Synthesize Training Data

Introduction

for TvD, Teaching via Data
fine-tune lightweight student model for running system

Limitations

1. 对一些像 NER 的 tasks（ICST, NER），不能生成 token-level label (slot tags) [1]
    - before：use a separate model to do work alignment. **Res:** noise ↑
2. scoped，不能 control 它的 output
3. LLM 都是在真实的大型数据集上进行训练。由 LLM 合成的 generated data 受到 origin dataset 的影响，進一步 worse performance of the fine-tune model. [2]
    1. limit the diversity
       size of vocabulary of synthesing's << size of ground truth's
    2. inherit systematic biases
       词的 frequency 两极分化更严重
       -> sol:

Approaches with gpt

Prompt Format

HTML/XML [1]

- target: control output

> [CLM] Sentence: example1
 Translation in French: ...
 Sentenc:target
 Translation in French:
>
> > [CLM] Sentence: example1
> > Translation in French: ...
> > Sentenc:target
> > Translation in French:

common

for multi-languages

- Back-translate
  English -> French ->English

> Senetence:...
> Translation in English:

- Paraphrase

> another way to say "...":

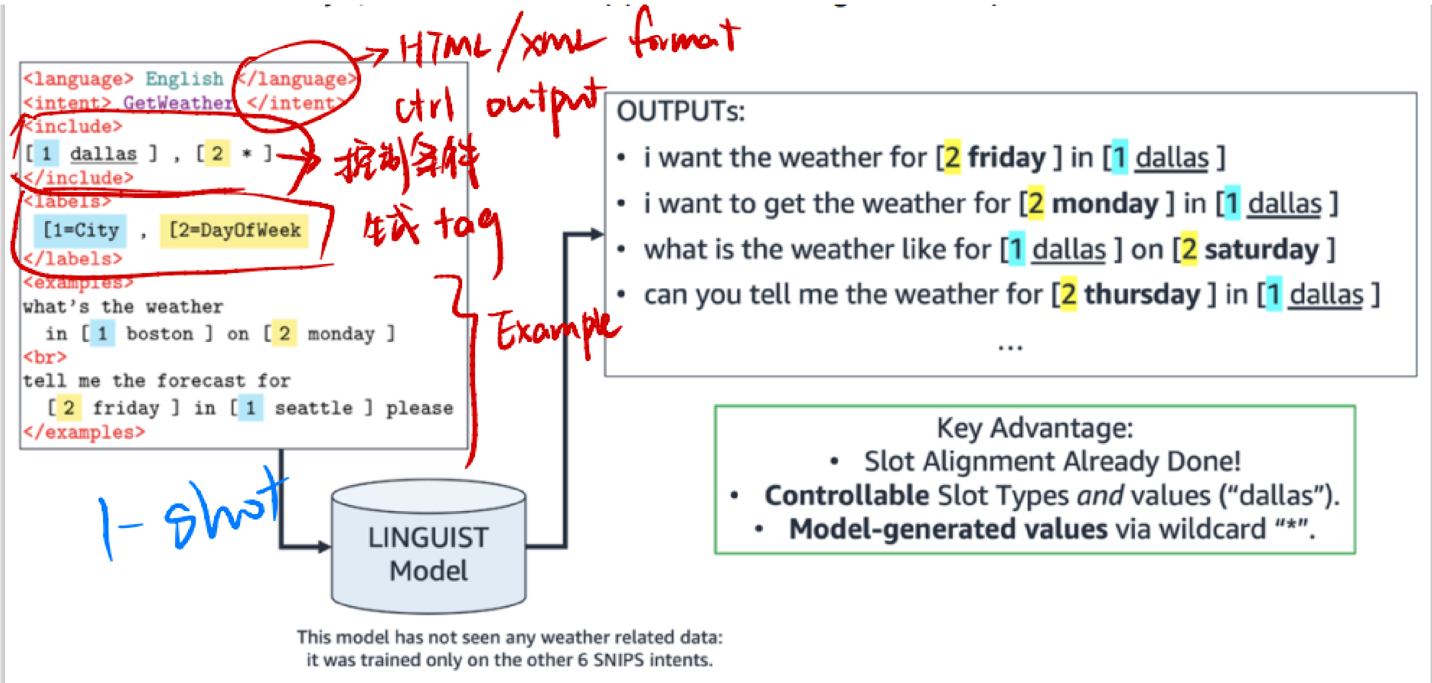
for slot tags

LINGUIST[1]

Language Model **In**struction Tuning to **G**enerate Annotated **U**tterances for **I**ntent Classification and **S**lot **T**agging

> **task: joint intent classification and slot tagging = IC+ST**
>
> outperforms state-of-the-art baselines like translation and paraphrasing

introduce an **output format** with **brackets and numbers** that enables the model to produce synthetic data with the slots already tagged.



