<단어 의미의 상호작용을 모델링하는 방법(박진호)>에 대한 토론문

이찬규(중앙대학교)

이 연구는 언어 처리 모델인 BERT가 동위소(isotopy)를 반영할 수 있는지를 확인하고, 다의어와 동음어를 지닌 문장을 벡터 처리했을 때 의미를 좀 더 명확히 구별할 수 있는 방안을 실험적으로 제시했다는 점에서 의의가 있습니다. 실험적인 연구이기는 하지만 논의를 진행하기 위해 몇 가지 질문과 의견을 제시해 보겠습니다.

1) 먼저 Word2vec에서는 문장에서 앞 뒤 단어만을 통해 목표(target) 단어를 예측해 내기 때문에 isotopy(이하 동위소)를 구별해 내기가 어려운 반면, BERT 모델은 양방향으로 학습하기 때문에 문맥을 더 잘 파악할 수 있어서 동위소를 더 잘 분석해 낼 수 있다고 보았습니다.

동위소는 그레마스(1970년) 제안한 개념으로 "담화의 기저에 있는 반복적 의미론적 범주들의 다발을 의미한다"고 정의하고 있다. 이 후 많은 학자들이 동위소 개념을 정교화하는데 기여해 왔지만 대부분이 텍스트의 기저에 깔린 의미소들의 연결성을 분석해내는 방식에 초점을 맞추어 오고 있습니다. 이 동위소 개념은 자연어 처리의 언어 의미 관계 분석에 매우 유용하게 사용될 수 있습니다.(이미 수학이나 물리학에서 사용되는 것으로 알고 있음.) 벡터 공간에서 비슷한 위치에 나타난다는 것은 의미상 관련이 있다는 증거라고 할 수 있습니다. 하지만 언어학의 관점에서 보면 관련되는 단어들이 어떤 동위소 의미를 지니는지 파악하는 것이 중요한데, 현재 나와 있는 언어모델이나 향후의 연구에서 이런 문제를 해결하려는 시도는 있는지요?

또 이 발표에서는 문장 내의 공기관계에 초점을 맞추어 동위소를 파악하고 있는데, 이러한 방법을 이용하여 텍스트의 범주로까지 확장하여 텍스트성을 확인할 수도 있다고 생각합니다. 이러한 확장이 가능할지에 대해서도 설명 부탁드립니다.

2) 또한 각 단어의 부면(facet)을 확인할 수 있으면 이를 활용하여 동음어나 다의어를 비교적 용이하게 구분할 수 있다고 보고 있습니다. '눈'이 '아프다', '떨어진다', '예쁘다' 등 어떤 단어와 결합하는냐에 따라 각각 특정 부면이 부각된다고 보고 이를 '눈_아프다', '눈_떨어진다', '눈_예쁘다'처럼 관련성이 깊은 요소만을 뽑아 벡터화하는 것이 훨씬 구분이 쉽다고 하였습니다. 여기서 Word2vec을 이용하여 단어를 벡터화하고, FSE로 문장을 벡터화할 경우 중의성 해소에 만족스러운 결과를 나타내지도 못하고, 시간도 많이 걸리는 문제가 있어 문장의 통사 구조에 대한 언어학적 고려를 제안하였습니다.

"문장의 모든 토큰을 포함해서 벡터화할 게 아니라, 중의성을 해소하고자 하는 타깃 단어와 통사적으로 긴밀한 관계에 있는 요소들만 뽑아서 벡터화하는 게 더 좋을 것이다."라고 하였는데, 그런데 이 경우 '통사적으로 긴밀한 관계가 있는 요소'들을 어떻게 모든 문장에서 하나하나 토큰화하느냐가 문제라고 봅니다.

예컨대 ‘재미있-는 책-을 읽-었-다’에서 ‘재미있-는’과 ‘읽-었-다’는 직접적인 관련이 없다. ‘책-을’은 ‘읽’과 관련을 맺고, ’었, 다’와는 직접적인 관계가 없다. ‘재미있-는’은 ‘책’과 관련을 맺고 ‘을’과는 직접적인 관계가 없다. 따라서 실제로 고려해야 할 관계는 ‘책, 을, 읽’ 셋 사이의 관계, ‘읽, 었, 다’ 셋 사이의 관계, ‘재미있, 는, 책’ 셋 사이의 관계뿐이다.

데이터가 제한적인 실험에서는 위와 같은 구문 분석 방식이 가능하겠지만 실험이 아닌 용량이 큰 실제 언어처리에서도 이러한 방식이 가능한지요?