Tsinghua University

Scalable, Controllable, and Interpretable Machine Learning for Natural Language Generation

Lei Li
ByteDance AI Lab

10/8/2020

Revolution in Information Creation and Sharing

- New media platforms

- Tremendous improvement in the efficiency and quality of content creation
- Massive distribution of personalized information

AI for Information Creation and Sharing

AI for Information Creation and Sharing

Information Creation & Sharing

AI Technology

Automated news writing

Sharing Content Globally

Filtering Misinformation

Natural Lang. Generation

Machine Translation

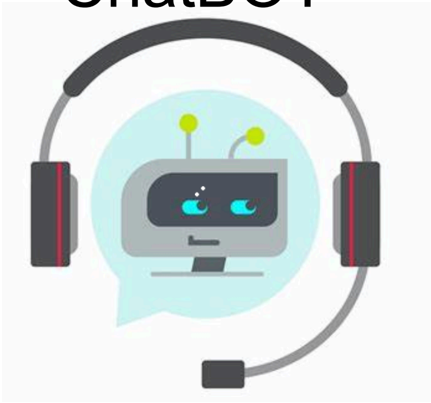
Classification/ Graph Neural Nets/ GANs

Why is NLG important?

Machine Writing

Question Answering

ChatBOT

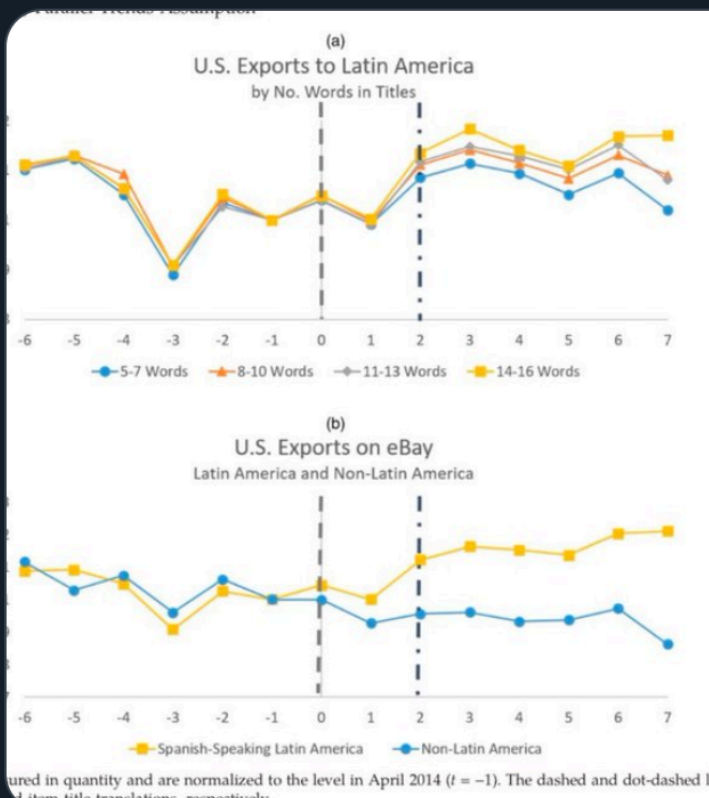


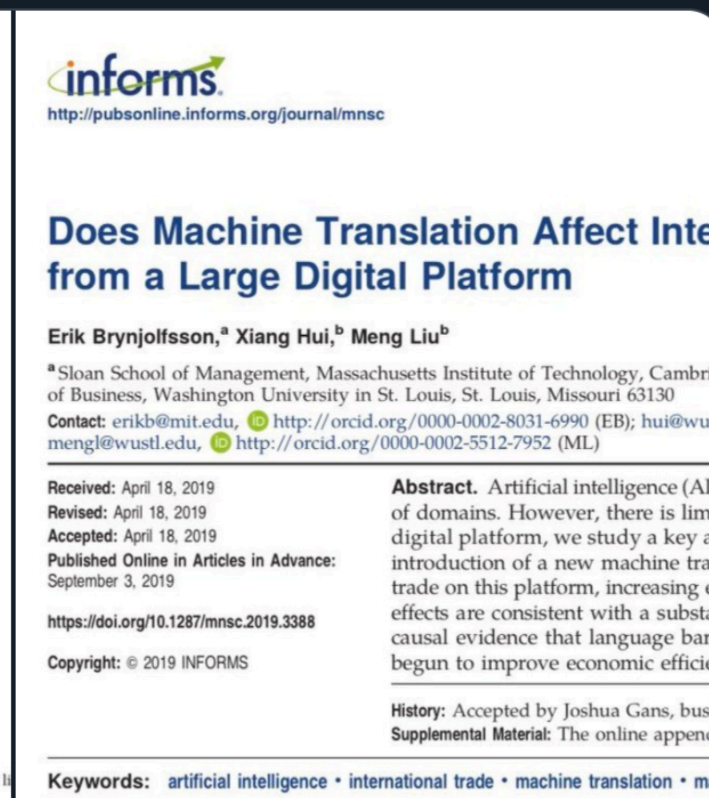
Machine Translation

Ethan Mollick @emollick

Replying to @emollick

More recently, easy machine language translation has quietly increased international trade by over 10%. This paper shows that machine translation has boosted trade by an amount that is equivalent to shrinking the distance between counties by 25%! 2/2





informs
http://pubsonline.informs.org/journal/mnsc

Does Machine Translation Affect Inte[r
from a Large Digital Platform

Erik Brynjolfsson,[a] Xiang Hui,[b] Meng Liu[b]

[a]Sloan School of Management, Massachusetts Institute of Technology, Cambrid
of Business, Washington University in St. Louis, St. Louis, Missouri 63130

Contact: erikb@mit.edu, http://orcid.org/0000-0002-8031-6990 (EB); hui@wust
mengl@wustl.edu, http://orcid.org/0000-0002-5512-7952 (ML)

Received: April 18, 2019
Revised: April 18, 2019
Accepted: April 18, 2019
Published Online in Articles in Advance: September 3, 2019

https://doi.org/10.1287/mnsc.2019.3388

Abstract. Artificial intelligence (AI)
of domains. However, there is limit
digital platform, we study a key ap
introduction of a new machine tran
trade on this platform, increasing ex
effects are consistent with a substan
causal evidence that language barri
begun to improve economic efficien

History: Accepted by Joshua Gans, busin
Supplemental Material: The online appendi

Keywords: artificial intelligence • international trade • machine translation • mac

AI to Improve Writing

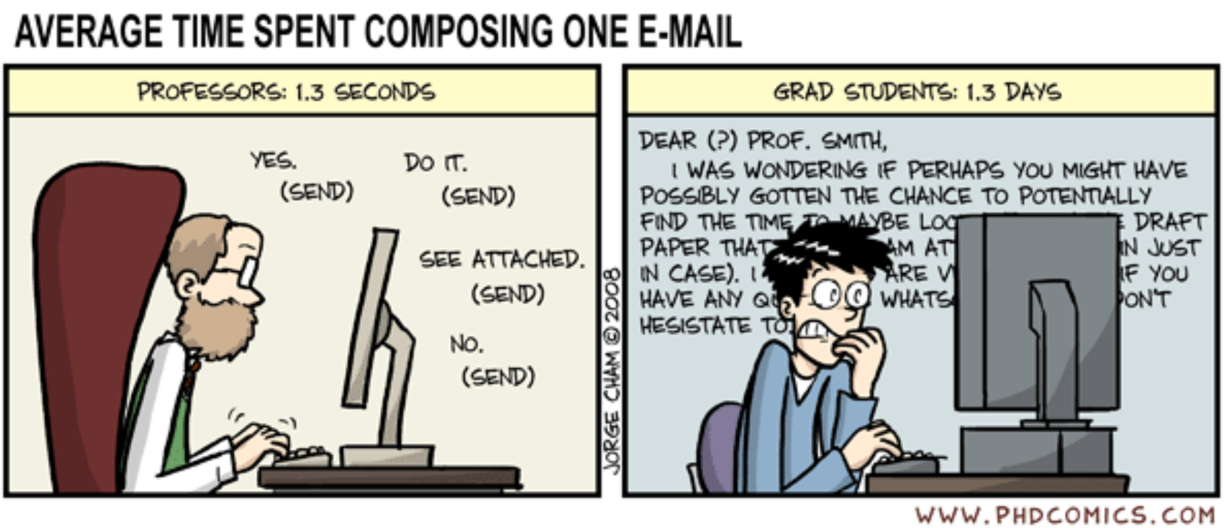
Text generation to rescue!

Gmail smart compose, smart reply

Humans Run Experiments, a Robot Writes the Paper

The future of automated scientific writing is upon us—and that's a good thing.

By Daniel Engber



The New York Times

Soon a Robot Will Be Writing This Headline

Gabriel Alcala

By Alana Semuels

Jan. 14, 2020

Automated News Writing

Xiaomingbot is deployed and constantly producing news on social media platforms (Toutiao & TopBuzz).

Xiaomingbot-European ⚽ | Following

202 Post | 4 Following | 1.1K Followers

La Liga: Real Betis suffered from an utterly embarassing ending in their 1: 4 fiasco against Barcelona

Mar 17, 2019   0

FIFA WORLD CUP RUSSIA 2018

LaLiga

Rio2016

A robot wrote this entire article. Are you scared yet, human?

Search International edition

The Guardian

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

human written

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

GPT3, edited by human

A New Working Style for Authors Human-AI Co-authoring

Outline

1. Basics of Deep Generative Models for Sequences
2. Deep Latent Variable Models
3. Monte-Carlo Methods for Constrained Text Generation
4. Multimodal machine writing: show case
5. Summary

Basics of Deep Generative Models for Sequences

How to generate a sentence?

Modeling a Sequence

The quick brown fox jumps over the lazy dog .

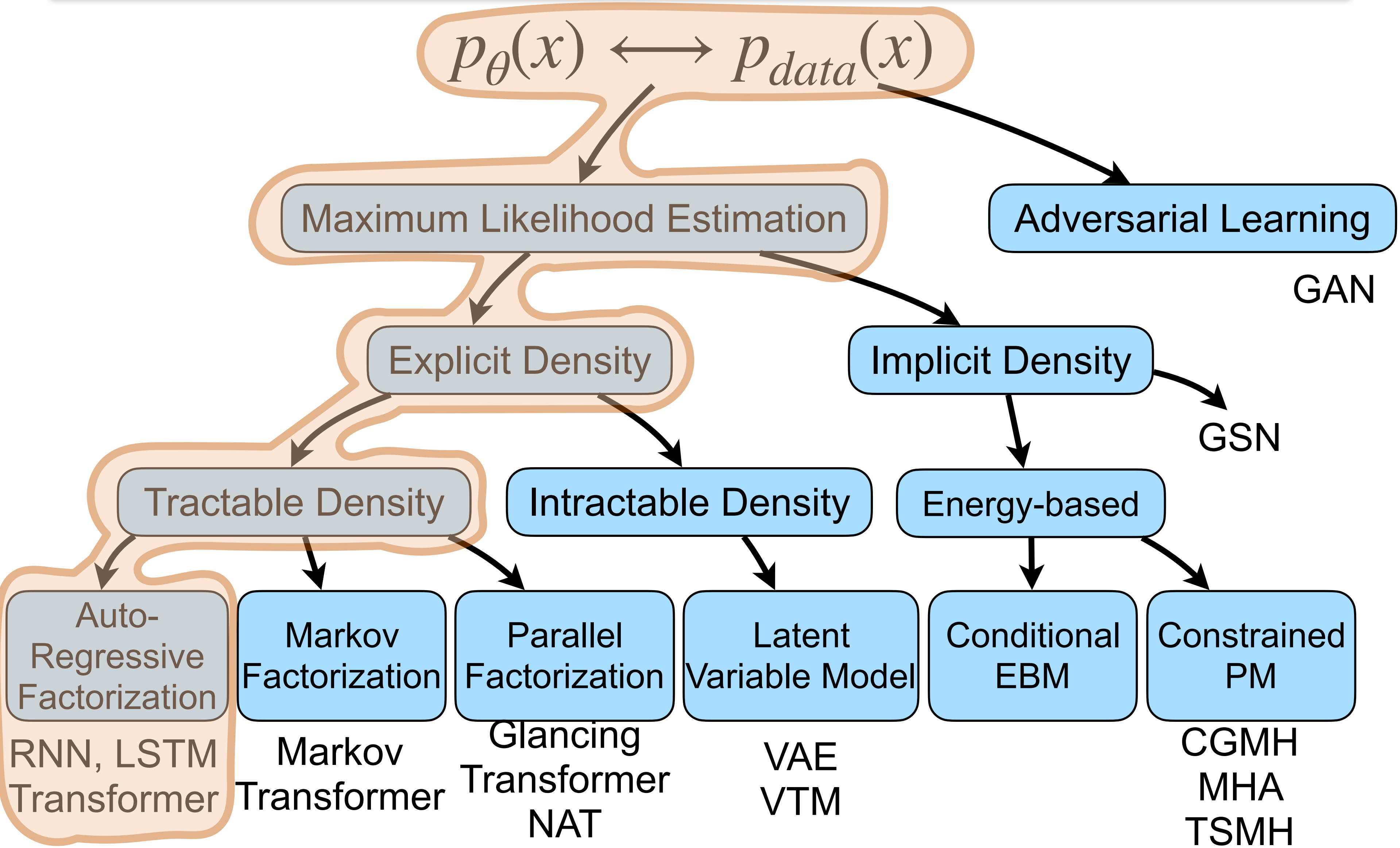
$$x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})$$

The central problem of *language modeling* is to find the *joint probability distribution*:

$$p_\theta(x) = p_\theta(x_1, \cdots, x_L)$$

There are many ways to represent and learn the joint probability model.

DGM Taxonomy

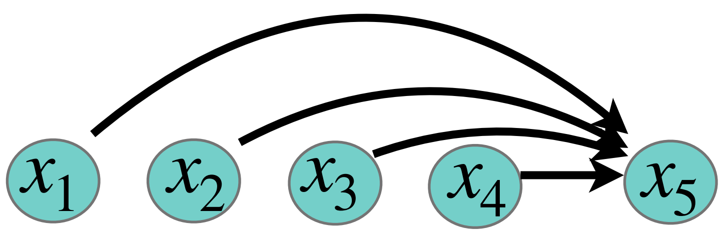


Auto-Regressive Language Model

Decompose the joint distribution as a product of tractable conditional probabilities:

Given $x = [x_1, x_2, x_3 \ldots, x_n]$

$$p_\theta = \prod_{i=1}^{n} p_\theta(x_i \mid x_1, x_2, \ldots, x_{i-1}) = \prod_{i=1}^{n} p_\theta(x_i \mid x_{<i})$$

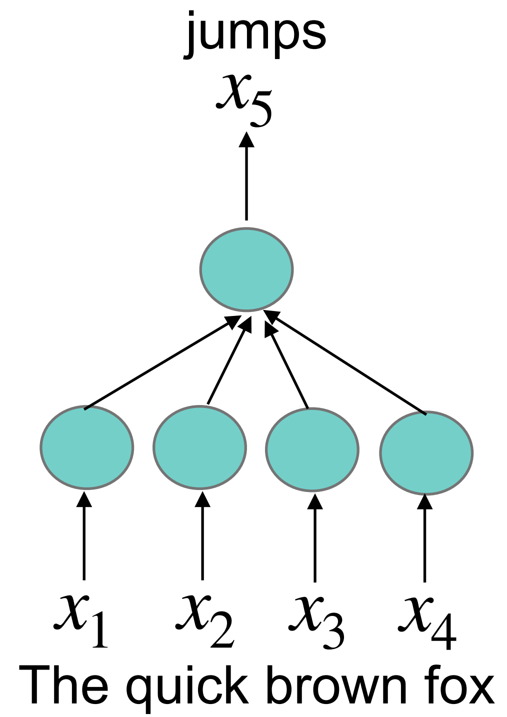


Auto-Regressive Factorization - Token Probability from a Neural Network

$$p_\theta(x_i \mid x_{<i}) = \text{Softmax}\left(f_\theta(x_{<i})\right)_{x_i}$$

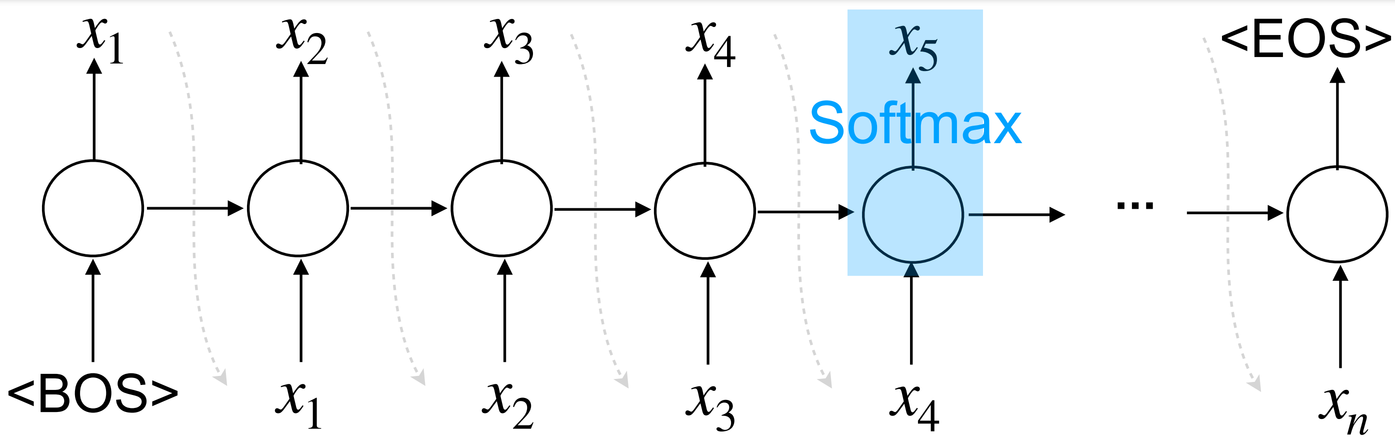
$$\text{Softmax}(\mathrm{x})_j = \frac{\exp x_j}{\sum_k \exp x_k}$$

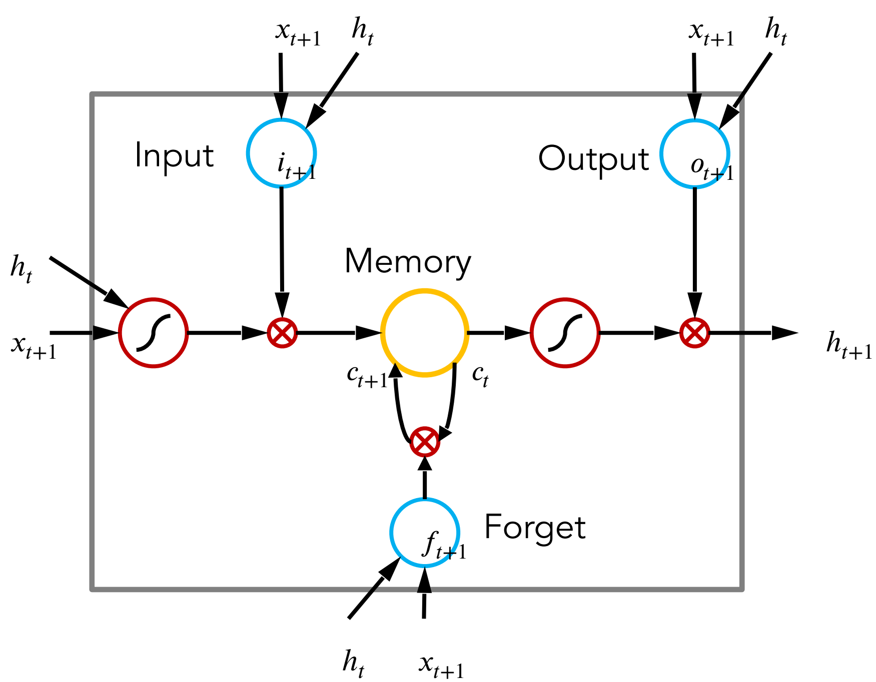
$$p_\theta(x_5 \mid x_1, x_2, x_3, x_4)$$



