문장 생성하기

NAVER Conversation
이기창

목차

- 과제 조망하기
- 모델 파인튜닝
- 문장 생성 인퍼런스

과제 조망하기

과제 조망하기

- 태스크 소개

  문장 생성하기

과제 조망하기

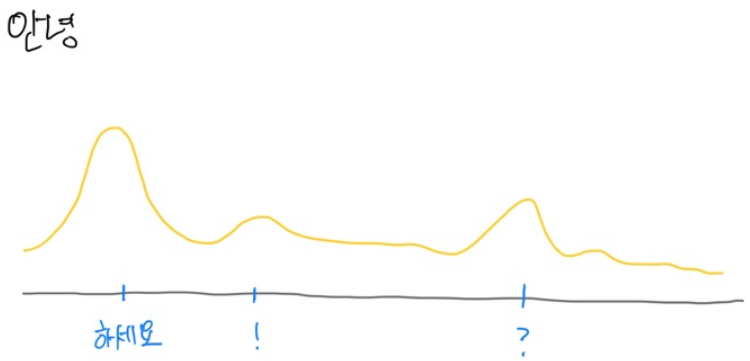
- 태스크 소개

  문장 생성하기 : 다음 단어 맞추기

과제 조망하기

- 태스크 소개

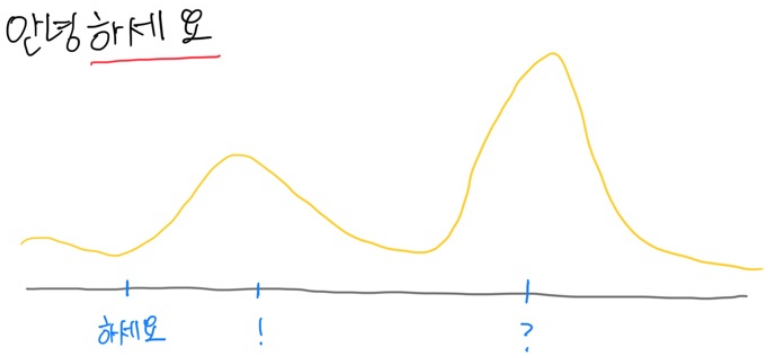
  문장 생성하기 : 다음 단어 맞추기



과제 조망하기

- 태스크 소개

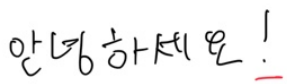
  문장 생성하기 : 다음 단어 맞추기



과제 조망하기

- 태스크 소개

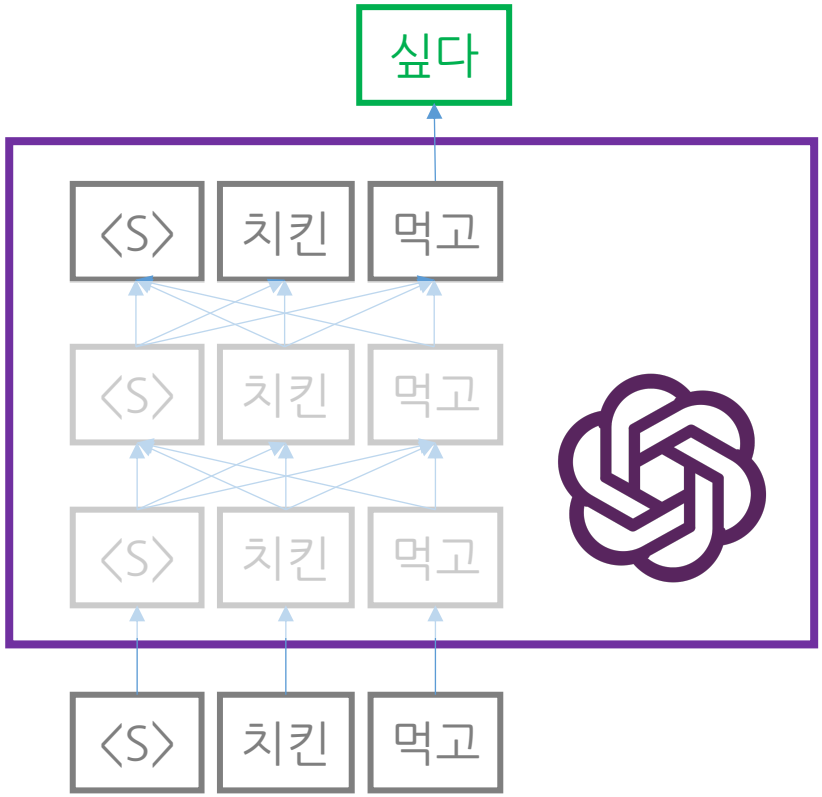
  문장 생성하기 : 다음 단어 맞추기



과제 조망하기

- 모델 전체 구조

GPT



모델 학습

(실습)

