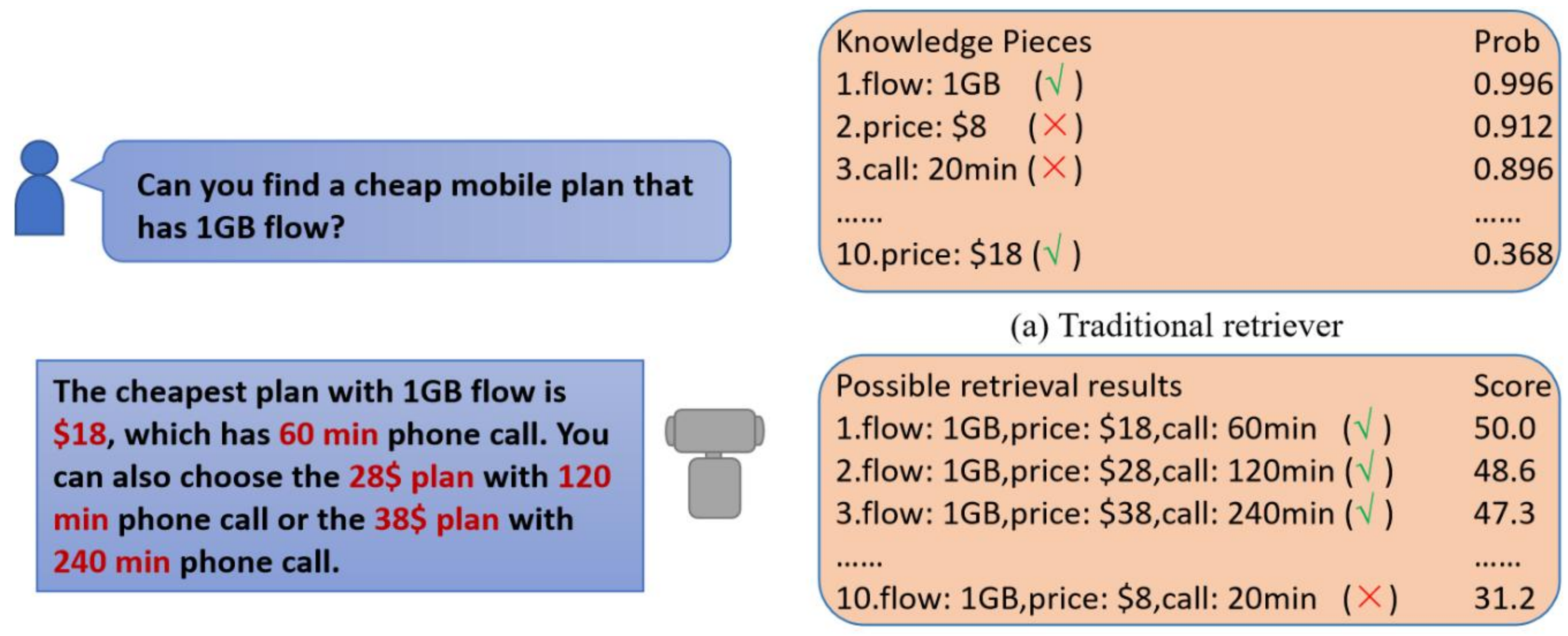
Entriever: Energy-based Retriever for Knowledge-Grounded Dialog Systems

Yucheng Cai, Ke Li, Yi Huang, Junlan Feng, Zhijian Ou
Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, China Mobile Research Institute, China
fengjunlan@chinamobile.com, ozj@tsinghua.edu.cn

Motivation

- Traditional retrievers assume conditional **independence** of knowledge pieces, ignoring inter-dependencies. This leads to **redundant retrievals** or **missing critical information** (e.g., price-flow package constraints) .
- In unlabeled data, **the knowledge base (KB) is unavailable**, making traditional methods unable to accurately compute retrieval probabilities, thus limiting **semi-supervised** dialog system performance **.**



(a) Traditional retriever

(b) Entriever (energy-based retriever)

Key Innovation

- **Holistic Modeling via Energy Function**: Treats candidate retrieval results (combinations of knowledge pieces) **as a whole**, calculating relevance scores through an energy function $U_\theta(c_t, u_t, \xi_t)$ to model inter-piece dependencies directly.
- **Residual Energy Design**: Constructs a residual form $p_\theta^{\mathrm{ret}} \propto p^{\mathrm{ref}} \cdot \exp(-U_\theta)$ based on traditional retrieval distribution p^{ref}, **reducing training difficulty**.
- **Semi-supervised Adaptability**: Enables retrieval probability calculation **without accessing the full KB**, suitable for pseudo-label filtering in unlabeled data.

