

한국어 Fact 검증을 위한 자동 Claim 데이터 생성

이종현⁰¹, 나승훈¹, 신동욱², 김선훈², 강인호²

¹전북대학교, ²네이버

1. 서론
2. 사실 확인을 위한 데이터 셋 구축
3. 제안 모델
4. 실험
5. 결론

1. 서론

■ 사실 검증 (Fact Verification)

- 웹(Web) 상에서 쉽고 빠르게 수많은 정보를 얻을 수 있고 이러한 정보를 쉽게 타인과 공유할 수 있는 현대 사회에서 사실인 정보와 그렇지 않은 정보를 판단하는 것은 매우 중요함
- 국내에는 아직 자동 사실 확인 모델 구축을 위한 데이터 셋이 많이 부족한 상황임

■ 사실 검증 절차

- 1) 주장(claim)을 뒷받침 할 만한 증거 문서 검색
- 2) 검색된 문서들 중 적절한 증거 문장 선택
- 3) 뉴럴 모델을 통한 주장의 사실 확인

2. 사실 확인을 위한 데이터 셋 구축

■ 사실 확인 데이터 셋 자동 구축

- 위키피디아에서 자체 수집한 10,000 개의 문서를 이용해 사실 확인 데이터 셋을 자동 구축함

Question	Answer	Label
<조금 이따 샤워해>는 개리의 첫 솔로 음반 MR .GAE의 몇 번째 타이틀 곡인가?	두 번째	positive
피구는 몇 팀으로 나뉘어서 진행되는 게임인가?	세 팀	negative
공소효력의 범위에 대한 형사소송법의 조문은 대한민국 형사소송법 몇 조 인가?	248조	positive
《야심작 정열맨》의 1부는 언제 완결 되었나?	2019년 9월 2일	negative

표: 사실 확인을 위한 데이터 예시

- 질문(question) 과 정답(answer) 쌍을 주장(claim) 이라 함

2. 사실 확인을 위한 데이터 셋 구축

- 질문 자동 생성을 위한 Bart 모델 학습
 - KorQuAD 1.0 데이터 셋을 사용

Answer

22일부터 28일

Masked Context

한국청소년단체협의회와 여성가족부는 <mask>까지 서울과 충북 괴산에서 '국제청소년포럼'을 연다고 21일 밝혔다...

Input Text

Answer + </s><s> + Masked Context

Bart

Generated Question

한국청소년단체협의회와 여성가족부는 언제부터 언제까지 국제청소년포럼을 열어?

Context

한국청소년단체협의회와 여성가족부는 22일부터 28일까지 서울과 충북 괴산에서 '국제청소년포럼'을 연다고 21일 밝혔다...

Question

한국청소년단체협의회와 여성가족부는 언제부터 언제까지 국제청소년포럼을 열어?

Answer

22일부터 28일

KorQuAD 1.0

2. 사실 확인을 위한 데이터 셋 구축

■ 자동 주장 생성

- 주장을 사실인 주장(Supported)과 그렇지 않은 거짓인 주장(Refuted)의 2가지로 다룸

· Supported 주장 생성

- Masking 할 텍스트는 NER을 통해 선택하고 이를 정답으로써 사용

열린우리당 <mask> 제5정조위원장은 15일 민주노총 대의원대회가 강경파의 원천봉쇄로 무산되는데 대해 "민주노총은 극좌 맹동주의자와 결별하고 온건.합리적인 노동운동을 통해 새 길을 가야한다"고 밝혔다. 이 위원장은 이날 국회에서 브리핑을 갖고 "수많은 사람들의 죽음과 투옥으로 만들어진 민주노총이 한줌도 안되는 극좌파에 의해 나락으로 떨어지는 모습이 안타깝고 민주노총 지도부에 연민을 느낀다"며 이 같이 말했다. ...

<mask> : 이목희

이목희

Answer + </s><s> + Masked Context



Generated Question

열린우리당 제5정조위원장은 누구인가?

2. 사실 확인을 위한 데이터 셋 구축

· Refuted 주장 생성

- Supported 주장과 같은 방식으로 NER을 이용해 얻어낸 정답을 가지고 질문을 생성한 후에 정답과 같은 Type 의 멘션으로 정답을 교체하는 방식

<mask> : 22일부터 28일

한국청소년단체협의회와 여성가족부는 <mask>까지 서울과 충북 괴산에서 '국제청소년포럼'을 열고 21일 밝혔다. 한국 미국 캐나다 호주 등 전 세계 32개국 75여명의 대학생, 청소년들이 모여 전 세계적 현안문제에 대한 대안과 해결책을 모색하는 자리다. 이번 포럼의 주제는 '청소년과 뉴미디어'다. 스마트폰 SNS 태블릿PC 등 새로운 커뮤니케이션 매체인 '뉴미디어'에 대한 성찰과 문제점에 대해 토론한다. 기조강연을 시작으로 국가별 주제관련 사례발표, 그룹 토론 및 전체총회, '청소년선언문' 작성 및 채택 등 다양한 프로그램을 운영한다.

