Paper Reading

Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. (*ICLR 2023*)

LLM Group: XUE Boyang

Nov. 20th, 2023

Introduction

- **Motivation**
  - Why do we need uncertainty estimates?

    Uncertainty estimation: how we can **trust** the natural language **generations** of LMs (Fig. 1).
  - What's the challenges in uncertainty estimation of LLM generations?

    LMs output token-likelihoods which represents lexical confidence, lacking of **meanings of sentences**.

    Measuring uncertainty in sentences is challenging because of '**semantic equivalence**'—different sentences can mean the same thing (Fig. 2).

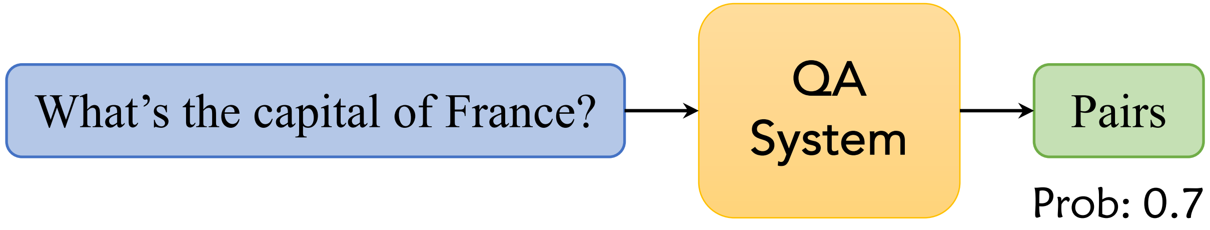


Fig. 1

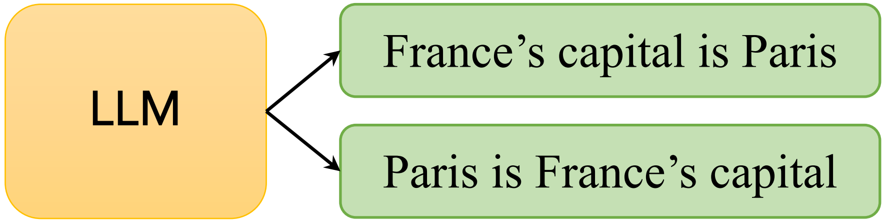


Fig. 2

Introduction

- **Framework**
  - **Background** on uncertainty estimation.
  - **Challenges** in uncertainty estimation for NLG.
  - **Methodology**: Semantic uncertainty.
  - **Experiments**: Empirical evaluation on free-form QA tasks.
- **Contributions**
  - Explain **why uncertainty in NLG is different** from other settings (Challenges in details).
  - Introduce **semantic entropy** —— uncertainty measure over semantically-equivalent samples.
  - Present how to balance the trade-off between sampling **diverse and accurate** generations.

Background

- **Uncertainty Estimation: Prediction Entropy**

$$PE(x) = H(Y|x) = -\int p(y|x) \ln p(y|x)\, dy$$

- **Uncertainty Type**
  - Aleatoric Uncertainty: **inherent variation** of inputs, parameters, or data distributions.
  - Epistemic Uncertainty: results from **missing information**.
- **Direct Application of LM Uncertainty**
  - Product of the conditional probabilities of new tokens given past tokens $\log p(\boldsymbol{s}|x) = \sum_i \log p(s_i|\boldsymbol{s}_{<i}, x)$.
  - Prompt the generative LLM itself to estimate its own uncertainty.

Challenges

- **Semantic Equivalence in Language Outputs**
  - Linguists distinguish text's meaning—its semantic content—from its syntactic and lexical form.