Auto-Regressive Factorization Parameterization by RNN/LSTM

$p_\theta(x_i \mid x_{<i})$





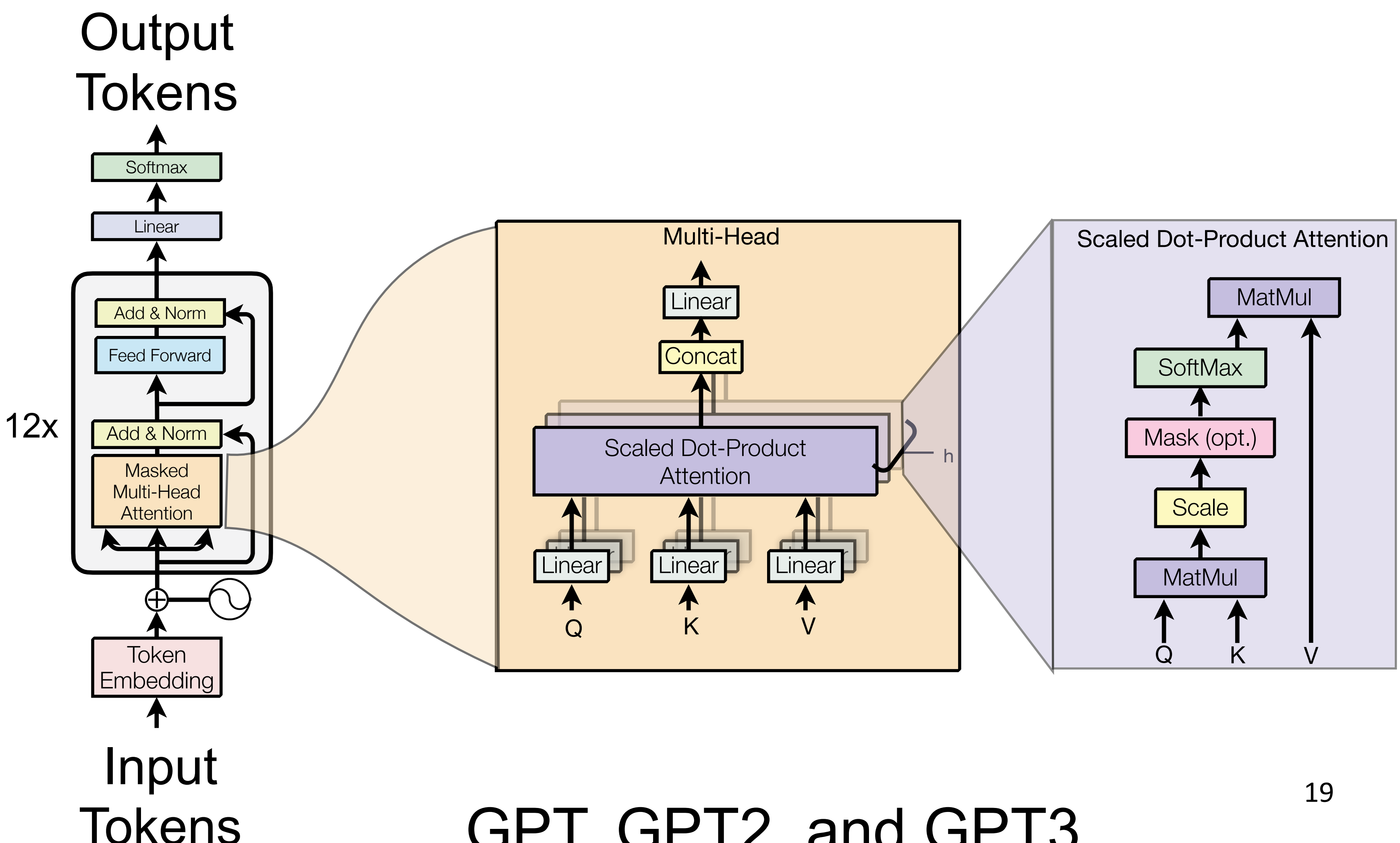
Adaptively memorize short and long term information

$$\begin{pmatrix} i_{t+1} \\ f_{t+1} \\ o_{t+1} \\ a_{t+1} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \odot \left(M \cdot \begin{pmatrix} x_{t+1} \\ h_t \end{pmatrix} + b \right)$$

$$c_{t+1} = f_{t+1} \otimes c_t + i_{t+1} \otimes a_{t+1}$$

$$h_{t+1} = o_{t+1} \otimes \tanh(c + t + 1)$$

Auto-Regressive Factorization Parameterization by Transformer

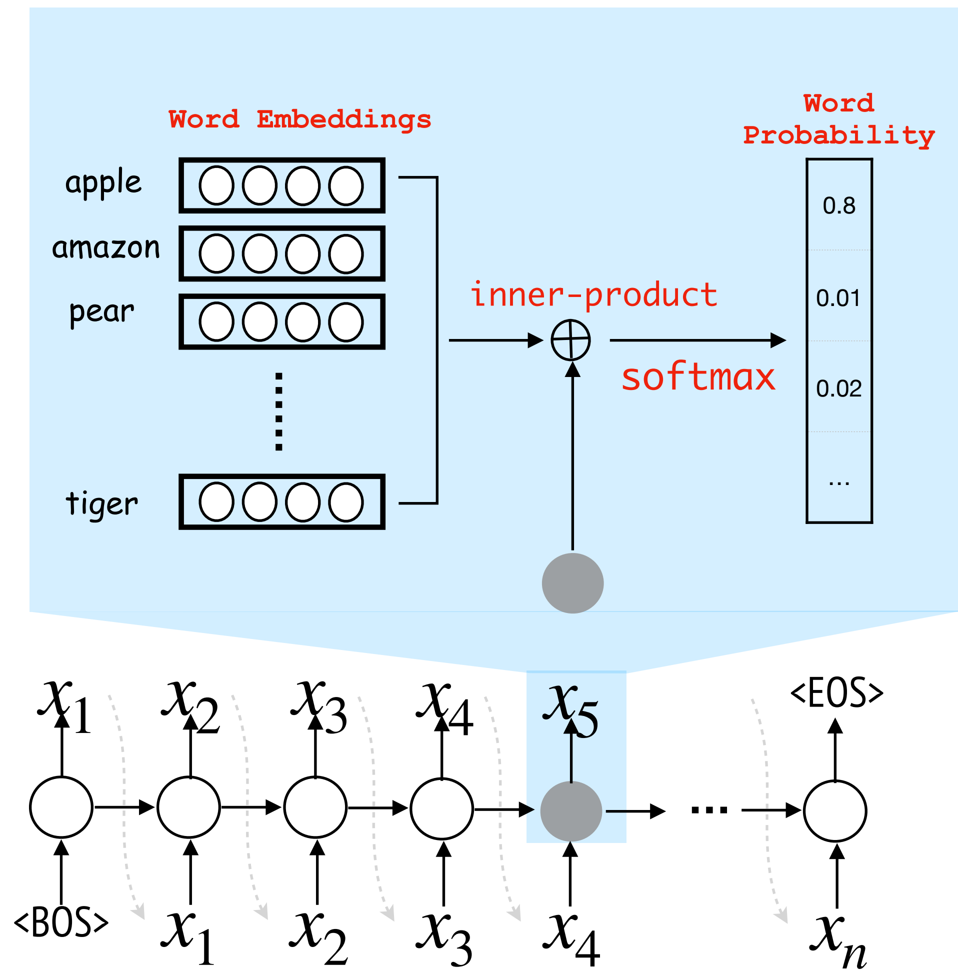


GPT, GPT2, and GPT3

What is Softmax essentially Computing?

softmax

$$p_\theta(x_i \mid x_{<i})$$



Training Objective

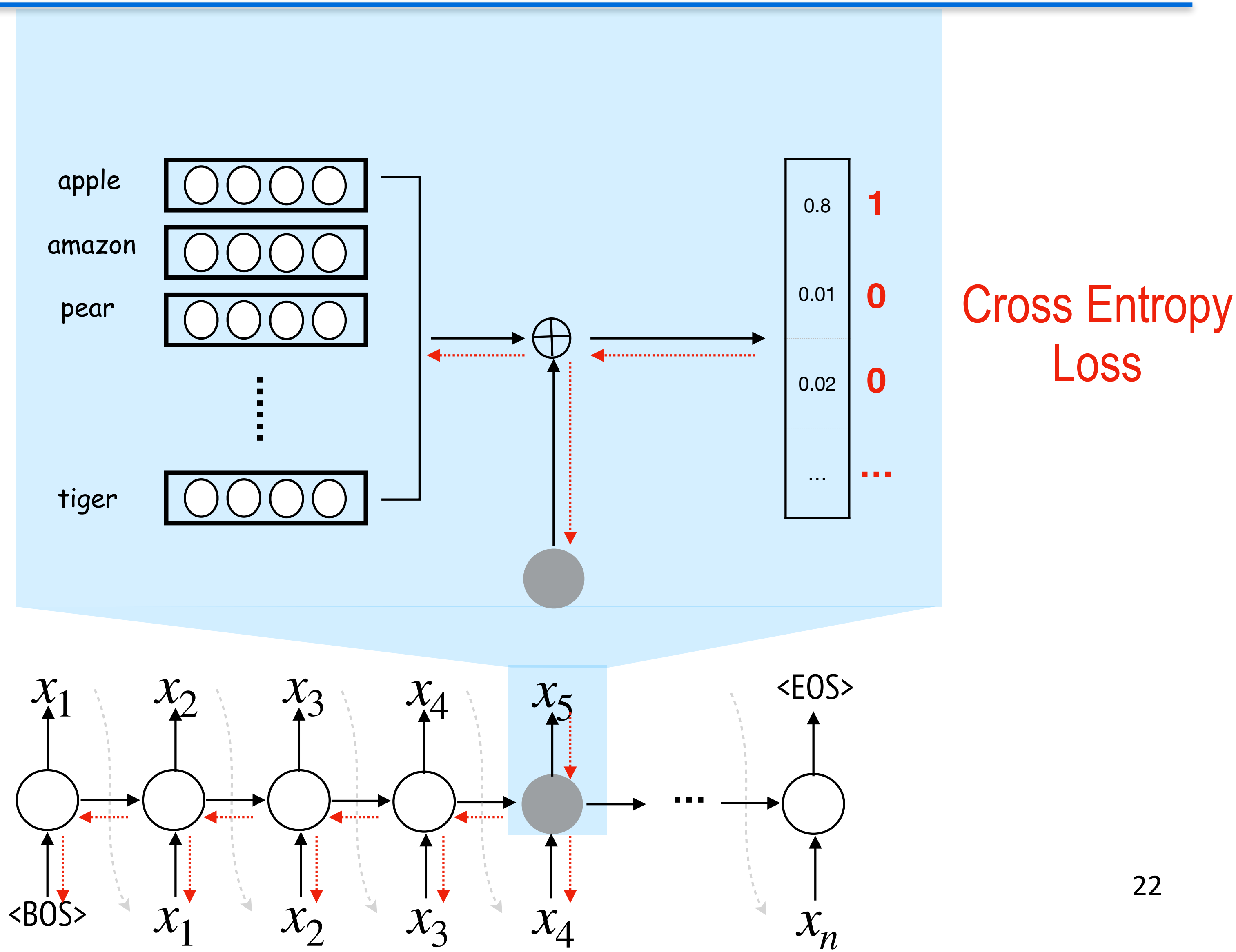
Maximum Likelihood Estimation (or Cross-Entropy loss):

$$\min \mathbb{E}_{x \sim p_{data}} \left[-\log p_\theta(x) \right]$$

$$p_\theta(x) = \prod_{i=1}^{n} p_\theta(x_i \,|\, x_1, x_2, \ldots, x_{i-1}) = \prod_{i=1}^{n} p_\theta(x_i \,|\, x_{<i})$$

Parameterization by RNN/LSTM/Transformer

Training: Back-propagation Algorithm



Conditional Sequence Generation

aka. sequence-to-sequence generation

- Machine Translation
- Dialog Generation
- Question Answering
- …

敏捷的棕狐跳过懒狗

Output

$$p_\theta(y \mid x)$$

Input

The quick brown fox jumps over the lazy dog .

Conditional Sequence Generation

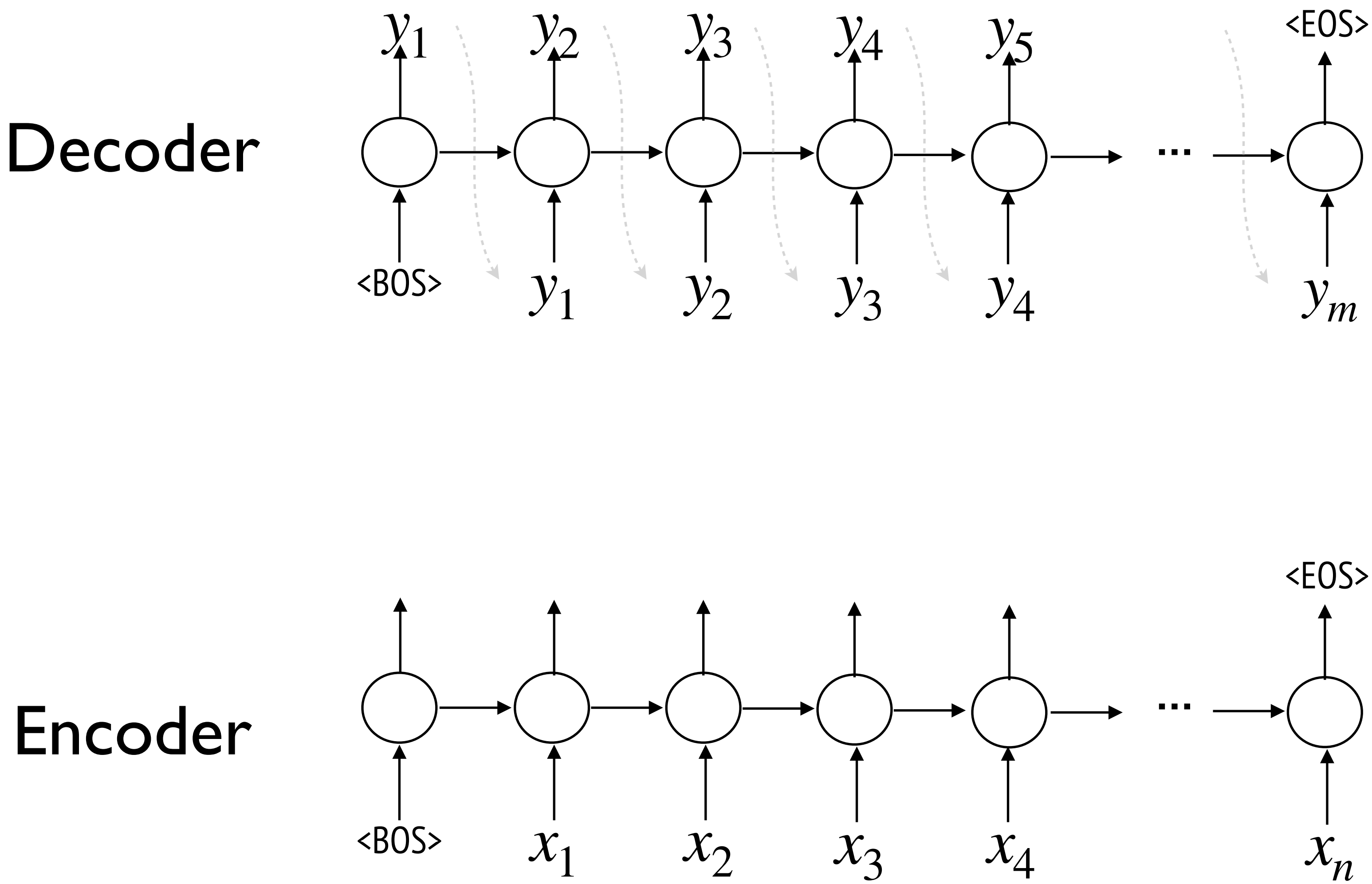
<u>Maximum Likelihood Estimation (or Cross-Entropy loss):</u>

$$\min \mathbb{E}_{x \sim p_{data}} \left[-\log p_\theta(y \mid x) \right]$$

$$p_\theta(y \mid x) = \prod_{i=1}^{n} p_\theta(y_i \mid y_1, y_2, \ldots, y_{i-1}, x) = \prod_{i=1}^{n} p_\theta(y_i \mid y_{<i}, x)$$

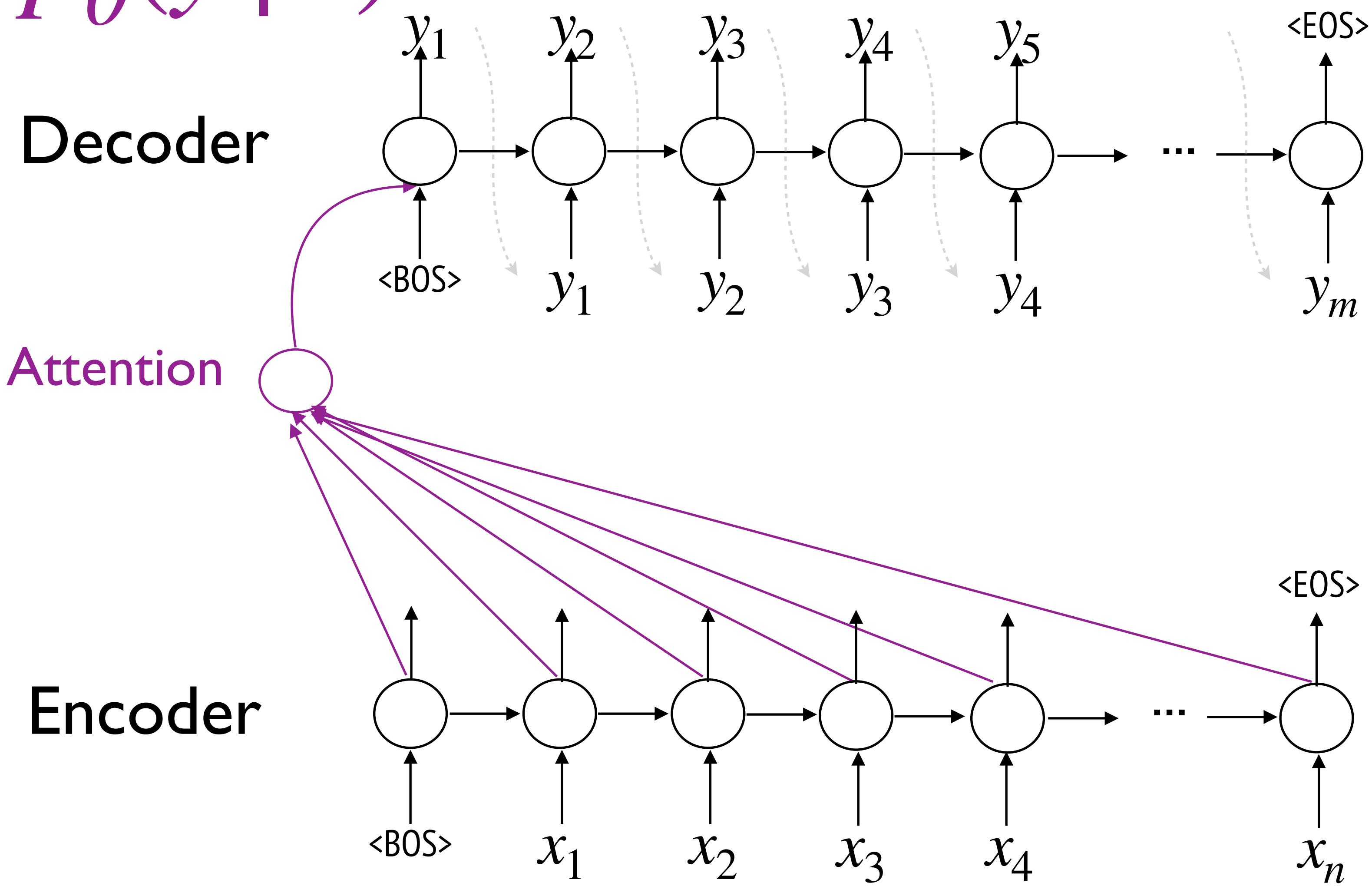
Parameterization by Transformer
or LSTM-seq2seq

