

Interpretable, Controllable and Scalable Natural Language Generation

高效可控可解释的文本生成算法

李磊

字节跳动人工智能实验室

2021/5/28

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 plays big role in content creation!

Why is NLG important?

Machine Translation



Machine Writing



ChatBOT



Question Answering



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



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.

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.

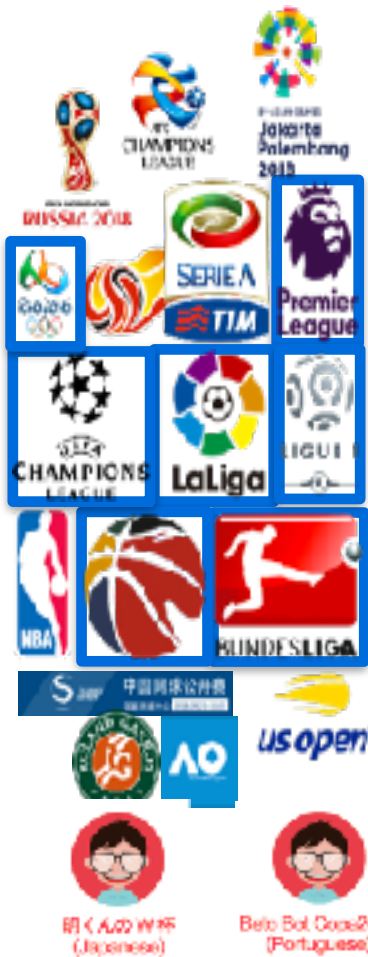
} human
written

} GPT3,
edited
by
human

Xiaomingbot

Automatic News Writing

Winning 2017 Wu Wen-tsün Award in AI from CAAI



北京时间2016年1月24日的下午, 世界杯小组赛, 比利时对阵俄罗斯。最终比利时以2比0战胜俄罗斯, 卢卡库、巴里奥斯, 阿扎尔为本队进球, 斯塔里、耶德瓦为客队进球。... 进球, 帮助为本队取得领先。



Post:
Thomas Strakosha's 4 saves did not stop Lazio from defeat against Inter Milan, final score 0: 3



Following - Xiaomingbot-European
Marseille dropped a 0: 2 decision against PSG in Ligue 1

Following - Xiaomingbot-European
Sevilla took away a victory against Huesca, 2: 1



600,000 articles

6 lang

150,000 Followers On Toutiao

Joint w/
Xiaojun Wan
@ PKU

Xiaomingbot : Multilingual Robot News Reporter



ByteDance AI Lab
字节跳动人工智能实验室

**MULTILINGUAL ROBOT
NEWS REPORTER**

--- Xiaomingbot ---



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 2020.⁶

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)

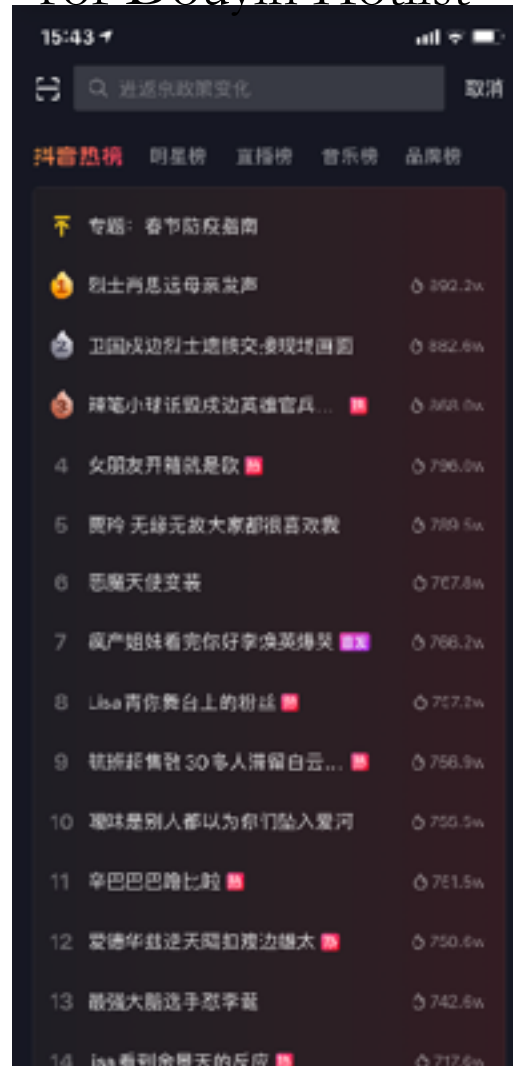
(positions after mapping)

Text Generation for Mobile Apps

Generating Entity Short Description



Generating Description for Douyin Hotlist



Description of E-Commerce Goods



Text Generation for Mobile Apps (Cont')

Explanation for Recommendation



小伙伴们还喜欢



E-commerce Product Subtitle



prich荷叶边金属扣半身裙

裙身的金属扣装饰, 打破了纯色的单调感

韩国品牌 PRICH官方旗舰店

¥349



ELAND衣恋夏装米色褶皱休闲

腰部抽褶设计, 丰富了裙身的层次感

韩国品牌 ELAND官方旗舰店

¥179



chuu韩版纯色高腰百褶裙

青春洋溢的少女气息, 减龄神器

韩国品牌 CHUU海外旗舰店

¥159



scofield通勤简约半身裙

简约的竖条纹设计, 优雅大方

韩国品牌

¥1422

Product highlights



华为智能全网通手机



李白写作

前置一体指纹, 圆润轻薄机身, 搭配华为巴黎美学研究所设计的优雅UI主题, 内外兼美。还可以通过双镜头特有的大光圈效果, 虚化纷繁背景, 你就是焦点。

Unified Framework: Conditional Sequence Generation

Unified encoder-decoder architectures

- Machine Translation
- Dialog Generation
- Table-to-Text
- Question Answering
- ...

Output 敏捷的棕狐跳过懒狗

$$P_{\theta}(y|x) = \prod_{i=1}^n P_{\theta}(y_i | y_{<i}, x)$$

Input

The quick brown fox jumps over the lazy dog .

(Partial) Challenges in NLG

- Blackbox - hard to interpret
- Data scarcity
 - Limited or none parallel data for most text generation tasks (except for high-resource MT)
- Controllability
 - surface form
 - Semantics
- Scalability
 - Inference complexity (\$\$)
 - Training convergence (\$)

Outline

- Learning Disentangled Representation for Text Generation
- Monte-Carlo Generation under Constraints
- GLAT: High-performant Parallel Generation

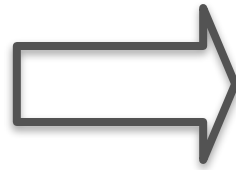
Variational Template Machine for Data-to-Text Generation

Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, Lei Li



Generating Natural Language Description from Data (Table)

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

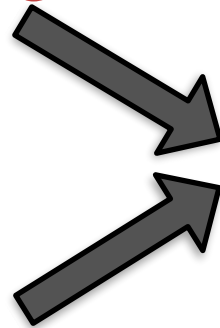
Person Biography

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

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

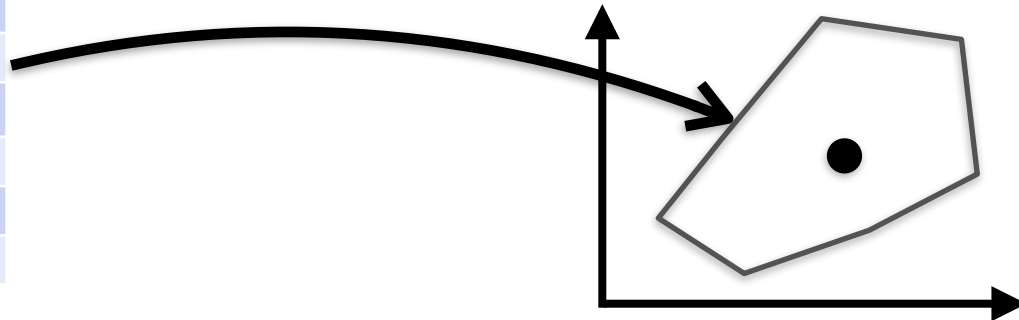
Why is it difficult?

