Sentence Generation

Lili Mou
lmou@ualberta.ca
lili-mou.github.io
Roapmap

• Motivation and examples

• Techniques
  - Generation from latent space
  - Generation from word space
Motivation

• Sentence generation
  - Dialogue systems
  - Paraphrase generation
  - Machine translation

• A seq2seq model may not suffice
  - No input
  - Constructing new information
  - Diversity needed

• Probabilistic sentence generation
  - Prior sampling, posterior sampling
Latent Space Sampling
Variational Autoencoder

- Humans’ sentence generation involves two steps
  - First, we have some “vague” idea of the sentence
  - Then, we flesh it out by words

- A sentence $x = (x_1, \cdots, x_T)$ is subject to some latent representation $z$

$$p(z, x) = p(z)p(x | z)$$

Variational Autoencoder

- Humans’ sentence generation involves two steps
  - First, we have some “vague” idea of the sentence
  - Then, we flesh it out by words

- A sentence $x = (x_1, \cdots, x_T)$ is subject to some latent representation $z$

\[
p(z, x) = p(z)p(x | z)
\]

How can we learn a model with latent variables?

E-step: $p(z | x) = \frac{p(z)p(x | z)}{p(x)} = \frac{p(z)p(x | z)}{\int p(z')p(x | z')dz'}$

M-step: maximize $\mathbb{E}_{z \sim p(z | x)} \log p(z, x)$
Variational Autoencoder

- Humans’ sentence generation involves two steps
  - First, we have some “vague” idea of the sentence
  - Then, we flesh it out by words

- A sentence $x = (x_1, \cdots, x_T)$ is subject to some latent representation $z$

\[ p(z, x) = p(z)p(x | z) \]

How can we learn a model with latent variables?

E-step: $p(z | x) = \frac{p(z)p(x | z)}{p(x)} = \frac{p(z)p(x | z)}{\int p(z')p(x | z')dz'}$

M-step: maximize $\mathbb{E}_{z \sim p(z|x)} \log p(z, x)$
Variational Autoencoder

- Humans’ sentence generation involves two steps
  - First, we have some “vague” idea of the sentence
  - Then, we flesh it out by words

- A sentence $x = (x_1, \ldots, x_T)$ is subject to some latent representation $z$

$$p(z, x) = p(z)p(x | z)$$

How can we learn a model with latent variables?

**Recognition**

- **E-step:** $p(z | x) = \frac{p(z)p(x | z)}{p(x)} = \frac{p(z)p(x | z)}{\int p(z')p(x | z')dz'}$

**Reconstruction**

- **M-step:** maximize $\mathbb{E}_{z \sim p(z|x)} \log p(z, x)$

(in a more general sense)
Variational Inference

\[
\log p(x; \theta) = \log \left( \int_z p(x, z; \theta) dz \right)
\]

\[
= \int q(z | x) \log \frac{p(y, z; \theta)}{q(z | x)} dz + \int q(z | x) \log \frac{q(z | x)}{p(z | x; \theta)} dz
\]

\[
L(q, \theta)\quad \text{KL}(q(Z | x) \| p(Z | x))
\]

Variational inference vs EM

<table>
<thead>
<tr>
<th>Variational family ( q \in Q )</th>
<th>( q ) can be any distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignore the KL-term</td>
<td>True posterior is the best</td>
</tr>
<tr>
<td></td>
<td>KL=0 after E step</td>
</tr>
</tbody>
</table>

CMPUT 651 (Fall 2019)

Variational Inference
Variational Inference

\[
\log p(x; \theta) = \log \left( \int_z p(x, z; \theta) dz \right)
\]

\[
= \int q(z|x) \log \frac{p(y, z; \theta)}{q(z|x)} dz + \int q(z|x) \log \frac{q(z|x)}{p(z|x; \theta)} dz
\]

\[
L(q, \theta) \quad \text{KL}(q(Z|x) || p(Z|x))
\]

- Two extremes
  - Q = any function \(\Rightarrow\) EM
    \(\Rightarrow\) powerful model; optimization intractable
  - Q = \{a fixed distribution\}
    \(\Rightarrow\) degenerated model; optimization easy
Variational Inference

\[
\log p(x; \theta) = \log \left( \int p(x, z; \theta) dz \right)
\]

\[
= \int q(z | x) \log \frac{p(y, z; \theta)}{q(z | x)} dz + \int q(z | x) \log \frac{q(z | x)}{p(z | x; \theta)} dz
\]

\[
L(q, \theta) \quad \text{KL}(q(Z | x) \| p(Z | x))
\]

