Semi-Supervised Task-Oriented Dialog Systems and Natural Language Labeling

Zhijian Ou

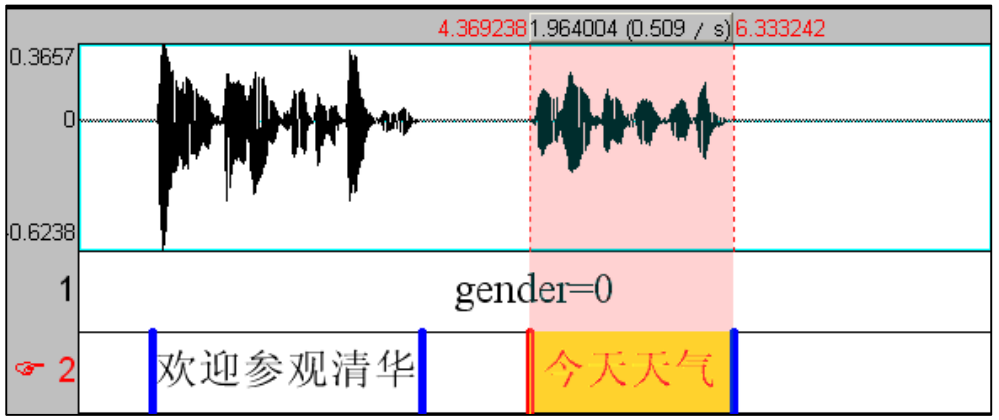
Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University

http://oa.ee.tsinghua.edu.cn/ouzhijian/

Present at UIUC NLP Seminar, September 24, 2021

Supervised learning from Labeled data $\{(x_j, y_j), j = 1, 2, \cdots\}$

Tremendous Success!

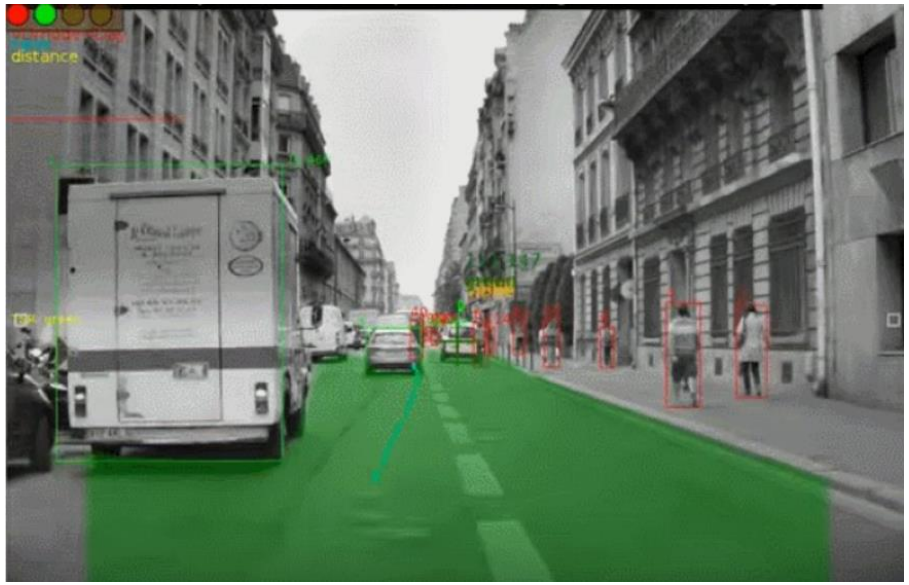


Speech Recognition

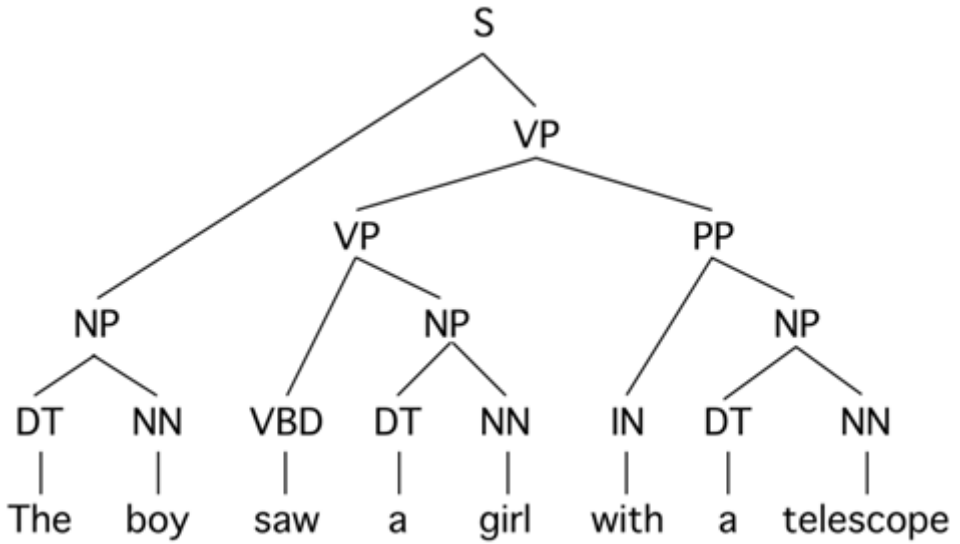
ATIS UTTERANCE EXAMPLE IOB REPRESENTATION

Sentence	*show*	*flights*	*from*	*Boston*	*To*	*New*	*York*	*today*
Slots/Concepts	O	O	O	B-dept	O	B-arr	I-arr	B-date
Named Entity	O	O	O	B-city	O	B-city	I-city	O
Intent	*Find_Flight*							
Domain	*Airline Travel*							

Intent Detection, Slot Filling, Named Entity Recognition



Object Detection and Tracking



Syntactic Parsing

Content

1. Related work

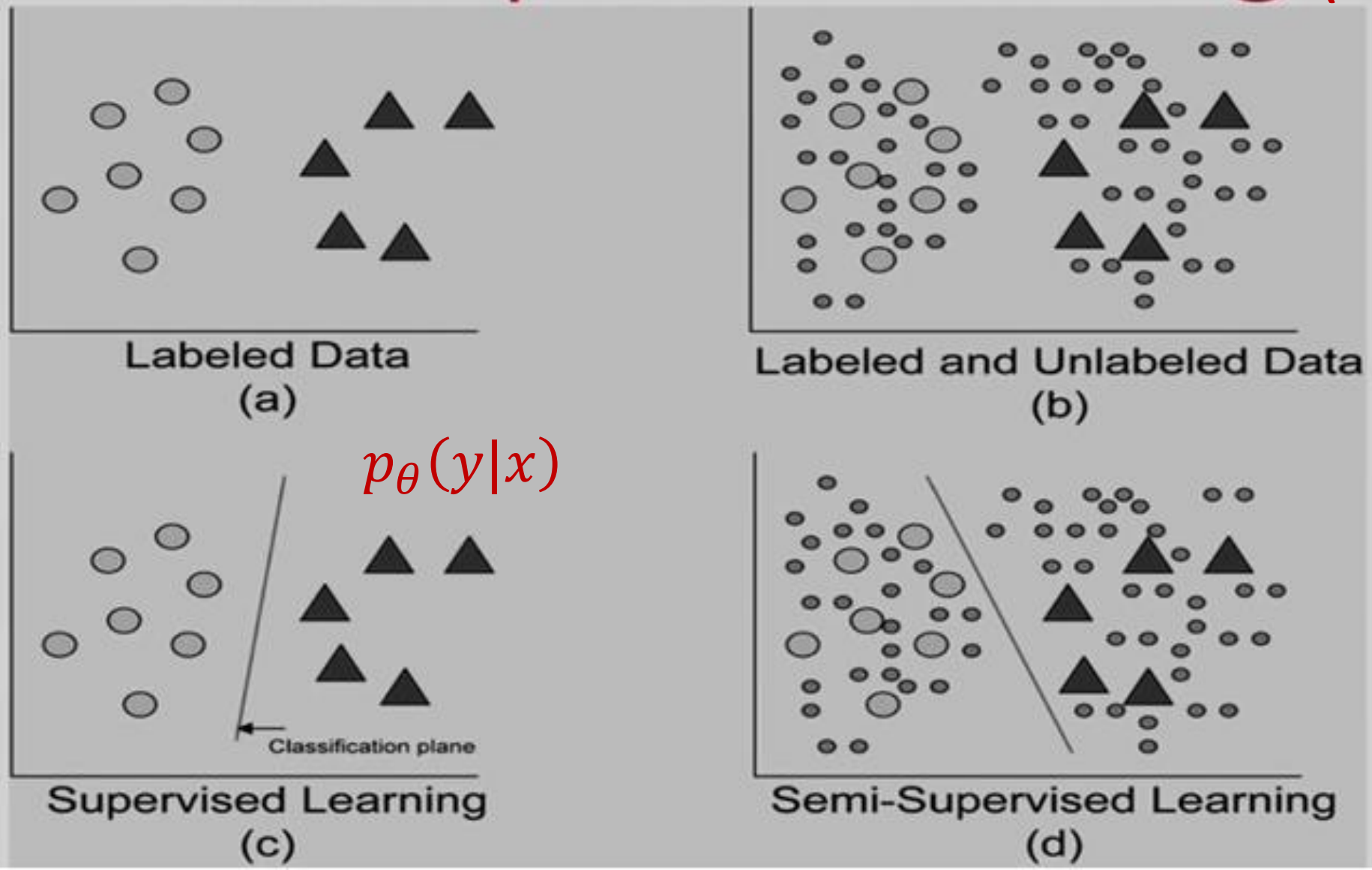
- Semi-Supervised Learning (SSL): Discriminative vs. Generative
- Generative SSL: Joint-training vs. Pre-training
- Probabilistic Graphical Models: Directed vs. Undirected (EBM)

2. Tasks and Methods

- Variational Latent-State GPT for Semi-supervised Task-Oriented Dialog Systems
- EBM based Semi-supervised Natural Language Labeling (POS, Chunking, NER)

3. Discussion and Conclusion

Semi-supervised learning (SSL)



The key to designing SSL methods is:
How to effectively exploit the information contained in the unlabeled data $\{x\}$,
which can provide
priors/regularizations/inductive biases
for finding the posterior $p_\theta(y|x)$.

Xiaojin Zhu, "Semi-supervised learning literature survey," Technical report, University of Wisconsin-Madison, 2006.

SSL methods (for using DNNs)

- Recent SSL methods with DNNs can be distinguished by the priors they adopt, and, can be divided into two classes.
  - **Generative SSL**
  - **Discriminative SSL**: The outputs from the discriminative classifier are smooth with respect to local and random perturbations of the inputs [1-5].

[1] Takeru Miyato, et al, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," TPAMI, 2018.
[2] Samuli Laine and Timo Aila, "Temporal ensembling for semisupervised learning," ICLR, 2017.
[3] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," NIPS, 2017.
[4] Kihyuk Sohn, David Berthelot, Chun-Liang Li, and et al, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," arXiv:2001.07685, 2020.
[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," arXiv:2002.05709, 2020.

Discriminative SSL

- Recent SSL methods with DNNs can be distinguished by the priors they adopt, and, can be divided into two classes.
  - **Generative SSL**
  - **Discriminative SSL**: The outputs from the discriminative classifier are smooth with respect to local and random perturbations of the inputs.

☹ heavily rely on domain-specific data augmentations, which are tuned intensively for images leading to impressive performance in some image domains
☹ less successful for other domains where these augmentations are less effective (e.g., medical images and text). For instance, random input perturbations are more difficult to apply to discrete data like text [6].

[6] Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le, "Semi-supervised sequence modeling with cross-view training," in EMNLP, 2018

Generative SSL - Basics

- Exploit unsupervised learning of generative models over unlabeled data, blend unsupervised learning and supervised learning.

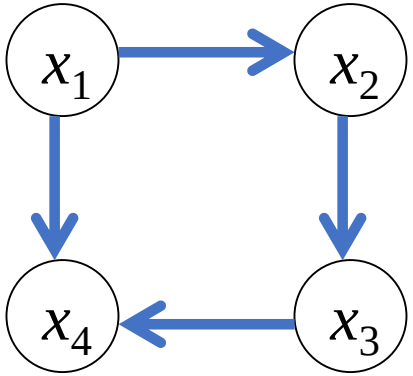
☺ inherently not require data augmentations and generally can be applied to a wider range of domains.
☺ make fewer domain-specific assumptions and tend to be domain-agnostic.

Generative SSL - Two Different Approaches

- **Joint-training**
  - A joint model of p(x,y) is defined.
  - When we have label y, we maximize p(y|x) (the supervised objective), and when the label is unobserved, we marginalize it out and maximize p(x) (the unsupervised objective).
  - Semi-supervised learning over a mix of labeled and unlabeled data is formulated as maximizing the (weighted) sum of log p(y|x) and log p(x).
- **Pre-training**
  - Only defines p(x) without y.
  - Perform unsupervised representation learning (called pre-training) on unlabeled data, followed by supervised training (called fine-tuning) on labeled data.
  - This manner of pre-training followed by fine-tuning has received increasing application in Natural Language Processing.

