Learning about Language with Normalizing Flows

Graham Neubig
Language Technologies Institute, Carnegie Mellon University

Chunting Zhou  Junxian He  Xuezhe Ma

Di Wang, Daniel Spokoyny, Xian Li, Taylor Berg-Kirkpatrick, Eduard Hovy
Learning about Language?

- **Syntactic structure**

  The cat sat on a green wall

  Parts-of-speech: DT NN VBD IN DT JJ NN

  Dependency:

  - **Cross-lingual correspondences**

    a cat green on sat the wall

    の は 上 壁 猫 緑 座った
Supervised Approaches

John passes the ball uphill to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.
Supervised Learning

\[ X \quad \theta \quad Y \]

- John passes the ball uphill to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.
- John passes the ball uphill to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.
- John passes the ball uphill to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.

John passes the ball uphill to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.
Supervised Approaches

Supervised Learning

\[ X \rightarrow \theta \rightarrow Y \]

John passes the ball upfield to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.

John passes the ball upfield to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.

John passes the ball upfield to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.

John passes the ball upfield to Peter, who shoots for the goal. The shot is deflected by Mary and the ball goes out of bounds.
Unsupervised Approaches

- Learning language models $P(X)$
- Learning continuous features from language models (e.g. word2vec, skipthought, BERT)
- But how do we turn this into *interpretable structure*?
- How do we do it while *taking advantage of continuous features*?
Latent Variable Approaches

Unsupervised

\[ X \quad \theta \quad Y \]

\[ ? \quad ? \quad ? \]

\[ X \quad Y \]
Latent Variable Approaches

Unsupervised

\[
\begin{array}{c}
X \\
\theta \\
Y \\
\end{array}
\]

\[
\begin{array}{c}
? \\
? \\
? \\
\end{array}
\]
Density Matching for Bilingual Word Embedding

Chunting Zhou, Xuezhe Ma, Di Wang, Graham Neubig
(NAACL 2019)
Bilingual Word Embedding

- Map word embeddings from different languages into a single vector space
  - Cross-lingual transfer
  - Cross-lingual NLP tasks
Previous Work on Unsupervised BWE

• Unsupervised methods of minimization some form of distance between distributions of discrete vector sets:

  (A) X
  cat feline
car
depth

  (B) Y
  Y profondo
  WX

dee

gatto

  (C) Y
  WX

gatto
cat

  (D) Y
  WX

gatto
cat

• No direct probabilistic interpretation, not a "typical" unsupervised generative model
Density Mapping for Bilingual Word Embedding (DeMa-BWE)

- Mapping function is learned with normalizing flow
Normalizing Flows

$X \sim P(X)$

$Z = f_\theta(X)$

$Z \sim N(0, I)$

Change of variable formula:

$p_\theta(x) = p_Z(f_\theta(x)) \left| \det \left( \frac{\partial f_\theta(x)}{\partial x} \right) \right|$

Intuitively, prevents degenerative mapping of everything to zero vector

**Normalizing Flow:** A series of such invertible transformations $f$
DeMa-BWE: Preliminaries

Notations:

\[ x \in \mathbb{R}^d, \quad y \in \mathbb{R}^d \]: denote vectors in the src and tgt embedding space

\[ x_i, \quad y_j \]: denote an actual word in src and tgt vocabularies

\[ f_{xy}, \quad f_{yx} \]: denote src->tgt, and tgt-src mapping functions
Prior Distribution

• Assumption on the monolingual word embedding space: Gaussian mixture model

\[ p(x) = \sum_{i \in \{1, \ldots, N_x\}} \pi(x_i) \tilde{p}(x|x_i) \]

\[ \tilde{p}(x|x_i) = \mathcal{N}(x|x_i, \sigma^2_x I) \]
DeMa-BWE: Density Matching

- Sampling a continuous vector from the GMM
  \[ x_i \sim \pi(x_i) \quad x \sim \tilde{p}(x|x_i) \]

- Apply the mapping function \( f_{xy} \) to obtain the transformed vector in the target space.
  \[ f_{xy}(\cdot) = W_{xy}. \]

- Computing the density of \( x \) in the mapped target space
  \[ \log p(x; W_{xy}) = \log p(y) + \log |\det(W_{xy})| \]

