환각 현상 완화를 위한 단위 사실 기반 사후 교정

이용환[1○], 신정완[2], 송현제[1]
전북대학교 컴퓨터인공지능학부[1], 경북대학교 컴퓨터학부[2]
{lyhthy6,hyunje.song}@jbnu.ac.kr jwshin@knu.ac.kr

Atomic Unit-based Post Editing for Hallucination Reduction

Yonghwan Lee[1○], Jeongwan Shin[2], Hyun-Je Song[1]
Department of Computer Science and Artificial Intelligence, Jeonbuk National University[1]
Graduate School of Computer Science Engineering, Kyungpook National University[2]

요 약

환각 현상이란 LLM이 생성 태스크에서 사실이 아닌 내용을 생성하거나 근거가 없는 내용을 생성하는 현상을 말한다. 환각 현상은 LLM이 생성한 출력물에 대한 사용자의 신뢰를 떨어뜨리기 때문에 환각을 완화할 수 있는 방법이 필요하다. 최근 사후 편집 모델 중 하나인 RARR는 입력 텍스트를 질문들 순서에 따라 순차적으로 편집하여 환각을 완화하였지만 이전 단계의 편집 오류가 전파되거나 같은 작업을 반복하는 등의 단점이 있었다. 본 논문은 환각 현상 완화를 위한 단위 사실 기반 사후 교정을 제안한다. 제안한 방법은 입력 텍스트를 단위 사실로 분해하고 각 사실에 대응하는 질문을 생성한 후 검색된 관련 문서로 환각 여부를 판단한다. 환각이라 판단되면 편집을 수행하여 환각을 완화한다. 병렬적으로 편집을 진행하기 때문에 기존 연구의 순차적인 오류 전파 문제를 해결하고 기존 연구에 비해 더 빠른 사후 편집을 진행할 수 있다. 실험 결과, 제안 방법이 RARR보다 Preservation Score, 원문과의 사실성 일치여부, 의도 보존 여부에서 모두 우수한 성능을 보인다.

주제어: 대규모 언어 모델(LLM), 환각 현상 완화, 사후 편집, 검색 기반 병렬 편집기

1. 서론

최근 대규모 언어 모델(Large Language Model, 이하 LLM)은 지속적으로 발전하여 자연어 처리 분야의 생성 요약 [1], 다단계 추론 [2, 3, 4] 등과 같은 텍스트 생성 작업에서 뛰어난 성능을 보여준다. 언어 모델의 발전에도 불구하고 LLM은 생성 태스크에서 사실이 아닌 내용을 생성하거나 근거가 없는 내용을 생성하는 '환각 현상(Hallucination)'을 발생시킨다 [5, 6]. 환각 현상은 LLM이 생성한 출력물에 대한 사용자의 신뢰를 떨어뜨리며, 결과적으로 사용자 만족도가 악화된다 [7].

기존 연구들은 LLM의 환각 현상을 완화하기 위해 웹에서 검색한 문서들이 사실이라는 가정을 두고 출력과 관련된 검색 문서를 참고하는 방법들을 제시하였다 [1, 2, 8]. 기존 연구들은 크게 검색 문서들을 조건으로 텍스트를 생성하는 검색 증강 모델(Retrieval Augmented Model) [9, 10]이나 LLM으로부터 이미 생성된 텍스트가 검색 문서들과 일치하는지 여부에 따라 단위 사실별로 수정하는 사후 편집 모델(Post-hoc Editing)로 나뉜다 [2, 8]. 사후 편집 모델은 LLM으로부터 이미 생성된 텍스트를 수정하기 때문에 언어 모델의 종류에 상관없이 편집된 출력을 입력으로 바로 사용할 수 있다는 장점이 있다. 대표적인 사후 편집 모델인 RARR [2]는 입력 텍스트를 바탕으로 질문을 생성하고 관련된 문서를 검색해 순차적으로 교정한다. RARR는 질문에 따라 입력 텍스트를 순차적으로 교정하기 때문에 이전 교정 작업에서 오류가 발생할 경우 다음 교정 작업에 오류가 전달될 수 있다. 그림 1은 RARR의 방법을 나타낸 것이다. 첫

