송상헌의 '인공지능 언어모델의 상식적 판단' 연구에 대한 토론문

조희련(중앙대학교 인문콘텐츠연구소)

먼저 발표문을 정말 재미있게 읽었다는 말씀을 드립니다. 발표문의 맺음말에 도달했을 때 아쉬운 기분마저 들었습니다. 발표문을 읽는 동안 즐거운 시간을 선사해 주셔서 감사합니다.

1. 토론자에 의한 발표문 요약

발표자는 인공지능 언어모델이 상식적 판단을 내리기 위해서는 가추 추론(abductive reasoning)이 필요하다고 설명하고, 이와 관련된 연구 동향을 소개하고 있습니다.

먼저 발표자는 '상식' 또는 '상식적'이란 말이 어떤 의미인지를 세 종류의 언어자원('우리말샘', '연세 20세기 한국어 말뭉치', '세종 말뭉치') 속 ('상식'이라는 단어를 포함하고 있는) 예문들을 분석하여 요약하고 있습니다. 그는 상식의 몇 가지 특성을 ①교과서와 같이 명문화된 지식이 아니라 공동체에 의해 암묵적으로 공유되는 지식, ②유추를 통해 도출될 수 있는 지식, ③'윤리'와 관련이 있는 지식, ④오류의 가능성을 내포하고 있어 수정될 수도 있는 지식 등으로 정리하고 있습니다.

발표자는 상식에 대한 개념을 정리하고 나서 (GPT-3와 같은) 초거대 언어모델의 대표적인 연구들을 소개합니다. 이들 연구는 언어모델로 하여금 언어로 표현된 정보와 지식을 활용하여 특정한 문제를 풀도록 하고 있으며, 예를 들어 언어모델이 사칙연산을 수행할 수 있는지, 체스의 체크메이트 상황을 이해할 수 있는지 등을 조사하고 있습니다. 또 언어모델의 한계를 지적한 연구도 소개하는데, 언어모델을 '확률적 앵무새(stochastic parrot)'라고 지칭한 연구가 그것입니다. 발표자는 언어모델이 이러한 한계를 넘기 위해서는 '행간의 의미, 문맥의 의미'를 파악해야 하며, 그러기 위해서는 언어모델이 상식적 판단을 할 수 있어야 한다고 말합니다. 그리고 상식적 판단을 할 수 있으려면 추론 능력과 상식이 중요하다고 지적합니다.

이후 발표자는 "상식 추론에 있어서 철학적 뼈대"를 형성하는 것이 가추 추론이라고 설명합니다. 발표문의 10~11쪽에 가추 추론에 관해 설명이 제시되어 있습니다만, 여기에 몇 가지 예를 더 들어 보겠습니다.[1)]

1) 이들 예시는 "Stanford Encyclopedia of Philosophy"에서 발췌.

(1) 팀과 해리가 크게 싸워 우정에 금이 갔다. 그런데 오늘 아침에 팀과 해리가 함께 조깅하는 모습을 보았다. 이에 둘이 아마 화해했고 다시 친구 사이로 돌아간 모양이라고 결론 내렸다.

(2) 아침에 부엌에 가보니 빵가루가 남겨진 접시와 우유가 묻은 컵, 잼 통이 테이블 위에 놓여 있었다. 이에 간밤에 룸메이트가 야식을 먹고 안 치운 모양이라고 결론 내렸다.

(3) 해변을 거닐고 있는데 모래사장에 윈스턴 처칠의 얼굴이 그려져 있었다. 이에 누군가가 재미로 모래 위에 그림을 그린 모양이라고 결론 내렸다.

위의 세 예문은 모두가 가추 추론을 이용하여 결론에 도달하고 있는 사례들입니다. 이들은 모두 전제(premises)에서 논리적인 결론을 도출하고 있지는 *않지만*, 다양한 결론 중에서 그 상황을 *가장 잘 설명할 수 있는 결론*을 도출하고 있습니다(Inference to the Best Explanation).

가령 (1)의 경우 또 다른 결론으로 팀과 해리가 비즈니스 파트너여서, 비록 둘이 절교했지만 사업 문제를 논의해야 해서 같이 조깅을 하면서 사업 논의를 하고 있었던 것이라는 결론을 내릴 수 있습니다. 또 (2)의 경우도 간밤에 도둑이 들어서 물건을 훔치다가 배가 고파 빵과 우유를 훔쳐 먹었을 것이라는 결론을 내릴 수 있습니다. (3)의 경우도 사람이 아니라 개미 군단이 모래사장에 윈스턴 처칠의 얼굴을 새겼다는 결론을 내릴 수 있습니다. 이밖에도 다양한 결론을 내릴 수 있는데, 우리 인간은 그 많은 결론 중에서 (1), (2), (3)에서 제시한 결론을 가장 그럴듯한, 즉 상식적인 결론으로 보고 있습니다. 이렇게 가장 그럴듯한 설명을 통해 상식적인 결론을 도출해 주는 추론 기법이 바로 가추 추론입니다.

발표자는 이어 인공지능 분야에서 진행되고 있는 최신 가추 추론 연구에 관해 소개하고 가추 추론의 필요성을 요약하면서 발표문을 마무리하고 있습니다.

---

https://plato.stanford.edu/entries/abduction/#AbdGenIde

2. 질문: 연역법, 귀납법이 언어모델의 상식 판단에 도움을 줄 수 있을까요?

이제 발표자께 질문을 드립니다. 인공지능 언어모델이 상식적 판단을 내릴 수 있으면 좋을 것이라는 점에는 이견이 없습니다. 그런데 언어모델이 상식적 판단을 내릴 때 가추 추론 말고 연역법이나 귀납법이 도움을 줄 수 있는 부분이 있을까요?

예를 들어 연역법(deduction)은 세 추론 기법 중에서 유일하게 전제가 참이면 결론이 참임을 보장하는 추론 기법(necessary inference)입니다[2]. 예를 들어,

모든 A 는 B 이다.
a 는 A 이다.
따라서 a 는 B 이다.

와 같이 결론이 논리적으로 도출되는 추론 기법이 상식 추론에 도움을 줄 수 있는 부분이 있을까 여쭙니다.

또 현재 초거대 언어모델이 연역 추론(deductive reasoning)을 얼마나 잘하는지 혹시 알고 계시는 부분이 있으면 소개해 주시면 고맙겠습니다.

질문 1. 연역법 and/or 귀납법이 언어모델 상식 추론에 도움을 줄 수 있을지?

질문 2. 최신 초거대 언어모델 연구에서 연역법을 다룬 연구가 있는지? 있다면 언어모델이 연역 추론을 얼마나 잘하는지?

감사합니다.

2) https://plato.stanford.edu/entries/abduction/#DedIndAbd
연역법을 제외한 나머지 귀납법과 가추법은 non-necessary inference로 구분되고, 전제가 참이라도 결론이 참임을 보장하지 않습니다. 윈스턴 처칠 그림이 모래사장에 그려져 있어도 그 그림을 사람이 그렸음을 보장하지는 않습니다. (바다사자나 외계인이 그렸을 수도 있습니다.)