# Maximum Reconstruction Estimation for Generative Latent-Variable Models

**Yong Cheng**

清華大學
Tsinghua University

交叉信息研究院
Institute for Interdisciplinary Information Sciences

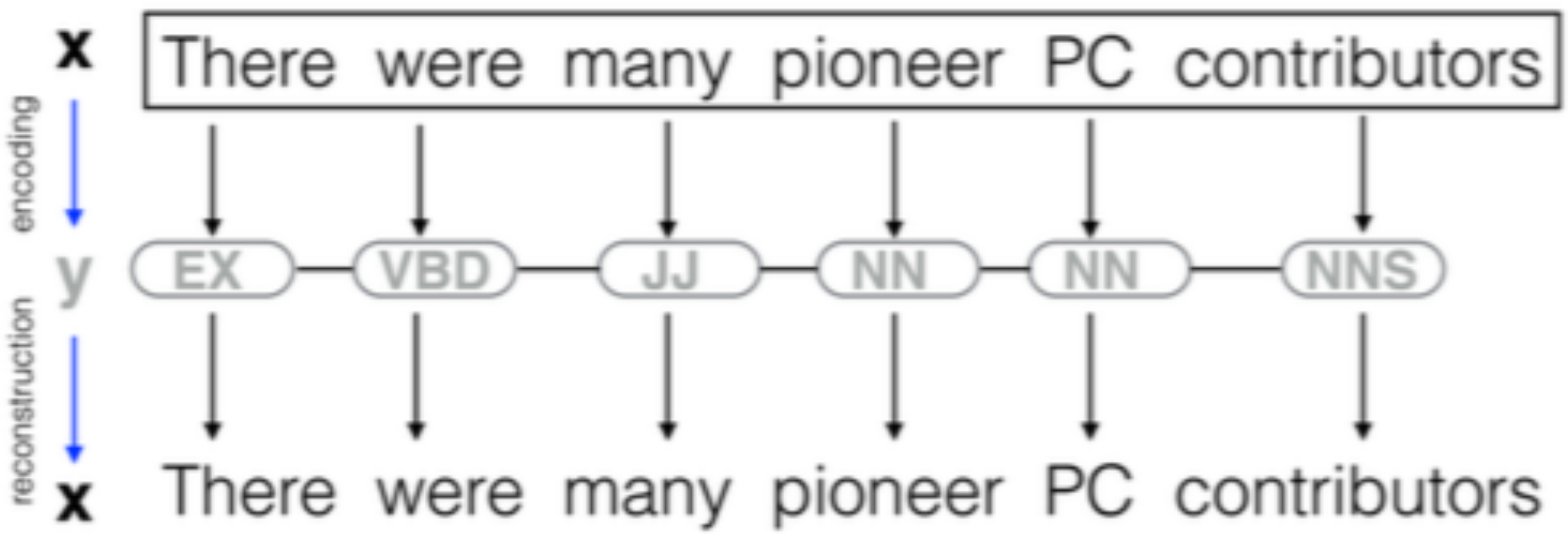
joint work with **Yang Liu, Wei Xu**

# Problem

- Generative latent-variable models are important for natural language processing due to their capability of providing compact representations of data.
- Maximum likelihood estimation suffers from a significant problem: it may guide the model to focus on explaining irrelevant but common correlations in the data.

# Maximum Reconstruction Estimation

- Circumvent irrelevant but common correlations by maximizing the probability of reconstructing observed data.



# Maximum Reconstruction Estimation

- Advantages:
    - Direct learning of model parameters.
    - Tractable inference.

# Maximum Likelihood Estimation

- A generative latent-variable model:

$$P(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$$

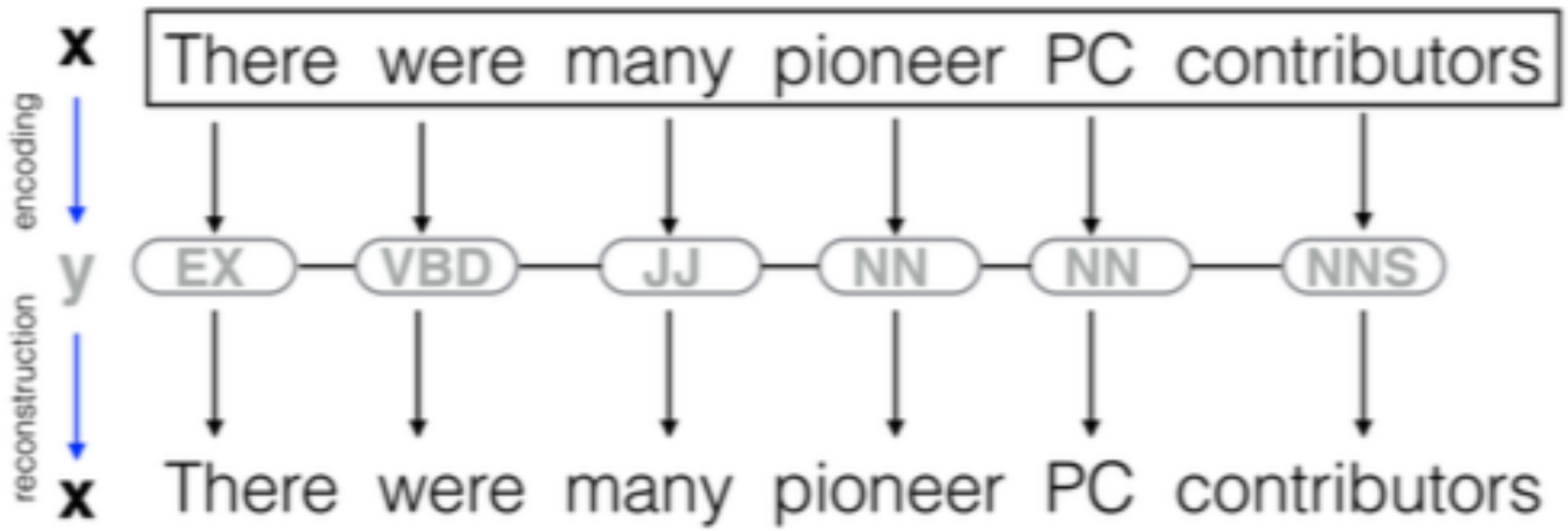
- Maximum likelihood estimation (MLE)

$$\boldsymbol{\theta}^*_{\mathrm{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \left\{ \sum_{s=1}^{S} \log P(\mathbf{x}^{(s)};\boldsymbol{\theta}) \right\}$$

- Inference

$$\mathbf{z}^*_{\mathrm{MLE}} = \operatorname*{argmax}_{\mathbf{z}} \left\{ P(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) \right\}$$

# Maximum Reconstruction Estimation

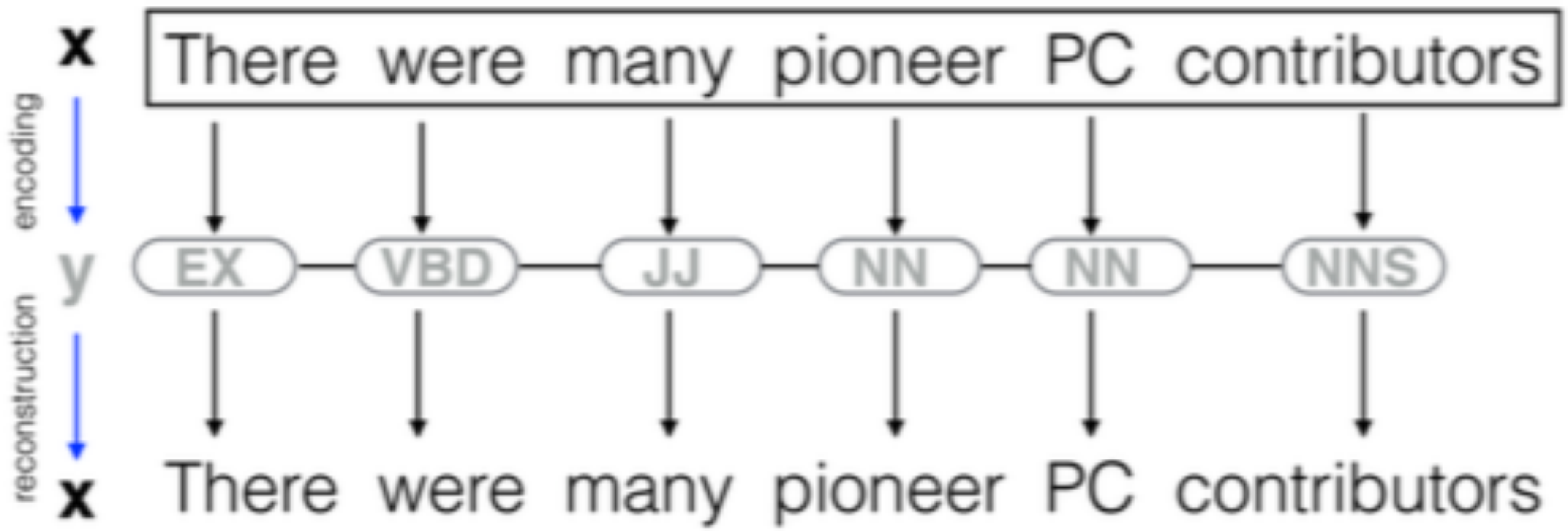



$$P(\hat{\mathbf{x}}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} \underbrace{P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})}_{encoding} \underbrace{P(\hat{\mathbf{x}}|\mathbf{z};\boldsymbol{\theta})}_{reconstruction}$$

$$= \mathbb{E}_{\mathbf{z}|\mathbf{x};\boldsymbol{\theta}}\left[P(\hat{\mathbf{x}}|\mathbf{z};\boldsymbol{\theta})\right]$$

Objective: $$\boldsymbol{\theta}^*_{\mathrm{MRE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \left\{ \sum_{s=1}^{S} \log P(\hat{\mathbf{x}}^{(s)}|\mathbf{x}^{(s)};\boldsymbol{\theta}) \right\}$$

# Maximum Reconstruction Estimation




$$P(\hat{\mathbf{x}}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} \underbrace{P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})}_{\text{encoding}} \underbrace{P(\hat{\mathbf{x}}|\mathbf{z};\boldsymbol{\theta})}_{\text{reconstruction}}$$

$$= \mathbb{E}_{\mathbf{z}|\mathbf{x};\boldsymbol{\theta}}\Big[P(\hat{\mathbf{x}}|\mathbf{z};\boldsymbol{\theta})\Big]$$

Objective: $$\boldsymbol{\theta}^*_{\text{MRE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \left\{ \sum_{s=1}^{S} \log P(\hat{\mathbf{x}}^{(s)}|\mathbf{x}^{(s)};\boldsymbol{\theta}) \right\}$$

Prediction: $$\mathbf{z}^*_{\text{MRE}} = \operatorname*{argmax}_{\mathbf{z}} \left\{ P(\hat{\mathbf{x}}|\mathbf{z};\boldsymbol{\theta}) P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) \right\}$$

# Maximum Reconstruction Estimation

- Two classical generative latent-variable models:
  - Hidden Markov models for unsupervised POS induction
  - IBM translation models for unsupervised word alignment

# Hidden Markov Models for Unsupervised POS Induction

| | | | | | |
|---|---|---|---|---|---|
| *latent structure* | NNP | VBD | DT | NN | NN |
| *observation* | **Obama** | **made** | **a** | **speech** | **yesterday** |

- Given an observed English sentence, the task is to induce the latent sequence of part-of-speech tags.

# Hidden Markov Models for Unsupervised POS Induction

| | | | | | |
|---|---|---|---|---|---|
| *latent structure* | NNP | VBD | DT | NN | NN |
| *observation* | **Obama** | **made** | **a** | **speech** | **yesterday** |

- Given an observed English sentence, the task is to induce the latent sequence of part-of-speech tags.

$$P(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1)\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)$$

# Hidden Markov Models for Unsupervised POS Induction

Maximum Reconstruction Estimation（MLE）

$$P(\mathbf{x}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$$

$$P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}) \; = \; \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

# Hidden Markov Models for Unsupervised POS Induction

Maximum Reconstruction Estimation（MLE）

$$P(\mathbf{x}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$$

$$P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}) \;=\; \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

$$P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) = \frac{P(\mathbf{z};\boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})}{P(\mathbf{x};\boldsymbol{\theta})}$$

# Hidden Markov Models for Unsupervised POS Induction

## Maximum Reconstruction Estimation（MLE）

$$P(\mathbf{x}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$$

$$P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}) \;=\; \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

$$P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) = \frac{P(\mathbf{z};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})}{P(\mathbf{x};\boldsymbol{\theta})}$$

# Hidden Markov Models for Unsupervised POS Induction

## Maximum Reconstruction Estimation（MLE）

$$P(\mathbf{x}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$$

$$P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

$$P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) = \frac{P(\mathbf{z};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})}{P(\mathbf{x};\boldsymbol{\theta})}$$

$$P(\mathbf{x}|\mathbf{x};\boldsymbol{\theta}) = \frac{\sum_{\mathbf{z}} P(\mathbf{z};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})^2}{P(\mathbf{x};\boldsymbol{\theta})}$$

# Hidden Markov Models for Unsupervised POS Induction

## Maximum Reconstruction Estimation（MLE）

$$\frac{\partial \log P(\mathbf{x};\boldsymbol{\theta})}{\partial p(z'|z)}$$
$$= \frac{1}{p(z'|z)}\mathbb{E}_{\mathbf{z}|\mathbf{x};\boldsymbol{\theta}}\left[\sum_{n=2}^{N}\delta(\mathbf{z}_{n-1},z)\delta(\mathbf{z}_n,z')\right]$$

## Maximum Reconstruction Estimation（MRE）

$$\frac{\partial \log P(\mathbf{x}|\mathbf{x};\boldsymbol{\theta})}{\partial p(z'|z)}$$
$$= \frac{1}{p(z'|z)}\Bigg(\mathbb{E}_{Q}\Big[\sum_{n=2}^{N}\delta(\mathbf{z}_{n-1},z)\delta(\mathbf{z}_n,z')\Big] - \mathbb{E}_{\mathbf{z}|\mathbf{x};\boldsymbol{\theta}}\Big[\sum_{n=2}^{N}\delta(\mathbf{z}_{n-1},z)\delta(\mathbf{z}_n,z')\Big]\Bigg)$$

Maximum Reconstruction Estimation（MLE）

$$\frac{\partial \log P(\mathbf{x};\boldsymbol{\theta})}{\partial p(z'|z)} = \frac{1}{p(z'|z)} \mathbb{E}_{\mathbf{z}|\mathbf{x};\boldsymbol{\theta}}\left[\sum_{n=2}^{N} \delta(\mathbf{z}_{n-1}, z)\delta(\mathbf{z}_n, z')\right]$$

Maximum Reconstruction Estimation（MRE）

$$\frac{\partial \log P(\mathbf{x}|\mathbf{x};\boldsymbol{\theta})}{\partial p(z'|z)} = \frac{1}{p(z'|z)}\left(\mathbb{E}_Q\left[\sum_{n=2}^{N} \delta(\mathbf{z}_{n-1}, z)\delta(\mathbf{z}_n, z')\right] - \mathbb{E}_{\mathbf{z}|\mathbf{x};\boldsymbol{\theta}}\left[\sum_{n=2}^{N} \delta(\mathbf{z}_{n-1}, z)\delta(\mathbf{z}_n, z')\right]\right)$$

$$Q(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \frac{P(\mathbf{z};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})^2}{\sum_{\mathbf{z}'} P(\mathbf{z}';\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z}';\boldsymbol{\theta})^2}$$

# Experiments

Comparison with MLE

| # state | MLE | | MRE | |
|---|---|---|---|---|
| | accuracy | VI | accuracy | VI |
| 10 | 0.4054 | 3.0575 | 0.3881 | 2.9322 |
| 20 | 0.4804 | 3.1119 | 0.5203 | 2.8879 |
| 30 | 0.5341 | 3.0835 | 0.5653 | 2.8199 |
| 40 | 0.5817 | 3.1780 | 0.6191 | 2.9255 |
| 50 | 0.6108 | 3.2087 | 0.6739 | 2.7522 |

# Experiments

## Comparison with MLE

| # sent. | MLE | | MRE | |
|---|---|---|---|---|
| | accuracy | VI | accuracy | VI |
| 10,000 | 0.5087 | 3.3471 | 0.5825 | 2.9018 |
| 20,000 | 0.5390 | 3.2387 | 0.5874 | 2.9217 |
| 30,000 | 0.5556 | 3.0764 | 0.6000 | 2.7904 |
| 40,000 | 0.5800 | 3.0117 | 0.6112 | 2.7403 |

# Experiments

## Comparison with CRF autoencoder

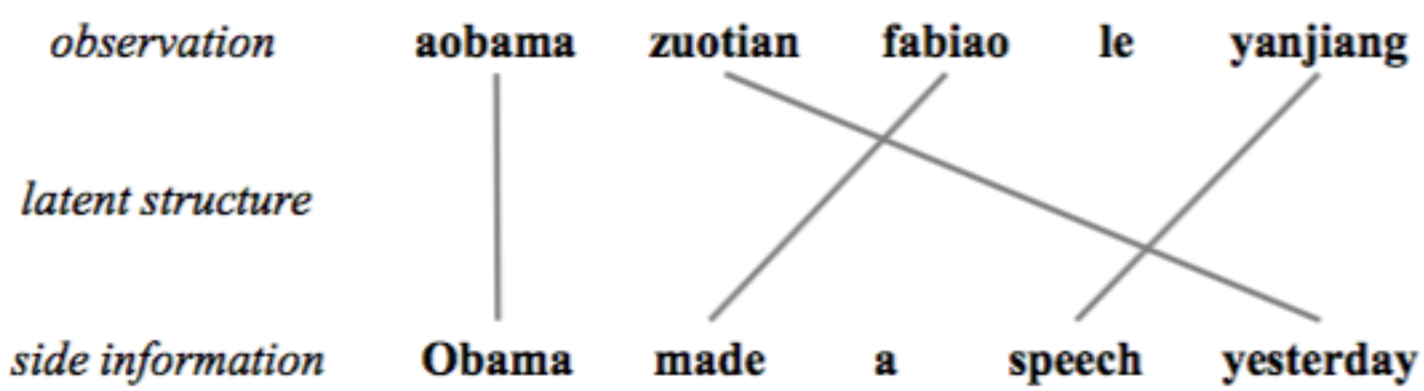
| # state | CRF Autoencoders | | MRE | |
|---|---|---|---|---|
| | accuracy | VI | accuracy | VI |
| 10 | 0.4059 | 2.7145 | 0.3881 | 2.9322 |
| 20 | 0.4657 | 2.7462 | 0.5203 | 2.8879 |
| 30 | 0.5479 | 2.9585 | 0.5653 | 2.8199 |
| 40 | 0.5377 | 3.1048 | 0.6191 | 2.9255 |
| 50 | 0.5662 | 2.8450 | 0.6739 | 2.7522 |

