

미디어 현장

이준한
서울대
언론정보학과 교수

인공지능 기반 팩트체크, 어디까지 왔나

월등히 빠른 검증 가능하나 신뢰성 있는 레퍼런스 확보가 관건

팩트체크의 중요성은 커지고 있지만, 다양한 플랫폼에서 실시간으로 생성되는 대량 정보를 사람이 일일이 검증하는 데는 한계가 있다. 이에 따라 인공지능을 이용한 팩트체크가 주목받고 있다. 국내에서 이제 막 시작된 한글 대상 인공지능 팩트체크 연구에 대해 알아본다. 편집자 주

가짜뉴스가 거대한 사회문제로 부상했다. 인터넷, SNS와 같이 정보를 유통할 수 있는 채널이 다양해지고 이를 통해 정보가 순식간에 생성되고 확산하면서 무분별한 정보에 대한 팩트체크 필요성이 크게 대두됐다. 많은 언론사도 팩트체크 팀을 설치해 팩트체크 뉴스를 보도에 적극 활용하고 있다.

하지만 ‘가짜뉴스’라는 용어의 정의에는 조금 더 주의를 기울여야 한다. 단순히 정보가 틀렸다고 해서 가짜뉴스라고 규정하기는 어렵다. 칼로바와 피셔(Karlova & Fisher)의 연구¹⁾에 따르면, 가짜뉴스의 큰 특성 중 하나는 ‘기만성’이다. 이것은 의도적으로 다른 사람을 속이려는 속성을 의미한다.

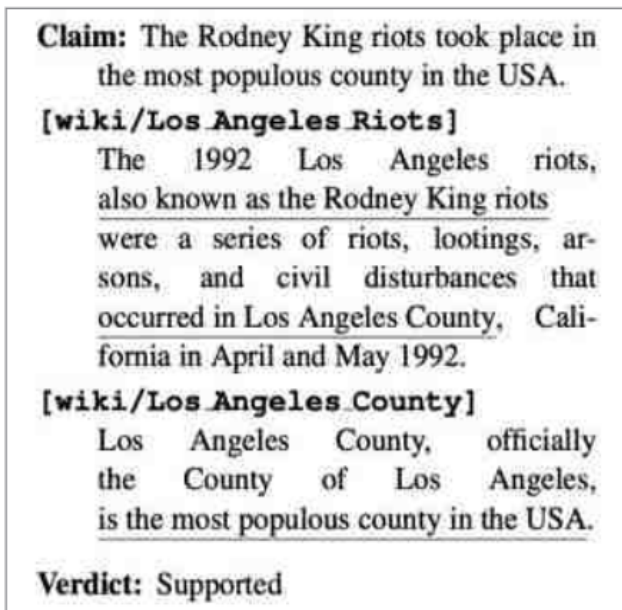
팩트체크의 역할은 이러한 잘못된 정보, 특히 기만성을 가진 정보를 찾아내는 것이다. 그래서 여러 언론사의 팩트체크 전문 기자들은 가짜 정보들을 정확하게 걸러내는 데 중추적인 역할을 하고 있다. 그러나 사람의 눈으로 직접 확인하는 팩트체크는 뛰어난 품질을 제공하지만, SNS와 같은 플랫폼에서 실시간으로 생성되는 대량의 정보를 모두 검증하기에는 시간과 비용의 한계가 있다.

이런 문제에 대한 해결책의 일환으로 최근 인공지

능 기반의 팩트체크 연구가 활발히 진행되고 있다. 자연어 처리(Natural Language Processing), 머신러닝(Machine Learning)을 넘어 인공지능(AI)을 활용한 연구가 수행 중이다. 이 중 Fever.ai²⁾는 대표적인 예로, ACL(Association for Computational Linguistics)이라는 컴퓨터 언어학 분야의 유명한 학회에서 시작한 프로젝트다. 이 프로젝트는 위키피디아를 활용해 만든 학습용 데이터 18만 건을 제공하며, 이 데이터를 활용한 연구자들은 각자의 알고리즘을 개발해 팩트체크 챌린지를 벌이고 있다. 아직은 가짜뉴스가 가진 기만성까지 파악하는 기술 개발은 어렵지만, 신뢰성 있는 소스를 통해 제시된 주장이 얼마나 지지받는지 파악하는 것은 어느 정도 가능하다. [그림 1]

