



# Joint Training for Neural Machine Translation Models with Monolingual Data

<sup>1</sup>Zhirui Zhang, <sup>2</sup>Shujie Liu, <sup>2</sup>Mu Li, <sup>2</sup>Ming Zhou, <sup>1</sup>Enhong Chen

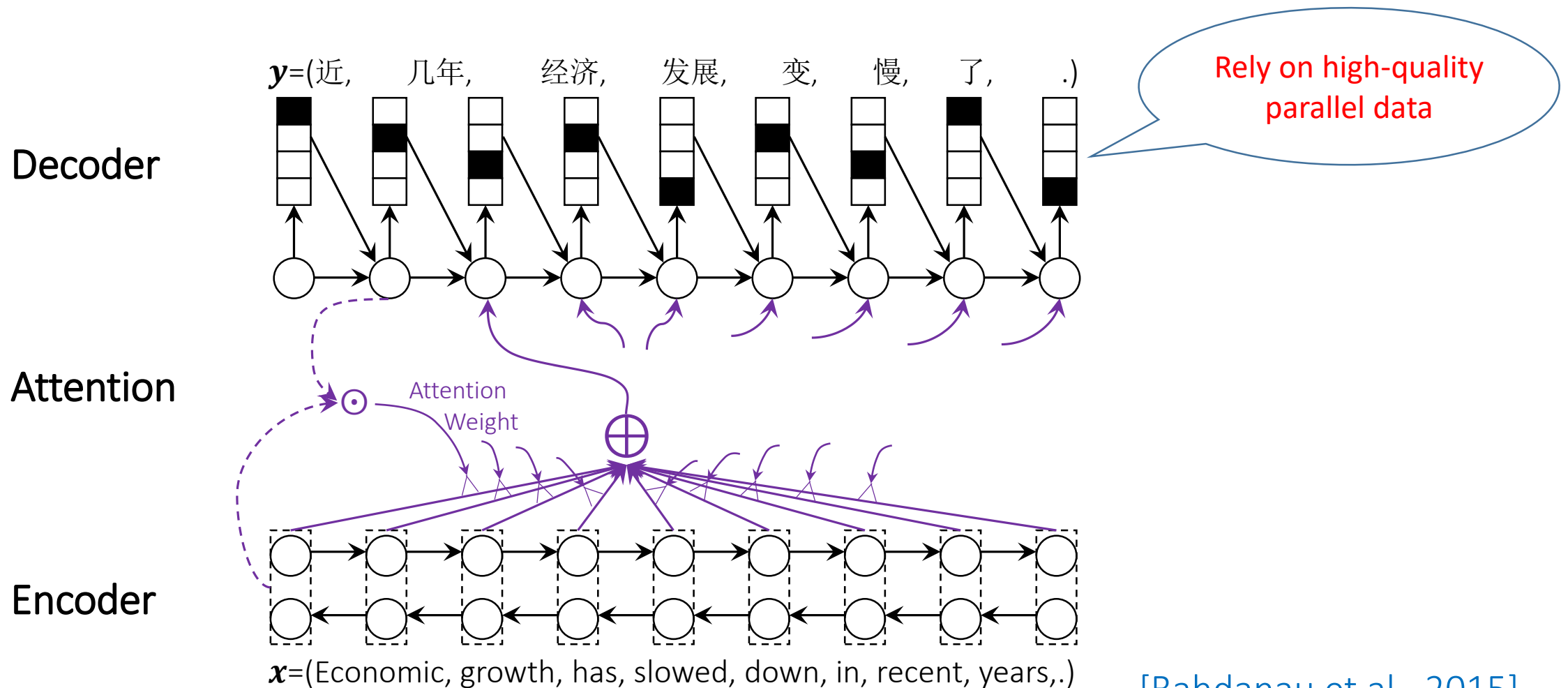
<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Microsoft Research

