Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models

조건희

Contents

- Introduction
- Choose Your Own Adventure
- Experiment

Introduction

- NAACL 2021 논문
- 최근 자연어 생성 모델이 떠오르고 있지만 이 모델들을 평가하는 것은 어려운 일이다.
- 이 논문에서는 pairwise system evaluation이란 방법으로 생성 모델들을 직접적으로 평가한다.

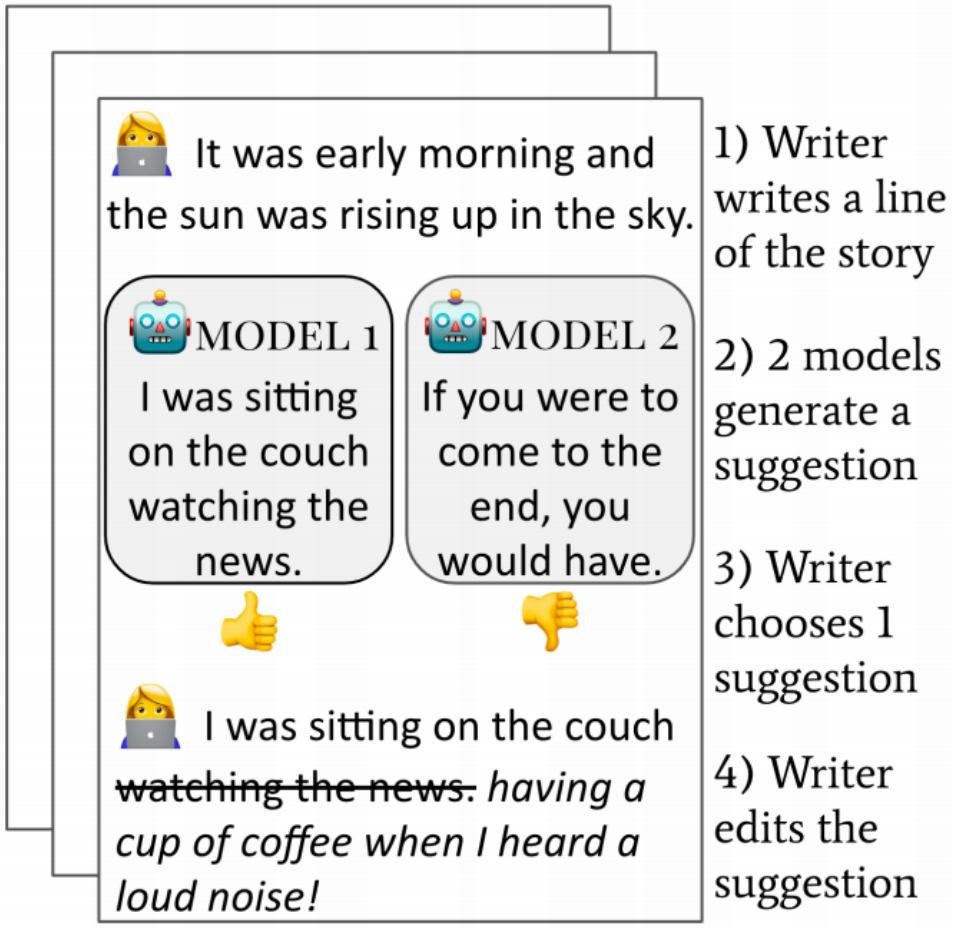
  -> 이 평가 방법을 'Choose Your Own Adventure(CYOA)'라고 한다.
- 생성 모델들의 비교 평가를 가능케 한다.

Choose Your Own Adventure(CYOA)

- Introduction에서 언급했다시피 생성 모델을 평가하는 방법

1) Writing Setup
2) Evaluation Setup

Writing Setup



X 5

Writing Setup

- 편집에는 제한이 없다.
- 한 번 story에 넣었으면 다시 편집할 수 없다.
- Story가 완성되면 참가자들에게 Likert-scale을 이용하여 질문한다.

Writing Setup - Likert Scale

- “Strongly Degree”에서 “Strongly Agree”까지를 말한다.
- 5가지 질문

  1) 스토리에 대해 만족한다.

  2) 시스템과 내가 협동적으로 스토리를 만들었다고 생각한다.

  3) 스토리를 만들면서 모델이 제안한 문장들은 도움이 되었다.

  4) 모델이 제안한 문장들은 스토리랑 연관이 있었다.

  5) 모델이 제안한 문장들은 새로운 아이디어 도출에 도움이 되었다.

Writing Setup – 주관식 질문

- 질문

  1) 어떤 요인이 그 문장을 선택하게 하였는가?

  2) Suggestions에서 어떤 것을 보았는가?
- 이 질문들은 없어지거나 조정 가능하다.