		Equivalence		
Sentence A	Sentence B	Lexical	Syntactic	Semantic
Paris is the capital of France.	Paris is the capital of France.	✓	✓	✓
	Berlin is the capital of France.		✓	
	France's capital is Paris.			✓

  - A system can be reliable even with many different ways to say the same thing but answering with inconsistent meanings shows poor reliability.
  - For the space of **semantic equivalence classes** $\mathcal{C}$ the sentences in the set $c \in \mathcal{C}$ that shares some meaning

$$p(c|x) = \sum_{\mathbf{s}\in c} p(\mathbf{s}|x) = \sum_{\mathbf{s}\in c} \prod_i p(s_i|\mathbf{s}_{<i}, x)$$

Challenges

- **Sampling The Extremely High-Dimensional Language-Space**
  - The output-space of natural language has $\mathcal{O}(|\mathcal{T}|^N)$ dimensions.
  - Lack a normalized probability density function over sentences.
- **Variable Length Generations**
  - The joint likelihood of a sequence of length N **shrinks exponentially in *N***, so longer sentences tend to contribute more to entropy.
  - Although length-normalizing the log-probabilities, sometimes **longer sentences may well be usually more uncertain** (when the goal is to exactly match a typically short reference answer).

Methodology

- **Semantic Uncertainty**
  - **Key Point**: Uncertainty over **meanings** is more important for most situations than uncertainty over the exact tokens used to express those meanings. (token event-space → semantic event-space)
  - A novel uncertainty estimation algorithm for **semantic entropy**
    - **a) Generation**: Sample M sequences $\{s^{(1)}, \ldots, s^{(M)}\}$ from the predictive distribution of a LLM given x.
    - **b) Clustering**: Cluster the sequences which mean the same thing by bi-directional entailment algorithm.
    - **c) Entropy Estimation**: Approximate semantic entropy by summing probabilities that share a meaning and compute resulting entropy.
  - **How The Semantic Entropy Addresses The Challenges of NLG**
    - Address the semantic invariance by converting uncertainty estimation into meaning-space.
    - Address unequal token importance by reducing the effect of the likelihoods of unimportant tokens.

Methodology

- **Semantic Uncertainty**
  - **Clustering by Semantic Equivalence**
    - Using bi-directional entailment (NLI model: Deberta-large model fine-tuned on the NLI data set MNLI), **if and only if** they entail (i.e. logically imply) each other given context.

Paris is France's capital $\nLeftrightarrow$ Paris

What's the capital of France? Paris is France's capital $\Leftrightarrow$ What's the capital of France? Paris

  - **Clustering by Semantic Equivalence**
    - **Computational Cost** - Even though requires requires $\binom{M}{2}$ comparisons in the worst-case, the computational cost is small compared to the cost of generating sequences (Generally $M < 20$ and Deberta-Large model only has 1.5B parameters).

Methodology

- **Semantic Entropy**
  - Compute the semantic entropy (SE) as the entropy over the meaning-distribution.

$$SE(x) = -\sum_{c} p(c|x) \log p(c|x) = -\sum_{c} \left(\left(\sum_{\mathbf{s} \in c} p(\mathbf{s}|x) \right) \log \left[\sum_{\mathbf{s} \in c} p(\mathbf{s}|x) \right] \right)$$

  - When some of the answers are semantically equivalent ("Paris" and "It's Paris") the semantic entropy does a better job of capturing the actually low uncertainty.

Answer $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5	0.5
Rome	0.4	0.4
London	0.1	0.1
Entropy	0.31	0.31

Answer $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5	0.9
It's Paris	0.4	
London	0.1	0.1
Entropy	0.31	0.16

