Retrieval Enhanced Model for Commonsense Generation

ACL-IJCNLP Findings 2021

ING Lab 논문 세미나
2021-06-24

발표자: 현지웅

Commonsense Generation

Commonsense Generation (Generative Commonsense Reasoning)

- 입력: Concept Set (a collection of objects/actions)
  ex.) dog | frisbee | catch | throw

- 출력: Everyday scenarios covering all given concepts
  ex.) The **dog catches** the **frisbee** when the boy **throws** it.

Task for testing machines for the ability of generative commonsense reasoning

Generate a coherent sentence describing an everyday scenario using these concepts

주어진 concept들을 이용하여 상식적으로 말이 되는 문장들을 생성해내는 것이 목표

Commonsense Generation

CommonGen [paper] [dataset] [leaderboard]

- 30k concepts set, 50k 문장들로 구성되어 있음

핵심 평가 지표: **SPICE**, BLUE-4, CIDEr

SOTA (in leaderboard): KFC (33.911), KGR (33.564)

SOTA (in publication): **RE-T5** (31.079), KG-BART (29.634), EKI-BART (29.583)

Figure 1: **An example of the dataset of COMMONGEN.** GPT-2, UniLM, BART and T5 are large pre-trained text generation models, *fine-tuned* on the proposed task.

Baseline in CommonGen

GPT2

학습 prompt: "c_1 c_2 ⋯ c_k = y" 추론: "c_1 c_2 ⋯ c_k =" + beam search

T5

prompt: "generate a sentence with c_1 c_2 ⋯ c_k."

추론: beam search

GPT-2 (Radford et al., 2019)	16.85	39.01	33.92	23.73	26.83	12.19	23.57	79.09
BERT-Gen (Bao et al., 2020)	17.78	40.21	33.29	23.47	28.25	12.61	24.82	86.06
UniLM (Dong et al., 2019)	21.20	**43.60**	41.82	30.73	30.62	14.89	27.43	89.19
UniLM-v2 (Bao et al., 2020)	18.11	40.51	34.31	24.53	29.04	13.19	25.52	89.13
BART (Lewis et al., 2019)	**22.02**	41.78	39.52	29.01	**31.83**	13.98	28.00	**97.35**
T5-Base (Raffel et al., 2019)	14.63	34.56	28.76	18.54	23.94	9.40	19.87	76.67
T5-Large (Raffel et al., 2019)	21.74	42.75	**43.01**	**31.96**	31.12	**15.13**	**28.86**	95.29
Human Performance (Upper Bound)	36.72	53.45	52.55	46.49	38.79	37.64	52.43	99.33

EKI-BART (Fan et al., 2020)

외부 corpus을 이용하여 주어진 concept들이 등장하는 문장들을 retrieve하고, 이를 추가적인 입력(auxiliary input)으로 이용하는 retrieve-and-generate 방식

Task를 주어진 concepts과 Prototype을 이용하여 문장을 생성하는 것으로 변형

외부 corpus: in-domain corpus (Image, Video captioning), out-of-domain corpus (Wikipedia)

Concepts	front, guitar, microphone, sit	ear, feel, pain, pierce
BART	guitar sits in front of a microphone in the front.	I can feel the pain in my ears and feel the pierce in my neck from the piercing.
Prototype	A singer performed the song standing in front of the audiences while playing guitar.	He expresses severe pain as he tries to pierce his hand.
BART+ Prototype	A singer sitting in front of the audiences while playing guitar.	He expresses severe pain as he pierces his ear.

Table 1: Example of *BART*, *Prototype* and *BART+Prototype*.

Motivation

Flaw in EKI-BART

EKI-BART의 retrieve 방식: 단순히 concept이 포함되어 있는 여부를 기준으로 (match)

→ retriever가 학습 불가능하고, fine-tuning 과정에서만 사용된다.

retriever를 학습할 수 있게 설계하고, pre-training 과정에서도 사용할 수 있으면 좋지 않을까?

