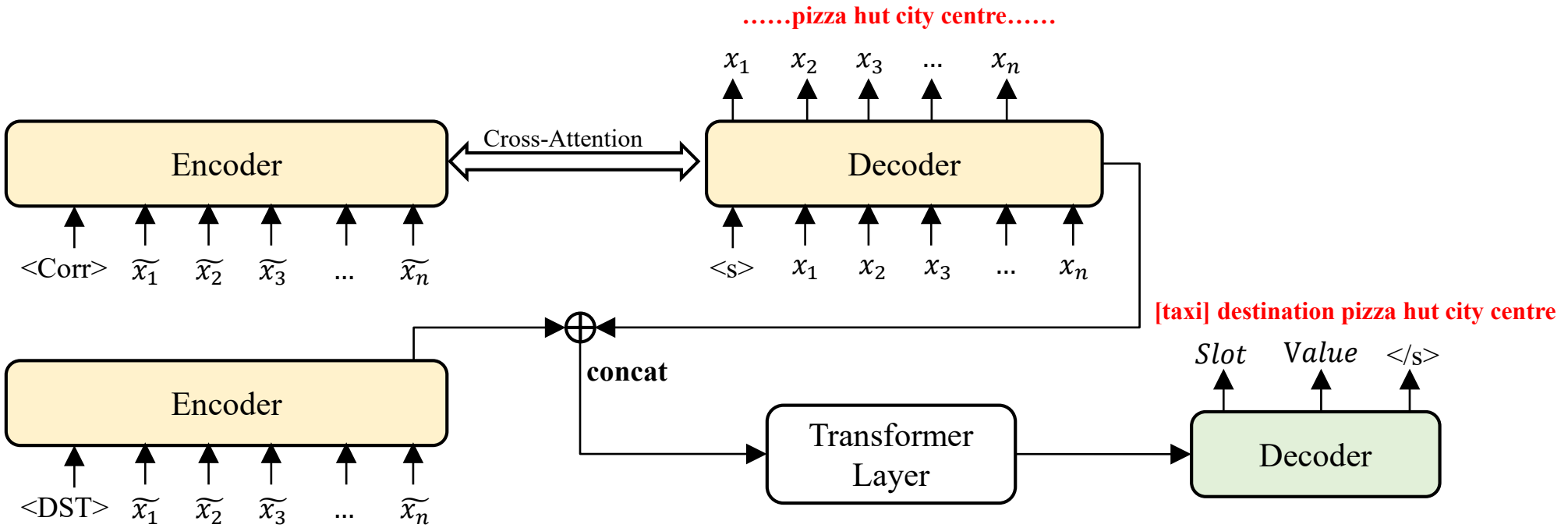
实验效果

DSTC2		
Method	**Slot-F1**	**Slot-Acc**
Nbest-BERT	**87.8**	**81.8**
WCN-BERT + HD	**87.33**	**80.9**
BART	**86.87**	**81.9**
T5	**87.00**	**82.01**
T5-correction	**87.96**	**82.84**
MultiWOZ-Noise		
Method	**Joint_acc**	**WER**
Pptod-base (ceiling)	**53.37**	**-**
Pptod-base	**36.55**	**11.4**
T5-base	**36.77**	**11.4**
T5-aug	**40.96**	**11.4**
Multi-Task Pre-train	**46.39**	**4.4**
w/o CL	**44.87**	**4.01**
w/o CL aug	**42.15**	**6.2**
Denoising DST	**47.01**	**3.9**
Denoising DST+EW	**47.62**	**3.52**

