Low Resource Neural Machine Translation

Shujie LIU

Natural Language Computing Group

Microsoft Research Asia

Microsoft

Outline

- Introduction to Machine Translation
- Neural Machine Translation
- Low Resource Neural Machine Translation
- Unsupervised Neural Machine Translation

Microsoft

Introduction to Machine Translation

Microsoft

Introduction to Machine Translation

- **Background of MT**
- Methods of MT
- Evaluation of MT

全球现存6909种语言在使用

普通话	885,000,000
SPANISH	332,000,000
ENGLISH	322,000,000
BENGALI	189,000,000
HINDI	182,000,000
PORTUGUESE	170,000,000
RUSSIAN	170,000,000
JAPANESE	125,000,000
GERMAN	98,000,000
吴 语(China)	77,175,000
JAVANESE	75,500,800
KOREAN	75,000,000
FRENCH	72,000,000
VIETNAMESE	67,662,000
TELUGU	66,350,000
粤语 (China)	66,000,000
MARATHI	64,783,000
TAMIL	63,075,000

TURKISH	59,000,000
URDU	58,000,000
MIN NAN (China)	49,000,000
晋语 (China)	45,000,000
GUJARATI	44,000,000
POLISH	44,000,000
ARABIC	42,500,000
UKRAINIAN	41,000,000
ITALIAN	37,000,000
XIANG (China)	36,015,000
MALAYALAM	34,022,000
客家话(China)	34,000,000
KANNADA	33,663,000
ORIYA	31,000,000
PANJABI	30,000,000
SUNDA	27,000,000

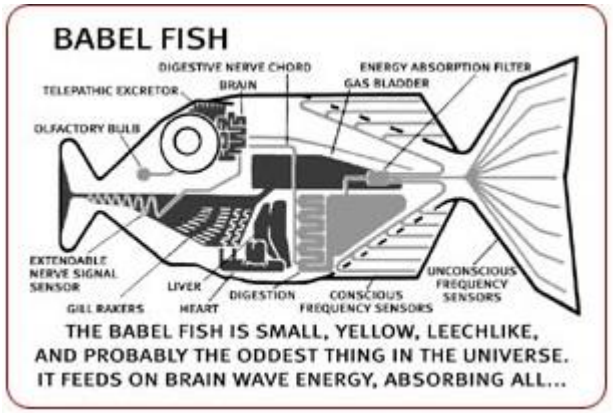
Source: Ethnologue

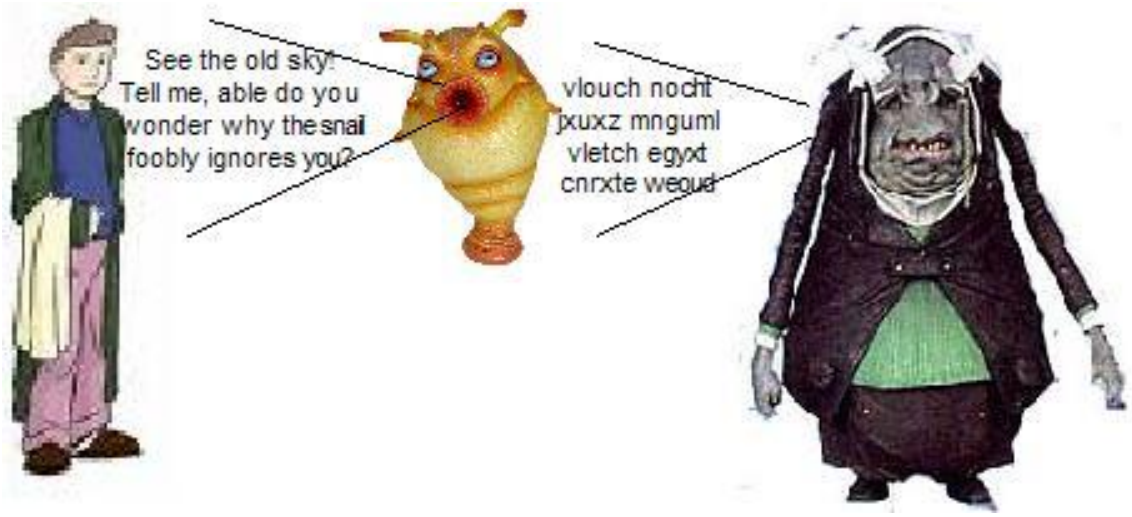
人类的梦想

- 《星际迷航》中的**万能翻译器**
  - 星际联邦广泛使用了宇宙翻译器。宇宙翻译器可以将所有的已知语言翻译为听者所懂的语言，对其余的未知语言亦可通过对简短几句话的分析而进行转换。“进取号”能与整个银河系中的智能生物沟通，当遇到讲不同语言的外星生物时，他们会借助“宇宙翻译机”的帮助。

人类的梦想

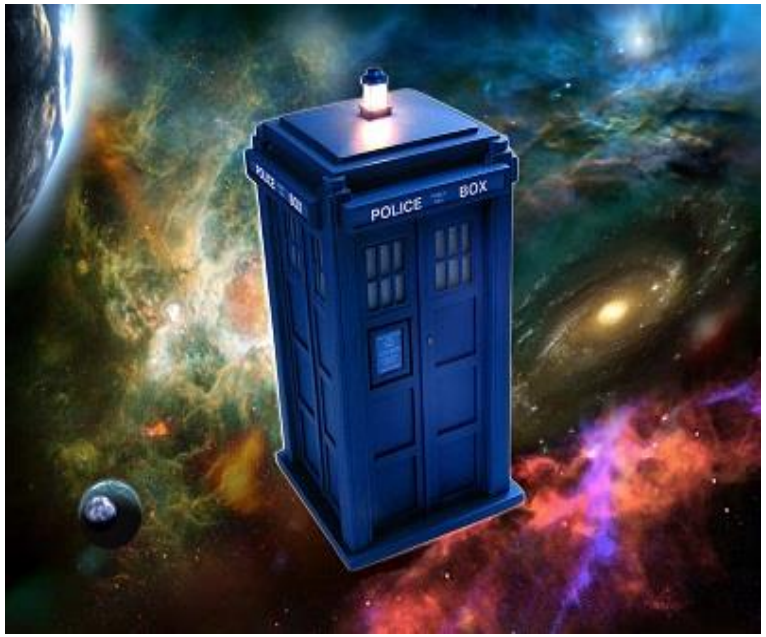
- 《银河系漫游指南》中的**巴别鱼**
  - “巴别鱼，”《银河系漫游指南》轻轻朗读着，“体型很小，黄色，外形像水蛭，很可能是宇宙中最奇异的事物。它靠接收脑电波的能量为生....所有这些过程的实际效果就是，如果你把一条巴别鱼塞进耳朵，你就能立刻理解以任何形式的语言对你说的任何事情。你所听到的解码信号就是巴别鱼向你的思想提供的脑电波矩阵。”





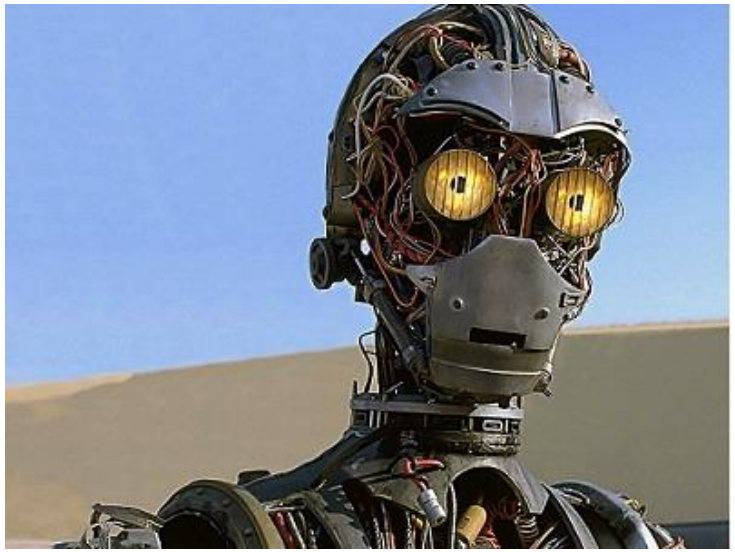
人类的梦想

- 《神秘博士》中的**TARDIS**
  - Time and Relative Dimension in Space
  - TARDIS的翻译功能可以让靠近tardis的人自动把看到的听到的文字语言自动翻译为自己听得懂的语言



人类的梦想

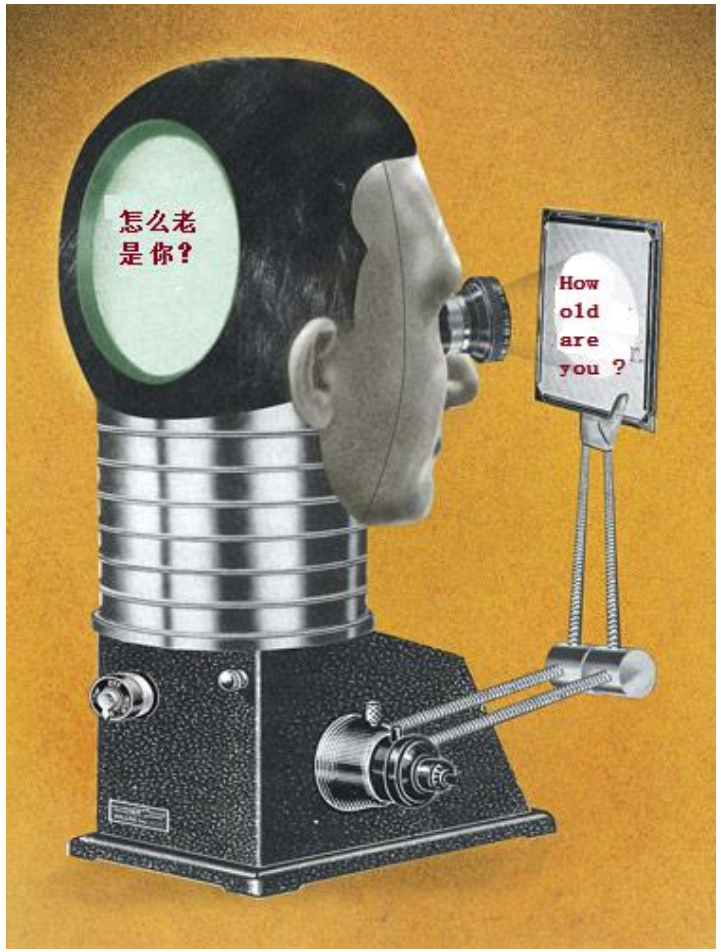
- 《星球大战》中的**斯瑞皮欧**
  - 作为一个神经质的、多愁善感的礼仪机器人，C-3PO经过多次的改造，他更成为一部精通六百万种沟通方式、懂得各地风俗的金色机器人。他不单只是像其他翻译机器人只是翻译，他也有人类的思想。



人类的梦想

- 《遥远星际》中的**翻译微生物**
  - 在电视剧《遥远星际》中，John Crichton被注射了一种称为翻译微生物的细菌，该微生物起到了万能翻译器的作用。翻译微生物克隆了寄主的脑干，并翻译听到的任何语言，并将翻译后的结果直接传给寄主的大脑。该翻译微生物有的时候并不会翻译一些俚语，而只能将其做简单的字面翻译。

What is Machine Translation



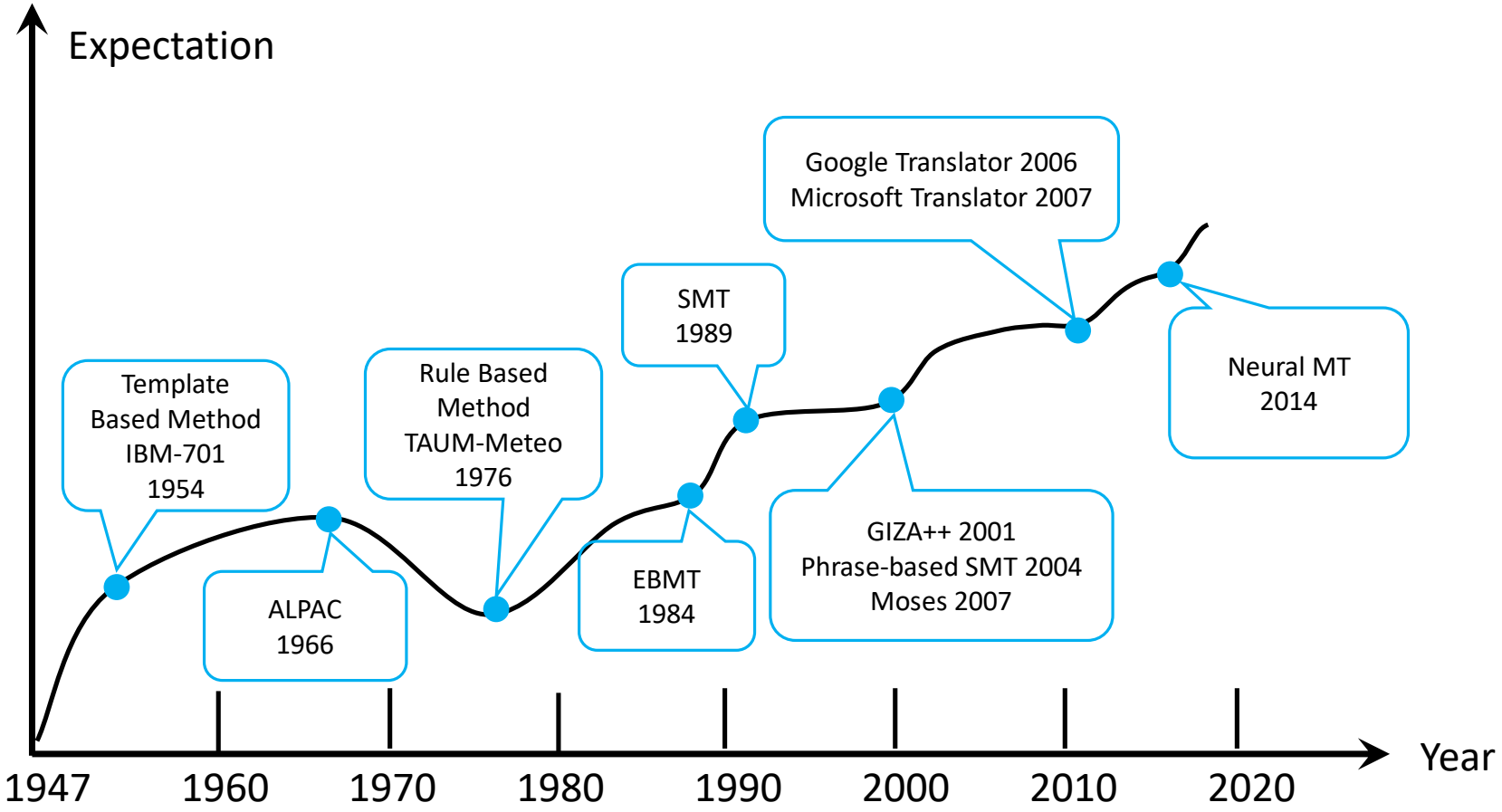
It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the Chinese code. If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

----------Warren Weaver, 1947

What is Machine Translation

- Machine Translation
  - Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

Roadmap of Machine Translation



Microsoft

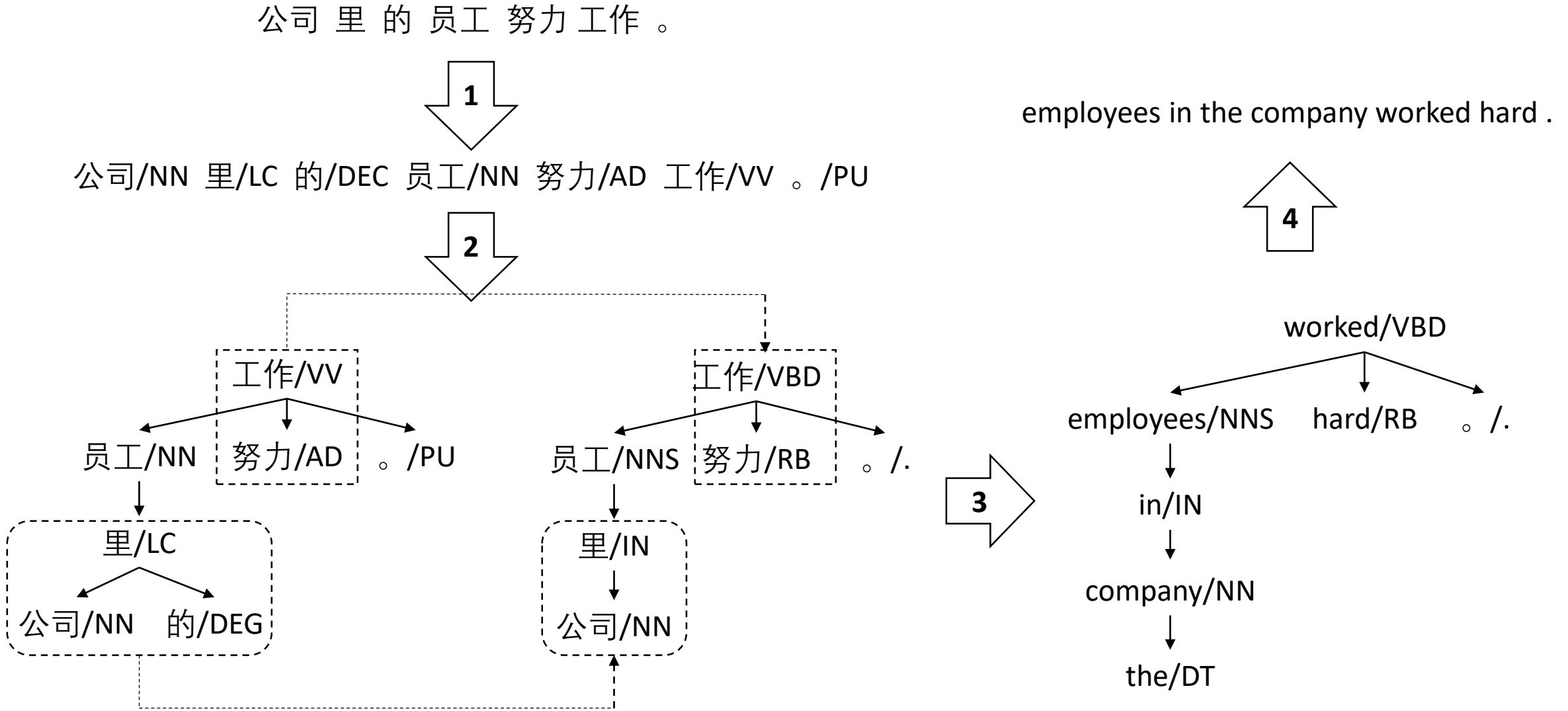
Introduction to Machine Translation

- Background of MT
- **Methods of MT**
- Evaluation of MT

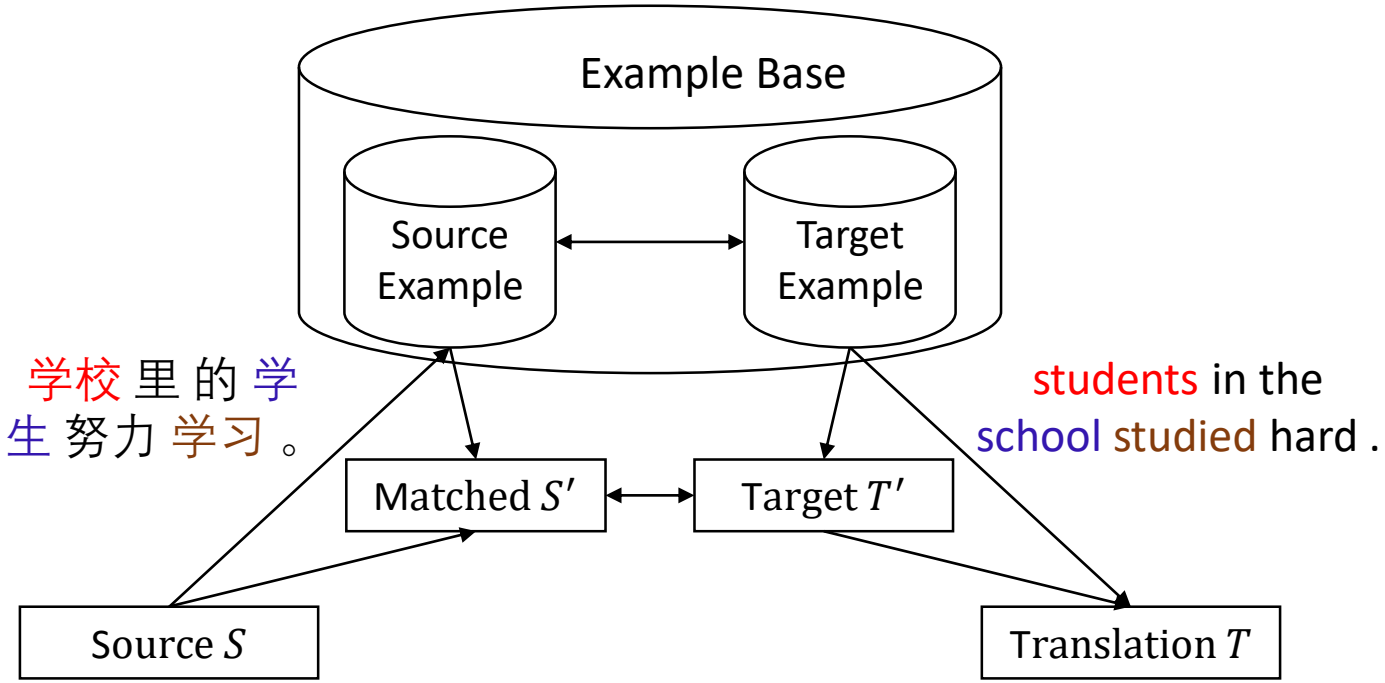
Machine Translation

- Modern Machine Translation
  - In 1946, Warren Weaver firstly proposed the idea of translate the text from one language to another language
- Rule Base Machine Translation
  - Based on rules designed by experts.
- Example Based Machine Translation
  - Based on example sentence pairs.
- Statistical Machine Translation
  - Based on statistical models trained with large corpus
- Neural Machine Translation
  - Based on neural network trained with huge corpus

Rule Based Machine Translation



Example based Machine Translation



公司 里 的 员工 努力 工作 。

employees in the company worked hard .

Statistical Machine Translation

- How to translate X to Y? → What is the translation probability of Y given X?
  - Search Space： the space of all the possible translation candidates.
  - Probability Model： How to define the probability of Y given X?

- SMT = linguistic modeling + statistical decision theory
  - Herman Ney
  - Data Driven： Learn translation rules from large bilingual data with human defined features.
  - Statistic Driven： The translation candidate with maximum probability is the best translation.

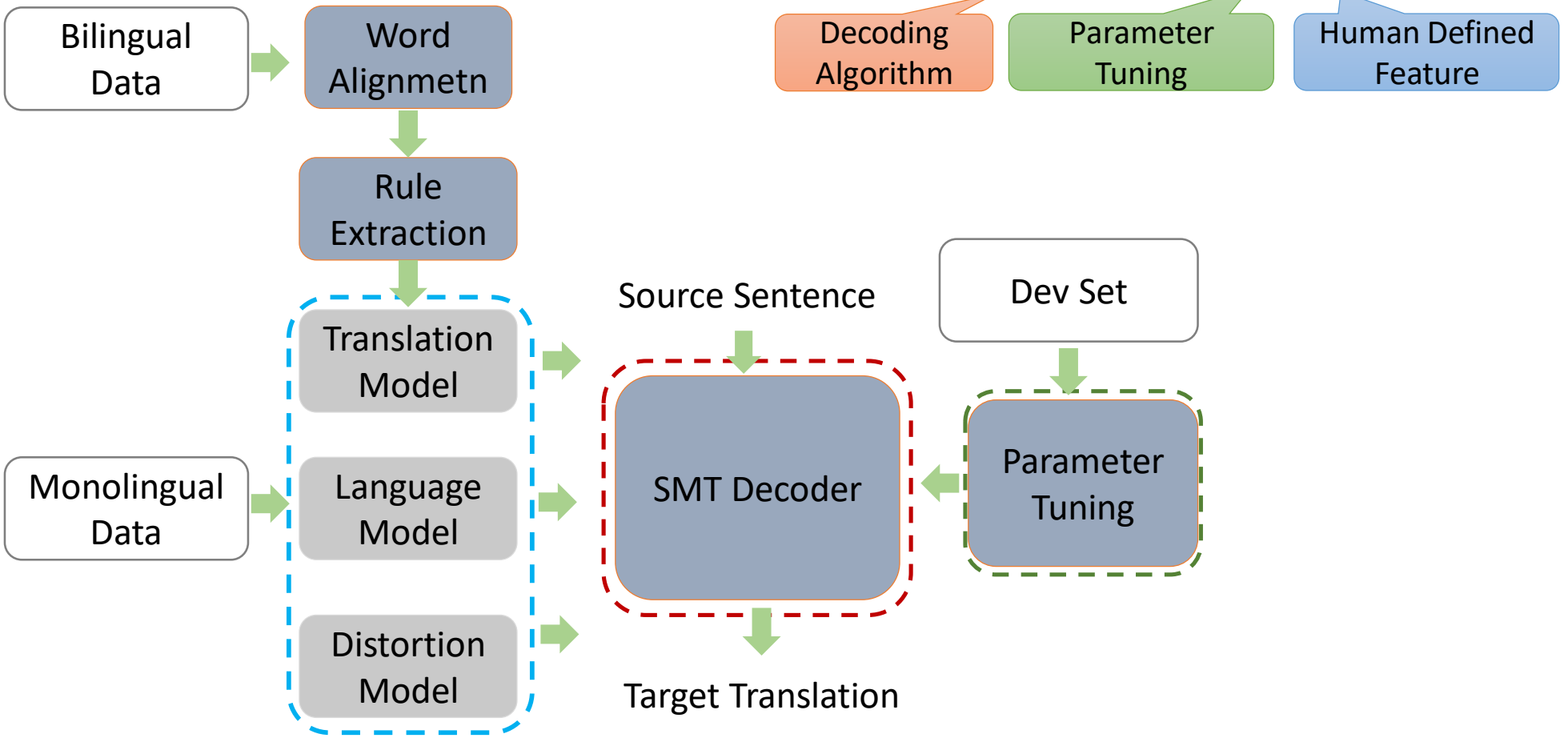
Framework of SMT

$$\hat{e} = \underset{e}{argmax}\{\sum_{m=1}^{M} \lambda_m h_m(e,f)\}$$

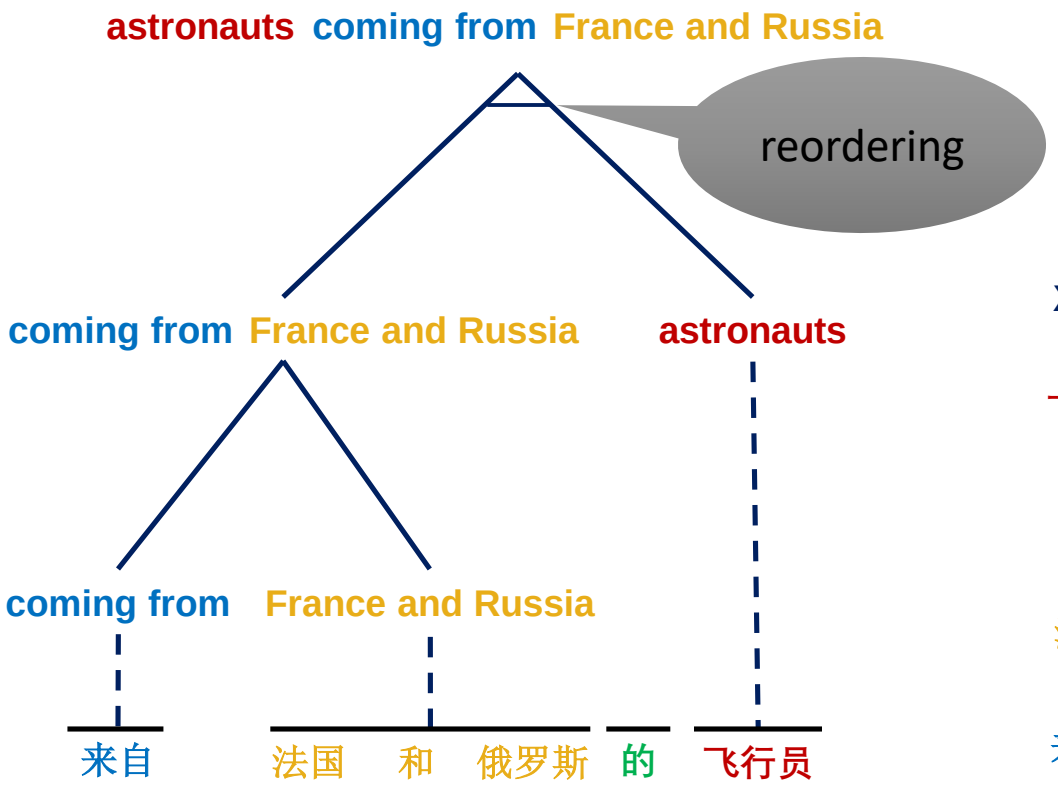
Decoding Algorithm

Parameter Tuning

Human Defined Feature



SMT Decoding： BTG



$X_1\ X_2 \mid\mid\mid X_2\ X_1$

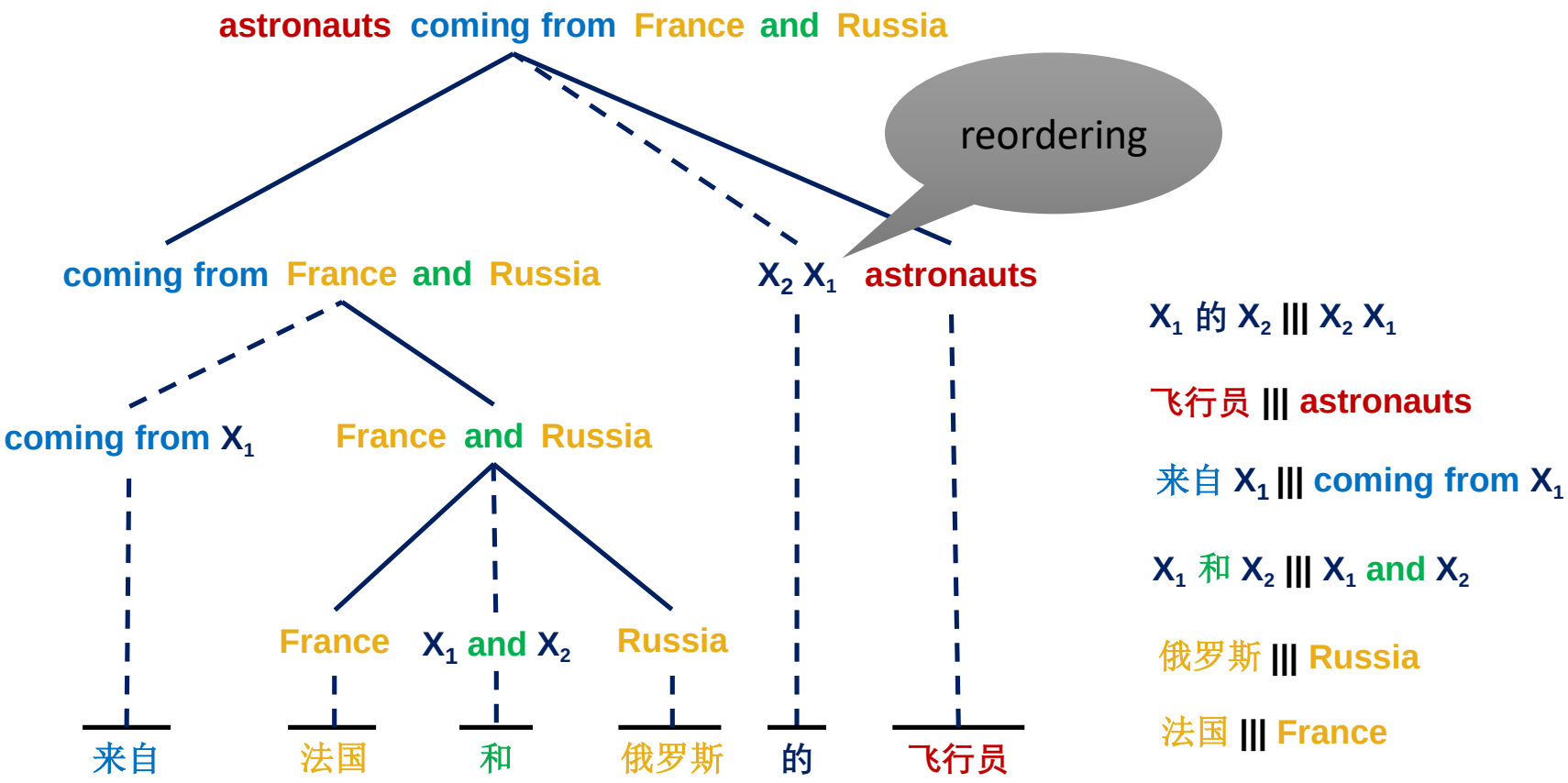
飞行员 ||| astronauts

$X_1\ X_2 \mid\mid\mid X_1\ X_2$

法国 和 俄罗斯 ||| France and Russia

来自 ||| coming from

SMT Decoding: Hiero



X_1 的 X_2 ||| X_2 X_1

飞行员 ||| astronauts

来自 X_1 ||| coming from X_1

X_1 和 X_2 ||| X_1 and X_2

俄罗斯 ||| Russia

法国 ||| France

Microsoft

Neural Machine Translation

- Still a statistical method
  - Try to model the translation probability of Y given X.
- Much larger search space
  - Translation candidates are all the sequences of the target words.
- No need human defined features
  - No translation model, language model and distortion model.
- Will be introduced in detail later.

