특집 01

딥러닝 기반의 한글 문장 생성 기법*

손성환 (로앤컴퍼니), 강승식 (국민대학교)

목 차

1. 서 론

인간의 사고 내용을 자연스러운 문장으로 생성하는 과정은 매우 복잡할 뿐만 아니라 문장을 구성하는데 사용되는 어휘들의 조합은 무한에 가깝다. 자연스러운 문장을 생성하는 연구로 딥러닝 기법 이전의 연구에서는 시나 노래 가사 등 길이가 짧고, 함축적인 표현이 허용되는 텍스트를 n-gram 언어 모델이나 규칙 기반으로 생성하는 연구들이 진행되었다[1,2,3,4]. 이 방식에서는 좌우 문맥에 따른 빈칸을 채우거나, 특정 패턴으로 생성하는 방법이 사용된다. 딥러닝 기법은 데이터의 순서정보 패턴을 학습하는 순환 신경망(RNN)으로 다양한 연구가 진행되었고, RNN의 발전된 모델인 LSTM과 GRU가 연구자들의 관심을 많이 끌었다[5,6].

LSTM은 순환 신경망에 과거 데이터의 정보를 유지하는 장기 메모리인 cell state와 이전 정보를 얼마나 cell state 반영할 것인지 정하는 forget gate, 입력으로 들어오는 새로운 정보를 얼마나 cell state에 반영할 것인지 결정하는 input gate를 추가하여, RNN이 가지고 있던 오래된 과거의 정보를 잊어버리는 그래디언트 소실(vanishing gradient) 문제를 완화시켰다. GRU는 LSTM을 간소화시킨 모델로 상대적으로 학습에 사용되는 데이터가 적은 경우 좋은 성능을 보였다. 이와 같은 RNN 모델을 이용한 중국의 한시 생성, 영문시 생성 랩 가사 생성 실험들이 진행되었으며, 문장의 구조까지 습득하여 생성하는 결과를 낳았다[7,8,9]. 그러나 여전히 의미적, 문맥적 오류와 그래디언트 소실 문제도 남아있었다.

딥러닝 모델의 발전으로 다양한 생성 모델들이 제안되었다[10]. GPT-2는 transformer의 디코더를 사용한 사전학습 모델로 생성 모델의 비약적인 발전을 보여주는 우수한 모델이다[11,12]. GPT-2는 약 40GB에 이르는 방대한 규모의 영어 텍스트를 학습하여 생성작업을 수행하였는데 해당 모델이 사람들에 의해 가짜 뉴스나 잘못된 정보 유포와 같이 악용될 수 있다는 우려 때문에 모델의 일부 부분만 공개하기도 하였다. GPT-2를 이용하여

* 이 연구는 손성환(2020)의 석사학위 논문의 주요 내용을 요약한 것임[33].

완성도 높은 텍스트를 생성하기 위해서는 기본적인 딥러닝 학습 외에 추가 학습이 필요하다. 본 연구에서는 카테고리별 특수 토큰을 입력 데이터에 추가함으로써 생성되는 문장이 카테고리별로 생성되는 결과를 보이고자 한다. 이처럼 GPT-2 기반의 생성모델을 한국어 신문 기사 생성에 적용하여 실험을 실시하였다.

2. 딥러닝 기반의 자연어처리 모델

통계적 언어 모델은 토큰의 순서 정보에 따른 확률분포라고 할 수 있다. 언어 모델은 크게 통계를 이용한 방법과 신경망 모델을 사용하는 방법으로 나누어진다. 통계를 사용한 방법은 학습 데이터의 토큰들의 빈도수 및 n-gram 확률 정보를 활용하여 이전 토큰들의 순서 정보를 토대로 다음 토큰을 예측하는 방법이다. 따라서 데이터를 토대로 이전에 등장한 어절 토큰들을 기반으로 다음에 등장할 어절 토큰을 예측할 수 있다. 이때 목표 어절 토큰 이전의 모든 토큰 어절 개수를 가지고 확률을 계산하기 보다 임의의 n개를 기준으로 토큰을 분할하여 확률을 계산하여 사용하는 방법이 있는데, 그것이 바로 n-gram을 이용하는 것이다.

바이그램 모델은 마코프 모델(Markov model)이라고 불린다. 마코프 모델은 1913년경 러시아의 수학자 마코프가 문헌에 나오는 문자들이 배열된 순서에 대한 모델을 구축하기 위해 고안된 마코프 체인(Markov chain)에서 기인한 것이다. 마코프 체인은 목표 문자의 등장 확률이 이전에 등장했던 문자들의 긴 순서 정보가 필요없이 바로 직전의 문자로 도출될 수 있다는 것이다. 통계적인 언어 모델인 n-gram 언어 모델은 이전 n-1개의 토큰 나열 확률에 따라 다음에 등장할 확률이 높은 토큰을 선택하는 방식으로 텍스트 생성이 가능하다. 그러나 생성하는 텍스트의 토큰 수가 많아질수록 토큰 배열 경우의 수가 무한에 가까운 언어의 특성 때문에 생성된 결과는 언어가 가지는 문법적인 패턴 및 의미를 전부 반영할 수 없다. 그리고 n이 커짐에 따라 n-gram 확률 모델의 크기가 기하급수적으로 비대해지고, 데이터 대부분의 n-gram 확률이 독립적으로 하나만 존재하게 되는 문제가 존재한다.

