Decoding Strategies

Background

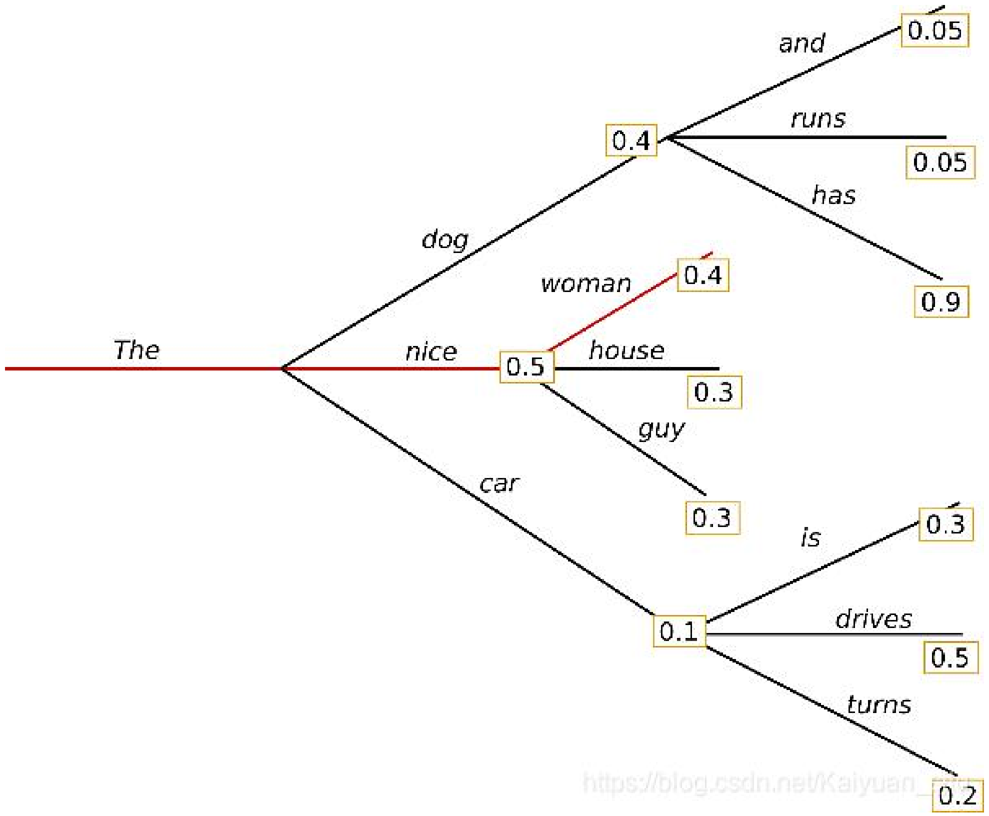
- Directed Generation
- Open—ended Generation

- Language Model $p_\theta(\mathbf{x}) = \prod_{t=1}^{|\mathbf{x}|} p_\theta(x_t|x_{<t})$

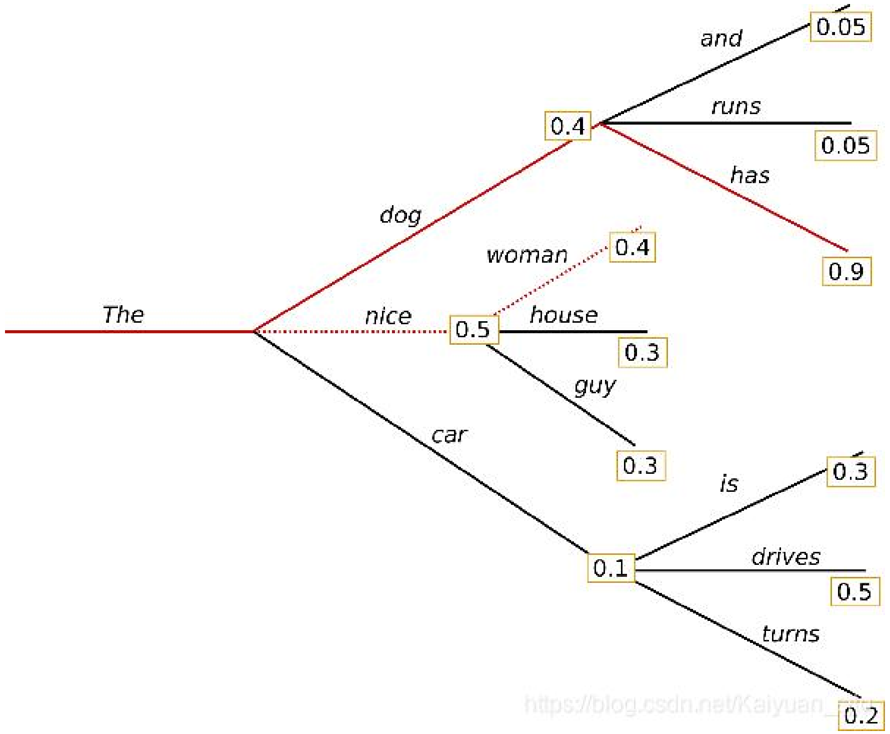
- Deterministic Decoding $x_t = \arg\max p_\theta(x_t|x_{<t})$
- Stochastic Decoding $x_t \sim q(x_t|x_{<t}, p_\theta)$

