효율적인 트랜스포머에 기반한 설명 가능한 팩트체크 모델

윤희승[1] · 정재은[1,*] · 이건주[1] · 정다희[1] · 김건오[2]

[1]중앙대학교 · [2]트윈워드

Explainable Fact Checking Model Based on Efficient Transformer

Heeseung Yun[1] · Jason J. Jung[1*] · Gunju Lee[1] · Dahee Jung[1] · Kono Kim[2]

[1]Chung Ang University · [2]Twinword

E-mail : {yunihg, j3ung, hanyd1012}@cau.ac.kr, wjgmu.d@gmail.com, Kono@twinworld.com

요 약

본 논문에서는 어텐션 메커니즘에 기반하여 정보 판단에 대한 근거를 제공하는, 이른바 설명 가능한 팩트체크 모델을 제안할 것이다. 최근 미디어의 발달에 따라 각종 뉴스가 쏟아지고 있는 바, 이와 더불어 뉴스에 대한 진위 여부 판단, 즉 팩트체크가 주목받고 있는 상황이다. 하지만 현재 팩트체크는 언론인이나 시민 단체 일원들의 검색 능력에 의존하고 있어서, 이를 자동적으로 하는 모델에 대한 연구가 진행되고 있다. 이에 본 논문에서 설명 가능한 자동 팩트체크 모델을 제안하고자 한다.

ABSTRACT

In this paper, we introduce the model so-called Explainable Fact-Checking model based on attention mechanism which shows both the result of fact check of the news and the evidence of verdict. Recently, several news surge on media, so fact check attracts much attentions. However, in present fact check relies on the search made by journalists and members of fact check orgainzation, so there is some researchs about automated fact checking. Therefore in this paper we propose explainable automated fact checking model.

키워드

Attention Mechanism, Transformer, Explainable Fact Checking, Hashing Function

Ⅰ. 서 론

최근 페이스북, 트위터 등 소셜 미디어를 이용하는 사람이 늘면서, 뉴스까지 소셜 미디어를 통해 소비하고 공유하는 사례가 늘고 있다. 그러나 이러한 경향과 더불어 소셜 미디어 내에서 가짜 뉴스가 확산하는 사례도 늘고 있다. 이러한 경향은 2017년 미래창조과학부에서 인공지능 R&D 챌린지 과제로 가짜 뉴스 찾기를 선정한 것에서 짐작할 수 있다[1].

이에 대중 사이에서 인터넷이나 TV, 그리고 소셜 미디어에서 나오는 뉴스의 신뢰도에 의문을 제기하는 경우가 많아졌다. 이에 부응하기 위해 일부 언론 및 단체들을 주축으로 해서 뉴스의 진위 여부를 판단하는, 이른바 팩트체크가 진행되고 있다. 이를 통해 뉴스에 대해서 보다 객관적인 판단이 가능해졌으나, 문제는 작업 속도였다. 기자나 단체 회원 등 인력의 검색 능력에 의존하면서 진행되는 지라, 실시간 단위의 작업이 어렵기 때문이다. 따라서 이를 자동화하는 모델들이 학계에서 연구되고 있다.

그러나 기존 모델들의 공통적인 문제점은 진위 판단에 대한 근거를 제시하지 못한다는 데 있다. 뉴스 데이터를 분류기가 학습하면 팩트 여부만 보여준다는 것이다. 이에 본 논문에서는 이러한 점을 보완한, 팩트체크에 대한 근거를 보여주는 모델을 제시하고자 한다.

* corresponding author

II. 관련 연구

팩트체크를 자동화시키게 도와주는 모델에 대한 연구는 주로 딥러닝이나 휴리스틱 모델 위주로 이뤄지고 있다[2]. 주로 쓰이는 모델은 LSTM과 같은 순환형 신경망 모델이다. 문장의 단어 배열 순서가 중요한 자연어처리 특성상 이를 학습 과정에서 반영할 수 있는 순환형 모델을 쓰게 되는 것이다. 그리고 딥러닝을 쓰지 않은 모델은 주로 문서 검색 등을 써서 얻은 특징을 토대로 가짜뉴스를 가려내는, 이른바 휴리스틱 기법을 이용하기도 한다[2].

이러한 모형의 핵심은 가짜 뉴스로 알려진 뉴스의 특징을 추출해서, 그것을 토대로 이를 분류하는 모델을 학습한다는 것이다. 또한, 단어 배열 순서가 학습에서 중요한 만큼, 문장을 시계열 데이터로 간주해서 학습하는 순환형 신경망을 쓴다는 것도 핵심이 될 수 있다. 이러한 방식을 통해 가짜 뉴스와 팩트를 가려내는 것이다.

하지만, 이러한 모델은 앞에서 언급한 바와 같이 팩트 여부를 판단하는 데는 도움을 주는 면이 있지만, 그 판단에 대한 근거를 제공하는 데에는 역부족이다. 이에 판단에 대한 근거까지 보여주는 모델이 연구되고 있는 것이다. 이러한 연구는 주로 어텐션 함수, BERT 등과 같은 트랜스포머 기반의 모델로 이루어져 있다. 문장의 단어 간의 관련성을 계산하는 어텐션 함수 특성상, 뉴스와 그것의 팩트 여부의 근거가 되는 정보 간의 관련성을 연결시킬 수 있다는 점 때문이다[3].

또한, 지식그래프 기반 모델도 연구되고 있는데, 이것은 팩트체크의 판단 근거가 되는 정보와 뉴스를 나이브 베이즈나 랜덤 포레스트와 같은 모델로 그래프로 연결짓는 방식이 개발되고 있다[4].

