Recent Advances in Machine Writing and Translation – Algorithms and Challenges

Lei Li
ByteDance AI Lab
Volctrans

12/19/2020
Revolution in Information Creation and Sharing

- New media platforms
- Tremendous improvement in the efficiency and quality of content creation
- Massive distribution of personalized information
Why is NLG important?

Machine Translation

ChatBOT

Machine Writing

Question Answering
Machine Translation has quietly increased international trade by over 10%
Machine Translation at ByteDance

- 50+ Clients
- 50+ languages
- Five champions in WMT 20 including Chinese-to-English, German-to-English, German-to-French

Public MT Corpus

BLEU scores for various language pairs:

Simultaneous Speech-to-Text Translation
Soon a Robot Will Be Writing This Headline

By Alana Semuels

Jan. 14, 2020
Xiaomingbot
Automatic News Writing System

Winning 2017 Wu Wen-tsün Award in AI from CAAI

600,000 articles
6 lang
150,000 followers
Xiaomingbot: Multilingual Robot News Reporter

--- Xiaomingbot ---
Snooker Commentary Generation
Combining Visual Understanding with Strategy Prediction

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: (184, 555)
- Red11: (197, 590)
- Red12: (241, 595)
- Red13: (155, 606)
- Red14: (327, 611)
- Brown: (183, 163)
- Green: (240, 163)
- Yellow: (127, 163)
- Blue: (183, 366)

1. Sequence Generation Problem
3. Monte-Carlo Methods for Constrained Text Generation
4. One model to acquire 4 language skills
   – Mirror Generative NMT [ICLR 20a]
5. mRASP: Multilingual Pretraining NMT
6. Summary
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:

\[ p_\theta(x) = p_\theta(x_1, \ldots, x_L) \]

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

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

\[p_\theta(x_i | x_{<i}) = \text{Softmax} \left( f_\theta(x_{<i}) \right)_{x_i}
\]

\[
\text{Softmax}(x)_j = \frac{\exp x_j}{\sum_k \exp x_k}
\]

\[p_\theta(x_5 | x_1, x_2, x_3, x_4)
\]

The quick brown fox
Conditional Sequence Generation

aka. sequence-to-sequence generation

- Machine Translation
- Dialog Generation
- Question Answering
- ...

\[ p_\theta(y|x) \]

敏捷的棕狐跳过懒狗

The quick brown fox jumps over the lazy dog.
Outline

1. Sequence Generation Problem
3. Monte-Carlo Methods for Constrained Text Generation
4. One model to acquire 4 language skills
   – Mirror Generative NMT [ICLR 20a]
5. mRASP: Multilingual Pretraining NMT
6. Summary
Deep Latent Variable Models for Text

- Interpretable Deep Latent Representation from Raw Text
  - Learning Exponential Family Mixture VAE [ICML 20]

- Disentangled Representation Learning for Text Generation
  - Data to Generation: VTM [ICLR 20b]
  - Learning syntax-semantic representation [ACL 19c]
Learning Interpretable Latent Representation

Generate Sentences with interpretable factors

Latent structure dialog actions

GENERATOR

Sampling

“Remind me about the football game.”
[action=remind]

“Will it be overcast tomorrow?”
[action=request]

……..
How to Interpret Latent Variables in VAEs?

Variational Auto-encoder (VAE)

Interpretable structure $\leftrightarrow Z \rightarrow X$

$Z$: continuous latent variables

Difficult to interpret discrete factors

Will it be humid in New York today?

Remind me about my meeting.

(Kingma & Welling, 2013)
Discrete Variables Could Enhance Interpretability - but one has to do it right!

Gaussian Mixture Variational Auto-encoder (GM-VAE)

\[ C \rightarrow Z \rightarrow X \]

Interpretable structure

**c**: discrete component

**z**: continuous latent variable

Why?

Mode-collapse

How to fix it?

Will it be overcast tomorrow?

Remind me about the football game.

(Dilokthanakul et al., 2016; Jiang et al., 2017)
Do it right for VAE w/ hierarchical priors -
Dispersed Exponential-family Mixture VAE

Exponential-family Mixture VAE

C → Z → X

adding dispersion term in training

Dispersed EM-VAE

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

\[ L_d = \mathbb{E}_{q_\phi(c|x)} A(\eta_c) - A(\mathbb{E}_{q_\phi(c|x)} \eta_c). \]

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 $c$. 

