형태소 분석 및 Perplexity를 활용한 한국어 텍스트 요약 모델

오유솔 (wina0817@khu.ac.kr), 박상근 (sk.park@khu.ac.kr)
Department of Software Convergence, Kyung Hee University

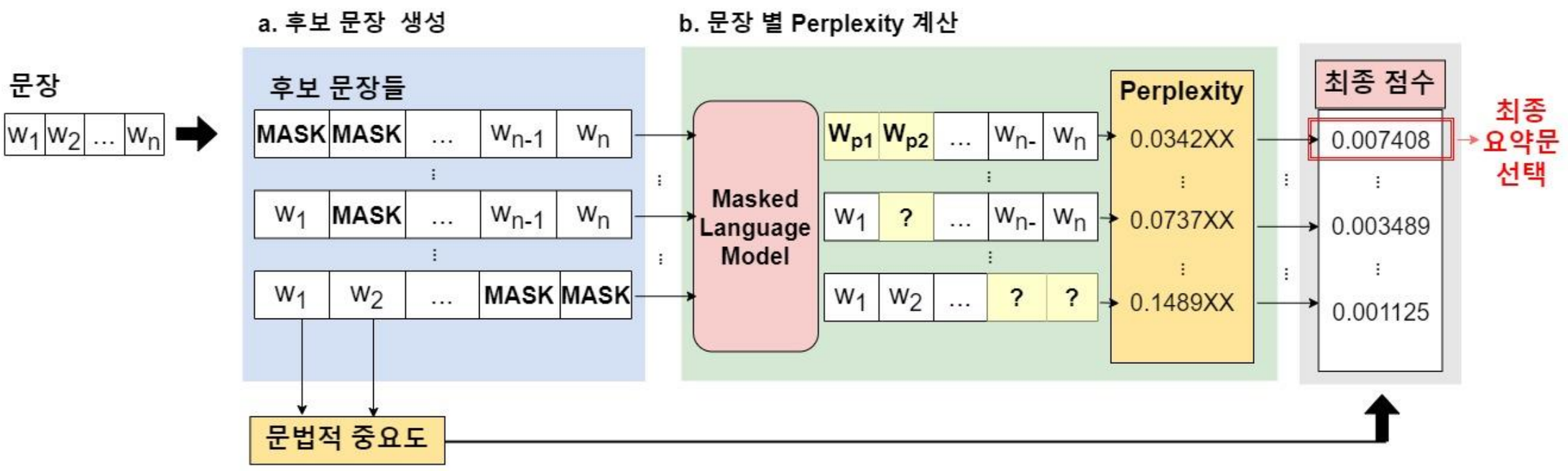
경희대학교 KYUNG HEE UNIVERSITY

서론

- 인터넷에 수많은 정보가 쏟아짐에 따라, 사람들은 다양한 정보를 더 빠르고 쉽게 습득하고자 한다.
  이에 따라, 짧은 영상과 글들을 선호하게 되었고, 이러한 경향에 맞게 긴 문서를 간결하게 요약해 주는 텍스트 요약 서비스가 증가하였다.
- 그러나, 단순히 텍스트를 짧게 요약하면 기존의 정보가 과장되거나 왜곡되어 소비자에게 전달 될 수 있기 때문에,
  많은 정보가 담긴 문서를 간결하게 요약하되, 문서의 전체 흐름을 정확하게 전달할 필요가 있다.
- 본 논문에서는
  ① **생성 요약 모델**이 거짓 정보 또는 허위 정보를 생성하는 **Hallucination을 발생**할 수 있다는 점과,
  ② **기존의 문장 추출 요약 모델**이 문서 전체 내용을 전부 대변하지 못해 내용이 **과장되거나 왜곡되어 전달**될 수 있다는 점을

해결하기 위해, **한국어의 문법적 특성을 고려한 단어 추출 요약 모델을 제안**한다.

텍스트 요약 모델



a. 후보 문장 생성

요약할 문서를 구성하는 모든 문장을 하나씩 구분 → N-gram 기법을 활용하여 후보 문장 생성

b. 문장 별 Perplexity 계산

KoBERT 데이터세트로 학습시킨 Masked Language Model을 사용해서 후보 문장마다 **Perplexity** 계산
- Perplexity : 문장의 혼란스러운 정도를 의미.
- Perplexity가 낮을 수록 문장의 혼잡도가 낮고, 문법적으로 오류가 적음을 의미

c. 문장 별 문법적 중요도 계산

KOMORAN을 사용해서 각 후보 문장에서 문법적으로 중요한 의미를 갖는 형태소(고유명사, 일반명사, 동사, 외국어, 주격 조사, 목적격 조사)에 해당하는 어절을 찾기
→ 해당 어절에 **문법적 중요도** 점수 부여

d. 최종 점수 계산 및 최종 요약문 선택

후보 문장마다 계산된 Perplexity와 문법적 중요도 값을 활용하여, Perplexity의 역수와 문법적 중요도를 곱한 값이 가장 높은 문장을 최종 요약 문장으로 선정

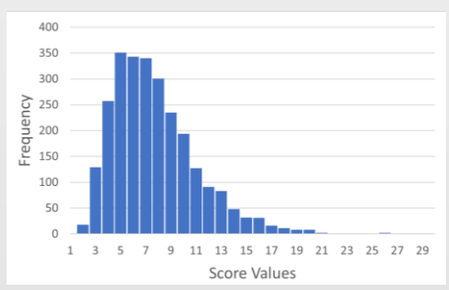
원문	만약 지방간의 원인이 되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 복용을 중단하거나 다른 약물로 대체해야 한다.			
순위	후보 문장	Perplexity	문법적 중요도	최종 점수
1	만약 지방간의 원인이 되는 약물을 복용하고 있다면 복용을 중단하거나 다른 약물로 대체해야 한다.	0.000030463	0.0014	45.957240
2	되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 복용을 중단하거나 다른 약물로 대체해야 한다.	0.000030556	0.0014	45.817026
3	만약 지방간의 원인이 되는 약물을 복용하고 있다면 주치의와 상의하여 약물의 약물로 대체해야 한다.	0.000030966	0.0014	45.209502
	…	…	…	…

텍스트 요약 모델 평가

- '원본 문서'와 '정답 요약 문서' 쌍으로 구성된 데이터셋을 사용하여 평가 진행

1. '모델이 생성한 요약문'과 '정답 요약문'의 **코사인 유사도** 평균 : 약 0.68
2. '모델이 생성한 요약문'과 '정답 요약문'의 **ROUGE-W 점수**가 0에 가깝게 분포

▼ **ROUGE-W 점수 분포**



3. **NAVER의 CLOVA Summary API와의 기사 요약 결과 비교**

결론

코사인 유사도 및 ROUGE-W를 통한 정량적 평가와, 네이버 기사 요약 서비스와의 직접적인 비교를 통해, 제안 모델이 **문장 내 단어 간의 자연스러운 정도를 고려하며**, 문서에서 삭제되는 문장 없이 **전체적인 흐름을 잘 대변**하고 있음을 확인함.

그러나, 다음과 같은 해결이 필요한 한계점을 발견.
① 문장 간의 연결이 부자연스러운 경우가 존재
② 어절의 삭제로 인해 문장 내에서 문맥적 완성도가↓

원문
"킹크랩 가격, 갑자기 4만원 뚝… 7만원대로 급락한 이유" 고급 식자재로 불리는 킹크랩 가격이 4년 만에 1㎏당 7만원대로 떨어졌다. • • • 지난달 18일까지 kg 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.** • • •