BERT 기반 Sequence-to-Sequence 모델을 이용한 한국어 질문 생성

이동헌O, 황현선, 이창기

강원대학교
[skh1578, hhs4322, leeck]@kangwon.ac.kr

Korean Question Generation using BERT-based Sequence-to-Sequence Model

Gil-Dong HongO, Gil-Su Kim
Hanguk University, Hanguk Research Institute

요 약

기계 독해는 입력 받은 질문과 문단의 관계를 파악하여 알맞은 정답을 예측하는 자연어처리 태스크로 양질의 많은 데이터 셋을 필요로 한다. 기계 독해 학습 데이터 구축은 어려운 작업으로, 문서에서 등장하는 정답과 정답을 도출할 수 있는 질문을 수작업으로 만들어야 한다. 이러한 문제를 해결하기 위하여, 본 논문에서는 정답이 속한 문서로부터 질문을 자동으로 생성해주는 BERT 기반의 Sequence-to-sequence 모델을 이용한 한국어 질문 생성 모델을 제안한다. 또한 정답이 속한 문서와 질문의 언어가 같고 정답이 속한 문장의 주변 단어가 질문에 등장할 확률이 크다는 특성에 따라 BERT 기반의 Sequence-to-sequence 모델에 복사 메카니즘을 추가한다. 실험 결과, BERT + Transformer 디코더 모델의 성능이 기존 모델과 BERT + GRU 디코더 모델보다 좋았다.

주제어: 질문 자동 생성, 기계 독해, BERT, transformer

1. 서론

기계 독해는 입력받은 질문과 문단의 관계를 파악하여 알맞은 정답을 예측하는 자연어처리 태스크이다. 기계독해 데이터셋으로는 SQuAD(Stanford Question Answering Dataset)[1]와 MS-MARCO[2], NQ(Natural Question)[3]등이 있으며, 한국어 데이터인 KorQuAD(Korean Question Answering Dataset)[4]가 있다. 기계 독해 학습을 위해서는 다량의 학습 데이터가 필요하지만, 기계 독해 학습 데이터 구축은 어려운 작업으로, 문서에서 등장하는 정답과 정답을 도출할 수 있는 질문을 수작업으로 만들어야 한다. 이러한 문제를 해결하기 위해 정답이 속한 문서로부터 질문을 자동으로 생성해주는 연구가 활발히 진행되고 있다[5,6].

복사 메커니즘(Copy Mechanism)은 자연어 생성 모델에서 입력 단어가 출력 단어에 복사 될 수 있게 하는 기술로 학습데이터에 자주 등장하지 않는 고유명사나 신조어 같은 단어들의 출력 확률이 작아지는 문제를 해결하기 위해 고안된 방법으로, 입력 문장과 출력 문장의 언어가 같은 문서 요약에 주로 사용되고 있다.

본 논문에서는 기계 독해 학습 데이터 부족 문제를 해결하기 위해 정답이 속한 문서로부터 질문을 자동으로 생성해주는 BERT 기반의 Sequence-to-sequence 모델을 이용한 한국어 질문 생성 모델을 제안한다. BERT 기반의 Sequence-to-sequence 모델은 BERT를 인코더로 사용하고 Transformer decoder를 사용하며, 정답이 속한 문서와 질문의 언어가 같고 정답이 속한 문장의 주변 단어가 질문에 등장할 확률이 크다는 특성에 따라 복사 메커니즘이 추가되었다.

2. 관련 연구

[5]에서는 기존 단순 seq2seq 모델의 입력에 문장과 정답이 함께 들어가 생성한 질문에 정답이 자주 등장하는 문제를 해결하기 위하여, 정답 분리 인코더와 복사 메커니즘을 이용하여 자동으로 한국어 질문을 생성한다. 정답 분리 모델은 정답이 마스킹된 문장과 정답을 각각 인코딩하기 위하여 2개의 인코더를 사용한다. 이는 생성된 질문에 정답이 등장하는 것을 억제한다.

