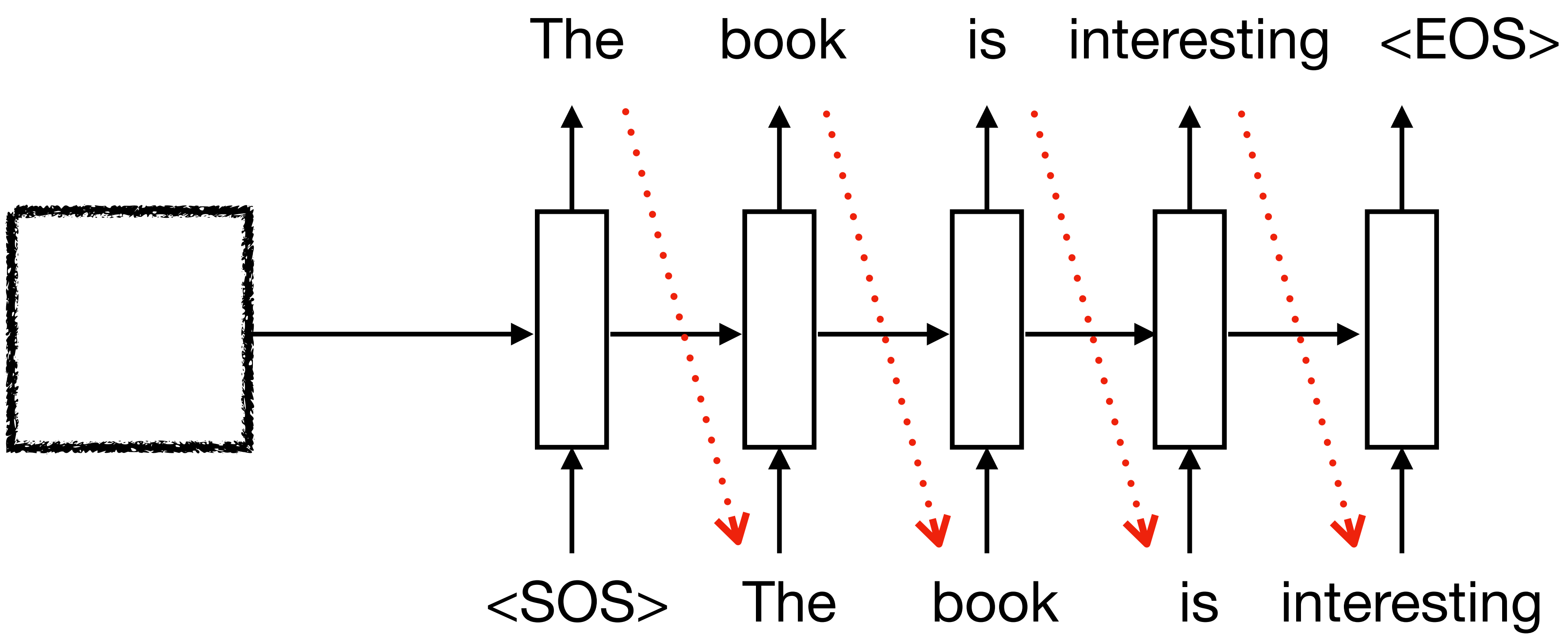
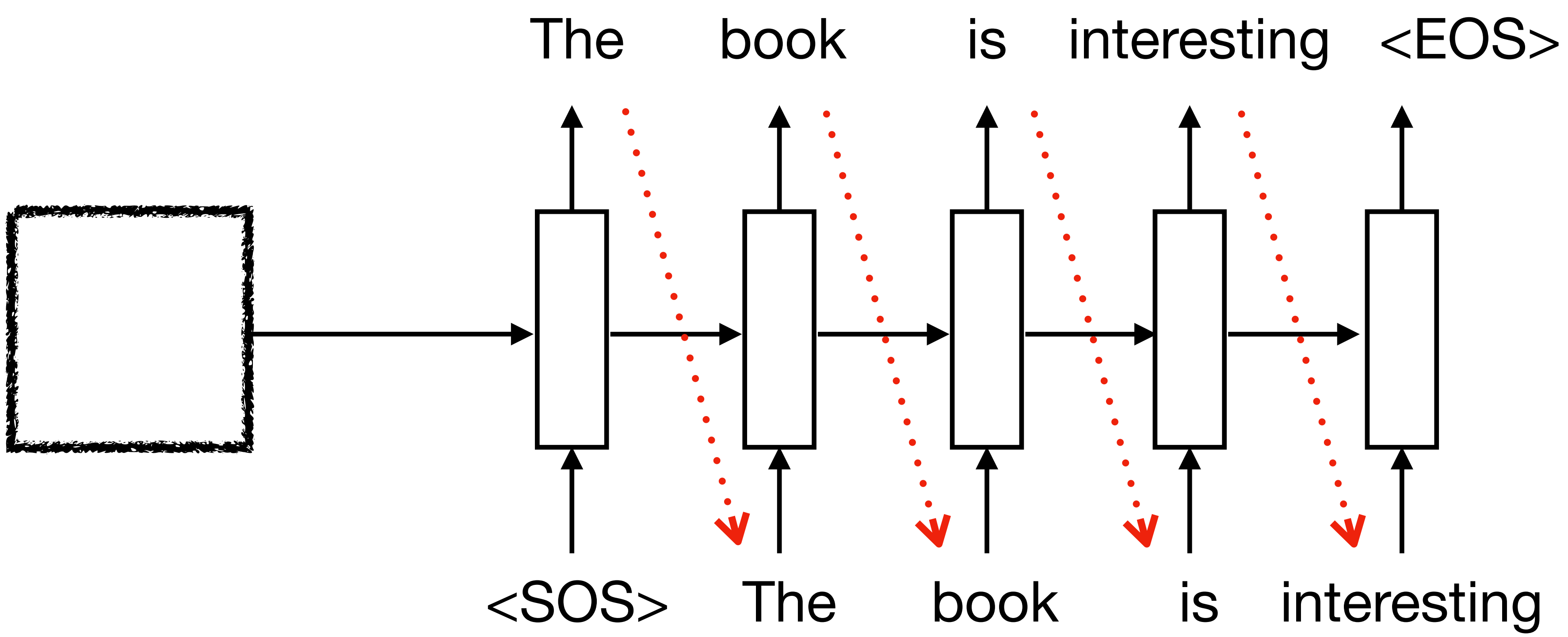
Sampling and Stochastic Search for Sentence Generation

