**Maximum Likelihood Estimation** The partial derivatives of $\log P(\mathbf{x}|\mathbf{y})$ with respect to model parameters are

$$\frac{\partial \log P(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})}{\partial p(x|y)} = \frac{1}{p(x|y)} \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}} \left[ \sum_{n=1}^{N} \delta(\mathbf{x}_n, x)\delta(\mathbf{y}_{\mathbf{z}_n}, y) \right] \quad (21)$$

$$\frac{\partial \log P(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})}{\partial p(m'|n', M', N')} = \frac{\delta(M', M)\delta(N', N)}{p(m'|n', M', N')} \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}} \left[ \sum_{n=1}^{N} \delta(\mathbf{z}_n, m')\delta(n, n') \right] \quad (22)$$

Please refer to (Brown et al. 1993) for details about calculating the expectations.

**Maximum Reconstruction Estimation** Although both $\mathbf{x}$ and $\mathbf{y}$ are observed, we are only interested in reconstructing $\mathbf{x}$ because of the goal of machine translation: translating $\mathbf{y}$ to $\mathbf{x}$. The reconstruction probability is defined as

$$P(\mathbf{x}|\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z}, \mathbf{y}; \boldsymbol{\theta}) \quad (23)$$

$$= \frac{\sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z}|\mathbf{y}; \boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z}, \mathbf{y}; \boldsymbol{\theta})}{P(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta})} \quad (24)$$

where the probability of re-generating $\mathbf{x}$ given $\mathbf{y}$ and $\mathbf{z}$ is given by

$$P(\mathbf{x}|\mathbf{z}, \mathbf{y}; \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{y}_{\mathbf{z}_n}) \quad (25)$$

Note that all terms in Eq. (24) are differentiable with respect to model parameters.

The partial derivatives of $\log P(\mathbf{x}|\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ with respect to model parameters are

$$\frac{\partial \log P(\mathbf{x}|\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})}{\partial p(x|y)} = \frac{1}{p(x|y)} \left( \mathbb{E}_Q \Big[ \sum_{n=1}^{N} 2\delta(\mathbf{x}_n, x)\delta(\mathbf{y}_{\mathbf{z}_n}, y) \Big] - \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}} \Big[ \sum_{n=1}^{N} \delta(\mathbf{x}_n, x)\delta(\mathbf{y}_{\mathbf{z}_n}, y) \Big] \right) \quad (26)$$

$$\frac{\partial \log P(\mathbf{x}|\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})}{\partial p(m'|n', M', N')} = \frac{\delta(M', M)\delta(N', N)}{p(m'|n', M', N')} \times \left( \mathbb{E}_Q \Big[ \sum_{n=1}^{N} \delta(\mathbf{z}_n, m')\delta(n, n') \Big] - \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y};\boldsymbol{\theta}} \Big[ \sum_{n=1}^{N} \delta(\mathbf{z}_n, m')\delta(n, n') \Big] \right) \quad (27)$$

| # state | MLE accuracy | MLE VI | MRE accuracy | MRE VI |
|---|---|---|---|---|
| 10 | 0.4054 | 3.0575 | 0.3381 | 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 |

Table 1: Comparison of MLE and MRE on HMMs for unsupervised part-of-speech induction. The evaluation metrics are many-to-one accuracy (accuracy) and variation of information (VI).

where the distribution $Q$ is defined as

$$Q(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) = \frac{P(\mathbf{x}, \mathbf{z}|\mathbf{y}; \boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z}, \mathbf{y}; \boldsymbol{\theta})}{\sum_{\mathbf{z}'} P(\mathbf{x}, \mathbf{z}'|\mathbf{y}; \boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z}', \mathbf{y}; \boldsymbol{\theta})} \quad (28)$$

Note that Eq. (28) is equivalent to

$$Q(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) = \frac{P(\mathbf{z}|\mathbf{y}; \boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z}, \mathbf{y}; \boldsymbol{\theta})^2}{\sum_{\mathbf{z}'} P(\mathbf{z}'|\mathbf{y}; \boldsymbol{\theta}) P(\mathbf{x}|\mathbf{z}', \mathbf{y}; \boldsymbol{\theta})^2} \quad (29)$$

The expectations with respect to $Q$ can also be exactly and efficiently calculated (see Appendix B).

## Experiments

We evaluated our approach on two unsupervised NLP tasks: part-of-speech induction and word alignment.

### Evaluation on Part-of-Speech Induction

**Setting** We split the English Penn Treebank into two parts: 46K sentences for training and test and 1K sentences for optimizing hyper-parameters of the exponentiated gradient (EG) algorithm with adaptive learning rate. Each word is manually labeled with a gold-standard part-of-speech tag. We used two evaluation metrics: *many-to-1 accuracy* (Johnson 2007) and *variation of information* (VI) (Beal 2003). The EM algorithm for maximum likelihood estimation runs for 100 iterations and the EG algorithm with adaptive learning rate runs for 50 iterations with initialization of a basic HMM (Ammar, Dyer, and Smith 2014). The number of hidden states in HMMs is set to 50, which is close to the size of the POS tag set.

**Comparison with MLE** Table 1 shows the comparison of MLE and MRE. With the increase of the number of hidden states, the expressiveness of HMMs generally improves accordingly. We find that MRE outperforms MLE for 50-state HMMs in terms of both *many-to-one accuracy* and *VI*, suggesting that our approach is capable of guiding the HMMs to use latent structures to find intended correlations in the data. The differences are statistically significant ($p < 0.01$). On average, the reconstruction probability of training examples using model parameters learned by MLE (i.e., $P(\mathbf{x}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{MLE}})$) is $e^{-105}$. In contrast, the average reconstruction probability by MRE (i.e., $P(\mathbf{x}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{MRE}})$) is $e^{-84}$.

Table 2 gives the results on training corpora with various sizes. Generally, the accuracy improves with the increase of