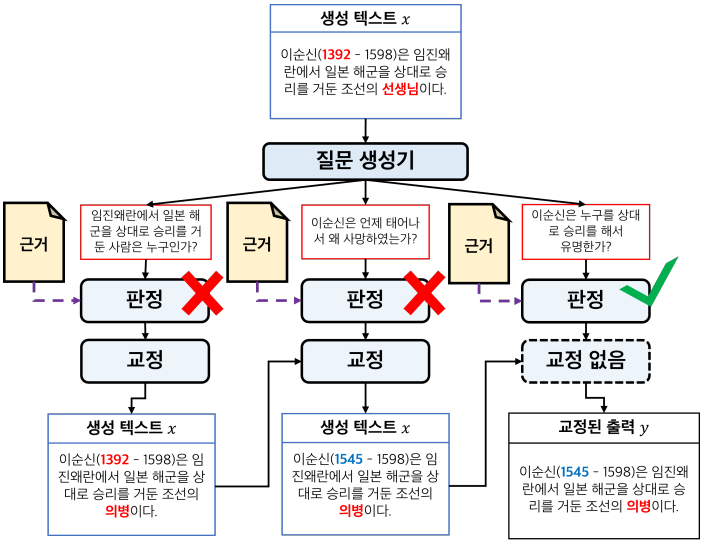

그림 1. RARR의 오류 예시

교정 단계를 보면 '임진왜란에서 일본 해군을 상대로 승리를 거둔 사람은 누구인가?'라는 질문에 '이순신은 조선의 의병이다.'라고 잘못된 교정이 이루어진다. 이는 여러 사실로 구성된 복잡한 질문으로 인해 의병이 임진왜란에서 승리하였다는 검색 문서를 근거로 삼았기 때문이다. 이 교정 오류는 다음 교정에서도 고쳐지지 않고 전파되는 것을 알 수 있다.

