단어 손실함수를 추가한 트랜스포머 인코더-디코더 기반의 제목 생성 모델

성수진 [O1] , 이승우 [2], 차정원 [3]
창원대학교 [1,3], 한국과학기술정보연구원 [2]
{sjseong[1], jcha[3]}@changwon.ac.kr, swlee[2]@kisti.re.kr

Transformer Encoder-Decoder based Title Generation Model with Word Loss Function

Su-Jin Seong[O1], Seung-Woo Lee[2], Jeong-Won Cha[3]
Changwon National University[1,3], Korea Institute of Science and Technology Information[2]

요 약

제목은 문서의 특징을 나타내는 어구나 문장을 의미한다, 우리는 문서의 제목을 결정하기 위해 트랜스포머 기반 인코더-디코더 구조를 제안한다. 대용량 문서를 이용하여 인코더-디코더 구조의 사전학습(pre-training)을 진행하고 본문과 제목 쌍으로 이루어진 문서를 이용하여 미세조정(fine-tuning)을 하였다. 또한 단어조각(wordpiece)으로 학습된 트랜스포머 모델이 부정확한 어절을 생성하는 문제를 개선하기 위해 단어 손실함수(word loss)를 추가하여 보완하였다. 그 결과 BLEU1 0.320, BLEU4 0.124, Rouge-L 0.316의 성능을 달성하였고, 모델의 출력 결과가 개선되도록 기여할 수 있음을 확인하였다.

1. 서론

방대한 양의 텍스트 문서들의 사용이 가능해지며, 유용한 정보를 찾아내는 것이 중요해졌다. 대다수의 사람들은 수많은 문서들 사이에서 제목을 통해 자신에게 필요한 문서를 선별한다. 따라서 문서의 요점을 함축적으로 표현하고 다른 문서와 차별되는 제목을 짓는 것이 중요하다. 이처럼 제목의 중요도가 높고, 소셜 미디어, 발췌문 등 제목이 없는 문서들의 양이 증가하며 자동으로 제목을 생성하고자 하는 요구 또한 높아져 이에 대한 연구가 진행되어왔다.

기존의 제목 생성 연구는 모델이 제목을 직접 생성하지 않고 클러스터링과 같은 방법을 사용하여 문서에 나타나는 핵심 키워드를 추출하고 이를 제목으로 사용하는 방식이 제안되었다[1,2]. 추출 기반 방식은 터무니없는 단어를 생성하지는 않으나 표현이 한정적이며 응집도나 가독성을 확보하기 어렵다. 이후 sequence-to-sequence, 트랜스포머와 같은 심층학습 기반 시스템이 제안되고 문서 요약[3,4] 등의 분야에서 우수한 성능을 보이며, 제목 생성 문제 또한 심층학습 모델을 활용한 자동 생성 방법으로 발전하였다[5, 6, 7].

본 논문에서는 트랜스포머를 이용하여 논문 초록에 대한 제목을 생성하고자 한다. 또한 텍스트 생성에서 나타나는 부정확한 어절 생성 문제를 완화시키기 위한 추가 손실함수를 제안한다.

2. 제안 방법

2.1 사전학습(pre-training) 모델

본 논문에서 모델의 구조 및 학습은 T5[8]의 small 모델 구조를 차용하였다. 이 모델의 구조는 트랜스포머 인코더 디코더 모델 구조를 기반으로 한다. 인코더와 디코더는 각각 6개의 레이어와 8개의 헤드를 갖는다.

사전학습에는 뉴스 크롤링 데이터 5,500만 문장을 사용하였고, 최대 문장 길이를 단어조각 (wordpiece) 기준 512로 제한하였다. 토크나이저 사전의 크기는 30,100개로 설정하였다.

