Neural Text Generation With Unlikelihood Training

Sean et al.

Presenter: Jungsoo Park

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

1 Introduction

"It's Not About Bigger Models,
But Bigger Ideas."

1 Introduction

Prefix	*Lyrically the song has excerpts of different languages including French , Spanish ,*
GPT2-greedy	German , Italian , Portuguese , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , ...
GPT2-top3	German , Italian , Portuguese , Spanish , Portuguese , Portuguese, Italian and Spanish. It is also available in English, French and Spanish. In addition to its lyrics, the album features a number of original songs, ...
GPT2-nucleus-0.3	German , Italian , Portuguese , Spanish , Portuguese , Italian , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , Spanish , Portuguese , ...

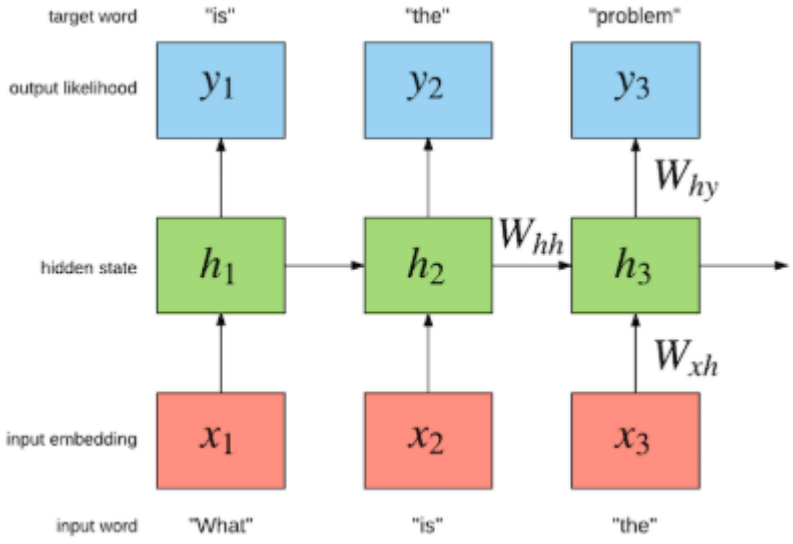
Degenerate repetition(single, phrase) in completion task(Radford et al. 2019)

Probably Due to

- By-product of transformer model
- Intrinsic property of human language
- Limitation of likelihood objective

Language models are unsupervised multitask learners(Radford et al. 2019)

1 Introduction



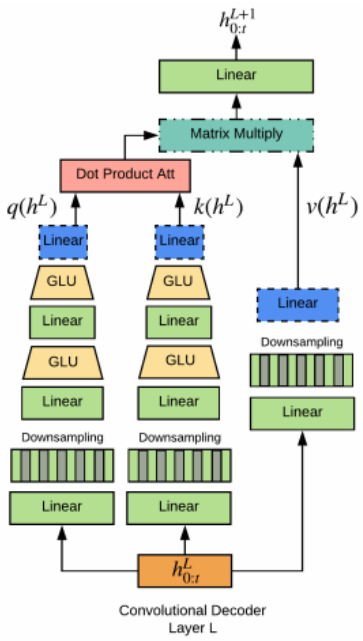
Language Model

Current neural text generation models

- Pay little attention to the argmax or the top of the ranked list of next token probability
- Not focused on optimizing sequence generation

2 Related Work

Hierarchical Neural Story Generation(Fan et al., 2018)

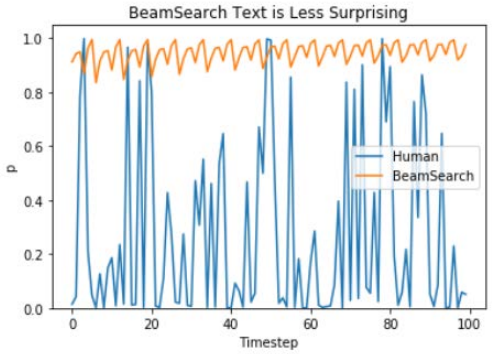


- Main contribution: collected a dataset of story generation(w/ prompts) and proposed a model
- Model consists of revised self attention head(left figure), convolutional decoder, and fusion module
- Top-K sampling method was used to generate "creative" story (not generic, not frequent)

