可解释性自然语言处理

Li Qintong

2022/03/10

第一步，收集人工标注的解释性文本

- 作为数据增强改进在预测任务上的效果
- 作为监督数据训练模型对预测结果做解释
- 作为真实标签评估模型的生成的解释

Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. 2021.

文本类解释的类型

Instance	Explanation
Premise: A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. *Label:* contradiction	`(highlight)` *Premise:* A white race dog wearing the number eight runs on the track. *Hypothesis:* A white race dog runs around his yard. `(free-text)` A race track is not usually in someone's yard.
Question: Who sang the theme song from Russia With Love? *Paragraph:* ...The theme song was composed by Lionel Bart of Oliver! fame and sung by Matt Monro... *Answer:* Matt Monro	`(structured)` *Sentence selection:* (not shown) *Referential equality:* "the theme song from russia with love" (from question) = "The theme song" (from paragraph) *Entailment:* X was composed by Lionel Bart of Oliver! fame and sung by ANSWER. ⊢ ANSWER sung X

explanations are implicitly or explicitly designed to answer the question "why is [input] assigned [label]?".

- Highlights
  - Compactness
  - Sufficiency
  - Comprehensiveness / selected
- Free-text explanations
  - not constrained to the words or modality of the input instance
  - Expressive / readable
- Structured explanations
  - there may be constraints placed on the explanation writing process, such as the required use of specific inference rules.
  - dataset-specific designs

Dataset	Task	Collection	# Instances
Jansen et al. [56]	science exam QA	authors	363
Ling et al. [76]	solving algebraic word problems	auto + crowd	~101K
Srivastava et al. [115]*	detecting phishing emails	crowd + authors	7 (30-35)
BABBLELABBLE [46]*	relation extraction	students + authors	200‡‡
E-SNLI [20]	natural language inference	crowd	~569K (1 or 3)
LIAR-PLUS [4]	verifying claims from text	auto	12,836
COS-E v1.0 [100]	commonsense QA	crowd	8,560
COS-E v1.11 [100]	commonsense QA	crowd	10,962
ECQA [2]	commonsense QA	crowd	10,962
SEN-MAKING [124]	commonsense validation	students + authors	2,021
CHANGEMYVIEW [10]	argument persuasiveness	crowd	37,718
WINOWHY [144]	pronoun coreference resolution	crowd	273 (5)
SBIC [111]	social bias inference	crowd	48,923 (1-3)
PUBHEALTH [64]	verifying claims from text	auto	11,832
Wang et al. [125]*	relation extraction	crowd + authors	373
Wang et al. [125]*	sentiment classification	crowd + authors	85
E-δ-NLI [18]	defeasible natural language inference	auto	92,298 (~8)
BDD-X†† [62]	vehicle control for self-driving cars	crowd	~26K
VQA-E†† [75]	visual QA	auto	~270K
VQA-X†† [94]	visual QA	crowd	28,180 (1 or 3)
ACT-X†† [94]	activity recognition	crowd	18,030 (3)
Ehsan et al. [34]††	playing arcade games	crowd	2000
VCR†† [143]	visual commonsense reasoning	crowd	~290K
E-SNLI-VE†† [32]	visual-textual entailment	crowd	11,335 (3)‡
ESPRIT†† [101]	reasoning about qualitative physics	crowd	2441 (2)
VLEP†† [72]	future event prediction	auto + crowd	28,726
EMU†† [27]	reasoning about manipulated images	crowd	48K

Table 4: Overview of EXNLP datasets with **free-text explanations** for textual and visual-textual tasks (marked with †† and placed in the lower part). Values in parentheses indicate number of explanations collected per instance (if > 1). ‡ A subset of the original dataset that is annotated. ‡‡ Subset publicly available. $*$ Authors semantically parse the collected explanations.

1. highlight input words and then formulate a free-text explanation from them, to control quality.

2. template-like explanations are discarded because they are deemed uninformative.

