A Semantic-based Method for Unsupervised Commonsense Question Answering

ACL-IJCNLP 2021

Abstract

Unsupervised Commonsense Question Answering의 기존 해결 방안

- Pre-trained LM이 주어진 Question이나 Context에 대해 각 선택지에 점수를 매긴 후, 제일 높은 점수를 받은 선택지를 Answer로 선택!
- But! 단어의 빈도, 문장의 구조 등 실제 Answer와 무관한 요인에 큰 영향을 받음

Semantic-based Question Answering method(SEQA)를 제시

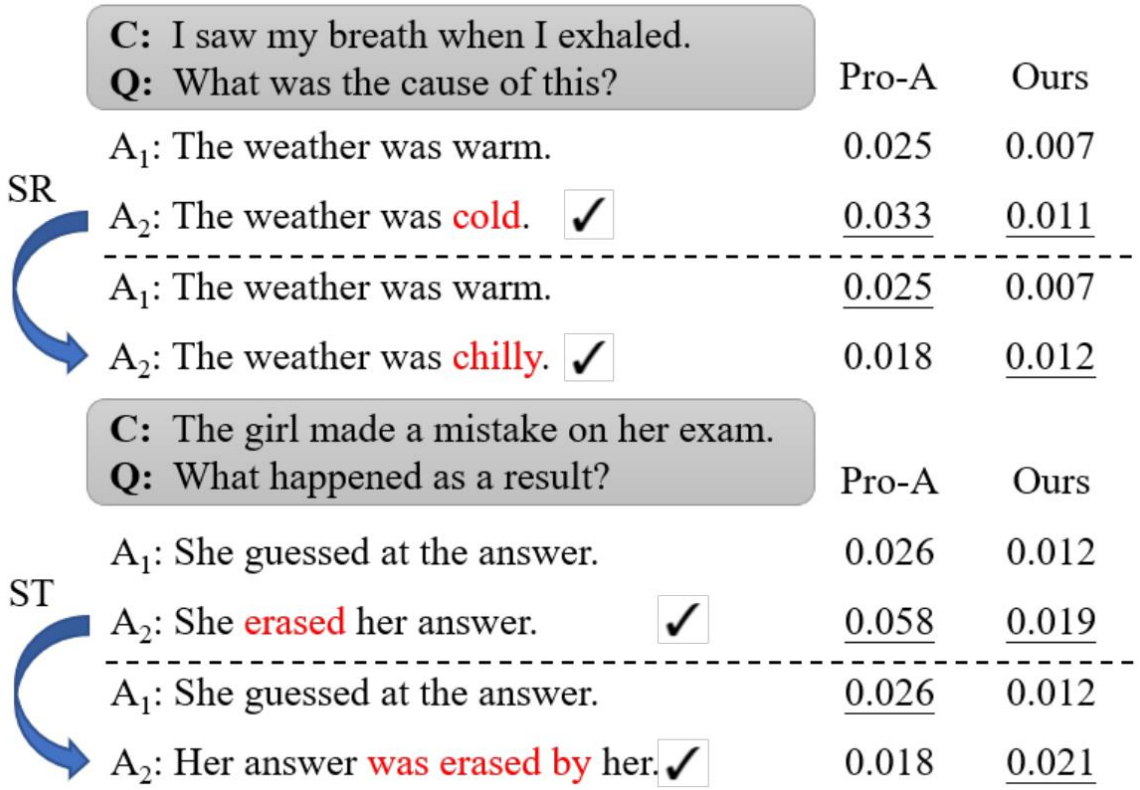
- Question에 이어지는 문장을 Generate, 이후 각 선택지와의 Semantic similarity를 비교해 제일 가까운 선택지 선택.
- 동의어 등의 lexical한 문제에서도 정답을 잘 찾아냄.

Introduction

- 기존의 많은 Unsupervised method는 LM을 이용해서 각 선택지에 점수를 매김.
  Ex) Question에 대한 각 Answer의 조건부 확률
- 단점: 동의어에 대해서 정답 선택의 변화가 있을 수 있음.
  ⇒ lexical한 변화에 oversensitive하다.