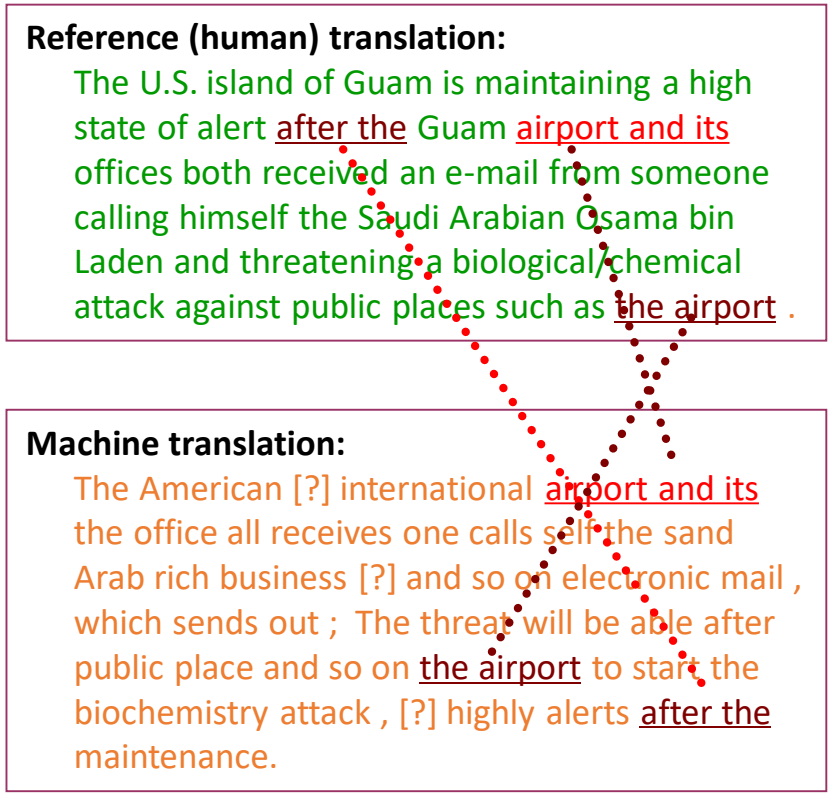
Microsoft

Introduction to Machine Translation

- Background of MT
- Methods of MT
- **Evaluation of MT**

Evaluation of Machine Translation

- Human Evaluation：信达雅
  - 信达： fidelity
    - I cannot agree with you more. → 我不能同意你更多。
  - 雅： fluency
    - How are you ? → 怎么是你?
    - How old are you ? → 怎么老是你?
  - Hight Cost
- Automatic Evaluation
  - Compare the translation with the reference
  - N-gram similarity-based methods
    - NIST
    - BLEU
    - ….
  - Convenient but may not be consistent with human preference



Evaluation of Machine Translation: BLEU

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
  - An n-gram is a sequence of n words.

$$BLEU = BP \bullet \exp\left(\sum_{1}^{N} w_n \log p_n\right)$$

- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

- Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

$$\log BLEU = \min(1 - \frac{r}{c}, 0) + \sum_{1}^{N} w_n \log p_n$$

$$N = 4, w_n = 1/N$$

Papineni, et al. ***BLEU: a Method for Automatic Evaluation of Machine Translation***. ACL, 2002

Evaluation of Machine Translation: BLEU

- Example:
  - **Reference**: “the Iraqi weapons are to be handed over to the army within two weeks”
  - **MT output**: “in two weeks Iraq’s weapons will give army”

- BLEU metric:
  - 1-gram precision 4/8
  - 2-gram precision: 1/7
  - 3-gram precision: 0/6
  - 4-gram precision: 0/5
  - BLEU score = 0

$$BLEU = BP \bullet \exp\left(\sum_{1}^{N} w_n \log p_n\right)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

$$\log BLEU = \min(1-\frac{r}{c},0) + \sum_{1}^{N} w_n \log p_n$$

$$N = 4, w_n = 1/N$$

Evaluation of Machine Translation: BLEU

- Example:
  - **Reference1**: "the Iraqi weapons are to be handed over to the army within two weeks"
  - **Reference2**: "the Iraqi weapons will be surrendered to the army in two weeks"
  - **MT output**: " the the the the"
- Clipping precision counts:
  - Precision count for "the" is clipped at 2: max count of the word in any reference.
  - Modified unigram score will be 2/4

$$BLEU = BP \bullet \exp\left(\sum_{1}^{N} w_n \log p_n\right)$$

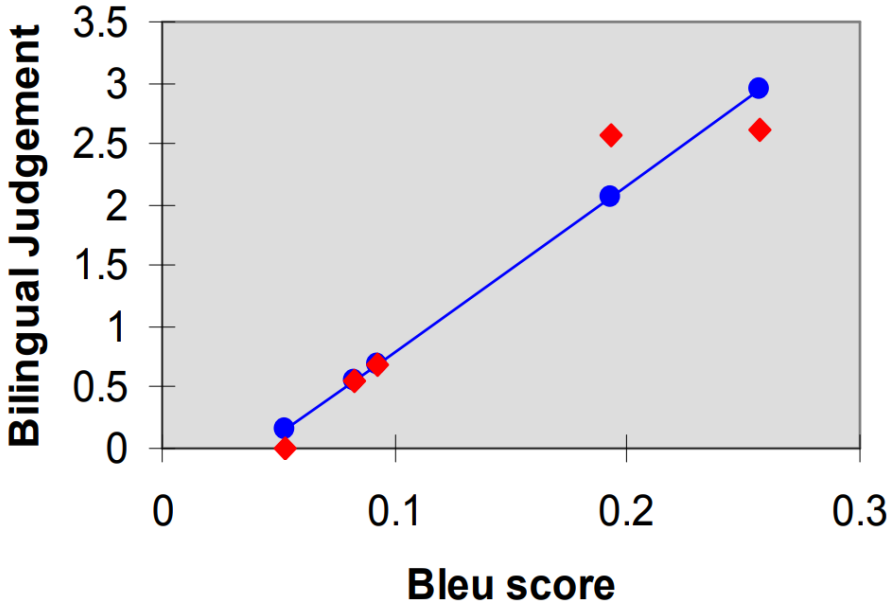
$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

$$\log BLEU = \min(1-\frac{r}{c},0) + \sum_{1}^{N} w_n \log p_n$$

$$N = 4, w_n = 1/N$$

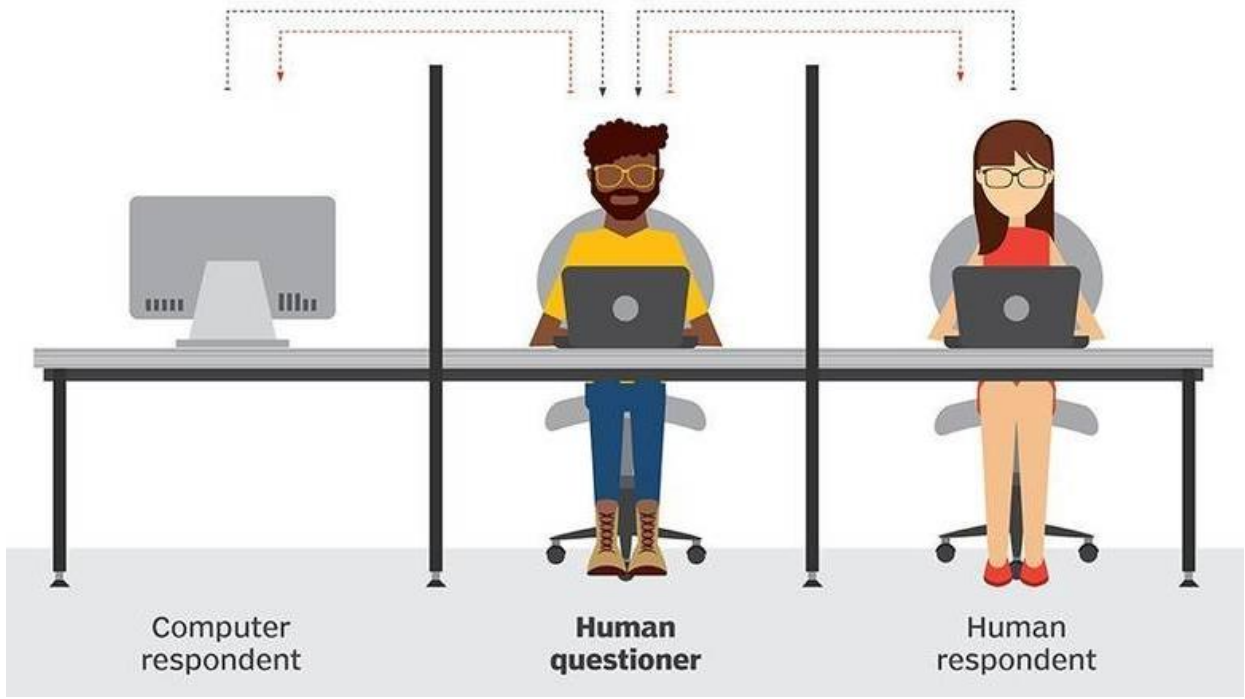
Evaluation of Machine Translation: BLEU

- BLEU vs The Human Evaluation
  - BLEU tends to be high correlated with human judgments when the translation quality is low.
  - Correlation may be not high when the translation quality is good enough.



Papineni et al., 2002

Weak AI: Turing Test vs Chinese Room



http://www.perspecsnews.com/read/tech/HJxOL8vHAM/HkxiJYKBRf
http://cognitivephilosophy.net/consciousness/human-cognition-and-the-chinese-room/

Microsoft

Neural Machine Translation

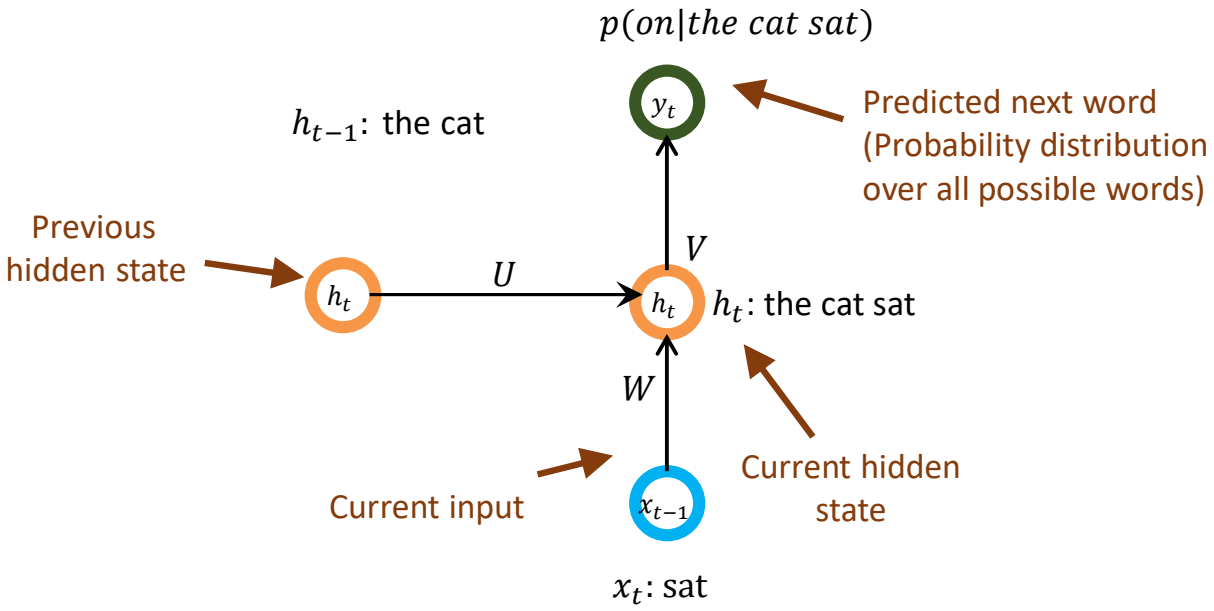
Microsoft

Neural Machine Translation

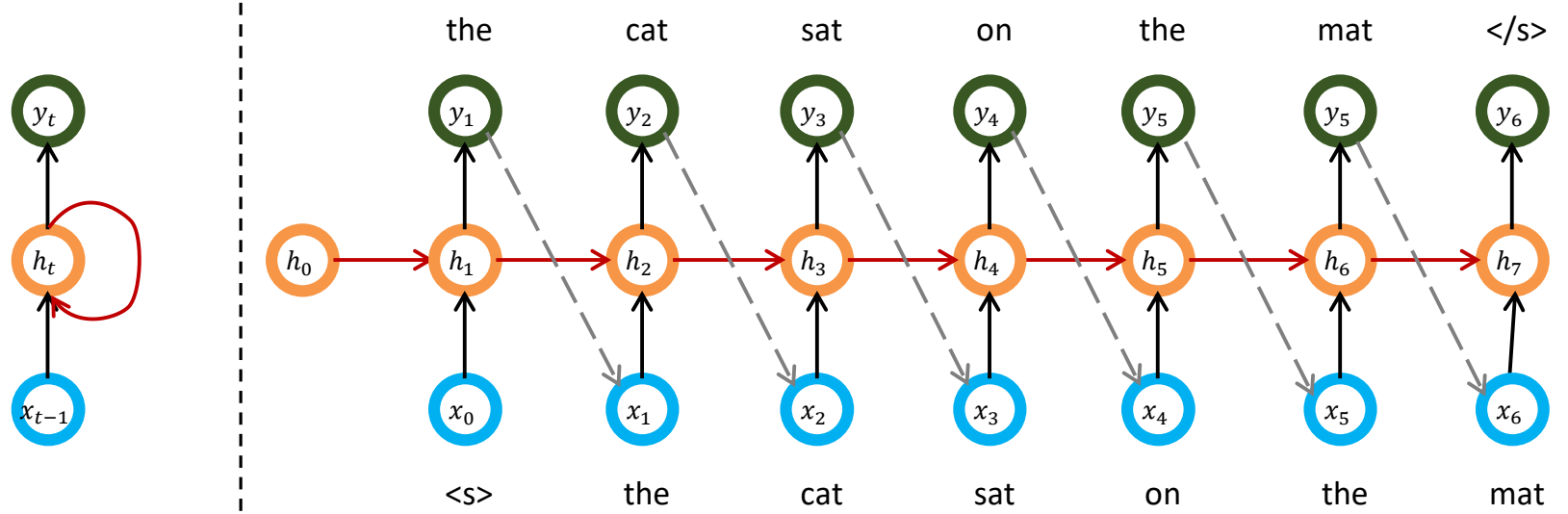
- **Recurrent Neural Networks**
- Naïve Neural Machine Translation
- RNN-based NMT
- Transformer-based NMT
- NMT Training

Recurrent Neural Network

- Inputs: History s_{t-1} at time $t-1$ and input w_t at time t
- Output: History s_t at time t and next input y_t at time t+1



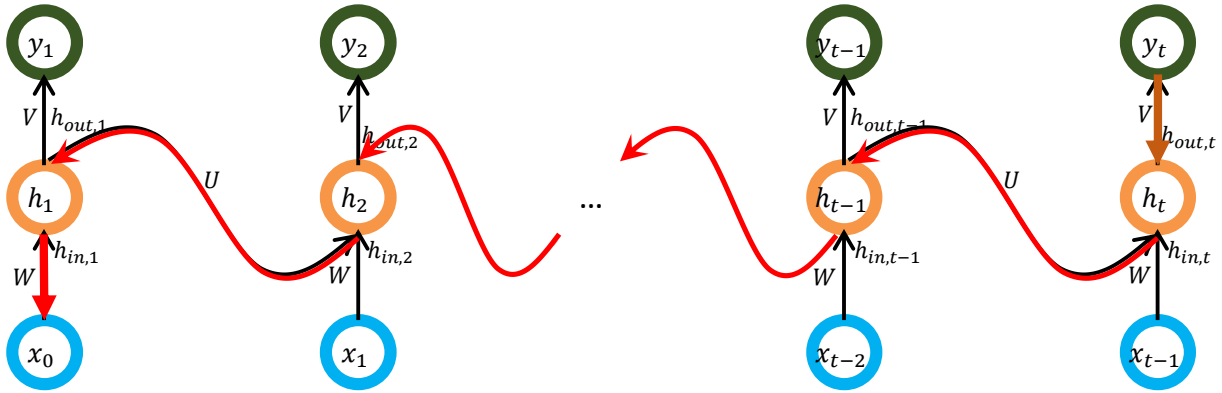
Unfolded RNN



RNN

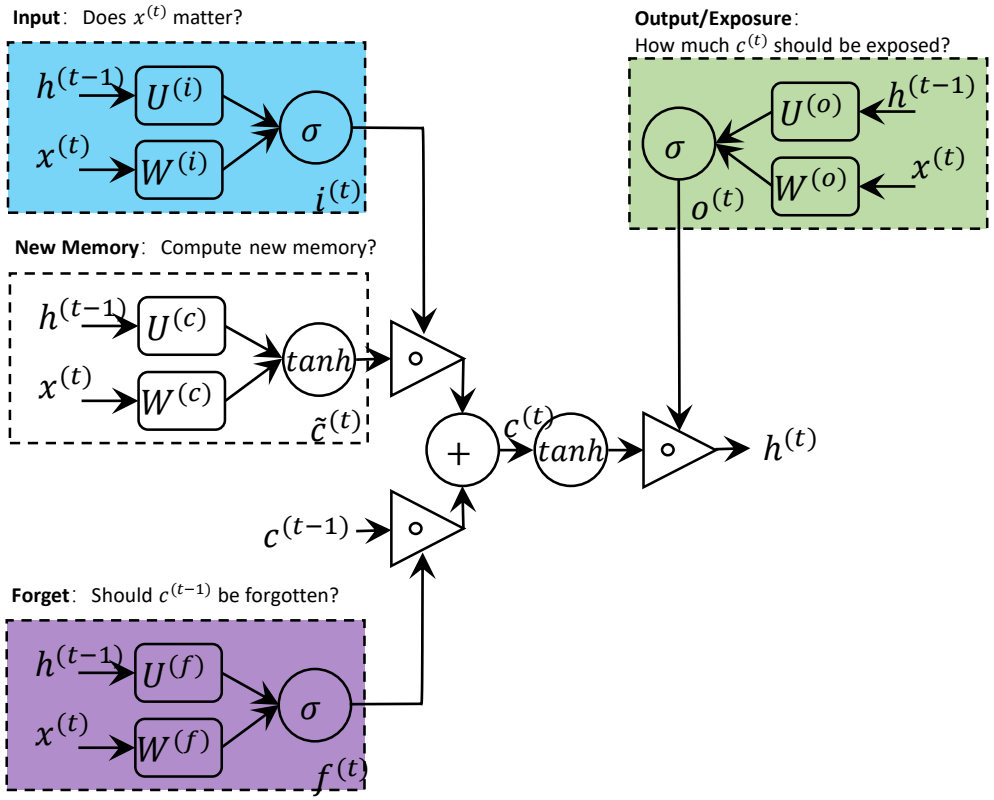
Unfolded RNN for Language Modeling

Vanishing Gradient Problem



$$\delta_{in,1} = \delta_{out,t} \times \frac{\partial h_{out,t}}{\partial h_{in,t}} \times \frac{\partial h_{in,t}}{\partial h_{out,t-1}} \times \frac{\partial h_{out,t-1}}{\partial h_{\mathrm{in},t-1}} \times \cdots \times \frac{\partial h_{in,2}}{\partial h_{out,1}} \times \frac{\partial h_{out,1}}{\partial h_{in,1}}$$

LSTM: Long Short Term Memory



$$i^{(t)} = \sigma(W^{(i)}x^{(t)} + U^{(i)}h^{(t-1)})$$

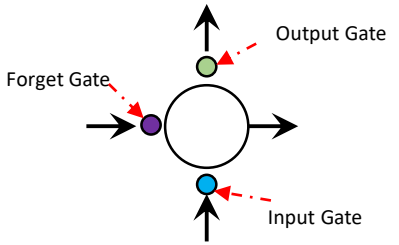
$$f^{(t)} = \sigma(W^{(f)}x^{(t)} + U^{(f)}h^{(t-1)})$$

$$o^{(t)} = \sigma(W^{(o)}x^{(t)} + U^{(o)}h^{(t-1)})$$

$$\tilde{c}^{(t)} = tanh(W^{(c)}x^{(t)} + U^{(c)}h^{(t-1)})$$

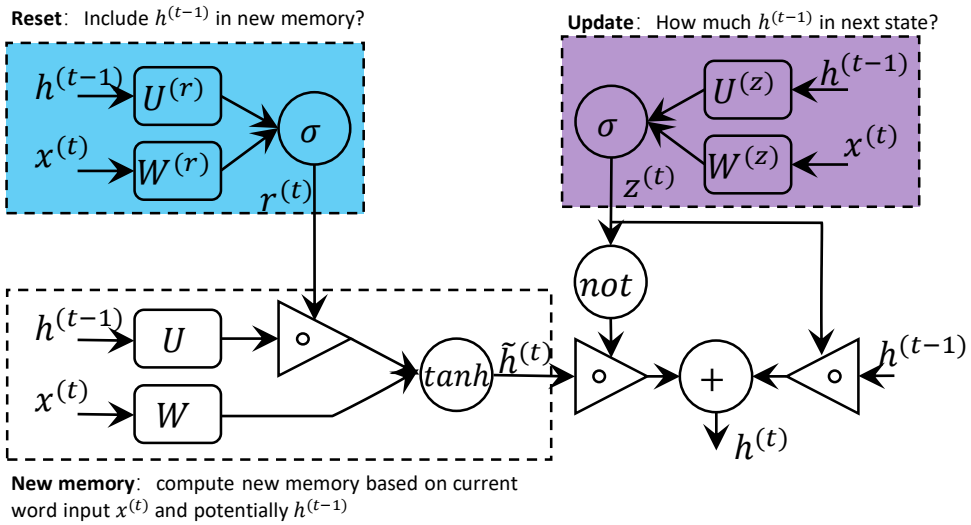
$$c^{(t)} = f^{(t)} \circ \tilde{c}^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)}$$

$$h^{(t)} = o^{(t)} \circ tanh(c^{(t)})$$



Sutskever, et al. ***Sequence to sequence learning with neural networks***. NIPS, 2014

GRU: Gated Recurrent Unit

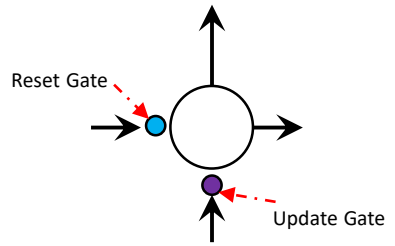


$$z^{(t)} = \sigma(W^{(z)}x^{(t)} + U^{(z)}h^{(t-1)})$$

$$r^{(t)} = \sigma(W^{(r)}x^{(t)} + U^{(r)}h^{(t-1)})$$

$$\tilde{h}^{(t)} = tanh(r^{(t)} \circ Uh^{(t-1)} + Wx^{(t)})$$

$$h^{(t)} = (1 - z^{(t)}) \circ \tilde{h}^{(t)} + z^{(t)} \circ h^{(t-1)}$$



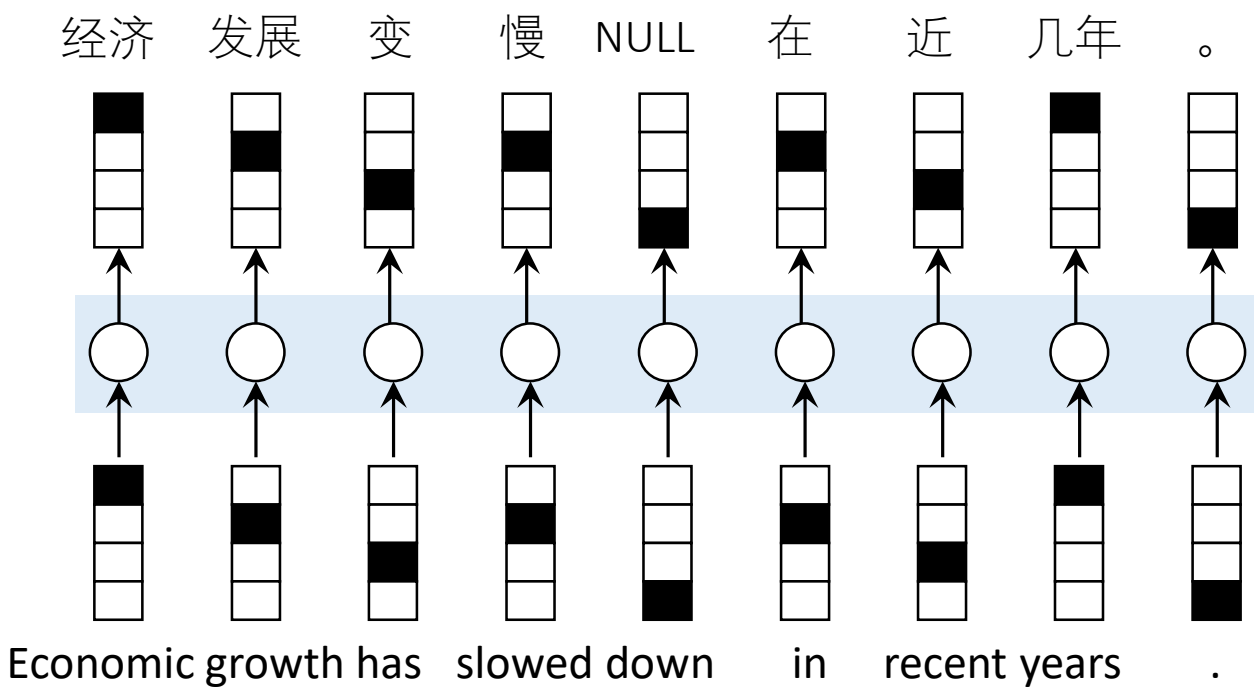
Bahdanau, et al. ***Neural Machine Translation by Jointly Learning to Align and Translate***. ICLR, 2015

Microsoft

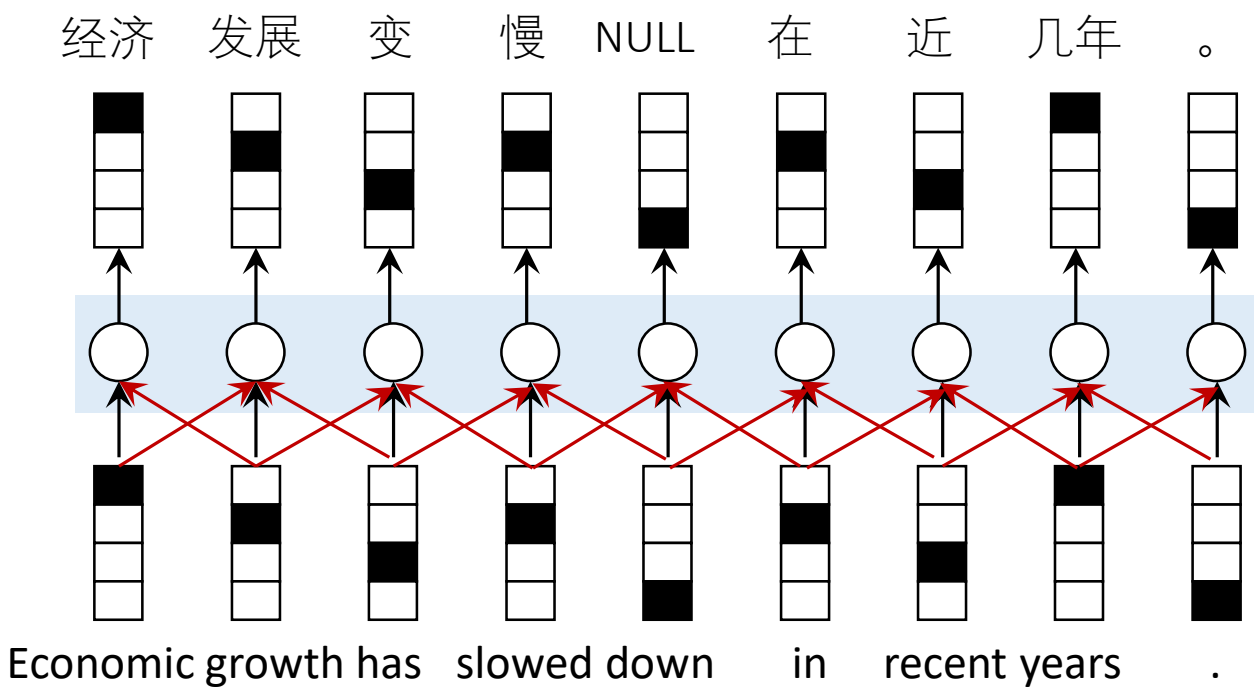
Neural Machine Translation

- Recurrent Neural Networks
- **Naïve Neural Machine Translation**
- RNN-based NMT
- Transformer-based NMT
- NMT Training

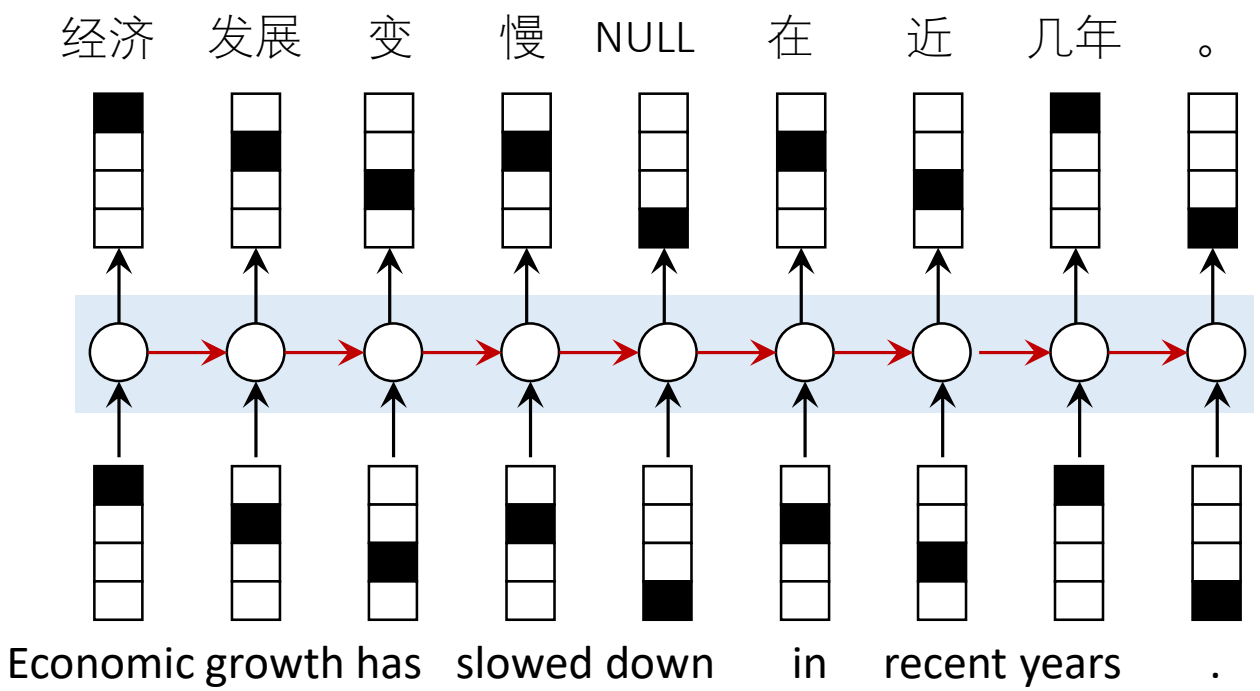
Naive Neural Machine Translation (1)



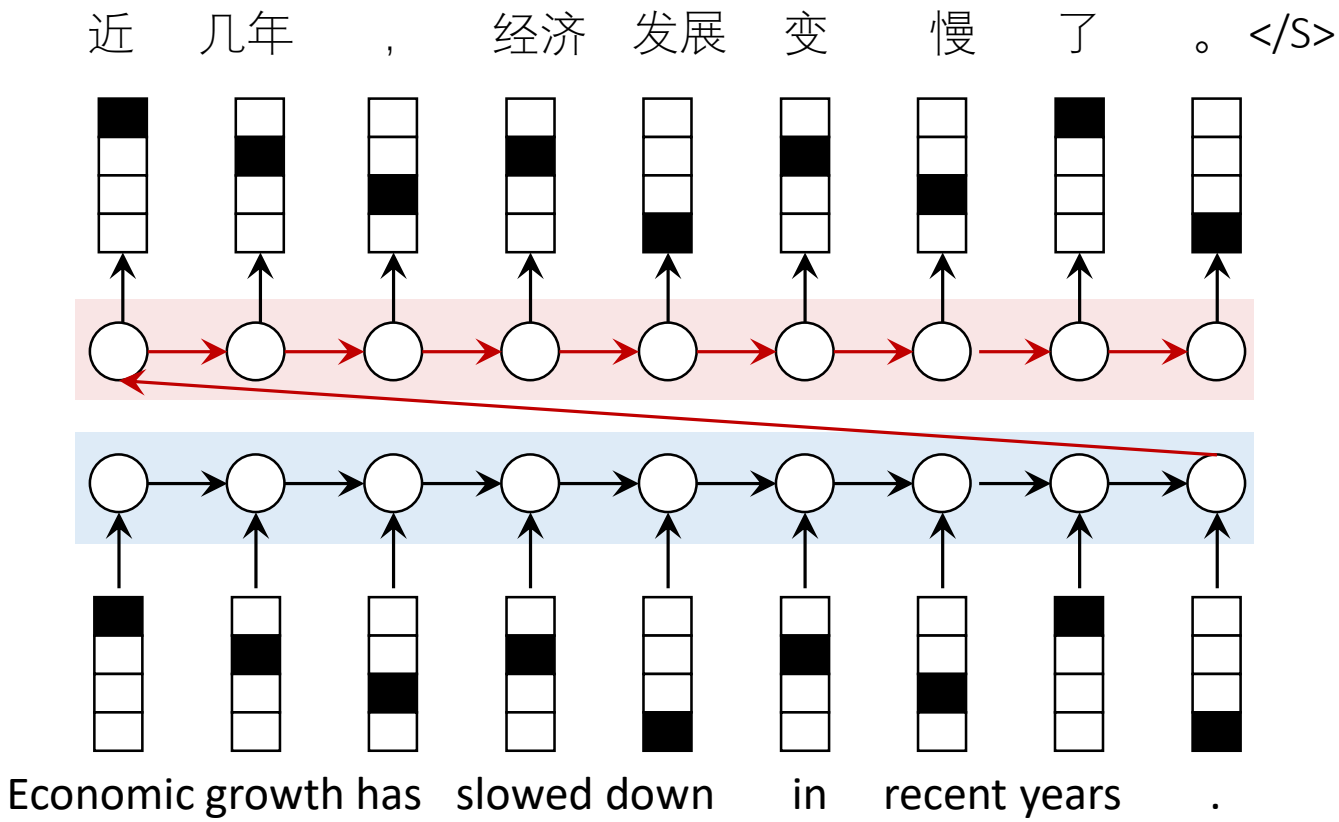
Naive Neural Machine Translation (2)



Naive Neural Machine Translation (3)

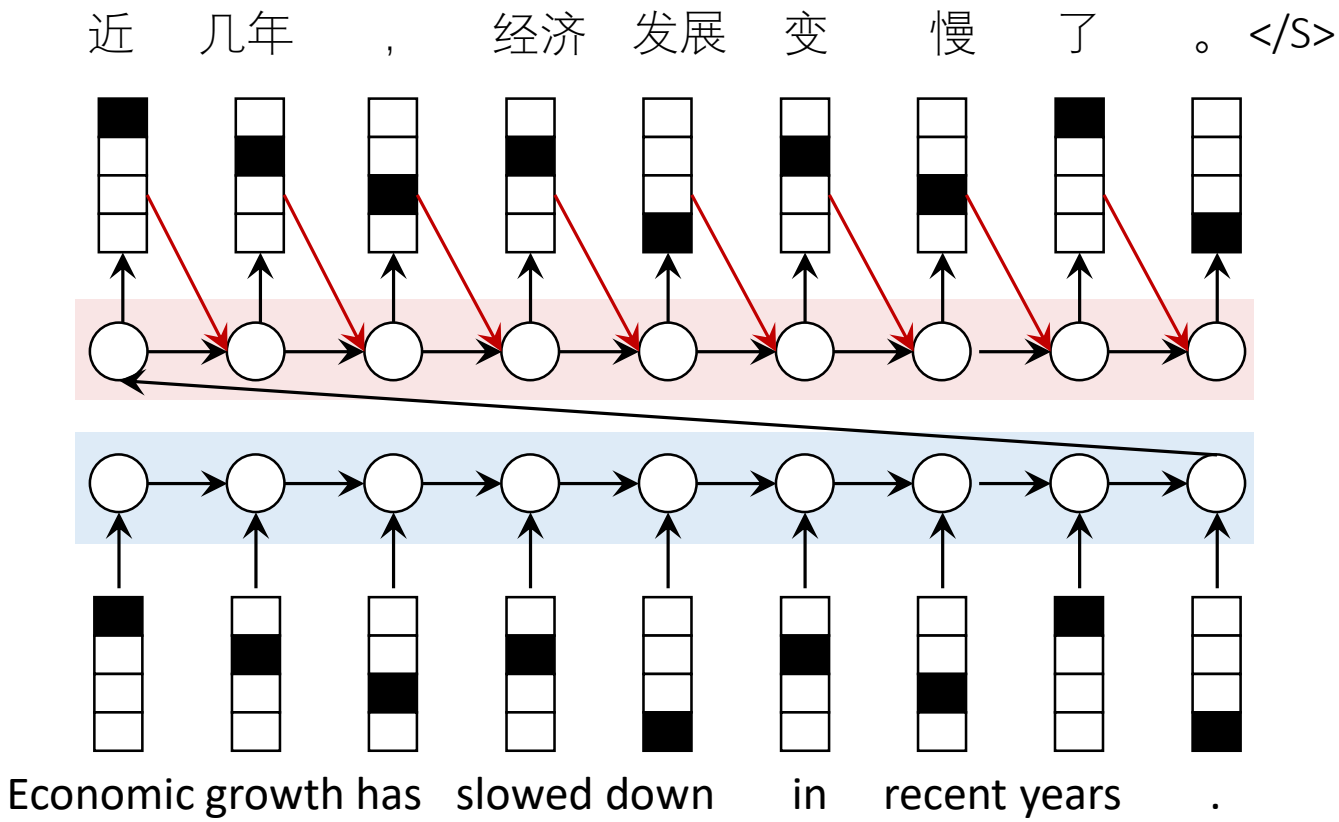


Naive Neural Machine Translation (4)