Greedy Search

$$x_t = \arg\max p_\theta(x_t|x_{<t})$$



Beam Search



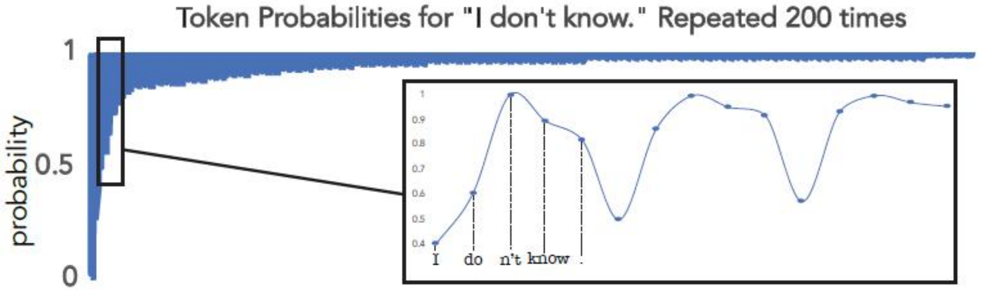
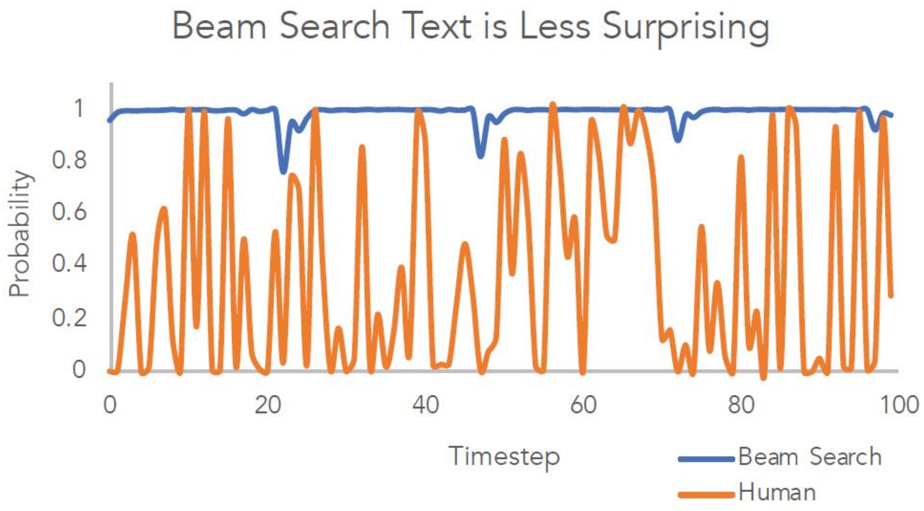


Figure 4: The probability of a repeated phrase increases with each repetition, creating a positive feedback loop. We found this effect to hold for the vast majority of phrases we tested, regardless of phrase length or if the phrases were sampled randomly rather than taken from human text.



Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

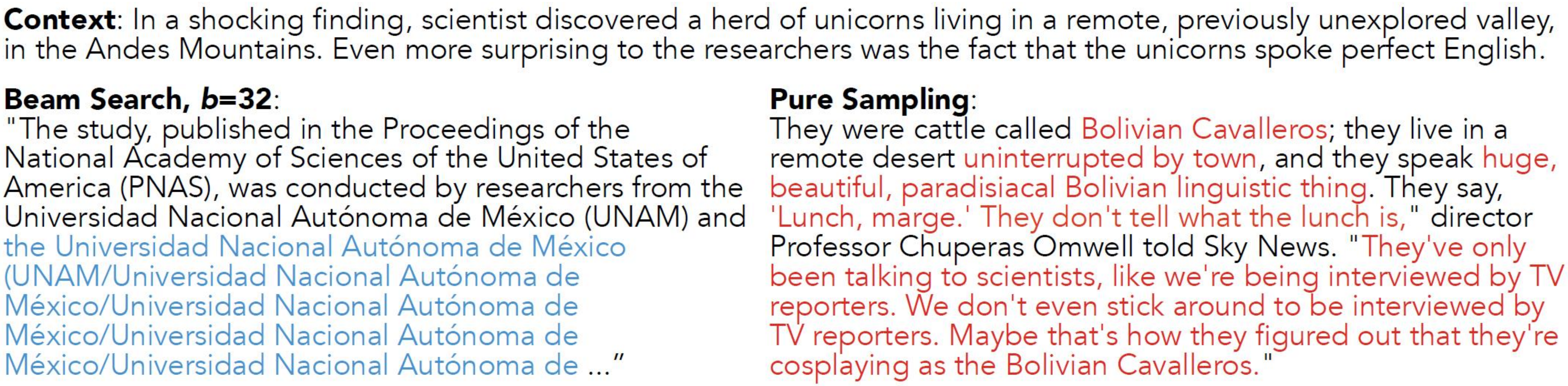
Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Pure Sampling

