ACL论文&投稿分享

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

姓　　名：王青悦
年　　级：2018级直博生
时　　间：2023年05月19日

主要内容

- 研究方向
- 论文工作
- 投稿经验

主要内容

- 研究方向
- 论文工作
- 投稿经验

研究方向

- 对话系统从功能上分为知识问答、聊天对话和任务型对话。其中，任务型对话主要针对具体的垂直应用领域，具有清楚的语义表达形式、明确的用户任务目标范围。

知识问答 | 闲聊对话 | 任务型对话

- 对话状态追踪（Dialogue State Tracking）主要负责理解用户目标，追踪用户需求。输入是对话文本，输出是当前轮对话的对话状态（槽-值对集合），评价指标：槽准确率（slot accuracy） 联合准确率 （joint slot accuracy）

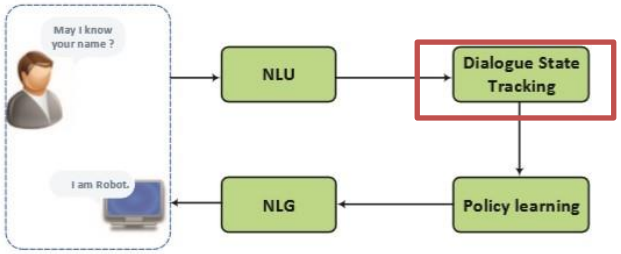


Figure 1: Traditional Pipeline for Task-oriented Systems.

Sys: Hi, what can I do for you?
User: Please find me a Chinese restaurant.
State: *restaurant-food=chinese*

Sys: Charlie Chan fits your criterion, can I book it for you?
User: Yes, I need a table on Monday at 12:15.
State: *restaurant-food=chinese; restaurant-name=charlie chan restaurant-book day=monday; restaurant-book time=12:15*

Sys: Booking is successful. Is there anything else I can assist you with today?
User: I also need a taxi to get me to the restaurant on time.
State: *restaurant-food=chinese; restaurant-name=charlie chan restaurant-book day=monday; restaurant-book time=12:15 taxi-destination=charlie chan; taxi-arriveby=12:15*

主要内容

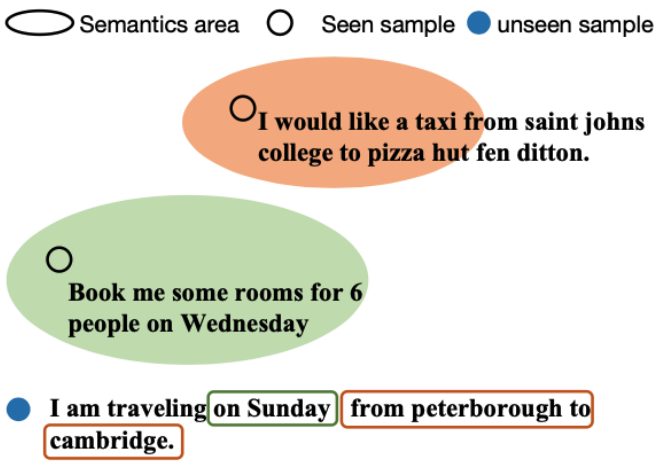
- 研究方向
- 论文工作
- 投稿经验

Divide, Conquer, and Combine: Mixture of Semantic-Independent Experts for Zero-Shot Dialogue State Tracking

Motivation:

- 已有工作往往尝试从**数据/模型层面**提升模型泛化性。数据层面：合成更多的对话样本或者利用其他任务的标注数据。模型层面：设计特定的组件或利用预训练模型。

- 这些方法并未探究**零样本泛化本质**，即缺乏语义解耦能力，无法将未见样本映射到已有空间。



Contribution：

- 提出来简单高效的三步模型框架——划分-解决-合并
- 仅仅使用**10M的可训练参数**，将模型在MultiWOZ数据集上的表现提升**5%~10%**

方法 (Method)