Lili Mou
doublepower.mou@gmail.com
lili-mou.github.io

RNN Generation



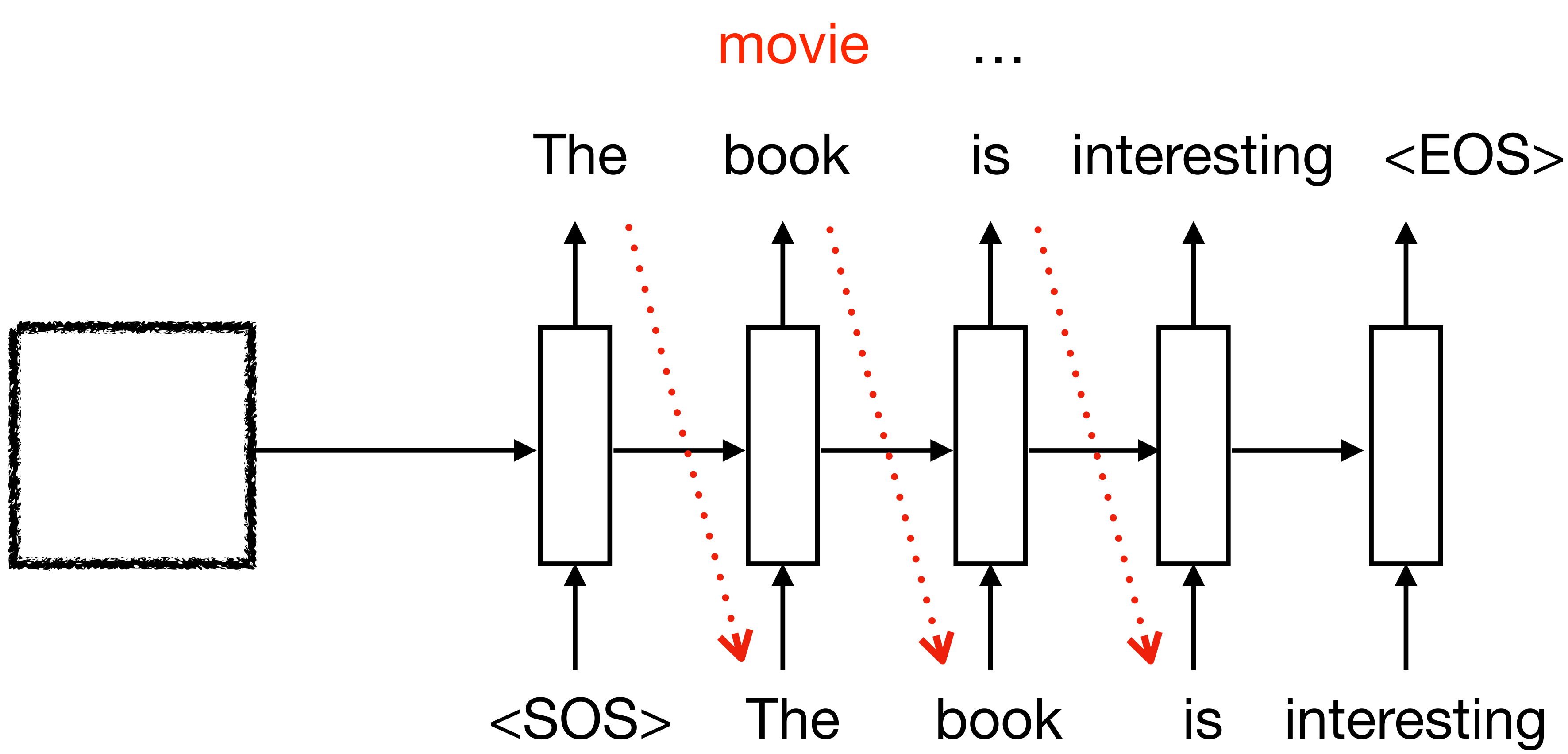
RNN Generation



Question: Can we generate a sentence right-to-left?

Issues with Single Directional Generation

- Information bottleneck
- Error cumulation
  - Due to sampling or incompetency of the RNN



Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

~~The~~ book is interesting <EOS>

This

Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

~~The~~ book is interesting <EOS>

This

quite

Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

fascinating

~~The~~ book is ~~interesting~~ <EOS>

This

quite

Applications

- Paraphrase generation
  - “Sample” a sentence with similar semantics but different wordings
- Summarization
  - “Sample” a sentence with similar semantics
- Grammatical error correction
  - “Sample” a more likely sentence with the same semantics

Sampling Methods

Independent Sampling

- Sampling from CDF
  - Probabilistic density function (PDF)

$$\Pr[a \le x \le b] = \int_a^b f(x)\,dx$$

  - Cumulative density function (CDF)

$$F(x) = \int_{-\infty}^{x} f(u)\,du = \Pr[u \le x]$$

  - Sampling procedure

$$u \sim \mathrm{U}[0,1]; \quad x = \mathrm{CDF}^{-1}(u)$$

- Problems
  - CDF not analytic
  - Especially, the conditional CDF in multivariate cases

Independent Sampling

- Reject Sampling
  - To sample from $p(x) = \frac{1}{Z}\widetilde{p}(x)$
  - We instead sample $x \sim q(x)$
  - Accept the sample x with probability $\dfrac{\widetilde{p}(x)}{k \cdot q(x)}$

    where k is a constant s.t. $kq(x) \geq \widetilde{p}(x), \forall x$
  - Reject x w.p. $1 - \dfrac{\widetilde{p}(x)}{k \cdot q(x)}$
- Many other sampling methods