Summary

1. XML format
2. 生成的 textdata 包括了 `[label, entity]`
3. 最好是 few shots，0-shot $\Longrightarrow$ more noise
4. wildcard instruction `*` 自由发挥，which did not appear in the original examples
5. 可以改 language

CLASP[1,7]

Few-shot **C**ross-**L**ingual Data **A**ugmentation for **S**emantic **P**arsing

task: few-shot multilingual semantic parsing, SP

machine translation

CLASP	replacing the slots with other values		generate both the parse and the text	
	只要句子		more flexibility，句子和解析都要	
Output examples	It is a **panda** from **China**.		(Introduction (Animal panda) (Nationality n China))=> It is a **panda** from **China**.	
approaches	**RS** replace slots	**TS** translate slots	**GB** generate both	**TB** translate both
	from a catalogue of options	via translation to a new language	in the same language	in a new language
	English	for others	English	for others

Summary

1. multi-language
2.

for diversity

AttrPrompt [2,8]

AttrPrompt github

- origin **SimPrompt**
  simple class-conditional prompt

Table 1: Prompt template for the NYT news dataset.

Method	Prompt
SimPrompt	Suppose you are a news writer. Please generate a {topic-class} news in NYT.
AttrPrompt	Suppose you are a news writer. Please generate a {topic-class} news in NYT following the requirements below: 1. Should focus on {subtopic}; 2. Should be in length between {length:min-words} and {length:max-words} words; 3. The writing style of the news should be {style}; 4. The location of the news should be in {location}.

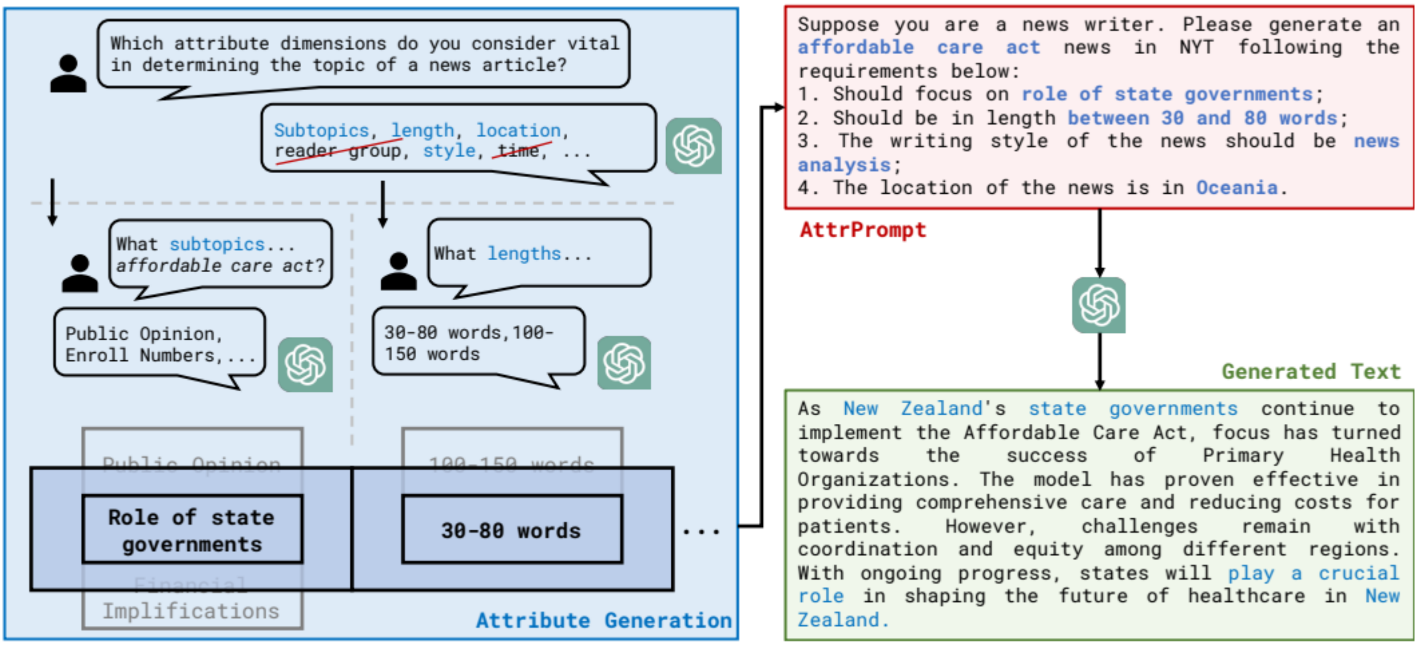


Figure 1: The overall workflow of AttrPrompt.

