Dialogue Generation on Infrequent Sentence Functions via Structured Meta-Learning

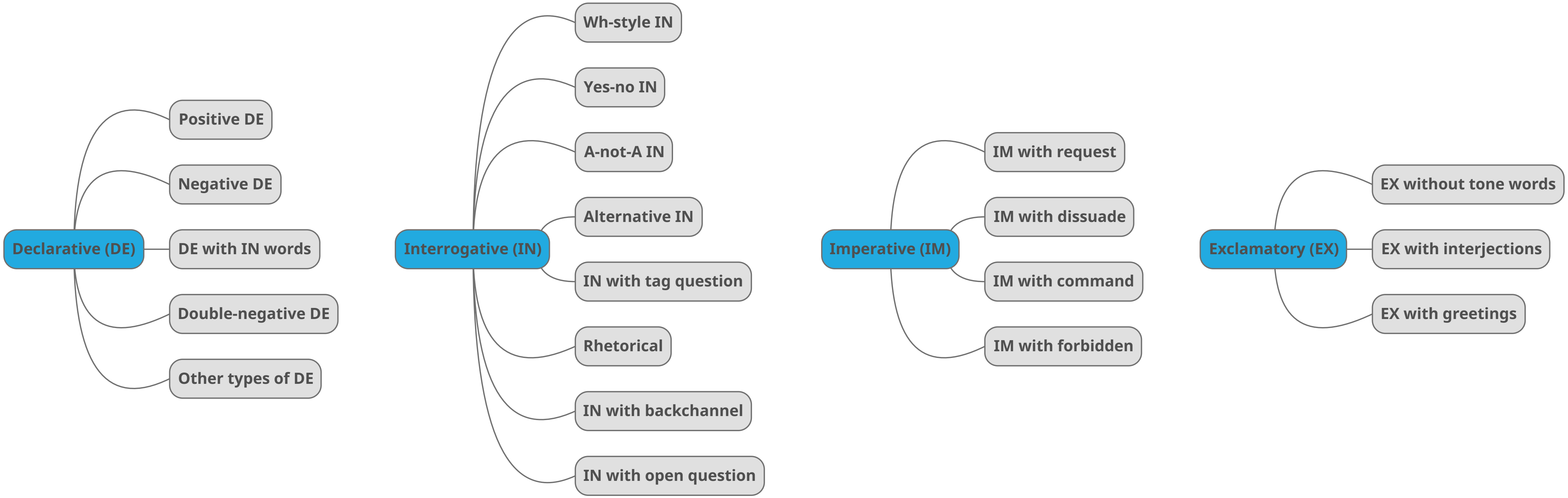
Yifan Gao[1], Piji Li[2], Wei Bi[2], Xiaojiang Liu[2], Michael R. Lyu[1], and Irwin King[1]

1. The Chinese University of Hong Kong   2. Tencent AI Lab

Introduction

Sentence Function

Definition: “In linguistics, a sentence function refers to a speaker's purpose in uttering a specific sentence, phrase, or clause.” [Wikipedia]



Introduction

Sentence Functions in Conversation

- Humans express intentions in conversations through sentence functions.
- Sentence functions have great influences on the structures of utterances in conversations including word orders, syntactic patterns, and other aspects.

Sentence Function	Frequent Patterns		Sentence Examples	
	Chinese	English	Chinese	English
Wh-style IN	x在哪y? 谁是x?	Where does x y? Who is x?	周末在哪过啊 谁是天蝎座	Where do you spend your weekend Who is a Scorpio
Yes-no IN	x是在y吗? x是指y吗?	Is x y? Does x y?	你是在云南吗? 你是指昨天的篮球比赛吗?	Are you in Yunnan? Do you mean the basketball match yesterday?
Alternative IN	x还是y x y哪个	x or y x y which	狮子和白羊真配还是假配? 香蕉和苹果哪个卖得比较好?	Leo and Aries go together or not? Which sells better, banana or apple?