LSTM과 GRU 이후 딥러닝을 이용한 다양한 언어 모델들이 등장하였다. 순환 신경망을 이용하여 인코더-디코더 구조를 구현한 Seq2Seq 모델은 인코더에서 입력된 정보가 축약되고, 축약된 정보를 기반으로 디코더에서 값을 출력하는 모델이다. 이 기법은 노래 가사 및 텍스트를 생성하거나, 텍스트 내용을 요약 또는 댓글의 긍정 부정 스타일 변환과 같은 연구에 적합하다[13-17]. Seq2Seq 모델은 인코더 부분에서 많은 데이터가 입력될수록 축약되어 소실되는 정보가 많아지는 병목 현상을 보완하기 위해 입력 데이터에서 상대적으로 중요한 정보를 더 많이 반영시키도록 하는 attention 기법을 적용하였다[18].

Kingma(2013)은 RNN 기반의 VAE(Variation Auto-Encoder) 생성 모델을 제안하였고, Miao (2016)는 Neural Topic Model을 제안하였다[19,20]. Wang(2019)은 이를 참고하고 VAE를 사용하여 특정 주제 카테고리에 맞는 생성하는 모델을 소개하였다[21]. 이 모델은 주제별 문서들에 등장하는 어휘들의 분포 및 빈도수를 입력으로 각 주제별 문서들의 특성들을 함축하여 생성 정보에 포함, 생성하게 하였다. Wang은 이 모델이 각 주제에 맞는 글을 생성하였으며, 모델 구조 특성상 문서 요약 모델로 확장하여 좋은 성능을 보인다고 서술하였다.

메모리 네트워크는 질의응답을 위해 제안되었는데 학습에서 데이터의 중요한 정보를 메모리라는 형태로 저장하고 작업을 수행할 때 attention을 사용하여 입력에 적합한 부분을 메모리에 저장된

정보에서 상대적으로 많이 반영하도록 하여 결과를 도출하는 모델이다[22]. Sukhbaatar(2015)는 스토리가 있는 텍스트를 입력으로 질의응답을 하는 작업을 메모리 네트워크 모델을 개선하여 종단간(end-to-end) 학습 모델을 제안하였는데 이는 깊은 LSTM 모델과 유사한 성능을 보였다[23]. 그리고 Lin(2019)은 메모리 네트워크를 이용하여 대화 문장을 생성하는 모델을 제안하였다[24]. 포인터 네트워크는 RNN을 기반으로 한 attention 매카니즘을 사용하여 입력값에 대응되는 위치들의 결과들을 출력하는 딥러닝 모델이다. See(2017)는 해당 모델을 사용하여 입력 텍스트에 대한 요약문을 생성하는 모델을 소개하기도 하였다[25,26].

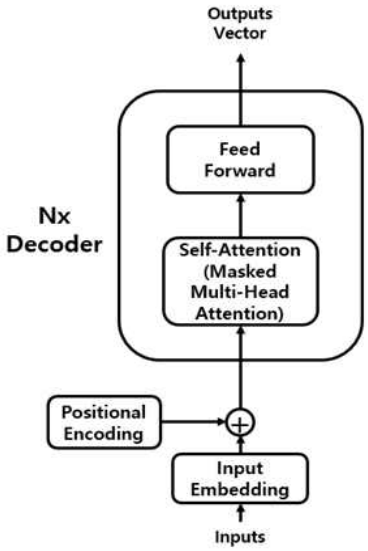
Transformer는 기존의 Seq2Seq의 인코더-디코더 구조를 지닌 사전학습 모델로 구글의 번역 모델로써 큰 성과를 보였다. 이 방식은 번역 대상 언어의 텍스트를 인코더에 입력으로 디코더에서 번역하고자 하는 언어의 출력으로 텍스트를 번역하였으며, 당시 상당한 성과를 입증했다. 이후 transformer 모델을 사용한 다양한 모델들이 등장하였으며, 그 중 하나가 BERT(Bidirectional Encoder Representations from Transformers)이다[27]. BERT는 transformer의 인코더만 사용하는 사전학습 모델로 대량의 데이터를 MLM(Masked Language Model), NSP(Next Sentence Prediction)을 통해 사전학습을 한다. 사전학습 모델은 대량의 데이터로 해당 언어가 가진 기본적인 패턴들을 학습하고, 이후에 사용자의 의도에 맞게 추가학습하여 사용자 의도에 맞는 작업의 성능을 한층 끌어올리는 방법으로 전이학습(transfer leaning) 모델이라고도 불린다. BERT는 다양한 자연 언어 처리 작업에서 그 당시 최상의 성과를 보였다.