- Objective: minimize: \( \text{KL}(p(x) || p(x; W_{xy})) \)
  \[ \mathcal{L}_{xy} = \mathbb{E}_{x \sim p(x)}[\log p(y) + \log |\det(W_{xy})|] \]
Method Details

- **Weak Orthogonality Constraint:** Try to make sure that the transformation is close to orthogonal

\[ L_{bt} = E_{x_i \sim \pi(x_i), x \sim \tilde{p}(x|x_i)} \left[ g(W_{yx} W_{xy} x, x) \right] + E_{y_j \sim \pi(y_j), y \sim \tilde{p}(y|x_j)} \left[ g(W_{xy} W_{yx} y, y) \right] \]

- **Weak Supervision w/ Identical Strings:** Take advantage of the fact that identical strings are usually the same word in both languages

\[ L_{sup} = \sum_{v \in W_{id}} g(v_x W_{xy}^T, v_y) + g(v_y W_{yx}^T, v_x) \]

- **Alignment Selection Methods:** Use cross-domain similarity local scaling (CSLS)

\[ CSLS(x', y) = 2\cos(x', y) - r_T(x') - r_S(y) \]
Experiments

• Dataset and Tasks
  • Bilingual Lexicon Induction Task: MUSE dataset (Conneau et al., 2017)
  • Cross-lingual Word Similarity Task: SemEval 2017

• Languages
  • Baseline languages: en - es, de, fr, ru, zh, ja
  • Morphologically rich languages: en - et, fi, el, hu, pl, tr
Main Results on BLI (close languages)

- en-de
- de-en
- en-es
- es-en

- Procrustes(R)
- MUSE (U+R)
- SL-unsup-ID
- DeMa-BWE

Scores: 70, 73.75, 77.5, 81.25, 85
Main Results on BLI (distant languages)

- Procrustes(R)
- MUSE (U+R)
- DeMa-BWE

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Procrustes(R)</th>
<th>MUSE (U+R)</th>
<th>DeMa-BWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-et</td>
<td>20</td>
<td>16.25</td>
<td>32.5</td>
</tr>
<tr>
<td>et-en</td>
<td>32.5</td>
<td>48.75</td>
<td>65</td>
</tr>
<tr>
<td>en-el</td>
<td>48.75</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>el-en</td>
<td>65</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>en-ja</td>
<td>16.25</td>
<td>32.5</td>
<td>48.75</td>
</tr>
<tr>
<td>ja-en</td>
<td>32.5</td>
<td>48.75</td>
<td>65</td>
</tr>
</tbody>
</table>
Unsupervised Learning of Syntactic Structure w/ Invertible Neural Projections

Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick
(EMNLP 2018)
HMM for Part-of-Speech Induction

The cat sat
Gaussian HMM for POS Induction

$$x_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$$

[Lin et al. 2015]
Latent Embeddings w/ Neural Projection

$z_i \sim \text{Markov Structure}$

$e_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$

$x_i = f_{\phi}(e_i)$

$z_1 \rightarrow z_2 \rightarrow z_3$

$e_1 \rightarrow e_2 \rightarrow e_3$

$x_1 \rightarrow x_2 \rightarrow x_3$
Dependency Model with Valence

The cat stopped a dog in Paris

[Klein and Manning 2004]
Dependency Model with Valence

[Klein and Manning 2004]
Dependency Parse Induction from POS
Grammar Induction from Raw Text
Grammar Induction from Raw Text

The cat stopped a dog in Paris
Latent Embeddings w/ Neural Projection

\[ z_i \sim \text{Syntax Model} \]
\[ e_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i}) \]

Neural Projector
\[ x_i = f_\phi(e_i) \]

Point mass at \( f_\phi(e_i) \)
Learning and Inference

\[ p(\mathbf{x}_i | z_i; \eta, \phi) \]
Learning and Inference

\[ p(\mathbf{x}_i | z_i; \eta, \phi) \]

Gaussian embedding parameters
Learning and Inference

\[ p(x_i | z_i; \eta, \phi) \]