# Experiments

Example emission probabilities for the POS “VBD” (verb past tense)

| MLE | | MRE | |
|---|---|---|---|
| , | 0.2077 | said | 0.4632 |
| said | 0.1514 | says | 0.0773 |
| is | 0.0371 | reported | 0.0326 |
| says | 0.0312 | officials | 0.0198 |
| say | 0.0307 | announced | 0.0195 |
| : | 0.0237 | unit | 0.0158 |
| ’s | 0.0203 | noted | 0.0119 |
| think | 0.0169 | gained | 0.0106 |
| added | 0.0129 | told | 0.0102 |
| was | 0.0129 | court | 0.0101 |

# IBM Translation Models for Unsupervised Word Alignment




$$P(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{x},\mathbf{z}|\mathbf{y};\boldsymbol{\theta})$$
$$= \sum_{\mathbf{z}} \epsilon \prod_{n=1}^{N} p(\mathbf{z}_n|n,M,N)p(\mathbf{x}_n|\mathbf{y}_{\mathbf{z}_n})$$

# IBM Translation Models for Unsupervised Word Alignment

## Maximum Likelihood Estimation (MLE)

$$\frac{\partial \log P(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})}{\partial p(x|y)} =$$

$$\frac{1}{p(x|y)} \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}} \left[ \sum_{n=1}^{N} \delta(\mathbf{x}_n, x) \delta(\mathbf{y}_{\mathbf{z}_n}, y) \right]$$

## Maximum Reconstruction Estimation (MRE)

$$\frac{\partial \log P(\mathbf{x}|\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{\partial p(x|y)}$$

$$= \frac{1}{p(x|y)} \Bigg( \mathbb{E}_Q \left[ \sum_{n=1}^{N} 2\delta(\mathbf{x}_n, x) \delta(\mathbf{y}_{\mathbf{z}_n}, y) \right] - \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}} \left[ \sum_{n=1}^{N} \delta(\mathbf{x}_n, x) \delta(\mathbf{y}_{\mathbf{z}_n}, y) \right] \Bigg)$$

# IBM Translation Models for Unsupervised Word Alignment

## Maximum Likelihood Estimation (MLE)

$$\frac{\partial \log P(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})}{\partial p(x|y)} = \frac{1}{p(x|y)}\mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}}\left[\sum_{n=1}^{N}\delta(\mathbf{x}_n, x)\delta(\mathbf{y}_{\mathbf{z}_n}, y)\right]$$

## Maximum Reconstruction Estimation (MRE)

$$\frac{\partial \log P(\mathbf{x}|\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{\partial p(x|y)} = \frac{1}{p(x|y)}\left(\mathbb{E}_Q\left[\sum_{n=1}^{N} 2\delta(\mathbf{x}_n, x)\delta(\mathbf{y}_{\mathbf{z}_n}, y)\right] - \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}}\left[\sum_{n=1}^{N}\delta(\mathbf{x}_n, x)\delta(\mathbf{y}_{\mathbf{z}_n}, y)\right]\right)$$

$$Q(\mathbf{x},\mathbf{y},\mathbf{z};\boldsymbol{\theta}) = \frac{P(\mathbf{x},\mathbf{z}|\mathbf{y};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z},\mathbf{y};\boldsymbol{\theta})}{\sum_{\mathbf{z}'} P(\mathbf{x},\mathbf{z}'|\mathbf{y};\boldsymbol{\theta})P(\mathbf{x}|\mathbf{z}',\mathbf{y};\boldsymbol{\theta})}$$

## Comparison with MLE

| criterion | model | C → E | E → C |
|---|---|---|---|
| MLE | Model 1 | 43.07 | 45.89 |
| | Model 2 | 40.28 | 42.38 |
| MRE | Model 1 | 41.90 | 45.39 |
| | Model 2 | 38.33 | 41.73 |

# Conclusion

- We have presented maximum reconstruction estimation for training generative latent-variable models such as hidden Markov models and IBM translation models.
- In the future, we plan to apply our approach to more generative latent-variable models such as probabilistic context-free grammars and explore the possibility of developing new training algorithms that minimize reconstruction errors.

# Thank you !