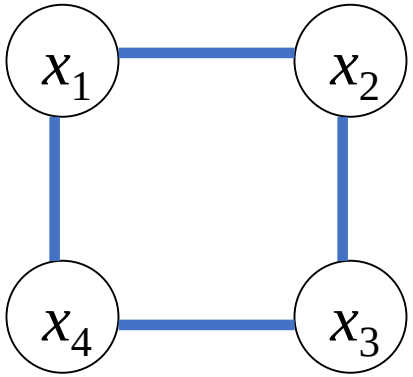
Generative SSL - Two Different Probabilistic Models

- **Directed Graphical Models / Bayesian Networks (BNs)**
  - Self-normalized
  - e.g. Hidden Markov Models (HMMs), Neural network (NN) based classifiers, Variational AutoEncoders (VAEs), Generative Adversarial Networks (GANs), auto-regressive models (e.g. RNNs/LSTMs)



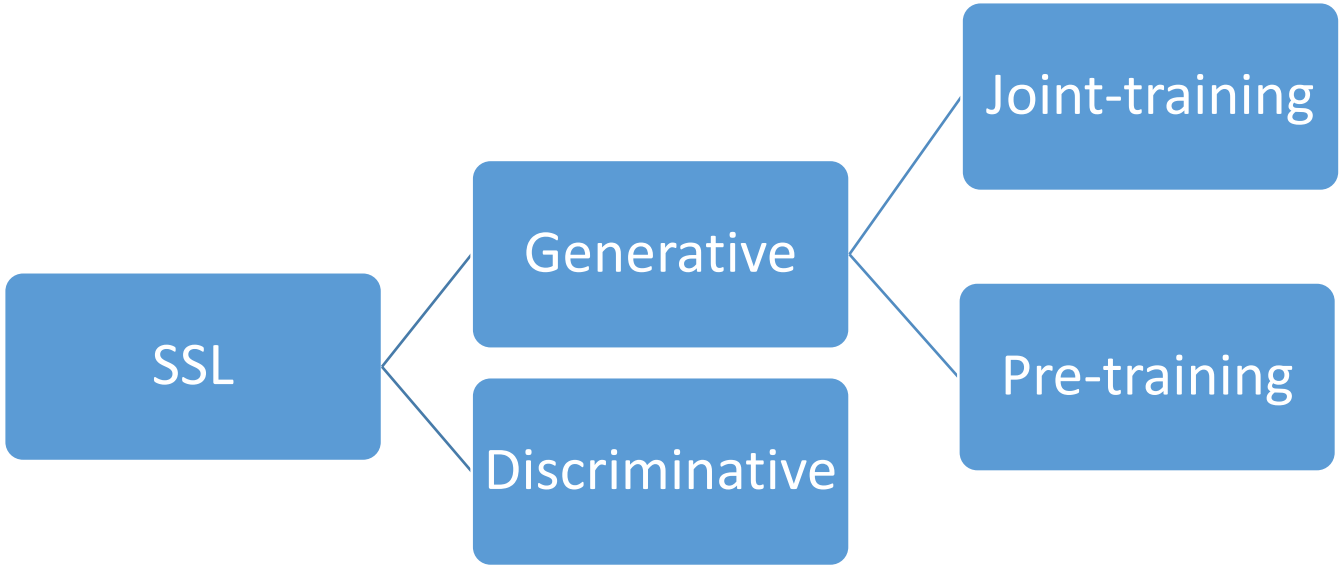
$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_1, x_3)$$

- **Undirected Graphical Models / Random Fields (RFs) / Energy-based models**
  - Involves the normalizing constant (the partition function) Z
  - e.g. Conditional Random Fields (CRFs)



$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z}\Phi(x_1, x_2)\Phi(x_2, x_3)\Phi(x_3, x_4)\Phi(x_1, x_4)$$

There are many open questions in designing semi-supervised methods for particular NLP tasks !



	Directed models	Undirected models
Joint-training	Latent Variable Model (LVM), e.g. LABES	Joint Random Field (JRF), Joint Energy-based model (JEM)
Pre-training	Auto-Regressive Language Model, e.g. GPT Masked Language Model, e.g. BERT	Random-field Language Model, e.g. TRF, Electric

[LABES] Y. Zhang, Z. Ou, et al. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. EMNLP, 2020.
[JRF] Y. Song, Z. Ou, et al. Upgrading CRFs to JRFs and its benefits to sequence modeling and labeling. ICASSP, 2020.
[JEM] S. Zhao, J.H. Jacobsen, et al. Joint energy-based models for semi-supervised classification. ICML Workshop on Uncertainty and Robustness in Deep Learning, 2020.
[TRF] B. Wang, Z. Ou. Improved training of neural trans-dimensional random field language models with dynamic noise-contrastive estimation. SLT, 2018.
[Electric] K. Clark, M.T. Luong, et al. Pre-Training Transformers as Energy-Based Cloze Models. EMNLP, 2020.

Content

1. Related work

- Semi-Supervised Learning (SSL): Discriminative vs. Generative
- Generative SSL: Joint-training vs. Pre-training
- Probabilistic Graphical Models: Directed vs. Undirected (EBM)

2. Tasks and Methods

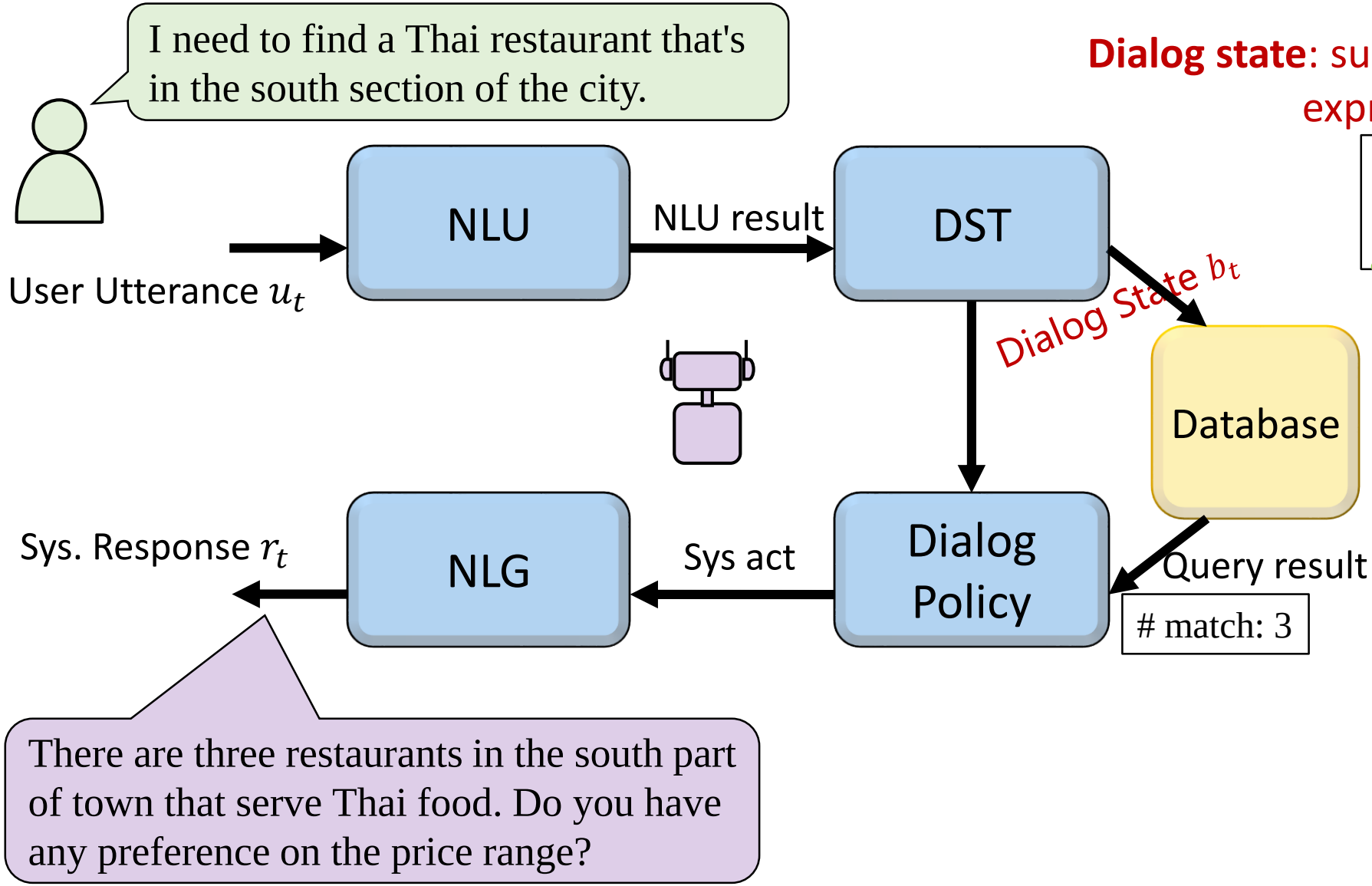
- Variational Latent-State GPT for Semi-supervised Task-Oriented Dialog Systems
- EBM based Semi-supervised Natural Language Labeling (POS, Chunking, NER)

3. Discussion and Conclusion

Hong Liu, Yucheng Cai, Zhenru Lin, Zhijian Ou, Yi Huang, Junlan Feng.
Variational Latent-State GPT for Semi-supervised Task-Oriented Dialog Systems,
arXiv:2109.04314, 2021.

1. Related work and Motivation
2. VLS-GPT
3. Experiments
4. Conclusion

Motivation: Task-Oriented Dialog (TOD)



Dialog state: summarize the dialog history, expressed as slot-value pairs

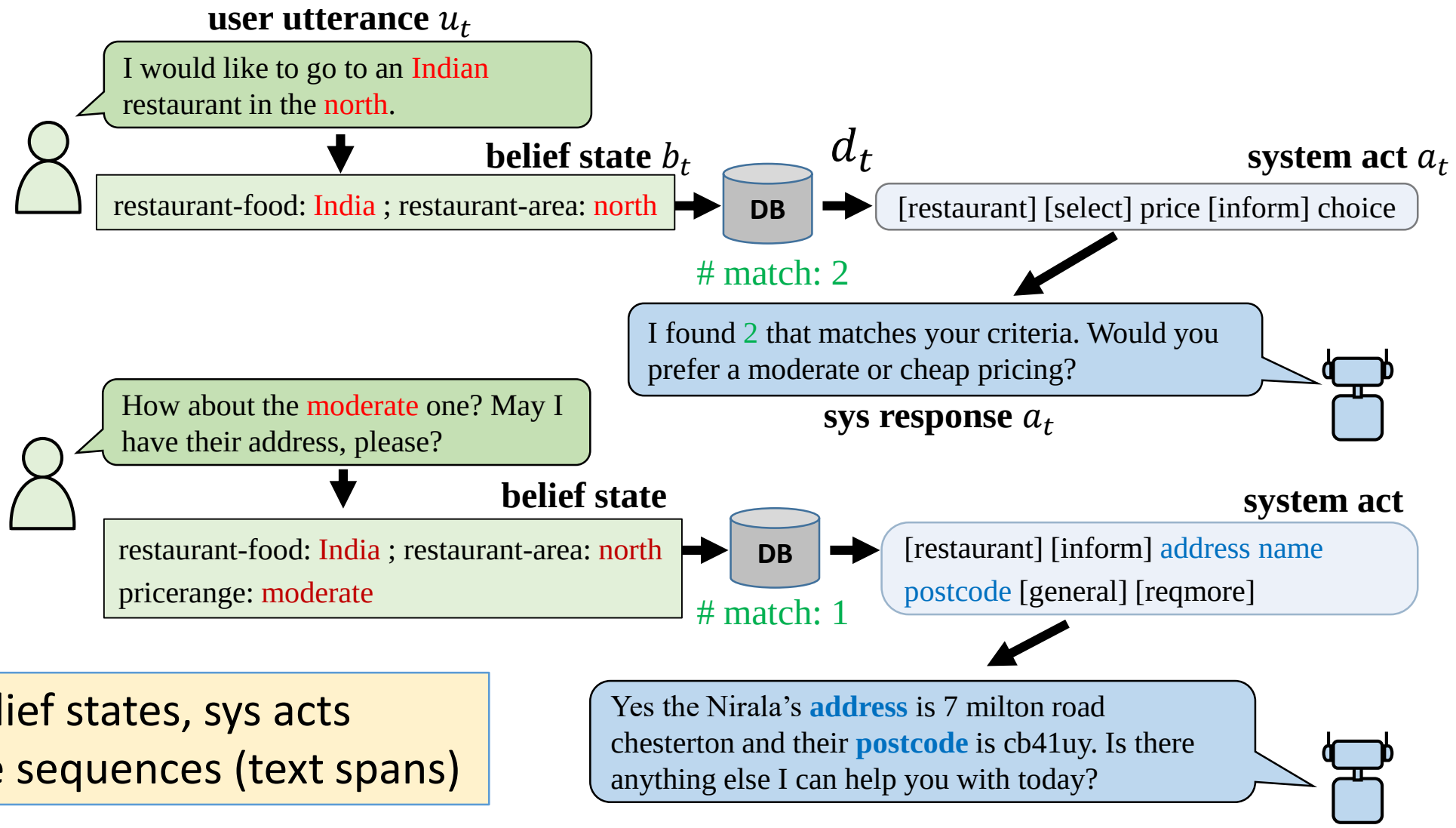
restaurant-food: Thai,
restaurant-area: south

