A Generative Approach for Script Event Prediction via Contrastive Fine-tuning

Fangqi Zhu Li [1] Jun Gao [1] Changlong Yu Wei Wang Chen Xu [3] Xin Mu [2] Min Yang [4] Ruifeng Xu[1,2,5]

[1]Harbin Institute of Technology (Shenzhen) [2]Peng Cheng Laboratory [3]Beijing University of Technology [4]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences [5]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies

Introduction

We want to learn a knowledge model that can predict the next events based on the sequence of previous events.

Many applications: - story generation - dialogue generation

Previous approaches:

- using pretrained language models - incorporating external knowledge (e.g., discourse relations)

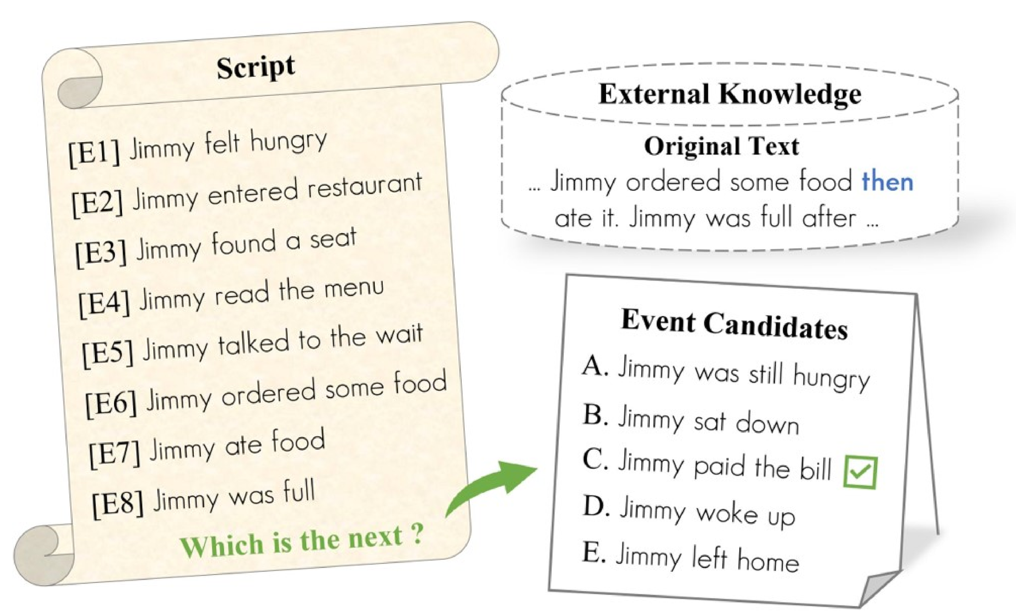
Remaining challenges:

- currently adopted pretrained language models ignore event-level knowledge
- modeling correlations between events with discourse relations is limited

Basic idea:

- Post-pretrain the pretrained language model with an event-centric pretraining objective to inject event-level knowledge into the pretrained language model
- Design a likelihood-based contrastive loss for fine-tuning the generative model

Task Definition



Script event prediction with external knowledge.

Giving a script, the task aims to select the subsequent event from candidate events.

Events E6~E8 are extracted from the original text where "then" is a discourse marker, which is used by previous methods to extract the discourse relation.

Note that our method is **free** with this kind of external knowledge (dashed box).

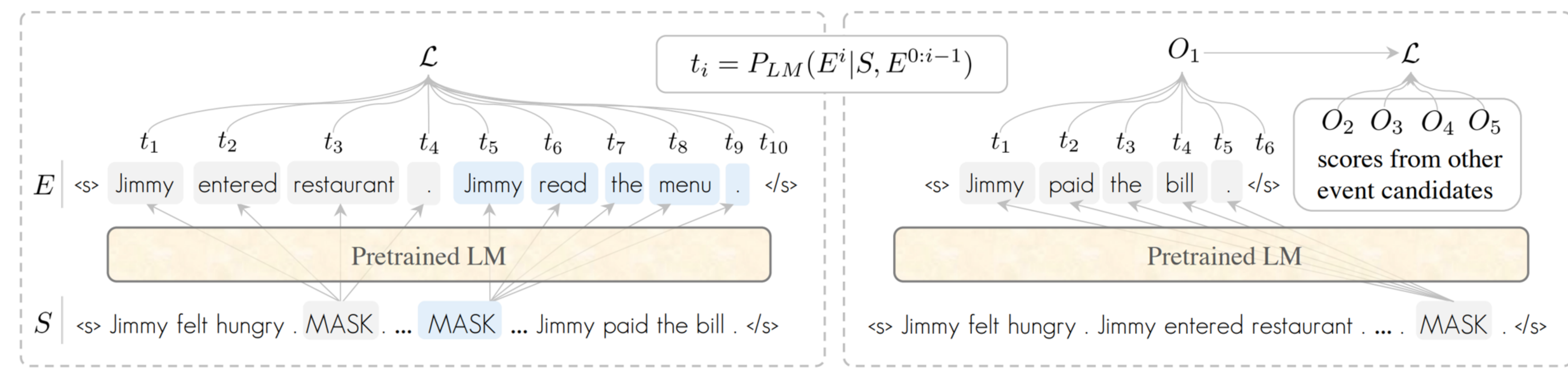
The Proposed Approach

Overview: our proposed approach for training generative model includes two stage: event-centric pretraining and task-specific contrastive fine-tuning.