Evaluation Setup

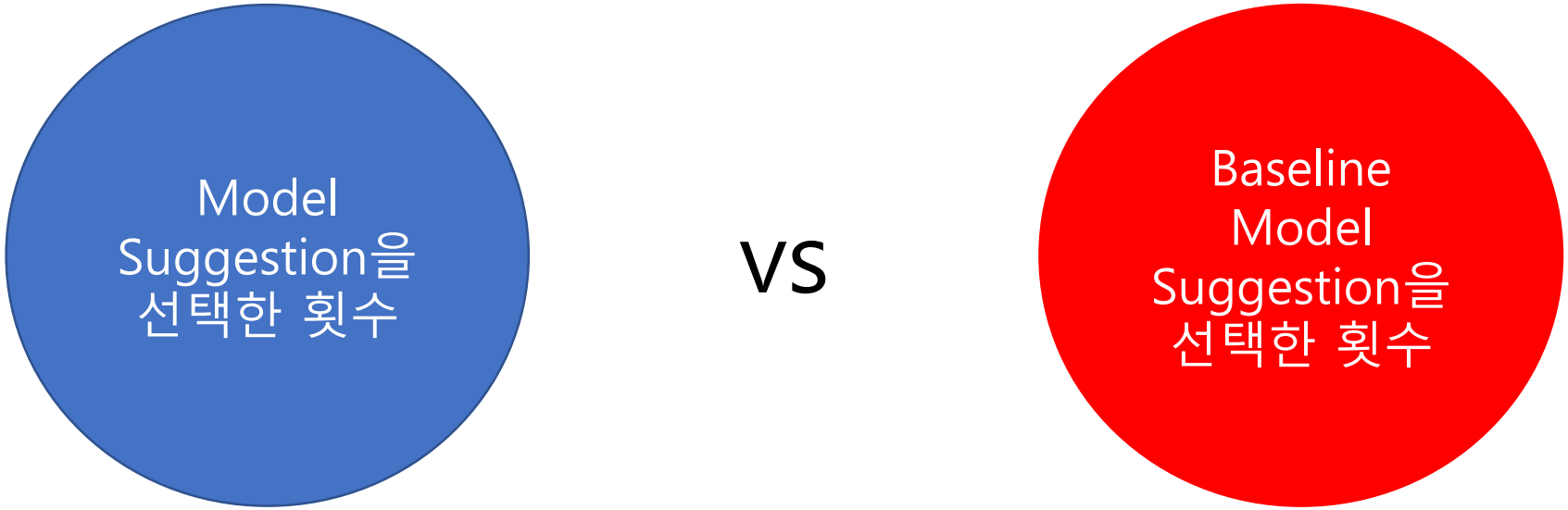
- Writing Setup에서 얻은 정보를 이용하는 단계

  -> 각 모델로부터 나온 suggestions, 두 모델 중 작가들이 선호하는 모델, 작가들이 한 편집들

- 여기서도 3가지 질문에 답한다.

Evaluation Setup

1. 모델이 baseline 모델보다 story 생성을 잘하는가?



- 또한 이것을 suggestion round(1 ~ 5)로 나누어서 작가의 선호도 변화도 보았다.

Evaluation Setup

2. 모델의 suggestions이 얼마나 도움이 되는가?

- 작가들의 suggestion 편집 정도에 주목한다.
- Metrics

  1) Levenshtein Edit Distance

     - 삽입, 삭제, 대체된 문자 개수 측정

  2) Jaccard Similarity

     - Orginal text와 편집된 text 사이에 공유된 토큰들의 비율 측정

  3) User Story Edit Ratings(USER)

     - Orginal text와 편집된 text 사이에 가장 긴 substrings의 개수 측정

Evaluation Setup

3. 모델이 생성한 text가 human-authored text에 비해 어떤가?

- Pairwise comparison이 이것의 답을 해준다.
- 작가가 직접 쓴 문장과 모델이 generate한 text를 비교한다.
- 평균 문장 길이, distinct-n(반복 빈도수), text의 명사와 동사의 concreteness를 측정한다.

  -> concreteness는 Brysbaert et al.(2014)에서 제시한 concreteness ratings로 계산

Experiment

Fusion vs. GPT2

- Fusion: 두 개의 convolutional seq-to-seq model을 합치는 방법
- GPT2: Top-K sampling을 사용하고 story data에 fine-tuning한 small GPT2
- Train dataset으로 WritingPrompts dataset을 사용
- 105명의 Turkers(사람들)을 고용

Fusion vs. GPT2 – Q1

	Total	#1	#2	#3	#4	#5
% GPT2	66	76	70	63	63	57

Fusion vs. GPT2 – Q2

	ED (↓)	JS (↑)	USER (↑)
FUSION	37.61	51.13	60.69
GPT2	**29.49**	**61.35**	**71.77**

Fusion vs. GPT2 – Q3

	FUSION	GPT2	HUMAN
avg. sent. len.	13.70	10.31	18.86
concrete N	4.04	4.35	4.17
concrete V	2.90	3.10	3.12
distinct-1	0.75	0.53	0.72
distinct-2	0.97	0.70	0.95
distinct-3	1.00	0.76	0.99