즉, Commonsense Question Answering에서는,

1. Answer의 Semantic에 더욱 초점을 맞추고
2. 동의어에 비슷한 수준의 점수를 부여해야 함



Introduction

Semantic-based Question Answering method(SEQA)

- 각 선택지에 대해 직접적으로 점수를 매기지 않음.
  ⇒ 각 선택의 Semantic을 관찰할 확률을 계산함.
- 각 선택지의 semantic score
  = LM에서 선택지와 semantic이 동일한 문장(=Supporter)의 생성 확률의 합

Supporter는 Answer 선택 시 표면적인 구조 보다는 의미론적인 방법으로 선택하도록 도와줌.

또한, 동의어는 같은 Supporter를 공유할 가능성이 큼. ⇒ Score가 근접할 것으로 예상됨.

Previous Work

- Pro-A: Question에 대한 Answer의 조건부 생성 확률.
  ⇒ 단어 빈도나 문장 길이 같은 통계적인 편향성을 보임.
- MI-QA: Pro-A의 통계적인 편향성을 완화하기 위해서 제안.
  Question과 각 Choice 간의 상호적인 정보를 포함해 계산.
- Pro-Q: Answer에 대한 Question의 조건부 확률.

Method	Score Function
Pro-A	$[P_{LM}(A\|Q)]^{\frac{1}{\|A\|}}$
Pro-Q	$[P_{LM}(Q\|A)]^{\frac{1}{\|Q\|}}$
MI-QA	$\left[\frac{P_{LM}(A\|Q)}{P_{LM}(A)}\right]^{\frac{1}{\|A\|}}$
SEQA (Ours)	$\sum_{S\in\mathbb{A}} \omega(S\|A) P_{LM}(S\|Q)$

Motivation

Question Q가 주어졌을 때, 주어진 Score function s를 최대로 하는 Answer A를 선택해야 함.

$$\hat{A} = \underset{A}{\operatorname{argmax}}\, s(A|Q),$$

기존 방법: Score function은 일반적으로 LM 점수를 기반으로 정의됨.

Ex) Pro-A: 주어진 Question을 평문으로 변환했을 때, 각 choice들이 생성될 확률을 계산.

- Q: I saw my breath when I exhaled. What was the cause of this? ⟶ Rewrite: I saw my breath when I exhaled because ___

⇒ 이러한 방법은 단어 빈도, 문장 구조 등의 방해 요소들에 큰 영향을 받음.

SEQA

- SEQA: 선택지 A의 Semantic 점수를 예측하는 방법
- 1개의 선택지 A에 대해 $P(A|Q)$를 직접적으로 계산하는 것이 아닌,
  M_A가 A의 semantic을 나타낸다고 했을 때의 확률 $P(M_A|Q)$에 초점을 맞춘다.
- 이때, $P(M_A|Q)$는 A의 supporter에 대한 조건부확률의 합으로 나타낼 수 있다.
- 즉, Semantic score는 다음과 같이 정의됨.

$$s(A|Q) \triangleq P(M_A|Q) = \sum_{S \in \mathbb{S}_A} P_{LM}(S|Q) \quad (1)$$

$$= \sum_{S \in \mathbb{A}} \mathbb{I}(S \in \mathbb{S}_A) P_{LM}(S|Q). \quad (2)$$

$\mathbb{S}_A$: 선택지 A에 대한 Supporter 집합.
$\mathbb{A}$: 가능한 모든 answer의 집합.
$\mathbb{I}(S \in \mathbb{S}_A)$: S가 A의 Supporter인지 여부를 나타내는 함수.

- 이때, Supporter 집합 $\mathbb{S}_A$를 얻기 위해, 문장 단위의 semantic 특징을 추출하는 모델을 채택.

SEQA

- 이때, $\mathbb{I}(S \in \mathbb{S}_A)$는 다음과 같이 정의됨.

$$\mathbb{I}(S \in \mathbb{S}_A) = \begin{cases} 1 & \text{if } \cos(h_S, h_A) = 1, \\ 0 & \text{if } \cos(h_S, h_A) < 1, \end{cases} \quad (3)$$

h_A: 문장 A의 Semantic 특징.
h_S: 문장 S의 Semantic 특징.

