사전훈련모델의 언어 지식 수준 평가

고사성어(故事成語) 편

중어중문학과 박사과정 이다연

배경

- 사전훈련모델의 언어 지식에 대한 연구가 많이 이루어졌음
- Syntax, semantic, word, commonsense…
- 고사성어는 압축적으로 많은 의미를 전달할 수 있는 효율적인 단어
- 문맥을 이해해야 알 수 있음 → commonsense 차원에서 평가 가능
- 문어(ex. 뉴스)와 구어(ex. 대화)에서 모두 빈번하게 등장
- 과연 사전훈련모델이 얼마나 이해하는지 의문

정의

- 고사성어: 옛이야기에서 유래한, 한자로 이루어진 말.
- 한자성어: 한자로 이루어진 말. 교훈이나 유래를 담고 있다.
- 사자성어: 한자 네 자로 이루어진 성어. 교훈이나 유래를 담고 있다.
- 성어: 옛사람들이 만든 말.

→옛 이야기를 담고 있는 단어 수준에서 쓰임.

고사성어를 자연스럽게 이해, 사용한다면?

- 해당 문맥을 완전히 이해했다는 것을 드러냄
- 박식함을 드러냄 (과하면 현학적)
- 자연스러운 문장 구사 가능
- 형태소 분류기와 비교 – 형태소를 분류할 때 하나의 토큰으로 분류 잘 못함

진단 (1) 문장 내 [MASK]

*모델: klue/roberta-large

그것은 비일비재한 [MASK] 입니다.

Compute

Computation time on cpu: 0.183 s

일	0.521
현실	0.067
현상	0.059
범죄	0.042
상황	0.019

[MASK]하게 일어나는 일입니다.

Compute

Computation time on cpu: 0.117 s

비일비재	0.519
당연	0.152
아이러니	0.022
빈번	0.022
어마어마	0.016

- 기본적으로 명사로 쓰임
- 다른 문장성분과 함께 결합되는 경우도 있음

비일비재한 일

비일비재하게 일어나는 일

흔히 일어나는 일

진단 (1)

사면초가의 [MASK]로 내몰렸다

Compute

Computation time on cpu: cached

위기	0.625
궁지	0.090
처지	0.076
상태	0.064
상황	0.026

[MASK]의 위기로 내몰렸다

Compute

Computation time on cpu: cached

IMF	0.064
[PAD]	0.045
[UNK]	0.042
는	0.026
가	0.025

사면초가의 위기로 내몰렸다

IMF의 위기로 내몰렸다

파산 위기로 내몰렸다

[MASK] 위기로 내몰렸다

파산	0.223
부도	0.050
존폐	0.041
경제	0.030
사면	0.027

진단 (2) Sentence Generation

*모델: skt/Ko-GPT-Trinity 1.2B (v0.5)

그가 금의환향해서 돌아왔다.

Compute

Computation time on cpu: 1.690 s

그가 금의환향해서 돌아왔다. 그리고 그 후로도 몇 차례 더 한국을 방문했다. 그는 한국 방문

오매불망

Compute

Computation time on cpu: 2.016 s

오매불망 기다리던 그 날.

주경야독하며

Compute

Computation time on cpu: 0.911 s

주경야독하며 학업에 정진하고 있다.

인생은 새옹지마다. 때로는

Compute

Computation time on cpu: 1.813 s

인생은 새옹지마다. 때로는 모든 것이 뜻대로 되지 않을 때도 있다. 하지만 좌절하지 말자.