Frequent word patterns of three level-2 Interrogative sentence functions. (Bi et al., ACL 2019)

Introduction

Imbalance Problem in Large-scale Conversation Dataset with Sentence Function Annotation

- Existing work shows that the use of sentence functions improves the overall quality of generated responses (Ke et al., ACL 2018).
- However, the number of utterances for different types of fine-grained sentence functions is usually extremely imbalanced. In the large-scale dataset STC-SeFun (Bi et al., ACL 2019):

Sentence Function	Query	Response
Declarative (DE)		
Positive DE	49,223 (48%)	67,540 (57%)
Negative DE	9,241(9%)	18,428(16%)
DE with IN words	887(.9%)	2,660(2%)
Double-negative DE	40(<.1%)	99(.1%)
Other types of DE	2,675(3%)	5,218(4%)

Dialogue generation models suffer from data deficiency for these infrequent sentence functions!

Proposed Approach

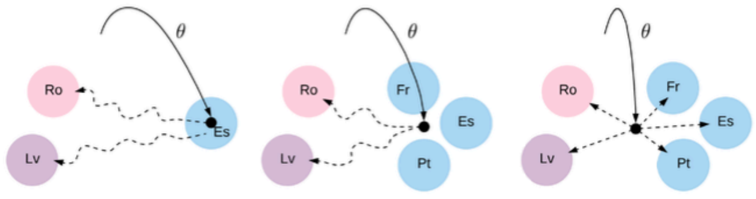
Model-Agnostic Meta-Learning (MAML)

Model-Agnostic Meta-Learning (MAML) can learn from a variety of **tasks** such that it can solve new learning tasks using only **a small number** of training samples.

- Training: Learn transferable internal representations across tasks (task=domain).
- Testing: Quickly adapt to a new task using only a few datapoints and training iterations.

Adapting to *new languages*

Low-Resource Neural Machine Translation

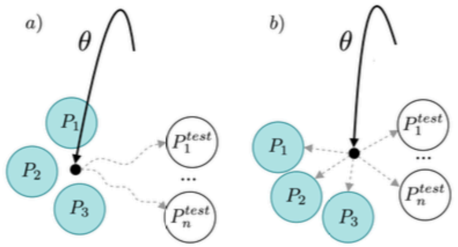


(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning

Learn to translate new language pair
w/o a lot of paired data?
Gu et al. EMNLP '18

Adapting to *new personas*

Personalizing Dialogue Agents



Adapt dialogue to a persona
with a few examples
Lin, Madotto* et al. ACL '19*

[Credit: Finn and Levine, 2019]

Adapting to new sentence functions?

Proposed Approach

Problem Formulation

A Single Task: response generation conditioned on a query-response sentence function pair (d_X, d_Y)

Training Data: K high-resource tasks: $D^k_{train} = \{(X^k_n, Y^k_n, d^k_X, d^k_Y), n = 1...N\}, k = 1...K$

Testing Data (Target): T tasks with infrequent sentence function: $D^t_{target} = \{(X^t_n, Y^t_n, d^t_X, d^t_Y), n = 1...N'\}, t = 1...T \ N' \ll N$

Training:

$$\underbrace{f_\theta}_{\text{Model}} : \underbrace{X^k}_{\text{Query}} \times (d^k_X, d^k_Y) \rightarrow \underbrace{Y^k}_{\text{Response}}, \underbrace{k = 1...K}_{\text{Task}}$$

Testing:

$$\underbrace{f_{\theta^*}}_{\text{Adapted Model}} = \arg\max_\theta \log p(f_\theta | D^t_{target}, \underbrace{f_{\theta_0}}_{\text{Trained Model}})$$

Proposed Approach

Base Model: C-Seq2Seq

We use a conditional sequence-to-sequence learning model as our base model.

- Attentional Sequence-to-Sequence Model
- Learn an additional query-response sentence function embedding for each query-response type
- The sentence function embedding is used at every decoding step:

$$\mathbf{u}_t = \mathrm{LSTM}(\mathbf{u}_{t-1}, [\underbrace{\mathbf{w}_t}_{\text{Word Emb}}; \underbrace{\mathbf{s}_k}_{\text{Sentence Function Emb}}])$$

Proposed Approach

MAML for C-Seq2Seq

Goal behind MAML: the conditions between task adaptation (fine-tuning) stage and training stage must match.

