CTRL: A CONDITIONAL TRANSFORMER LANGUAGE MODEL FOR CONTROLLABLE GENERATION

이명재

ABSTRACT

- 이전 text generation
  - 특정 측면에 대한 생성 제어가 쉽지 않음
- 우리 text generation
  - Control code (제어코드) 로 스타일, 콘텐츠, task 별 동작을 제어 할 수 있음
- 비지도 학습

- 입력 형식
  - **BPE 사용**
  - **Control code 는 시퀀스의 첫 번째 토큰으로 존재**

- 모델
  - Transformer encoder – hugging face (Tensorflow)

Control code 종류 [55개]

1	0.000296793	Pregnancy
2	0.000127197	Christianity
3	0.003084531	Explain
4	0.000180196	Fitness
5	6.88985E-05	Saving
6	0.000217295	Ask
7	8.47981E-05	Ass
8	0.000143097	Joke
9	0.000196096	Questions
10	0.000127197	Thoughts
11	0.000169596	Retail
12	0.000270294	Feminism
13	0.000111298	Writing
14	0.000402791	Atheism
15	1.05998E-06	Netflix
16	0.000365692	Computing
17	0.000132497	Opinion
18	0.000169596	Alone
19	0.000323293	Funny
20	0.000249094	Gaming
21	0.000402791	Human
22	0.000132497	India
23	2.11995E-08	Joker
24	0.000201395	Diet
25	0.000238495	Legal
26	6.35986E-06	Norman
27	3.60392E-07	Tip
28	0.000302093	Weight
29	0.000132497	Movies
30	0.000111298	Running
31	7.41983E-05	Science
32	0.00135147	Horror
33	0.000291493	Confession
34	0.000190796	Finance
35	0.000413391	Politics
36	7.41983E-05	Scary
37	0.000206695	Support
38	6.35986E-05	Technologies
39	0.000243795	Teenage
40	0.000217295	Event
41	0.000206695	Learned
42	0.000121897	Notion
43	0.0847981	Wikipedia
44	0.095927851	Books
45	0.001176574	Extract
46	0.000127197	Confessions
47	0.000227895	Conspiracy
48	0.365691808	Links
49	0.000423991	Narcissus
50	0.000280894	Relationship
51	0.000922179	Relationships
52	0.153696557	Reviews
53	0.043877717	News
54	0.129847091	Translation
55	0.111297507	multilingual

Control code [Style by domain]

- 제어코드를 특정 훈련데이터 영역표시로 사용
  - <표1> 동일한 프롬프트의 경우, 제어 코드가 예측 가능한 생성 변동을 허용함을 보여줌
  - <표2> 프롬프트 없이 domain-specific 텍스트를 생성을 보여줌