1) Karlova, N. A. & Fisher, K. E., <A social diffusion model of misinformation and disinformation for understanding human information behaviour>, Information Research, 18(1), pp.1-17, 2013

2) ACL(Association for Computational Linguistics)에서 주관하는 팩트 확인(Fact Extraction and Verification) 워크숍으로 18만여 건의 팩트체크 학습용 데이터를 제공하며 이를 활용한 다양한 연구가 공개된다. <https://fever.ai/>



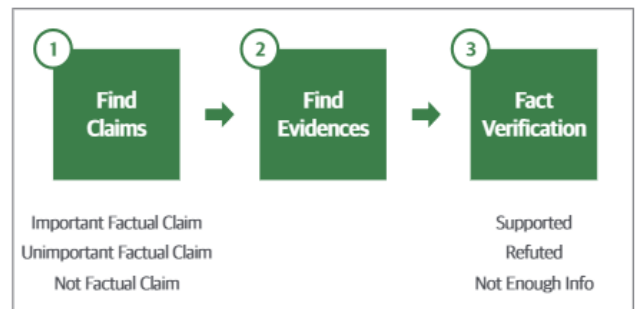
[그림 1] Fever.ai를 활용한 팩트체크 사례. 주장(claim)이 주어지면 위키 피디아 문서에서 근거를 찾아 주장의 참/거짓 여부를 판단한다. <출처-
https://aclanthology.org/N18-1074.pdf>

한글을 대상으로 한 인공지능 팩트체크 연구는 이제 막 시작됐다. 연구의 가장 큰 걸림돌은 역시 학습용 데이터의 부재에 있다. 이 문제를 해결하기 위해 이준환·김진희 서울대 교수 연구팀은 AI 팩트체크 연구에 활용 가능한 학습용 데이터 10만 건을 제작·공개하고 한글 기반 인공지능 AI 팩트체크의 방향을 제시하는 인공지능 모델을 소개했다. 이 글에서는 이 연구의 결과를 간략하게 소개하고자 한다.

● 사람과 유사한 AI 팩트체크의 과정

AI를 활용한 팩트체크는 사람이 수행하는 팩트체크와 여러모로 유사한 과정을 거친다. [그림 2] 이 과정은 정보의 정확성을 확보하고 오류나 편향을 방지하는 데 중요하다.

먼저, 팩트체크의 첫 단계는 주장이 무엇인지 파악하는 것이다. 일반적으로 문장은 다양한 정보를 포함하고 있어 알고리즘이 검증 가능한 정보를 가진 ‘주장’이 무엇인지 식별하는 것이 중요하다. 예를 들



[그림 2] AI 팩트체크의 세 가지 단계. 첫 단계에서는 주장을 파악하고 두 번째 단계에서는 근거를 찾는다. 세 번째 팩트 검증(Fact Verification) 단계에서 주장과 근거를 바탕으로 판단이 이루어지는데, 인공지능 알고리즘이 이 단계에 활용된다.

어 ‘어제 대전에서 올라왔어요’와 같은 주장은 특별한 검증이 필요 없는 반면, ‘4년 전과 비교해서 최근의 실업률이 훨씬 더 나쁘다’는 주장은 검증이 가능하다. 따라서 두 번째와 같은 문장 선별을 첫 단계에서 수행해야 한다.

주장이 정확히 정의되면, 두 번째 단계는 이 주장을 검증하기 위한 적절한 근거를 찾는다. 언론사의 팩트체크에서도 근거 제시는 팩트체크 뉴스의 핵심 요소 중 하나다. AI 팩트체크에서도 마찬가지로 근거를 찾아야 하는데 보통은 신뢰도가 높은 인터넷 정보 소스를 레퍼런스 데이터베이스로 활용한다. 이 과정은 마치 구글 검색을 자동화하는 것과 비슷하다. 알고리즘은 먼저 문서를 찾고(Document Retrieval), 찾은 문서 중 주장과 가장 유사한 문장을 찾는다(Sentence Selection). 서울대에서 개발한 알고리즘은 근거 문장으로 가장 유사한 다섯 개의 문장을 선택한 후 판단의 과정으로 넘어간다.