1. 划分(dividing)：根据可见样本的语义表示聚类划分语义空间
2. 解决(conquering)：为每一个语义独立的空间学习特定的专家
3. 合并(combining)：通过关系挖掘和聚合推理预测结果

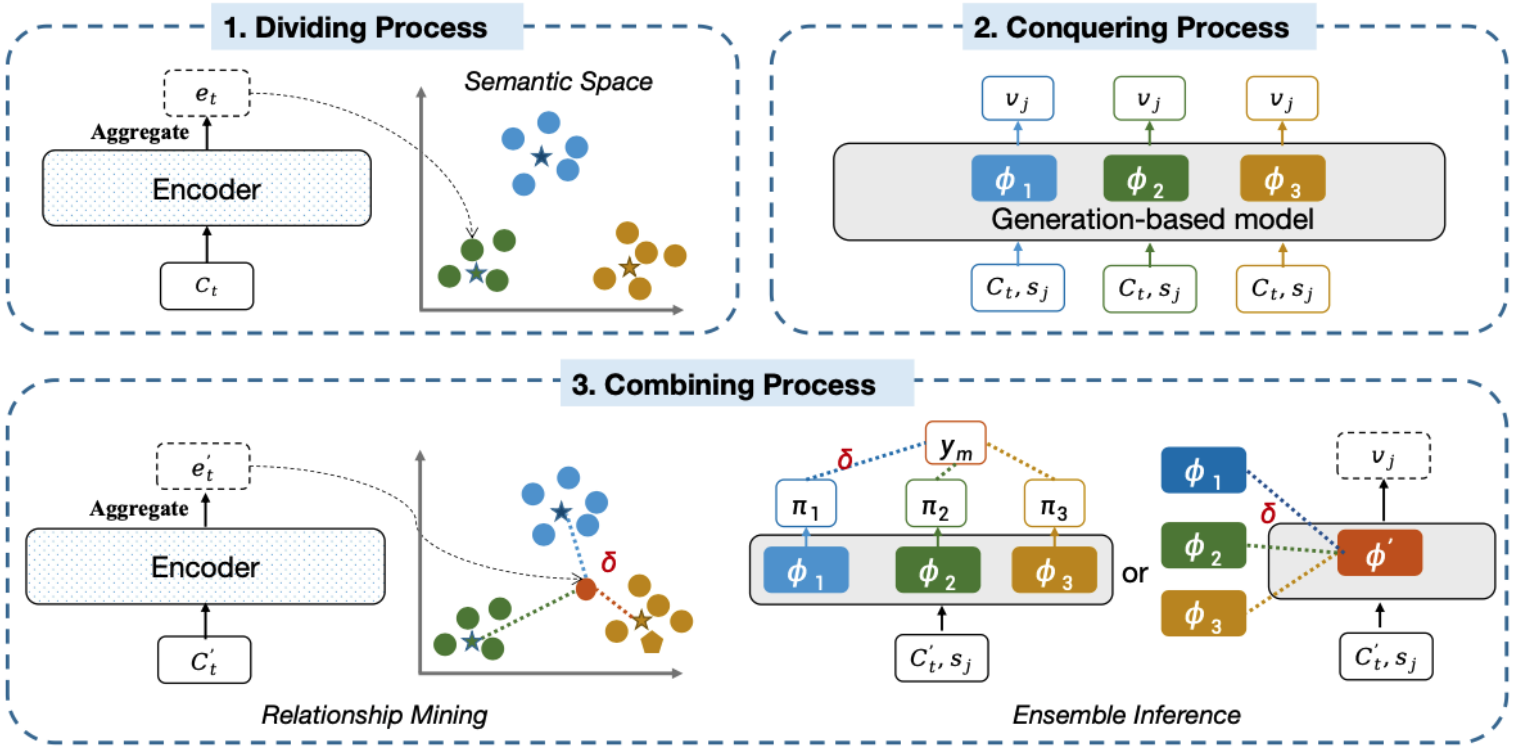


Figure 2: Illustration of our proposed schema (best viewed in color).