Wikipedia	*Anarchism is* a political philosophy that advocates the abolition of all forms of hierarchy and domination, including capitalism, patriarchy, racism, sexism, heterosexism and other oppressive social structures.\n\nThe term "anarchism" was coined by Pierre-Joseph Proudhon in his book "The Poverty of Philosophy" (1844). It has been used to describe various movements within anarchism since then. In its modern sense, it refers to an ideology or movement advocating for social, political, economic and/or cultural change through direct action against existing institutions and practices. Anarchists are often associated with radical left-wing politics, but they have also advocated many different types of libertarian socialism. Some anarchists advocate anarchist communism, others support anarcho-syndicalism, while still others reject both socialism and communism. Many anarchists believe that capitalism should be abolished entirely, though some argue that this would require overthrowing capitalism. Other anarchists oppose any form of state control over private property, such as land ownership, which they see as inherently oppressive. They may favor limited government regulation on certain aspects of life, such as drug use, prostitution, gambling, pornography, abortion, euthanasia, etc.\n\nHistory.\n\nEarly history.\n\nAnarchist ideas were first developed during the French Revolution (17891799). During these years, there was widespread opposition to traditional society based upon religion, class, gender, race, ethnicity, nationality, language, education, wealth, and so on. This led to the development of new ideologies like socialism, liberalism, republicanism, communitarianism, individualism/egoism/. . .
Books	*Anarchism is* the \n only true and practical form of Socialism. It has been said that Socialism \n will never be established until it can be proved to work in practice. The \n same may be said, with equal truth, about Anarchism. But if we are not able \n to prove its practicability at present, then let us wait for a better time \n when our present conditions shall have become so much more favorable than \n they now are, that we shall no longer need any such proof. Let us, then, \n patiently await this time, which I believe will come sooner or later. In \n the meantime, as long as there remains one man who believes in it, he must \n continue to do so. If you wish to see him succeed, try to make him feel his \n own impotence. You cannot succeed by force alone. He must know himself to \n be powerless before he begins to work. And what does he do? Why, he goes on \n working just as hard as ever, because he knows that he is powerless, but \n still he continues to work, because, after all, he thinks that he is doing \n right. That is why men like Mr. Owen and myself are called "cranks." We say \n things that seem absurd to others, but we mean them all the same. When we \n speak of "the right of labor," we really mean something very different from \n what most . . .
Horror	*A knife* handle pulled through the open hole in the front. I jumped when the knife hit.\n\nEyes widened in horror. Her scream was the only sound I heard besides her sobs.\n\nThe spider touched her feet as it started to dig into the top of her arch. The creature's tears began to flow.\n\nThe spider looked up to her and looked back at me with eyes filled with burning tears. My heart started to race. . .
Reviews	*A knife* is a tool and this one does the job well.\n\nRating: 4.0\n\nI bought these for my husband who has been using them to cut up his own meat since he got them. He says they are very sharp so be careful when you use them, but that doesn't seem like much of an issue because he's used it on everything from chicken breasts to beef tenderloin. . .
Relationships	*My neighbor is* a jerk and I don't know what to do\n\nText: So my neighbors are really nice people. They have been for years. We live in an apartment complex so we get along great.\n\nBut recently they started acting like jerks. . .
Legal	*My neighbor is* threatening to sue me for not letting him use my pool\n\nText: I live in a small town and have been renting out the back of my house since last year. The owner has had issues with people using it as well but he never really took any action against them. Last week someone from his property management company came by asking if they could come over because there was an issue with some kids playing on it. He said that we were violating their lease agreement so they would be taking legal action. . .

Table 1: Even for identical *prompts*, `control codes` allow for predictable variation in generation.

Science Title:	Scientists have discovered a new type of bacteria that can survive in the presence of high levels of carbon dioxide\nText:[Link to article] (http://news.nationalgeographic.com/news/2011/07/110717-methane-bacteria-carbon-dioxide-environment-science/)\n\nThe researchers found that this particular strain of methane bacterium is able to thrive at concentrations up to 10 times higher than normal atmospheric levels.\n\nThey also note that these findings are consistent with previous research showing that some types of microbes may be capable of surviving under extreme conditions such as those associated with global warming.
Politics Title:	The US is the only country in history to have a national debt of more than $20 trillion.\nText: I know this isn't exactly an accurate representation but it gives you some idea how much money we are talking about.\n\nhttp://en.wikipedia.org/wiki/National_debt#United_States
Running Text:	I have been running for about a year and a half now but never really got into it.\n\n This past weekend was my first time doing it in the last few months.\n\n My goal is to run at least 3 miles every day this week.
Horror Text:	I was a little girl when my parents got divorced. My dad had been in the military for years and he left me with my mom. She worked as an RN at a hospital so she could take care of me.\n\n When we moved to our new house it took some time before things settled down. We were still living together but there wasn't much going on. It didn't help that my mom would get mad if someone came over or even just walked by her house.\n\n One day while walking through the yard I noticed something out of place...
Reviews Rating: 5.0\n\n	I have been using this product for a few years and it is the best thing on my hair. It makes your hair soft but not greasy.\n\nRating: 5.0\n\nI've used this product before and love it. It's light weight yet keeps my curls defined without making them crunchy!
Reviews Rating: 1.0\n\n	I bought this for my daughter and it was a complete waste of money. The case is very flimsy and the screen protector doesn't fit properly.\n\nRating: 1.0\n\nThis product did not work at all. It would charge but when you unplug it, it does nothing. Very disappointed.