- 상용되는 고사성어를 주면 문장을 잘 만들어 냄

사전모델이 학습한 지식에 대한 의문

- 고사성어는 몇 자의 글자로 풍부한 의미를 포함하기 때문에 사용된 문장의 길이가 짧은 경향이 있음: 전형적인 단어와의 결합만 학습되었을 수 있음

ex.이는 비일비재하게 일어나는 일입니다

- 고사성어가 포함된 문장에서 고사성어를 제외한 다른 단어들에 문장의 의미를 알 수 있는 힌트가 많음: 고사성어가 수식성분 수준에서 활용되는 단어일 수 있음

ex. 그가 금의환향해서 돌아왔다

→ 전형적인 단어와의 결합을 고려하고, 문맥을 이해해야 사용할 수 있는 문장 형식으로 만든 데이터셋으로 평가 필요

실험 계획

- 논문 참조: <Evaluating Commonsense in Pre-trained Language Models(2019)>
- 한국어 문장에 맞게 변형
- 사전훈련모델 선정: GPT 계열 / BERT 계열 모델
- 점수 계산

n개의 단어의 문장 S = {w_1 , …, w_{k-1}, w_k, w_{k+1}, …, w_n}일 때,

$$Score(S) = \frac{\sum_{k=1}^{n} log(P_\theta(w_k|context_k)}{n}$$

실험에서 예상되는 문제

- 논문처럼 Bi-directional vs Uni-directional 의 문제가 아니라 사전훈련모델 자체의 개별적인 능력으로 평가될 수 있음: 사전학습 데이터의 scale 문제

평가 데이터셋

- 고사성어를 다루고 있는 데이터셋은 없음
- 빈도가 높은 고사성어를 선정하고, 크롤링을 통해 실제 문장을 가져옴
- 문장을 수정하여 평가가 가능한 문장으로 만들기
- CATs 데이터셋을 참고해서 종류별 데이터셋 100개씩 구성
- Sense Making(SM): 말이 되는 문장인지 아닌지
- Conjunction Acceptability(CA): 접속사가 제대로 맞게 쓰였는지 아닌지
- SWAG: 주어진 문장의 상황과 관계된 문장인지 아닌지

1. Sense Making(SM) - 말이 되는 문장인지 아닌지

- 삼고초려

-신동빈 롯데그룹 회장이 오랜 전통의 순혈주의를 과감히 깨고 삼고초려 끝에 김상현 롯데 유통군HQ 총괄대표를 직접 영입하였다. (o)

-신동빈 롯데그룹 회장이 오랜 전통의 순혈주의를 과감히 깨고 고진감래 끝에 김상현 롯데 유통군HQ 총괄대표를 직접 영입하였다. (x)

- 유비무환

- 철저한 대비와 늘 상황 유지에 온 힘을 다하고 앞으로도 시민 안전을 위해 유비무환의 자세로 폭설과 한파 등을 대비하겠다. (o)

-철저한 대비와 늘 상황 유지에 온 힘을 다하고 앞으로도 시민 안전을 위해 우공이산의 자세로 폭설과 한파 등을 대비하겠다. (x)

2. Conjunction Acceptability(CA) - 접속사가 맞게 쓰였는지

- 일장춘몽

-최근 미국 나스닥 지수는 훈풍이 불었다. 그러나 24일에 다시 급락세를 보이며 일장춘몽의 장세를 연출했다. (o)

-최근 미국 나스닥 지수는 훈풍이 불었다. 그래서 24일에 다시 급락세를 보이며 일장춘몽의 장세를 연출했다. (x)

- 고립무원

- 미국과 서방국가들이 직접적인 군사 개입에 선을 그었다. 그래서 우크라이나는 고립무원 상태에서 홀로 러시아 대군과 싸우고 있다. (o)
- 미국과 서방국가들이 직접적인 군사 개입에 선을 그었다. 왜냐면 우크라이나는 고립무원 상태에서 홀로 러시아 대군과 싸우고 있다. (x)

3. SWAG - 주어진 문장의 상황과 관계된 문장인지

- 새옹지마

- 인생은 새옹지마다. 그러니 좋은 일이 있어도 취하지 말며, 나쁜 일이 있어도 낙담할 일이 아니다. (o)
- 인생은 새옹지마다. 삶은 영원하지 않으며, 남은 행복을 소중히 여긴다. (x)
- 인생은 새옹지마다. 남이 모른다고 해서 그냥 넘어가는 게 아니다. (x)
- 인생은 새옹지마다. 짐은 무겁고 갈 길은 멀다. (x)

결론