형태소 분석 및 Perplexity를 활용한 한국어 텍스트 요약 모델*

오유솔○ 박상근
경희대학교 소프트웨어융합학과
wina0817@khu.ac.kr , sk.park@khu.ac.kr

Korean Sentence Summary Model: Language Scoring with Morphological Analysis and Perplexity

Yusol Oh○ Sangkeun Park
Department of Software Convergence, Kyung Hee University

요 약

최근 정보량의 증가로 긴 문서를 짧게 요약해주는 텍스트 요약 서비스가 증가하고 있다. 그러나 단순히 텍스트를 짧게 요약하면 문서의 흐름이 모두 담기지 않아 내용이 왜곡되어 전달될 수 있다. 따라서 본 논문에서는 문서의 전체 흐름이 정확하게 전달되도록 한국어의 문법적 특성을 고려한 요약 모델을 제안한다. 그 후 모델의 원문, 정답 요약문 쌍으로 구성된 텍스트 데이터를 통해 모델이 생성한 요약문의 정확도를 평가하고, 네이버 기사 요약 서비스와의 비교를 통해 제안 모델의 성능을 입증한다.

1. 서 론

인터넷에서는 수많은 정보가 쏟아지고 있으며, 사람들은 다양한 정보를 더 빠르고 쉽게 습득하고자 한다. 이에 따라, 긴 영상보다는 1분 내의 짧은 영상을, 긴 글보다는 짧은 글을 선호하게 되었다. 이러한 경향에 맞게 긴 문서를 간결하게 요약해 주는 텍스트 요약 서비스도 증가하였다. 신문 기사, 보험사 약관, 은행 문서, 쇼핑몰 리뷰 등과 같은 다양한 곳에 텍스트 요약 서비스가 적용되어 많이 사용되고 있지만, 단순히 텍스트를 짧게 요약하면 기존의 정보가 과장되거나 왜곡되어 소비자에게 전달될 수 있다는 위험이 있다. 따라서 많은 정보가 담긴 문서를 간결하게 요약하되, 문서의 전체 흐름을 정확하게 전달할 필요가 있다.

중요한 정보는 그대로 보존하면서 텍스트를 요약하는 다양한 연구가 수행되었다. 텍스트 요약 알고리즘 방식은 크게 '생성 요약'과 '추출 요약' 두 가지로 나눌 수 있다. 생성 요약 방식은 핵심 문맥을 반영한 새로운 문장을 생성해서 문서를 요약하는 방식(예: ChatGPT[1]) 이고, 추출 요약 방식(예: CLOVA Summary API)은 원문에서 중요한 핵심 문장 또는 단어 구를 추출하여 이들로 구성된 요약문을 만드는 방식이다.

본 논문에서는 기존의 생성 요약 방식이 거짓 정보가 담긴 요약 문장을 생성할 수 있다는 문제점과 추출 요약 방식이 문서의 흐름을 모두 담지 못해 내용에 왜곡이 생길 수 있다는 문제점을 해결하고자 했다. 형태소 분석을 통해 한국어의 문법적 특성을 고려하여 문장에서 각 단어의 중요도를 판단하고, Perplexity[2] 계산을 통해 문법적으로 오류가 적도록 단어를 추출하여 요약한다. 그 후, 네이버 CLOVA Summary API와의 직접적인 문서 요약 결과 비교를 통해 기존 요약 방식의 문제점을 해결함을 보이고, 데이터 세트를 활용한 작성된 요약문의 정확도 평가를 통해 성능을 테스트하였다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2023년도 SW중심대학사업의 결과로 수행되었음(2023-0-00042).

2. 관련 연구

2.1 생성 요약

생성 요약 모델은 원문으로부터 내용이 요약된 새로운 문장을 생성하는 모델이다. 따라서 모델을 훈련하기 위해서는 '원문' 뿐만 아니라 '실제 요약문'이라는 Label data가 필요하기 때문에 모델을 위한 데이터를 구성하는 것이 힘들다. 또한, 데이터를 학습하여 확률상 가장 높은 문장을 생성하기 때문에, 각 문장의 진위를 확인하지 못하여 거짓 정보나 허위 정보를 생성하는 Hallucination[3]이 발생할 수 있다는 문제점이 존재한다.

2.2 추출 요약

기존 추출 요약 모델의 경우 원문에서 문장을 그대로 가져오기 때문에 생성 요약의 문제점인 Hallucination이 발생하지 않아 진위에 대한 문제는 존재하지 않지만, 추출된 일부 문장들이 문서 전체 내용을 전부 대변할 수 없기 때문에 내용이 과장되거나 왜곡되어 전달될 수 있다는 문제점이 존재한다.