- h_S와 h_A가 같은 방향이면, S와 A가 정확하게 동일한 Semantic을 갖는다고 가정.
- 그러나, 위의 조건은 너무 엄격함 ⇒ 조건을 만족하는 Supporter를 찾는 것이 어려움.
- 그래서 위의 조건을 완화해서 Semantic score를 재정의.

$$s(A|Q) \triangleq \sum_{S \in \mathbb{A}} \omega(S|A) P_{LM}(S|Q), \quad (4)$$

- 새로운 함수 $\omega(S|A)$가 $\mathbb{I}(S \in \mathbb{S}_A)$를 따라하기 위해서, 3가지 요구사항을 만족해야 함.
  1. $\omega(S|A) \in [0, 1]$
  2. $\omega(S|A) = 1$ if $\cos(h_S, h_A) = 1$
  3. $\omega(S|A)$는 $\cos(h_S, h_A)$에 따라 단조증가.

SEQA

- 앞선 요구 사항을 만족하는 $\omega(S|A)$를 본 논문에서는 다음과 같이 정의.

$$\omega(S|A) = \frac{1}{Z(T)} \exp\left[\frac{\cos(h_S, h_A)}{T}\right]. \quad (5)$$

T: temperature.
$Z(T)$=exp($\frac{1}{T}$): $\omega(A|A) = 1$을 만드는 normalization 항.

  - ✓ 만약 $T \to 0$이면, $\mathbb{I}(S \in \mathbb{S}_A)$와 거의 일치.
  - ✓ 만약 $T > 0$이면, feature space의 unit sphere에 대한 von Mises-Fisher 분포를 따름.
    이때, 허용가능한 feature vector는 평균 방향인 $\frac{h_A}{\|h_A\|}$ 주위에 분포.

- 그리고, A에 가능한 모든 Answer를 나열하는 것은 어렵기 때문에,
  앞선 수식을 $P_{LM}(S|Q)$에 대한 기댓값으로 변환
  - ✓ 이때 $S_1, \ldots, S_K$는 $P_{LM}(\cdot|Q)$에서 Sampling된 문장. (K=Sample 크기)
  - ✓ h_A, h_{S_i}는 pre-trained model에서 추출할 수 있음. (ex. SentenceBERT)

$$s(A|Q) = \mathbb{E}_{S\sim P_{LM}(S|Q)}[\omega(S|A)]$$
$$\approx \frac{1}{K}\sum_{i=1}^{K}\omega(S_i|A) \quad (6)$$
$$= \frac{1}{K\cdot Z(T)}\sum_{i=1}^{K}\exp\left[\frac{\cos(h_{S_i}, h_A)}{T}\right], \quad (7)$$

- 이때, Semantic Score $s(A|Q)$가 A의 표면적 형태와 상관없이,
  A의 semantic 특징 h에만 의존한다는 것을 알 수 있음.

The Voting View of SEQA

- Semantic score = Supporter들에 대한 조건부확률의 합.
- But! 앞서 Sampling된 문장 $S_1, \dots, S_K$는 A와 semantic한 유사성이 적을 수 있으므로, A의 Supporter는 아님.
- 이러한 차이점을 해결하기 위해, $S_1, \dots, S_K$ 를 Voter라고 지정함.
- Voter들은 Supporter는 아니지만, Question Q에 대해 "그럴듯한" 답변으로 구성.

두 개의 선택지 A_1, A_2 가 있을 때, Semantic score $s(A_1|Q)$, $s(A_2|Q)$에 따라 정답을 찾는 법.

