

논문 서론 작성 AI

CUAI 5기 T12 NLP팀

김서린(응용통계학과), 서희재(컴퓨터공학과), 이재용(통계학과)

[요약] 본 연구는 논문의 서론을 대신 작성하는 모형을 구현하였다. 주어진 문장에 대해 한 문단을 생성하고, 생성된 문장들을 paraphrase 한다. 연구 결과, 이러한 구조를 가진 모형은 충분히 주어진 문장을 기반으로 관련이 있는 문장을 생성할 수 있었고 paraphrase를 통하여 생성된 문장들에 대해 표절 문제를 해결할 수 있음을 보였다.

1. 서론

논문은 크게 서론, 본론, 결론으로 구성된다. 본론에서는 선행연구에 대한 설명과 진행하고자 하는 연구와 그의 결과를 작성하고 결론은 본론의 내용을 요약하고 연구 결과에 대한 시사점 및 제언한다. 따라서 본론과 결론은 연구의 성과가 중요하다고 볼 수 있다. 하지만 서론은 진행하고자 하는 연구의 목적과 이유를 독자에게 설득해야 하며 독자의 흥미를 끌어야 한다. 이처럼 서론에서는 사실상 연구보다는 글쓰기의 실력이 더 중요하다고 말하여도 과언이 아니다.

박순우(2004)에 따르면 서론에서는 문제 제기, 연구 배경 및 필요성, 연구목적을 제시하는 단계라고 한다. 이때 문제 제기는 독자들의 흥미를 끌어야 하며 본 연구가 왜 중요한지를 호소할 수 있어야 한다고 한다. 하지만 이러한 역량은 글쓰기에 해당한다. 만약 연구자가 글쓰기 역량이 부족한 경우 연구 성과와 별개로 작성한 논문의 질이 좋지 않을 수 있는 문제가 있다.

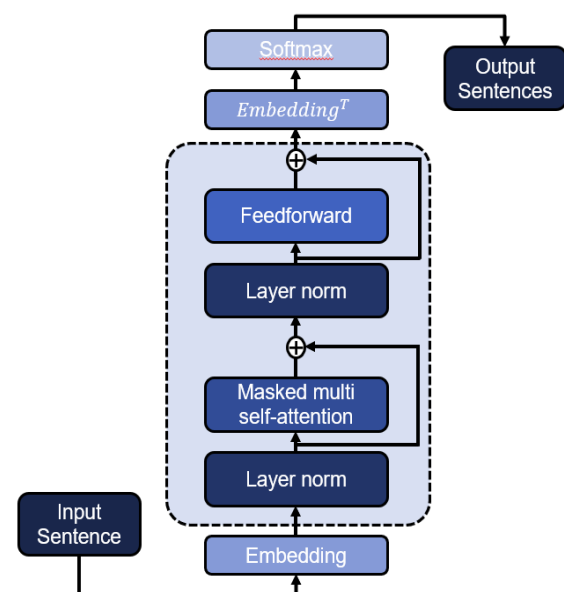
이처럼 논문에서 서론을 작성하기 위해서는 큰 노력이 필요함을 알 수 있다. 여기서 본 연구는 이러한 수고를 덜 방법을 구상하고자 하였다. 서론을 대신 작성해 주는 인공지능이 개발된다면 연구자는 서론을 금방 완성할 수 있고, 여기에서 절약한 시간을 연구에 더 투자할 수 있을 것이다. 이러한 점에서 서론을 대신 작성해 주는 인공지능의 이점을 보고 본 연구에서는 기존의 문장 생성 모형을 응용하여 논문의 서론을 작성해 주는 모형을 구현하고자 한다.