- 从整个词典分布中随机采样

$$x_t \sim q(x_t | x_{<t}, p_\theta)$$



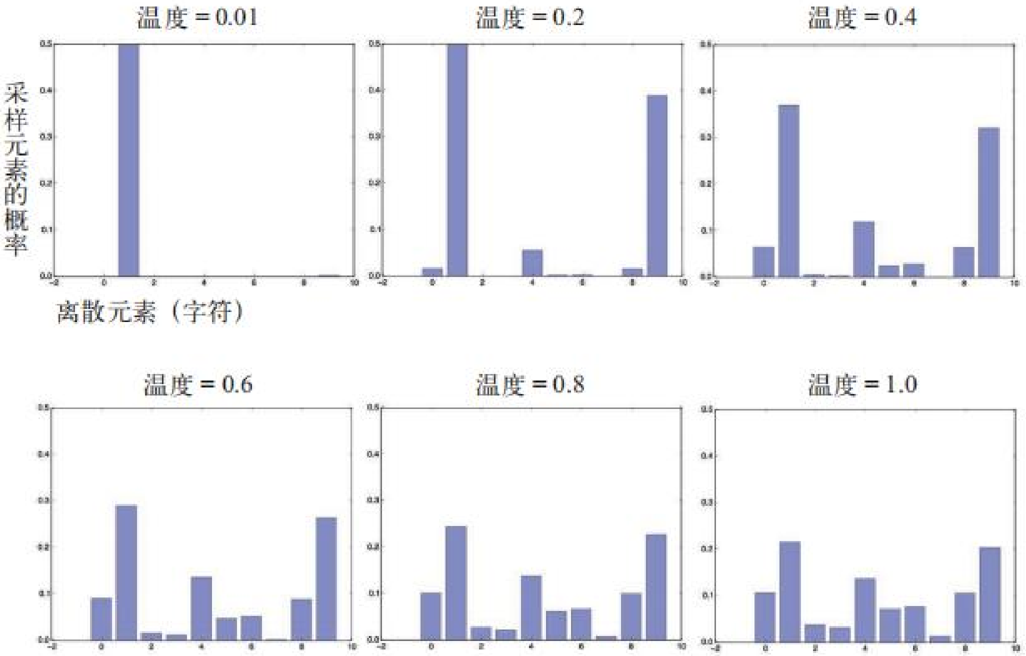
Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

Sampling with Temperature

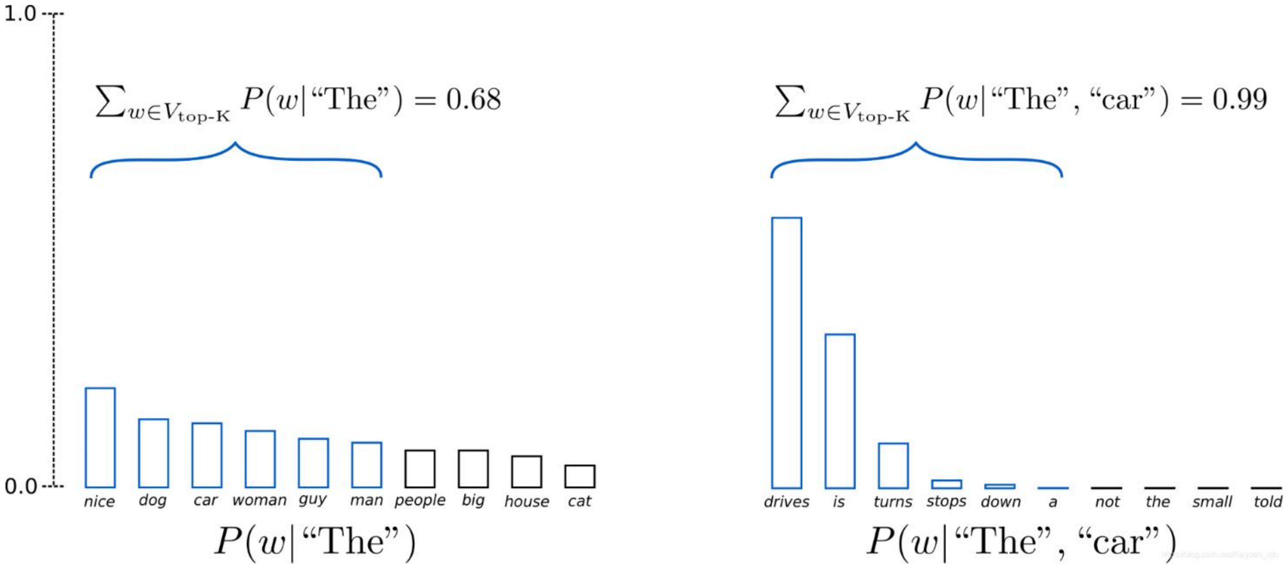
$$p(x = V_l | x_{1:i-1}) = \frac{\exp(u_l/t)}{\sum_{l'} \exp(u'_l/t)}$$



Top-k Sampling

- Top-k vocabulary $V^{(k)} \subset V$
- 最大化 $\sum_{x\in V^{(k)}} P(x|x_{1:i-1})$ 的大小为k的集合

$$P'(x|x_{1:i-1}) = \begin{cases} P(x|x_{1:i-1})/p' & \text{if } x \in V^{(p)} \\ 0 & \text{otherwise.} \end{cases}$$