- Two extremes
  - \( Q = \text{any function} \Rightarrow \text{EM} \)
    \[
    \Rightarrow \text{powerful model; optimization intractable}
    \]
  - \( Q = \{\text{a fixed distribution}\} \)
    \[
    \Rightarrow \text{degenerated model; optimization easy}
    \]

Trade-off, e.g.,
- Independent assumption
- Gaussian assumption
Example  

Variational family: factorized distribution

\[
p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp \left\{ -\frac{\tau}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}
\]

\[
p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0 \tau)^{-1})
\]

\[
p(\tau) = \text{Gam}(\tau|a_0, b_0)
\]

\[
q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)
\]

Figure 10.4 Illustration of variational inference for the mean \(\mu\) and precision \(\tau\) of a univariate Gaussian distribution. Contours of the true posterior distribution \(p(\mu, \tau|\mathcal{D})\) are shown in green. (a) Contours of the initial factorized approximation \(q_\mu(\mu)q_\tau(\tau)\) are shown in blue. (b) After re-estimating the factor \(q_\mu(\mu)\). (c) After re-estimating the factor \(q_\tau(\tau)\). (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.
Variational Autoencoder

- Variational autoencoder
  - Variational family: $Q = \{ \mathcal{N}(\mu, \text{diag } \sigma^2) : \mu \in \mathbb{R}^d, \sigma \in \mathbb{R}^d_{++} \}$
  - Recognizing $\mu, \sigma$ by NN
  - Modeling $x$ also by NN (need a little bit more efforts)
Variational Autoencoder

\[
\log p_\theta(x) = \log \left( \int_z p_\theta(x, z)dz \right)
\]

\[
= \int q_\phi(z \mid x) \log \frac{p_\theta(y, z)}{q_\phi(z \mid x)} dz + \int q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} dz
\]

\[
\geq \int q_\phi(z \mid x) \log \frac{p_\theta(y, z; \theta)}{q_\phi(z \mid x)} dz
\]

\[
= \int q_\phi(z \mid x) \log p_\theta(y \mid z) dz + \int q_\phi(z \mid x) \log \frac{p_\theta(z)}{q_\phi(z \mid x)} dz
\]

\[
= \mathbb{E}_{z \sim q_\phi(z \mid x)} \log p_\theta(y \mid z) - KL(q_\phi(z \mid x) \| p_\theta(z))
\]
Formula Zoo

\[
\begin{align*}
E_{x \sim p_{\text{data}}(x)} \left[ \log p(x) \right] & \geq E_{x \sim p_{\text{decoder}}(x)} \left[ E_{q(z|x)} \left[ \log p(x|z) \right] \right] - \underbrace{E_{x \sim p_{\text{data}}(x)} \left[ \text{KL} \left( q(z|x) \parallel p(z) \right) \right]}_{\text{VAE Reconstruction}} \\
& \quad - \underbrace{\text{VAE Regularization}}_{(1)} \\
& = E_{x \sim p_{\text{data}}(x)} \left[ E_{q(z|x)} \left[ \log p(x|z) \right] \right] - \overbrace{\text{KL} \left( q(z|x) \parallel p(z) \right) p_{\text{data}}(x)}_{\text{AVB Reconstruction}} \\
& \quad - \underbrace{\text{AVB Regularization}}_{(2)} \\
& = E_{x \sim p_{\text{data}}(x)} \left[ E_{q(z|x)} \left[ \log p(x|z) \right] \right] - \overbrace{\text{KL} \left( q(z) \parallel p(z) \right) - I(z; x)}_{\text{AAE Reconstruction}} \\
& \quad - \underbrace{\text{AAE Regularization}}_{\text{Natural Info.}}_{(3)} \\
& = - E_{z \sim q(z)} \left[ \text{KL} \left( q(z|x) \parallel p(z) \right) \right] - \overbrace{\text{KL} \left( q(z) \parallel p(z) \right) - H_{\text{data}}(x)}_{\text{IAE Reconstruction}} \\
& \quad - \underbrace{\text{IAE Regularization}}_{\text{Entropy of data}}_{(4)} \\
& = - \overbrace{\text{KL} \left( q(z|x) \parallel r(z|x) \right) - \text{KL} \left( q(z) \parallel p(z) \right) - H_{\text{data}}(x)}_{\text{IAE Reconstruction}} \\
& \quad - \underbrace{\text{IAE Regularization}}_{\text{Entropy of data}}_{(5)} \\
& = - \overbrace{\text{KL} \left( q(z|x) \parallel p(z) \right) - H_{\text{data}}(x)}_{\text{ALL BiGAN Cost}} \\
& \quad - \underbrace{\text{Entropy of data}}_{(6)}
\end{align*}
\]
Adversarial/Wasserstein Autoencoder