Nucleus vs. Top-K

- 둘 다 GPT2
- Nucleus: Nucleus sampling을 사용한 GPT2
- Top-K: Top-k sampling을 사용한 GPT2
- 103명의 Turkers(사람들)을 고용

Nucleus vs. Top-K – Q1

	Total	#1	#2	#3	#4	#5
% TOP-K	53	58	53	53	53	49

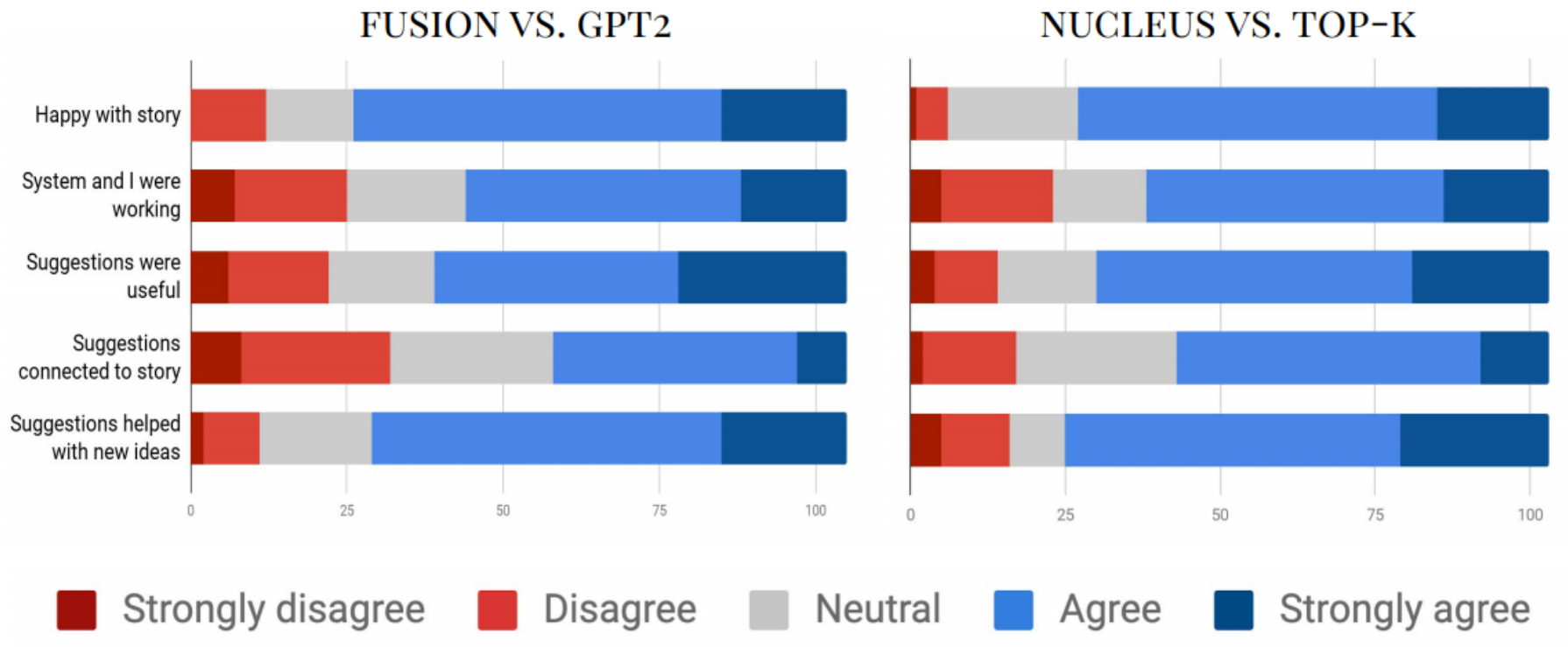
Nucleus vs. Top-K – Q2

	ED (↓)	JS (↑)	USER (↑)
NUCLEUS	**34.65**	**53.64**	**63.64**
TOP-K	36.69	50.96	62.18

Nucleus vs. Top-K – Q3

	NUCLEUS	TOP-K	HUMAN
avg. sent. len.	12.76	10.53	19.28
concrete N	4.15	4.34	4.23
concrete V	3.08	3.08	3.11
distinct-1	0.77	0.60	0.72
distinct-2	0.96	0.78	0.96
distinct-3	0.99	0.84	0.99

Writer's Feedback



Contribution

- CYOA는 story generation model에 대한 자동적인 평가가 가능하다.
- Paired Suggestion을 통해서 두 모델 사이를 직접적으로 비교할 수 있다.