데이터

- 원본 데이터

NSMC

Quick peek

```
$ head ratings_train.txt
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리        0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나        1
10265843        너무재밓었다그래서보는것을추천한다       0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정        0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.      0
7797314 원작의 긴장감을 제대로 살려내지못했다.  0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫겟다 납치.감금만반복반
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```

데이터

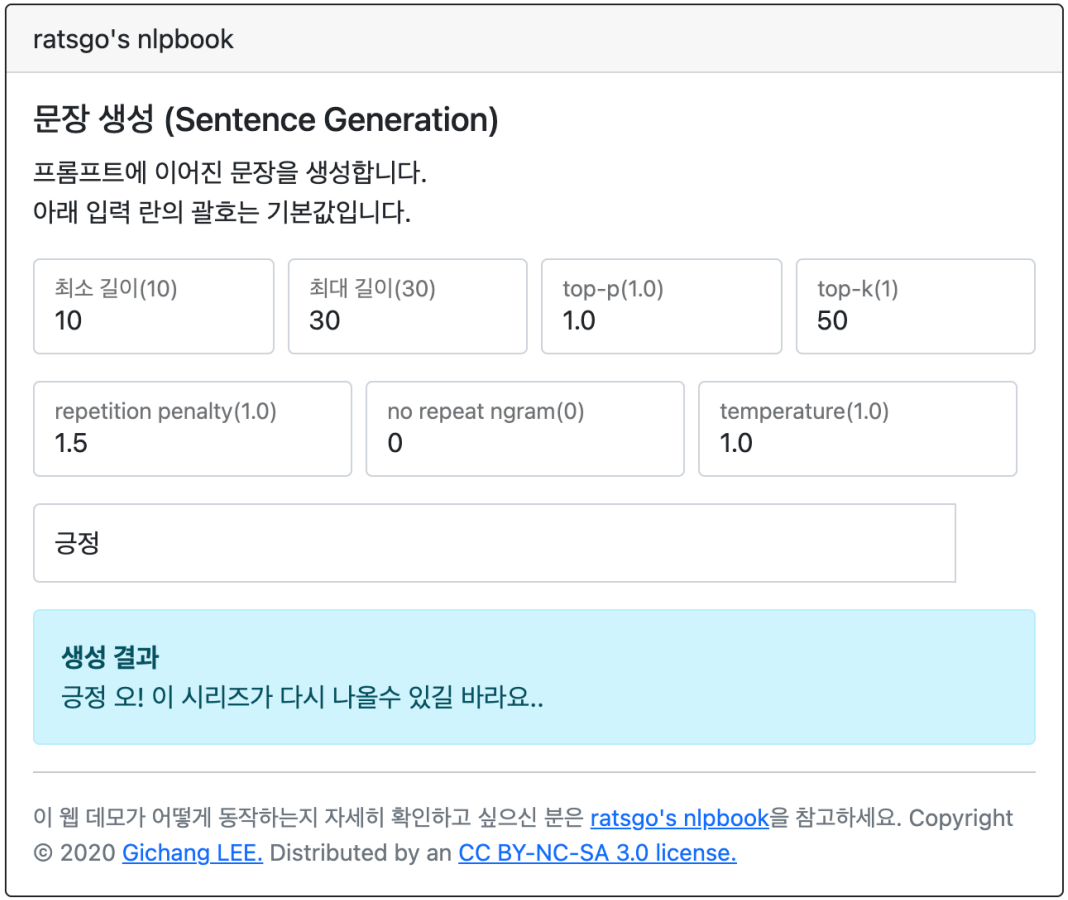
- 파인튜닝 데이터 처리 결과

NSMC

