

“Language modeling via stochastic
processes” &
“On Homophony and Rényi Entropy”

持橋大地

統計数理研究所 数理・推論研究系

daichi@ism.ac.jp

最先端NLP2022

2022-9-26 (月)

1本目

- “Language Modeling via Stochastic Processes”, Rose E. Wang, Esin Durmus, Noah Goodman, Tatsunori B. Hashimoto, ICLR 2022.
- 選んだ理由：
OpenReviewでこの論文の査読者だった (採択を推薦した)
- ICLR 2022のOralだったようです
- 本人達の発表:
<https://www.youtube.com/watch?v=AwnoASlXels>

3行まとめ

- 言語モデルからのランダム生成は、長い文章になると支離滅裂
- 「始め」と「終わり」を指定して、潜在的なブラウン橋 (Brownian bridge) から生成
- 人手および自動評価により、より consistent な文章が生成できることを確認した

GPT-2からの長文生成

[USER] I'm interested in buying some tickets for a movie tonight. [ASSISTANT] OK. Can you tell me the city and state where the theater is located? [USER] It's been so long since we've been to the movies that I want to get out. It can't get any better. [ASSISTANT] Great. Can you tell me the city and state where you plan to see the movie? [USER] I'll be in Santa Barbara, California at the moment. [ASSISTANT] Got it. Are you interested in a particular movie? [USER] There's a couple of nice theaters in town. [ASSISTANT] OK. Is there a movie you have in mind? [USER] I'm thinking action. [ASSISTANT] Ok. Let's see what action films are out in that area. [ASSISTANT] I found one action movie at AMC Mountain 16: No Time To Die. Are you interested in that one? [USER] My best friend Sally has been wanting to

Can I help with anything else? [USER] No. [ASSISTANT] OK. Enjoy your movie! As always, have a good day. **javascript ive been! havn bout ive heard it. dont think to much. in there. lmao. id appreciate it. reccommended. The email address for your tickets is theater. There's no problem at all with anything else. thanks! sign off. Your tickets are confirmed and they've been sent to your email address. thanks a lot. your welcome. enjoy your movie. -lbye. I'm very, very sorry for leaving you behind. Thanks for calling. Is there anything else I can help with? dont remember the last time we went out. I'm very sorry. Thank you very much. arn. Thanks for all of your help. Enjoy the movie. www, bye. -en Chinese girlfriend, bye. can you enjoy your movie. Bye. remaining showtimes for No Time To Die at AMC Mountain 16 are 4:30pm, 6:40pm and 9:10pm. This is all for right now. N/A Thanks for all of your assistance. @ @ b/r/r/r/r abys mam. @ cornyyyyyyyyyyyyy. BULLY BULLY MOVIE FOR YOU! Thank you for calling. Enjoy your movie! Your movie. Aww, bye. happy movie. thanks for calling. -en . Thanks for using our services. > .**

ブラウン橋からの長文生成

[USER] Hey! I wanna go tonight to watch a movie! [ASSISTANT] Hi, what movie would you like to watch? [USER] I want to know about Birds of Prey! [ASSISTANT] Sure, is there any particular theater you would like to watch Birds of Prey at? [USER] Can you tell me what theaters are showing the movie? [ASSISTANT] What are you looking at? [USER] Can you tell me about what theaters are showing Birds of Prey at? [ASSISTANT] Are you interested in one theater? [USER] What theaters are showing Birds of Prey at? [ASSISTANT] Okay, what movie theatre are you going to? [USER] I want to go to AMC. [ASSISTANT] Very well, what time would you like to watch the movie? [USER] What is the movie playing tonight at 1:00pm and 3:00 pm? [ASSISTANT] Can you look up times for Birds of Prey at AMC theaters? [USER] Can you tell me about the time and the theater where Birds of Prey is playing? [ASSISTANT] Sure,