<sup>1</sup>zrustc11@gmail.com cheneh@ustc.edu.cn

<sup>2</sup>{shujliu, muli, mingzhou}@microsoft.com

# Neural Machine Translation (NMT)



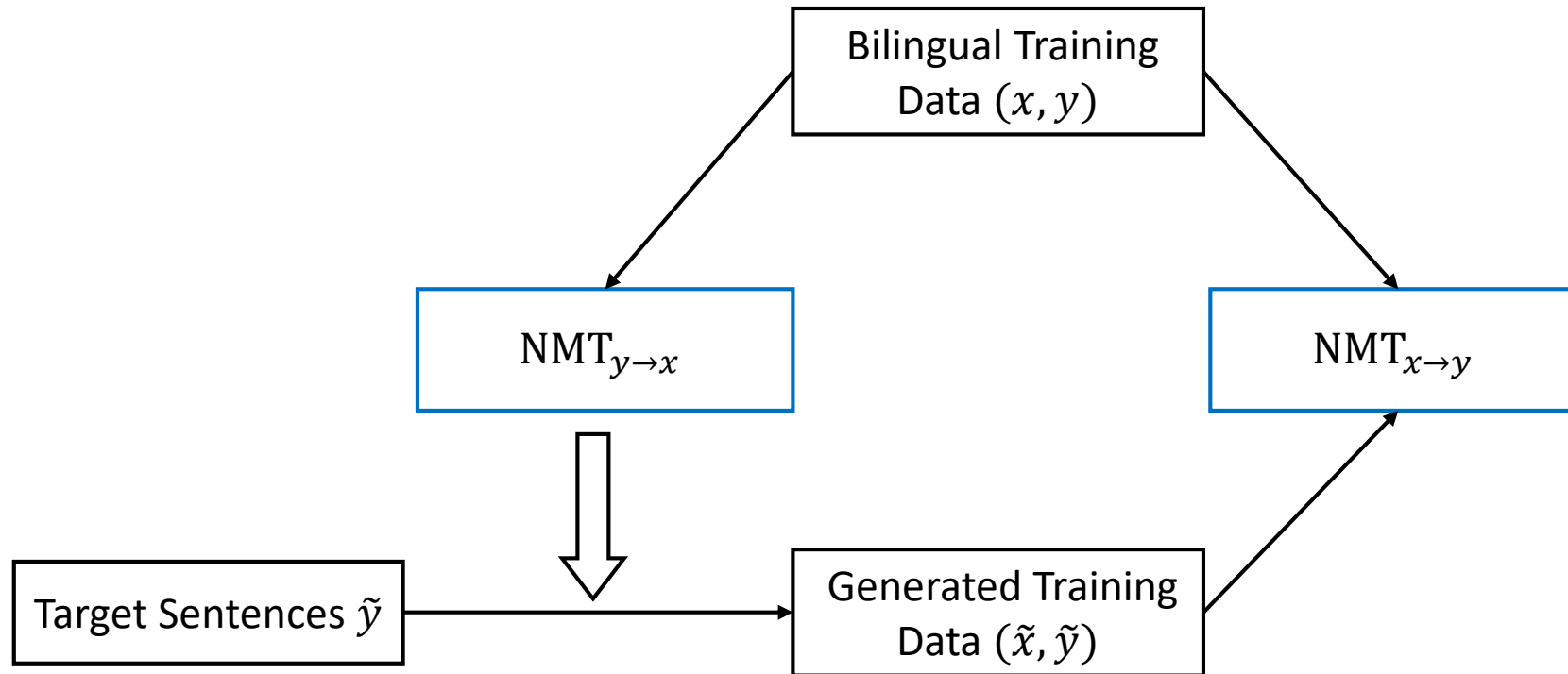
[Bahdanau et al., 2015]

# Monolingual Training Data

- Why monolingual data for NMT?
  - Easy to collect and more diverse
  - More training data
- Solutions
  - Deep fusion (Gulcehre et al., 2015)
  - Auto-Encoder (Cheng et al., 2016)
  - Back-Translation (Sennrich et al., 2016)
  - ...