→ **RE-T5**

Method

Retrieval

외부 corpora에서 최소 2개 이상의 concept이 포함된 문장을 후보로 선정하여 candidate set (Z)를 구축

Matching Retriever: concept이 포함된 개수를 기준으로 정렬하여 top-k개의 문장을 선택하여 이를 auxiliary 입력으로 사용

Trainable Retriever: fine-tuned BERT "[CLS] X [SEP] z_i [SEP]"

어떻게 학습이 가능하나요?

CommonGen 학습 데이터셋의 문장을 positive example, 그 외 문장을 랜덤으로 샘플링한 문장을 negative example

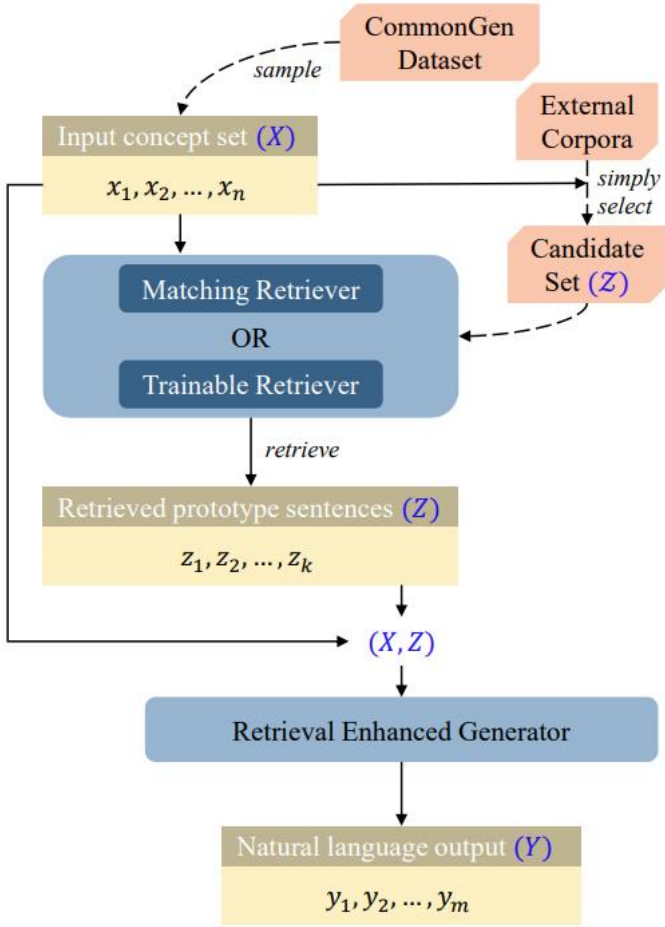

Figure 1: The overall framework of Retrieved Enhanced Model for Commonsense Generation.