- Desired Properties:
 - Accuracy: semantically consistent with the content in the table
 - Diversity: Ability to generate infinite varying utterances
- Writing many templates is labor intensive
- Limited <table entry, text> pairs
 - Inadequate to train a seq2seq/Transformer model
 - But large raw text can be mined

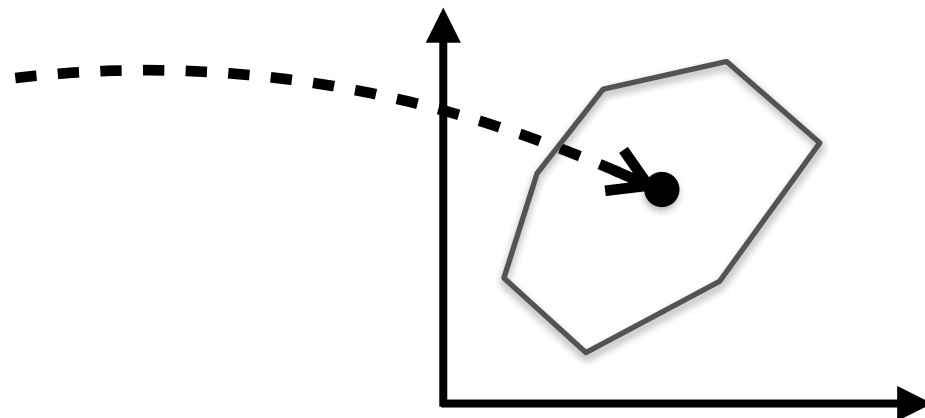
Motivation: Generating from (semantic meaningful) latent representations

- Disentangled latent representations for semantic content and templates.

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



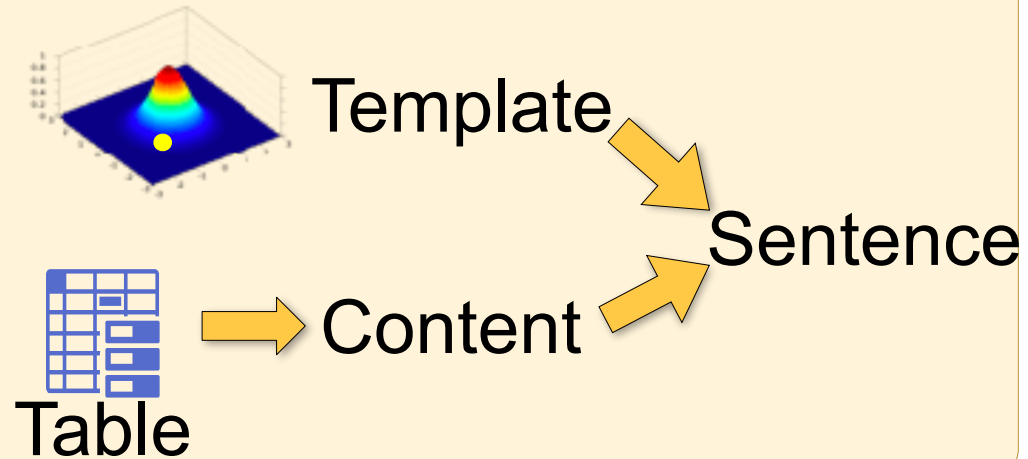
[name] is a [food] restaurant.
It is a [eatType] and it has
a [price] cost and [rating]
rating. It is in [area].



Generating from (semantic meaningful) Latent Factors

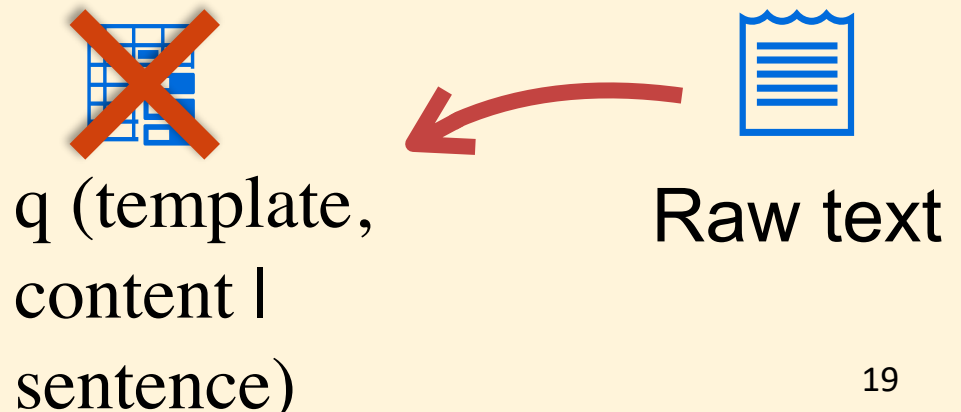
Motivation 1:

Continuous and **disentangled** representation for template and content



Motivation 2:

Incorporate **raw text corpus** to learn good representation.



Variational Template Machine

Input: triples of <field_name, position, value>

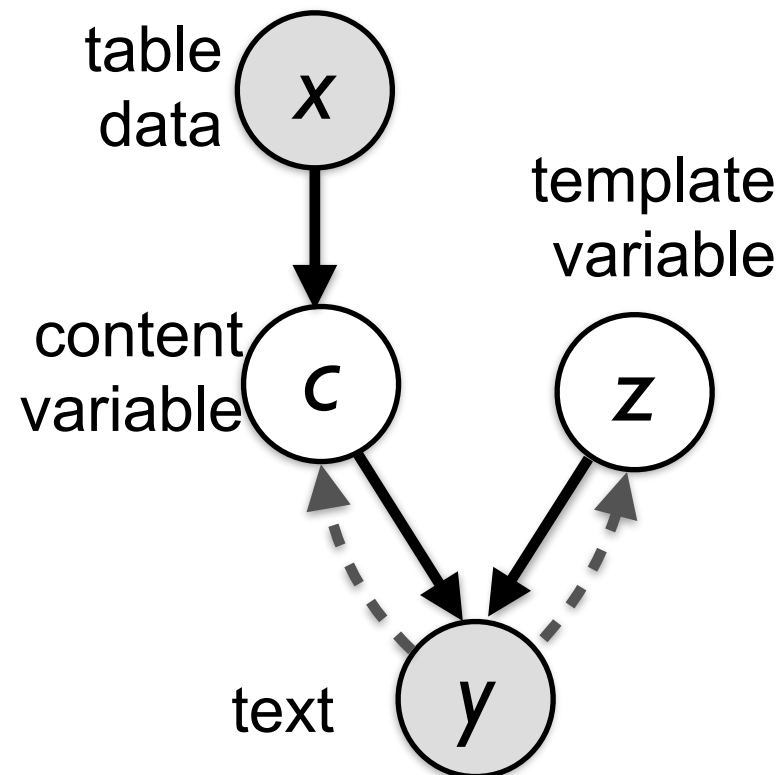
$$\{x_k^f, x_k^p, x_k^v\}_{k=1}^K$$

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

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

2. Sample $z \sim p_0(z)$, e.g. Gaussian

3. Decode y from $[c, z]$ using another NN (e.g. Transformer)

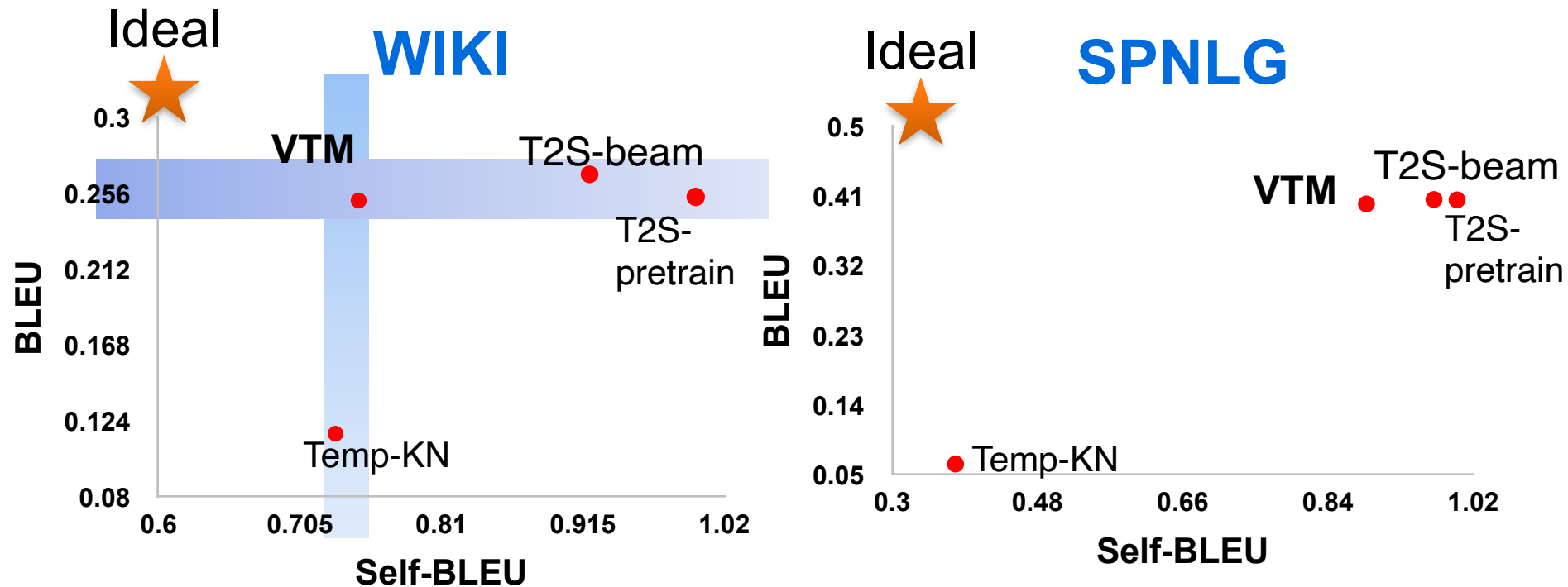


Learning with Raw Corpus

- Semi-supervised learning: “Back-translate” corpus to obtain pseudo-parallel pairs $\langle \text{table}, \text{text} \rangle$, to enrich the learning

