알파이솝

-SeqGAN을 이용한 동화 generating-

2018. 07. 14.

투빅스 6th 컨퍼런스

신용재, 임소정, 최재훈, 최희정

목차

주제 선정 배경

- **예술 창작과 딥러닝의 접목에 대한 관심 대두**
  - 음악, 미술, 문학 등 다양한 예술 분야에서 딥러닝 기반 예술 창작에 관한 연구가 활발히 진행 중임

< 음악 분야 >

AI 래퍼 딥비트(Deep Beat)는
딥러닝 기반으로
랩 가사를 생성함

< 미술 분야 >

한국에서도 AI Art Lab이
설립되어 인공지능으로 생성한
그림으로 전시회를 개최함

주제 선정 배경

- **문학창작 분야에서는 주로 시와 소설을 대상으로 연구가 진행됨**
  - 딥러닝 기반의 문학창작을 동화 분야로 확장하고자 함

< 소설 창작 >

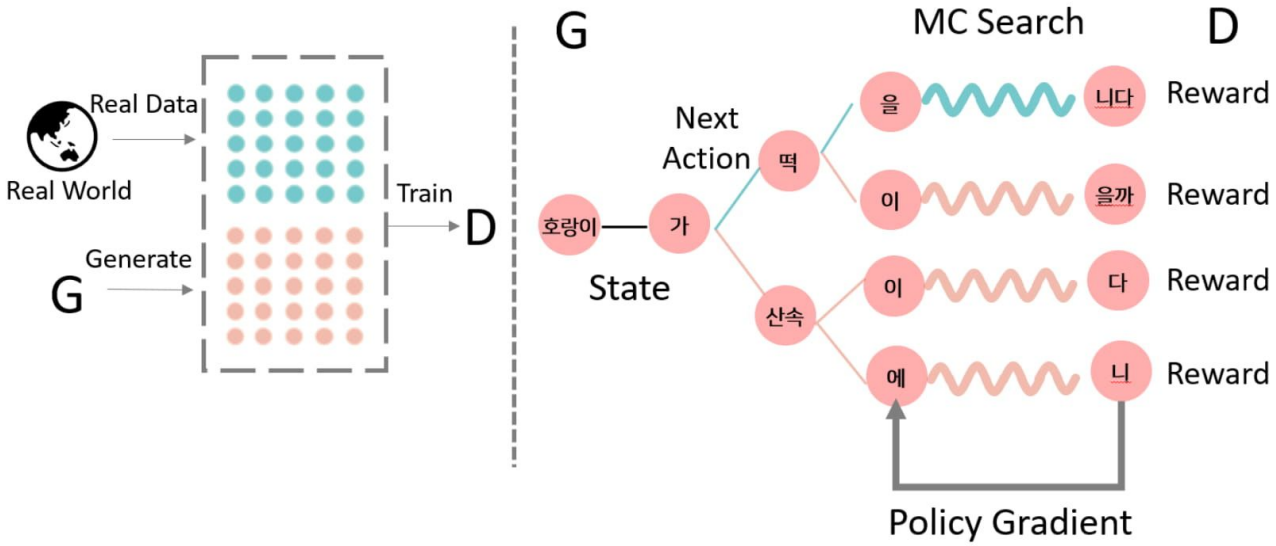
보트닉 스튜디오(Botnik studios)에
서는 해리포터 소설책을 학습시켜
새로운 챕터를 작성함