Conditional Sequence Generation

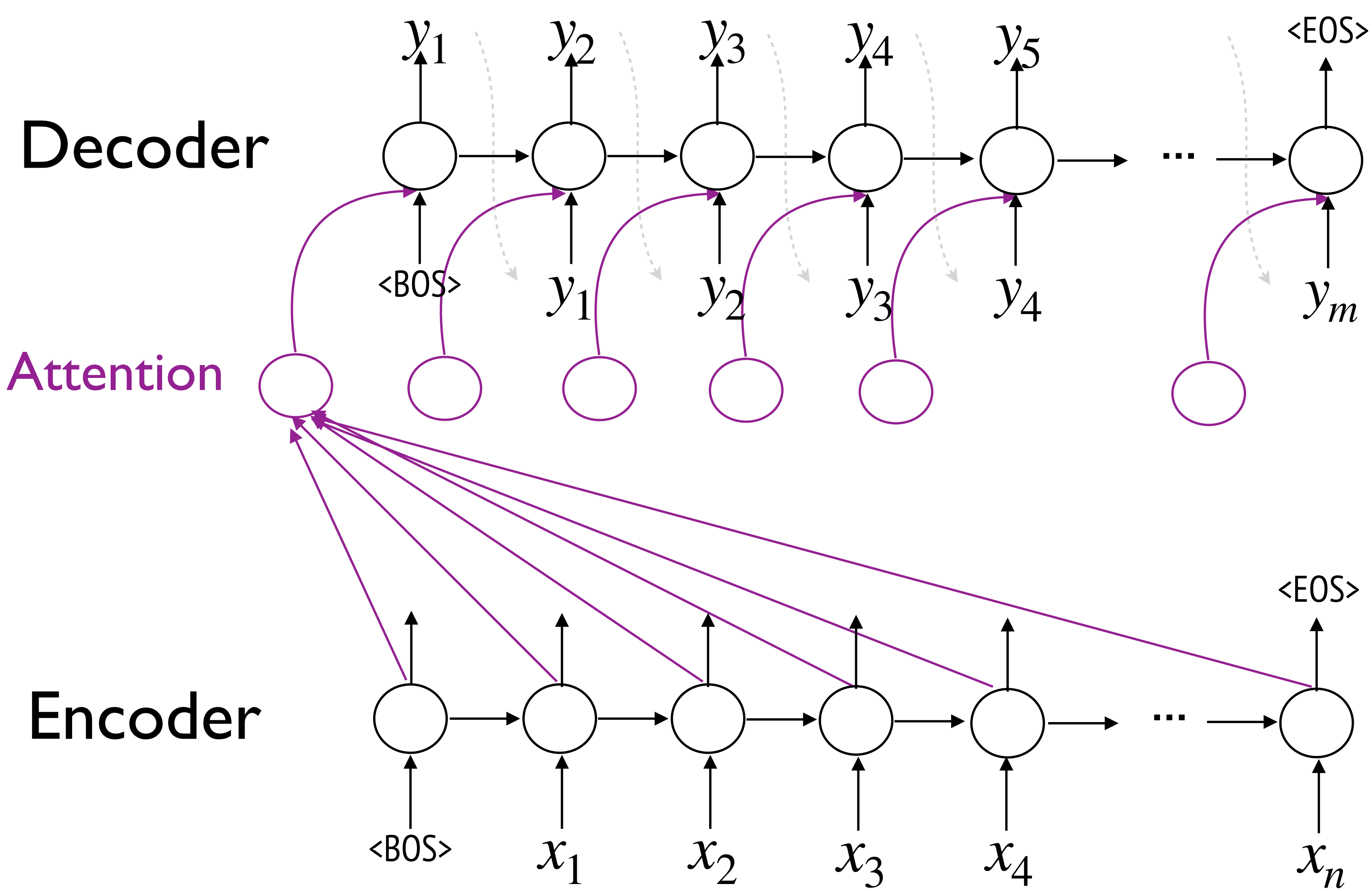


Conditional Sequence Generation

$p_\theta(y \mid x)$

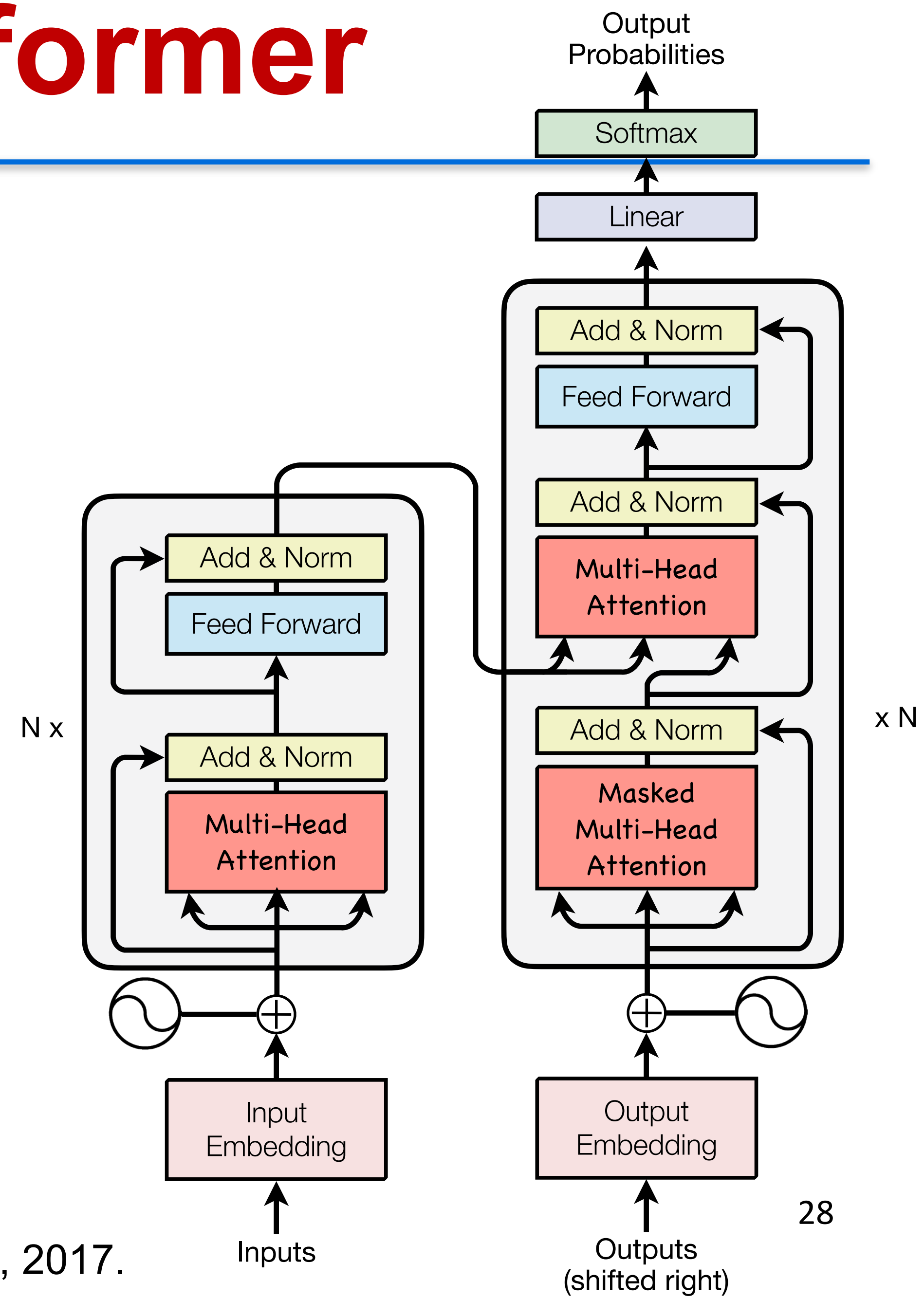


Conditional Sequence Generation



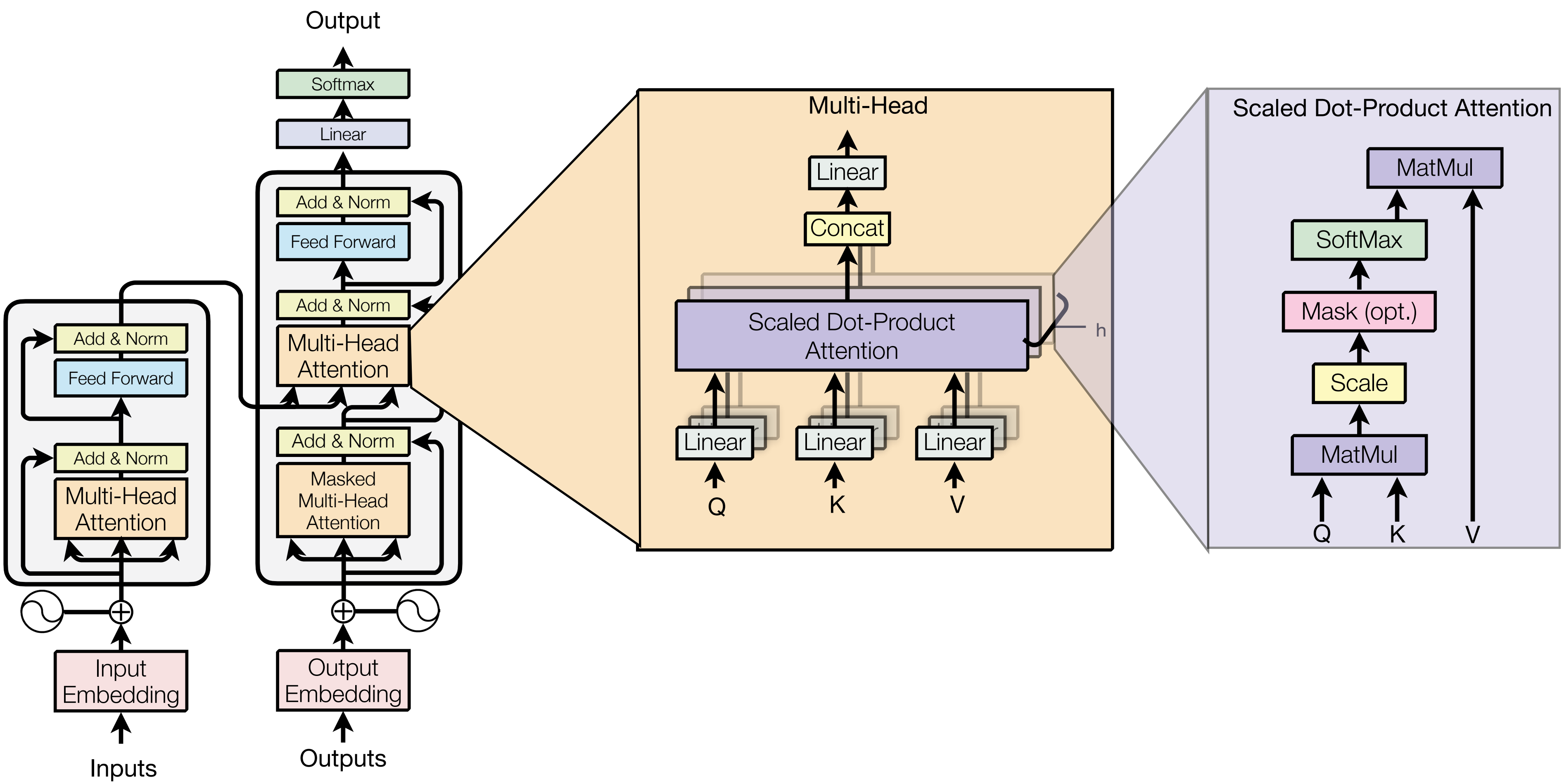
Transformer

Transformer abandons RNN by using Multi-head Self-Attention!



Vaswani et al., Attention is all you need, in NIPS, 2017.

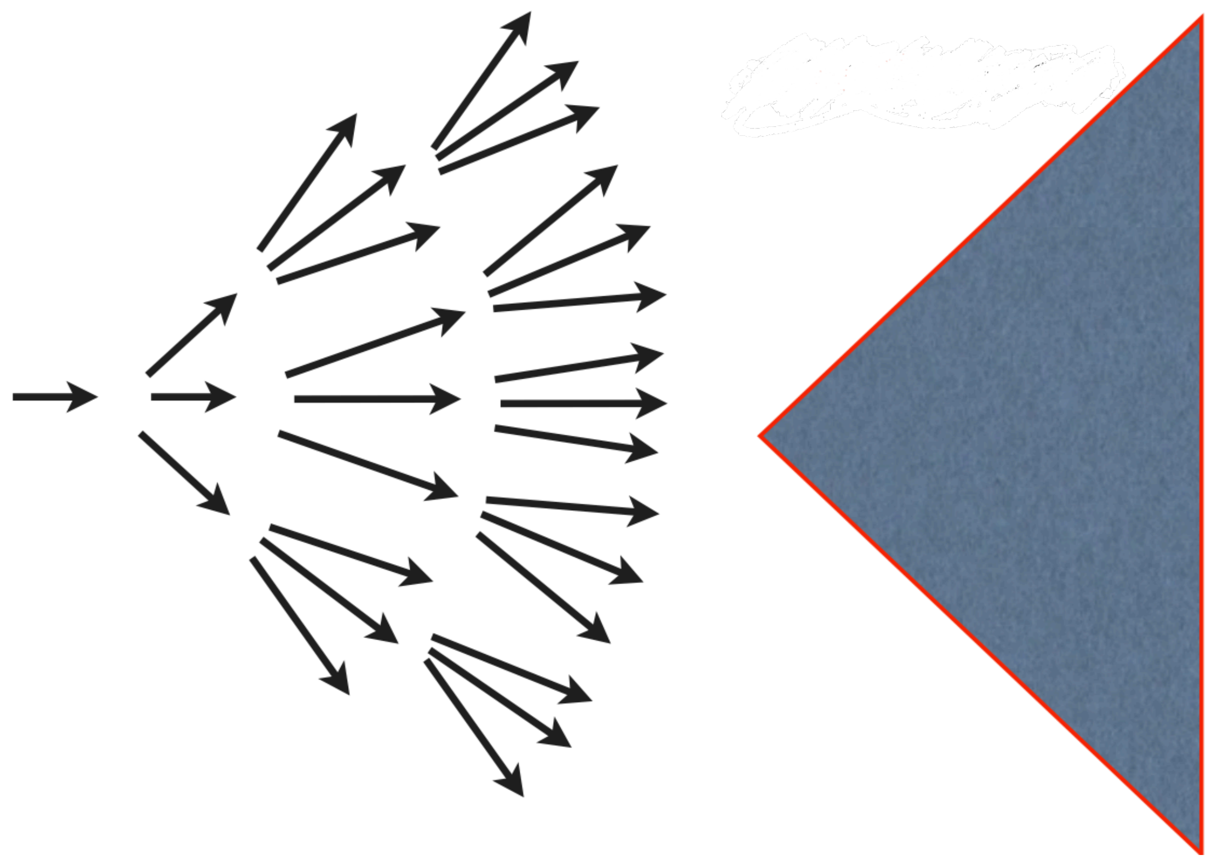
Multi-Head Attention



The Decoding Problem

$$\log p_\theta(x \mid y) = \sum_{i=1}^{n} \log p_\theta(x_i \mid x_1, x_2, \ldots, x_{i-1}, y) = \sum_{i=1}^{n} \log p_\theta(x_i \mid x_{<i}, y)$$

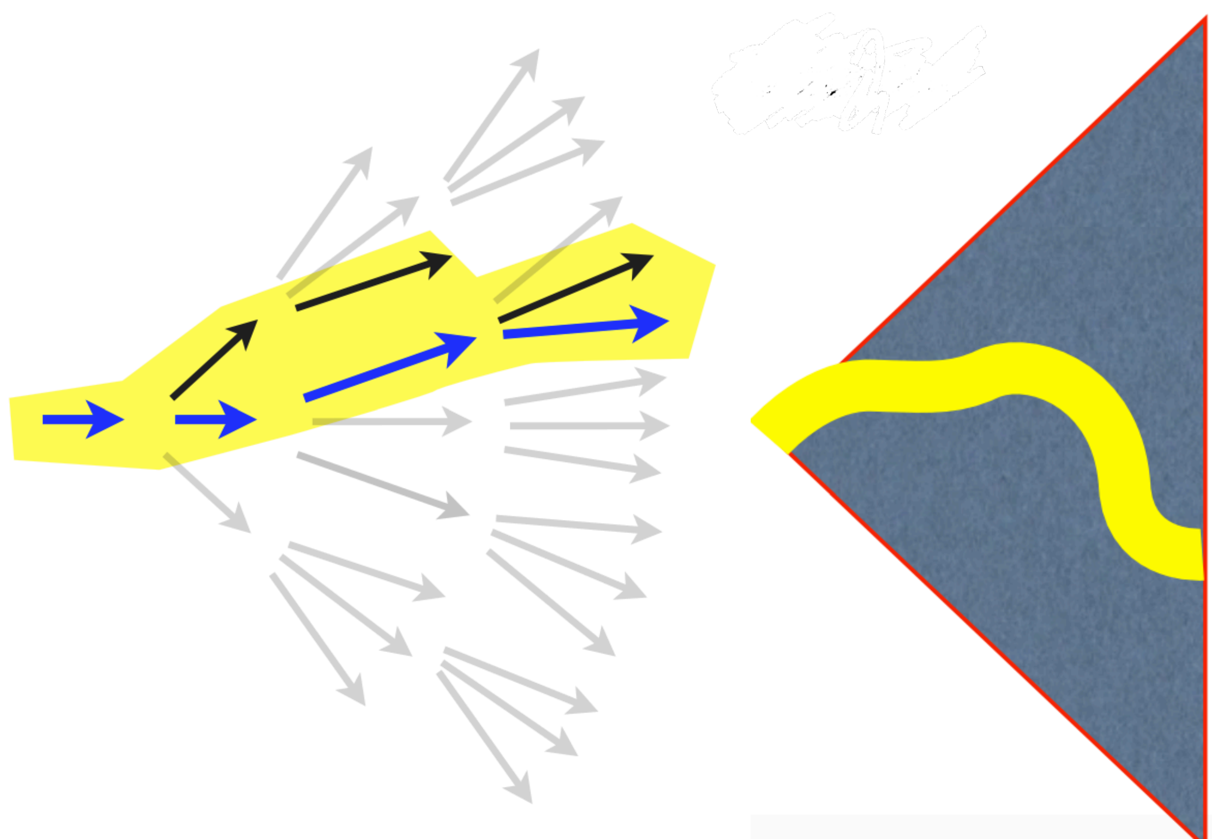
Decoding space is still exponential



Approximate Decoding: Beam Search

$$\log p_\theta(x \mid y) = \sum_{i=1}^{n} \log p_\theta(x_i \mid x_1, x_2, \ldots, x_{i-1}, y) = \sum_{i=1}^{n} \log p_\theta(x_i \mid x_{<i}, y)$$

Heuristic decoding by beam search: keeping k-best at each step and incrementally updating



Machine Translation Performance

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [17]	23.75			
Deep-Att + PosUnk [37]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [36]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [31]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [37]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [36]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Though no long the state-of-the-art result today, Transformer is the default backbone model.

Attention is all you need, Vaswani et al, NIPS 2017

Outline

1. Basics of Deep Generative Models for Sequences
2. Deep Latent Variable Models
3. Monte-Carlo Methods for Constrained Text Generation
4. Multimodal machine writing: show case
5. Summary

Deep Latent Variable Models for Text

VTM [R. Ye, W. Shi, H. Zhou, Z. Wei, **Lei Li**, ICLR20b]
DSS-VAE [Y. Bao, H. Zhou, S. Huang, **Lei Li**, L. Mou, O. Vechtomova, X. Dai, J. Chen, ACL19c]
DEM-VAE [W. Shi, H. Zhou, N. Miao, **Lei Li**, ICML 2020]
MGNMT [Z. Zheng, H. Zhou, S. Huang, **Lei Li**, X. Dai, J. Chen, ICLR 2020a]

Outline

- Disentangled Representation Learning for Text Generation
- Interpretable Deep Latent Representation from Raw Text
- Mirror Generative Model for Neural Machine Translation

Natural Language Descriptions

name	Sukiyaki
eatType	pub
food	Japanese
price	average
rating	good
area	seattle

Sukiyaki is a Japanese restaurant. It is a pub and it has a average cost and good rating. It is based in seattle.

Data to Text Generation

Data Table
<key, value>

Sentence

Medical Reports

The blood pressure is higher than normal and may expose to the risk of hypertension

Style	long dress
Painting	bamboo ink
Texture	poplin
Feel	smooth

Fashion Product Description

Made of poplin, this long dress has an ink painting of bamboo and feels fresh and smooth.

Name: Sia Kate Isobelle Furler
DoB: 12/18/1975
Nationality: Australia
Occupation: Singer, Songwriter

Person Biography

Sia Kate Isobelle Furler (born 18 December 1975) is an Australian singer, songwriter, voice actress and music video director.

[1] The E2E Dataset: New Challenges For End-to-End Generation. https://github.com/tuetschek/e2e-dataset
[2] Can Neural Generators for Dialogue Learn Sentence Planning and Discourse Structuring?. https://nlds.soe.ucsc.edu/sentence-planning-

Problem Setup

- Inference:
  - Given: table data x, as key-position-value triples.
  - e.g. Name: Jim Green => (Name, 0, Jim), (Name, 1, Green)
  - Output: **fluent**, **accurate** and **diverse** text sequences y
- Training:
  - $\{\langle x_i, y_i \rangle\}_{i=1}^{N}$: pairs of table data and text.
  - $\{y_j\}_{j=1}^{M}$: raw text corpus. $M \gg N$

Why is Data-to-Text Hard?

- Desired Properties:
  - Accuracy: semantically consistent with the content in the table
  - Diversity: Ability to generate infinite varying utterances
- Scalability: real-time generation, latency, throughput (QPS)
- Training: limited table-text pairs

Previous Idea: Templates

[name] is a [food] restaurant. It is a [eatType] and it has a [price] cost and [rating] rating. It is in [area].