实验(Experiments)

主实验（**Main Results**）

我们的方法提升了零样本/全监督场景下的性能。

Model	#Trainable Parameters	Pretrained-model	Joint Goal Accuracy					
			Attraction	Hotel	Restaurant	Taxi	Train	Average
TRADE (Wu et al., 2019)	-	N	19.87	13.70	11.52	60.58	22.37	25.76
MA-DST (Kumar et al., 2020)	-	N	22.46	16.28	13.56	59.27	22.76	26.87
SUMBT (Lee et al., 2019)	440M	Bert-base	22.60	19.80	16.50	59.50	22.50	28.18
T5DST (Lin et al., 2021b)	60M	T5-small	33.09	21.21	21.65	64.62	35.42	35.20
T5DST †(Lin et al., 2021b)	220M	T5-base	35.51	22.48	25.04	65.93	37.82	37.36
SlotDM-DST (Wang et al., 2022)	60M	T5-small	33.92	19.18	20.75	66.25	36.96	35.55
SlotDM-DST (Wang et al., 2022)	220M	T5-base	37.83	26.50	27.05	**69.23**	40.27	40.18
TransferQA (Lin et al., 2021a)	770M	T5-large	31.25	22.72	26.28	61.87	36.72	35.77
T5-Adapter†	0.8M	T5-small	33.85	18.22	19.62	64.93	32.25	33.77
	3.6M	T5-base	39.98	23.28	28.58	65.03	36.98	38.77
Ours (Param-level)	0.8M×K	T5-small	34.63	24.22	22.07	65.41	33.88	36.02
Ours (Token-level)			35.82	24.78	22.86	65.87	40.27	**37.92**
Ours (Param-level)	3.6M×K	T5-base	41.28	26.15	31.05	66.64	38.72	40.76
Ours (Token-level)			**41.35**	**27.72**	**33.76**	66.90	**43.81**	**42.71**

Table 1: Zero-shot results on MultiWOZ 2.1 dataset. All numbers are reported in joint goal accuracy (%) and the best results among each setting are bolded. K is a hyper-parameter and refers to the number of sub-sets. Expect for †, all results of baselines come from the original papers.

Model	#Trainable Parameter	Pre-trained Model	JGA
TRADE	-	N	45.60
STARC (Gao et al., 2020)	440M	Bert-base	49.48
SGD-baseline	440M	Bert-base	43.40
T5DST	220M	T5-base	53.15
T5-Adapter	3.6M	T5-base	52.14
Ours (Param-level)	3.6M×K	T5-base	52.54
Ours (Token-level)	3.6M×K	T5-base	54.35

Table 3: Full data results on MultiWOZ 2.1 dataset. For a fair comparison, only those generative models with the ability of zero-shot inference are listed here.

实验

消融实验（Ablation Study）

1. 聚类算法的影响：

（1） 使用任何聚类算法都比不聚类要好

（2）GMM算法要优于Kmeans,更好的聚类带来更高的提升

2. 聚类个数的影响

3. 聚合时温度的影响

4. 聚合时权重的影响

（1） 使用average权重

（2） 使用argmax权重

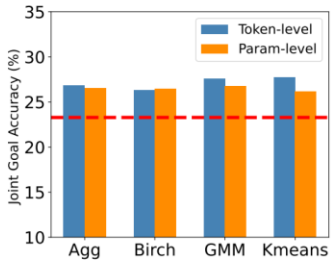


Figure 3: Impact of different clustering algorithms.

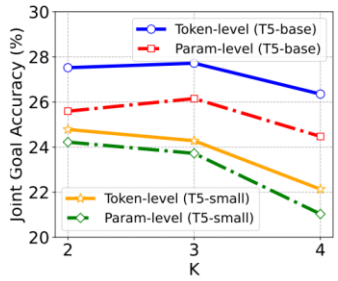


Figure 4: Impact of different numbers of sub-sets.