Method

Energy Function Architecture:

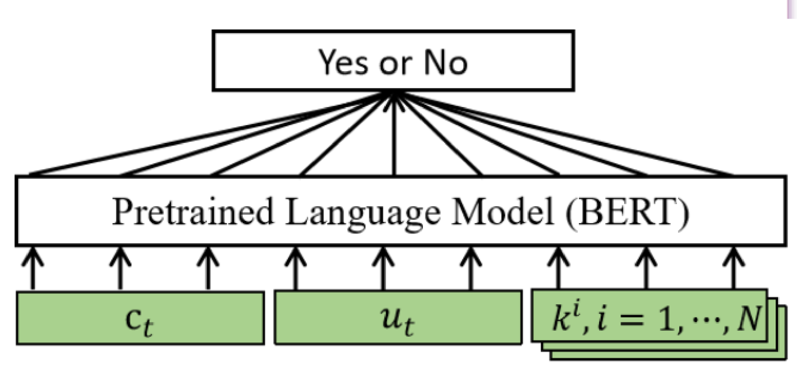
- **Inputs:** Dialog context $\mathrm{c_t}$ + user query $\mathrm{u_t}$ + knowledge piece combination ξ_t.

$$x \triangleq c_t \oplus u_t \oplus \xi_t$$

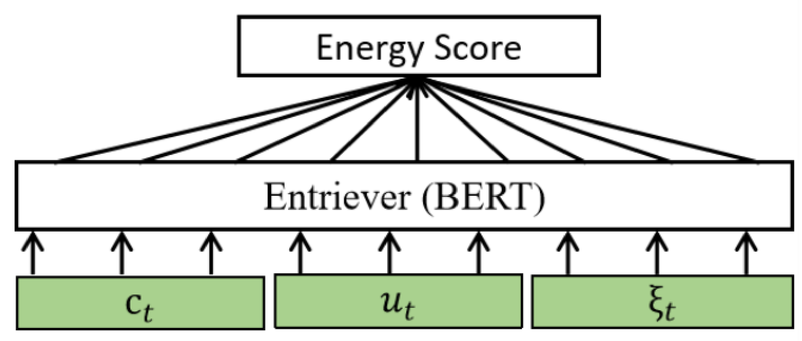
- **Architecture:** BERT bidirectional encoding + linear layer output

$$U_\theta(c_t, u_t, \xi_t)$$

$$= -\mathrm{Linear}\left(\sum_{i=1}^{|x|} \mathrm{enc}_\theta(x)[i]\right)$$



(a) Cross-encoder retrieval model



(b) Entriever

Training Methods:

- **Target**: Maximum Likelihood Estimation: $\mathcal{J}_\theta = -\log p_\theta^{\mathrm{ret}}(\xi_t | c_t, u_t)$

$$\frac{\partial \mathcal{J}_\theta(\xi_t|c_t,u_t)}{\partial \theta} = -\frac{\partial U_\theta(c_t,u_t,\xi_t)}{\partial \theta} + \mathbb{E}_{\xi_t \sim p_\theta^{\mathrm{ret}}}\left[\frac{\partial U_\theta(c_t,u_t,\xi_t)}{\partial \theta}\right]$$

- **Sampling Methods**
  - Importance Sampling (IS)
  - Metropolis Independence Sampling (MIS)

Retrieval Pipeline:

- **Retrieval Inference Flow:** Viterbi algorithm, to generation K candidates from the 2^N choices
- **Semi-supervised Application**

Weights for unlabeled data:

$$w(\xi_t) \propto \frac{\exp(-U_\theta(c_t,u_t,\xi_t)) \times p_\theta^{\mathrm{gen}}(r_t|c_t,u_t,\xi_t)}{q_\phi(\xi_t|c_t,u_t,r_t)}$$