Takeaway:

1. study how people **define and generate** explanations for the task before collecting free-text explanations
2. explanations are naturally structured, **embrace the structure**.
3. No all-encompassing definition of explanation

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Table 1: Examples from our CoS-E dataset.

Explain Yourself! Leveraging Language Models for Commonsense Reasoning. 2019.

Dataset	Task	Explanation Type	Collection	# Instances
WORLDTREE V1 [57]	science exam QA	explanation graphs	authors	1,680
OPENBOOKQA [81]	open-book science QA	1 fact from WORLDTREE	crowd	5,957
Yang et al. [135]††	action recognition	lists of relations + attributes	crowd	853
WORLDTREE V2 [132]	science exam QA	explanation graphs	experts	5,100
QED [70]	reading comp. QA	inference rules	authors	8,991
QASC [61]	science exam QA	2-fact chain	authors + crowd	9,980
EQASC [58]	science exam QA	2-fact chain	auto + crowd	9,980 (∼10)
+ PERTURBED	science exam QA	2-fact chain	auto + crowd	n/a‡
EOBQA [58]	open-book science QA	2-fact chain	auto + crowd	n/a‡
Ye et al. [138]*	SQUAD QA	semi-structured text	crowd + authors	164
Ye et al. [138]*	NATURALQUESTIONS QA	semi-structured text	crowd + authors	109
R^4C [53]	reading comp. QA	chains of facts	crowd	4,588 (3)
STRATEGYQA [41]	implicit reasoning QA	reasoning steps w/ highlights	crowd	2,780 (3)
TRIGGERNER	named entity recognition	groups of highlighted tokens	crowd	∼7K (2)

place constraints (e.g., phrase) on the textual explanations that annotators can write,

construct a chain of facts that detail the reasoning steps

Table 5: Overview of EXNLP datasets with **structured explanations** (§5). Values in parentheses indicate number of explanations collected per instance (if > 1). †† Visual-textual dataset. ∗ Authors semantically parse the collected explanations. ‡ Subset of instances annotated with explanations is not reported. Total # of explanations is 855 for EQASC PERTURBED and 998 for EOBQA.

Dataset-specific forms

第二步，评估解释性文本

- Head-to-head evaluations：对同一数据集实例在不同条件下生成的两种解释进行直接比较
- Understand the fine-grained aspects：收集每个解释的绝对Likert-scale评分

Reframing Human-AI Collaboration for Generating Free-Text Explanations. 2021.