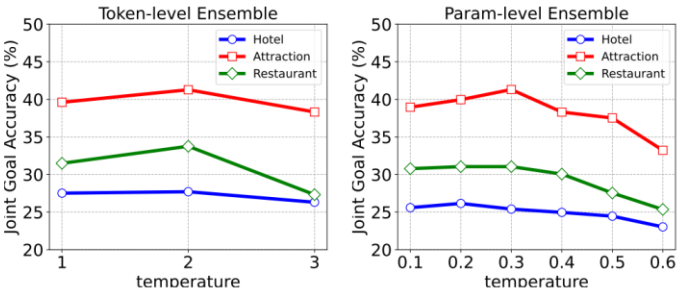


Figure 5: Impact of different temperatures τ.

Weights	Hotel		Taxi	
	Param-level	Token-level	Param-level	Token-level
Ours	26.15	27.72	66.64	66.90
Argmax	24.47	24.85	65.09	66.38
Average	20.62	25.87	59.61	65.51

Table 4: The Impact of weight in combing process.

实验

- 对聚类的理解（Analysis on Clustering）

1. 对不同的编码器健壮
2. 对数据有明确的语义划分
3. 优于按领域划分数据的方法

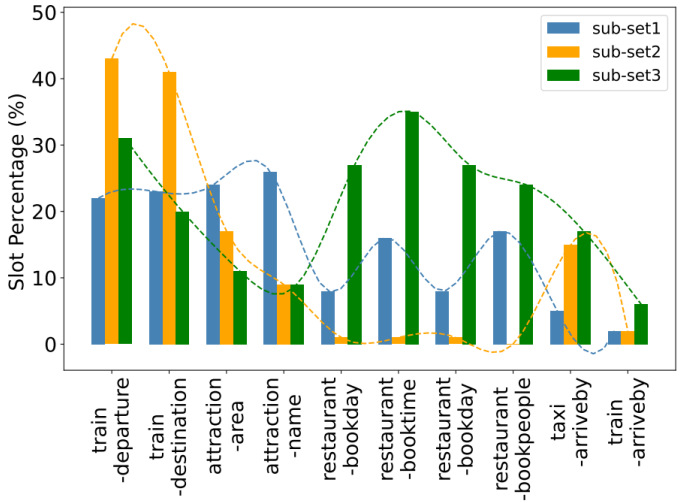


Figure 7: Statistics of slot distribution across sub-sets.

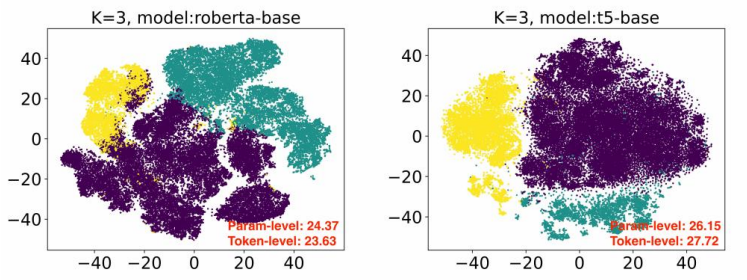


Figure 6: The t-SNE visualizations of clustered subsets represented with T5 and RoBERTa. "Token-level" and "Param-level" show the zero-shot performance.

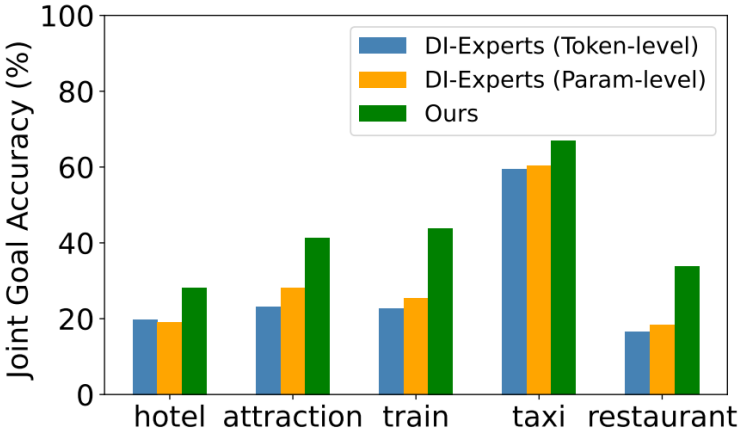


Figure 8: The zero-shot performance of DI-Experts.