해외 연구에서는 저작권 등으로부터 자유로운 위키피디아를 정보의 소스로 활용하고 있는데, 서울대에서 수행된 연구에서도 같은 이유로 한글 위키피디아를 정보 소스로 선택했다. 다만, 한글 위키피디아와 영문 위키피디아는 신뢰도 측면에서 큰 차이가 있어 향후 연구에서 좀 더 고민이 필요한 부분이다.

근거를 찾은 후, 세 번째 단계에서는 그 근거가 주장을 어떻게 지지하거나 반박하는지 파악하는 과정

이 이어진다. 서울대 연구팀은 이를 위해 RTE(Recognizing Textual Entailment)라는 기술을 활용했다. 이 기술은 자연어 처리 과정에서의 다양한 임베딩 모델을 활용해 두 개의 텍스트 문장(주로 주장과 가설)이 주어졌을 때, 한 문장이 다른 문장의 논리적 추론 또는 함축을 의미하는지 여부를 판단하는 태스크다. 이를 바탕으로 근거가 주장을 지지(supported)하거나 반박(reputed)하는지, 아니면 정보가 부족하여 결론을 내릴 수 없는지(not enough information)를 판단한다.

(RTE 예시)

주장: “사자는 고양이과 동물입니다.”

가설: “사자는 동물입니다.”

주장이 가설을 함축하므로 entailment(논리적 함의) 관계가 성립

최근에는 RTE 대신 챗GPT와 같은 거대언어모델(LLM, Large Language Model)을 활용하는 방법 또한 모색되고 있다.

● 사람보다 빠른 AI 팩트체크, 의미와 맥락은?

서울대 연구팀은 직접 생성한 10만 건의 학습 데이터를 바탕으로 인공지능 팩트체크를 위한 베이스라인 모델을 개발했다. 초기 모델은 0.59의 F1 스코어를 보여주었는데 이는 실제 서비스에 활용하기는 아직 다소 부족한 성능이다. 그러나 이후 지속된 모델 개선을 통해 조금씩 성능이 개선되고 있다. 다음은 서울대가 제작한 베이스라인 모델을 활용한 실제 팩트체크 사례다.

[그림 3]에서 입력된 ‘임진왜란에서 이억기가 사망한 전투는 칠천량 해전이다’라는 주장에 대해 팩트



[그림 3] 인공지능 팩트체크 사례 1



[그림 4] 인공지능 팩트체크 사례 2

체크 모델은 63%의 확률로 ‘참’이라 판단했다. 이 주장을 검증하기 위해 알고리즘은 ‘이억기’와 ‘칠천량 해전’이라는 두 개의 문서에서 찾은 5개의 문장을 근거로 판단을 내렸다.

[그림 4]는 ‘2021년 서귀포에서는 지진이 발생하지 않았다’는 문장을 검증한 사례다. ‘2021년 서귀포 해역 지진’이라는 문서에서 추출한 5개의 문장을 근거로 알고리즘은 주장을 거짓으로 판단했다.

사례가 보여주듯이 이 시스템은 인공지능 기반의 최신 자연어 처리 기법을 통해 팩트체크를 자동화한다. 본 연구에서는 한국어 위키피디아를 검증의 대상으로 사용했지만, 백과사전이나 실시간으로 업데이트되는 뉴스 데이터 등을 활용한 팩트체크 검증 또한 가능하다.