Method

Pre-training

외부 문장에 spaCy 라이브러리를 이용하여 {동사, 명사, 대명사}를 추출하여 pseudo concept set을 구축

* CommonGen test set에 존재하는 concept set은 제외
* 추출된 concept은 ConceptNet에 존재하는 형태로만

구축된 concept set과 문장을 이용해서 pre-training을 진행

Pre-training 단계에서는 matching retriever를 사용

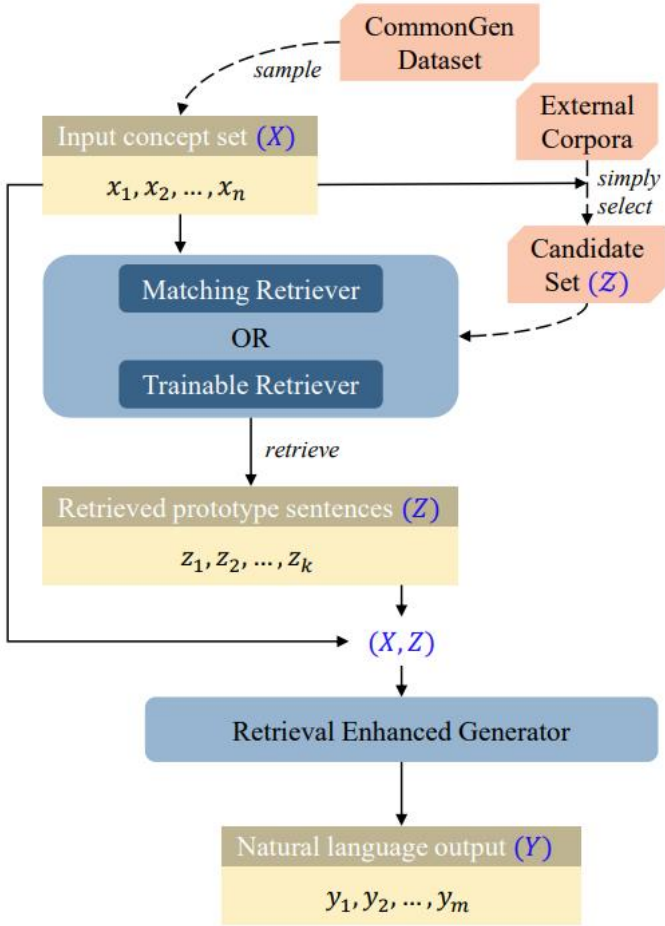

Figure 1: The overall framework of Retrieved Enhanced Model for Commonsense Generation.

Method

Fine-tuning

trainable retriever로 Z로부터 top-k 문장을 선택

concepts과 검색된 prototype 문장들을 입력으로 하여 원본 문장을 생성하도록 fine-tuning

Pre-training과 Fine-tuning의 차이?

Pre-training: 외부 corpora의 concept set으로 문장 생성 학습

Fine-tuning: concept set과 외부 corpora의 문장으로 문장 생성

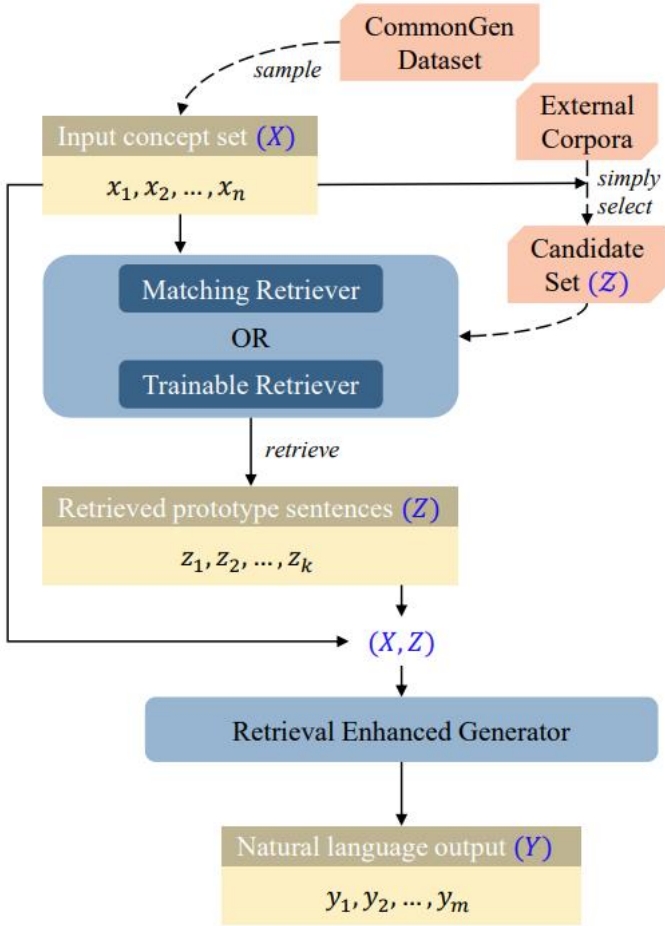

Figure 1: The overall framework of Retrieved Enhanced Model for Commonsense Generation.