Table 2: With CTRL, no *prompt* is necessary as long as a `control code` is provided. Control codes can be combined (`Reviews`, `Rating:`, and `VALUE`) to provide finer-grained control.

Control code [More complex]

Science Title: Scientists have discovered a new type of bacteria that can survive in the presence of high levels of carbon dioxide\nText:[Link to article] (http://news.nationalgeographic.com/news/2011/07/110717-methane-bacteria-carbon-dioxide-environment-science/)\n\nThe researchers found that this particular strain of methane bacterium is able to thrive at concentrations up to 10 times higher than normal atmospheric levels.\n\nThey also note that these findings are consistent with previous research showing that some types of microbes may be capable of surviving under extreme conditions such as those associated with global warming.
Politics Title: The US is the only country in history to have a national debt of more than $20 trillion.\nText: I know this isn't exactly an accurate representation but it gives you some idea how much money we are talking about.\n\nhttp://en.wikipedia.org/wiki/National_debt#United_States
Running Text: I have been running for about a year and a half now but never really got into it.\n\n This past weekend was my first time doing it in the last few months.\n\n My goal is to run at least 3 miles every day this week.
Horror Text: I was a little girl when my parents got divorced. My dad had been in the military for years and he left me with my mom. She worked as an RN at a hospital so she could take care of me.\n\n When we moved to our new house it took some time before things settled down. We were still living together but there wasn't much going on. It didn't help that my mom would get mad if someone came over or even just walked by her house.\n\n One day while walking through the yard I noticed something out of place...
Reviews Rating: 5.0\n\n I have been using this product for a few years and it is the best thing on my hair. It makes your hair soft but not greasy.\n\nRating: 5.0\n\nI've used this product before and love it. It's light weight yet keeps my curls defined without making them crunchy!
Reviews Rating: 1.0\n\nI bought this for my daughter and it was a complete waste of money. The case is very flimsy and the screen protector doesn't fit properly.\n\nRating: 1.0\n\nThis product did not work at all. It would charge but when you unplug it, it does nothing. Very disappointed.

Table 2: With CTRL, no *prompt* is necessary as long as a control code is provided. Control codes can be combined (Reviews, Rating:, and VALUE) to provide finer-grained control.

- 생성을 더 제한 → 더 원하는 방향으로 생성
- domain control code 뒤에, 추가 control code 를 붙임 → 제어코드가 2개

Control code [More complex]