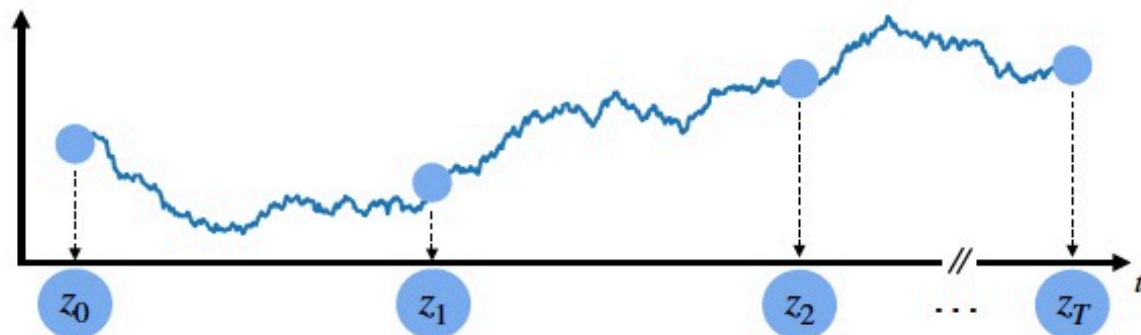
.

Not a problem, can you look up tickets to that showing instead instead? [USER] Can you think of anything else? [ASSISTANT] Sure, I'll check out the tickets at AMC showing time showing in 3D as well, if there is anything else? [USER] Yea, cool, I just need to get those booked and I'll get more information to help buy those tickets! [ASSISTANT] Awesome, I have just booked three tickets for Dolittle showing in 3D at AMC showing time in 3D at 10:20 pm tonight, does that work for you today? [USER] I want to upgrade to 3D, and then use my Apple Pay. [ASSISTANT] Okay, I'm turning this over and getting those tickets online now. Back to you, I want to purchase your movie tickets to purchase the movie tickets tonight at 10:20 pm at AMC showing in 3D. Correct? [USER] Sure, so I want to purchase a movie ticket for the movie showing in 3D instead of 3D instead, and then use my Apple Pay. [ASSISTANT] Okay, I will now see if you can purchase the movie tickets to purchase your movie tickets to purchase the movie tickets tonight at 10:20 pm at AMC showing in 3D. Is there anything else you need help with? [USER] Not really, not really. Thanks!

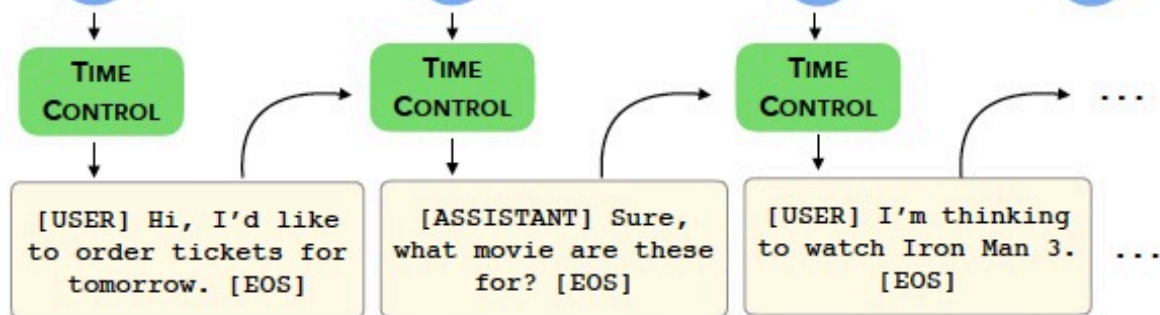
中でやっていること

ブラウン橋 (Brownian bridge) という

① Generate latent plan pinned at z_0 and z_T .



② Conditionally generate each sentence using the latent plan.



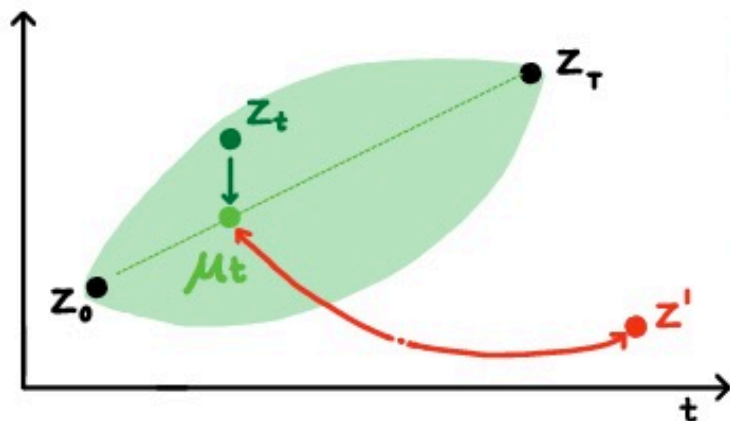
- “始点”と“終点”を決めて、その間を結ぶようにブラウン運動 (ランダムウォーク) する
- これは潜在空間での軌跡なので、デコーダーで言語表現にデコードする