2. 본론

I. 이론적 배경

1) GPT2

Text generation에서 가장 강력한 모형은 GPT(Generative pretrained transformer)이다. GPT2의 구조는 <그림 1>과 같다. 기본적으로 GPT1과 구조가 동일하지만, layer normalization이 각 sub block의 input으로 옮겨졌다는 점과 마지막 self-attention block 이후에 추가되었다는 점에서 차이점이 있다 (Radford et al., 2019). GPT2는 byte pair encoding을 활용하였기 때문에 byte 수준에서 동작한다. 이는 전 처리와 토큰화 과정에서 정보가 손실되는 것을 방지하고, 다른 언어 모형 benchmark에서도 evaluation을 할 수 있다는 점에서 강점을 갖는다.



<그림 1> GPT2의 전반적인 구조

2) Paraphrasing

Paraphrasing은 텍스트나 구절의 의미를 다른 단어를 사용하여 다르게 표현하는 것을 말한다. Wieting과 Gimpel(2017)에 따르면 주어진 문장과 의미가 비슷한 문장을 만들어내는 paraphrase generation은 서로 의미가 비슷한 두 문장으로 이루어진 데이터셋(paraphrase pairs)을 학습하여 구현할 수 있다고 한다. 즉, 문장 생성에서 자주 사용되는 GPT와 같은 모형으로 페어(paired) 데이터를 학습시키면 된다. 또한 페어 데이터를 encoder-decoder 모형으로 학습한 경우 생성된 paraphrase 문장을 “canonicalize”할 수 있다고 한다. 즉, 문장 구조가 복잡하거나 뒤죽박죽인 문장을 표준화된 문장 구조를 가지게끔 바꿔준다는 것이다. 마지막으로, 더욱 정교한 paraphrase 문장을 만들기 위해선 적절한 문장 구조를 가진 템플릿(template)을 활용할 수 있다. 이를 통해 paraphrase 문장이 원하는 문장 구조를 가지게끔 하여 더 취지에 맞는 문장을 만들 수 있다는 것이다(Chen, Tang, Wiseman & Gimpel, 2019).

3) Evaluation

Evaluation 방법으로는 Perplexity(이하 PPL)를 사용하였다. PPL은 text generation 모형의 성능 평가 지표 중 하나로, 수치가 낮을수록 우수한 성능을 의미한다. PPL은 문장의 길이로 정규화된 문장 확률의 역수로, 수식으로 나타내면 다음과 같다.

$$PPL(W) = P(w_1, w_2, w_3, \dots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \dots, w_N)}}$$

〈수식 1〉 Perplexity

PPL은 언어 모형의 분기 계수인데, 간단히 설명하면 다음 단어를 예측할 때 몇 개의 후보를 고려하는지를 의미한다.

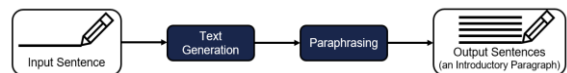
II. 연구 방법

1) 논문 작성 AI 구조

본 연구에서 구상한 논문 작성 AI의 구조는 다음과 같다. 정언학(2014)에 따르면 두괄식 문장이 가장 설득력이 있는 글쓰기 방법이다. 따라서 한 문단의 첫 번째 문장이 가장 중요하기 때문에 사용자에게 첫 번째

문장을 입력받는다. 입력받은 문장에 기반하여 text generation을 통해 여러 문장을 생성하면 하나의 문단이 완성된다. 이때 표절 문제를 해결하고 생성된 문장의 불안정한 문법 구조를 보완하고자 생성된 문장에 대해 paraphrase를 한다.

이를 통해 사용자는 적은 개수의 문장을 작성하기만 하여도 여러 문단을 얻게 되어 서론을 금방 작성할 수 있을 것이다. 〈그림 2〉는 이러한 논문 작성 AI의 순서도를 나타낸다.



〈그림 2〉 논문 작성 AI 순서도

2) Text generation 구현

본 연구에서는 서론의 첫 문장을 제시하면 제시된 문장을 기반으로 문단을 완성하는 것을 목표로 한다. 주어진 단어를 통해 다음 단어를 예측하는 GPT는 연구의 목표와 부합하는 모형으로 text generation에 있어 가장 유명한 모형 중 하나이다. 본 연구에서는 한국어로 사전 학습된 KoGPT2 모형에 한국어로 된 논문 서론 데이터셋을 이용하여 finetuning 하였다. 본 모형의 학습은 PPL을 통해 평가되었다.