- **VAE:**
  \[ q(z \mid x) \rightarrow p(z) \]

- **WAE:**
  \[ q(z) = \int p_{\mathcal{D}}(x)q(z \mid x)dx \xrightarrow{close} p(z) \]

\[
J = \mathbb{E}_{x \in p_{\mathcal{D}}(x)}\mathbb{E}_{z \in q(z \mid x)} \log p(z \mid x) + \mathbb{D}(q(z), p(z))
\]

Implicit Distributions

- We penalize some distance between \( q(z) \) and \( p(x) \)
- We do not have an explicit form or \( q(z) \)
- But samples from \( q(z) := \int p_\mathcal{D}(x)q(z \mid x)dx \)
  \[
  x^{(i)} \sim p_\mathcal{D}(x), \quad z \sim q(z \mid x^{(i)})
  \]
Adversarial Training

- We penalize some distance between $q(z)$ and $p(z)$
- We deliberately introduce a classifier (discriminator/adversary) to distinguish samples from $q(z)$ and $p(z)$
- We train the model to fool the discriminator
  - Flipping gradient
  - Maximizing predicted entropy
Algorithm

foreach mini-batch do

minimize $J_{\text{dis}}(\theta_{\text{dis}})$ w.r.t. $\theta_{\text{dis}}$

minimize $J_{\text{ovr}}$ w.r.t. $\theta_{E}, \theta_{D}$

end

- Modeling $J_{\text{ovr}}$
  - Flipping gradient
    \[ J_{\text{ovr}} = J_{\text{rec}} - J_{\text{dis}} \]
  - Maximizing predicted entropy
    \[ J_{\text{ovr}} = J_{\text{rec}} - \mathcal{H}(y_{\text{dis}}) \]

$q$ or $p$?

\[ \theta_{\text{dis}} \]

\[ \mathcal{X} \rightarrow \theta_{E} \rightarrow Z \rightarrow \theta_{D} \rightarrow \mathcal{X} \]
Applications of Adv Training

• Adversarial/Wasserstein autoencoder
  \[ q(z) \ vs \ p(z) \]

• Generative adversarial network
  \[ p_{\text{gen}}(z) \ vs \ p_{\mathcal{D}}(z) \]

• Domain adaptation
  \[ p_{D1}(z) \ vs \ p_{D2}(z) \]
A Few Fundamental Questions

- VAE: KL collapse

\[ J = \mathbb{E}_{z \sim q(z|x)}[\log p(x | z)] + \text{KL}(q(z | x) \| p(z)) \]

- WAE alleviates this problem

A Few Fundamental Questions

• WAE: Stochasticity collapse

\[ J = \mathbb{E}_{z \sim q(z|x)}[\log p(x \mid z)] + \mathcal{D}(q(z), p(z)) \]

• If
  - Gaussian encoder
  - Gradient comes from samples
  - Sampling var << Data var

