Relevance-Promoting Language Model for Short-Text Conversation

Xin Li† Piji Li‡ Wei Bi‡ Xiaojiang Liu‡ Wai Lam†

†Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong

‡Tencent AI Lab

Overview

◆ **Potential issues in the existing model**:

(i) Training data under-exploitation: decoder-only word-by-word prediction.
(ii) Explanation-away issue: **recency-bias** in language model.
(iii) Copy and Word Repetition: maximization-based decoding strategy.

◆ **Contributions**:

(i) A non-encoder-decoder paradigm for the STC task .
(ii) Two relevance-promoting components for language model.
(iii) Generation with sampling-based decoding strategy.

Our Framework

◆ Transformer Language Model rather than Seq2Seq Encoder-Decoder.

◆ Top-k Sampling rather than Beam Search.

◆ Promote relevance modeling:

- SSA: Supervised Source Attention component.
- TI: Topic Inference component.

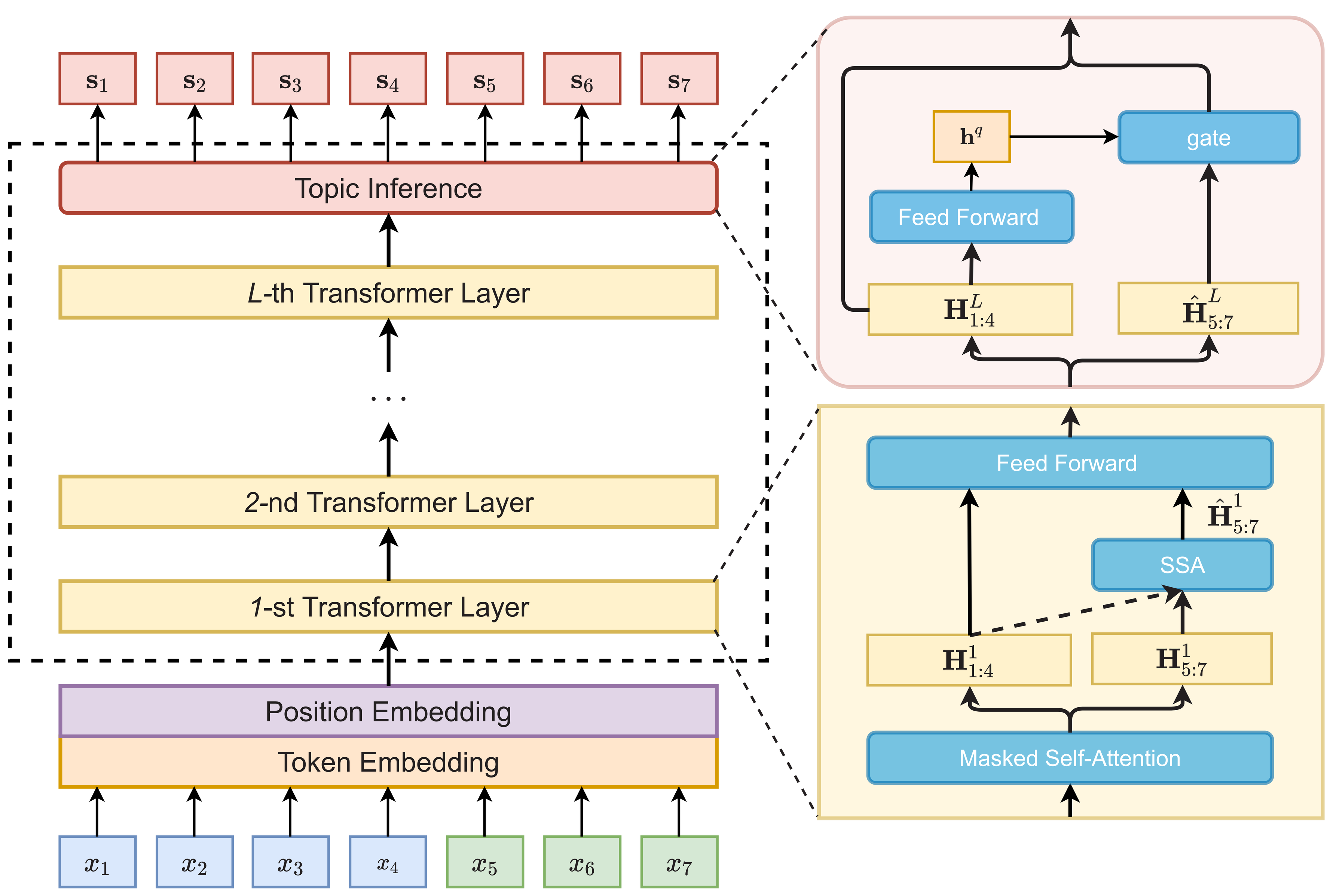

Figure 1: Our Framework.

