가짜연구소 NLP 논문 리뷰 및 구현

Content Planning for Neural Story Generation with Aristotelian Rescoring

Date 2021.10.15

Presenter 남궁민상

Introduction
서론

기존 story generation 연구의 한계

모델이 유창한 텍스트를 만들 수는 있지만, long-range coherence를 학습하는 데 어려움을 겪고 있음

이렇게 global structure가 없는 글은 읽는 이로 하여금 글에 별 내용이 없다는 인상을 준다

이런 generic, repetitive text 문제는 massive pre-training을 이용해도 크게 개선되지 않는다 (See et al., 2019)

이 연구의 제안

아리스토텔레스는 <시학>에서 이야기의 주요 구성 요소로 다음과 같은 것들을 들었다.

이들 요소들을 반영한 rescorer를 통해, '좋은' plot structure를 만드는 모델을 설계하자

1. Event (사건의 선택과 배열)
2. Character (캐릭터)
3. Relevant Content (적절한 컨텐츠)

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C. D. (2019).
Do massively pretrained language models make better storytellers?. arXiv preprint arXiv:1909.10705.

Background
배경

Content Planning for Neural Story Generation

Neural Story Generation은 **prompt → plot**과 **prompt + plot → story**의 두 개 모델로 이루어진 파이프라인

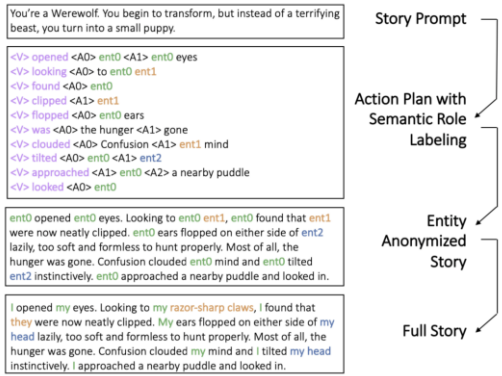
source(prompt)가 $\boldsymbol{x} = x_1 \dots x_n$, target(story)이 $\mathbf{y} = y_1 \dots y_n$일 때 intermediate representation(plot)이 $\mathbf{z}$에 대하여

$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x})$를 모델링한다

Plot Representation

Plot representation은 SRL(Semantic Role Labelling)을 사용 (Fan et al., 2019)

SRL은 문장의 술어(predicate)와 논항(argument) 구조를 밝혀내 각 논항의 semantic role을 표시한다.



참조

https://ratsgo.github.io/Korean%20linguistics/2017/07/19/valency/

https://arxiv.org/abs/1902.01109

Fan, A., Lewis, M., & Dauphin, Y. (2019).
Strategies for structuring story generation. arXiv preprint arXiv:1902.01109.

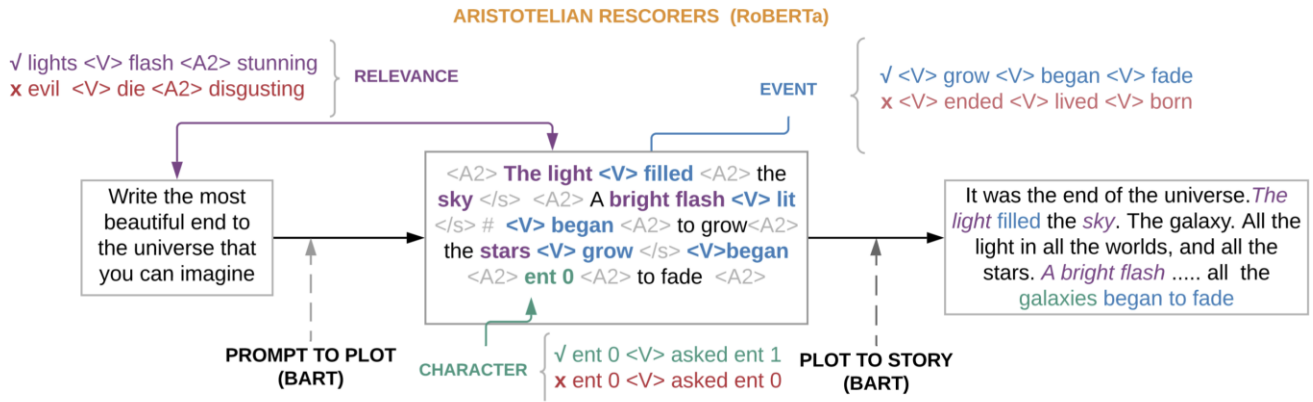
Approach

발상

Aristotelian Rescoring Model

<시학>에서 제시한 원리를 학습할 수 있는 rescoring model을 제시

1. Event rescorer
2. Character rescorer
3. Relevance rescorer



Approach

발상

Event Rescorers

Event ← Composition of an action-based verb and its subject and object (a verb, subject, object tuple)