process

For a given classification task

1. initail step
   identify attribute dimensions and their corresponding attribute values in an interactive, semi-automated process facilitated by the LLM.
   1. use `gpt` help establish both attribute dimensions and attribute values.

      > Which attribute dimensions do you consider vital in determining the topic of a news article?”
      >
      > > “subtopics, length, location, reader group, style, time”

   2. adopt the human-ai collaboration scheme to interactively select the attribute dimensions of **the highest quality** that best suit the dataset. **人为地选择** Best Top-N attributes.
   3. generate values corresponding to selected attributes similarly
      !!! quote ""
      List 10 diverse subtopics for {class_name} news on NYT.

atrrs	class-depe	class-indepe
	need value filtering	remain unchanged across different classes
examples	subtopic	length

4. **Class-Dependent Attribute Value Filtering, CAF**
   - target： avoid ambiguity and potential connections to multiple classes

   对 gpt 根据任务给出的 Top-5 个相似 classes，$\forall \text{value} \in \text{class}$进行询问：是否和别的类相关。相关就 remove.

   > List 5 similar classes for {class-name} news on NYT. The set of classes is listed as: {[a list of class-names]}.
   >
   > if the answer is positive which indicates a potential ambiguity, we remove that attribute value for the specific class.

2. generate diverse prompts by combining attributes randomly.

> Suppose you are a review writer. Please write a review for {product-class} product in Amazon following the requirements below:
>
> 1. The review should be about the product of {subtopic};
> 2. The brand for the product should be {brand};
> 3. Should be in length between {length:min-words} and {length:max-words} words;
> 4. Should describe the usage experience {usage-experience}
> 5. The writing style of the review should be {style};
> 6. the review must be relevant to {product-class} and irrelevant to: {similar-class}.

> **Summary**
>
> 设计一种使用不同的attributed prompt（带有特征的prompt）生成训练数据的方法（比如限制长度、风格）
> 展望：
>
> - exploring automated or semi-automated methods for identifying high-quality attribute dimensions and values
> - Domain Limitation 只在 text classification 中
> - 生成的数据继承了 LLM 的 hallucination 幻觉问题(生成的文本中在语义或句法上看似合理但实际上不正确或无意义的错误)

increasing diversity wihile maintain accuracy [3]

> Write a movie review (text type) to cover all following elements
> Elements: positive sentiment (label)
> Movie review (text type): "This is a great movie"