Dependent Sampling

- Problem: Sample from $p(x)$
- MCMC sampling
  - Start from an arbitrary initial sample $x^{(0)}$
  - Sample $x^{(1)} \sim p(x^{(1)} | x^{(0)}), \quad x^{(2)} \sim p(x^{(2)} | x^{(1)}), \quad \cdots$
  - Hope $p(x^{(n)}) \to p(x)$ as $n \to \infty$

Markov Chain

- States: $S = \{s_1, s_2, \cdots\}$
- Initial distribution $\pi^{(0)}$
- Transition probability: $\mathcal{T}_{i \to j} = p(x^{(t+1)} = s_j \,|\, x^{(t)} = s_i)$
  - $x^{(t+1)}$ is independent of $x^{(t-1)}$, given $x^{(t)}$
  - $\mathcal{T}_{i \to j}$ works for all time steps t
- **Thm**: Starting from an arbitrary initial distribution, a Markov Chain converges to a **unique** stationary distribution (under mild assumptions).

Markov Chain Monte Carlo

- Problem: Sample from $p(x)$
- MCMC sampling
  - Start from an arbitrary initial sample $x^{(0)}$
  - Sample $x^{(1)} \sim p(x^{(1)} \mid x^{(0)}), \quad x^{(2)} \sim p(x^{(2)} \mid x^{(1)}), \quad \cdots$
  - Hope $p(x^{(n)}) \to p(x)$ as $n \to \infty$

Markov Chain Monte Carlo

- Problem: Sample from $p(x)$
- MCMC sampling
  - Start from an arbitrary initial sample $x^{(0)}$
  - Sample $x^{(1)} \sim p(x^{(1)} \mid x^{(0)}), \quad x^{(2)} \sim p(x^{(2)} \mid x^{(1)}), \quad \cdots$
    by following a carefully designed Markov chain
  - ~~Hope~~ $p(x^{(n)}) \rightarrow p(x)$ as $n \rightarrow \infty$
    Guaranteed that

Metropolis—Hastings Sampler

- **Input**
  - An arbitrary desired distribution $p(x)$
- **Output**
  - An unbiased sample $x \sim p(x)$
- **Algorithm**
  - Start from an arbitrary initial state $x^{(0)}$
  - For every step t

    $g(x'|x)$: arbitrary proposal distribution

    Propose a new state $x' \sim g(x'|x^{(t)})$

    Accept x' w.p. $A(x'|x) = \min\left\{1, \frac{p(x')g(x^{(t)}|x')}{p(x)g(x'|x^{(t)})}\right\}$, i.e., $x^{(t)} = x'$

    Reject x' otherwise, i.e., $x^{(t+1)} = x^{(t)}$
  - Return $x^{(t)}$ with a large t

Proof Sketch

- Detailed balance property = > Stationary distribution

If

$$\forall x, y, \qquad \pi(x) \cdot \mathcal{T}_{x \to y} = \pi(y) \cdot \mathcal{T}_{y \to x}$$

Then

$\pi(x)$ is a stationary distribution

Because

$$\forall x, \qquad \pi(x) = \sum_y \pi(y)\mathcal{T}_{y \to x} = \sum_y \pi(x)\mathcal{T}_{x \to y} = \pi(x)$$

Proof Sketch (Cont.)

- MH Sampler satisfies detailed balance
  - $\forall x, y,$ $$p(x) \cdot \mathcal{T}_{x \to y} = p(x) \cdot g(y \mid x) \cdot \min \left\{ 1, \frac{p(y)g(x \mid y)}{p(x)g(y \mid x)} \right\} \quad (1)$$
    $$p(y) \cdot \mathcal{T}_{y \to x} = p(y) \cdot g(x \mid y) \cdot \min \left\{ 1, \frac{p(x)g(y \mid x)}{p(y)g(x \mid y)} \right\} \quad (2)$$
  - W.L.O.G., we assume $p(x)g(y \mid x) \geq p(y)g(x \mid y)$

