그래프 기반 증거 추론을 이용한 질의 응답에 대한 사실 여부 검증 연구

전북대학교[1, 2], 네이버[3, 4, 5]
박은환[1], 나승훈[2], 신동욱[3], 김선훈[4], 강인호[5]

Content

1. 서 론

출처: https://factcheck.snu.ac.kr/

- 인터넷의 발전에 따라 검증되지 않은 방대한 양의 정보가 매일 쏟아지고 있음.
  - 사실 여부 검증을 위해선 **그림 1과 같이 많은 시간과 비용이 발생한다는 문제점이 존재**함.

1. 서 론

- 질의 응답에 대한 사실 여부 검증

  적절한 증거를 검색하고 이를 기반으로 사실 여부를 검증하는 태스크

질문: 영국의 록 밴드 오아시스가 결성된 도시는 ?

답변: 런던

증거 문장 1: 오아시스는 1991년 **영국 맨체스터**에서 결성된 록 밴드이다.

증거 문장 2: 오아시스가 앤디 벨, 겜 아처와 함께 작업한 첫 앨범인 Heathen Chemistry는 2002년 7월 발매되었다.

사실 여부: 사실이 아님

2. 관련 연구 [FEVER, Thorne et al. 2018]

- **위키피디아 문헌들을 기반으로 수동으로 구축된 FEVER 데이터 집합**

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]

Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

Figure 1: Manually verified claim requiring evidence from multiple Wikipedia pages.

- **Claim Generation**: 위키피디아 문헌들로부터 정보를 추출하고 Claim 을 생성
- **Claim Labeling**: Claim 이 Supported 혹은 Refuted 인지 증거를 바탕으로 분류하고 증거가 충분하지 않을 때는 NotEnoughInfo (NEI) 로 분류
- **Issue**: 한국어에는 FEVER 와 같은 **데이터 집합이 존재하지 않을 뿐**만 아니라 **질의에 대한 사실 여부 검증을 위한 데이터 집합 또한 존재하지 않음**.
  - BART를 이용한 질의에 대한 사실 여부 검증을 위한 데이터 집합 자동 구축

2. 관련 연구 [GEAR, Zhou et al. 2019]

- **GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification**

"SUPPORTED" Example	
Claim	The Rodney King riots took place in the most populous county in the USA.
Evidence	(1) The 1992 Los Angeles riots, ***also known as the Rodney King riots*** were a series of riots, lootings, arsons, and civil disturbances that ***occurred in Los Angeles County***, California in April and May 1992. (2) ***Los Angeles County***, officially the County of Los Angeles, ***is the most populous county in the USA***.

"REFUTED" Example	
Claim	Giada at Home was only available on DVD.
Evidence	(1) ***Giada at Home*** is a television show and first ***aired*** on October 18, 2008, ***on the Food Network***. (2) ***Food Network*** is an American ***basic cable and satellite television channel***.

- 사실 여부 검증은 **여러 증거를 보고 추론될 필요성**이 존재함.
- 검색된 증거를 바탕으로 완전 연결 그래프를 구축한 후 **그래프 기반 추론 신경망을 통하여 사실 여부 검증 태스크를 수행**함.

2. 관련 연구 [KorNLI and KorSTS, Ham et al. 2020]