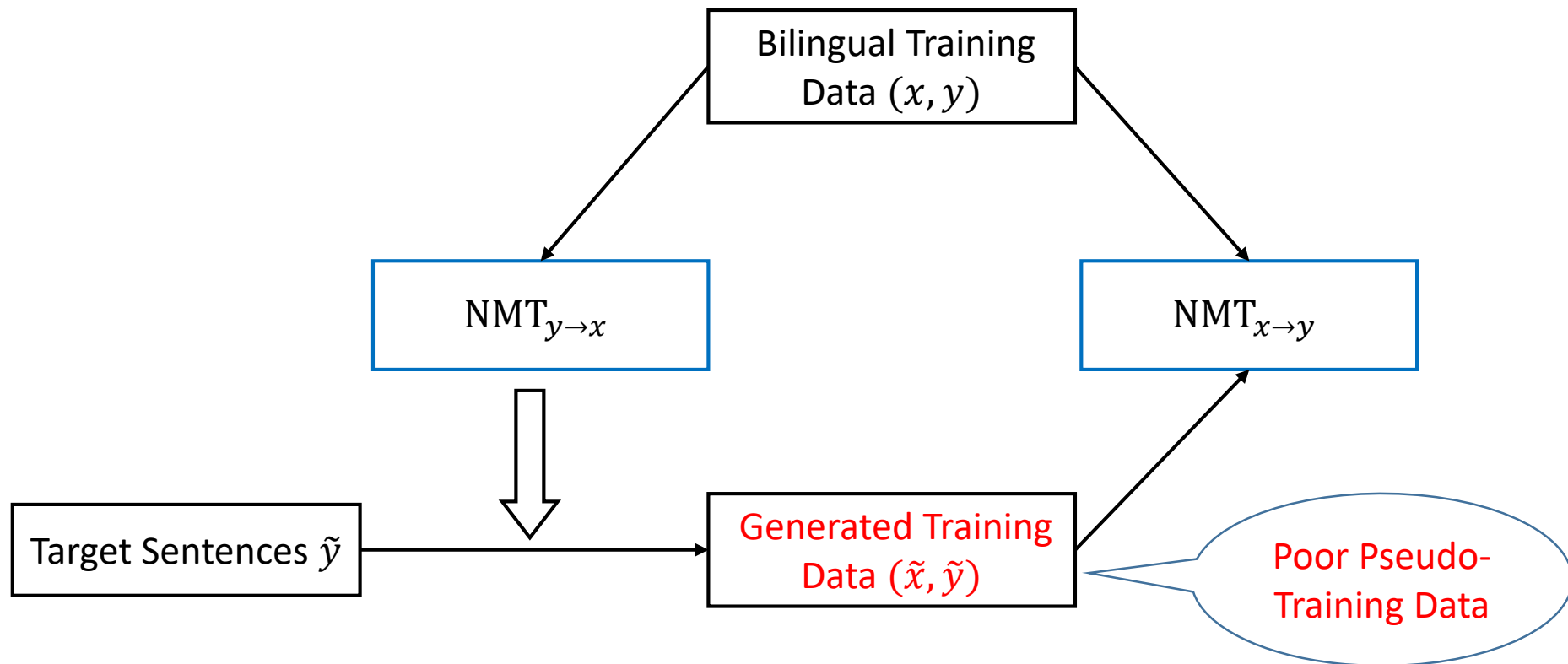
# Back-Translation (Sennrich et al., 2016)