생성된 질문 : 한국청소년단체협의회와 여성가족부는 언제부터 언제까지 국제청소년포럼을 열어?

Answer 를 '22일부터 28일' 이 아닌 같은 [DATE] type 인 '21일'로 교체

정답 : 21일

3. 제안 모델

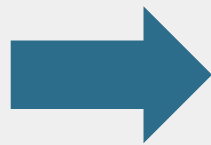
■ 증거 문서 검색 및 증거 문장 선택

- 주어진 주장이 사실인지 아닌지를 판단하기 위해서는 주장에 대한 증거(evidence)가 뒷받침 되어야 함
- 주장에 대한 증거 문서를 검색하기 위해 검색 모델인 REALM 을 이용함
- 검색된 증거 문서 중 적절한 증거 문장 선택 과정에는 TF-IDF 를 이용함



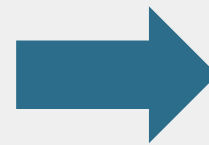
전체 문서 셋

REALM



검색된 증거 문서

TF-IDF

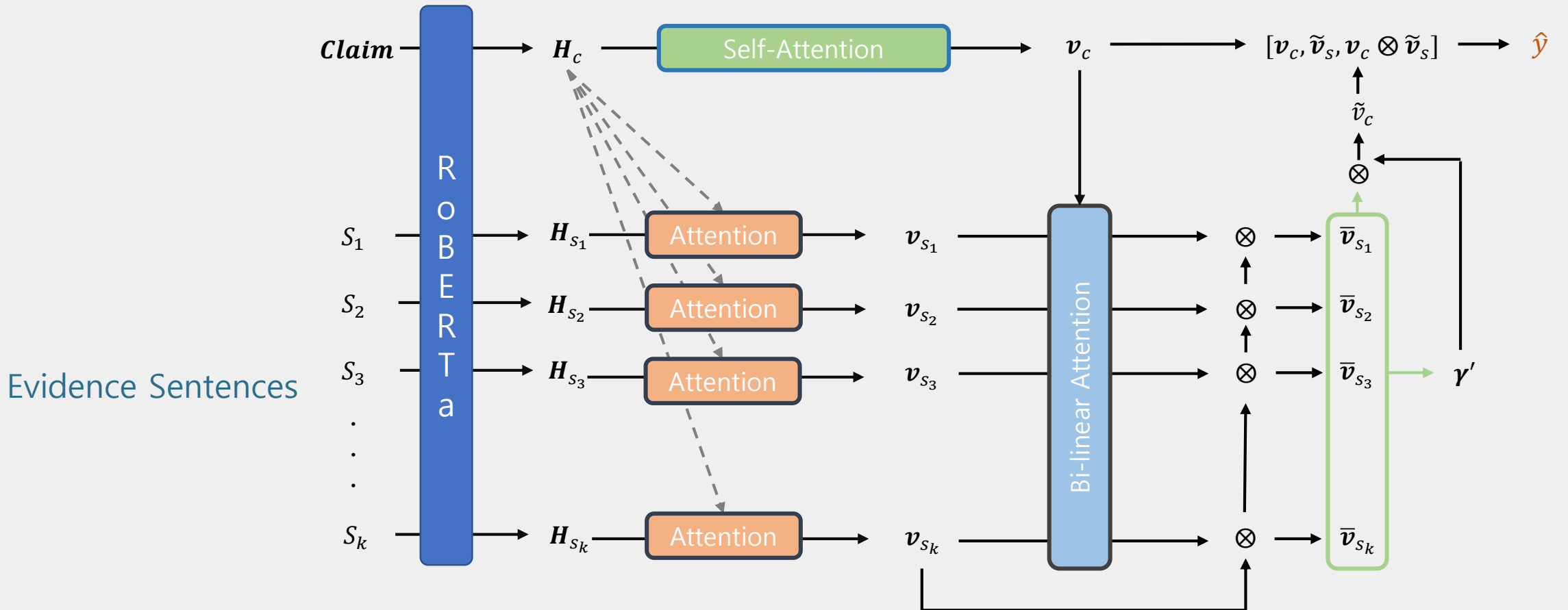


선택된 증거 문장들

3. 제안 모델

■ 사실 확인을 위한 뉴럴 모델

· $Claim = [< sos > Question < eos > < eos > Answer < eos >]$



3. 제안 모델

■ 사실 확인을 위한 뉴럴 모델

· RoBERTa 를 이용해 얻어진 claim 의 token embedding 을 $H_c \in \mathbb{R}^{l_c \times d_h}$ 라 할 때, claim 에 대한 representation v_c 는 다음과 같이 얻어냄

$$\mathbf{S} = \mathbf{H}_c \cdot \mathbf{H}_c^\top \in \mathbb{R}^{l_c \times l_c}$$

$$\boldsymbol{\alpha}_c = \text{softmax}(\max_{\text{row}}(\mathbf{S})) \in \mathbb{R}^{l_c}$$

$$\mathbf{v}_c = \mathbf{H}_c^\top \cdot \boldsymbol{\alpha}_c \in \mathbb{R}^{d_h}$$