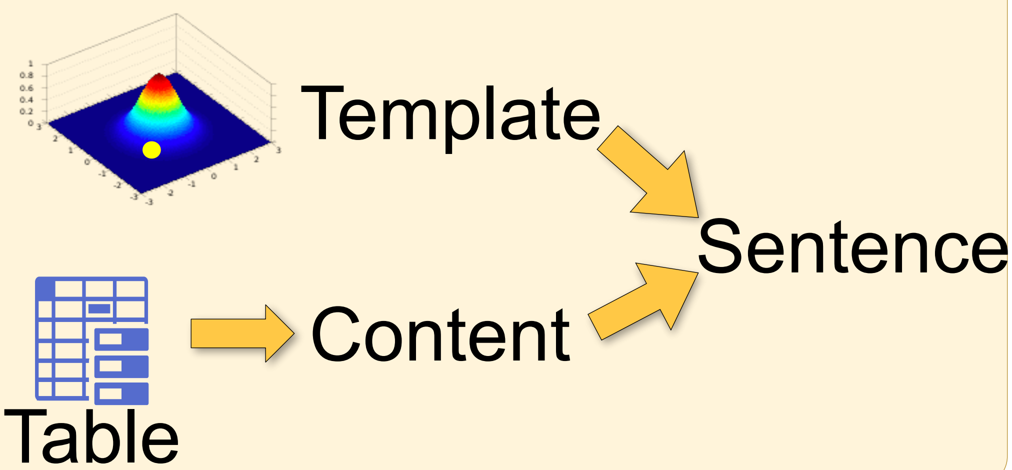
name	Sukiyaki
eatType	pub
food	Japanese
price	average
rating	good
area	seattle

Sukiyaki is a Japanese restaurant. It is a pub and it has a average cost and good rating. It is in seattle.

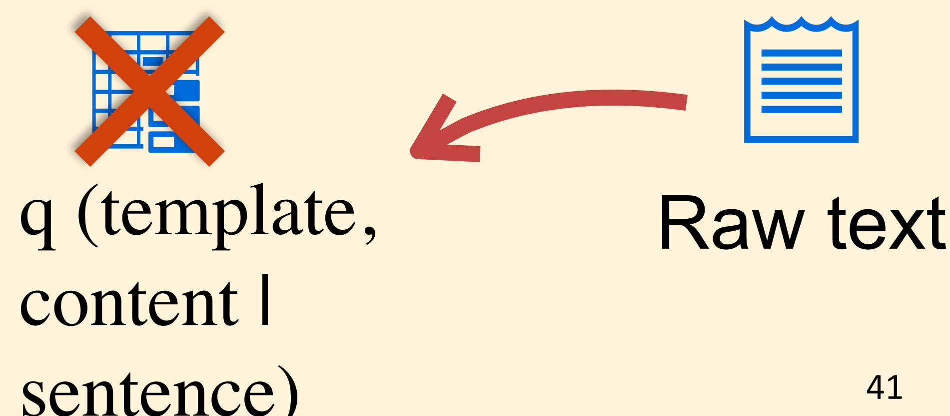
But manually creation of templates are tedious

Our Motivation for Variational Template Machine

Motivation 1:
Continuous and disentangled representation for template and content

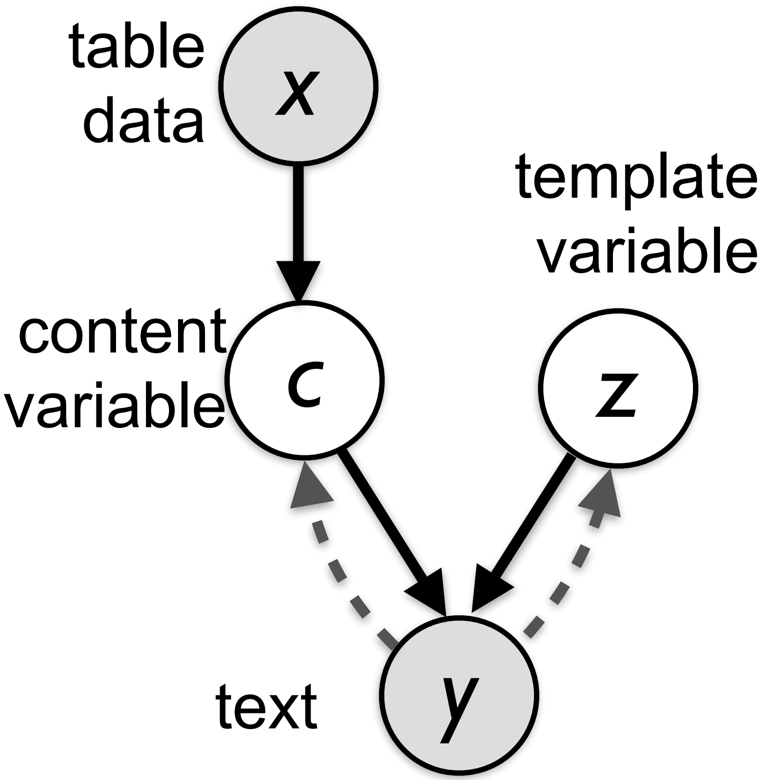


Motivation 2:
Incorporate raw text corpus to learn good representation.



VTM [R. Ye, W. Shi, H. Zhou, Z. Wei, **Lei Li**, ICLR20b]

Variational Template Machine



Input: triples of <field_name, position, value>

$$\{x_k^f, x_k^p, x_k^v\}_{k=1}^K$$

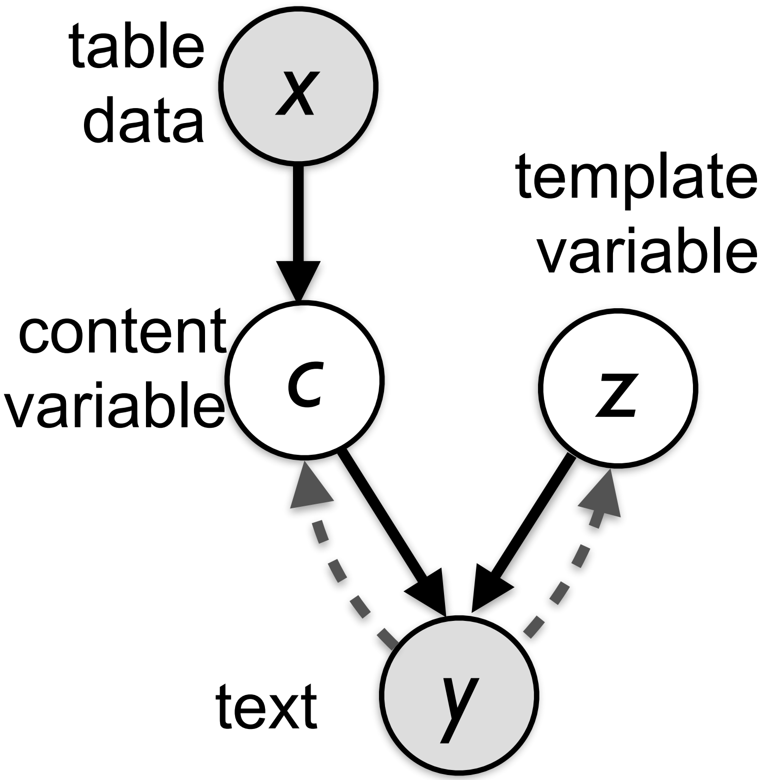
1. $p(c \mid x) \sim$ Neural Net

$$\mathrm{maxpool}\big(\tanh(W \cdot [x_f^k, x_p^k, x_v^k] + b)\big)$$

2. Sample $z \sim p_0(z)$, e.g. Gaussian
3. Decode y from $[c, z]$ using another NN (e.g. Transformer)

VTM [R. Ye, W. Shi, H. Zhou, Z. Wei, **Lei Li**, ICLR20b]

Training VTM



Key idea: Disentangling content and templates while preserving as much information as possible!

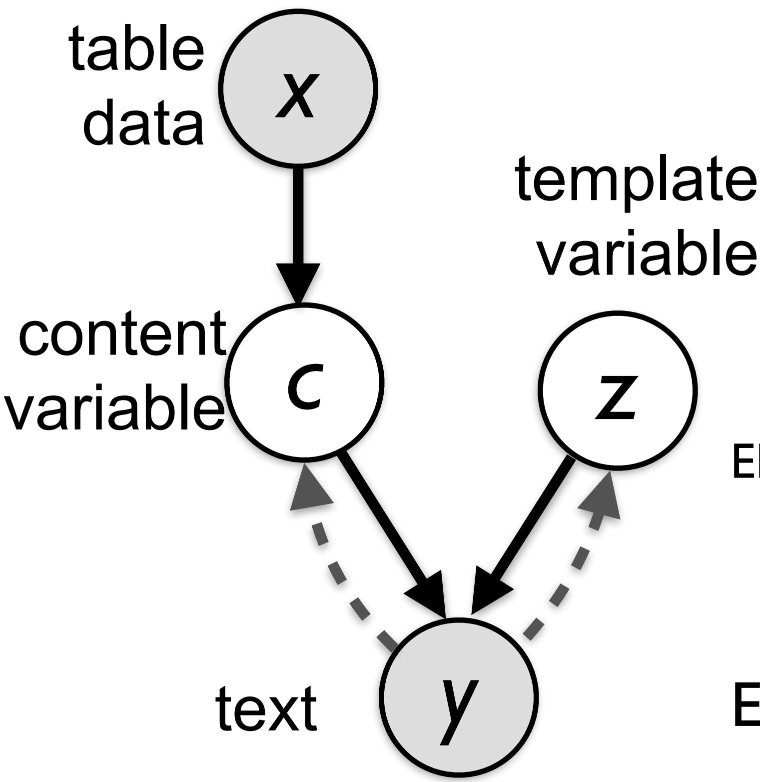
Total loss =

Reconstruction loss

+

Information-Preserving loss

Variational Inference



Instead of optimizing exact and intractable expected likelihood, minimizing the (tractable) variational lower bounds.

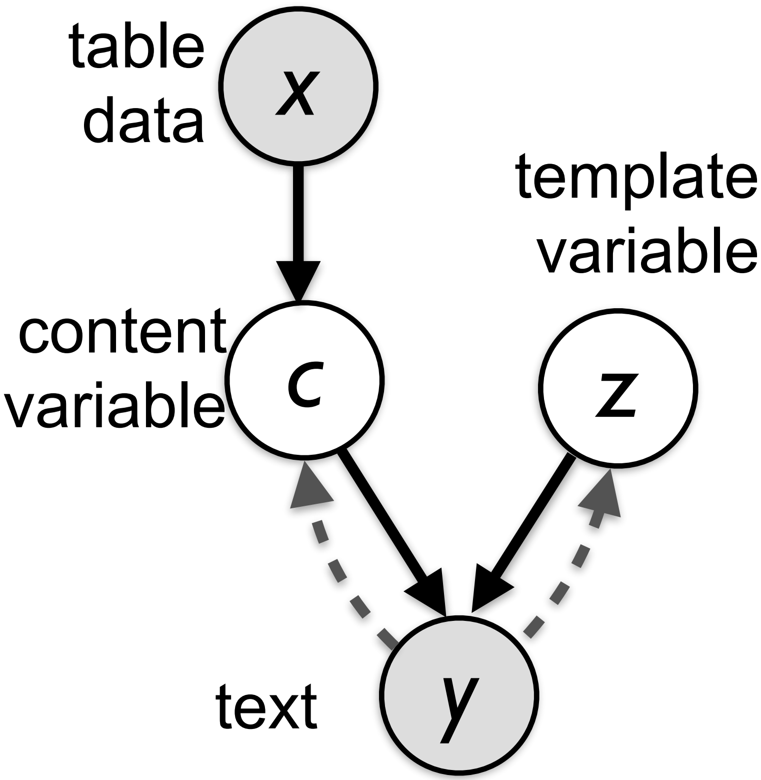
$$\cancel{l_p = -\mathrm{E}\log\int p\left(y \middle| c(x), z\right) p(z)\mathrm{d}z}$$

$$\mathrm{ELBO_p} = -\mathrm{E}_{q(z|y)}\log p\left(y \middle| c(x), z\right) + \mathrm{KL}[q(z|y)\,||\,p(z)]$$

$$\cancel{l_r = -\mathrm{E}\log\iint p\left(y \middle| c, z\right) p(z)p(c)\mathrm{d}z\mathrm{d}c}$$

$$\mathrm{ELBO_r} = -\mathrm{E}_{q(z|y)q(c|y)}\log p\left(y \middle| c, z\right) + \mathrm{KL}[q(z|y)\,||\,p(z)] + \mathrm{KL}[q(c|y)\,||\,q(c)]$$

Preserving Content & Template



1. Content preserving loss

$$l_{cp} = \mathbb{E}_{q(c|y)} | c - f(x) |^2 + D_{KL}\left(q(c \mid y) \parallel p(c)\right)$$

2. Template preserving loss of pairs

$$l_{tp} = -\,\mathbb{E}_{q(z|y)}\left[\log p(\tilde{y} \mid z, x)\right]$$

$\tilde{y}$ is the text sketch by removing table entry

i.e. cross entropy of variational prediction from templates

Preserving Template

Ensure the template variable could recover the text sketch

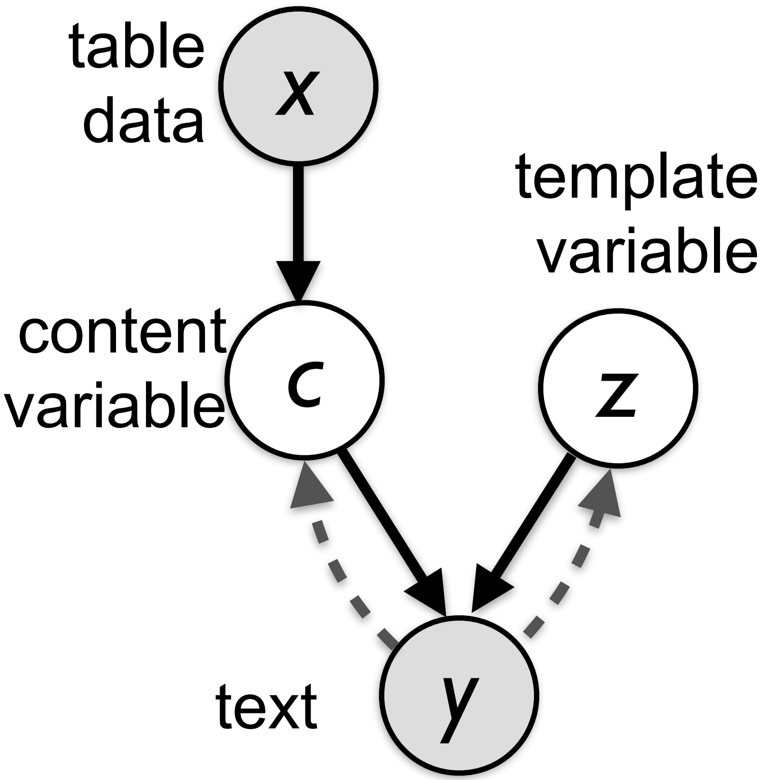


Table data x:

{name[Loch Fyne], eatType[restaurant], food[French] price[below $20]}

Text y:

Loch Fyne is a French restaurant catering to a budget of below $20.

Text Sketch $\tilde{y}$:

<ent> is a *<ent>* *<ent>* catering to a budget of *<ent>*.

Learning with Raw Corpus

- Semi-supervised learning: “Back-translate” corpus to obtain pseudo-parallel pairs <table, text>, to enrich the learning

Table	Text
name: Sukiyaki eatType: pub food: Japanese price: average rating: good area: seattle	Sukiyaki is a Japanese restaurant. It is a pub and it has a average cost and good rating. It is in seattle.
? $q(<c,z> \mid y)$	Known for its creative flavours, Holycrab's signatures are the Hokkien crab.

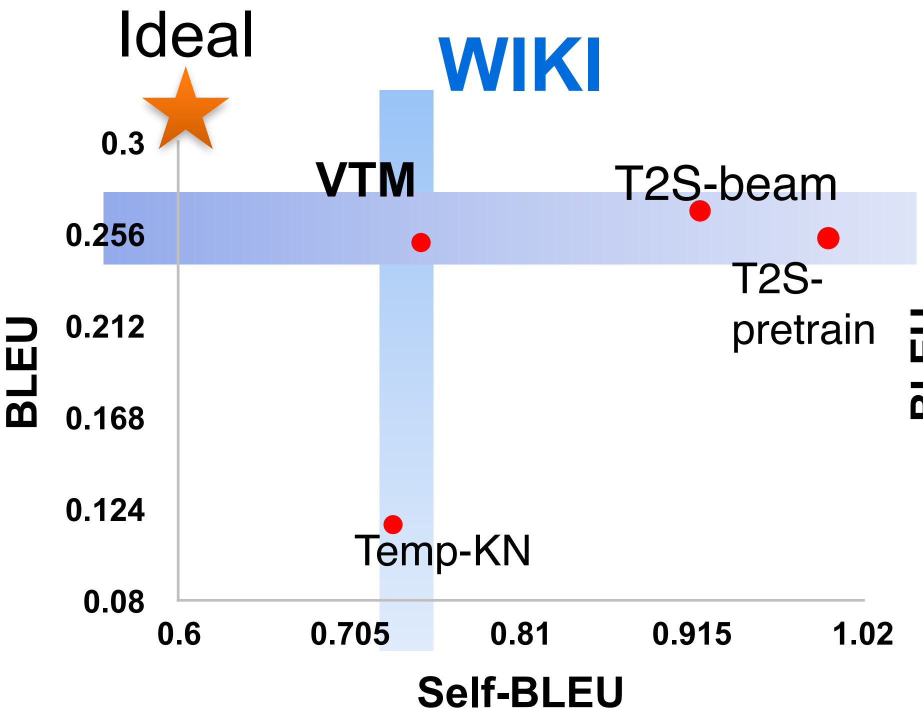
Evaluation Setup

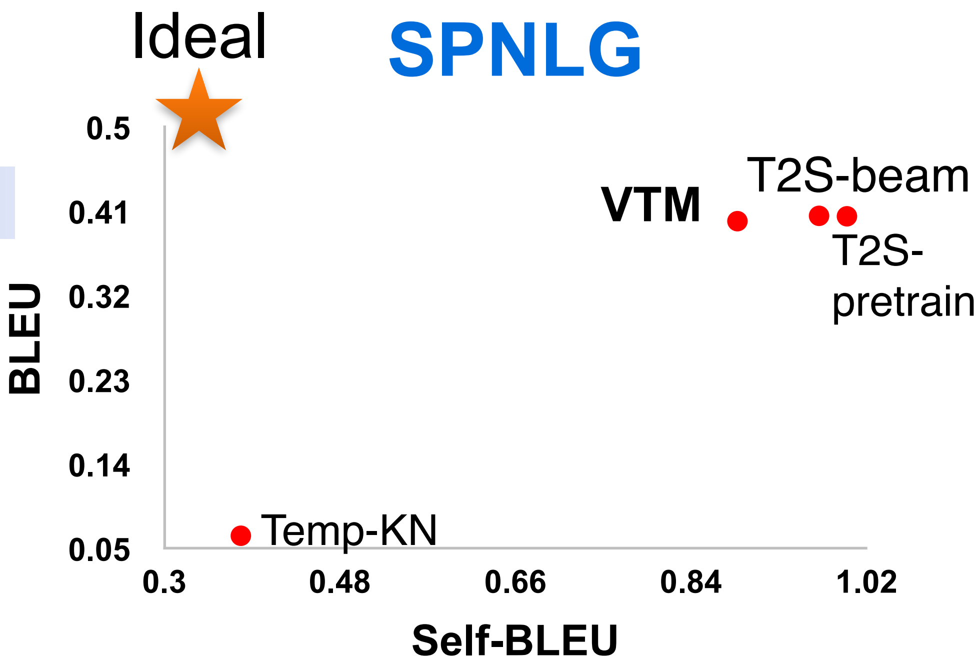
- Tasks
  - WIKI: generating short-bio from person profile.
  - SPNLG: generating restaurant description from attributes

Dataset	Train		Valid		Test
	table-text pairs	raw text	table-text pairs	raw text	table-text pairs
WIKI	84k	842k	73k	43k	73k
SPNLG	14k	150k	21k	/	21k

- Evaluation Metric:
  - Quality (Accuracy): BLEU score to ground-truth
  - Diversity: self-BLEU (lower is better)

VTM Produces High-quality and Diverse Text



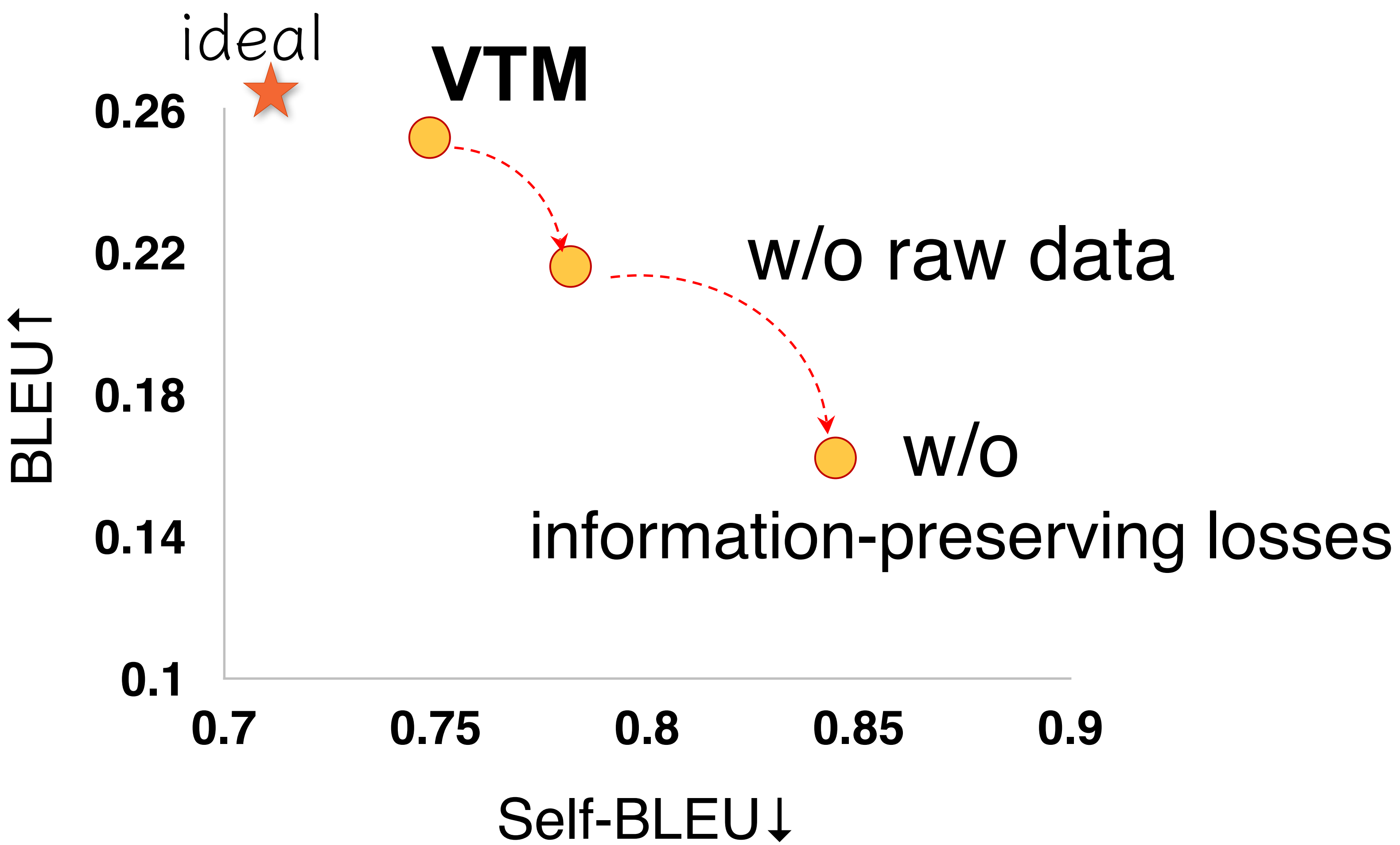


VTM uses beam-search decoding.