- Back-translate target monolingual data into source language



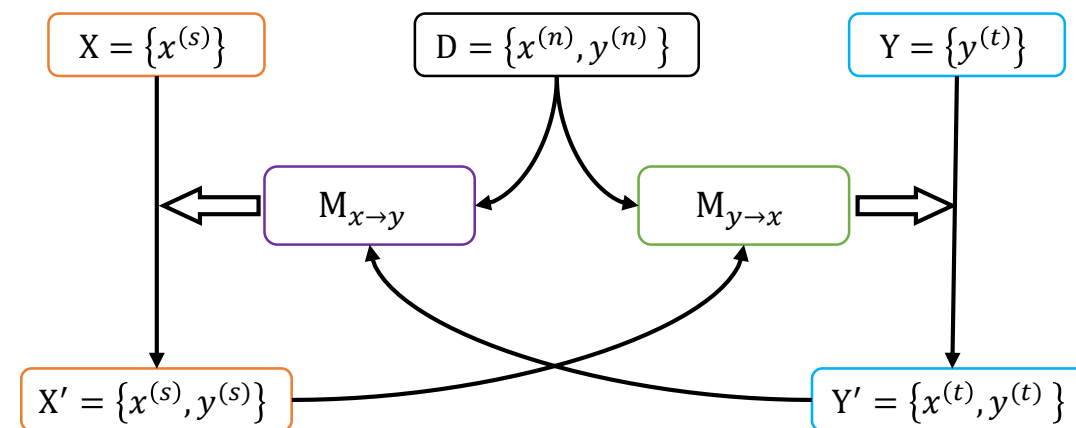
# Back-Translation (Sennrich et al., 2016)

- Back-translate target monolingual data into source language



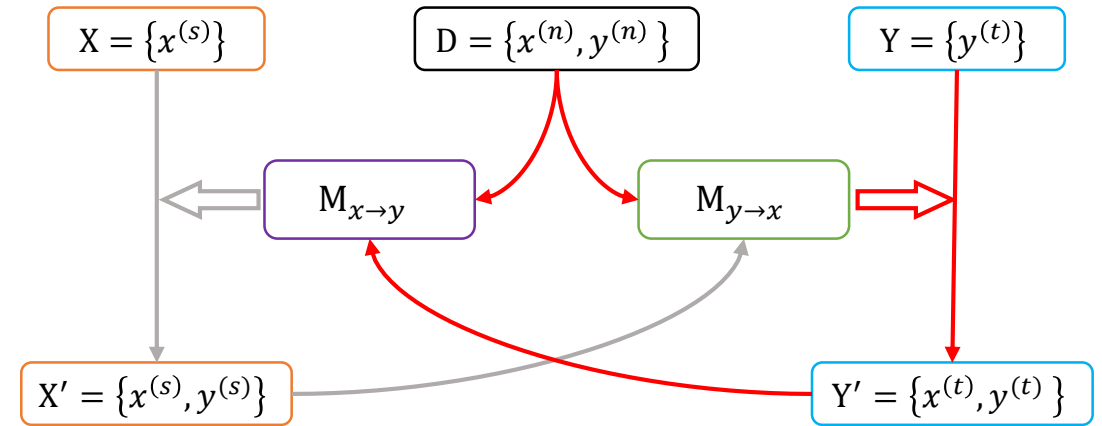
# Our Approach

- We extend this task setting to a paired one:
  - Jointly optimizing source-to-target NMT model  $M_{x \rightarrow y}$  and target-to-source NMT model  $M_{y \rightarrow x}$  with the aid of monolingual data from both source language  $X$  and target language  $Y$