또한 사후 편집 모델은 같은 편집을 여러 번 반복해야 하기 때문에 비효율적이다. 그림 1을 보면 입력 텍스트 x가 순차적

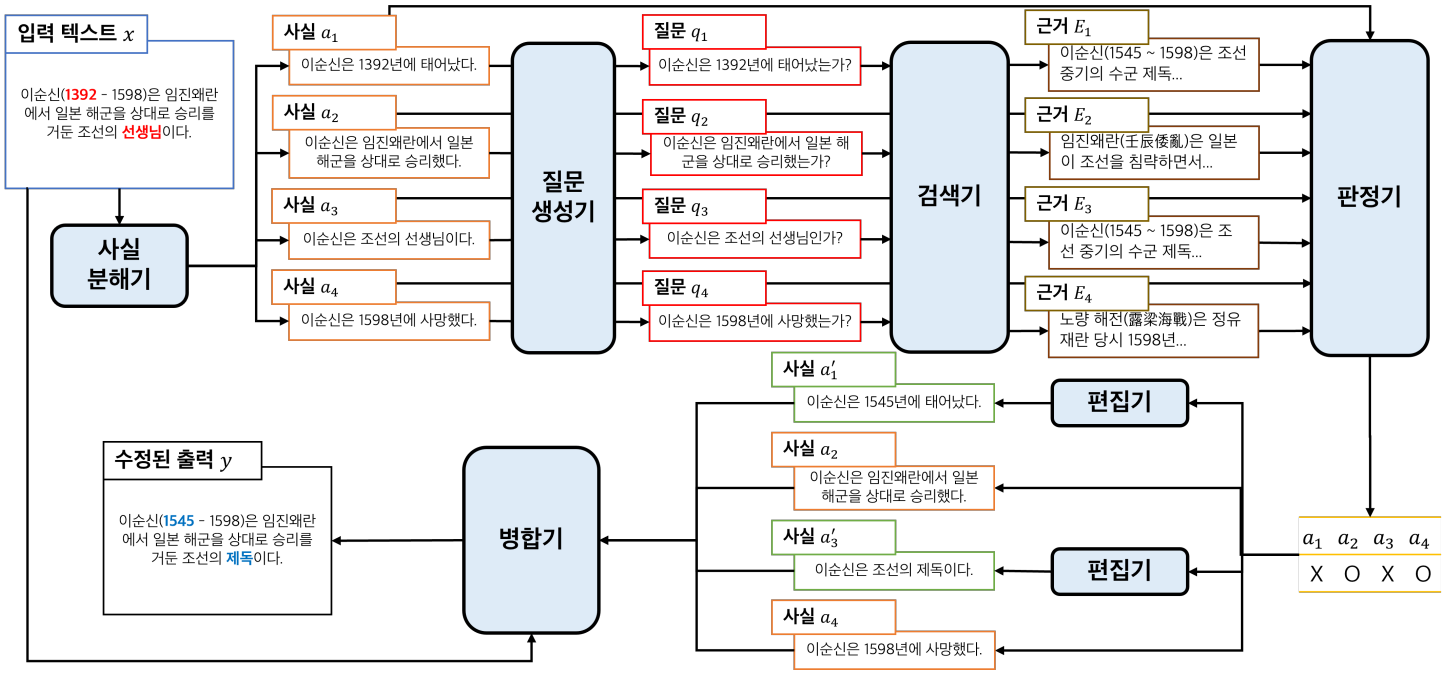

그림 2. 분해된 단위 사실을 활용한 검색 기반 병렬 편집기

으로 같은 편집 과정을 반복해서 거치고 있음을 알 수 있다. 그럼에도 이순신의 출생연도에 대한 환각은 편집이 이루어졌지만 이순신의 직업에 대해서는 여전히 환각이 남아있다.

기존에는 입력 텍스트의 여러 사실들을 따로 살펴보지 않고 그대로 편집하였기 때문에 환각을 놓치거나 잘못 편집하는 경우가 발생했다. 만약 입력 텍스트를 단위 사실로 나누어 확인한다면 부분별로 모든 사실을 확인하기 때문에 편집 오류가 줄어들 수 있다. 또한 단위 사실로 나누면 각 사실간 중첩이 없어 병렬적으로 편집이 가능하므로 순차적으로 편집해 발생했던 오류 전파의 문제가 해결된다. 본 논문은 환각 현상 완화를 위한 단위 사실 기반 사후 교정을 제안한다. 제안 방법은 입력 텍스트를 단위 사실로 분해하기 때문에 단위 사실별로 환각 완화가 가능하며 병렬적으로 편집하기 때문에 한번에 편집할 수 있다. 먼저 사실 분해기로 입력 텍스트의 하나 이상의 사실들을 단위 사실들로 분해하여 질문 생성기로 각 사실과 1:1로 대응하는 질문들을 생성한다. 각 질문들은 검색기를 통해 단위 사실과 관련된 문서들을 검색하여 단위 사실의 사실 여부를 판별한다. 이를 위해 판정기는 각 사실들이 검색된 문서와 사실 여부가 일치하는지 판단하고 일치하지 않는다면 편집기가 사실을 검색된 문서에 맞게 수정한다. 마지막으로 환각이 완화된 텍스트를 생성하기 위해 병합기는 교정이 끝난 모든 사실들을 병합하여 편집을 마무리한다. 제안 방법은 위 과정을 병렬적으로 진행하기 때문에 순차적 편집보다 빠르다.

본 논문은 실험에서 사용하기 위해 국립국어원의 문서 요약 말뭉치를 이용하였다. 환각 데이터를 만들기 위해 요약문의 개체 중 일부를 원문의 개체와 교체하였다. 말뭉치의 모든 원문 데이터는 검색기가 관련 문서를 찾아올 때 검색할 소스 말뭉치로 사용하였다. 제안 방법이 얼마나 환각을 완화하는지 알아보기 위해 RARR [2]와 비교하는 실험을 진행하였다. 불필요한 편집을 하는지 평가하기 위해 레벤슈타인 편집 거리 [11]를 활용한 Preservation Score [2, 8]를 계산하였고 편집된 텍스트의 사실성 일치, 의도 보존 여부는 3명의 평가자를 통해 평가하였다. 실험 결과, 제안 방법은 RARR에 비해 Preservation Score, 사실성 일치 여부, 의도 보존 여부 모두 높은 수치를 보였다.

