Semi-supervised Learning for Neural Machine Translation

Yong Cheng

清華大學
Tsinghua University

交叉信息研究院
Institute for Interdisciplinary Information Sciences

joint work with Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, Yang Liu

Machine Translation

Automated translation using computer software

Machine Translation

- Rule-based Machine Translation 1970s
- Example-based Machine Translation 1984
- Statical Machine Translation (SMT) 1993
- **Neural Machine Translation （NMT） 2014**

Trends: learning to translate from **DATA**

Machine Translation

Parallel corpora are usually limited in

quantity & **quality** & **coverage**

Monolingual Corpora

Parallel Corpora

Monolingual Corpora Used in SMT and NMT

- N-gram language model in SMT
  Koehn et al., [2007]

- Monolingual corpora as decipherment
  Ravi and Knight [2011]

- Integrate a neural language model into NMT.
  Gulccehre et al. [2015]

- Additional pseudo parallel corpus.
  Sennrich et al. [2016]

Supervised Training

Parallel Corpus $$\mathcal{D} = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$$

Objective $$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}; \boldsymbol{\theta})$$

Unsupervised Training

Monolingual Corpus $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$

Our Approach — Autoencoders

bushi yu shalong juxing le huitan

$\mathbf{x}$

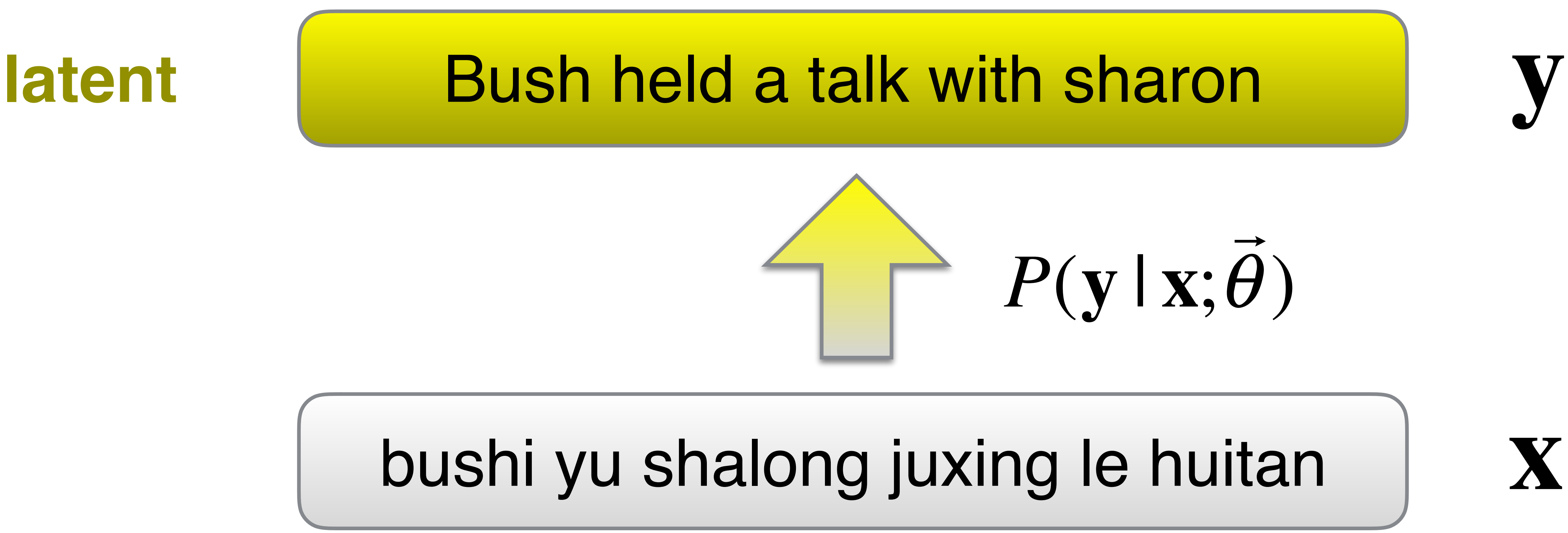
Our Approach — Autoencoders

$$P(\mathbf{y} \mid \mathbf{x}; \vec{\theta})$$

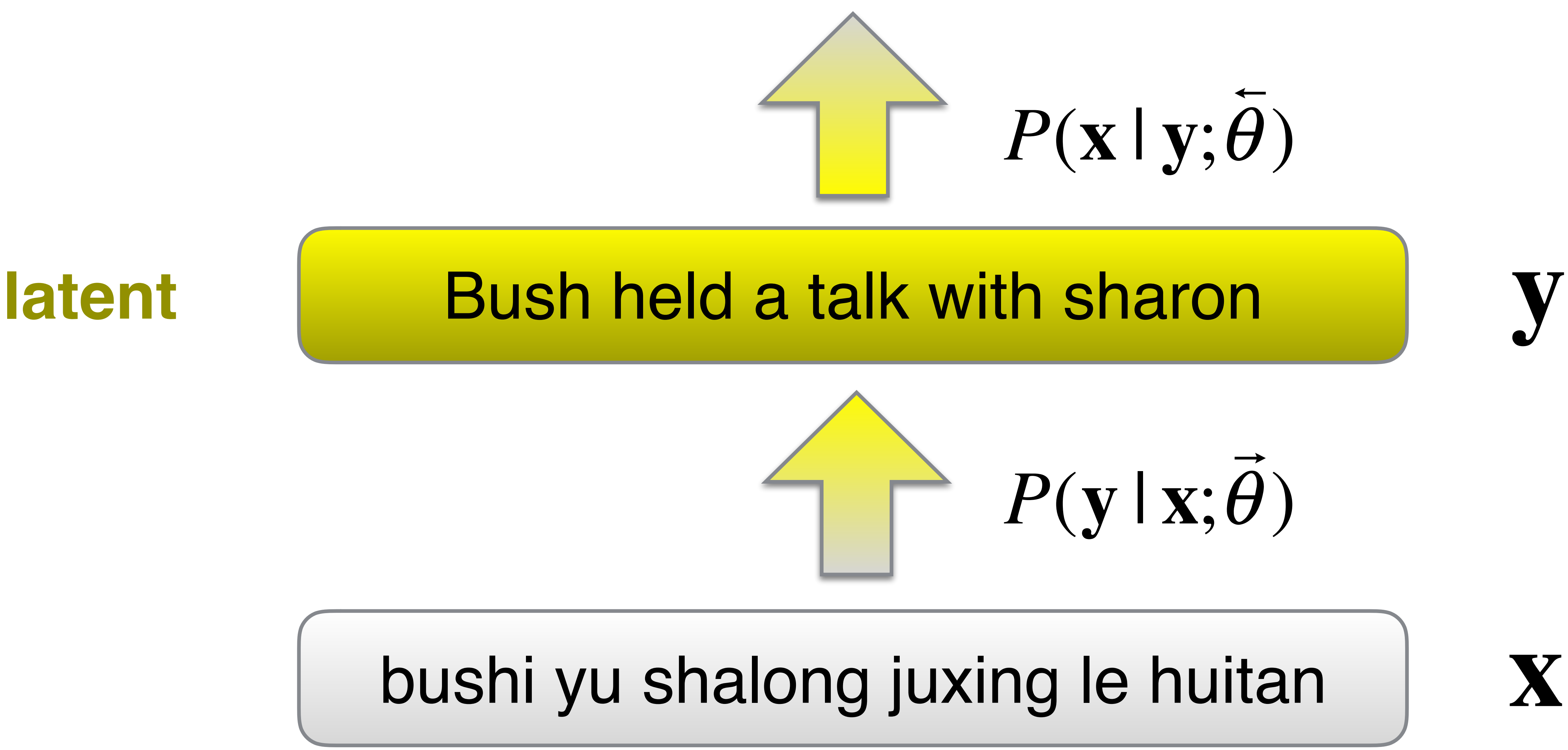
bushi yu shalong juxing le huitan

$\mathbf{x}$

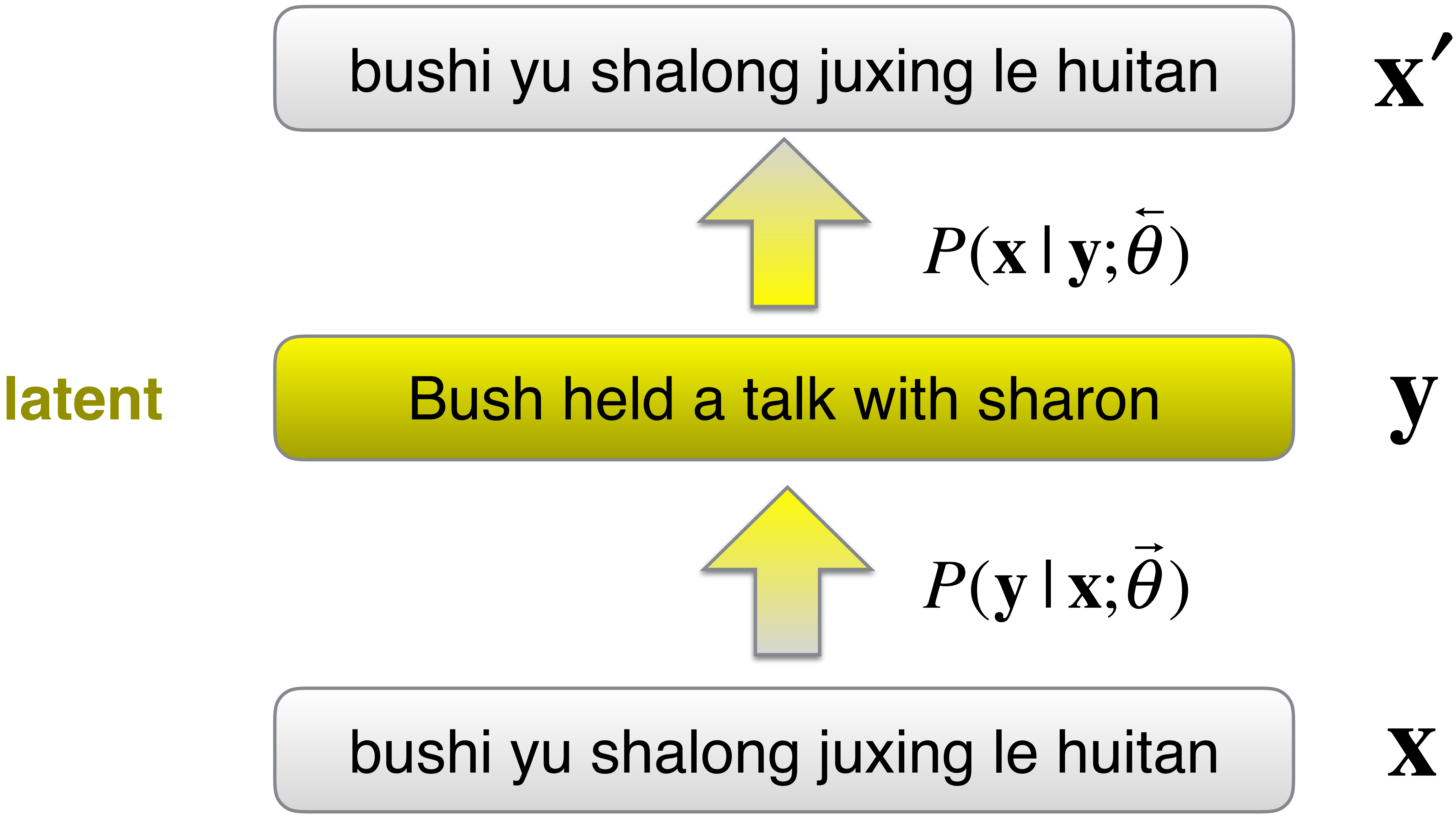
Our Approach — Autoencoders



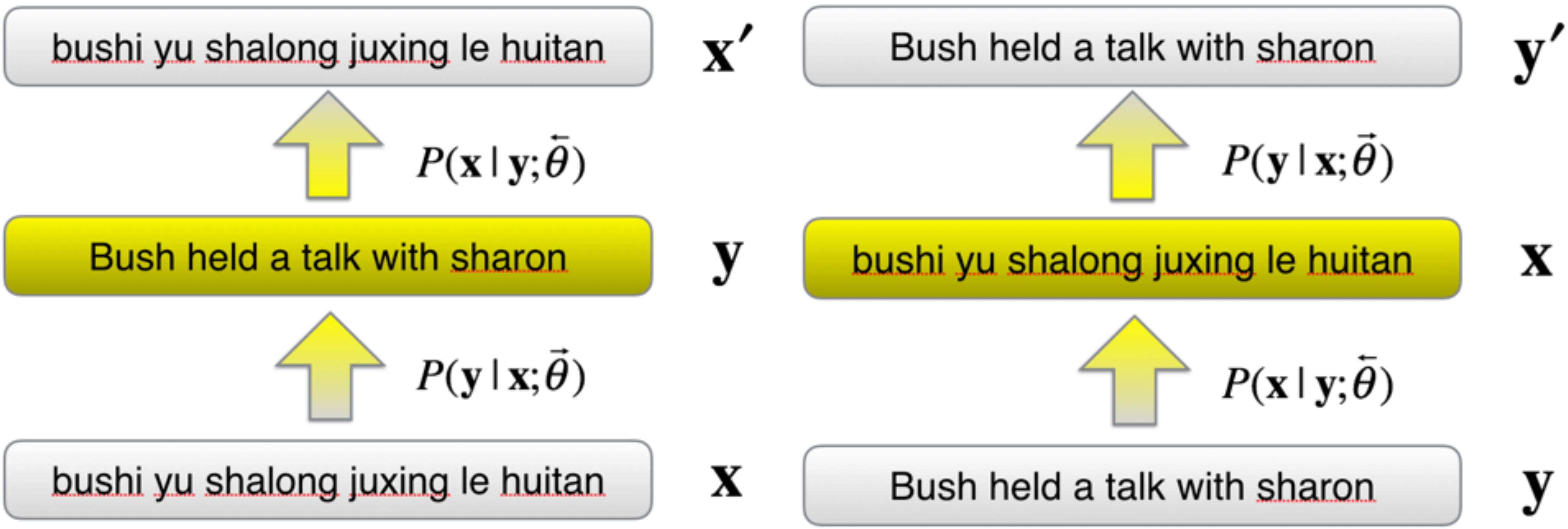
Our Approach — Autoencoders