- **KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding**

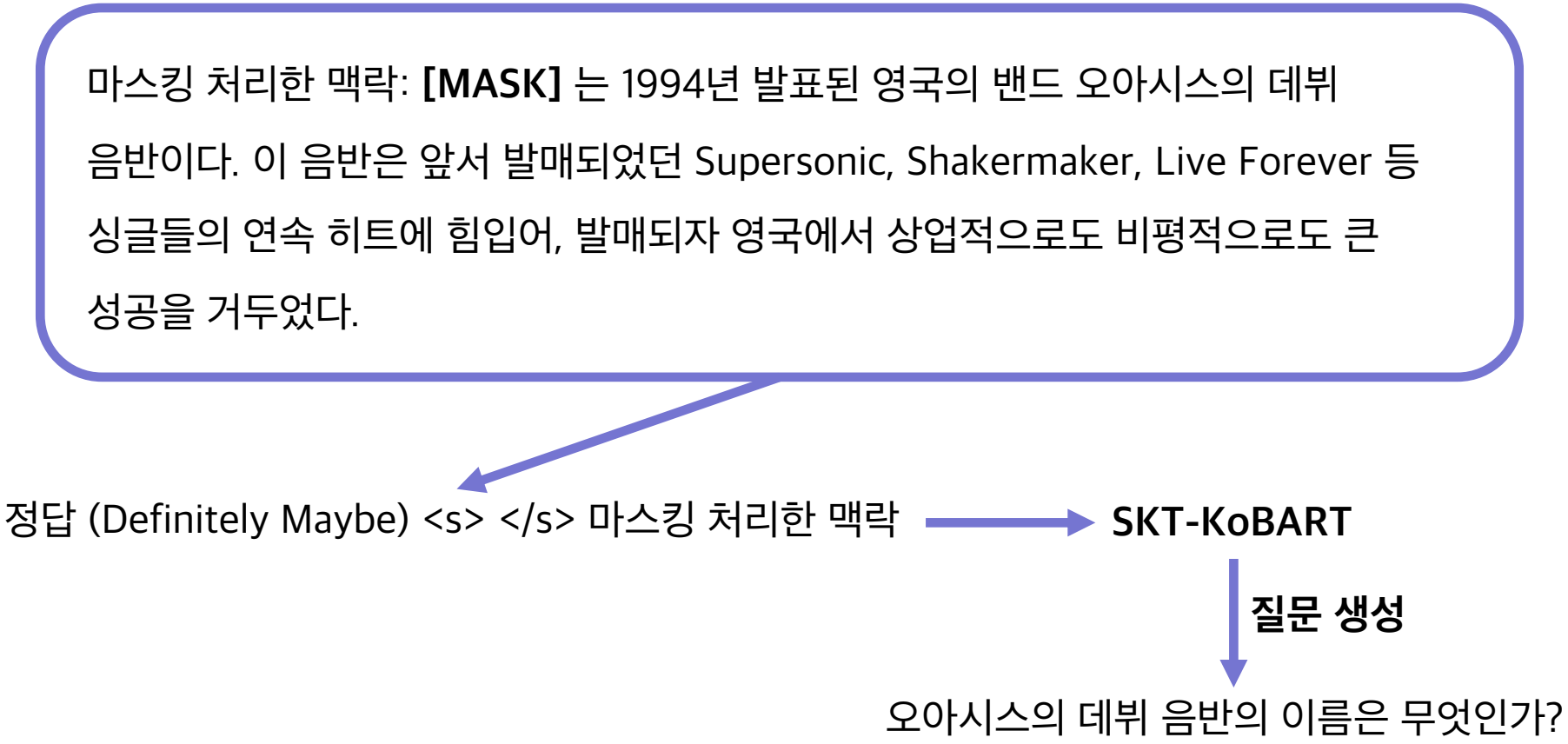
Examples	Label
P: 너는 거기에 있을 필요 없어. "You don't have to stay there." **H**: 가도 돼. "You can leave."	E
P: 너는 거기에 있을 필요 없어. "You don't have to stay there." **H**: 넌 정확히 그 자리에 있어야 해! "You need to stay in this place exactly!"	C
P: 너는 거기에 있을 필요 없어. "You don't have to stay there." **H**: 네가 원하면 넌 집에 가도 돼. "You can go home if you like."	N

Table 2: Examples from KorNLI dataset. **P**: Premise, **H**: Hypothesis. E: Entailment, C: Contradiction, N: Neutral.

- **KorNLI**: 전제와 가설로 구성한 한 쌍의 관계를 얽힘, 모순, 중립 중 하나로 분류
- 본 연구에서는 **이러한 태스크를 미세 조정한 언어 모델의 경우 질의와 증거 간의 관계성 보강에 좋은 영향을 끼칠 수 있다고 가정**함.

3. 제안 모델

- **BART 기반 질의 응답 데이터 집합 자동 생성**



- 형태소 분석기를 기반으로 **명사, 숫자 등을 마스킹**함
- **거짓의 경우 정답을 같은 형태소 타입의 단어로 교체**함.
  - 예: Definitely Maybe -> Supersonic

3. 제안 모델

- **BART 기반 질의 응답 데이터 집합 자동 생성**

생성된 질문	정답	라벨
비토 볼테라는 언제 이탈리아의 수학자이자, 물리학자, 생물학자이며 수리생물학과 적분방정식에 대한 업적에 의해 잘 알려져 있나?	1883년	Unsupport
1923년 위싱 파턴에서 개최된 여자기독교청년회세계대회(YWCA 세계대회)에 참석한 사람이 누구야?	오하이오 웨슬리언	Support

- **테스트 데이터 집합 구성**

테스트 집합 질문	정답	라벨
2014년 에베레스트 눈사태 때 사상자는 몇 명인가?	1883년	Unsupport
조금 이따 샤워해는 개리의 첫 솔로 음반 MR.GAE의 몇 번째 타이틀 곡인가?	두 번째	Support

- 위키피디아 문헌을 참고하여 100개의 테스트 데이터 집합을 구성함.