Projection parameters
Learning and Inference

\[ \dim(x) = \dim(e) \text{ and } f \text{ is invertible} \]

\[
p(x_i | z_i; \eta, \phi) = p(f^{-1}_\phi(x_i) | z_i; \eta) \left| \det \frac{\partial f^{-1}}{\partial x_i} \right|
\]
Learning and Inference

\[
dim(x) = \dim(e) \text{ and } f \text{ is invertible}
\]

\[
p(x_i | z_i; \eta, \phi) = p(f^{-1}_\phi(x_i) | z_i; \eta) \left| \det \frac{\partial f^{-1}}{\partial x_i} \right|
\]

Determinant of Jacobian matrix
Learning and Inference

\[ \text{dim}(x) = \text{dim}(e) \text{ and } f \text{ is invertible} \]

\[ p(x_i | z_i; \eta, \phi) = p(f^{-1}_\phi(x_i) | z_i; \eta) \det \frac{\partial f^{-1}}{\partial x_i} \]

Gaussian distribution

Determinant of Jacobian matrix
Example of Markov prior

\[ \log p(x) = \log p_{\text{GHMM}}(f_\phi^{-1}(x)) \]

\[ + \sum \log \left| \det \frac{\partial f_\phi^{-1}}{\partial x_i} \right| \]

\[ -\infty \text{ when } f \text{ is not invertible} \]
Learning with Inverse Projection
Learning with Inverse Projection

\[ h^{(1)}_{i,l} = x_{i,l} \]

\[ h^{(1)}_{i,r} = x_{i,r} + g(x_{i,l}) \]

[Dinh et al. 2014]
Learning with Inverse Projection

\[ f^{-1}(x_i) \]

\[ h_{i,l}^{(1)} = x_{i,l} \]

\[ h_{i,r}^{(1)} = x_{i,r} + g(x_{i,l}) \]

\[ \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 \end{bmatrix} \]

[Dinh et al. 2014]
Experiments

- Dataset: English Penn Treebank
- POS tagging
  Trained and tested on whole PTB
- Grammar induction
  Trained on sentences of length $\leq 10$ in section 2-21
  Tested on sentences in section 23
Part-of-speech Induction

Outperform feature-based SOTA
Dependency Parse Induction

- **Oracle POS**
  - Directed % on len <= 10:
    - DM: 49.6
    - Gaussi: 55.4
    - Neural: 60.2
  - Directed % on all:
    - DM: 35.8
    - Gaussi: 43.1
    - Neural: 47.9
Original Embedding Space

- adjective
- adverb
- noun singular
- noun proper
- noun plural
- verb base
- verb gerund
- verb past tense
- verb past participle
- verb 3rd singular
- cardinal number
Projected Embedding Space w/ Markov Prior
Projected Embedding Space w/ DMV Prior

- adjective
- adverb
- noun singular
- noun proper
- noun plural
- verb base
- verb gerund
- verb past tense
- verb past participle
- verb 3rd singular
- cardinal number
Projected Embedding Space w/ DMV Prior

- adjective
- adverb
- noun singular
- noun proper
- noun plural
- verb base
- verb gerund
- verb past tense
- verb past participle
- verb 3rd singular
- cardinal number

Language Technologies Institute
Projected Embedding Space w/ DMV Prior

- adjective
- adverb
- noun singular
- noun proper
- noun plural
- verb base
- verb gerund
- verb past tense
- verb past participle
- verb 3rd singular
- cardinal number

- smokers
- advertisers
- riders
- performers
- issuers
- foes
- principals
- furriers
- aides
- specialists
- technicians
- authorities

- aide
- resident
- owner
- politician
- associate
- examiner
- attorney
- actress
- commissioner
- singer
FlowSeq: Non-Autoregressive Conditional Sequence Generation with Generative

Xuezhe Ma*, Chunting Zhou*, Xian Li, Graham Neubig, Eduard Hovy
Background

• Autoregressive Sequence Generation

\[ P_\theta(y|x) = \prod_{t=1}^{T} P_\theta(y_t|y_{<t}, x). \]

• Left-to-right factorization is not optimal
• Generation is not easily parallelizable on GPUs

• Non-autoregressive Sequence Generation

\[ P_\theta(y|x) = \prod_{t=1}^{T} P_\theta(y_t|x). \]
Motivation

- **Latent Variable Model**

\[
P_\theta(y|x) = \int_z P_\theta(y|z, x)p_\theta(z|x)dz,
\]