```
Creating features from dataset file at /root/Korpora/nsmc
loading train data... LOOKING AT /root/Korpora/nsmc/ratings_train.txt
tokenize sentences, it could take a lot of time...
tokenize sentences [took 7.762 s]
*** Example ***
sentence: 부정 아 더빙.. 진짜 짜증나네요 목소리
tokens: _부정 _아 _더 빙 .. _진짜 _짜 증 나 네 요 _목소리 </s> </s> </s> </s> </s> </s> </s> </s> </s> </s
features: GenerationFeatures(input_ids=[11775, 9050, 9267, 7700, 9705, 23971, 12870, 8262, 7055,
*** Example ***
sentence: 긍정 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나
tokens: _긍정 _흠 ... 포 스터 보고 _초 딩 영화 줄 .... 오 버 연기 조차 _가볍 지 _않 구나 </s> </s> </s> </s> </
features: GenerationFeatures(input_ids=[16420, 19243, 29045, 8658, 11211, 11213, 9206, 7301, 1455
*** Example ***
sentence: 부정 너무재밎었다그래서보는것을추천한다
tokens: _부정 _너무 재 <unk> 었 다 그래 서 보는 것 을 추천 한다 </s> </s> </s> </s> </s> </s> </s> </s> </s
features: GenerationFeatures(input_ids=[11775, 12371, 8170, 5, 8017, 7182, 19561, 7788, 12399, 68
*** Example ***
sentence: 부정 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정
tokens: _부정 _교도 소 _이야기 구 먼 _ .. 솔 직 히 _재 미는 _없다. . 평 점 _조정 </s> </s> </s> </s> </s> </s
features: GenerationFeatures(input_ids=[11775, 25365, 7824, 11120, 6919, 7514, 739, 9705, 7828, 8
*** Example ***
sentence: 긍정 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다
tokens: _긍정 _사이 몬 페 그의 _익살 스런 _연기가 _돋보 였던 _영화 ! 스파 이 더 맨 에서 _늙 어 보 이기 만 _했던 _커 스
features: GenerationFeatures(input_ids=[16420, 9435, 7543, 8643, 28657, 34499, 19912, 43245, 2177
```

모델 학습

- 형상


ratsgo's nlpbook
문장 생성 (Sentence Generation)
프롬프트에 이어진 문장을 생성합니다.
아래 입력 란의 괄호는 기본값입니다.
최소 길이(10)
10
최대 길이(30)
30
top-p(1.0)
1.0
top-k(1)
50
repetition penalty(1.0)
1.5
no repeat ngram(0)
0
temperature(1.0)
1.0
긍정
생성 결과
긍정 오! 이 시리즈가 다시 나올수 있길 바라요..