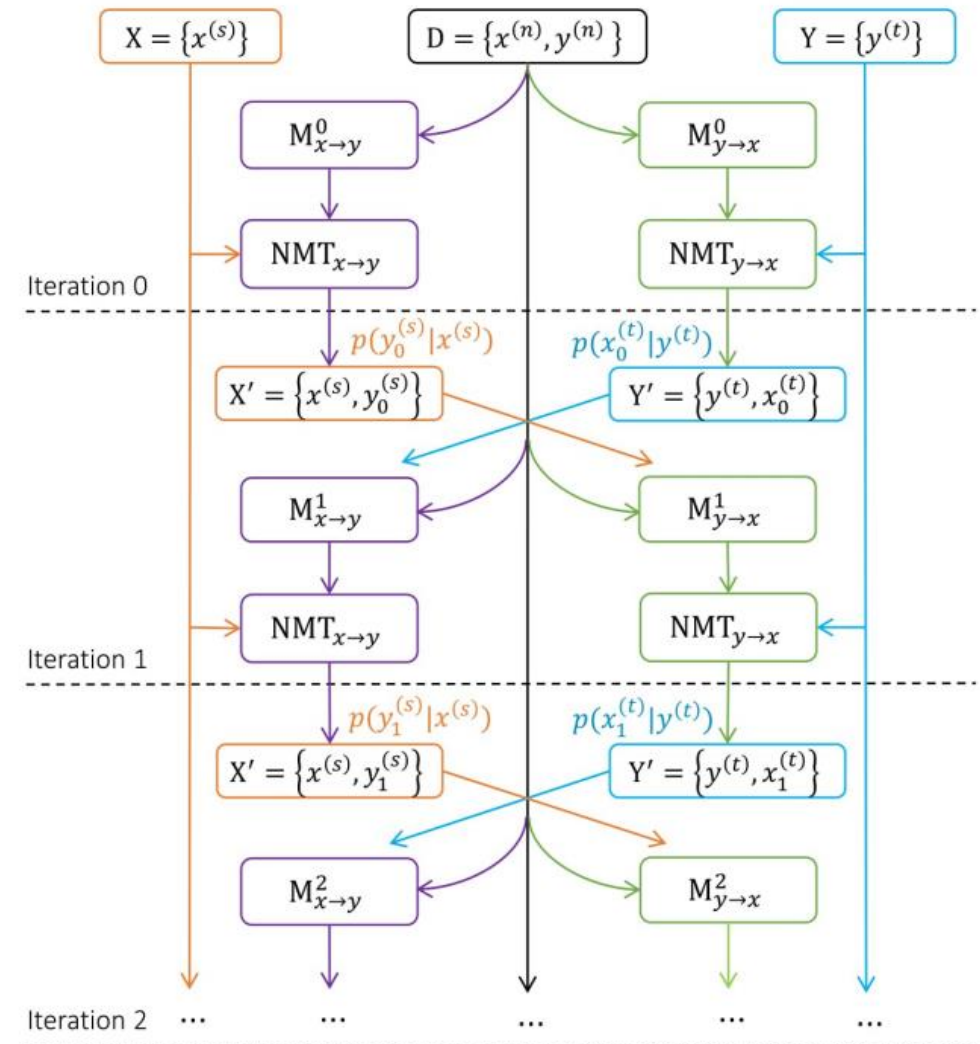
# Our Approach

- We extend this task setting to a paired one:
  - Jointly optimizing source-to-target NMT model  $M_{x \rightarrow y}$  and target-to-source NMT model  $M_{y \rightarrow x}$  with the aid of monolingual data from both source language  $X$  and target language  $Y$
- Back-Translation is a special case:
  - Using  $Y$  and  $M_{y \rightarrow x}$  to improve  $M_{x \rightarrow y}$



# Joint Training for Paired NMT Models

- **Iteration 0:** pre-train two direction models  $M_{x \rightarrow y}^0$  and  $M_{y \rightarrow x}^0$  with bilingual data  $D = \{x^{(n)}, y^{(n)}\}$
- **Iteration 1:** two NMT systems based on  $M_{x \rightarrow y}^0$  and  $M_{y \rightarrow x}^0$  are used to translate monolingual data  $X = \{x^{(s)}\}$  and  $Y = \{y^{(t)}\}$ ; two synthetic training data sets  $X'$  and  $Y'$  combined with bilingual data  $D$  are used to train  $M_{x \rightarrow y}^1$  and  $M_{y \rightarrow x}^1$  respectively.
- **Iteration 2:** repeat the above process





# Training Objective

- Given parallel corpus  $D = \{(x^{(n)}, y^{(n)})\}_{n=1}^N$  and monolingual corpus in target language  $Y = \{y^{(t)}\}_{t=1}^T$ , the semi-supervised training objective as follows:

$$L(\theta_{x \rightarrow y}) = \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}) + \sum_{t=1}^T \log p(y^{(t)})$$