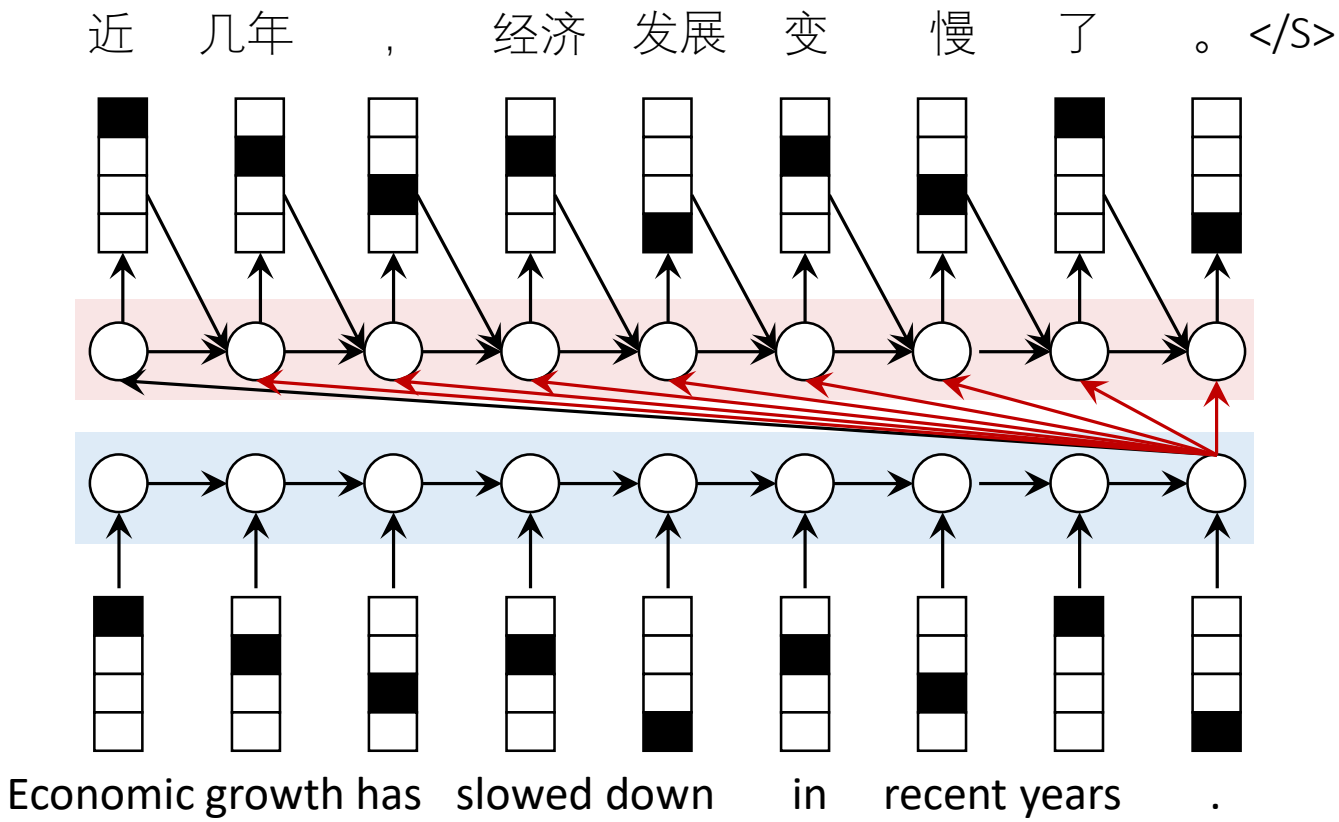


Sutskever, et al. ***Sequence to sequence learning with neural networksslate***. NIPS, 2014

Naive Neural Machine Translation (6)



Naive Neural Machine Translation (5)

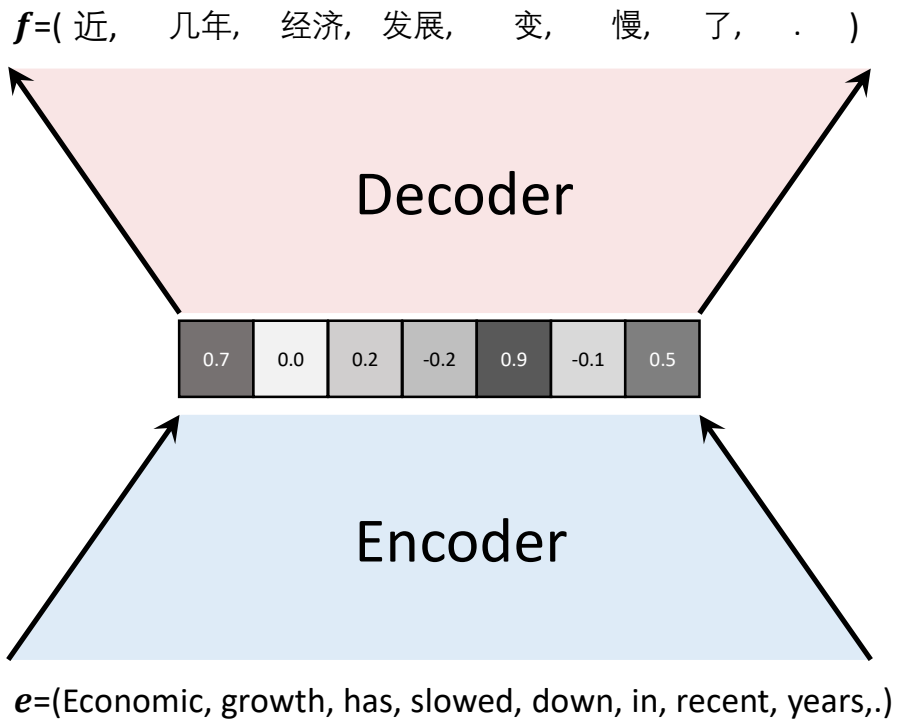


Microsoft

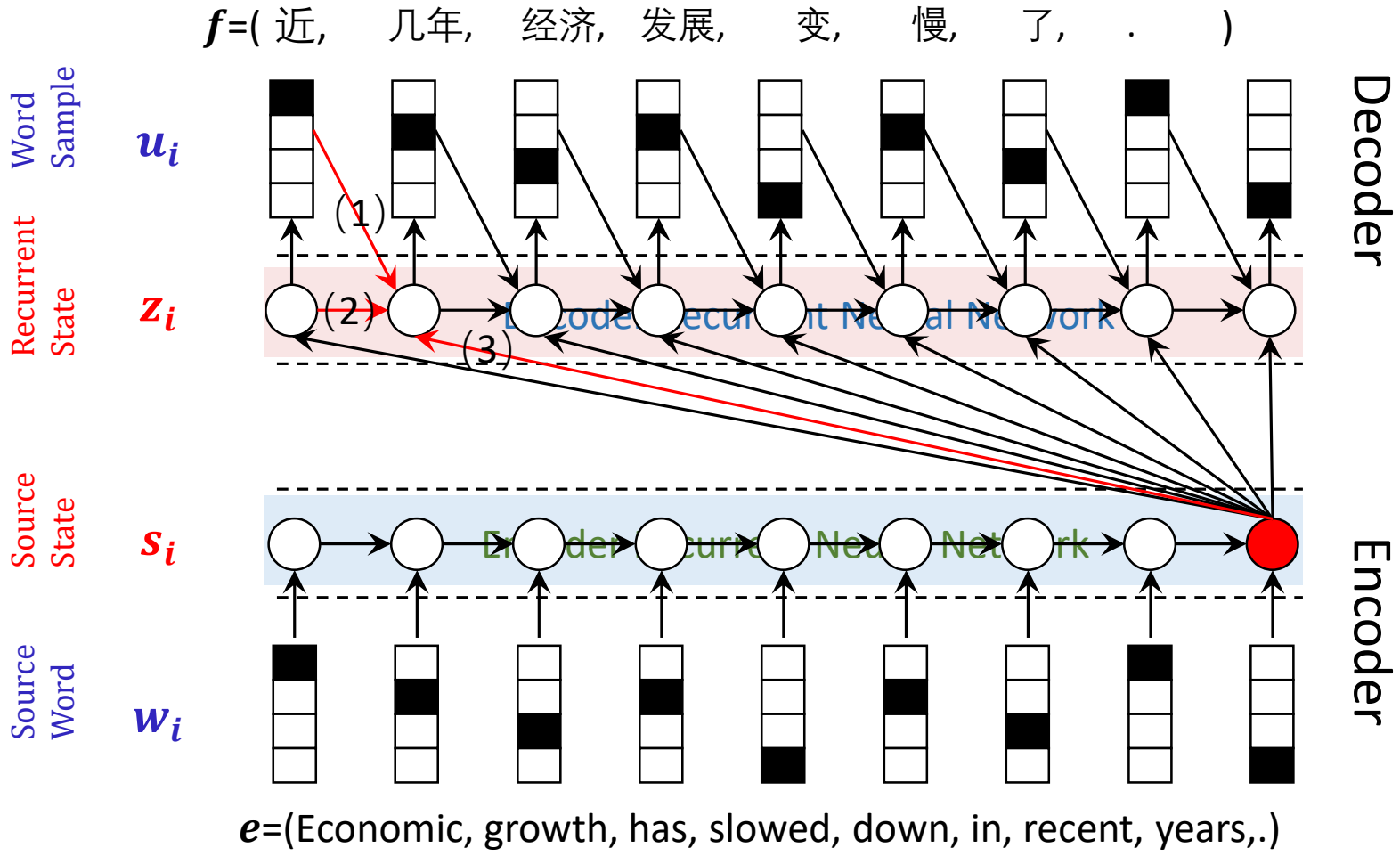
Neural Machine Translation

- Recurrent Neural Networks
- Naïve Neural Machine Translation
- **RNN-based NMT**
- Transformer-based NMT
- NMT Training

Encoder-Decoder for NMT



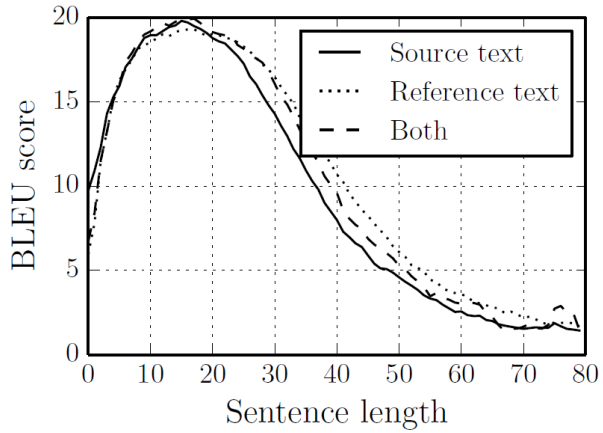
Encoder-Decoder for NMT



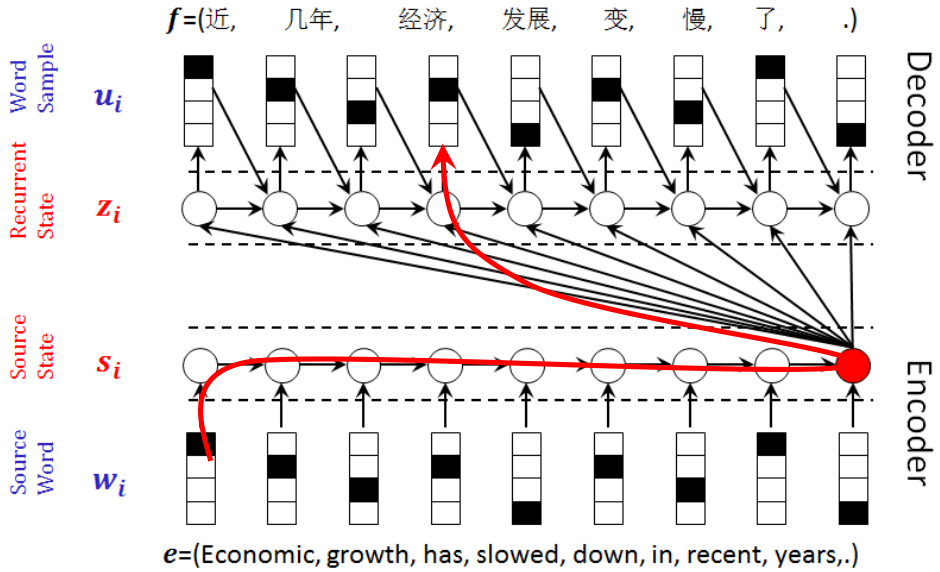
Bahdanau, et al. ***Neural Machine Translation by Jointly Learning to Align and Translate***. ICLR, 2015

Motivation of Attention

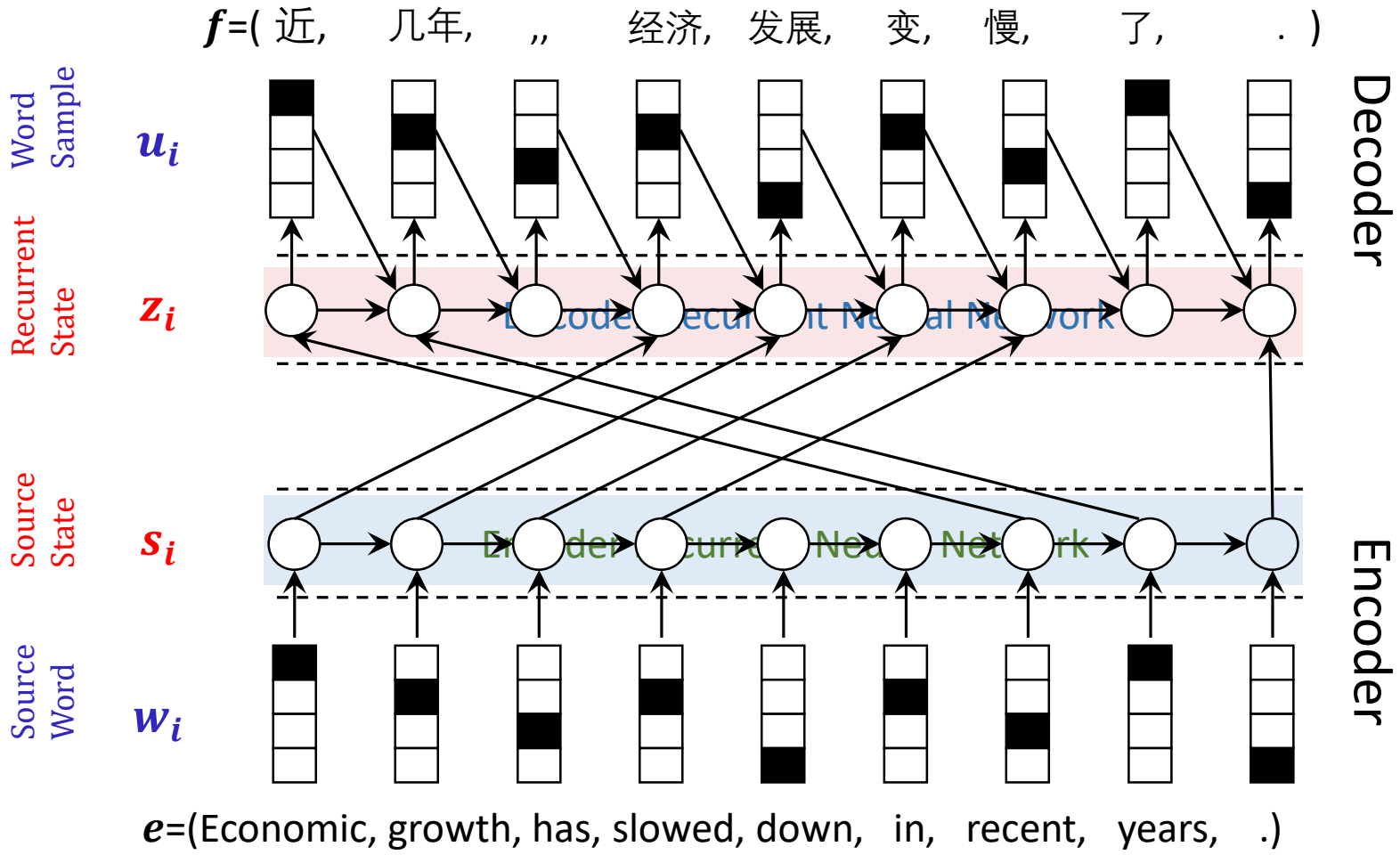
- NMT performs bad for long sentences
  - A long way from source to target
  - Only last hidden vector for decoding
  - Fixed-length hidden state is not enough



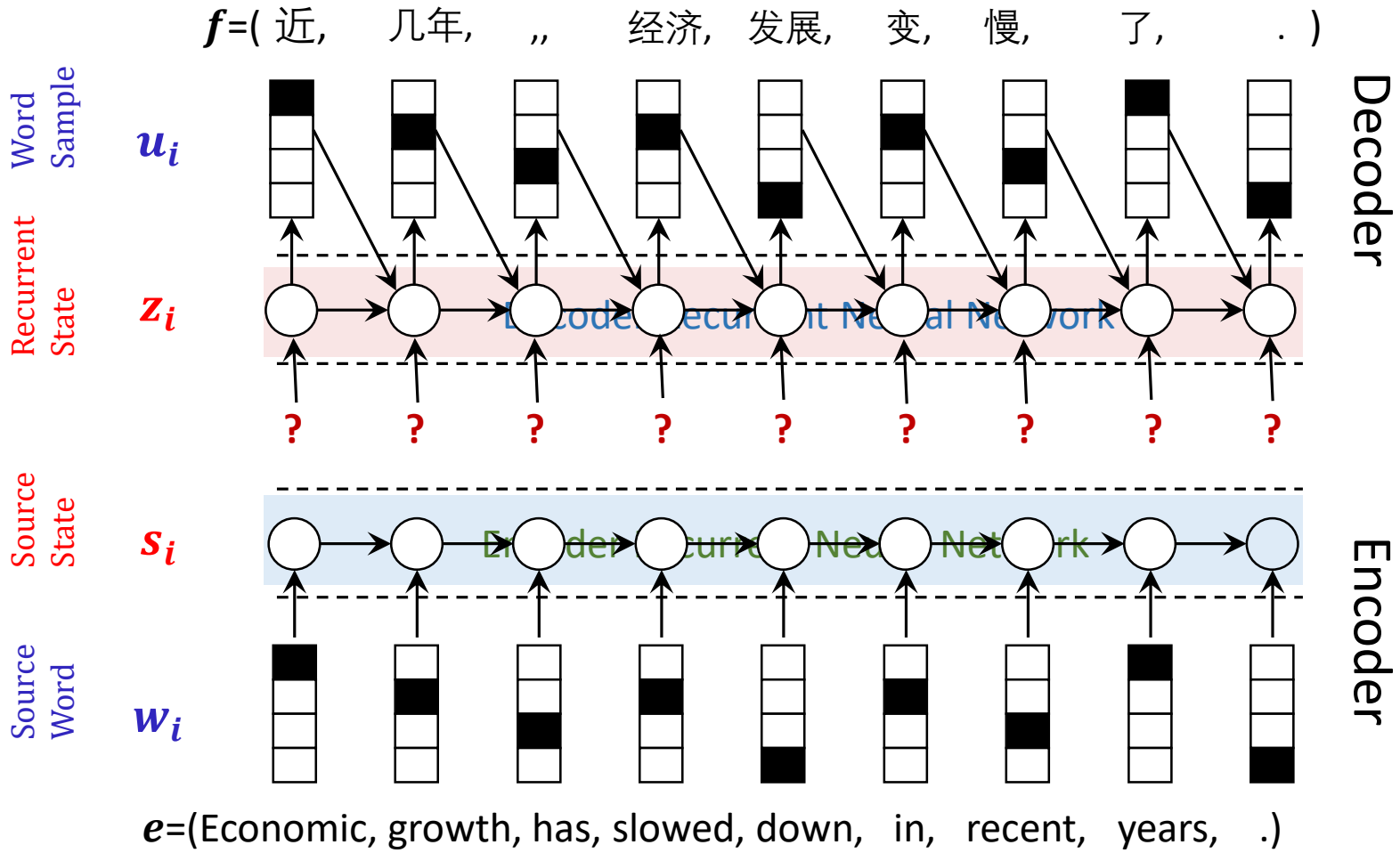
- Solution
  - Connect the source and target directly
  - Use all the hidden states for decoding



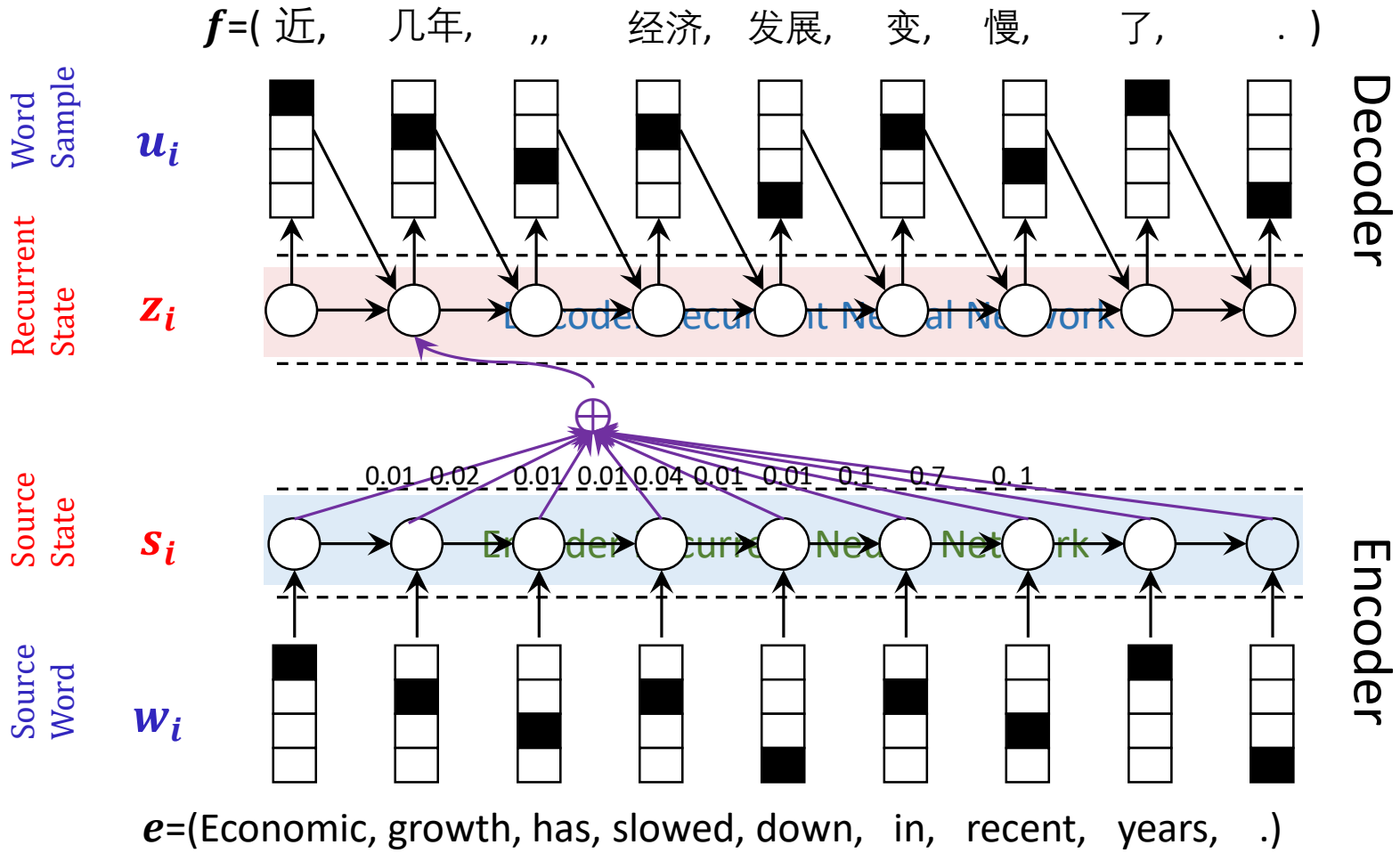
Encoder-Decoder for NMT



Encoder-Decoder for NMT

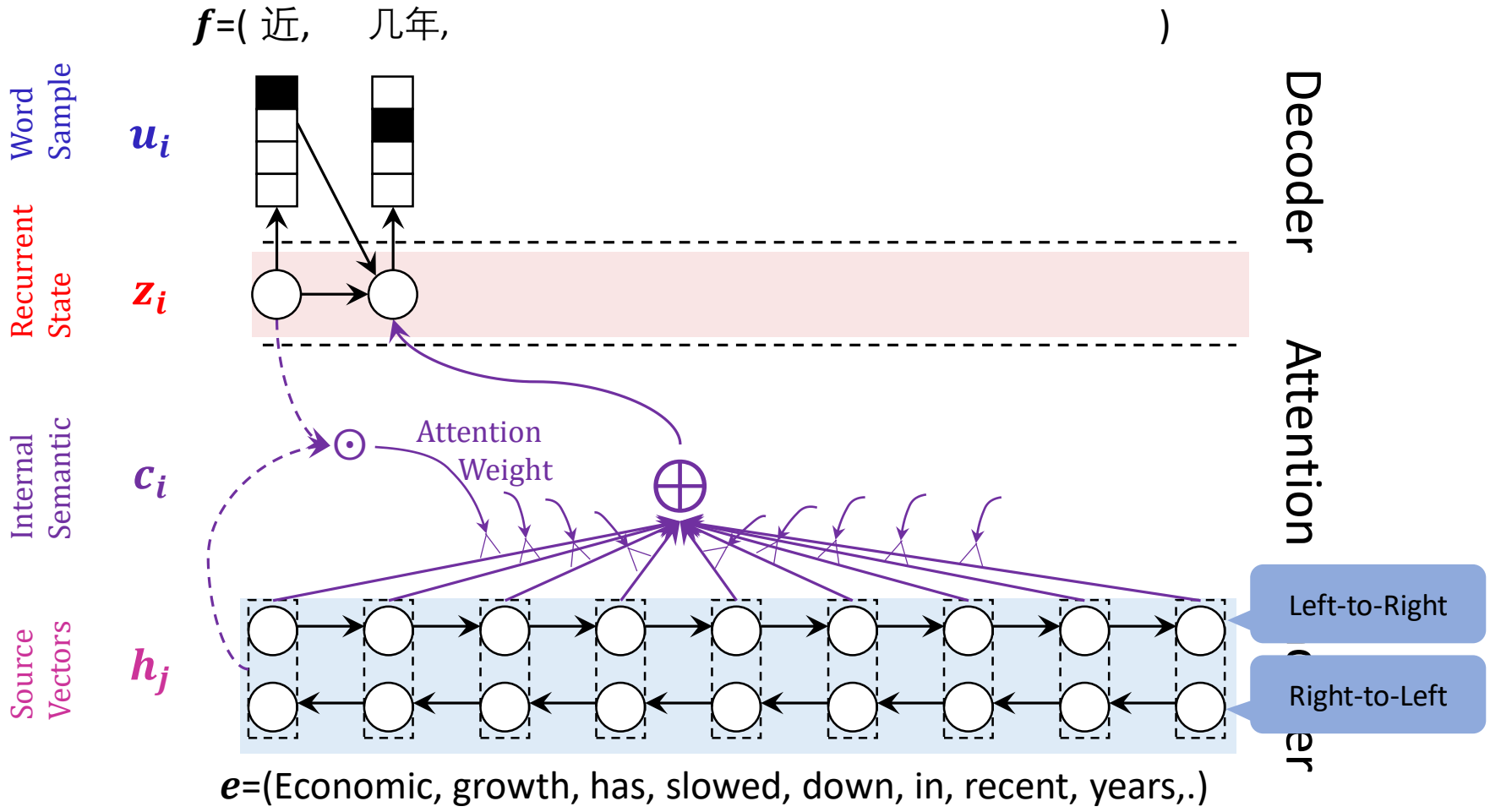


Encoder-Decoder for NMT



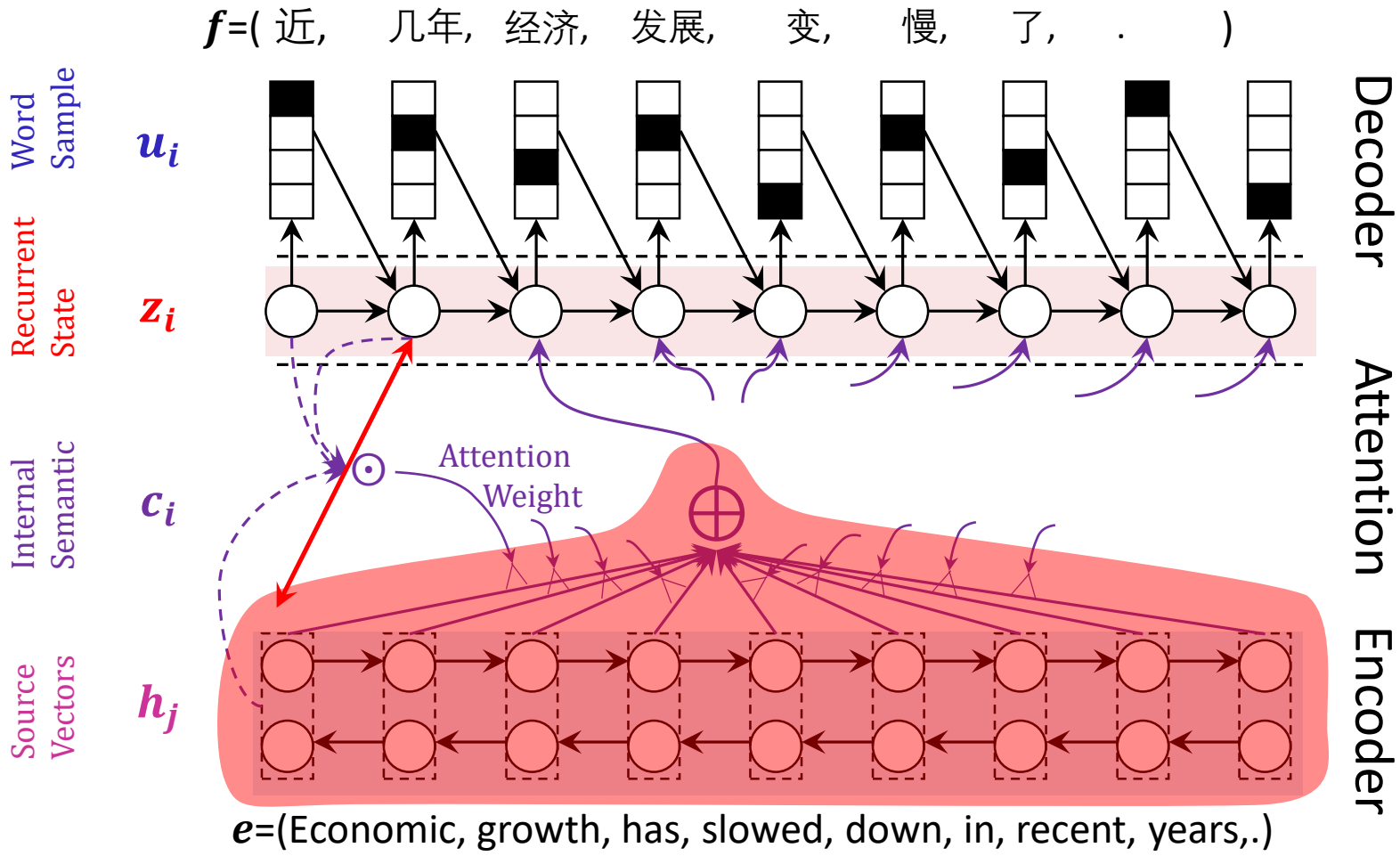
Bahdanau, et al. ***Neural Machine Translation by Jointly Learning to Align and Translate***. ICLR, 2015

Attention based Encoder-Decoder



Bahdanau, et al. ***Neural Machine Translation by Jointly Learning to Align and Translate***. ICLR, 2015

Attention based Encoder-Decoder



Bahdanau, et al. ***Neural Machine Translation by Jointly Learning to Align and Translate***. ICLR, 2015

Attention in Detail

- Compute matching score

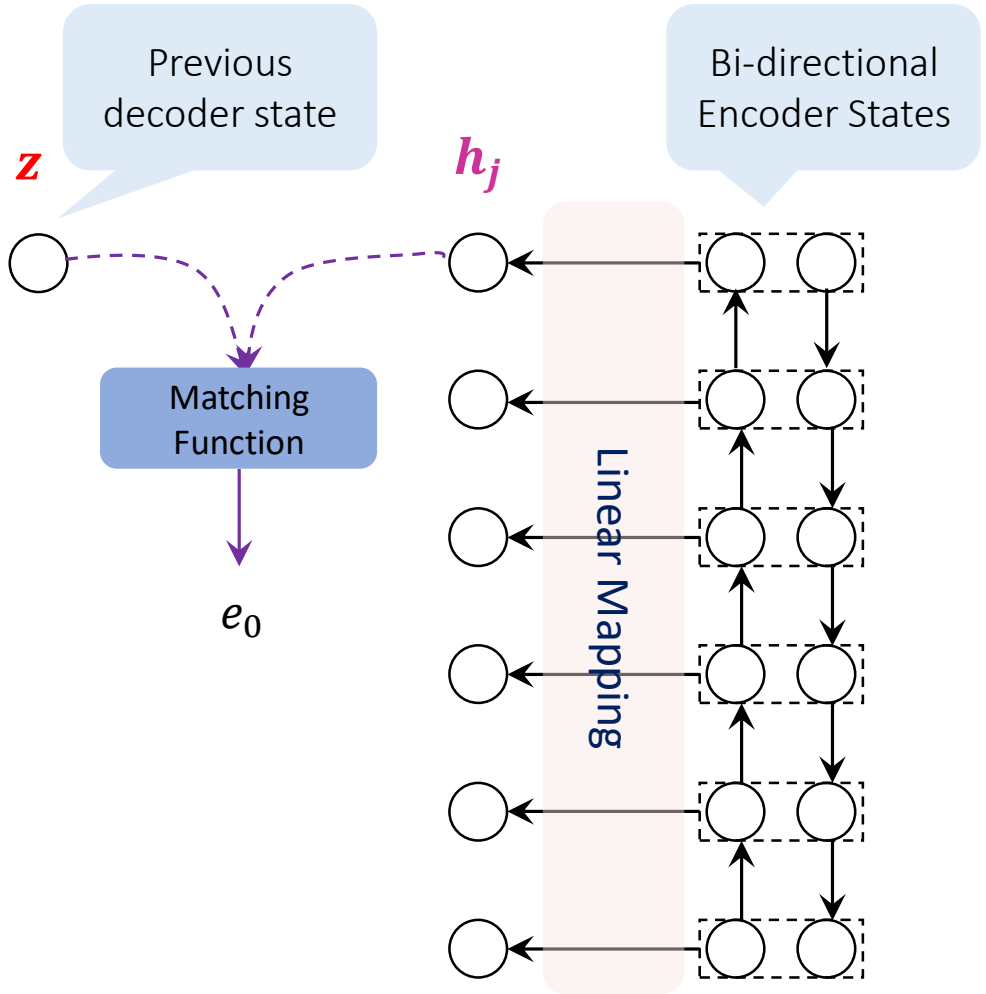
$$e_j = \begin{cases} z^T h_j & dot \\ z^T W h_j & general \\ V^T tanh\left(W\left[z; h_j\right]\right) & concat \end{cases}$$

- Normalize to be probability

$$a_j = e_j - \log(\sum_{j'} \exp(e_{j'}))$$

- Sum all the source hidden states

$$c = \sum_j \exp(a_j) h_j$$



Bahdanau, et al. ***Neural Machine Translation by Jointly Learning to Align and Translate***. ICLR, 2015

Attention in Detail

- Compute matching score

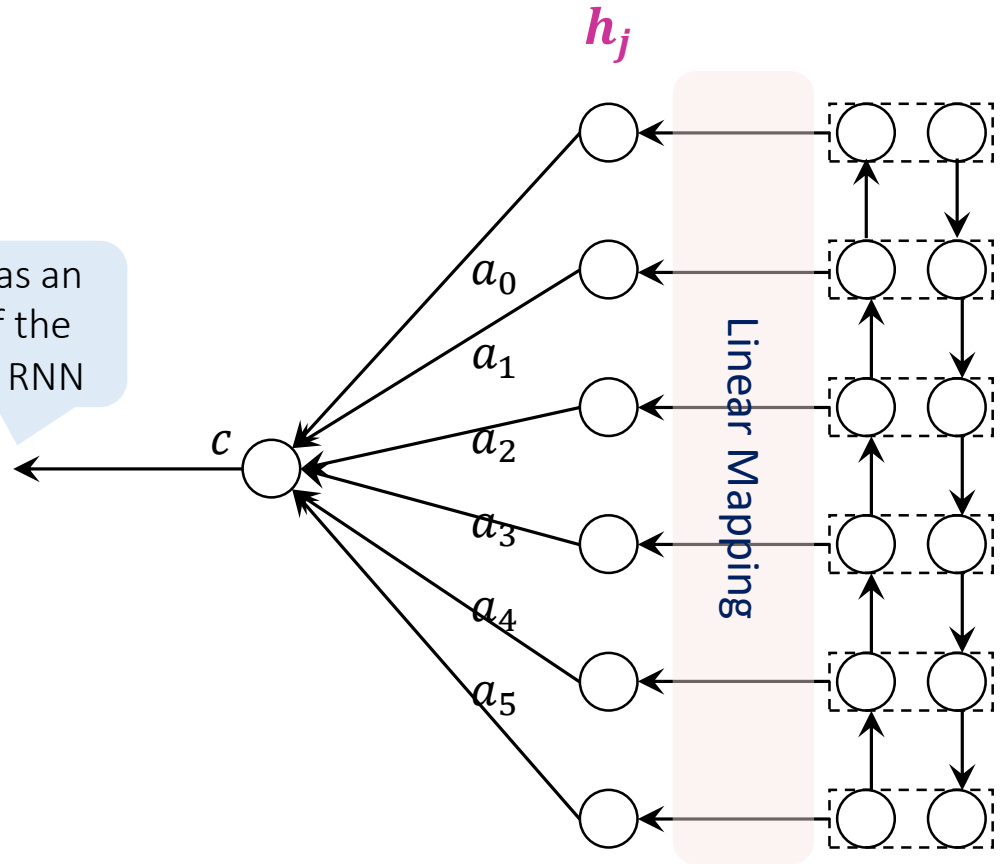
$$e_j = \begin{cases} z^T h_j & dot \\ z^T W h_j & general \\ V^T tanh\left(W\left[z; h_j\right]\right) & concat \end{cases}$$

