Text Style Transfer

Hung-yi Lee 李宏毅
**Image Style**

**Audio Style**

**Text Style Transfer**
Text Style Transfer

你真笨 (negative) -> Seq2seq -> 你真聰明 (positive)
Text Style Transfer

Cycle GAN

You are not very smart
(negative)

You are very smart
(positive)

minimize the reconstruction error
Can we use gradient ascent? **NO!**

**Seq2seq**
- **Generator**
  - Obtained by attention

**Update Parameters**

**Discriminator**
- scalar

---

**Can we use gradient ascent?**

**NO!**
Can we use gradient ascent? **NO!**

Having non-differentiable part: obtained by attention
Three Categories of Solutions

Gumbel-softmax
- [Matt J. Kusner, et al., arXiv, 2016]
- [Weili Nie, et al. ICLR, 2019]

Continuous Input for Discriminator
- [Sai Rajeswar, et al., arXiv, 2017]
- [Ofir Press, et al., ICML workshop, 2017]
- [Zhen Xu, et al., EMNLP, 2017]
- [Alex Lamb, et al., NIPS, 2016]
- [Yizhe Zhang, et al., ICML, 2017]

Reinforcement Learning
- [Yu, et al., AAAI, 2017]
- [Li, et al., EMNLP, 2017]
- [Tong Che, et al, arXiv, 2017]
- [Jiaxian Guo, et al., AAAI, 2018]
- [Kevin Lin, et al, NIPS, 2017]
- [William Fedus, et al., ICLR, 2018]
Gumbel-softmax

Using the reparameterization trick

As what people do for training VAE
Three Categories of Solutions

**Gumbel-softmax**

**Continuous Input for Discriminator**

**Reinforcement Learning**
Use the distribution as the input of discriminator

Avoid the sampling process

We can do backpropagation now.
What is the problem?

- Real sentence

- Generated

Discriminator with constraint (e.g. WGAN) can be helpful.

Discriminator can immediately find the difference.

Can never be 1-hot
Three Categories of Solutions

Gumbel-softmax

Continuous Input for Discriminator

Reinforcement Learning
The reward function may change

→ Different from typical RL

Environment

Actions taken

Generator

= Agent in RL

Trained by RL algorithm (e.g. Policy Gradient)
Disaster ......

RL is difficult to train  GAN is difficult to train

RL+GAN
Tips?

- ScratchGAN

[Cyprien de Masson d'Autume, et al., arXiv 2019]
Tips?

- Typical
  
  En | De
  Gen

  You is good → Discriminator → 0.1

  I don’t know which part is wrong ...

- Reward for Every Generation Step

  En | De
  Gen

  You → You is → You is good → Discriminator

  0.9 → 0.1 → 0.1
Tips?

• Reward for Every Generation Step


Method 2. Discriminator For Partially Decoded Sequences  [Li, et al., EMNLP, 2017]

Text Style Transfer

[Lee, et al., ICASSP’18]

- From negative sentence to positive one

胃疼, 沒睡醒, 各種不舒服

我都想去上班了, 真夠賤的!

暈死了, 吃燒烤、竟然遇到個變態狂

我肚子痛的厲害
### Relaxed ↔ Annoyed

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relaxed</td>
<td>Sitting by the Christmas tree and watching Star Wars after cooking dinner. What a nice night 🎄🌲🎄</td>
</tr>
<tr>
<td>Annoyed</td>
<td>Sitting by the computer and watching The Voice for the second time tonight. What a horrible way to start the weekend 😞😞😞</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annoyed</td>
<td>Getting a speeding ticket 50 feet in front of work is not how I wanted to start this month 😞</td>
</tr>
<tr>
<td>Relaxed</td>
<td>Getting a haircut followed by a cold foot massage in the morning is how I wanted to start this month 😊</td>
</tr>
</tbody>
</table>

### Male ↔ Female

<table>
<thead>
<tr>
<th>Gender</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Gotta say that beard makes you look like a Viking...</td>
</tr>
<tr>
<td>Female</td>
<td>Gotta say that hair makes you look like a Mermaid...</td>
</tr>
<tr>
<td>Female</td>
<td>Awww he’s so gorgeous 😍 can’t wait for a cuddle. Well done 😞 xxx</td>
</tr>
<tr>
<td>Male</td>
<td>Bro he’s so f***ing dope can’t wait for a cuddle. Well done bro</td>
</tr>
</tbody>
</table>