세 가지의 positive / negative example을 만들 수 있는 방법을 고안해 실험

- Inter-sentence shuffled events

문장 자체를 셔플 / 문장 내의 구조는 그대로

Global scope에서의 구조를 학습할 수 있지만, 그만큼 어려운 task

- Intra-sentence shuffled events

문장 내의 event tuple을 셔플 / predicate-argument 관계는 그대로

Clearer and more learnable, 하지만 long-distance pattern은 학습할 수 없음

- Verb-shuffled events

문장 순서, argument는 그대로 / predicate(동사)는 셔플

문맥상 verb ordering과 verb suitability를 학습

Approach

발상

Event Rescorers

논문에서 주어진 예시

– Inter-sentence shuffled events

Positive: <A0> ent2 <V> faced <A1> ent3 </s> <A0> ent2 <V> felt <A1> the cold <A2> ent2 their backs #

Negative: <A0> ent 2 eyes <V> stayed <A2> upon the saving light </s> <A0> ent2 <V> faced <A1> ent3 #

– Intra-sentence shuffled events

Positive: <A0> ent2 <V> felt <A1> the cold <A2> ent2 their backs #

Negative: <A0> ent2 their backs <V> felt <A1> the cold <A2> ent2 #

– Verb-shuffled events

Positive: predicate의 순서가 grow → began → fade

Negative: predicate의 순서가 ended → lived → born

Approach

발상

Character Rescorers

Training example: each entity + preceding plot tokens

기존에 등장했던 entity가 다시 등장해야 하는지, 새로운 entity가 등장해야 하는지 결정

Relevance Rescorers

Training example: prompt와 plot이 pair로 존재

주어진 promp에 대해 적절한 plot words, verbs, SRL pattern을 결정

Approach

발상

Character Rescorers

Positive: ent0 <V> asked ent1

Negative: ent0 <V> asked ent0

Relevance Rescorers

‘아름다운 우주’라는 프롬프트에 대해,

Positive: lights <V> flash <A2> stunning

Negative: evil <V> die <A2> disgusting

Model Architecture

XGBoost, CNN, RoBERTa-large의 3가지 아키텍처로 실험

그 중 RoBERTa-large가 가장 뛰어난 성능을 보임

Experiment
실험

Dataset

Writing Prompts - 레딧에서 prompt를 주었을 때, 유저들이 작성한 스토리

Baseline

- **Targeted Common Sense Grounding Model:** Mao et al. (2019)
- **Knowledge-Enhanced Commonsense Model:** Guan et al. (2020)
- **Prompt to Story:** plot structure (intermediate representation) 없이 promp와 story만 파인튜닝시킨 BART 모델
- **Naïve Plot:** plot structure를 학습시키긴 하지만, Aristotelian rescoring model을 적용하지 않은 모델

Mao, H. H., Majumder, B. P., McAuley, J., & Cottrell, G. W. (2019). Improving neural story generation by targeted common sense grounding. arXiv preprint arXiv:1908.09451.
Guan, J., Huang, F., Zhao, Z., Zhu, X., & Huang, M. (2020). A knowledge-enhanced pretraining model for commonsense story generation. Transactions of the Association for Computational Linguistics, 8, 93-108.

Experiment
실험

Metrics

– Plot Structure Metrics

두 metric 모두 repetition과 관련된 지표

- Vocab:Token ratio – 스토리 컨텐츠의 originality, diversity 측정하는 데 사용
- Entities per plot – reasonableness check

– Automatic Story Metrics

- Vocab:Token ratio – 스토리 컨텐츠의 originality, diversity 측정하는 데 사용
- Inter-story trigram repetition rates – diversity metric
- Intra-story trigram repetition rates – fluency metric
- Unique verbs & % of diverse verbs

– Human Metrics

사람들에게 1~5 Likert scale을 매기도록 함

- Relevance – 프롬프트와 스토리가 밀접한 관련을 가지고 있는지 평가
- Overall quality – coherence, relevance, interestingness를 종합적으로 고려하여 이야기를 평가

Results
결과

System	Voc:Tok ratio	Entities	Avg Tok
Naive Plot	1.52	8.25	199
Aristotelian Plot	1.81	7.49	168
Gold Plot	3.59	9.26	371