$$(1) = p(y) \cdot g(x \mid y)$$

$$(2) = p(y) \cdot g(x \mid y)$$

Gibbs Sampler

- Suppose $\mathbf{x} = (x_1, x_2, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n)$
- If the proposal distribution is $x_i' \sim p(x_i \,|\, \mathbf{x}_{-i})$
- Then, the acceptance rate is $A(\mathbf{x}' \,|\, \mathbf{x}) = \min \left\{ 1, \dfrac{p(\mathbf{x}')g(\mathbf{x} \,|\, \mathbf{x}')}{p(\mathbf{x})g(\mathbf{x}' \,|\, \mathbf{x})} \right\}$
  - Notice that $\mathbf{x}' = (x_1, x_2, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)$
  - Thus, $\dfrac{p(\mathbf{x}')g(\mathbf{x} \,|\, \mathbf{x}')}{p(\mathbf{x})g(\mathbf{x}' \,|\, \mathbf{x})} = \dfrac{p(\mathbf{x}_{-i})p(x_i' \,|\, \mathbf{x}_{-i}) \cdot p(x_i \,|\, \mathbf{x}_{-i})}{p(\mathbf{x}_{-i})p(x_i \,|\, \mathbf{x}_{-i}) \cdot p(x_i' \,|\, \mathbf{x}_{-i})} = 1$

=> Gibbs step is a special case of an MH step, with AC rate = 1

Applying MH to Sentence Generation

MH Components

- State: Every sentence
- Target distribution: Depend on the task
- Proposal distribution
  - Task agnostic, or task specific
- Compute acceptance rate
  - We can’t do anything here

Target distribution

- General formula
  - $p(\mathbf{x}) \propto p_{\text{LM}}(\mathbf{x}) \cdot s_1(\mathbf{x}) \cdots s_n(\mathbf{x})$
  - $s_i(\mathbf{x})$: scoring functions specific to the task

Target distribution

- General formula
  - $p(\mathbf{x}) \propto p_{\mathrm{LM}}(\mathbf{x}) \cdot s_1(\mathbf{x}) \cdots s_n(\mathbf{x})$
  - $s_i(\mathbf{x})$: scoring functions specific to the task

- Keywords-to-sentence generation

$$s(\mathbf{x}) = \begin{cases} 1, & \text{if keywords in } \mathbf{x} \\ 0, & \text{otherwise} \end{cases}$$

- Paraphrase generation/Grammatical error correction
  - $s(\mathbf{x}) = \mathrm{sim}_{\mathrm{semantic}}(\mathbf{x}, \mathbf{x}_0) + \mathrm{diff}_{\mathrm{word}}(\mathbf{x}, \mathbf{x}_0)$

Proposal Distribution

- Replace

$$g_{\text{replace}}(\mathrm{x}'|\mathrm{x}) = \pi(w_m^* = w^c|\mathrm{x}_{-m}) =$$
$$\frac{\pi(w_1, \cdots, w_{m-1}, w^c, w_{m+1}, \cdots, w_n)}{\sum_{w \in \mathcal{V}} \pi(w_1, \cdots, w_{m-1}, w, w_{m+1}, \cdots, w_n)}$$

- Delete

- Insert

  - Also sample from posterior

Examples: Keywords-to-Sentence

Keyword(s)	**Generated Sentences**
friends	My good **friends** were in danger .
project	The first **project** of the scheme .
have, trip	But many people **have** never made the **trip** .
lottery, scholarships	But the **lottery** has provided **scholarships** .
decision, build, home	The **decision** is to **build** a new **home** .
attempt, copy, painting, denounced	The first **attempt** to **copy** the **painting** was **denounced** .