Language Model as Response Generator

The predictions of words in source sentence are also considered:

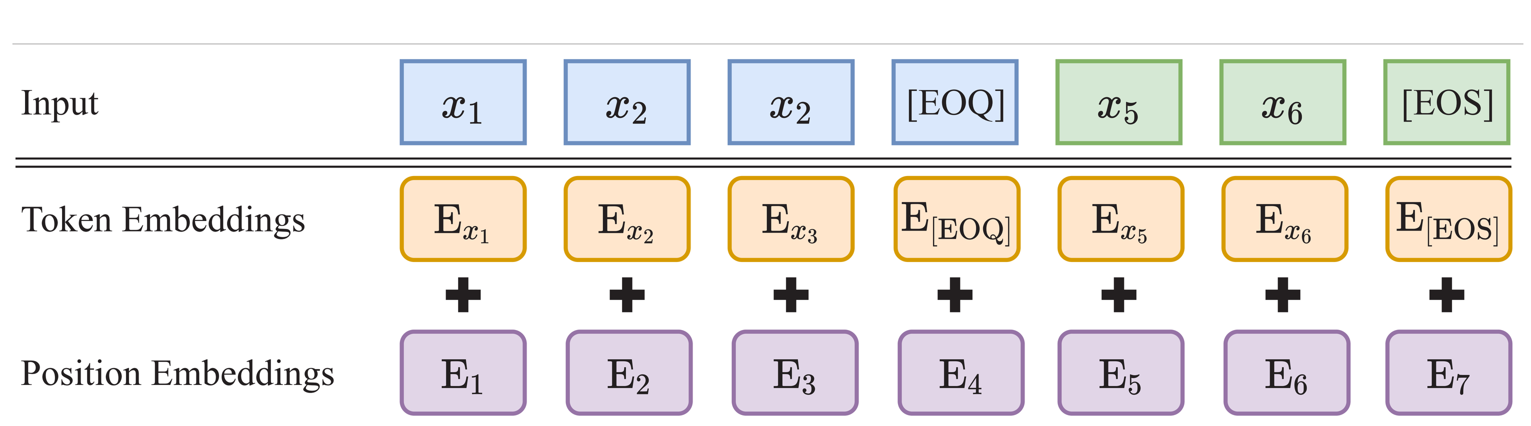

Figure 2: Language model input during training.

Promote Relevance Modeling

◆ **Vanilla Transformer Language Model**:

$$\begin{aligned} \mathbf{h}_t^l, \alpha_t^l &= \text{SLF-ATT}(\mathbf{q}_t^{l-1}, \mathbf{K}_{\le t}^{l-1}, \mathbf{V}_{\le t}^{l-1}) \\ \mathbf{Q}^{l-1} &= \mathbf{H}^{l-1}\mathbf{W}^Q \\ \mathbf{K}^{l-1}, \mathbf{V}^{l-1} &= \mathbf{H}^{l-1}\mathbf{W}^K, \mathbf{H}^{l-1}\mathbf{W}^V \end{aligned} \quad (1)$$

Transformer suffers from **explanation away** issue.

◆ **Supervised Source Attention**

Reconsider source-side information:

$$\begin{aligned} \hat{\mathbf{h}}_{t'}^l, \beta_{t'}^l &= \text{SRC-ATT}(\hat{\mathbf{q}}_{t'}^l, \hat{\mathbf{K}}^l, \hat{\mathbf{V}}^l) \\ \hat{\mathbf{Q}}^l &= \mathbf{H}^l\mathbf{W}^Q \\ \hat{\mathbf{K}}^l, \hat{\mathbf{V}}^l &= \mathbf{H}^l_{1:m}\mathbf{W}^K, \mathbf{H}^l_{1:m}\mathbf{W}^V \end{aligned} \quad (2)$$

Guide attention learning with source-side keywords:

$$\hat{\mathbf{y}}_i^{\text{src}} = \max\{\beta^L_{m+1,i}, \cdots, \beta^L_{n,i}\} \quad (3)$$

$$\mathcal{L}^{\text{src}} = \frac{1}{m}||\hat{\mathbf{y}}_i^{\text{src}} - \mathbf{y}^{\text{src}}||_2^2 \quad (4)$$