- Normalize to be probability

$$a_j = e_j - \log\left(\sum_{j'} \exp(e_{j'})\right)$$

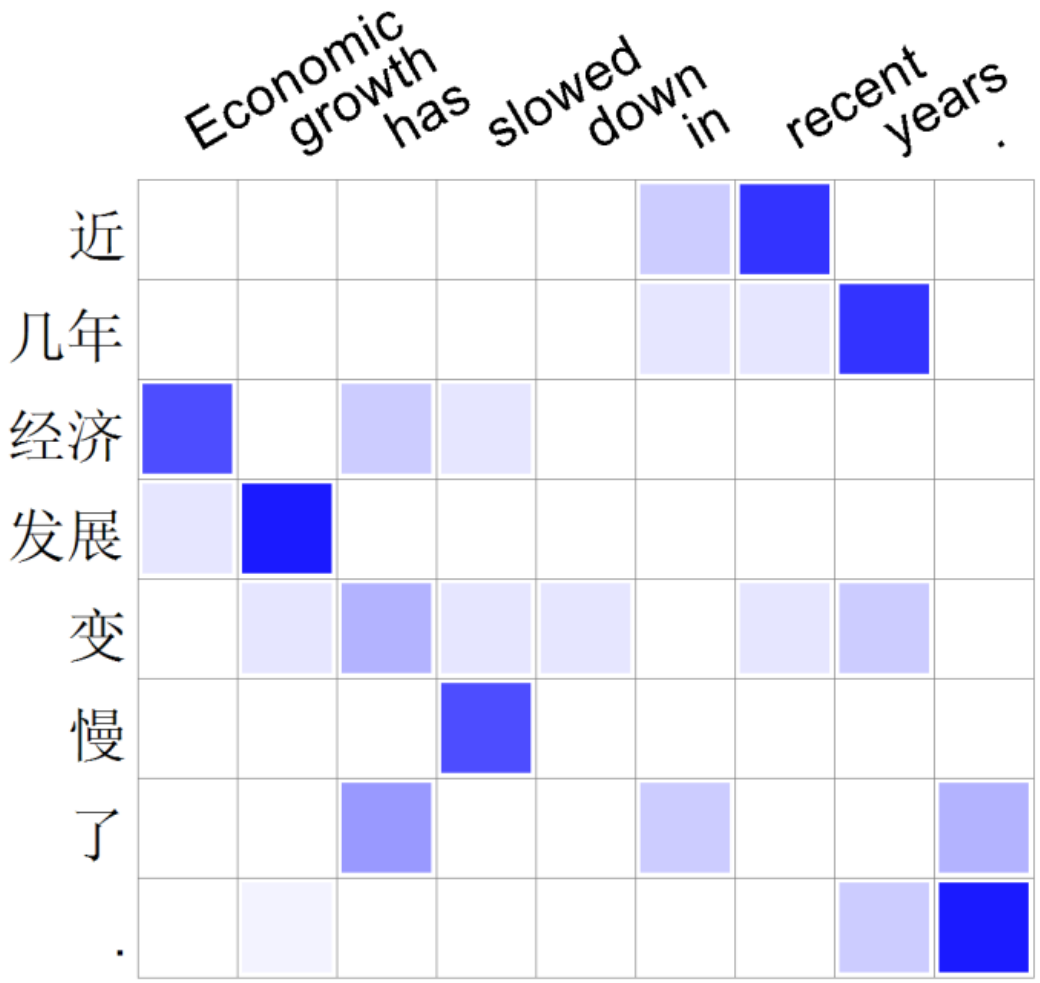
- Sum all the source hidden states

$$c = \sum_j \exp(a_j) h_j$$



Bahdanau, et al. ***Neural Machine Translation by Jointly Learning to Align and Translate***. ICLR, 2015

Case Study of Attention



Microsoft

Neural Machine Translation

- Recurrent Neural Networks
- Naïve Neural Machine Translation
- RNN-based NMT
- **Transformer-based NMT**
- NMT Training

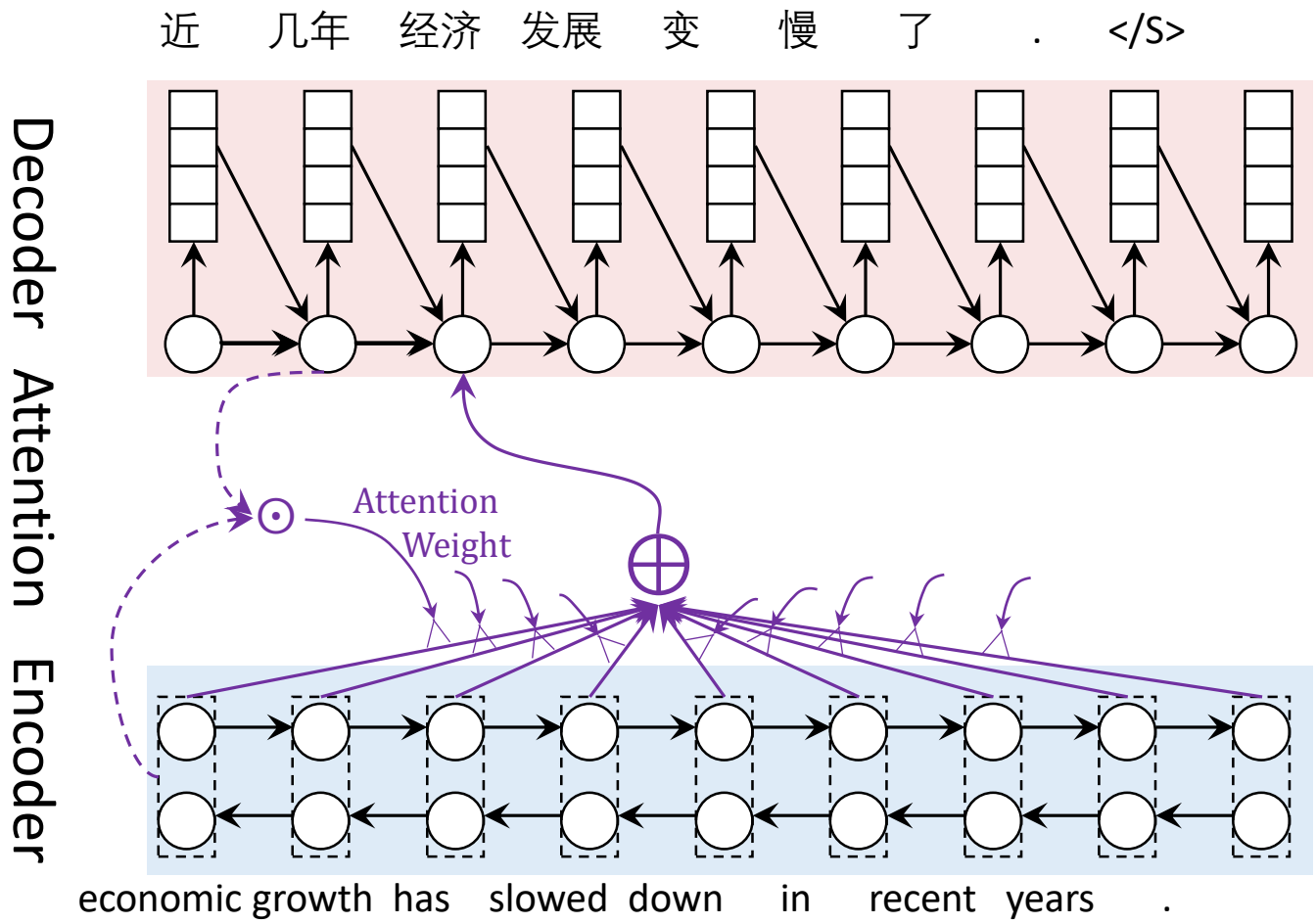
Merits and Demerits of RNN-based NMT

Merits

- Not only the last one but all the source hidden states are used.
- To generate the target word, the corresponding source word can receive more attention.
- Source and target hidden states are directly connected.

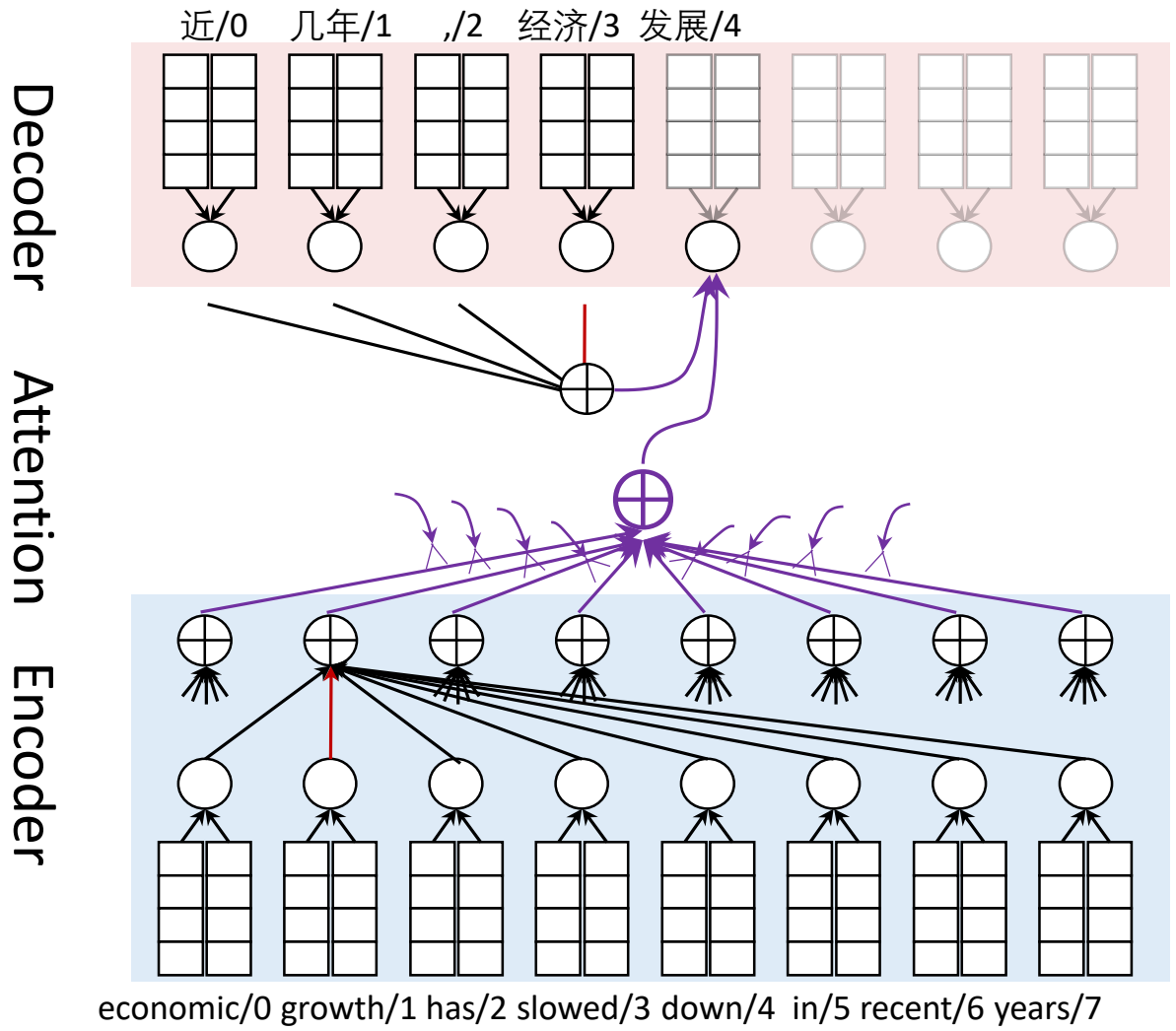
Demerits

- Hard to model long dependency using RNNs.
- Can only pay attention to the source hidden states from one perspective using the one-head attention mechanism.



Bahdanau, et al. ***Neural machine translation by jointly learning to align and translate***. arXiv, 2014

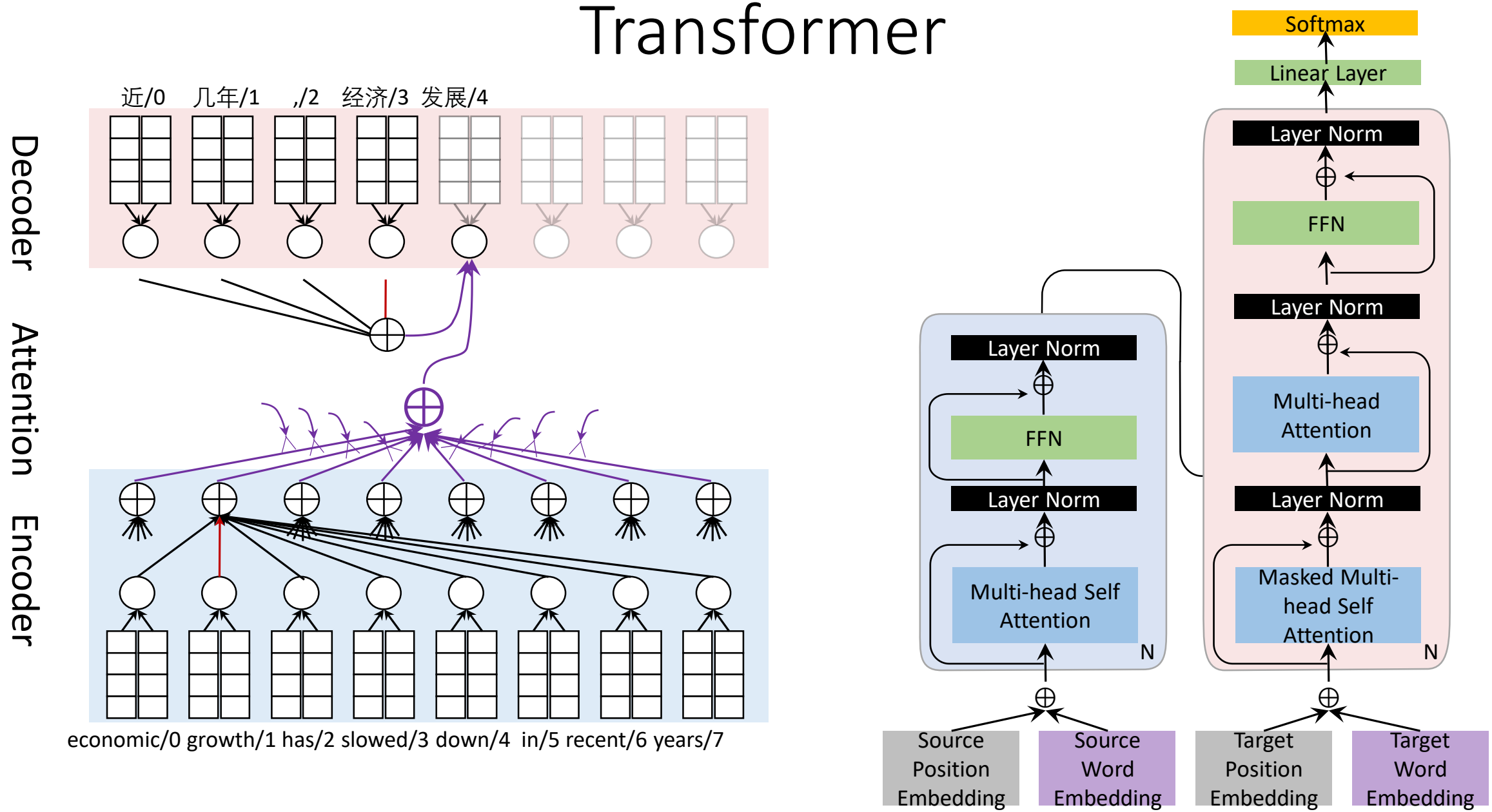
Transformer



- Self-attention is used to replace the RNNs in the encoder and decoder.
- Words with a long distance can be connected directly with self-attention.
- Multi-head attention is used to replace the one-head attention.
- Context information from different aspects can be modeled with multi-head attention.

Vaswani, et al. ***Attention is all you need***. NeurIPS, 2017

Transformer

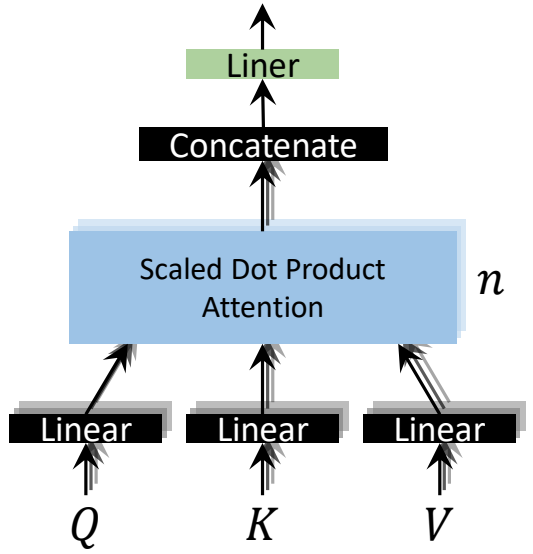


Transformer: Multi-Head Self-Attention

- Given the current hidden state as query Q, all the hidden states as K and V, multi-head self-attention is calculated as:

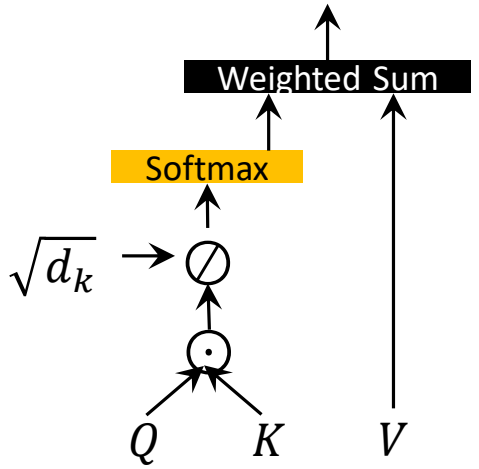
$$\text{MultiHead}(Q,K,V) = \text{Concat}(head_1, head_2, \dots, head_n)W^o$$

$$\text{head}_i = Attention(\text{Q}W_i^Q, \text{K}W_i^K, \text{V}W_i^V)$$



- Scaled dot product attention is calculated as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



d_k is the vector size of K. $\sqrt{d_k}$ is the temperature factor. the larger $\sqrt{d_k}$, the more even distribution.

Transformer: Position Embedding

- Why Position Embedding?
  - No recurrent structure in the encoder and decoder network.
  - Cannot model the ordering information of words in a sentence.
- Definition of Position Embedding

$$PE(pos, 2i) = sin(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$

$$PE(pos, 2i+1) = cos(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$

$$PE(pos+k, 2i) = sin\left(\frac{pos+k}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$= sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) cos\left(\frac{k}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$+ cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) sin\left(\frac{k}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$= PE(pos, 2i)PE(k, 2i+1)$$

$$+ PE(pos, 2i+1)PE(k, 2i)$$

Microsoft

Neural Machine Translation

- Recurrent Neural Networks
- Naïve Neural Machine Translation
- RNN-based NMT
- Transformer-based NMT
- **NMT Training**

Cross Entropy Training of NMT

Minimize This !

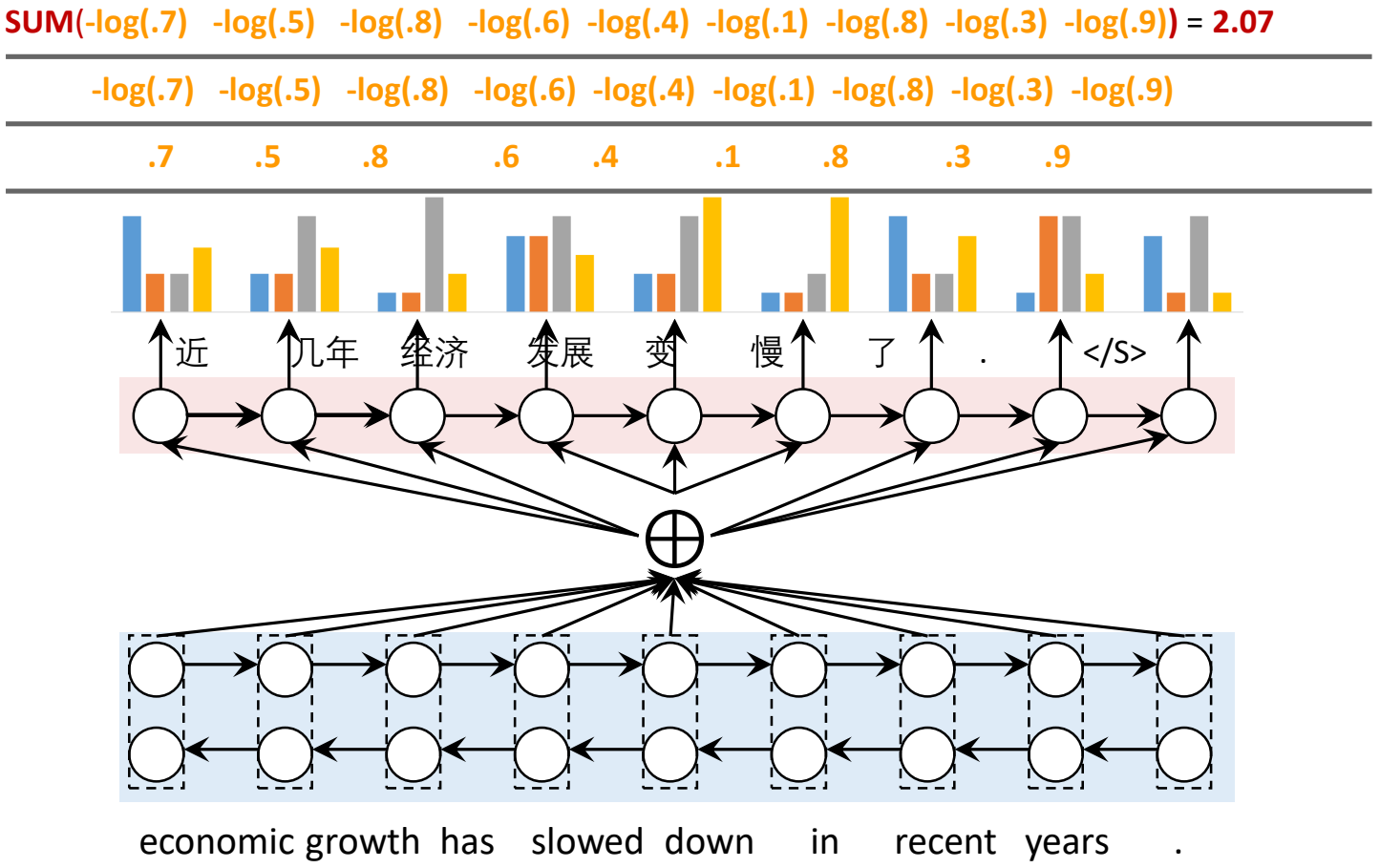
- Given the input sentence $x = (x_1, x_2, \dots, x_{|x|})$ and the output sentence $y = (y_1, y_2, \dots, y_{|y|})$, the likelihood of the conditional probability is:

$$p_\theta(y|x) = \prod_{i=1}^{|y|} p_\theta(y_i|y_{i-1}, \dots, y_1, x)$$

- The loss is defined as:

$$LOSS = \sum_{(x,y)\in T} -\log p_\theta(y|x)$$

$$= \sum_{(x,y)\in T} \sum_{i=1}^{|y|} -\log p_\theta(y_i|y_{i-1}, \dots, y_1, x)$$



SGD (Stochastic Gradient Descent)

- Perform a parameter update for each training example $x^{(i)}$ and label $y^{(i)}$, where θ is the parameter to be updated and η is learning rate.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^{(i)}; y^{(i)})$$

- When using mini-batch, $\nabla_\theta J(\theta)$ will be the average of gradient w.r.t θ over batches.

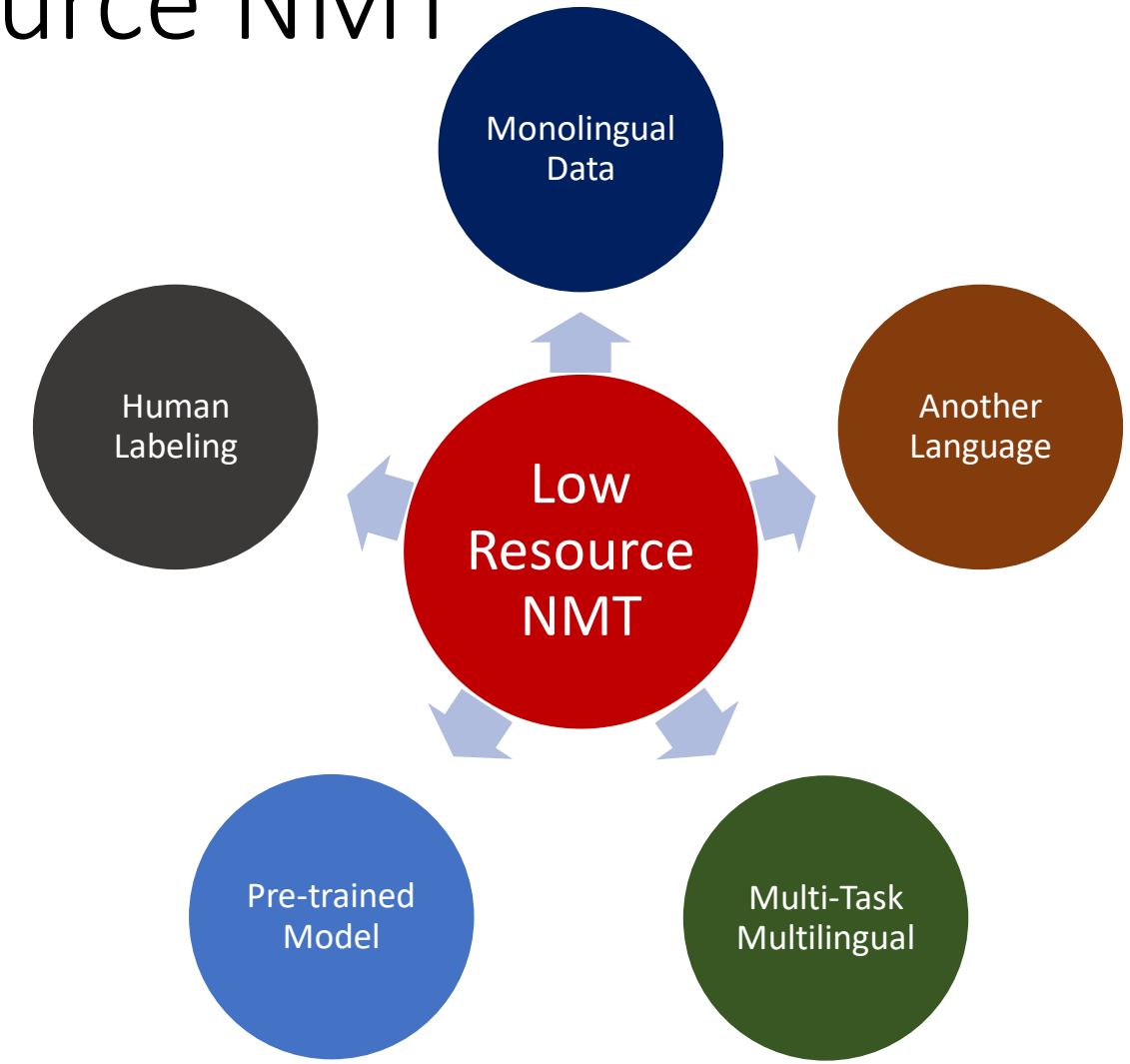
$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta)$$

- Other adaptive learning rate training methods: Adgrad, Adadelta, Adam.

Microsoft

Low Resource Neural Machine Translation

Low Resource NMT

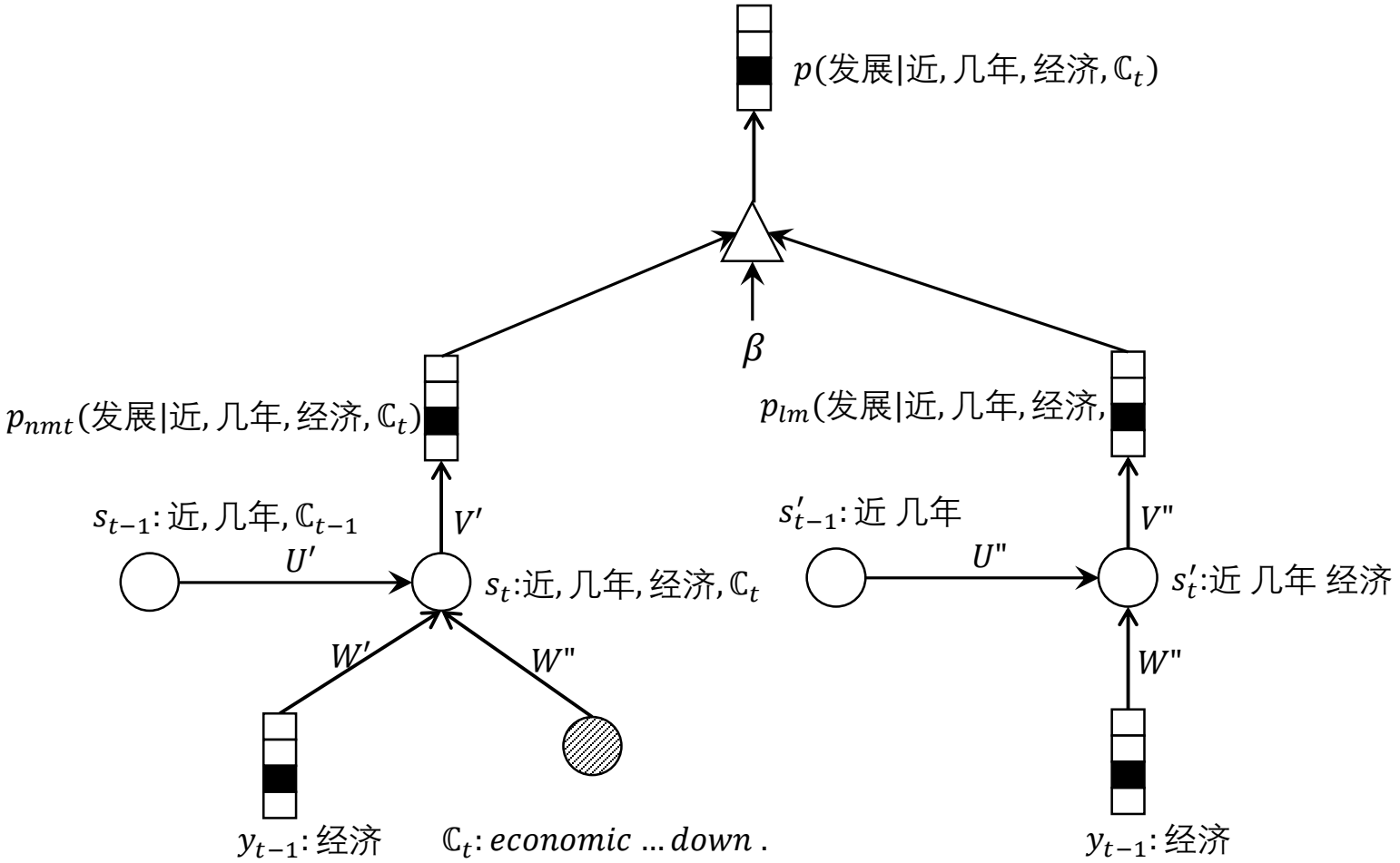


Microsoft

Leverage Monolingual Data for Low Resource NMT

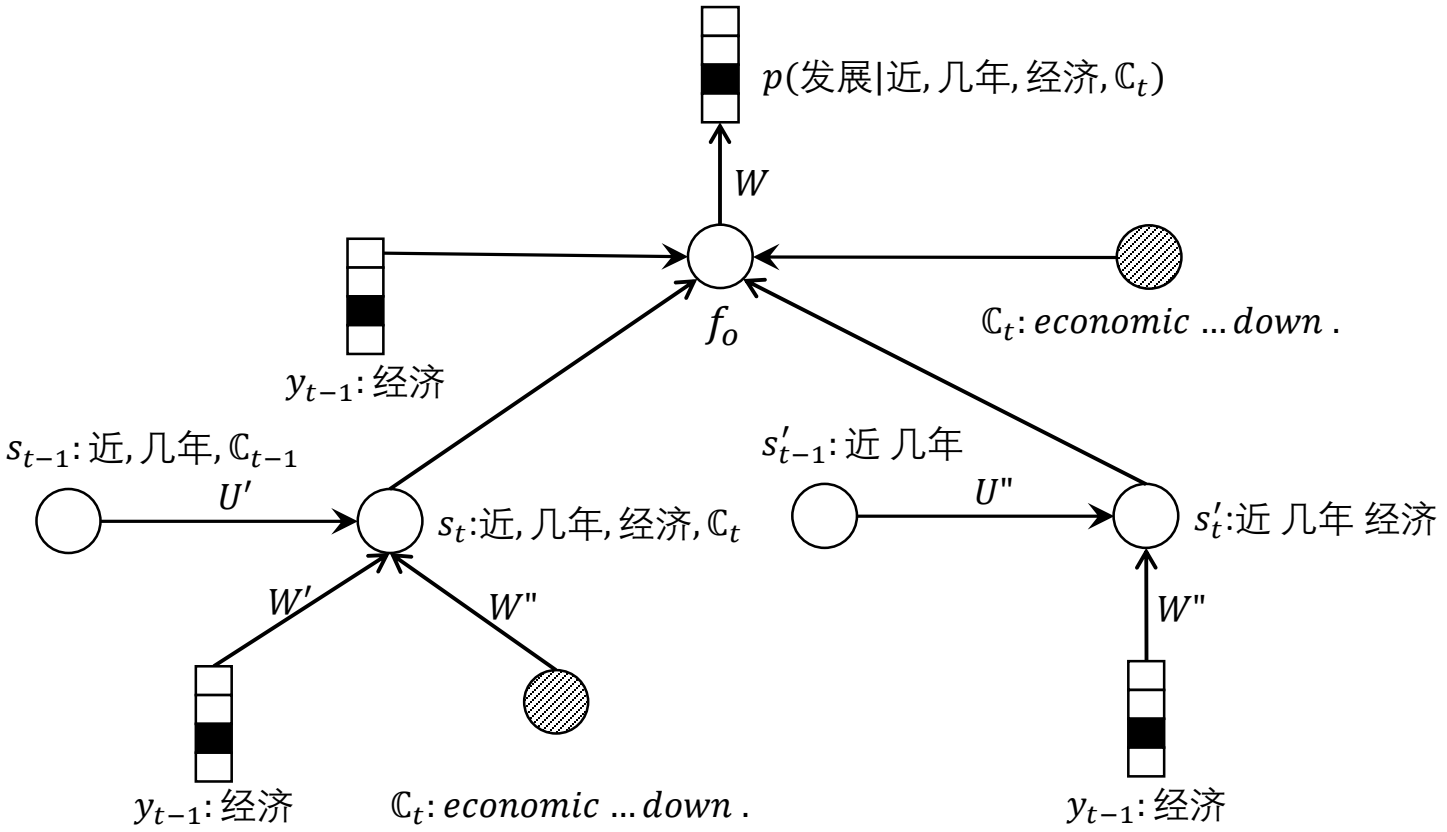
- **Shallow/Deep Fusion with Language Model**
- Autoencoder Method with Monolingual Data
- Dual Learning for Neural Machine Translation
- Back-Translation with Target Monolingual Data
- Joint Training for S2T and T2S Models

Shallow Fusion with Language Model



Gulcehre et al., **On Using Monolingual Corpora in Neural Machine Translation**. arxiv, 2015.

Deep Fusion with Language Model



Gulcehre et al., **On Using Monolingual Corpora in Neural Machine Translation**. arxiv, 2015.