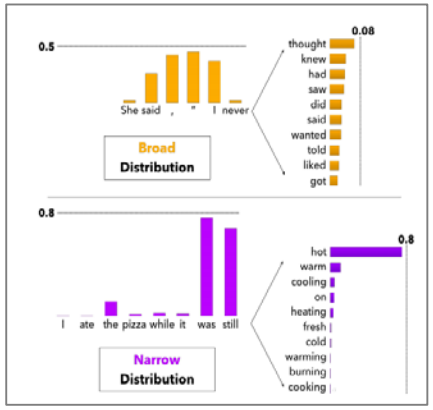
Self attention(single head) with GLU gating and downsampling

2 Related Work

The Curious Case of Neural Text *De*generation (Holtzman et al., 2019)



Lack of Diversity



Broad vs Narrow

- Main contribution: Identify the limitation of beam search based decoding in open-ended generation task, and propose nucleus sampling
- Nucleus (Top-p) sampling

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

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

3 Method

Formal Definition

Objective

$$\mathcal{L}_{\text{MLE}}(p_\theta, \mathcal{D}) = -\sum_{i=1}^{|\mathcal{D}|} \sum_{t=1}^{|\mathbf{x}^{(i)}|} \log p_\theta(x_t^{(i)} | x_{<t}^{(i)}).$$

Task(Sequence Completion)

$$\mathbf{x}_{1:k} \sim p_*, \quad \hat{\mathbf{x}}_{k+1:N} \sim p_\theta(\cdot | \mathbf{x}_{1:k})$$

$$(x_1, \ldots, x_k, \hat{x}_{k+1}, \ldots, \hat{x}_N) \sim p_*,$$

Deterministic Decoding

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

Stochastic Decoding

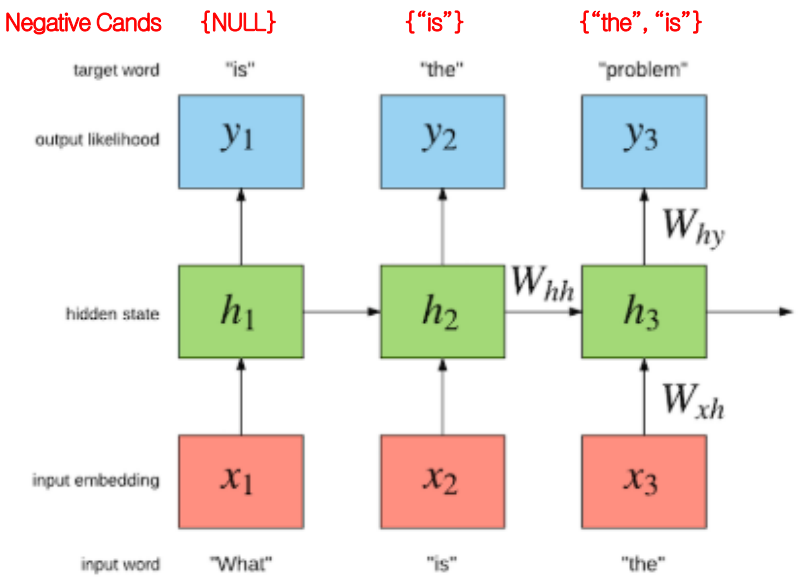
$$q(x_t | x_{<t}, p_\theta) = \begin{cases} p_\theta(x_t | x_{<t})/Z & x_t \in U \\ 0 & \text{otherwise,} \end{cases}$$