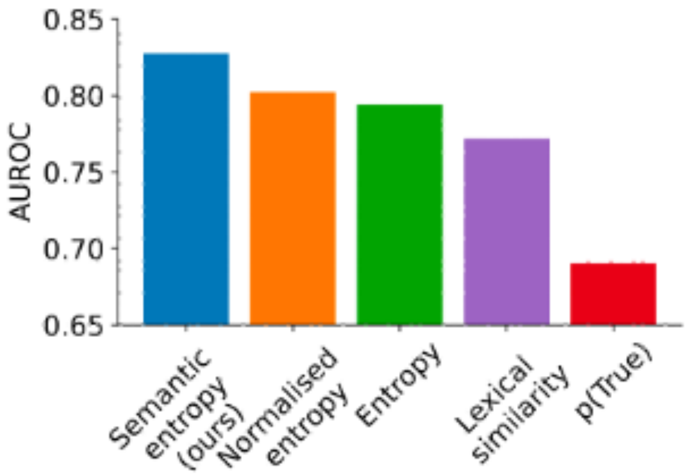
Experiments

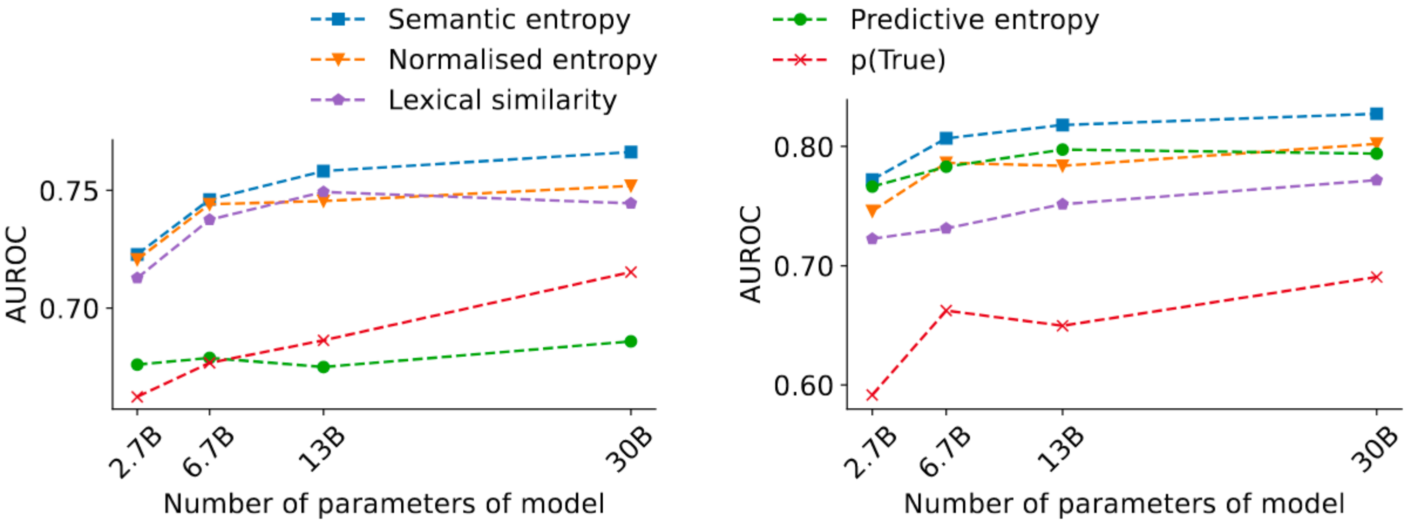
- **Experimental Settings**
  - **Performance Evaluation:** Receiver operator characteristic curve (AUROC) - The correct answers has a higher uncertainty score —— whether to trust an answer to a question.
  - **Models:** GPT-like OPT varying from 2.7B, 6.7B, 13B and 30B parameters.
  - **Datasets:** CoQA (Open-book QA); TriviaQA (Closed-book QA).

    Correctness of QA evaluation: fuzzy matching criterion $\mathcal{L}(\mathbf{s}, \mathbf{s}') = \mathbf{1}(\mathrm{RougeL}(\mathbf{s}, \boldsymbol{s}') > 0.3)$
  - **Baselines Methods:**
    - **Predictive entropy:** The predictive entropy of the output distribution $PE(x) = H(Y|x)$.
    - **Length-normalized predictive entropy:** Divides the joint log-probability by the sequence length.
    - ***p*(True):** By 'asking' the model if its answer is correct and measuring the probability of being `True`.
    - **Lexical similarity:** Average answer similarity of the answer set $\mathbb{A}$: $\frac{1}{C}\sum_{i=1}^{|\mathbb{A}|}\sum_{j=1}^{|\mathbb{A}|} \mathrm{RougeL}(s_i, s_j)$.

Experiments

- **Semantic Entropy Uncertainty**
  - Semantic entropy predicts model accuracy better than baselines on the free-form question answering data set TriviaQA and CoQA.