◆ **Topic Inference**

Incorporate source representation in prediction:

$$\mathbf{h}^q = f(\mathbf{x}_{1:m}), P(z|\mathbf{x}_{1:m}) = \text{Softmax}(\mathbf{W}^o\mathbf{h}^q) \quad (5)$$

$$\mathbf{s}_t = \begin{cases} (1-g_t) * \mathbf{h}_t^L + g_t * \mathbf{h}^q & \text{, if } t > m \\ \mathbf{h}_t^L & \text{, Otherwise} \end{cases}$$
$$g_t = \sigma(\mathbf{W}^g\mathbf{h}^q + \mathbf{W}'\mathbf{h}_t^L + \mathbf{b}), \quad (6)$$

Infer relevance-clues from references:

$$\mathcal{L}^{\text{kwd}} = -\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\mathbf{y}_i^{\text{kwd}} \cdot \log P_i(z|\mathbf{x}_{1:m}) \quad (7)$$

Randomization-Over-Maximization Decoding Strategy

◆ **Top-k Sampling**:

Given a conditional distribution $P(x|x_{1:i-1})$

(1) Obtain top-k vocabulary $V^{(k)}$.

(2) Re-scale the distribution:

$$P'(x|x_{1:i-1}) = \begin{cases} P(x|x_{1:i-1}) & \text{, if } x \in V^{(k)} \\ 0 & \text{, Otherwise} \end{cases} \quad (8)$$

(3) Do sampling: $x_i \sim P'(x|x_{1:i-1})$.

Training

Jointly consider auto-regressive LM loss ($\mathcal{L}^{\text{mle}}$), source attention loss ($\mathcal{L}^{\text{src}}$) and topic-based loss ($\mathcal{L}^{\text{kwd}}$):

$$\begin{aligned} \mathcal{L}(\theta) &= \frac{1}{|\mathbb{D}|}\sum_{(\mathbf{x},\mathbf{y}^{\text{src}},\mathbf{y}^{\text{kwd}})\in\mathbb{D}} \mathcal{L}(\mathbf{x}, \mathbf{y}^{\text{src}}, \mathbf{y}^{\text{kwd}}) \\ \mathcal{L}(\mathbf{x}, \mathbf{y}^{\text{src}}, \mathbf{y}^{\text{kwd}}) &= \mathcal{L}^{\text{mle}} + \gamma_1\mathcal{L}^{\text{src}} + \gamma_2\mathcal{L}^{\text{kwd}} \end{aligned} \quad (9)$$

Experiment

◆ Dataset: a large dataset built from **Baidu Baike**, **Douban** and **Weibo**.

◆ Evaluations: automatic evaluations and human evaluations.

Model	Relevance					Diversity	
	BLEU-2	BLEU-3	BLEU-4	HIT-Q	HIT-R	DIST-1	DIST-2
LSTM-LM	3.8	0.9	0.3	0.084	0.066	0.028	0.094
LSTM-S2S	5.6	2.8	1.8	0.293	0.145	0.039	0.137
TFM-LM	6.9	3.2	2.1	0.295	0.144	0.058	0.259
TFM-S2S	7.3	3.5	2.3	0.369	0.172	0.078	0.290
MMI	7.9	2.5	1.0	0.197	0.145	0.093	0.349
CVAE	5.8	1.5	0.4	0.211	0.135	0.060	0.211
MMPMS	6.7	3.0	1.8	0.151	0.102	0.057	0.220
OURS-tk w/o SSA & TI	4.9	1.0	0.3	0.119	0.076	0.086	0.441
OURS-tk w/o SSA	5.5	2.1	1.5	0.150	0.146	0.102	0.521
OURS-tk w/o TI	5.1	2.1	1.4	0.171	0.132	0.090	0.445
OURS-bm	**10.3**	**5.3**	**3.4**	**0.510**	**0.193**	0.102	0.398
OURS-tk	6.0	3.6	2.5	0.191	0.152	**0.107**	**0.544**

Table 1: Automatic evaluations.

Model	Evaluation Metrics		
	Relevance	*Fluency*	*Acceptance*
LSTM-LM	1.206	1.297	0.26
LSTM-S2S	1.386	1.285	0.37
TFM-LM	1.412	1.328	0.39
TFM-S2S	1.475	1.306	0.43
MMI	1.432	1.301	0.34
CVAE	1.316	1.274	0.33
MMPMS	1.528	1.396	0.42
OURS-tk w/o SSA & TI	1.273	1.368	0.28
OURS-tk w/o SSA	1.485	**1.407**	0.39
OURS-tk w/o TI	1.503	1.303	0.36
OURS-bm	1.515	1.359	0.38
OURS-tk	**1.606**	1.346	**0.44**

Table 2: Human evaluations.