实验

- **对合并的理解 （Analysis on Ensemble Inference）**

（1）整合了独立专家优势

（2）轻量级的计算开销

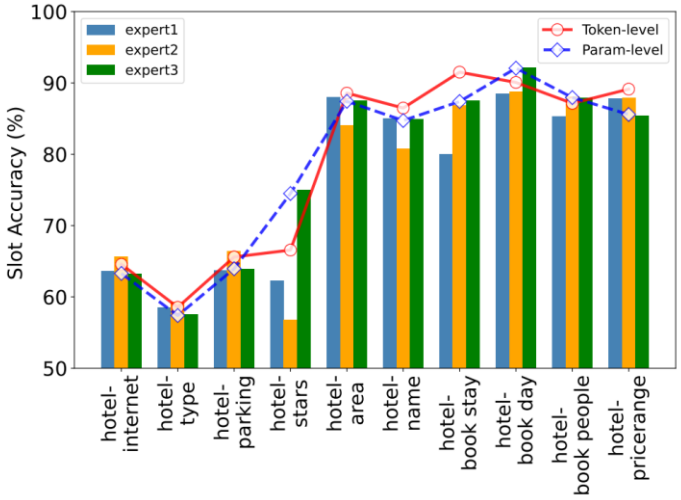


Figure 9: Slot accuracy of different single experts and ensemble models on hotel domain.

Model	Training $\|\Theta\|$	Inference $\|\Theta\|$	Average (%)
T5DST	100%	100%	37.36
T5-Adapter	1.6%	+1.6%	37.92
Ours (Param-level)	4.9%	+1.6%	40.76$^{\Uparrow +3.4}$
Ours (Token-level)	4.9%	+4.9%	42.71$^{\Uparrow +5.4}$

Table 5: Costs for training and inference of methods. $|\Theta|$ denotes the number of trained/ deployed parameters for training and inference, respectively.

实验

- **对已有工作的补充**

(1) 数据层面： 使用数据增强的结果 (2) 模型层面： 使用特殊结构的模型

Model	Raw Data	Augmented Data
TRADE	19.50	28.30
Ours (Param-level)	26.15	27.56$^{\Uparrow +1.4}$
Ours (Token-level)	27.71	29.36$^{\Uparrow +1.7}$

Table 6: Complementarity between ours and data augmentation methods, in terms of zero-shot performance on hotel domain.

Model	Attraction	Hotel	Taxi
SlotDM	36.38	25.45	67.21
+Our Framework	37.41$^{\Uparrow +1.0}$	26.58$^{\Uparrow +1.1}$	68.02$^{\Uparrow +0.8}$

Table 7: Complementarity between ours and competitive model-level methods “SlotDM”, in terms of zero-shot performance on three domains.

总结

1. 我们提出了简单的“划分-解决-合并”来提高DST任务中零样本的泛化性。

2. 实验结果表面我们的模型使用少量的可训练数据实现显著的性能提升。

主要内容

- 研究方向
- 论文工作
- 投稿经验

投稿经历

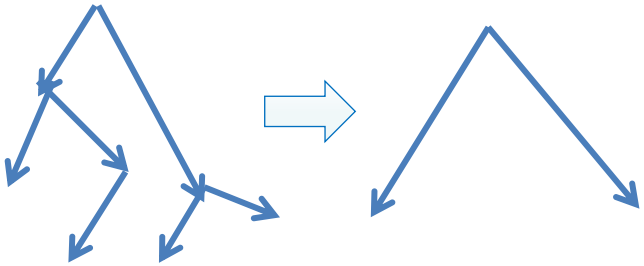
- 2018 KSEM Accept （CCF-C）
- 2019 IJCAI Reject
- 2020 COLING Reject
- 2020 ICASSP Reject
- 2021 IJCNN Accept （CCF-C）
- 2021 AAAI Reject
- 2021 ACL Reject
- 2021 SIGIR Reject
- 2022 COLING Accept （CCF-B）
- 2022 TASLP Reject
- 2023 CogSci Accept (CCF-B)
- 2023 ACL Accept (CCF-A)