### Age 18-24 ↔ 65+

<table>
<thead>
<tr>
<th>Age</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-24</td>
<td>You cheated on me but now I know nothing about loyalty 😅 ok</td>
</tr>
<tr>
<td>65+</td>
<td>You cheated on America but now I know nothing about patriotism. So ok.</td>
</tr>
<tr>
<td>65+</td>
<td>Ah! Sweet photo of the sisters. So happy to see them together today 😊</td>
</tr>
<tr>
<td>18-24</td>
<td>Ah 😊 Thankyou 💖 #sisters 💖 happy to see them together today</td>
</tr>
</tbody>
</table>

1 Note that using “gender” (or any other attribute for that matter) as a differentiating attribute between several bodies of text implies that there are indeed signatures of gender in the data. These signatures could be as innocuous as some first names like Mary being usually associated with women, or disheartening like biases and stereotypes exposed by statistical methods, (e.g., “man is to computer programmer as woman is to homemaker” (Bolukbasi et al., 2016)). We certainly do not condone those stereotypes, and on the contrary, we hope that showing that our models can uncover these biases might down the line turn them into powerful tools for researchers who study fairness and debiasing (Reddy & Knight, 2016).

Source of image: https://openreview.net/forum?id=H1g2NhC5KQ

[Lample, et al., ICLR’19]
StarGAN

[Kaneko, et al., INTERSPEECH’19]

For CycleGAN:
If there are N speakers, you need N x (N-1) generators.
Style Transformer (Text version of StarGAN)

Voice Conversion

as close as possible

Content Encoder → Decoder → reconstructed

input audio

Speaker Encoder

Text Style Transfer

as close as possible

The food is delicious.

Content Encoder → Style Encoder → Decoder → The food is delicious.

The food is delicious.
Text Style Transfer

The food is delicious.

The food is ??? style.

The food is delicious.

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]

[Li, et al., NAACL’18]
Image Style

Audio Style

male

female

positive
sentences

negative
sentences

Text Style Transfer

Unsupervised Abstractive Summarization
Unsupervised Abstractive Summarization

document \rightarrow \text{Seq2seq} \rightarrow \text{summary?}
Unsupervised Abstractive Summarization

Seq2seq

human-written summary

summary or not?

Discriminator

document

Seq2seq

summary?

Seq2seq

minimize the reconstruction error

document
Summarization

English Gigaword (Document title as summary)

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>33.2</td>
<td>14.2</td>
<td>30.5</td>
</tr>
<tr>
<td>Trivial</td>
<td>21.9</td>
<td>7.7</td>
<td>20.5</td>
</tr>
<tr>
<td>Unsupervised (matched data)</td>
<td>28.1</td>
<td>10.0</td>
<td>25.4</td>
</tr>
<tr>
<td>Unsupervised (no matched data)</td>
<td>27.2</td>
<td>9.1</td>
<td>24.1</td>
</tr>
</tbody>
</table>

- Matched data: using the title of English Gigaword to train Discriminator
- No matched data: using the title of CNN/Diary Mail to train Discriminator

[Wang, Lee, EMNLP 2018]
More Unsupervised Summarization

• Unsupervised summarization with language prior

[Baziotis, et al., NAACL 2019]

• Unsupervised multi-document summarization

[Chu, et al., ICML 2019]
Mapping of Word Embedding
Mapping of Word Embedding

- rabbit
- fish
- jump
- swim
- 跳
- 兔
- 游
- 魚

<table>
<thead>
<tr>
<th>Authors</th>
<th>Unsupervised step</th>
<th>Supervised step</th>
<th>Extras</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barone (2016)</td>
<td>GAN</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Zhang et al. (2017)</td>
<td>Wasserstein GAN</td>
<td>Procrustes</td>
<td></td>
</tr>
<tr>
<td>Conneau et al. (2018)</td>
<td>GAN</td>
<td>Procrustes</td>
<td></td>
</tr>
<tr>
<td>Hoshen and Wolf (2018)</td>
<td>ICP</td>
<td>Procrustes</td>
<td></td>
</tr>
<tr>
<td>Artetxe et al. (2018)</td>
<td>Gromov-Wasserstein</td>
<td>Stochastic</td>
<td></td>
</tr>
<tr>
<td>Yang et al. (2018)</td>
<td>Gromov-Wasserstein</td>
<td>MMD</td>
<td></td>
</tr>
<tr>
<td>Xu et al. (2018)</td>
<td>GAN</td>
<td>Sinkhorn</td>
<td>Back-translation</td>
</tr>
<tr>
<td>Grave et al. (2018)</td>
<td>Gold-Rangarajan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Hartmann, et al., NeurIPS’19]
## VecMap