存在的问题

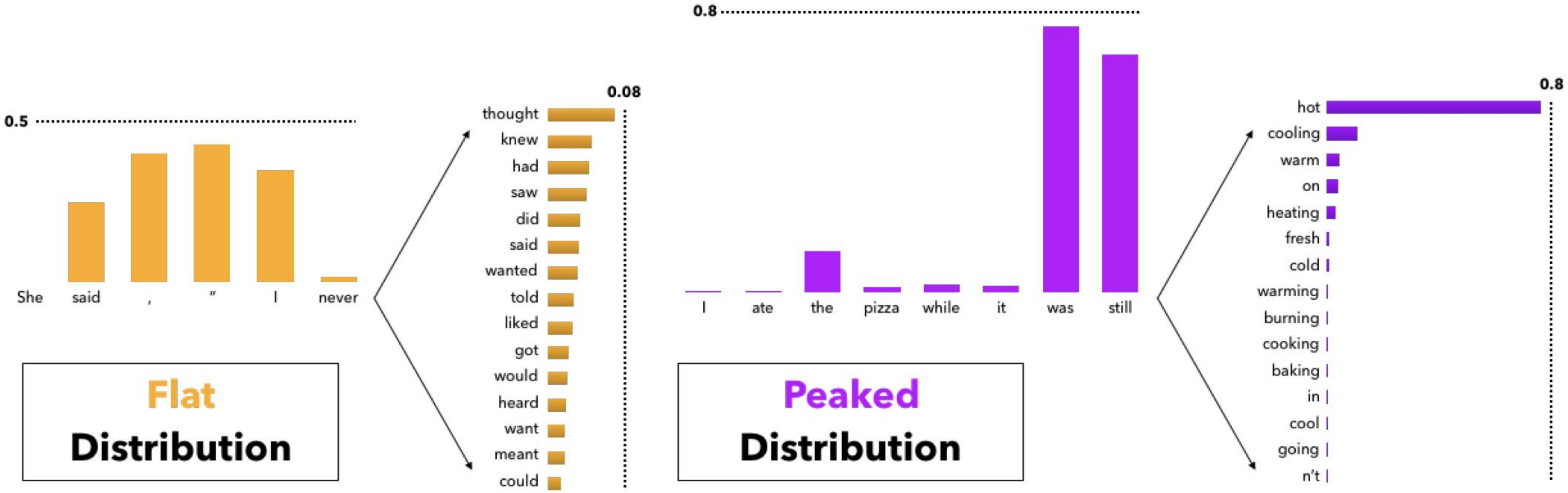


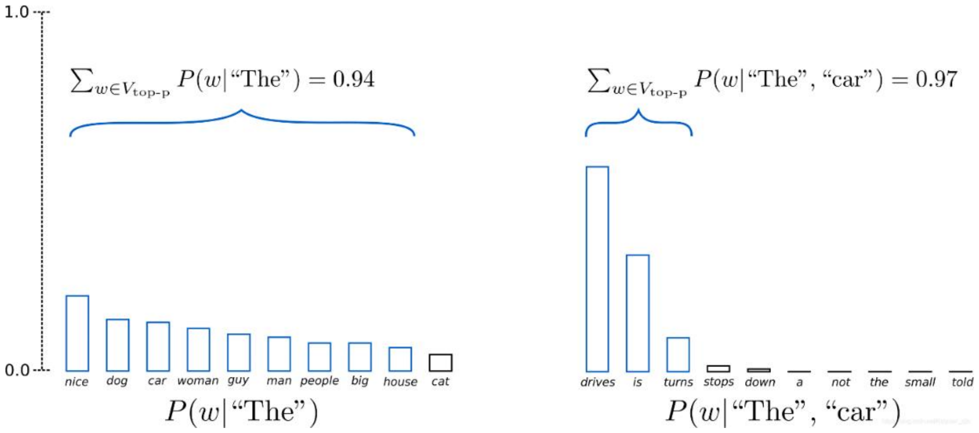
Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