2. 관련 연구

2.1 검색 증강 모델

언어 생성 과제에서 환각 현상이 나타나는 문제를 완화하기 위해 입력 텍스트에 대해 검색 문서를 조건으로 지정하여 텍스트를 생성하는 검색 증강 모델이 제안되었다 [9, 10, 2]. 검색 증강 모델은 검색 문서의 사실 정보를 추가적인 정보로 반영할 수 있기 때문에 환각 문제가 완화된 텍스트를 생성한다. Guu et al [9]은 언어 모델이 입력 텍스트뿐만 아니라 검색한 문서도 조건으로 주고 텍스트를 생성한다. Guu et al은 입력 텍스트와 관련된 문서를 검색하고 이를 입력 텍스트와 순차적으로 연결한다. 연결된 텍스트의 일부를 마스크 토큰으로 대체하고 이를 예측하도록 훈련시킨다. 훈련된 인코더를 추론 시 입력된 질문과 관련된 문서를 검색하고 조건으로 삼아 답변을 생성한다. Borgeaud et al [10]은 검색된 문서의 임베딩 벡터를 크로스 어텐션의 키와 벨류 값으로 반영하여 텍스트를 생성한다. 검색

증강 모델은 환각을 완화시키는데 좋은 성능을 보였지만 모델에 의존적이기 때문에 다양한 모델에 적용하기는 어렵다는 단점이 있다.

2.2 사후 편집 모델

사후 편집 모델(Post-hoc Editing)은 입력 텍스트와 관련된 문서를 검색하고 이를 근거로 편집을 진행한다. Thorne et al [12]은 입력 텍스트가 사실과 일치하도록 수정하는 Corrector를 Masking을 통해 학습시키고 추론 시 학습된 Corrector가 Wikipedia 근거를 통해 입력 텍스트를 수정한다. Gao et al [2]은 Thorne et al로부터 영감을 받아 입력 텍스트를 관련된 검색 문서들을 근거로 교정하는 방식이다. 먼저 입력과 관련된 질문들을 생성하고 이 질문과 관련된 문서들을 검색한다. 검색된 문서와 입력된 내용이 일치하는지 여부를 순차적으로 판단하여 입력을 교정한다. Gao et al이 제시한 방식은 입력 문장이 복잡한 사실들로 이루어진 경우 내용이 중복되는 질문을 생성할 수 있고 순차적으로 편집을 진행하기 때문에 이전 단계의 편집 오류가 다음 단계로 전달될 수 있다는 단점이 있다.

3. 환각 현상 완화를 위한 단위 사실 기반 사후 교정

본 논문에서는 환각 현상 완화를 위한 단위 사실 기반 사후 교정을 제안한다. 그림 2는 본 논문에서 제안한 환각 현상 완화를 위한 단위 사실 기반 사후 교정의 전체 아키텍쳐이다. 먼저 입력 텍스트의 복잡한 사실관계를 나누어 편집하기 위해 사실 분해기가 입력 텍스트를 단위 사실로 분해한다. 단위 사실들은 질문 생성 과정에서 질문 생성기를 통해 각 단위 사실에 1:1로 대응하는 질문을 생성한다. 생성된 질문으로 검색기는 관련된 근거 문서를 검색하고 판정기가 각 단위 사실이 근거 문서와 일치하는지 판단한다. 마지막으로 근거와 일치하지 않은 단위 사실들만 편집기가 편집을 진행하고 병합기가 모든 단위 사실을 병합하여 환각이 완화된 텍스트를 생성한다.

본 논문의 제안방법은 질문 생성 과정부터 편집 과정까지는 단위 사실에 따라 이루어지므로 병렬적으로 실행이 가능하다. 병렬적으로 진행하면 순차적으로 생성하는 $O(n)$에 비해 상대적으로 $O(1)$에 해당하는 시간이 걸리므로 효율적이고, 순차적 편집 중 이전 순서의 편집이 잘못되었을때 다음 순서의 편집으로 오류가 전파되는 문제점이 발생하지 않는다.