- Introduce source translations as hidden states for monolingual target sentences


$$\begin{aligned} \log p(y^{(t)}) &= \log \sum_x Q(x) \frac{p(x, y^{(t)})}{Q(x)} \geq \sum_x Q(x) \log \frac{p(x, y^{(t)})}{Q(x)} \text{ (Jensen's inequality)} \\ &= \sum_x [Q(x) \log p(y^{(t)} | x) - KL(Q(x) || p(x))] \end{aligned}$$

- The equal condition is  $Q(x) = p^*(x | y^{(t)})$

# Training Objective

- The semi-supervised training objective can be simplified as

$$L(\theta_{x \rightarrow y}) = \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}) + \sum_{t=1}^T \sum_x p(x | y^{(t)}) \log p(y^{(t)} | x)$$



The weight of the  
pseudo sentence pairs

# Training Objective

- The semi-supervised training objective can be simplified as

$$L(\theta_{x \rightarrow y}) = \sum_{n=1}^N \log p(y^{(n)} | x^{(n)}) + \sum_{t=1}^T \sum_x p(x | y^{(t)}) \log p(y^{(t)} | x)$$

- The inverse direction translation is similar

$$L(\theta_{y \rightarrow x}) = \sum_{n=1}^N \log p(x^{(n)} | y^{(n)}) + \sum_{s=1}^S \sum_y p(y | x^{(s)}) \log p(x^{(s)} | y)$$

- The overall training objective is the sum of likelihood in both directions

$$L(\theta) = L(\theta_{x \rightarrow y}) + L(\theta_{y \rightarrow x})$$

# Experiment

- Evaluate our approach on Chinese-English and English-German translation tasks
- NMT systems
  - **RNNSearch**: attention-based NMT system (Bahdanau et al., 2015)
  - **RNNSearch+M**: back-translation method (Sennrich et al., 2016)
  - **SS-NMT**: auto-encoder method (Cheng et al. 2016)
  - **JT-NMT**: our method

# Experiment: Chinese-English Translation

- Dataset
  - Bilingual data: 2.6M sentence pairs from LDC corpus
  - Monolingual data: 8M Chinese and English sentences extracted from Gigaword
  - Development data: NIST2006
  - Test data: NIST2003, NIST2005, NIST2008, NIST2012
- Result

Direction	System	NIST2006	NIST2003	NIST2005	NIST2008	NIST2012	Average
C→E	RNNSearch	38.61	39.39	38.31	30.04	28.48	34.97
	RNNSearch+M	40.66	43.26	41.61	32.48	31.16	37.83
	SS-NMT	41.53	44.03	42.24	33.40	31.58	38.56
	JT-NMT	<b>42.56</b>	<b>45.10</b>	<b>44.36</b>	<b>34.10</b>	<b>32.26</b>	<b>39.67</b>
E→C	RNNSearch	17.75	18.37	17.10	13.14	12.85	15.84
	RNNSearch+M	21.28	21.19	19.53	16.47	15.86	18.87
	SS-NMT	21.62	22.00	19.70	17.06	16.48	19.37
	JT-NMT	<b>22.56</b>	<b>22.98</b>	<b>20.95</b>	<b>17.62</b>	<b>17.39</b>	<b>20.30</b>

Table 1: Case-insensitive BLEU scores (%) on Chinese↔English translation. The “Average” denotes the average BLEU score of all datasets in the same setting. The “C” and “E” denote Chinese and English respectively.

# Experiment: English-German Translation

- Dataset
  - Bilingual data: 4.5M sentence pairs from WMT 2014 training corpus
  - Monolingual data: 8M German and English sentences extracted from "New Crawl: articles from 2012"
  - Development data: the concatenation of news-test 2012 and news-test 2013
  - Test data: news-test2014

- Result

System	Architecture	E→D	D→E
Jean et al. (2015)	Gated RNN with search + PosUnk	18.97	-
Jean et al. (2015)	Gated RNN with search + PosUnk + 500K vocabs	19.40	-
Shen et al. (2016)	Gated RNN with search + PosUnk + MRT	20.45	-
Luong, Pham, and Manning (2015)	LSTM with 4 layers + dropout + local att. + PosUnk	20.90	-
RNNSearch	Gated RNN with search + BPE	19.78	24.91
RNNSearch+M	Gated RNN with search + BPE + monolingual data	21.89	26.81
SS-NMT	Gated RNN with search + BPE + monolingual data	22.64	27.30
JT-NMT	Gated RNN with search + BPE + monolingual data	<b>23.60</b>	<b>27.98</b>