마스크 방법은 T5 모델의 방식을 사용한다. 입력 문장이 주어지면 해당 문장 길이의 15%에 해당하는 개수만큼 마스크 개수를 정하고, 지정된 마스크의 길이만큼의 연속된 단어조각을 하나의 마스크로 치환한다. 이 때 사전학습 모델은 개별적인 마스크로 치환된 원래 단어조각을 예측하여 마스크와 쌍이 맞도록 생성하는 것을 목표로 한다. 예를 들어 입력 문장은 '_아침 <A> 을 _먹 었고 _점심 으로는 <B> _것 이다 .'이고 출력은 '<A> 으로 밥 <B> _빵 을 _먹을'이 될 수 있다. 마스크를 여러개 사용하고 각 마스크에 개별적인 토큰을 부여함으로써 모델이 각 토큰이 등장하는 상황을 구분하고 그에 따라 적절한 결과를 생성해내는 방법을 학습할 수 있을 것이다.

이렇게 학습한 한국어 사전학습 모델을 KoT5로 명명한다.

2.1 논문 제목 생성 모델

사전학습된 모델을 기반으로 논문 제목 생성 분야에

대한 학습을 진행한다. 텍스트 생성 방법으로 학습된 모델이기 때문에 별도의 코드 수정 및 레이어 추가 없이 학습한다. Learning rate는 1e-05를 사용한다.

사전학습은 다양한 태스크에 모델이 적용될 수 있도록 일반화를 잘 하는 것이 목적이기 때문에 학습과정에 제약을 가하기 어렵다. 하지만 미세조정(fine-tuning) 단계에서는 태스크가 정해짐에 따라 생성 결과의 범위를 한정시킬 수 있으므로 미세조정 과정에 어절 생성 결과를 개선하기 위한 단어 손실함수(word loss function)를 추가한다.

단어조각 단위로 생성을 수행했을 때 가장 큰 비율을 차지하는 오류는 반복되는 단어조각과 부정확한 어절을 생성하는 것이다. 이 중 부정확한 어절 생성 문제를 완화하기 위해 단어조각 단위 생성된 출력을 어절로 복구하여 그 결과를 기반으로 단어 손실함수를 설계하고 기존 교차 엔트로피 손실함수(cross-entropy loss function)와 가중치 합하여 모델에 적용한다.

출력으로 생성된 어절이 올바르게 생성되었는지를 측정하는 방법으로 예측 어절이 입력과 정답에 나타나는 어절 목록에 포함되는 비율을 측정한다. 이 때 입력과 정답을 구성하는 어절은 오류가 없는 정확한 어절의 형태라고 가정한다. 이를 통해 모델은 올바른 어절의 형태와 유사한 어절을 생성하도록 학습할 것이며, 입력과 정답을 대상으로 비교하기 때문에 입력의 중요한 단어의 가중치를 높여 이를 출력에 반영할 것이라 기대된다.

$$L_{word} = \frac{\#\ words\ in\ Input\ or\ Gold}{\#\ words\ of\ Pred} \quad (1)$$

$$Loss = \alpha L_{CE} + (1-\alpha) L_{word} \quad (2)$$

식(1)은 단어 손실함수를 나타내고, 식(2)는 전체 손실함수를 나타낸다. 식(2)의 L_{CE} 는 교차 엔트로피 손실값(Cross-Entropy loss)를 의미하며 α를 0.5로 두고 학습하였다.

3. 실험 및 성능

3.1 학습 코퍼스

본 연구는 DBpia[9]에서 직접 수집한 25,564개의 한국어로 작성된 논문의 초록과 제목을 대상으로 수행된다. 이 중에서 학습 데이터로 17,895개, 평가 데이터로 5,113개, 검증 데이터로 2,556개를 사용한다.

3.2 평가 방법

BLEU (Bilingual Evaluation understudy) score[10]와 Rouge (Recall-Oriented Understudy for Gisting Evaluation)[11]을 사용하여 모델의 성능을 측정한다.

