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
Soon a Robot Will Be Writing This Headline

By Alana Semuels

Jan. 14, 2020
Xiaomingbot is deployed and constantly producing news on social media platforms (TopBuzz & Toutiao).

La Liga: Real Betis suffered from an utterly embarassing ending in their 1:4 fiasco against Barcelona.
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]
**Natural Language Descriptions**

<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 an average cost and good rating. It is based in Seattle.
Made of poplin, this long dress has an ink painting of bamboo and feels fresh and smooth.

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

<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 an 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>
\[ \{x^f_k, x^p_k, x^v_k\}_{k=1}^K \]

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

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

Total loss = 

Reconstruction loss + Information-Preserving loss
Instead of optimizing exact and intractable expected likelihood, minimizing the (tractable) variational lower bounds.

\[ l_p = -\mathbb{E} \log \int p(y \mid c(x), z) p(z) \, dz \]

\[ \text{ELBO}_p = -\mathbb{E}_{q(z \mid y)} \log p(y \mid c(x), z) + \text{KL}[q(z \mid y) \mid \mid p(z)] \]

\[ l_r = -\mathbb{E} \log \int \int p(y \mid c, z) p(z) p(c) \, dz \, dc \]

\[ \text{ELBO}_r = -\mathbb{E}_{q(z \mid y) q(c \mid y)} \log p(y \mid c, z) + \text{KL}[q(z \mid y) \mid \mid p(z)] + \text{KL}[q(c \mid y) \mid \mid q(c)] \]
1. Content preserving loss
\[ l_{cp} = \mathbb{E}_{q(c|y)} |c - f(x)|^2 + D_{KL}(q(c|y) \mid \mid p(c)) \]

2. Template preserving loss of pairs
\[ l_{tp} = - \mathbb{E}_{q(z|y)} \left[ \log p(\tilde{y} | z, x) \right] \]

\( \tilde{y} \) 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]\}\n
Text $y$:
Loch Fyne is a French restaurant catering to a budget of below $20$.

Text Sketch $\tilde{y}$:
$<\text{ent}>$ is a $<\text{ent}>$ $<\text{ent}>$ catering to a budget of $<\text{ent}>$. 
Semi-supervised learning: “Back-translate” corpus to obtain pseudo-parallel pairs $<\text{table, text}>$, to enrich the learning.

<table>
<thead>
<tr>
<th>Table</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>Sukiyaki</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>

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

Known for its creative flavours, Holycrab's signatures are the Hokkien crab.
Tasks
– WIKI: generating short-bio from person profile.
– SPNLF: generating restaurant description from attributes

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

- **Ideal**
- **VTM**
- **w/o raw data**
- **w/o information-preserving losses**

```
<table>
<thead>
<tr>
<th>BLEU↑</th>
<th>Self-BLEU↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.26</td>
<td>0.7</td>
</tr>
<tr>
<td>0.22</td>
<td>0.75</td>
</tr>
<tr>
<td>0.18</td>
<td>0.8</td>
</tr>
<tr>
<td>0.14</td>
<td>0.85</td>
</tr>
<tr>
<td>0.10</td>
<td>0.9</td>
</tr>
</tbody>
</table>
```
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

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 100 million active users.
- 10% of query suggestion phrases from the generation algorithm.
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

  \[ p(y|x; \theta_{xy}) \]

  \[ p(x|y; \theta_{yx}) \]

- **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

MGNM

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**

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

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

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

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

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

\[\text{mirror}\]
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

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

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

$$\mathcal{L}(x^{(s)}; \theta_x, \theta_{yx}, \phi) = \mathbb{E}_{p(y|x^{(s)})} \left[ \mathbb{E}_{q(z|x^{(s)},y;\phi)} \left[ \frac{1}{2} \{ \log p(x^{(s)}|z; \theta_x) + \log p(x^{(s)}|y,z; \theta_{yx}) \} \right] - D_{KL}[q(z|x^{(s)},y;\phi)||p(z)] \right]$$
Decoding: TM&LM work as a whole