Table 2: Case-sensitive BLEU scores (%) on English↔German translation. "PosUnk" denotes Luong et al. (2015)'s technique of handling rare words. "MRT" denotes minimum risk training proposed in Shen et al. (2016). "BPE" denotes Byte Pair Encoding proposed by Sennrich, Haddow, and Birch (2016b) for word segmentation. The "D" and "E" denote German and English respectively.

# Effect of Joint Training

- Performance on every iteration

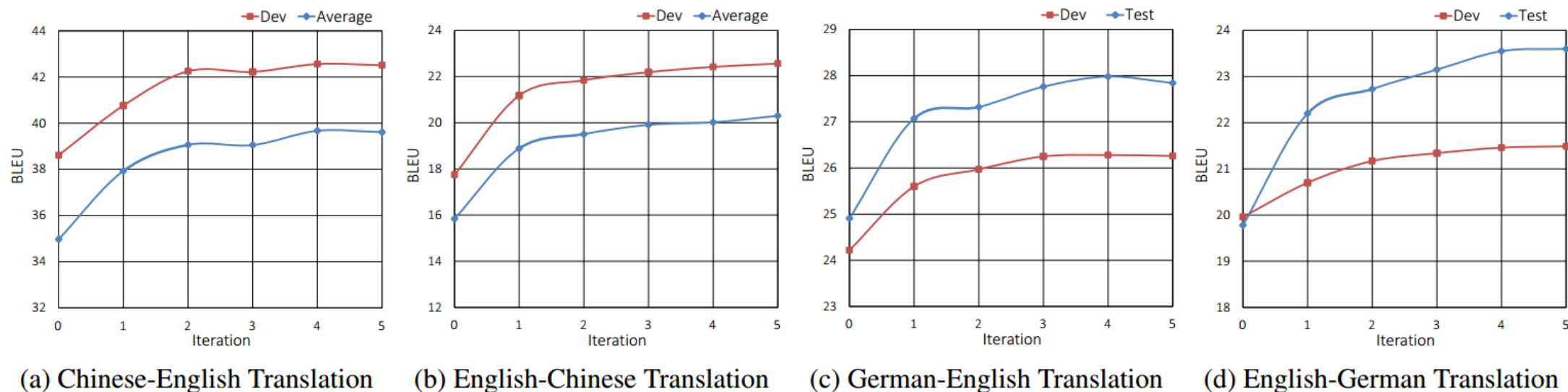


Figure 2: BLEU scores (%) on Chinese $\leftrightarrow$ English and English $\leftrightarrow$ German validation and test sets for JT-NMT during training process. “Dev” denotes the results of validation datasets, while “Test” denotes the results of test datasets.

# Effect of Joint Training

Monolingual	当 终场 哨声 响起 , 意大利 首都 罗马 沸腾 了 。 <i>dang zhongchang shaosheng xiang qi , yidali shoudu luoma feiteng le .</i>
Reference	<b>when the final whistle sounded</b> , the italian capital of rome boiled .
Translation	[Iteration 0]: the italian capital of rome was boiling <b>with the rome</b> .
	[Iteration 1]: the italian capital of rome was boiling <b>with the sound of the end of the door</b> .
	[Iteration 4]: <b>when the final whistle sounded</b> , the italian capital of rome was boiling .

Table 4: Example translations of a Chinese sentence in different iterations.



# Conclusion

- Proposed a new semi-supervised training approach
  - A joint-EM training algorithm
  - Bidirectional models are able to mutually boost their translation performance
- Significant improvements on Chinese-English/English-German tasks
- Future work
  - Extend this method to jointly train multiple NMT systems for 3+ languages

Thanks!