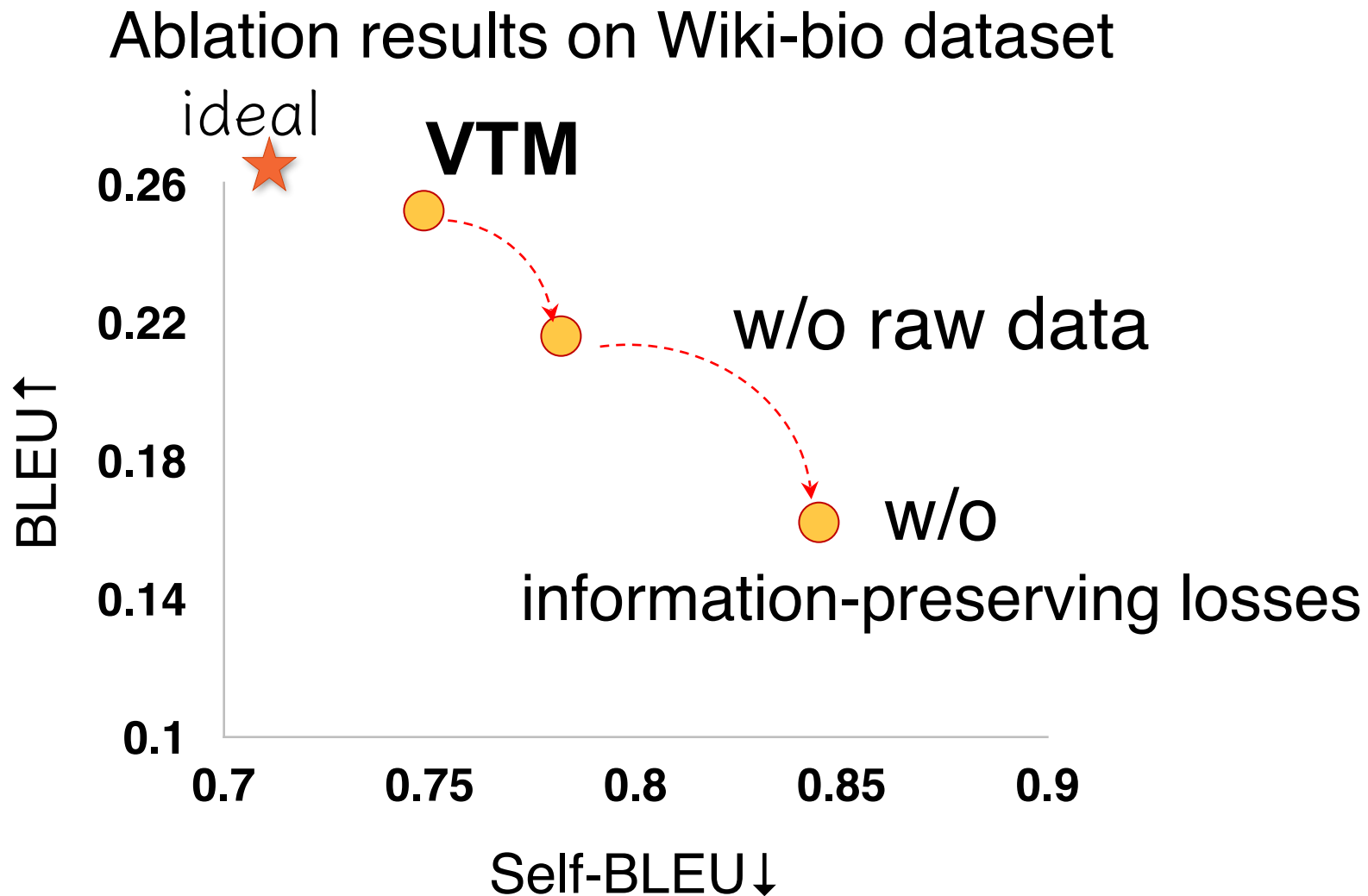
Table		Text
name	Sukiyaki	Sukiyaki is a Japanese restaurant. It is a pub and it has a average cost and good rating. It is in seattle .
eatType	pub	
food	Japanese	
price	average	
rating	good	
area	seattle	
?		Known for its creative flavours, Holycrab's signatures are the Hokkien crab.
$q(\langle c, z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

VTM Produces High-quality and Diverse Text



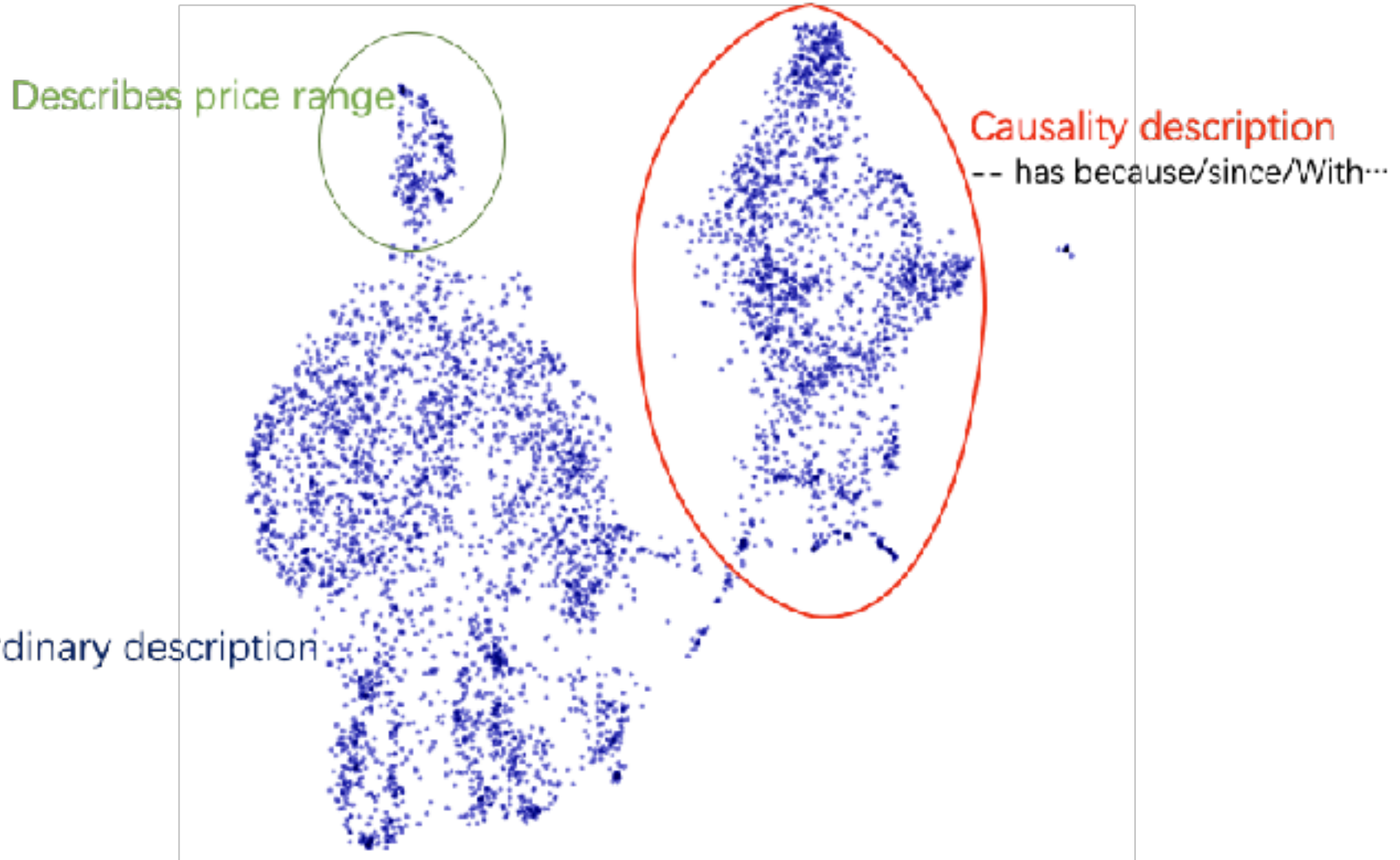
VTM uses beam-search decoding.