Links https://www.cnn.com/2007/09/20/us-president-meets-british-pm	LONDON, England (CNN) – U.S. President George W. Bush met with British Prime Minister Tony Blair on Monday to discuss the war in Iraq, according to a statement from Blair's office.\n\nThe meeting was held at 10 Downing Street and lasted about an hour.\n\n"The two leaders discussed Iraq and other international issues of mutual concern," said Blair.\n\nBush, who is scheduled to meet Wednesday with Russian President Vladimir Putin, will also visit Germany for talks later this week.\n\nIn his statement, Blair said, "We agreed that we should continue our efforts together to bring peace and stability to Iraq. We both reaffirmed our commitment to working closely together, as well as to continuing to work constructively toward achieving lasting security and prosperity throughout the Middle East region."\n\nBush's trip comes after he visited Britain last week where he spoke out against terrorism while visiting Buckingham Palace.\n\nHe has been criticized by some lawmakers over what they say are insufficient military resources being devoted to fighting terrorism.
Links https://www.cnn.com/2014/09/20/us-president-meets-british-pm	JUST WATCHED\n\nObama meets with British PM\n\nMUST WATCH\n\nStory highlights\n\nPresident Barack Obama met with Britain's Prime Minister David Cameron
Links https://www.cnn.com/2018/09/20/us-president-meets-british-pm	JUST WATCHED\n\nTrump and May meet for first time\n\nMUST WATCH\n\nWashington (CNN) President Donald Trump, who has been criticized by some in the UK over his decision to leave the European Union, met with British Prime Minister Theresa May, a White House official said on Thursday.
Links https://www.cnn.com/09/20/2018/star-spotted	JUST WATCHED\n\nStar spotted: The world's most elusive star\n\nMUST WATCH\n\nStory highlights\n\nA new study suggests that the Milky Way is home to a supermassive black hole\n\nIt may be one of our closest galactic neighbors\n\n(CNN) Astronomers have discovered what they believe could be an extremely rare astronomical event: An object so bright it can only be seen with powerful telescopes.\n\nThis object, dubbed SN2014J, was first detected by NASA's Spitzer Space Telescope in 2014. . . .
Links https://www.etonline.com/09/20/2018/star-spotted	\n\nEXCLUSIVE: Star Spotted in New York City\n\nBy Kelly Marot — September 20, 2018 10:51 AM PDT\n\nGetty Images\n\nThe star of the new season of Stranger Things is making a splash on her own turf and its not even Halloween.\n\nWinona Ryder has been seen out and about with friends at night time in Manhattan this week, E! News can exclusively reveal.\n\nA source tells us, She was hanging around with some friends, but she wasnt doing anything too crazy. She just wanted to get away from it all for a little bit.. . .
Links https://www.cnn.com/style/09/20/2018/george-clooney-interview	George Clooney on the future of his acting career\n\nBy\n\nUpdated 10:51 AM ET, Thu September 20, 2018\n\nChat with us in Facebook Messenger. Find out what's happening in the world as it unfolds.\n\nPhotos:George Clooney, 'Ocean's 8'\n\nActor George Clooney attends a photocall for "Ocean's 8" at Grauman's Chinese Theatre on August 31, 2018, in Los Angeles.\n\n. . .
Links https://www.cnn.com/politics/09/20/2018/george-clooney-interview	JUST WATCHED\n\nGeorge Clooney on the Trump administration\n\nMUST WATCH\n\n(CNN) Actor and activist George Clooney, who has been a vocal critic of President Donald Trump, said he is "ready to go back into the political arena" after his role in an anti-Trump documentary was cut from theaters this week.\n\n. . .

Table 3: CTRL is trained with links as control codes. Links provide a way to specify domain, subdomain, entities, entity relations, and even date. The links in these examples do not actually link to text; users can mimic the structure of the URLs that appear during training to create novel content during generation. Note that us-president is interpreted differently by the model depending on the date used (2007,2014,2018). Similarly, star is interpreted differently based on the domain (cnn vs. etonline) and topic (style vs. politics) can be varied even for identical entities (george-clooney).

- <표3> Advance 제어코드
- 입력 시작 부분에 URL 포함
  - 각 문서를 다운로드하는 데 사용되는 **URL**
- URL 구조와 이후 텍스트 간의 관계 학습

Control code [Triggering specific tasks]

Questions	*Q: What is the capital of India?* A: New Delhi\n\nQ: Which city in India has a large population of people from Gujarat?\n\nA: Ahmedabad\n\nQ: How many languages are spoken by Indians living abroad?\n\nA: more than 100...
Translation English :	*We release a new model for coherent language generation* ; French : Nous publions un nouveau modéle de génération cohérente du langage\n\nEnglish : The European Parliament has just adopted the report on the proposal for a regulation of the European Parliament and of the Council establishing an action programme to promote multilingualism in Europe. ; French : Le Parlement européen vient d'adopter le rapport sur la proposition de règlement du Parlement et du Conseil établissant un programme-cadre pour promouvoir le pluriling uisme dans l'Union européenne...

Table 4: More complex templatized control codes are used for task-specific generation.