[4]에서는 영어 원문에서 단어를 하나씩 제거해가며, '문장의 혼란스러운 정도'를 의미하는 Perplexity를 계산 후, 가장 혼잡도가 낮은 단어 조합 경로를 선택하는 방식으로 문장 요약을 수행했다. [5]에서는 [4]와 마찬가지로 Perplexity를 고려하되, 단어들의 품사 정보, 의존 관계 정보 등을 고려하여 원문에서 중요하지 않은 단어를 삭제해 가며 문장 요약을 수행한다. 그러나 [4]와 [5]에서 제시한 요약 모델은 영어 문장을 요약하는 것으로, 영어와 달리 어순이 중요하지 않고, 접사, 조사에 따라 문장 성분 및 품사가 달라지는 한국어에 적용은 적절하지 않다. 본 논문에서는 이러한 문제점을 해결하기 위해 한국어의 언어적 특성을 고려하여 단어 추출 방식을 활용한 문장 요약 모델을 제안한다.

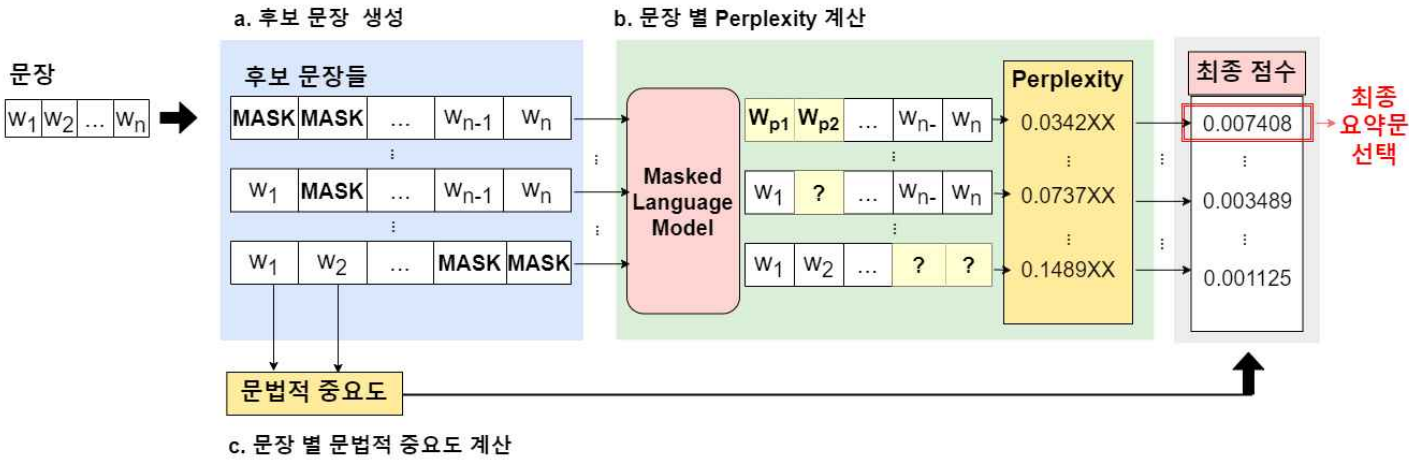

그림 1. 요약 모델 구조도

원문 : 만약 지방간의 원인이 되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 복용을 중단하거나 다른 약물로 대체해야 한다.

순위	후보 문장	Perplexity	문법적 중요도	최종 점수
1	만약 지방간의 원인이 되는 약물을 복용하고 있다면 복용을 중단하거나 다른 약물로 대체해야 한다.	0.000030463	0.0014	45.957240
2	되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 복용을 중단하거나 다른 약물로 대체해야 한다.	0.000030556	0.0014	45.817026
3	만약 지방간의 원인이 되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 약물로 대체해야 한다.	0.000030966	0.0014	45.209502
4	만약 지방간의 원인이 되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 복용을 한다.	0.000033615	0.0015	44.622761
5	만약 지방간의 원인이 되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 복용을 중단하거나 다른 약물로	0.000034387	0.0015	43.620515

그림 2. 후보 문장 TOP 5

3. 텍스트 요약 모델

본 논문에서 제안하는 텍스트 요약 모델은 [그림 1]과 같이 구성된다. 입력한 문장에 대해 요약 문장 후보들을 생성하고, 문법적 중요도와 Perplexity를 고려한 최종 점수를 통해 문장 후보 중 최종 요약문을 선택하는 방식으로 문서 요약이 진행된다.