VTM [Ye, …, **Lei Li**, ICLR20b]

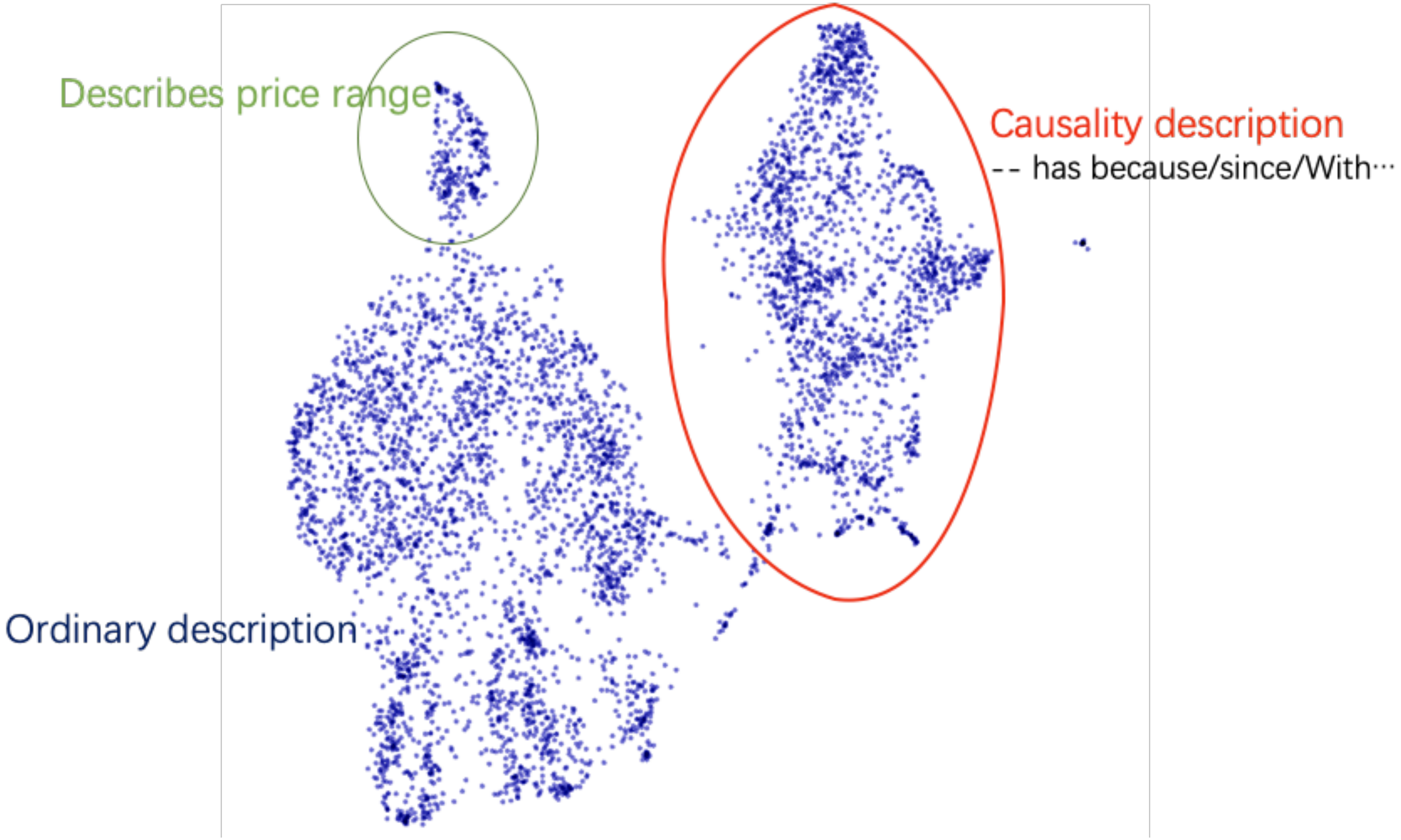
Raw data and loss terms are necessary

Ablation results on Wiki-bio dataset



Interpreting VTM

Template variable project to 2D



VTM Generates Diverse Text

Input Data Table

Jack Ryder

Ryder in about 1930

Personal information	
Full name	John Ryder
Born	8 August 1889 Collingwood, Victoria, Australia
Died	3 April 1977 (aged 87) Fitzroy, Victoria, Australia
Nickname	The King of Collingwood
Height	1.85 m (6 ft 1 in)
Batting	Right-handed
Bowling	Right-arm medium pace
Role	All-rounder

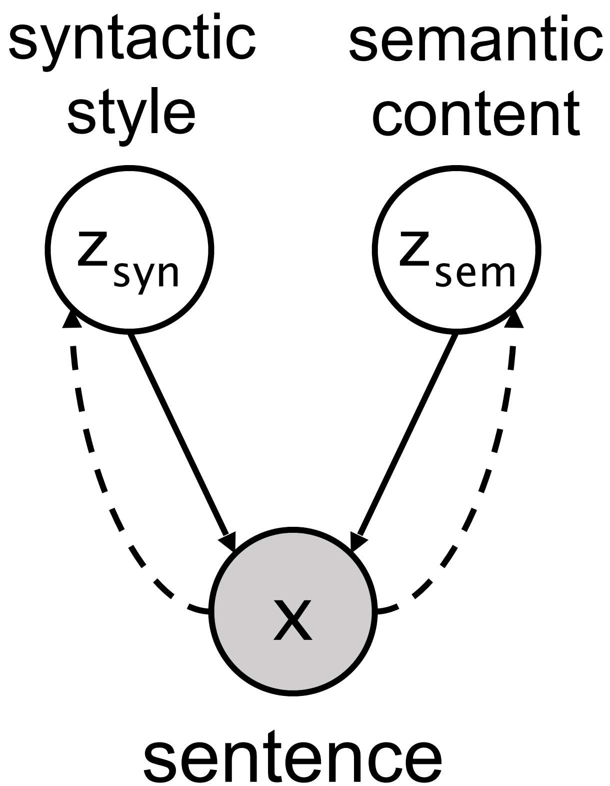
Generated Text

1: John Ryder (8 August 1889 – 4 April 1977) was an Australian cricketer.

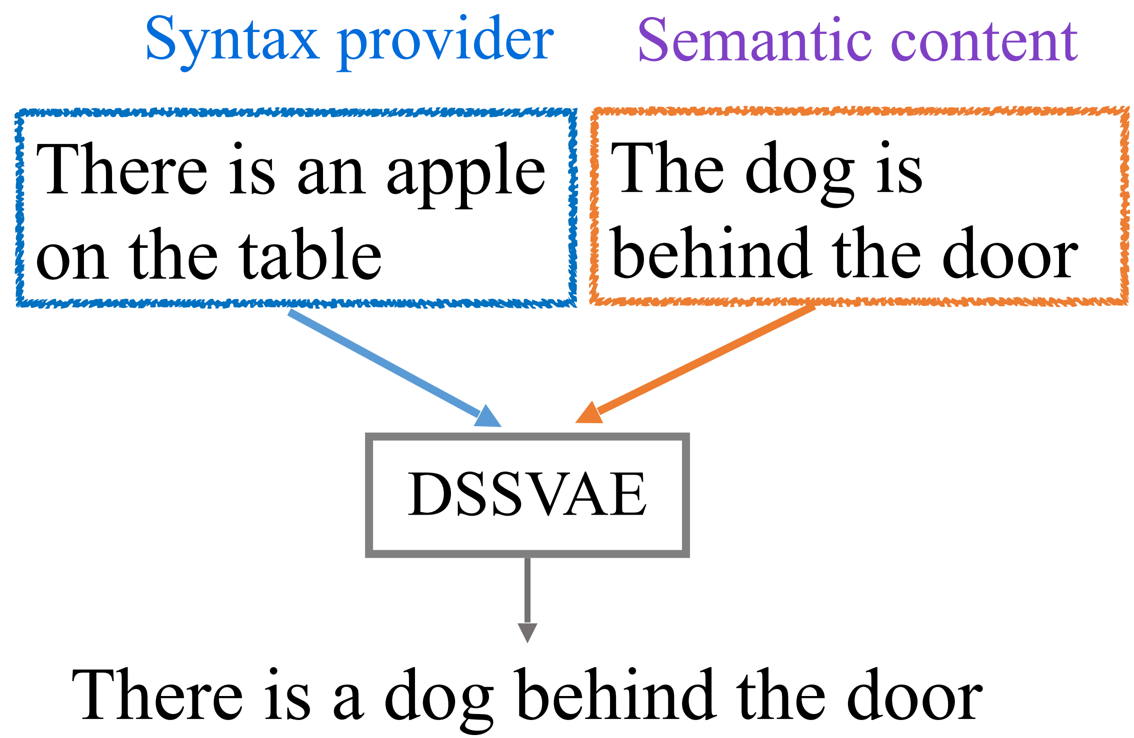
2: Jack Ryder (born August 9, 1889 in Victoria, Australia) was an Australian cricketer.

3: John Ryder, also known as the king of Collingwood (8 August 1889 – 4 April 1977) was an Australian cricketer.

Learning Disentangled Representation of Syntax and Semantics



DSSVAE enables learning and transferring sentence-writing styles



DSS-VAE [Y. Bao, H. Zhou, S. Huang, **Lei Li**, L. Mou, O. Vechtomova, X. Dai, J. Chen, ACL19c]

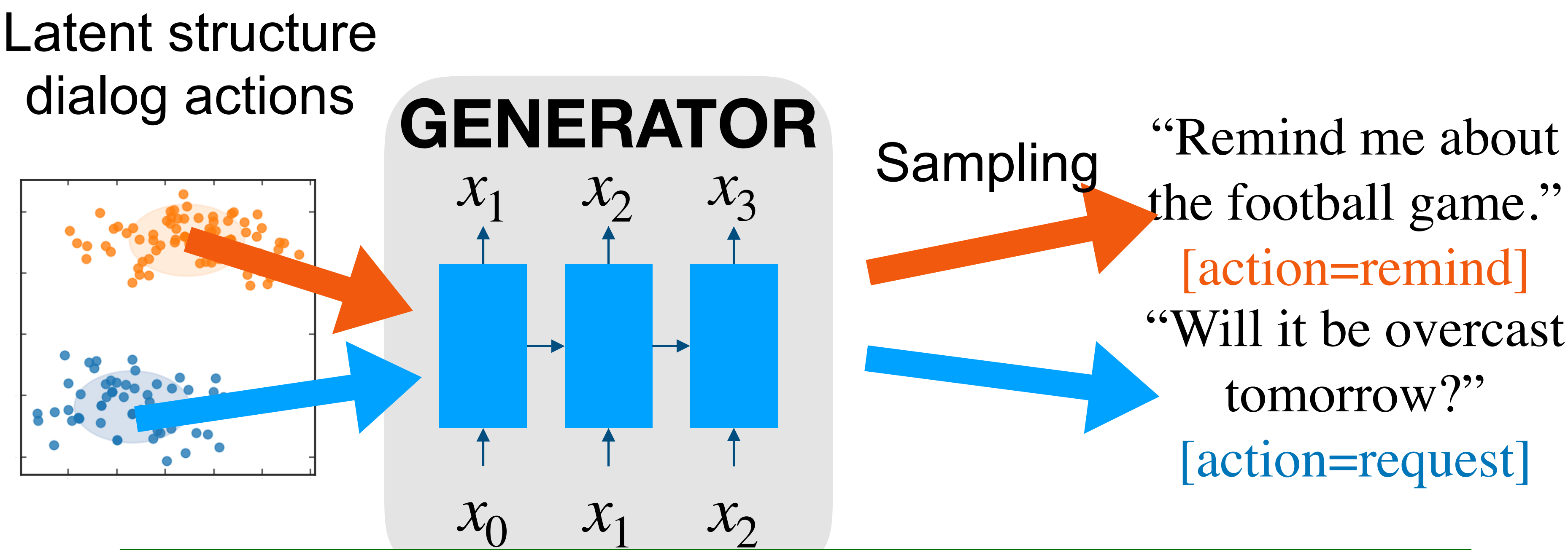
Impact

- VTM and its extensions have been applied to multiple online systems on Toutiao including query suggestion generation, ads bid-word generation, etc.
- Serving over 100million active users.
- 10% of query suggestion phrases from the generation algorithm.

Takeaway

- Deep latent models enable learning with both table-text pairs and unpaired text, with high accuracy
- Disentangling approach for model composition
- Variational technique to speed up inference

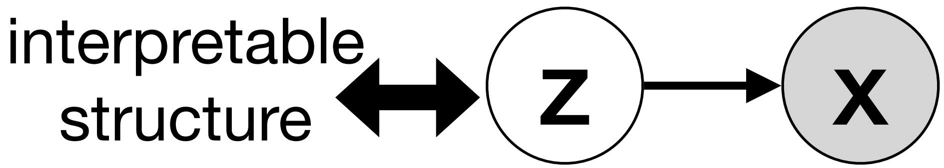
Interpretable Text Generation



Generate Sentences with interpretable factors

How to Interpret Latent Variables in VAEs?

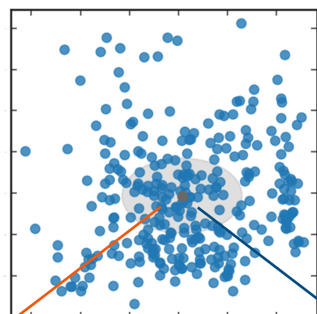
Variational Auto-encoder (VAE)



(Kingma & Welling, 2013)

difficult to interpret discrete factors

z: **continuous** latent variables



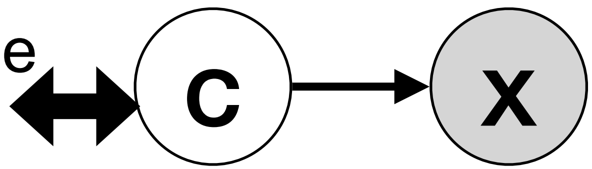
Will it be humid in New York today?

Remind me about my meeting.

VAEs Introduce Latent Variables

Variational Auto-encoder (VAE)

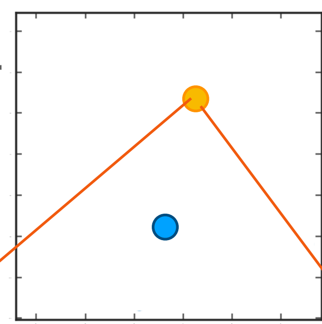
interpretable structure



(Zhao et al, 2018b)

C: **discrete**

latent variables



Remind me about my meeting.

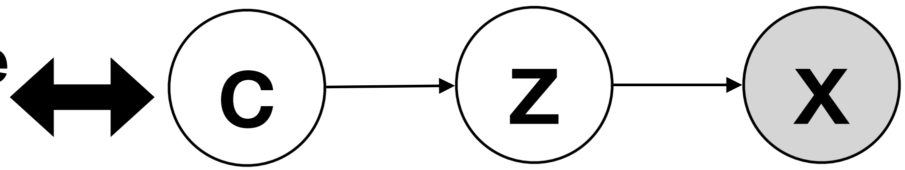
Remind me about the football game.

expressiveness is limited.

Discrete Variables Could Enhance Interpretability - but one has to do it right!

Gaussian Mixture Variational Auto-encoder (GM-VAE)

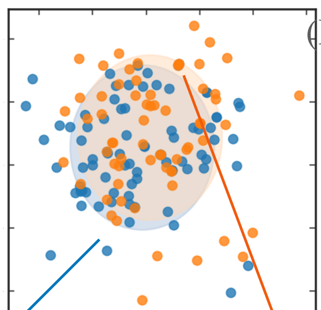
interpretable structure



(Dilokthanakul et al., 2016; Jiang et al., 2017)

C: **discrete** component

Z: **continuous** latent variable



Will it be overcast tomorrow?

Remind me about the football game.

Why?