Experiments for Shallow/Deep Fusion of LM

	SMS/CHAT		CTS	
	Dev	Test	Dev	Test
NMT	17.32	17.36	23.4	**23.59**
Shallow	16.59	16.42	22.7	22.83
Deep	**17.58**	**17.64**	**23.78**	23.5

OpenMT15:Zh-En

(Bilingual: 436K sentence pairs)
(Monolingual: Gigaword)
(SMS/CHAT: chat data,
CTS: conversational telephone speech)

	De-En		Cs-En	
	Dev	Test	Dev	Test
NMT Baseline	25.51	23.61	21.47	21.89
Shallow Fusion	25.53	23.69	21.95	22.18
Deep Fusion	**25.88**	**24.00**	**22.49**	**22.36**

WMT15:De-En, Cs-En

(Bilingual/De-En: 4.1M sentence pairs)
(Bilingual/Cs-En : 12.1M sentence pairs)
(Monolingual: Gigaword)

Gulcehre et al., **On Using Monolingual Corpora in Neural Machine Translation**. arxiv, 2015.

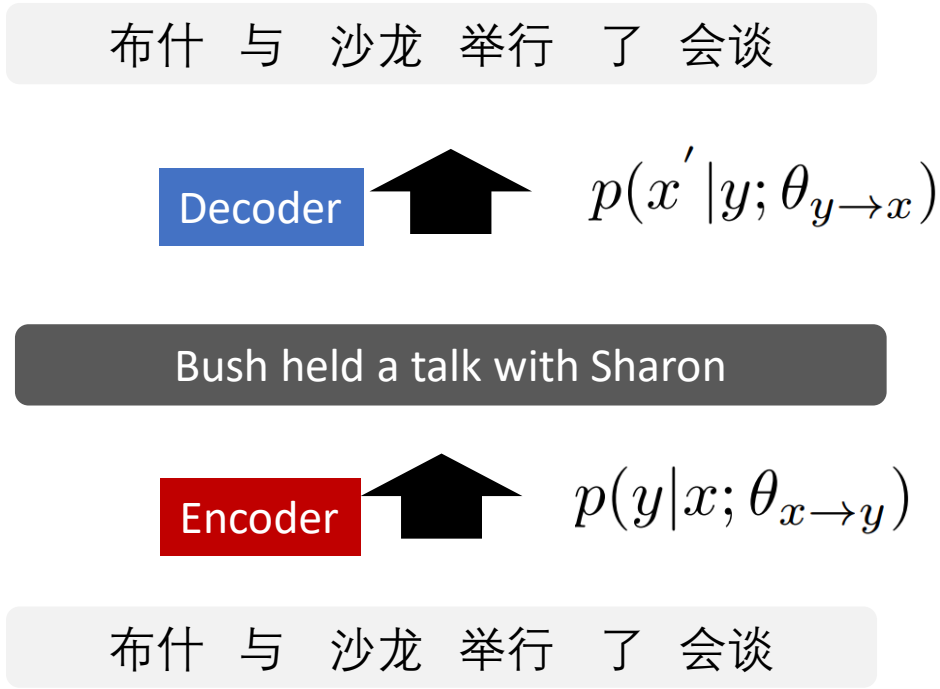
Microsoft

Leverage Monolingual Data for Low Resource NMT

- Shallow/Deep Fusion with Language Model
- **Autoencoder Method with Monolingual Data**
- Dual Learning for Neural Machine Translation
- Back-Translation with Target Monolingual Data
- Joint Training for S2T and T2S Models

Autoencoder Method with Monolingual Data

Source monolingual corpus: $\mathcal{X} = \{x^{(s)}\}_{s=1}^{S}$



$$p(x'|x;\theta_{x\to y},\theta_{y\to x})$$
$$= \sum_y p(x', y|x;\theta_{x\to y},\theta_{y\to x})$$
$$= \sum_y p(x'|y;\theta_{y\to x})p(y|x;\theta_{x\to y})$$

Encoder Decoder

Cheng et al., **Semi-Supervised Learning for Neural Machine Translation**. ACL, 2016.

Autoencoder Method with Monolingual Data

Target monolingual corpus: $\mathcal{Y} = \{y^{(t)}\}_{t=1}^{T}$

Bush held a talk with Sharon

Decoder $p(y^{'}|x;\theta_{x\to y})$

布什 与 沙龙 举行 了 会谈

Encoder $p(x|y;\theta_{y\to x})$

Bush held a talk with Sharon

$$p(y^{'}|y;\theta_{x\to y},\theta_{y\to x})$$
$$= \sum_{x} p(y^{'},x|y;\theta_{x\to y},\theta_{y\to x})$$
$$= \sum_{x} p(y^{'}|x;\theta_{x\to y})p(x|y;\theta_{y\to x})$$

Encoder Decoder

Cheng et al., **Semi-Supervised Learning for Neural Machine Translation**. ACL, 2016.

Autoencoder Method with Monolingual Data

$$L(\theta_{x\to y}) = \sum_{n=1}^{N} \log p(y^{(n)}|x^{(n)};\theta_{x\to y}) + \sum_{t=1}^{T} \log \sum_{x^{(t)}} p(y^{(t)'}|x^{(t)};\theta_{x\to y}) p(x^{(t)}|y^{(t)})$$
$$+ \sum_{s=1}^{S} \log \sum_{y^{(s)}} p(x^{(s)'}|y^{(s)}) p(y^{(s)}|x^{(s)};\theta_{x\to y})$$

$$\frac{\partial L(\theta_{x\to y})}{\partial \theta_{x\to y}} = \sum_{n=1}^{N} \frac{\partial \log p(y^{(n)}|x^{(n)};\theta_{x\to y})}{\partial \theta_{x\to y}} + \sum_{t=1}^{T} \frac{\partial \log \sum_{x^{(t)}} p(y^{(t)'}|x^{(t)};\theta_{x\to y}) p(x^{(t)}|y^{(t)})}{\partial \theta_{x\to y}}$$
$$+ \sum_{s=1}^{S} \frac{\partial \log \sum_{y^{(s)}} p(x^{(s)'}|y^{(s)}) p(y^{(s)}|x^{(s)};\theta_{x\to y})}{\partial \theta_{x\to y}}$$
$$= \sum_{n=1}^{N} \frac{\partial \log p(y^{(n)}|x^{(n)};\theta_{x\to y})}{\partial \theta_{x\to y}} + \sum_{t=1}^{T} \frac{\sum_{x^{(t)}\sim p(x^{(t)}|y^{(t)})} \sum_{y^{(t')}\sim p(y^{(t)'}|x^{(t)})} \partial \log p(y^{(t)'}|x^{(t)};\theta_{x\to y})\}}{\sum_{x^{(t)}} p(x^{(t)}|y^{(t)}) p(y^{(t)'}|x^{(t)}) \partial \theta_{x\to y}}$$
$$+ \sum_{s=1}^{S} \frac{\sum_{y^{(s)}\sim p(y^{(s)}|x^{(s)})} p(x^{(s)'}|y^{(s)}) \partial \log p(y^{(s)}|x^{(s)};\theta_{x\to y})}{\sum_{y^{(s)}} p(y^{(s)}|x^{(s)}) p(x^{(s)'}|y^{(s)}) \partial \theta_{x\to y}}$$

Intractable due to the exponential search space. Top-*k* list used as a substitution.

Cheng et al., **Semi-Supervised Learning for Neural Machine Translation**. ACL, 2016.

Autoencoder Method with Monolingual Data

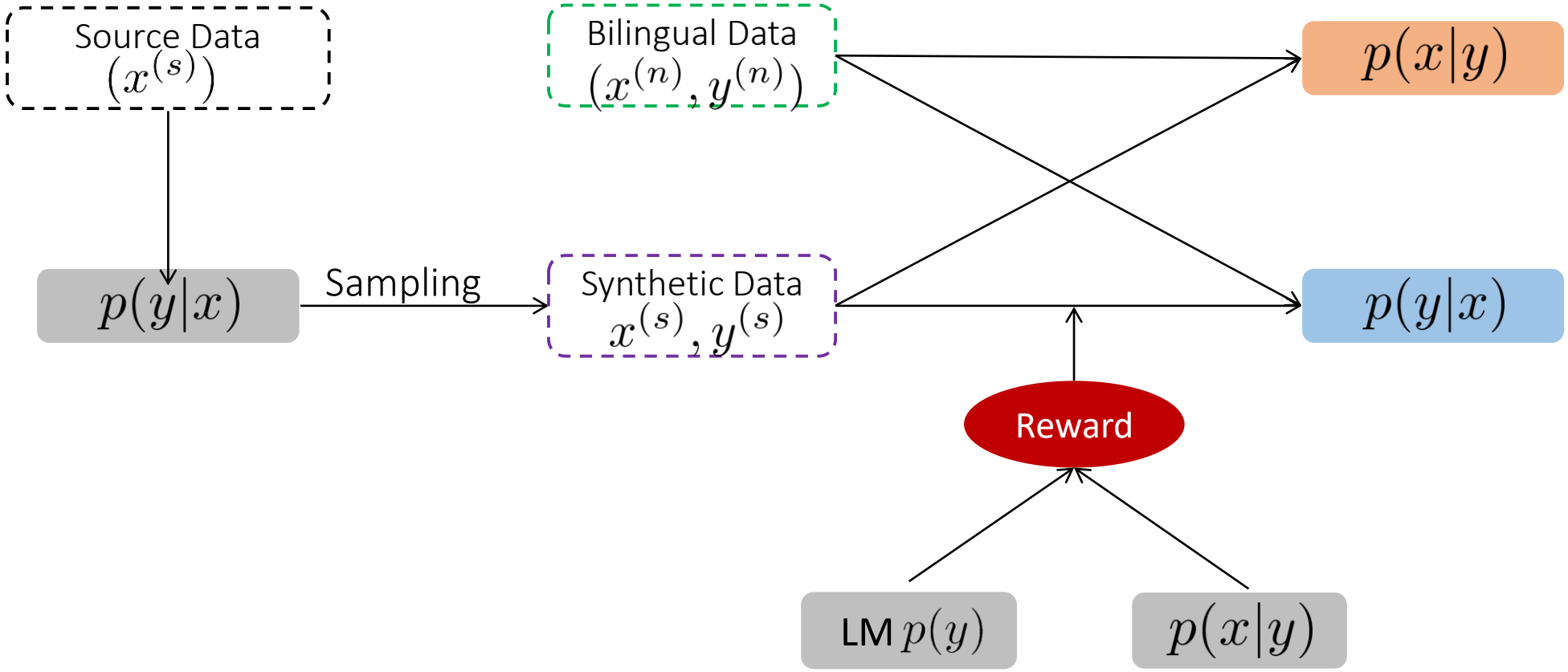
System	Training Data			Direction	NIST06	NIST02	NIST03	NIST04	NIST05
	CE	C	E						
MOSES	√	×	×	C → E	32.48	32.69	32.39	33.62	30.23
				E → C	14.27	18.28	15.36	13.96	14.11
	√	×	√	C → E	34.59	35.21	35.71	35.56	33.74
	√	√	×	E → C	20.69	25.85	19.76	18.77	19.74
RNNSEARCH	√	×	×	C →E	30.74	35.16	33.75	34.63	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^{++}

Cheng et al., **Semi-Supervised Learning for Neural Machine Translation**. ACL, 2016.

Leverage Monolingual Data for Low Resource NMT

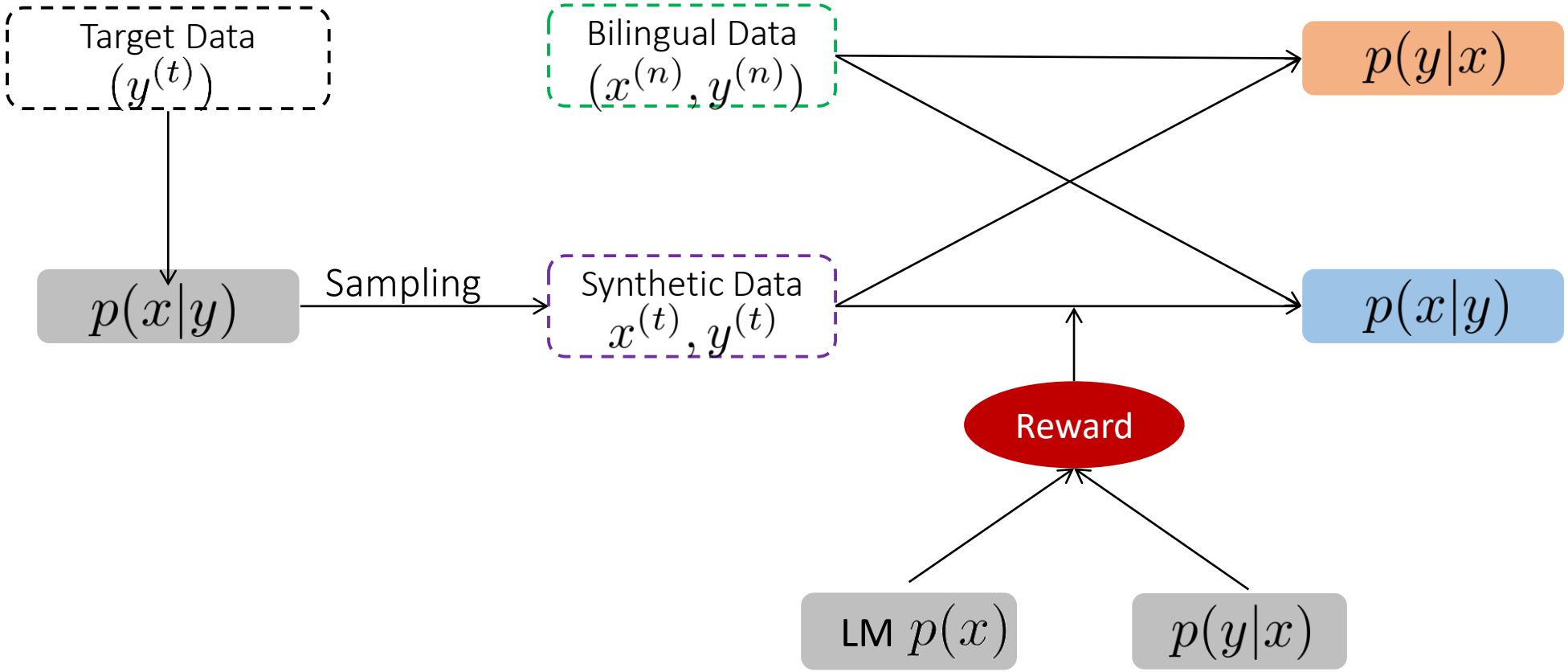
- Shallow/Deep Fusion with Language Model
- Autoencoder Method with Monolingual Data
- **Dual Learning for Neural Machine Translation**
- Back-Translation with Target Monolingual Data
- Joint Training for S2T and T2S Models

Dual Learning for Neural Machine Translation



He et al., **Dual Learning for Machine Translation**. NIPS, 2016.

Dual Learning for Neural Machine Translation



He et al., **Dual Learning for Machine Translation**. NIPS, 2016.

Dual Learning for Neural Machine Translation

Table 1: Translation results of En↔Fr task. The results of the experiments using all the parallel data for training are provided in the first two columns (marked by "Large"), and the results using 10% parallel data for training are in the last two columns (marked by "Small").

	En→Fr (Large)	Fr→En (Large)	En→Fr (Small)	Fr→En (Small)
NMT	29.92	27.49	25.32	22.27
pseudo-NMT	30.40	27.66	25.63	23.24
dual-NMT	**32.06**	**29.78**	**28.73**	**27.50**

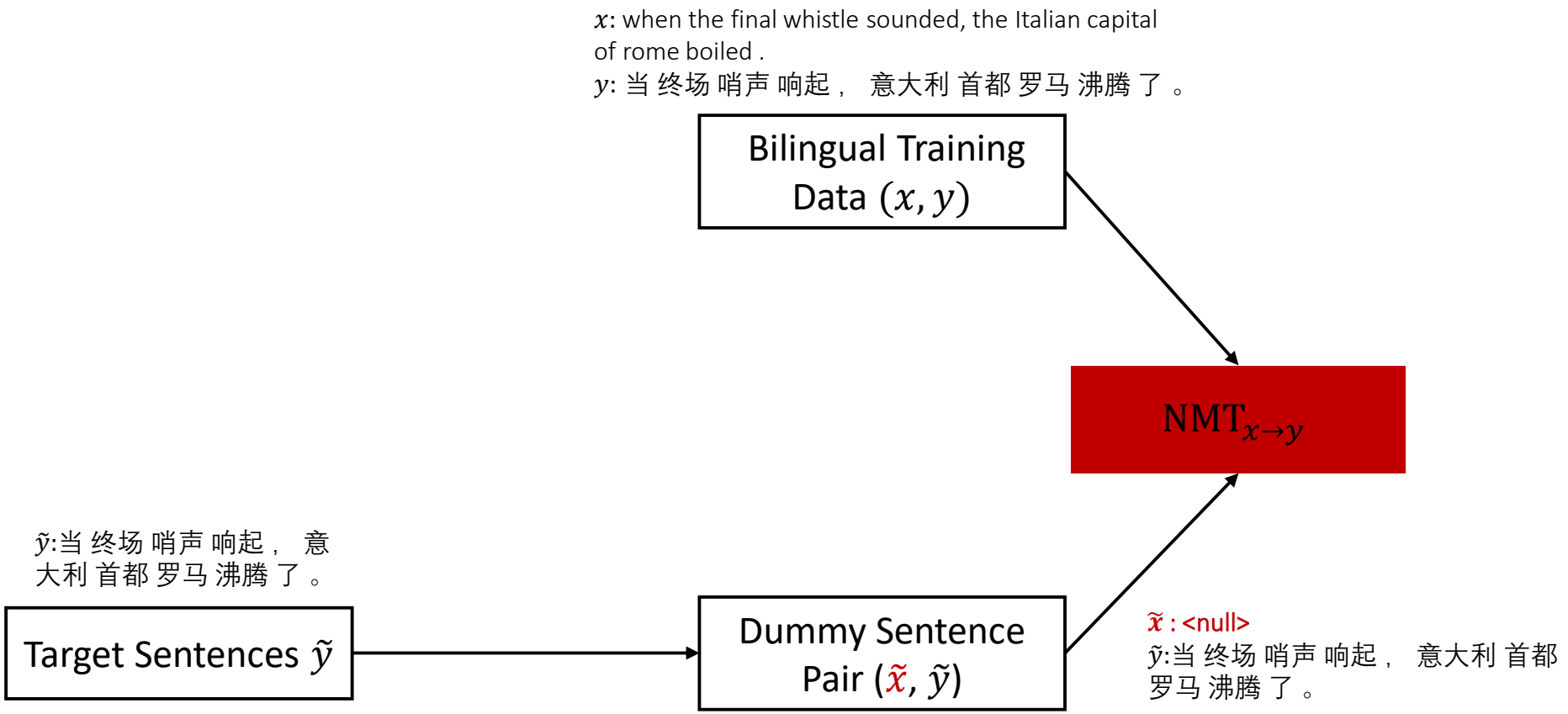
He et al., **Dual Learning for Machine Translation**. NIPS, 2016.

Microsoft

Leverage Monolingual Data for Low Resource NMT

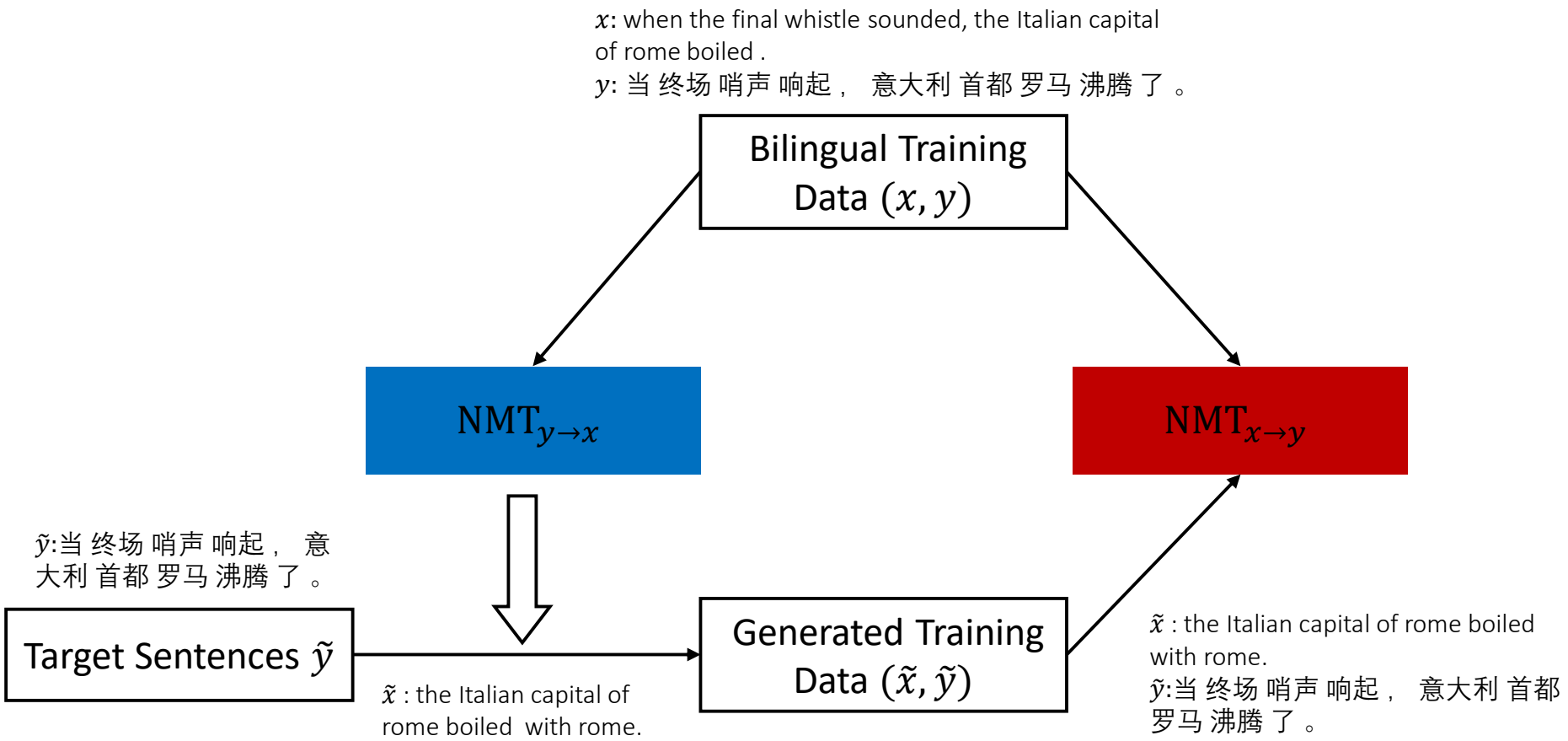
- Shallow/Deep Fusion with Language Model
- Autoencoder Method with Monolingual Data
- Dual Learning for Neural Machine Translation
- **Back-Translation with Target Monolingual Data**
- Joint Training for S2T and T2S Models

Back-Translation with Target Monolingual Data



Sennrich et al., **Improving Neural Machine Translation Models with Monolingual Data**. ACL, 2016.

Back-Translation with Target Monolingual Data



Sennrich et al., **Improving Neural Machine Translation Models with Monolingual Data**. ACL, 2016.

Back-Translation with Target Monolingual Data

name	training instances	BLEU newstest2014 single	BLEU newstest2014 ens-4	BLEU newstest2015 single	BLEU newstest2015 ens-4
syntax-based (Sennrich and Haddow, 2015)		22.6	-	24.4	-
Neural MT (Jean et al., 2015b)		-	-	22.4	-
parallel	37m (parallel)	19.9	20.4	22.8	23.6
+monolingual	49m (parallel) / 49m (monolingual)	20.4	21.4	23.2	24.6
+synthetic	44m (parallel) / 36m (synthetic)	**22.7**	**23.8**	**25.7**	**26.5**

Table 3: English→German translation performance (BLEU) on WMT training/test sets. Ens-4: ensemble of 4 models. Number of training instances varies due to differences in training time and speed.

Sennrich et al., **Improving Neural Machine Translation Models with Monolingual Data**. ACL, 2016.

Leverage Monolingual Data for Low Resource NMT

- Shallow/Deep Fusion with Language Model
- Autoencoder Method with Monolingual Data
- Dual Learning for Neural Machine Translation
- Back-Translation with Target Monolingual Data
- **Joint Training for S2T and T2S Models**

Joint Training for S2T and T2S Models

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

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

- Introduce source translations as hidden states for monolingual target sentences

$$\log p(y^{(t)}) = \log \sum_x Q(x) \frac{p(x, y^{(t)})}{Q(x)} \geq \sum_x Q(x) \log \frac{p(x, y^{(t)})}{Q(x)} \text{(Jensen's inequality)}$$

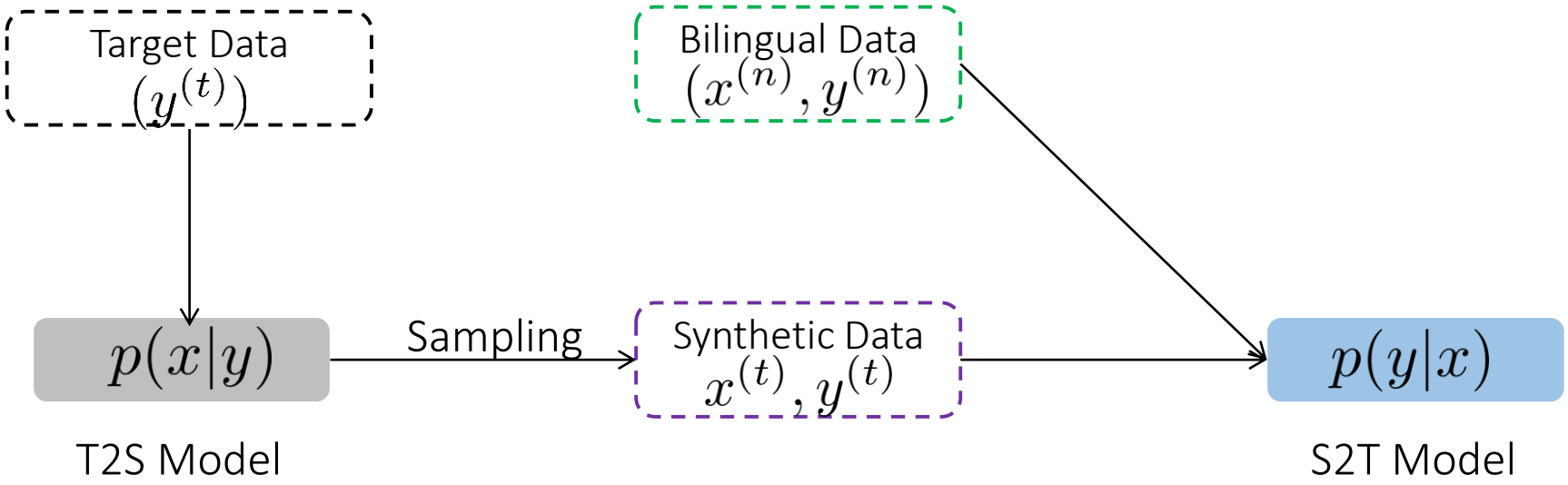
$$= \sum_x [Q(x) \log p(y^{(t)}|x) - KL(Q(x)||p(x))]$$

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

Zhang et al., **Joint Training for Neural Machine Translation Models with Monolingual Data**. AAAI, 2018.

Joint Training for S2T and T2S Models

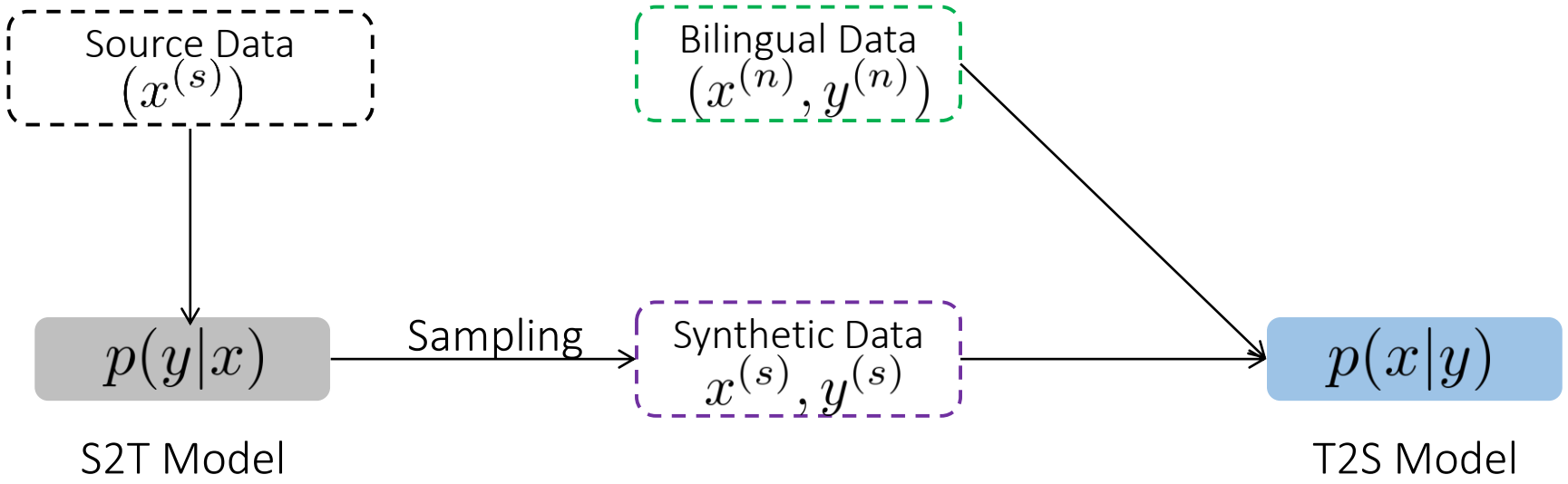
$$\frac{\partial L(\theta_{x\to y})}{\partial \theta_{x\to y}} = \sum_{n=1}^{N} \frac{\partial \log P(y^{(n)}|x^{(n)})}{\partial \theta_{x\to y}} + \sum_{t=1}^{T} E_{x^{(t)}\sim p(x|y^{(t)})} \frac{\partial \log P(y^{(t)}|x^{(t)})}{\partial \theta_{x\to y}}$$



Zhang et al., **Joint Training for Neural Machine Translation Models with Monolingual Data**. AAAI, 2018.

Joint Training for S2T and T2S Models

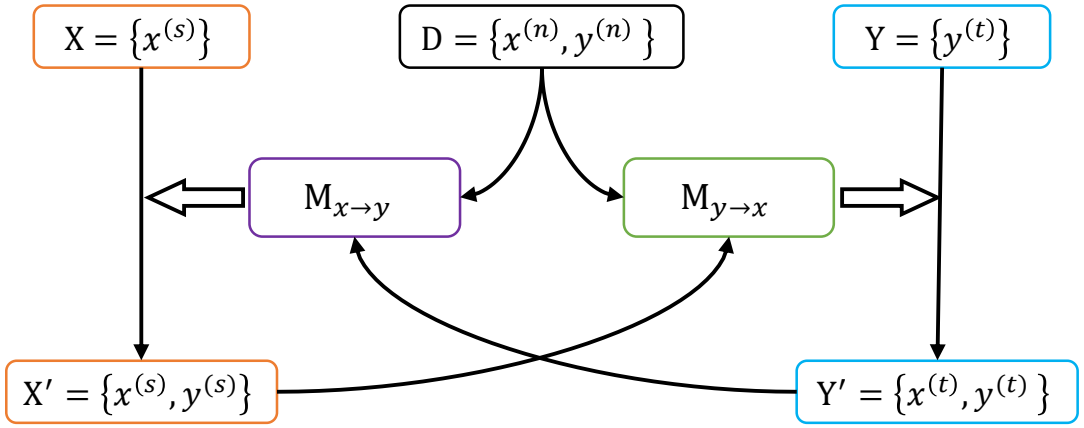
$$\frac{\partial L(\theta_{y\to x})}{\partial \theta_{y\to x}} = \sum_{n=1}^{N} \frac{\partial \log P(y^{(n)}|x^{(n)})}{\partial \theta_{y\to x}} + \sum_{s=1}^{S} E_{y^{(s)}\sim p(y|x^{(s)})} \frac{\partial \log P(x^{(s)}|y^{(s)})}{\partial \theta_{y\to x}}$$



Zhang et al., **Joint Training for Neural Machine Translation Models with Monolingual Data**. AAAI, 2018.

Joint Training for S2T and T2S Models

- Extend back-translation to paired one:
  - Jointly optimizing source-to-target NMT model $\mathrm{M}_{x\to y}$ and target-to-source NMT model $\mathrm{M}_{y\to x}$ with the help of monolingual data from both source language X and target language Y



Zhang et al., **Joint Training for Neural Machine Translation Models with Monolingual Data**. AAAI, 2018.

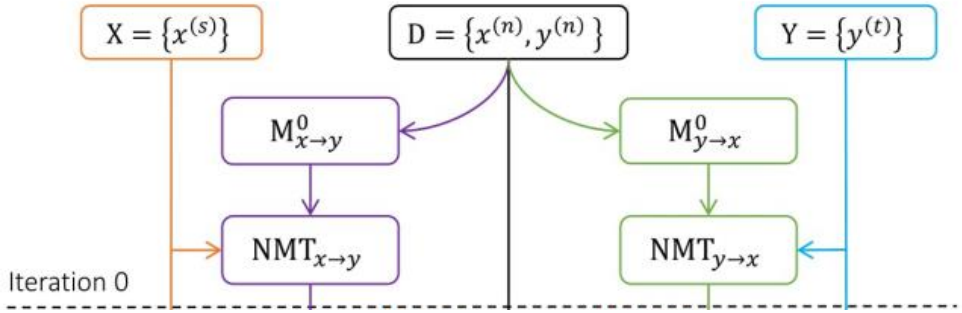
Joint Training for S2T and T2S Models

- **Iteration 0:** Pre-train two direction models $\mathrm{M}^{0}_{\mathrm{x}\to\mathrm{y}}$ and $\mathrm{M}^{0}_{\mathrm{y}\to\mathrm{x}}$ with bilingual data $\mathrm{D} = \{\mathrm{x}^{(\mathrm{n})}, \mathrm{y}^{(\mathrm{n})}\}$

- **Iteration 1:** Two NMT systems based on $\mathrm{M}^{0}_{\mathrm{x}\to\mathrm{y}}$ and $\mathrm{M}^{0}_{\mathrm{y}\to\mathrm{x}}$ are used to translate monolingual data $\mathrm{X} = \{\mathrm{x}^{(\mathrm{s})}\}$ and $\mathrm{Y} = \{\mathrm{y}^{(\mathrm{t})}\}$; two synthetic training data sets X' and Y' combined with bilingual data D are used to train $\mathrm{M}^{1}_{\mathrm{x}\to\mathrm{y}}$ and $\mathrm{M}^{1}_{\mathrm{y}\to\mathrm{x}}$ respectively.

- **Iteration 2:** Repeat the above process

- ….



Zhang et al., **Joint Training for Neural Machine Translation Models with Monolingual Data**. AAAI, 2018.