BLEU score는 n-gram에 기반하여 생성 문장의 정밀도(precision)을 측정하는 방법이며, Rouge-N는 n-gram에 기반하여 정답 문장에 대한 생성 문장의 단어 재현율(recall)을 측정하며, Rouge-L은 LCS(Longest Common Subsequence)를 사용하여 측정한 방법이다.

3.3 실험 결과 및 분석

표 1은 논문 초록에 대한 제목 생성 결과의 성능을 측정하여 나타낸 것이다. KoT5는 사전학습된 모델을 사용하여 학습한 것이고, KoT5+wordLoss는 사전학습된 KoT5를 학습할 때 단어 손실함수를 추가한 결과의 성능을 나타낸다. Base KoT5와 비교했을 때 사전학습된 모델에 미세조정을 거친 모델의 결과가 훨씬 높은 성능을 달성하며, 단어 손실함수를 추가한 실험의 성능이 BLEU와 Rouge 모두에서 향상됨을 보인다.

표 1 제목 생성 결과 성능표

Model	BLEU1	BLEU4	Rouge-L
Base KoT5	0.057	0.017	0.067
KoT5	0.297	0.114	0.311
KoT5+wordLoss	0.320	0.124	0.316

단어 손실함수를 추가한 후 생성된 예측 문장의 어절이 입력에 포함되는 비율은 전체 예측 어절의 39.7% (13,255개)에서 40.4% (13,553개)로 하였다. 전체 코퍼스에 나타나는 어절 목록에 오류가 없다고 가정하고 예측 어절의 포함 비율을 측정하였을 때 비율은 75.7% (33,918개)에서 76.8% (34,448개)로 향상되었다. 더하여 단어 손실함수가 의도한 방향으로 모델의 학습에 영향을 주었는지 확인하기 위해 과다 생성, 과소 생성 오류 유형의 개수를 측정한다. 과다 생성은 정답을 포함하지만 잉여 단어조각을 생성하여 오답이 된 경우이고, 과소 생성은 정답에 포함되지만 구 혹은 절이 완전하지 못한 경우를 나타낸다. 측정 결과는 표 2와 같다.

표 2 과다/과소 생성 오류 발생 개수

Error Type		KoT5	KoT5 +wordLoss
기존 출력	과다 생성	18	26
	과소 생성	32	44
연속 반복 제거 후	과다 생성	5	3
	과소 생성	25	28

단어 손실함수를 추가하였을 때 정답 제목과 똑같이 생성한 출력의 수와 정답은 아니지만 정답에 포함된 혹은 정답을 포함하는 구를 생성한 출력의 수가 증가하였다. 또한 제목이 본문에 포함되어있을 때 정답과 완전히 동일하거나 정답 문장에 포함되는 출력이 33개로 KoT5보다 10개 많았다. 이러한 결과는

추가된 단어 손실함수가 의도한 목적대로 학습에 영향을 주어 모델이 올바른 어절을 생성하는데 기여하였음을 보여준다.

표 3과 4는 각각 단어 손실함수를 추가한 모델 결과의 좋은 예와 나쁜 예를 보여준다.

표 3의 KoT5의 출력에서 '참가가', '정신건강이의'의 경우 본문에 나타나지않은 단어로, 단어 손실함수를 추가한 모델에서는 이러한 단어가 본문에 나타나는 단어로 생성되었음을 확인할 수 있었다.

하지만 표 4의 예시에서는 정답이 '경험 기반의 물리적인 …'임에도 '게임'이 본문에 높은 빈도로 등장하기 때문에 모델에서 잘못 생성한 것으로 예상되며 이는 단어 손실함수의 추가로 발생할 수 있는 문제점으로 보인다.