롤링 특유의 문체를 흉내내어
호그와트에 찾아온 손님에 대한
덤블도어 이야기를 만들어 냄

목차

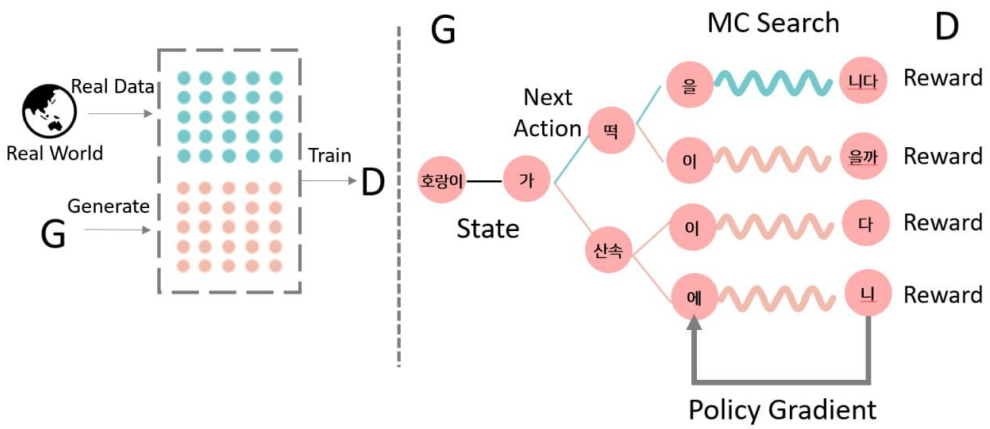
SeqGAN: Sequence Generative Adversarial Nets

- SeqGAN
  - 기존 GAN을 discrete한 text data에 적용하기 위해 기존 Generator에 강화학습을 접목한 model로 text generation을 실행하는 역할



SeqGAN: Sequence Generative Adversarial Nets

- SeqGAN



모델	Network	역할
Generator	LSTM	discrete한 token(단어)을 생성하는 역할
Discriminator	Text-CNN	Generator가 생성한 문장(sequence)이 real인지 판단해 reward를 주는 역할

목차

데이터 전처리

- Data 수집
  - 청와대 어린이 홈페이지 전래동화 200여개 크롤링 및 그 외 동화 text data 100여개 추가 수집

No.96 호랑이 형님

옛날 한 나무꾼이 가난한 나머지 호랑이한테 잡아먹히겠다며 산을 넘어가는데, 도중에 진짜 호랑이를 만났습니다. 막상 호랑이를 만나 겁이 난 나무꾼은 위기를 모면하기 위해 호랑이에게 "아이고 형님! 어머니께서 말씀하시길 저에게 형이 하나 있는데 죽어서 호랑이가 되었다고 하더니 바로 그 형님이시군요!" 그러면서 어머님이 형님을 그리워하니 당장 뵈러 가자고 하였습니다. 이에 호랑이가 그 말을 믿고서 "지금 당장 어머니를 뵙고 싶지만, 호랑이의 탈을 쓰고 그럴 수 없다"라고 거절했습니다. 그 뒤로 꼬박꼬박 호랑이가 돼지를 가져다 놓으니 그것으로 나무꾼과 어머니는 부자로 살게 되었고 나무꾼은 호랑이가 얻어준 색시에게 장가도 들었습니다.

몇 년 뒤 어머니께서 돌아가시자 호랑이가 돼지를 가져다 놓는 일도 사라졌는데 궁금해진 나무꾼이 예전 호랑이가 살던 굴에 가 보니 새끼 호랑이들이 꼬리에 흰 베를 두르고 있었습니다. 이유를 물으니 "우리 할머니는 인간인데 할머니께서 돌아가시자 아버지도 식음을 전폐하다 돌아가셨어요. 그래서 꼬리에 흰 댕기를 드리고 있는 거예요."라고 하였습니다. 나무꾼은 호랑이의 효성에 감동해서 어머니 산소 옆에 나란히 묘를 써 주었다고 합니다.

나무꾼은 위험한 상황에 순간의 지혜로 위기를 모면하게 됩니다. '호랑이 굴에 들어가도 정신만 바짝 차리면 된다'라는 말이 와 닿는 이야기입니다. 더불어 호랑이가 어머니에게 효도하는 모습이 감동적이기도 합니다. 사람도 불효자가 있는 마당에 호랑이가 효심이 지극한 모습이 우리의 부끄러운 부분을 되돌아보게 합니다.