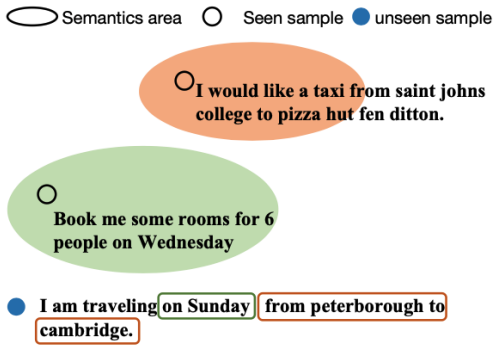
13次投稿，8次被拒

Introduction

1. （有效）总结已有工作和存在问题
2. （清晰）论述本文动机和方法
3. （尽可能）突出本文方法优势

切忌平铺直叙，内容有的放矢





pected to improve the zero-shot performance. In practice, we design a three-step framework, where stages 1&2 are for training and stage 3 is for inference: ❶dividing: encode and cluster the semantics of seen data into subsets, ❷conquering: train expert for each subset with dialogue state labels, and ❸combining: mine the relationship between newly unseen sample and seen semantics, and perform ensemble inference with weighted experts.

Experimentally, we implement our framework upon T5-Adapter and demonstrate the effectiveness and universality of our proposed schema. Specifically, we achieve averaging 5%∼10% improvement on the MultiWOZ benchmark with negligible training and deployment costs, achieving state-of-the-art zero-shot performance under settings without external information. Comprehensive analyses are reported to provide some insights to better understand our method.

Method

1. **整体结构直观**，根据情况考虑对应题目和配图。
2. **方法表述清晰**（相信是可以说清楚的），适当弱化/抹去繁琐的实现细节（会给审稿人带来更多困惑）。

4 Methodology

Overviews Figure 2 illustrates the overview of our method following three steps. In the ❶dividing process, a context encoder f encodes seen dialogue contexts into representations to construct semantic space $\mathcal{E}$. These samples are then divided into several sub-sets by clustering. After that, We train semantic-independent DST experts using labeled states of sub-sets, also called the ❷conquering process. During ❸combining, we first estimate the relationships δ between seen data and unseen sample C_t', and perform the weighted mixture-of-experts inference conditioned on δ for the unseen sample.

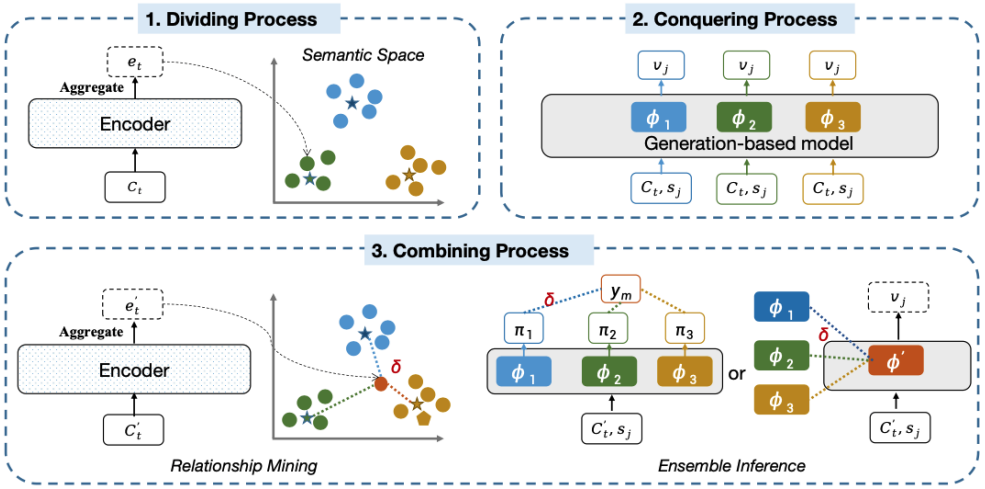


Figure 2: Illustration of our proposed schema (best viewed in color).