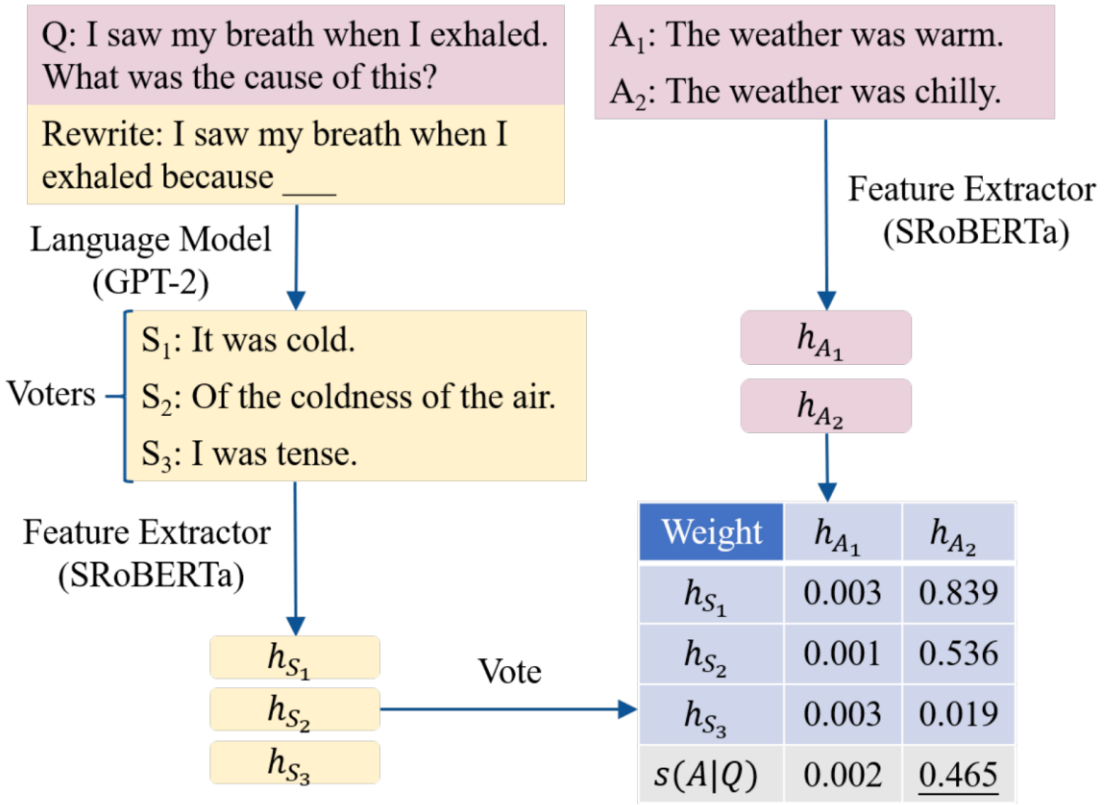
1. $P_{LM}(\cdot|Q)$에서 Question Q만 고려하여 Voter $S_1, \dots, S_K$ 를 Sampling.
2. 각 Voter는 semantic 유사도 가중치가 있는 선택지에 Vote.
   Ex) S_i는 가중치 $\omega(S_i|A_j)$를 가진 A_j에 Vote

더 많은 표를 얻은 선택지는 더 높은 Semantic score를 가짐. ⇒ 최종 답안으로 선택!

The Voting View of SEQA

SEQA의 과정:

1. Question을 평문으로 Rewrite.
2. GPT-2를 사용해 Voter S_i 생성.
3. 각 선택지 A와 Voter S는 SRoBERTa를 통해 인코딩되어 semantic feature h_{A_j}, h_{S_i} 얻음.
4. Voting weights $\omega(S_i|A_j)$ 를 계산해서, $s(A_j|Q)$가 제일 높은 A_j를 Answer로 선택.