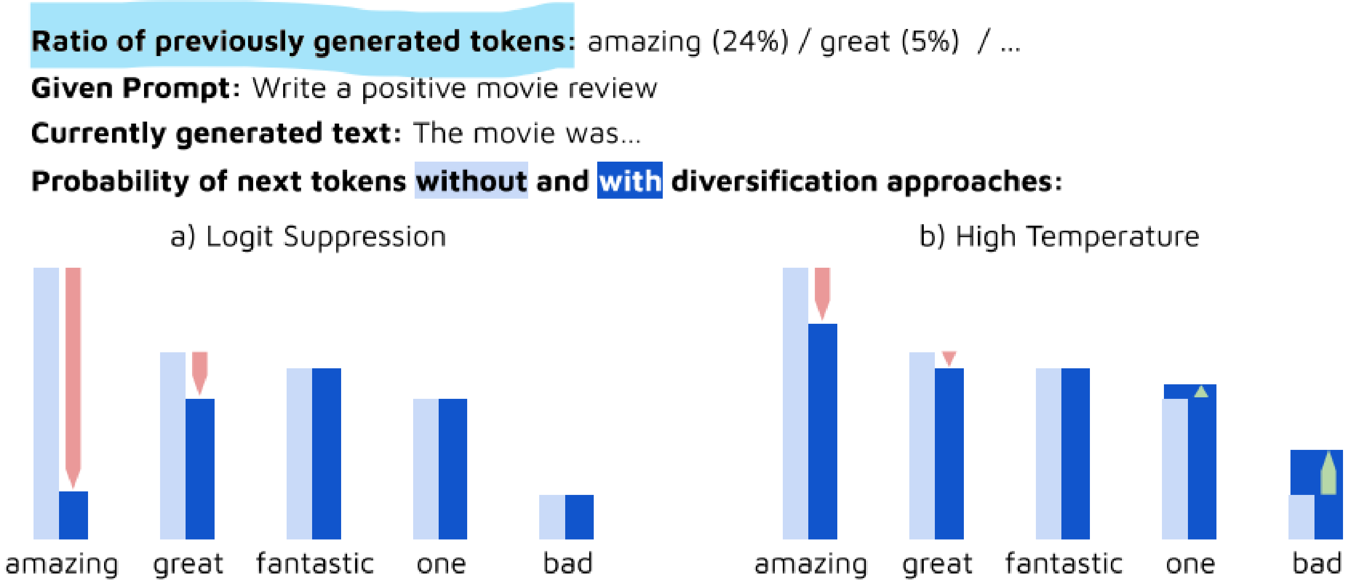


Figure 1: Examples of Diversification Approaches.

```
openai.Completion.create(
    engine='davinci',
    prompt='q: What is the capital of france?\na:',
    logprobs = 5,  # TopN the natural log of the probability
    stop = '\n',
    temperature=0,
    logit_bias={Token_ID:logprob} # map: {6342:-1, 1582:-10}
    )

"""

- logit_bias:
    Accepts a json object that
    maps tokensto an associated bias value from -100 to 100
    token_ID: in the GPT tokenizer
"""
```

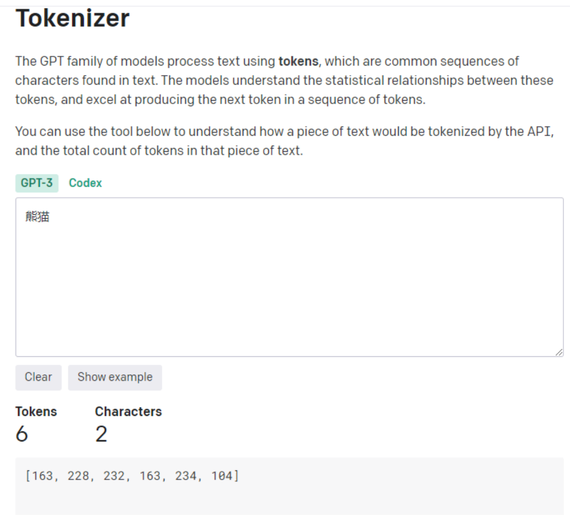
logit supression[9]

OpenAI API

- **Logit bias** parameter
  GPT3 的一个很有用的参数。通过 modify the likelihood of tokens 控制 token in GPT Tokenizer(convert text to token IDs) 的生成，unwanted tokens ↓， wanted tokens ↑.[9] **bias 会直接加到 gpt 生成的 logprob 上。**

  $$\text{logprob}\begin{cases} -1|1 & \uparrow\downarrow \text{ the likelihood of tokens} \\ -100|100 & \text{禁止或者直接指定} \end{cases}$$

  create-logit_bias in openAI Docs

> **Question**
>
> - 中文？会有在那50000
>
> 
>
>
> - only 100 tokens for logit biasing

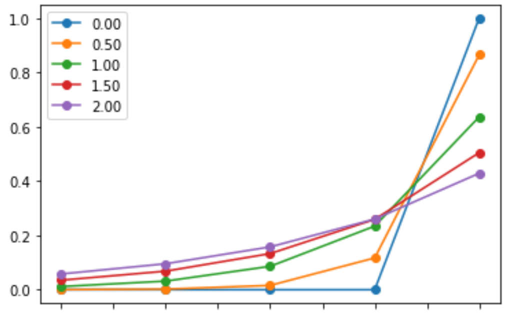
- how gpt generate tokens
  When run, GPT-3 takes the prompt and predicts the probabilities of the token that is going to occur next. [9]
  Rather than the percentages, logprobs is used. $\text{logprob} \to 0 \iff \text{prob} \uparrow$. [9]

> Specifically, for the logit bias weights, we multiplied the token appearance ratio (in percentage) by -7.5 while capping the minimum weight at –7.5.[9]
>
> 1. 统计 tokens 的 frequency
> 2. logprob = 出现的 freq * -7.5（也就是说最低不可能超过 -7.5

temperature-based sampling[5,6]

