Selective Generation for Controllable Language Models (NeurIPS 2024 Spotlight Paper)

Kyungmin Kim

Graduate School of Artificial Intelligence
Pohang University of Science and Technology

Joint work with Minjae Lee (POSTECH),

Taesoo Kim (Georgia Tech), and Sangdon Park (POSTECH)

TL;DR

- Learn an **“entailment-aware” selective generator** to control the **rate of hallucination** for a given language model under a specific downstream language generation task.

Contributions

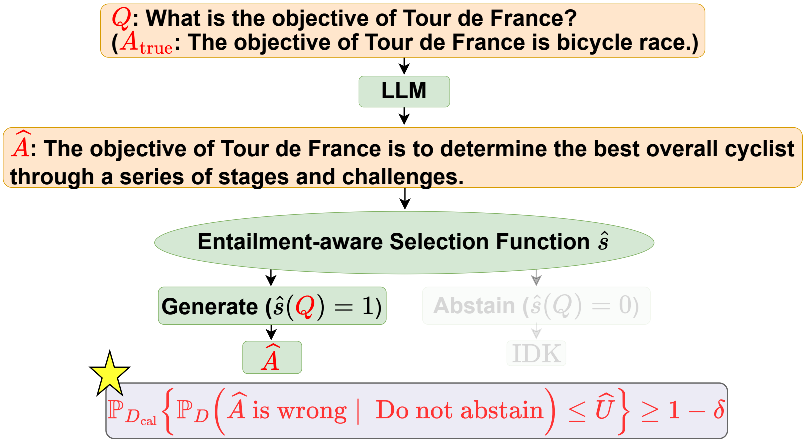
1. Propose the first “certified” selective generator learning algorithm for language generation.

2. Leverage textual entailment as a correctness metric.

3. Design a cost-efficient semi-supervised learning algorithm.

4. Prove a controllability guarantee of the proposed algorithm.

Overview



Q: **What is the objective of Tour de France?**
(A_{true}**: The objective of Tour de France is bicycle race.)**

LLM

$\widehat{A}$: **The objective of Tour de France is to determine the best overall cyclist through a series of stages and challenges.**

Entailment-aware Selection Function $\hat{s}$

Generate $(\hat{s}(Q) = 1)$

Abstain $(\hat{s}(Q) = 0)$

$\widehat{A}$

IDK

$$\mathbb{P}_{D_{\text{cal}}}\left\{\mathbb{P}_{D}\left(\widehat{A} \text{ is wrong} \mid \text{ Do not abstain}\right) \leq \widehat{U}\right\} \geq 1-\delta$$

Figure 1: An illustration of selective generation in the inference time

Related Work

- Selective classifier[1]

$$\hat{S}(\mathbf{X}) = \begin{cases} \hat{\mathbf{Y}} & \hat{s}(\mathbf{X}) = 1, \\ \texttt{IDK} & \text{o.w.} \end{cases}$$

- Selective generator

$$\hat{S}(\mathbf{Q}) = \begin{cases} \hat{\mathbf{A}} & \hat{s}(\mathbf{Q}) = 1, \\ \texttt{IDK} & \text{o.w.} \end{cases}$$

[1] Geifman and El-Yaniv. 2017. Selective Classification for Deep Neural Networks. *NeurIPS*.

Q. Why don’t we just directly apply
selective classification to language generation task?

Main Challenge: Metric Misalignment

Definition. Metric Misalignment

Learning Metric (*e.g.* EM) $\neq$ Evaluation Metric (*e.g.* SC)

- Example:
  - **Q**: **Where in the bible does it mention Sodom and Gomorrah?**
  - $\mathbf{A}_{\text{true}}$: **The book of Genesis mentions Sodom and Gomorrah.**
  - **A**: **The story of Sodom and Gomorrah is found in Genesis 19.**
- A standard learning metric on correct answers, *i.e.* Exact Match (EM), assumes a single correct answer (*i.e.* $\mathbf{A} =_{\texttt{EM}} \mathbf{A}_{\text{true}}?$)
- As $\mathbf{A} \neq_{\text{EM}} \mathbf{A}_{\text{true}}$, $\mathbf{A}$ is wrong even if it is semantically correct (SC) ☹.

Idea 1: Textual Entailment as a Correctness Metric

Definition. Correctness Metric by Entailment

A generated answer $\mathbf{A}$ is correct if

$$\mathbf{A} \in E_{\text{true}}(\mathbf{A}_{\text{true}}) := \{\tilde{\mathbf{A}} \mid \tilde{\mathbf{A}} \text{ entails } \mathbf{A}_{\text{true}}\}.$$

Definition. False Discovery Rate w.r.t. Entailment (FDR-E)

$$\text{Learning Metric:} \quad \mathbb{P}_{\mathcal{D}}\left(\mathbf{A} \notin E_{\text{true}}(\mathbf{A}_{\text{true}}) \,\middle|\, \hat{S}(\mathbf{Q}) \neq \texttt{IDK}\right)$$

- We find a learning algorithm to control the FDR-E.