데이터 전처리

- Word & Index Dictionary
  - word와 word vector의 상호변환을 위해 word에 고유한 index를 부여한 Word to Index와 Index to Word 사전을 구축함

✓ <PAD> : 지정한 sequence length보다 size가 작을 경우 채워주는 token
<S> : 문장의 처음에 붙이는 신호

옛날	시골	마을	에	별나	ㄴ	재주	세	형제	가	살	았	습니다
22	97	579	523	213	60	766	674	24	17	19	440	983

Word to Index ↓ ↑ Index to Word

데이터 전처리

- Tokenization
  - KoNLPy의 kkma 형태소 분석을 통해 tokenization을 진행함

< data >

옛날 시골 마을에 별난 재주 세 형제가 살았습니다.
눈이 밝은 맏이는 별별 것을 다 봅니다.
둘째는 천하장사 바윗돌도 번쩍번쩍 듭니다.

< tokenize >

[옛날, 시골, 마을, 에, 별나, ㄴ, 재주, 세, 형제, 가, 살, 았, 습니다]
[눈, 이, 밝, 은, 맏이, 는, 별, 별, 것, 을, 다, 보, ㅂ니다]
[둘째, 는, 천하장사, 바윗돌, 도, 번쩍번쩍, 드, ㅂ니다]

데이터 전처리

- Initial Word Embedding Matrix
  - pre-trained Word2Vec을 이용하여 initial word embedding matrix를 구성하고 존재하지 않는 단어들은 random initialization으로 구성함

word embedding matrix : 각 행이 특정 word의 word vector로 이루어진 array

$$\begin{pmatrix} \text{옛날(22)} & : & [\,-0.04721257, & 0.06758873, & \cdots\,] \\ \text{시골(97)} & : & [\,0.11269119, & -0.04476307, & \cdots\,] \\ & & \vdots & & \\ \text{마을(679)} & : & [\,-0.01550575, & 0.10764064, & \cdots\,] \end{pmatrix}$$

목차

SeqGAN 결과

- 동화 generation
  - train data로 학습한 SeqGAN으로 부터 문장을 생성한 결과
    ① train data와 동일한 문장
    ② 새로운 문장
    ③ paraphrasing 문장을 생성함

  - 하지만, 문장 단위로 동화 text를 생성하므로 문맥이 반영되지 못함

① 나무꾼 이 대답 하 었 다
② 마침 지나가 더 ㄴ 나무꾼 이 부채 를 줍 게 되 었 지요
③ 수염 이 허옇 ㄴ 신령님 이 나타나 이렇게 말하 는 것 이 아니 겠 어요

그때 연못물이 흔들리더니 수염이 하얀 신령님이 나타나셨어요.
신령님은 번쩍거리는 금도끼를 보여주며 물으셨어요.

