ICLR 2020

Learning Deep Latent Models for Text Sequences

Lei Li

ByteDance AI Lab

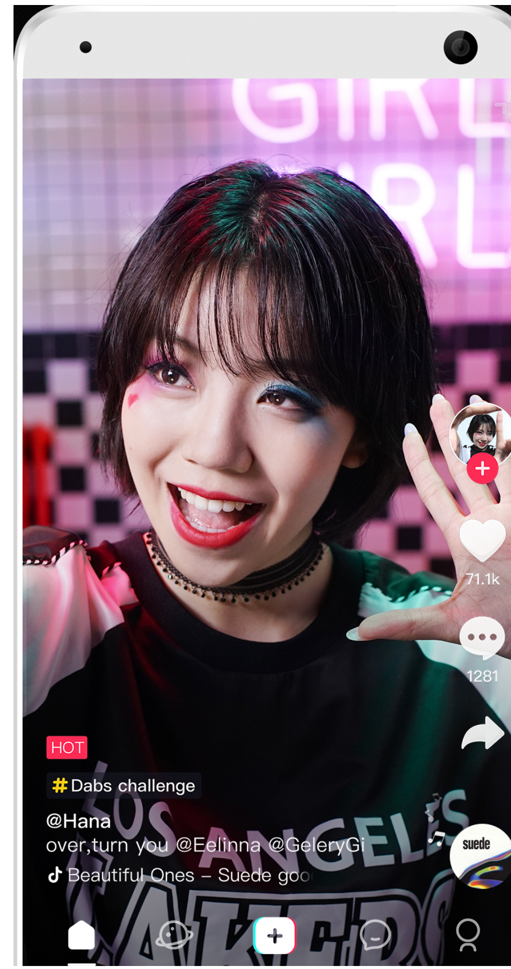
4/29/2020

The Rise of New Media Platforms

Toutiao

Helo

Douyin/Tiktok



Huge Demand for Automatic Content Generation Technologies

- Automatic News Writing
- Author writing assist tools
  - Title generation and text summarization
- Automatic Creative Advertisement Design
- Dialog Robots w/ response generation
- Translation of content across multiple languages
- Story Generation

The New York Times

Soon a Robot Will Be Writing This Headline

Gabriel Alcala

BUY BOOK ▾ When you purchase an independently reviewed book through our site, we earn an affiliate commission.

By Alana Semuels

Jan. 14, 2020

Automated News Writing

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

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

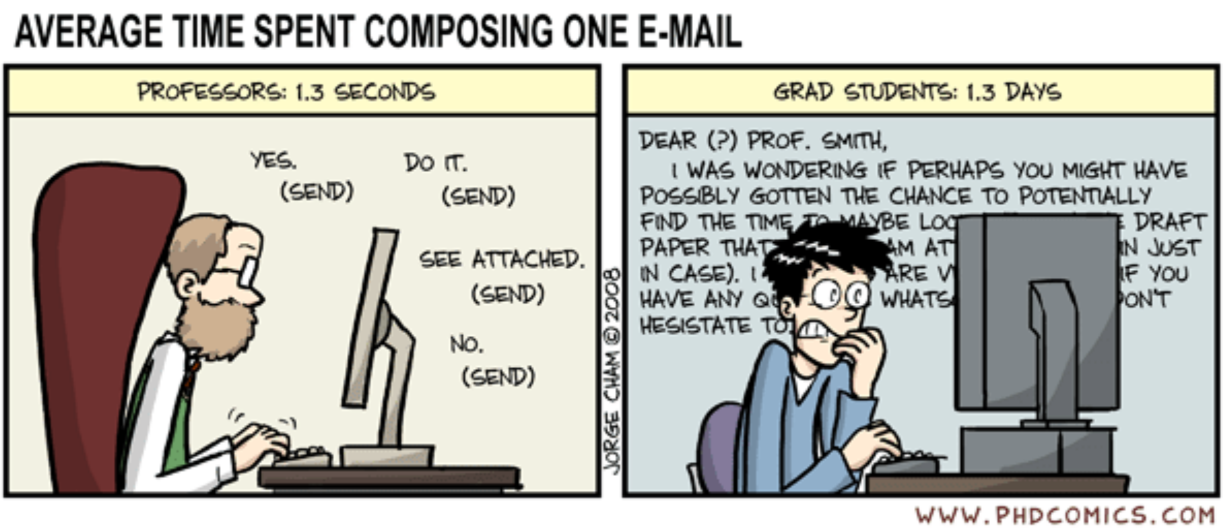
AI to Improve Writing

Text generation to rescue!

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



Outline

1. Overview
2. Learning disentangled latent representation for text
3. Mirror-Generative NMT
4. Multimodal machine writing
5. Summary

Disentangled Latent Representation 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]

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].

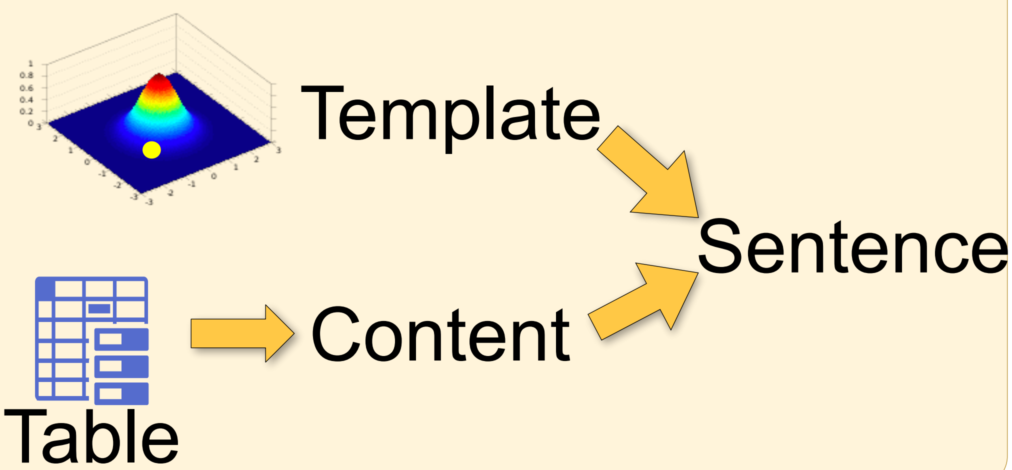
name	Sukiyaki
eatType	pub
food	Japanese
price	average
rating	good
area	seattle