Our Approach — Autoencoders



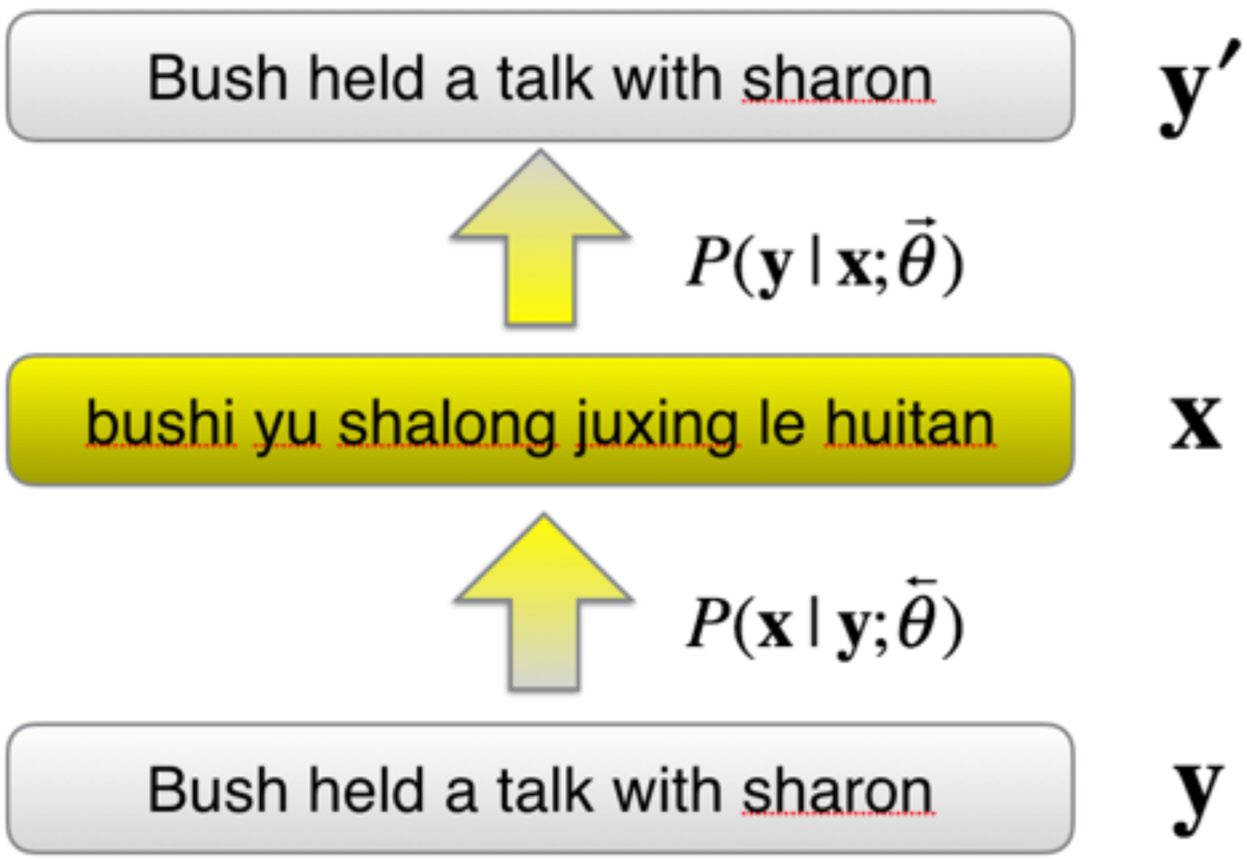
Our Approach — Autoencoders



source autoencoder **target autoencoder**

Unsupervised Training (Autoencoders)

Monolingual Corpus $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$



target autoencoder

$$
\begin{aligned}
& P(\mathbf{y}'|\mathbf{y}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) \\
= & \sum_{\mathbf{x}} P(\mathbf{y}', \mathbf{x}|\mathbf{y}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) \\
= & \sum_{\mathbf{x}} \underbrace{P(\mathbf{x}|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}})}_{encoder} \underbrace{P(\mathbf{y}'|\mathbf{x}; \overrightarrow{\boldsymbol{\theta}})}_{decoder}
\end{aligned}
$$

Semi-supervised Training

Training Objective

$$
\begin{aligned}
& J(\overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) \\
= & \underbrace{\sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \overrightarrow{\boldsymbol{\theta}})}_{\textit{source-to-target likelihood}} + \underbrace{\sum_{n=1}^{N} \log P(\mathbf{x}^{(n)}|\mathbf{y}^{(n)}; \overleftarrow{\boldsymbol{\theta}})}_{\textit{target-to-source likelihood}} \\
& \boxed{+\lambda_1 \underbrace{\sum_{t=1}^{T} \log P(\mathbf{y}'|\mathbf{y}^{(t)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})}_{\textit{target autoencoder}} + \lambda_2 \underbrace{\sum_{s=1}^{S} \log P(\mathbf{x}'|\mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})}_{\textit{source autoencoder}},}
\end{aligned}
$$

Translation Results

Compared with Moses (SMT) and RNNSearch (NMT)

System	Training Data			Direction	NIST06	NIST02
	CE	C	E			
MOSES	√	×	×	C → E E → C	32.48 14.27	32.69 18.28
RNNSEARCH	√	×	×	C → E E → C	30.74 15.71	35.16 20.76