☹ Need human annotations

name	food	area	pricerange
Bangkok City	Thai	south	Cheap
Sala Thong	Thai	south	Medium
Saigon City	Thai	south	expansive
…	…	…	…

Related work: TOD systems

For dialog turn t:
u_t - user utterance
b_t - belief state
d_t - DB search result
a_t - system act
r_t - sys response

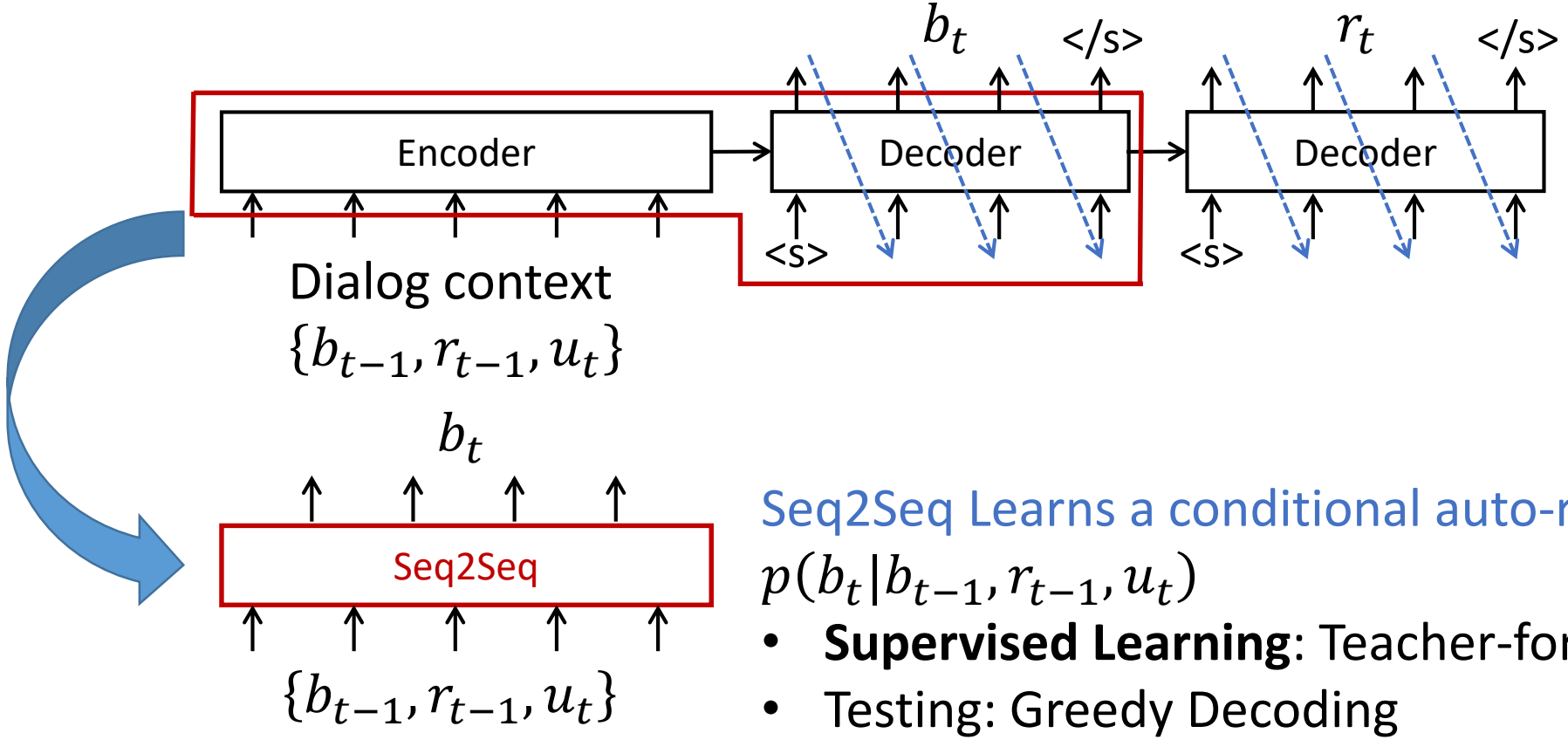


Represent belief states, sys acts
as natural language sequences (text spans)

W. Lei, X. Jin, et al. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. ACL 2018.

Related work: End-to-End TOD systems

- The methodology for building TOD systems is advancing from separate training of individual modules to the end-to-end (E2E) trainable approach
  - Employ encoder-decoder Seq2Seq architecture to connect modules and train them together



For dialog turn t:
u_t - user utterance
b_t - belief state
d_t - DB search result
a_t - system act
r_t - sys response

Seq2Seq Learns a conditional auto-regressive LM:

$p(b_t|b_{t-1}, r_{t-1}, u_t)$

- **Supervised Learning**: Teacher-forcing
- Testing: Greedy Decoding

Y. Zhang, Z. Ou, Z. Yu. Task-Oriented Dialog Systems that Consider Multiple Appropriate Responses under the Same Context. AAAI, 2020.

Related work: Two approaches for semi-supervised TOD systems

Joint-training

- Formulating a latent variable model (**LVM**) of observations and labels, blending unsupervised and supervised learning
- Unsupervised learning with LVM usually maximizes p(x) via variational learning over unlabeled in-domain data
- Previous works typically use LSTM based Seq2Seq architectures [1, 2]

Pre-training

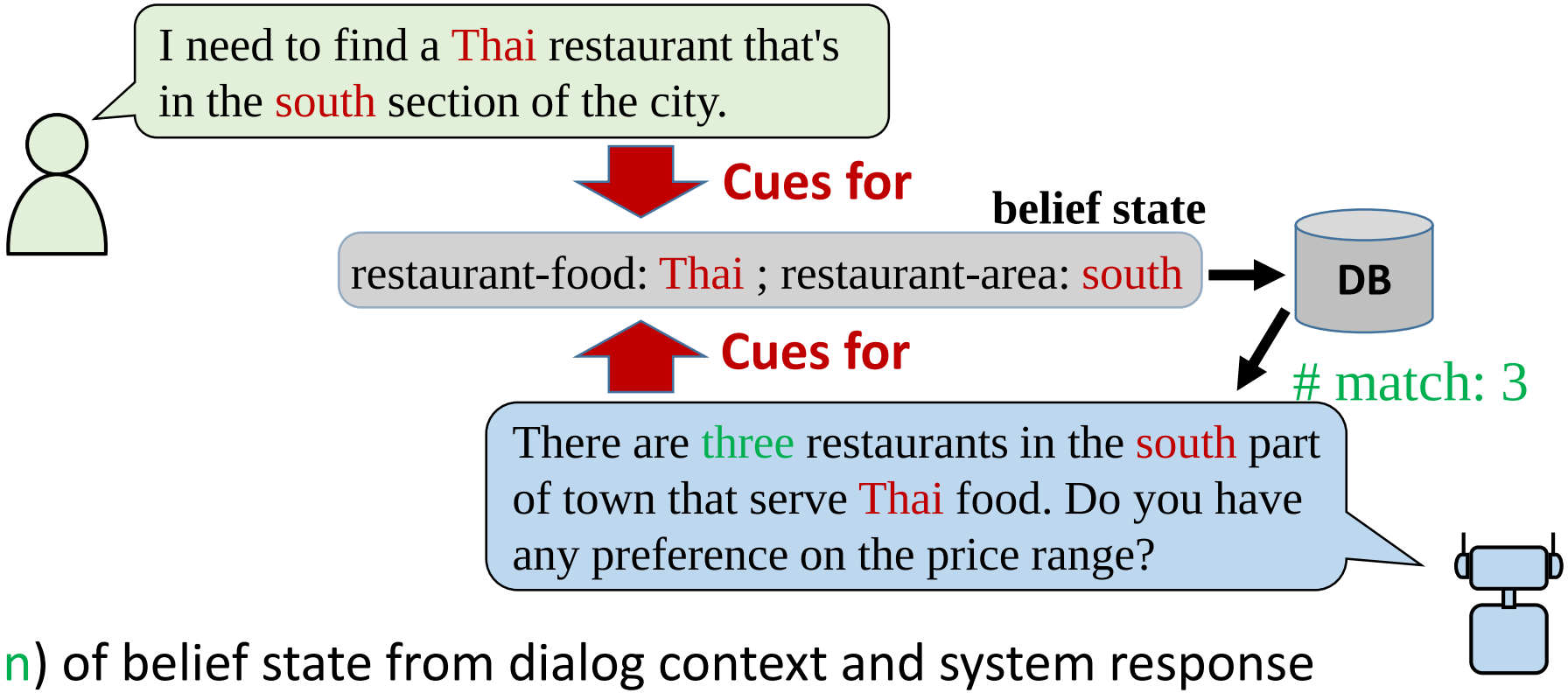
- Unsupervised pre-training followed by supervised fine-tuning
- Large-scale language models pre-trained on open-domain texts are fine-tuned with in-domain labels
- Transformer based autoregressive language models, like GPT-2, learn a strong distribution for next-token prediction, which makes them particularly useful for generative TOD systems [3,4]

1. Jin, X.; Lei, W.; et al. Explicit State Tracking with Semi-Supervision for Neural Dialogue Generation. CIKM, 2018.
2. Zhang, Y.; Ou, Z.; et al. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. EMNLP, 2020.
3. Budzianowski, P.; Vulic, I; Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. The 3rd Workshop on Neural Generation and Translation, 2019.
4. Hosseini-Asl, E.; McCann, B.; et al. A simple language model for task-oriented dialogue. arXiv:2005.00796, 2020.

The Underlying Idea for Joint-training

For building end-to-end task-oriented systems,

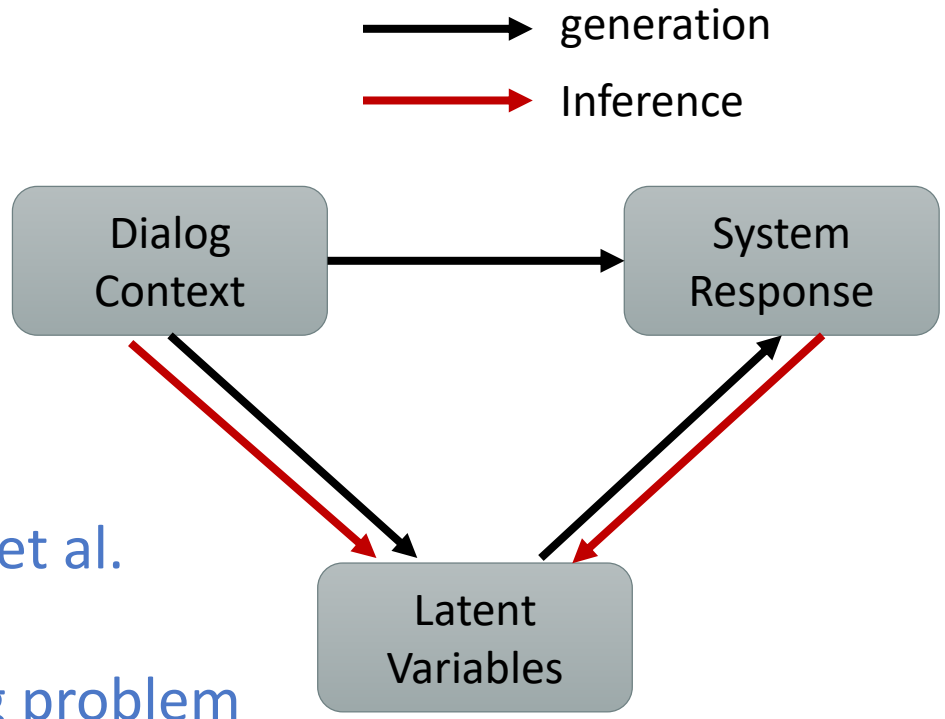
- Goal: reducing the needs of belief state labels
- Idea: inferring belief state from unlabeled dialog data