Examples: Paraphrase Generation

Model	BLEU-ref	BLEU-ori	NLL
Origin Sentence	30.49	100.00	7.73
VAE-SVG (100k)	22.50	-	-
VAE-SVG-eq (100k)	**22.90**	-	-
VAE-SVG (50k)	17.10	-	-
VAE-SVG-eq (50k)	17.40	-	-
Seq2seq (100k)	22.79	33.83	6.37
Seq2seq (50k)	20.18	27.59	**6.71**
Seq2seq (20k)	16.77	22.44	6.67
VAE (unsupervised)	9.25	27.23	7.74
CGMH *w/o matching*	18.85	50.28	7.52
w/ KW	20.17	53.15	7.57
w/ KW + WVA	20.41	53.64	7.57
w/ KW + WVM	20.89	54.96	7.46
w/ KW + ST	20.70	54.50	7.78

Type	Examples
Ori	what 's the best plan to lose weight
Ref	what is a good diet to lose weight
Gen	what 's the best way to slim down quickly
Ori	how should i control my emotion
Ref	how do i control anger and impulsive emotions
Gen	how do i control my anger
Ori	why do my dogs love to eat tuna fish
Ref	why do my dogs love eating tuna fish
Gen	why do some dogs like to eat raw tuna and raw fish

Examples: Paraphrase Generation

Model	**#parallel data**	**GLEU**
AMU	2.3M	44.85
CAMB-14	155k	46.04
MLE	720k	52.75
NRL	720k	**53.98**
CGMH	0	45.5

Ori	Even if **we are failed** , We have to try to get **a new things** .
Ref	Even if we all failed , we have to try to get new things .
Gen	Even if we are failing , We have to try to get some new things .

Ori	In the world **oil price very high** right now .
Ref	In today 's world , oil prices are very high right now .
Gen	In the world , oil prices are very high right now .

Analysis

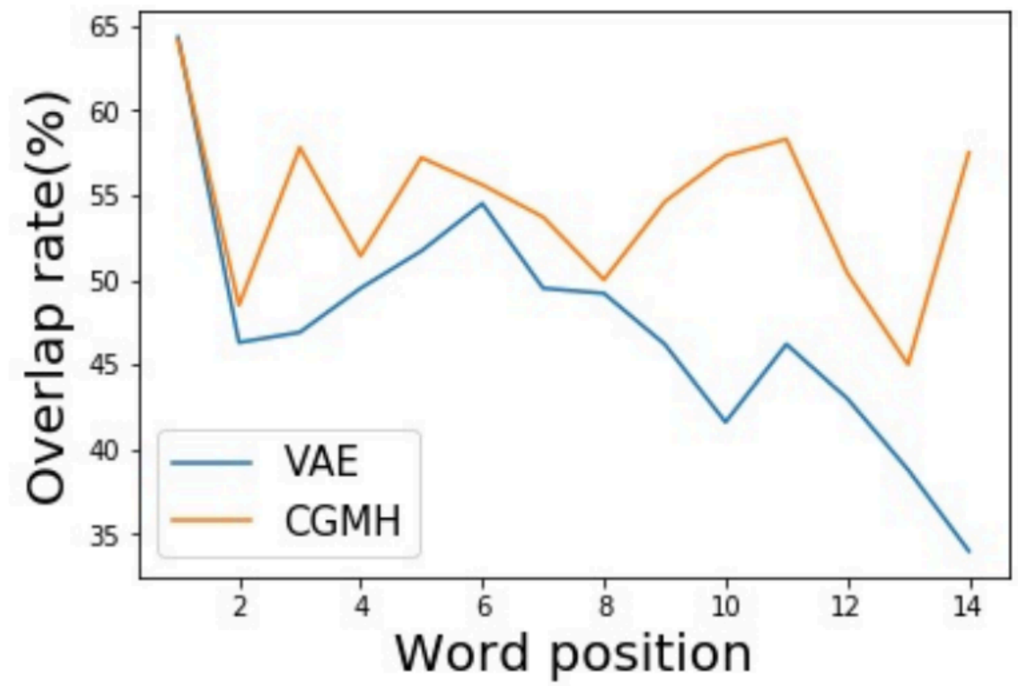


Figure 3: Overlap rates of CGMH and VAE for each word position of sentences.