이후 BERT를 이용한 생성 기법 연구가 많이 등장하였고, 이주성(2019)은 BERT를 이용하여 한국어의 긍정, 부정 문장 스타일을 서로 변환시켜 생성하는 모델을 소개하였다[28]. Transformer 기반의 언어 모델로 구글은 T5 모델을 공개하였는데, 해당 모델은 110억개의 파라미터를 가진 모델로써 TensorFlow가 공개한 약 6.9TB의 방대한 데이터를 사용하였다[29]. 2020년에 마이크로소프트는 T-NLG(Turing-NLG) 모델을 발표하였으며, 파라미터 개수는 약 170억개이다. OpenAI는 2019년 GPT-2 이후 2020년 GPT-3는 파라미터 개수가 지금까지 나온 딥러닝 언어 모델을 통틀어 가장 큰 약 1,750억개라고 소개했다[30]. 최근 발표되는 transformer 기반의 딥러닝 언어 모델은 파라미터 개수가 기하급수적으로 증가하는 추세를 보이는 것을 확인할 수 있다.

국내에서는 BERT를 한국어 말뭉치로 학습한 KoBERT와 함께 GPT-2 모델을 한국어에 적용한 KoGPT2가 공개되었으며, 서울대학교 신효필 교수팀은 한국어 원시 말뭉치를 학습시킨 KR-BERT와 한국어 감성 관련 작업을 위해 KOSAC이라는 감성 말뭉치를 적용한 KR-KOSAC-BERT를 공개하였다.

3. GPT-2 기반의 문장 생성 기법

GPT-2는 transformer의 디코더 부분을 사용한 self-attention 기반의 사전학습 및 전이학습 기법의 딥러닝 모델이다. (그림 1)은 transformer의 디코더 부분에 대한 구조도이다. GPT-2는 미등록어 문제를 완화시키고 통계적으로 빈번하게 등장하는 내부단어를 기준으로 토큰화하기 위해 sentencePiece 토큰화 방식을 사용하였다[31]. SentencePiece는 한국어와 같이 토큰 사전이 비대해지는 교착어를 효율적으로 토큰화할 수 있는 특징을 가지고 있다. 학습 데이터는 sentencePiece 모델에 의해 토큰화되어 input에 입력으로 주어진다.



(그림 1) 트랜스포머 디코더 구조

3.1 Positional Encoding

순환 신경망이 토큰의 순서 정보를 토큰을 순차적으로 입력하면서 특성(feature)을 누적시키는 방식을 사용했다면, transformer 구조를 사용하는 BERT나 GPT-2의 경우 positional encoding 벡터를 사용한다. 입력되는 토큰들은 정해진 크기의 embedding 벡터로 입력되고, 토큰 위치에 따라 각기 다른 동일한 크기의 positional encoding 벡터를 더하여 순서 정보를 추가한다. 수식 (1)은 토큰 위치(pos)에 따라 positional encoding 벡터를 계산하는 식이다.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (1)$$

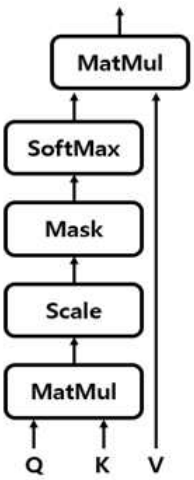
d_{model}은 embedding 벡터의 크기이고, i는 positional embedding 벡터의 차원이다. Positional embedding 벡터의 짝수 부분은 sine 함수를 사용하여 계산하고, 홀수 부분은 cosine 함수로 계산하여 토큰 위치마다 각기 다른 벡터를 만들어낸다. 따라서 같은 토큰이라도 토큰 위치에 따라 다른 embedding 벡터로 표현될 수 있다. 이 positional embedding은 pos에 따른 one-hot 벡터를 사용하거나 positional embedding 자체를 학습시켜서 사용하기도 하는데, GPT-2에서는 후자의 방식으로 토큰 위치에 따른 정보를 반영한다.