III. 트랜스포머

이 장에서는 앞으로 제시할 모델의 기반이 되는 트랜스포머[5]를 소개할 것이다. 트랜스포머란, 어텐션 매커니즘을 기반으로 한, 인코더-디코더 기반의 기계번역 모델을 말한다. 어텐션 메커니즘에 대해서 간략하게 설명을 하자면, 어텐션 메커니즘이란 소프트맥스 함수를 변형하여 만든 어텐션 함수를 통해 두 단어 간의 관련성을 수치로 나타내는 것을 말한다. 이것을 이용하여 모델이 단어의 문맥을 보다 잘 이해하게 만드는 것이다. 어텐션 함수를 수식으로 표현하면 아래와 같다.

$$A(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \quad (1)$$

여기서 Q는 쿼리(Query) 벡터, K는 키(Key) 벡터, V는 밸류(Value) 벡터를 말하고, d는 임베딩 차원을 의미한다. 즉, 어텐션 함수는 Q와 K 간의 관련성을 구한 후, $\sqrt{d}$를 나누는 정규화 과정을 거쳐 V와 곱하는 형식으로 값을 계산한다.

이러한 트랜스포머 모델은 기존의 여러 시퀀스투 시퀀스 모델의 성능을 압도하면서 자연어처리의 주류로 자리 잡았고, 이것에 기반한 BERT와 GPT 모델이 탄생하게 되었다. 그러나 이것에도 문제점이 있었으니 바로 계산 복잡도가 높다는 것이다. 따라서 이러한 문제를 해결하기 위한 연구도 진행되고 있다[6][7].

IV. 제안 모델

본 논문에서는 팩트 체크에 대한 근거를 제시하고, 더 나아가 실제 웹사이트에 구현할 것에 대비하여 보다 계산 복잡도를 줄인 트랜스포머 모델에 기반한 설명 가능한 팩트체크 모델을 제시하고자 한다. 간략한 모델 구조는 그림 1과 같다.

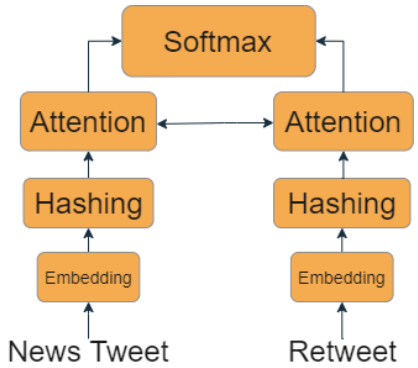

그림 1. 제안 모델 구조도

여기에서 핵심은 해싱 레이어(Hashing layer)이다. 해싱 레이어에서는 의미가 유사한 단어끼리 같은 해시 값으로 묶음으로써 어텐션 함수 계산에 필요한 데이터 양을 줄인다. 같은 해시 값을 가진 단어끼리 묶어서 어텐션 값을 계산하기 때문이다.

이후 어텐션 함수에서는 뉴스 트윗과 그것의 리트윗을 서로 참조하면서 값을 계산한다. 이를 통해 뉴스와 관련 있는 리트윗의 단어에 대한 관련성을 계산할 수 있는데, 이것을 이용해 뉴스에 대한 가치 판단의 근거를 보여줄 수 있다는 것이다.

V. 결론 및 향후 계획

본 논문에서는 트랜스포머 모델에 대한 간략한 소개 및 이것을 이용한 설명 가능한 팩트체크 모델에 대해서 제안했다.

이러한 모델을 기반으로 하여 한국어 트위터 데이터를 수집한 후, 팩트체크 모델을 학습시킨 후 이를 이용해서 실시간으로 팩트체크를 하는 웹사이트를 구현할 계획이다. 비록 한국어 전처리 문제라든지 데이터 수집 문제가 있으나, 연구하면서 효율적인 전처리 방식을 고민할 것이고 데이터의 경우 트위터에서 제공하는 학술용 API를 이용해서 수집할 것이다.

이번 연구를 통해 한국어 데이터에 기반한 팩트체크 모델을 구축할 뿐만 아니라, 이를 기반으로 한 팩트체크 자동화 사이트를 운영함으로써, 보다 신뢰성 있는 언론을 이용자들이 선별해서 볼 수 있을 것을 기대할 수 있다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2021년도 SW중심대학지원사업의 연구결과로 수행되었음 (20170001000051001).

References

[1] S. S. Park, K. C. Lee, "A Comparative Study of Text analysis and Network embedding Methods for Effective Fake News Detection" 『Journal of Digital Convergence』, Vol. 17, No. 5, pp.137-143, 2019

[2] Kotonya N, Toni F, "Explainable Automated Fact-Checking: A Survey", *Proceedings of the 28th International Conference on Computational Linguistics,* Barcelona (online), pp.5430-5443, 2020.

[3] Naeemul H, Fatma A, Chengkai L, Mark T. "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster". *In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, halifax NS Canada, 2017.

[4] Lianwei W, Yuan R, Yongqiang Z, Hao L, Ambreen N, "DTCA: Decision tree-based co-attention networks for explainable claim verification". *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, online, pp.1024–1035, 2020.

[5] Vaswani A, Noam S, Niki P, Jakob U, Llion J, Aidan N. Gomez, Lukasz K, Illia P. "Attention is all you need". *In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp.6000–6010, 2017.

[6] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. "Efficient transformers: A survey". *arXiv preprint arXiv:2009.06732*, 2020.

[7] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. "Reformer: The efficient transformer". *In International Conference on Learning Representations*, online, 2020.