Cues (red, green) of belief state from dialog context and system response

Related Work: Latent Variable Models for Dialog Modeling

- Chit-chat
  - Diversity (Serban et al. 2017; Zhao et al. 2017)
  - Language style (Gao et al. 2019)
  - Selected knowledge (Kim et al. 2020)
- Task-oriented
  - Model dialog structure (Zhai and Williams 2014; Shi et al. 2019)
  - Latent sys act, for tackling the one-to-many mapping problem in response generation (Wen et al. 2017; Zhao et al. 2019)
  - Latent belief state, for semi-supervised learning (Jin, Lei, et al. 2018†; Zhang, Ou, et al. 2020)



† uses a combination of posterior regularization and auto-encoding

LAtent BElief State Dialog Model - LABES

For dialog turn t:
u_t - user utterance
$c_t = \{r_{t-1}, u_t\}$
b_t - belief state
d_t - DB search result
r_t - sys response

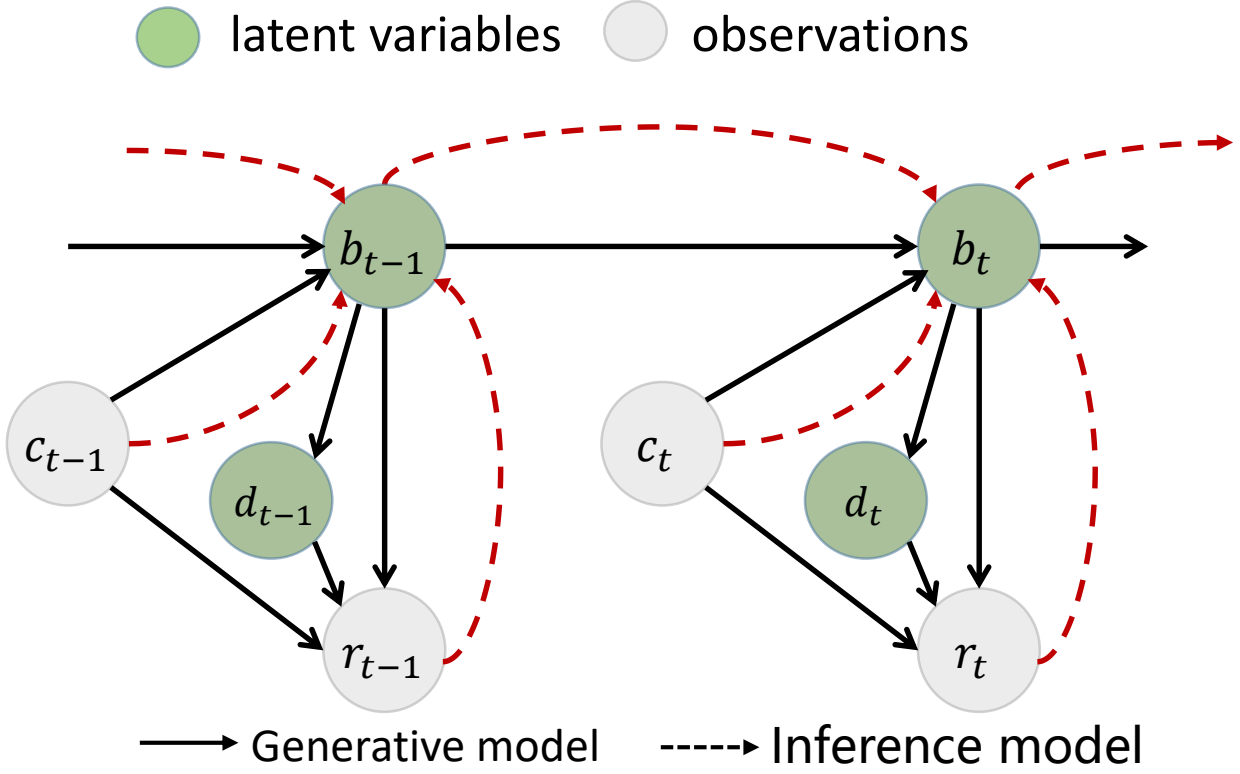
Generative model

$$p_\theta(b_{1:T}, r_{1:T}|u_{1:T}) = p_\theta(b_{1:T}|u_{1:T}) p_\theta(r_{1:T}|b_{1:T}, u_{1:T})$$
$$= \prod_{t=1}^{T} p_\theta(b_t|b_{t-1}, c_t) p_\theta(r_t|c_t, b_t, d_t)$$

Variational Learning

$$\max_{\theta,\phi} \mathcal{J}_{un} = \mathbb{E}_{q_\phi(b_{1:T})}\left[\log \frac{p_\theta(b_{1:T}, r_{1:T}|u_{1:T})}{q_\phi(b_{1:T}|u_{1:T}, r_{1:T})}\right]$$
$$= \sum_{t=1}^{T} \mathbb{E}_{q_\phi(b_{1:T})}\left[\log p_\theta(r_t|c_t, b_t, d_t)\right]$$
$$-\alpha \mathrm{KL}\left[q_\phi(b_t|b_{t-1}, c_t, r_t) \| p_\theta(b_t|b_{t-1}, c_t)\right]$$

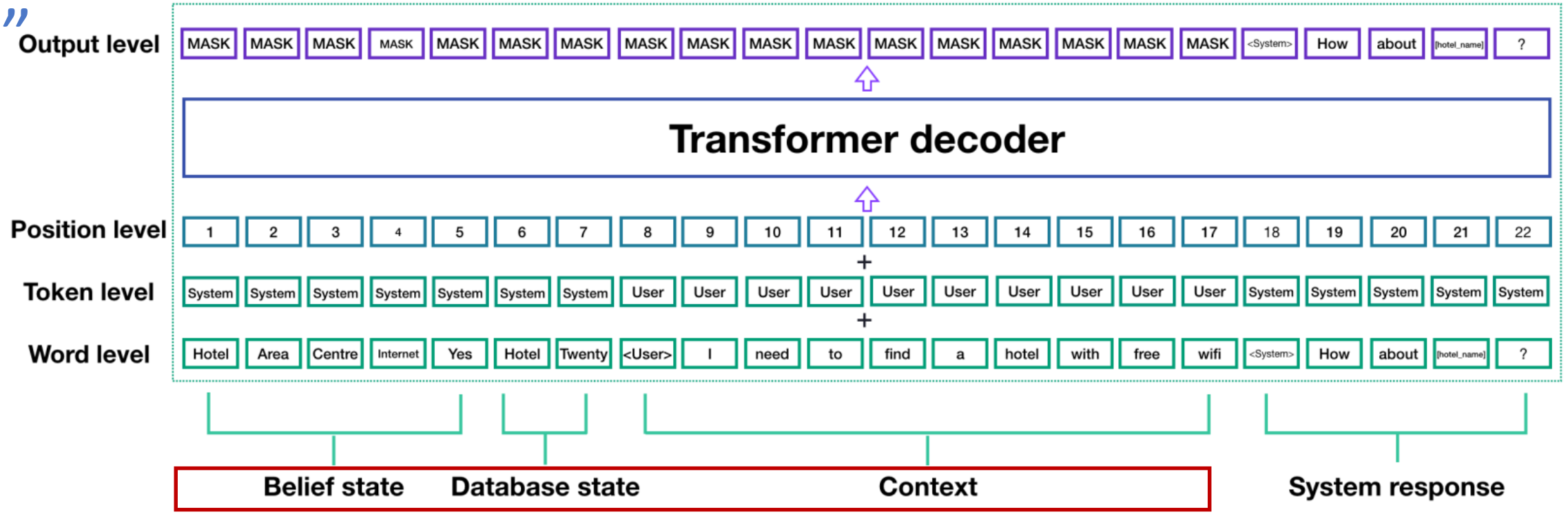
$$q_\phi(b_{1:T}) \triangleq \prod_{t=1}^{T} q_\phi(b_t|b_{t-1}, c_t, r_t)$$



Zhang, Y.; Ou, Z.; et al. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. EMNLP, 2020.

The Underlying Idea for Pre-training

"Task-oriented dialogue modeling requires substantial amounts of domain-specific manually labeled data. A natural question to ask is: Can we leverage transfer learning through generative pretraining on large unlabelled corpora to enable task-oriented dialogue modeling."

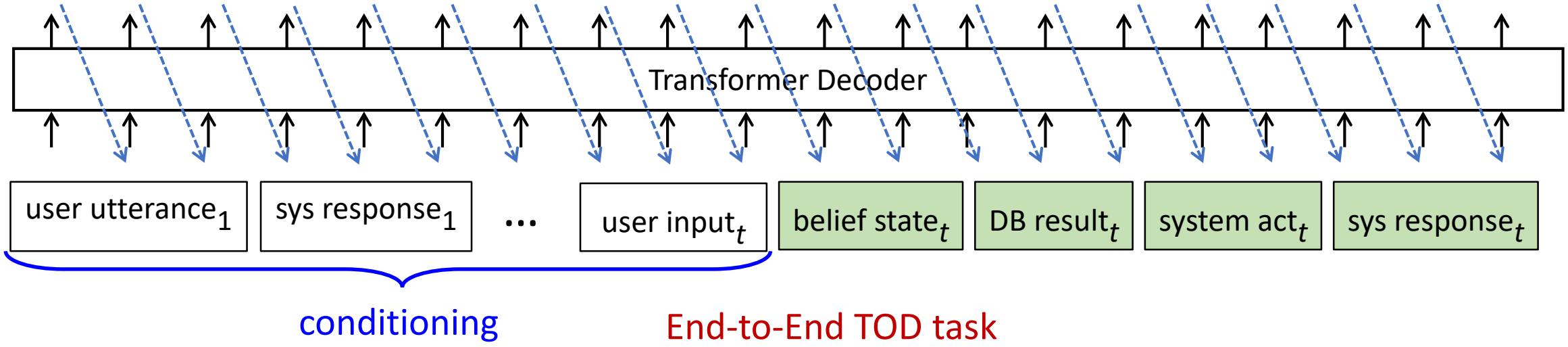


Dialogue-context-to-response task

Budzianowski, P.; Vulic, I; Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. The 3rd Workshop on Neural Generation and Translation, 2019.

The Underlying Idea for Pre-training

“ SimpleTOD is a simple approach to task-oriented dialogue that uses a single causal language model to generate all outputs given the dialogue history and database search results. ”



Hosseini-Asl, E.; McCann, B.; et al. A simple language model for task-oriented dialogue. arXiv:2005.00796, 2020.

Related Work: Semi-supervised TOD systems with pre-trained GPT-2

Comparison of existing GPT-based TOD methods by their training objectives

Method	Training Objective
B&V	$\prod_{t=1}^{T} p(r_t \vert \{u, b, d\}_t)$
Ham et al. (2020)	$\prod_{t=1}^{T} p(\{b, a, r\}_t \vert \{u, r\}_1, \cdots, \{u, r\}_{t-1}, u_t)$
SOLOIST, AuGPT	$\prod_{t=1}^{T} p(\{b, d, r\}_t \vert \{u, r\}_1, \cdots, \{u, r\}_{t-1}, u_t)$
SimpleTOD	$\prod_{t=1}^{T} p(\{u, r\}_1, \cdots, \{u, r\}_{t-1}, \{u, b, d, a, r\}_t)$
UBAR	$p(\{u, b, d, a, r\}_1, \cdots, \{u, b, d, a, r\}_T) = \prod_{t=1}^{T} p(\{u, b, d, a, r\}_t \vert \{u, b, d, a, r\}_1, \cdots, \{u, b, d, a, r\}_{t-1})$
LABES	$p(\{b, d, r\}_1, \cdots, \{b, d, r\}_T \vert u_1, \cdots, u_T) = \prod_{t=1}^{T} p(\{b, d, r\}_t \vert r_{t-1}, b_{t-1}, u_t)$
VLS-GPT	$p(\{b, d, a, r\}_1, \cdots, \{b, d, a, r\}_T \vert u_1, \cdots, u_T) = \prod_{t=1}^{T} p(\{b, d, a, r\}_t \vert \{u, b, d, a, r\}_1, \cdots, \{u, b, d, a, r\}_{t-1}, u_t)$

- UBAR proposes the session-level finetuning of GPT-2, namely on the whole sequence of the entire dialog session which is composed of user utterances, belief states, DB results, system acts and responses of all dialog turns.
- This is different from the turn-level training, employed in all previous works.

For dialog turn t:
u_t - user utterance
b_t - belief state
d_t - DB search result
a_t - system act
r_t - sys response

Yang, Y.; Li, Y.; and Quan, X. 2021. UBAR: Towards Fully End-to End Task-Oriented Dialog System with GPT-2. AAAI 2021.

Remarkably, the two approaches, pre-training-and-fine-tuning and LVM based variational training, are not exclusive to take and could be jointly used, and, can complement each other.