Experiments

- Datasets: 4개의 multiple-choice commonsense question answering task에 대해서 experiment 진행. ⇒ COPA, StoryClozeTest, SocialIQA, CosmosQA
- Baselines: 앞서 소개했던 Pro-A, Pro-Q, MI-QA + CGA, Self-Talk
- Settings: 각 method에 대해 서로 다른 pre-trained LM을 시도 후, 제일 정확도가 높은 LM을 선택.

ex) SEQA: GPT-2를 사용해 p=0.9인 Nucleus Sampling으로 500개의 Voter 생성

Experiments

Dataset	Method	Pre-trained Models	Original Accuracy (↑)	After-Attack Accuracy (↑)	Attack Success Rate (↓)	Percentage of Perturbed Words	Semantic Similarity
COPA	Pro-A	GPT-2	73.6	4.6	93.8	17.3	0.883
	Pro-Q	RoBERTa	**79.4**	23.0	71.0	22.9	0.828
	MI-QA	GPT-2	74.6	16.2	78.3	19.9	0.865
	Self-talk	COMET+GPT-2	68.6	8.4	87.8	19.8	0.855
	CGA	GPT-2	72.2	4.8	93.4	17.1	0.886
	SEQA	GPT-2+SRoBERTa	**79.4**	**59.0**	**25.7**	21.7	0.827
SCT	Pro-A	GPT-2	72.3	4.8	93.3	14.3	0.917
	Pro-Q	RoBERTa	56.3	22.3	60.3	18.1	0.872
	MI-QA	GPT-2	66.1	29.2	55.8	16.2	0.885
	Self-talk	COMET+GPT-2	70.4	4.7	93.3	14.2	0.915
	CGA	GPT-2	71.5	4.8	93.2	14.3	0.916
	SEQA	GPT-2+SRoBERTa	**83.2**	**69.4**	**16.5**	18.3	0.856
SocialIQA	Pro-A	GPT-2	46.0	16.2	64.7	21.0	0.876
	Pro-Q	RoBERTa	42.2	27.8	34.2	23.2	0.843
	MI-QA	GPT-2	41.2	24.6	40.4	25.3	0.866
	Self-talk	COMET+GPT-2	**47.5**	12.3	74.0	22.2	0.872
	CGA	COMET	45.4	18.4	59.4	22.3	0.867
	SEQA	GPT-2+SRoBERTa	**47.5**	**38.2**	**19.5**	23.5	0.839
CosmosQA	Pro-A	GPT-2	36.8	1.3	96.4	9.2	0.927
	Pro-Q	RoBERTa	21.5	5.0	76.6	13.7	0.859
	MI-QA	GPT-2	29.3	7.4	74.8	12.1	0.886
	Self-talk	COMET+GPT-2	36.1	1.2	96.7	8.9	0.928
	CGA	GPT-2	42.4	1.7	96.0	9.6	0.924
	SEQA	GPT-2+SRoBERTa	**56.1**	**32.6**	**41.8**	13.9	0.859

Experiments

Accuracy

- 모든 Dataset에서 SEQA가 제일 좋은 accuracy를 보임.
- Semantic score를 사용하는 것이 방해 요소들의 감소로 인해 Commonsense QA에 유리하다고 생각됨.

Robustness

- 동의어 대체 공격에서 robustness를 확인하기 위해 TextFooler를 사용.
- SEQA의 attack success rate는 다른 방법보다 매우 낮음.
- CosmosQA에서의 성공률이 조금 높았는데, 이는 CosmosQA의 context가 매우 복잡해서 GPT-2로는 더 좋은 품질의 answer를 생성하기가 어렵기 때문.

Experiments

Consistency Testing

- SEQA가 동의어 선택지에 비슷한 점수를 할당하는지 확인.
  1. 1개의 정답과 19개의 오답을 가진 예제를 제시.
  2. 각 선택지를 다른 언어(중국어, 스페인어, 러시아어)로 번역 후, 역번역.
  3. 동의어 선택지를 포함한 모든 선택지에 대한 점수를 계산 후, 점수에 대해 정렬.
- 각 방법의 score 방식이 다르므로, 정답인 선택지와, 그것의 동의어 선택지 순위 간의 표준편차를 계산.
- SEQA의 평균 표준편차가 다른 방법 보다 훨씬 낮음.