Table 4: Metrics for plots

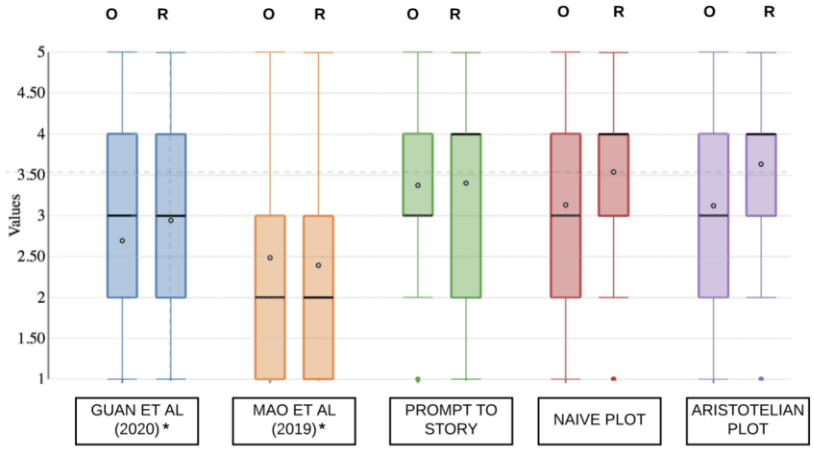
- Voc:Ratio 측면에서, Aristotelian plot이 gold plot에 더 가까움. 하지만 그 차이가 여전히 큼
- 다만, 그러면서 entities per plot은 약간 멀어짐
- entities per plot이 줄어든 것은 Aristotelian plot이 더 짧기 때문일지도

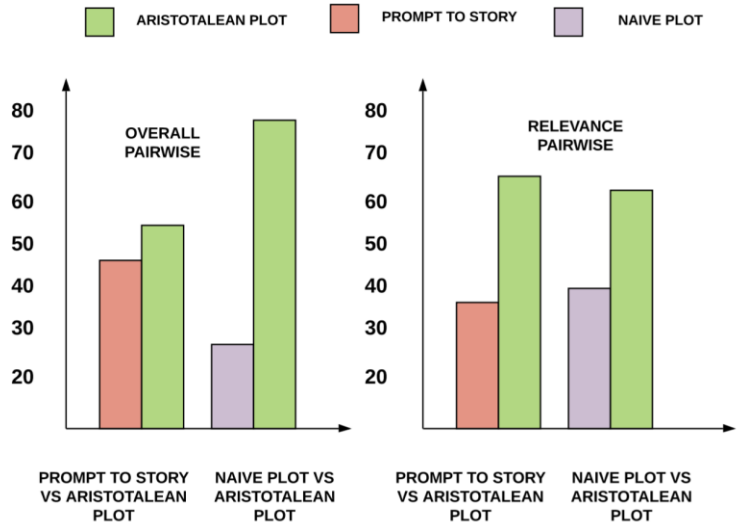
	Automatic Evaluation			
System	Voc: Tok↑	Diverse Verb %↑	Intra-Rep↓	Inter-Rep↓
Mao et al. (2019)	6.6	81.8	5.68	27.5
Guan et al. (2020)	2.3	71.9	0.60	56.1
Prompt-to-Story	1.5	68.9	0.22	65.1
Naive Plot	1.4	76.4	0.11	63.8
Aristotelian Plot	1.5	74.8	0.12	64.1

Table 5: Automated metrics for all models

- 전반적으로, 기존 모델이 diversity 관련 metric에서는 높은 성능을 보이나 fluency 관련 metric은 낮은 성능을 보임
- lexical metric에 대해서는 기존 연구들이 더 높은 성능을 보임
- 하지만 intra-rep의 결과는 기존 모델들이 심각한 lack of fluency를 보임을 시사

Results
결과





- Relevance 측면에서는 Aristotelian plot의 성능이 좋지만, overal은 prompt to story의 성능이 높다
- 다만, BART를 이용한 모델 3개는 서로 비슷비슷하다
- 선행연구의 모델은 lexical score는 높지만 human metric에서는 좋은 점수를 받지 못했다

- Pairwise comparison을 했을 때, Aristotelian plot이 다른 BART-based plot에 비해 높은 점수를 보인다

Conclusion
결론

의의 및 제언

아리스토텔레스의 story-writing principle을 반영한 rescoring model을 제안

이렇게 했을 때, 기존 모델이 가지고 있던 global structure의 부재, repetitive text 문제가 해결된 more relevant, higher quality story를 얻을 수 있었다

앞으로, Plot generation에 story principle을 적용한 연구를 할 수 있을 것이다

개인적 감상: Story NLP에서 Narratology의 역할은...?

기존 narratology 연구들은 Story NLP 분야에서 어떻게 작용할까?

이야기의 구조에 대한 연구가 loss function, 또는 rescoring function을 설계하는 데 적용될 수 있을까?

또는 대규모 데이터셋을 이용하면 신경 쓰지 않아도 되는 문제일까?

기존 narratology 연구를 어떻게 수학적/전산적 모델로 나타낼 수 있을까?