Raw data and loss terms are necessary



Interpreting VTM

Template variable project to 2D



VTM Generates Diverse Text

Input Data Table

Jack Ryder



Ryder in about 1930

Personal information

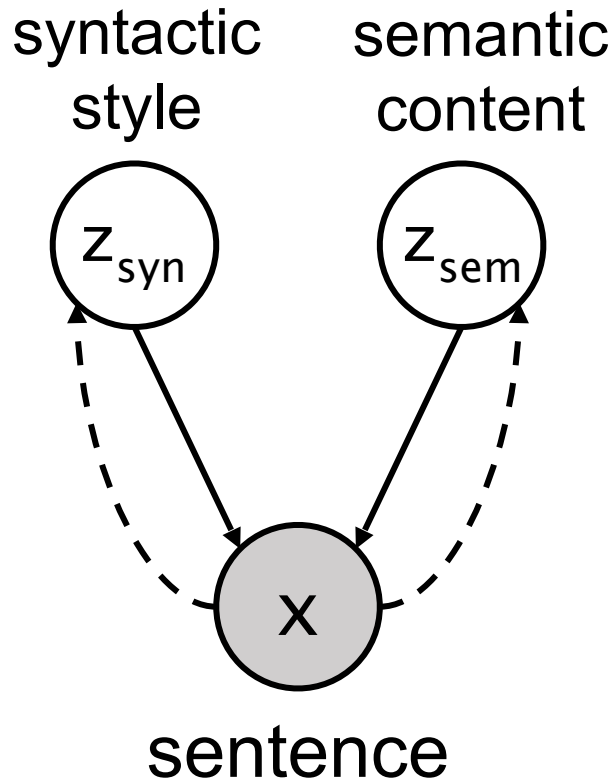
Full name	John Ryder
Born	8 August 1889 Collingwood, Victoria, Australia
Died	3 April 1977 (aged 87) Fitzroy, Victoria, Australia
Nickname	The King of Collingwood
Height	1.85 m (6 ft 1 in)
Batting	Right-handed
Bowling	Right-arm medium pace
Role	All-rounder

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



Syntax provider

Semantic content

There is an apple
on the table

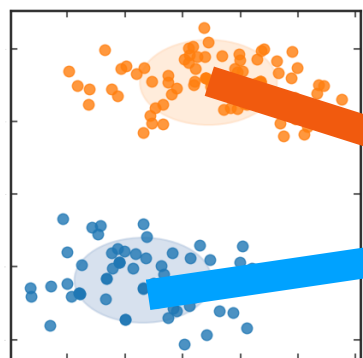
The dog is
behind the door

DSSVAE

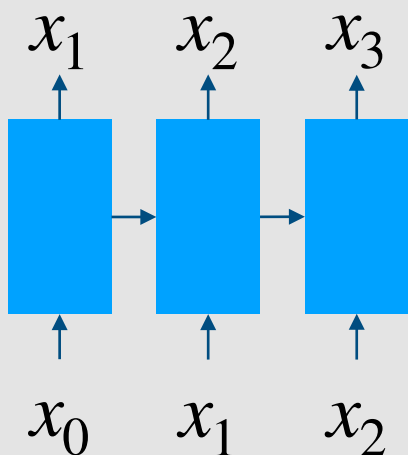
There is a dog behind the door

Discrete Latent Variables Enhance Interpretability

Latent structure
dialog actions



GENERATOR



Sampling

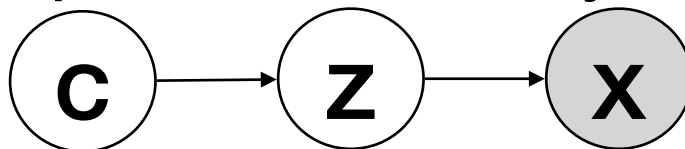
“Remind me about
the football game.”

[action=remind]

“Will it be overcast
tomorrow?”

[action=request]

Dispersed Exponential-family Mixture VAE

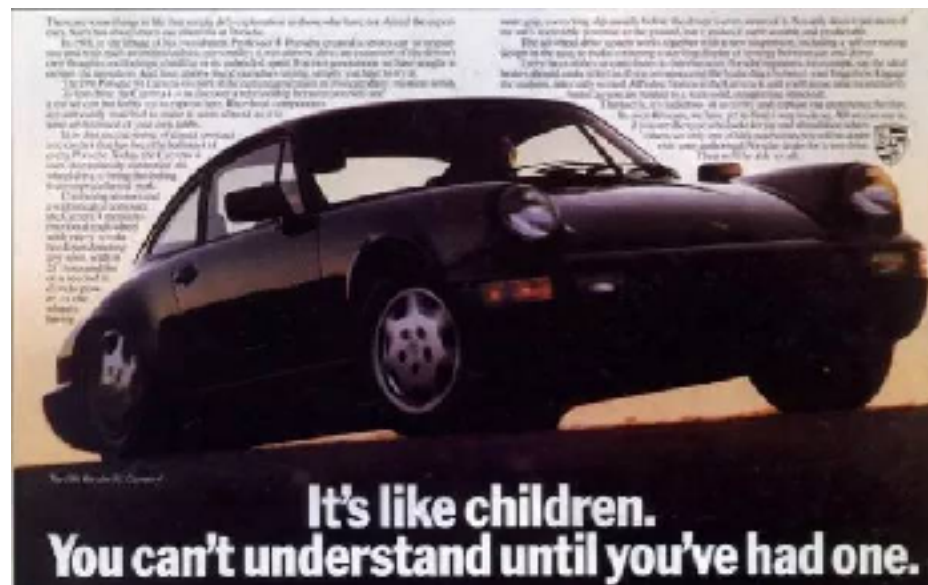


CGMH: Constrained Sentence Generation by Metropolis-Hastings Sampling

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, Lei Li



Automate Creative Advertisement Design



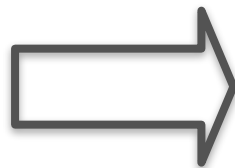
Constrained Text Generation

To generate sentences that are:

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

“Autumn”

“Sports shoes”



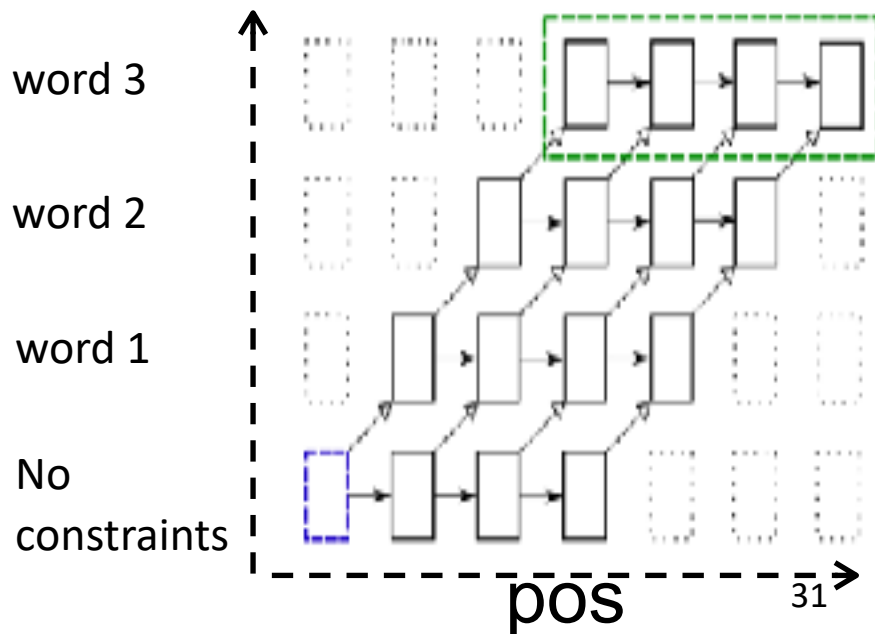
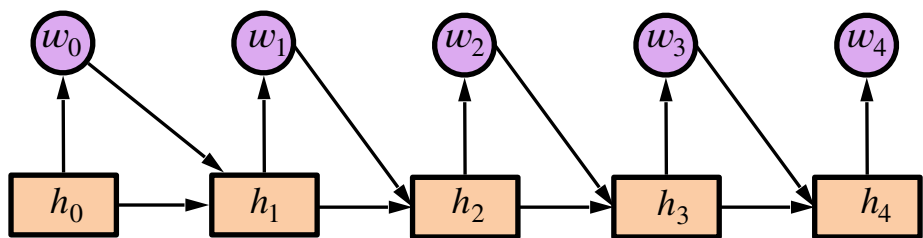
Comfortable **sports shoes**,
a breathing pair of man's
shoes, accompanying you
in **autumn**

Why is Constrained Text Generation difficult?