3.1. 후보 문장 생성

먼저 요약할 문서를 구성하는 모든 문장을 하나씩 구분한다. 각 문장마다 N-gram[6] 기법을 활용해서 1~N개의 크기만큼 연속된 어절들을 반복적으로 삭제하며 다수의 후보 문장을 생성한다. 문장마다 삭제된 n개의 어절 자리에는 [MASK]라는 토큰을 채워 넣는다. 이 때, 요약 문장 후보들이 문장의 주성분인 주어, 목적어, 보어, 서술어는 반드시 포함되어야 한다고 판단하여, 후보 문장들의 길이가 단어 4개 미만인 경우에는 후보 문장을 생성하지 않는다.

3.2. 문장 별 Perplexity 계산

생성된 후보 문장이 문맥적으로 자연스러운지 판단하기 위해, KoBERT 데이터 세트로 학습시킨 마스크 언어모델(Masked Language Model)을 사용해서 후보 문장마다 Perplexity를 계산한다.

3.3. 문장 별 문법적 중요도 계산

형태소 분석을 활용해 생성된 여러 후보 문장마다 문법적 중요도를 계산한다. 후보 문장을 구성하는 각 어절을 확인하면서, [MASK] 토큰이 아닌 본래의 어절이 그대로 남아 있다면 KoNLPy에서 제공하는 SHINEWARE의 KOMORAN[7]을 사용해서 해당 어절의 형태소 성분을 확인한다. 문법적으로 중요한 의미를 갖는 것으로 판단되는 형태소 성분(고유 명사, 일반 명사, 동사, 외국어, 주격 조사, 목적격 조사)에 해당하는 어절을 찾을 때마다 해당 어절에 0.0001씩 중요도 점수를 부여한다. 해당 문장에 부여된 모든 중요도 점수를 합산한 값이 해당 문장의 문법적 중요도가 된다.

3.4. 요약 문서 생성

문장 별로 생성된 여러 후보 문장마다 계산된 Perplexity와 문법적 중요도 값을 활용해서 최종 요약 문장을 선정한다. Perplexity가 낮을수록 문장의 혼잡도가 낮다는 의미이고, 문법적 중요도 값이 클수록 문법적으로 완성도가 높다는 의미이므로, Perplexity의 역수와 문법적 중요도를 곱한 값이 가장 높은 문장을 원 문장의 최종 요약 문장으로 선정했다 [그림 2]. 이렇게 문서를 구성하는 모든 문장을 각기 더 짧은 문장으로 요약하고, 이 문장들을 다시 병합하여 최종적으로 요약된 문서를 완성한다.

원문 [10]
“킹크랩 가격, 갑자기 4만원 뚝... 7만원대로 급락한 이유”
고급 식자재로 불리는 킹크랩 가격이 4년 만에 1㎏당 7만원대로 떨어졌다. · · · 지난달 18일까지 ㎏ 당 11만5000원이던 레드 킹크랩 가격은 하루 만에 7만7400원으로 4만원 가까이 하락했다. 이에 따라 한때 30만원까지 치솟았던 킹크랩 한 마리 가격도 17만5000원으로, 20만원 밑으로 하락했다. 킹크랩 시세 하락 배경으로는 러시아·우크라이나 전쟁으로 인한 물량 증가가 꼽힌다. · · ·
CLOVA Summary API
고급 식자재로 불리는 킹크랩 가격이 4년 만에 1㎏당 7만원대로 떨어졌다. 지난달 18일까지 ㎏당 11만5000원이던 레드 킹크랩 가격은 하루 만에 7만7400원으로 4만원 가까이 하락했다. 이에 따라 한때 30만원까지 치솟았던 킹크랩 한 마리 가격도 17만5000원으로, 20만원 밑으로 하락했다.
제안 모델
킹크랩 가격이 4년 만에 1㎏당 7만원대로 떨어졌다. · · · 이에 따라 킹크랩 한 마리 가격도 17만5000원으로, 20만원 밑으로 하락했다. 하락 배경으로는 러시아·우크라이나 전쟁으로 인한 물량 증가가 꼽힌다. 전쟁 이후 미국과 유럽이 러시아산 해산물 수입을 금지하면서 러시아는 자국 냉동 창고가 포화 상태에 이르렀다고 한다. · · ·