3.1 사실 분해기

사실 분해기는 입력 텍스트를 단위 사실로 세분화하여 편집하기 위해 LLM Few-shot 프롬프트 [13]로 입력 텍스트 x를 최소 단위 사실들로 분해한다($x = \{a_1, ..., a_N\}$, 단, N은 입력 텍스트에 따라 달라진다.). 각 단위 사실들은 하나의 사실로 이루어져 길이가 짧아 긴 입력 텍스트를 바로 편집하는 것에 비해 쉽게 편집할 수 있다. 또한 단위 사실들은 서로 중첩되지 않도록 생성된 사실이므로 사실 분해기 이후부터 병합기 전까지 병렬적으로 편집이 가능하다. 단위 사실을 추출할 때 사용한 프롬프트는 표 1에서 확인할 수 있다. 그림 2의 예시를 보면 입력 텍스트가 '이순신(1392 - 1598)은 임진왜란에서 일본 해군을 상대로 승리를 거둔 조선의 선생님이다.'로 주어지면, 사실 분해기는 이를 '이순신은 1392년에 태어났다.', '이순신은 임진왜란에서 일본 해군을 상대로 승리했다.', '이순신은 조선의 선생님이다.', '이순신은 1598년에 사망했다.'라는 단위 사실들로 분해한다.

3.2 질문 생성기

질문 생성기는 사실 분해기에서 분해된 각 단위 사실들과 관련된 문서를 검색하기 위해 질문을 생성한다. 사실 분해 과정에서 사실 분해기가 분해한 단위 사실들 이 질문 생성기에 입력으로 들어간다. 질문 생성기는 단위 사실 a_i와 1:1로 대응되는 질문 q_i를 생성한다. 사실 분해기와 마찬가지로 LLM Few-shot 프롬프트를 통해 질문을 생성한다. 단위 사실들로 부터 질문을 생성할 때 사용한 프롬프트는 표 1과 같다. 예를 들어, 그림 2의 질문 생성기는 '이순신은 1392년에 태어났다.'라는 사실에 1:1로 대응하는 '이순신은 1392년에 태어났는가?'라는 질문을 생성한다.

3.3 검색기

검색기는 생성된 질문을 통해 각 사실과 관련된 문서를 검색한다. 본 논문에서는 BM25를 사용하여 질문 생성기가 생성한 질문 q_i를 입력받아 근거 e_i를 찾는다. 근거 e_i는 단위 사실의 사실성을 판단하고 편집할 때 기준이 되어야하므로 각 사실과의 연관성을 기준으로 찾는다. 그림 2에서 보면 검색기는 '이순신은 1392년에 태어났는가?'라는 질문을 받아 '이순신(1545 1598)은 조선 중기의 수군 제독...'이라는 연관된 근거를 찾아낸다.

3.4 판정기

판정기는 검색된 문서를 바탕으로 단위 사실의 환각 여부를 판단한다. 이를 위해 단위 사실 a_i, 질문 q_i, 근거 e_i를 입력받아 불일치 여부 $edit_i$와 추론 $reason_i$를 생성한다. 불일치 여부의 동사가 '다르다'로 나올 경우($edit_i = True$) 편집기에 추론과 질문, 사실을 입력하고, '일치한다'로 나올 경우($edit_i = False$) 편집 과정을 생략한다. 판정기가 사용한 프롬프트는 표 1과 같다. 예를 들어, 그림 2를 보면 '이순신은 조선의 선생님인가?'이라는 질문에 검색기는 '이순신(1545 1598)은 조선 중기의 수군 제독...'이라는 내용의 관련 문서를 찾았다. 추론 $reason_i$는 '근거는 이순신이 조선 중기의 수군 제독이라 말했고, 입력된 사실은 이순신이 조선의 선생님이라고 말했다.'이고 결론이

