언어 모델의 사실 지식을 이용한 한국어 사실 확인 모델

이종현[01], 나승훈[1], 신동욱[2], 김선훈[2], 강인호[2]

[1]전북대학교, [2]네이버

Contents

1. 서론

2. 제안 방법

3. 실험

4. 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

- 언어 모델(LM, Language Model)은 단어 시퀀스(sequence)의 확률 분포를 모델링하는 법을 학습하고, 당면한 언어의 구문론(syntax) 또는 의미론(semantic)의 다양한 측면에 관한 정보를 얻어냄

- 최근 언어 모델에 관한 연구들[1, 2, 3]은 현재까지 공개된 언어 모델들이 상당한 사실 지식(factual knowledge)를 보유하고 있다는 결과를 제시함

[1] Language models as knowledge bases? [F. Petroni et al., 2019]

[2] How can we know what language models know? [Z. Jiang et al., 2020]

[3] E-bert: Efficient-yet-effective entity embeddings for bert [N. Poerner., 2019]

1. 서론

■ 사실 검증 (Fact Verification)

- 증거 기반의 사실 확인 모델 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

- 증거 기반 사실 확인 모델은 주장을 뒷받침 할 만한 증거가 올바르게 검색되지 않았을 경우에는 사실 확인 여부 예측이 힘듦

- 본 연구에서는 이를 해결하기 위해 언어 모델이 보유한 상당한 양의 사실 지식을 이용하여 증거 기반 사실 확인 모델의 성능을 향상시키고자 함

- 최종적으로 기존 한국어 사실 확인 모델의 최고 성능인 65.93% F1 score 에서 0.41%p 향상된 66.34% F1 score 를 얻어 냄

2. 제안 방법

■ 사전 학습된 언어 모델을 이용한 사실 확인

- n 개의 토큰(token)으로 이루어진 문장에 대한 언어 모델의 조건부 확률 분포

$$p(x_k|x'_1, \ldots, x'_{k-1}, < mask >_k, x'_{x+1}, \ldots, x'_n)$$

- 언어 모델의 사실 지식 추론을 위해 개체명 인식(NER) 모델을 이용

흥부전은 조선 시대의 작자 미상의 고전 소설로 빈부 격차에 대한 비판 내용을 담고 있으며, 한국에서 널리 알려진 이야기 입니다.

Masking

흥부전은 [MASK]의 작자 미상의 고전 소설로 빈부 격차에 대한 비판 내용을 담고 있으며, 한국에서 널리 알려진 이야기 입니다.

Masked LM Prediction

흥부전은 조선 시대의 작자 미상의 고전 소설로 빈부 격차에 대한 비판 내용을 담고 있으며, 한국에서 널리 알려진 이야기 입니다.

2. 제안 방법

■ 사전 학습된 언어 모델을 이용한 사실 확인

- Multi-token (개체명이 반드시 하나의 토큰으로 이루어진 것은 아님)

$$s_{i:j} = x_1, \dots, x_{i-1}, < mask >_i, \dots, < maxk >_j, x_{j+1}, \dots, x_n$$

$s_{i:j}$ 는 i 번 째 부터 j 번 째 단어가 마스크 토큰으로 대체된 문장을 의미

- Multi-token decoding 방법을 이용하여 연속된 마스크 토큰에서 원래의 토큰을 예측

1. 가장 확률이 높은 마스크 토큰에 대한 예측 값을 먼저 선택하고 이를 기반으로 다시 문장에 대한 확률 분포를 계산

2. 가장 확률이 높은 예측 값을 선택하는 과정을 반복하여 전체 마스크 토큰들에 대한 초기 예측 $\hat{s}_{i:j} = x_1, \dots, \hat{y}_i, \dots, \hat{y}_j, \dots, x_n$ 을 얻어냄

$$\hat{y}_k = \underset{i \le k \le j, y_k}{\operatorname{argmax}} p\left(y_k | s_{i:j}\right)$$