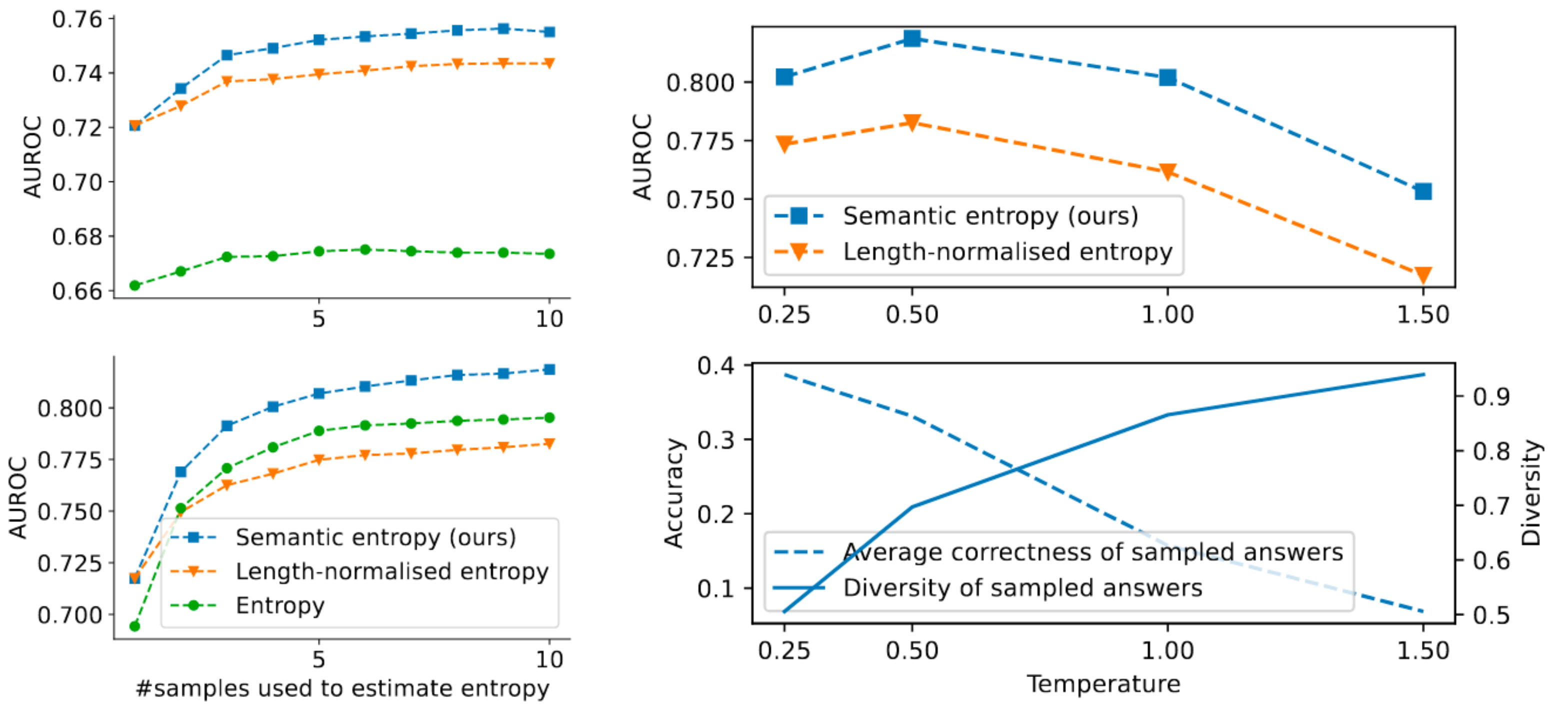




(a) CoQA (b) TriviaQA

Experiments

- **Hyperparameters for Effective Sampling**
  - Temperature: average correctness vs. diversity of samples.



Discussion & Conclusion

- Many natural language problems display a crucial invariance: sequences of distinct tokens mean the same thing.
- Introduce semantic entropy —— the entropy of the distribution over meanings rather than sequences—and show that this is more predictive of model accuracy on QA than strong baselines.
- For semantic entropy, this work introduces a novel bidirectional entailment clustering algorithm which uses a smaller natural language inference model.
- Prospective 1 - semantic equivalence can pave the way towards progress in settings like summarization where correctness requires more human evaluation.
- Prospective 2 - semantic likelihoods could also be extended to other tools for probabilistic uncertainty like mutual information, potentially offering new strategies for NLG uncertainty.

Thank you !