- The pre-training approach is powerful at leveraging unlabeled open-domain data
- The variational approach is suited to exploiting unlabeled in-domain data

How we can leverage both pre-trained GPT and variational learning is not clear, requires new design and has not ever been examined.

Hong Liu, Yucheng Cai, Zhenru Lin, Zhijian Ou, Yi Huang, Junlan Feng.
Variational Latent-State GPT for Semi-supervised Task-Oriented Dialog Systems,
arXiv:2109.04314, 2021.

1.Related work and Motivation

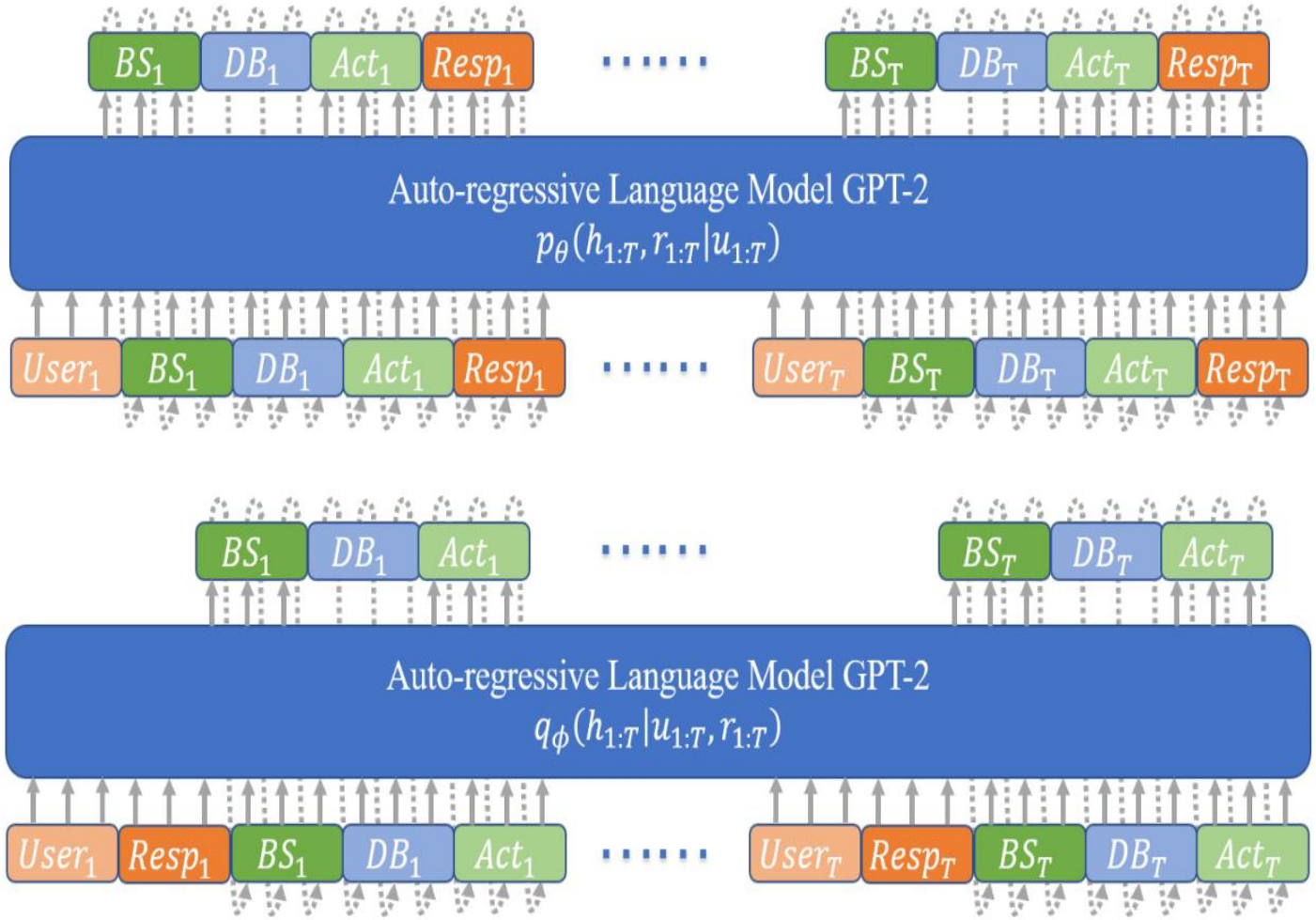
2.VLS-GPT

3.Experiments

4.Conclusion

Variational Latent-State GPT (VLS-GPT)

We unify the dialog flow (belief state tracking, action and response generation) into a single sequence prediction problem, which can be accomplished by an auto-regressive LM.



Generative model

Latent State $h_{1:T} = \{b, d, a\}_{1:T}$

$$p_\theta(h_{1:T}, r_{1:T}|u_{1:T})$$
$$= \Pi_{t=1}^{T} p_\theta(h_t|\{u,h,r\}_1, \cdots, \{u,h,r\}_{t-1}, u_t) \times p_\theta(r_t|\{u,h,r\}_1, \cdots, \{u,h,r\}_{t-1}, \{u,h\}_t)$$
$$\triangleq \Pi_{t=1}^{T} p_\theta(h_t|\prec) p_\theta(r_t|\prec)$$

Latent State Prior

Response Probability

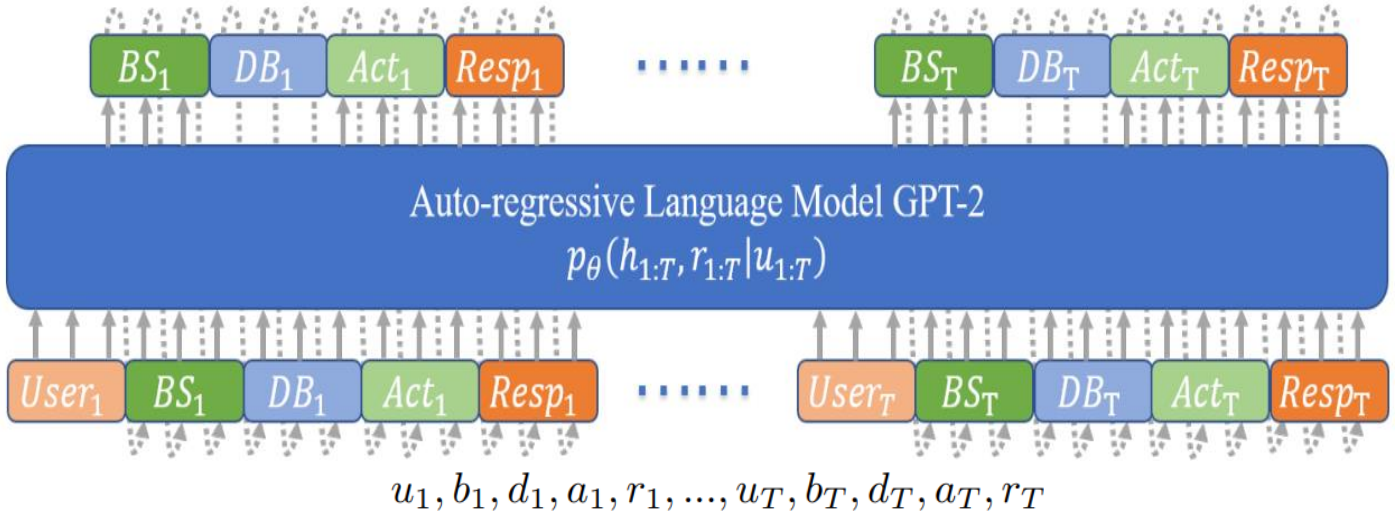
Inference model

$$q_\phi(h_{1:T}|u_{1:T}, r_{1:T})$$
$$= \Pi_{t=1}^{T} q_\phi(h_t|\{u,r,h\}_1, \cdots, \{u,r,h\}_{t-1}, \{u,r\}_t)$$
$$\triangleq \Pi_{t=1}^{T} q_\phi(h_t|\prec)$$

Latent State Approx. Posterior

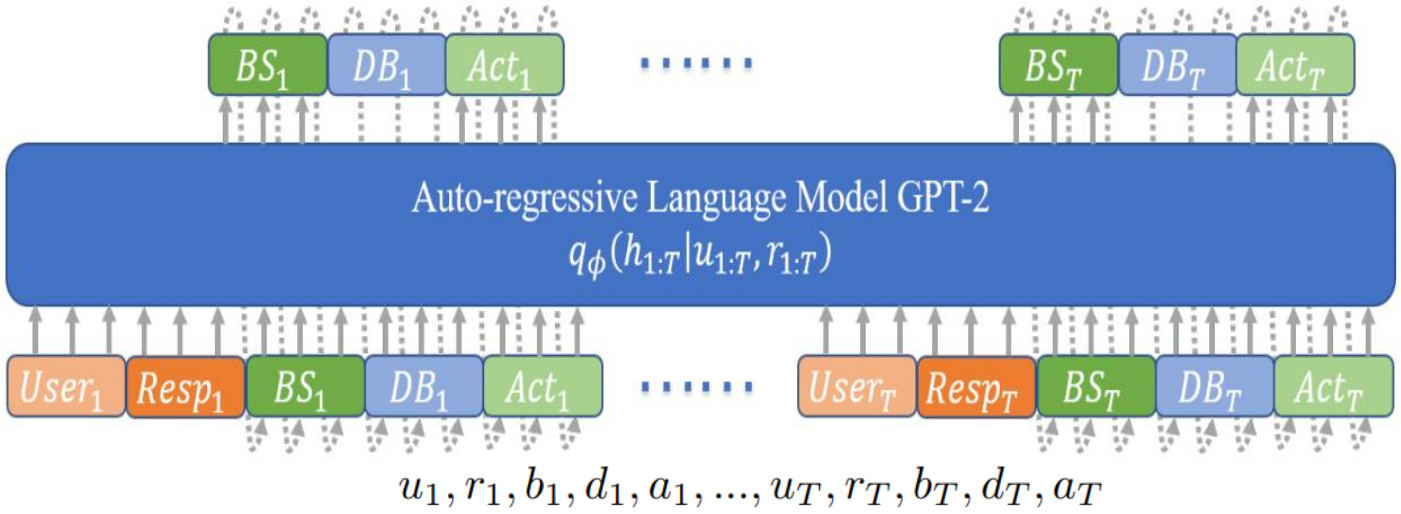
Variational Latent-State GPT (VLS-GPT)

VLS-GPT consists of two auto-regressive models - generative model and inference model, both initialized from GPT-2 but trained with different training sequences.



...<sos_u> thanks for your help. in addition, i am looking for a thursday train departing cambridge, please. <eos_u> <sos_b> [train] day thursday departure cambridge [restaurant] pricerange cheap area centre food indian people 8 day monday time 16:30 <eos_b> <sos_db> [db_3] <eos_db> <sos_a> [train] [inform] day choice departure [request] destination <eos_a> <sos_r> there are [value_choice] trains leaving [value_departure] on [value_day]. where is your destination? <eos_r> <sos_u> i am looking to travel to ely departing after 13:15 if possible. <eos_u>...

(a) Training sequence for the generative model



...<sos_u> thanks for your help. in addition, i am looking for a thursday train departing cambridge, please. <eos_u> <sos_r> there are [value_choice] trains leaving [value_departure] on [value_day]. where is your destination? <eos_r> <sos_b> [train] day thursday departure cambridge [restaurant] pricerange cheap area centre food indian people 8 day monday time 16:30 <eos_b> <sos_db> [db_3] <eos_db> <sos_a> [train] [inform] day choice departure [request] destination <eos_a> <sos_u> i am looking to travel to ely departing after 13:15 if possible. <eos_u>...

(b) Training sequence for the inference model

Unsupervised Learning for VLS-GPT

- Maximizing the variational evidence lower bound (ELBO):

Latent State
$h_{1:T} = \{b, d, a\}_{1:T}$

$$\mathcal{J}_{\text{VL}} = \mathbb{E}_{q_\phi(h_{1:T}|u_{1:T},r_{1:T})}\left[\log \frac{p_\theta(h_{1:T}, r_{1:T}|u_{1:T})}{q_\phi(h_{1:T}|u_{1:T}, r_{1:T})}\right]$$

  - Optimize such ELBO objective for sequential discrete latent variable models is challenging!
  - Previous studies used Gumbel-Softmax or Straight-Through to back-propagate gradients through discrete variables.

