All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text

ACL-IJCNLP 2021

ING Lab 논문 세미나
2021-09-13

발표자: 현지웅

Paper Abstract

자연어 생성에서의 Human Evaluation은 아직까지도 Gold standard

하지만, Human Evaluation은 결국 평가자의 재량으로 평가하게 됨

GPT3과 같은 최신 언어 생성 모델이 등장한 지금, 인간은 모델이 생성한 텍스트와 사람이 쓴 텍스트를 잘 구분할 수 있을 것인가? 구분하는 기준은 무엇인가?

모델이 생성한 텍스트를 잘 인지하기 위하여 평가자들을 훈련할 수 있을까? 그에 대한 3가지 방법론 적용

저자가 제시하는 NLG 평가에 대한 Recommendation

모델이 생성한 텍스트를 얼마나 잘 식별하는가?

주어진 text passage에 대하여 평가 (4-point scale)

1. Definitely human-written
2. Possibly human-written
3. Possibly machine-generated
4. Definitely machine-generated

3가지 도메인 (스토리, 뉴스, 레시피)에서 사람이 쓴 텍스트 50개와 모델이 생성한 텍스트 50개에 대하여 평가

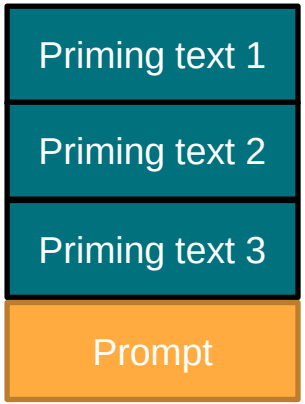
각 평가자는 한 도메인과 한 모델에 한정되어 할당 (?)

텍스트 생성에 이용되는 모델은 GPT2-XL과 GPT3

'three-shot' setting으로 텍스트 생성

‘three-shot’ setting

해당 task에 대하여 따로 fine-tuning을 진행하는 것이
아니라 해당 task에 대한 예제 3개를 보여주고 해당 prompt에
대하여 생성하도록 하는 방식



Example (스토리 도메인의 priming text)

Once upon a time magic was an art. The great magicians of the time would take on apprentices who would train for a decade or more, slaving under their tutelage for the hope of one day being the next great magician. They were free from the bonds of apprenticeship only once they had made the perfect magic circle. Only then could they harness the power of the mages and go off on their own. But times changed, magicians were persecuted, and few longed to toil away for the long years of study necessary. In the course of a century, magic became extinct.

For the machine-generated text, we conditioned the models on the three priming texts and on the phrase Once upon a time.

모델이 생성한 텍스트를 얼마나 잘 식별하는가?

Data

스토리 (WritingPrompts)
Once upon a time으로 시작하는 스토리 255개 중 랜덤으로 50개 선택 (human)
특징: 가장 open-ended, 제한이 없고 창의적인 도메인, 가장 noisy한 데이터셋

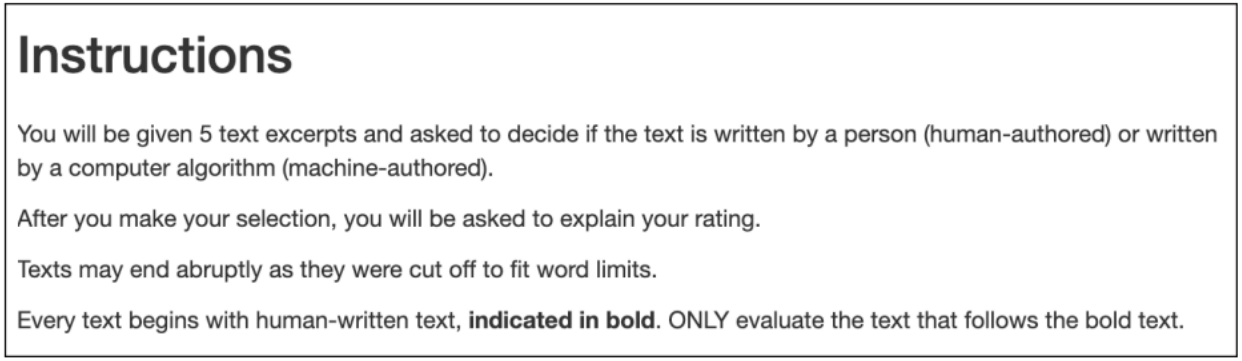
뉴스 (Newspaper3k)
2,111개의 지역 뉴스 기사 중 랜덤으로 50개 선택(human), 50개 선택 (prompt)
기사 제목과 첫 문장을 prompt로 주었음 - 이는 첫 문장과 제목이 전체 내용을 요약하는 것을 감안함