• Then
  - \( \sigma^2 \to 0 \) by SGD

Applications of *AEs

- Regularization

Especially good for interpolation
Applications of *AEs

• Prior sampling

<table>
<thead>
<tr>
<th>Training Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>a mother and her child are outdoors.</td>
</tr>
<tr>
<td>the people are opening presents.</td>
</tr>
<tr>
<td>the girls are looking toward the water.</td>
</tr>
<tr>
<td>a small boy walks down a wooden path in the woods.</td>
</tr>
<tr>
<td>a person in a green jacket it surfing while holding on to a line.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>two families walking in a towel down alaska sands a cot.</td>
</tr>
<tr>
<td>a blade is rolling its nose furiously paper.</td>
</tr>
<tr>
<td>a woman in blue shirts is passing by a some beach</td>
</tr>
<tr>
<td>transporting his child are wearing overalls.</td>
</tr>
<tr>
<td>a guys are blowing on professional thinks the horse.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WAE-D ($\lambda_{\text{WAE}} = 10$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>the lone man is working.</td>
</tr>
<tr>
<td>the group of men is using ice at the sunset.</td>
</tr>
<tr>
<td>a family is outside in the background.</td>
</tr>
<tr>
<td>two women are standing on a busy street outside a fair</td>
</tr>
<tr>
<td>a tourists is having fun on a sunny day</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WAE-S ($\lambda_{\text{WAE}} = 10, \lambda_{\text{KL}} = 0.01$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>an asian man is dancing in a highland house.</td>
</tr>
<tr>
<td>a person wearing a purple snowsuit jumps over the tree.</td>
</tr>
<tr>
<td>the vocalist is at the music and dancing with a microphone.</td>
</tr>
<tr>
<td>a young man is dressed in a white shirt cleaning clothes.</td>
</tr>
<tr>
<td>three children lie together and a woman falls in a plane.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VAE without Annealing</th>
</tr>
</thead>
<tbody>
<tr>
<td>a man is playing a guitar.</td>
</tr>
<tr>
<td>a man is playing with a dog.</td>
</tr>
<tr>
<td>a man is playing with a dog.</td>
</tr>
<tr>
<td>a man is playing a guitar.</td>
</tr>
<tr>
<td>a man is playing with a dog.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VAE with Annealing</th>
</tr>
</thead>
<tbody>
<tr>
<td>the band is sitting on the main street.</td>
</tr>
<tr>
<td>couple dance on stage in a crowded room.</td>
</tr>
<tr>
<td>two people run alone in an empty field.</td>
</tr>
<tr>
<td>the group of people have gathered in a picture.</td>
</tr>
<tr>
<td>a cruise ship is docking a boat ship.</td>
</tr>
</tbody>
</table>
Applications of *AEs

- Posterior sampling
- Paraphrase generation
- Style-transfer generation

<table>
<thead>
<tr>
<th>Semantic and Syntactic Providers</th>
<th>Syntax-Transfer Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ref\text{syn}:</strong> The child is playing in the garden.</td>
<td>VAE: There is a person in the garden.</td>
</tr>
<tr>
<td><strong>Ref\text{sem}:</strong> There is a dog behind the door.</td>
<td>DSS-VAE: A dog is walking behind the door.</td>
</tr>
<tr>
<td><strong>Ref\text{syn}:</strong> The stadium was packed with people.</td>
<td>DSS-VAE: The stadium was packed with people.</td>
</tr>
<tr>
<td><strong>Ref\text{sem}:</strong> The shellfish was cooked in a wok.</td>
<td>VAE: The man was filled with people.</td>
</tr>
<tr>
<td><strong>Ref\text{syn}:</strong> The airplane is in the sky.</td>
<td>DSS-VAE: There is a airplane in the sky.</td>
</tr>
<tr>
<td><strong>Ref\text{sem}:</strong> There is an apple on the table.</td>
<td>VAE: The man is in the kitchen.</td>
</tr>
</tbody>
</table>
Word Space Sampling

RNN Generation

The book is interesting

<EOS>
RNN Generation

The book is interesting <EOS>

The book is interesting <SOS>

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>

This
Generation by Local Changes

• Suppose we have a blueprint

The book is interesting <EOS>

This book is quite interesting <EOS>
Generation by Local Changes

- Suppose we have a blueprint

The book is interesting <EOS>

This book is quite fascinating <EOS>
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)
    \[ \Pr[a \leq x \leq b] = \int_a^b f(x) \, dx \]
  - Cumulative density function (CDF)
    \[ F(x) = \int_{-\infty}^x f(u) \, du = \Pr[u \leq x] \]
  - Sampling procedure
    \[ u \sim U[0,1]; \quad x = \text{CDF}^{-1}(u) \]

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

• Rejection sampling

  - To sample from \( p(x) = \frac{1}{Z} \tilde{p}(x) \)
  
  - We instead sample \( x \sim q(x) \)
  