How to fix it?

mode-collapse

Do it right for VAE w/ hierarchical priors - Dispersed Exponential-family Mixture VAE

The *negative* **dispersion term** in ELBO encourages the parameters of all mixture components in-distinguishable and induces the **mode-collapse.**

Dispersed EM-VAE

$$L(\theta; x) = \text{ELBO} + \beta \cdot \boxed{L_d},$$

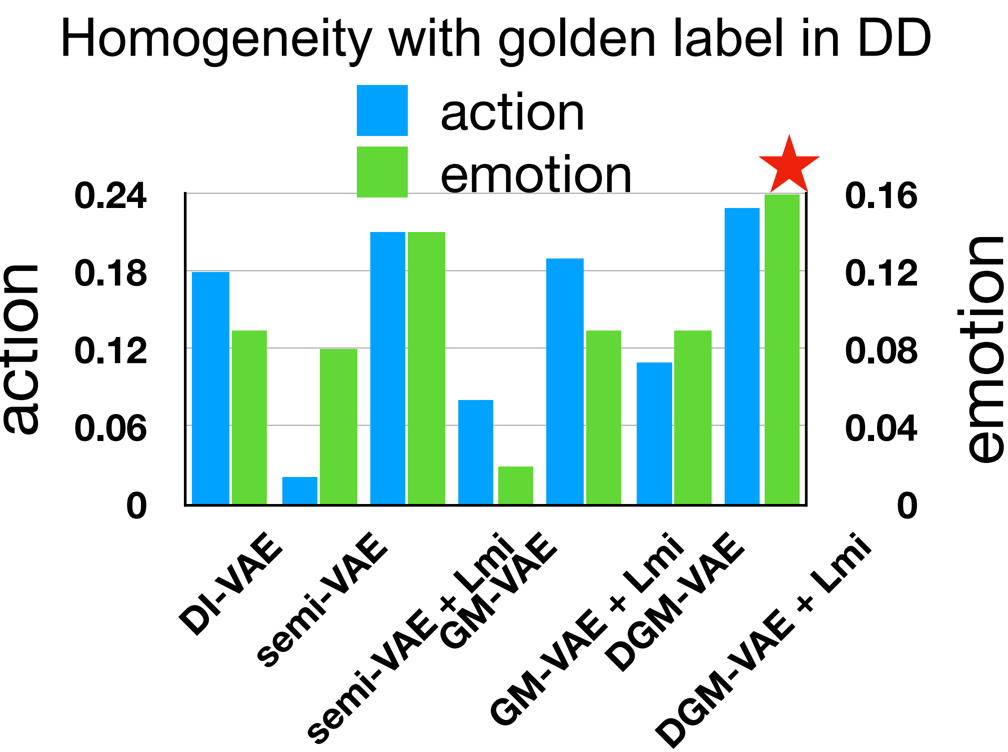
$$L_d = \mathbb{E}_{q_\phi(c|x)} A(\boldsymbol{\eta}_c) - A(\mathbb{E}_{q_\phi(c|x)} \boldsymbol{\eta}_c) .$$

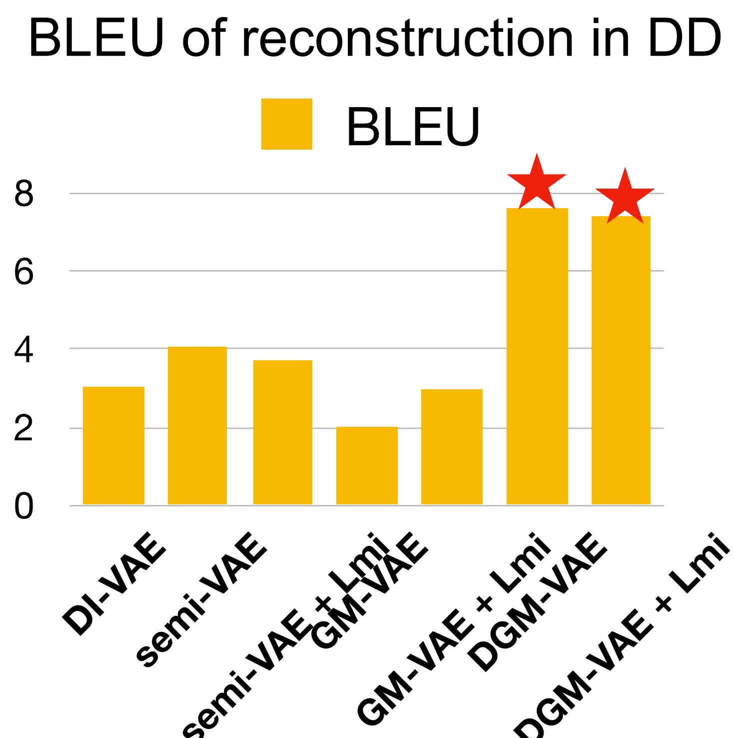
Include an extra *positive* dispersion term to balance the mode collapse from ELBO

DEM-VAE [W. Shi, H. Zhou, N. Miao, **Lei Li**, ICML 2020]

Generation Quality and Interpretability

DGM-VAE obtains the best performance in interpretability and reconstruction





Best interpretability

Best reconstruction

DEM-VAE [W. Shi, H. Zhou, N. Miao, **Lei Li**, ICML 2020]

Latent Variables Learned by DEM-VAE are Semantically Meaningful

Example actions and corresponding utterances (classified by $q_\phi(c \mid x)$)

Inferred action=Inform-route/address

"There is a Safeway 4 miles away."
"There are no hospitals within 2 miles."
"There is Jing Jing and PF Changs."
…

Inferred action =Request-weather

"What is the weather today?"
"What is the weather like in the city?"
"What's the weather forecast in New York?"
…

Utterances of the same actions could be assigned with the same discrete latent variable c.

DEM-VAE [W. Shi, H. Zhou, N. Miao, **Lei Li**, ICML 2020]

Generate Sensible Dialog Response with DEM-VAE

Input Context
Sys: "Taking you to Chevron."

sampling different values of discrete latent variables

(action = thanks)

Predict
User: "Thank you car, let's go there!"

(action = request-address)

Predict
User: "What is the address?"

Responses with different actions are generated by sampling different values of discrete latent variables.

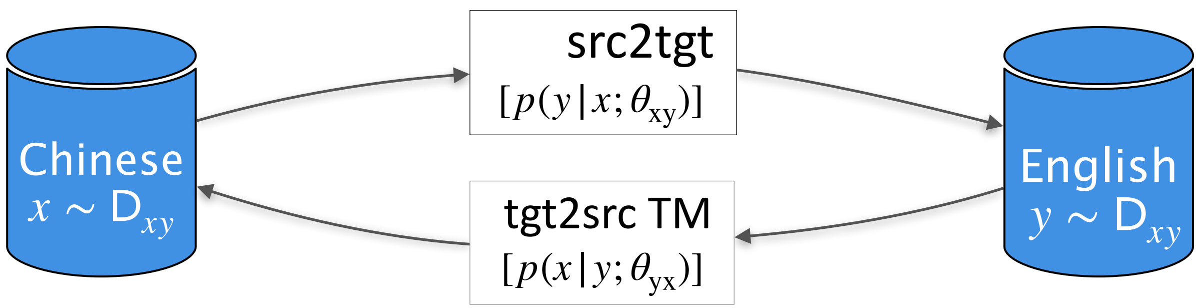
DEM-VAE [W. Shi, H. Zhou, N. Miao, **Lei Li**, ICML 2020]

Mirror Generative Model for Neural Machine Translation

MGNMT [Z. Zheng, H. Zhou, S. Huang, **Lei Li**, X. Dai, J. Chen, ICLR 2020a]

Neural Machine Translation

- Neural machine translation (NMT) systems are super good when you have large amount of **parallel bilingual data**



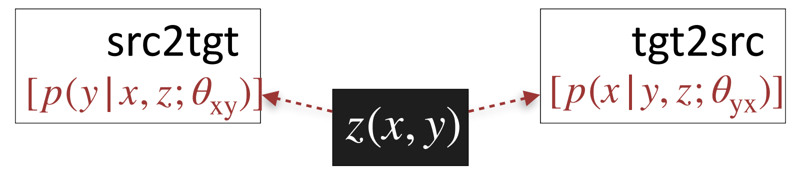
- **BUT**, very expensive/non-trivial to obtain
  - Low resource language pairs (*e.g.*, English-to-Tamil)
  - Low resource domains (*e.g.*, social network)
- Large-scale mono-lingual data are not fully utilized

Existing approaches to exploit non-parallel data

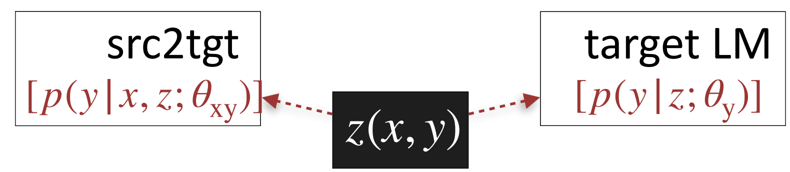
- There are two categories of methods using non-parallel data
  - Training
    - Back-translation, Joint Back-translation, dual learning…
  - Decoding
    - Interpolation w/ external LM …
- Still not the best

So, what we expect?

- A pair of **relevant** TMs so that they can directly boost each other in training

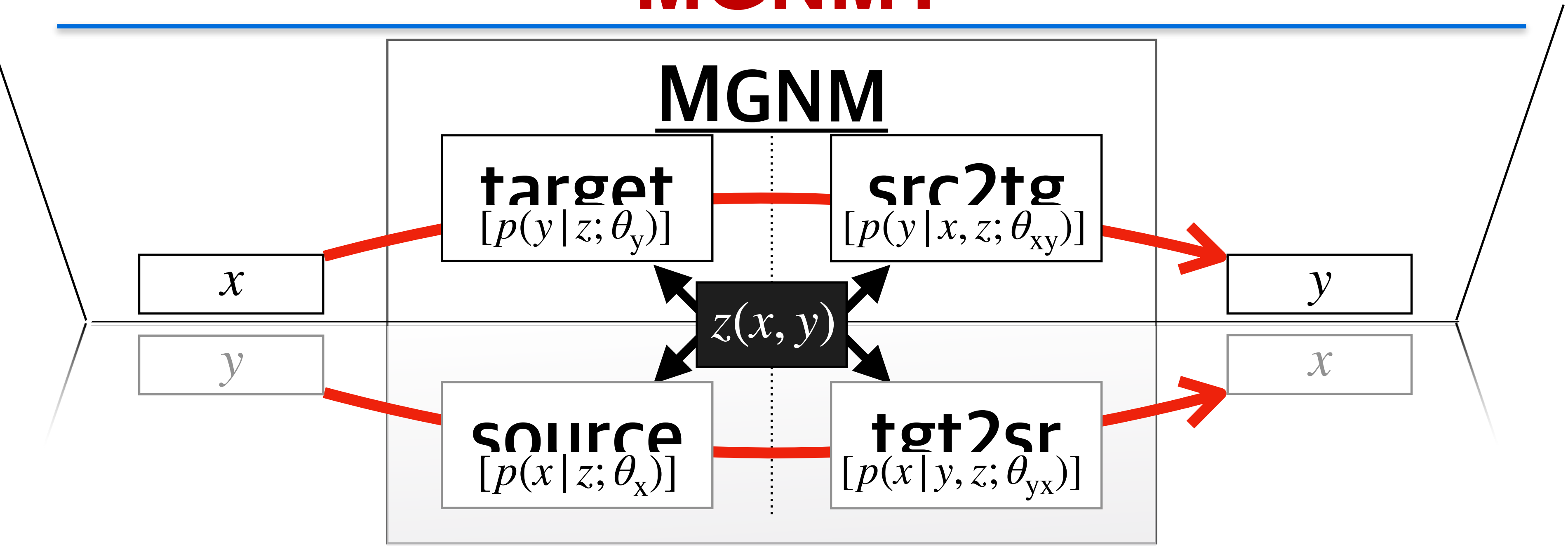


- A pair of **relevant** TM & LM so that they can cooperate more effectively for better decoding



We need a bridge

Integrating Four Language Skills with MGNMT



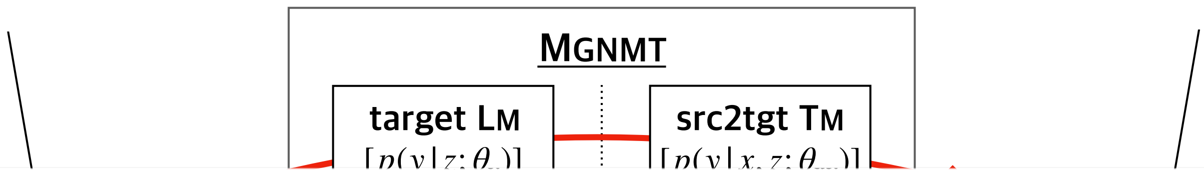
1. composing sentence in Source lang
2. composing sentence in Target lang
3. translating from source to target
4. translating from target to source

Benefits utilizing both parallel bilingual data and non-parallel corpus

MGNMT [Z. Zheng, H. Zhou, S. Huang, **L. Li**, X. Dai, J. Chen, ICLR 2020a]

Approach: Mirror-Generative NMT

- The **mirror** property to decompose



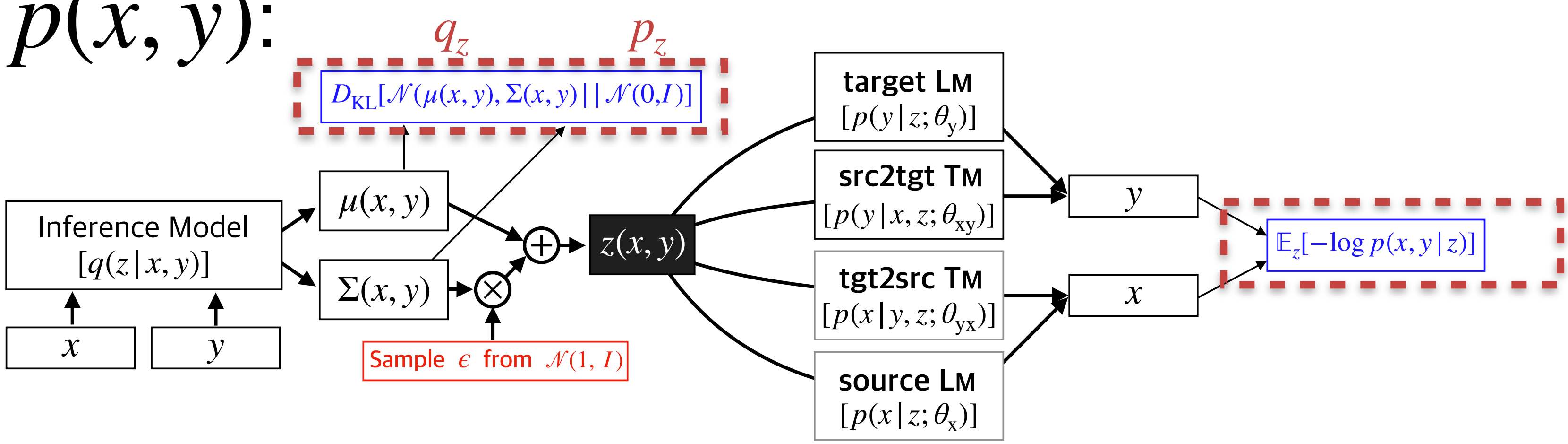
$$\begin{aligned}\log p(x,y|z) &= \log p(x|z) + \log p(y|x,z) = \log p(y|z) + \log p(x|y,z) \\ &= \frac{1}{2}[\underbrace{\log p(y|x,z)}_{src2tgt\ \mathrm{TM}_{x\to y}} + \underbrace{\log p(y|z)}_{target\ \mathrm{LM}_{y}} + \underbrace{\log p(x|y,z)}_{tgt2src\ \mathrm{TM}_{y\to x}} + \underbrace{\log p(x|z)}_{source\ \mathrm{LM}_{x}}]\end{aligned}$$

$$p(x, y \mid z) = p(y \mid x, z)p(x \mid z) = p(x \mid y, z)p(x \mid z)$$

- **Relevant** TMs & LMs under a **unified probabilistic framework!**
  - Enables the **aforementioned advantages**

Training w/ parallel data

- Given: a parallel bilingual sentence pair $\langle x, y \rangle$
- Goal: maximize the ELBO of the joint dist. $p(x, y)$:



$$\log p(x,y) \geq \mathcal{L}(x,y;\boldsymbol{\theta},\phi) = \mathbb{E}_{q(z|x,y;\phi)}[\frac{1}{2}\{\log p(y|x,z;\theta_{\mathrm{xy}}) + \log p(y|z;\theta_{\mathrm{y}}) + \log p(x|y,z;\theta_{\mathrm{yx}}) + \log p(x|z;\theta_{\mathrm{x}})\}] - D_{\mathrm{KL}}[q(z|x,y;\phi)||p(z)]$$

mirror

Training w/ non-parallel data

- Given: monolingual source sentence $x^{(s)}$ and target sentence $y^{(t)}$
- Goal: maximize the lower-bounds of source & target marginals

$$\log p(x^{(s)}) + \log p(y^{(t)}) \geq \mathcal{L}(x^{(s)}; \theta_{\mathrm{x}}, \theta_{\mathrm{yx}}, \phi) + \mathcal{L}(y^{(t)}; \theta_{\mathrm{y}}, \theta_{\mathrm{xy}}, \phi)$$

$$\begin{aligned}
\mathcal{L}(y^{(t)}; \theta_{\mathrm{y}}, \theta_{\mathrm{xy}}, \phi) = \mathbb{E}_{p(x|y^{(t)})}\big[\mathbb{E}_{q(z|x,y^{(t)};\phi)}[&\frac{1}{2}\{\log p(y^{(t)}|z;\theta_{\mathrm{y}}) + \log p(y^{(t)}|x,z;\theta_{\mathrm{xy}})\}] \\
&- D_{\mathrm{KL}}[q(z|x,y^{(t)};\phi)||p(z)]\big] \\
\mathcal{L}(x^{(s)}; \theta_{\mathrm{x}}, \theta_{\mathrm{yx}}, \phi) = \mathbb{E}_{p(y|x^{(s)})}\big[\mathbb{E}_{q(z|x^{(s)},y;\phi)}[&\frac{1}{2}\{\log p(x^{(s)}|z;\theta_{\mathrm{x}}) + \log p(x^{(s)}|y,z;\theta_{\mathrm{yx}})\}] \\
&- D_{\mathrm{KL}}[q(z|x^{(s)},y;\phi)||p(z)]\big]
\end{aligned}$$

Decoding: TM&LM work as a whole

- Iterative EM decoding
  - Given source sentence x, find a translation

$$y = \mathrm{argmax}_y \; p(y|x) = \mathrm{argmax}_y \; p(x, y) \approx \mathrm{argmax}_y \; \mathcal{L}(x, y; \boldsymbol{\theta}, \phi)$$

  - **Initialization**: get a draft translation
  - **Iterative refinement**: resampling z from inference model and redecoding by maximizing ELBO

$$\begin{aligned}\tilde{y} \leftarrow & \mathrm{argmax}_y \; \mathcal{L}(x, \tilde{y}; \boldsymbol{\theta}, \phi) \\ &= \mathrm{argmax}_y \; \mathbb{E}_{q(z|x,\tilde{y};\phi)}[\log p(y|x,z) + \log p(y|z) + \log p(x|z) + \log p(x|y,z)] \\ &= \mathrm{argmax}_y \; \mathbb{E}_{q(z|x,\tilde{y};\phi)}\Big[\sum_i \underbrace{[\log p(y_i|y_{<i},x,z) + \log p(y_i|y_{<i},z)]}_{\text{Decoding Score}} + \underbrace{\log p(x|z) + \log p(x|y,z)}_{\text{Reconstructive Reranking Score}}\Big]\end{aligned}$$

Experiments

- Datasets
  - Low resource
    - WMT16 EN-RO
    - IWSLT16 EN-DE: domain adaptation (from TED to News)
  - High resource:
    - WMT14 EN-DE, NIST EN-ZH
- Avoiding **posterior collapse** (Important!)
  - KL-annealing
  - Word dropout

MGNMT makes better use of non-parallel data

- Low resource results

Model	Low-Resource		Cross-Domain			
	WMT16 En↔Ro		In-Domain (TED)		Out-Domain (News)	
	En-Ro	Ro-En	En-De	De-En	En-De	De-En
Transformer (Vaswani et al., 2017)	32.1	33.2	27.5	32.8	17.1	19.9
GNMT (Shah & Barber, 2018)	32.4	33.6	28.0	33.2	17.4	20.1
GNMT-M-SSL + *non-parallel* (Shah & Barber, 2018)	34.1	35.3	28.4	33.7	22.0	24.9
Transformer+BT + *non-parallel* (Sennrich et al., 2016b)	33.9	35.0	27.8	33.3	20.9	24.3
Transformer+JBT + *non-parallel* (Zhang et al., 2018)	34.5	35.7	28.4	33.8	21.9	25.1
Transformer+Dual + *non-parallel* (He et al., 2016a)	34.6	35.7	28.5	34.0	21.8	25.3
MGNMT	32.7	33.9	28.2	33.6	17.6	20.2
MGNMT + *non-parallel*	**34.9**	**36.1**	28.5	34.2	**22.8**	**26.1**

MGNMT makes better use of non-parallel data

- High resource results

Model	WMT14		NIST	
	EN-DE	DE-EN	EN-ZH	ZH-EN
Transformer (Vaswani et al., 2017)	27.2	30.8	39.02	45.72
GNMT (Shah & Barber, 2018)	27.5	31.1	40.10	46.69
GNMT-M-SSL + *non-parallel* (Shah & Barber, 2018)	29.7	33.5	41.73	47.70
Transformer+BT + *non-parallel* (Sennrich et al., 2016b)	29.6	33.2	41.98	48.35
Transformer+JBT + *non-parallel* (Zhang et al., 2018)	30.0	33.6	42.43	48.75
Transformer+Dual + *non-parallel* (He et al., 2016b)	29.6	33.2	42.13	48.60
MGNMT	27.7	31.4	40.42	46.98
MGNMT + *non-parallel*	30.3	33.8	42.56	49.05

  - Non-parallel data is helpful
  - MGNMT works well especially on **low resource** settings

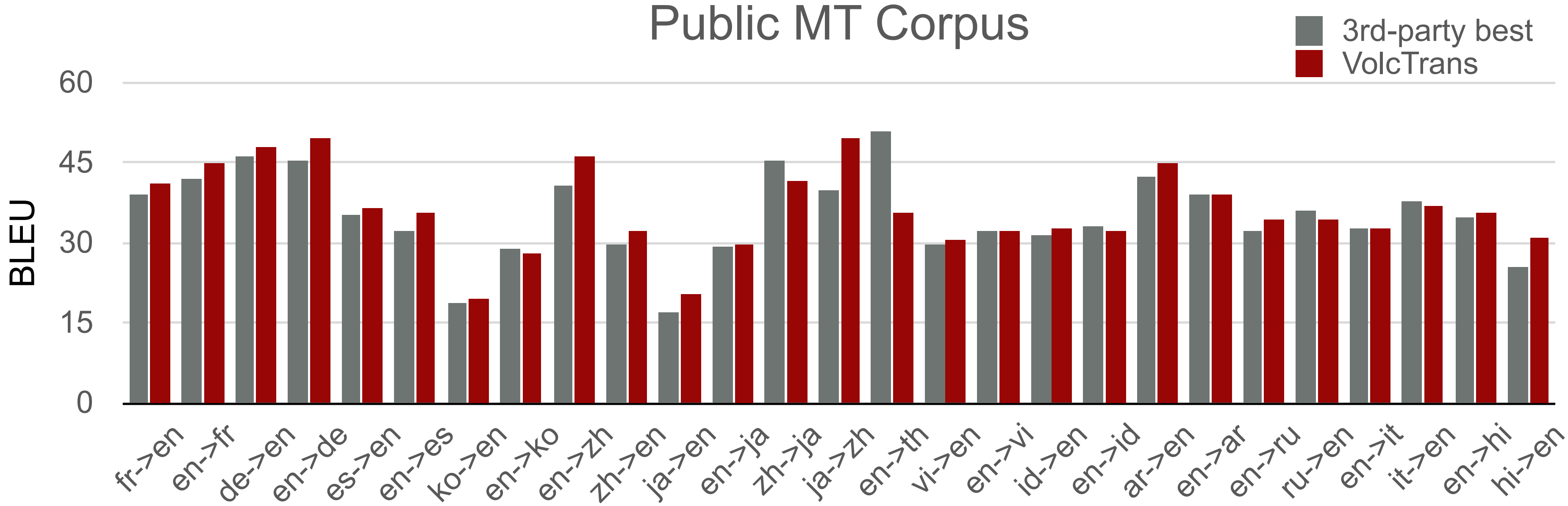
Machine Translation at Bytedance (VolcTrans)