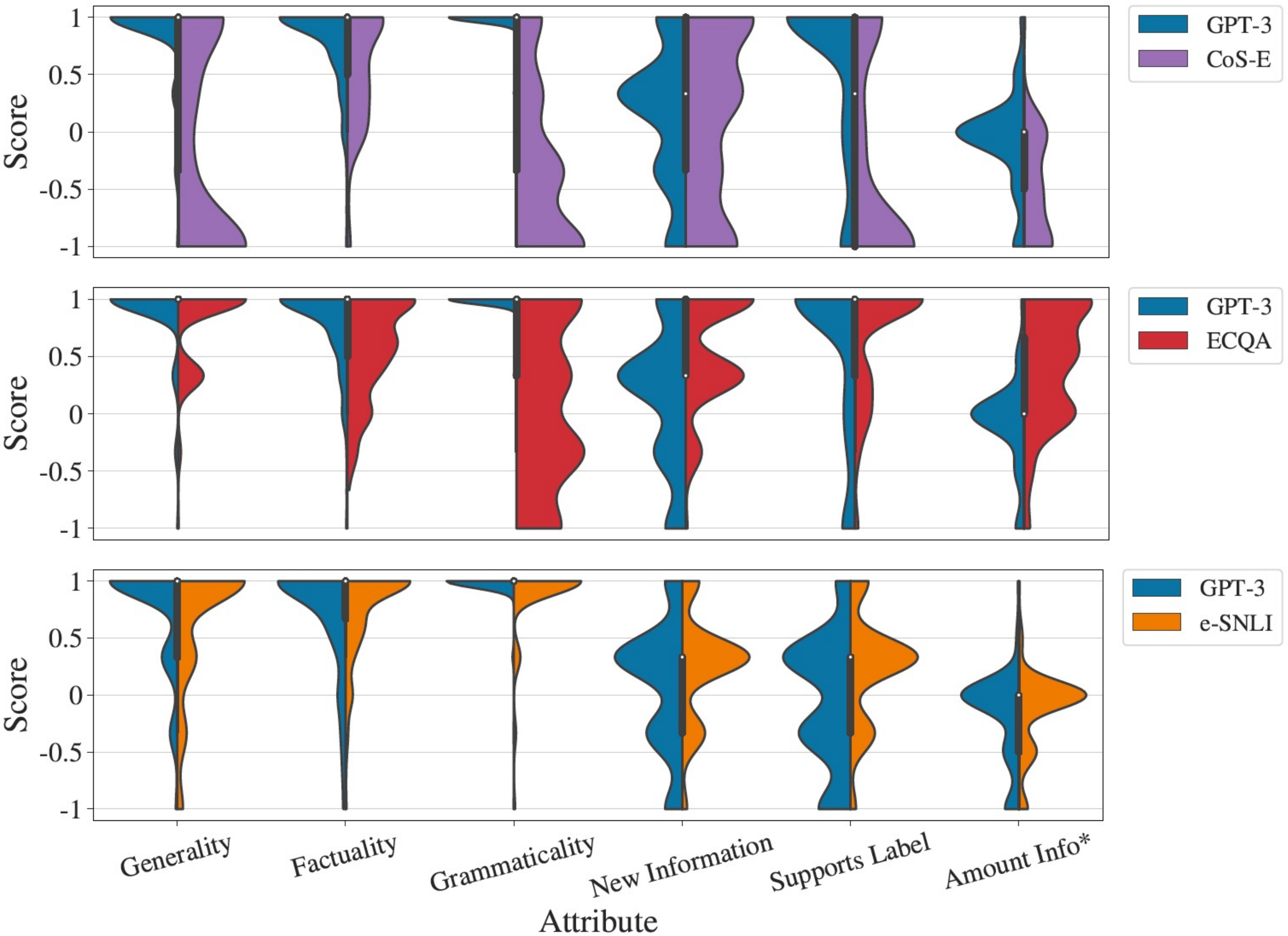
Two dimensions:

1. Surface-level features:
   generality
   grammaticality
   factuality

2. Explanation quality:
   <u>New information</u>
   <u>Support the label</u>
   The information is sufficient

Ensuring explanations are not vacuous and are on-topic.

In an ideal setting, machine-generated explanation quality should be unambiguous enough to elicit high scores across a group of annotators.