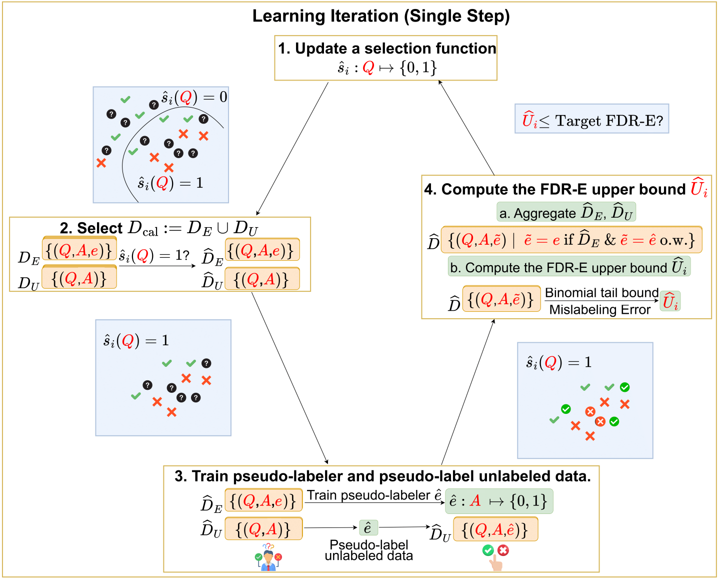
Idea 2: Pseudo-labeling Textual Entailment

Calibration Set

$$\{(\mathbf{Q}, \mathbf{A}_{\text{true}}, \underbrace{\mathbf{A} \in E_{\text{true}}(\mathbf{A}_{\text{true}})}_{\text{additional labels}})\} \cup \{(\mathbf{Q}, \mathbf{A}_{\text{true}}, \underbrace{\mathbf{A} \in \hat{E}(\mathbf{A}_{\text{true}})}_{\text{pseudo labels}})\}$$

- We propose a label efficient semi-supervised learning algorithm.

Solution: Semi-supervised Selective Generator Learning Algorithm



Learning Iteration (Single Step)

1. Update a selection function $\hat{s}_i : Q \mapsto \{0,1\}$

$\widehat{U}_i \leq$ Target FDR-E?

2. Select $D_{\text{cal}} := D_E \cup D_U$

D_E $\{(Q,A,e)\}$ $\xrightarrow{\hat{s}_i(Q)=1?}$ $\widehat{D}_E$ $\{(Q,A,e)\}$

D_U $\{(Q,A)\}$ $\widehat{D}_U$ $\{(Q,A)\}$

3. Train pseudo-labeler and pseudo-label unlabeled data.

$\widehat{D}_E$ $\{(Q,A,e)\}$ $\xrightarrow{\text{Train pseudo-labeler } \hat{e}}$ $\hat{e} : A \mapsto \{0,1\}$

$\widehat{D}_U$ $\{(Q,A)\}$ $\rightarrow \hat{e} \rightarrow$ $\widehat{D}_U$ $\{(Q,A,\hat{e})\}$

Pseudo-label unlabeled data

4. Compute the FDR-E upper bound $\widehat{U}_i$

a. Aggregate $\widehat{D}_E$, $\widehat{D}_U$

$\widehat{D}$ $\{(Q,A,\tilde{e}) \mid \tilde{e} = e \text{ if } \widehat{D}_E \;\&\; \tilde{e} = \hat{e} \text{ o.w.}\}$

b. Compute the FDR-E upper bound $\widehat{U}_i$

$\widehat{D}$ $\{(Q,A,\tilde{e})\}$ $\xrightarrow[\text{Mislabeling Error}]{\text{Binomial tail bound}}$ $\widehat{U}_i$

Figure 2: A single iteration of the proposed semi-supervised learning algorithm, SGen$^{\text{Semi}}$.

Theoretical Result

Theorem. Controllability Guarantee on the FDR-E

For any LLMs and downstream language generation tasks, the following model-agnostic and task-free controllability guarantee holds:

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}\left\{\overbrace{\mathbb{P}_{\mathcal{D}}(\underbrace{\mathbf{A} \notin E_{\text{true}}(\mathbf{A}_{\text{true}})}_{\mathbf{A}\text{ is ``wrong''}} \mid \underbrace{\hat{S}(\mathbf{Q}) \neq \texttt{IDK}}_{\text{Do not abstain}})}^{\text{FDR-E}} \leq \hat{U}\right\} \geq 1 - \delta,$$

where δ is the confidence level and $(\hat{s}, \hat{U})$ is the algorithm output.

Experimental Result: Benefit of Textual Entailment

- Our entailment-based learning metric shows better selection efficiency.
  - Selection efficiency: The proportion of non-abstained samples

Q	Who is the actor that plays Draco Malfoy?	When did the movie Benjamin Button come out?
A$_{\text{true}}$	Thomas Andrew Felton plays Draco Malfoy in the Harry Potter movies.	The movie Benjamin Button come out December 25, 2008.
$\hat{\mathbf{A}}$	The actor who plays Draco Malfoy is Tom Felton. (correct)	The Curious Journey of Benjamin Button was released in 2008. (correct)
EM (Baseline)	rejected	rejected
Textual Entailment (Ours)	accepted	accepted

Thank You!