50+
Clients

9 Billion

16
languages

Public MT Corpus



Speech-to-Text Translation Demo

Simultaneous Speech-to-text Translation @ VolcTrans

Takeaway

- MGNMT is a unified probabilistic framework which jointly models TMs and LMs and enables their cooperation in a better way.
- In low-resource settings, MGNMT works better than in high-resource settings
- Training of MGNMT is somewhat tricky and inefficient
- Could be extended to multilingual or unsupervised scenarios.
- Our VolcTrans system already serves > 100million active users

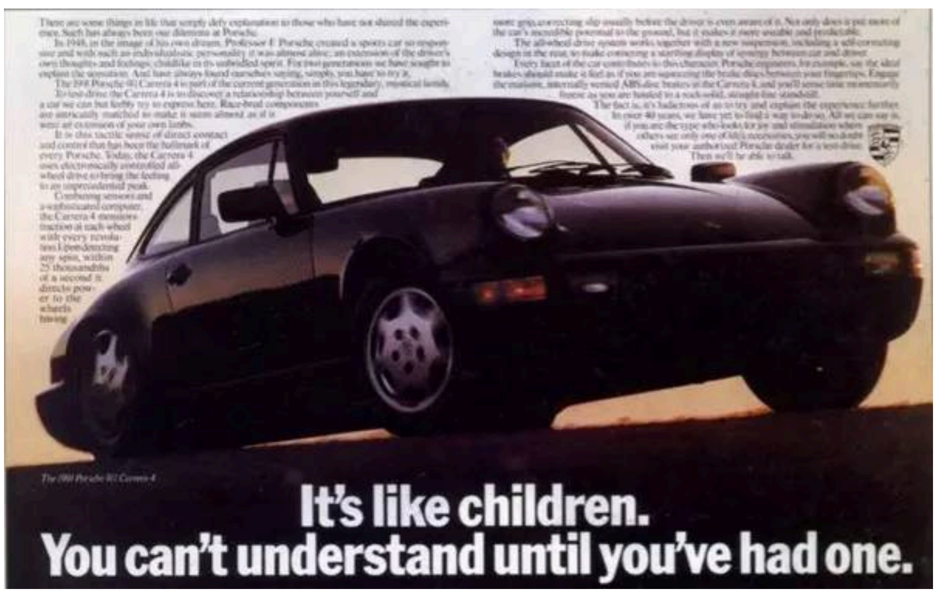
Outline

1. Basics of Deep Generative Models for Sequences
2. Deep Latent Variable Models
3. Monte-Carlo Methods for Constrained Text Generation
4. Multimodal machine writing: show case
5. Summary

Monte-Carlo Methods for Constrained Text Generation

CGMH [N. Miao, H. Zhou, L. Mou, R. Yan, **Lei Li**, AAAI19]
MHA [H. Zhang, N. Miao, H. Zhou, **Lei Li**, ACL19a]
TSMH [M. Zhang, N. Jiang, **Lei Li**, Yexiang Xue, EMNLP20e]

Automate Creative Advertisement Design



Constrained Text Generation

To generate sentences that are:

- Fluent
- Constraint-satisfying
  - e.g. keyword-occurrence constraint

"Autumn"
"Sports shoes"

Comfortable **sports shoes**, a breathing pair of man's shoes, accompanying you in **autumn**

Why is Constrained Text Generation important?

- One generic formulation for many tasks
- Ads creative slogan design given product highlighting attributes
- Title generation for articles given keywords
- Writer assistant: automatic sentence error correction
- Machine translation with bilingual entity-dictionary

Why is Text Generation difficult?

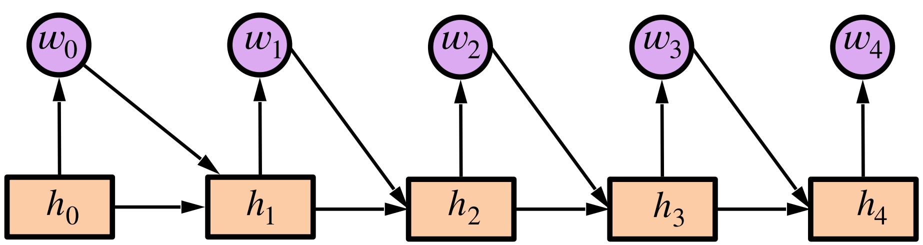
- Text space is discrete
  - Interpolation and smoothing in the surface level would not work
- High-dimensional space: exponential search space for sentence
- Controlling the generation with desired properties is challenging
- The lack of labeled data pairs <constraint, ground-truth sentence> ➔ learning without supervision!

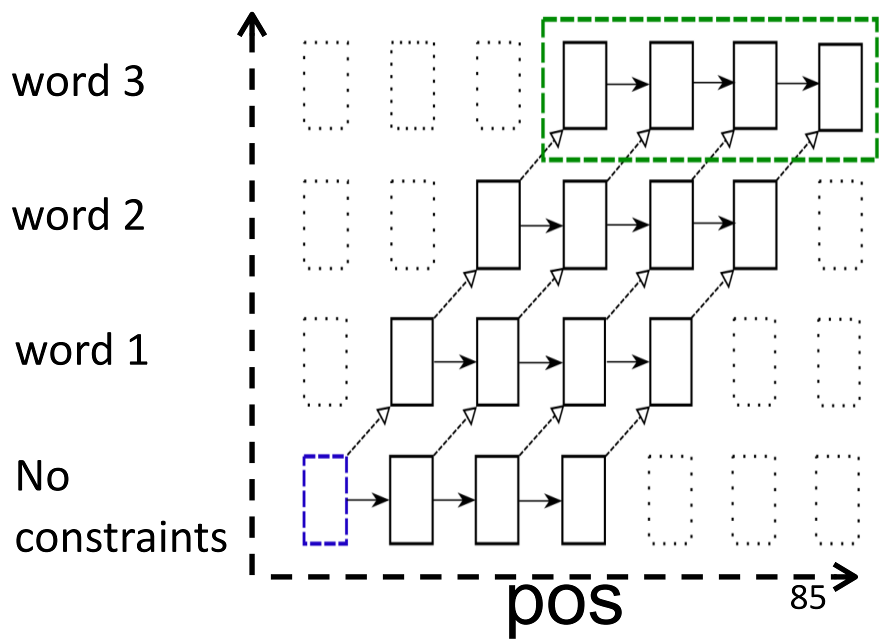
Why is Constrained Text Generation difficult?

Exponential search space, $O((N-k)^V)$

RNN grid beam search [Hokamp & Liu 2017]

does not usually produce high quality sentences





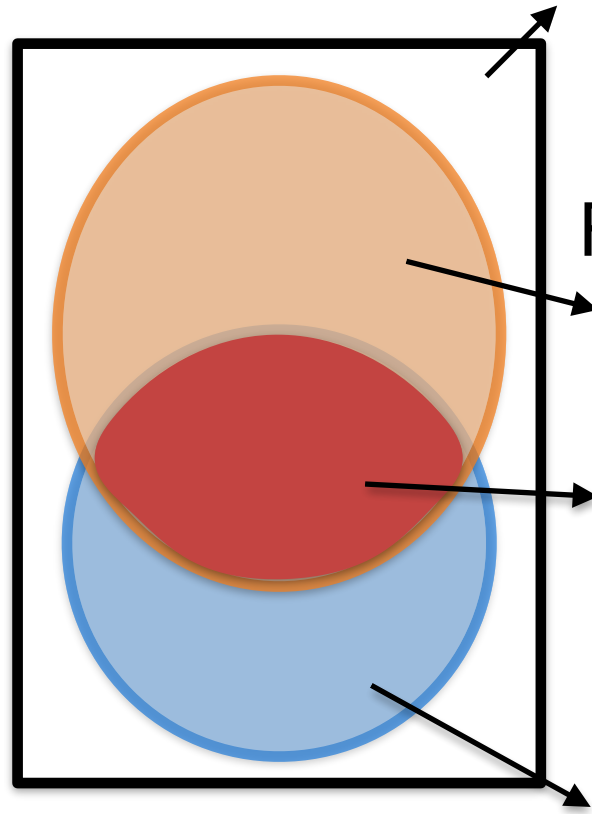
Constrained Sentence Generation via Metropolis-Hastings Sampling

- Key idea: To generation samples from the *implicit* distribution by iterative editing (MH sampling)

$$\pi(x) = \prod_i P(x_i \mid x_{0:i-1}) \cdot \prod_j P_C^j(x)$$

pre-trained language model prob.

indicator (0-1) function for constraints

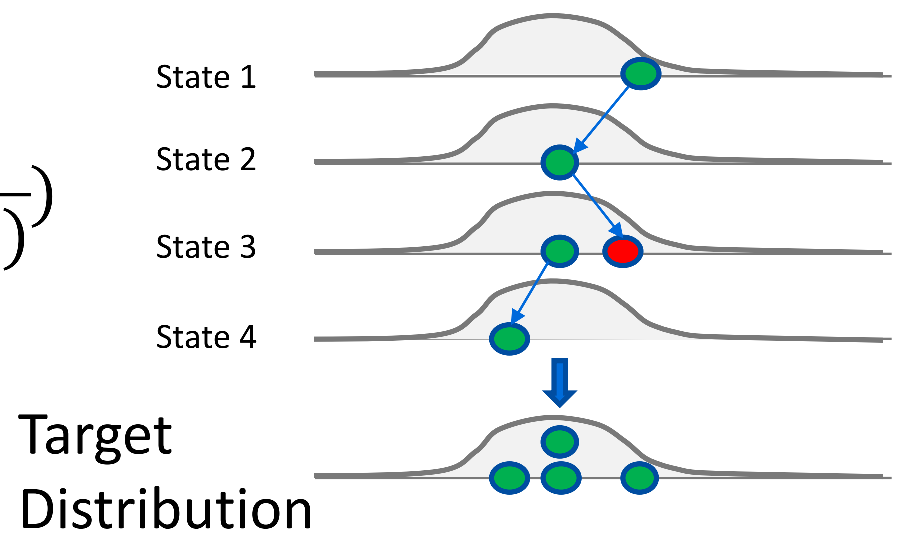


CGMH [N. Miao, H. Zhou, L. Mou, R. Yan, **Lei Li**, AAAI19]

Metropolis-Hastings Sampling

One case of Markov chain Monte Carlo methods, Metropolis-Hastings(MH) performs sampling by first **proposes** a transition, and then **accepts or rejects** the transition.

$$A(x'|x_{t-1}) = \min\left(1, \frac{\pi(x') \cdot g(x_{t-1}|x')}{\pi(x_{t-1}) \cdot g(x'|x_{t-1})}\right)$$



π is the target density,

g is proposal distribution, which is easy to sample

CGMH: Main Idea

- CGMH performs constrained generation by:
  1. Pretrain Neural Language Model (e.g. GPT2);
  2. Iterative Editing:
     1) Start from a initial sentence x_0;
     2) Propose a new sentence x_t from x_{t-1}, and accept/reject the action. Action proposal include:
        I. Replacement: change a word to another one
        II. Insertion: add a word
        III. Deletion: remove a word

BMW, the sports car of daily life

BMW, the sports car of today's life
BMW, the sports car of future life
BMW, the sports car of new life
BMW, the sports car of happy life
…

CGMH [N. Miao, H. Zhou, L. Mou, R. Yan, **Lei Li**, AAAI19]

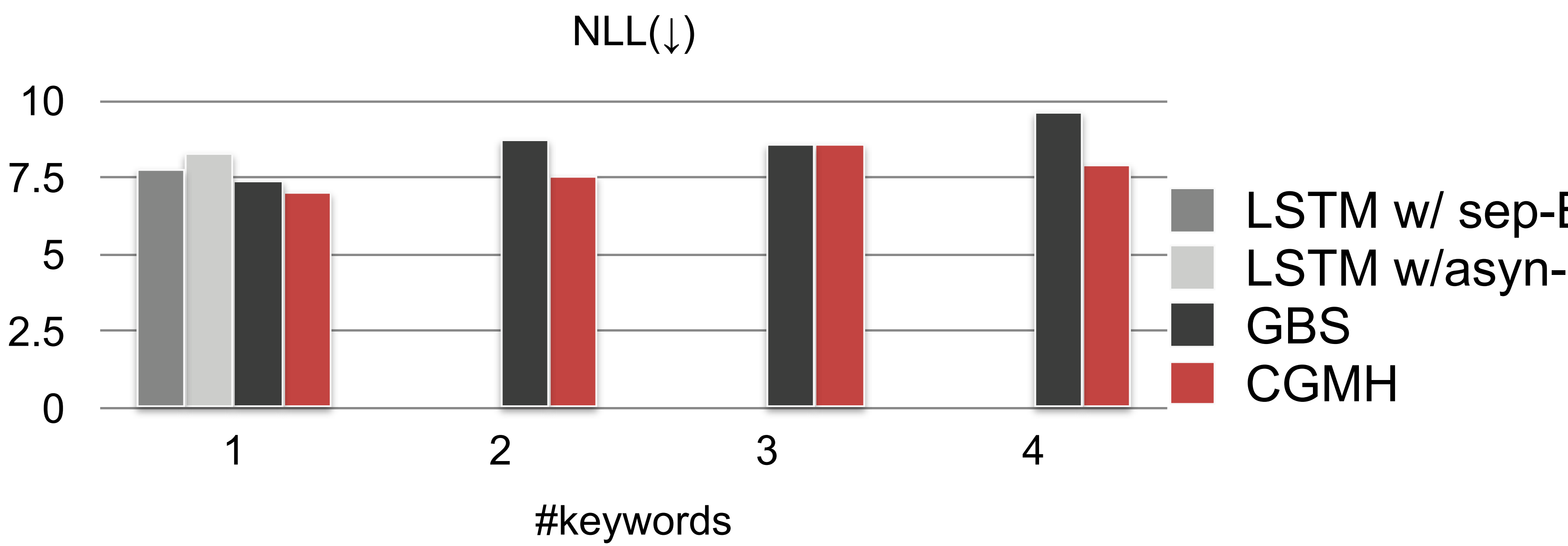
CGMH Iteratively Edits Candidates

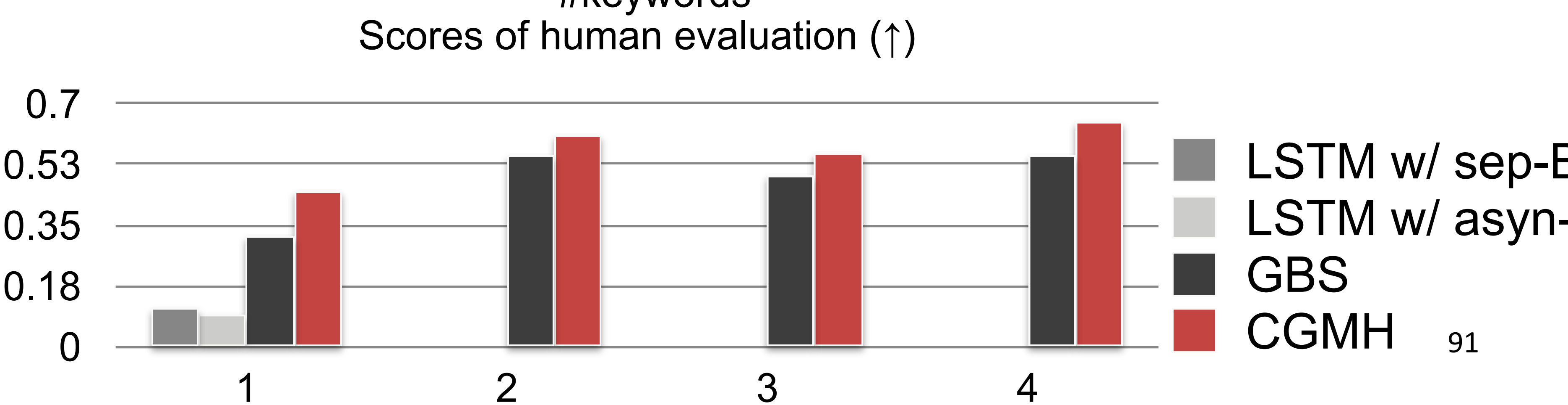
Step	Action	Acc/Rej	Sentences
0	[Input]		BMW sports
1	Insert	Accept	BMW sports car
2	Insert	Accept	BMW the sports car
…	…	…	…
6	Insert	Accept	BMW , the sports car of daily life
7	Replace	Accept	BMW , the sports car of ~~daily~~future life
8	Insert	Accept	BMW , the sports car of the future life
9	Delete	Reject	BMW , the sports ~~car~~ of the future life
10	Delete	Accept	BMW , the sports car of the future ~~life~~
11	[Output]		BMW , the sports car of the future

Evaluation 1: Keyword to Sentence

- Keywords to sentence generation (hard constraints)
  - Aim: To generate fluent sentences containing the given set of words.
  - Dataset: A subset of one-billion-word corpus (5M)
  - Input: Keywords random selected from the target sentence.
  - Constraint: $1_{\text{keywords occur in sentence}}$

CGMH generates better sentences from keywords





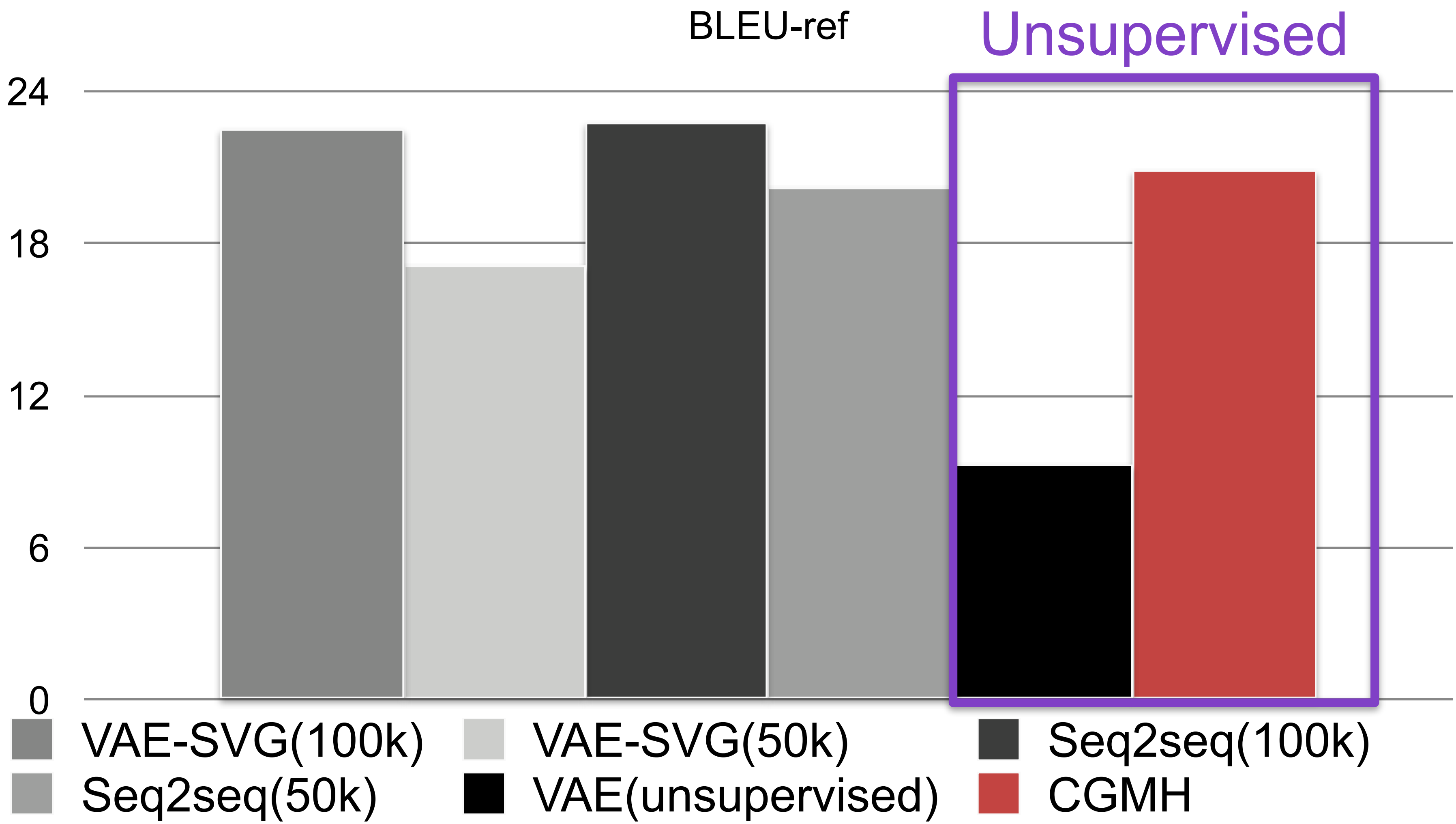
Evaluation 2: Paraphrase Generation

- Unsupervised paraphrase generation (soft constraints)
  - Aim: To generate sentences with similar meaning of the given one.