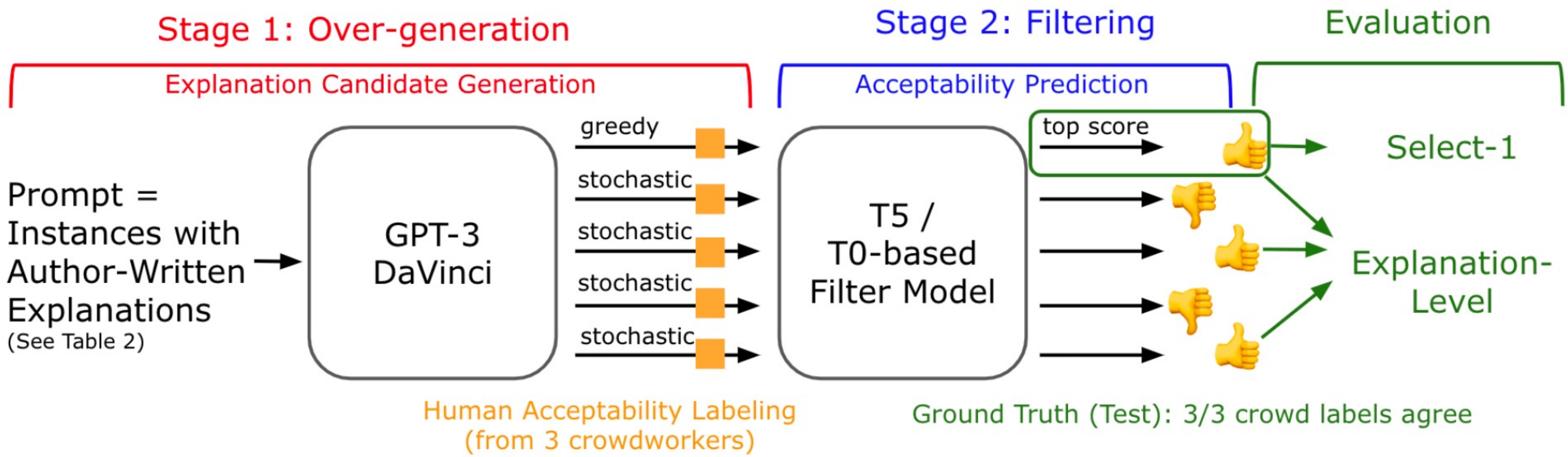
第三步，设计方法生成可解释性文本

- 基于 Prompt + large-scale language model 的方法
- 引入外部知识图谱

Reframing Human-AI Collaboration for Generating Free-Text Explanations. 2021.
Event Transition Planning for Open-ended Text Generation. 2022.

人工和模型合作改进解释文本质量:

将标注人员的职责由标注可解释性文本（生成任务）降低难度转变为判断文本的可接受性（二分类任务）。



Reframing Human-AI Collaboration for Generating Free-Text Explanations

1. In-context learning

We prompt the model with several (question, answer and explanation) triplets, followed by an unexplained question-answer instance for which we expect the model to generate an explanation, without updating any parameters

115 randomly sampled train instances to create our prompts;
Each prompt consists of 8-24 randomly selected examples from this set.

"A dog cannot carry something while asleep".

Let's explain classification decisions.
A young boy wearing a tank-top is climbing a tree.
question: A boy was showing off for a girl.
true, false, or neither? neither
why? A boy might climb a tree to show off for a girl, but he also might do it for fun or for other reasons.
###
A person on a horse jumps over a broken down airplane.
question: A person is outdoors, on a horse.
true, false, or neither? true
why? Horse riding is an activity almost always done outdoors. Additionally, a plane is a large object and is most likely to be found outdoors.
###
There is a red truck behind the horses.
question: The horses are becoming suspicious of my apples.
true, false, or neither? false
why? The presence of a red truck does not imply there are apples, nor does it imply the horses are suspicious.
###
A dog carries an object in the snow.
question: A dog is asleep in its dog house.
true, false, or neither? false
why?