표 1. 제안 방법에서 사용된 프롬프트 예제

사실 분해기
다음 문장을 독립적인 사실로 분해해 주세요: 사상 처음 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵은 국민을 하나로 묶고 정치가 해결할 수 없는 일을 해낼 것으로 기대된다.
- 사상 처음 아프리카대륙에서 남아프리카공화국 월드컵이 열린다. - 남아프리카공화국 월드컵은 2010년에 열린다. - 남아프리카공화국 월드컵은 국민을 하나로 묶을 것으로 기대된다. - 남아프리카공화국 월드컵은 정치가 해결할 수 없는 일을 해낼 것으로 기대된다.
질문 생성기
확인할 내용: 사상 처음 아프리카대륙에서 남아프리카공화국 월드컵이 열린다., 남아프리카공화국 월드컵은 2010년에 열린다., 남아프리카공화국 월드컵은 국민을 하나로 묶을 것으로 기대된다., 남아프리카공화국 월드컵은 정치가 해결할 수 없는 일을 해낼 것으로 기대된다. **이것을 확인하기 위해,**
1. 질문: 아프리카대륙에서 사상 처음으로 열리는 월드컵이 남아프리카공화국 월드컵인가? **2. 질문:** 남아프리카공화국 월드컵은 2010년에 열리는가? **3. 질문:** 남아프리카공화국 월드컵은 국민을 하나로 묶을 것으로 기대되는가? **4. 질문:** 남아프리카공화국 월드컵은 정치가 해결할 수 없는 일을 해낼 것으로 기대되는가?
판정기
1. 사실: 사상 처음 아프리카대륙에서 남아프리카공화국 월드컵이 열린다. **2. 질문:** 아프리카대륙에서 사상 처음으로 열리는 월드컵이 남아프리카공화국 월드컵인가? **3. 근거:** ... 중략 ... 사상 처음 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵도 축구 이상의 큰 의미를 담고 있다. **4. 추론:** 근거는 사상 처음 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵이라 말했고, 입력된 사실은 사상 처음 아프리카대륙에서 남아프리카공화국 월드컵이 열린다고 말했다. **5. 결론:** 이 사실은 근거와 일치한다.
편집기
1. 입력된 문장: 사상 두 번째로 아프리카대륙에서 열리는 월드컵은 남아프리카공화국 월드컵이다. **2. 근거:** ... 중략 ... 사상 처음 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵도 축구 이상의 큰 의미를 담고 있다. **3. 추론:** 근거는 사상 처음 아프리카대륙에서 열리는 월드컵은 남아프리카공화국 월드컵이라고 말했고, 입력된 문장은 사상 두 번째로 아프리카대륙에서 열리는 월드컵은 남아프리카공화국 월드컵이라고 말했다. **4. 수정된 문장:** 사상 처음 아프리카대륙에서 열리는 월드컵은 남아프리카공화국 월드컵이다.
병합기
1. 입력된 문장: 사상 두번째로 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵은 국민을 하나로 묶고 경제가 해결할 수 없는 일을 해낼 것으로 기대된다. **2. 사실들:** 사상 처음 아프리카대륙에서 남아프리카공화국 월드컵이 열린다., 남아프리카공화국 월드컵은 2010년에 열린다., 남아프리카공화국 월드컵은 국민을 하나로 묶을 것으로 기대된다., 남아프리카공화국 월드컵은 정치가 해결할 수 없는 일을 해낼 것으로 기대된다. **3. 병합된 문장:** 사상 처음 아프리카대륙에서 열리는 2010년 남아프리카공화국 월드컵은 국민을 하나로 묶고 정치가 해결할 수 없는 일을 해낼 것으로 기대된다.

'이 사실은 근거와 일치한다'이므로 결론 $edit_i$는 False이다.

3.5 편집기

편집기는 각 단위 사실들을 편집하여 환각을 완화한다. 편집기는 단위 사실 a_i와 근거 e_i, 추론 $reason_i$를 입력받아 단위 사실 a_i가 근거 e_i와 다른 부분을 편집하고 편집된 단위 사실 a'_i를 생성한다. 편집기도 LLM Few-shot 프롬프트를 이용하였고 사용한 판정기가 사용한 프롬프트는 표 1에서 확인할 수 있다. 예시로, 그림 2를 보면 a_1, a_3는 근거와 불일치하므로 편집기가 편집을 진행한다. 단위 사실 '이순신은 1392년에 태어났다.'는 '이순신은 1545년에 태어났다.'로, '이순신은 조선의 선생님이다.'는 '이순신은 조선의 제독이다.'로 편집되어 환각이 완화됨을 볼 수 있다.