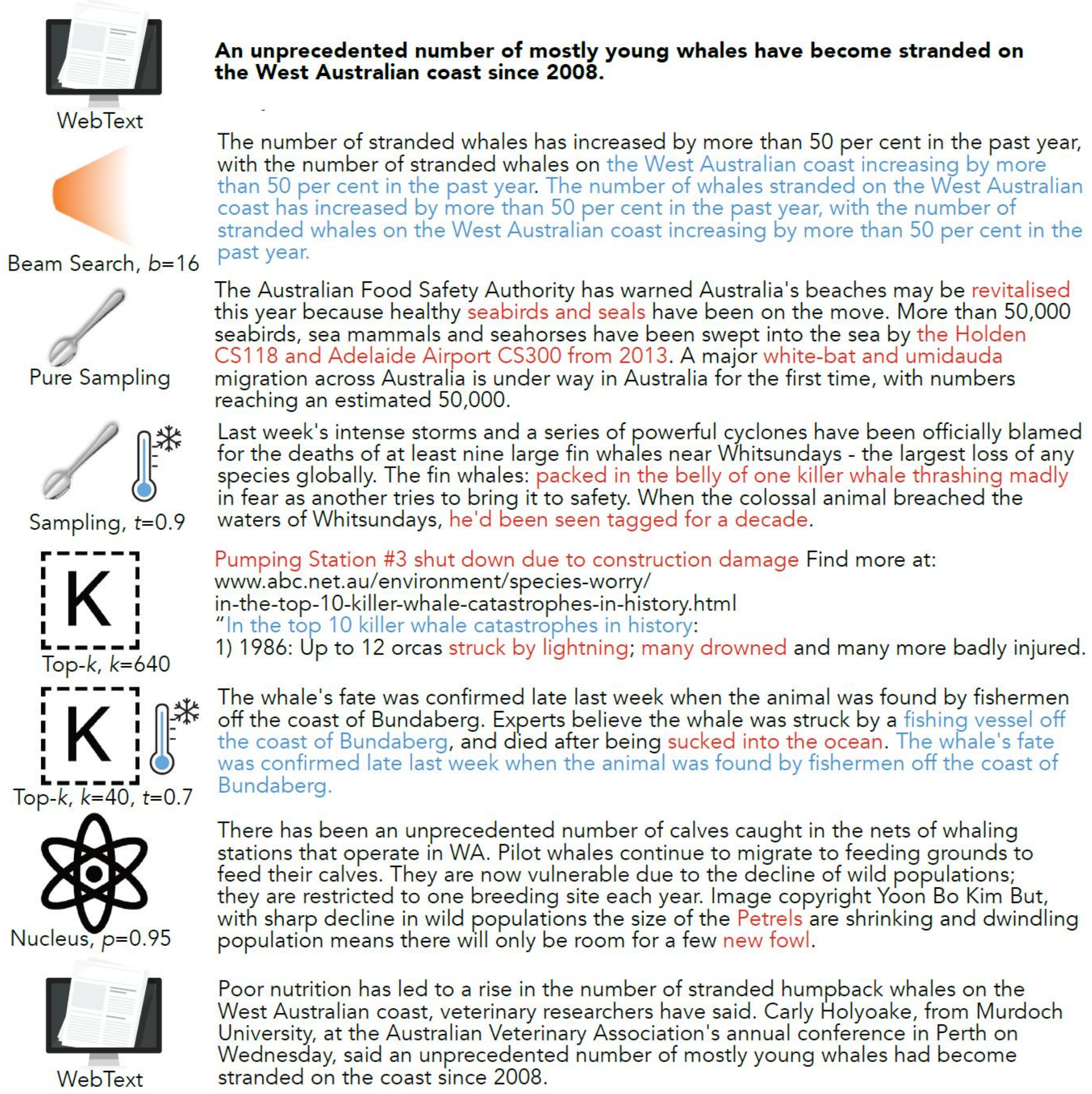
Nucleus Sampling

- 绝大多数的概率集中在整个词典的一个小子集（核）中，核的大小可以从一个到成百上千不等。
- 设定一个概率阈值p ，将现有词概率由大到小排序，取累积概率值达到p的前N个词，重新计算softmax，再从中采样。

$$\sum_{x\in V^{(p)}} P(x|x_{1:i-1}) \geq p$$

- 对比Top–k采样策略，Top–p限制了生成低频词带来的语句错误





An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Pumping Station #3 shut down due to construction damage Find more at: www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html
“In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

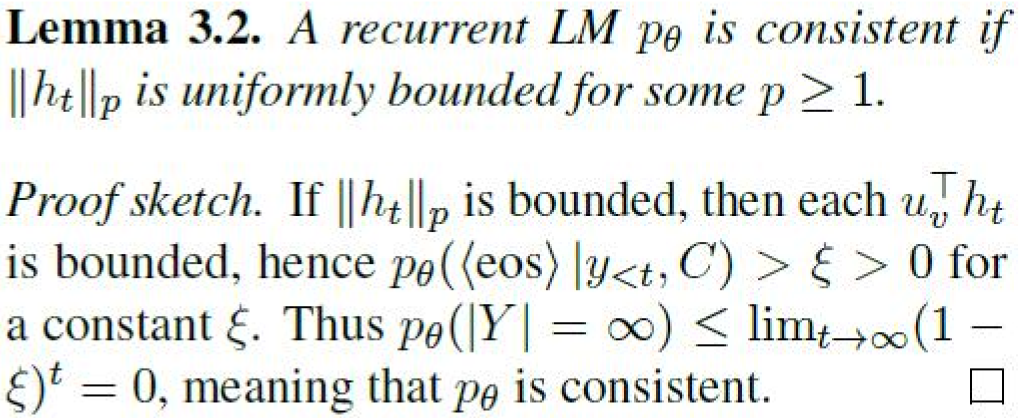
There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Figure 3: Example generations continuing an initial sentence. Maximization and top-k truncation methods lead to copious repetition (highlighted in blue), while sampling with and without temperature tends to lead to incoherence (highlighted in red). Nucleus Sampling largely avoids both issues.