3. 제안 모델

- **증거 문서 검색 및 증거 문장 선택**

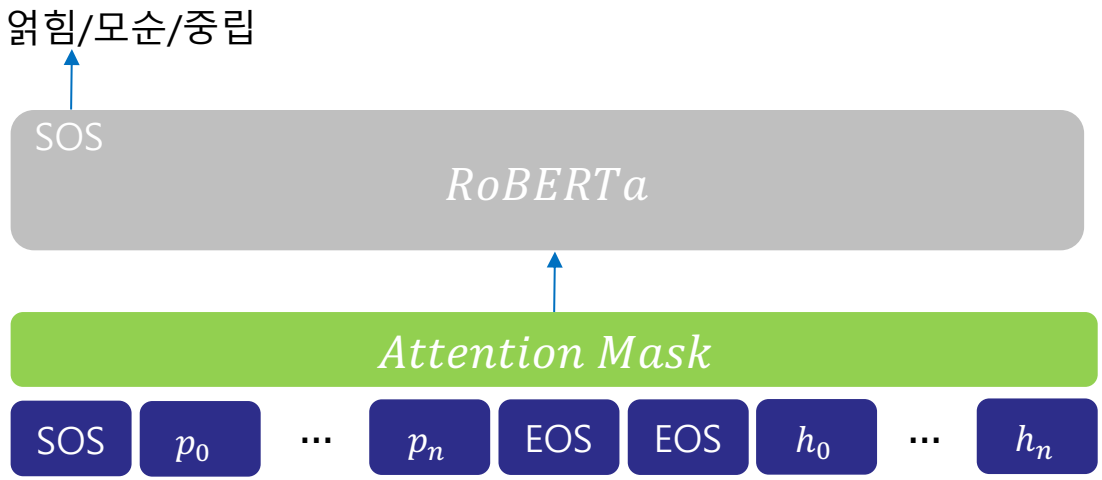
증거 문서 검색 결과 성능

K	Num Documents	MRR
10	437,299	65.34%
5	**437,299**	**65.06%**
1	437,299	58.00%

- REALM(Dense Retrieval) 로 437,299개의 문서를 검색하여 **질문과 관련된 top-k 개의 문서**를 추출함.
- TF-IDF 로 추출한 top-k 개의 문서로부터 **질문과 연관된 top-j 개의 문장**을 추출함.

3. 제안 모델

- **KorNLI 미세 조정을 이용한 언어 모델 표상 보강**

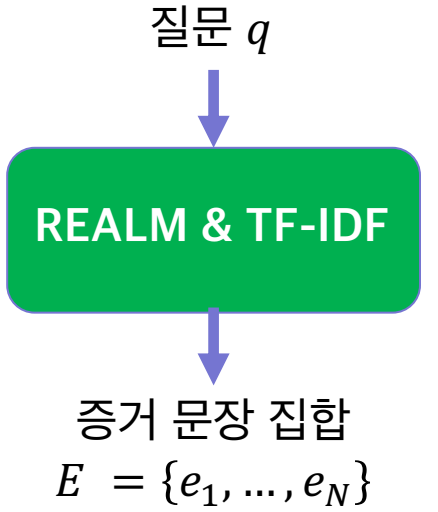


Model	Label	Precision	Recall	F1 Score	Accuracy	Macro F1 Score
RoBERTa-KorNLI	얽힘	83.72%	72.39%	77.64%	77.38%	77.49%
	중립	69.36%	80.53%	74.53%		
	모순	81.31%	79.22%	80.25%		

- **한국어 RoBERTa 언어 모델 기반 KorNLI 태스크에 대하여 미세 조정**을 진행
  - RoBERTa-KorNLI로 명칭함.
- 기존 RoBERTa 모델보다 **RoBERTa-KorNLI 모델로 사실 여부 검증 태스크를 실험하였을 때 더 좋은 성능**을 거두었음.