- **소수의 제어 코드 → 특정 작업과 관련**
  - 질문 응답
  - 번역

Control code [zero-shot code-mixing]

A DATA SOURCES AND BREAKDOWN

Control Code	Description
Wikipedia	English Wikipedia
Books	Books from Project Gutenberg
Reviews	Amazon Reviews data (McAuley et al., 2015)
Links	OpenWebText (See Sec. 3.2)
Translation	WMT translation date (Barrault et al., 2019)
News	News articles from CNN/DailyMail Nallapati et al. (2016), New York Times and Newsroom (Grusky et al., 2018)
multilingual	Wikipedias in German, Spanish and French
Questions	(Questions and answers only) MRQA shared task (See Section 3.1)
Explain	(Only main post) (Fan et al., 2019)
	Sub-reddit data (Title, Text and Score/Karma) collected from `pushshift.io`.
Alone	`r/childfree`
Atheism	`r/atheism`
Christianity	`r/christianity`
Computing	`r/computing`
Confession	`r/offmychest`
Confessions	`r/confession`
Conspiracy	`r/conspiracy`
Diet	`r/keto`
Extract	`r/childfree`
Feminism	`r/twoxchromosome`
Finance	`r/personalfinance`
Fitness	`r/fitness`
Funny	`r/funny`
Gaming	`r/gaming`
Horror	`r/nosleep`
Human	`r/nfy`
India	`r/india`
Joke	`r/jokes`
Joker	`r/joke`
Learned	`r/todayilearned`
Legal	`r/legaladvice`
Movies	`r/movies`
Netflix	`r/netflix`
Norman	`r/lifeofnorman`
Notion	`r/unpopularopinion`
Opinion	`r/changemyview`
Politics	`r/politics`
Pregnancy	`r/babybumps`
Relationship	`r/relationshipadvice`
Relationships	`r/relationships`
Retail	`r/talesfromretail`
Running	`r/running`
Saving	`r/frugal`
Scary	`r/scaryshortstories`
Science	`r/science`
Technologies	`r/technology`
Teenage	`r/teenager`
Thoughts	`r/showerthoughts`
Tip	`r/lifeprotips`
Weight	`r/loseit`
Writing	`r/writingprompts`

Table 7: Data and control codes. Wikipedia, Books, News and multilingual have no secondary code. `Reviews` can be followed by `Rating:` and a value of {`1.0, 2.0, 3.0, 4.0, 5.0`}. For Links, a full or partial URL can be provided (See Table 3). For all the Reddit data, the secondary code can be `Title:` or `Text:`, which is the title and text of the article, respectively.

- 첫 번째 예
  - Diet - (r / keto)
  - 영어 및 독일어의 기계 번역 제어 코드 혼합
- 두 번째 예
  - Politics - (r / politics)
  - 프랑스어로 시작하는 프롬프트와 혼합
- 독일어와 프랑스어를 ZERO SHOT

`Diet English:` *I lost 10 kgs!* ; `German:` Ich habe 10 Kilogramm verloren!\n\nEnglish : It is a great achievement for me and my family. ; German : Das ist eine groe Leistung fr mich und meine Familie.\n\nEnglish : Thank you very much to all the people who have helped me. ; German : Vielen Dank an alle, die mir geholfen haben.. . .

`Politics Title:` *Les Etats-Unis sont un* pays de droite\n\nText: Je suis Américain et je ne peux pas supporter ce que font les USA en matière d'immigration\n\nJ'ai déjà écrit sur le sujet mais il n'y a aucune réaction `http://www.reddit.com/r/politics/comments/1tqjzw/les_etats_unes_son_tres-de-gauche/ ...`

Table 5: Some codes can be mixed to generate text with novel cross-over behavior. In Table 5, we present two examples. In the first example, we mix translation codes into the `Diet` domain. By doing so, the model continues alternatively generates English and German sentences while respecting the `Diet` domain and remains coherent across translations. In the second example, the `Politics` domain is mixed with a French prompt despite never seeing this combination in training.