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

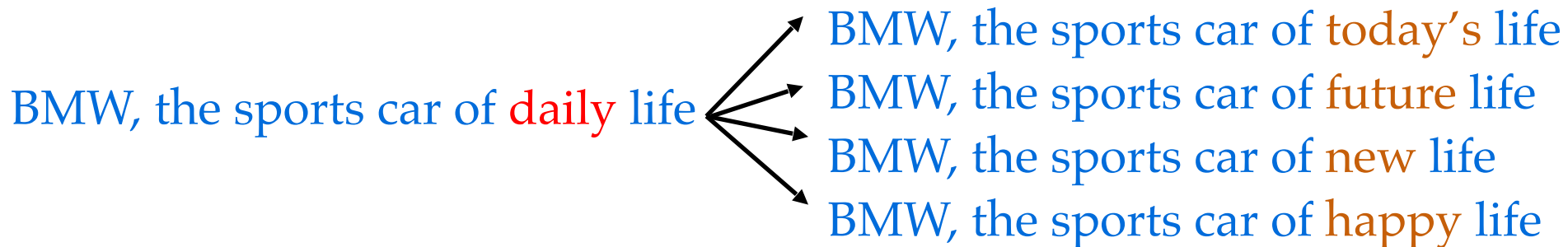
RNN grid beam search [Hokamp & Liu 2017]

does not usually produce high quality sentences



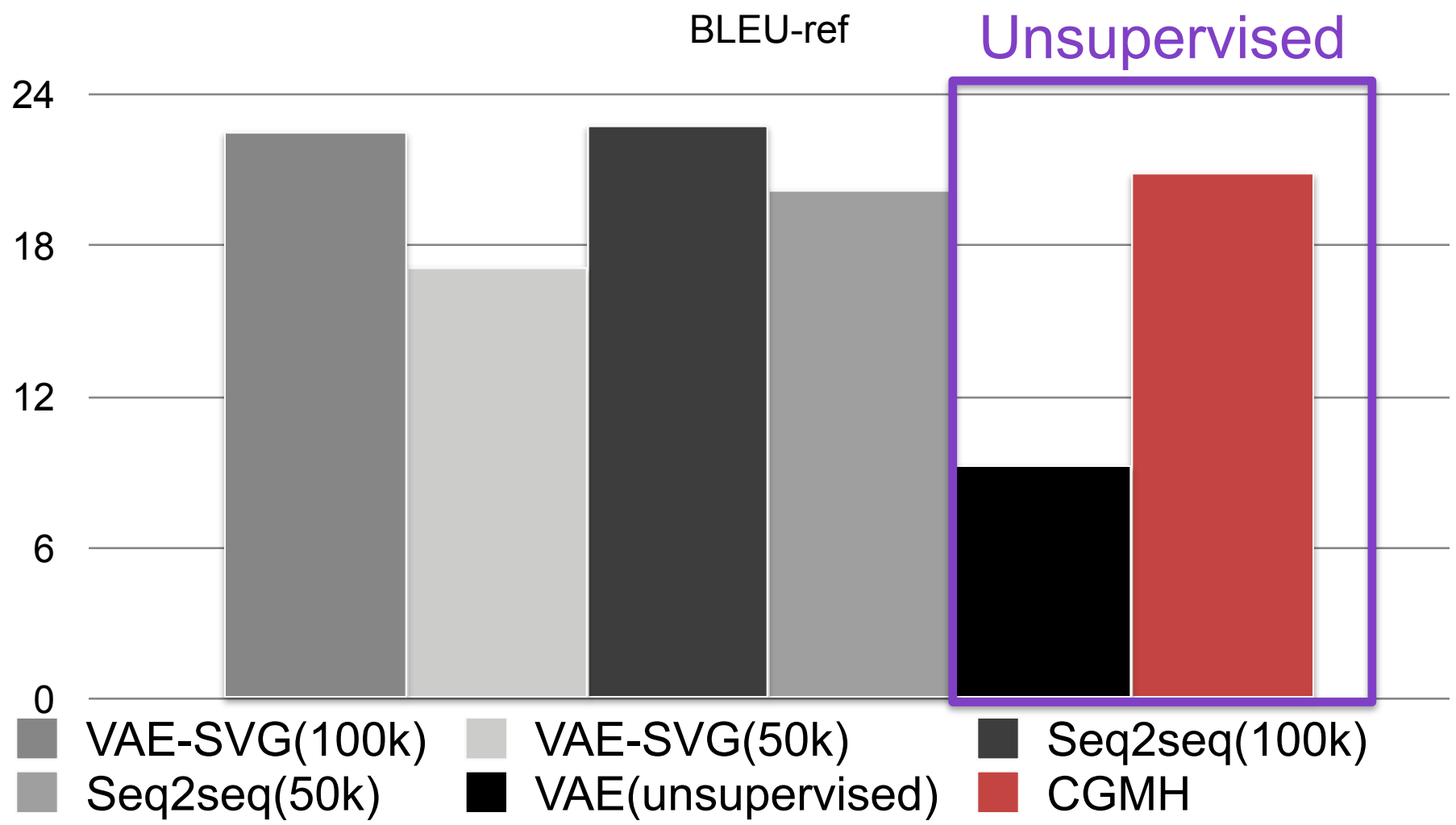
CGMH: Main Idea

- 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



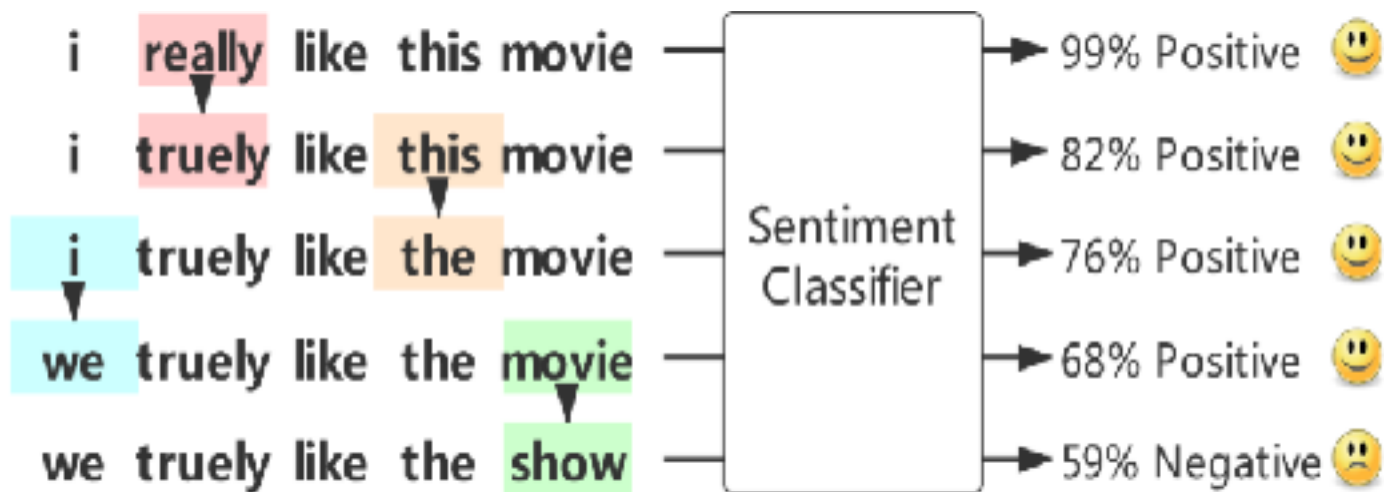
...

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



Generating Adversarial Fluent Sentence Generation

- 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



Generation under Combinatorial Constraints

- Logical and Combinatorial constraints

$$\pi(x) = \underbrace{P_{\text{LM}}(x; \theta)}_{\text{Language Model}} \cdot \underbrace{\phi(x)}_{\text{Constraint}}$$

$$\phi(x) = \beta^{M - \sum_i c_i(x)}, \quad 0 < \beta < 1$$

$c_i(x)$ is a formula or logical constraint. e.g. the first word must be Wh- words.