한국어 논문 서론 데이터셋은 한국과학기술정보연구원(KISTI)¹에서 제공하는 국내 논문 전문 텍스트 데이터셋을 활용하였다. 해당 데이터셋에서 영어로 쓰인 논문을 제거한 후 서론만을 추출하여 총 283,893개로 이루어진 논문 서론 데이터셋을 구성하였다. 또한, 논문의 인용 표시는 학습에 방해가 된다고 판단하여 인용 표시를 제거하였다.

3) Paraphrasing 구현

본 연구에서는 KoGPT2로 작성된 문장이 특정 논문의 문장과 똑같이 표절의 문제가 발생할 상황을 대비하여, 생성된 문장을 paraphrasing 하는 과정을 추가하였다. 또한, Wieting과 Gimpel(2017)가 제시하였듯이 encoder-decoder 모형으로 학습한 경우 생성된 paraphrase 문장은 표준화된 문장 구조를 가지기 때문에 text generation 단계에서 생성된 문장의 구조가 좋지 않더라도 이 단계에서 한번 다듬을 기회를 마련하고자 하였다. Paraphrasing 모형은 한국어로 사전 학습된 GPT3-small 모형을 이용하여 한국어로 된 페어

¹ <https://aida.kisti.re.kr>

데이터셋을 학습시켰다. Paraphrase 모형의 학습도 PPL을 통해 평가되었다.

본 연구에서는 한국어 paraphrase 페어 데이터셋을 직접 만들었는데, AI-Hub²에서 제공하는 한국학술정보 등에서 가져온 사회 과학 분야(법률, 교육, 경제, 문화/관광/예술)와 기술 과학 분야(ICT, 전기/전자/기계, 의학)의 한-영 번역 말뭉치 데이터셋을 활용하였다. 사회 과학 분야와 기술 과학 분야에서 데이터를 각각 6,500개씩 뽑아 총 13,000개의 한-영 페어 데이터셋을 구성한 후, 이것의 영어 문장을 Papago API로 번역하여 한-한 페어 데이터셋으로 변형하였다. 한-한 페어 데이터셋의 일부는 <표 1>과 같다.

<표 1> 한-한 페어 데이터셋의 일부

A	B
오른 클리니에 포함된 모든 피부 부식/자극 물질도 심각한 눈 손상/눈 자극의 영향을 받습니다.	오른 세척제에 포함된 피부 부식성/자극성 물질은 모두 심한 눈손상성/눈 자극성에도 해당된다.
공기 중의 오일 입자가 기준치 이하인지 아닌지는 센싱으로 판단할 수 있다.	센싱에 의해 공기 중 오일입자가 기준치이하인지 여부를 판단할 수 있다.
각 처리 후 세포 배지를 채취하여 저장하였다.	각각의 치료 후에, 세포 배양 배지는 수집되고 저장되었다.

III. 연구 결과

<표 2> 서론 논문 작성 AI의 구현 결과

Generated text	After paraphrasing
최근 들어 스마트폰의 보급이 확산됨에 따라 모바일 인터넷에 대한 관심이 높아지고 있다. 특히 스마트폰은 기존의 휴대폰과 달리 언제 어디서나 인터넷을 사용할 수 있는 편리함과 이동성을 바탕으로 다양한 서비스를 제공받을 수 있다는 장점을 가지고 있어 많은 사람들이 이용하고 있다. 하지만 이러한 장점에도 불구하고 악의적인 공격자에 의해 개인정보가 유출될 경우 심각한 사회문제로	최근 들어 스마트폰의 보급이 확산됨에 따라 모바일 인터넷에 대한 관심이 높아지고 있다. 특히 스마트폰의 경우 언제 어디서든 인터넷을 이용할 수 있다는 편리성과 이동성, 그리고 이동성이라는 특징을 가지고 있기 때문에 많은 사람이 이용하고 있는 추세이다. 그럼에도 불구하고 이러한 장점을 가지고 있음에도 불구하고 악용자에 의한 개인정보 유출로 인한 사회적인

대두되고 있으며, 이에 따른 보안 위협도 증가하고 있다. 따라서 본 논문에서는 스마트폰에서 사용되는 인증 프로토콜을 분석하고 이를 기반으로 한 안전한 인증 기법을 제안한다.