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

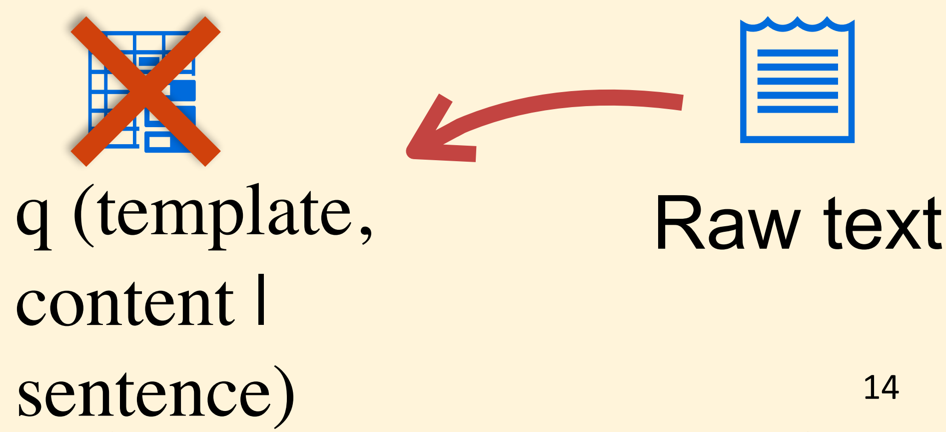
But manually creation of templates are tedious

Our Motivation for Variational Template Machine

Motivation 1:
Continuous and disentangled representation for template and content

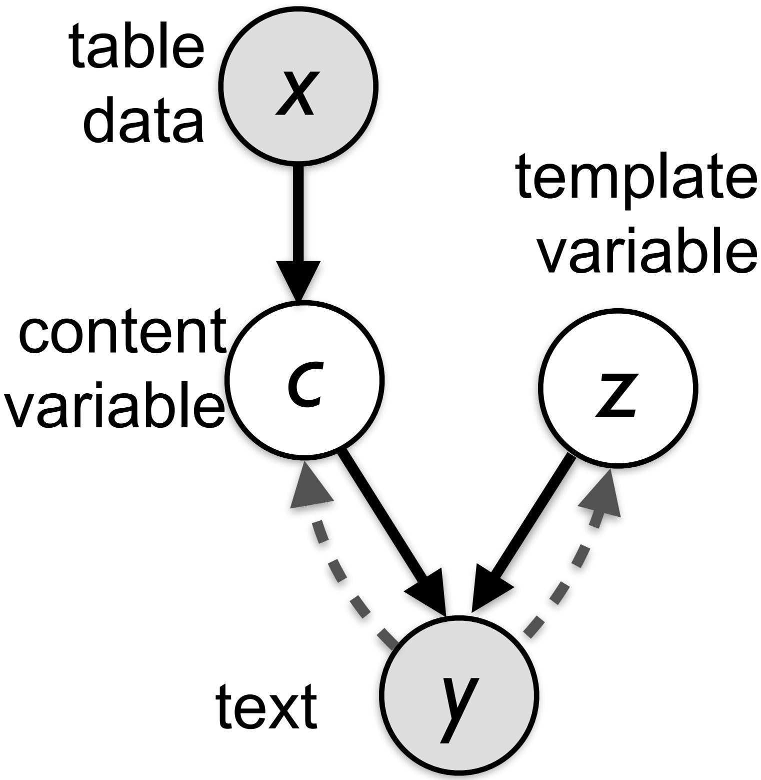


Motivation 2:
Incorporate raw text corpus to learn good representation.



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

Variational Template Machine



Input: triples of <field_name, position, value>

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

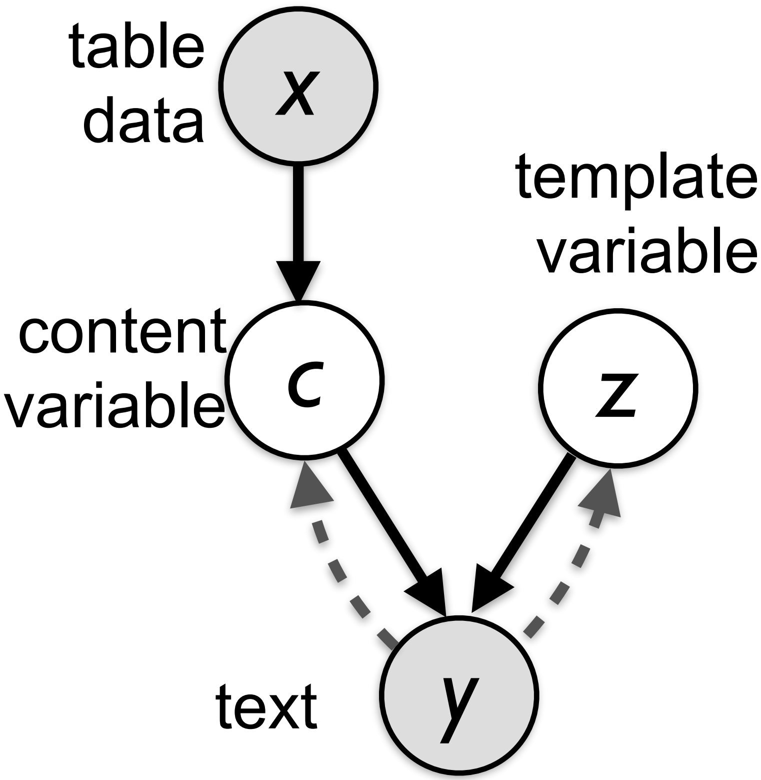
1. $p(c \mid x) \sim$ Neural Net

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

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!