3.2 Self-Attention

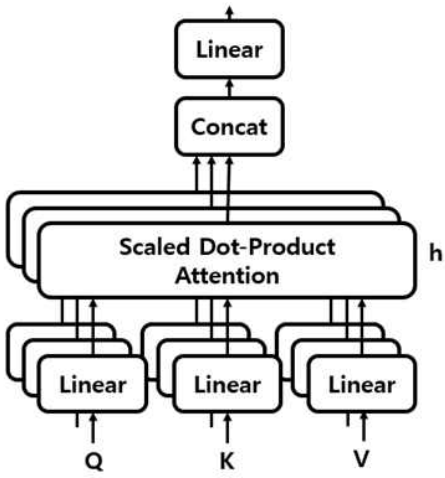
Attention은 입력되는 토큰들의 정보 중에서 손실(loss)을 줄이기 위해 특정 정보를 더 반영하도록 하는 방법이다. 즉, 입력되는 값의 중요한 특징을 반영하여 입력되는 값의 특성을 embedding 공간에 표현하고, 이로써 학습의 정확도를 높이는 역할을 한다[32].

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Masked multi-head attention은 BERT에서 사용된 transformer 인코더의 multi-head attention에 mask 기법을 적용한 것이다. (그림 2)는 multi-head attention을 구성하는 scaled dot-product attention 구조이다. 입력값은 d_k차원인 Q(query), K(key), 그리고 d_v 차원의 V(value)로 이루어져 있다. Q와 K에 대한 내적(dot-product)을 계산하고 각각을 $\sqrt{d_k}$ 로 나누어 준다. 즉, $\sqrt{d_k}$ 로 Scaling을 진행하는 것이다. 그리고 softmax를 적용하고 그 결과값과 V와의 내적을 진행한다. Softmax 결과값과 V를 내적하면 Q, K와 유사한 V일수록 결과값이



(그림 2) Scaled dot-product 어텐션



(그림 3) 다중 헤드 어텐션 구조

더 큰 값을 가지므로 중요한 정보를 더 많이 반영하게 되어 attention 역할을 하게 된다.

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \dots, head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

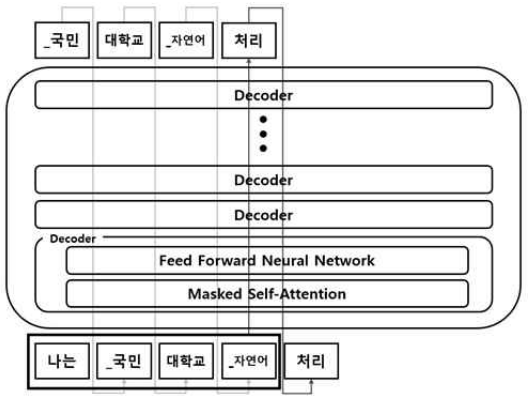
(그림 3)은 multi-head attention의 구조도이다. Multi-head attention에서 Q, K, V는 동일한 값이다. 그러나 각기 다른 가중치 매트릭스 (W_i^Q, W_i^K, W_i^V)를 곱한(linear projection) 값이 scaled dot-product attention에 입력으로 주어지게 된다. 이 과정을 각기 다른 가중치로 h번 진행하는데, 이는 곧 다양한 관점에서 attention을 진행하는 것이라고 볼 수 있다. h번 반복한 결과들은 concat을 통해 선형 투영을 거쳐 Q, K, V와 동일한 차원으로 도출된다.

Transformer 디코더에 사용되는 masked multi-head attention은 수식 (2)의 softmax의 입력값을 $-\infty$으로 설정(masking out)한 scaled dot-product attention을 적용한 것으로 i 번째 위치에 대한 attention을 계산할 때, multi-head attention이 전체 모든 데이터를 전부 반영한다면, masked multi-head attention은 i 번째 이후에 있는 모든 위치의 입력값을 무시하도록 하는 것이다. 즉, 실제 입력값이 존재하는 부분에만 집중하도록 하여, 토큰을 하나씩 생성할 때 생성된 토큰이 다시 입력값으로 주어져서 첫 토큰부터 이전에 생성된 토큰까지 입력값으로 모델에 입력하여 다음 토큰을 생성하는 과정을 반복하는 것으로 볼 수 있다.

Masked multi-head attention에서 도출된 결과가 FFNN(Feed-Forward Neural Network)을 거치면, 하나의 transformer 디코더를 통과한 것이다. 하나의 디코더만 보더라도 상당한 크기의 파라미터(parameter)가 필요한 것을 알 수 있다.