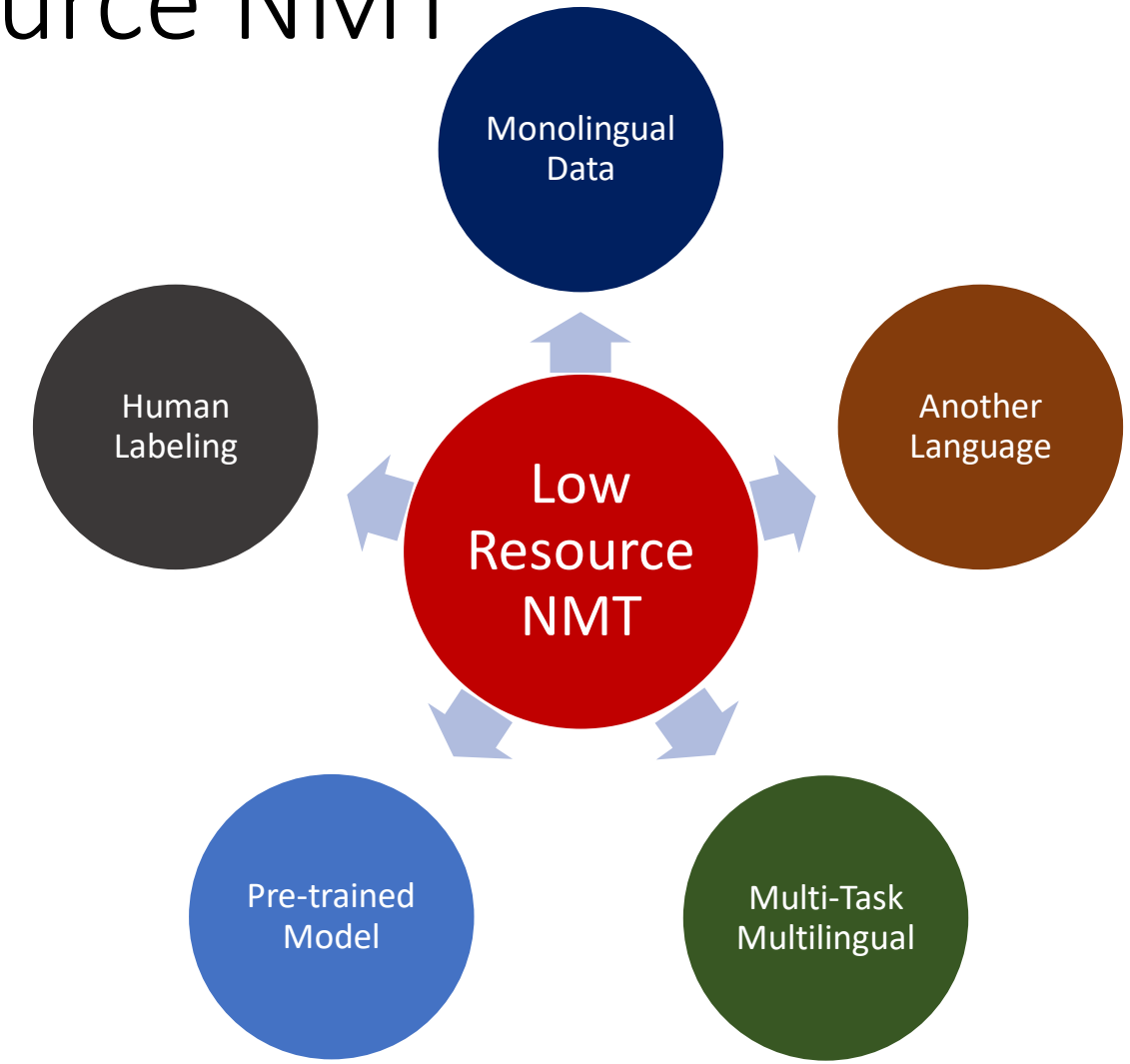
Joint Training for S2T and T2S Models

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	**42.56**	**45.10**	**44.36**	**34.10**	**32.26**	**39.67**
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	**22.56**	**22.98**	**20.95**	**17.62**	**17.39**	**20.30**

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.

Zhang et al., **Joint Training for Neural Machine Translation Models with Monolingual Data**. AAAI, 2018.

Low Resource NMT

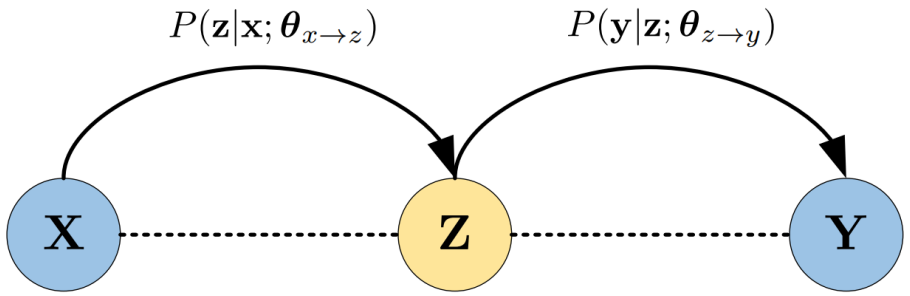


Microsoft

Tri-Language Learning for Low Resource NMT

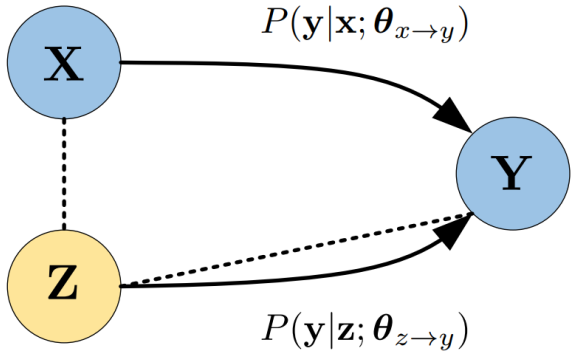
- **Teacher-Student Framework for Zero-Resource NMT**
- Triangular Architecture for Rare Language NMT
- Consistency by Agreement in Zero-shot NMT
- Generalized Data Augmentation for Low-Resource Translation

Teacher-Student Framework for Zero-Resource NMT



(a)

Pivot-based Method



(b)

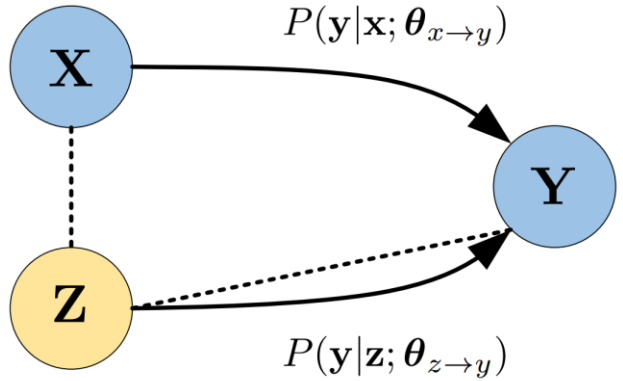
Teacher Student Method

Chen et al., **A Teacher-Student Framework for Zero-Resource Neural Machine Translation**. ACL, 2017.

Teacher-Student Framework for Zero-Resource NMT

Assumption 1 *If a source sentence* $\mathbf{x}$ *is a translation of a pivot sentence* $\mathbf{z}$*, then the probability of generating a target sentence* $\mathbf{y}$ *from* $\mathbf{x}$ *should be close to that from its counterpart* $\mathbf{z}$*.*

$$
\begin{aligned}
& \mathcal{J}_{\mathrm{SENT}}(\boldsymbol{\theta}_{x\to y}) \\
= & \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathrm{KL}\Big(P(\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}) \Big|\Big| P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y}) \Big) \\
= & -\sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}} \Big[\log P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y}) \Big]
\end{aligned}
$$

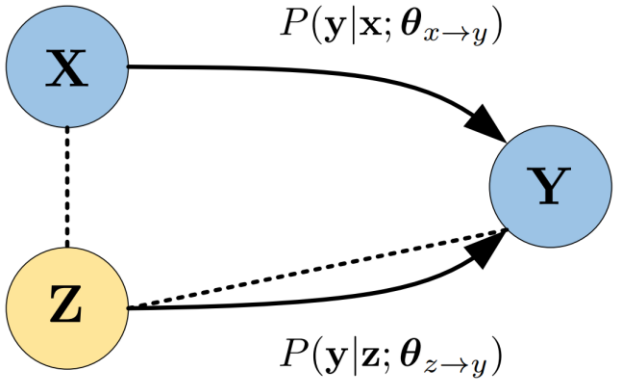


Chen et al., **A Teacher-Student Framework for Zero-Resource Neural Machine Translation**. ACL, 2017.

Teacher-Student Framework for Zero-Resource NMT

Assumption 2 *If a source sentence* $\mathbf{x}$ *is a translation of a pivot sentence* $\mathbf{z}$*, then the probability of generating a target word* y *from* $\mathbf{x}$ *should be close to that from its counterpart* $\mathbf{z}$*, given the already obtained partial translation* $\mathbf{y}_{<j}$*.*

$$
\begin{aligned}
&\mathcal{J}_{\text{WORD}}(\boldsymbol{\theta}_{x\to y}) \\
&= \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}} \Big[\sum_{j=1}^{|\mathbf{y}|} \mathrm{KL}\Big(P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\boldsymbol{\theta}}_{z\to y}) \Big|\Big| P(y|\mathbf{x},\mathbf{y}_{<j};\boldsymbol{\theta}_{x\to y}) \Big) \Big] \\
&= \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}} \Big[\sum_{j=1}^{|\mathbf{y}|} \sum_{y\in\mathcal{V}_y} P(y|\mathbf{z},\mathbf{y}_{<j};\hat{\boldsymbol{\theta}}_{z\to y}) \times \log P(y|\mathbf{x},\mathbf{y}_{<j};\boldsymbol{\theta}_{x\to y}) \Big]
\end{aligned}
$$



Chen et al., **A Teacher-Student Framework for Zero-Resource Neural Machine Translation**. ACL, 2017.

Teacher-Student Framework for Zero-Resource NMT

Method	Es→ Fr	De→ Fr
pivot	29.79	23.70
sent-beam	31.64	24.39
word-sampling	33.86	27.03

Chen et al., **A Teacher-Student Framework for Zero-Resource Neural Machine Translation**. ACL, 2017.

Microsoft

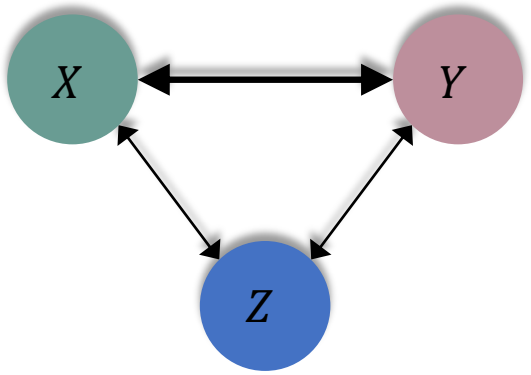
Tri-Language Learning for Low Resource NMT

- Teacher-Student Framework for Zero-Resource NMT
- **Triangular Architecture for Rare Language NMT**
- Consistency by Agreement in Zero-shot NMT
- Generalized Data Augmentation for Low-Resource Translation

Triangular Architecture for Rare Language NMT

Problem

- Large bilingual corpus (X, Y) between rich languages X and Y.
- Small bilingual corpus (X, Z) and (Z, Y) between rare language Z and rich languages X and Y.



X: English
Y: French
Z: Hebrew

Method

- Language Z is used as a hidden space to translate a sentence from language X to language Y, and from Y to X.

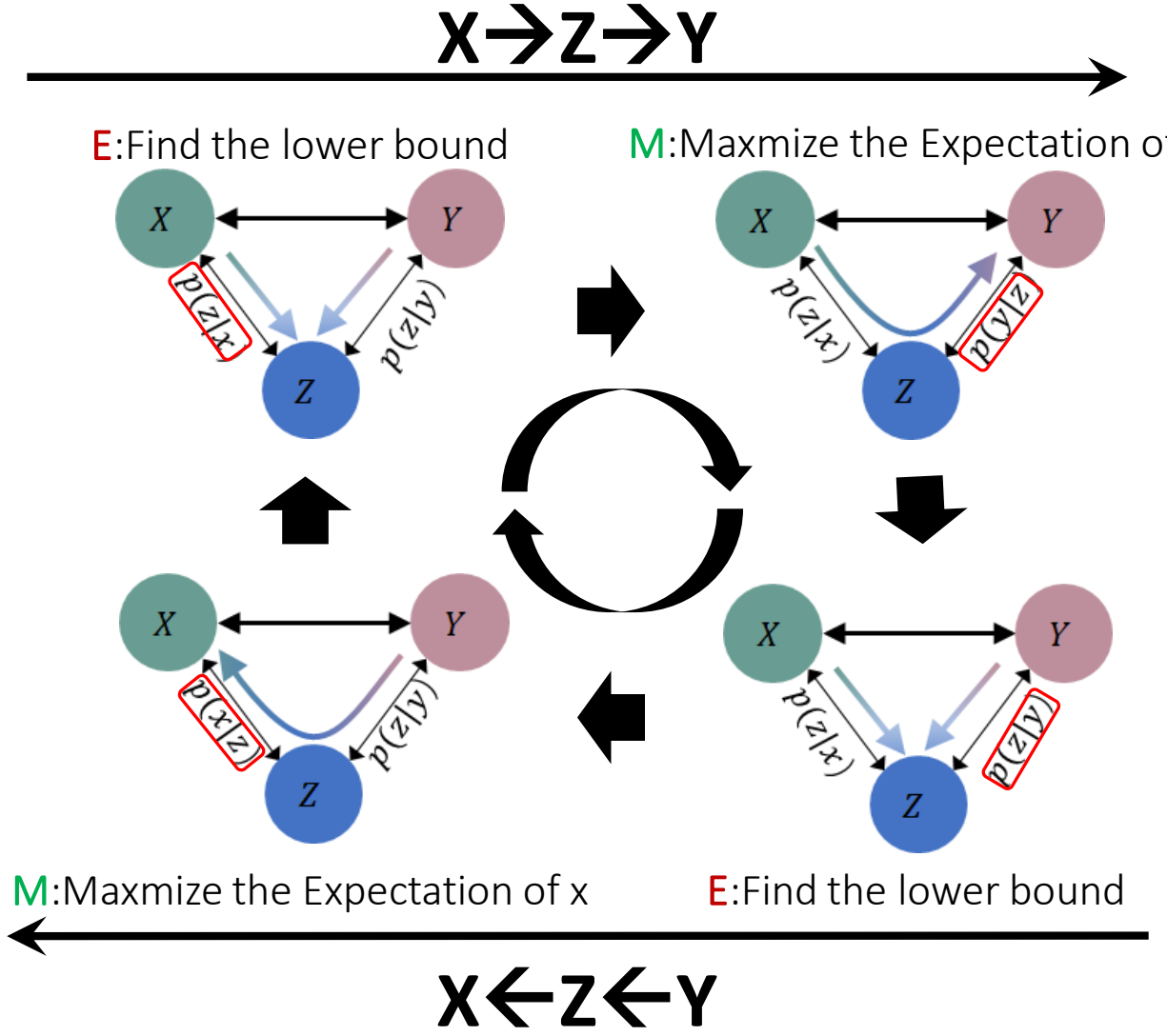
$$p(y|x) = \sum_{z \in Z} p(z|x)p(y|z)$$

$$p(x|y) = \sum_{z \in Z} p(z|y)p(x|z)$$

- Small bilingual corpora are used to initialize $p(z|x), p(y|z), p(z|y)$ and $p(x|z)$.
- EM training is leveraged for fine tuning.

Ren et al., **Triangular Architecture for Rare Language Translation**. ACL, 2018.

Triangular Architecture for Rare Language NMT

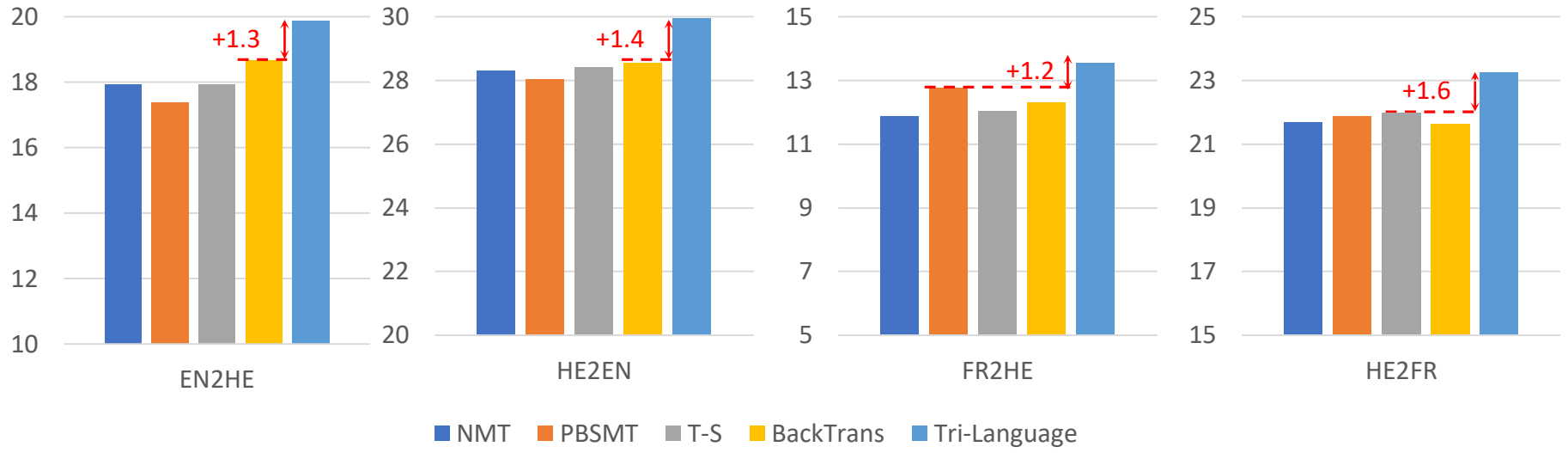


E-Step of X→Z→Y:
Update $p(z|x)$ using $p(z|y)$
M-Step of X→Z→Y:
Update $p(y|z)$ using $p(z|x)$
E-Step of Y→Z→X:
Update $p(z|y)$ using $p(z|x)$
M-Step of X→Z→Y:
Update $p(x|z)$ using $p(z|y)$
.....

Ren et al., **Triangular Architecture for Rare Language Translation**. ACL, 2018.

Triangular Architecture for Rare Language NMT

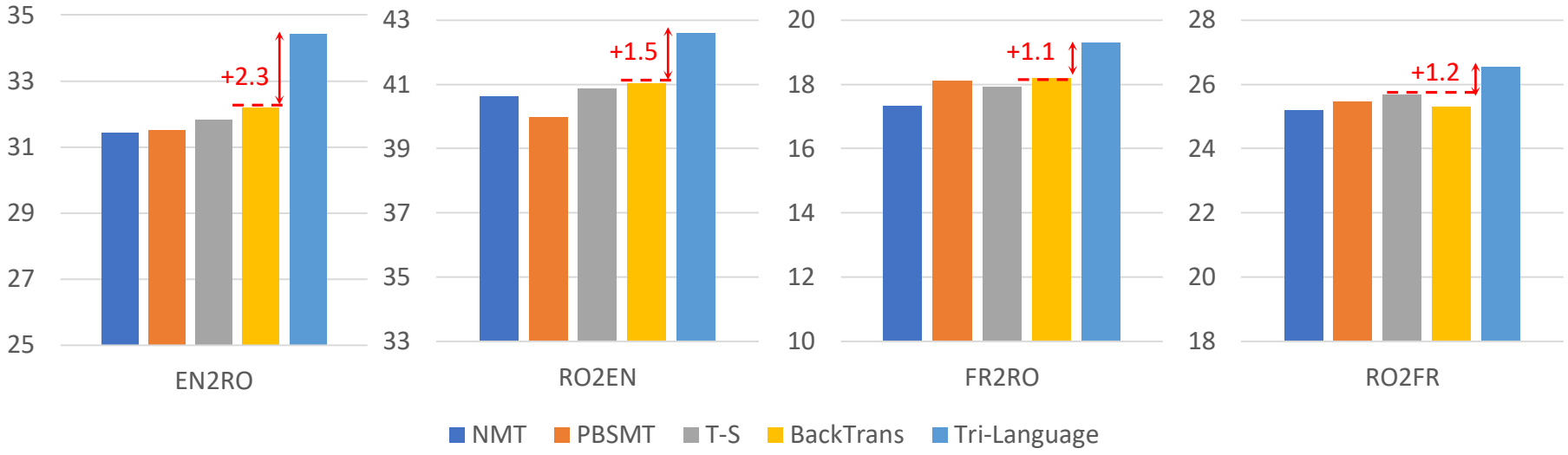
- Experiment Setting: Single layer GRU-based NMT, with hidden size 512, embedding size 256.
- Baseline
  - NMT: NMT system trained only with bilingual data.
  - PBSMT: Phrase based statistical machine translation system.
- Comparison System
  - T-S: Teach student training method.
  - BackTrans: Back translation with monolingual data.
- Data: IWSLT (EN-FR:7.9M, EN-HE:112.6k, FR-HE:116.3K, HE:512.5K)



Ren et al., **Triangular Architecture for Rare Language Translation**. ACL, 2018.

Triangular Architecture for Rare Language NMT

- Experiment Setting: Single layer GRU-based NMT, with hidden size 512, embedding size 256.
- Baseline
  - NMT: NMT system trained only with bilingual data.
  - PBSMT: Phrase based statistical machine translation system.
- Comparison System
  - T-S: Teach student training method.
  - BackTrans: Back translation with monolingual data.
- Data: IWSLT (EN-FR:7.9M, EN-RO:467.3k, FR-RO:111.6K, RO:885.5K)



Ren et al., **Triangular Architecture for Rare Language Translation**. ACL, 2018.

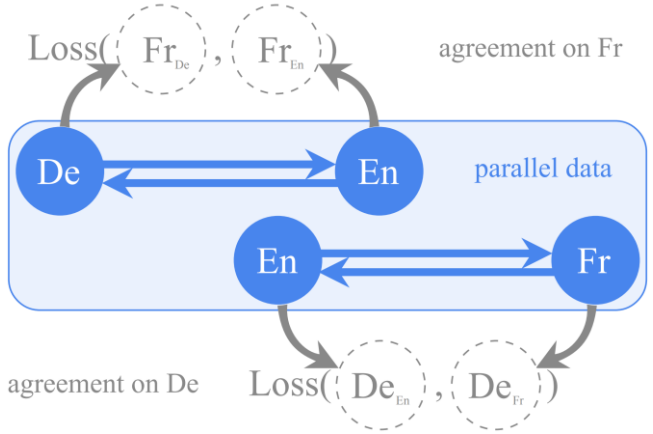
Microsoft

Tri-Language Learning for Low Resource NMT

- Teacher-Student Framework for Zero-Resource NMT
- Triangular Architecture for Rare Language NMT
- **Consistency by Agreement in Zero-shot NMT**
- Generalized Data Augmentation for Low-Resource Translation

Consistency by Agreement in Zero-shot NMT

- Motivation
  - The performance on zero-shot tasks is often unstable and significantly lags behind the supervised directions.
  - Traditional multilingual objective misses the terms that correspond to zero-shot models, and hence has no statistical guarantee for performance on zero-shot tasks
- Method
  - Agreement-based learning:



$$\mathcal{L}^{\text{agree}}_{\text{EnFr}}(\theta) \approx \underbrace{\log \mathbb{P}_\theta\left(\mathbf{x}_{\text{Fr}} \mid \mathbf{x}_{\text{En}}\right) + \log \mathbb{P}_\theta\left(\mathbf{x}_{\text{En}} \mid \mathbf{x}_{\text{Fr}}\right)}_{\text{composite likelihood terms}} + \underbrace{\log \sum_{\mathbf{z}} \mathbb{P}_\theta\left(\mathbf{z} \mid \mathbf{x}_{\text{En}}\right) \mathbb{P}_\theta\left(\mathbf{z} \mid \mathbf{x}_{\text{Fr}}\right)}_{\text{agreement term}}$$

Al-Shedivat and P. Parikh, **Consistency by Agreement in Zero-shot Neural Machine Translation** . NAACL, 2019.

Consistency by Agreement in Zero-shot NMT

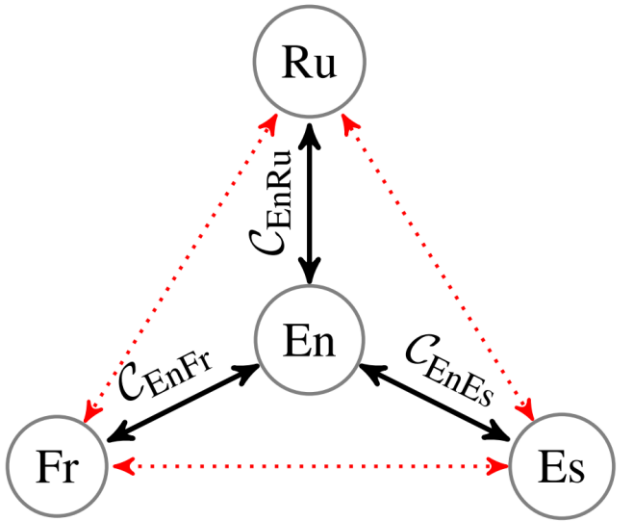


Figure 2: Translation graph: Languages (nodes), parallel corpora (solid edges), and zero-shot directions (dotted edges).

Algorithm 1 Agreement-based M-NMT training

input Architecture (GNMT), agreement coefficient (γ)

Initialize: $\theta \leftarrow \theta_0$

while not (converged or step limit reached) **do**

  Get a mini-batch of parallel src-tgt pairs, $(\mathbf{X}_s, \mathbf{X}_t)$

  Supervised loss: $\mathcal{L}^{\text{sup}}(\theta) \leftarrow \log \mathbb{P}_\theta\left(\mathbf{X}_t \mid \mathbf{X}_s\right)$

  Auxiliary languages: $L_a \sim \text{Unif}(\{1, \ldots, k\})$

  Auxiliary translations:

$$\mathbf{Z}_{a \leftarrow s} \leftarrow \text{Decode}\left(\mathbf{Z}_a \mid f_\theta^{\text{enc}}(\mathbf{X}_s, L_a)\right)$$
$$\mathbf{Z}_{a \leftarrow t} \leftarrow \text{Decode}\left(\mathbf{Z}_a \mid f_\theta^{\text{enc}}(\mathbf{X}_t, L_a)\right)$$

  Agreement log-probabilities:

$$\ell^t_{a \leftarrow s} \leftarrow \log \mathbb{P}_\theta\left(\mathbf{Z}_{a \leftarrow s} \mid \mathbf{X}_t\right)$$
$$\ell^s_{a \leftarrow t} \leftarrow \log \mathbb{P}_\theta\left(\mathbf{Z}_{a \leftarrow t} \mid \mathbf{X}_s\right)$$

  Apply stop-gradients to supervised $\ell^t_{a \leftarrow s}$ and $\ell^s_{a \leftarrow t}$

  Total loss: $\mathcal{L}^{\text{total}}(\theta) \leftarrow \mathcal{L}^{\text{sup}}(\theta) + \gamma(\ell^t_{a \leftarrow s} + \ell^s_{a \leftarrow t})$

  Update: $\theta \leftarrow \text{optimizer\_update}(\mathcal{L}^{\text{total}}, \theta)$

end while

output θ

Al-Shedivat and P. Parikh, **Consistency by Agreement in Zero-shot Neural Machine Translation** . NAACL, 2019.

Consistency by Agreement in Zero-shot NMT

	Sestorain et al. (2018)†			Our baselines		
	PBSMT	NMT-0	Dual-0	Basic	Pivot	Agree
En → Es	61.26	51.93	—	56.58	56.58	56.36
En → Fr	50.09	40.56	—	44.27	44.27	44.80
Es → En	59.89	51.58	—	55.70	55.70	55.24
Fr → En	52.22	43.33	—	46.46	46.46	46.17
Supervised (avg.)	55.87	46.85	—	50.75	50.75	50.64
Es → Fr	52.44	20.29	36.68	34.75	**38.10**	37.54
Fr → Es	49.79	19.01	39.19	37.67	**40.84**	40.02
Zero-shot (avg.)	51.11	19.69	37.93	36.21	**39.47**	38.78

†Source: https://openreview.net/forum?id=ByecAoAqK7.

Table 1: Results on UNCorpus-1.

	Sestorain et al. (2018)			Our baselines		
	PBSMT	NMT-0	Dual-0	Basic	Pivot	Agree
En → Es	61.26	47.51	44.30	55.15	55.15	54.30
En → Fr	50.09	36.70	34.34	43.42	43.42	42.57
En → Ru	43.25	30.45	29.47	36.26	36.26	35.89
Es → En	59.89	48.56	45.55	54.35	54.35	54.33
Fr → En	52.22	40.75	37.75	45.55	45.55	45.87
Ru → En	52.59	39.35	37.96	45.52	45.52	44.67
Supervised (avg.)	53.22	40.55	36.74	46.71	46.71	46.27
Es → Fr	52.44	25.85	34.51	34.73	35.93	**36.02**
Fr → Es	49.79	22.68	37.71	38.20	39.51	**39.94**
Es → Ru	39.69	9.36	24.55	26.29	27.15	**28.08**
Ru → Es	49.61	26.26	33.23	33.43	**37.17**	35.01
Fr → Ru	36.48	9.35	22.76	23.88	24.99	**25.13**
Ru → Fr	43.37	22.43	26.49	28.52	**30.06**	29.53
Zero-shot (avg.)	45.23	26.26	29.88	30.84	**32.47**	32.29

Table 2: Results on UNCorpus-2.

	Previous work		Our baselines		
	Soft‡	Distill†	Basic	Pivot	Agree
En → Es	—	—	34.69	34.69	33.80
En → De	—	—	23.06	23.06	22.44
En → Fr	31.40	—	33.87	33.87	32.55
Es → En	31.96	—	34.77	34.77	34.53
De → En	26.55	—	29.06	29.06	29.07
Fr → En	—	—	33.67	33.67	33.30
Supervised (avg.)	—	—	31.52	31.52	30.95
Es → De	—	—	18.23	20.14	**20.70**
De → Es	—	—	20.28	**26.50**	22.45
Es → Fr	30.57	**33.86**	27.99	32.56	30.94
Fr → Es	—	—	27.12	**32.96**	29.91
De → Fr	23.79	**27.03**	21.36	25.67	24.45
Fr → De	—	—	18.57	**19.86**	19.15
Zero-shot (avg.)	—	—	22.25	26.28	24.60

†Soft pivoting (Cheng et al., 2017). ‡Distillation (Chen et al., 2017).

Table 3: Zero-shot results on Europarl. Note that *Soft* and *Distill* are not multilingual systems.

	Previous work		Our baselines		
	SOTA†	CPG‡	Basic	Pivot	Agree
Supervised (avg.)	24.10	19.75	24.63	24.63	23.97
Zero-shot (avg.)	20.55	11.69	19.86	19.26	**20.58**

†Table 2 from Dabre et al. (2017). ‡Table 2 from Platanios et al. (2018).

Table 4: Results on the official IWSLT17 multilingual task.

	Basic	Pivot	Agree
Supervised (avg.)	28.72	28.72	**29.17**
Zero-shot (avg.)	12.61	**17.68**	15.23

Table 5: Results on our proposed IWSLT17⋆.

Al-Shedivat and P. Parikh, **Consistency by Agreement in Zero-shot Neural Machine Translation** . NAACL, 2019.

Microsoft

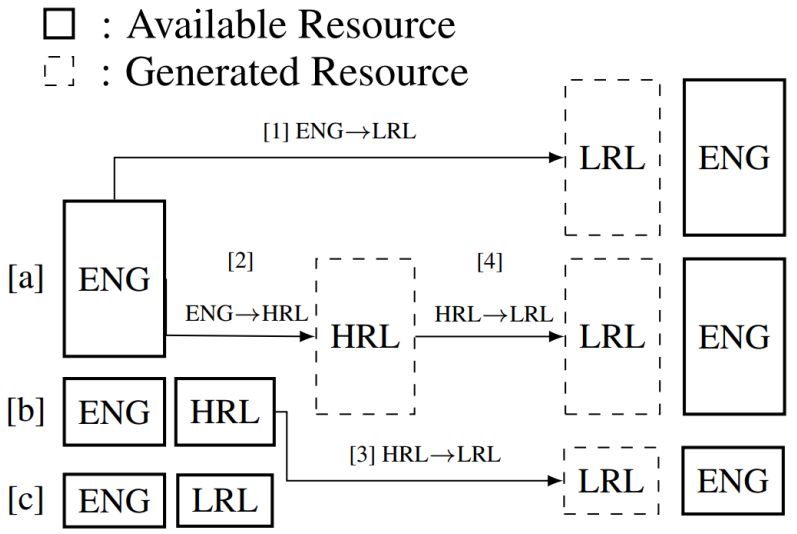
Tri-Language Learning for Low Resource NMT

- Teacher-Student Framework for Zero-Resource NMT
- Triangular Architecture for Rare Language NMT
- Consistency by Agreement in Zero-shot NMT
- **Generalized Data Augmentation for Low-Resource Translation**

Generalized Data Augmentation for Low-Resource NMT

- Data Available
  - $\{S_{LE}, T_{LE}\}$, $\{S_{HE}, T_{HE}\}$, $\{S_{HL}, T_{HL}\}$, M_L, M_H, M_E
- Four types of augmented parallel data:
  - ENG-LRL: $\{S_{E\to L}, M_E\}$
  - ENG-HRL: $\{S_{E\to H}, M_E\}$
  - HRL-LRL: $\{S_{H\to L}, T_{HE}\}$
  - ENG-HRL-LRL: $\{S_{E\to H\to L}, M_E\}$
- LRL-HRL translation models
  - Convert the words of HRL words in $\{S_{HE}, T_{HE}\}$ and $\{S_{E\to H}, M_E\}$ into LRL words with unsupervised BLI to get pseudo LRL data
  - Translate pseudo LRL data to true LRL data with UMT system (trained with $\{\text{pseudo } M_L, M_L\}$)

Datasets	LRL (HRL) AZE (TUR)	BEL (RUS)	GLG (POR)	SLK (CES)
$\mathcal{S}_{\text{LE}}, \mathcal{T}_{\text{LE}}$	5.9K	4.5K	10K	61K
$\mathcal{S}_{\text{HE}}, \mathcal{T}_{\text{HE}}$	182K	208K	185K	103K
$\mathcal{S}_{\text{LH}}, \mathcal{T}_{\text{LH}}$	5.7K	4.2K	3.8K	44K
$\mathcal{M}_{\text{L}}$	2.02M	1.95M	1.98M	2M
$\mathcal{M}_{\text{H}}$	2M	2M	2M	2M
$\mathcal{M}_{\text{E}}$	2M/ 200K			



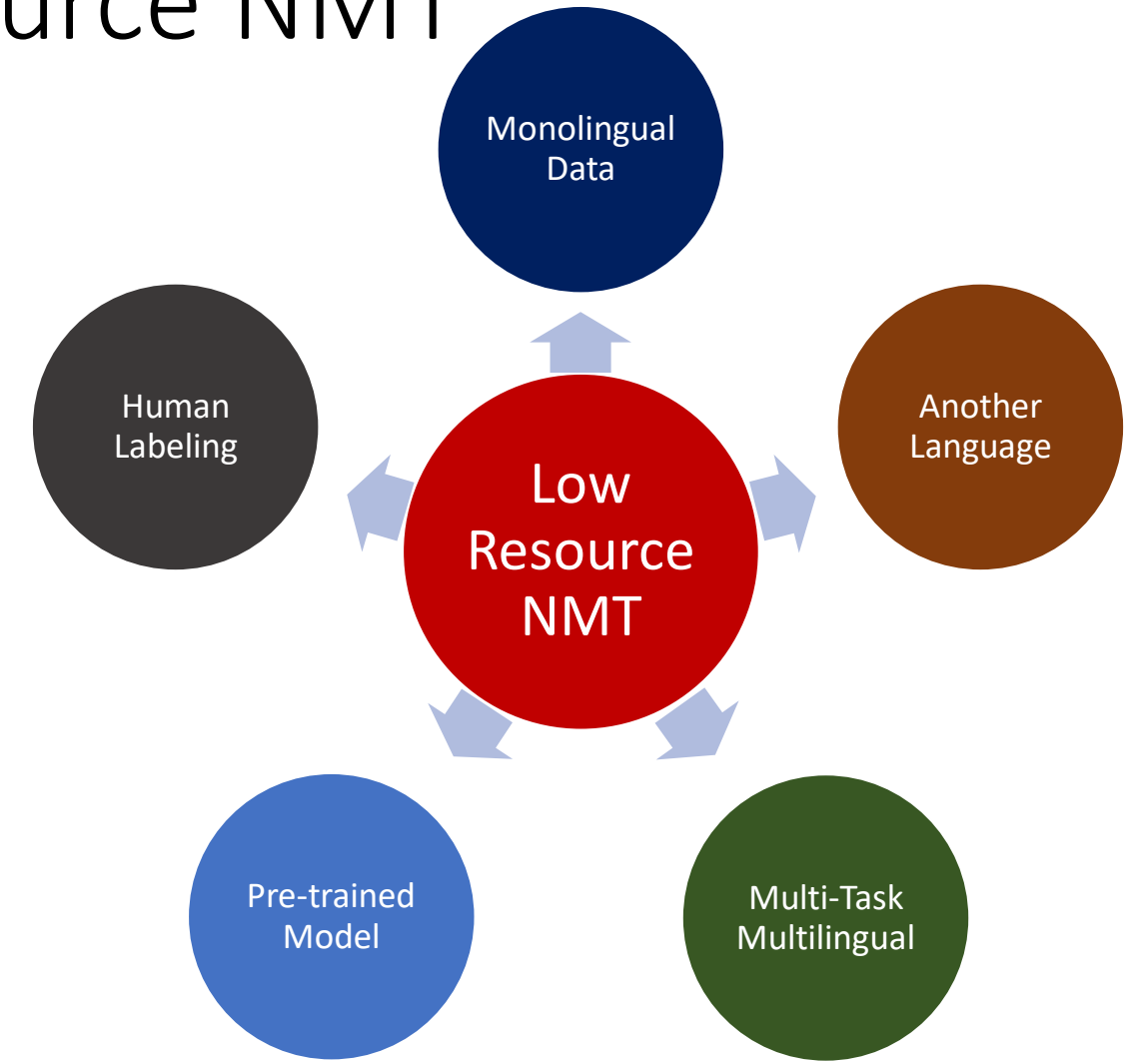
Xia et al., **Generalized Data Augmentation for Low-resource Translation** . ACL, 2019.