Consistent Sampling Algorithms

- 有界隐向量：



Lemma 3.2. *A recurrent LM p_θ is consistent if $\|h_t\|_p$ is uniformly bounded for some $p \geq 1$.*

Proof sketch. If $\|h_t\|_p$ is bounded, then each $u_v^\top h_t$ is bounded, hence $p_\theta(\langle \text{eos} \rangle \,|y_{<t}, C) > \xi > 0$ for a constant ξ. Thus $p_\theta(|Y| = \infty) \leq \lim_{t\to\infty}(1 - \xi)^t = 0$, meaning that p_θ is consistent. □

- Consistent Sampling Algorithms
  - Top-k sampling

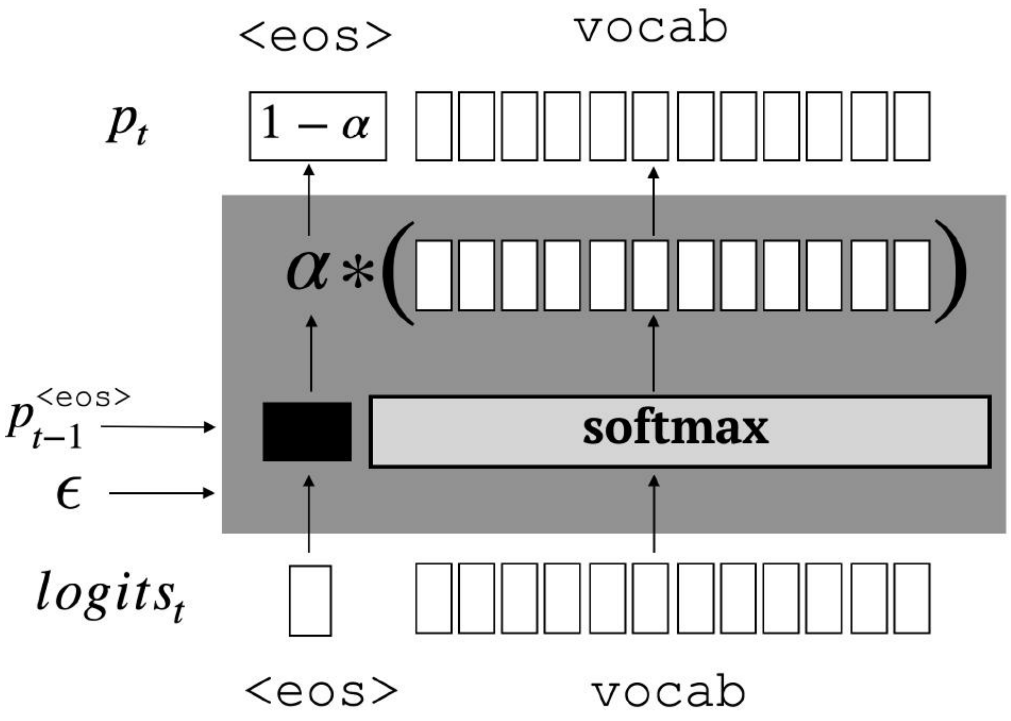
$$q(v) \propto \begin{cases} p_\theta(v|y_{<t}, C), & \textit{if } v \in V', \\ 0, & \textit{otherwise}, \end{cases}$$

$$\textit{where } V' = \{\langle eos \rangle\} \cup \arg \underset{v'}{\textit{top-k}}\, p_\theta(v' \,|\, y_{<t}, C)$$

  - Nucleus sampling

$$q(v) \propto \begin{cases} p_\theta(v \,|\, y_{<t}, C), & \textit{if } v \in V_\mu \cup \{\langle eos \rangle\} \\ 0, & \textit{otherwise}. \end{cases}$$

Self-Terminating Recurrent LM



$$p_\theta(v \mid y_{<t}, C) = \begin{cases} 1-\alpha(h_t), & v = \langle eos \rangle, \\ \frac{\alpha(h_t)\exp(u_v^\top h_t + c_v)}{\sum_{v' \in V'} \exp(u_{v'}^\top h_t + c_{v'})}, & \end{cases}$$

$$\alpha(h_0) = \sigma(u_{\langle eos \rangle}^\top h_0),$$

$$\alpha(h_t) = \sigma(u_{\langle eos \rangle}^\top h_t)\left[1 - p_\theta(\langle eos \rangle \mid y_{<t-1}, C)\right],$$

$$p_t^{\langle \text{eos} \rangle} = 1 - \prod_{t'=0}^{t} \sigma(u_{\langle \text{eos} \rangle}^\top h_{t'})$$

Unlikelihood Loss

- Unlikelihood $\mathcal{L}_{\mathrm{UL}}^t(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = -\sum_{c\in\mathcal{C}^t} \log(1 - p_\theta(c|x_{<t})).$

- Token−level

$$\mathcal{L}_{\text{UL-token}}^t(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = \underbrace{-\alpha \cdot \sum_{c\in\mathcal{C}^t} \log(1 - p_\theta(c|x_{<t}))}_{\mathcal{C}^t_{\text{prev-context}} = \{x_1,\ldots,x_{t-1}\} \setminus \{x_t\}} - \underbrace{\log p_\theta(x_t|x_{<t})}_{\text{likelihood}}.$$