- \(p_\theta(z|x)\) is the prior distribution over latent z
- \(P_\theta(y|z, x)\) is the generative distribution (a.k.a decoder)
- non-autoregressive generation

\[
P_\theta(y|z, x) = \prod_{t=1}^{T} P_\theta(y_t|z, x).
\]
Reminder: Flow-based Generative Models

- **What is Generative Flows:**
  - Transform a simple distribution into a complex one through a chain of invertible transformations

- Change of variable formula:
  \[ p_\theta(z) = p_\gamma(f_\theta(z)) \left| \det \left( \frac{\partial f_\theta(z)}{\partial z} \right) \right|. \]

- Generative Flow:
  \[ z \xleftrightarrow{g_1} H_1 \xleftrightarrow{g_2} H_2 \xleftrightarrow{g_3} \cdots \xleftrightarrow{g_K} \nu, \]
FlowSeq

- Variational Training: FlowSeq optimizes the evidence lower bound (ELBO)

\[
\log P_\theta(y|x) \geq E_{q_\phi(z|y,x)}[\log P_\theta(y|z, x)] - KL(q_\phi(z|y,x) \| p_\theta(z|x)).
\]
FlowSeq Architecture

- **Source Encoder**
  - Standard Transformer encoder

- **Posterior: diagonal Gaussian**
  - The latent variables $z$ are represented as a sequence of continuous random variables with the same length as the target sequence $y$:

$$z = \{ z_1, \ldots, z_T \}$$

- **Decoder: Transformer decoder w/o causal masking**

$$q_\phi(z|y, x) = \prod_{t=1}^{T} \mathcal{N}(z_t | \mu_t(x, y), \sigma_t^2(x, y))$$
Flow Architecture for Prior

- Actnorm (activation normalization layer): \( z'_t = s \odot z_t + b \).
- Invertible Multi-head Linear Layers: \( z'_t = z_t W \), \((W: [dz \times dz])\)
- Affine Coupling Layers

\[
\begin{align*}
  z_a, z_b &= \text{split}(z) \\
  z'_a &= z_a \\
  z'_b &= s(z_a, x) \odot z_b + b(z_a, x) \\
  z' &= \text{concat}(z'_a, z'_b),
\end{align*}
\]
Decoding Process

- **Argmax Decoding**
  \[ z^* = \arg\max_{z \in Z} p_\theta(z|x) \]
  \[ y^* = \arg\max_y P_\theta(y|z^*, x) \]

- **Noisy Parallel Decoding (NPD):** rescoring multiple samples by a pre-trained auto-regressive model.

- **Importance Weighted Decoding (IWD):** rescoring multiple candidates by importance samples.

\[ z_i \sim p_\theta(z|x), \forall i = 1, \ldots, N \]
\[ \hat{y}_i = \arg\max_y P_\theta(y|z_i, x) \]
\[ z_i^{(k)} \sim q_\phi(z|\hat{y}_i, x), \forall k = 1, \ldots, K \]
\[ P(\hat{y}_i | x) \approx \frac{1}{K} \sum_{k=1}^{K} \frac{P_\theta(\hat{y}_i | z_i^{(k)} , x)p_\theta(z_i^{(k)} | x)}{q_\phi(z_i^{(k)} | \hat{y}_i, x)} \]
Experiments

- **MT benchmark datasets:**
  - IWSLT 2014 EN-DE
  - WMT14 EN-DE, DE-EN
  - WMT16 EN-RO, RO-EN