$$c_k = p(\hat{y}_k | s_{i:j})$$

2. 제안 방법

■ 사전 학습된 언어 모델을 이용한 사실 확인

- 정제(refinement) 단계

$\backslash k$ 는 k 번 째 토큰이 마스크 토큰인 것을 의미

$$\hat{y}_k = \underset{y_k}{\operatorname{argmax}}\, p\left(y_k \middle| \hat{s}_{i:j} \backslash k\right)$$

$$c_k = p\left(\hat{y}_k \middle| \hat{s}_{i:j} \backslash k\right), \qquad k = \underset{i \le k \le j}{\operatorname{argmin}}\, c_k$$

- 가장 낮은 확률값을 가지는 토큰을 선택하여 이를 다시 예측하는 과정을 거침
- 재 예측된 토큰이 이전 예측 값과 같을 때까지 이를 최대 N 번 반복

- 길이 정규화

$$v(j-i+1) = \frac{1}{j-i+1} \sum_{k=i}^{j} \log c_k$$

- c_k 는 k 번 째 토큰에 대한 확률을 의미하고 v_m 은 m 개의 마스크 토큰에 대한 전체 신뢰도를 의미함

2. 제안 방법

■ 증거 기반 사실 확인 모델과 언어 모델 기반 사실 확인 모델의 결합

- 언어 모델 기반 사실 확인의 최종 출력

$$o = \begin{cases} SUPPORETD, & \text{if } \exp\big(v(j-i+1)\big) \geq p \\ REFUTED, & \text{otherwise} \end{cases}$$

- 증거 기반 사실 확인 모델과의 결합

  • 언어 모델 기반 사실 확인 모델의 출력인 o 가 REFUTED 인 경우에만 증거 기반 사실 확인 모델을 이용해 재 예측

  • 즉, 언어 모델 기반 사실 확인 모델의 결과로 SUPPORTED 가 얻어진 주장은 전체 모델의 최종 예측 결과를 SUPPORTED 로 고정함

3. 실험

■ 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[3] 을 이용함
-
- 검색된 증거 문서 중 적절한 증거 문장 선택 과정에는 TF-IDF 를 이용함

[3] Realm: Retrieval-augmented language model pre-training, [K. Guu et al., 20]

3. 실험

■ 실험 및 데이터 세팅

- 한국어 사실 확인 작업을 위한 데이터 셋은 [4]에서 제안한 데이터 셋을 그대로 사용

	학습 데이터	개발 데이터	평가 데이터
데이터 수	17,979	1,400	213

- 하이퍼파라미터

  - 언어 모델 기반 사실 확인에는 한국어 RoBERTa 를 이용
  - $N = 10, \mathrm{p} = 0.75$
  - 증거 기반 한국어 사실 확인 모델로는 [4]에서 65.94% F1 score 를 달성한 KGAT 모델을 이용

[4] Hyperclover를 이용한 한국어 fact 검증을 위한 자동 데이터 생성, [이종현, 나승훈, 신동욱, 김선훈, 강인호., 20]

3. 실험

■ 실험 결과

- 기존의 증거 기반 사실 확인 모델에 언어 모델 기반 사실 확인 모델을 결합하여 0.41%p 개선된 결과를 얻어냄

표 1: 최종 실험 결과

모델	Precision	Recall	F1 score
KGAT	82.19%	55.04%	65.93%
KGAT + LM	69.69%	63.30%	**66.34%**

4. 결론

- 증거 기반 사실 확인 모델에 비해 언어 모델 기반 사실 확인 모델의 단일 성능은 상당히 낮게 측정되어 단일 모델로써의 활용에 한계가 있음

- 따라서, 본 연구에서는 언어 모델이 보유한 상당한 양의 사실 지식을 토대로 기존 한국어 자동 사실 확인 모델의 성능을 향상시키고자 함

감사합니다