Method: Tree search enhanced Metropolis-Hastings

details in TSMH [M. Zhang, N. Jiang, **Lei Li**, Yexiang Xue, EMNLP20e]₃₆

Impact

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

“Autumn”

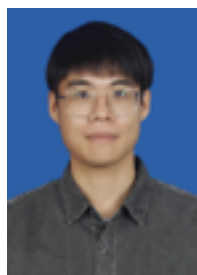
“Sports shoes”



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

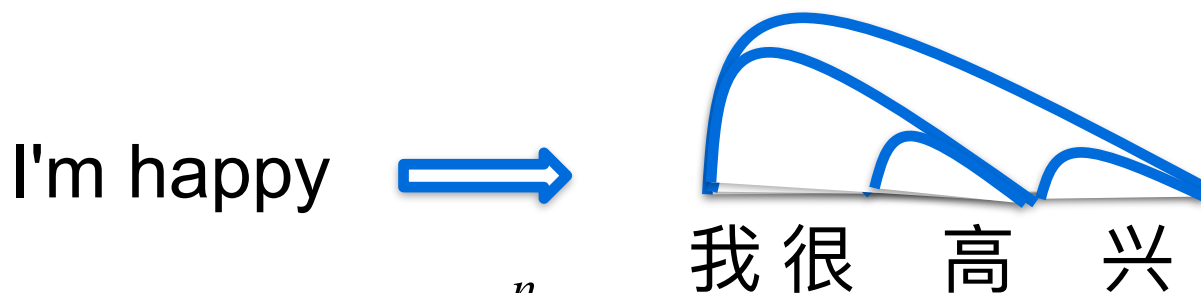
Glancing Transformer (GLAT)

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang,
Lin Qiu, Weinan Zhang, Yong Yu, Lei Li



Autoregressive Seq Generation Model

LSTM, Transformer, GPT all generates token by token sequentially



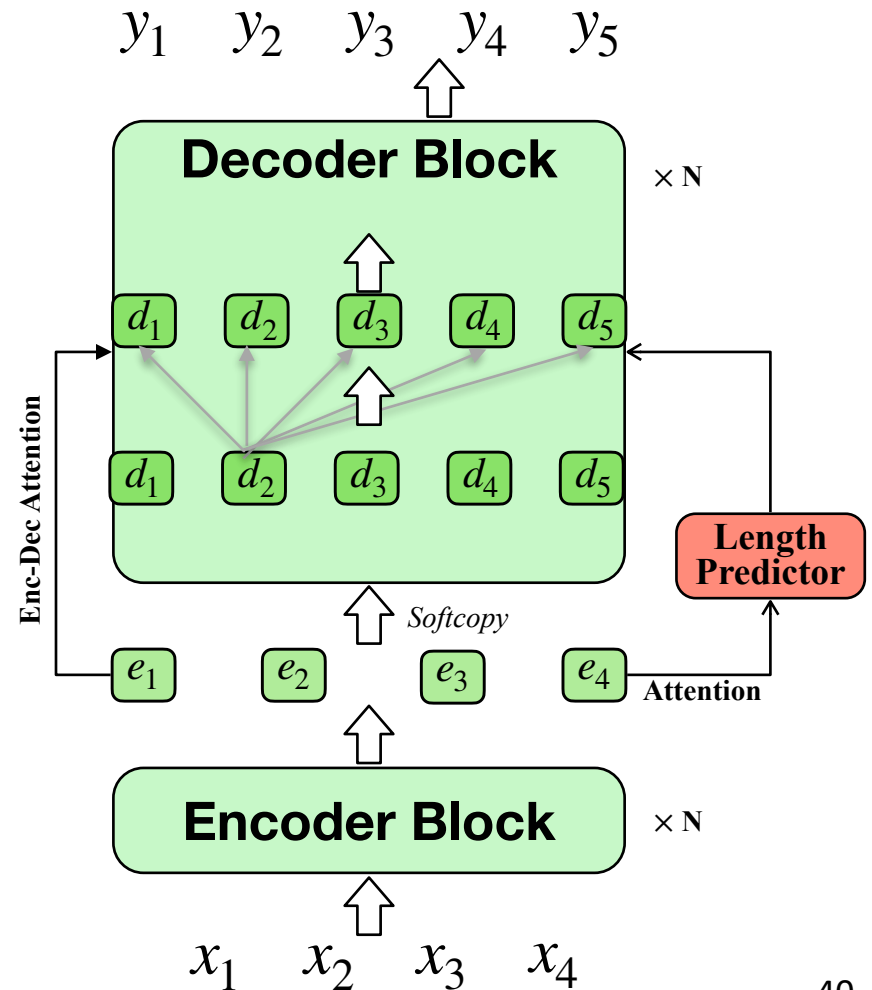
$$p_{\theta}(y | x) = \prod_{i=1}^n p_{\theta}(y_i | y_{<i}, x)$$

Issues:

1. Auto-regressive factorization may not be ideal for computer text generation (teacher forcing??)
2. Inefficient to generate long sequence

Non-autoregressive Decoding

- Various techniques proposed since 2018's vanilla Non-autoregressive Transformer (NAT)
- NAT+CTC [Libovicky 2018]
- Flowseq [Ma 2019]
- PNAT [Bao 2019]
- Iterative NAT
 - CMLM [Ghazvininejad 2019]
- ...

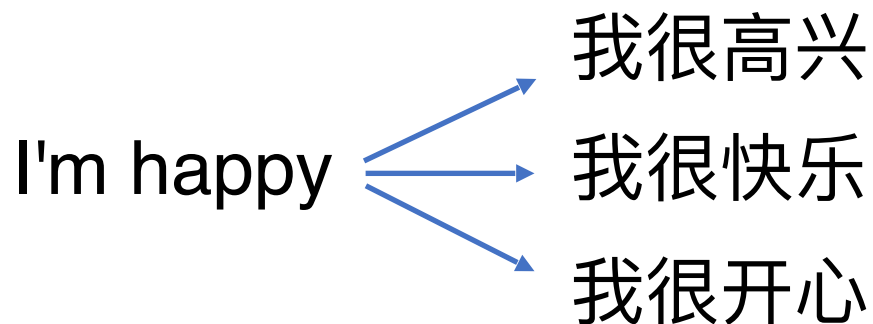


Limitations of current NAT

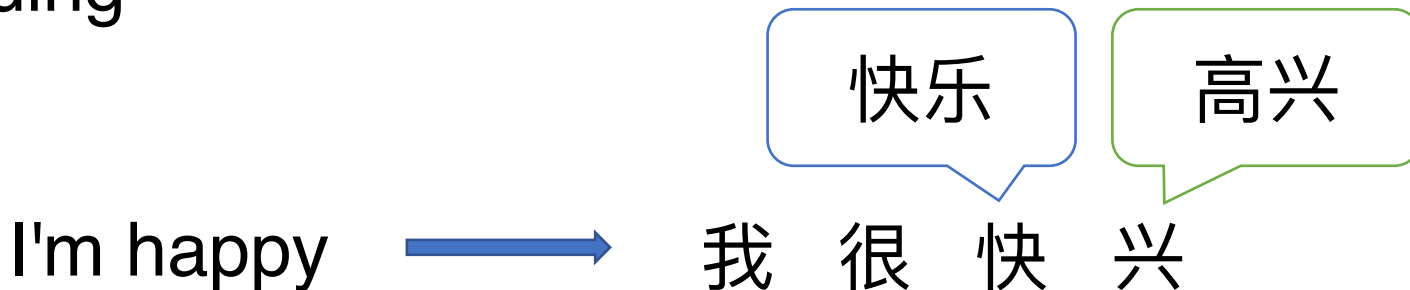
- Vanilla NAT
 - performance gap of 7 BLEU
- Iterative NAT
 - Need decode multiple times
 - None or limited speedup
- Relies on additional separately trained Transformer (AT)
 - Knowledge distillation
 - Re-ranking

Why is it Difficult?

- One input -> multiple candidate targets



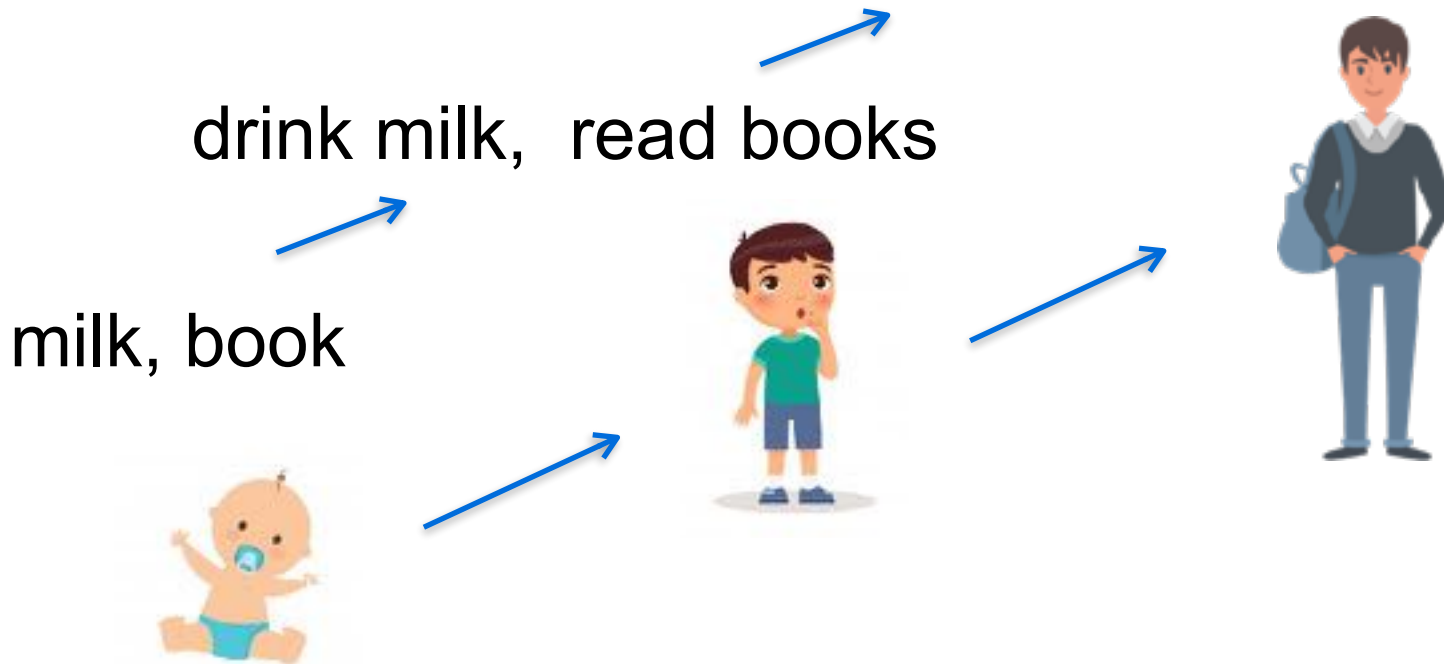
- The inconsistency during non-autoregressive decoding



