초거대 언어모델의 현재와 미래(이기창)에 대한 토론

신서인(한림대)

이기창 선생님의 발표 잘 들었습니다. 자연언어처리 분야에서 언어모델이 작동하는 원리와 초거대 언어모델이 등장하게 된 배경을 설명해주시고, Hyper CLOVA를 이용하여 수행할 수 있는 다양한 과제들의 예시를 보여주셔서 무척 흥미로웠습니다. 한 문장을 구성하고 있는 각각의 요소들은 여러 층위의 의미와 기능을 가지고 있는데 이런 것들이 과연 벡터라고 하는 숫자 값의 집합으로 변환이 가능한 것인가 하는 근본적인 의문이 들기도 하였고, 그럼에도 불구하고 훌륭한 성능을 보이고 있는 예들을 보면 그 과정을 어떻게 설명할 수 있을까 하는 생각도 들었습니다. 제가 이 분야의 전문가가 아니라서 미처 이해를 하지 못한 부분들이 많이 있었는데 그에 대한 질문을 드림으로써 토론을 대신하고자 합니다.

1. 현재의 언어모델은 입력된 단어 시퀀스를 바탕으로 다음 단어를 맞추는 방식으로 훈련을 한다고 하였습니다. 이 경우 인접한 단어들이 무엇인지가 중요한 정보일 것입니다. 언어는 선조적으로 실현되지만 그 내부에는 위계적인 구조가 있는 것도 사실입니다. 멀리 떨어져 있는 단어가 더 중요한 정보를 담고 있을 수도 있고, 인접한 단어가 덜 중요한 정보를 담고 있을 수도 있습니다. 예를 들어 '철수가 파란 옷을 입고 왔다.'와 같은 문장에서는 '철수가'와 '파란'이 인접해 있지만, '철수가 파랗다.'라는 의미를 가지고 있지는 않습니다. 물론 셀프 어텐션의 방법으로 단어-단어 쌍의 가중치를 달리하는 방법이 있지만, 계산에 많은 자원이 소요될 것입니다. 문장의 구조적인 정보를 언어모델의 입력에 효과적인 반영할 수 있는 방법은 없는지 궁금합니다.

2. 언어모델이 문장을 생성하는 것은 이전 단어들을 바탕으로 다음 단어에 어떤 단어가 오는 것이 가장 그럴듯한지 예측하는 것이라고 하였습니다. 그런데 한국어에서는 문장의 구조를 결정하는 서술어가 뒤쪽에 나옵니다. '철수가 그 집을 샀다.'와 '철수가 그 집에 산다.'라는 두 문장에서는 '철수-가-그-집'으로 이어지는 단어 연쇄가 동일합니다. 앞의 문장에서 '집' 다음에 '을'이 나오는 것은 '사다'가 결정하는 것이고, 뒤의 문장에서 '집' 다음에 '에'가 나오는 것은 '살다'가 결정하는 것입니다. 또한 한국어는 수식어-피수식어의 어순이 고정적입니다. 문장을 뒤에서부터 생성하는 방법은 없는지 궁금합니다.

3. 최근의 경향은 초거대 언어모델이 있기만 하면 제로샷 러닝, 원샷 러닝, 퓨샷 러닝 등 인컨텍스트 방식으로 태스크 데이터를 거의 사용하지 않고 구체적인 과제(다운스트림 태스크)를 수행할 수 있다고 하였습니다. 파인튜닝을 통해 모델을 업데이트 하지 않고도 요약, 번역, 대화 등의 과제가 어떻게 가능한지에 대한 알려진 설명이 있는지 궁금합니다.

4. 자유대화에서 대화가 자연스러운 정도를 어떻게 평가하는지, 번역의 성능을 측정하는 공인된 지표가 있는지 궁금합니다.

5. 종래의 자연언어처리는 규칙 기반의 방식으로서 형태 분석, 구문 분석, 의미 분석 등의 단계로 나눌 수 있었고 각 단계에 국어학의 연구 성과를 어떻게 반영할 수 있을지가 비교적 분명했습니다. 그러나 최근의 자연언어처리는 딥러닝 방식으로서 중간과정을 알기 어려운 블랙박스의 형태이고 엔드-투-엔드 방식이라 결국 입력을 어떻게 줄 것인가 하는 것에만 관여할 수 있는 것 같습니다. 초거대 언어모델의 발전에 기여하기 위해서 국어학자들이 할 일이 무엇인지, 즉 기존의 국어학의 연구 성과를 어떠한 형식으로 가공해야 하는지 알려주시면 감사하겠습니다.

흥미로운 발표를 해주신 데 다시 한번 감사드립니다.