3. 제안 모델

- **그래프 기반 증거 추론을 이용한 질의 응답에 대한 사실 여부 검증 연구**
  - BART 기반 자동 구축 데이터 집합과 REALM & TFIDF 통한 증거 추출 및 RoBERTa-NLI 그래프 기반 증거 추론 신경망 구축



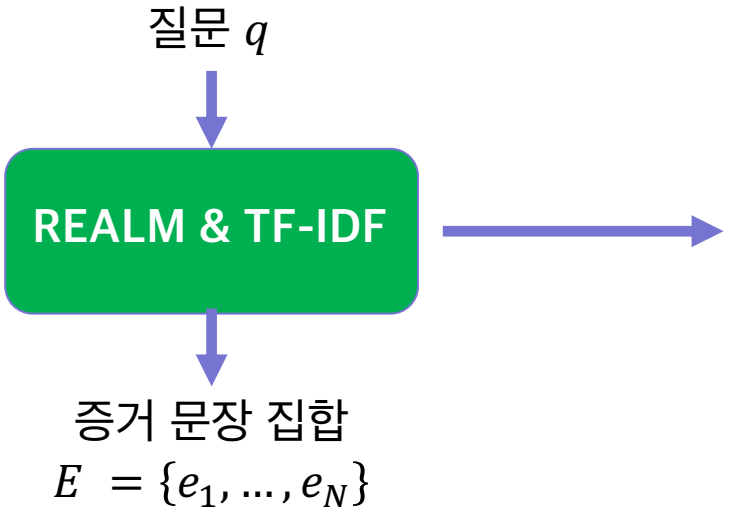
1. $c_i = [q; a; e_i]$ 로 입력 구성
   1. $[\cdot;\cdot]$은 병합 연산을 의미
2. $h_0^t = RoBERTa(c_i) \in \mathbb{R}^{F \times 1}$
   1. [SOS]에 해당하는 표상 추출
   2. 증거 문장의 개수만큼 표상 구성
      1. $h^t = \{h_1^t, \ldots, h_N^t\}$
3. 이웃 노드간 Attention Coefficients 계산 및 증거 추론 표상 추출
   1. $p_{ij} = W_1^{t-1}\left(ReLU\left(W_0^{t-1}\left(h_i^{t-1}; h_j^{t-1}\right)\right)\right)$

      $\alpha_{ij} = softmax(p_{ij})$

      $h_i^t = \sum_{j \in N} \alpha_{ij} h_j^{t-1}$
   2. $W_1^{t-1} \in \mathbb{R}^{H \times 2F}, W_0^{t-1} \in \mathbb{R}^{1 \times H}$

3. 제안 모델

- **그래프 기반 증거 추론을 이용한 질의 응답에 대한 사실 여부 검증 연구**
  - BART 기반 자동 구축 데이터 집합과 REALM & TFIDF 통한 증거 추출 및 RoBERTa-NLI 그래프 기반 증거 추론 신경망 구축



1. 추출한 그래프 증거 추론 표상으로 Mean Aggregation 연산
   1. $o = mean([h_1^t, \ldots, h_n^t])$
2. 사실 여부 예측
   1. $l = softmax\left(W_4\left(\tanh\left(W_3(o)\right)\right)\right)$
      1. $W_3 \in \mathbb{R}^{F \times F}, W_4 \in \mathbb{R}^{F \times 2}$

4. 실험 결과

Model	Graph Layer t	Precision	Recall	F1
RoBERTa(Q+A)	–	55.97%	56.00%	54.85%
RoBERTa+Mean	–	75.15%	75.00%	74.98%
RoBERTa+Max	–	78.51%	78.00%	77.97%
RoBERTa+GR+Max	1	78.00%	78.00%	77.97%
RoBERTa+GR+Mean	1	79.52%	78.00%	77.56%
RoBERTa+KorNLI+Mean	–	76.83%	76.00%	75.92%
RoBERTa+KorNLI+Max	–	77.02%	73.49%	73.49%
RoBERTa-KorNLI+GR+Max	1	78.44%	76.00%	75.66%
RoBERTa-KorNLI+GR+Mean	1	82.27%	80.00%	79.71%
RoBERTa-KorNLI+GR+Mean	**2**	**83.00%**	**83.00%**	**83.00%**
RoBERTa-KorNLI+GR+Mean	3	78.00%	78.00%	78.00%

5. 결 론

1. 질의 응답에 대한 사실 여부 검증 태스크를 위해 BART를 이용하여 데이터를 자동 구축.
2. KorNLI 를 미세 조정한 RoBERTa 모델로 그래프 기반 증거 추론 신경망으로 사실 여부 검증 태스크를 수행.
3. 향후 Aggregation 등을 보강한 연구를 진행할 예정.

감사합니다