Method / Dataset	COPA	SCT	SocialIQA	CosmosQA
Pro-A	9.1	11.0	11.7	9.4
Pro-Q	6.9	8.5	11.6	12.3
MI-QA	7.5	5.8	11.1	7.9
Self-Talk	13.3	9.5	10.7	10.1
CGA	9.7	11.0	10.9	9.5
SEQA	**4.1**	**3.2**	**5.8**	**4.7**

Experiments

Ablation Study

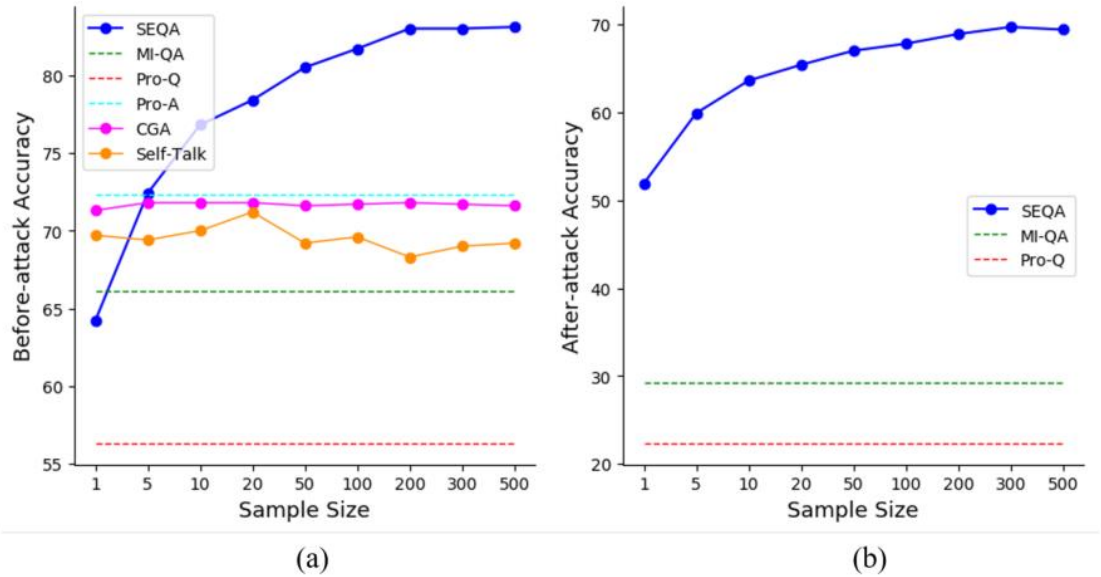
1. Temperature

- 기본적으로 SEQA의 T=0.1.
- T를 크게 변화시켜도 SEQA의 성능은 안정적.

T	COPA		SCT		SocialIQA		CosmosQA	
	Bef	Aft	Bef	Aft	Bef	Aft	Bef	Aft
10	75.6	48.8	82.0	64.7	46.3	35.9	52.7	22.3
1	76.4	48.8	82.4	64.5	46.6	36.1	53.3	22.4
0.2	77.0	52.8	**83.6**	66.3	46.9	36.8	54.8	26.1
0.1	79.4	**59.0**	83.2	**69.4**	**47.5**	**38.2**	**56.1**	**32.6**
0.05	**80.2**	54.6	80.8	61.4	46.0	36.5	55.1	28.8

2. Sample Size

- Sample size에 따라서 정확도가 증가함.
- 다른 방법은 Sample size에 크게 영향을 받지 않음.



Experiments

Ablation Study

3. $\omega(S|A)$

- 3가지 요구사항을 만족하는 다른 형식으로 정의 가능.
- 어떠한 정의라도 Baseline 보다 좋은 성능.
- SEQA의 고유한 Robustness를 유지.

$\omega(S\|A) = \frac{1}{f(1)} f\left(\cos(h_S, h_A)\right)$	Bef	Aft
$f(x) = \mathbb{I}(x > \alpha)$	77.2	47.2
$f(x) = \mathrm{ReLU}(x - \beta)$	77.6	45.2
$f(x) = \mathrm{sigmoid}(\frac{x}{T})$	75.6	48.6
$f(x) = \exp\left(\frac{x}{T}\right)$	79.4	59.0

4. Pre-trained LM and Feature Extractor

- SEQA는 pre-trained LM과 Feature Extractor 선택에 제약이 없음.
- 동일한 LM에서 더 강력한 Extractor는 더 높은 정확도.