数学的表現と学習

- 始点 z_0 と終点 z_T が決まった上での時刻 $0 < t < T$ での潜在状態の確率分布

$$p(z_t | z_0, z_T) = \mathcal{N}\left(\left(1 - \frac{t}{T}\right)z_0 + \frac{t}{T}z_T, \frac{t(T-t)}{T}\right).$$

- イメージ



x_0 : [USER] Hello, I'd like to buy t

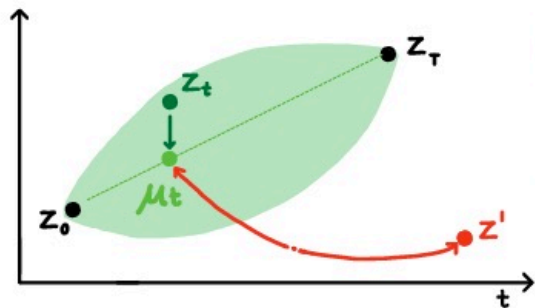
x_t : [ASSISTANT] What movie theater

x_T : [USER] Could you confirm my tic

x' : [USER] Hi, I'm looking to purch

$$\mathcal{L} = -\log \frac{\exp(d(z_t, \mu_t))}{\exp(d(z_t, \mu_t)) + \exp(d(z_t, z'))}$$

数学的表現と学習 (2)



`x_0`: [USER] Hello, I'd like to buy tickets for tomorrow.

`x_t`: [ASSISTANT] What movie theater do you prefer?

`x_T`: [USER] Could you confirm my tickets just in case?

`x'`: [USER] Hi, I'm looking to purchase tickets for my family.

$$\mathcal{L} = -\log \frac{\exp(d(z_t, \mu_t))}{\exp(d(z_t, \mu_t)) + \exp(d(z', \mu_t))}$$

- ランダムにサンプリングした3つ組 $(x_0, x_{t'}, x_T)$ よりも、観測された (x_0, x_t, x_T) の確率が高くなるようにエンコーダー/デコーダーを学習

$$\text{Maximize } \mathcal{L}_N = \mathbb{E}_X \left[-\log \frac{\exp(d(x_0, x_t, x_T; f_\theta))}{\sum_{(x_0, x_{t'}, x_T) \in \mathcal{B}} \exp(d(x_0, x_{t'}, x_T; f_\theta))} \right], \text{ where}$$

$$d(x_0, x_t, x_T; f_\theta) = -\frac{1}{2\sigma^2} \left\| \underbrace{f_\theta(x_t)}_{z_t} - \underbrace{\left(\left(1 - \frac{t}{T}\right) f_\theta(x_0) - \frac{t}{T} f_\theta(x_T) \right)}_{\text{mean in Equation 1}} \right\|_2^2$$

生成された例の比較

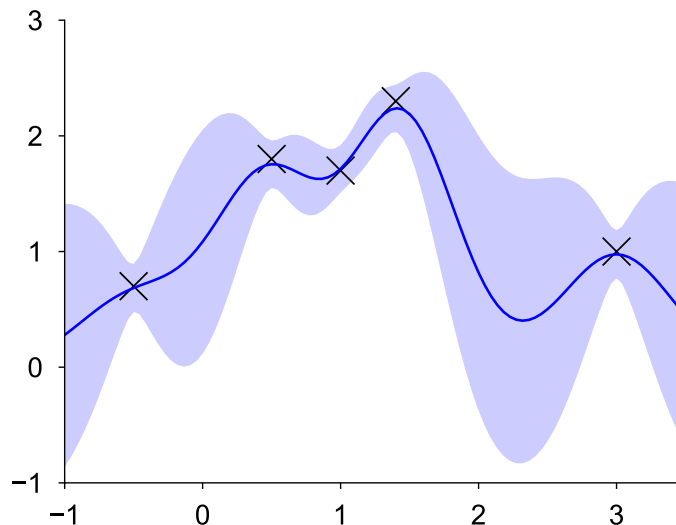
Time Control : [USER] I'm looking for movie tickets please. [ASSISTANT] OK, where would you like to see the movie? [USER] I'd like to see it at Creek's End, Oregon please. [...] [ASSISTANT] Is it OK to go ahead and purchase these tickets? [USER] Yeah, that would be great. [ASSISTANT] OK. I understand that you are interested in tickets for Emma at AMC Mercado 24 tonight at 7:10pm. Is that OK? [USER] Yes, please do that. [ASSISTANT] OK