Algorithm 2 Training of MAML

Require: $\mathcal{E}$: distribution over tasks $\{\mathcal{T}_1, ..., \mathcal{T}_K\}$

Require: α, β: step size hyperparameters

Randomly initialize θ

while not done **do**

  Sample a batch of tasks $\mathcal{T}_k \sim \mathcal{E}$

  for all $\mathcal{T}_k$ **do**

    Sample $D_{\mathcal{T}_k}, D'_{\mathcal{T}_k}$ from $\mathcal{T}_k$

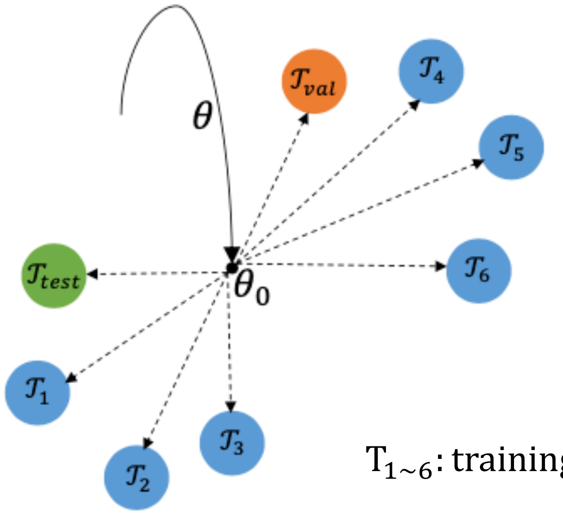
    Evaluate $\nabla_{\theta_k}\mathcal{L}(f_{\theta_k})$ with respect to $D_{\mathcal{T}_k}$

    Update $\theta'_k = \theta_k - \alpha\nabla_{\theta_k}\mathcal{L}(f_{\theta_k})$

  end for

  Update $\theta \leftarrow \theta - \beta\sum_k \boxed{\nabla_\theta}\mathcal{L}(f_{\theta'_k})$ with respect to all $D'_{\mathcal{T}_k}$

end while



$T_{1\sim6}$: training tasks

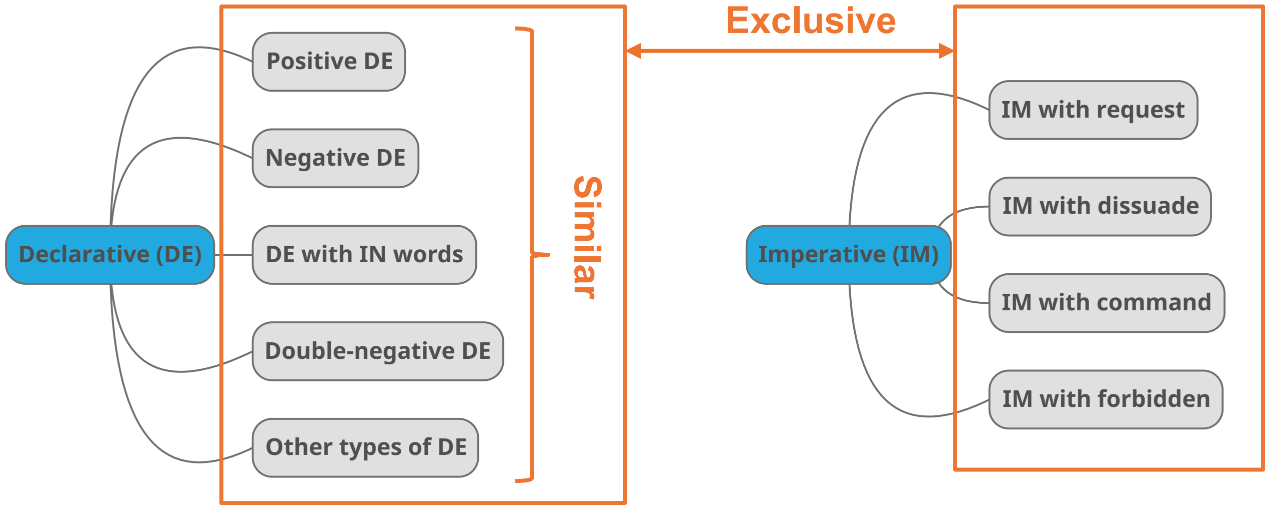
Findings of ACL: EMNLP 2020, Fine-Grained Sentence Functions for Short-Text Conversation

Proposed Approach

Exploring Structured Modeling

MAML assumes all tasks in training and adaptation stages distributed uniformly.

In conditioned response generation, some tasks may share some similarities while some are exclusive to each other.