표 3 KoT5+ wordLoss 결과 좋은 예

본문	
본 연구는 **재즈댄스 참가와 신체상 및 정신건강과의 관계**를 규명하여 **재즈댄스**의 … (생략) … 하체요인은 **재즈댄스 참가** 빈도와 **정신건강과의** 관계에서 매개변수로 중요한 작용을 하는것으로 나타났다.	
Gold	재즈댄스 참가와 신체상 및 정신건강과의 관계
KoT5	**댄스 참가가** 신체상 및 **정신건강이의** 관계
KoT5+wordLoss	**재즈댄스 참가와** 신체상 및 **정신건강과의** 관계

표 4 KoT5+ wordLoss 결과 나쁜 예

본문	
… (생략) … 연구에서는 **게임** 환경에서 **경험** 기반의 물리적인 인터페이스가 **게임** 몰입에 미치는 긍정적인 영향을 알아보고자 한다.	
Gold	**경험** 기반의 물리적인 인터페이스가 게임 몰입에 미치는 영향
KoT5	**경험** 기반의 물리적인 인터페이스가 게임 몰입에 미치는 영향
KoT5+wordLoss	**게임** 기반의 물리적인 인터페이스가 게임 몰입에 미치는 영향

4. 결론

본 논문에서는 트랜스포머 기반 모델 중 T5 모델 구조를 바탕으로 사전학습을 진행하였고, 논문 초록과 제목 데이터로 사전학습된 모델을 미세조정하여 논문 제목 생성 모델을 구축하였다. 미세조정을 수행할 때 단어조각 단위로 출력을 생성하는 모델이 초록 내용과 연관있고 정확한 어절을 생성하도록 추가 손실함수를 설계하여 기존 교차 엔트로피 손실함수와 함께 사용하였고, 그 결과 모델의 출력에서 입력 혹은 정답에 포함되는 어구의 수가 늘어 제목 생성 결과를 개선할 수 있는 가능성을 보였다.

제안한 단어 손실함수는 모델의 출력 어절과 입력 문서 혹은 정답의 어절을 비교하여 출력 어절이 입력 문서나 정답에 나타나는 어절과 유사한 어절로 생성되도록 하는 것을 목적으로 한다. 때문에 정답에 포함이 되지 않음에도 출현 빈도가 높은 어절이 결과로 생성될 수 있다. 따라서 발생할 수 있는 문제에 대한 패널티를 추가하거나 연산 방식을 수정하여 태스크에 적절하게 손실함수를 개선하는 것이 필요하다.

참 고 문 헌

[1] 한규열, 안영민, "LDA로 형성된 한국어 문서 클러스 터의 자동 제목 생성", 한국정보과학회 학술발표논문집, pp. 616-618, 2013.

[2] 김태현, 맹성현, "계층구조를 이용한 문서 클러스터 제목의 자동생성", 한국정보과학회 언어공학연구회 학술 발표 논문집, pp. 163-170, 2001.

[3] 이태석, 강승식, "Bert 임베딩과 선택적 OOV 복사 방법을 사용한 문서 요약", 정보과학회논문지, 47(1), pp. 36-44, 2020

[4] 조명현, 김희성, 구명완, "주의집중을 가진 Sequence-to-Sequence 순환신경망을 이용한 문서요약 구현", 한국정보과학회 학술발표논문집, pp. 587-589, 2019

[5] 이현구, 김학수, "주의집중 및 복사 작용을 가진 Sequence-to-Sequence 순환신경망을 이용한 제목 생성 모델", 정보과학회논문지, 44(7), pp. 674-659, 2017

[6] 유홍연, 이승우, 고영중, "단어 관련성 추정과 바이트 페어 인코딩(Byte Pair Encoding)을 이용한 요약 기반 다중 뉴스 기사 제목 추출", 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp. 115-119, 2018

[7] 조성민, 김우생, "RNN과 강화 학습을 이용한 자동 문서 제목 생성", Journal of Information Technology Applications & Management, 27(1), pp. 49-58, 2020

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", arXiv preprint arXiv:1910.10683, 2019

[9] Dbpia, https://www.dbpia.co.kr/, accessed May 8, 2020

[10] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, "BLUE: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the ACL, pp. 311-318, 2002

[11] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Proceedings of the ACL-04 workshop, 8, 2004