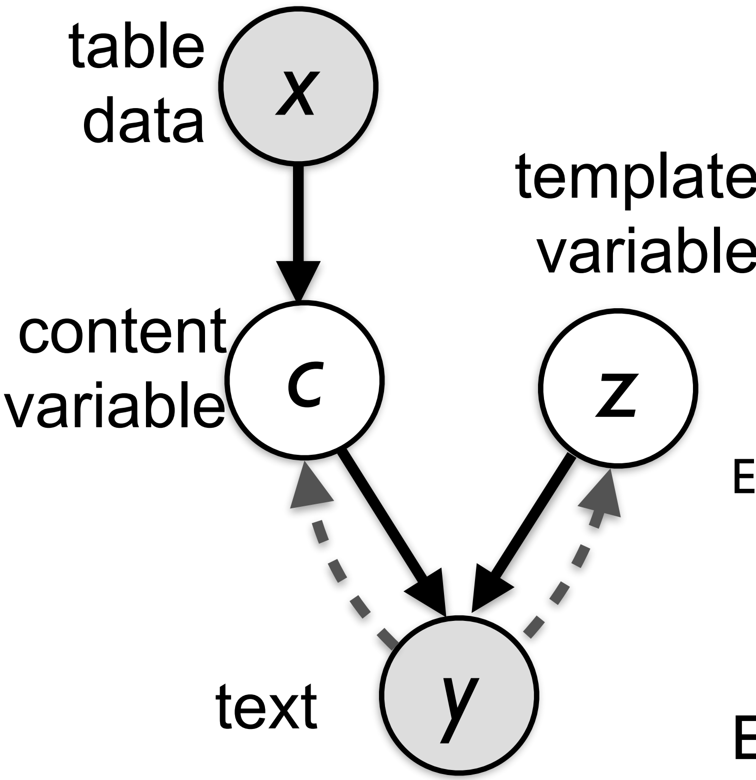
Total loss =

Reconstruction loss

+

Information-Preserving loss

Variational Inference



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

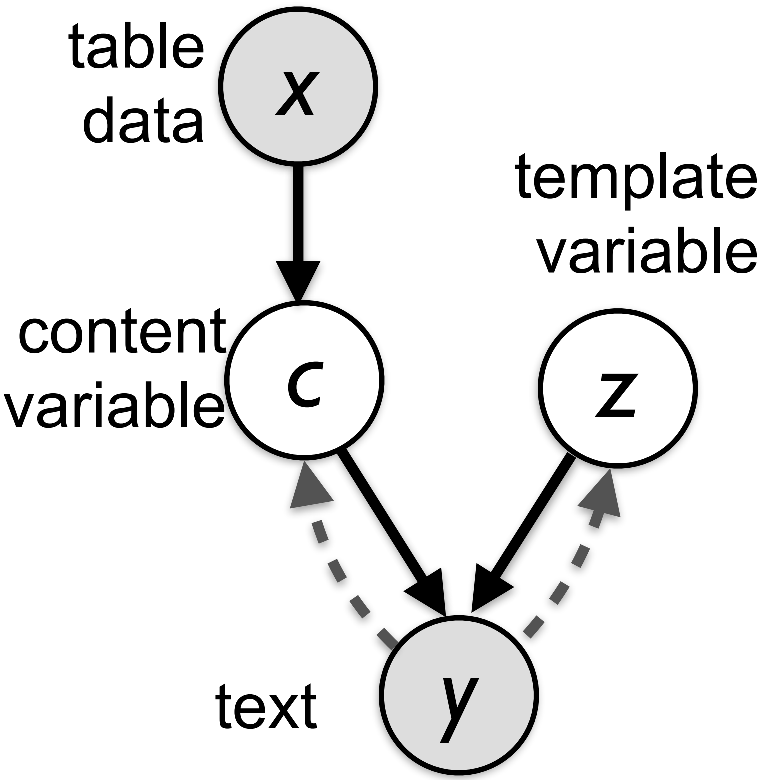
$$\cancel{l_p = -\mathrm{E}\log\int p\left(y\middle|c(x),z\right)p(z)\mathrm{d}z}$$

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

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

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

Preserving Content & Template



1. Content preserving loss

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

2. Template preserving loss of pairs

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

$\tilde{y}$ 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

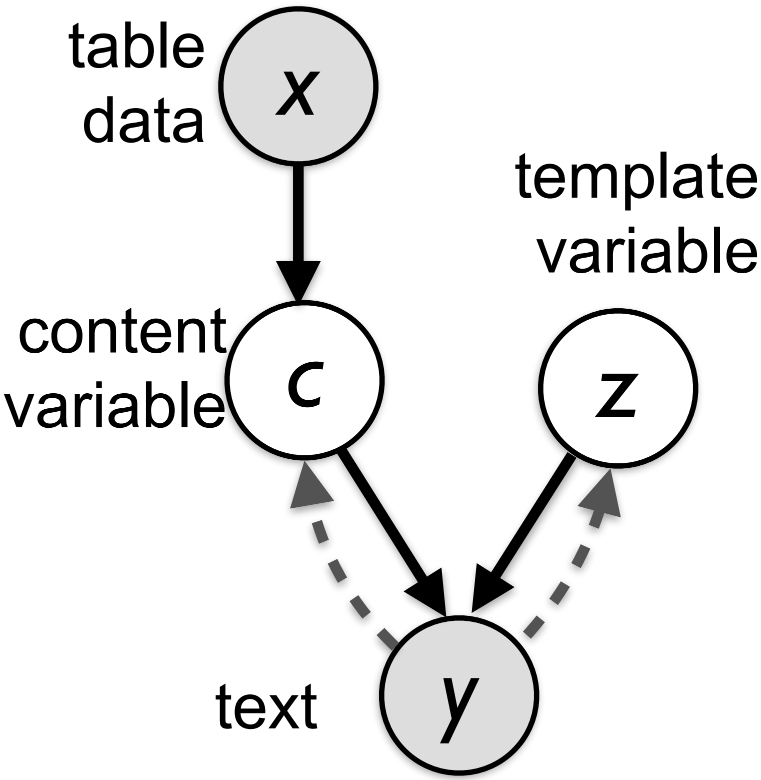


Table data x:

{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.