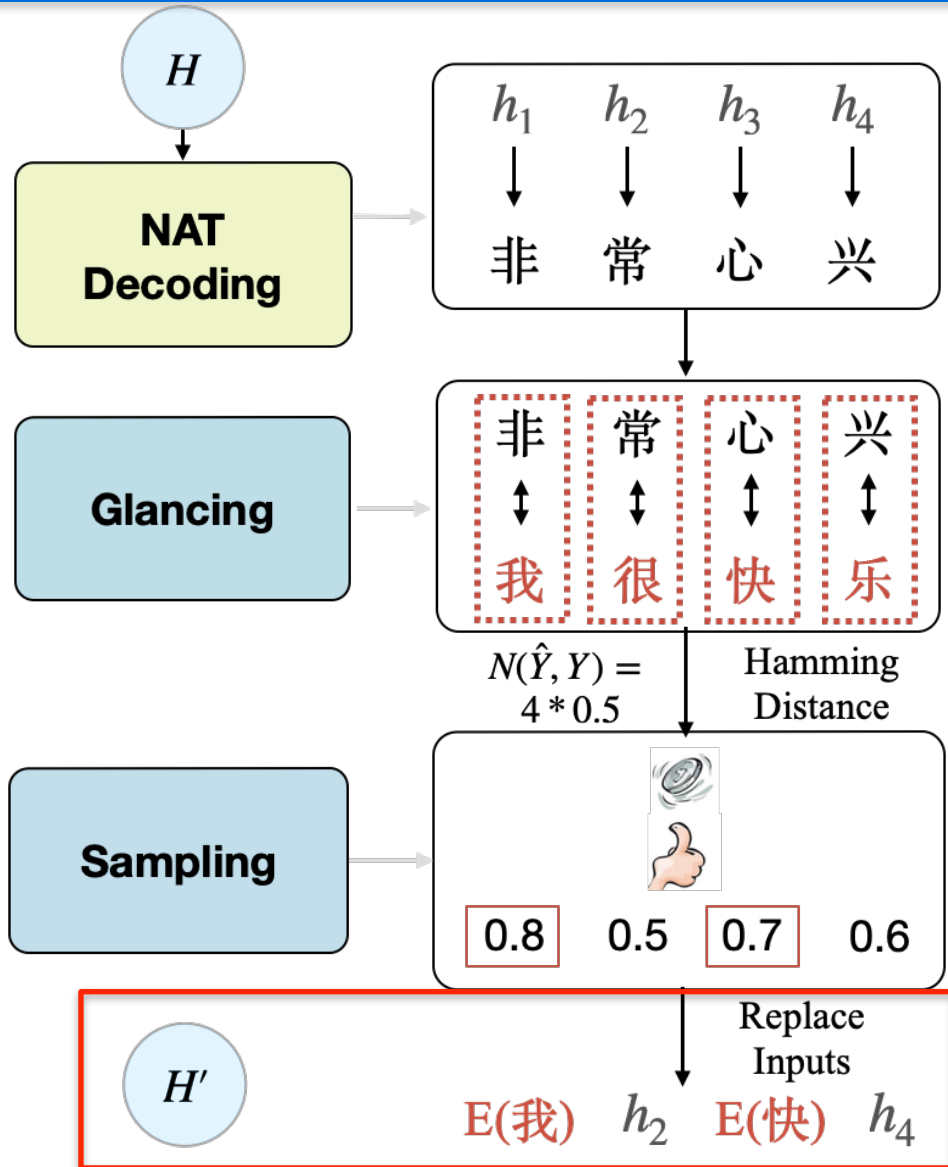
Key Idea: Progressive Training by Glancing Language Model

Progressively learn to generate more complex fragments or sequences from partial target hints

I sit on my chair, reading books and drinking milk.



Glancing Transformer



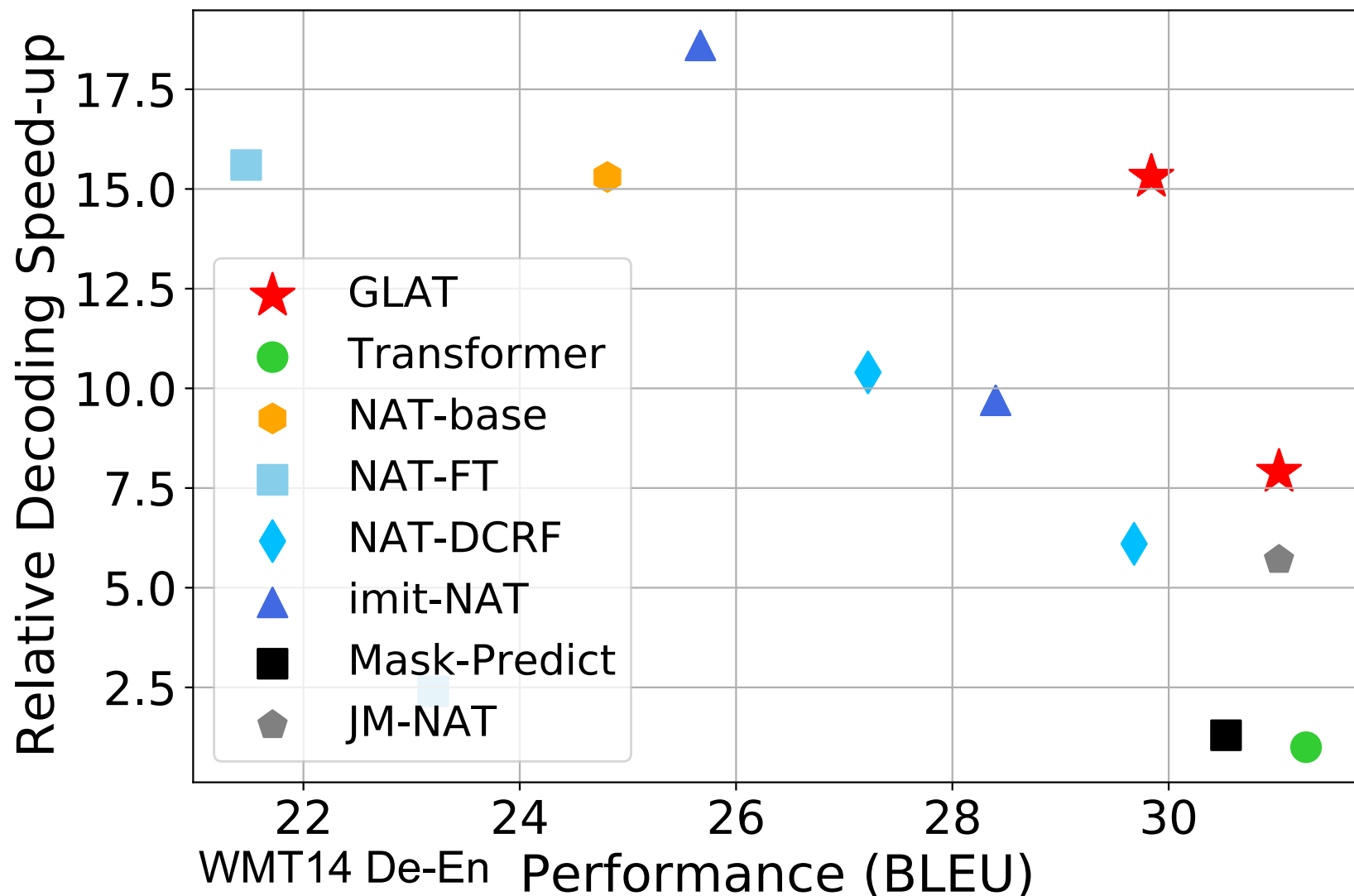
Training GLAT

1. Copy encoder embedding as decoder input
2. Generate with current model
3. Select tokens randomly proportional to difference
4. Replace the original decoder inputs with the embedding of sampled target words

