Preference-grounded Token-level Guidance for Language Model Fine-tuning

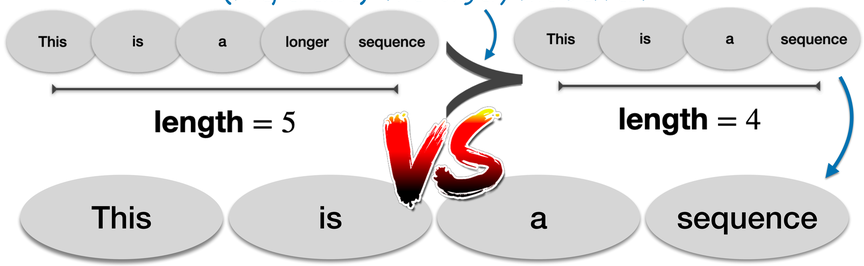
Shentao Yang, Shujian Zhang, Congying Xia, Yihao Feng, Caiming Xiong, Mingyuan Zhou

Full Paper | GitHub Repo

Motivation

- Sequence-level **preference** *v.s.* Token-level **training loss**

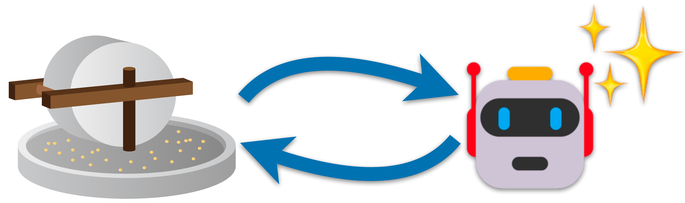
(Preference) The longer, the better!



$\Pr(\textbf{This} \mid \textbf{<sos>}) \times \Pr(\textbf{is} \mid \textbf{This}) \times \Pr(\textbf{a} \mid \textbf{This is}) \times \Pr(\textbf{sequence} \mid \textbf{This is a})$

- **Problem**: **Granularity mismatch** ⟷ **Delayed feedback**

Proposed Method Sketch



- Alternate training process

① **Ground** sequence-level preference into token-level training guidance → token-level "reward" $r_\phi(s_t, a_t)$

> *" Is it good to select this word here? "*

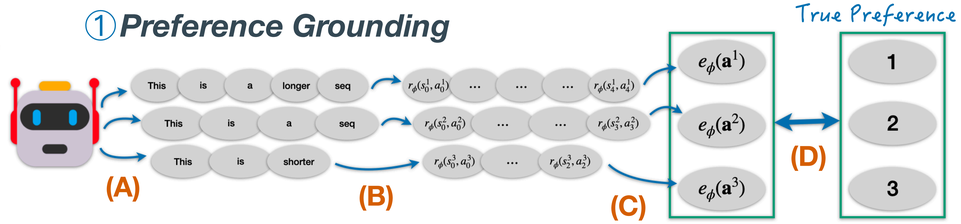
② **Improve** the LM using the learned guidance

> *" Select the next word that has a high reward ! "*

③ **Alternate** ⟶ more up-to-date guidance

Preference-grounded Token-level Guidance

① *Preference Grounding*



(A) **Sample:** Use LM π_θ to generate $K \geq 2$ text sequences $(\mathbf{a}^1, \ldots, \mathbf{a}^K)$

(B) **Score:** Let r_ϕ to produce reward for each step of each sequence

(C) **Aggregate:** Use $f(\cdot)$ to aggregate the token-level rewards into parametrized sequence evaluation $e_\phi(\mathbf{a}^k) = f(\{r_\phi(s_t^k, a_t^k)\}_{t=0}^{T^k-1})$

(D) **Align:** Align the ordering of all $\{e_\phi(\mathbf{a}^k)\}_{k=1}^K$ with true preference

Aggregation function $f(\cdot)$: LM specific, ~~*Summation* $\sum(\cdot)$~~

- Our proposal: *Average*, *Soft Maximum*, *Soft Minimum*

Alignment Loss: Plackett–Luce choice model

$$\min_\phi \mathscr{L}(\phi) \triangleq -\log P(\text{ord} \mid \{e_\phi(\mathbf{a}^k)\}_{k=1}^K),$$

$$P\left(\text{ord} \,\middle|\, \{e_\phi(\mathbf{a}^k)\}_{k=1}^K\right) = \prod_{k=1}^K \left\{ \exp\left(e_\phi(\mathbf{a}^k)\right) \Big/ \sum_{i=k}^K \exp\left(e_\phi(\mathbf{a}^i)\right) \right\}$$

where $\text{ord} = (1, \ldots, K)$ is the assumed true preference

② *LM π_θ Improving*

Setting 1: **No Supervised Data** — REINFORCE + Max-Entropy

$$\mathbb{E}_{t\sim\text{Uniform}\{0,\ldots,T-1\}} \left\{ \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}[r_\phi(s_t, a_t) \times \nabla_\theta \log \pi_\theta(a_t \mid s_t)] + \alpha \times \nabla_\theta \mathscr{H}(\pi_\theta(\cdot \mid s_t)) \right\}$$

where $\mathscr{H}(\pi_\theta(\cdot \mid s_t))$ denotes entropy and α is a hyperparameter