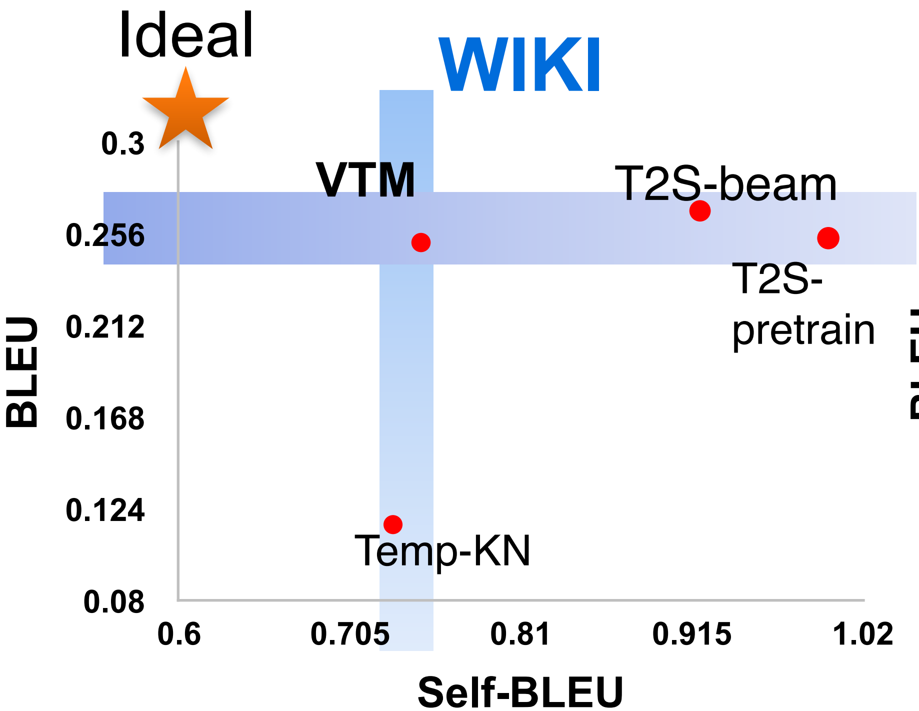
Evaluation Setup

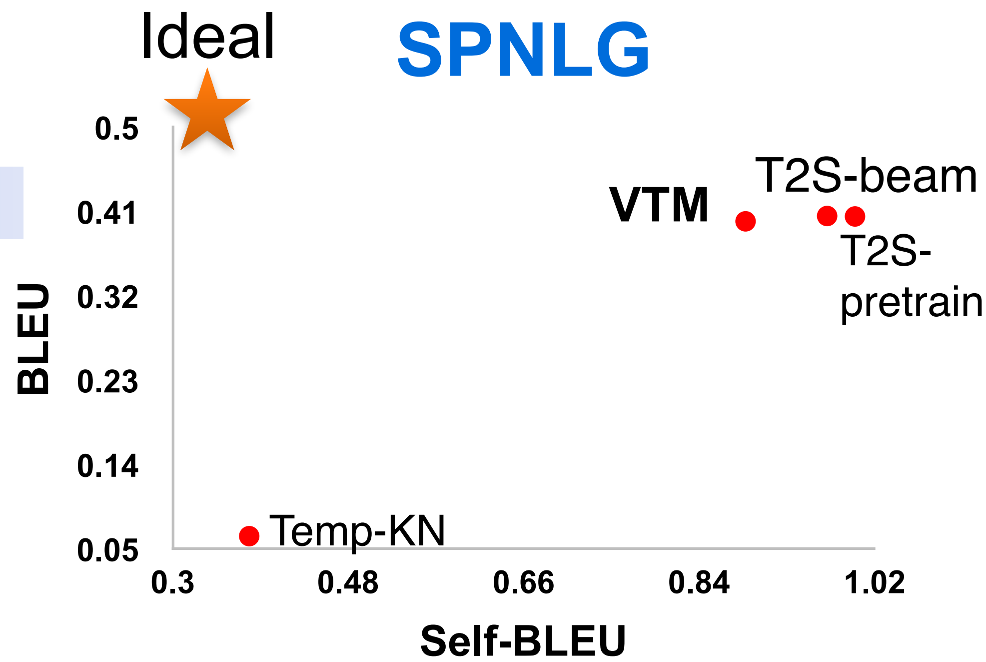
- Tasks
  - WIKI: generating short-bio from person profile.
  - SPNLG: generating restaurant description from attributes

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

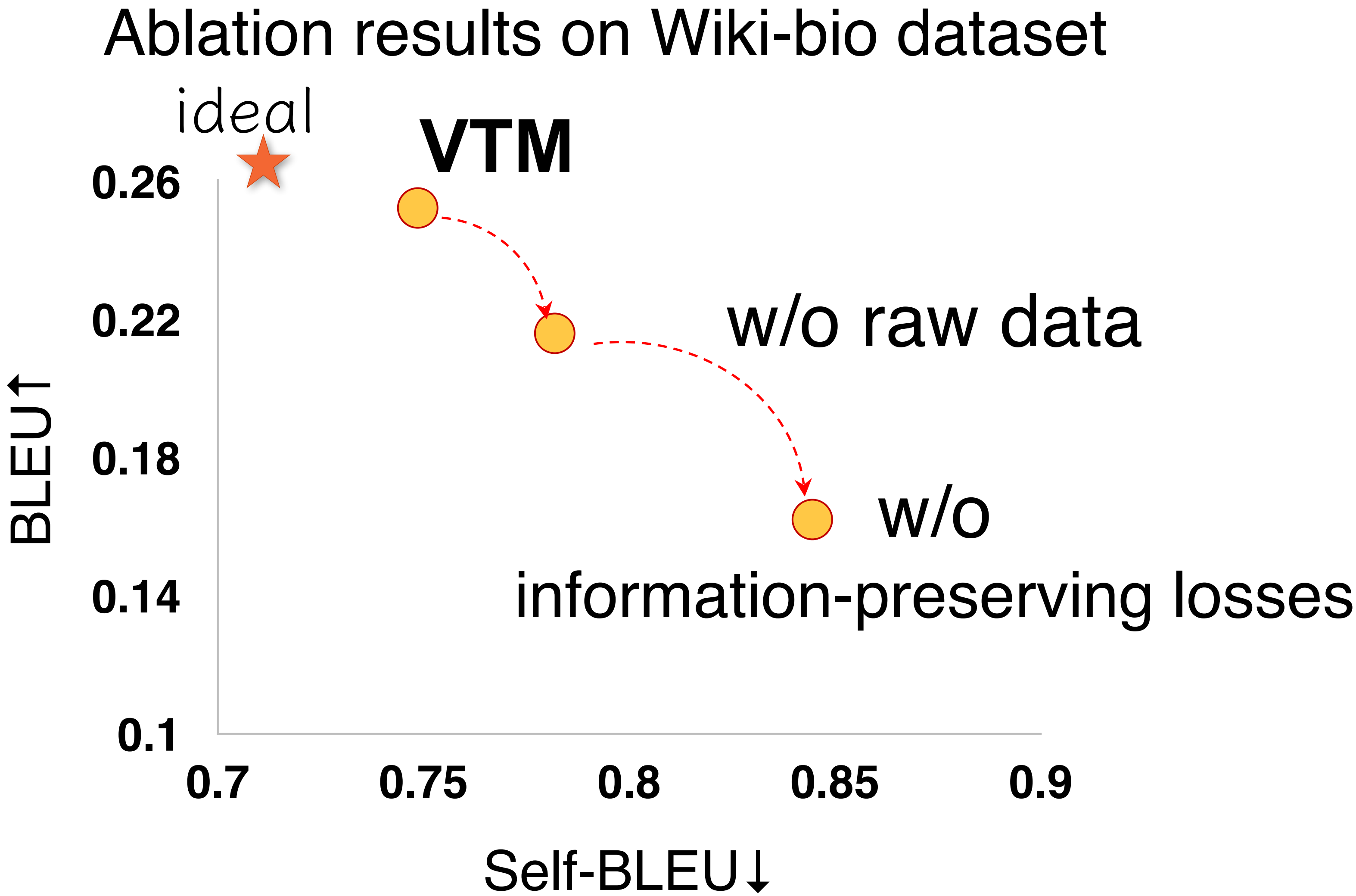




VTM uses beam-search decoding.

