A Comparative Study on Schema-Guided Dialogue State Tracking

Jie Cao[†] , Yi Zhang[‡]
[†]School of Computing, University of Utah
[‡]AWS AI, Amazon

Introduction

Flight Service 1 and 2 shares overlapping functionalities while using different intent/slot tags.

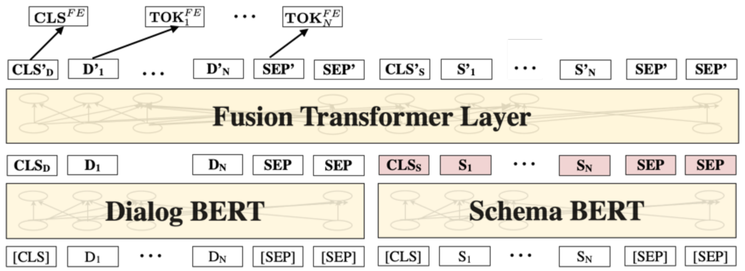
Without retraining, can a dialog model trained on Flight Service 1 also support Flight Service 2?

Schema-guided Dialog uses natural language description to explain each intent and slot, thus it may share knowledges across multiple services in multiple domains.

We study the following three research problems **(Q1, Q2, Q3)** on four subtasks:

- Intent
- Requested Slot
- Categorical Slot
- Non-Categorical slot

Q1: Dialog & Schema Description Encoding



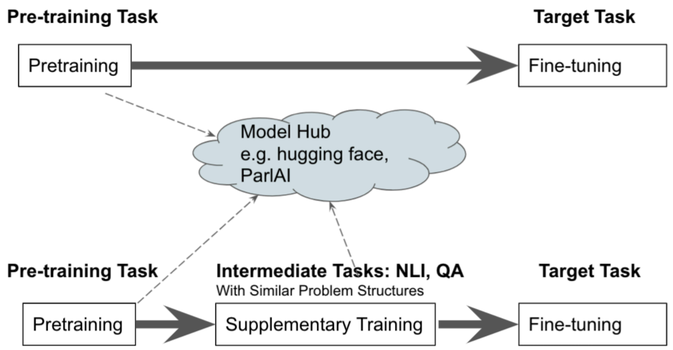
Pink boxes means the representation are cached

Method/Task	SG-DST Acc Intent	SG-DST F1 Req	SG-DST Joint Acc Cat	SG-DST Joint Acc NonCat	SG-DST Joint Acc All	MULTIWOZ 2.2 Joint Acc Cat	MULTIWOZ 2.2 Joint Acc NonCat	MULTIWOZ 2.2 Joint Acc All
Seen Services								
Dual-Encoder	94.51	99.62	87.92	47.77	43.20	79.20	79.34	65.64
Fusion-Encoder	94.90	**99.69**	88.94	48.78	58.52	81.37	80.58	67.43
Cross-Encoder	**95.55**	99.59	**93.68**	**91.85**	**87.58**	**85.99**	**81.02**	**71.93**
Unseen Services								
Dual-Encoder	89.73	95.20	42.44	31.62	19.51	56.92	50.82	31.83
Fusion-Encoder	90.47	95.95	48.79	35.91	22.85	57.01	52.23	33.64
Cross-Encoder	**93.84**	**98.26**	**71.55**	**74.13**	**54.54**	**59.85**	**59.62**	**38.46**

Partial-attention balance between speed and accuracy ?

By caching the token embedding instead of the single CLS embedding, a simple partial-attention ***Fusion-Encoder*** can achieve much better performance than ***Dual-Encoder***, while still infers two times faster than ***Cross-Encoder***

Q2. Supplementary Training



	SG-DST Intent seen	SG-DST Intent unseen	SG-DST Req seen	SG-DST Req unseen	SG-DST Cat seen	SG-DST Cat unseen	SG-DST NonCat seen	SG-DST NonCat unseen
Δ_{SNLI}	+0.02	**+0.68**	+0.38	-0.38	-2.87	-1.23	-0.1	-6.25
Δ_{SQuAD}	-0.17	-1.32	-0.01	-0.33	-3.02	-5.17	-1.79	**+3.25**

How supplementary training helps?

- **SNLI only** helps for Intent (emphasizing the whole sentence entailment), although Req and Cat are also sentence-pair classification tasks.
- **SQuAD** consistently helps for non-categorical slot identification tasks, due to span-based retrieving
- Supplementary training helps more on unseen services.

Q3. Impact of Description Styles

style	Intent Description	Slot Description
Identifer	intent_1	slot_4
NameOnly	CheckBalance	account_type
Q-Name	Is the user intending to CheckBalance?	What is the value of acctount_type ?
Orig	Check the amount of money in a user's bank account	The account type of the user
Q-Orig	Does the user want to check the amount of money in the bank account ?	What is the account type of the user ?
Name-Para	CheckAccountBalance	user_account_type
Orig-Para	Check the balance of the user's bank account	Type of the user account

Homogeneous Evaluation

Style\Task	SG-DST Intent	SG-DST Req	SG-DST Cat	SG-DST NonCat	MULTIWOZ 2.2 Cat	MULTIWOZ 2.2 NonCat
Identifer	61.16	91.48	62.47	30.19	34.25	52.28
NameOnly	**94.24**	98.84	74.01	75.63	53.72	56.18
Q-Name	93.31	**98.86**	74.36	74.86	**54.19**	56.17
Orig	93.01	98.55	74.51	75.76	52.19	57.20
Q-Orig	93.42	98.51	**76.64**	**76.60**	53.61	**57.80**

Is named-based description enough?

- Most name are meaningful, and perform **not bad,** especially **on** Intent/Req subtasks
- Rich description outperforms the name-based on **NonCat**, but inconsistent on other tasks.

Style/Dataset	SG-DST seen	SG-DST unseen	MULTIWOZ 2.2 seen	MULTIWOZ 2.2 unseen
Orig	-1.79	**+3.25**	-2.21	**+4.27**
Q-Orig	-2.01	**+8.84**	-1.28	**+3.06**
NameOnly	-1.49	-0.11	+0.58	**+1.77**
Q-Name	-2.98	**+1.04**	-0.32	**+1.25**

Is question format helpful?

- It generally helps on Cat/NonCat
- Adding it to rich description will benefit more from SQuAD2 supplementary training on unseen. However, not on MultiWOZ.

Heterogeneous Evaluation

Style\Task	SG-DST Intent(Acc) mean	Δ	Req(F1) mean	Δ	Cat(Joint Acc) mean	Δ	NonCat(Joint Acc) mean	Δ
NameOnly	82.47	-11.47	96.92	-1.64	61.37	-5.54	56.53	-14.68
Q-Name	**93.27**	+0.58	**97.88**	-0.76	68.55	+2.63	62.92	-6.30
Orig	79.47	-12.70	97.42	-0.74	**68.58**	-0.3	**66.72**	-3.11
Q-Orig	84.57	-8.24	96.70	-1.45	68.40	-2.89	56.17	-15.00
	para	Δ	para	Δ	para	Δ	para	Δ
NameOnly	**92.22**	-1.74	97.69	-0.87	67.39	-0.7	67.17	-4.04
Orig	91.54	**-0.63**	**98.42**	**+0.26**	**71.74**	**+2.86**	**67.68**	**-2.16**

What if unseen service in different description styles?

- For unseen styles, all tasks surfer from inconsistencies, though to varying degrees
- For paraphrased styles, richer description are relatively more robust than named-based descriptions.