  - Accept the sample \( x \) with probability
    \[
    \frac{\tilde{p}(x)}{k \cdot q(x)}
    \]
    where \( k \) is a constant s.t. \( kq(x) \geq \tilde{p}(x), \forall x \)

  - Reject \( x \) w.p. \( 1 - \frac{\tilde{p}(x)}{k \cdot q(x)} \)

• Many other sampling methods
Dependent Sampling

- Goal: Sample from $p(x)$

- MCMC sampling
  - Start from an arbitrary initial sample $x^{(0)}$
  - Sample $x^{(1)} \sim p(x^{(1)} | x^{(0)})$, $x^{(2)} \sim p(x^{(2)} | x^{(1)})$, ...
  - Hope $p(x^{(n)}) \rightarrow p(x)$ as $n \rightarrow \infty$
Markov Chain

• States: \( S = \{s_1, s_2, \ldots \} \)

• Initial distribution \( \pi^{(0)} \)

• Transition probability: \( \mathcal{T}_{i \rightarrow j} = p(x^{(t+1)} = s_j \mid x^{(t)} = s_i) \)
  
  - \( x^{(t+1)} \) is independent of \( x^{(t-1)} \), given \( x^{(t)} \)

  - \( \mathcal{T}_{i \rightarrow j} \) 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

- Goal: Sample from \( p(x) \)
- MCMC sampling
  - Start from an arbitrary initial sample \( x^{(0)} \)
  - Sample \( x^{(1)} \sim p(x^{(1)} | x^{(0)}), \quad x^{(2)} \sim p(x^{(2)} | x^{(1)}), \quad \ldots \)
  - Hope \( p(x^{(n)}) \rightarrow p(x) \) as \( n \rightarrow \infty \)
Markov Chain Monte Carlo

• Goal: Sample from $p(x)$

• MCMC sampling
  
  - Start from an arbitrary initial sample $x^{(0)}$

  - Sample $x^{(1)} \sim p(x^{(1)} | x^{(0)})$, $x^{(2)} \sim p(x^{(2)} | x^{(1)})$, ... 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 $t$
    - 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., $x^{(t+1)} = x'$
    - Reject $x'$ otherwise, i.e., $x^{(t+1)} = x^{(t)}$
  - Return $x^{(t)}$ with a large $t$
Proof Sketch

- Detailed balance property $\Rightarrow$ Stationary distribution

  If

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

  Then

  $\pi(x)$ is a stationary distribution

  Because

  $\forall x, \quad \pi(x) = \sum_{y} \pi(y) \mathcal{T}_{y \rightarrow x} = \sum_{y} \pi(x) \mathcal{T}_{x \rightarrow y} = \pi(x)$
Proof Sketch (Cont.)

- MH Sampler satisfies detailed balance

\[ \forall x, y, \text{ if } x \neq y, \ p(x) \cdot \mathcal{T}_{x \rightarrow y} = p(x) \cdot g(y \mid x) \cdot \min \left\{ 1, \frac{p(y)g(x \mid y)}{p(x)g(y \mid x)} \right\} \] \hspace{1cm} (1)

\[ p(y) \cdot \mathcal{T}_{y \rightarrow x} = p(y) \cdot g(x \mid y) \cdot \min \left\{ 1, \frac{p(x)g(y \mid x)}{p(y)g(x \mid y)} \right\} \] \hspace{1cm} (2)

- W.L.O.G., we assume \( p(x)g(y \mid x) \geq p(y)g(x \mid y) \)

\[ (1) = p(y) \cdot g(x \mid y) \]

\[ (2) = p(y) \cdot g(x \mid y) \]

- \( \forall x, y, \text{ if } x = y, \ p(x)\mathcal{T}_{x \rightarrow y} = p(y)\mathcal{T}_{y \rightarrow x} \) also holds
• 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 \( A(x'|x) = \min \left\{ 1, \frac{p(x')g(x | x')} {p(x)g(x' | x)} \right\} \)

  - Notice that \( \mathbf{x}' = (x_1, x_2, \cdots, x_{i-1}, x'_i, x_{i+1}, \cdots, x_n) \)

  - Thus, \( \frac{p(x')g(x | x')} {p(x)g(x' | x)} = \frac{p(x_{-i})p(x'_i | x_{-i}) \cdot p(x_i | x_{-i})} {p(x_{-i})p(x_i | x_{-i}) \cdot p(x'_i | x_{-i})} = 1 \)