4. GPT-2 모델을 이용한 신문기사 생성 실험

GPT-2는 transformer 디코더를 n 층으로 쌓아서 모델을 구현한다. 12층을 쌓으면 GPT-2 small, 24층 쌓으면 GPT-2 medium, 36층 쌓으면 GPT-2 large, 마지막으로 48층을 쌓으면 GPT-2 extra large로 구분한다. (그림 4)는 GPT-2의 모델 구조를 나타낸다. '나는 국민대학교 자연어처리 연구



(그림 4) GPT-2 모델 구조와 생성 과정

실 학생이다.'라는 문장 생성을 가정해 보자. 문장이 ['나는', '_국민', '대학교', '_자연어', '처리', '_연구실', '_학생', '이다', '.']로 토큰화 되고 '나는'을 입력으로 넣는다면, '_국민' 토큰은 직관적으로 ['나는']을 입력으로 생성하는 것이고, '대학교' 토큰은 이전에 생성된 '_국민'이 입력값으로 추가되고 ['나는', '_국민']을 입력으로 '대학교' 토큰이 생성되는 것으로 볼 수 있다. (그림 4)는 ['나는', '_국민', '대학교', '_자연어']라는 입력으로 '처리' 토큰을 생성하는 것을 보여준다. 따라서 GPT-2의 입출력 구조는 LSTM의 입출력 구조와 일정부분 일치하는 것을 확인할 수 있다.

GPT-2의 extra large 모델은 파라미터가 약 15억개로 이루어져 있다. BERT large 모델의 파라미터가 약 3.4억인데 비해 4배 이상의 파라미터 수를 가지고 있는 것이다. GPT-2는 사전학습 모델로 미리 대용량 데이터를 학습하고, 이후에 fine-tuning을 거쳐 사용자가 원하는 결과를 위한 추가학습을 진행한다. 즉, 사전학습으로 기본적인 언어적 특성을 익히고 이후에 구체적인 작업을 위해 조정 학습을 진행하도록 하여 성능을 높이는 구조이다. 실제 GPT-2 extra large 모델은 약 4천500만 링크의 텍스트 데이터와 신문 기사, 소설 데이터를 합하여 약 40GB의 대용량 데이터를

〈표 1〉 학습 말뭉치의 카테고리별 주요 키워드 15개

	경제	국제	문화	사회	스포츠	정치
1	공시	일본	서울	경찰	경기	의원
2	한국	중국	배우	검찰	리그	대통령
3	사업	미국	오후	혐의	감독	북한
4	기업	대통령	영화	서울	선수	한국
5	시장	홍콩	포토	조사	서울	대표
6	서비스	트럼프	사람	장관	한국	장관
7	기술	시위	한국	사건	오후	후보자
8	스톡	한국	진행	교수	두산	일본
9	개발	경찰	방송	지난	승리	미국
10	통해	지난	무대	수사	시즌	정부
11	변동	정부	작품	사고	인천	검찰
12	지난	현지	모습	한국	체육관	조국
13	전자	북한	통해	학생	안타	대해
14	미국	시간	사진	오후	라운드	국회
15	국내	총리	지난	재판	기록	우리

〈표 2〉 생성된 문장의 카테고리별 주요 키워드 15개

	경제	국제	문화	사회	스포츠	정치
1	공시	중국	포토	검찰	경기	검찰
2	보이	미국	무대	경찰	선수	후보자
3	종목	일본	타임	장관	두산	장관
4	스톡	한국	배우	교수	한국	대통령
5	매도	트럼프	오후	후보자	감독	조국
6	변동	대통령	서울	조사	승리	북한
7	매수	홍콩	지난	조국	안타	한국
8	기술	정부	참석	수사	일본	일본
9	매출	북한	브랜드	사건	기쁨	대표
10	발표	시위	한국	혐의	서울	수사
11	전자	협상	트와이스	오후	시즌	의원
12	미국	무역	행사	지난	리그	개혁
13	지난	수출	진행	서울	오후	국민
14	기관	경찰	일본	관련	타점	법무부
15	한국	지난	그룹	의혹	김경문	미국