Generate Sensible Dialog Response with DEM-VAE

**Input Context**

Sys: “Taking you to Chevron.”

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

(action = thanks)

(action = request-address)

**Predict**

User: “Thank you car, let's go there!”

User: “What is the address?”

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

<table>
<thead>
<tr>
<th>name</th>
<th>Sukiyaki</th>
</tr>
</thead>
<tbody>
<tr>
<td>eatType</td>
<td>pub</td>
</tr>
<tr>
<td>food</td>
<td>Japanese</td>
</tr>
<tr>
<td>price</td>
<td>average</td>
</tr>
<tr>
<td>rating</td>
<td>good</td>
</tr>
<tr>
<td>area</td>
<td>Seattle</td>
</tr>
</tbody>
</table>
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].

<table>
<thead>
<tr>
<th>name</th>
<th>Sukiyaki</th>
</tr>
</thead>
<tbody>
<tr>
<td>eatType</td>
<td>pub</td>
</tr>
<tr>
<td>food</td>
<td>Japanese</td>
</tr>
<tr>
<td>price</td>
<td>average</td>
</tr>
<tr>
<td>rating</td>
<td>good</td>
</tr>
<tr>
<td>area</td>
<td>seattle</td>
</tr>
</tbody>
</table>

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.
Generating from Latent Factors

**Motivation 1:** Continuous and disentangled representation for template and content

<table>
<thead>
<tr>
<th>Template</th>
<th>Content</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

\[
q \text{ (template, content | sentence)}
\]

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

Input: triples of \(<field\_name, position, value>\)
\[
\{x^f_k, x^p_k, x^v_k\}_{k=1}^K
\]

1. \(p(c | x) \sim \text{Neural Net} \)
\[
\text{maxpool}(\tanh(W \cdot [x^f_k, x^p_k, x^v_k] + b))
\]

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

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

VTM [R. Ye, W. Shi, H. Zhou, Z. Wei, Lei Li, ICLR20b]
### Learning with Raw Corpus

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

<table>
<thead>
<tr>
<th>Table</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>Sukiyaki is a Japanese restaurant. It is a pub and it has a average cost and good rating. It is in seattle.</td>
</tr>
<tr>
<td>eatType</td>
<td>pub</td>
</tr>
<tr>
<td>food</td>
<td>Japanese</td>
</tr>
<tr>
<td>price</td>
<td>average</td>
</tr>
<tr>
<td>rating</td>
<td>good</td>
</tr>
<tr>
<td>area</td>
<td>seattle</td>
</tr>
</tbody>
</table>

Known for its creative flavours, Holycrab's signatures are the Hokkien crab.
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

ideal

VTM

w/o raw data

w/o information-preserving losses
Interpreting VTM

Template variable project to 2D

- Describes price range
- Causality description -- has because/since/With...
- Ordinary description
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

There is a dog behind the door

Outline

1. Sequence Generation Problem
3. Monte-Carlo Methods for Constrained Text Generation
4. One model to acquire 4 language skills
   – Mirror Generative NMT [ICLR 20a]
5. mRASP: Multilingual Pretraining NMT
6. Summary
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.
Constrained Sentence Generation via Metropolis-Hastings Sampling

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

\[
\pi(x) = \prod_{i} P(x_i | x_{0:i-1}) \cdot \prod_{j} P^j_C(x)
\]

- All token seq’s
- Fluent Text
- Ideal Text
- Constrained Text

CGMH [N. Miao, H. Zhou, L. Mou, R. Yan, Lei Li, AAAI19]
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 generates better sentences from keywords

Scores of human evaluation (↑)

NLL(↓)
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
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

MHA [H. Zhang, N. Miao, H. Zhou, Lei Li, ACL19a]
Generation under Combinatorial Constraints

- Logical and Combinatorial constraints
- E.g. generating a question for the following statement.
  - Paris is located in France.
  - ==> Is Paris located in France?
  - ==> Which country is Paris located in?
Generation under Combinatorial Constraints