温度 采样受到统计热力学的启发，其中高温意味着更有可能遇到低能态。在概率模型中，logits 扮演着能量的角色，我们可以通过将 logits 除以温度来实现温度采样，然后将其输入到 softmax 中并获得采样概率

```
>>> import torch
>>> import torch.nn.functional as F
>>> a = torch.tensor([1,2,3,4.])
>>> F.softmax(a, dim=0)
tensor([0.0321, 0.0871, 0.2369, 0.6439])
>>> F.softmax(a/.5, dim=0)
tensor([0.0021, 0.0158, 0.1171, 0.8650])
>>> F.softmax(a/1.5, dim=0)
tensor([0.0708, 0.1378, 0.2685, 0.5229])
>>> F.softmax(a/1e-6, dim=0)
tensor([0., 0., 0., 1.])
```



> 0.3, 0.7, 0.9, and 1.3[3]

- create-temperature in openAI Docs

$$\text{temperature} \in [0,2] \begin{cases} \uparrow \geq 0.8 & \text{more random} \\ \downarrow \leq 0.2 & \text{more focused and deterministic} \end{cases}$$

> **more about sampling**
>
> The Curious Case of Neural Text Degeneration
> We generally recommend altering this or top_p but not both.

	diversify text generation		
approaches	logit suppression	temperature sampling	
	minimizes the generation of that have already been frequently generated. 減少已頻繁生成的	flattens the token sampling probability 展平概率	
	the **minimum weight** at –7.5	four temperature values, 0.3, 0.7, 0.9, and 1.3	
process	1. logs the frequency 2. Logprob = -7.5 * freq Every time we complete a single generation iteration, we recorded the frequency of tokens.	seeding examples	zero-shot generation
		1. 从原始数据集 均匀采样 as an **initial example pool**, with a balanced number of labels.	1. 第一轮就是 a zero-shot generation. 第一轮后都每个label 都有 instance.
		2. 新生成的 instance 加入到 它的 **example pool**	
		3. 每一次都从**example pool** 拿 one example for each label	
Res	Diversity ↑ accuracy ↓ 根据每个 task，有不同的比较，参见附录C		
	human utterance to a chatbot	模型有比 GPT few-shot 好	

	accuracy	
approaches	**LR** label replacement	**OOSF** out-of-scope filtering
	correct misaligned labels	remove instances that are out of the user's domain of interest or to which no considerred label applies.
Res	Good √	Not good ×
	Too complex	

metrics

the quality of synthesized training data [4]

- fidelity
  how closely the synthetic data matches with the original data
- utility
  synthetic data performs well on common tasks in data science
- privacy
  protect sensitive information，此處沒管
- diversity

- 【fidelity】

	Evaluating fidelity			
	Statistical comparisons	**Histogram similarity score**	**Mutual information score**	**Correlation score**
	mean, median, standard deviation, number of distinct values, ...	how similar the distribution of each feature (or category of data)	how dependent two features are on each other	how well relationships between two or more columns of data
	for each category of data	→1, similarity ↑		
Notes	we can look at autocorrelation and partial autocorrelation scores to see how well the synthetic data has preserved significant correlations from the original dataset			

- 【utility】**Feature importance score**[4]
  檢查順序
- 【utility】**QScore? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?** :
  This score is used to check if a model trained on synthetic data will give the same results as a model trained on original data. It does this by running random aggregation-based queries on both datasets and comparing the results. If the results are similar, it means the synthetic data has good utility.
- 【utility】the accuracies of models [3]

> We compared the accuracies of models trained with generated data to 1) models trained with oracle datasets (oracle model) and 2) GPT-3's few-/zero-shot classifications

- label accuracy[3]
  the accuracy of the alignment between the generated texts and the specified labels
- 【diversity】average mean pairwise distances[3]
  - Remote-Clique metric cox2021directed, which is the average mean pairwise distances. S
  - we embedded generated data with BERT devlin2019bert, then calculated the distances
- 【utility】similarity between dataset [3]
  We also measured the similarity of the generated dataset to the oracle dataset with the average mean pairwise distances between the two. For similarity, we also used BERT to embed the generated texts.