[6]에서는 정방향 포인터 네트워크와 역방향 포인터 네트워크의 포인터를(듀얼 포인터 네트워크) 사용하여 정답 후보를 추출하고 포인터 제너레이터를 사용해 질문을 생성한다. 포인터 제너레이터는 생성 확률과 복사 확률을 사용하여 문맥 표현을 더 잘 표현하도록 한다.

BERT(Bidirectional Encoder Representations from Transformer)[7]는 구글에서 공개한 언어 모델로 최근 다양한 자연어처리 태스크에서 뛰어난 성능을 보이고 있다. BERT는 양방향 Transformer[8]의 인코더를 사용하여 입력 문서의 정보를 양방향으로 확인할 수 있으며, Masked LM(masked language modeling)방법과 다음 문장 예측 방법을 사용하여 언어 모델을 학습한다. BERT로 다른 자연어처리 태스크를 학습할 시, fine-tuning을 적용하며, 다양한 자연어 처리 작업에서 높은 성능을 보이고

있다.

복사 메커니즘(copying mechanism)[9,10]은 자연어 생성 모델에서 입력 단어가 출력 단어에 복사 될 수 있게 하는 기술로 입력 문장과 출력 문장의 언어가 같은 문서 요약에서 주로 연구되고 있다. [11]에서는 sequence-to-sequence 모델에 copy mechanism을 적용한 문서 생성 요약 모델을 제안하였다. 제안된 모델은 기존의 모델보다 고유 명사 등의 문제에 높은 성능을 보였다.

3. BERT 기반 한국어 질문 생성 모델

본 논문에서는 한국어 질문 생성을 위해서 한국어 기계 독해 데이터셋인 KorQuAD 데이터를 사용하였다. KorQuAD 데이터를 가공하여 src(문장, 정답), tgt(질문)의 형태로 전처리를 수행하였다. 이때 문단에서 정답을 포함한 문장만을 사용하였으며, [5]과 동일하게 문장에서 정답이 위치한 자리에는 [ANS] 태그로 대체하고 [SEP] 태그로 문장과 정답을 연결하고, 질문의 끝 위치를 나타내는 [EOS] 태그를 질문 마지막 위치에 추가하였다. 형태소 분석과 BPE(Byte Pair Encoding)을 적용하였으며, 그 예는 표 1과 같다.

표 1. KorQuAD 질문 생성 데이터 예제

입력	조사/NNG_ 단/XSN_ 의/JKG_ 구성/NNG_ 원/XSN_ 문제/NNG_ 나/JC_ 은폐/NNG_ 가능/NNG_ 성/XSN_ 을/JKO_ 제기/NNG_ 하/XSV_ 었/EP_ 던/ETM_ **[ANS]** _ (/SS_ 서 프라이즈/NNG_ 대표/NNG_)/SS_ 에/JKB_ 대하/VV_ ㄴ/ETM_ 비판/NNG_ 도/JX_ 존재/NNG_ 하/XSV_ ㄴ다/EF_ ./SF_ **[SEP]** _ 신 상 철/NNP_
출력	천안함/NNG_ 사건/NNG_ 조사/NNG_ 단/XSN_ 의/JKG_ 구성/NNG_ 원/XSN_ 문제/NNG_ 나/JC_ 은폐/NNG_ 가능/NNG_ 성/XSN_ 을/JKO_ 제기/NNG_ 하/XSV_ ㄴ/ETM_ 것/NNB_ 은/JX_ 누구/NP_ 이/VCP_ ㄴ가/EF_ ?/SF_ **[EOS]** _