[Artetxe, et al., ACL’18]

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mikolov et al. (2013)</td>
<td>34.93†</td>
<td>35.00†</td>
<td>25.91†</td>
<td>27.73†</td>
</tr>
<tr>
<td></td>
<td>Faruqui and Dyer (2014)</td>
<td>38.40*</td>
<td>37.13*</td>
<td>27.60*</td>
<td>26.80*</td>
</tr>
<tr>
<td></td>
<td>Shigeto et al. (2015)</td>
<td>41.53†</td>
<td>43.07†</td>
<td>31.04†</td>
<td>33.73†</td>
</tr>
<tr>
<td></td>
<td>Dinu et al. (2015)</td>
<td>37.7</td>
<td>38.93*</td>
<td>29.14*</td>
<td>30.40*</td>
</tr>
<tr>
<td></td>
<td>Lazaridou et al. (2015)</td>
<td>40.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5k dict.</td>
<td>Xing et al. (2015)</td>
<td>36.87†</td>
<td>41.27†</td>
<td>28.23†</td>
<td>31.20†</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. (2016)</td>
<td>36.73†</td>
<td>40.80†</td>
<td>28.16†</td>
<td>31.07†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2016)</td>
<td>39.27†</td>
<td>41.87*</td>
<td>30.62*</td>
<td>31.40*</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2017)</td>
<td>39.67†</td>
<td>40.87</td>
<td>28.72</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Smith et al. (2017)</td>
<td>43.1</td>
<td>43.33†</td>
<td>29.42†</td>
<td>35.13†</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2018a)</td>
<td>45.27†</td>
<td>44.13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>25 dict.</td>
<td>Artetxe et al. (2017)</td>
<td>37.27</td>
<td>39.60</td>
<td>28.16</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Init. heurist.</th>
<th>Method</th>
<th>EN-IT</th>
<th>EN-DE</th>
<th>EN-FI</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Smith et al. (2017), cognates</td>
<td>39.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Artetxe et al. (2017), num.</td>
<td>39.40</td>
<td>40.27</td>
<td>26.47</td>
<td>-</td>
</tr>
<tr>
<td>None</td>
<td>Zhang et al. (2017a), $\lambda = 1$</td>
<td>0.00*</td>
<td>0.00*</td>
<td>0.00*</td>
<td>0.00*</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. (2017a), $\lambda = 10$</td>
<td>0.00*</td>
<td>0.00*</td>
<td>0.01*</td>
<td>0.01*</td>
</tr>
<tr>
<td></td>
<td>Conneau et al. (2018), code†</td>
<td>45.15*</td>
<td>46.83*</td>
<td>0.38*</td>
<td>35.38*</td>
</tr>
<tr>
<td></td>
<td>Conneau et al. (2018), paper†</td>
<td>45.1</td>
<td>0.01*</td>
<td>0.01*</td>
<td>35.44*</td>
</tr>
<tr>
<td></td>
<td>Proposed method</td>
<td>48.13</td>
<td>48.19</td>
<td>32.63</td>
<td>37.33</td>
</tr>
</tbody>
</table>
Unsupervised Translation

[Corresponding text content extracted from the image]

[Lample, et al., ICLR, 2018]
Unsupervised Translation

Sentence A \( \rightarrow \) \( EN_A \) \( \rightarrow \) \( DE_A \) \( \rightarrow \) Sentence A

再見 \( \rightarrow \) \( EN_B \) \( \rightarrow \) \( DE_B \) \( \rightarrow \) 再見

How are you?

再見

Sentence B

[1] Lample, et al., ICLR, 2018
Unsupervised Translation

Sentence A

Another Model

Sentence B

好嗎你

$EN_B$

$DE_A$

Sentence A

How are you?