Translation Results

Compared with Moses (SMT) and RNNSearch (NMT)

System	Training Data			Direction	NIST06	NIST02
	CE	C	E			
Moses	√	×	×	C → E	32.48	32.69
				E → C	14.27	18.28
	√	×	√	C → E	34.59	35.21
	√	√	×	E → C	20.69	25.85
RNNSearch	√	×	×	C → E	30.74	35.16
				E → C	15.71	20.76

Translation Results

Compared with Moses (SMT) and RNNSearch (NMT)

System	Training Data CE	C	E	Direction	NIST06	NIST02
MOSES	√	×	×	C → E	32.48	32.69
				E → C	14.27	18.28
	√	×	√	C → E	34.59	35.21
	√	√	×	E → C	20.69	25.85
RNNSEARCH	√	×	×	C → E	30.74	35.16
				E → C	15.71	20.76
	√	×	√	C → E	35.61**++	38.78**++
				E → C	17.59++	23.99 ++

Translation Results

Compared with Moses (SMT) and RNNSearch (NMT)

System	Training Data			Direction	NIST06	NIST02
	CE	C	E			
MOSES	√	×	×	C → E	32.48	32.69
				E → C	14.27	18.28
	√	×	√	C → E	34.59	35.21
	√	√	×	E → C	20.69	25.85
RNNSEARCH	√	×	×	C → E	30.74	35.16
				E → C	15.71	20.76
	√	×	√	C → E	35.61**++	38.78**++
				E → C	17.59++	23.99 ++
	√	√	×	C → E	35.01++	38.20**++
				E → C	21.12*++	29.52**++

Translation Results

Compared with Moses (SMT) and RNNSearch (NMT)

System	Training Data CE	C	E	Direction	NIST06	NIST02	NIST03	NIST04	NIST05
Moses	√	×	×	C → E	32.48	32.69	32.39	33.62	30.23
				E → C	14.27	18.28	15.36	13.96	14.11
	√	×	√	C → E	34.59	35.21	35.71	35.56	33.74
	√	√	×	E → C	20.69	25.85	19.76	18.77	19.74
RNNSearch	√	×	×	C → E	30.74	35.16	33.75	34.63	31.74
				E → C	15.71	20.76	16.56	16.85	15.14
	√	×	√	C → E	35.61**++	38.78**++	38.32**++	38.49**++	36.45**++
				E → C	17.59++	23.99 ++	18.95++	18.85++	17.91++
	√	√	×	C → E	35.01++	38.20**++	37.99**++	38.16**++	36.07**++
				E → C	21.12*++	29.52**++	20.49**++	21.59**++	19.97++

Translation Results

Compared with *Sennrich et al. [2015a]*

Method	Training Data CE	C	E	Direction	NIST06	NIST02	NIST03	NIST04	NIST05
Sennrich *et al.* [2015a]	√	×	√	C →E	34.10	36.95	36.80	37.99	35.33
	√	√	×	E →C	19.85	28.83	20.61	20.54	19.17
this work	√	×	√	C →E	35.61**	38.78**	38.32**	38.49*	36.45**
				E →C	17.59	23.99	18.95	18.85	17.91
	√	√	×	C →E	35.01**	38.20**	37.99**	38.16	36.07**
				E →C	21.12**	29.52**	20.49	21.59**	19.97**

Example Translation of Monolingual Corpus

Monolingual	hongsen shuo , ruguo you na jia famu gongsi dangan yishenshifa , name tamen jiang zihui qiancheng .
Reference	hongsen said, if any *logging companies* dare to defy the law, then they will *destroy their own future* .
Translation	hun sen said , if any of *those companies* dare defy the law , then they will *have their own fate* . [*iteration 0*]
	hun sen said if any *tree felling company* dared to break the law , then they would *kill themselves* . [*iteration 40K*]
	hun sen said if any *logging companies* dare to defy the law , they would *destroy the future themselves* . [*iteration 240K*]