Proposed Approach

Exploring Structured Modeling: Structured Meta-Learning (SML)

Task Representation Learning: sentence function embeddings are used to interact with each other via a gated self-attention mechanism

Task-Specific Knowledge Adaptation: the self-attended representations of these sentence functions are used as parameter gates to tailor the transferable knowledge of the meta-learned prior parameters.

Proposed Approach

Task Representation Learning

1. Sentence functions (tasks) seen in training: $\boldsymbol{S} = [\boldsymbol{s}_1, \dots, \boldsymbol{s}_K]$
2. Self-matching Operation:

$$\mathbf{a}_k = \mathrm{softmax}(\mathbf{S}^\top \mathbf{s}_k),\ \mathbf{m}_k = \mathbf{S}\mathbf{a}_k$$
$$\mathbf{f}_k = \tanh(\mathbf{W}_f[\mathbf{s}_k; \mathbf{m}_k]),$$

3. Gated summation for the final sentence function representation:

$$\mathbf{g}_k = \mathrm{sigmoid}(\mathbf{W}_g[\mathbf{s}_k; \mathbf{m}_k])$$
$$\tilde{\mathbf{s}}_k = \mathbf{g}_k \odot \mathbf{f}_k + (1 - \mathbf{g}_k) \odot \mathbf{s}_k$$

4. $\tilde{\boldsymbol{s}}_k$ replaces $\boldsymbol{s}_k$ as input at each decoding time step
5. In the testing stage, new sentence functions can benefit from the learned sentence functions for fast adaptation.

Proposed Approach

Task-Specific Knowledge Adaptation

To adapt globally transferable knowledge θ_0 to each sentence function, we design a parameter gate $\mathbf{o}_k$ for θ_0:

$$\mathbf{o}_k = \mathrm{FC}^{\sigma}_{\mathbf{W}_p}(\tilde{\mathbf{s}}_k), \qquad \theta_{0k} = \theta_0 \odot \mathbf{o}_k$$

fully connected layer parameterized by W_p and activated by a sigmoid function σ

element-wise multiplication

Intuitively, sentence functions with similar representations will activate similar initial parameters while dissimilar sentence functions trigger different ones.

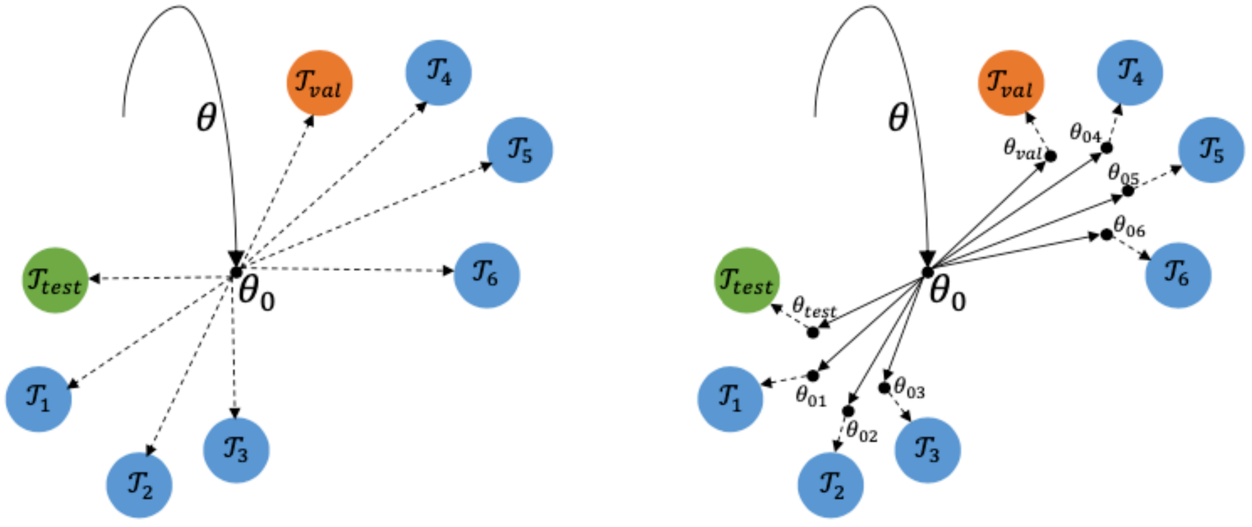
Proposed Approach

MAML vs. SML

Algorithm 3 Training of SML

Require: $\mathcal{E}$: distribution over tasks $\{\mathcal{T}_1, ..., \mathcal{T}_K\}$
Require: α, β: step size hyperparameters