好嗎你

Sentence B

Start from another unsupervised translation model (word embedding translation model)

[Lample, et al., ICLR, 2018]
Unsupervised learning with 10M sentences = Supervised learning with 100K sentence pairs

[Lample, et al., ICLR, 2018]
### Unsupervised PBSMT

<table>
<thead>
<tr>
<th></th>
<th>en→fr</th>
<th>fr→en</th>
<th>en→de</th>
<th>de→en</th>
<th>en→ro</th>
<th>ro→en</th>
<th>en→ru</th>
<th>ru→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised phrase table</td>
<td>-</td>
<td>17.50</td>
<td>-</td>
<td>15.63</td>
<td>-</td>
<td>14.10</td>
<td>-</td>
<td>8.08</td>
</tr>
<tr>
<td>Back-translation - Iter. 1</td>
<td>24.79</td>
<td>26.16</td>
<td>15.92</td>
<td>22.43</td>
<td>18.21</td>
<td>21.49</td>
<td>11.04</td>
<td>15.16</td>
</tr>
<tr>
<td>Back-translation - Iter. 2</td>
<td>27.32</td>
<td>26.80</td>
<td>17.65</td>
<td>22.85</td>
<td>20.61</td>
<td>22.52</td>
<td>12.87</td>
<td>16.42</td>
</tr>
<tr>
<td>Back-translation - Iter. 4</td>
<td>27.84</td>
<td>27.20</td>
<td>17.77</td>
<td>22.68</td>
<td>21.33</td>
<td>23.01</td>
<td>13.37</td>
<td><strong>16.62</strong></td>
</tr>
<tr>
<td>Back-translation - Iter. 5</td>
<td><strong>28.11</strong></td>
<td>27.16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Unsupervised NMT

<table>
<thead>
<tr>
<th></th>
<th>en→fr</th>
<th>fr→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>24.48</td>
<td>23.74</td>
</tr>
<tr>
<td>Transformer</td>
<td>25.14</td>
<td>24.18</td>
</tr>
</tbody>
</table>

### Phrase-based + Neural network

<table>
<thead>
<tr>
<th></th>
<th>en→fr</th>
<th>fr→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT + PBSMT</td>
<td>27.12</td>
<td>26.29</td>
</tr>
<tr>
<td>PBSMT + NMT</td>
<td>27.60</td>
<td><strong>27.68</strong></td>
</tr>
</tbody>
</table>

### Previous state-of-the-art - [Lample et al. (2018b)]

<table>
<thead>
<tr>
<th></th>
<th>en→fr</th>
<th>fr→en</th>
<th>en→de</th>
<th>de→en</th>
<th>en→ro</th>
<th>ro→en</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>25.1</td>
<td>24.2</td>
<td>17.2</td>
<td>21.0</td>
<td>21.2</td>
<td>19.4</td>
</tr>
<tr>
<td>PBSMT</td>
<td>28.1</td>
<td>27.2</td>
<td>17.8</td>
<td>22.7</td>
<td>21.3</td>
<td>23.0</td>
</tr>
<tr>
<td>PBSMT + NMT</td>
<td>27.6</td>
<td>27.7</td>
<td>20.2</td>
<td>25.2</td>
<td>25.1</td>
<td>23.9</td>
</tr>
</tbody>
</table>

### Our results for different encoder and decoder initializations

<table>
<thead>
<tr>
<th></th>
<th>EMB</th>
<th>EMB</th>
<th>EMB</th>
<th>EMB</th>
<th>EMB</th>
<th>EMB</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>13.0</td>
<td>15.8</td>
<td>6.7</td>
<td>15.3</td>
<td>18.9</td>
<td>18.3</td>
</tr>
<tr>
<td>-</td>
<td>25.3</td>
<td>26.4</td>
<td>19.2</td>
<td>26.0</td>
<td>25.7</td>
<td>24.6</td>
</tr>
<tr>
<td>-</td>
<td>29.2</td>
<td>29.1</td>
<td>21.6</td>
<td>28.6</td>
<td>28.2</td>
<td>27.3</td>
</tr>
<tr>
<td>CLM</td>
<td>28.7</td>
<td>28.2</td>
<td>24.4</td>
<td>30.3</td>
<td>29.2</td>
<td>28.0</td>
</tr>
<tr>
<td>CLM</td>
<td>30.4</td>
<td>30.0</td>
<td>22.7</td>
<td>30.5</td>
<td>29.0</td>
<td>27.8</td>
</tr>
<tr>
<td>CLM</td>
<td>32.3</td>
<td>31.6</td>
<td>24.3</td>
<td>32.5</td>
<td>31.6</td>
<td>29.8</td>
</tr>
<tr>
<td>MLM</td>
<td>31.6</td>
<td>32.1</td>
<td>27.0</td>
<td>33.2</td>
<td>31.8</td>
<td>30.5</td>
</tr>
<tr>
<td>MLM</td>
<td>33.4</td>
<td>32.3</td>
<td>24.9</td>
<td>32.9</td>
<td>31.7</td>
<td>30.4</td>
</tr>
<tr>
<td>MLM</td>
<td>33.4</td>
<td>33.3</td>
<td>26.4</td>
<td>34.3</td>
<td>33.3</td>
<td>31.8</td>
</tr>
</tbody>
</table>