Generalized Data Augmentation for Low-Resource NMT

	Training Data		BLEU for X→ENG AZE (TUR)	BEL (RUS)	GLG (POR)	SLK (CES)
Results from Literature						
	SDE (Wang et al., 2019)		12.89	18.71	31.16	29.16
	many-to-many (Aharoni et al., 2019)		12.78	21.73	30.65	29.54
Standard NMT						
1	$\{\mathcal{S}_{\text{LE}}\mathcal{S}_{\text{HE}}\ ,\ \mathcal{T}_{\text{LE}}\mathcal{T}_{\text{HE}}\}$	(supervised MT)	11.83	16.34	29.51	28.12
2	$\{\mathcal{M}_{\text{L}}, \mathcal{M}_{\text{E}}\}$	(unsupervised MT)	0.47	0.18	1.15	0.75
Standard Supervised Back-translation						
3	$+\{\hat{\mathcal{S}}^{s}_{\text{E}\to\text{L}}\ ,\ \mathcal{M}_{\text{E}}\}$		11.84	15.72	29.19	29.79
4	$+\{\hat{\mathcal{S}}^{s}_{\text{E}\to\text{H}}\ ,\ \mathcal{M}_{\text{E}}\}$		12.46	16.40	30.07	30.60
Augmentation from HRL-ENG						
5	$+\{\hat{\mathcal{S}}^{s}_{\text{H}\to\text{L}}\ ,\ \mathcal{T}_{\text{HE}}\}$	(supervised MT)	11.92	15.79	29.91	28.52
6	$+\{\hat{\mathcal{S}}^{u}_{\text{H}\to\text{L}}\ ,\ \mathcal{T}_{\text{HE}}\}$	(unsupervised MT)	11.86	13.83	29.80	28.69
7	$+\{\hat{\mathcal{S}}^{w}_{\text{H}\to\text{L}}\ ,\ \mathcal{T}_{\text{HE}}\}$	(word subst.)	14.87	23.56	32.02	29.60
8	$+\{\hat{\mathcal{S}}^{m}_{\text{H}\to\text{L}}\ ,\ \mathcal{T}_{\text{HE}}\}$	(modified UMT)	14.72	23.31	32.27	29.55
9	$+\{\hat{\mathcal{S}}^{w}_{\text{H}\to\text{L}}\hat{\mathcal{S}}^{m}_{\text{H}\to\text{L}}\ ,\ \mathcal{T}_{\text{HE}}\mathcal{T}_{\text{HE}}\}$		15.24	**24.25**	32.30	30.00
Augmention from ENG by pivoting						
10	$+\{\hat{\mathcal{S}}^{w}_{\text{E}\to\text{H}\to\text{L}}\ ,\ \mathcal{M}_{\text{E}}\}$	(word subst.)	14.18	21.74	31.72	30.90
11	$+\{\hat{\mathcal{S}}^{m}_{\text{E}\to\text{H}\to\text{L}}\ ,\ \mathcal{M}_{\text{E}}\}$	(modified UMT)	13.71	19.94	31.39	30.22
Combinations						
12	$+\{\hat{\mathcal{S}}^{w}_{\text{H}\to\text{L}}\hat{\mathcal{S}}^{w}_{\text{E}\to\text{H}\to\text{L}}\ ,\ \mathcal{T}_{\text{HE}}\mathcal{M}_{\text{E}}\}$	(word subst.)	**15.74**	**24.51**	**33.16**	**32.07**
13	$+\{\hat{\mathcal{S}}^{w}_{\text{H}\to\text{L}}\hat{\mathcal{S}}^{m}_{\text{H}\to\text{L}}\ ,\ \mathcal{T}_{\text{HE}}\mathcal{T}_{\text{HE}}\}$ $+\{\hat{\mathcal{S}}^{w}_{\text{E}\to\text{H}\to\text{L}}\hat{\mathcal{S}}^{m}_{\text{E}\to\text{H}\to\text{L}}\ ,\ \mathcal{M}_{\text{E}}\mathcal{M}_{\text{E}}\}$		**15.91**	23.69	32.55	31.58

Xia et al., **Generalized Data Augmentation for Low-resource Translation** . ACL, 2019.

Low Resource NMT

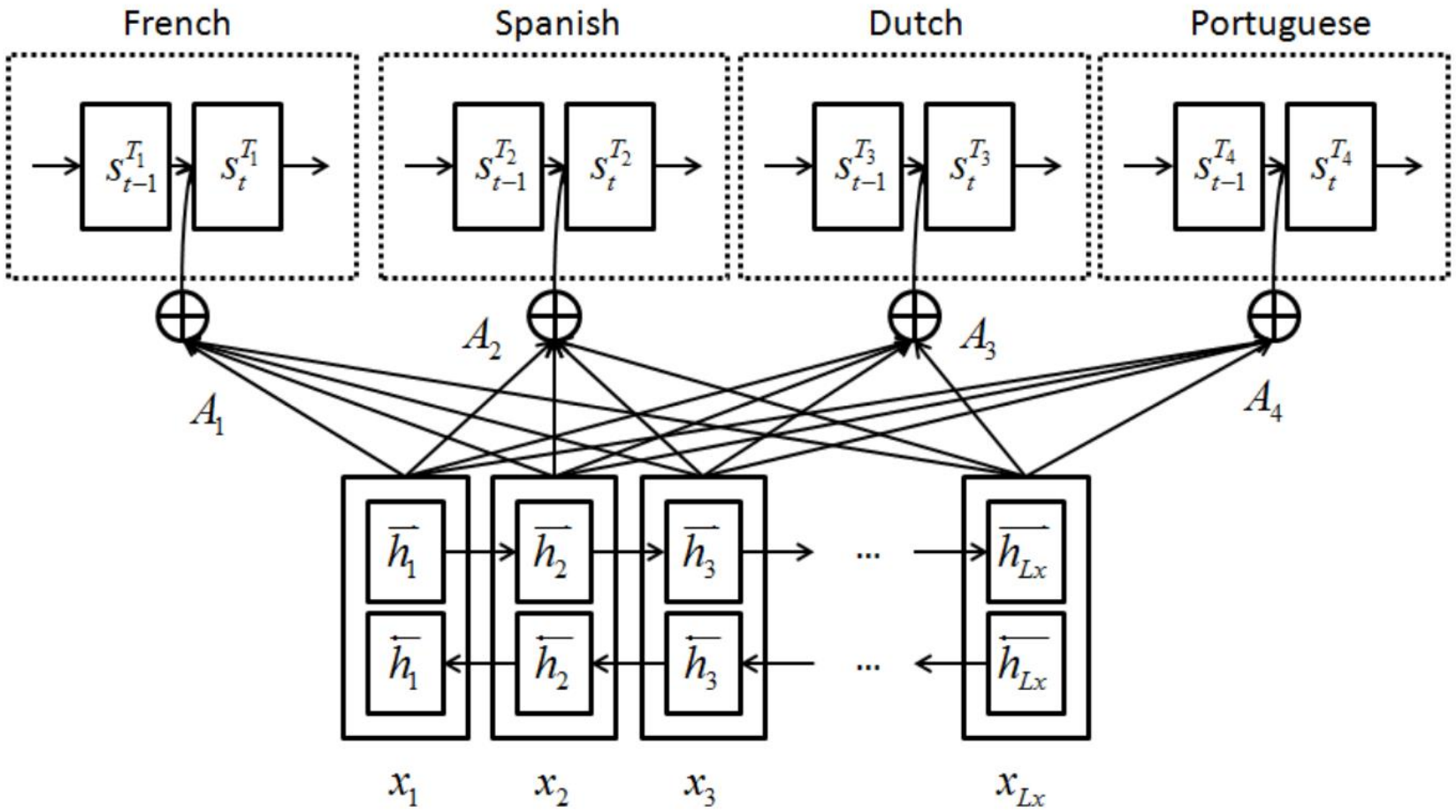


Microsoft

Multi-Task Learning for Low Resource NMT

- **Multi-Task Learning for Multiple Language Translation**
- Google's Multilingual Neural Machine Translation System
- Universal Neural Machine Translation for Extremely Low Resource Languages
- Agreement Regularization of L2R and R2L Models

Multi-Task Learning for Multiple Language Translation



Dong et al., **Multi-Task Learning for Multiple Language Translation**. ACL, 2015.

Multi-Task Learning for Multiple Language Translation

Training Data Information						
Lang	En-Es	En-Fr	En-Nl	En-Pt	En-Nl-sub	En-Pt-sub
Sent size	1,965,734	2,007,723	1,997,775	1,960,407	300,000	300,000
Src tokens	49,158,635	50,263,003	49,533,217	49,283,373	8,362,323	8,260,690
Trg tokens	51,622,215	52,525,000	50,661,711	54,996,139	8,590,245	8,334,454

Table 1: Size of training corpus for different language pairs

Lang-Pair	En-Es	En-Fr	En-Nl*	En-Pt*
Single NMT	26.65	21.22	26.59	18.26
Multi Task	28.29	21.89	27.85	19.32
Delta	**+1.64**	**+0.67**	**+1.26**	**+1.06**

Table 4: Multi-task neural translation v.s. single model with a small-scale training corpus on some language pairs. * means that the language pair is sub-sampled.

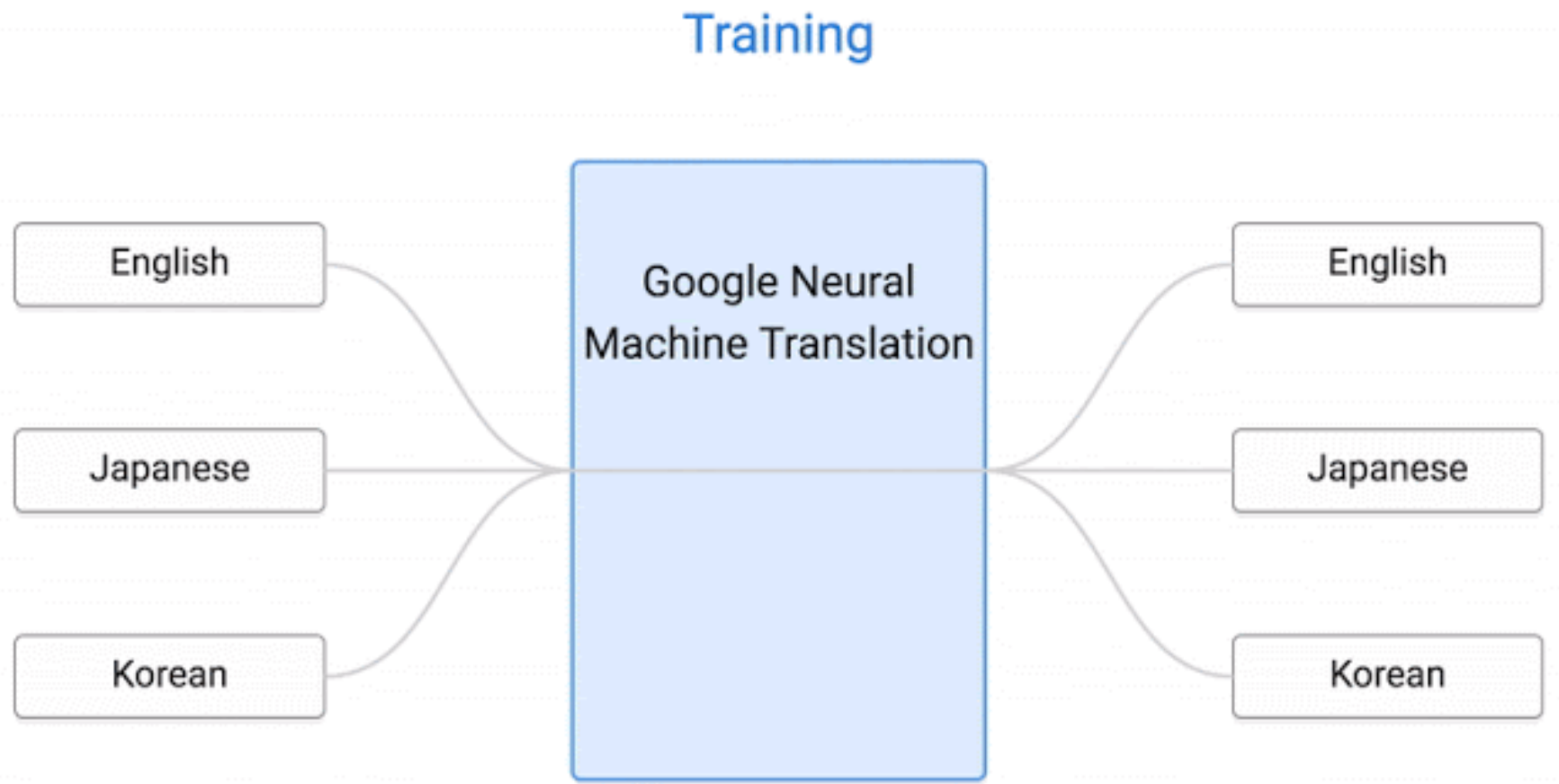
Dong et al., **Multi-Task Learning for Multiple Language Translation**. ACL, 2015.

Microsoft

Multi-Task Learning for Low Resource NMT

- Multi-Task Learning for Multiple Language Translation
- **Google's Multilingual Neural Machine Translation System**
- Universal Neural Machine Translation for Extremely Low Resource Languages
- Agreement Regularization of L2R and R2L Models

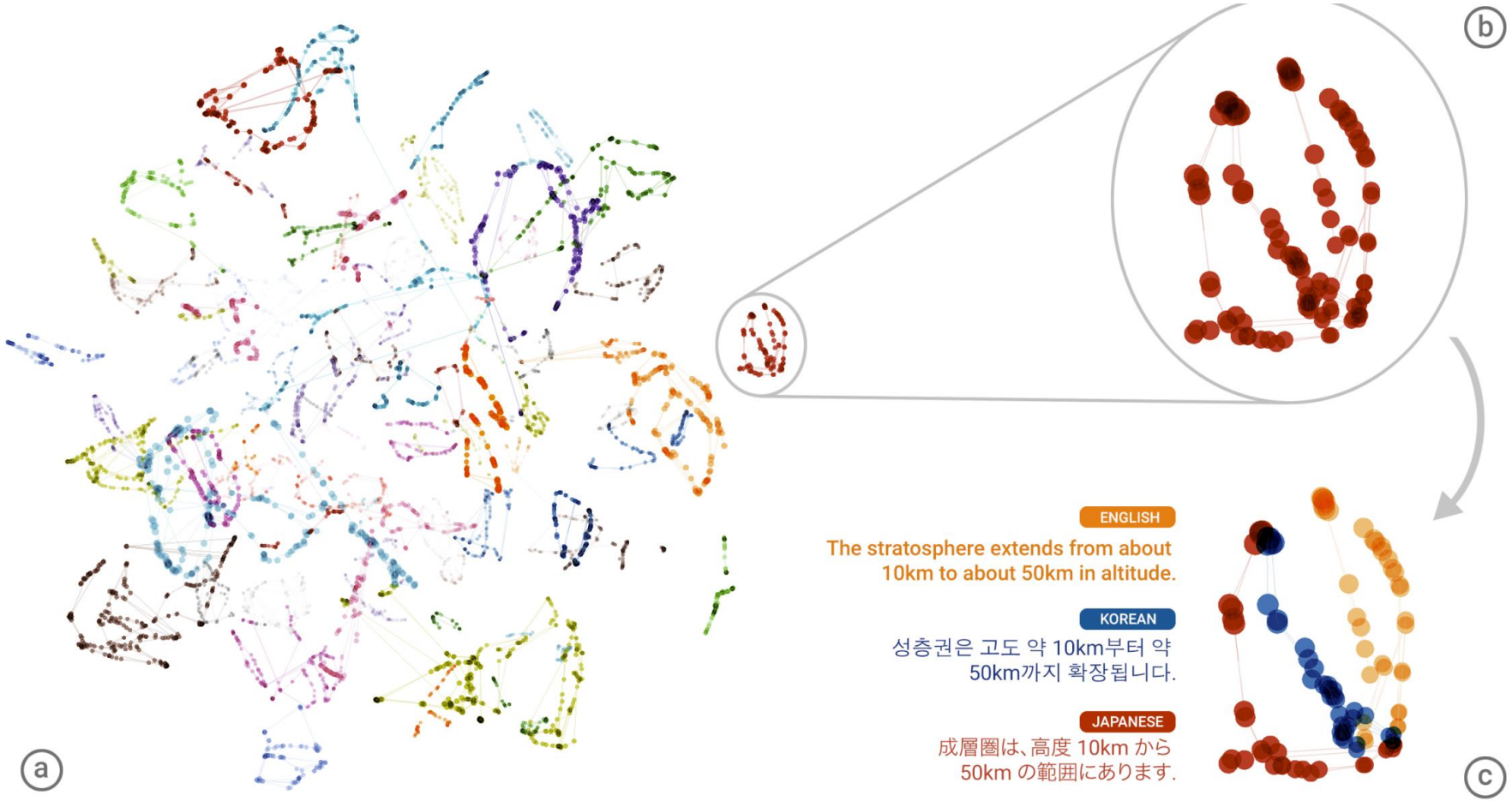
Google’s Multilingual Neural Machine Translation System



https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html

Johnson et al., **Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation**. TACL, 2017.

Google’s Multilingual Neural Machine Translation System



Johnson et al., **Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation**. TACL, 2017.

Google's Multilingual Neural Machine Translation System

Table 4: Large-scale experiments: BLEU scores for single language pair and multilingual models.

Model	Single	Multi	Multi	Multi	Multi
#nodes	1024	1024	1280	1536	1792
#params	3B	255M	367M	499M	650M
En→Ja	23.66	21.10	21.17	21.72	21.70
En→Ko	19.75	18.41	18.36	18.30	18.28
Ja→En	23.41	21.62	22.03	22.51	23.18
Ko→En	25.42	22.87	23.46	24.00	24.67
En→Es	34.50	34.25	34.40	34.77	34.70
En→Pt	38.40	37.35	37.42	37.80	37.92
Es→En	38.00	36.04	36.50	37.26	37.45
Pt→En	44.40	42.53	42.82	43.64	43.87
En→De	26.43	23.15	23.77	23.63	24.01
En→Fr	35.37	34.00	34.19	34.91	34.81
De→En	31.77	31.17	31.65	32.24	32.32
Fr→En	36.47	34.40	34.56	35.35	35.52
ave diff	-	-1.72	-1.43	-0.95	-0.76
vs single	-	-5.6%	-4.7%	-3.1%	-2.5%

Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	Zero-shot	BLEU
(a)	PBMT bridged	no	28.99
(b)	NMT bridged	no	30.91
(c)	NMT Pt→Es	no	31.50
(d)	Model 1 (Pt→En, En→Es)	yes	21.62
(e)	Model 2 (En↔{Es, Pt})	yes	24.75
(f)	Model 2 + incremental training	no	31.77

Johnson et al., **Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation**. TACL, 2017.

Microsoft

Multi-Task Learning for Low Resource NMT

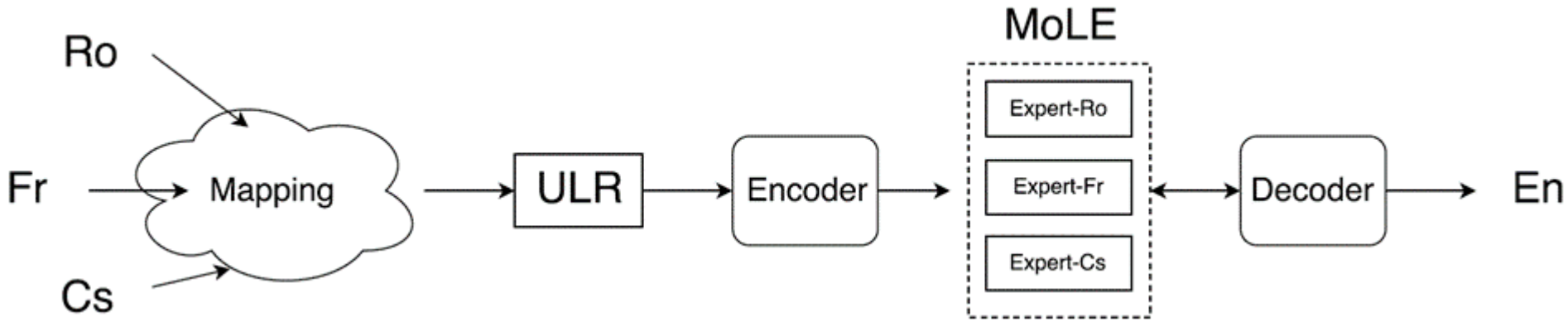
- Multi-Task Learning for Multiple Language Translation
- Google's Multilingual Neural Machine Translation System
- **Universal Neural Machine Translation for Extremely Low Resource Languages**
- Agreement Regularization of L2R and R2L Models

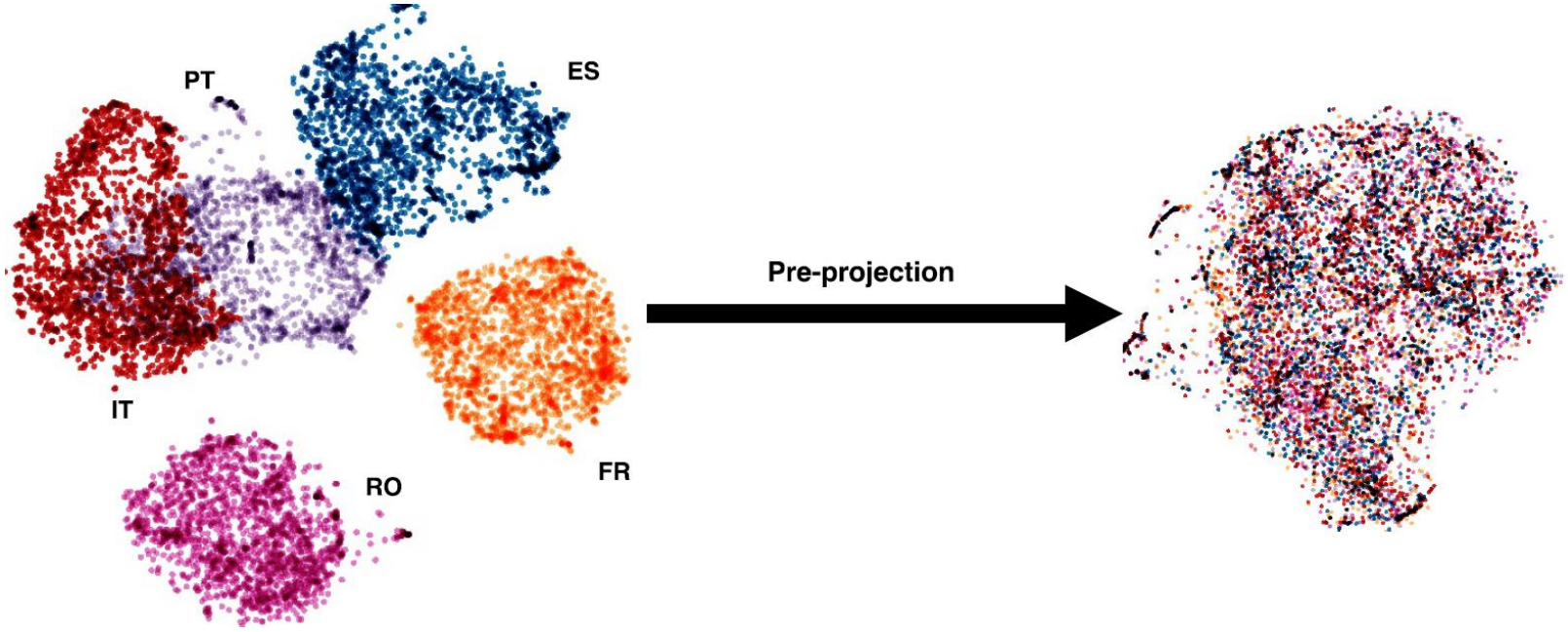
Universal Neural Machine Translation

- Motivation
  - For extremely low-resource language pairs, NMT does not perform well
  - Challenges for multilingual MT
    - Lexical-level sharing: low-resource languages may not even share the same character set as any high-resource language. → create shared semantic representation
    - Sentence-level sharing: share source sentence representation with other similar languages.
- Method
  - Universal lexical representation (ULR): a novel representation for multilingual embedding where each word from any language is represented as a probabilistic mixture of universal space word embeddings.
  - Mixture of language experts (MoLE): model the sentence level universal encoder

Gu et al., **Universal Neural Machine Translation for Extremely Low Resource Languages**. NAACL-HLT, 2018.

Universal Neural Machine Translation



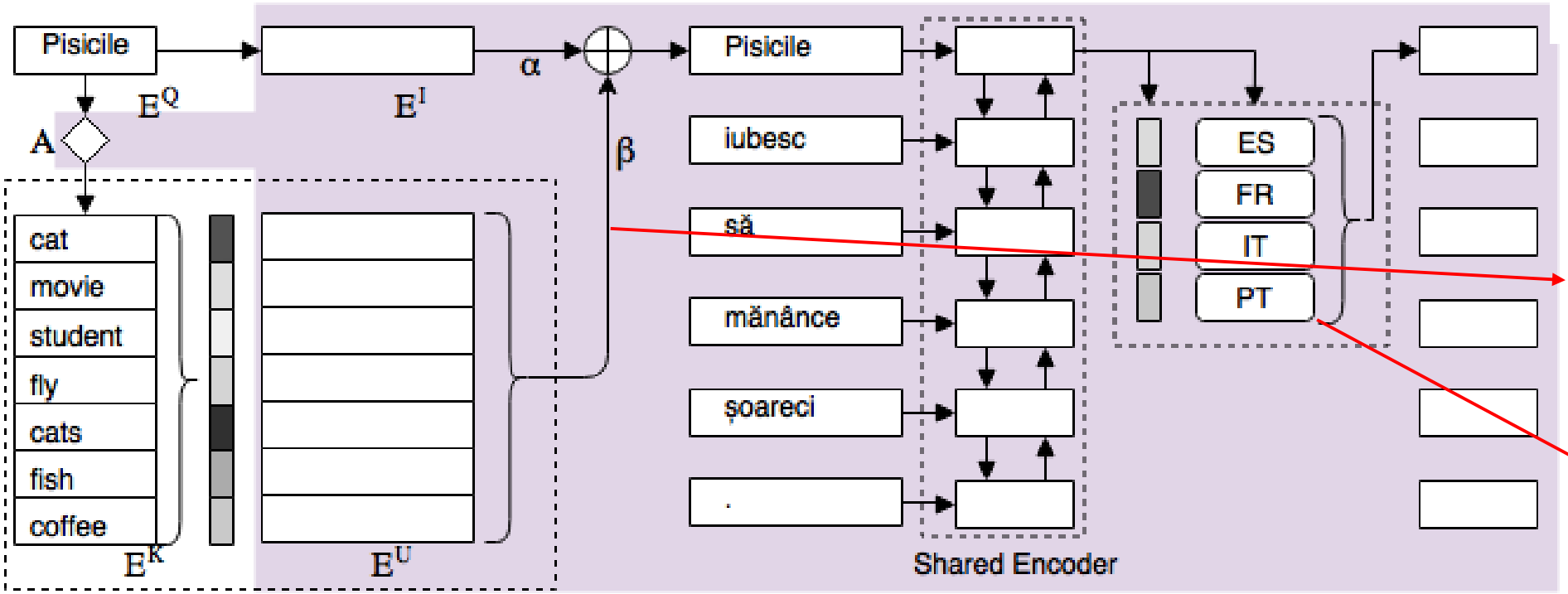


Gu et al., **Universal Neural Machine Translation for Extremely Low Resource Languages**. NAACL-HLT, 2018.

Universal Neural Machine Translation

$$\alpha(x)E^I(x) + \beta(x)\sum_{i=1}^{M} E^U(u_i)\cdot q(u_i|x) \quad (8)$$

for top-500 freq words



$$D(u,x) = E^K(u)\cdot A\cdot E^Q(x)^T$$

$$q(u_i|x) = \frac{e^{D(u_i,x)/\tau}}{\sum_{u_j} e^{D(u_j,x)/\tau}}$$

$$e_x = \sum_{i=1}^{M} E^U(u_i)\cdot q(u_i|x)$$

$$h' = \sum_{k=1}^{K} f_k(h)\cdot \text{softmax}(g(h))_k,$$

E^K and E^Q are learned cross-lingual embeddings

$$\mathcal{L}^{\text{gate}} = \sum_{k=1}^{K-1}\sum_{n=1}^{N_k} \log\left[\text{softmax}\left(g(h)\right)_k\right]$$

Gu et al., **Universal Neural Machine Translation for Extremely Low Resource Languages**. NAACL-HLT, 2018.

Universal Neural Machine Translation

- Monolingual corpus to train embeddings: wiki dump

	Zero-Resource Translation			Auxiliary High-Resource Translation								
source	Ro	Ko	Lv	Cs	De	El	Es	Fi	Fr	It	Pt	Ru
corpora	WMT16[1]	KPD[2]	Europarl v8[3]									UN[4]
size	612k	97k	638k	645k	1.91m	1.23m	1.96m	1.92m	2.00m	1.90m	1.96m	11.7m
subset	0/6k/60k	10k	6k	/								2.00m

Table 1: Statistics of the available parallel resource in our experiments. All the languages are translated to English.

Src	Aux	Multi	+ULR	+ MoLE
Ro	Cs De El Fi		18.02	18.37
	Cs De El Fr		19.48	19.52
	De El Fi It		19.11	19.33
	Es Fr It Pt	14.83	20.01	**20.51**
Lv	Es Fr It Pt	7.68	10.86	11.02
	Es Fr It Pt Ru	7.88	12.40	**13.16**
Ko	Es Fr It Pt	2.45	5.49	**6.14**

Table 2: Scores over variant source languages (6k sentences for Ro & Lv, and 10k for Ko). "Multi" means the Multi-lingual NMT baseline.

Models	BLEU
Vanilla	1.21
Multi-NMT	14.94
Closest Uni-Token Only	5.83
Multi-NMT + ULR + (A=I)	18.61
Multi-NMT + ULR	**20.01**
Multi-NMT + BT	17.91
Multi-NMT + ULR + BT	**22.35**
Multi-NMT + ULR + MoLE	20.51
Multi-NMT + ULR + MoLE + BT	**22.92**
Full data (612k) NMT	**28.34**

Table 3: BLEU scores evaluated on test set (6k), compared with ULR and MoLE. "vanilla" is the standard NMT system trained only on Ro-En training set

Gu et al., **Universal Neural Machine Translation for Extremely Low Resource Languages**. NAACL-HLT, 2018.

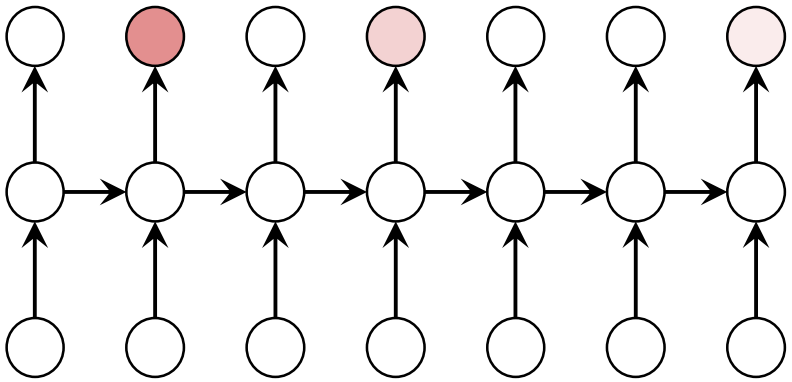
Microsoft

Multi-Task Learning for Low Resource NMT

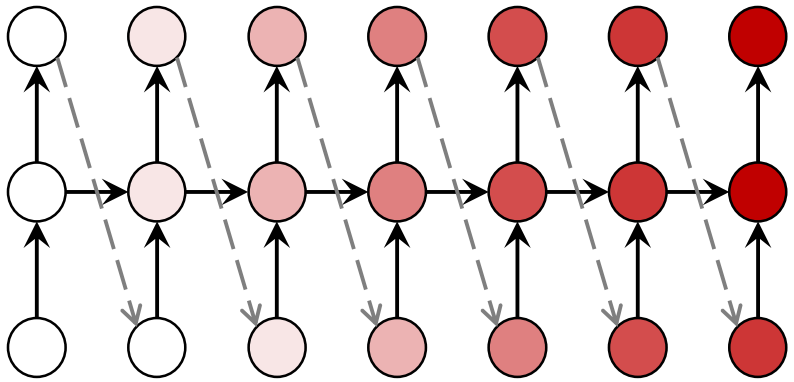
- Multi-Task Learning for Multiple Language Translation
- Google's Multilingual Neural Machine Translation System
- Universal Neural Machine Translation for Extremely Low Resource Languages
- **Agreement Regularization of L2R and R2L Models**

Agreement Regularization of L2R and R2L Models

- NMT model is only trained with golden bilingual corpus
- Translation sentence is auto-regressively generated word by word
- Previous errors will mislead the generation of the subsequences
- Errors will be quickly amplified



Training



Decoding

Zhang et al., **Regularizing Neural Machine Translation by Target-Bidirectional Agreement**. AAAI, 2019.

Agreement Regularization of L2R and R2L Models

- Introduce two Kullback-Leibler (KL) divergence regularization terms

$$L(\overrightarrow{\theta}) = \sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)}; \overrightarrow{\theta})$$
$$- \lambda \sum_{n=1}^{N} \mathrm{KL}(P(y|x^{(n)}; \overleftarrow{\theta}) || P(y|x^{(n)}; \overrightarrow{\theta}))$$
$$- \lambda \sum_{n=1}^{N} \mathrm{KL}(P(y|x^{(n)}; \overrightarrow{\theta}) || P(y|x^{(n)}; \overleftarrow{\theta}))$$

$$\frac{\partial \mathrm{KL}(P(y|x^{(n)}; \overleftarrow{\theta}) || P(y|x^{(n)}; \overrightarrow{\theta}))}{\partial \overrightarrow{\theta}} = -\mathrm{E}_{y \sim P(y|x^{(n)}; \overleftarrow{\theta})} \frac{\partial \log P(y|x^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}}$$

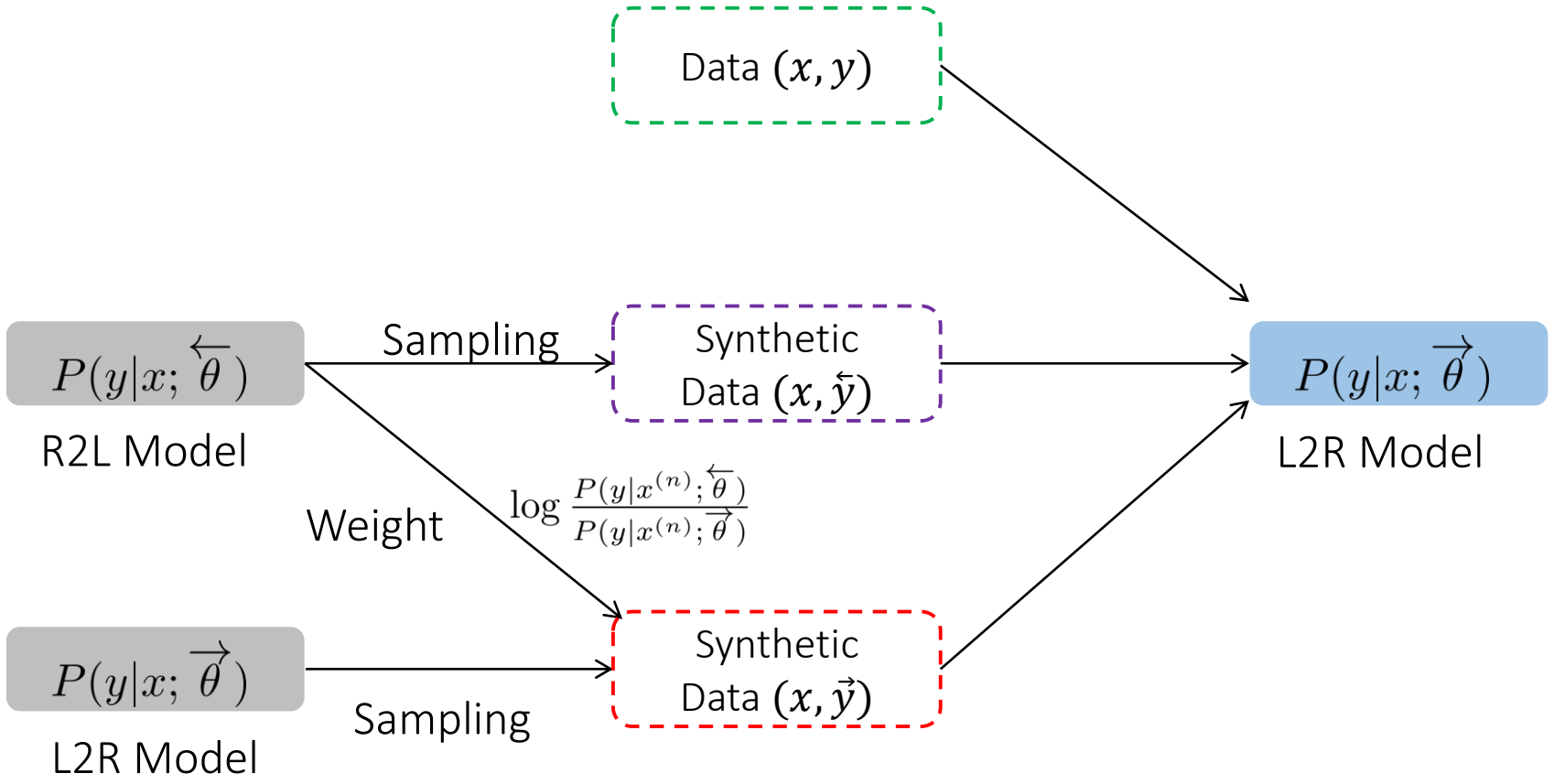
$$\frac{\partial \mathrm{KL}(P(y|x^{(n)}; \overrightarrow{\theta}) || P(y|x^{(n)}; \overleftarrow{\theta}))}{\partial \overrightarrow{\theta}} = -\mathrm{E}_{y \sim P(y|x^{(n)}; \overrightarrow{\theta})} \left(\log \frac{P(y|x^{(n)}; \overleftarrow{\theta})}{P(y|x^{(n)}; \overrightarrow{\theta})} \frac{\partial \log P(y|x^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}} \right)$$

Weight

Zhang et al., **Regularizing Neural Machine Translation by Target-Bidirectional Agreement**. AAAI, 2019.