Randomly initialize θ
while not done **do**
  Sample a batch of tasks $\mathcal{T}_k \sim \mathcal{E}$
  for all $\mathcal{T}_k$ **do**
    Sample $D_{\mathcal{T}_k}, D'_{\mathcal{T}_k}$ from $\mathcal{T}_k$
    Compute task representation $\tilde{\mathbf{s}}_k = \mathbf{g}_k \odot \mathbf{f}_k + (1 - \mathbf{g}_k) \odot \mathbf{s}_k$
    Compute $\mathbf{o}_k = \mathrm{FC}^{\sigma}_{\mathbf{W}_p}(\tilde{\mathbf{s}}_k), \theta_{0k} = \theta_0 \odot \mathbf{o}_k$
    Evaluate $\nabla_{\theta_{0k}} \mathcal{L}(f_{\theta_{0k}})$ with respect to $D_{\mathcal{T}_k}$
    Update $\theta'_{0k} = \theta_{0k} - \alpha \nabla_{\theta_{0k}} \mathcal{L}(f_{\theta_{0k}})$
  end for
  Update $\theta \leftarrow \theta - \beta \sum_k \nabla_{\theta'_{0k}} \mathcal{L}(f_{\theta'_{0k}})$ with respect to all $D'_{\mathcal{T}_k}$
end while



(b) MAML (c) SML

$T_{1\sim 6}$: training tasks

Experiment

Dataset

- STC-SeFun dataset (Bi et al., 2019)
- A large-scale Chinese short text conversation dataset with manually labeled sentence functions
- We select 9 high-resource tasks for meta-training, 4 tasks for meta-validation and 5 tasks for testing (adaptation).

	Query SF	Response SF	# Samples		
Meta Train	Positive DE	Positive DE	27058		
	Wh-style IN	Positive DE	12854		
	Positive DE	Negative DE	5831		
	Negative DE	Positive DE	4006		
	Positive DE	Wh-style IN	3935		
	A-not-A IN	Positive DE	3508		
	Wh-style IN	Negative DE	3367		
	Yes-no IN	Positive DE	3267		
	Negative DE	Negative DE	2466		
Meta Val	Wh-style IN	DE w/ IN words	271	100	500
	Negative DE	Wh-style IN	161	100	500
	Positive DE	EX w/ interjections	134	100	500
	Positive DE	DE w/ IN words	120	100	500
Meta Test	Positive DE	Yes-no IN	1314	100	500
	Yes-no IN	Negative DE	893	100	500
	Positive DE	EX w/o tone words	846	100	500
	A-not-A IN	Negative DE	684	100	500
	Wh-style IN	Wh-style IN	488	100	500

Findings of ACL: EMNLP 2020, Fine-Grained Sentence Functions for Short-Text Conversation

Experiment

Result

MTL: Multi-Task Learning

FT: Fine-tuning

SML: Structured Meta-Learning

Flue: Fluency measures the grammatical correctness of responses (1-5)

Rele: Relevance measures whether the response is a relevant reply to the query (1-5)