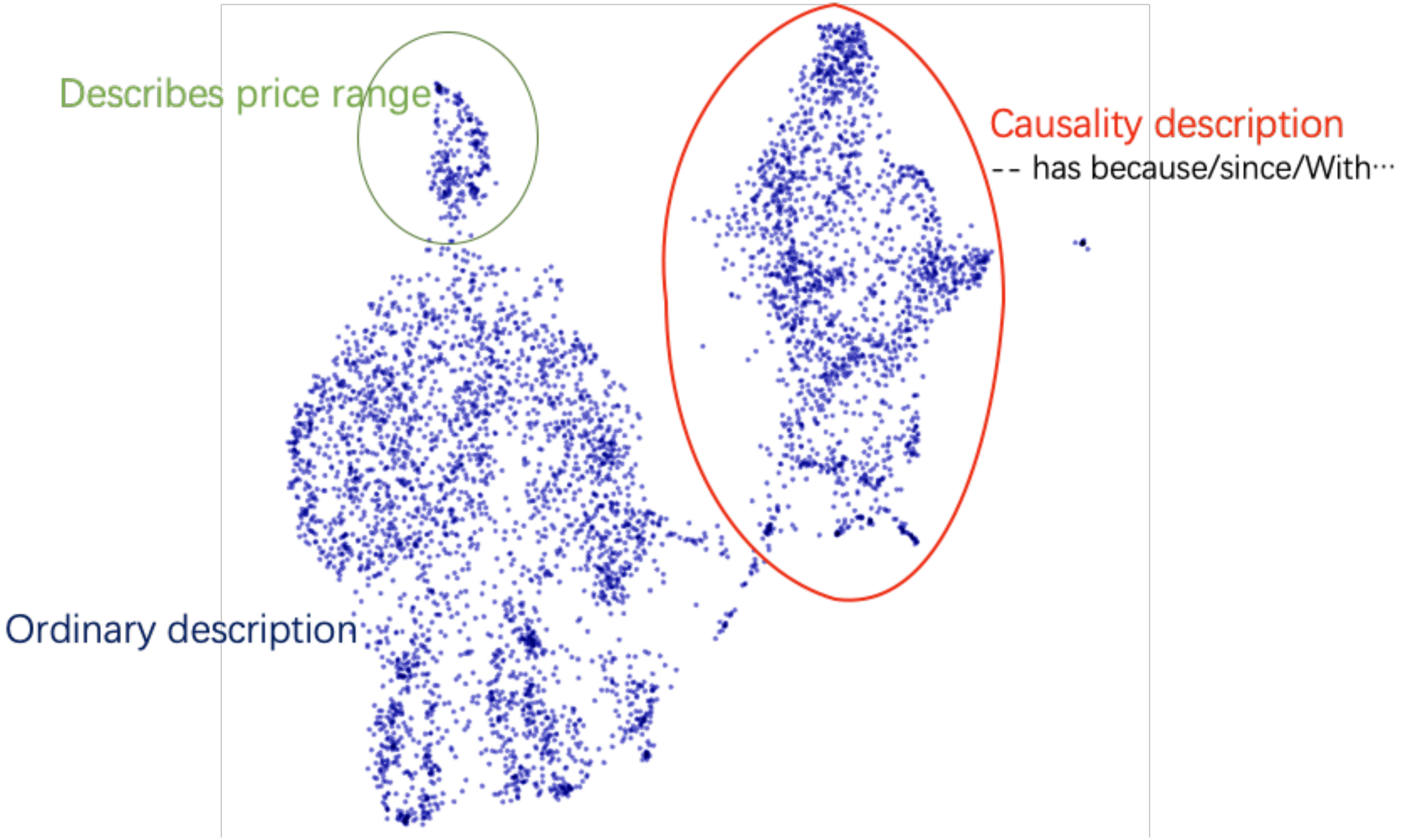
VTM [Ye, …, **Lei Li**, ICLR20b]

Raw data and loss terms are necessary



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

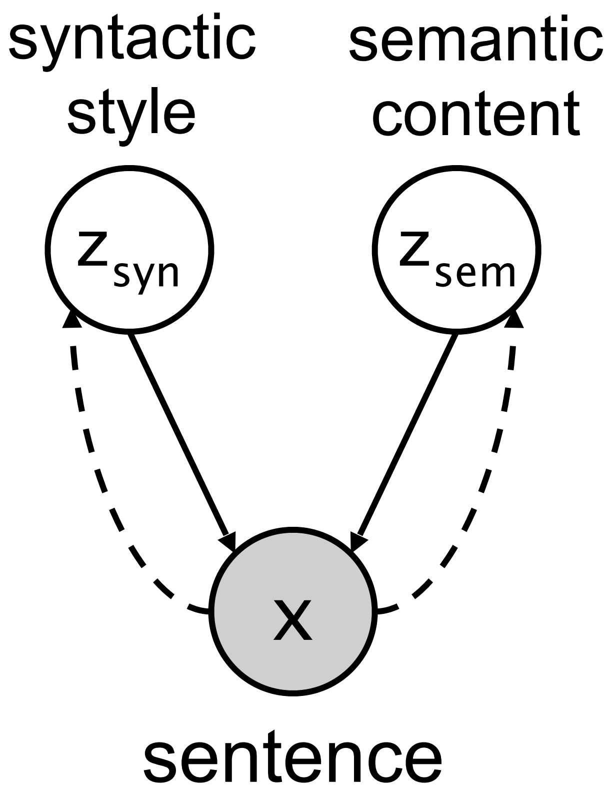
Generated Text

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: There is an apple on the table