Inference:

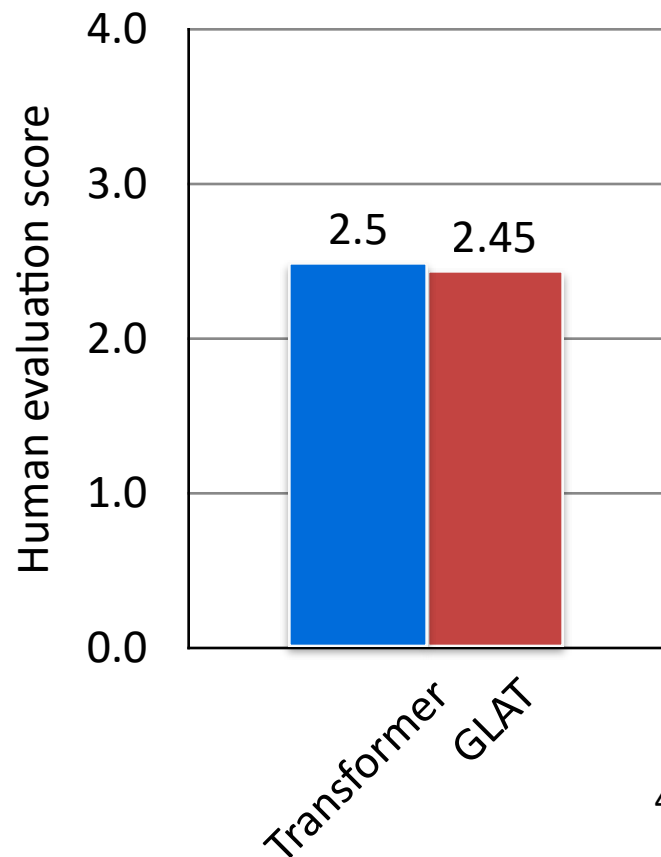
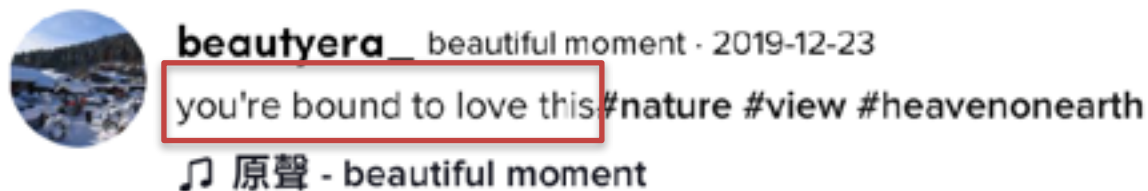
- Only single-pass predict

GLAT is the first Non-autoregressive Model to get on-par with AT (<0.3 BLEU)



Impact: the 1st deployed Non-Auto model 5-15x speedup with comparable quality!

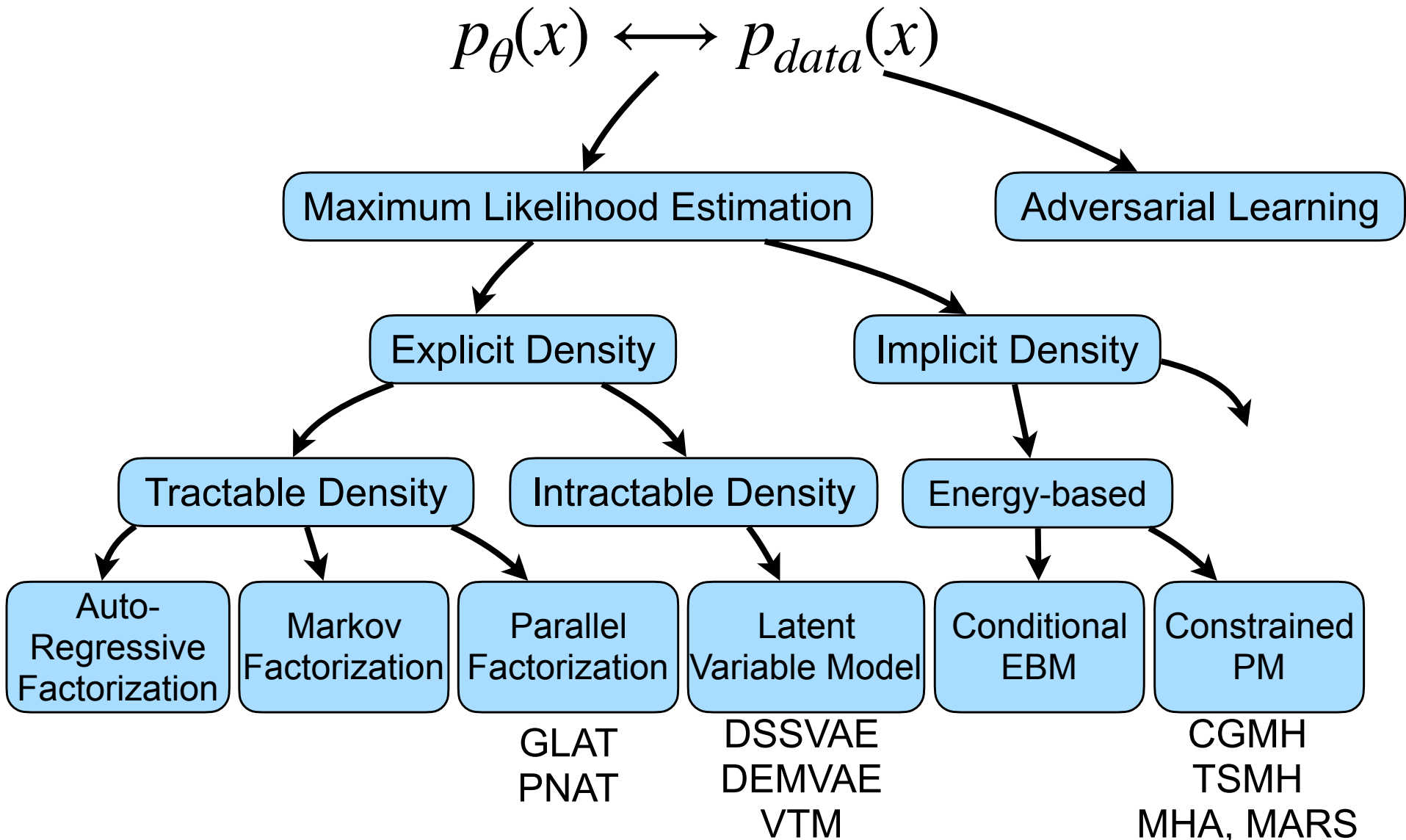
- Deployed on VolcTrans, for TikTok en->ja



Summary

- 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
- Bayesian approach to constrained text generation
 - CGMH: generic framework to specify constraints and generate
 - MHA, TSMH
- Parallel Generation
 - GLAT: 10x speedup with comparable performance

DGM Taxonomy



Takeaway Message

- Generation is fundamental for understanding data
- Interpretability from probabilistic (discrete) latent variable
 - Variational methods to learn the intractable objective
- Explicit control by enforcing constraints in generation
 - a generic method to adapt any pre-trained language model
- Thinking about serving: Need for Speed
 - Glancing techniques will transform the text generation: better and faster, both! How to achieve?

For the Community

Variational Template Machine

<https://github.com/ReneeYe/VariationalTemplateMachine>

CGMH code and examples

<https://github.com/NingMiao/CGMH>



<https://github.com/bytedance/lightseq>

A high performance sequence processing lib



<https://translate.volcengine.cn>

火山翻译

Thanks

- Joint w/ Hao Zhou, Rong Ye, Ning Miao, Wenxian Shi, Huangzhao Zhang, Yu Bao, Mingxuan Wang, Shujian Huang (NJU), Lili Mou (U. Alberta), Rui Yan (PKU), Maosen Zhang (Purdue), Yexiang Xue (Purdue), Nan Jiang (Purdue), Lin Qiu, Weinan Zhang (SJTU), Yong Yu (SJTU)
- Papers / code / datasets can be found at <https://lileicc.github.io>

Reference

1. 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.
2. 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.
3. 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 Meeting of the Association for Computational Linguistics (ACL). July 2019.
4. 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.
5. 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.
6. 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.
7. Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, Lei Li. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.