Event-Centric Pretraining: we transform the event chain into the format of natural language and then randomly replace some events with [MASK] tokens (each masked event corresponding to a [MASK] token). Finally, we force the generative model to generate the masked events sequentially according to the masked event chain (both input and output event sequences are in the natural language format).

Task-Specific Contrastive Fine-tuning: we calculate the conditional generation probability of each candidate event given the script and optimize the scores to make the score of the correct candidate higher and the scores of incorrect candidates lower. We also design a special COT loss to make the scores of incorrect candidates equally low, resulting in a higher performance of our model.

Model Architecture



Event-Centric Pretraining (a) Task-Specific Contrastive Fine-tuning (b)

As shown above, for both stages, the generation probability of each token t_i is calculated by

$$t_i = P_{LM}(E^i|S, E^{0:i-1}) \tag{1}$$

For the event-centric pretraining stage, our training objective is to **maximize the sum of the generation probabilities of each token** in the masked event sequence E given masked script S. For the task-specific contrastive fine-tuning stage, we **calculate the average of the generation probabilities of each token in each candidate event as the score** for each candidate event, and subsequently optimize these scores with the COT Loss we designed. We also try Cross Entropy Loss and Margin Ranking Loss as alternatives.

Loss Functions

We introduce COT (Complement Objective Training) Loss to optimize the scores of each candidate event and try two other loss functions, including Cross Entropy Loss and Margin Ranking Loss.

COT Loss

$$\mathcal{L}_{cot} = -\log(s_t) + \frac{1}{M-1}\sum_{\substack{i=1\\ i\neq t}}^{M}\left(\frac{s_i}{1-s_t}\right)\log\left(\frac{s_i}{1-s_t}\right) \tag{2}$$

In Equation (2), s_i denotes the score (normalized) of candidate event i, t denotes the subscript of the correct candidate event, and M denotes the number of candidate events. COT Loss improves Cross Entropy Loss by maximizing the likelihood of the ground truth class while neutralizing the probabilities of the complement(incorrect) classes.

Cross Entropy Loss

$$\mathcal{L}_{cross} = -\log(s_t) \tag{3}$$

Margin Ranking Loss

$$\mathcal{L}_{margin} = \sum_{1\leq i\leq M, i\neq t} \max\left(m - (s_i - s_t), 0\right) \tag{4}$$

In Equation (4), hyperparameter m controls the interval between the score of correct candidate and incorrect.

Dataset Statistics

	Original Dataset	Public Dataset
Train set	1,440,295	140,331
Dev set	10,000	10,000
Test set	10,000	10,000

Table: The statistics of the reproduced original dataset and the public dataset.

Experiments

Experiments on small public dataset

Methods	Acc. (%)	ext.
w/ external knowledge		
SGNN + Int&Senti	56.03	Int & Senti
RoBERTa$_{base}$ + Rep. Fusion	58.66	ASER
RoBERTa$_{base}$ + Know. Model	59.99	ASER
w/o external knowledge		
Random	20.00	w/o ext.
Event-Comp	49.57	w/o ext.
PairLSTM	50.83	w/o ext.
SGNN	52.45	w/o ext.
GraphBERT	<u>60.72</u>	w/o ext.
BART$_{base}$	60.00	w/o ext.
Ours (BART$_{base}$)	**62.94**	w/o ext.

Methods	Acc. (%)	ext.
w/ external knowledge		
EventBERT	63.50	BookCorpus
RoBERTa$_{large}$ + Know. Model	63.62	ASER
ClarET	64.61	BookCorpus
w/o external knowledge		
Ours (BART$_{large}$)	<u>64.82</u>	w/o ext.
Ours (BART$_{large}$) + NYT	**65.88**	NYT

Experiments on big original dataset

Methods	Acc. (%)	ext.
w/ external knowledge		
NG	63.59	discourse relation
MCPredictor	<u>67.14</u>*	original text
w/o external knowledge		
SAM-Net	55.60	w/o ext.
SCPredictor-s	58.79*	w/o ext.
BERT-based SCPredictor-s	59.13	w/o ext.
Ours (BART$_{base}$)	**67.21**	w/o ext.

Ablation Study

Methods	Acc. (%)
Ours (BART$_{base}$)	62.94
w/o event-centric pretraining	61.08
w/o task-specific contrastive fine-tuning	40.00
replace with a linear classifier	61.77
replace with random span mask	61.44
replace with sum of log-probabilities	60.84
replace with Cross Entroy Loss	62.71
replace with Margin Ranking Loss	61.18

Paper & Code

- **Preprint**: `arxiv.org/abs/2212.03496v3`
- **Code**: `github.com/zhufq00/mcnc`