Semantic content: The dog is behind the door

DSSVAE

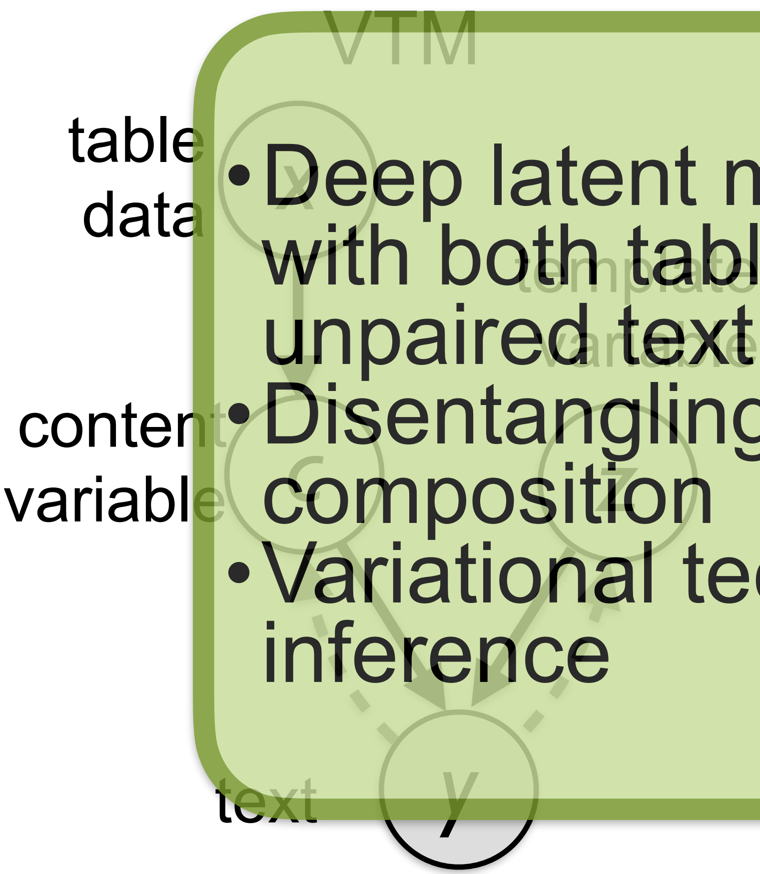
There is a dog behind the door

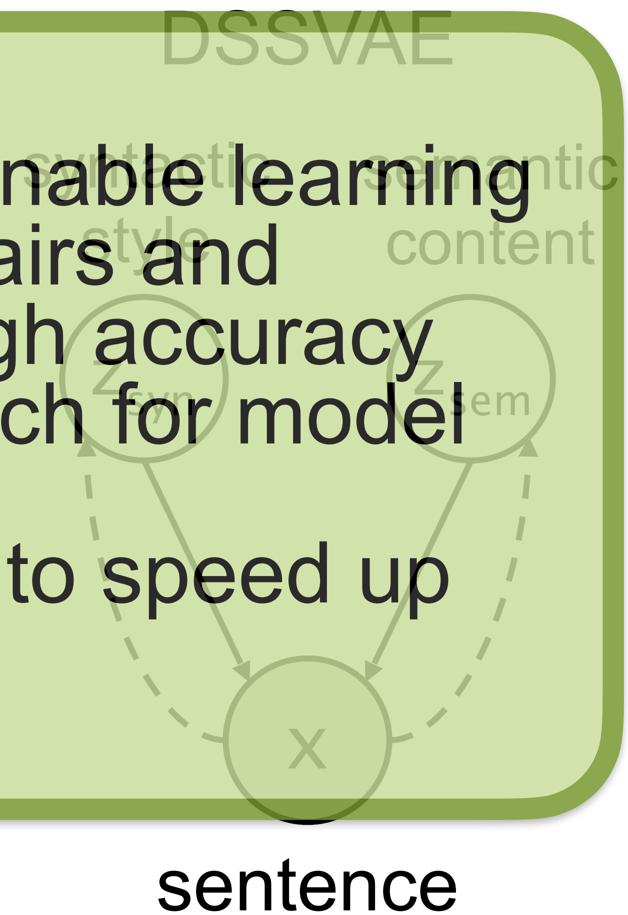
DSS-VAE [Y. Bao, H. Zhou, S. Huang, **Lei Li**, L. Mou, O. Vechtomova, X. Dai, J. Chen, ACL19c]

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.

Part I Takeaway





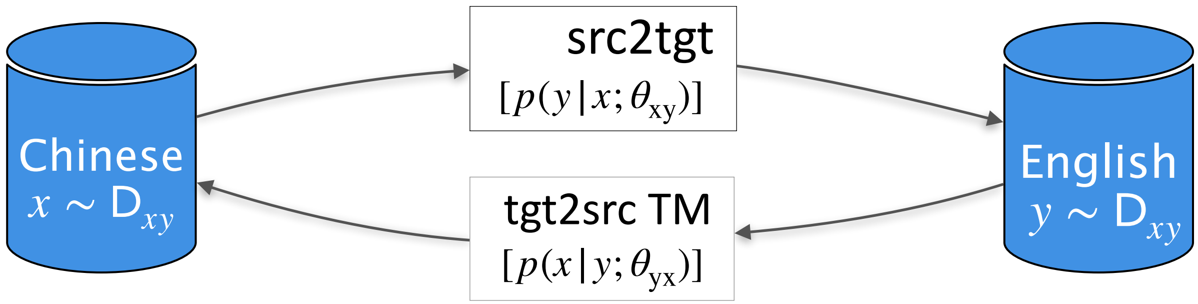
- Deep latent models enable learning with both table-text pairs and unpaired text, with high accuracy
- Disentangling approach for model composition
- Variational technique to speed up inference

Outline

1. Overview of Intelligent Information Assistant
2. Learning disentangled latent representation for text
3. Mirror-Generative NMT
4. Multimodal machine writing
5. Summary and Future Directions

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

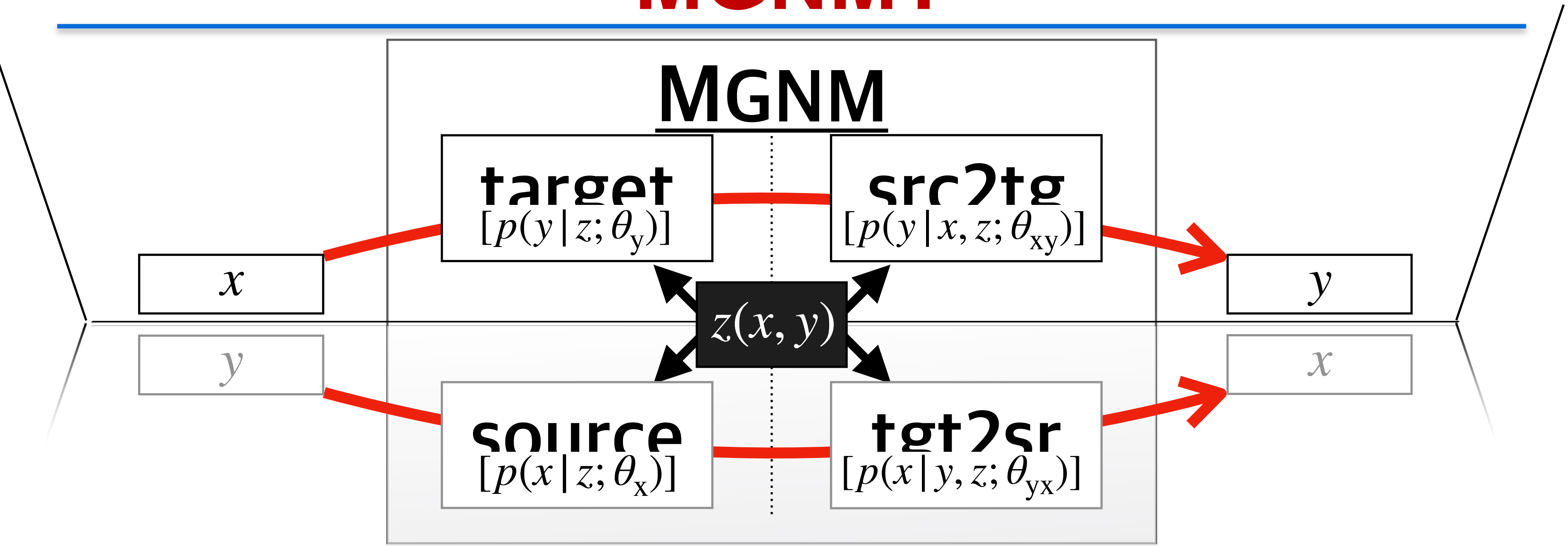
src2tgt $[p(y|x,z;\theta_{xy})]$ ← $z(x,y)$ → tgt2src $[p(x|y,z;\theta_{yx})]$