[Lample, et al., NeurIPS’19]
Unsupervised Speech Recognition

How are you. 
\textit{hh aw aa r y uw}

He think it’s...
\textit{hh ih hh ih ng kcl ...}

Thanks for...
\textit{th ae ng k s f ao r ...}

text

G

ASR

D

Discriminator

text or transcription?

Phoneme Sequences
Acoustic Token Discovery
Acoustic tokens can be discovered from audio collection without text annotation.

Acoustic tokens: chunks of acoustically similar audio segments with token IDs
**Generator (v1)**

Acoustic Feature

Segmental Audio word2vec

Audio embedding sequence

ID sequence: 2, 16, 25, ... 2

Generator

Lookup Table

Generated phoneme sequence: sil, hh, ih, ... sil

K-means

Clusters:
- Cluster index 1
- Cluster index 2
- Cluster index 16
- Cluster index 25

[Wang, et al., ICASSP 2018]
Experiment

Matched Case (Oracle)

Nonmatched Case

4620 (TIMIT)

4620 (TIMIT)

3620 (TIMIT)

1000 (TIMIT)
## Experimental Results [Liu, et al., INTERSPEECH, 2018]

<table>
<thead>
<tr>
<th>Approaches</th>
<th>PER</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Matched</td>
<td>Nonmatched</td>
<td></td>
</tr>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN Transducer</td>
<td>17.7</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Standard HMMs</td>
<td>21.5</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

[Table] Experimental Results [Liu, et al., INTERSPEECH, 2018]
Generator (v2)

Sampling one output from each segment

Phoneme boundaries obtained by Gate Activation Signals (GAS)

[Wang, et al., INTERSPEECH 2017]

Inspired From [Yeh, et al., ICLR 2019]
### Experimental Results [Chen, et al., INTERSPEECH, 2019]

<table>
<thead>
<tr>
<th>Approaches</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Matched</td>
</tr>
<tr>
<td><strong>Supervised</strong></td>
<td></td>
</tr>
<tr>
<td>RNN Transducer</td>
<td>17.7</td>
</tr>
<tr>
<td>Standard HMMs</td>
<td>21.5</td>
</tr>
<tr>
<td><strong>Completely unsupervised (no label at all)</strong></td>
<td></td>
</tr>
<tr>
<td>Generator (v1)</td>
<td>76.0</td>
</tr>
<tr>
<td><strong>Iteration 1</strong></td>
<td></td>
</tr>
<tr>
<td>GAN</td>
<td>48.6</td>
</tr>
<tr>
<td><strong>Iteration 2</strong></td>
<td></td>
</tr>
<tr>
<td>GAN</td>
<td>41.0</td>
</tr>
<tr>
<td><strong>Iteration 3</strong></td>
<td></td>
</tr>
<tr>
<td>GAN</td>
<td>38.4</td>
</tr>
</tbody>
</table>
Refining Boundaries

1. Initial boundaries

2. Unsupervised train the GAN

3. Obtain transcriptions

4. Train the HMMs

5. Obtain new boundaries by HMMs

Phoneme Boundaries

GAN

Transcriptions (Pseudo label)

HMM
## Experimental Results

[Chen, et al., INTERSPEECH, 2019]

<table>
<thead>
<tr>
<th>Approaches</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervised</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN Transducer</td>
<td>17.7</td>
<td>-</td>
</tr>
<tr>
<td>Standard HMMs</td>
<td>21.5</td>
<td>-</td>
</tr>
<tr>
<td><strong>Completely unsupervised (no label at all)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Generator (v1)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iteration 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAN</td>
<td>48.6</td>
<td>50.0</td>
</tr>
<tr>
<td>HMM</td>
<td>30.7</td>
<td>39.5</td>
</tr>
<tr>
<td><strong>Iteration 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAN</td>
<td>41.0</td>
<td>44.3</td>
</tr>
<tr>
<td>HMM</td>
<td>27.0</td>
<td>35.5</td>
</tr>
<tr>
<td><strong>Iteration 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAN</td>
<td>38.4</td>
<td>44.2</td>
</tr>
<tr>
<td>HMM</td>
<td>26.1</td>
<td>33.1</td>
</tr>
</tbody>
</table>
The progress of supervised learning

Unsupervised learning today (2019) is as good as supervised learning 30 years ago.