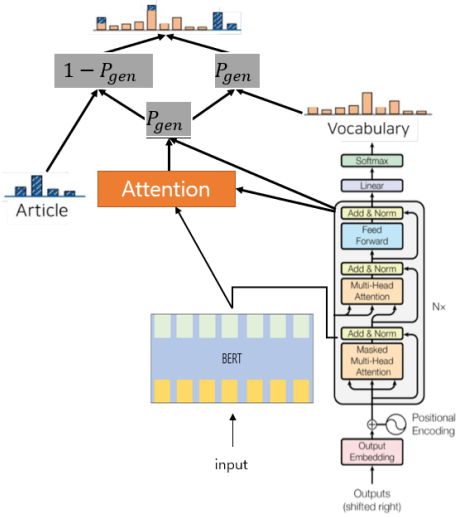

그림 1. BERT + Transformer 디코더 모델

그림 1은 BERT와 Transformer 디코더를 이용한 BERT 기반 seq2seq 모델의 구조이다. 입력 문장은 BERT 모델을 통해 인코딩 되며 BERT의 마지막 레이어를 입력 문장의 인코딩 정보로서 활용한다. 해당 정보를 가지고 Transformer 디코더가 디코딩을 수행하게 되며 마지막 출력 단어 생성부분에서 복사 메커니즘이 적용되어 최종 출력 단어를 결정하게 된다.

복사 메커니즘(copying mechanism)은 [10]의 방법을 적용하였다. Pointer-generator 모델의 수식은 다음과 같다.

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

$$p_{\text{gen}} = \sigma\left(W_g\left[h_{dec}, \text{Attention}\left(W_c^Q h_{dec}, W_c^K h_{enc}, W_c^V h_{enc}\right)\right]\right)$$

$$P_{vocab}(w) = \text{softmax}\left(W_{output} h_{dec}\right)$$

$$P(w) = P_{gen} P_{vocab}(w) + (1 - P_{gen}) \sum_{i:w_i=w} a_i^t$$

h_{dec}는 디코더의 히든 스테이트(hidden state)이며 복사 메커니즘을 적용하지 않는 모델의 경우 여기에 가중치 W_{output}을 곱하고 softmax 함수를 적용하여 최종 출력 단어의 확률 $P_{vocab}(w)$ 을 계산하게 된다. Pointer-generator 모델은 여기에 입력 단어들을 복사할 것인가를 결정할 확률 P_{gen}을 계산하여, 디코딩 시간 t에서 복사 스코어 a^t를 또 다른 attention network로 구하여 해당하는 출력 단어 확률에 P_{gen}을 곱하고 더하여 최종 출력 단어 확률인 $P(w)$를 계산한다.

4. 실험 및 결과

본 논문에서는 BERT 기반 seq2seq 모델을 학습하기 위하여 KorQuad 1.0 데이터를 변환하여 사용했으며, 학습셋(train set) 60,406개의 문서와 개발 셋(dev set) 5,773개의 문서를 사용하였다. BERT 모델은 ETRI에서 배포한 KorBERT(형태소 단위)를 사용하였으며, BERT 학습 시의 하이퍼파리미터는 BERT-base(L = 12, H = 768, A = 12)를 따른다. Transformer 디코더 모델은 다음과 같은 하이퍼파라미터로 fine-tuning하였다. 입력(질문+정답)의 최대 길이는 512로 설정하였고 출력(질문)의 최대 길이는 200으로 설정하였다. Transformer 디코더의 layer 수는 6으로 설정하였으며, 히든 사이즈는 768로 설정하였다. 각 레이어의 드랍아웃은 0.1로 설정하였고 학습률은 5e-5로 설정하였다.

본 논문에서는 비교 실험을 위해 추가적으로 BERT + GRU 디코더 모델(복사 메카니즘 포함)을 추가로 구현하여 실험하였으며, GRU 디코더의 fine-tuning 수행 시의 하이퍼파라미터는 다음과 같다. 최대 입력 문장길이는 200으로 설정하였으며, 학습률(earning rate)은 5e-5로 설정하였다. RNN 히든 레이어의 차원 수는 768로 설정하였고 각 레이어의 드랍아웃(dropout)은 0.1로 설정하였다.