Analysis (Cont.)

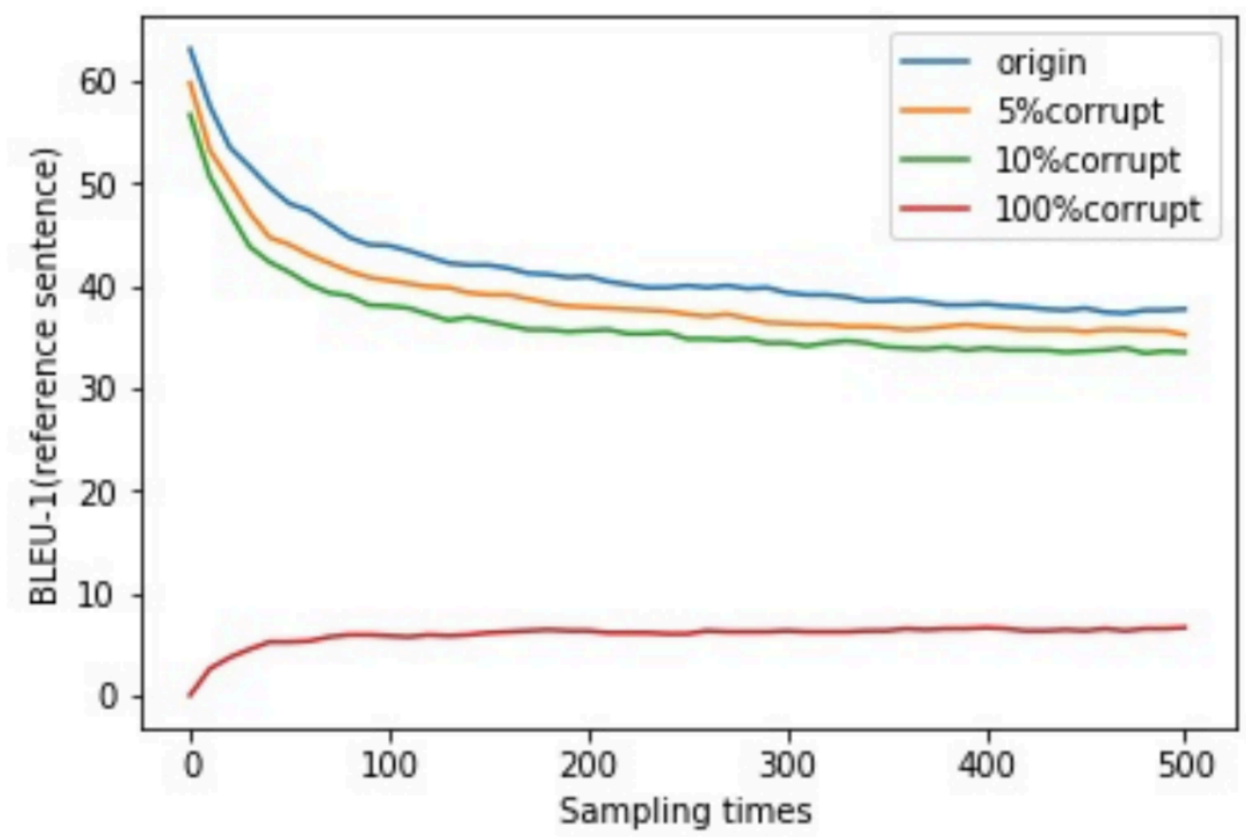


Figure 2: Generation quality with corrupted initial states. At each situation, 0/5%/10%/100% of the words in initial sentences are randomly replaced with other words.

Summary (Take-Home Msg)

- MCMC: Dependent sampling with a Markov chain
  - Gibbs sampler: posterior sampling
  - MH sampler: propose-and-reject sampling
- MH sentence generation
  - The framework is the same
  - Need to design target and proposal distributions
  - Various applications: paraphrase generation, grammatical error correction, keywords-to-sentence generation, etc.

    (A blueprint is needed for MH generation)

Thank you!

Q&A