结果展示

仍然没解决

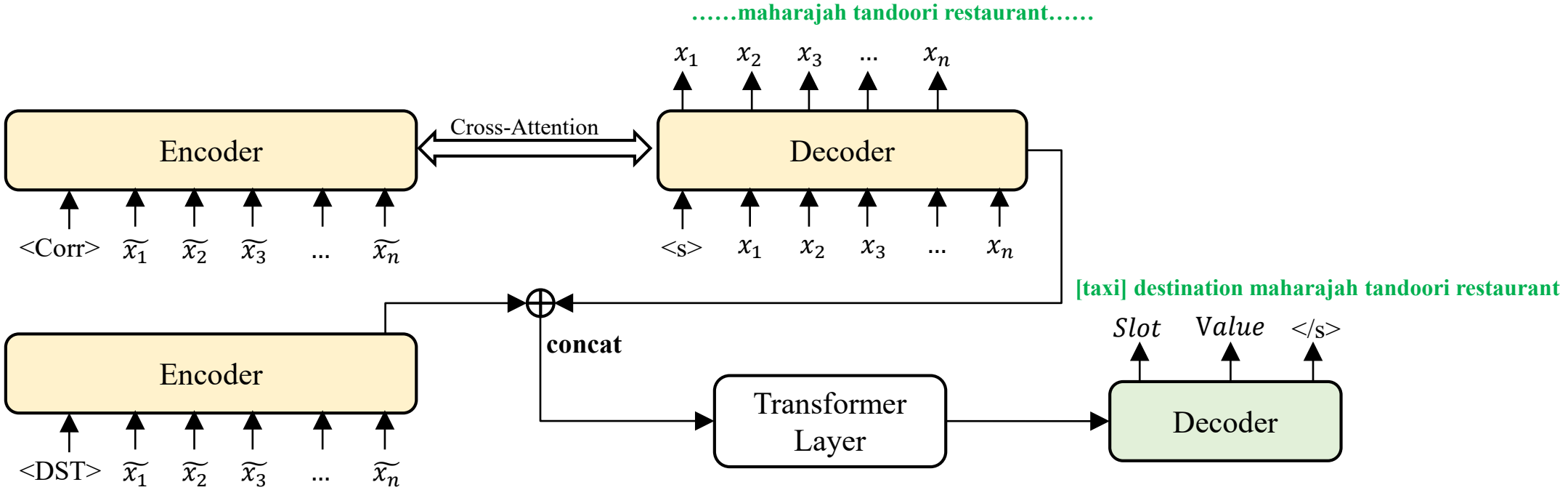
User: i would like a taxi from saint john s college to **petza hut fandidden**
Sys: what time do you want to leave and what time do you want to arrive by?
User: i want to leave after seventeen fifteen
Sys: you are welcome . is there anything else i can help you with today ?
User: thank you for all the help i appreciate it **pizza hut city centre**



结果展示

成功解决

User: i need to find information about a certain restaurant can you help with that
Sys: yes i can. what restaurant are you looking for?
User: it is called **maheraji tandori restaurant** **mahal of cambridge**
Sys: i have got your booking set , the reference number is [value_reference] .
User: can you book a table for seven people at twelve thirty on tuesday



目录

目前工作进度

数据增强
利用多种数据增强方式,产生更多的训练数据.

降噪DST
通过加载第一阶段预训练的模型参数，进一步微调对话状态追踪的效果

基于Prompt多任务预训练
基于 T5 设计了一个端到端的编解码器。通过不同的Prompt来进行多任务学习。

规律探究
通过手动调节噪声的功率，测试不同模型的极限性能。

目前工作进度

时间	研究内容	预期效果
2021.7-2021.11	前期调研	确定研究方向与目标，制订实施方案
2021.11–2021.12	基于多任务的降噪对话状态追踪预训练算法的调研，完成初期基线模型搭建，完成预训练所基于的数据集的构建	完成开题报告
2022.01-2022.07	优化基于多任务的降噪对话状态追踪预训练算法的效果	在数据集上达到主流水平
2022.07-2022.10	进行端到端的降噪对话状态追踪的初步探究	阶段性总结，完成阶段报告
2022.10-2022.11	小论文撰写	完成小论文
2022.11-2022.12	专利撰写	完成专利
2022.02-2022.05	学位论文的撰写	完成学位论文

目录

未来计划

1、小论文的撰写

大纲的制定和部分章节的中文版。（摘要、介绍和模型的一小结）

预计7月30日前能完成整个小论文的中文版。

8月份完成英文版的小论文，9月投稿。

2、整理系统：开机组网那边的工作。