이 웹 데모가 어떻게 동작하는지 자세히 확인하고 싶으신 분은 ratsgo's nlpbook을 참고하세요. Copyright
© 2020 Gichang LEE. Distributed by an CC BY-NC-SA 3.0 license.

모델 학습

- ## 모델 (SKT 공개 KoGPT2)

김태윤님(TED) <tae.y.kim@sk.com> 5월 7일 (금) 오전 9:23 (2일 전)
이기창, 전희원님(HAVEN)에게

안녕하세요.

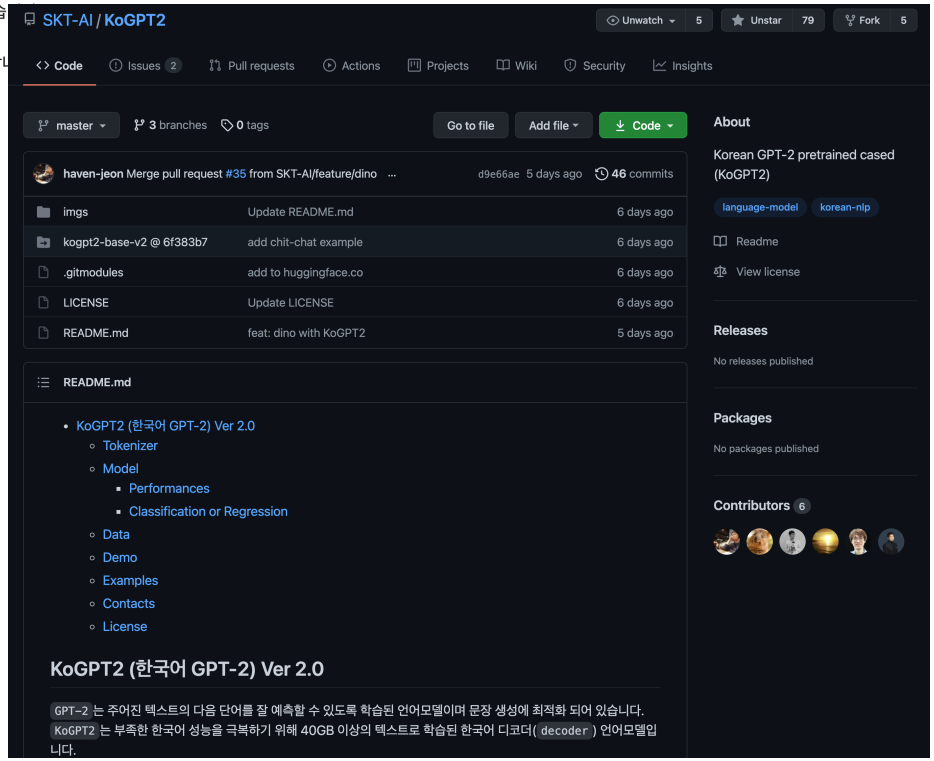
KoGPT2 에 관심을 가져 주셔서 감사 드립니다.
이 모델은 본래 많은 사람들이 사용하고 그것을 통해서 자연어 처리에 관심을 가지도록 하는 것을 목적으로 공개 하였습니다.

이창기님께서 준비 하시는 책에 저희 모델이 사용 될 수 있다면 저희로서도 고맙고 즐거운 일이 될 것 같습니다.

라이센스에 대해서는 저희 모델을 응용하는 법을 안내 하는 것이 라이센스에 위배되지 않는다고 이해하고 있습

이창기님의 책을 통해서 자연어 처리에 관심이 있는 분들께 저희 소프트웨어가 조금이라도 도움이 되기를 바랍니

김태윤 드림



모델 학습

- Colab 실습

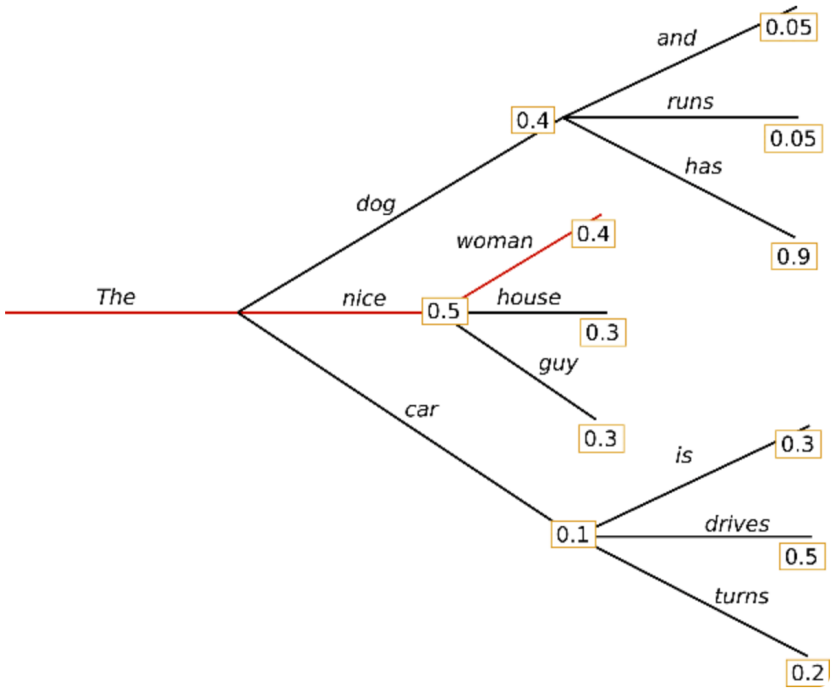
  https://colab.research.google.com/github/ratsgo/nlpbook/blob/master/examples/sentence_generation/train_colab.ipynb

인퍼런스

이론

- Greedy Search

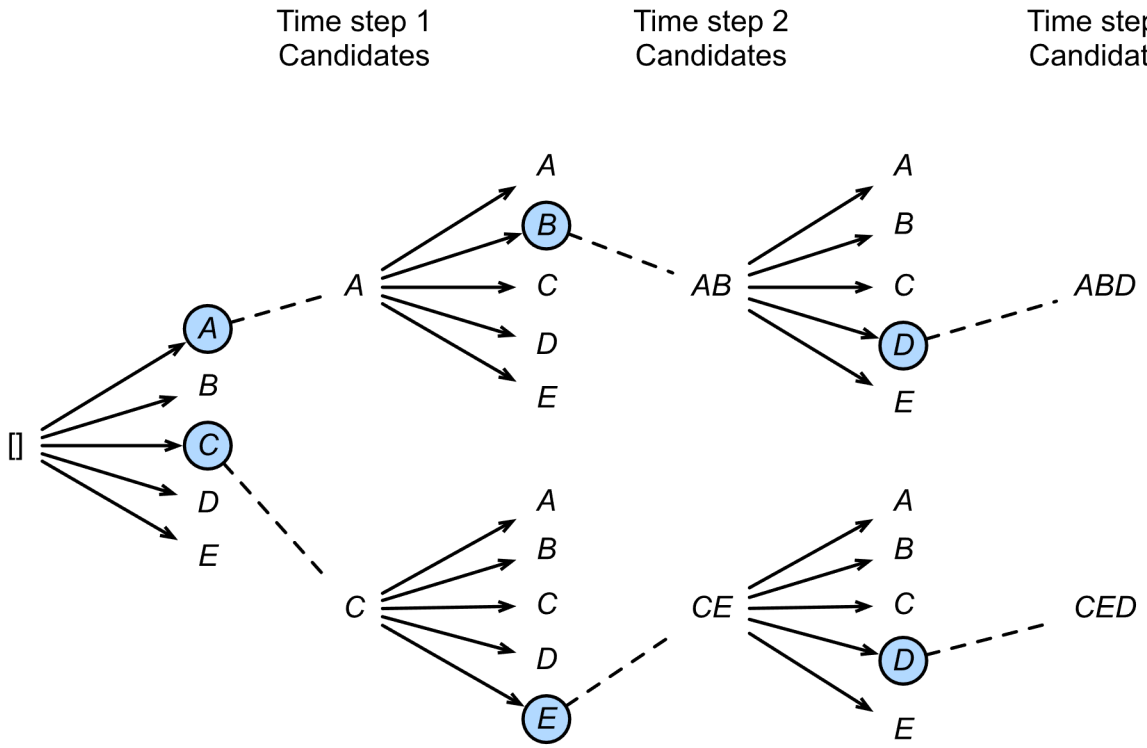
  다음 단어 선택시 가장 높은 확률 가진 단어 선택 (deterministic)



이론

- Beam Search