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

src2tgt $[p(y|x,z;\theta_{xy})]$ ← $z(x,y)$ → target LM $[p(y|z;\theta_{y})]$

We need a bridge

Integrating Four Language Skills with MGNMT



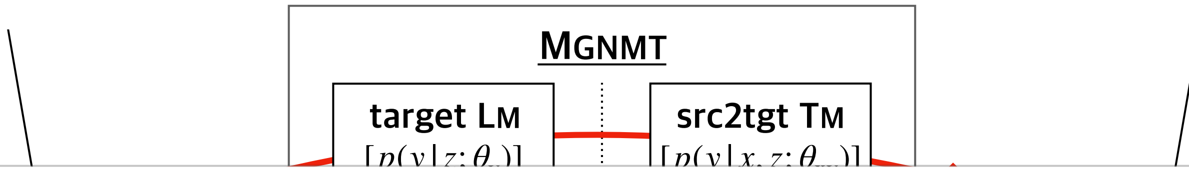
1. composing sentence in Source lang
2. composing sentence in Target lang
3. translating from source to target
4. translating from target to source

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



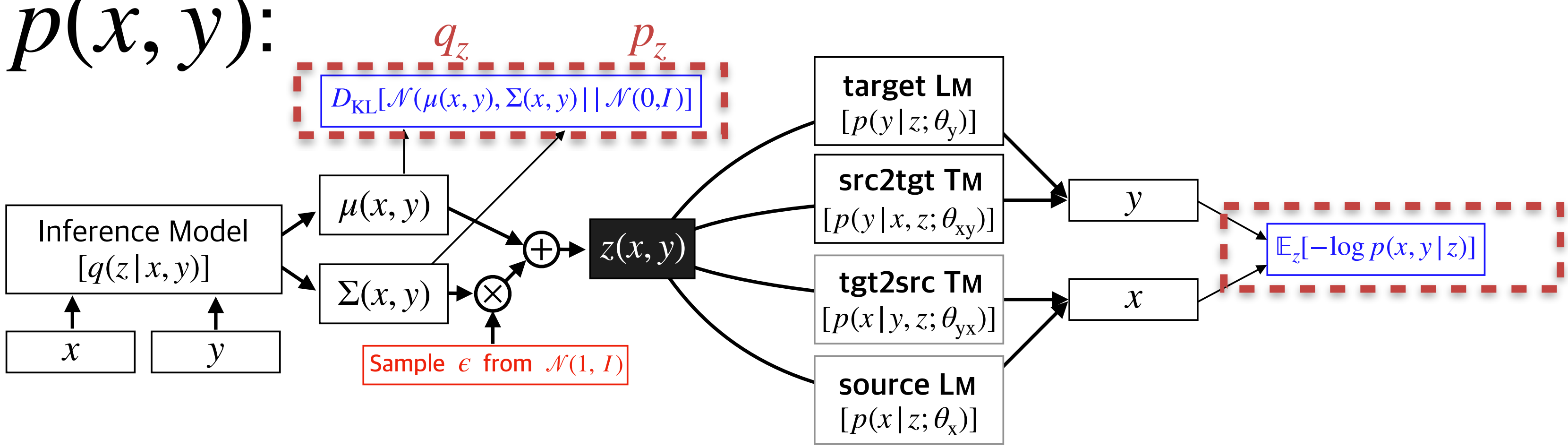
$$\log p(x,y|z) = \log p(x|z) + \log p(y|x,z) = \log p(y|z) + \log p(x|y,z)$$
$$= \frac{1}{2}[\underbrace{\log p(y|x,z)}_{src2tgt\ \mathrm{TM}_{x\to y}} + \underbrace{\log p(y|z)}_{target\ \mathrm{LM}_{y}} + \underbrace{\log p(x|y,z)}_{tgt2src\ \mathrm{TM}_{y\to x}} + \underbrace{\log p(x|z)}_{source\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

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

- **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. $p(x, y)$:



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

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

MT Technology Innovation

- Solving data scarcity
  - BERT for NMT [Yang et al, AAAI 2020]
  - Mirror Generative NMT [Zheng et al ICLR 2020a]
- Enhancing discourse coherence
  - Document-to-document translation [Sun et al, 2020, in submission]
- Speedup and Scaling NMT
  - Capsule NMT [Wang et al, EMNLP 2019]
  - Non-autoregressive NMT [Wang et al, ACL 2019]
  - Human-machine co-operative translation, CAMIT [Weng et al, IJCAI 2019]
- Cross-modal Translation
  - Visually guided MT [Wang et al, ICCV 2019]

Part II 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.
- ByteTrans system already serves > 100million active users

Outline

1. Overview of Intelligent Information Assistant
2. Learning disentangled latent representation for text
3. Mirror-Generative NMT
4. Multimodal machine writing
5. Summary and Future Directions