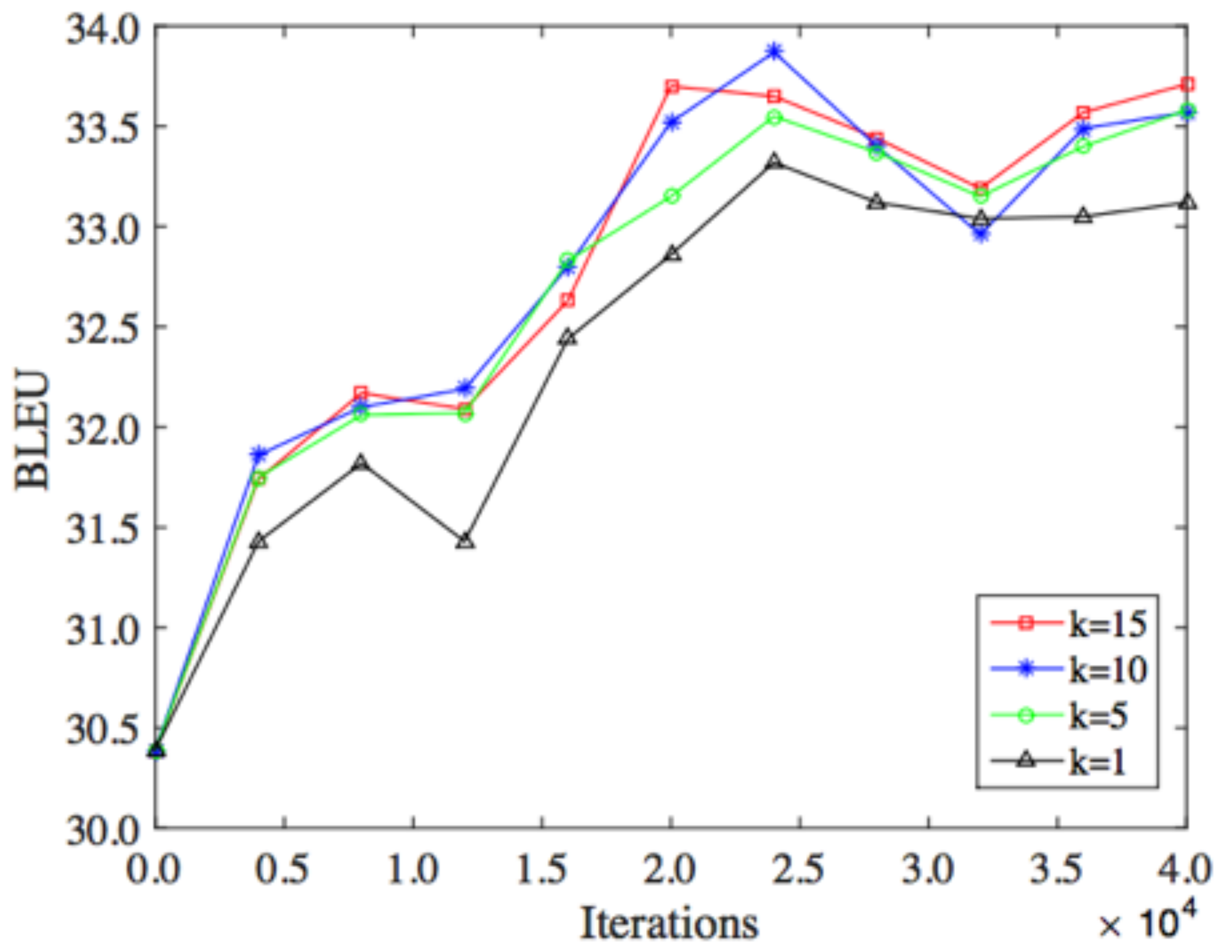
$$\mathbf{y}^* = \operatorname*{argmax}_{\mathbf{x}} \left\{ P(\mathbf{y}|\mathbf{x}; \overrightarrow{\boldsymbol{\theta}}) P(\mathbf{x}'|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}}) \right\}$$