데이터 셋

A DATA SOURCES AND BREAKDOWN

Control Code	Description
Wikipedia	English Wikipedia
Books	Books from Project Gutenberg
Reviews	Amazon Reviews data (McAuley et al., 2015)
Links	OpenWebText (See Sec. 3.2)
Translation	WMT translation date (Barrault et al., 2019)
News	News articles from CNN/DailyMail Nallapati et al. (2016), New York Times and Newsroom (Grusky et al., 2018)
multilingual	Wikipedias in German, Spanish and French
Questions	(Questions and answers only) MRQA shared task (See Section 3.1)
Explain	(Only main post) (Fan et al., 2019)
	Sub-reddit data (Title, Text and Score/Karma) collected from `pushshift.io`.
Alone	`r/childfree`
Atheism	`r/atheism`
Christianity	`r/christianity`
Computing	`r/computing`
Confession	`r/offmychest`
Confessions	`r/confession`
Conspiracy	`r/conspiracy`
Diet	`r/keto`
Extract	`r/childfree`
Feminism	`r/twoxchromosome`
Finance	`r/personalfinance`
Fitness	`r/fitness`
Funny	`r/funny`
Gaming	`r/gaming`
Horror	`r/nosleep`
Human	`r/nfy`
India	`r/india`
Joke	`r/jokes`
Joker	`r/joke`
Learned	`r/todayilearned`
Legal	`r/legaladvice`
Movies	`r/movies`
Netflix	`r/netflix`
Norman	`r/lifeofnorman`
Notion	`r/unpopularopinion`
Opinion	`r/changemyview`
Politics	`r/politics`
Pregnancy	`r/babybumps`
Relationship	`r/relationshipadvice`
Relationships	`r/relationships`
Retail	`r/talesfromretail`
Running	`r/running`
Saving	`r/frugal`
Scary	`r/scaryshortstories`
Science	`r/science`
Technologies	`r/technology`
Teenage	`r/teenager`
Thoughts	`r/showerthoughts`
Tip	`r/lifeprotips`
Weight	`r/loseit`
Writing	`r/writingprompts`

Table 7: Data and control codes. Wikipedia, Books, News and multilingual have no secondary code. `Reviews` can be followed by `Rating:` and a value of {`1.0, 2.0, 3.0, 4.0, 5.0`}. For Links, a full or partial URL can be provided (See Table 3). For all the Reddit data, the secondary code can be `Title:` or `Text:`, which is the title and text of the article, respectively.

- 다양한 도메인에서 140GB의 text drawing 을 학습
  - Wikipedia (En, De, Es, Fr),
  - Project Gutenberg1
  - submissions from 45 subreddits
  - OpenWebText2
  - a large collection of news data
  - Amazon Reviews
  - Europarl and UN data from WMT (En-De, En-Es, En-Fr)
  - question-answer pairs (no context documents) from ELI5
  - MRQA
    - Stanford Question Answering Dataset
    - NewsQA
    - TriviaQA
    - SearchQA
    - HotpotQA
    - Natural Questions

LANGUAGE MODELING WITH CTRL

$$p(x|c) = \prod_{i=1}^{n} p(x_i|x_{<i}, c) \qquad \mathcal{L}(D) = -\sum_{k=1}^{|D|} \log p_\theta(x_i^k|x_{<i}^k, c^k)$$

- c = control code
- 왼쪽 식
  - distribution p(x|c) → decompose → chain rule →
- 오른쪽 식
  - → chain rule → log → control code 에 대한 loss 가 train 에 포함

- C 가 어떻게 condition 되는지는 위의 여러 개 예시로 보았음

Transformer encoder [1번째 블록]

- Encoder 에 붙어있는 causal mask 사용
- future tokens 에 참석하지 못하게 함 → multi head 의 k 이용

$$\text{Attention}(X, Y, Z) = \text{softmax}\left(\frac{\text{mask}(XY^\top)}{\sqrt{d}}\right) Z$$

$$\text{MultiHead}(X, k) = [h_1; \cdots; h_k] W_o$$

$$\text{where } h_j = \text{Attention}(XW_j^1, XW_j^2, XW_j^3)$$

Transformer encoder [2번째 블록]

- 보다 먼저 layer normalization 를 수행

Block 1

$$\bar{X}_i = \text{LayerNorm}(X_i)$$
$$H_i = \text{MultiHead}(\bar{X}_i) + \bar{X}_i$$