〈표 3〉 신문기사 생성 실험 결과

경제 기사 생성 결과
한국농수산식품유통공사가 올해 3분기에 영업손실 8억원으로 전년 동기에 비해 적자 폭이 줄었다. 한국농수산식품유통공사(aT)는 올해 3분기 기준 영업이익이 지난해 같은 기간보다 17.3% 증가한 15조5489억원에 그쳐 전체 매출액 증가율 약 46.4%를 기록했다고 8일 밝혔다. 지난해 3분기와 비교해 적자로 전환한 것은 매출 증가로 이어졌는데, 이 역시 같은 기간 영업이익이 전년 동기 대비 23.7%나 증가해 손실이 늘어났다. 2018년 3분기부터 올해 1분기까지 5분기까지 누적 실적도 1.6%에 달했다. 올 3분기까지 누적 누적 매출액은 2조6469억원으로 2017년 3분기 1413억원대비 38억원(5.4%) 줄었다. 전년 동기 대비 적자폭은 18.4%로 작년 3분기 2356억원 보다 32억원 증가했다. 같은 기간 매출액은 2조4971억원으로 전년 동기 대비 5611억원(17.4%) 증가했다. 내수 증가율은 2016년 3월(1.4%) 이후 3년 만에 가장 높은 수준이다.
국제 기사 생성 결과
도널드 트럼프 미 대통령이 '인종 혐오론' 비판을 받는 인종주의자들을 위한 '우크라이나 스캔들'에 대해 직접 입장을 낼 뜻을 내비쳤다. 지난 19일(현지 시각) 트위터 계정 '실화이트(Future)' 등에 게재된 영상 속 민주당 상원의원인 팀 모리슨 주니어라는 이름의 영상을 보면, '우크라이나 스캔들'을 비판하는 글이 이어졌다. 트위터에는 "나쁜 사람들 때문에 우리의 총기규제를 지지하는 사람이 많다. 그들이 어떤 생각을 할 때 우리가 더 많은 것을 얻을 수 있는지를 보여준다"는 내용의 동영상이 올라왔다. 이날 트럼프 재선 캠프 관계자, 트럼프 선거 캠프 관계자 등이 이 영상에 '저걸(No Kid)은 공산주의자다'라는 해시태그도 달렸다. 페이스북은 "트럼프의 '우크라이나 스캔들'을 가장 많이 보여주는 인물은 미국 공화당 소속 제임스 서먼(No John)과 커트 볼츠(Dover)"이라고 소개했다. 해당 영상은 도널드 트럼프 행정부 내에서는 '우크라이나 의혹'으로 불리며 진보 성향의 정치학자들도 '왜냐' '우크라이나 스캔들' 같은 비판을 받았다고 밝혔다. 트럼프 대통령은 이날 트위터에 "우리나라에 매우 좋지 않은 일이 계속되고 있다"며 "미국 내 증오주의자들을 위해 그들이 하는 일에 대해 말하고 있다"고 썼다.
문화 기사 생성 결과
'동풍'은 이날 오후 5시 서울 강남구 청담동 플래그십 스토어에서 일제히 '2019 스탠드' 진행된다. 강남에서 진행된 새로운 스페셜 이벤트에 대한 자세한 내용은 공식 홈페이지에서 확인할 수 있다. 전날 오후 서울 강남구 청담동 플래그십 스토어에서 진행된 한 브랜드 행사에 참석한 두 사람은 '어벤저스 인디아'에서 최초로 공개된 '2019 스탠드'를 위해 무대에 섰다. 배우 김희애가 포토타임을 갖고 있다. 그룹 CIX가 1일 오후 서울 강남구 신사동 플래그십 스토어에서 진행된 한 브랜드 행사에 참석해 포토타임을 갖고 있다. '2019 스탠드'는 한 공식 뷰티 브랜드 포토콜이 17일 오후 서울 강남구 청담동 플래그십 스토어에서 열렸다. 리처드 리빗과 크리스티의 일상이 펼쳐졌다. 이날 론칭행사에서 세 사람은 '매운맛을 내 운명이라는 메시지로 뭉클함을 더했다'고 포즈를 선물했다.
사회 기사 생성 결과
전남도, 청년내일 돌봄 지원단 워크숍'신북방여성센터' 지원 본격화전남도 청년내일 돌봄 지원단 워크숍이 도내 청년들에게 일자리 연계와 공공 지원 정책 수립 기회를 제공하는 자리가 될 전망이다. 전남도는 16일 "'함께 일하는 마을' 사업 운영을 위해 이날 오전 9시부터 11시까지 2시간 5분 동안 광주전남전북전남지역 청년내일 돌봄 전담관을 운영한다고 밝혔다. 이번 워크숍으로는 전남도에 사는 주민 70여명, 광주전남지역 20여명이 참여해 직접 방문해 다양한 취업 관련 체험을 하고 구직활동에 나서는 자리였다. 이에 따라 청년내일 돌봄에 필요한 각종 서비스 제공기관 및 공공지원 사업 진행 과정에서 겪는 어려움을 해소하고, 청년이 원하는 일자리 사업 추진에 도움이 되길 기대했다. 문신언 전남도 일자리정책과장은 "청년내일 돌봄 전담관을 통해 현장 의견을 직접 듣는 좋은 기회가 되도록 하겠다"며 "이번 워크숍은 도내 청년을 중심으로 한 청년내일 돌봄 사업이 지역의 청년들의 꿈과 희망을 키우는 힘이 될 것"이라고 말했다.
스포츠 기사 생성 결과
2019 KBO리그 한화생명이 KT 위즈와 경기에서 승리했다. 한화는 1일 수원 KT 위즈파크에서 열린 2019 신한은행 마이카(WBSC) 4강 2연승을 거뒀다. 한화는 4강행 마지노선인 지난 2일 서울 삼성과의 경기에서 선발 투수 브록 다익손의 호투와 카를로스 페게로를 나란히 6이닝 동안 7안타 3탈삼진 2실점으로 틀어막았다. 시즌 7승째(6패)를 수확했던 한화는 5연승을 달리며 5강로에 안착했다. 4위 키움은 두산 베어스(00승1패)를 꺾었다. 한화는 2연승과 함께 시즌 9승째(9패)를 수확했다. 한화는 선발 등판해 5이닝 1실점(1자책점)으로 호투했으나 팀 패배를 막지 못했다.

sentencePiece로 약 130,000개의 내부단어 토큰을 만들고, 사전학습을 진행하였다.