Experiment

Dataset
CommmonGen

평가 지표
BLEU, ROUGE, METEOR, CIDEr, SPICE

외부 Corpora
VATEX, Activity, SNLI, MNLI
이 들 중 500k 문장을 sample하여 사용

비교군
GPT-2, BERT-Gen, UniLM, BART, T5
EKI-BART, KG-BART, CALM

Concept Set: trailer shirt side sit road
T5: A man sits on the side of a trailer and a shirt.
Matching Retriever: (1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer. (2)Two men, one wearing a straw cone hat, blue shirt, talking with a guy in a tan sunhat, red plaid shirt, both with baskets in front of them, sitting on the side of a dirt road. (3)An older guy with a tan shirt and hat sitting on the side of a road with bricks all around him and a small green bowl on the side. **RE-T5(matching retriever)**: a man in a tan shirt sits on the side of a road.
Trainable Retriever: (1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer. (2)Teenagers in matching shirts stand at the side of the road holding trash bags. (3)A man in a white shirt and black pants standing at the side or the road. **RE-T5(trainable retriever)**: a man in a white shirt and black pants sits on the side of a trailer on the road.

Experiment

Model	BLEU-4	CIDEr	**SPICE**	SPICE(v1.0)
GPT-2 (Radford et al., 2019)	26.833	12.187	23.567	25.90
BERT-Gen (Bao et al., 2020)	23.468	12.606	24.822	27.30
UniLM (Dong et al., 2019)	30.616	14.889	27.429	30.20
BART (Lewis et al., 2020)	31.827	13.976	27.995	30.60
T5-base (Raffel et al., 2020)	18.546	9.399	19.871	22.00
T5-large (Raffel et al., 2020)	31.962	15.128	28.855	31.60
EKI-BART (Fan et al., 2020)	35.945	16.999	29.583	32.40
KG-BART (Liu et al., 2021)	33.867	16.927	29.634	32.70
CALM(T5-base) (Zhou et al., 2021)	-	-	-	33.00
RE-T5 (ours)	**40.863**	**17.663**	**31.079**	**34.30**

Ablation Study

Retriever의 효과를 검증하기 위한 실험

Retrieve only vs. T5
large-scale pretrained LM (T5)이 많은 양의 지식을 포함하고 있듯이, 외부 corpora에서 검색된 문장 역시 concept으로부터 충분한 지식을 뽑아낼 수 있다.

T5 + MR
Data augmentation의 중요성

T5 + TR + pretrain
trainable scorer가 commonsense generation에서 좀 더 knowledge를 잘 capture한다

Model	SPICE
Retrieve (only)	29.60
T5	30.80[3]
T5 + *MR*	33.60
T5 + *MR* + pretrain	33.90
RE-T5 (T5 + *TR* + pretrain)	**34.30**

Example Analysis

T5 - road 빼먹음, 비상식적인 shirt 사용
Matching Retriever - 대개 긴 문장이 선택되어, 일부 concept을 제외하고 생성할 수 있다(trailer가 제외)
Trainable Retriever - 모든 concept 사용 및 상식적인 문장 생성

Concept Set: trailer shirt side sit road
T5: A man sits on the side of a trailer and a shirt.
Matching Retriever: (1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer. (2)Two men, one wearing a straw cone hat, blue shirt, talking with a guy in a tan sunhat, red plaid shirt, both with baskets in front of them, sitting on the side of a dirt road. (3)An older guy with a tan shirt and hat sitting on the side of a road with bricks all around him and a small green bowl on the side. **RE-T5(matching retriever)**: a man in a tan shirt sits on the side of a road.
Trainable Retriever: (1)Two guys in red shirts are sitting on chairs, by the side of the road, behind that open trailer. (2)Teenagers in matching shirts stand at the side of the road holding trash bags. (3)A man in a white shirt and black pants standing at the side or the road. **RE-T5(trainable retriever)**: a man in a white shirt and black pants sits on the side of a trailer on the road.

Conclusion

- 간단한 아이디어로 훌륭한 성능을 뽐낸 케이스
- 얼마 안된 따끈따끈한 데이터셋에 얼른 좋은 성능을 내서 논문을 내자! (물론 SOTA가 되어야 겠지만)