Conclusion

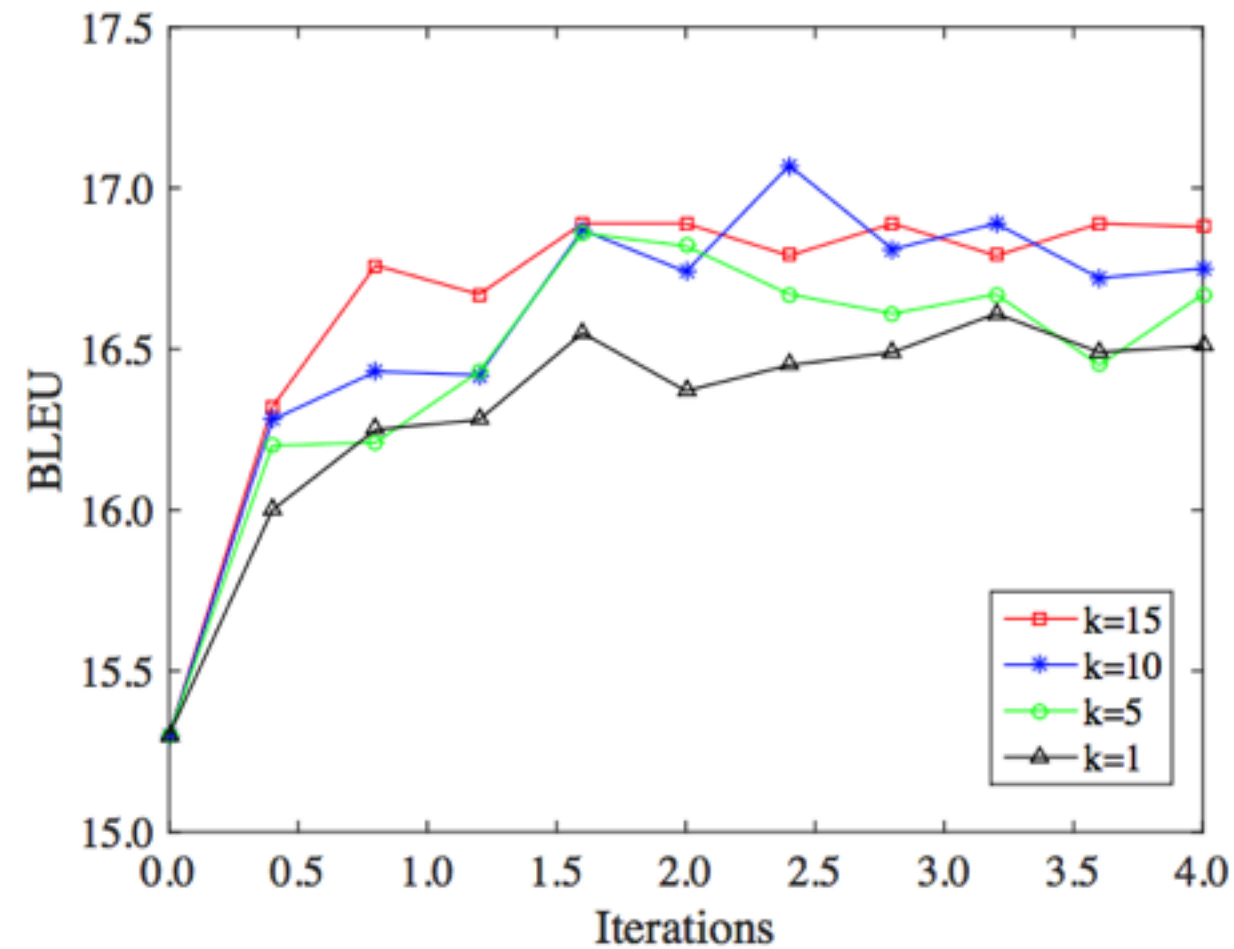
- Monolingual corpora is an important resource for neural machine translation.
- We have proposed a semi-supervised approach to training bidirectional neural machine translation models for exploiting monolingual corpora.
- As our method is sensitive to the OOVs present in monolingual corpora, we plan to integrate Jean et al. (2015)'s technique on using very large vocabulary into our approach.

Thank You !