· i 번째 evidence sentence 에 대한 representation v_{s_i} 는 bi-linear attention 을 이용해 다음과 같이 얻어냄

$$\mathbf{T} = \mathbf{H}_{s_i} \mathbf{W}_1 \mathbf{H}_c^\top \in \mathbb{R}^{l_{s_i} \times l_c}$$

$$\mathbf{v}_{s_i} = \mathbf{H}_{s_i}^\top \cdot \text{softmax}(\max_{\text{row}}(\mathbf{T})) \in \mathbb{R}^{d_h}$$

3. 제안 모델

■ 사실 확인을 위한 뉴럴 모델

· 전체 Evidence sentence 들에 대한 representation $v_s = [v_{s_1}, \dots, v_{s_k}] \in \mathbb{R}^{k \times d_h}$ 을 이용해 다음과 같이 evidence sentence 들에 대한 신뢰도를 계산하여 evidence sentence 들의 새로운 representation \bar{v}_s 를 얻어 낸다.

$$\beta = \text{softmax}(v_s W_2 v_c) \in \mathbb{R}^k$$

$$\bar{v}_s = \beta \otimes v_s \in \mathbb{R}^{k \times d_h}$$

· 모든 evidence sentences 들이 fact-checking 을 위해 필요한 문장은 아니므로 이들을 제한해야 할 필요가 있음. 따라서 다음과 같이 일부 evidence sentence 들의 정보를 제한함

$$\Gamma = W_4 \tanh(W_5 \cdot \bar{v}_s^T) \in \mathbb{R}^{h_a \times k}$$

$$\gamma' = \text{softmax}(\max_{col}(\Gamma)) \in \mathbb{R}^k$$

$$\tilde{v}_s = \sum_{i=1}^k \bar{v}_s \cdot \gamma'_i$$

3. 제안 모델

■ 사실 확인을 위한 뉴럴 모델

- 최종적으로 MLP layer 를 이용해 주어진 claim 이 사실인지 아닌지를 판단함

$$\hat{y} = MLP([v_s, \tilde{v}_s; v_c \otimes \tilde{v}_s])$$

4. 실험

■ 자동 구축한 사실 확인 데이터 셋

	Supported claim	Refuted claim
Train	8,057	6,608
Dev	1,602	1,135
Test (수동 구축)	53	47

- 자동으로 구축한 train, dev set 은 위키피디아 문서 10,000 개를 이용해 자동으로 구축함
- Evidence 검색을 위한 전체 검색 대상 위키 문서 셋은 총 437,299 개

4. 실험

■ 증거 검색 성능

- 사실 확인을 위한 증거 문서 검색 성능 (평가 셋)

Method	Hit@1	Hit@3	Hit@5	Hit@10	Hit@20
REALM	0.55	0.68	0.75	0.80	0.90

- 사실 확인을 위한 증거 문장 검색 성능 (평가 셋)

Method	Hit@1	Hit@3	Hit@5	Hit@10	Hit@20
TF-IDF / Top 10	0.56	0.68	0.72	0.76	-
TF-IDF / Top 20	0.62	0.73	0.76	0.85	0.86

- 증거 문서의 수를 각각 top 10 과 top 20 으로 세팅한 후 평가
- Evidence 검색 시에는 [Question; Answer] (pair)를 이용해 검색
- [Question; Answer] (pair)에서 Answer 가 변경된 Negative claim 의 경우는 Answer 가 변경되지 않은 Positive claim 에 비해 검색 성능이 낮은 결과를 보임

4. 실험

■ 실험 결과

	Precision	Recall	F1
Baseline	50.01%	88.31%	63.85%
제안 모델	72.47%	98.11%	83.36%

- 베이스라인 모델은 증거 없이 RoBERTa 모델에 질문과 정답 쌍 만을 입력으로 하여 얻어 낸 결과임

5. 결론

- 최근 활발히 연구되고 있는 사실 확인 작업을 수행하기 위해 Bart 모델을 이용하여 사실 확인 데이터 셋을 자동으로 구축하여 사실 확인 모델의 학습에 이용함
- 사실 확인을 위한 뉴럴 모델을 제안 함
- 결과적으로 83.36% 의 F1 score 를 달성 함
- 향후에는 사실 확인을 위한 데이터 셋 자동 구축 방식을 좀 더 정교하게 진행해 볼 예정임