문제제기가 증가하고 있으며 이에 대한 대책 마련이 요구되고 있다. 따라서 이 논문은 스마트폰에서의 인증 프로세스를 살펴보고 이를 바탕으로 안전한 인증을 위한 기법들을 제시할 것이다.

<표 2>는 주어진 문장을 기반으로 생성된 문장들과 생성된 문장들의 paraphrase가 된 후의 결과를 보인다. 생성된 문장들을 보면 첫 번째 문장과 매우 관련이 있고 글의 흐름도 상당히 자연스러운 것을 볼 수 있다. Paraphrase가 되기 전과 후를 비교하면 전의 문장들이 이미 문법이 괜찮고 표현이 오히려 더 자연스러운 것을 볼 수 있다. 즉, paraphrase를 통하여 문법적인 오류가 덜한 문장을 생성하지는 못하였다. 하지만, 서론 데이터를 전이학습시킨 KoGPT2가 생성한 문장들에 대하여 표절 문제를 주의할 수밖에 없는데 이를 paraphrase를 통해 해결할 수 있음을 볼 수 있다.

3. 결 론

본 연구는 논문의 서론을 작성해 주는 모형을 text generation과 paraphrase를 활용하여 구현하였다. 연구 결과 한 문장을 입력받으면 문장 생성 모형이 충분히 그에 기반하여 관련이 있는 문장을 생성할 수 있다는 것을 보였다. 또한 paraphrase를 통해 표절 문제를 해결할 수 있음을 보였다.

서론은 연구자가 작성하기 가장 힘들 수 있는 부분 이기에 이 부분을 AI가 대신한다면 절약된 시간을 연구에 더 투자할 수 있다는 이점이 있다. 또한 본 연구의 논문 작성 AI는 실제로 사용자에게 생성된 문장들의 paraphrase가 되기 전과 후를 모두 보여주기 때문에 사용자는 사실상 2가지의 사용할 수 있는 문단을 얻게 된다.

본 연구의 한계점은 다음과 같다. Text generation에 있어 제시된 문장이 충분한 정보를 담고 있지 않은 경우에는 실제 목적과 무관한 문장을 만들어 내는 경우가 비일비재하다. 또한, 생성된 문장들을 paraphrase 하면 오히려 문법적으로 오류가 더 생기는 경우가 존재한다. 하지만 이는 모형을 충분히 학습시키지 못하

² <https://aihub.or.kr>

였기 때문에 판단이 되며 시간과 자원이 충분하다면 더 성능이 좋은 논문 서론 작성 AI를 구현할 수 있을 것으로 기대한다. 특히나 paraphrase 같은 경우 Chen et al.(2019)이 제시하였듯이 템플릿 기반으로 하면 문법적인 오류가 덜한 결과를 보일 수 있을 것이다.

참고 문헌

- 1) 박순우, "연구논문 작성법," *대한예방의학회: 학술대회논문집*: 57-68, 2004.
- 2) Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I., "Language models are unsupervised multitask learners," *OpenAI blog*, 1(8), 9, 2019.
- 3) Wieting, J., & Gimpel, K., "ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations," arXiv preprint arXiv:1711.05732, 2017.
- 4) Chen, M., Tang, Q., Wiseman, S., & Gimpel, K., "Controllable paraphrase generation with a syntactic exemplar," arXiv preprint arXiv:1906.00565, 2019.
- 5) 정언학, "대학생 글쓰기의 형식적 오류 분석: 구, 문장, 단락, 글 차원의 문제," *교양교육연구* 8.1: 113-169, 2014.