    what’s the best plan to lose weight

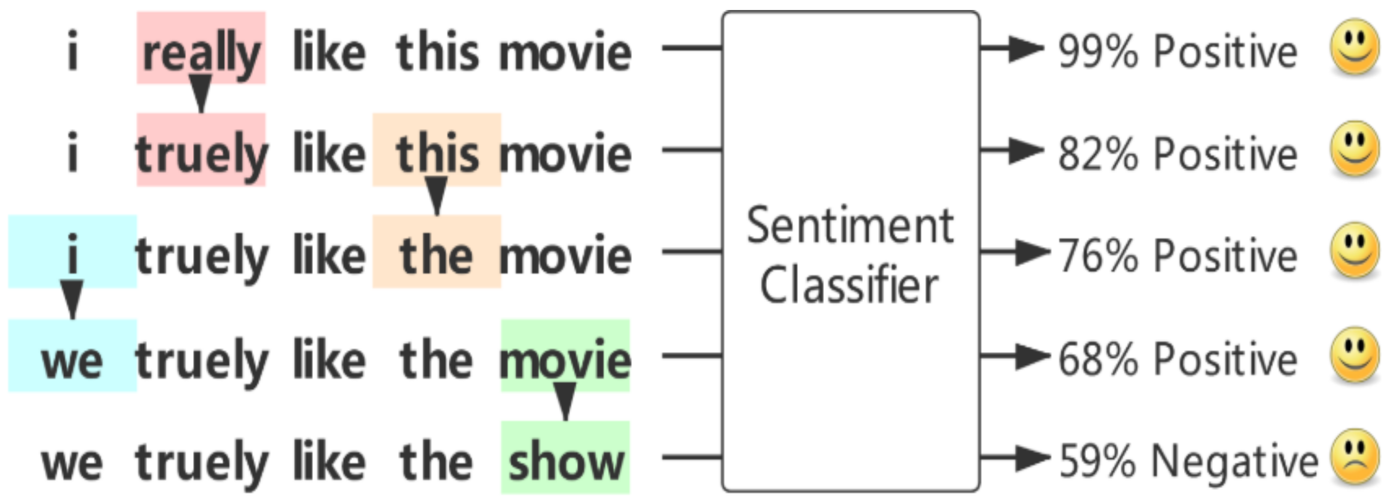
    ➜ what’s the best way to slim down quickly

CGMH is the first unsupervised model to achieve comparable results with supervised models.



Extension: Adversarial Fluent Sentence Generation w/ Iterative Editing

- Machine learning models are vulnerable to noises and attacks.
- Generating fluent adversarial text is challenging, due to the discreteness in text! (Ebrahimi et al., 2018; Alzantot et al., 2018)
- Our MHA achieves higher attack success rate

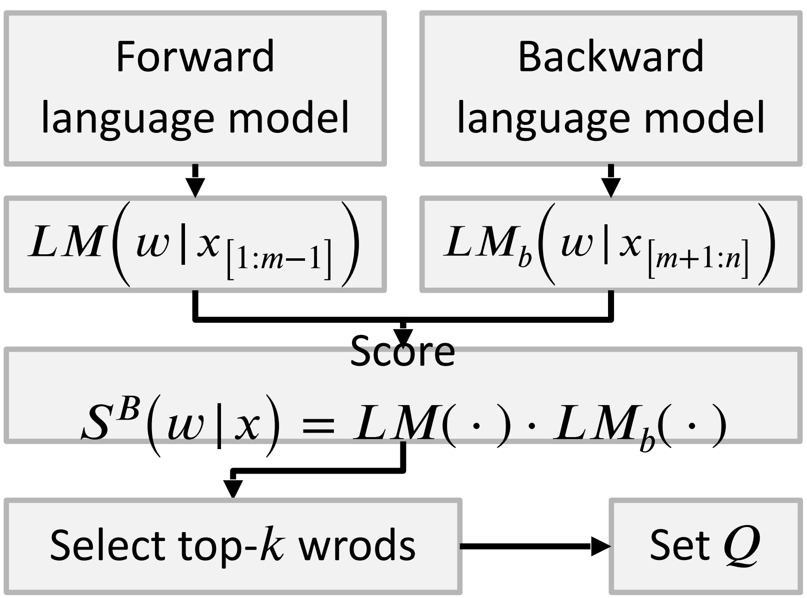


MHA [H. Zhang, N. Miao, H. Zhou, **Lei Li**, ACL19a]

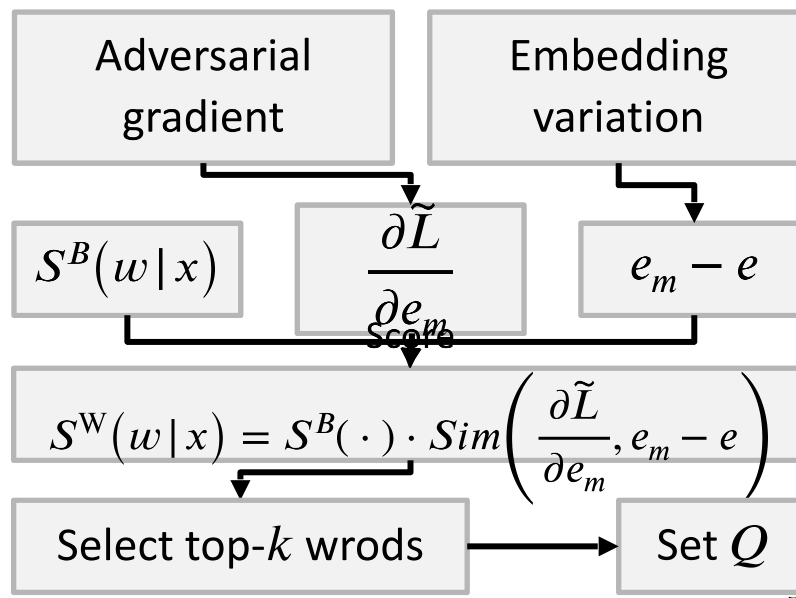
Adversarial Sentence Generation via MCMC

Reuse the CGMH algorithm

- *Blackbox b*-MHA
  - Black-box setting
  - Pre-select set Q with a forward language model and a backward language model



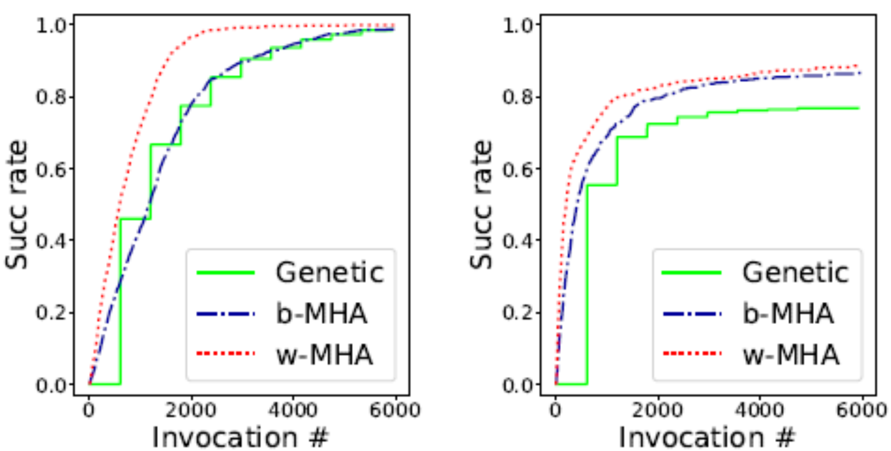
- *Whitebox w*-MHA
  - White-box setting
  - Pre-select set Q with a forward language model, a backward language model and the similarity of embedding variation and adversarial gradients.



Higher Attack Success Rate and Improved Text Classifier!

- MHA achieves higher attack success rate with fewer invocations, and gives lower perplexity, than the genetic approach (Alzantot et al., 2018) baseline.
- Examples generated by MHA may improve the adversarial robustness and the classification accuracy after adversarial training.

Attack Success Rate



(a) IMDB (b) SNLI

Accuracy w/ Adversaries

Model	Acc (%) Train # = 10K	30K	100K
Victim model	58.9	65.8	73.0
+ Genetic adv training	58.8	66.1	**73.6**
+ *w*-MHA adv training	**60.0**	**66.9**	**73.5**

Zhang et al., Generating Fluent Adversarial Examples for Natural Languages, in ACL, 2019, short paper.

Impact

- CGMH is deployed in a large-scale online ads creation platform
- Active used by 100,000 merchants and organizations
- Adoption rate: ~75%

“Autumn”
“Sports shoes”

Comfortable **sports shoes**, a breathing pair of man's shoes, accompanying you in **autumn**

Outline

1. Basics of Deep Generative Models for Sequences
2. Deep Latent Variable Models
3. Monte-Carlo Methods for Constrained Text Generation
4. Multimodal machine writing: show case
5. Summary

Multimodal Machine Writing

Xiaomingbot [R. Xu, J. Cao, M. Wang, J. Chen, H. Zhou, Y. Zeng, Y. Wang, L. Chen, X. Yin, X. Zhang, S. Jiang, Y. Wang, **Lei Li**, ACL 2020]
GraspSnooker [Z. Sun, J. Chen, H. Zhou, D. Zhou, **Lei Li**, M. Jiang, IJCAI19b]

Jersey Number Recognition with Semi-Supervised Spatial Transformer Network [G. Li, S. Xu, X. Liu, **Lei Li**, C. Wang, CVPR-CVS18]

Automatic News Writing in Real-world

- Tencent: Dreamwriter, started in 2015.9
- Fast Writer Xiaoxin: Xinhuanet, started in 2015.11
- Xiaomingbot: ByteDance, started in 2016.8
- Xiaonan: Southern Weekend, started 2017.1
- Wibbitz: USA Today
- Heliograf: Washington Post

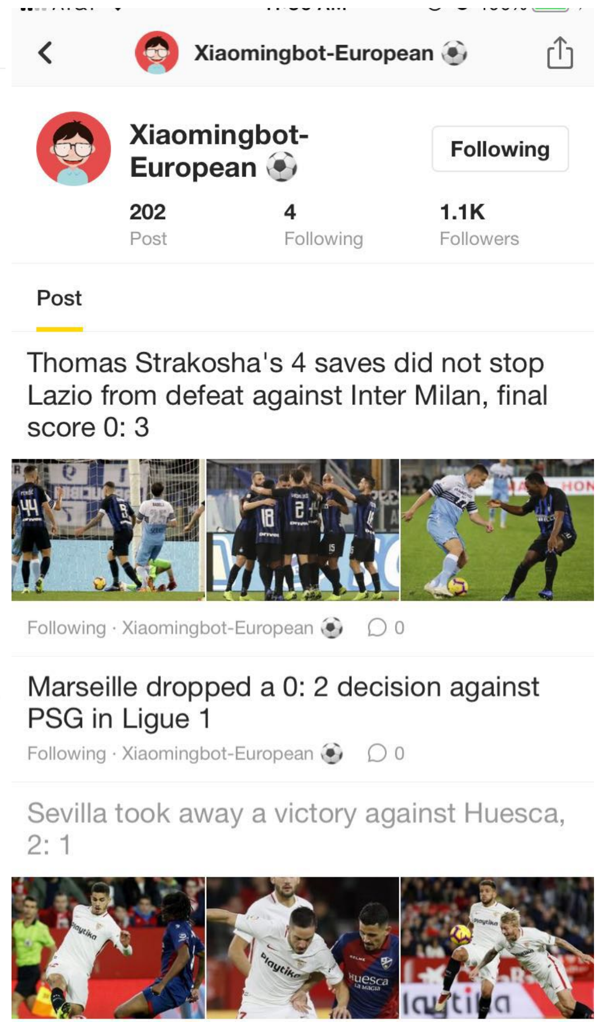
Landon beat Whitman 34-0; https://t.co/V6zVPi7a9O @LandonSports @koachkuhn
— WashPost HS Sports (@WashPostHS) September 2, 2017

Xiaomingbot
Automatic News Writing System

Winning 2017 Wu Wen-tsün Award in AI from CAAI

明くんのW杯
(Japanese)

Beto Bot Copa2018
(Portuguese)



600,000 articles

6 lang

150,000 followers

Xiaomingbot : Multilingual Robot News Reporter

Runxin Xu, Jun Cao, Mingxuan Wang, Jiaze Chen, Hao Zhou, Ying Zeng, Yuping Wang, Li Chen, Xiang Yin, Xijin Zhang, Songcheng Jiang, Yuxuan Wang, Lei Li, ACL 2020.

Snooker Commentary Generation
Combining Visual Understanding with Strategy Prediction

Balls Detection

Balls' Positions at the Beginning

Red0: (180, 542)
Red1: (189, 552)
Red2: (179, 555)
Red3: (184, 561)
Red4: (202, 563)
Red5: (174, 564)
Red6: (189, 569)
Red7:
Red11:(197, 590)
Red12:(241, 595)
Red13:(155, 606)
Red14:(327, 611)
Brown: (183, 163)
Green: (240, 163)
Yellow: (127, 163)
Blue: (183, 366

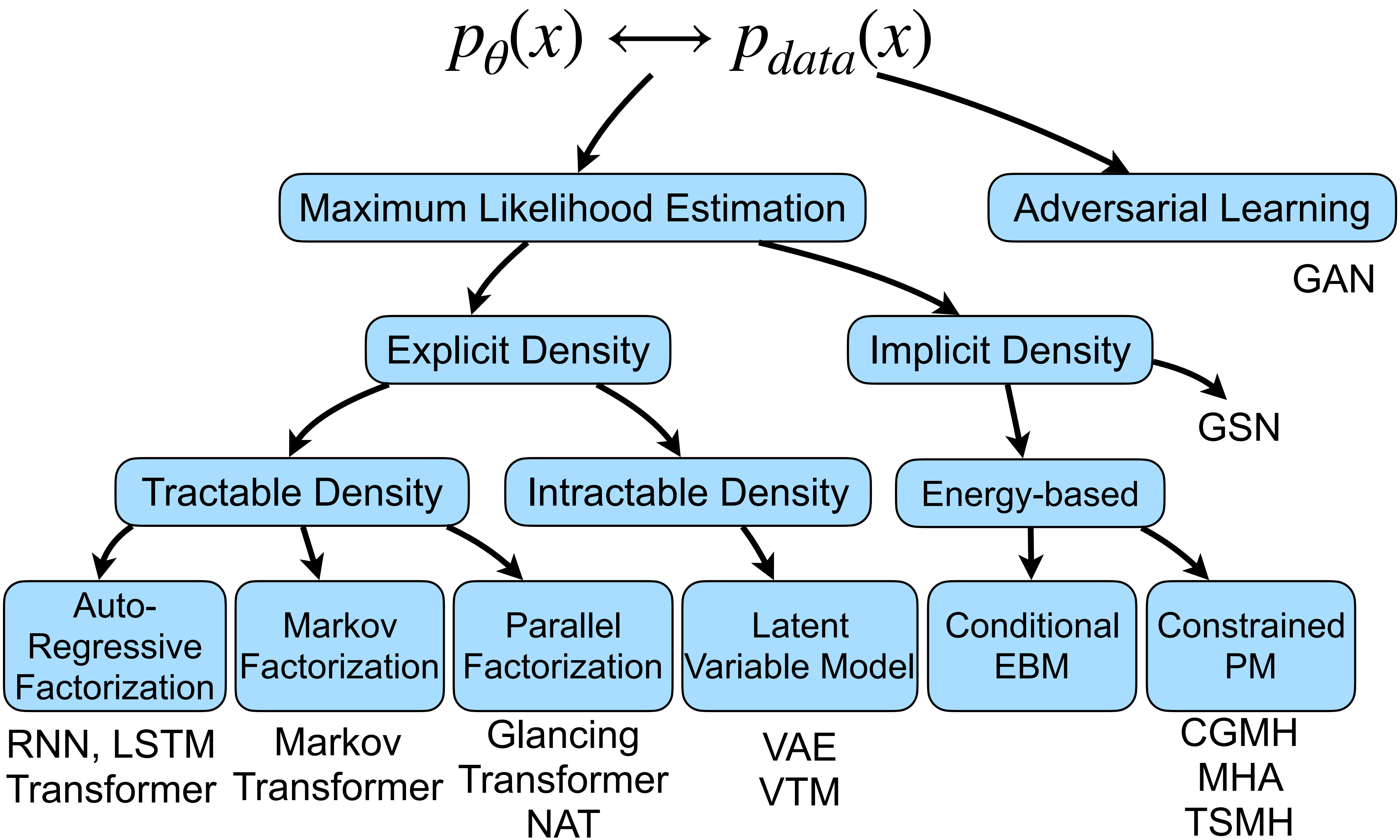
(positions after mapping)

GraspSnooker [Z. Sun, J. Chen, H. Zhou, D. Zhou, **Lei Li**, M. Jiang, IJCAI19b]

Summary

- Transformer, LSTM & Softmax: Basic neural generation nets for text
- Disentangled Latent Representation
  - VTM: Learning Latent Templates in Variational Space
  - DSS-VAE: Disentangled syntax and semantic representation
- DEM-VAE: Self identifying meaningful clusters with corpus
- MGNMT:
  - integrate four language capabilities together
  - Utilize both parallel and non-parallel corpus
- CGMH: Bayesian approach to constrained text generation
  - Able to learn with raw data only
- Multimodal Machine Writing
  - Xiaomingbot system: 600k articles and 150k followers
- Deployed in multiple online platforms and used by over 100 millions of users

Recap: DGM Taxonomy



Thanks

- Joint w/ Hao Zhou, Rong Ye, Ning Miao, Wenxian Shi, Zaixiang Zheng, Huangzhao Zhang, Ying Zeng, Jiaze Chen, Han Zhang
- Contact: lileilab@bytedance.com

Reference

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. NeurIPS 2017.
2. Ning Miao, Hao Zhou, Lili Mou, Rui Yan, Lei Li. “CGMH: Constrained Sentence Generation by Metropolis-Hastings Sampling”. In: the 33rd AAAI Conference on Artificial Intelligence (AAAI). Jan. 2019.
3. Huangzhao Zhang, Ning Miao, Hao Zhou, Lei Li. “Generating Fluent Adversarial Examples for Natural Languages”. In: the 57th Annual Meeting of the Association for Computational Linguistics (ACL) - short papers. July 2019.
4. Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, Jiajun Chen. “Generating Sentences from Disentangled Syntactic and Semantic Spaces”. In: the 57th Annual Meet- ing of the Association for Computational Linguistics (ACL). July 2019.
5. Ning Miao, Hao Zhou, Chengqi Zhao, Wenxian Shi, Lei Li. “Kernelized Bayesian Softmax for Text Generation”. In: the 33rd Conference on Neural Information Processing Systems (NeurIPS). Dec. 2019.
6. Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xinyu Dai, Jiajun Chen. “Mirror Generative Models for Neural Machine Translation”. In: International Conference on Learning Representations (ICLR). Apr. 2020.
7. Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, Lei Li. “Variational Template Machine for Data- to-Text Generation”. In: International Conference on Learning Representations (ICLR). Apr. 2020.
8. Ning Miao, Yuxuan Song, Hao Zhou, Lei Li. “Do you have the right scissors? Tailoring Pre-trained Language Models via Monte-Carlo Methods”. In: the 58th Annual Meeting of the Association for Computational Linguistics (ACL) - short papers. July 2020.
9. Wenxian Shi, Hao Zhou, Ning Miao, Lei Li. “Dispersing Exponential Family Mixture VAEs for Interpretable Text Generation”. In: Proceedings of the 37th International Conference on Machine learning (ICML). July 2020.
10. Maosen Zhang, Nan Jiang, Lei Li, Yexiang Xue. “Constraint Satisfaction Driven Natural Language Generation: A Tree Search Embedded MCMC Approach”. In: the Conference on Empirical Methods in Natural Language Processing (EMNLP) - Findings. Nov. 2020.