Experiment

TFM-LM	MMI	TFM-S2S	MMPMS	OURS-tk
Query #1: 有了这样的花梯，回家的感觉真好 (With this flower ladder, it feels so good going back home.)				
好漂亮的花，好喜欢 The flower is beautiful! Like it!	我也有这样的感觉 I have the same feeling!	我很喜欢！ I really like it!	我擦，这也太有才了吧 Oh my god! It is really funny	美得令人窒息的花梯 What a beautiful flower ladder!!
Query #2: 抹茶与奶油的结合，精致得不忍下口 (This is the combination of matcha and cream. It is quite delicate such that I couldn't bear to eat.)				
不忍下口，不忍下口。 Couldn't bear to eat, couldn't bear to eat.	好想吃啊！ I really want to have a try.	我能接受。 The food is acceptable for me.	我喜欢抹茶味的抹茶。 I like the matcha of matcha flavor.	我喜欢抹茶和奶油 I love matcha and cream!
Query #3: 移动的别墅。介个太强大了 (This moving cottage is awesome!)				
我也想要这样的房子 I wanna a house like this, too.	有创意，有创意啊 It is quite creative.	介个太强大了，房间装不下 It is too strong to fit in the room.	移动和联通 Chinese Mobile and Chinese Unicom	这是哪儿啊，怎么拍的，好厉害好酷哦 Where is it? How do you shoot this video. It look cool!

◆ OUR model performs better on relevance metrics, diversity metrics and also generates fluent & natural responses.

◆ Analysis:

- Supervised Source Attention component and Topic Inference component are beneficial for the generation of informative topical words related to the query.
- Top-k sampling is simple yet effective to achieve diverse response generation but we should be careful with its uncertainty on relevance and fluency.
- Decoder trained on more data can give more fluent output.

Further Discussion about Top-k Sampling

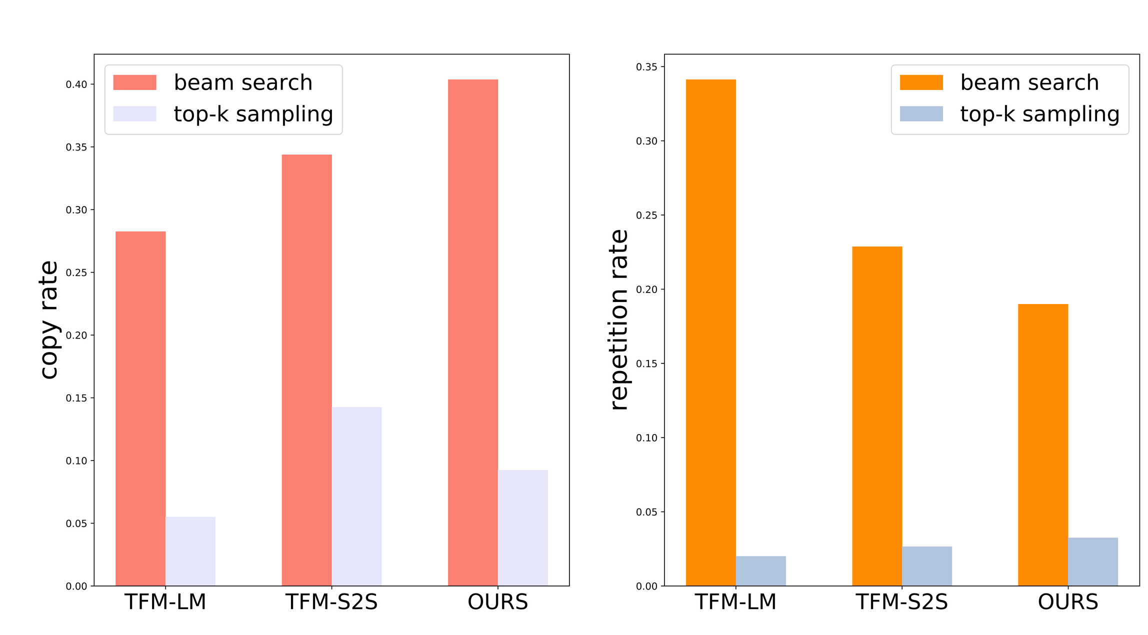

Figure 3: Beam Search v.s. Top-k Sampling.

◆ Top-k sampling greatly reduces the query copy rate.

◆ Top-k sampling almost eliminates the phrase repetition phenomenon.

Conclusion

◆ An alternative LM-based solution is proposed for STC task.

◆ Relevance-promoting components make up for the LM in conditional generation.

◆ Top-k Sampling consistently improves the naturalness and diversity of generation but it may hurt the results on relevance metrics.