3 Method

The Unlikelihood Training Objective

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

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



3 Method

The Unlikelihood Training Objective(Gradient Analysis)

$$\nabla \mathcal{L}_a = x^* - m \odot p, \quad m_i = \begin{cases} (1 - \alpha \frac{p_{\text{neg}}}{1-p_{\text{neg}}}) & \text{if } i \neq i_{\text{neg}} \\ (1+\alpha) & \text{if } i = i_{\text{neg}}, \end{cases}$$

True Next-Token ($i = i^*$)

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_*} \frac{\partial p_*}{\partial a_{i^*}} &= (1 - p_*) - \alpha \frac{p_{\text{neg}}}{1 - p_{\text{neg}}} (0 - p_*) \\ &= 1 - p_*(1 - \alpha \frac{p_{\text{neg}}}{1 - \cdots}). \end{aligned}$$

Negative Candidate ($i = i_{\text{neg}}$)

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_{\text{neg}}} \frac{\partial p_{\text{neg}}}{\partial a_{\text{neg}}} &= (0 - p_{\text{neg}}) - \alpha \frac{p_{\text{neg}}}{1 - p_{\text{neg}}} (1 - p_{\text{neg}}) \\ &= -p_{\text{neg}}(1 + \alpha). \end{aligned}$$

Other Token ($i \notin \{i^*, i_{\text{neg}}\}$)

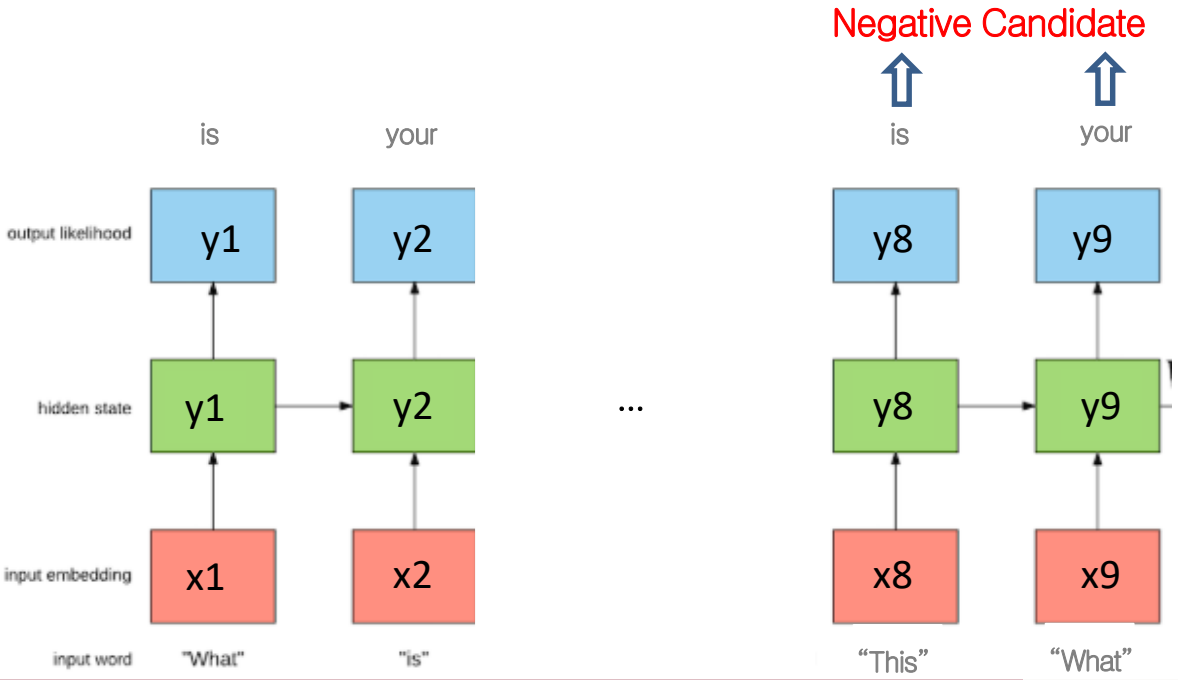
$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \tilde{p}_i} \frac{\partial \tilde{p}_i}{\partial a_i} &= (0 - \tilde{p}_i) - \alpha \frac{p_{\text{neg}}}{1 - p_{\text{neg}}} (0 - \tilde{p}_i) \\ &= -\tilde{p}_i (1 - \alpha \frac{p_{\text{neg}}}{1 - p_{\text{neg}}}). \end{aligned}$$