3.6 병합기

병합기는 편집기가 편집한 단위 사실들을 처음 입력 텍스트와 같은 스타일로 병합하여 편집 작업을 마무리한다. 편집된 단위 사실 a'_i들을 병합기가 입력으로 받고 편집이 완료된 텍스트 y를 출력한다. 제안 방법은 병합기가 입력 텍스트 x도 입력으로 받아 입력 텍스트와 편집이 완료된 텍스트 y가 유사하도록 병합시켰다. 병합기가 사용한 프롬프트는 표 1에서 확인할 수 있다. 예를 들어, 그림 2를 보면 병합기가 편집이 완료된 단위 사실들 '이순신은 1545년에 태어났다.', '이순신은 임진왜란에서 일본 해군을 상대로 승리했다.', '이순신은 제독이다.', '이순신은 1598년에 사망했다.'를 '이순신(1545 - 1598)은 임진왜란에서 일본 해군을 상대로 승리를 거둔 조선의 제독이다.'라는 텍스트로 병합하여 출력함을 볼 수 있다.

4. 실험

4.1 데이터셋 구성

실험에 사용할 데이터셋으로 국립국어원의 문서 요약 말뭉치를 이용하였다. 데이터의 모든 원문들을 검색기가 검색할 소스 말뭉치로 간주하고 요약문은 입력 텍스트로 사용하였다. 환각 데이터를 만들기 위해 입력 텍스트의 개체를 원문의 개체로 교체하였다. 실험에 사용하는 데이터는 총 50개이며, 훈련 과정

표 2. 실험 결과 비교

비교 방법	Preservation	사실성 일치	의도 보존
RARR(Gao et al.)	0.61	3.6	2.14
제안 방법	**0.68**	**3.77**	**2.17**

없이 추론에만 사용된다.

4.2 실험 설정

제안 방법의 성능을 알아보기 위해 RARR [2]와 비교를 진행하였다. RARR과 제안 방법 모두 LLaMa-2 [14] 기반 Upstage 사의 SOLAR-0-70b-8bit 모델을 사용하였다. Few-shot 프롬프트를 위해 Few-shot 예제는 사실 분해기, 질문 생성기, 판정기, 편집기, 병합기에 4개의 예제를 제공하였다. 비교 방법인 RARR는 Gao et al의 논문에 제시된 프롬프트를 한국어로 번역하여 사용하였고 각 과정을 논문에 따라 재현하였다. RARR와 제안 방법의 성능을 비교하기 위해 RARR의 논문에서 제시한 Preservation을 수식 (1)과 같이 재현하여 불필요한 편집이 발생하였는지 여부를 비교하였다. Preservation은 레벤슈타인 편집 거리($Lev(x, y)$) [11]를 이용하여 계산한다.

$$Pres_{(x,y)} = max\left(1 - \frac{Lev(x, y)}{length(x)}, 0\right) \quad (1)$$

편집된 텍스트와 근거 사이 사실성 일치 여부와 원래 의도 보존 여부는 3명의 사람이 평가하였다. [2] 사실성 일치 여부를 평가하기 위해 평가자에게 근거와 사실성이 전혀 일치하지 않을 경우 1점, 근거와 사실이 완벽히 일치할 경우 5점으로 두어 점수를 매기도록 하였다. 의도 보존 여부를 평가하기 위해 평가자에게 의도가 전혀 보존되지 않을 경우 1점, 의도가 완벽히 보존될 경우를 3점으로 점수를 매기도록 하였다.

4.3 실험 결과

실험 결과는 표 2를 통해 확인할 수 있다. 실험 결과, 제안 방법은 68.04, RARR는 61.51로 RARR에 비해 Preservation 점수가 6.53 만큼 높다. 이는 제안 방법이 RARR보다 불필요한 편집은 하지 않는다는 것을 보여준다. RARR와 달리 제안 방법은 단위 사실들을 각각 수정하여 병합하였기 때문에 중복되는 사실이 없어 편집 거리가 줄어들었기 때문이다.