이렇게 자동화된 팩트체크 시스템의 주요 장점 중 하나는 신속성이다. 사람이 수행하는 팩트체크

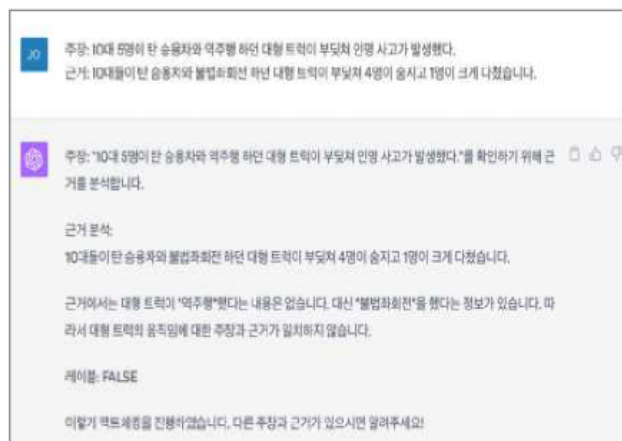
에 비해 월등히 빠른 검증 시간을 보여준다. 이로 인해 정확도가 좀 더 확보된다면 웹상에서 퍼져나가는 잘못된 정보를 빠르게 발견하고 확인하는 서비스 개발이 가능하다. 특히 SNS나 다양한 웹사이트에서 대량으로 발생하는 댓글 등의 정보를 실시간으로 처리하거나 정치 토론에서 정치인의 발언을 바로 검증하는 것도 가능할 것이다.

그러나 이 시스템은 주장의 의미와 맥락을 파악하는 팩트체크이라기 보다는, 기존에 저장된 데이터를 통해 주장의 진위를 검증하는 팩트 검증에 가깝다. 주장을 검증하기 위해서는 정해진 소스 데이터베이스 내에서 검색을 수행한다. 따라서 새로운 사실의 검증은 가능하지 않다. 주장의 근거가 되는 문장이 검증의 대상이 되는 레퍼런스 데이터베이스에 존재하지 않기 때문이다. 이러한 이유로 팩트 검증에 활용되는 레퍼런스 데이터베이스는 매우 중요한 역할을 한다. 현재 한국어 위키피디아는 연구의 편의성 측면에서는 좋은 자료로 간주되지만, 정보의 정확도와 신뢰성 측면에서는 의문을 가지는 연구자들이 많이 있다. 보다 신뢰성 있는 레퍼런스 데이터베이스의 확보가 필요한 이유다.

● LLM, AI 팩트체크 진보의 열쇠

검증의 정확도를 높이는 것은 인공지능 팩트체크의 핵심 과제 중 하나다. 이를 위한 방법은 크게 두 가지로 요약할 수 있다. 첫째, 보다 양질의 데이터 확보와 둘째, 고도화된 합의 분석 언어모델과 알고리즘의 적용이다. 최근 인공지능 기반의 자연어 처리 기술은 놀라운 속도로 발전하고 있어, 새롭게 등장하는 언어모델과 알고리즘의 적용은 팩트체크 시스템의 성능 향상에 결정적인 역할을 할 것으로 예상된다.

특히 챗GPT로 대표되는 LLM 기술의 등장이 주목받고 있다. 이미 다수의 연구에서는 GPT, 라마



[그림 5] 챗GPT를 활용한 팩트체크 사례 <출처 - 챗GPT 사용 화면 갈무리>

(LLaMA), 바드(Bard) 등의 언어모델을 성공적으로 활용하며 연구 성과를 내고 있다. 이처럼 진보된 언어모델들은 팩트체크의 정확도를 한 단계 끌어올릴 수 있는 중요한 열쇠가 될 것이다.

앞서 언급한 RTE는 다양한 임베딩 모델, 예를 들면 워드투벡터(Word2Vec), 글로브(GloVe), 트랜스포머(Transformer) 등을 통합해 성과를 내고 있다. 그리고 이제 GPT-4 아키텍처를 포함한 LLM이 자연어 처리 분야에 활발하게 도입되고 있다. GPT-4와 같은 언어모델은 기존의 언어모델과는 비교할 수 없는 대규모의 텍스트 데이터로 학습되어 문장 간의 논리적 관계나 함축을 파악하는 능력이 뛰어나다. 챗GPT를 활용하는 연구는 현재 진행형이지만 다음의 사례를 통해 어느 정도 그 가능성을 확인할 수 있다. [그림 5] ‘프롬프트 엔지니어링’이라고 부르는 태스크를 통해 챗GPT에게 주장과 근거의 관계, 학습용 데이터를 분석하는 방법을 입력한 후 수행한 팩트체크 태스크에서 챗GPT는 근거가 제시한 잘못된 사례를 정확히 판단해 냈다. 앞으로 이러한 LLM의 활용이 인공지능 팩트체크의 가능성을 보다 높일 것으로 예상된다. 📌