3 Method

The Unlikelihood Training Objective

$$(\mathcal{C}^{k+1}, \ldots, \mathcal{C}^{k+N}) \qquad \mathcal{L}_{\text{ULS}}^{t}(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = -\sum_{c \in \mathcal{C}^t} \log(1 - p_\theta(c|x_{<t})).$$

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



4 Experiments

Prefix	*... Lyrically the song has excerpts of different languages including French , Spanish*
$\mathcal{L}_{\text{MLE}}$	, Italian , Spanish , Italian , Spanish , Italian , Spanish , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese
$\mathcal{L}_{\text{UL-token+seq}}$	, and German . In the first verse , the protagonist sings about being a " girl who 's been in love with someone else " , while the second verse describes the relationship between the protagonist and her lover . In the third verse , the protagonist sings
Prefix	*... starboard engines and was going to crash . " We 're going in ,"*
$\mathcal{L}_{\text{MLE}}$	he said . " We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to crash . We 're going to
$\mathcal{L}_{\text{UL-token+seq}}$	Hood said . " I 'm going to make sure we 're going to get back to the water . " The order to abandon ship was given by Admiral Beatty , who ordered the remaining two battlecruisers to turn away . At 18 : 25 , Hood turned his
Prefix	*... career - high 27 points on 8 - for - 11 shooting with three rebounds*
$\mathcal{L}_{\text{MLE}}$	and two assists . On January 3 , 2012 , he was named to the 2012 13 All - Atlantic 10 first team . On February 3 , 2012 , he was named to the Atlantic 10 first team . On February 5 , 2012 , he was named
$\mathcal{L}_{\text{UL-token+seq}}$	and a career - high 7 assists against the Minnesota Timberwolves . On February 3 , 2012 , he was named to the 2012 All - NBA First Team . On March 7 , 2012 , he was named one of five finalists for the Naismith Award , which is

4 Experiments

Model	search	seq-rep-4	uniq-seq	ppl	acc	rep	wrep	uniq
$\mathcal{L}_{\text{MLE}}$	greedy	.442	10.8k	**25.64**	**.395**	.627	.352	11.8k
	beam	.523	9.5k					
$\mathcal{L}_{\text{UL-token}}$	greedy	**.283**	**13.2k**	26.91	.390	**.577**	**.311**	**12.7k**
	beam	**.336**	**11.7k**					
$\mathcal{L}_{\text{UL-seq}}$	greedy	.137	13.1k	**25.42**	**.399**	.609	.335	12.8k
	beam	.019	18.3k					
$\mathcal{L}_{\text{UL-token+seq}}$	greedy	**.058**	**15.4k**	26.72	.395	**.559**	**.293**	**13.8k**
	beam	**.013**	**19.1k**					
Human	-	.006	19.8k	-	-	.487	-	19.8k

$$\text{rep}/\ell = \frac{1}{|\mathcal{D}|T} \sum_{\mathbf{x} \in \mathcal{D}} \sum_{t=1}^{T} \mathbb{I}\left[\arg\max p_\theta(x|\mathbf{x}_{<t}) \in \mathbf{x}_{t-\ell-1:t-1}\right].$$

$$\text{seq-rep-n} = 1.0 - \frac{|\text{unique n-grams}(\mathbf{x}_{k+1:k+N})|}{|\text{n-grams}|},$$