그림 3. 텍스트 요약 모델 비교

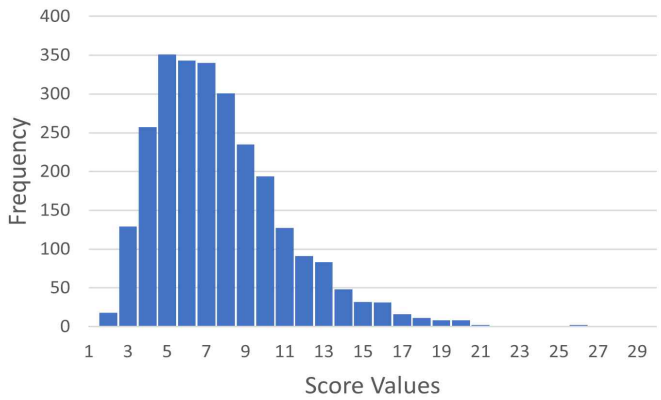

그림 4. Rouge Score 분포

4. 텍스트 요약 모델 평가

텍스트 요약 모델 성능 평가를 위해, 사건 단어 주의 집중 메커니즘을 활용해서 문장 요약을 시도한 연구[8]에서 사용된 2,865개의 원본 문서와 정답 요약 문서 쌍으로 구성된 데이터 세트를 사용하였다. 요약된 문서가 원본 문서의 핵심 어절을 포함하면서 내용을 적절히 요약하고 있는지 평가하기 위해, 모델이 생성한 요약문과 정답 요약문의 유사 정도를 평가하였다. 전체 데이터에 대해서 코사인 유사도를 계산 후 평균을 내었을 때 약 0.68의 유사도를 보였다.

또한, 모델이 생성한 요약문과 정답 요약문 사이의 연속된 일치를 고려하여 가중치를 할당하고, 값이 작을수록 두 문장이 유사함을 의미하는 ROUGE-W[9] 점수를 계산하였다. 전체 데이터의 ROUGE-W 점수 분포를 보았을 때, [그림 4]와 같이 0에 가깝게 분포함을 확인할 수 있다.

[그림 3]과같이, 본 논문에서 제시한 문서 요약 모델과 네이버 기사 요약 서비스의 요약 결과를 비교했다. 원문 기사[10]의 제목을 보면, 사건의 현상에 대한 이유를 전달하는 것이 이 기사의 목표임을 알 수 있다. 그러나 네이버 기사 요약 서비스가 생성한 요약문은 AI가 판단했을 때 중요하다고 생각되는 문장 3개만을 추출하기 때문에, 현상에 관한 내용만을 포함하고 원인에 관한 내용은 포함하지 않고 있다. 반면 제안 모델을 통해 요약된 문서는 현상의 원인에 대한 핵심 문장을 모두 포함하여 기존 글의 전체적인 흐름을 잘 대변하고 있음을 확인할 수 있다.

5. 결론

본 논문에서는 단어 추출 방식을 기반으로 한국어의 문법적 특성과 문장 내 단어 간의 자연스러운 정도를 고려하며, 삭제되는 문장 없이 문서의 모든 흐름을 담을 수 있는 요약 모델을 제안하였다. 본 연구의 요약문 생성 결과는 다른 요약 모델과 비교했을 때, 허위 정보를 생성할 가능성이 없고 문서의 모든 흐름이 담긴다는 결과를 보였으나, 문장 간의 연결이 부자연스러운 경우가 존재하고 어절의 삭제로 인해 각 문장 내에서 문법적, 문맥적 완성도가 떨어지는 경우가 있음을 확인하였다. 또한, 본 연구에서 제안한 모델은 개체명 인식에 약하다는 한계를 보였다. 따라서 향후 연구로 국립국어원이 공개한 개체명 데이터를 활용하여 요약 모델이 개체를 정확히 인식할 수 있도록 개선할 예정이다.

7. 참고 문헌

[1] Brown, Tom, et al. “Language models are few-shot learners.“ Advances in Neural Information Processing Systems. 33 (2020)
[2] F.Jelinek et al. “Perplexity —a measure of the difficulty of speech recognition tasks” ASA 1977
[3] Ji et al. “Survey of Hallucination in Natural Language Generation.” ACM Computing Surveys (2022)
[4] Niu et al. “Deleter: Leveraging BERT to Perform Unsupervised Successive Text Compression.” Salesforce Research (2019)
[5] 이준범 et al. “언어 정보를 반영한 문장 점수 측정 기반의 문장 압축.” 한국정보처리학회 춘계학술발표대회(2021)
[6] Brown et al. “Class-Based n-gram Models of Natural Language” ACL Anthology (1992)
[7] 박은정, 조성준. “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지” 2014년도 제26회 한글 및 한국어 정보처리 학술대회/제 2014권 제10호
[8] 정이안 et al. “사건 단어 주의 집중 메커니즘을 적용한 단일 문장 요약 생성,” 정보과학회논문지 47(2), p. 155-161, 2020.
[9] Lin, Chin-Yew. “Rouge: A package for automatic evaluation of summaries.“ Text summarization branches out. 2004.
[10] 이가영. “킹크랩 가격, 갑자기 4만원 뚝... 7만원대로 급락한 이유” 조선일보. 2023.10.12.