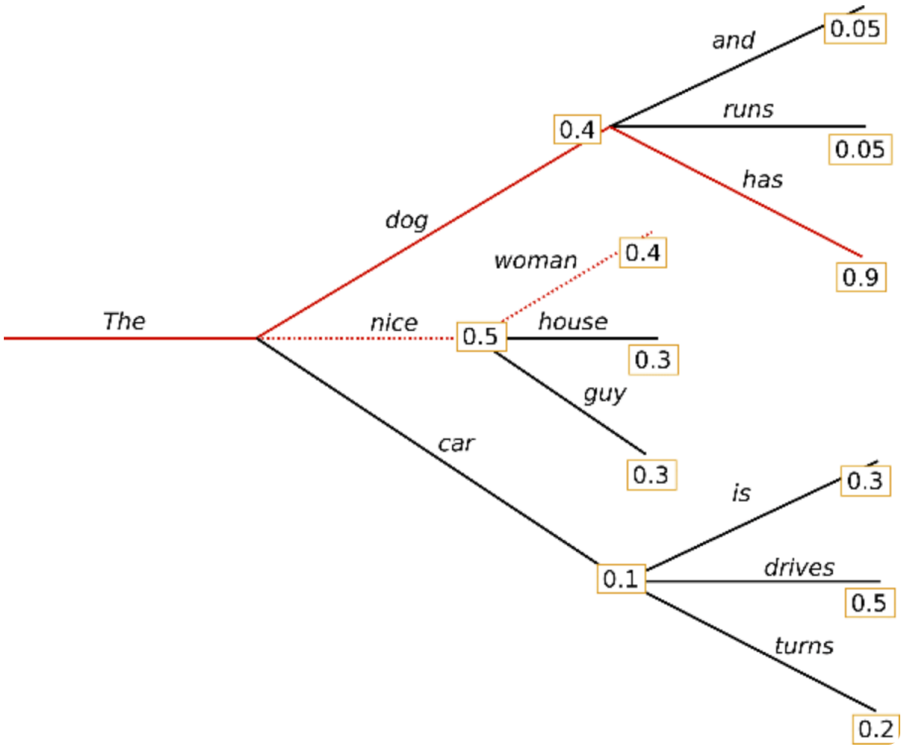
순간의 최적이 전체의 최적은 아님
Beam size(=2)만큼의 경우의 수 탐색 (deterministic)



이론

- Beam Search

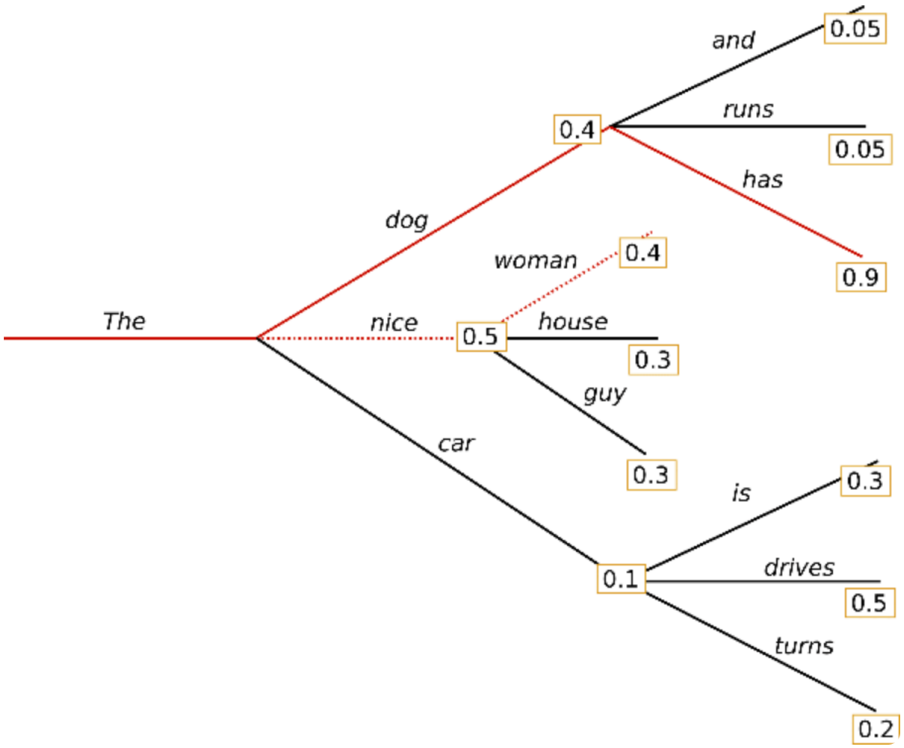
순간의 최적이 전체의 최적은 아님
Beam size(=2)만큼의 경우의 수 탐색 (deterministic)



이론

- Beam Search

순간의 최적이 전체의 최적은 아님
Beam size(=2)만큼의 경우의 수 탐색 (deterministic)

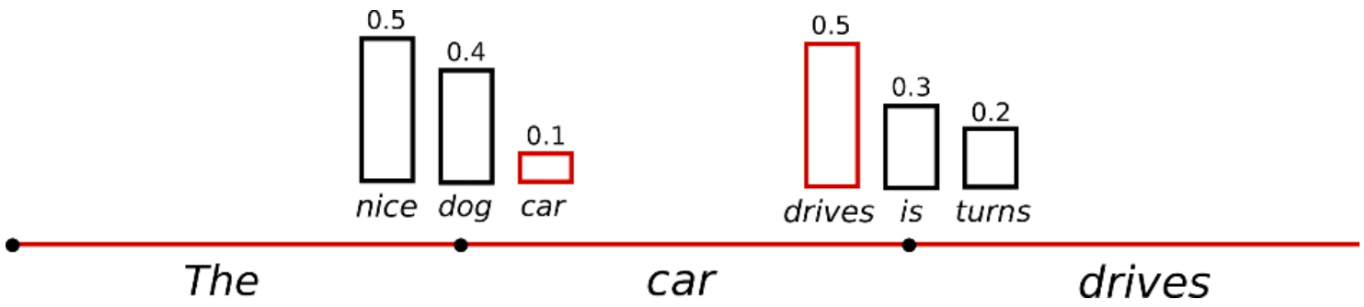


Greedy Search : 0.5 × 0.4 = 0.2
Beam Search : 0.4 × 0.9 = 0.36

이론

- sampling

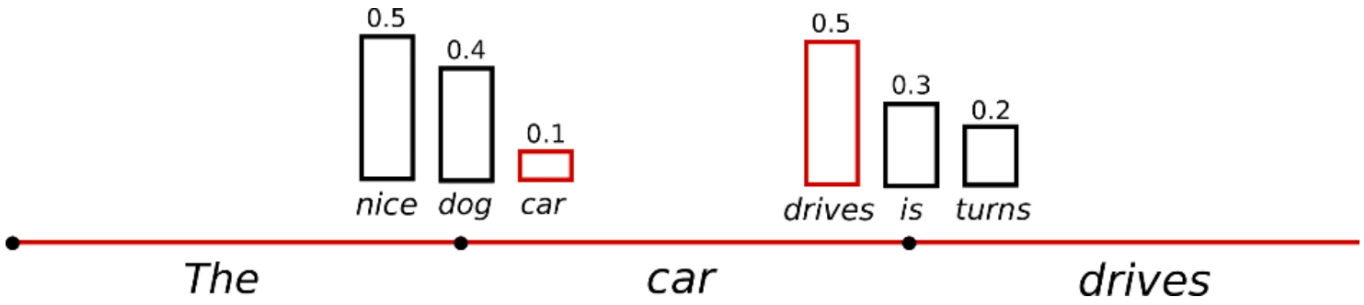
  다음 단어를 선택할 때 확률적으로 선택 (probabilistic)



이론

- sampling

다음 단어를 선택할 때 확률적으로 선택 (probabilistic)

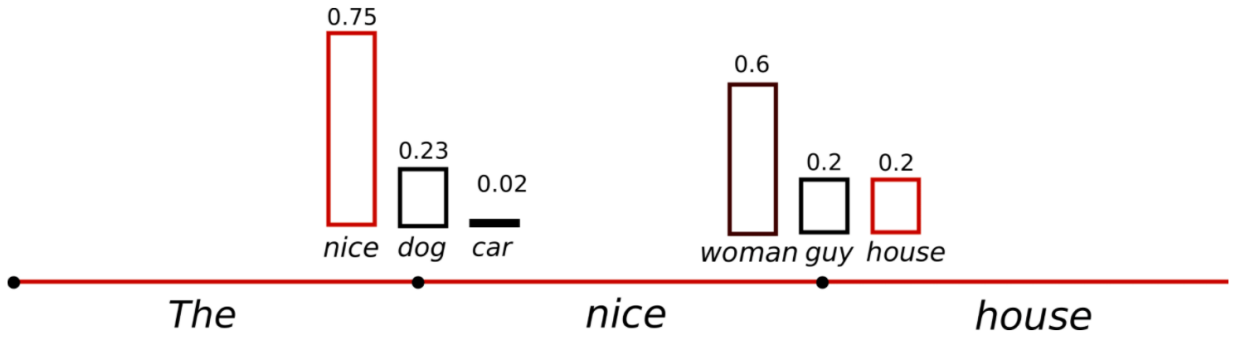


➔ 0.1로 가장 낮았지만 우연히 선택됨
➔ car가 선택될 가능성은 prob에 비례

이론

- sampling with temperature scaling

다음 단어 확률 분포를 변화시킨다 (probabilistic)



➔ temperature를 1보다 작게 하면

➔ 확률분포가 뾰족해짐 (대소 관계는 안변함)

➔ 원래 확률값이 높은 단어 위주로 선택

이론

- sampling with temperature scaling