사람이 평가한 사실성 일치의 경우 제안 방법은 3.77, RARR는 3.6으로 제안 방법이 RARR에 비해 사실과 일치하도록 편집을 수행한다. 제안 방법은 단순한 단위 사실로 분해하고 사실성을 검증하였기 때문에 입력 텍스트의 복잡한 사실관계를 판단해야 하는 RARR에 비해 모든 단위 사실에 대해서 편집을 수행하기 때문이다. 또한 RARR는 순차적으로 편집하면서 이전 편집의 오류가 다음 편집으로 전달될 수 있는 반면, 본 논문의 제안 방법은 단위 사실이 서로에게 영향을 주지 않고 병렬적으로 편집되기 때문에 RARR에 비해 사실성 일치 점수가 더 높다.

사람이 평가한 의도 보존의 경우에도 제안 방법은 2.17, RARR는 2.14로 평가되어 제안 방법의 편집은 RARR의 편집에 비해 원래 의도를 보존하면서 편집을 적절히 수행한다는 것을 알 수 있다. 제안 방법이 의도 보존도 뛰어난 이유는 단위 사실로 입력 텍스트를 분해하면서 원래 의도가 훼손 되지 않고 그대로 병합되는데 비해 RARR는 질문에 따라 순차적으로 편집하면서 처음 입력 텍스트의 의도도 함께 편집되기 때문이다.

4.4 어블레이션 연구

추가로 단위 사실이 편집에 미치는 영향을 알아보기 위해 단위 사실로 분해하지 않고 바로 질문을 생성하되, 제안 방법과 같이 병렬적으로 검색, 판정, 편집을 거치는 방법(Non-Atomic, Parallel)도 실험하였다. 단위 사실로 나누지 않고 병렬성만을 적용한 방법은 Preservation이 25.25로 제안 방법과 RARR에 비해 매우 낮았다. 이는 단위 사실로 나누지 않아 입력 텍스트의 모든 사실을 편집하는데 어려움을 겪었고 질문에 중첩되는 사실이 많았기 때문이다.

5. 결론

본 논문은 환각 현상 완화를 위한 단위 사실 기반 사후 교정을 제안하였다. 제안된 방법은 먼저 입력 텍스트를 단위 사실로 분해하고 각 사실에 대응하는 질문을 통해 관련 문서를 찾아와 편집을 수행하여 환각을 완화한다. 분해된 사실을 이용하여 부분적으로 모든 사실의 환각 여부를 확인하여 편집하기 때문에 기존 연구의 순차적인 오류 전파 문제를 해결하였다. 실험을 통해 제안 방법이 기존 RARR보다 Preservation Score는 6.53의 성능 향상을, 사람이 평가한 원문과의 사실성 일치여부는 0.17, 의도 보존 여부는 0.03 만큼의 성능 향상을 보였다. 이를 통해 본 논문이 제안한 단위 사실 중심의 병렬 편집 방식이 효과적임을 확인할 수 있었다.

감사의 글

본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1F1A1048181)

참고문헌

[1] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, Vol. 55, No. 12, mar 2023. [Online]. Available: https://doi.org/10.1145/3571730

[2] L. Gao *et al.*, "RARR: Researching and revising what language models say, using language models,"

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 16 477–16 508, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.910

[3] M. Nye *et al.*, "Show your work: Scratchpads for intermediate computation with language models," 2021.

[4] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Vol. 35, pp. 24 824–24 837, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[5] J. Maynez *et al.*, "On faithfulness and factuality in abstractive summarization," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.173

[6] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, and N. McAleese, "Teaching language models to support answers with verified quotes," 2022.

[7] D. Dale *et al.*, "Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–50, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.3

[8] A. Chen *et al.*, "Purr: Efficiently editing language model hallucinations by denoising language model corruptions," 2023.

[9] K. Guu *et al.*, "Realm: Retrieval-augmented language model pre-training," *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[10] S. Borgeaud *et al.*, "Improving language models by retrieving from trillions of tokens," *Proceedings of the 39th International Conference on Machine Learning*, pp. 2206–2240, 2022. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html

[11] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics. Doklady*, Vol. 10, pp. 707–710, 1965. [Online]. Available: https://api.semanticscholar.org/CorpusID:60827152

[12] J. Thorne *et al.*, "Evidence-based factual error correction," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3298–3309, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.acl-long.256

[13] T. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[14] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023.