증거와 Claim의 LM Perplexity를 이용한 Zero-Shot 사실 검증

박은환, 나승훈, 신동욱, 전동현, 강인호
전북대학교, 네이버

- References

[1]. FEVER: A Large-Scale Dataset for Fact Extraction and VERification [Thorne et al, 18’]

[2]. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification [Hanselowski et al, 18’]

[3]. GEAR: Graph-Based Evidence Aggregating and Reasoning for Fact Verification [Zhou et al, 19’]

[4]. A DQN-Based Approach to Finding Precise Evidences for Fact Verification [Wan et al, 21’]

[5]. Fine-Grained Fact Verification with Kernel Graph Attention Network [Liu et al, 21’]

[6]. A Multi-Level Attention Model for Evidence-Based Fact Checking [Kruengkrai et al, 21’]

[7]. 그래프 기반 증거 추론을 이용한 질의응답에 대한 사실 여부 검증 연구 [박은환 et al, 21’]

[8]. 한국어 Fact 검증을 위한 자동 Claim 데이터 생성 [이종현 et al, 21’]

[9]. Language Model as Fact Checkers? [Lee et al, 20’]

[10]. Toward Few-Shot Fact Checking via Perplexity [Lee et al, 21’]

[11]. Graph Attention Networks [Velickovic et al, 18’]

[12]. Language Models as Knowledge Bases? [Petroni et al, 19’]

[13]. RoBERTa를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱 [민진우 et al, 19’]

- 연구 개요
  - 최근 사실 검증(Fact Verification) 연구는 지속적으로 활성화되고 있음.
    - 영어권에서는 [1]을 데이터 집합으로 사용 중임.
  - 한국어의 경우:
    - [7], [8]은 BART 언어 모델을 이용하여 Claim 을 생성하는 방법을 제안, **하지만 생성 모델의 특성상 많은 노이즈가 존재함**.
      또한, 사람 평가(Human Evaluation)이 되지 않았다는 점에서 **퀄리티의 한계가 존재함.**

- 연구 개요

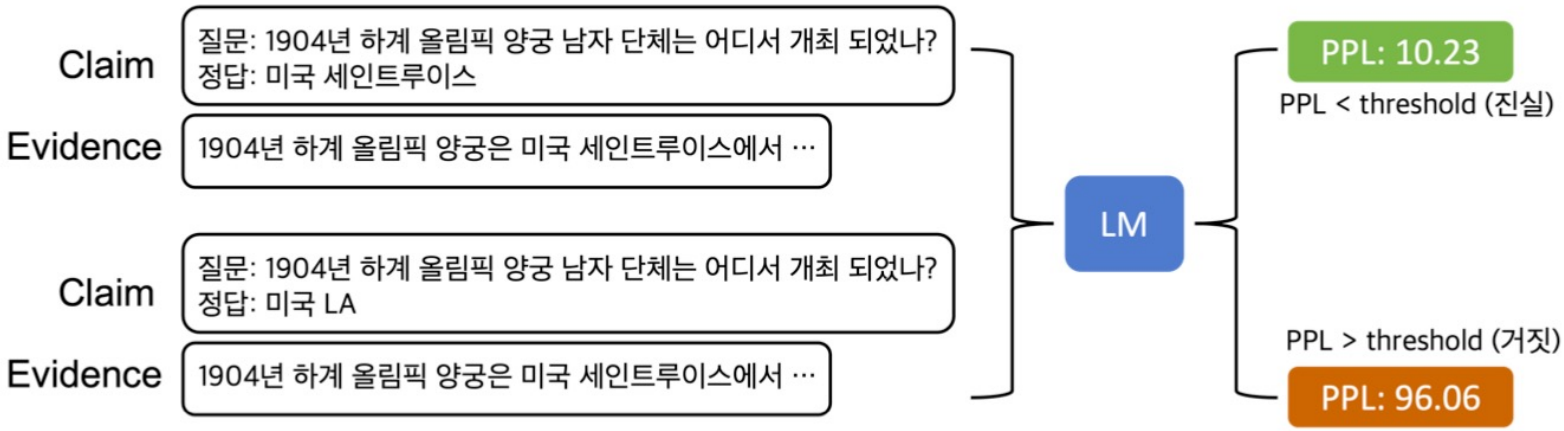

그림 1. [10]을 질의 응답에 대한 사실 검증 연구에 적용했을 때의 모습이다. 질문과 정답을 하나의 Claim으로 설정하고 일정 임계점 (Threshold)보다 PPL이 낮다면 진실이고 높다면 거짓임을 의미한다.

  - [10]은 Perplexity (PPL)을 이용한 Few-Shot 사실 검증 연구를 제안함.
    - PPL을 확장한 Evidence-Conditioned PPL을 연구에 사용함.
  - 본 연구에서는 **Evidence-Conditioned PPL을 확장한 PPL을 이용하여 언어 모델의 지식 (Knowledge)만을 이용한 Zero-Shot 사실 검증 연구를 제안**함.

- 제안 방법
  - 증거 기반 PPL [10]

$$PPL_E(X) \sqrt[C]{\prod_{i=1}^{C} \frac{1}{p\left(x_{c_i} | x_{e_0}, \ldots, x_{e_E}, \ldots, x_{c_{i-1}}\right)}}$$

  - Claim 기반 PPL

$$PPL_C(X) \sqrt[E]{\prod_{i=1}^{E} \frac{1}{p\left(x_{e_i} | x_{c_0}, \ldots, x_{c_C}, \ldots, x_{e_{i-1}}\right)}}$$

  - 증거 및 Claim 기반 PPL

$$PPL_{Mixture}(X) = PPL_E(X) \cdot PPL_C(X)^{-\beta}$$

  - 증거 기반 Claim 기반 PPL 뿐만이 아닌 Claim 기반 증거 PPL을 결합하는 방향으로 기존 [10]을 확장함.

- 실험 결과

표 1. 평가 데이터 집합에 대한 실험 결과

Model	Method	Precision	Recall	F1
LM(Q+A)[7]	Fine-tune	55.97	56.00	54.85
SKT-KoGPT2	$PPL(X)$	53.64	51.00	43.76
SKT-KoGPT2	$PPL_E(X)$	57.01	57.00	56.72
SKT-KoGPT2	$PPL_C(X)$	58.21	58.00	57.89
SKT-KoGPT2	$PPL_{Mixture}(X)$	**60.04**	**60.00**	**60.00**

  - 평가 데이터 집합은 100개로 한국어 위키피디아 백과를 바탕으로 수동으로 구축함.
  - RoBERTa[13](LM(Q+A)으로 미세 조정한 것보다 SKT-KoGPT2로 Zero-Shot 실험을 했을 때 더 좋은 성능을 보여줌.

- 결론
  - [10]을 확장한 PPL을 바탕으로 Zero-Shot 사실 검증 연구를 진행
  - 추후 연구
    - 100 건의 적은 데이터를 보충
    - RoBERTa 등, 다른 언어 모델로 Zero-Shot 사실 검증 실험

감사합니다