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Autoregressive Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer-base</td>
<td>27.30</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Our Implementation</td>
<td>27.16</td>
<td>31.44</td>
<td>32.92</td>
<td>33.09</td>
</tr>
<tr>
<td>Raw Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMLLM (refinement 1)</td>
<td>10.88</td>
<td>–</td>
<td>20.24</td>
<td>–</td>
</tr>
<tr>
<td>CMLLM (refinement 4)</td>
<td>22.06</td>
<td>–</td>
<td>30.89</td>
<td>–</td>
</tr>
<tr>
<td>CMLLM (refinement 10)</td>
<td><strong>24.65</strong></td>
<td>–</td>
<td><strong>32.53</strong></td>
<td>–</td>
</tr>
<tr>
<td>FlowSeq-large (Argmax)</td>
<td>20.85</td>
<td>25.40</td>
<td>29.86</td>
<td>30.59</td>
</tr>
<tr>
<td>FlowSeq-large (IWD n = 15)</td>
<td>22.94</td>
<td>27.16</td>
<td>31.08</td>
<td>32.03</td>
</tr>
<tr>
<td>FlowSeq-large (NPD n = 15)</td>
<td>23.14</td>
<td>27.71</td>
<td>31.97</td>
<td>32.46</td>
</tr>
<tr>
<td>FlowSeq-large (NPD n = 30)</td>
<td>23.64</td>
<td><strong>28.29</strong></td>
<td>32.35</td>
<td><strong>32.91</strong></td>
</tr>
<tr>
<td>Knowledge Distillation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAT w/ FT (Argmax)</td>
<td>17.69</td>
<td>21.47</td>
<td>27.29</td>
<td>29.06</td>
</tr>
<tr>
<td>NAT w/ FT (NPD n = 10)</td>
<td>18.66</td>
<td>22.42</td>
<td>29.02</td>
<td>31.14</td>
</tr>
<tr>
<td>NAT-IR (refinement 1)</td>
<td>13.91</td>
<td>16.77</td>
<td>24.45</td>
<td>25.73</td>
</tr>
<tr>
<td>NAT-IR (refinement 10)</td>
<td>21.61</td>
<td>25.48</td>
<td>29.32</td>
<td>30.19</td>
</tr>
<tr>
<td>NAT-REG (NPD n = 9)</td>
<td>24.61</td>
<td>28.90</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CMLLM (refinement 1)</td>
<td>15.12</td>
<td>22.26</td>
<td>23.65</td>
<td>22.78</td>
</tr>
<tr>
<td>CMLLM (refinement 4)</td>
<td>26.08</td>
<td>30.11</td>
<td>31.78</td>
<td>31.76</td>
</tr>
<tr>
<td>CMLLM (refinement 10)</td>
<td><strong>26.92</strong></td>
<td><strong>30.86</strong></td>
<td><strong>32.42</strong></td>
<td><strong>33.06</strong></td>
</tr>
<tr>
<td>FlowSeq-large (Argmax)</td>
<td>23.72</td>
<td>28.39</td>
<td>29.73</td>
<td>30.72</td>
</tr>
<tr>
<td>FlowSeq-large (IWD n = 15)</td>
<td>24.70</td>
<td>29.44</td>
<td>31.02</td>
<td>31.97</td>
</tr>
<tr>
<td>FlowSeq-large (NPD n = 15)</td>
<td>25.03</td>
<td>30.48</td>
<td>31.89</td>
<td>32.43</td>
</tr>
<tr>
<td>FlowSeq-large (NPD n = 30)</td>
<td>25.31</td>
<td>30.68</td>
<td>32.20</td>
<td>32.84</td>
</tr>
</tbody>
</table>
Decoding Speed

(a) batch size

(b) target length
Analysis of Translations

(c) Rescoring

(d) Diversity
An Example

| Source | Es kann nicht erklären, weshalb die National Security Agency Daten ber das Privatleben von Amerikanern sammelt und warum Whistleblower bestraft werden, die staatliches Fehlverhalten offenlegen. |
| Ground Truth | And, most recently, it cannot excuse the failure to design a simple website more than three years since the Affordable Care Act was signed into law. |
| Sample 1 | And recently, it cannot apologise for the inability to design a simple website in the more than three years since the adoption of Affordable Care Act. |
| Sample 2 | And recently, it cannot excuse the inability to design a simple website in more than three years since the adoption of Affordable Care Act. |
| Sample 3 | Recently, it cannot excuse the inability to design a simple website in more than three years since the Affordable Care Act has passed. |
Conclusion
Conclusion

• Normalizing flows for unsupervised learning

\[ X = f_\theta^{-1}(Z) \]
\[ Z = f_\theta(X) \]

• Learning of bilingual lexicons
• Learning of latent structure
• Learning of sequence-to-sequence models
Thank You! Questions?

https://github.com/violet-zct/DeMa-BWE

The cat sat on a green wall

https://github.com/jxhe/struct-learning-with-flow

https://github.com/XuezheMax/flowseq