Block 2

$$\bar{H}_i = \text{LayerNorm}(H_i)$$
$$X_{i+1} = \text{FF}(\bar{H}_i) + \bar{H}_i$$

Transformer encoder [출력]

$$\text{Scores}(X_0) = \text{LayerNorm}(X_l)W_{vocab}$$

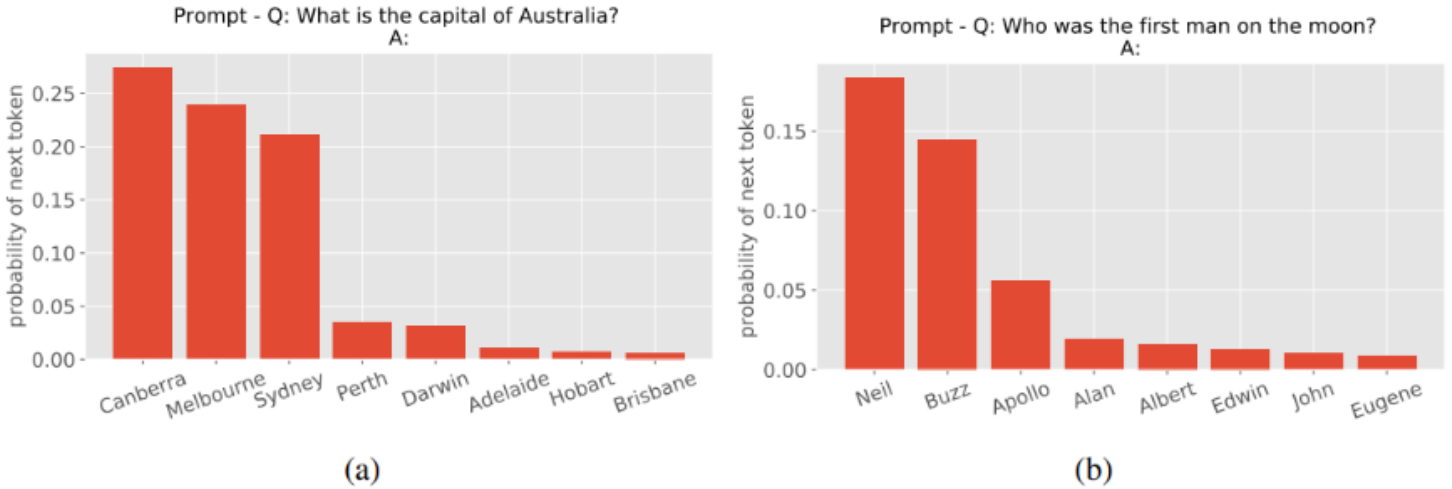
- 어휘의 각 토큰에 대한 score 는 마지막 레이어의 출력에서 계산
- Training 할때
  - score 값은 cross-entropy 의 입력 값
- Generation 할때
  - Score 값은 softmax 로 정규화 되어 새 토큰을 샘플링하기 위한 분포로 산출

Controllable Generation – Sampling

$$p_i = \frac{\exp(x_i/T)}{\sum_j \exp(x_j/T)}. \quad (1)$$

- temperature-controlled stochastic sampling methods
  - language model 로 generating text 할 때 사용
  - 오직 top k 로 샘플링을 제한
  - T > 0 (temperature = T)
  - x_i 는 d차원을 가진 i 번째 토큰의 score
  - p_i 는 i 번째 토큰 예측 확률
  - 다음 토큰은 top k 토큰에서 확률 pi가 클리핑된 다항분포를 통해 샘플링 하여 선택
- T 해석
  - T → 0은 탐욕 분포에 근사
  - T → ∞는 분포를 flat 시켜 분포를 uniform 하게 만듦
- Top k 에서 고정된 k 가 아닌 휴리스틱 k 를 제안 → **nucleus sampling**
  - 확률 임계값 p_t를 선택하고 k를 ∑ sort (p_i) > p_t가 되도록 하는 가장 낮은 값으로 설정
  - 다음 단어 예측에 확신이 있으면 k는 낮아짐, 확신이 없으면 k는 커짐
- 하지만 trade-off
  - 휴리스틱을 사용하는 모델의 generative capabilities 이 향상
  - There still exists a trade-off these parameters depending on generation intended

Given a prompt: Q: What is the capital of Australia?