面向开放式文本生成的事件转移规划

Event Transition Planning for Open-ended Text Generation

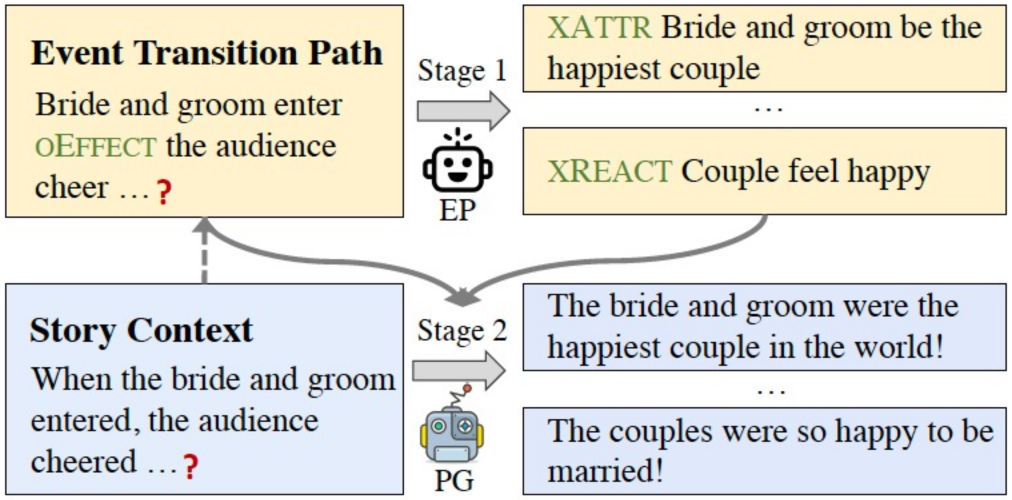


Figure 1: An illustration of our planning based framework in story completion task. Given story context, we extract corresponding event transition path, and use model EP to develop potential ensuing event transition paths. The planned paths accordingly guide the path-aware text generation model PG.

两步模型

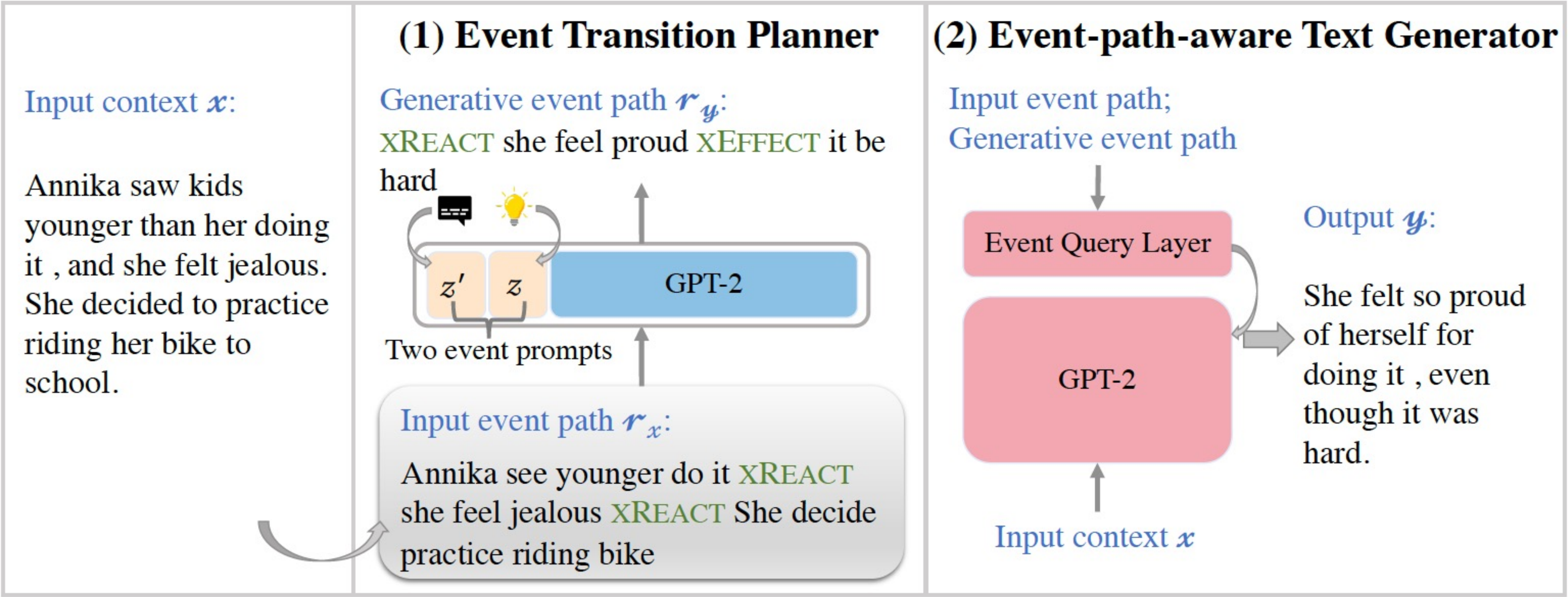


Figure 2: Overall architecture of the proposed coarse-to-fine framework. It consists of two components. (1) **Event Transition Planner**: given a input context, it first extracts corresponding event path and then generates possible ensuing event path. The planner directly inherits the pre-trained parameters from GPT-2; (2) **Event-path-aware Text Generator**: another GPT-2-based generator is applied to generate a natural language sentence by attending to input context and explicit event transition path.