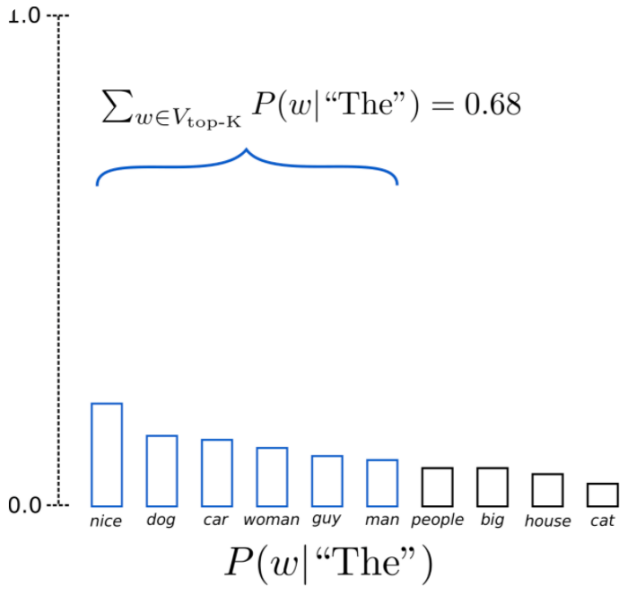
다음 단어 확률 분포를 변화시킨다 (probabilistic)

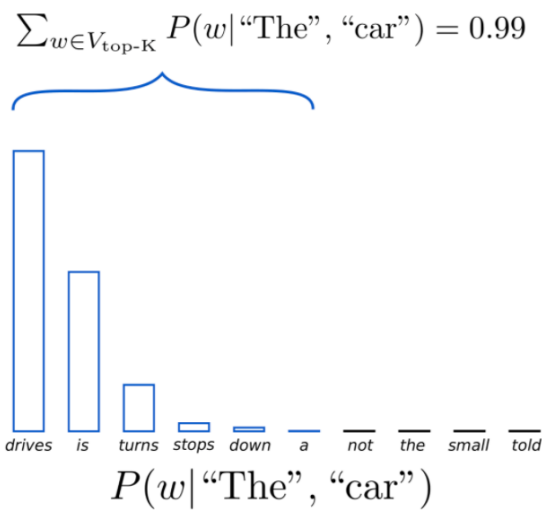
➔ temperature를 1보다 키우면
  ➔ 확률분포가 뭉툭해짐 (대소 관계는 안변함)
  ➔ 원래 확률값이 낮은 단어도 선택될 수 있음

이론

- top-k sampling

다음 단어 선택시 확률값 기준 top-k개 단어 내에서만 선택 (probabilistic)

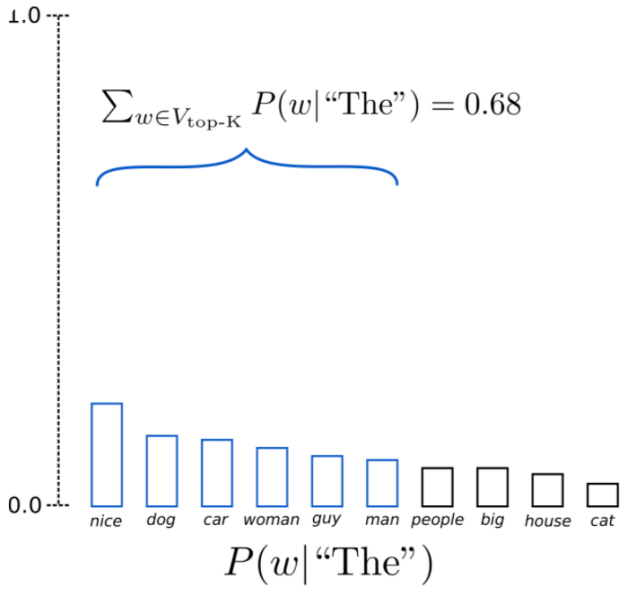




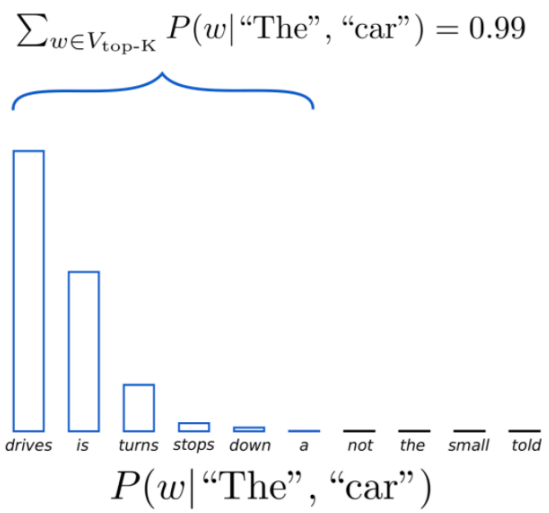
이론

- top-k sampling

다음 단어 선택시 확률값 기준 top-k개 단어 내에서만 선택 (probabilistic)



➔ The 다음 단어는 top-6개 중 선택



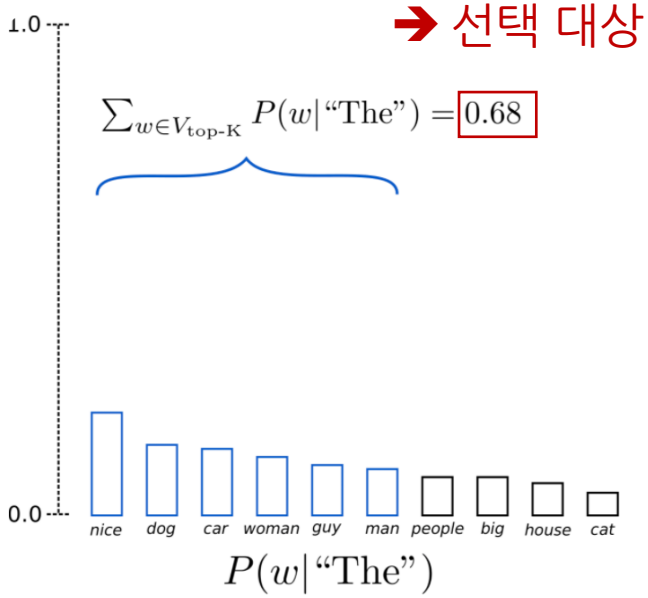
➔ The, car 다음 단어는 6개 중 선택

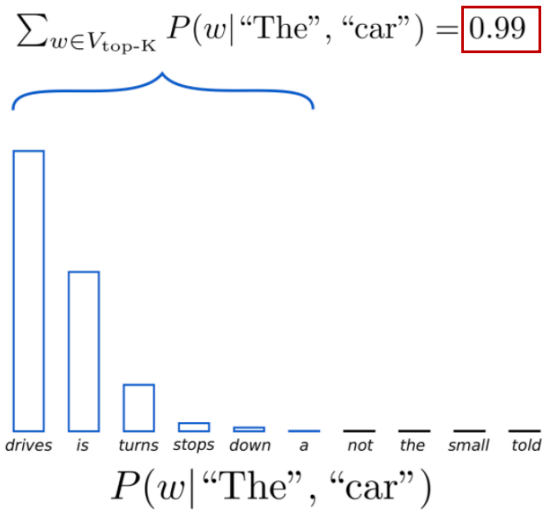
이론

- top-k sampling

다음 단어 선택시 확률값 기준 top-k개 단어 내에서만 선택 (probabilistic)

➔ 선택 대상 단어들의 확률합은 다를 수 있음





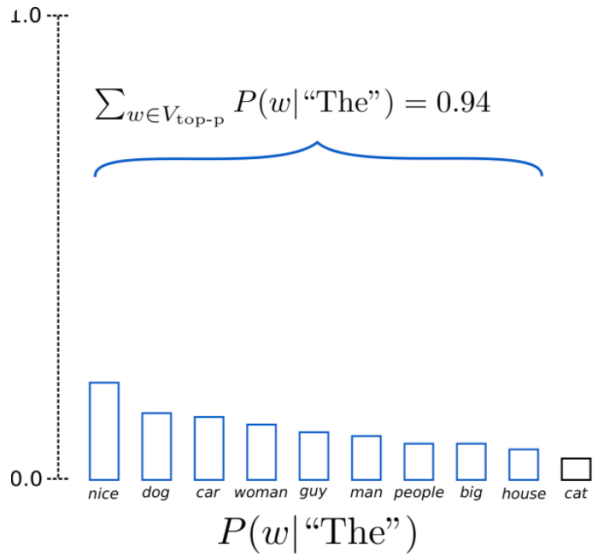
➔ The 다음 단어는 top-6개 중 선택

➔ The, car 다음 단어는 6개 중 선택

이론

- top-p sampling

다음 단어 선택시 누적확률 top-p개 단어 내에서만 선택 (probabilistic)

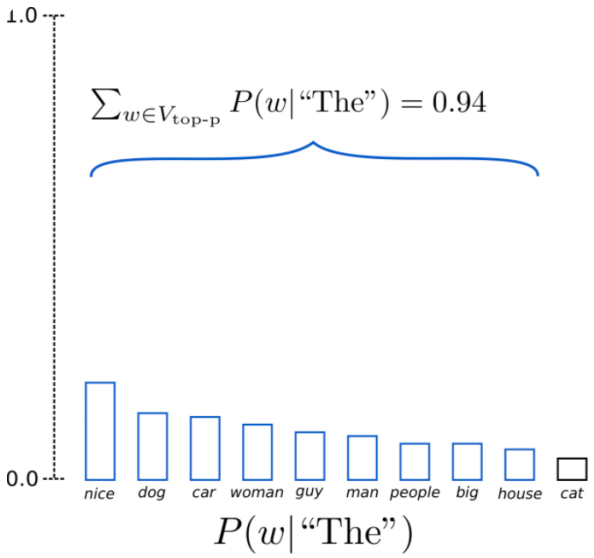


