北京大學
PEKING UNIVERSITY

AUTOPROMPT: Eliciting Knowledge from Language Models with Automatically Generated Prompts

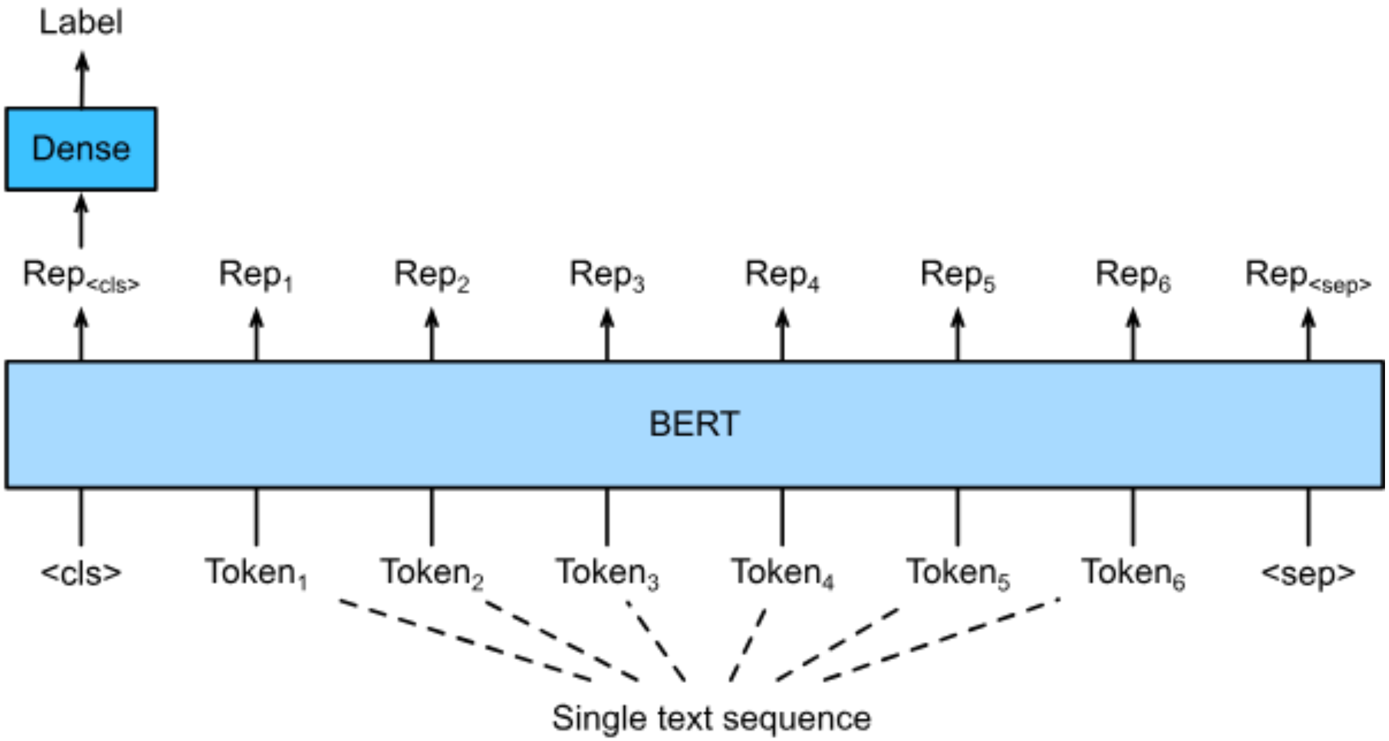
Jiayu Yao

Motivation

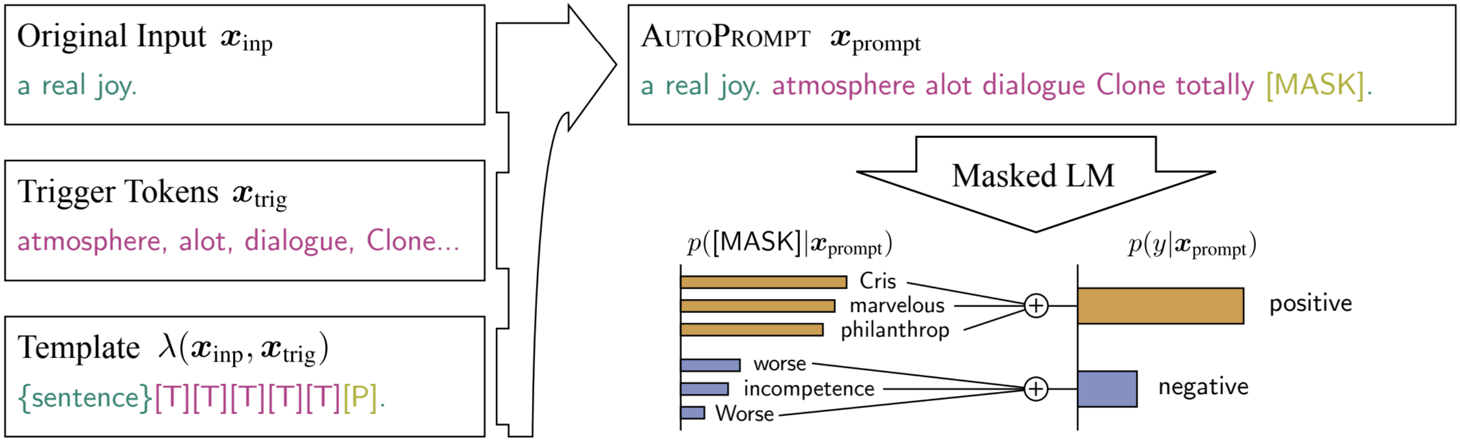
Determine whether knowledge are learned during language model training process

PROMPTING TO ELICIT KNOWLEDGE FROM LM

Method


Label
Dense
Rep<cls>
Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep<sep>
BERT
<cls>
Token1
Token2
Token3
Token4
Token5
Token6
<sep>
Single text sequence

Method


Original Input $\boldsymbol{x}_{\text{inp}}$
a real joy.
Trigger Tokens $\boldsymbol{x}_{\text{trig}}$
atmosphere, alot, dialogue, Clone...
Template $\lambda(\boldsymbol{x}_{\text{inp}}, \boldsymbol{x}_{\text{trig}})$
{sentence}[T][T][T][T][T][P].
AUTOPROMPT $\boldsymbol{x}_{\text{prompt}}$
a real joy. atmosphere alot dialogue Clone totally [MASK].
Masked LM
$p([\text{MASK}]|\boldsymbol{x}_{\text{prompt}})$
Cris
marvelous
philanthrop
worse
incompetence
Worse
$p(y|\boldsymbol{x}_{\text{prompt}})$
positive
negative

- Find Trigger Prompt
- Find Label Token Set

Method

$$p(y|\boldsymbol{x}_{\text{prompt}}) = \sum_{w \in \mathcal{V}_y} p(\text{[MASK]} = w|\boldsymbol{x}_{\text{prompt}})$$

$$\mathcal{V}_{\text{cand}} = \underset{w \in \mathcal{V}}{\text{top-}k} \left[\boldsymbol{w}_{\text{in}}^T \nabla \log p(y|\boldsymbol{x}_{\text{prompt}}) \right]$$