\( \Rightarrow \) 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
  - \( p(x) \propto p_{LM}(x) \cdot s_1(x) \cdots s_n(x) \)
  - \( s_i(x) \): scoring functions specific to the task
Target distribution

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

- Keywords-to-sentence generation
  \[
  s(x) = \begin{cases} 
  1, & \text{if keywords in } x \\ 
  0, & \text{otherwise} 
  \end{cases}
  \]

- Paraphrase generation/Grammatical error correction
  - \( s(x) = \text{sim}_{\text{semantic}}(x, x_0) + \text{diff}_{\text{word}}(x, x_0) \)
Proposal Distribution

• Replace

\[ g_{\text{replace}}(x'|x) = \pi(w^*_m = w^c|x_{-m}) = \frac{\pi(w_1, \cdots, w_{m-1}, w^c, w_{m+1}, \cdots, w_n)}{\sum_{w \in \mathcal{V}} \pi(w_1, \cdots, w_{m-1}, w, w_{m+1}, \cdots, w_n)} \]

• Delete

• Insert

  - Also sample from posterior
Examples: Keywords-to-Sentence

<table>
<thead>
<tr>
<th>Keyword(s)</th>
<th>Generated Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>friends</td>
<td>My good friends were in danger.</td>
</tr>
<tr>
<td>project</td>
<td>The first project of the scheme.</td>
</tr>
<tr>
<td>have, trip</td>
<td>But many people have never made the trip.</td>
</tr>
<tr>
<td>lottery, scholarships</td>
<td>But the lottery has provided scholarships.</td>
</tr>
<tr>
<td>decision, build, home</td>
<td>The decision is to build a new home.</td>
</tr>
<tr>
<td>attempt, copy, painting, denounced</td>
<td>The first attempt to copy the painting was denounced.</td>
</tr>
</tbody>
</table>
Examples: Paraphrase Generation

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ori</td>
<td>what ’s the best plan to lose weight</td>
</tr>
<tr>
<td>Ref</td>
<td>what is a good diet to lose weight</td>
</tr>
<tr>
<td>Gen</td>
<td>what ’s the best way to slim down quickly</td>
</tr>
<tr>
<td>Ori</td>
<td>how should i control my emotion</td>
</tr>
<tr>
<td>Ref</td>
<td>how do i control anger and impulsive emotions</td>
</tr>
<tr>
<td>Gen</td>
<td>how do i control my anger</td>
</tr>
<tr>
<td>Ori</td>
<td>why do my dogs love to eat tuna fish</td>
</tr>
<tr>
<td>Ref</td>
<td>why do my dogs love eating tuna fish</td>
</tr>
<tr>
<td>Gen</td>
<td>why do some dogs like to eat raw tuna and raw fish</td>
</tr>
</tbody>
</table>
Examples: Paraphrase Generation

<table>
<thead>
<tr>
<th>Model</th>
<th>#parallel data</th>
<th>GLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMU</td>
<td>2.3M</td>
<td>44.85</td>
</tr>
<tr>
<td>CAMB-14</td>
<td>155k</td>
<td>46.04</td>
</tr>
<tr>
<td>MLE</td>
<td>720k</td>
<td>52.75</td>
</tr>
<tr>
<td>NRL</td>
<td>720k</td>
<td><strong>53.98</strong></td>
</tr>
<tr>
<td>CGMH</td>
<td>0</td>
<td>45.5</td>
</tr>
</tbody>
</table>

| 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. |
Figure 3: Overlap rates of CGMH and VAE for each word position of sentences.
The Markov Chain never mixes. We mainly use MH as SA.

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.
Simulated Annealing

- MH: Sampling from $\propto \exp\{s(x)\}$

- SA: Searching the optimum of $s(x)$
  - Define $p_\tau(x) \propto \exp\{s(x)/\tau\}$
  - Start from high temperature, but cool it down gradually
  - With $\tau \to 0$, $p_\tau(x) = 1$ if $x = \arg\max s(x)$, or 0 otherwise

$$p(\text{accept}|x_*,x_t,T) = \min\left(1, e^{\frac{f(x_*) - f(x_t)}{T}}\right)$$

$$T = \max(0, T_{\text{init}} - C \cdot t)$$

References

Thank you!

Q&A