표 2. 질문 생성 성능 비교(%)

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
정답 분리 + 복사 [5]	38.61	30.58	25.18	21.12
BERT + GRU decoder	**48.63**	**37.93**	30.81	25.57
BERT + Transformer decoder	46.54	35.71	28.61	23.48
BERT + Transformer decoder + Copy	46.84	37.35	**30.91**	**26.16**

표 2는 본 논문에서 제안한 한국어 질문 생성 모델의 BLEU score를 나타낸다. BERT 기반의 seq2seq 모델들(GRU decoder와 Transformer 디코더)이 모두 기존 [5] 모델보다 높은 성능을 보였으며, copy mechanism을 적용한 BERT + Transformer 디코더 모델의 BLEU-3,4의 성능이 가장 높음을 알 수 있다.

표 3. BERT + GRU 디코더 모델 질문 생성 예제 1

문장	또 법률 연맹 [ANS]은 한명숙을 국감 우수 의원으로 선정하기도 하였다.
정답	국감 모니터단
기존 질문	한명숙을 국감 우수 의원으로 뽑은 법률 연맹의 이름은?
생성 질문	한명숙을 국감 우수 의원으로 선정한 단체는?

표 4. BERT + GRU 디코더 모델 질문 생성 예제 2

문장	[ANS]가 인도양 해역에서 해적들에게 납치되었다.
정답	삼호주얼리호
기존 질문	인도양에서 해적에게 납치된 한국 선박의 명칭은 무엇인가?
생성 질문	인도양 해역에서 해적들에게 납치 된 배의 이름은?

표 5. BERT + Transformer 디코더 질문 생성 예제 1

문장	전투 마지막날, 스튜어트는 적군의 후방으로 침투하여 피켓의 돌격이 세메터리츠에서 벌어지는 동시에 적의 [ANS]을 절단하라는 임무를 받았으나, 이스트 킴볼리 플디에서의 … 실패 하였다.
정답	통신선
기존 질문	전투 마지막날 스튜어트는 무엇을 절단하라는 임무를 받았는가?
생성 질문	슈트어트는 적군의 후방으로 침투하여 무엇을 절단하라는 임무를 받았는가?

표 6. BERT + Transformer 디코더 질문 생성 예제 2

문장	오늘날의 용어로 보면, [ANS]은 분석 철학과 대륙 철학이라는 두 가지 주요한 전통을 지니고 있다.
정답	서양 철학
기존 질문	분석 철학과 대륙 철학이라는 두 가지 전통을 지니고 있는 것은?
생성 질문	분석 철학과 대륙 철학이라는 두 가지의 주요한 전통을 지니고 있는 것은?

표 3,4,5,6은 BERT + GRU 디코더 모델과 BERT + Transformer 디코더 모델의 질문 생성 예제를 나타낸다. 두 모델이 기존 질문과 유사하게 동일한 정답을 도출해 낼 수 있는 질문을 생성한 것을 알 수 있다.

표 7. BERT + GRU 디코더 모델의 잘못된 질문 생성 예제

문장	2012년 SBS 월화 드라마 <패션왕>에서 출연하였으며, [ANS]의 뉴욕, 라스베이거스, 에리조나, 사막 등에서 촬영하였다.
정답	미국
기존 질문	유아인이 출연한 패션왕 드라마의 해외 촬영지는?
생성 질문	패션왕에서 패션왕에서 출연한 나라는 어디인가?

표 8. BERT + Transformer 디코더 모델의 잘못된 질문 생성 예제

문장	[ANS]에 이영애의 아역으로 출연하면서 연기자로 입문하였고, 2002년 독립 영화 《신 도시인》에 출연하였다.
정답	《선물》
기존 질문	김태희가 연기자로 입문한 작품은 무엇인가?
생성 질문	이영애의 연기 데뷔작은?

표 7,8은 두 모델의 잘못된 생성 예제이다. 생성한 질문이 정답을 도출할 수 있을 거라 판단되지만 BERT + GRU 디코더는 표 7에서 '패션왕에서'와 같이 같은 단어들이 반복되어 나타나는 문제가 다수 발생하였다. 표 ?는 BERT + Transformer 디코더 모델이 생성한 예제로 정답 '《선물》'에 대응하는 대상이 잘못된 것을 알 수 있다.