GPT2: [USER] Hi! Tonight, I'd like to go to the movies. [ASSISTANT] Okay. What theater would you like to go to? [USER] Center City. [...] [ASSISTANT] That will be all for now. Thank you for all your help. N/A [USER] Bye Bye. [ASSISTANT] N/A [ASSISTANT] N/A N/A N/A N/A N/A N/A N/A N/A [USER] N/A [ASSISTANT] N/A N/A N/A [USER] N/A [ASSISTANT] N/A N/A N/A [USER] N/A N/A N/A N/A N/A N/A N/A N/A N/A N/A [USER] N/A N/A N/A N/A [...]

Table 8: Example of forced long text generation on TicketTalk with Time Control vs. fine-tuned GPT2. Both models are forced to extrapolate when generating long texts. They start coherently, but only Time Control extrapolates coherently. For space reasons, some of the text has been removed, marked with [...].

議論

- (査読の際にも指摘しましたが) 始点と終点だけを指定するモデルなので、柔軟性には欠ける
- 文章の「終点」は単なる挨拶や全体のまとめだったりすることもあり、「終点」が内容的に重要とは限らない
- 途中の「チェックポイント」も経由するような条件付き生成を考えると、結局ガウス過程回帰になるのでは？



2本目

- “On Homophony and Rényi Entropy”,
Tiago Pimentel, ..., Ryan Cottrell, EMNLP 2021.
- 選んだ理由：Ryan Cottrellらしい、面白い計算言語学
の研究
 - ニューラル時代で言語処理が柔軟になり、こうした話が
増えてきた
- レニーエントロピーの話は一種の後付けなので、
恐れる必要はありません

実験コード:

<https://github.com/rycolab/homophony-as-renyi-entropy>

3行まとめ

- 同音異義語 (Homophony) が、言語でどれくらい起こりうるか?
- Trott and Bergen (2020)ではあまりにナイーブな言語モデルを仮定していたため、強すぎる結論
→ LSTM言語モデルで再評価すると、そうでもない
- 使った指標は、結果的に情報理論で知られているレニーエントロピーと同じになった

同音異義語 (homonym, homophony)

- knight <-> night
- add <-> ad
- ate <-> eight
- bear <-> bare
- week <-> weak

など、発音は同じだが単語としては異なる単語

- 言語はこうした同音異義語を積極的に利用することで、コミュニケーション効率を高めているという研究がある (Piantadosi+ 2012)

問題意識

- 言語に同音異義語があるのは、単なる偶然なのか？
- それとも、
 - 同音異義語を積極的に使うバイアスがあるのか？
(Piantadosi+ 2012)
 - 同音異義語を使わないようにするバイアスがあるのか？
(Trott and Bergen 2020)

方法

- 発音列 \mathbf{w} の単語が確率 $p(\mathbf{w})$ で生成されるとき、ランダムな2個が一致する確率は $p(\mathbf{w})^2$
- よって、衝突の起こる情報量は

$$H_2(p) = -\log \sum_{\mathbf{w} \in \mathcal{W}} p(\mathbf{w})^2$$

- これを、情報理論ではレニーエントロピー (Rényi Entropy) と呼んでいる
- 下記の一般的なレニーエントロピーで $\alpha=2$ の場合

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \sum_{\mathbf{w} \in \mathcal{W}} p(\mathbf{w})^\alpha$$

方法 (2)

- 注意：

$$H_2(p) = -\log \sum_{\mathbf{w} \in \mathcal{W}} p(\mathbf{w})^2$$

は綴りの確率で、単語自体の確率を考えていないが、
確率の高い綴りの単語は高頻度で使われる
という研究がある (Mahowald+ 2018, *Cognition*)

実際と同音異義語率

- M個の語彙 $\tilde{\mathbf{W}}$ で実際に同音異義語が生じている確率(情報量)は、

$$R(\tilde{\mathbf{W}}) = -\log \frac{\sum_{m \neq m'}^M \mathbb{I}[\tilde{\mathbf{w}}^{(m)} = \tilde{\mathbf{w}}^{(m')}]}{M(M-1)}$$

- これを、標本レニーエントロピーと呼ぶことにする
- 言語の確率モデル p から得られる同音異義語の確率と、実際の確率はどう異なるか?

$$H_2(p) = R(\tilde{\mathbf{W}})$$

になるか?