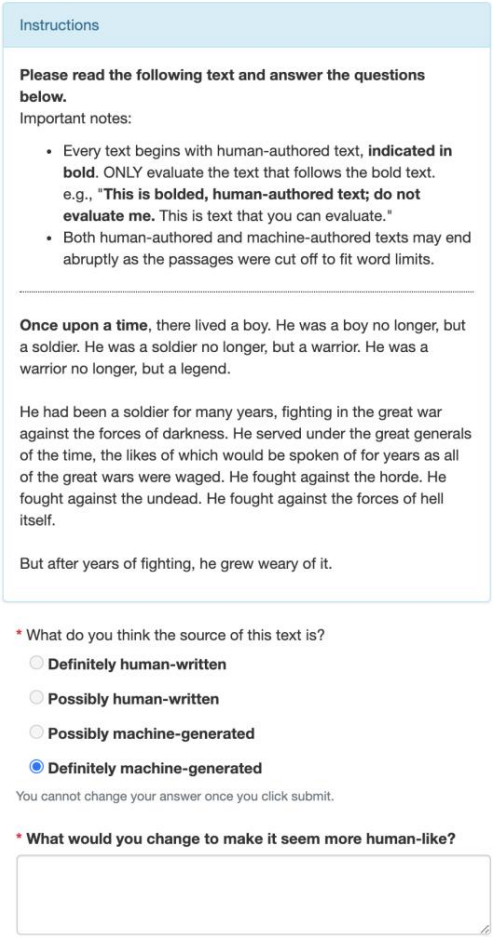
레시피 (RecipeNLG)
랜덤으로 50개 선택. 레시피 제목과 재료 리스트를 prompt로 주었음
특징: 가장 closed 도메인 (정형화 되어있음)

모델이 생성한 텍스트를 얼마나 잘 식별하는가?

Participants (AMT)

각 task setting 별 130명 선정 (미국 국적인 95% 이상의 acceptance rate & 1000 HITs 이상)
각 참가자는 5개의 text를 평가함
지시 사항을 어긴 참가자를 필터링 (445명 걸러짐 - 이 참가자들의 결과는 제외)





모델이 생성한 텍스트를 얼마나 잘 식별하는가?

Model	Overall Acc.	Domain	Acc.	F_1	Prec.	Recall	Kripp. α	% human	% confident
GPT2	*0.58	Stories	*0.62	0.60	0.64	0.56	0.10	55.23	52.00
		News	*0.57	0.52	0.60	0.47	0.09	60.46	51.38
		Recipes	0.55	0.48	0.59	0.40	0.03	65.08	50.31
GPT3	0.50	Stories	0.48	0.40	0.47	0.36	0.03	62.15	47.69
		News	0.51	0.44	0.54	0.37	0.05	65.54	52.46
		Recipes	0.50	0.41	0.50	0.34	0.00	66.15	50.62

Table 1: §2 results, broken down by domain and model, along with the F_1, precision, and recall at identifying machine-generated text, Krippendorff's α, % human-written guesses, and % confident guesses (i.e., *Definitely* machine- or human-authored). * indicates the accuracies significantly better than random (two-sided t-test, for Bonferroni-corrected $p < 0.00333$).

모델이 생성한 텍스트를 얼마나 잘 식별하는가?

결과 분석

평가자가 어느 기준에 맞추어 평가했는지 알아보기 위하여 응답 중 랜덤으로 150개를 골라 3가지 카테고리로 나눔

form (47%): format, style, tone of the text

content (25%): 텍스트 안 의미

machine capabilities: 모델이 생성한 텍스트에 대한 평가자들의 인식 (아직 이정도는 못하겠지)

대부분의 평가가 spelling, grammar, punctuation (150개 중 45개) 및 style, tone of the text (24개)에 집중

모델이 생성한 텍스트를 잘 식별하도록 훈련할 수 있을까?