Experiments

深度学习的可解释性不太强，多做（有用）实验

1. 实验不仅要有更要有用，审稿人可能会对哪些模块的效果产生疑问，添加实验消除疑惑。
2. 配图简单清晰，巧用特殊的标记（下划线，加粗，颜色等等）
3. 文字表述有目的性，知道每一段表述试图说明什么，切忌泛泛而谈。

Impact of Clustering Algorithms We study the effect of different clustering algorithms, including Kmeans (Hartigan and Wong, 1979), Birch (Zhang et al., 1996), Agglomerative (Gowda and Krishna, 1978), and GMM (Yang et al., 2012) on hotel domain in Figure 3. As shown, 1) all clustering algorithms perform better than the T5-Adapter (Red dotted line), showing the effectiveness and stability of our framework; and 2) GMM achieves the best performance on parameter-level ensemble inference while our chosen Kmeans wins on token-level ones. We believe advanced clustering may bring better division, thus achieving further improvement, which will be investigated in future work.

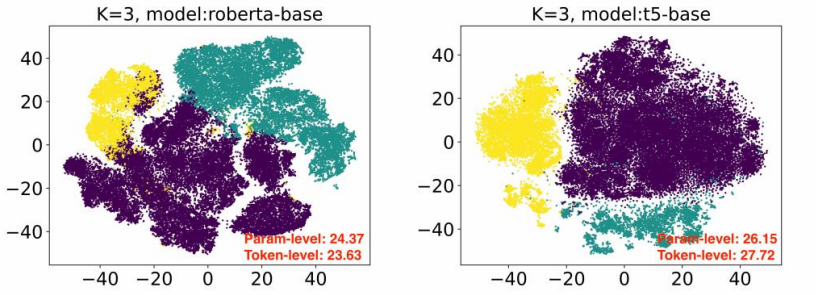


Figure 6: The t-SNE visualizations of clustered subsets represented with T5 and RoBERTa. "Token-level" and "Param-level" show the zero-shot performance.

Rebuttal

考场上，有人想问你借一只铅笔，你不仅仅给了他铅笔，还给橡皮，尺子，草稿纸，那么你的态度就到位了。

- “Can you please submit Figures 4 and 5 for the SGD dataset in the rebuttal?”

- “ In their response to the reviewers' questions, **the author provided some additional useful information** that I encourage them to include in the camera-ready version. I would recommend to accept the paper to ACL.”

- soundness 3/3/4 ->3/4/4
- excitement 3.5/4/4.5 -> 3.5/4/4.5

总结

日常科研：

1. 工作动机（多读论文 & 多想故事）
2. 实验分析 （多分析 & 多思考 & 多总结）

写作原则：

1. 黄金比例：好的写作远远重要于代码实验。
2. 去做个艺术品：从头到脚（文字+配图）重视细节。
3. 努力好好表现：不要废话和啰嗦，珍惜论文的每一处空间。

总结

日常科研：

1. 工作动机（多读论文 & 多想故事）
2. 实验分析 （多分析 & 多思考 & 多总结）

写作原则：

1. 黄金比例：好的写作远远重要于代码实验。
2. 去做个艺术品：从头到脚（文字+配图）重视细节。
3. 努力好好表现：不要废话和啰嗦，珍惜论文的每一处空间。

路漫漫其修远兮，祝我们都能做出漂亮的工作～

谢谢

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS