Language Modeling & Grammar Induction

- **Goal of Language Modeling**: assign high likelihood to held-out data
- **Goal of Grammar Induction**: learn linguistically meaningful tree structures without supervision

**Incompatible?**
- For good language modeling performance, need little independence assumptions and make use of flexible models (e.g. deep networks)
- For grammar induction, need strong independence assumptions for tractable training and to imbue inductive bias (e.g. context-freeness grammars)
Language Modeling & Grammar Induction

- **Goal of *Language Modeling***: assign high likelihood to held-out data
- **Goal of *Grammar Induction***: learn linguistically meaningful tree structures without supervision
- **Incompatible?**
  - For good language modeling performance, need little independence assumptions and make use of flexible models (e.g. deep networks)
  - For grammar induction, need strong independence assumptions for tractable training and to imbue inductive bias (e.g. context-freeness grammars)
This Work: Unsupervised Recurrent Neural Network Grammars

- Use a flexible generative model without any explicit independence assumptions (RNNG) $\implies$ good LM performance
- Variational inference with a structured inference network (CRF parser) to regularize the posterior $\implies$ learn linguistically meaningful trees
Background: Recurrent Neural Network Grammars [Dyer et al. 2016]

- Structured joint generative model of sentence $x$ and tree $z$

$$p_\theta(x, z)$$

- Generate next word conditioned on partially-completed syntax tree

- Hierarchical generative process (cf. flat generative process of RNN)
Background: Recurrent Neural Network Language Models

Standard RNNLMs: flat left-to-right generation

\[ x_t \sim p_{\theta}(x \mid x_1, \ldots, x_{t-1}) = \text{softmax}(W h_{t-1} + b) \]
Introduce binary variables $z = [z_1, \ldots, z_{2T-1}]$ (unlabeled binary tree)

Sample action $z_t \in \{\text{GENERATE, REDUCE}\}$ at each time step:

$$z_t \sim \text{Bernoulli}(p_t)$$

$$p_t = \sigma(w^\top h_{\text{prev}} + b)$$
Background: RNNG [Dyer et al. 2016]

If $z_t = \text{GENERATE}$

Sample word from context representation
Background: RNNG [Dyer et al. 2016]

(Similar to standard RNNLMs)

\[ x \sim \text{softmax}(Wh_{\text{prev}} + b) \]
Background: RNNG [Dyer et al. 2016]

Obtain new context representation with $e_{\text{hungry}}$

$$h_{\text{new}} = \text{LSTM}(e_{\text{hungry}}, h_{\text{prev}})$$
Background: RNNG [Dyer et al. 2016]

\[ h_{\text{new}} = \text{LSTM}(e_{\text{cat}}, h_{\text{prev}}) \]
Background: RNNG [Dyer et al. 2016]

If $z_t = \text{REDUCE}$
If $z_t = \text{REDUCE}$

Pop last two elements

Background: RNNG [Dyer et al. 2016]
Background: RNNG [Dyer et al. 2016]

Obtain new representation of constituent

\[ e_{\text{hungry cat}} = \text{TreeLSTM}(e_{\text{hungry}}, e_{\text{cat}}) \]
Background: RNNG [Dyer et al. 2016]

Move the new representation onto the stack

\[ h_{\text{new}} = \text{LSTM}(e_{(\text{hungry cat})}, h_{\text{prev}}) \]
Background: RNNG [Dyer et al. 2016]

Different inductive biases from RNN LMs $\Rightarrow$ learn different generalizations about the observed sequence of terminal symbols in language

- Lower perplexity than neural language models [Dyer et al. 2016]
- Better at syntactic evaluation tasks (e.g. grammaticality judgment) [Kuncoro et al. 2018; Wilcox et al. 2019]
- Correlate with electrophysiological responses in the brain [Hale et al. 2018]

(All require supervised training on annotated treebanks)
Unsupervised Recurrent Neural Network Grammars

- RNNG as a tool to learn structured, syntax-aware generative model of language
- Variational inference for tractable training and to imbue inductive bias
**URNNG: Issues**

Approach to unsupervised learning: maximize log marginal likelihood

\[
\log p_\theta(x) = \log \sum_{z \in Z_T} p_\theta(x, z)
\]

**Intractability**

- $Z_T$: exponentially large space
- No dynamic program

$$z_j \sim p_\theta(z \mid x_{\text{all previous words}}, z_{\text{all previous actions}})$$
URNNG: Issues

Approach to unsupervised learning: maximize log marginal likelihood

\[ \log p_\theta(x) = \log \sum_{z \in \mathcal{Z}_T} p_\theta(x, z) \]

**Intractability**

- \( \mathcal{Z}_T \): exponentially large space
- No dynamic program

\[ z_j \sim p_\theta(z \mid x_{\text{all previous words}}, z_{\text{all previous actions}}) \]
Approach to unsupervised learning: maximize log marginal likelihood

$$\log p_\theta(x) = \log \sum_{z \in Z_T} p_\theta(x, z)$$

Unconstrained Latent Space

- Little **inductive bias** for meaningful trees to emerge through maximizing likelihood (cf. PCFGs)
- Preliminary experiments on exhaustive marginalization on short sentences (length < 10) were not successful
URNNG: Overview

Inference Network $q_\phi(z \mid x)$  Generative Model $p_\theta(x, z)$
Tractable Training

\[
\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] = \text{ELBO}(\theta, \phi; x)
\]

- Define variational posterior \( q_\phi(z|x) \) with an inference network \( \phi \)
- Maximize lower bound on \( \log p_\theta(x) \) with sampled gradient estimators
URNNG: Structured Inference Network

Unconstrained Latent Space

\[
\max_{\theta} \text{ELBO}(\theta, \phi; x) = \\
\min_{\theta} -\log p_\theta(x) + \text{KL}[q_\phi(z \mid x) \parallel p_\theta(z \mid x)]
\]

- Structured inference network with context-free assumptions (CRF parser)
- Combination of language modeling and posterior regularization objectives

Inference Network \( q_\phi(z \mid x) \)  
Generative Model \( p_\theta(x, z) \)
Posterior Regularization [Ganchev et al. 2010]

$$\min_{\theta} - \log p_\theta(\mathbf{x}) + \text{KL}[q_\phi(z \mid x) \| p_\theta(z \mid x)]$$
Inference Network Parameterization

Inference network: CRF constituency parser [Finkel et al. 2008; Durrett and Klein 2015]

- Bidirectional LSTM over $x$ to get hidden states

$$\overrightarrow{h}, \overleftarrow{h} = \text{BiLSTM}(x)$$

- Score $s_{ij} \in \mathbb{R}$ for an unlabeled constituent spanning $x_i$ to $x_j$

$$s_{ij} = \text{MLP}(\overrightarrow{h}_{j+1} - \overrightarrow{h}_i, \overleftarrow{h}_{i-1} - \overleftarrow{h}_j)$$

- Similar score parameterization to recent works [Wang and Chang 2016; Stern et al. 2017; Kitaev and Klein 2018]
Training

\[ \text{ELBO}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(z \mid x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right] \]

\[ = \mathbb{E}_{q_\phi(z \mid x)} [\log p_\theta(x, z)] + \mathbb{H}[q_\phi(z \mid x)] \]

Gradient-based optimization with Monte Carlo estimators

\[ \nabla_\theta \text{ELBO}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(z \mid x)} [\nabla_\theta \log p(x, z)] \]

\[ \nabla_\phi \text{ELBO}(\theta, \phi; \mathbf{x}) = \nabla_\phi \mathbb{E}_{q_\phi(z \mid x)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(z \mid x)} \right] \]

\[ = \mathbb{E}_{q_\phi(z \mid x)} [\log p_\theta(x, z) \nabla_\phi \log q_\phi(z \mid x)] + \nabla_\phi \mathbb{H}[q_\phi(z \mid x)] \]

- score function gradient estimator
- \( O(T^3) \) dynamic program

Sampling from \( q_\phi(z \mid x) \) with forward-filtering backward-sampling in \( O(T^3) \)
Experimental Setup

Tasks and Evaluation

- Language Modeling: Perplexity
- Unsupervised Parsing: Unlabeled $F_1$

Data

- English: Penn Treebank (40K sents, 24K word types). Different from standard LM setup from Mikolov et al. [2010].
- Chinese: Chinese Treebank (15K sents, 17K word types)
- Preprocessing: Singletons replaced with UNK. Punctuation is retained
Experimental Setup: Baselines

- LSTM Language Model: same size as the RNNG
- Parsing Predict Reading Network (PRPN) [Shen et al. 2018]: neural language model with gated layers to induce binary trees
- Supervised RNNG: RNNG trained on binarized gold trees
<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PTB</td>
</tr>
<tr>
<td>LSTM LM</td>
<td>93.2</td>
</tr>
</tbody>
</table>
# Language Modeling

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PTB</td>
</tr>
<tr>
<td>LSTM LM</td>
<td>93.2</td>
</tr>
<tr>
<td>PRPN (default)</td>
<td>126.2</td>
</tr>
<tr>
<td>PRPN (tuned)</td>
<td>96.7</td>
</tr>
</tbody>
</table>
## Language Modeling

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PTB</td>
</tr>
<tr>
<td>LSTM LM</td>
<td>93.2</td>
</tr>
<tr>
<td>PRPN (default)</td>
<td>126.2</td>
</tr>
<tr>
<td>PRPN (tuned)</td>
<td>96.7</td>
</tr>
<tr>
<td>Unsupervised RNNG</td>
<td>90.6</td>
</tr>
<tr>
<td>Supervised RNNG</td>
<td>88.7</td>
</tr>
</tbody>
</table>
Language Modeling

Perplexity on PTB by Sentence Length
### Grammar Induction

**Unlabeled $F_1$ with evalb**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unlabeled $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PTB</td>
</tr>
<tr>
<td>Right Branching Trees</td>
<td>34.8</td>
</tr>
<tr>
<td>Random Trees</td>
<td>17.0</td>
</tr>
<tr>
<td>PRPN (default)</td>
<td>32.9</td>
</tr>
<tr>
<td>PRPN (tuned)</td>
<td>41.2</td>
</tr>
<tr>
<td>Unsupervised RNNG</td>
<td>40.7</td>
</tr>
<tr>
<td>Oracle Binary Trees</td>
<td>82.5</td>
</tr>
</tbody>
</table>
Grammar Induction

Using evaluation setup from Drozdov et al. [2019]

<table>
<thead>
<tr>
<th>Model</th>
<th>$F_1$</th>
<th>+PP Heuristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPN-LM [Shen et al. 2018]</td>
<td>42.8</td>
<td>42.4</td>
</tr>
<tr>
<td>ON-LSTM [Shen et al. 2019]</td>
<td>49.4</td>
<td>–</td>
</tr>
<tr>
<td>DIORA [Drozdov et al. 2019]</td>
<td>49.6</td>
<td>56.2</td>
</tr>
<tr>
<td>PRPN (tuned)</td>
<td>49.0</td>
<td>49.9</td>
</tr>
<tr>
<td>Unsupervised RNNG</td>
<td>52.4</td>
<td>52.4</td>
</tr>
</tbody>
</table>

+PP Heuristic attaches trailing punctuation directly to root
Grammar Induction

Label Recall

<table>
<thead>
<tr>
<th>Label</th>
<th>URNNG</th>
<th>PRPN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBAR</td>
<td>74.8%</td>
<td>28.9%</td>
</tr>
<tr>
<td>NP</td>
<td>39.5%</td>
<td>63.9%</td>
</tr>
<tr>
<td>VP</td>
<td>76.6%</td>
<td>27.3%</td>
</tr>
<tr>
<td>PP</td>
<td>55.8%</td>
<td>55.1%</td>
</tr>
<tr>
<td>ADJP</td>
<td>33.9%</td>
<td>42.5%</td>
</tr>
<tr>
<td>ADVP</td>
<td>50.4%</td>
<td>45.1%</td>
</tr>
</tbody>
</table>
Syntactic Evaluation [Marvin and Linzen 2018]

Two minimally different sentences:

The senators near the assistant are old

*The senators near the assistant is old

Model must assign higher probability to the correct one

<table>
<thead>
<tr>
<th></th>
<th>RNNLM</th>
<th>PRPN</th>
<th>URNNG</th>
<th>RNNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perplexity</td>
<td>93.2</td>
<td>96.7</td>
<td>90.6</td>
<td>88.7</td>
</tr>
<tr>
<td>Syntactic Eval.</td>
<td>62.5%</td>
<td>61.9%</td>
<td>64.6%</td>
<td>69.3%</td>
</tr>
</tbody>
</table>
Syntactic Evaluation [Marvin and Linzen 2018]

Two minimally different sentences:

The senators near the assistant are old

*The senators near the assistant is old

Model must assign higher probability to the correct one

<table>
<thead>
<tr>
<th></th>
<th>RNLM</th>
<th>PRPN</th>
<th>URNNG</th>
<th>RNNG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perplexity</td>
<td>93.2</td>
<td>96.7</td>
<td>90.6</td>
<td>88.7</td>
</tr>
<tr>
<td>Syntactic Eval.</td>
<td>62.5%</td>
<td>61.9%</td>
<td>64.6%</td>
<td>69.3%</td>
</tr>
</tbody>
</table>
Limitations

- Unable to improve on right-branching baseline on unpunctuated corpus
- Slower to train due to the $O(T^3)$ dynamic program and multiple samples for gradient estimators
- Requires various optimization strategies: KL annealing, different optimizers for $\theta$ and $\phi$, etc.
Conclusion

- Flexible generative model + structured inference network = low perplexity + meaningful structure
- Role of language structure & latent variable modeling in deep learning?


Wenhui Wang and Baobao Chang. 2016. Graph-based Dependency Parsing with Bidirectional LSTM. In Proceedings of ACL.