- Straight-Through Trick (STT): for token i in h_t

$$STT(h_t^{(i)}) = onehot(h_t^{(i)}) + softmax(h_t^{(i)}) - \text{softmax}(h_t^{(i)}).detach()$$

☺ Applying the above $STT(h_t^{(i)})$ in the forward direction successfully accomplish gradient propagation through $h_{1:T}$ in the backward direction

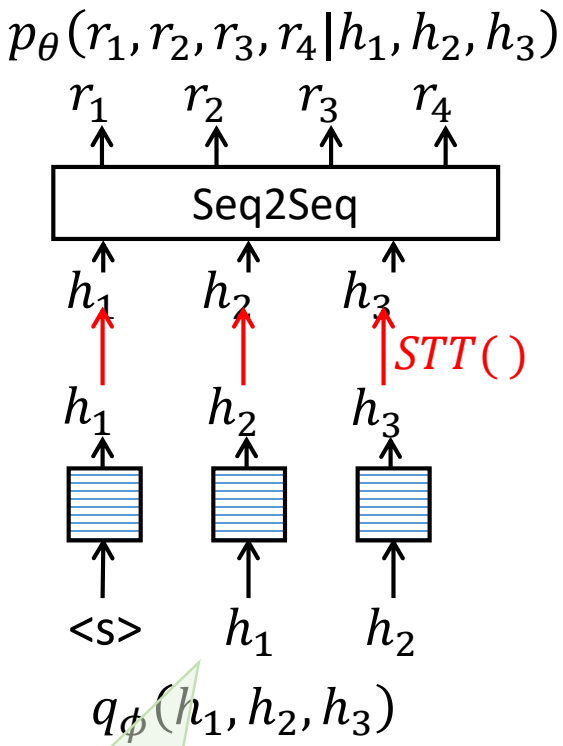
  - However, optimizing the above ELBO, which basically is an expectation under the GPT-based inference model, presents new challenge!

Sampling-then-forward-computation strategy

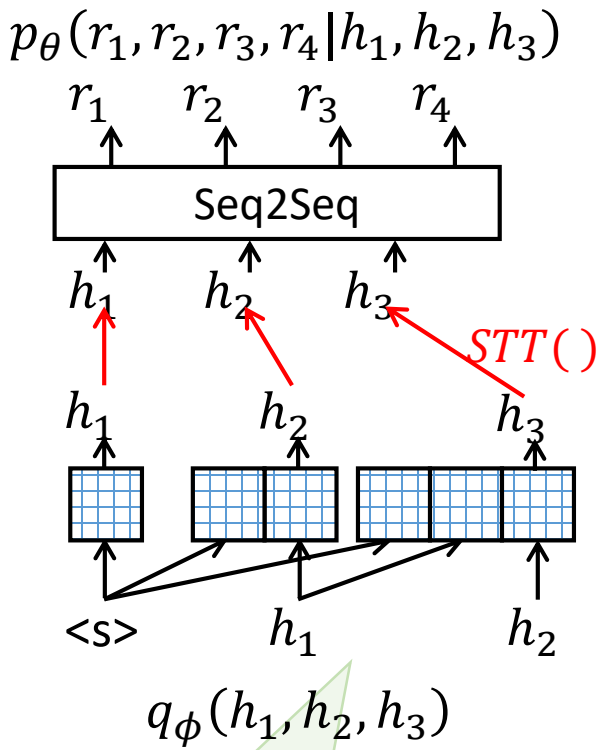
$$\max_{\theta,\phi} E_{q_\phi(h_1,h_2,h_3)}\left[\log p_\theta(r_1,r_2,r_3|h_1,h_2,h_3)\right]$$

LSTM

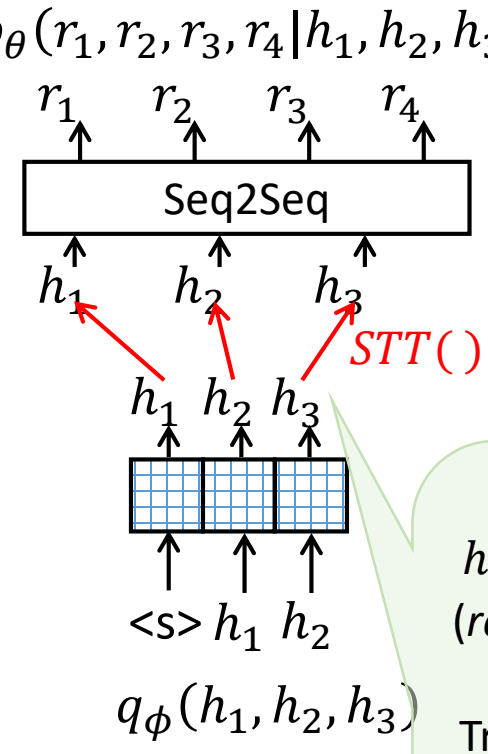
Transformer



(a)



(b)



(c)

The inference models in previous work are **first-order Markov models**. Memory Complexity: $O(T)$

The session-level GPT-based inference model in VLS-GPT is **non-Markovian**. Memory Complexity: $O(T(T+1)/2)$

First sampling $h_{1:3} \sim q_\phi(h_1,h_2,h_3)$ (*requires_grad=false*)

Treat $h_{1:3}$ as known, forward compute $q_\phi(h_1,h_2,h_3)$, feed $h_{1:3}$ forward further (*requires_grad=true*)

VLS-GPT: Algorithm Formulation

- Maximizing the variational evidence lower bound (ELBO):

Latent State
$h_{1:T} = \{b, d, a\}_{1:T}$

$$\mathcal{J}_{\mathrm{VL}} = \mathbb{E}_{q_\phi(h_{1:T}|u_{1:T},r_{1:T})} \sum_{t=1}^{T} \left[\log p_\theta(r_t|\prec) + \log \frac{p_\theta(h_t|\prec)}{q_\phi(h_t|\prec)} \right]$$

- An iteration consists of three steps:
  - Latent state generation
    Sampling $h_{1:T} \sim q_\phi(h_{1:T}|u_{1:T}, r_{1:T})$
  - Forward computation
    Treat $h_{1:T}$ as known, run the forward pass to compute the objective function :

$$\mathcal{J}_{\mathrm{VL}} \approx \sum_{t=1}^{T} \log p_\theta(r_t|\prec) - \sum_{t=1}^{T} KL\left[q_\phi(h_t|\prec)||p_\theta(h_t|\prec)\right]$$

  - Backward computation

$$KL\left[q_\phi(h_t^{(i)}|\prec)||p_\theta(h_t^{(i)}|\prec)\right] = \sum_{i=1}^{|h_t|} \sum_{h_t^{(i)}} q_\phi(h_t^{(i)}|\prec) \log \frac{q_\phi(h_t^{(i)}|\prec)}{p_\theta(h_t^{(i)}|\prec)}$$

Hong Liu, Yucheng Cai, Zhenru Lin, Zhijian Ou, Yi Huang, Junlan Feng.
Variational Latent-State GPT for Semi-supervised Task-Oriented Dialog Systems,
arXiv:2109.04314, 2021.

1.Related work and Motivation

2.VLS-GPT

3.Experiments

4.Conclusion

Datasets and Metrics

- MultiWOZ2.1
  - A English multi-domain TOD dataset from human-human WOZ conversations
  - 8438 multi-turn dialogues with 13.68 average turns, spanning over seven domains (restaurant, train, attraction, hotel, taxi, hospital, police)
  - Metrics: Inform, Success, BLEU

    Combined score=0.5*(Inform + Success) + BLEU
- CrossWOZ
  - A Chinese Cross-Domain WOZ TOD dataset
  - 6K dialogue sessions and 102K utterances for 5 domains (hotel, restaurant, attraction, metro, and taxi)
  - Metrics (original): Match, Request Success, BLEU

    Combined score=0.5*(Match + Request Success) + BLEU

Fully-supervised experiments

- End-to-end evaluation on MultiWOZ2.1

Model name	Pretrained LM	Inform	Success	BLEU	Combined
DAMD	-	76.4	60.4	16.6	85.0
LABES-S2S	-	76.89	63.3	17.92	88.01
SimpleTOD	DistilGPT-2	85.00	70.05	15.23	92.98
UBAR*	DistilGPT-2	90.19	79.38	17.64	102.43
AuGPT	GPT-2	91.4	72.9	17.2	99.35
VLS-GPT	DistilGPT-2	90.29	81.58	17.27	103.21

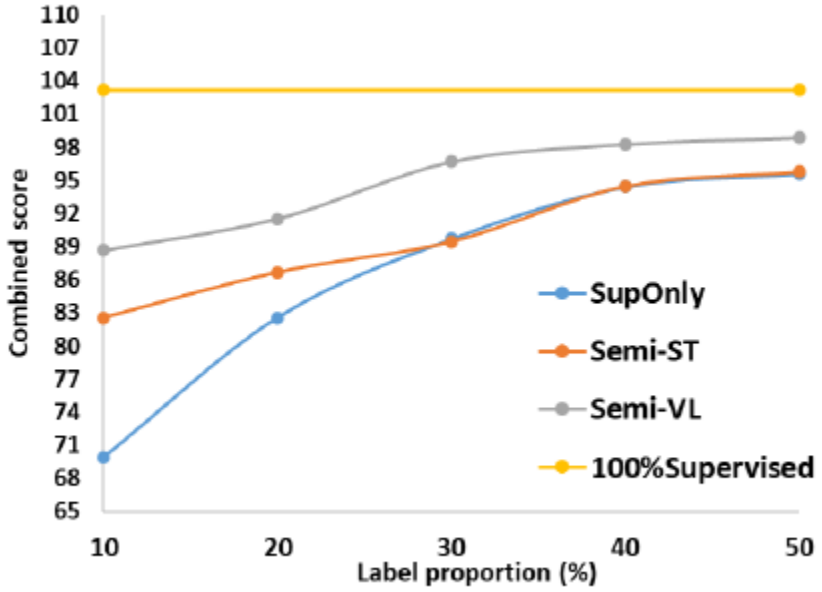
Table 2: End-to-end evaluation results on fully-supervised MultiWOZ2.1. * denotes results obtained by our run of the open-source code.

VLS-GPT outperforms all other models in end-to-end evaluation, reaching a new state-of-the-art on MultiWOZ2.1

Semi-supervised experiments

Model Configuration		MultiWOZ2.1				CrossWOZ			
Label Proportion	Method	Inform	Success	BLEU	Combined	Match	Req-Suc	BLEU	Combined
100%	VLS-GPT	90.29	81.58	17.27	103.21	63.93	77.33	37.31	107.94
50%	VLS-GPT+SupOnly	85.09	73.07	16.52	95.60	62.83	76.53	34.37	104.05
	VLS-GPT+Semi-ST	85.69	73.77	16.06	95.79	61.99	75.23	38.10	106.71
	VLS-GPT+Semi-VL	87.59	77.78	16.19	98.87	63.19	78.05	37.25	107.87
40%	VLS-GPT+SupOnly	84.48	72.37	16.00	94.43	60.78	74.03	33.44	100.84
	VLS-GPT+Semi-ST	83.78	71.87	16.66	94.49	60.38	78.42	38.16	107.56
	VLS-GPT+Semi-VL	89.09	75.98	15.75	98.28	63.05	76.71	38.09	107.97
30%	VLS-GPT+SupOnly	79.78	67.67	15.99	89.71	58.55	72.90	33.48	99.21
	VLS-GPT+Semi-ST	78.48	68.07	16.19	89.46	59.03	72.24	37.65	103.29
	VLS-GPT+Semi-VL	85.29	74.87	16.62	96.70	62.04	73.68	37.14	105.00
20%	VLS-GPT+SupOnly	73.47	60.46	15.62	82.59	58.21	64.53	31.32	92.69
	VLS-GPT+Semi-ST	75.08	64.56	16.86	86.68	58.04	71.32	38.99	103.67
	VLS-GPT+Semi-VL	80.38	69.17	16.77	91.54	60.68	76.31	37.43	105.92
10%	VLS-GPT+SupOnly	62.66	49.55	13.80	69.91	54.92	60.60	29.12	86.88
	VLS-GPT+Semi-ST	73.27	60.26	15.78	82.55	54.67	77.23	37.64	103.59
	VLS-GPT+Semi-VL	77.78	68.47	15.55	88.67	58.67	79.47	36.99	106.06

Table 3: Semi-supervised results on MultiWOZ2.1 and CrossWOZ.



- Two semi-supervised methods (Semi-ST and Semi-VL) generally outperform SupOnly
- Semi-VL generally performs better than Semi-ST (Self-Training) significantly
- Semi-VL VLS-GPT with 10% labeled > fully-supervised LABES which uses variational learning alone
- Semi-VL VLS-GPT with 50% labeled ≈ fully-supervised VLS-GPT

Conclusion

- We propose VLS-GPT, which is the first to combine the strengths of large pre-trained language model and variational learning for semi-supervised TOD systems.
- The inference model in VLS-GPT is non-Markovian due to the use of the Transformer architecture. We propose a sampling-then-forward-computation strategy, which enables successful variational training of VLS-GPT.
  - This strategy is useful in general for variational training involving large Transformer based models.
- Experiments on MultiWOZ2.1 (English) and CrossWOZ (Chinese) show that: VLS-GPT outperform the supervised-only baseline, a strong semi-supervised GPT-based self-training baseline, and a variational learning only baseline, across languages.
  - Given the capability of VLS-GPT for unsupervised learning, it is interesting in the future to extend VLS-GPT for leveraging unlabeled open-domain data together with in-domain data.

Yunfu Song, Huahuan Zheng, Zhijian Ou.
An empirical comparison of joint-training and pre-training for domain-agnostic semi-supervised learning via energy-based models,
IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2021.

1. Related work and Motivation
2. Methods and Tasks
   - Pre-training vs. Joint-training
   - Across domains: image classification and natural language labeling
3. Experiments
4. Conclusion

EBM models can be very flexibly defined for SSL, by either of joint-training and pre-training.

… previously known in the literature†, but it is unclear which is better when evaluated in a common experimental setup.

To the best of our knowledge, this paper is the first to systematically compare joint-training and pre-training for EBM-based for SSL, across domains (image classification and natural language labeling).

† EBM based SSL results have been reported across different data modalities (images, natural languages, an protein structure prediction and year prediction from the UCI dataset repository) [12,13,14].

Yunfu Song, Huahuan Zheng, Zhijian Ou.
An empirical comparison of joint-training and pre-training for domain-agnostic semi-supervised learning via energy-based models,
IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2021.

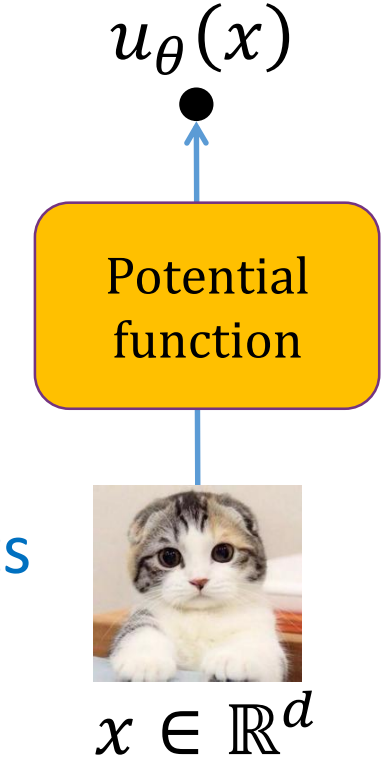
1. Related work and Motivation
2. Methods and Tasks
   - Pre-training vs. Joint-training
   - Across domains: image classification and natural language labeling
3. Experiments
4. Conclusion

Neural Random Fields (NRFs) - Basics

- NRFs are defined by using NNs to implement $u_\theta(x)$: $\mathbb{R}^d \rightarrow \mathbb{R}$

$$p_\theta(x) = \frac{1}{Z(\theta)} exp[u_\theta(x)]$$

  - $u_\theta(x)$ can be very flexibly defined; allows a close connection between p(y|x) and p(x,y).



- This type of RFs has been studied several times in different contexts
  - Deep energy models (DEMs)
    - Ngiam et al., 2012
    - Kim & Bengio, 2016 - includes linear and squared terms in $u_\theta(x)$
  - Descriptive models / Generative ConvNet
    - Xie et al., 2016 / Dai et al., 2014 - defines in the form of exponential tilting of a reference distribution (Gaussian white noise)
  - Neural random field language models
    - Wang & Ou, 2017 - defines over sequences

Learning NRFs - Basics

$$p_\theta(x) = \frac{1}{Z(\theta)} e^{u_\theta(x)}$$

- Maximum-likelihood training

$$\min_\theta KL[\tilde{p}(\tilde{x})||p_\theta(\tilde{x})]$$

$$\nabla_\theta = E_{\tilde{p}(\tilde{x})}[\nabla_\theta log p_\theta(\tilde{x})] = E_{\tilde{p}(\tilde{x})}[\nabla_\theta u_\theta(\tilde{x})] - E_{p_\theta(x)}[\nabla_\theta u_\theta(x)]$$

Expectation under empirical distribution $\tilde{p}(\tilde{x})$

Expectation under model distribution $p_\theta(x)$

- Stochastic maximum likelihood (SML) (Younes, 1989)
  - Approximate the model expectations by Monte Carlo sampling for calculating the gradient.
  - Examples: contrastive divergence (CD) 2002, persistent contrastive divergence (PCD) 2008

L. Younes, “Parametric inference for imperfectly observed gibbsian fields,” Probability Theory and Related Fields, 1989.

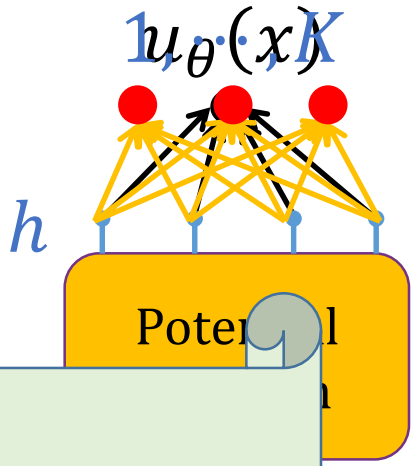
Table 1. Applications of EBMs across different domains: comparison and connection (See text for details).

	Image classification	Natural language labeling
Observation	$x \in \mathbb{R}^D$ continuous, fixed-dimensional	$x \in \bigcup_l \mathbb{V}^l$ discrete, sequence
Label	$y \in \{1, 2, \cdots, K\}$	$y \in \bigcup_l \{1, 2, \cdots, K\}^l$
Pre-training	① $u_\theta(x) = w^T h$	② $u_\theta(x)$ in Eq.(3)
Joint-training	③ $u_\theta(x, y) = \Psi_\theta(x)[y]$	④ $u_\theta(x, y)$ in Eq.(6)

① Pre-training of an EBM for semi-supervised image classification

1) Pre-training: estimate $p_\theta(x)$ over unlabeled images

$$p_\theta(x) = \frac{1}{Z(\theta)} exp[u_\theta(x)]$$



Use a feedforward NN to implement $u_\theta(x): \mathbb{R}^d \to \mathbb{R}$

wh

It can be seen that pre-training aims to learn representations that may be useful for multiple downstream tasks, and any information about the labels is not utilized until the fine-tuning stage.

2)

followed by $\text{softmax}(Wh)$, to predict $y \in \{1, \cdots, K\}$, where $W \in \mathbb{R}^{K \times H}$

[12] Yunfu Song, Zhijian Ou. Learning Neural Random Fields with Inclusive Auxiliary Generators. arXiv:1806.00271, 2018.

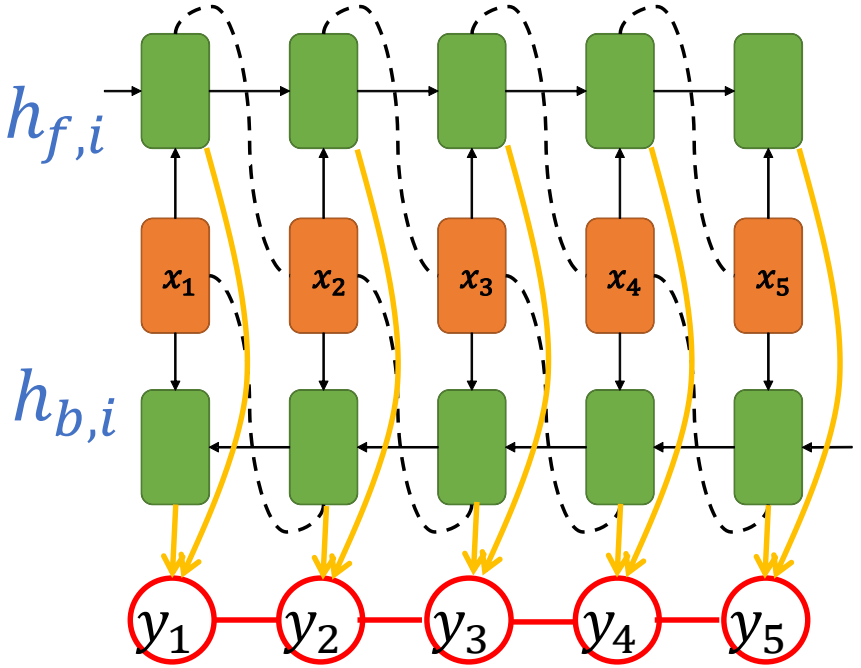
② Pre-training of an EBM for semi-supervised natural language labeling

1) Pre-training: estimate $p_\theta(x)$ over unlabeled sentences $x = (x_1, \cdots, x_l)$

$$p_\theta(x) = \frac{1}{Z(\theta)} exp[u_\theta(x)]$$

Use a B-LSTM to implement $u_\theta(x)$: $\mathbb{V}^l \to \mathbb{R}$

$$u_\theta(x) = \sum_{i=1}^{l-1} h_{f,i}^T e_{i+1} + \sum_{i=2}^{l} h_{b,i}^T e_{i-1}$$



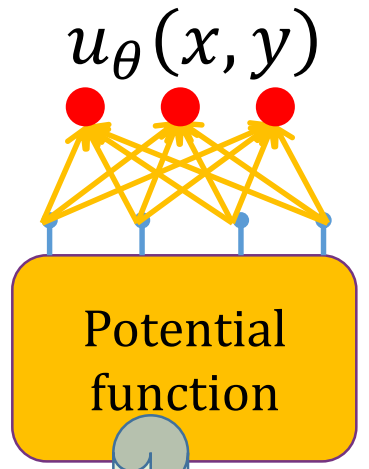
2) Fine-tuning: we add a CRF on top of the extracted representations $\{(h_{f,i}, h_{b,i}), i = 1, \cdots, l\}$ to predict label sequence $y = (y_1, \cdots, y_l)$.

[20] Bin Wang, Zhijian Ou. Improved training of neural trans-dimensional random field language models with dynamic noise-contrastive estimation. IEEE Workshop on Spoken Language Technology (SLT), 2018.

③ Joint-training of an EBM for semi-supervised image classification

- Joint modeling of observation $x \in \mathbb{R}^d$ and class label $y \in \{1, \cdots, K\}$:

$$p_\theta(x, y) = \frac{1}{Z(\theta)} exp[u_\theta(x, y)]$$



- Consider a NN $\Psi_\theta(x)$: $\mathbb{R}^d \to \mathbb{R}^K$ and define:

Different from pre-training, the unsupervised objective $p_\theta(x)$ in joint-training depends on the targeted task.

$$\begin{cases} \min_\theta KL[\tilde{p}(\tilde{x})||p_\theta(\tilde{x})] - \alpha \sum_{(\tilde{x},\tilde{y})\sim\mathcal{L}} log p_\theta(\tilde{y}|\tilde{x}) \\ \min_\phi KL[p_\theta(x)||q_\phi(x)] \end{cases}$$

[12] Yunfu Song, Zhijian Ou. Learning Neural Random Fields with Inclusive Auxiliary Generators. arXiv:1806.00271, 2018.

④ Joint-training of an EBM for semi-supervised natural language labeling

- JRF: Define a joint distribution over $x = (x_1, \cdots, x_l)$ and $y = (y_1, \cdots, y_l)$

$$p_\theta(l, x^l, y^l) = \pi_l p_\theta(x^l, y^l; l) = \frac{\pi_l}{Z_\theta(l)} \exp\left(u_\theta(x^l, y^l)\right)$$

- Consider a NN $\Psi_\theta(x)$: $\mathbb{V}^l \to \mathbb{R}^{l \times K}$ and define:

$$u_\theta(x, y) = \sum_{i=1}^{l} \Psi_\theta(x)[i, y_i] + \sum_{i=1}^{l} A[y_{i-1}, y_i]$$