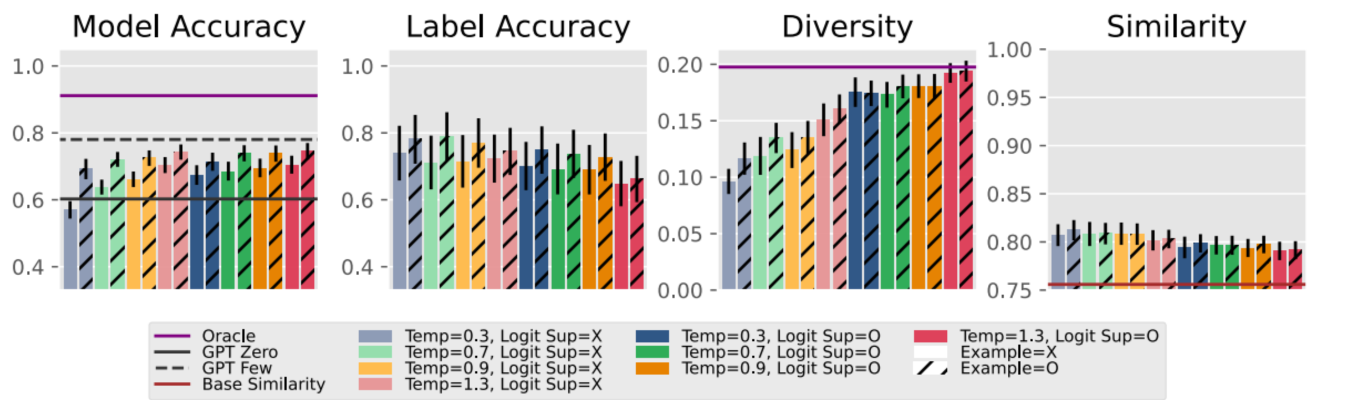


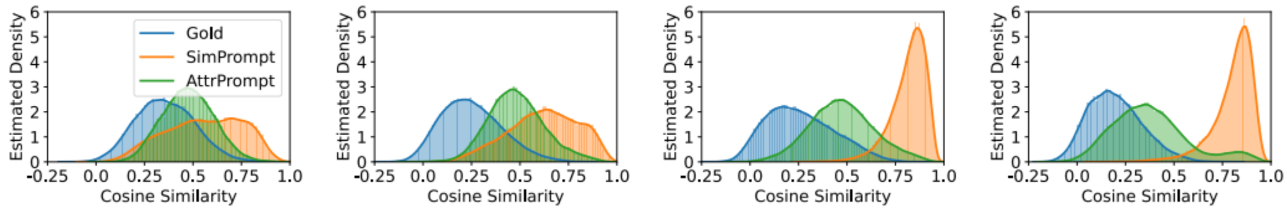
Figure 2: Impact of logit suppression and high temperatures on model accuracy, label accuracy, diversity, and similarity to the oracle dataset, averaged across eight tasks. Bars without hatches start generation without examples while those with hatches start with few-shot generation. Throughout this paper, error bars indicate 95% confidence interval.

- 【diversity】**vocalbulary size** for lexical diversity of datasets[2]

Table 5: Comparison of the vocabulary size of different datasets.

Method	NYT		Amazon		Reddit		StackExchange	
	All	Class Avg.	All	Class Avg.	All	Class Avg.	All	Class Avg.
Gold	70.8k	11.3k	44.7k	6.64k	50.8k	4.62k	52.3k	3.60k
SimPrompt	20.6k	3.13k	11.6k	2.50k	19.9k	3.06k	13.3k	2.20k
AttrPrompt	21.4k	3.50k	14.0k	2.76k	25.4k	3.64k	17.8k	2.93k

- 【diversity】**cosine similarity** for the diversity from the semantic perspective[2]
  - the cosine similarity is calculated based on the embedding of Sentence-BERT Reimers and Gurevych
  - cosine similarity ↓ diversity ↑



(a) NYT (b) Amazon (c) Reddit (d) StackExchange

Figure 2: The distribution of cosine similarity of text pairs sampled from the same class.

- 开销

> attributed prompt只需要simple prompt 5%的开销（主要用于query chatgpt）就可以达到和后者一样的效果。

Reference

1. Using large language models LLMs to synthesize training data
2. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias
3. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions
4. HOW TO USE LLMS IN SYNTHESIZING TRAINING DATA
5. temperature-based sampling（基于温度系数的采样）
6. How to sample from language models
7. CLASP: Few-Shot Cross-Lingual Data Augmentation for Semantic Parsing
8. AttrPrompt：一个关于多样性与偏见的故事
9. Controlling GPT-3 with Logit Bias
10. The need for sampling temperature and differences between whisper, GPT-3, and probabilistic model's temperature

未完待续