Setting 2: **With Supervised Data** — Reward-weighted MLE

$$\min_\theta -\mathbb{E}_{(x,y)\sim\mathscr{D}} \left[\sum_{t=0}^{|y|-1} w_t \times \log \pi_\theta(y_t \mid s_t) \right], \text{ with } w_t = \frac{r_\phi(s_t, y_t)}{\sum_{t'=0}^{|y|-1} r_\phi(s_{t'}, y_{t'})}$$

Takeaway

Train a sequential-decision-making model: **dense** guidance can be more effective 💪💪💪 than **sparse** delayed feedback 😞😞😞

Results: Prompt Generation

- ***Task***: Generate text prompts to ask a LLM to classify texts

- No supervised data → Use REINFORCE + Max-Entropy

		SST-2	Yelp P.	AG News
Finetuning	Few-shot Finetuning	80.6 (3.9)	88.7 (4.7)	**84.9** (3.6)
Continuous Prompt	Soft Prompt Tuning	73.8 (10.9)	88.6 (2.1)	82.6 (0.9)
	BB Tuning-50	89.1 (0.9)	93.2 (0.5)	83.5 (0.9)
	AutoPrompt	75.0 (7.6)	79.8 (8.3)	65.7 (1.9)
Discrete Prompt	Manual Prompt	82.8	83.0	76.9
	In-Context Demo	85.9 (0.7)	89.6 (0.4)	74.9 (0.8)
	Instructions	89.0	84.4	54.8
	GrIPS	87.1 (1.5)	88.2 (0.1)	65.4 (9.8)
	RLPrompt	90.5 (1.5)	94.2 (0.7)	79.7 (2.1)
	Ours (AVG)	**92.6** (1.7)	94.7 (0.6)	82.8 (1.5)
	Ours (MIN)	91.9 (1.8)	94.4 (0.8)	82.4 (1.1)
	Ours (MAX)	91.2 (2.5)	**94.8** (0.5)	83.3 (1.4)

- Competitive and stable results on all datasets 🎉🎉

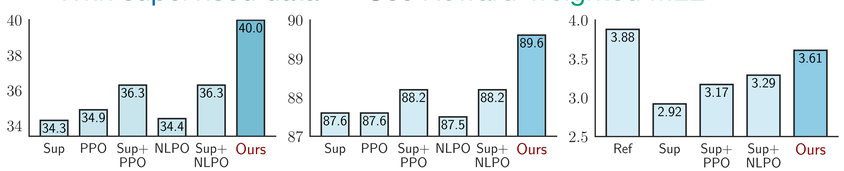
- Examples of generated prompt:

  - "`newsIntroduction Comments Tags Search`"
  - "`newsTopic Blog Support Category`"

- Topic-classifying keyword → *soft-maximum* aggregation ✓

Results: Text Summarization

- With supervised data → Use Reward-weighted MLE



- Perform favorably against strong baselines (under T5-base)

§ ***Performance under different aggregation functions $f(\cdot)$***

Algorithm	ROUGE-1	ROUGE-2	ROUGE-L	Meteor	BertScore
Ours - AVG	**43.09** (0.06)	**20.17** (0.04)	**39.99** (0.07)	**35.23** (0.06)	**89.61** (0.12)
Ours - SUM	42.86 (0.08)	19.92 (0.08)	39.76 (0.11)	34.74 (0.37)	89.24 (0.11)
Ours - MIN	42.92 (0.14)	20.01 (0.02)	39.84 (0.08)	34.88 (0.13)	89.33 (0.07)
Ours - MAX	42.38 (0.17)	19.49 (0.02)	39.34 (0.09)	34.13 (0.32)	89.09 (0.19)

- Unequal sequence-lengths → *Average* ✓ *Summation* ✗

- Overall text-quality → *Soft Minimum* ✓ *Soft Maximum* ✗

🔥🔥 *Importance of customizing $f(\cdot)$ for specific LM task* 🔥🔥