Allowing for scoring pseudo knowledge without KB

Experiment Set up

Datasets: 4 TOD Datasets with Extensive Knowledge Interaction

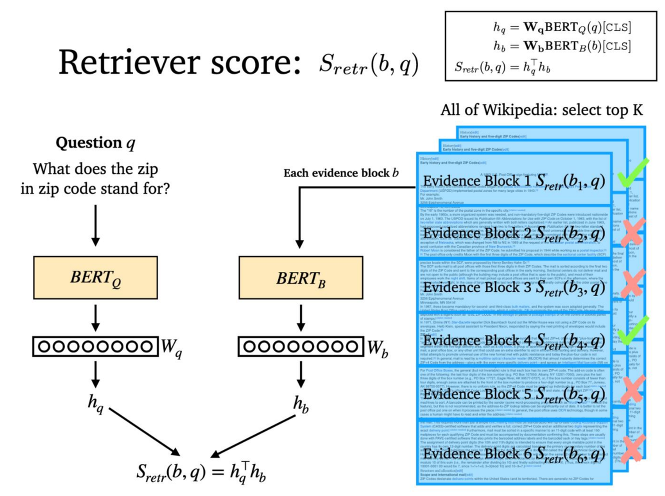
- MobileCS (Chinese)
- CamRest (English)
- In-Car (English)
- Woz2.1 (English)

Baselines:

- **Retrieval:** Dual-encoder (DPR), Cross-encoder
- **Semi-supervised dialog system:** JSA-KRTOD

Evaluation metrics:

- **retrieval:** Joint-acc / Inform / F1
- **dialog:** Success / BLEU (dialog)



DPR

Experiment Main Results

Results for **knowledge retrieval task**:

Method	MobileCS			Camrest			In-Car			Woz2.1		
	Joint-acc	Inform	F1	Joint-acc	Inform	F1	Joint-acc	Inform	F1	Joint-acc	Inform	F1
Cross-encoder	73.15	35.95	0.589	81.38	63.84	0.816	74.70	42.16	0.870	75.00	32.86	0.508
Entriever (MIS)	76.67	39.81	0.620	**83.17**	68.05	0.824	**78.66**	49.64	**0.875**	**80.24**	43.78	0.524
Entriever (IS)	**77.21**	**42.45**	**0.628**	**83.17**	**68.28**	**0.825**	78.51	**50.53**	**0.875**	79.72	**45.02**	**0.530**

Comparison over the MobileCS dataset for different **semi-supervision methods** (pseudo labeling (PL) and JSA)

Ratio	Method	Success	BLEU-4	Combined	p-value	
1:1	PL	87.5	8.853	105.21	0.025	0.013
	JSA	88.0	8.713	105.43		
	JSA + Entriever	90.6	9.816	**110.23**		
2:1	PL	87.8	9.196	106.19	0.006	0.018
	JSA	88.7	9.490	107.68		
	JSA + Entriever	92.1	9.725	**111.55**		
4:1	PL	88.5	9.341	107.18	0.049	0.088
	JSA	90.9	9.398	109.70		
	JSA + Entriever	92.8	9.554	**111.91**		
9:1	PL	89.4	9.532	108.46	0.083	0.192
	JSA	91.8	9.677	111.15		
	JSA + Entriever	93.0	9.627	**112.25**		

Semi-supervised response generation results on the MobileCS dataset

Method	Success	BLEU-4	Combined
Baseline (Liu et al., 2022)	31.5	4.170	39.84
Passion (Lu et al., 2022)	43.2	6.790	56.78
TJU-LMC (Yang et al., 2022)	68.9	7.54	83.98
PRIS (Zeng et al., 2022)	78.9	14.51	107.92
JSA-KRTOD (Cai et al., 2023)	91.8	9.677	111.15
JSA-KRTOD+Entriever (ours)	93.0	9.627	**112.25**

Ablation Results

Residual Structure: Joint-acc improved by 4.5%, significantly enhancing stability

Setting	Joint-acc	Inform	F1
Dual-encoder (Karpukhin et al., 2020b)	65.60	32.17	0.563
Cross-encoder (Cai et al., 2023)	73.15	35.95	0.589
Entriever (Non-residual, MIS)	76.94	31.89	0.593
Entriever (Non-residual, IS)	72.19	32.22	0.596
Entriever (Residual, MIS)	76.67	39.81	0.620
Entriever (Residual, IS)	**77.21**	**42.45**	**0.628**

Candidate number K: K=16 balances performance and computation

Config	Joint-acc	Inform	Precision	Recall	F1
$K = 4$	76.02	39.33	0.7162	0.5376	0.6142
$K = 8$	76.73	40.70	0.7054	0.5580	0.6231
$K = 16$	**77.21**	42.45	0.6855	0.5789	**0.6277**
$K = 32$	76.79	**42.60**	0.6455	0.6076	0.6260

Conclusion

- **Contributions:**
  - Apply energy-based language model to **retrieval**, modeling candidate retrieval results holistically
  - Extensive experiments demonstrate the efficacy of the energy-based retrieval model, and its potential in improving **semi-supervised dialog system**
- **Limitations:** BERT-based retriever can be substituted by LLM