- 정답 Canberra 에 높은 확률 질량을 할당
- Melbourne, Sydney, Brisbane, Darwin, Perth와 같은 다른 도시에는 0이 아닌 확률 질량을 할당
- 애매한 확률 질량에 불신, 하지만 정확한 다음 토큰을 예측
- 자연스럽게 (보통 우리가 해오던) greedy 하게 추출하면 **같은 문장, 문구 반복적 생성 문제 발생**

penalized sampling

$$p_i = \frac{\exp(x_i/(T \cdot I(i \in g))}{\sum_j \exp(x_j/(T \cdot I(j \in g))} \qquad I(c) = \theta \text{ if c is True else } 1$$

- 패널티를 통한 반복되는 문구, 문장을 방지하는 새로운 샘플링 체계를 제안
- 이전에 생성된 토큰의 점수를 할인하여 작동
- 방정식 1의 표기법을 사용
- 생성된 토큰 g의 목록이 주어지면 다음 토큰에 대한 확률 분포 pi

- greedy sampling 과 θ ≈ 1.2를 사용하면 good balance between truthful generation and lack of repetition
- θ = 1은 방정식 1과 같음
- 이 접근법은 모델이 충분히 신뢰할 수 있는 분포를 학습한 경우에만 성공한다는 점에 유의

Source Attribution

$$p_\theta(c|x) \propto p_\theta(x|c)p(c)$$

- 도메인 제어 코드는 훈련 데이터를 상호 배타적인 세트로 분할하는데 사용가능
- 언어 모델이 분포 pθ (x | c)를 배웠음을 상기
- p(c)에 대한 사전분포 도메인 제어코드를 지정하면 도메인 순위(ranking)를 계산하는 것이 간단

- 사전분포가, 많은 데이터가 있는 도메인에 높은 가중치를 부여하는 것을 발견
- p(c) 사전분포에 control code 대신 uniform prior 를 사용

Source Attribution

- p(c) 사전분포에 control code 대신 uniform prior 사용한 예

Query Prompt	Attributed Sources
Global warming is a lie.	r/unpopularopinion, r/conspiracy, r/science
Global warming is a lie	r/eli5, r/science, r/unpopularopinion
Global warming is a real phenomenon	r/eli5, r/science, r/changemyview
Global warming is a real phenomenon.	OpenWebText, r/changemyview, r/science
I don't think women should be allowed to vote.	r/christianity, r/atheism, r/unpopularopinion
Carbs are your enemy when you want to get lean.	r/fitness, r/loseit, r/keto
I just want to be a fun aunt. I'm not interested in babies.	r/babybumps, r/childfree, r/twoxchromosome
My landlord is suing me for unpaid rent.	r/legaladvice, r/personalfinance, r/frugal
FROM fairest creatures we desire increase,\n\nThat thereby beauty's rose might never die	Gutenberg, Wikipedia, OpenWebText

Table 6: We probe CTRL for learned correlations between sequences and domains. Note that this procedure is sensitive to small changes in the prompt. For example, "Global warming is a lie" differs from "Global warming is a lie." r/eli5 stands for "Explain like I'm five". Attribution experiments use the model trained on sequences of length 256; it was trained longer and provided better estimation of source. Source attribution cannot be considered a measure of veracity, but only a measure of how much each domain token influences a given sequence.

Source Attribution

- 사용한 the 데이터 → 보편성을 띄지 않음 → 원래 출처에 존재하는 문화적 연관성을 띔 → 예측을 할 땐 기존 연관성 (original associations) 에 의존 → 정리하면, 언어 모델과 학습 데이터 간의 관계를 설정하기 위해 original associations 을 이용

- CTRL 은 문화적 연관성과 도메인 간의 상관 관계만 배움 -> 많은 양의 텍스트에서 상관 관계를 분석하기위한 설명 도구로 사용할 수 있음