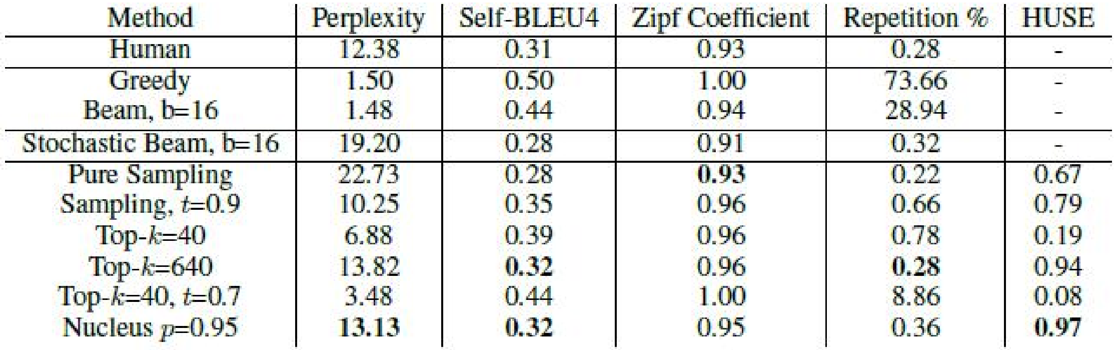
- Sequence−level

$$\mathcal{L}_{\mathrm{ULS}}^t(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = -\sum_{c\in\mathcal{C}^t} \log(1 - p_\theta(c|x_{<t})),$$

$$\mathcal{C}^t_{\text{repeat-n}} = \{x_t\} \text{ if } (x_{t-i},\ldots,x_t,\ldots,x_{t+j}) \in x_{<t-i} \text{ for any } (j-i) = n, i \le n \le j,$$

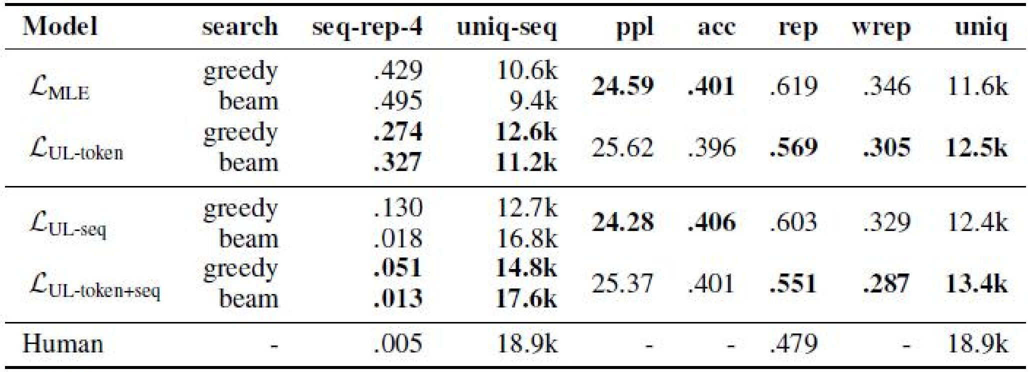
Experiment

- Nucleus



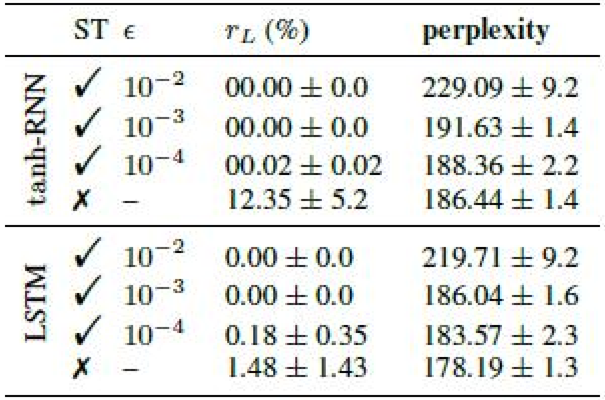
Method	Perplexity	Self-BLEU4	Zipf Coefficient	Repetition %	HUSE
Human	12.38	0.31	0.93	0.28	-
Greedy	1.50	0.50	1.00	73.66	-
Beam, b=16	1.48	0.44	0.94	28.94	-
Stochastic Beam, b=16	19.20	0.28	0.91	0.32	-
Pure Sampling	22.73	0.28	**0.93**	0.22	0.67
Sampling, t=0.9	10.25	0.35	0.96	0.66	0.79
Top-k=40	6.88	0.39	0.96	0.78	0.19
Top-k=640	13.82	**0.32**	0.96	**0.28**	0.94
Top-k=40, t=0.7	3.48	0.44	1.00	8.86	0.08
Nucleus p=0.95	**13.13**	**0.32**	0.95	0.36	**0.97**

- Unlikelihood



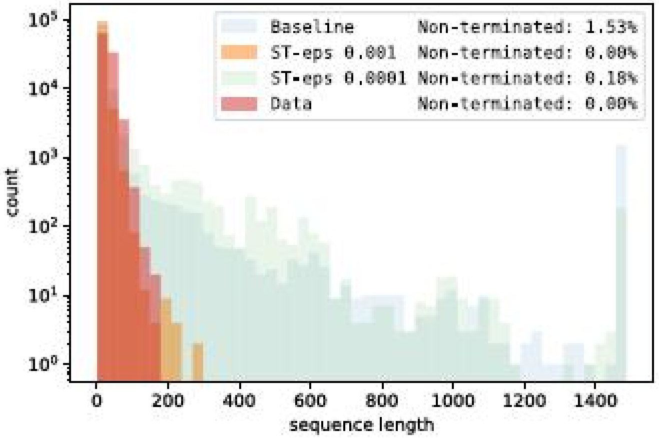
Model	search	seq-rep-4	uniq-seq	ppl	acc	rep	wrep	uniq
$\mathcal{L}_{\text{MLE}}$	greedy	.429	10.6k	**24.59**	**.401**	.619	.346	11.6k
	beam	.495	9.4k					
$\mathcal{L}_{\text{UL-token}}$	greedy	**.274**	**12.6k**	25.62	.396	**.569**	**.305**	**12.5k**
	beam	**.327**	**11.2k**					
$\mathcal{L}_{\text{UL-seq}}$	greedy	.130	12.7k	**24.28**	**.406**	.603	.329	12.4k
	beam	.018	16.8k					
$\mathcal{L}_{\text{UL-token+seq}}$	greedy	**.051**	**14.8k**	25.37	.401	**.551**	**.287**	**13.4k**
	beam	**.013**	**17.6k**					
Human	-	.005	18.9k	-	-	.479	-	18.9k