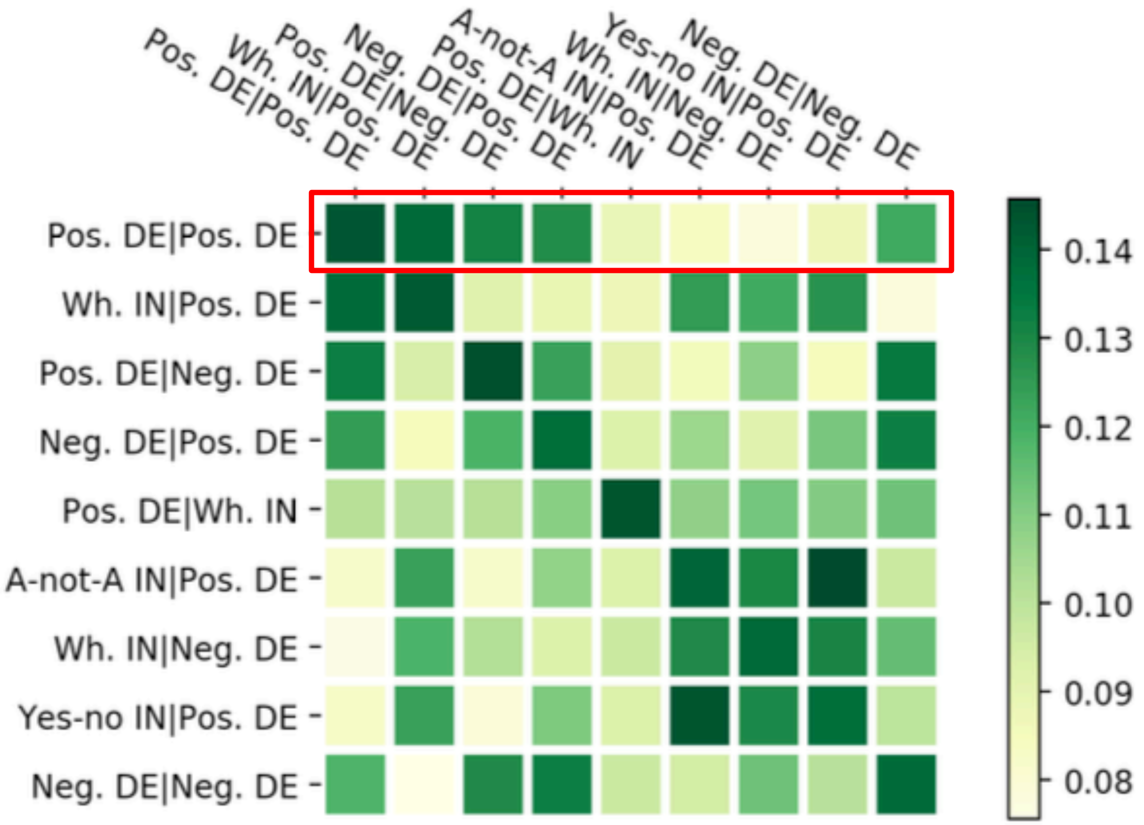
Info: Informativeness evaluates whether the response provides any meaningful information with regard to the query (1-5)

Accu: Accuracy evaluates whether the response is coherent with the given response sentence function (0-1)

Query SF	Response SF	Model	Human Evaluation			
			Flue	Rele	Info	Accu
Positive DE	Yes-no IN	MTL	59.40	50.40	36.53	0.33
		MTL+FT	62.47	53.93	39.87	34.00
		MAML	63.40	55.87	39.87	57.67
		SML	**64.27**	**56.00**	**40.13**	**69.00**
Yes-no IN	Negative DE	MTL	60.47	57.07	49.87	6.00
		MTL+FT	61.07	56.80	54.00	73.00
		MAML	62.00	**59.53**	53.67	**91.00**
		SML	**64.93**	57.80	**55.93**	**91.00**
Positive DE	EX without tone words	MTL	57.13	53.53	35.40	1.00
		MTL+FT	56.40	53.67	36.67	39.00
		MAML	65.33	56.20	39.27	**71.00**
		SML	**65.80**	**57.13**	**40.93**	68.00
A-not-A IN	Negative DE	MTL	60.33	54.93	49.73	4.33
		MTL+FT	62.13	55.13	51.47	53.67
		MAML	62.60	55.20	51.27	89.33
		SML	**63.27**	**56.00**	**52.80**	**96.00**
Wh-style IN	Wh-style IN	MTL	62.47	51.67	38.33	1.00
		MTL+FT	63.60	52.60	39.13	22.33
		MAML	64.07	53.13	43.33	85.00
		SML	**64.13**	**53.80**	**45.20**	**88.00**

Experiment

Effect of Structure Modeling



Heatmap of the self-attention weight matrix. Each row shows the attention distribution for a given query-response sentence function pair (denoted in “Query|Response” format).

Conclusion

- We apply model-agnostic meta-learning (MAML) for open domain dialogue generation on infrequent sentence functions.
- We further explore the structure across fine-grained sentence functions and such that the model can balance knowledge generalization and knowledge customization.
- Extensive experiments show that our structured meta-learning (SML) algorithm outperforms existing approaches under the low-resource setting.

Thanks!