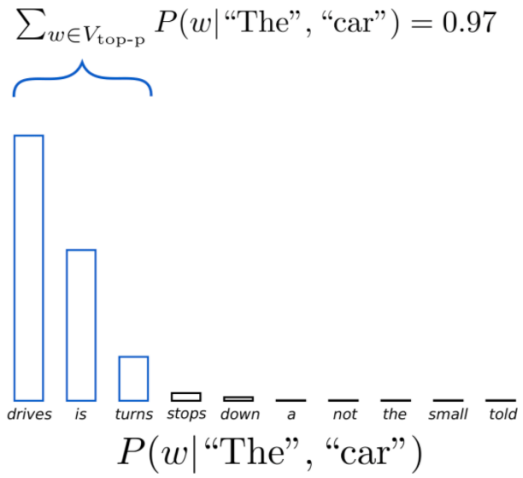
➔ 확률값 기준으로 단어들을 내림차순 정렬
➔ 만약 p를 0.92로 정했다면
누적확률 합이 0.92가 되는 단어들이 선택 대상

이론

- top-p sampling

다음 단어 선택시 누적확률 top-p개 단어 내에서만 선택 (probabilistic)





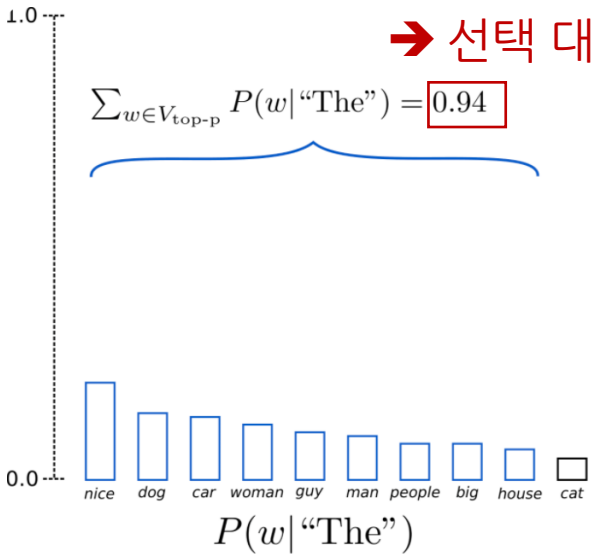
➔ The 다음 단어는 이들 중 선택

➔ The, car 다음 단어는 이들 중 선택

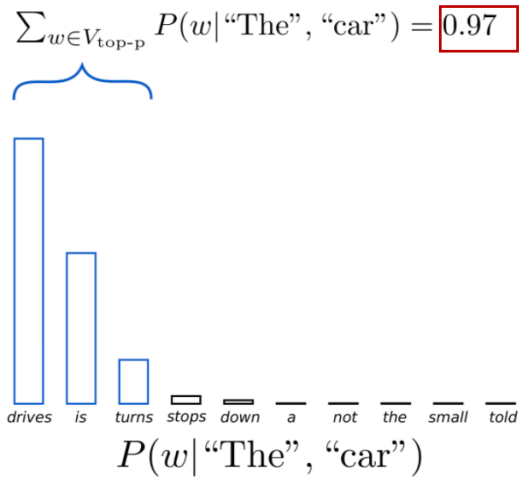
이론

- top-p sampling

다음 단어 선택시 누적확률 top-p개 단어 내에서만 선택 (probabilistic)



➔ 선택 대상 단어들의 확률합은 거의 비슷



➔ The 다음 단어는 이들 중 선택

➔ The, car 다음 단어는 이들 중 선택

이론

- top-k, top-p sampling 효과

  샘플링하면 다양한 문장 생성
  확률값이 낮은 단어들을 생성에서 배제함으로써 생성 품질 제고 효과

이론

- 반복 방지 기법

  반복되는 토큰의 등장 확률을 0으로 수정
  Repetition penalty

인퍼런스

- KoGPT2 모델 그대로 Colab에서 인퍼런스하기

  https://colab.research.google.com/drive/1Re0URacG2ocXi6BhUNCpAJ7M-gGjrcNV?usp=sharing

- 파인튜닝 모델을 Colab에서 인퍼런스하기

  https://colab.research.google.com/github/ratsgo/nlpbook/blob/master/examples/sentence_generation/deploy_colab.ipynb