SeqGAN 결과 개선

- **Bi-gram 단위의 문장 학습**
  - 기존 train data를 Bi-gram 단위로 학습시켜 문맥을 반영하고자 함

< data >

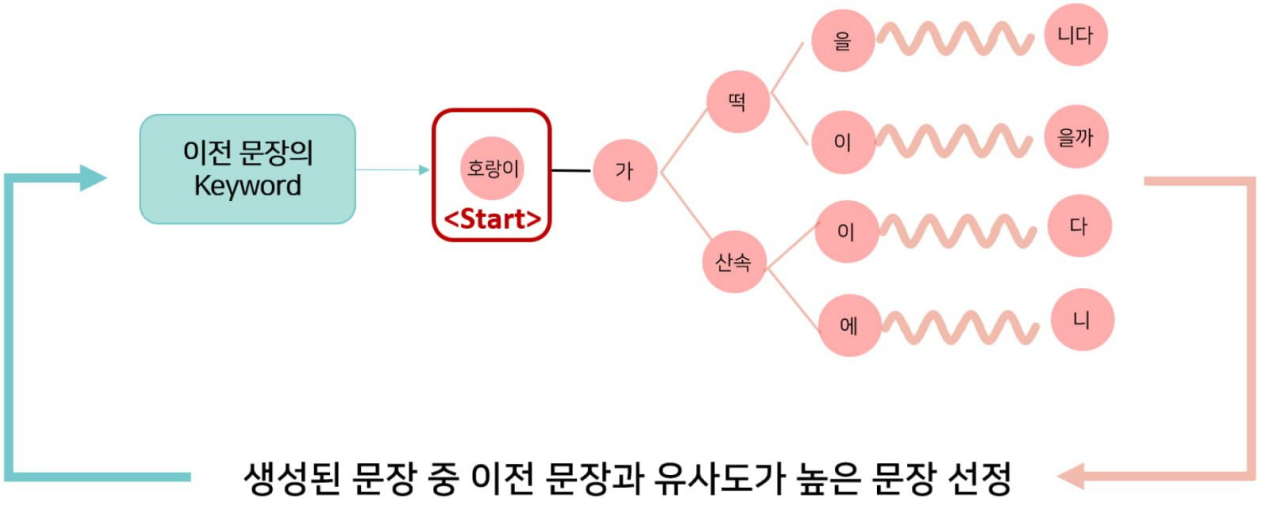
옛날 시골 마을에 별난 재주 세 형제가 살았습니다.
눈이 밝은 맏이는 별별 것을 다 봅니다.
둘째는 천하장사 바윗돌도 번쩍번쩍 듭니다.
개구쟁이 막내는 희한한 재주인데 매 맞는 재주입니다.

< Bi-gram >

[옛날 시골 마을에 별난 재주 세 형제가 살았습니다. 눈이 밝은 맏이는 별별 것을 다 봅니다.]
[눈이 밝은 맏이는 별별 것을 다 봅니다. 둘째는 천하장사 바윗돌도 번쩍번쩍 듭니다.]
[둘째는 천하장사 바윗돌도 번쩍번쩍 듭니다. 개구쟁이 막내는 희한한 재주인데 매 맞는 재주입니다.]
[개구쟁이 막내는 희한한 재주인데 매 맞는 재주입니다.]

SeqGAN 결과 개선

- **이전 문장의 키워드를 기반으로 문장 생성**
  - <S>를 input으로 받아 문장을 생성하는 SeqGAN의 특징을 이용해 이전 문장의 키워드로 <S>를 대체해 문맥을 반영하고자 함
  - 이전 문장 Sentence Vector는 문장을 구성하는 word vectors의 tf-idf 가중합으로 표현하고, 이전 문장의 키워드는 tf-idf 값을 기반으로 추출함



SeqGAN 결과 개선

- 개선된 동화 generation

키워드	생성된 문장
호랑이	호랑이 가 잡아먹 으려고 덤비 **거북이** 대답 하 였 어요
거북이	거북이가 **연못** 옆 에 있 는 저 어 나무 한참 봤 어요
연못	**산신령** 가 나타나 시 었 어요 신령님 이 웃 으며 말하 었 습니다
산신령	산신령 님 이 **회초리** 를 높이 쳐들 었 어요 아기 야
회초리	회초리 로 때리 어 버리 려고 발 을 구하 어 내 는 **동물** 들 이 모이 어 수군거리 었 습니다
동물	동물 들 이 모이 어 수군거리 었 습니다 여우 가 **목청껏** 소리치 자 구경꾼 이 모여들 어
목청껏	목청껏 소리 말하 자 자신 착하 게 하 었 는데 혼자 서 **중얼거리** 곤 이 었 다
중얼거리	중얼거리 며 호랑이 의 이야기 를 듣 어 오 았 어요

SeqGAN 결과 개선

- 개선된 동화 generation(Cont`d)

호랑이가 잡아 먹으려고 덤비자 거북이가 대답했어요.

거북이가 연못 옆에 있는 나무를 한참 봤어요.

산신령이 나타나더니 웃으며 말했어요.

산신령님이 회초리를 높이 쳐들었어요.

회초리로 때리려고 하자 동물들이 모여 수군거렸어요.

여우가 목청껏 소리치자 구경꾼들이 모여들었고

산신령은 자신이 착하다고 혼자 중얼거리면서 호랑이의 이야기를 들었어요.

감사합니다