기존의 평가 방식은 유지하면서 (짧고 비용이 적게 드는) + 평가자들의 성능을 향상할 수 있는 세가지 방법론에 대하여 테스트를 진행

Training with Instructions

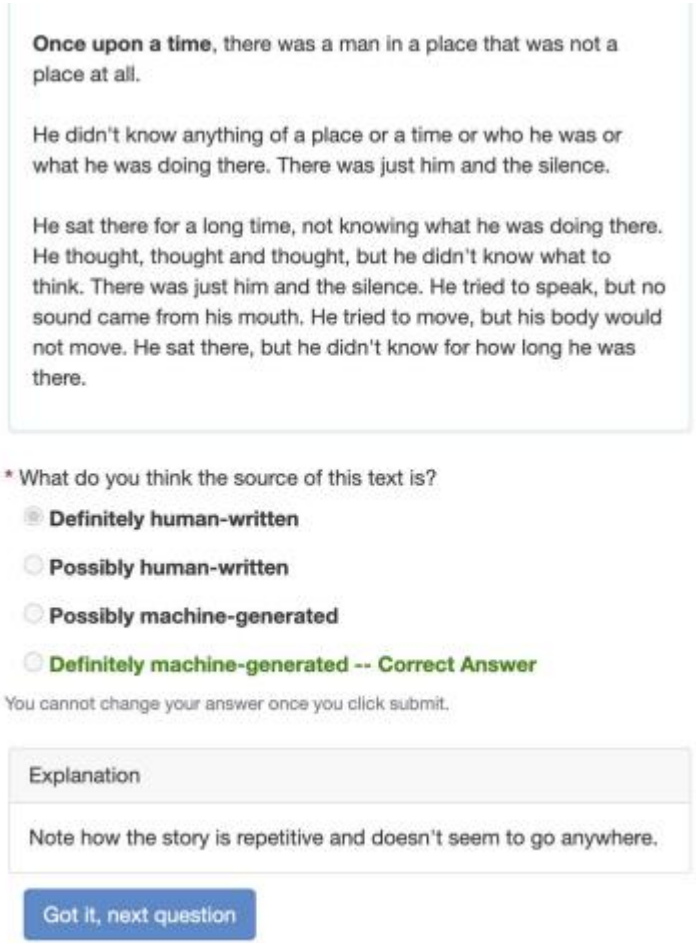
모델이 생성한 텍스트가 어느 특징을 갖는 지, 오해할 수 있는 부분을 추가 지시사항으로 알려줌

> We recommend you pay special attention to the following characteristics:
>
> - **Repetition**: Machine-generated text often repeats words or phrases or contains redundant information.
> - **Factuality**: Machine-generated text can contain text that is inaccurate or contradictory.
>
> On the other hand, be careful with these characteristics, as they may be misleading:
>
> - **Grammar and spelling**: While machine-generated text can contain these types of errors, human-authored text often contains them as well.
> - **Style**: Current AI systems can generally mimic style fairly well, so a text that "looks right" or matches the expected style of the text isn't necessarily human-authored.

모델이 생성한 텍스트를 잘 식별하도록 훈련할 수 있을까?