- Iterative EM decoding
  - Given source sentence $x$, find a translation
    \[
y = \arg\max_y p(y|x) = \arg\max_y p(x, y) \approx \arg\max_y \mathcal{L}(x, y; \theta, \phi)
    \]
  - Initialization: get a draft translation
  - Iterative refinement: resampling $z$ from inference model and redecoding by maximizing ELBO
    \[
    \tilde{y} \leftarrow \arg\max_y \mathcal{L}(x, \tilde{y}; \theta, \phi)
    \]
    \[
    = \arg\max_y \mathbb{E}_{q(z|x, \tilde{y}; \phi)} \left[ \log p(y|x, z) + \log p(y|z) + \log p(x|z) + \log p(x|y, z) \right]
    \]
    \[
    = \arg\max_y \mathbb{E}_{q(z|x, \tilde{y}; \phi)} \left[ \sum_i \left[ \log p(y_i|y_{<i}, x, z) + \log p(y_i|y_{<i}, z) \right] + \log p(x|z) + \log p(x|y, z) \right]
    \]
Experiments

• Datasets
  – Low resource
    ‣ WMT\textsuperscript{16} EN-RO
    ‣ IWSLT\textsuperscript{16} EN-DE: domain adaptation (from TED to News)
  – High resource:
    ‣ WMT\textsuperscript{14} EN-DE, NIST EN-ZH

• Avoiding posterior collapse (Important!)
  – KL-annealing
  – Word dropout
MGNMT makes better use of non-parallel data

- Low resource results

<table>
<thead>
<tr>
<th>Model</th>
<th>Low-Resource</th>
<th>Cross-Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WMT16 En↔Ro</td>
<td>In-Domain (TED)</td>
</tr>
<tr>
<td></td>
<td>En-Ro</td>
<td>Ro-En</td>
</tr>
<tr>
<td>Transformer (Vaswani et al., 2017)</td>
<td>32.1</td>
<td>33.2</td>
</tr>
<tr>
<td>GNMT (Shah &amp; Barber, 2018)</td>
<td>32.4</td>
<td>33.6</td>
</tr>
<tr>
<td>GNMT-M-SSL + non-parallel (Shah &amp; Barber, 2018)</td>
<td>34.1</td>
<td>35.3</td>
</tr>
<tr>
<td>Transformer+BT + non-parallel (Sennrich et al., 2016b)</td>
<td>33.9</td>
<td>35.0</td>
</tr>
<tr>
<td>Transformer+JBT + non-parallel (Zhang et al., 2018)</td>
<td>34.5</td>
<td>35.7</td>
</tr>
<tr>
<td>Transformer+Dual + non-parallel (He et al., 2016a)</td>
<td>34.6</td>
<td>35.7</td>
</tr>
<tr>
<td>MGNMT</td>
<td>32.7</td>
<td>33.9</td>
</tr>
<tr>
<td>MGNMT + non-parallel</td>
<td>34.9</td>
<td>36.1</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
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.
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


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

580,000 articles
6 lang
150,000 followers
Soccer News Generation from Multimodal Data

Machine Writing
Xiaomingbot

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’ 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)

Balls Detection

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

Xiaoming
News Editor, Anchor

Introduction
Newbie as news editor, news anchor, mastering languages: English, Chinese, Japanese.

Pick up the news you are interested in

阿拉维斯0-0西班牙人！双方握手言和

北京时间2019年8月25日23点，西甲第2轮，阿拉维斯主场对阵西班牙人。比赛开始后，Tomas在一次动作比较大的犯规下，被裁判出示了黄牌。Mubarak在一次动作比较大的犯规下，被裁判出示了黄牌。对阵双方，都有一定的机会威胁到对方的球门，但由于运气原因双方都没有破门，双方打成了平手。易边再战，Didac在一次动作比较大的犯规下，被裁判出示了黄牌。Aleix在一次动作比较大的犯规下，被裁判出示了黄...
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