본 연구에서는 GPT-2 모델을 한국어에 맞게 수정하여 신문기사 데이터 KCC150, KCCq28, KCC940을 정제하여, GPT-2 medium 모델에 사전학습을 진행하였다. KCC 데이터는 온전한 문장을 모아 놓은 데이터의 특성상 문맥정보가 고려되지 않는다는 문제와 데이터가 충분하지 않다는 문제가 있다. 따라서 추가로 수집한 문맥정보가 유지된 신문기사 데이터를 정제하여 추가 사전학습을 진행하였다. 이후 경제, 국제, 문화, 사회, 스포츠 그리고 정치 6가지 카테고리의 신문기사를 개별적인 카테고리 특수 토큰을 입력으로 학습을 진행하여, 카테고리별 한국어 신문기사 생성 모델을 구현하였다. 생성 실험 결과는 〈표 1〉~〈표 3〉과 같다.

5. 결 론

LSTM 기법과 GPT-2 생성 모델을 사용하여 카테고리별로 신문기사를 생성하는 연구를 진행하였다. 본 논문에서는 문장 완성도가 높은 데이터와 문맥 정보가 남아 있는 데이터를 이어서 사전학습을 수행하였고, 이후 카테고리별로 특수 토큰을 부여하여 학습을 진행하였다. TF-IDF 상위 단어 결과와 문서 분류 모델을 통하여 생성된 결과가 훈련 데이터에 맞춰 카테고리별로 생성하는 양상을 확인하였고, 인간 평가를 통해 생성 결과의 문장 단위 완성도와 전체 문맥 완성도가 LSTM 모델보다 더 높은 성능을 보인다는 것을 확인하였다. 그러나 특정 카테고리에서 반복되는 어구 및 문장구조가 많은 경우 일부 생성 결과에서도 훈련 데이터와 중복되는 결과나 특정 어구를 반복해서 생성하는 문제를 보였다. 또한, 정치/사회 카테고리에서는 그 당시 크게 이슈가 됐던 주제 위주로 생성되는 경향을 보이며, 문서 분류 모델 정확도가 상대적으로 낮은 양상을 보였다. 이것은 3개 신문사의 카테고리 분류 기준이 다르고, 카테고리별 데이터가 충분하지 않기 때문이다. GPT-2의 extra large 모델은 약 15억 개의 파라미터에 약 40GB의 다양한 종류의 텍스트 데이터를 사전 학습한다. 향후 연구에서는 데이터에 대한 정밀한 정제와 함께 충분한 대량의 데이터를 수집하여 학습할 필요가 있다.

참 고 문 헌

[1] Barbieri, G., F. Pachet, P. Roy, M. Degli Esposti, “Markov constraints for generating lyrics with style,” 20th European Conference on Artificial Intelligence, pp.115-120, 2012.

[2] Oliveira, H. G., R. Hervas, A. Diaz, P. Gervas, “Adapting a generic platform for poetry generation to produce spanish poems,” 5th International Conference on Computational Creativity, pp.63-71, 2014.

[3] Addanki, K., D. Wu, “Unsupervised rhyme scheme identification in hip hop lyrics using hidden markov models,” Statistical Language and Speech Processing, pp.39-50, 2013.

[4] Malmi, E., P. Takala, H. Toivonen, T. Raiko, A. Gionis, “DopeLearning: A computational approach to rap lyrics generation,” arXiv preprint arXiv: 1505.04771, pp.195-204, 2015.

[5] Hochreiter, S., J. Schmidhuber, "Long short-term memory," Neural computation, Vol.9(8), pp.1735-1780, 1997.

[6] Chung, J., C. Gulcehre, K. Cho, Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence

modeling," arXiv preprint arXiv:1412.3555, pp.1-9, 2014.
[7] Zhang, X., M. Lapata, “Chinese poetry generation with recurrent neural networks,” EMNLP 2014, pp.670-680, 2014.
[8] Potash, P., A. Romanov, A. Rumshisky, “GhostWriter: Using an LSTM for automatic rap lyric generation,” EMNLP 2015, pp.1919-1924, 2015.
[9] Yan, R., “i, poet: automatic poetry composition through recurrent neural networks with iterative polishing schema,” 25th International Joint Conference on Artificial Intelligence, pp.2238-2244, 2016.
[10] Bowman, S. R., L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, S. Bengio, "Generating sentences from a continuous space," arXiv preprint arXiv:1511.06349, 2015.
[11] Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI Blog, Vol.1(8), 2019.
[12] Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, S. Agarwal, et al. “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, 2020.
[13] Sutskever, I., O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks," In Advances in neural information processing systems, pp.3104-3112, 2014.
[14] Fan, H., J. Wang, B. Zhuang, S. Wang, J. Xiao, "A hierarchical attention based seq2seq model for Chinese lyrics generation," In Pacific Rim International Conference on Artificial Intelligence, pp.279-288, 2019.
[15] Egonmwan, E., Y. Chali, "Transformer and seq2seq model for paraphrase generation," In Proceedings of the 3rd Workshop on Neural Generation and Translation, pp.249-255, 2019.
[16] Zhang, Y., Y. Wang, J. Liao, W. Xiao, "A hierarchical attention seq2seq model with copynet for text summarization," In 2018 International Conference on Robots & Intelligent System(ICRIS), IEEE, pp.316-320, 2018.
[17] 최형준, 나승훈, "Delete-MASS Gen: MASS를 이용한 단어 n-gram 삭제 및 생성 기반 한국어 스타일 변환", 한국정보과학회 학술발표논문집, pp. 1433-1435, 2019.
[18] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, et al. "Attention is all you need," In Advances in neural information processing systems, pp.5998-6008, 2017.
[19] Kingma, D. P., M. Welling, “Auto-encoding variational Bayes,” arXiv preprint arXiv:1312.6114, pp.1-14, 2013.
[20] Miao, Y., L. Yu, P. Blunsom, “Neural variational inference for text processing,” In International conference on machine learning, pp.1727-1736, 2016.
[21] Wang, W., Z. Gan, H. Xu, R. Zhang, G. Wang, D. Shen, L. Carin, et al. “Topic-Guided variational autoencoders for text generation,” arXiv preprint arXiv:1903.07137, 2019.
[22] Weston, J., S. Chopra, A. Bordes, "Memory networks," arXiv preprint arXiv:1410.3916, pp.1-15, 2014.
[23] Sukhbaatar, S., J. Weston, R. Fergus, "End-to-end memory networks," In Advances in neural information processing systems, pp.2440-2448, 2015.
[24] Lin, Z., X. Huang, F. Ji, H. Chen, Y.

Zhang, "Task-Oriented conversation generation using heterogeneous memory networks," arXiv preprint arXiv:1909.11287, 2019.

[25] Vinyals, O., M. Fortunato, N. Jaitly, "Pointer networks," In Advances in neural information processing systems, pp.2692-2700, 2015.

[26] See, A., P. J. Liu, C. D. Manning, "Get to the point: Summarization with pointer-generator networks," arXiv preprint arXiv:1704.04368, 2017.

[27] Devlin, J., M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[28] 이주성, 오연택, 변현진, 민경구, “BERT를 이용한 한국어 문장의 스타일 변화”, 제31회 HCLT, pp.395-399, 2019.

[29] Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, P. J. Liu, et al. “Exploring the limits of transfer learning with a unified text-to-text transformer,” arXiv preprint arXiv:1910.10683, 2019.

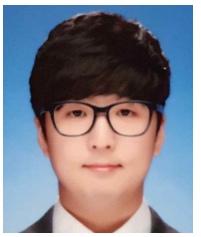
[30] Habert, B., G. Adda, M. Adda-Decker, P. B. de Marëuil, S. Ferrari, O. Ferret, P. Paroubek, et al. “Towards tokenization evaluation,” In Proceedings of Conference on Language Resources and Evaluation, Vol.98, pp.427-431, 1998.

[31] Kudo, T., J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," arXiv preprint arXiv:1808.06226, 2018.

[32] Bahdanau, D., K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, pp.1-15, 2014.

[33] 손성환, GPT-2 모델을 이용한 카테고리별 텍스트 생성, 국민대학교 석사학위 논문, 2020.

저 자 약 력



손 성 환

이메일 : sh.son@lawcompany.co.kr

- 2018년 한성대학교 컴퓨터공학부 (학사)
- 2020년 국민대학교 컴퓨터공학과 (석사)
- 2020년~현재 ㈜로앤컴퍼니 연구원
- 관심분야 : 자연어처리, 정보검색, 텍스트마이닝, 워드 임베딩, 딥러닝

강 승 식

이메일 : sskang@kookmin.ac.kr

- 1986년 서울대학교 전자계산기공학과 (학사)
- 1988년 서울대학교 전자계산기공학과 (석사)
- 1993년 서울대학교 전자계산기공학과 (박사)
- 1994년~2001년 한성대학교 부교수
- 2001년~현재 국민대학교 인공지능학부 교수
- 관심분야 : 자연어처리, 정보검색, 텍스트마이닝, 워드 임베딩, 딥러닝