技術的な話

$$H_2(p) = -\log \sum_{\mathbf{w} \in \mathcal{W}} p(\mathbf{w})^2$$

- を計算するとき、無限の語彙すべてについての和は実際には取れない
- 確率 δ 以上の単語 \mathbf{w} だけの和を取ることになると、上式のバウンドが得られる

$$H_2(p) \leq \hat{H}_2(p) \leq H_2(p) + \log \left(1 + \frac{(1 - \xi) \delta}{\eta} \right)$$

– ただし、 $\xi = \sum_{\mathbf{w} \in \mathcal{W}_\delta} p(\mathbf{w})$ 、 $\eta = \sum_{\mathbf{w} \in \mathcal{W}_\delta} p(\mathbf{w})^2$

実験方法

- 音素綴りの言語モデル (nグラムまたはLSTM)から単語をランダムに生成し、仮想語彙 \mathbf{W} を作る
- 実際の語彙 $\tilde{\mathbf{W}}$ と両方で、標本レニーエントロピー (=衝突確率)を計算して大小を比較

$$p(R(\tilde{\mathbf{W}})) < p(R(\mathbf{W})) ?$$

- 両辺が等しいかどうかについての検定を行う
- (サンプルサイズに左右されるのでは?)

実験結果

		Cross-entropy				
		Train	Test	$H_1(p)$	$H_2(p)$	$R(\mathbf{W})$
English	<i>n</i> -gram	13.61	28.10	30.45	13.89	13.90*
	LSTM	18.75	19.89	26.46	14.77	14.77*
	Lexicon	-	-	-	-	15.02
German	<i>n</i> -gram	14.08	29.25	30.44	14.27	14.27*
	LSTM	20.26	21.35	27.74	15.87	15.88
	Lexicon	-	-	-	-	15.67
Dutch	<i>n</i> -gram	13.89	26.08	30.45	14.07	14.06*
	LSTM	18.37	18.94	26.81	15.16	15.16*
	Lexicon	-	-	-	-	14.60

*Statistically different from lexicon's Rényi entropy ($p < 0.01$).

- CELEXデータセット (英米蘭) の語彙で実験
- 綴りのモデルをLSTMにすると、Trott&Bergen(2020)と異なる結果 (T&Bでは均一な0.1のラプラス平滑化!)

実験結果 (2)

- 言語モデルをLSTMにすると、実際の衝突率と近い結果→言語は同音異義語を促進しているとも、抑制しているとも言えない
- 適当な平滑化ではなく、正確な確率モデルを使うことが重要

		Cross-entropy				
		Train	Test	$H_1(p)$	$H_2(p)$	$R(\mathbf{W})$
English	<i>n</i> -gram	13.61	28.10	30.45	13.89	13.90*
	LSTM	18.75	19.89	26.46	14.77	14.77*
	Lexicon	-	-	-	-	15.02
German	<i>n</i> -gram	14.08	29.25	30.44	14.27	14.27*
	LSTM	20.26	21.35	27.74	15.87	15.88
	Lexicon	-	-	-	-	15.67
Dutch	<i>n</i> -gram	13.89	26.08	30.45	14.07	14.06*
	LSTM	18.37	18.94	26.81	15.16	15.16*
	Lexicon	-	-	-	-	14.60

* Statistically different from lexicon's Rényi entropy ($p < 0.01$).

議論

- 論文では実験していないが、Kneser-Ney nグラムでも Trott&Bergen (2020)よりはかなり良くなるはず
 - T&B(2020)は*Cognition*の論文
 - NLPの人がきちんと関わることが大事
- 綴りの話と、音素の話が混同されているのでは？
- 検定の結果はサンプル数に依存するので、若干疑問