표 9. 개체명 인식 질문 생성 예제

문장	[ANS] 정도전은 비로소 소환되어 정치일선에 나서서 새 왕조 창업을 위한 정지 작업을 단행하여 7월 17일 고양왕의 선양을 이끌어내어 이성계를 임금으로 추대하여 새 왕조 조선을 건국하였다.
정답	6월
BERT + GRU decoder	이성계가 이성계를 건국한 것은 몇월 몇월인가? 하였는가?
BERT + Transformer decoder	정도전이 이성계를 임금으로 추대한 것은 몇월인가?

표 9는 개체명 인식을 적용하여 새로 추출한 개체를 정답으로 지정하여 질문을 생성한 두 모델의 결과 예제이다. 새로운 정답을 사용하기 때문에 기존 질문은 존재하지 않는다. BERT + GRU 디코더 모델의 경우, 단어가 반복적으로 등장하며 의문사가 한번 더 등장한 것을 알 수 있다.

위의 예시와 약 10개(숫자는 동헌이 고쳐라) 결과를 정성 평가한 결과, BERT + GRU 디코더 모델은 앞서 언급한 문제의 발생 빈도가 높았다. ??개의 문서에 대하여 BERT + Transformer 디코더 모델의 경우 2개, BERT + GRU 디코더 모델의 경우 10개의 잘못된 질문을 생성하였다.

5. 결론

본 논문에서는 기계 독해 학습 데이터 부족 문제를 해결하기 위해 정답이 속한 문서로부터 질문을 자동으로 생성해주는 BERT 기반의 Sequence-to-sequence 모델을 이용한 한국어 질문 생성 모델을 제안하였다. 실험 결과, BERT + Transformer 디코더 모델의 성능이 기존 모델과 BERT + GRU 디코더 모델보다 좋았다.

향후 연구로는 본 논문에서 개발한 한국어 질문 생성 모델을 이용하여 생성한 데이터로 기계 독해 학습 데이터를 추가하여 기계 독해 성능을 향상 시킬 예정이다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 Wise QA 플랫폼 기술 개발)

참고문헌

[1] P. Rajpurkar, et al. Squad: 100,000+ question for machine comprehension of text arXiv preprint arXiv:1606.05250, 2016.

[2] T. Nguyen, et al. MS MARCO: A Human Generated Machine Reading Comprehension Dataset, arXiv preprint arXiv:1611.09268, 2016.

[3] Tom Kwiatkowski, et al. “Natural Questions: a Benchmark for Question Answering Research”, Transactions of the Associate of Computational Linguistics, 2019.

[4] 임승영, 김영지, 이주열. “KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋”, 한국정보과학회 학술발표논문집, pp.539-541, 2018.

[5] 김건영 et al. 정답 분리 인코더와 복사 메커니즘을 이용한 한국어 질문 생성. 한국정보과학회 학술발표논문집, pp.419-423, 2019.

[6] 이현구 et al. 추가 데이터 및 도메인 적응을 위한 기계독해 질의 생성, 한국정보과학회 학술발표논문집, pp.415-418, 2019.

[7] Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 2018.

[8] Vaswani et al. Attention Is ALL You Need, arXiv preprint arXiv:1706.03768, 2017.

[9] Gu Jiatao, et al, "Incorporating copying mechanism in sequence-to-sequence learning." arXiv preprint arXiv:1603.06393 (2016).

[10] Abigail See et al, Get To The Point: Summarization with Pointer-Generator Networks, arXiv preprint arXiv:1704.04368, 2017.

[11] 최경호, 이창기. Copy Mechanism과 Input Feeding을 이용한 End-to-End 한국어 문서요약, 제28회 한글 및 한국어 정보처리 학술대회, 2016

[12] 전동현 et al. 복사-메커니즘과 추론 단계의 페널티를 이용한 Copy-Transformer 기반 문서 생성 요약, 한국정보과학회 학술발표논문집, pp.301-306, 2019.