**Semi-supervised Speech Recognition**

Using **100 hours pairs annotated audio** from Librispeech, and **text without audio**

21.7% WER → 18.7% WER
Shared Latent Space

- Initial attempt [Chen, et al., SLT, 2018]
- 76.3% WER on Librispeech [Chung, et al., NIPS 2018]
- WSJ with 2.5 hours paired data: 64.6% WER [Jennifer Drexler, et al., SLT 2018]
- LJ speech with 20 mins paired data: 11.7% PER [Ren, et al., ICML 2019]
- Unsupervised speech translation is possible [Chung, et al., ICASSP 2019]
**Image Style**

- Male
- Female

**Audio Style**

- Male
- Female

---

**Text Style Transfer**

- Positive sentences
- Negative sentences

---

**Unsupervised Abstractive Summarization**

- Document
- Summary

---

**Unsupervised Translation**

- Language 1
- Language 2

---

**Unsupervised ASR**

- Audio
- Text
Image Style

Audio Style

male

female

positive sentences

negative sentences

Text Style Transfer

Language 1

Language 2

Unsupervised Translation

Unsupervised Abstractive Summarization

Unsupervised ASR

Unsupervised Translation
Reference


• [Ning Dai, et al., ACL’19] Ning Dai, Jianze Liang, Xipeng Qiu, Xuanjing Huang, Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation, ACL, 2019


• [Lample, et al., EMNLP’18] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, Marc’Aurelio Ranzato, Phrase-Based & Neural Unsupervised Machine Translation, EMNLP, 2018
Reference

• [Hu, et al., ICML’17] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing, Toward Controlled Generation of Text, ICML, 2017


• [Shen, et al., NIPS’17] Tianxiao Shen, Tao Lei, Regina Barzilay, Tommi Jaakkola, Style Transfer from Non-Parallel Text by Cross-Alignment, NIPS, 2017

• [Li, et al., NAACL’18] Juncen Li, Robin Jia, He He, Percy Liang, Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer, NAACL, 2018

• [Xu, et al., ACL’18] Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, Wenjie Li, Unpaired Sentiment-to-Sentiment Translation: A Cycled Reinforcement Learning Approach, ACL, 2018
Reference


Reference

- [Conneau, et al., ICLR’18] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou, Word Translation Without Parallel Data, ICLR 2018
- [Lample, et al., ICLR’18] Guillaume Lample, Ludovic Denoyer, Marc'Aurelio Ranzato, Unsupervised Machine Translation Using Monolingual Corpora Only, ICLR, 2018
- [Artetxe, et al., ACL’18] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, ACL, 2018
Reference


Reference

• [Zhen Xu, et al., EMNLP, 2017] Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, Xiaolong Wang, Zhuoran Wang, Chao Qi, Neural Response Generation via GAN with an Approximate Embedding Layer, EMNLP, 2017


• [Kevin Lin, et al, NIPS, 2017] Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, Ming-Ting Sun, Adversarial Ranking for Language Generation, NIPS, 2017
Reference


• [Xu, et al., EMNLP, 2018] Jingjing Xu, Xuancheng Ren, Junyang Lin, Xu Sun, Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation, EMNLP, 2018


• [Cyprien de Masson d'Autume, et al., arXiv 2019] Cyprien de Masson d'Autume, Mihaela Rosca, Jack Rae, Shakir Mohamed, Training language GANs from Scratch, arXiv 2019
Reference

- [Yeh, et al., ICLR’19] Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, Dong Yu, Unsupervised Speech Recognition via Segmental Empirical Output Distribution Matching, ICLR, 2019
Reference

• [Chung, et al., NIPS 2018] Yu-An Chung, Wei-Hung Weng, Schrasing Tong, James Glass, Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces, NIPS, 2018


• [Ren, et al., ICML 2019] Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, Almost Unsupervised Text to Speech and Automatic Speech Recognition, ICML 2019