怎么构建一个更好的事件转移模型

Tasks	Methods	BELU-1	BLEU-2	BLEU-4	DIST-1	DIST-2
Dialogue Generation	GPT-2	23.43	11.50	3.31	1.57	4.18
	PLANGENERATION (Ours)	**26.52**	**12.38**	3.29	**1.88**	**5.52**
	w/o PROMPT	23.58	11.85	**3.58**	1.80	5.13
	w/o TUNING ON ATOMIC	19.82	7.90	1.81	1.16	2.54
	PLANRETRIEVAL	0.75	0.14	0.00	13.05	39.52
Story Completion	GPT-2	15.98	7.19	1.08	5.53	17.44
	PLANGENERATION (Ours)	**19.51**	**9.01**	**1.35**	**5.83**	**17.48**
	w/o PROMPT	13.64	6.14	1.12	4.71	15.77
	w/o TUNING ON ATOMIC	12.74	4.61	0.47	6.08	12.27
	PLANRETRIEVAL	1.28	0.15	0.00	11.88	37.70

Table 2: Experimental results on event transition planning. For detailed description about the compared models, please refer to §4.2.

引入事件转移模型辅助下游任务

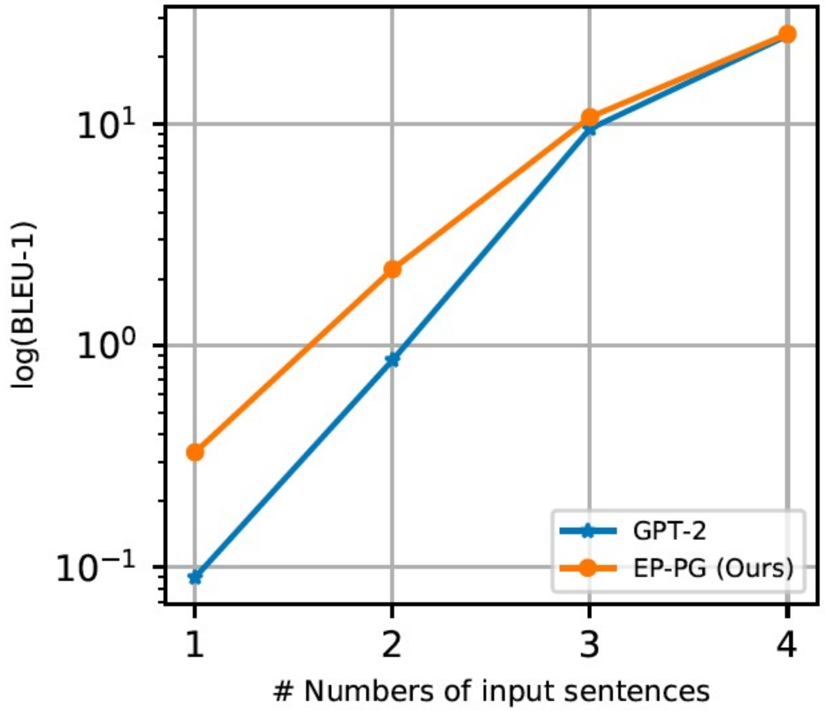


Figure 3: The log of BLEU-1 scores on story completion with different numbers of sentences as input.

解释性自然语言处理 GO AHEAD

- Contrastive explanations: justify why a prediction was made instead of another. [There is no dataset that contains contrastive free-text or structured explanations.]
  - "why...instead of...",
  - Collecting explanations for other labels besides the gold label
- Negative explanations: providing supervision of what is not a correct explanation
  - human JUDGE (low-scoring instances)
  - EDIT phase (instances pre-editing)

谢谢大家