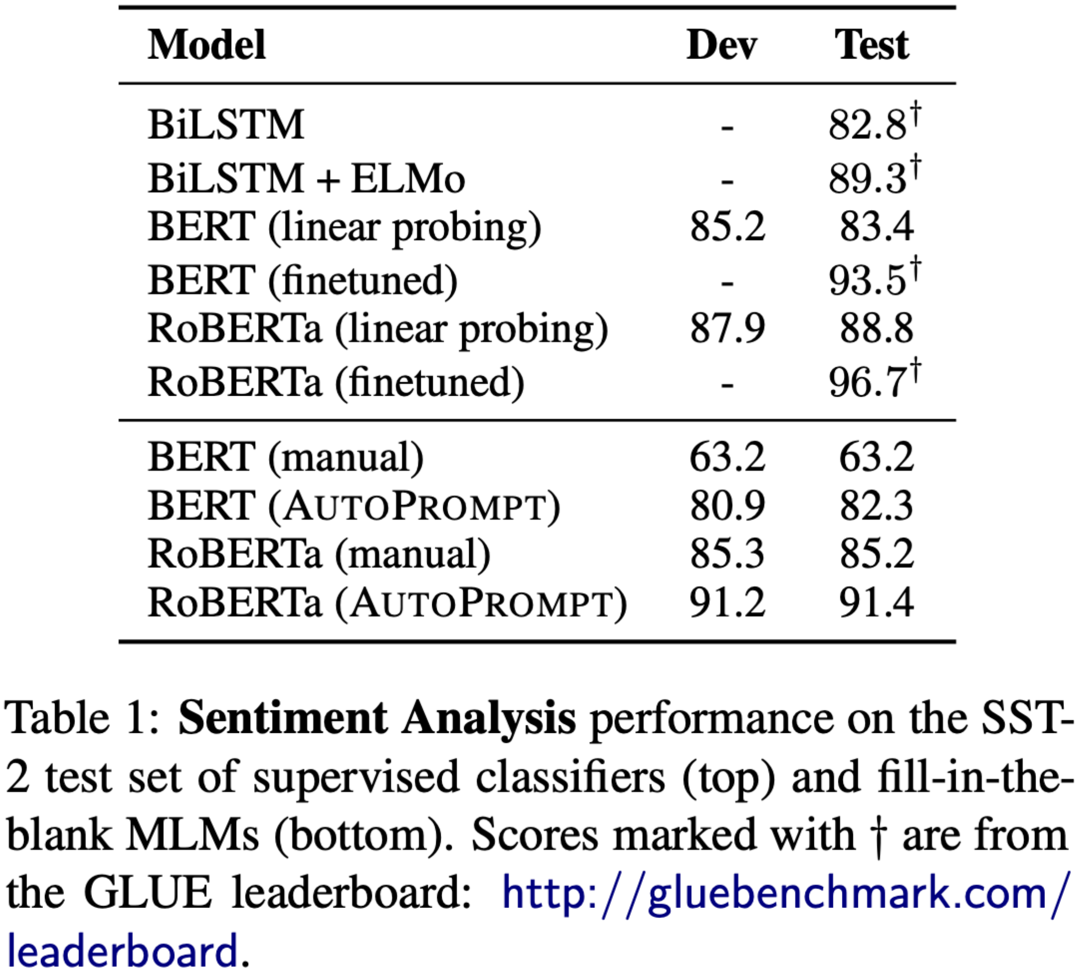
Label Projection

Replace $\boldsymbol{y}$ with $\boldsymbol{w}_{out}$ (token output embedding)

$$s(y, w) \;=\; p(y|\boldsymbol{w}_{\text{out}}).$$

$$\mathcal{V}_y = \underset{w \in \mathcal{V}}{\text{top-}k} \left[s(y, w) \right]$$

Experiment



Model	Dev	Test
BiLSTM	-	82.8†
BiLSTM + ELMo	-	89.3†
BERT (linear probing)	85.2	83.4
BERT (finetuned)	-	93.5†
RoBERTa (linear probing)	87.9	88.8
RoBERTa (finetuned)	-	96.7†
BERT (manual)	63.2	63.2
BERT (AUTOPROMPT)	80.9	82.3
RoBERTa (manual)	85.3	85.2
RoBERTa (AUTOPROMPT)	91.2	91.4

Table 1: **Sentiment Analysis** performance on the SST-2 test set of supervised classifiers (top) and fill-in-the-blank MLMs (bottom). Scores marked with † are from the GLUE leaderboard: http://gluebenchmark.com/leaderboard.

Discussion

- LLM Evaluation
    - Conventional Evaluation
    - Co-evaluation
- Prompt Guidence

Thanks