Effect of Sample Size

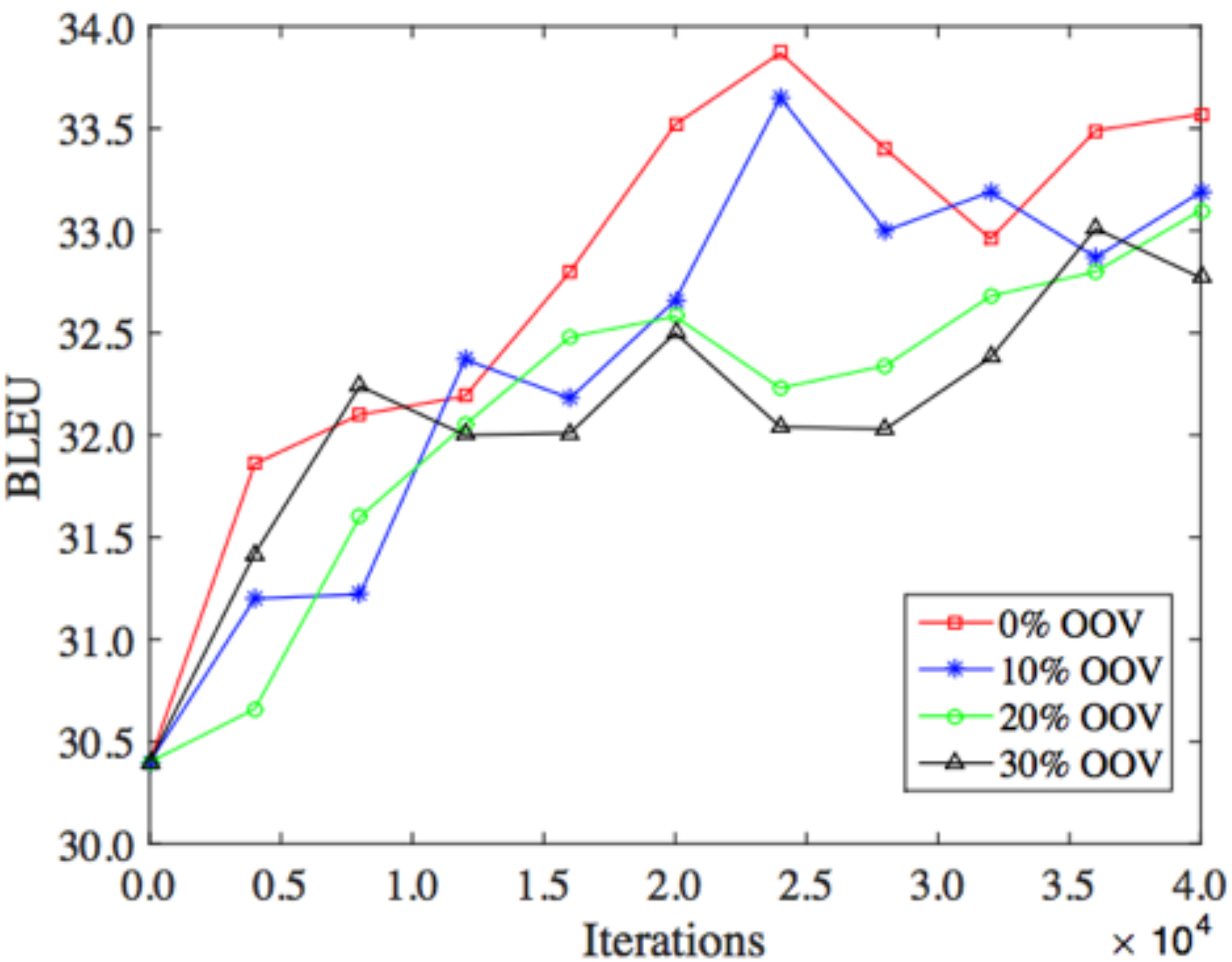


ZH-EN

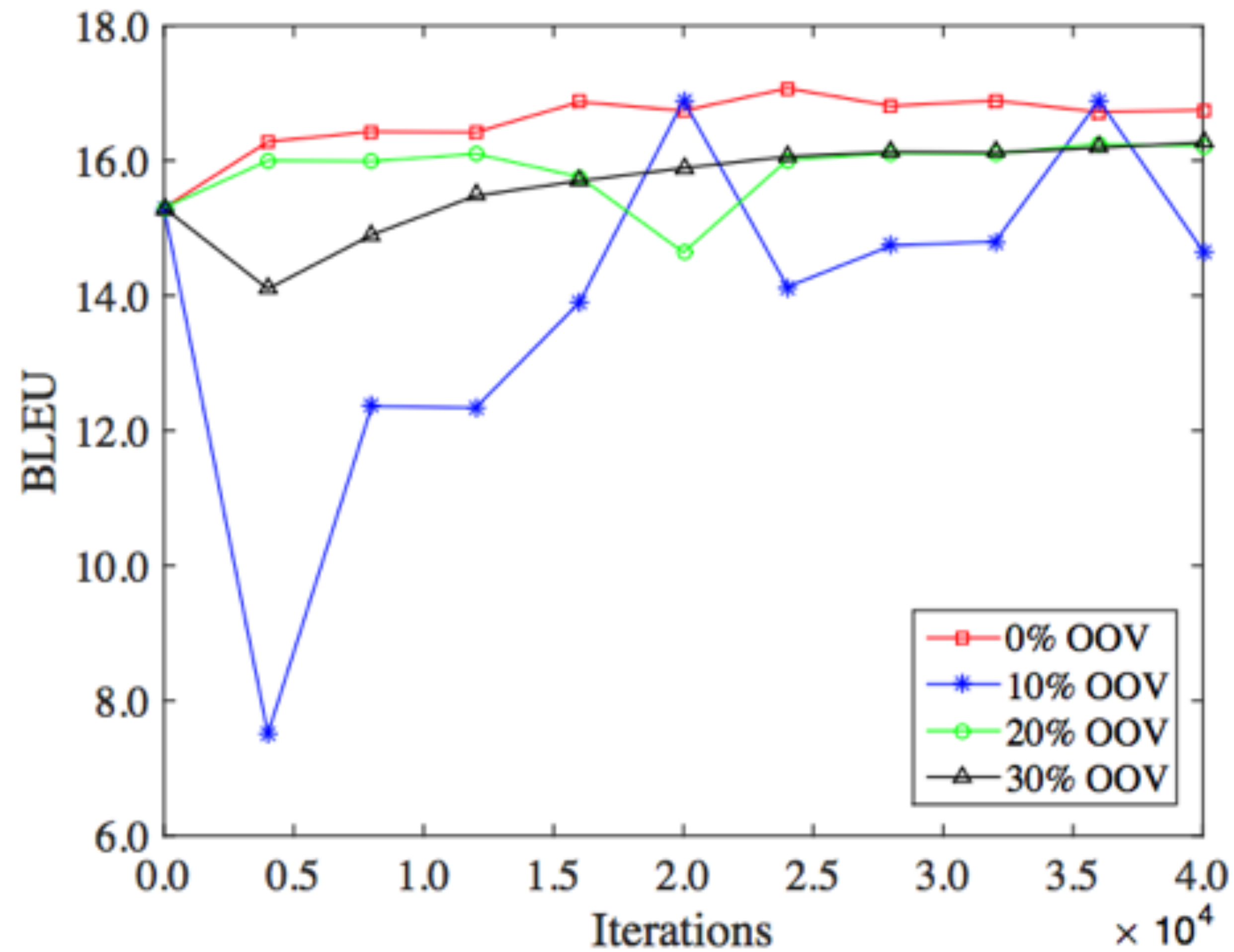


EN-ZH

Effect of OOV ratio



ZH-EN



EN-ZH