- Logical and Combinatorial constraints

\[ \pi(x) = P_{LM}(x; \theta) \cdot \phi(x) \]

\[ \phi(x) = \beta^M \sum_i c_i(x), \quad 0 < \beta < 1 \]

\( c_i(x) \) 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]
Mirror Generative Model for Neural Machine Translation

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
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

Approach: Mirror-Generative NMT

- The **mirror** property to decompose

\[
\log p(x, y | z) = \log p(x | z) + \log p(y | x, z) = \log p(y | z) + \log p(x | y, z)
\]

\[
= \frac{1}{2} [\log p(y | x, z) + \log p(y | z) + \log p(x | y, z) + \log p(x | z)]
\]

- **Relevant** TMs & LMs under a **unified probabilistic framework**!
  - Enables the **aforementioned advantages**
MGNMT makes better use of non-parallel data

- Low resource results

<table>
<thead>
<tr>
<th>Model</th>
<th>Low-Resource WMT16 En↔Ro</th>
<th>Cross-Domain IN-DOMAIN (TED)</th>
<th>OUT-DOMAIN (NEWS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-RO</td>
<td>EN-DE</td>
<td>DE-EN</td>
</tr>
<tr>
<td>Transformer (Vaswani et al., 2017)</td>
<td>32.1</td>
<td>27.5</td>
<td>17.1</td>
</tr>
<tr>
<td>GNMT (Shah &amp; Barber, 2018)</td>
<td>32.4</td>
<td>28.0</td>
<td>19.9</td>
</tr>
<tr>
<td>GNMT-M-SSL + non-parallel (Shah &amp; Barber, 2018)</td>
<td>34.1</td>
<td>28.4</td>
<td>22.0</td>
</tr>
<tr>
<td>Transformer+BT + non-parallel (Sennrich et al, 2016b)</td>
<td>33.9</td>
<td>27.8</td>
<td>20.9</td>
</tr>
<tr>
<td>Transformer+JBT + non-parallel (Zhang et al., 2018)</td>
<td>34.5</td>
<td>28.4</td>
<td>21.9</td>
</tr>
<tr>
<td>Transformer+Dual + non-parallel (He et al., 2016a)</td>
<td>34.6</td>
<td>28.5</td>
<td>21.8</td>
</tr>
<tr>
<td>MGNMT</td>
<td>32.7</td>
<td>28.2</td>
<td>17.6</td>
</tr>
<tr>
<td>MGNMT + non-parallel</td>
<td>34.9</td>
<td>28.5</td>
<td>22.8</td>
</tr>
</tbody>
</table>
MGNMT makes better use of non-parallel data

- High resource results

<table>
<thead>
<tr>
<th>Model</th>
<th>Wmt14</th>
<th>Nist</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN-DE</td>
<td>DE-EN</td>
</tr>
<tr>
<td>Transformer (Vaswani et al., 2017)</td>
<td>27.2</td>
<td>30.8</td>
</tr>
<tr>
<td>GNMT (Shah &amp; Barber, 2018)</td>
<td>27.5</td>
<td>31.1</td>
</tr>
<tr>
<td>GNMT-M-SSL + non-parallel (Shah &amp; Barber, 2018)</td>
<td>29.7</td>
<td>33.5</td>
</tr>
<tr>
<td>Transformer+BT + non-parallel (Sennrich et al., 2016b)</td>
<td>29.6</td>
<td>33.2</td>
</tr>
<tr>
<td>Transformer+JBT + non-parallel (Zhang et al., 2018)</td>
<td>30.0</td>
<td>33.6</td>
</tr>
<tr>
<td>Transformer+Dual + non-parallel (He et al., 2016b)</td>
<td>29.6</td>
<td>33.2</td>
</tr>
<tr>
<td>MGNMT</td>
<td>27.7</td>
<td>31.4</td>
</tr>
<tr>
<td>MGNMT + non-parallel</td>
<td>30.3</td>
<td>33.8</td>
</tr>
</tbody>
</table>

– Non-parallel data is helpful
– MGNMT works well especially on low resource settings.
Multilingual Pretraining NMT

mRASP [Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, Lei Li, EMNLP 2020]
The Ultimate Quest of Machine Translation

- # of human languages: >6900.
- How to build a universal MT system that is capable of translating any source language into a target one?
Why Training Multilingual MT Jointly?