	GPT-2		
	medium	large	xlarge
Avg. GloVe	56.6	59.6	61.2
SBERT-base	71.2	72.6	74.8
SRoBERTa-base	72.4	72.0	75.4
SRoBERTa-large	74.2	75.2	79.4

Experiments

Analysis on the Quality of Voters

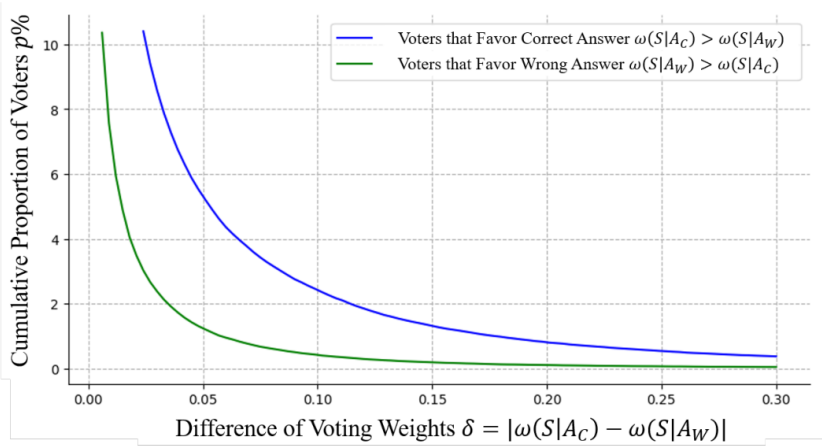
- Voter에 대한 평가를 사람들을 통해 진행
  1. Voter가 문법적인가? 완벽히 문법적이지는 않지만 이해할 수 있는 수준인가? 완전히 이해할 수 없는가?
  2. Voter가 Question에 대한 합리적인 답변인가? 합리적이지는 않지만 관련이 있는가? 완전히 무관한가?
- 대부분의 사람들이 문법적이거나, 최소한 이해할 수 있다고 답변.
- 논리적인 측면에서는 40% 정도 만이 합리적인 답변이라고 생각함.

Score	3	2	1
Grammar	84.8%	12.8%	2.4%
Logic	40.8%	25.6%	33.6%

Experiments

Voting Weight Distribution

- 그래프에서 Voter의 몇 가지 특징을 알 수 있음:
  1. Voter는 틀린 선택보다 맞는 선택을 선호.
  2. Voter의 93.5%는 어떠한 선택지도 강하게 선호하지 않음. ⇒ 두 후보 모두와 의미적으로 무관.
     앞서 Voter의 40.8%가 논리적으로 합리적이므로, 많은 Voter는 합리적이지만, 두 답변 모두와 무관.
     ⇒ 1개의 Question에 여러가지 합리적인 answer가 존재할 수 있음. Voter는 semantic의 다양성을 가짐.
  3. 맞는 답을 강하게 선호하는 Voter는 5.3%, 잘못된 답을 강하게 선호하는 Voter는 1.2%
- 논리적으로 합리적이지만 모두 비선호, 비합리적이며 모두 비선호. 모두 영향을 미치지 않음.



Q: The car ran out of gas. What happened as a result?
A_C: The driver was stranded on the road. (✓)
A_W: The driver picked up a hitchhiker. (✘)

$\omega(S_i\|A_C)$	*voter*	$\omega(S_i\|A_W)$
0.161	I had to park on a dead end road.	0.008
0.008	We picked up a hitchhiker and she drove us to the diner.	0.137
0.013	We stopped at a gas station.	0.011
0.018	It was time to hit the road again.	0.010

Conclusion

- SEQA는 Unsupervised 환경에서 Commonsense Question에 더 정확하고 강력하게 대답할 수 있는 Semantic-based 방법임.
- Question에 대해 이어지는 문장을 생성하고, 그 문장의 문맥과 유사한 답변을 선택.