- STRLM



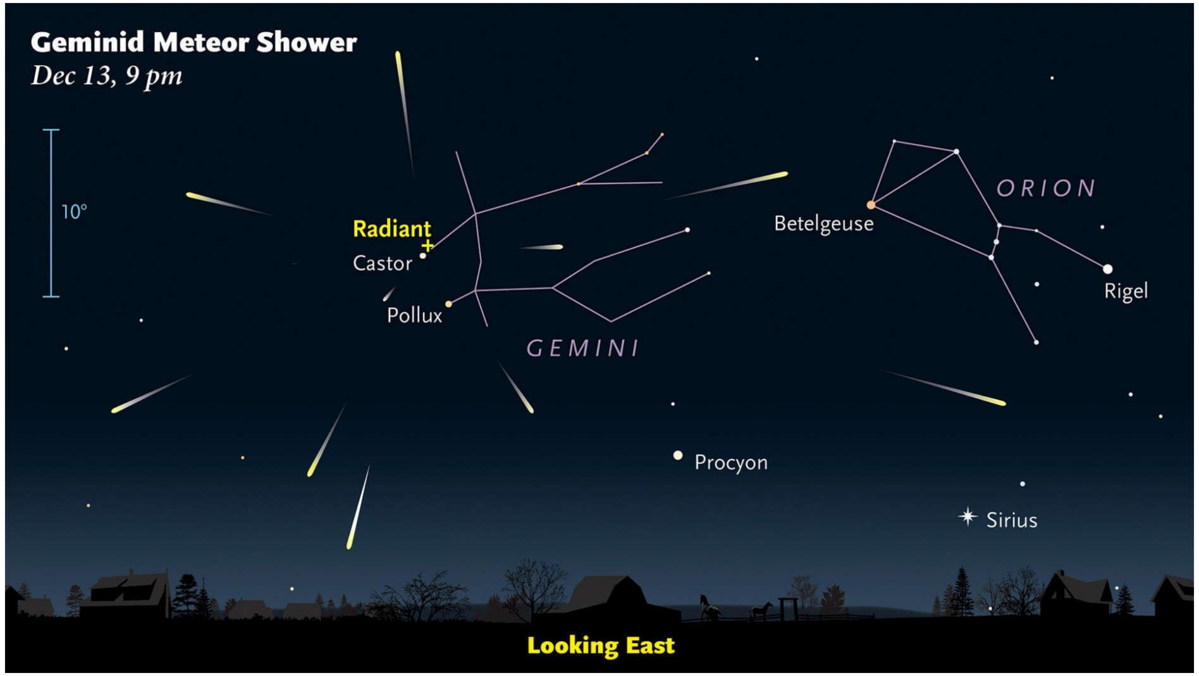
	ST	ϵ	r_L (%)	**perplexity**
tanh-RNN	✓	10^{-2}	00.00 ± 0.0	229.09 ± 9.2
	✓	10^{-3}	00.00 ± 0.0	191.63 ± 1.4
	✓	10^{-4}	00.02 ± 0.02	188.36 ± 2.2
	✗	–	12.35 ± 5.2	186.44 ± 1.4
LSTM	✓	10^{-2}	0.00 ± 0.0	219.71 ± 9.2
	✓	10^{-3}	0.00 ± 0.0	186.04 ± 1.6
	✓	10^{-4}	0.18 ± 0.35	183.57 ± 2.3
	✗	–	1.48 ± 1.43	178.19 ± 1.3

Table 3: Non-termination ratio (r_L (%)) of greedy-decoded sequences and test perplexity for STRLMs.



Geminids

- 高峰期：12月13至14日期间
- 流星雨的天顶每小时出现率可达到150，为今年所有流星雨现象中最高
- 不受月光影响，观测条件极佳。
- 然而，受制于其他因素包括城市光害等，即使在天气情况良好下，在偏远地区的观测人士于高峰期间每小时可能会看到大约20–30颗流星。
- 在12月13日，双子座流星雨的辐射点会于晚上约7时30分从东北方升起。当辐射点升出地平线后，即有机会欣赏到来自双子座的流星，而最佳的观赏时间是12月14日凌晨2时至3时。
- 由于流星不一定出现在辐射点附近，观测人士应选择天空视野广阔及光害程度低的地点进行观测。
- 有关流星雨高峰期及流星数量的预测有一定的不确定性，有兴趣观测的人士亦可于高峰期前后1至2天进行观测。



- Reference：
  - Hierarchical Neural Story Generation
  - The Curious Case of Neural Text Degeneration
  - Consistency of a Recurrent Language Model With Respect to Incomplete Decoding
  - Neural Text ~~de~~Generation with Unlikelihood Training
- Thanks~