- Data scarcity for low/zero resource languages.
- Transfer knowledge between languages.
Further Pursuit: Unified Multilingual Representation

- Further: It is expected to bridge distributional representation of different languages.
- Utterances in different languages with the same semantics will be mapped to adjacent embedding spaces.

<En> I love you.           <Fr> Je t’aime.
<De> Ich liebe dich.     <Es> Te quiero.
<It> ti amo.
Overview of mRASP

Pre-training

Encoder

Decoder

Random Aligned Substitution

<En> I love you.
<Fr> Je t’aime.
<De> Ich liebe dich.
<Es> Te quiero.
<It> ti amo.
(Extremely) Low Resource Directions

**Extremely-Low Resource Directions**

- **Direct**
  - En2Be: 8.5
  - Be2En: 9.6
  - En2My: 10.2
  - My2En: 5.4
  - En2Af: 8.3
  - Af2En: 7.2
  - En2Eo: 4.9
  - Eo2En: 6.7

- **mRASP**
  - En2Be: 25.8
  - Be2En: 32.3
  - En2My: 28.6
  - My2En: 25.3
  - En2Af: 31.1
  - Af2En: 27
  - En2Eo: 30.4
  - Eo2En: 35.8

**Low Resource Directions**

- **Direct**
  - En2He: 19
  - He2En: 27.6
  - En2Tr: 21
  - Tr2En: 19.4
  - En2Ro: 30.5
  - Ro2En: 29.2
  - En2Cs: 19
  - Cs2En: 22.7

- **mRASP**
  - En2He: 32.4
  - He2En: 44.6
  - En2Tr: 33.3
  - Tr2En: 39
  - En2Ro: 37.4
  - Ro2En: 23.2
  - En2Cs: 29.8
Medium & Rich Resource (Popular Benchmark)

- Rich resource benchmarks can be further improved (En->Fr +1.1BLEU).

En2De(wmt2016) vs. En2Fr(wmt2014)
Does mRASP boost MT performance for Exotic Languages?

- mRASP generalizes on all exotic scenarios.

<table>
<thead>
<tr>
<th>Exotic Pair</th>
<th>Fr-Zh(20K)</th>
<th>De-Fr(9M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>—&gt;</td>
<td>&lt;--</td>
</tr>
<tr>
<td>Direct</td>
<td>0.7</td>
<td>3</td>
</tr>
<tr>
<td>mRASP</td>
<td>25.8</td>
<td>26.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exotic Full</th>
<th>NI-Pt(12K)</th>
<th>Da-El(1.2M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>—&gt;</td>
<td>&lt;--</td>
</tr>
<tr>
<td>Direct</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mRASP</td>
<td>14.1</td>
<td>13.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exotic Source/Target</th>
<th>En-Mr(11K)</th>
<th>En-Gl(1.2M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>—&gt;</td>
<td>&lt;--</td>
</tr>
<tr>
<td>Direct</td>
<td>6.4</td>
<td>6.8</td>
</tr>
<tr>
<td>mRASP</td>
<td>22.7</td>
<td>22.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exotic Source/Target</th>
<th>En-Eu(726k)</th>
<th>En-Sl(2M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>—&gt;</td>
<td>&lt;--</td>
</tr>
<tr>
<td>Direct</td>
<td>7.1</td>
<td>10.9</td>
</tr>
<tr>
<td>mRASP</td>
<td>19.1</td>
<td>28.4</td>
</tr>
</tbody>
</table>

12k: **Direct** not work VS **mRASP** achieves 10+ BLEU!!
Summary

• Multimodal Machine Writing
  – Xiaomingbot system: 600k articles and 150k followers

• 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

• MGNMT:
  – integrate four language capabilities together
  – Utilize both parallel and non-parallel corpus

• mRASP: a new pre-trained model for many translation directions
For the Community

mRASP Multilingual MT Pretraining
https://github.com/linzehui/mRASP

Lightseq A high performance sequence processing lib
https://github.com/bytedance/lightseq

https://translate.volcengine.cn
Thanks

- ByteDance AI Lab MLNLC Group and many collaborators
- Contact: lileilab@bytedance.com