- From JRF we have:

$$p_\theta(y^l | x^l) = \frac{1}{\sum_{y^l} \exp\left(u_\theta(x^l, y^l)\right)} \exp\left(u_\theta(x^l, y^l)\right)$$

which is a CRF

- From JRF we have:

$$p_\theta(l, x^l) = \frac{\pi_l}{Z_\theta(l)} \sum_{y^l} \exp\left(u_\theta(x^l, y^l)\right)$$

$$= \frac{\pi_l}{Z_\theta(l)} \exp\left(u_\theta(x^l)\right)$$

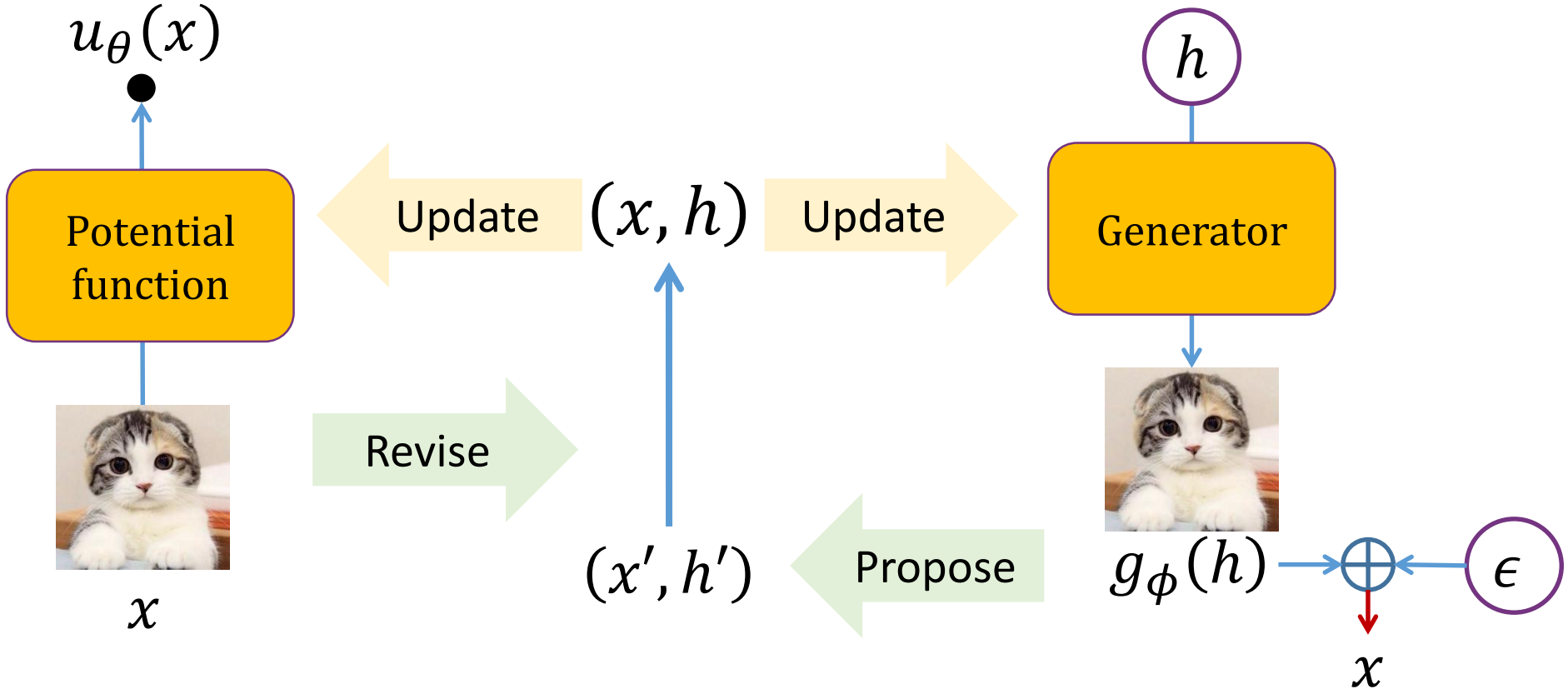
where $u_\theta(x^l) = \log \sum_{y^l} \exp\left(u_\theta(x^l, y^l)\right)$

which is a trans-dimensional random field (TRF)

[14] Yunfu Song, Zhijian Ou, et al. Upgrading CRFs to JRFs and its benefits to sequence modeling and labeling. ICASSP, 2020.

Inclusive-NRF algo. for learning from continuous data, e.g., Images.

simultaneously training a random field and a generator.



$$\begin{cases} \min_\theta KL[\tilde{p}(\tilde{x})||p_\theta(\tilde{x})] \\ \min_\phi KL[p_\theta(x)||q_\phi(x)] \end{cases} \Rightarrow \begin{cases} \nabla_\theta = E_{\tilde{p}(\tilde{x})}[\nabla_\theta log p_\theta(\tilde{x})] = E_{\tilde{p}(\tilde{x})}[\nabla_\theta u_\theta(\tilde{x})] - E_{p_\theta(x)}[\nabla_\theta u_\theta(x)] \\ \nabla_\phi = E_{p_\theta(x)}[\nabla_\phi log q_\phi(x)] = E_{p_\theta(x)q_\phi(h|x)}[\nabla_\phi log q_\phi(x,h)] \end{cases}$$

[12] Yunfu Song, Zhijian Ou. Learning Neural Random Fields with Inclusive Auxiliary Generators. arXiv:1806.00271, 2018.

Dynamic NCE algo. for learning from discrete data, e.g., texts.

Simultaneously train a random field and a generator.

- The target RF model $p_\theta(x) = \frac{1}{Z(\theta)} e^{u_\theta(x)}$
- Treat $\log Z(\theta)$ as a parameter ζ and rewrite $p_{\theta,\zeta}(x) \propto e^{u_\theta(x)-\zeta}$
- Introduce a **noise distribution** $q_n(x)$, and consider a binary classification

$$P(C=0|x) = \frac{p_{\theta,\zeta}(x)}{p_{\theta,\zeta}(x) + \nu q_n(x)}, where\ \nu = \frac{P(C=1)}{P(C=0)}$$

$$P(C=1|x) = 1 - P(C=0|x)$$

- Noise Contrastive Estimation (NCE):

$$\max_{\theta,\zeta} E_{x\sim p_0(x)}[\log P(C=0|x)] + E_{x\sim q_n(x)}[\log P(C=1|x)]$$

☺ $p_\theta \to p_0$ (oracle), under infinite amount of data and infinite capacity of p_θ.

☹ Reliable NCE needs a large $\nu \approx 20$; Overfitting. Dynamic-NCE in (Wang&Ou, SLT 2018).

Yunfu Song, Huahuan Zheng, Zhijian Ou.
An empirical comparison of joint-training and pre-training for domain-agnostic semi-supervised learning via energy-based models,
IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2021.

1. Related work and Motivation
2. Methods and Tasks
   - Pre-training vs. Joint-training
   - Across domains: image classification and natural language labeling
3. Experiments
4. Conclusion

Table 2. SSL for image classification over CIFAR-10 with 4,000 labels. The upper/lower blocks show generative/discriminative SSL methods respectively. The means and standard deviations are calculated over ten independent runs with randomly sampled labels.

Methods	error (%)
CatGAN [30]	19.58±0.46
Ladder network [31]	20.40±0.47
Improved-GAN [32]	18.63±2.32
BadGAN [33]	14.41±0.30
Sobolev-GAN [34]	15.77±0.19
Supervised baseline	25.72±0.44
Pre-training+fine-tuning EBM	21.40±0.38
Joint-training EBM	15.12±0.36
Results below this line cannot be directly compared to those above.	
VAT small [1]	14.87
Temporal Ensembling [2]	12.16±0.31
Mean Teacher [3]	12.31±0.28

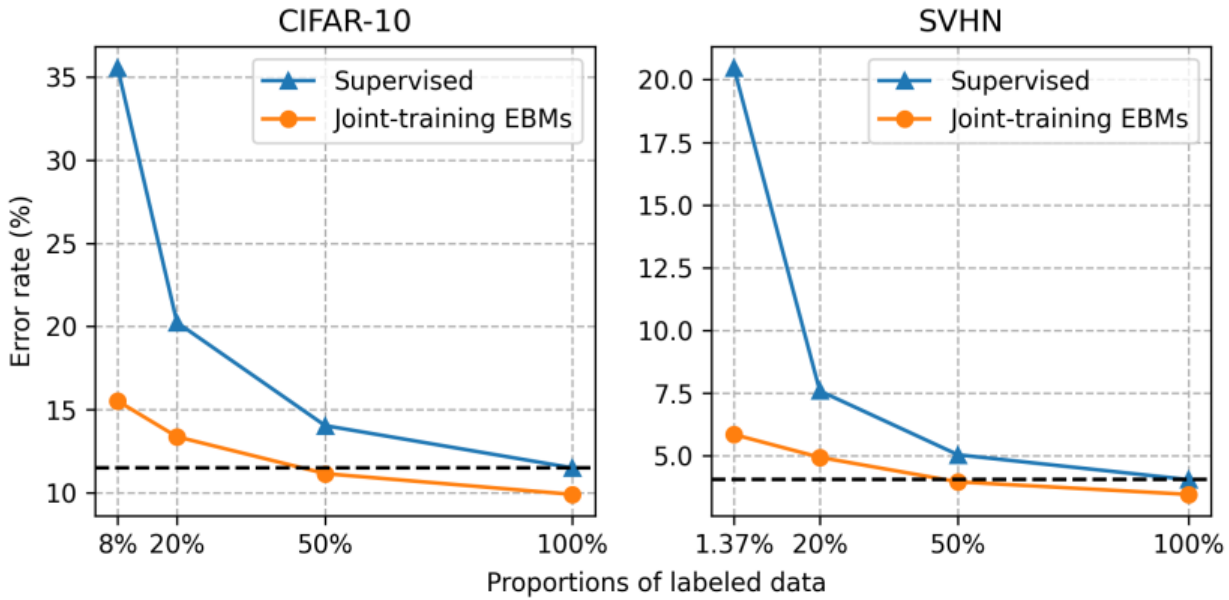


Fig. 1. Error rates of supervised baseline and joint-training EBMs as the amount of labels varies on SVHN and CIFAR-10 datasets. The dash line is the supervised result trained with 100% labeled data.

Table 3. Natural language labeling results. The evaluation metric is accuracy for POS and F_1 for chunking and NER. "Labeled" denotes the amount of labels in terms of the proportions w.r.t. the full set of labels. "U/L" denotes the ratio between the amount of unlabeled and labeled data. "U/L=0" denotes the supervised baseline. "pre." and "joint" denote the results by pre-training+fine-tuning EBMs and joint-training EBMs, respectively.

Labeled	U/L	POS tagging		Chunking		NER	
		pre.	joint	pre.	joint	pre.	joint
2%	0	95.57		78.73		78.19	
	50	95.72	95.92	81.62	82.24	76.74	77.61
	250	95.96	96.13	82.10	82.26	78.49	78.51
	500	96.08	96.24	83.10	83.05	79.47	79.17
10%	0	96.81		90.06		86.93	
	50	96.87	96.99	91.60	91.85	86.37	87.05
	250	96.88	97.00	91.09	91.93	86.86	86.77
	500	96.92	97.08	91.93	92.23	87.57	87.06
100%	0	97.41		94.77		90.74	
	50	97.40	97.49	95.05	95.31	91.24	91.34
	250	97.45	97.54	95.12	95.48	91.19	91.51
	500	97.46	97.57	95.19	95.50	91.30	91.52

Table 4. Relative improvements by joint-training EBMs compared to the supervised baseline (abbreviated as sup.) and pretraining+fine-tuning EBMs respectively. Refer to Table 3 for notations.

		joint over sup.			joint over pre.		
Labeled	U/L	POS	Chunking	NER	POS	Chunking	NER
2%	50	7.9	16.5	-2.7	4.7	3.4	3.7
	250	12.6	16.6	1.5	4.2	0.9	0.1
	500	15.1	20.3	4.5	4.1	-0.3	-1.5
10%	50	5.6	18.0	0.9	3.8	3.0	5.0
	250	6.0	18.3	-1.2	3.8	9.4	-0.7
	500	8.5	21.8	1.0	5.2	3.7	-4.1
100%	50	3.1	10.3	6.5	3.5	5.3	1.1
	250	5.0	13.6	8.3	3.5	7.4	3.6
	500	6.2	14.0	8.4	4.3	6.4	2.5

Conclusion

- We systematically evaluate and compare joint-training and pre-training for EBM-based domain-agnostic SSL, through a suite of experiments across a variety of domains such as image classification and natural language labeling.
- **Joint-training EBMs outperform pre-training EBMs marginally but nearly consistently.**
  - ▶Presumably, this is because that the optimization of joint-training is directly related to the targeted task, but pre-training is not aware of the labels for the targeted task.
- We hope this new finding would be helpful for future work to further explore better methods to leverage unlabeled data.

Reproducible code is at https://github.com/thu-spmi/semi-EBM

Final Notes

- Simply put, Data -Efficiency means using less Labels to get the same job done!
  - We need a spectrum of Data-Efficient methods, including but not limited to *Semi-Supervised-, Reinforcement-, Active-, Meta-, Lifelong- Learning*
    Model architectures
    …

- We plan to run a challenge/workshop:

Towards Semi-Supervised Task-Oriented Dialog Systems

- ~100K dialog transcripts between real users and customer agents from China Mobile
- Track1: Information Extraction and Knowledge Modeling from dialog transcripts
- Track2: Semi-Supervised Task-Oriented Dialog Systems
- Organizers: Tsinghua (me, Juanzi Li), China Mobile (Junlan Feng), …

Thanks !

Thanks to my collaborators and students :

Hong Liu, Yucheng Cai, Zhenru Lin, Yi Huang, Junlan Feng,

Yunfu Song, Huahuan Zheng

Supported by

NSFC 61976122, Ministry of Education and China Mobile joint funding MCM20170301, Apple,
Beijing National Research Center for Information Science and Technology