Training with Examples

평가 이전에 3개의 예제 평가를 보여주고 어느 특징이 모델이 생성한 것인지에 대한 설명도 주어짐

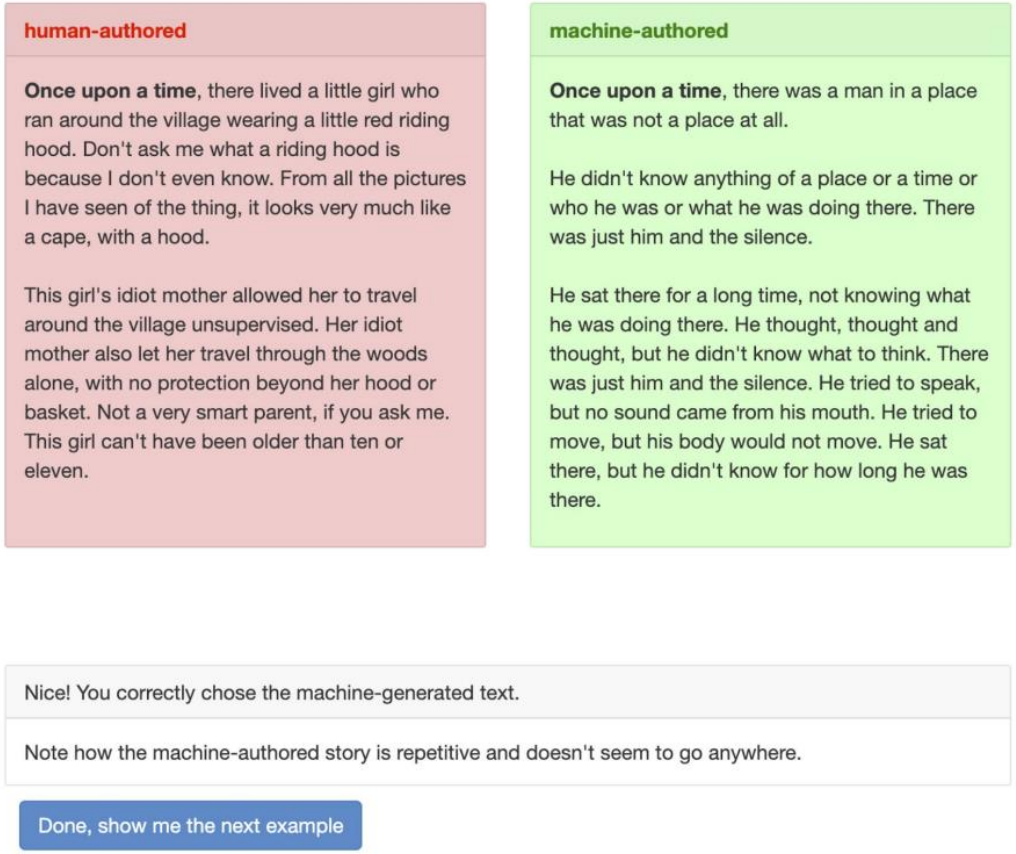


모델이 생성한 텍스트를 잘 식별하도록 훈련할 수 있을까?

Training with Comparison

사람이 쓴 텍스트와 언어가 생성한 텍스트를 둘 다 동시에 보여주고 모델이 생성한 텍스트를 고르도록 지시

Example training과 마찬가지로 정답 및 설명 또한 제공



모델이 생성한 텍스트를 잘 식별하도록 훈련할 수 있을까?

Training	Overall Acc.	Domain	Acc.	F_1	Prec.	Recall	Kripp. α	% human	% confident
None	0.50	Stories	0.48	0.40	0.47	0.36	0.03	62.15	47.69
		News	0.51	0.44	0.54	0.37	0.05	65.54	52.46
		Recipes	0.50	0.41	0.50	0.34	0.00	66.15	50.62
Instructions	0.52	Stories	0.50	0.45	0.49	0.42	0.11	57.69	45.54
		News	0.56	0.48	0.55	0.43	0.05	62.77	52.15
		Recipes	0.50	0.41	0.52	0.33	0.07	67.69	49.85
Examples	*0.55	Stories	0.57	0.55	0.58	0.53	0.06	53.69	64.31
		News	0.53	0.48	0.52	0.45	0.05	58.00	65.69
		Recipes	0.56	0.56	0.61	0.51	0.06	55.23	64.00
Comparison	0.53	Stories	0.56	0.56	0.55	0.57	0.07	48.46	56.62
		News	0.52	0.51	0.53	0.48	0.08	53.85	50.31
		Recipes	0.51	0.49	0.52	0.46	0.06	54.31	53.54

Recommendations for NLG evaluation

평가자가 어느 부분에 중점을 두어 평가했는지 조사

form과 machine capabilities에 대한 비율이 줄고 content에 대한 비율 증가

→ 이젠 평가에 내용적인 측면에 중점을 두어야 한다

Training	Form	Content	Machine capabilities
None	47.1	24.6	28.3
Examples	32.5	50.0	17.5

Recommendations

Move away from intrinsic evaluations for qualities like “humanlikeness”

Develop human evaluation that motivate evaluators to carefully read the generated text