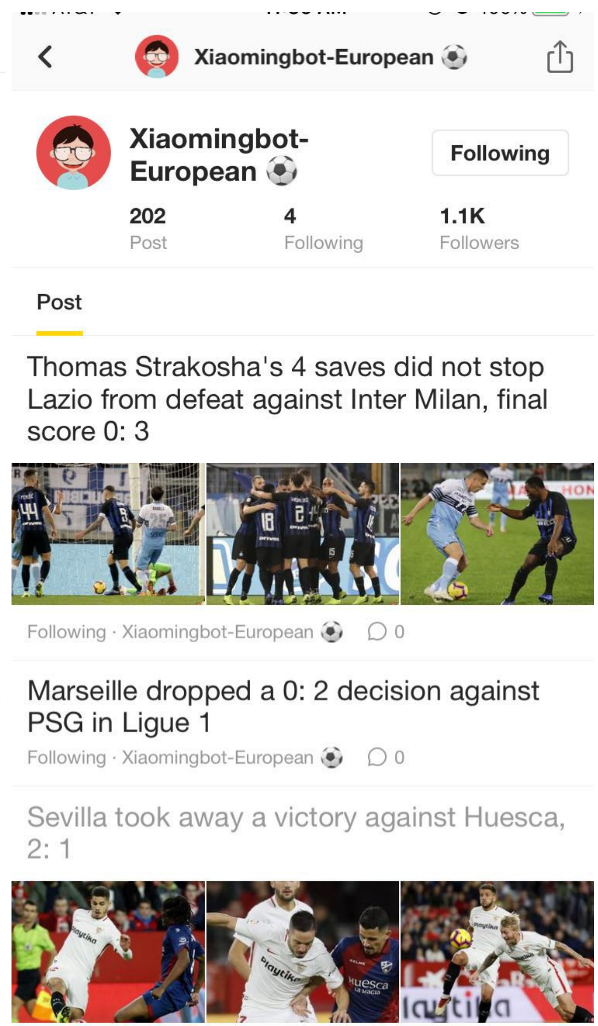
Multimodal Machine Writing

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]

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)

580,000 articles

6 lang

150,000 followers

Soccer News Generation from Multimodal Data

Lei Li, Han Zhang, Lifeng Hua, Jiaze Chen, Ying Zeng, Yuzhang Du, Yujie Li, Shikun Xu, Gen Li, Zhenqi Xu, Yandong Zhu, Siyi Gao, Changhu Wang, Weiying Ma

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]

AI Writing for Under-developed Region

Help farmers from rural countryside to sell agriculture products and promote culture through Toutiao and Douyin.
Certain product articles are semi-automatically generated by AI.

Promote Rural Products on Toutiao

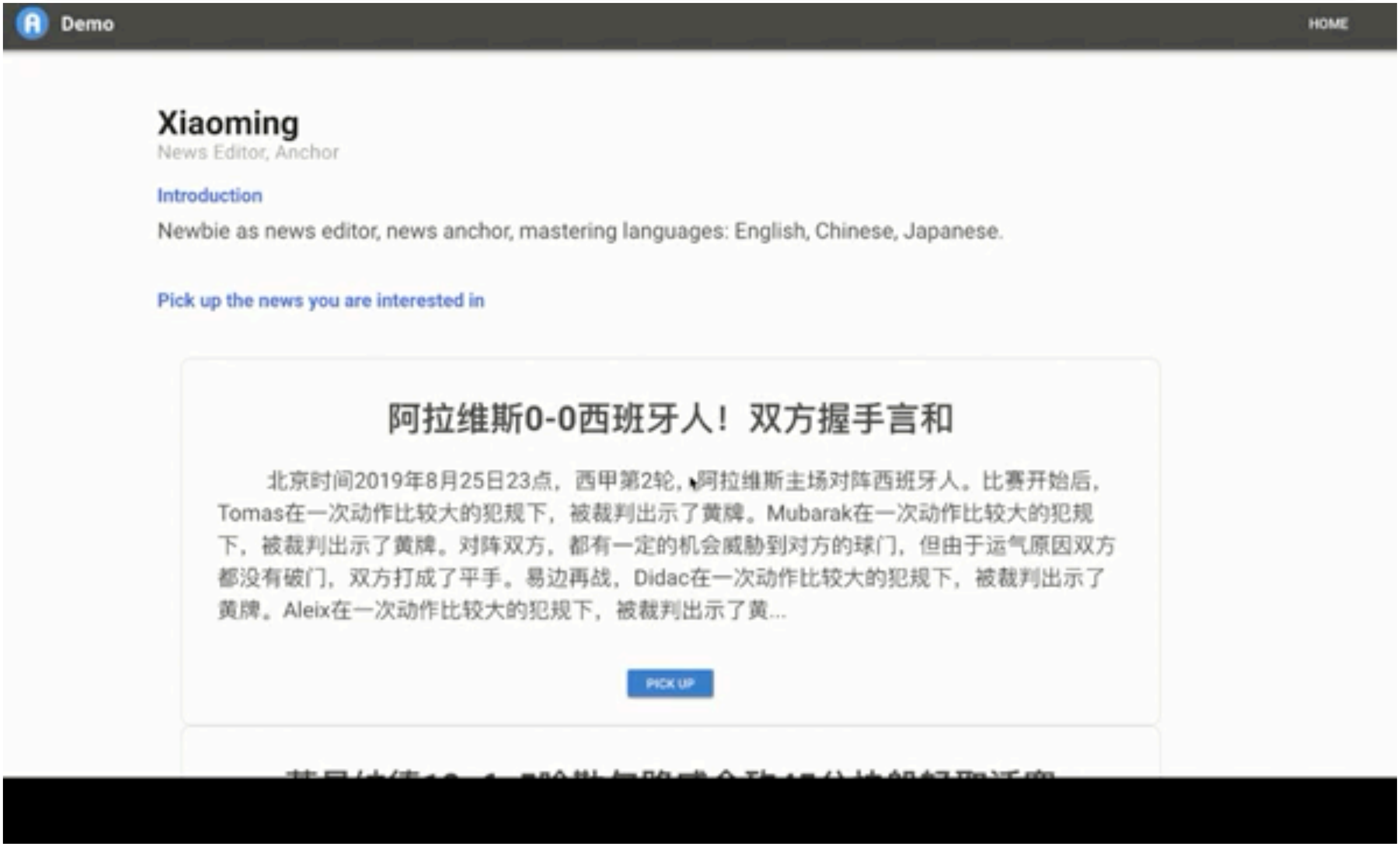
Gulin, Sichuan

Till 2018/7/15
Sold 27.5 tons of plum on Toutiao

Xiahe, Gansu

Boost beef selling by 4x after promotion on Toutiao

The Xiaoming Multilingual Reporter
News Writing + Summarization + Translation + TTS w/ Speech Cloning



Summary

- Goal: building intelligent information assistant
- Disentangled Latent Representation
  - VTM: Learning Latent Templates in Variational Space
  - DSS-VAE: Disentangled syntax and semantic representation
- MGNMT:
  - integrate four language capabilities together
  - Utilize both parallel and non-parallel corpus
- Multimodal Machine Writing
  - Xiaomingbot system: 600k articles and 150k followers
- Deployed in multiple online platforms and used by over 100 millions of users

Thanks

- We are hiring researchers, software engineers, and interns at Silicon Valley, Beijing, Shanghai.
- contact: lileilab@bytedance.com