Agreement Regularization of L2R and R2L Models

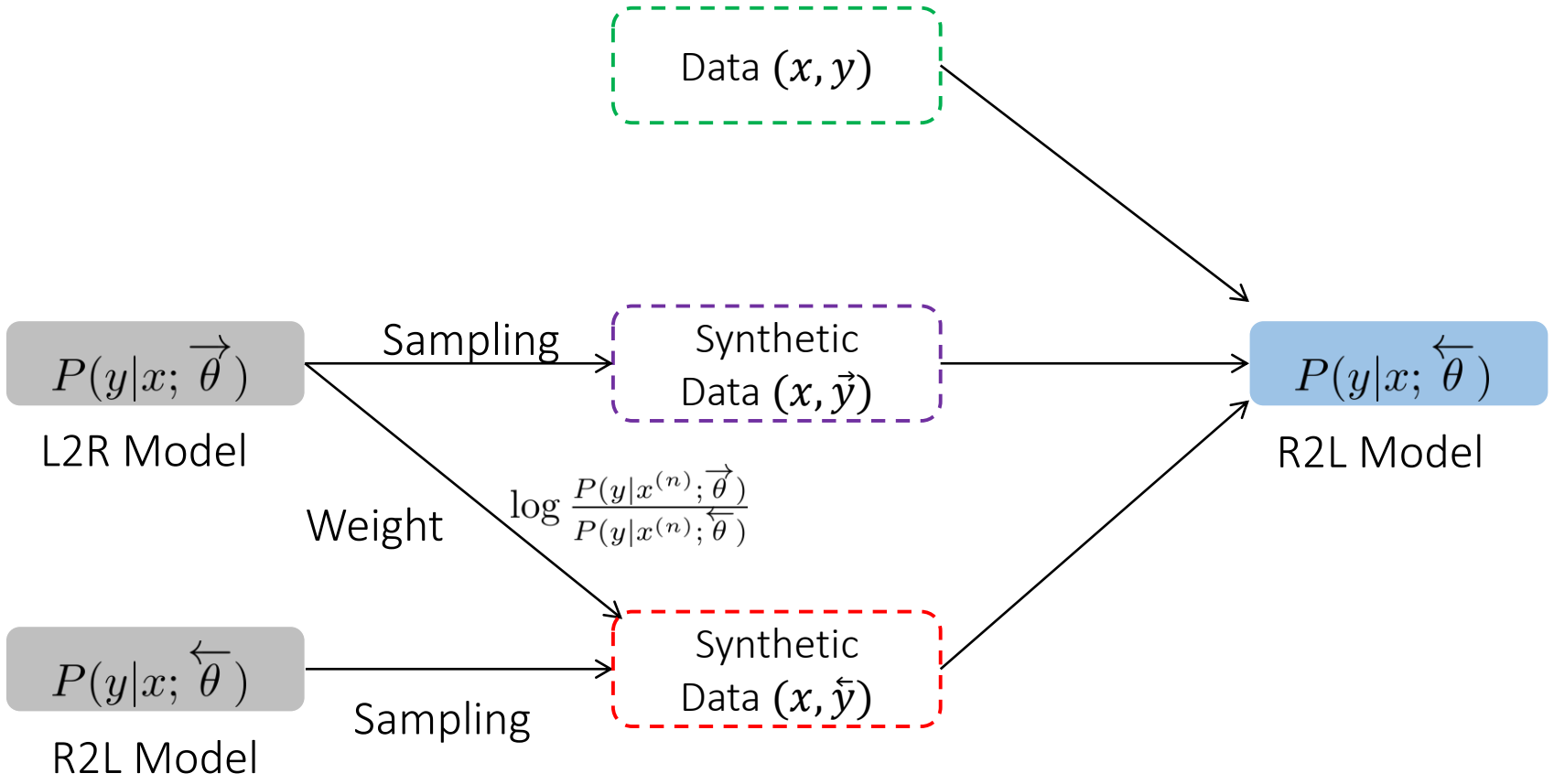
$$\frac{\partial L(\overrightarrow{\theta})}{\partial \overrightarrow{\theta}} = \sum_{n=1}^{N} \left(\frac{\partial \log P(y^{(n)}|x^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}} + \lambda \mathrm{E}_{y \sim P(y|x^{(n)}; \overleftarrow{\theta})} \frac{\partial \log P(y|x^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}} + \lambda \mathrm{E}_{y \sim P(y|x^{(n)}; \overrightarrow{\theta})} \left(\log \frac{P(y|x^{(n)}; \overleftarrow{\theta})}{P(y|x^{(n)}; \overrightarrow{\theta})} \frac{\partial \log P(y|x^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}} \right) \right)$$



Zhang et al., **Regularizing Neural Machine Translation by Target-Bidirectional Agreement**. AAAI, 2019.

Agreement Regularization of L2R and R2L Models

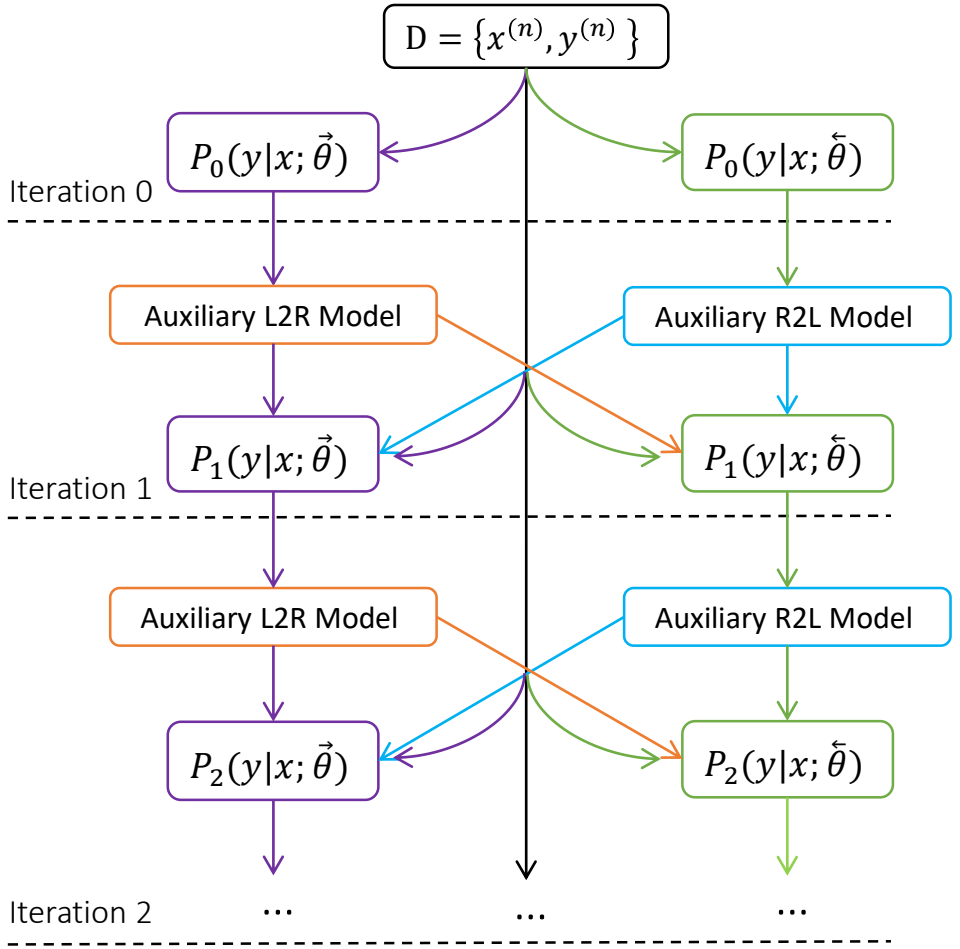
$$\frac{\partial L(\overleftarrow{\theta})}{\partial \overleftarrow{\theta}} = \sum_{n=1}^{N} \left(\frac{\partial \log P(y^{(n)}|x^{(n)}; \overleftarrow{\theta})}{\partial \overleftarrow{\theta}} + \lambda \mathrm{E}_{y \sim P(y|x^{(n)}; \overrightarrow{\theta})} \frac{\partial \log P(y|x^{(n)}; \overleftarrow{\theta})}{\partial \overleftarrow{\theta}} + \lambda \mathrm{E}_{y \sim P(y|x^{(n)}; \overleftarrow{\theta})} \left(\log \frac{P(y|x^{(n)}; \overrightarrow{\theta})}{P(y|x^{(n)}; \overleftarrow{\theta})} \frac{\partial \log P(y|x^{(n)}; \overleftarrow{\theta})}{\partial \overleftarrow{\theta}} \right) \right)$$



Zhang et al., **Regularizing Neural Machine Translation by Target-Bidirectional Agreement**. AAAI, 2019.

Agreement Regularization of L2R and R2L Models

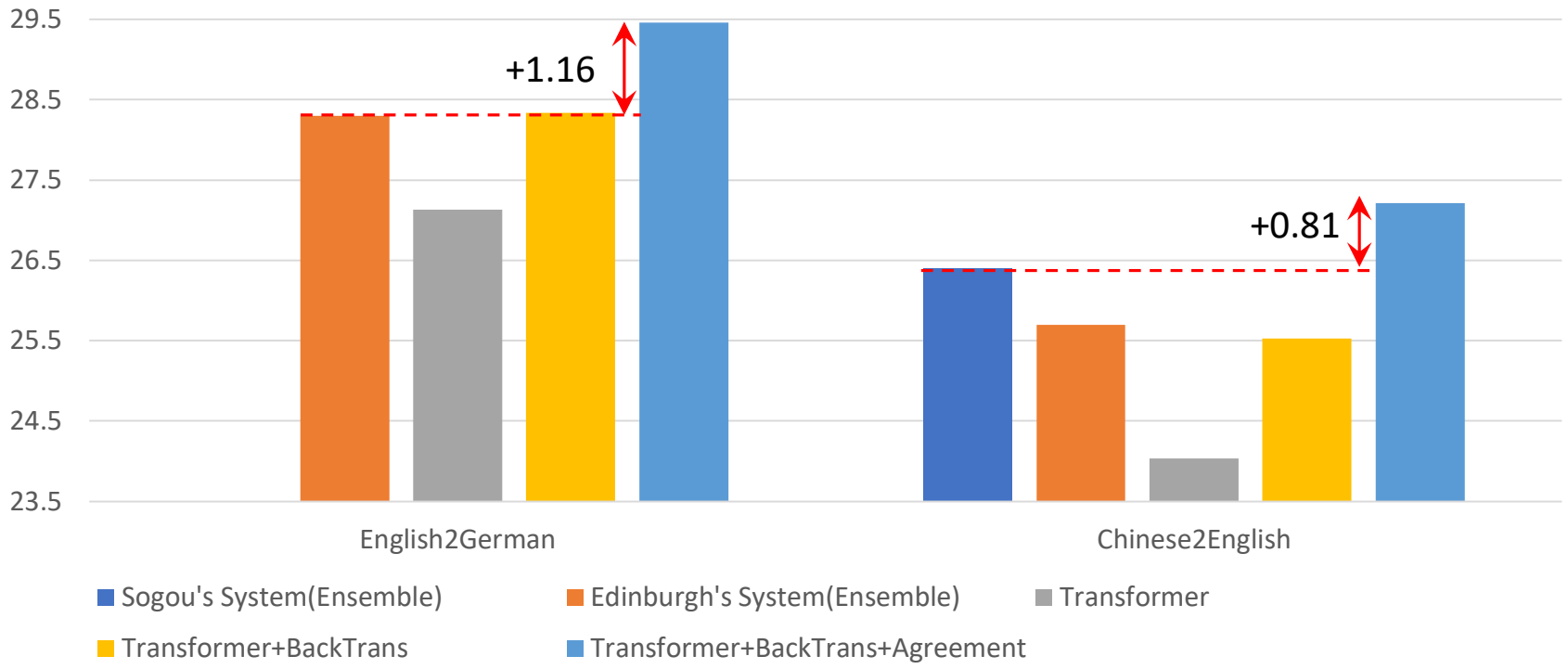
- **Iteration 0**: Pre-train L2R and R2L models $P_0(y|x;\vec{\theta})$ and $P_0(y|x;\overleftarrow{\theta})$ with bilingual data $D = \{x^{(n)}, y^{(n)}\}$
- **Iteration 1**: Two NMT systems based on $P_0(y|x;\vec{\theta})$ and $P_0(y|x;\overleftarrow{\theta})$ are used to generate pseudo bilingual data based on $\mathbf{X} = \{x^{(n)}\}$ and $\mathbf{Y} = \{y^{(n)}\}$; Three data sets(two pseudo and one true) are used to update L2R and R2L models.
- **Iteration 2:** Repeat the above process
- ....



Zhang et al., **Regularizing Neural Machine Translation by Target-Bidirectional Agreement**. AAAI, 2019.

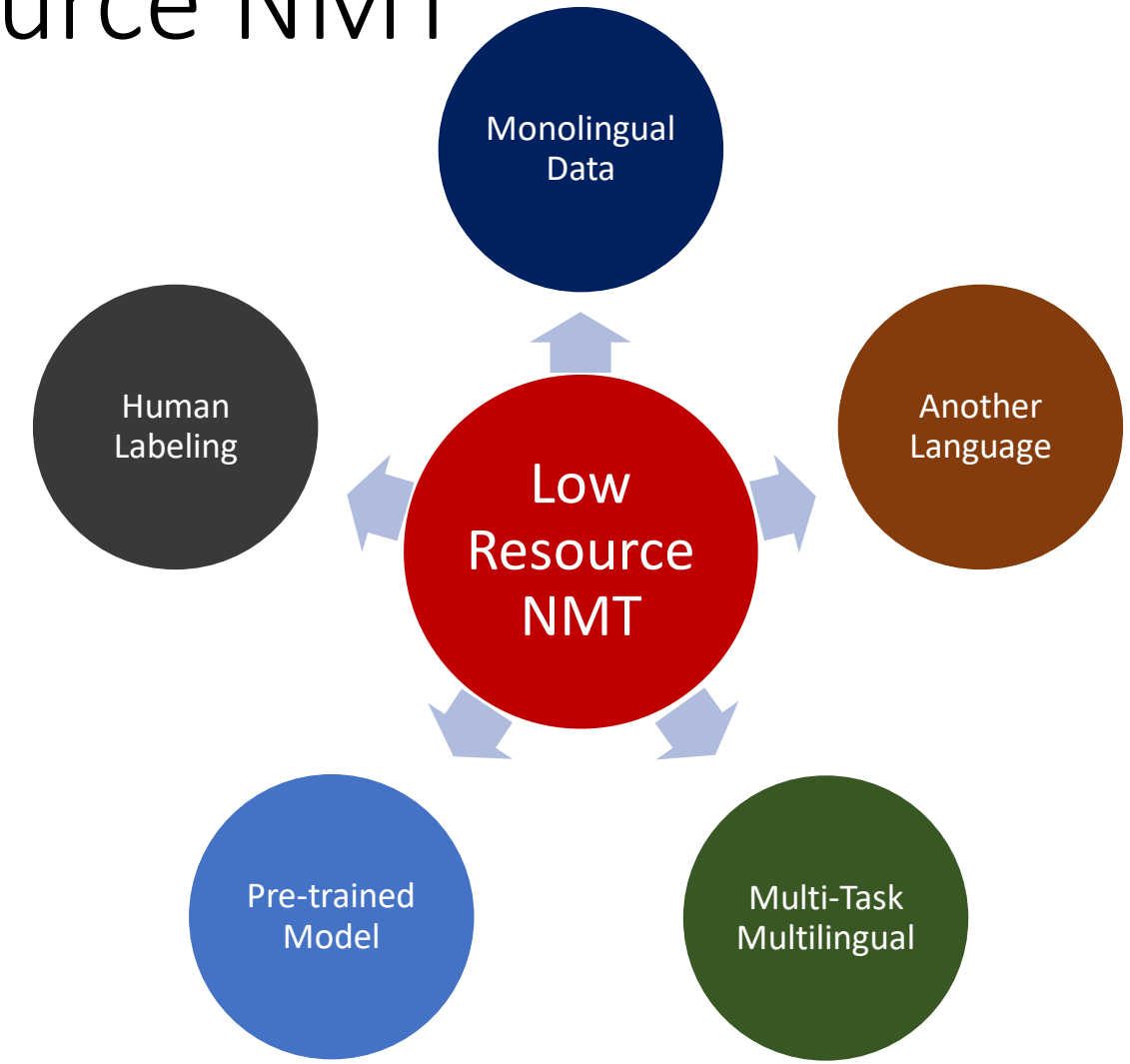
Agreement Regularization of L2R and R2L Models

- WMT17 English-to-German and Chinese-to-English Translation Tasks



Zhang et al., **Regularizing Neural Machine Translation by Target-Bidirectional Agreement**. AAAI, 2019.

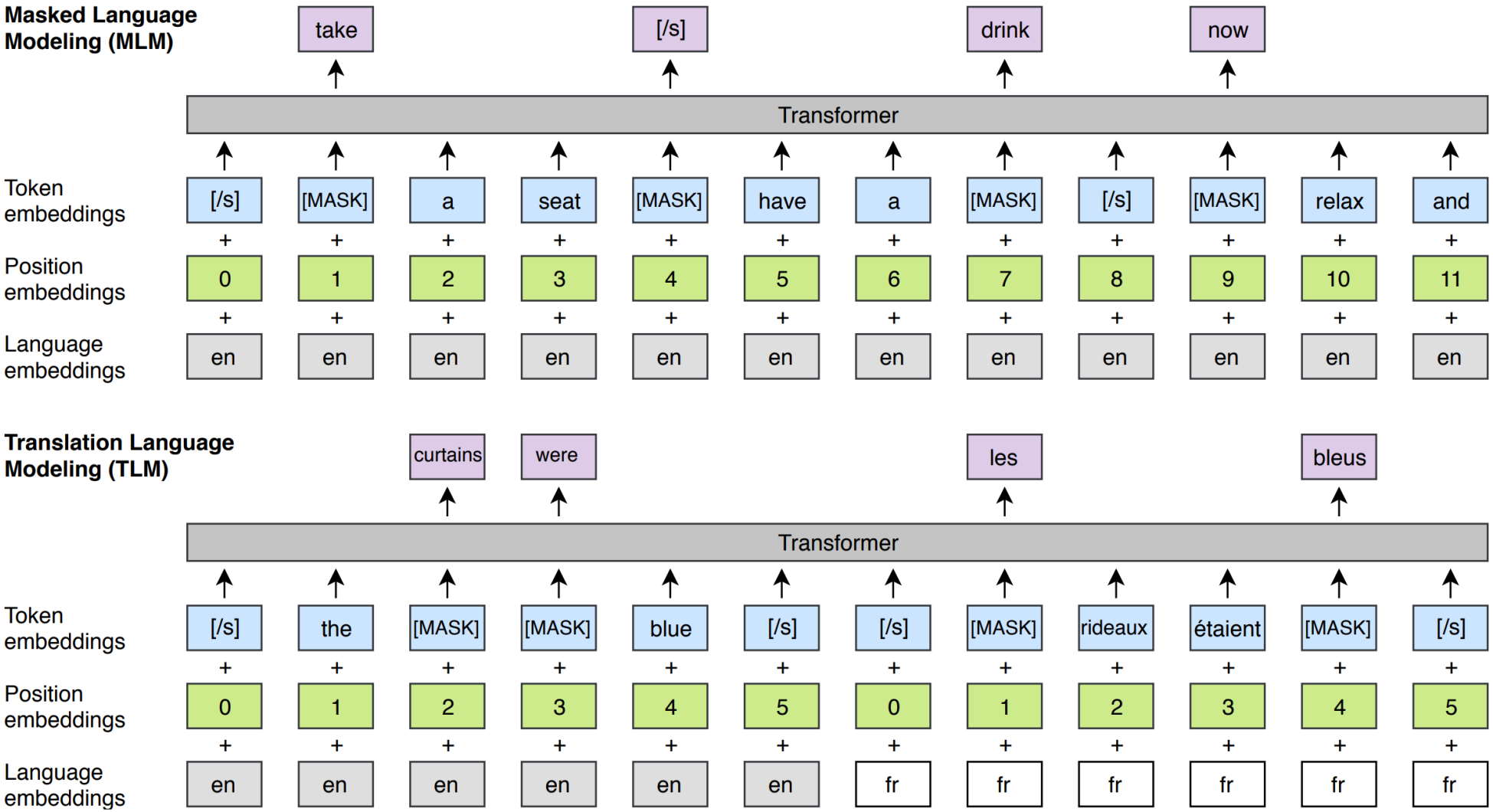
Low Resource NMT



Pre-trained Models and Transfer Learning

- **Cross-lingual Language Model Pretraining**
- Towards Making the Most of BERT in Neural Machine Translation
- Improving Neural Machine Translation with Pre-trained Representation
- Cross-Lingual Transfer Learning
- Meta-learning for Low-resource Neural Machine Translation

Cross-lingual Language Model Pretraining



Lample et al., **Cross-lingual Language Model Pretraining**, arxiv, 2019.

Cross-lingual Language Model Pretraining

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Table 1: **Results on cross-lingual classification accuracy.** Test accuracy on the 15 XNLI languages.

Lample et al., **Cross-lingual Language Model Pretraining**, arxiv, 2019.

Cross-lingual Language Model Pretraining

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Table 2: **Results on unsupervised MT.**

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro $\rightarrow$ en	28.4	31.5	35.3
ro $\leftrightarrow$ en	28.5	31.5	35.6
ro $\leftrightarrow$ en + BT	34.4	37.0	**38.5**

Table 3: **Results on supervised MT.**

Lample et al., **Cross-lingual Language Model Pretraining**, arxiv, 2019.

Microsoft

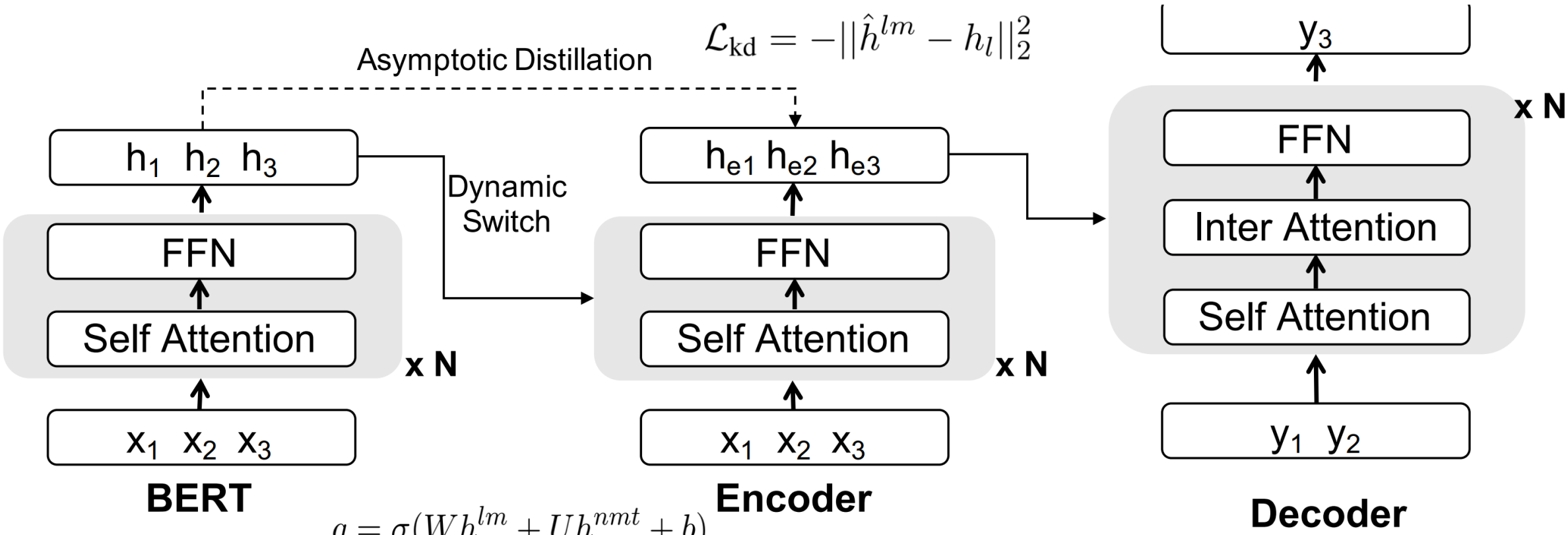
Pre-trained Models and Transfer Learning

- Cross-lingual Language Model Pretraining
- **Towards Making the Most of BERT in Neural Machine Translation**
- Improving Neural Machine Translation with Pre-trained Representation
- Cross-Lingual Transfer Learning
- Meta-learning for Low-resource Neural Machine Translation

Towards Making the Most of BERT in NMT

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{nmt}} + (1 - \alpha) \cdot \mathcal{L}_{\text{kd}}$$

$$\mathcal{L}_{\text{kd}} = -||\hat{h}^{lm} - h_l||_2^2$$



$$g = \sigma(Wh^{lm} + Uh^{nmt} + b)$$

$$h = g \odot h^{lm} + (1 - g) \odot h^{nmt}$$

Yang et al., **Towards Making the Most of BERT in Neural Machine Translation**, arxiv, 2019.

Towards Making the Most of BERT in NMT

System	Architecture	En-De	En-Fr	En-Zh
Existing systems				
Vaswani et al. (2017)	Transformer base	27.3	38.1	-
Vaswani et al. (2017)	Transformer big	28.4	41.0	-
Lample and Conneau (2019)	Transformer big + Fine-tuning	27.7	-	-
Lample and Conneau (2019)	Transformer big + Frozen Feature	28.7	-	-
Chen et al. (2018)	RNMT+ + MultiCol	41.7	28.7 -	
Our NMT systems				
CTNMT	Transformer (base)	27.2	41.0	37.3
CTNMT	Rate-scheduling	29.7	41.6	38.4
CTNMT	Dynamic Switch	29.4	41.4	38.6
CTNMT	Asymptotic Distillation	29.2	41.6	38.3
CTNMT	+ ALL	**30.1**	**42.3**	**38.9**

Table 1: Case-sensitive BLEU scores on English-German, English-French and English-Chinese translation. The best performance comes from the fusion of rate-scheduling, dynamic switch and asymptotic distillation.

Yang et al., **Towards Making the Most of BERT in Neural Machine Translation**, arxiv, 2019.

Microsoft

Pre-trained Models and Transfer Learning

- Cross-lingual Language Model Pretraining
- Towards Making the Most of BERT in Neural Machine Translation
- **Improving Neural Machine Translation with Pre-trained Representation**
- Cross-Lingual Transfer Learning
- Meta-learning for Low-resource Neural Machine Translation

Improving NMT with Pre-trained Representation

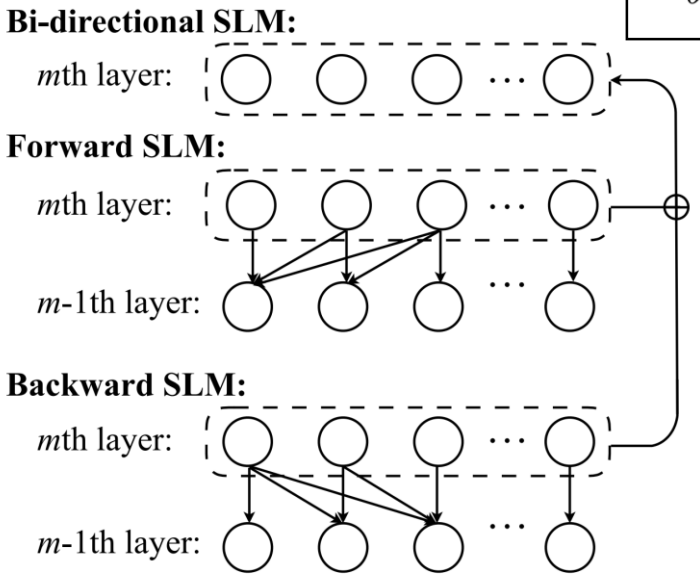


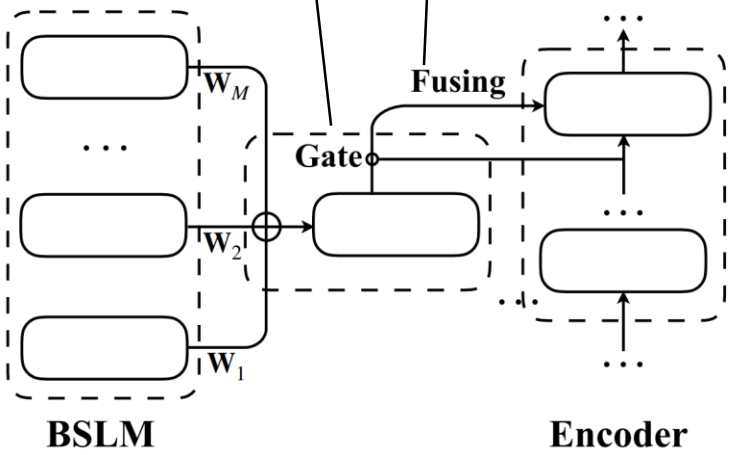
Figure 1: Overview of the bidirectional self-attention language model (BSLM).

$$\bar{\mathbf{r}}_n^S = \frac{1}{I}\sum_{i=1}^{I}\mathbf{r}_{n,i}^S$$

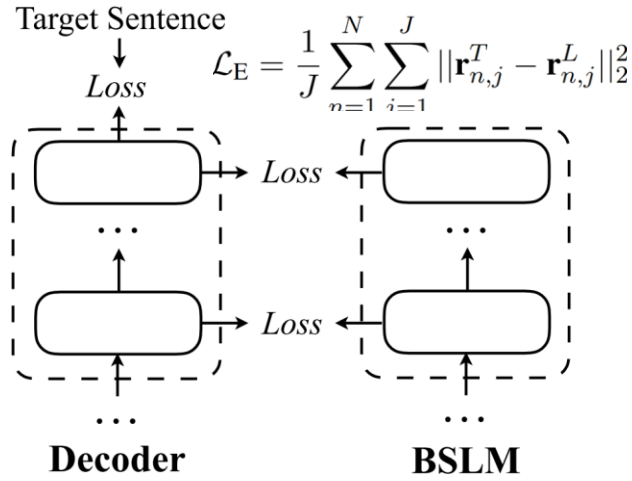
$$\theta_n = \text{sigmoid}(\bar{\mathbf{r}}_n^S)$$

$$\mathbf{R}_n^W = \sum_{m=1}^{M}(\mathbf{W}_{n,m} * \mathbf{R}_m^L)$$

$$\mathbf{R}_n^S = \mathbf{R}_n^S + \theta_n * \mathbf{R}_n^W$$



(a) The weighted-fusion mechanism used in the encoder.



(b) The knoledge transfer paradigm used in the decoder.

Figure 2: Overview of the proposed integration framework.

Improving NMT with Pre-trained Representation

#	Model	MT02	MT03	MT04	MT05	Average	Δ
1	RNNSearch (Luong et al., 2015)	N/A	28.38	30.85	26.78	–	–
2	(Sennrich et al., 2016)	36.95	36.80	37.99	35.33	–	–
3	(Zhang and Zong, 2016)	N/A	33.38	34.30	31.57	–	–
4	(Cheng et al., 2016)	38.78	38.32	38.49	36.45	–	–
5	(Zhang et al., 2018)	N/A	43.26	N/A	41.61	–	–
6	Transformer	44.77	44.93	45.81	43.04	44.59	–
7	+ ELMo (Peters et al., 2018)	45.23	45.60	46.26	43.61	45.16	+0.57
8	+ GPT (Radford et al., 2018)	44.89	45.22	45.99	43.31	44.84	+0.25
9	+ BERT (Devlin et al., 2018)	45.02	45.53	46.02	43.52	45.02	+0.43
	Effectiveness of weighted-fusion mechanism used in the different layers						
10	+ Weighted-fusion (*shallow*)	44.97	45.21	46.19	43.23	44.88	+0.29
11	+ Weighted-fusion (*deep*)	45.46	45.62	46.57	43.82	45.34	+0.75
	Effectiveness of knowledge transfer paradigm used in the different layers						
12	+ Knowledge Transfer (*shallow*)	45.61	45.63	46.54	43.86	45.34	+0.75
13	+ Knowledge Transfer (*deep*)	45.71	45.78	46.62	43.94	45.45	+0.85
	Our proposed model						
14	+ Our Approach	**45.82**	**45.86**	**46.83**	**44.13**	**45.61**	**+1.01**

Table 1: Translation qualities on the ZH→EN experiments. *deep* and *shallow* mean employing proposed methods on the all layers or the first layer, respectively.

Weng et al., **Improving Neural Machine Translation with Pre-trained Representation**, arxiv, 2019.

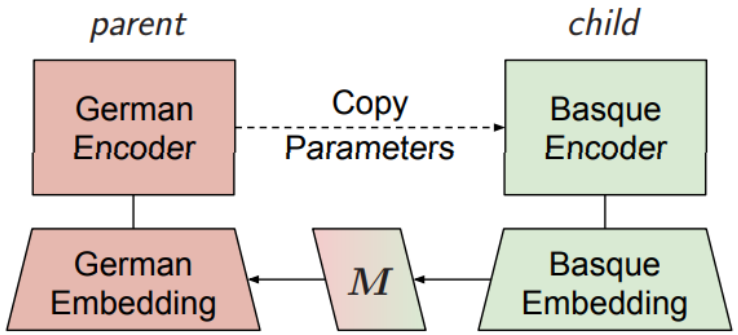
Microsoft

Pre-trained Models and Transfer Learning

- Cross-lingual Language Model Pretraining
- Towards Making the Most of BERT in Neural Machine Translation
- Improving Neural Machine Translation with Pre-trained Representation
- **Cross-Lingual Transfer Learning**
- Meta-learning for Low-resource Neural Machine Translation

Cross-Lingual Transfer Learning

Problem: Vocabulary mismatch between *parent* / *child* languages
Solution: Shared word embedding space

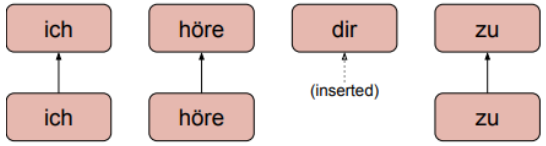


1. $\boldsymbol{E}_{\text{child}}$ = monolingual skip-gram, $\boldsymbol{E}_{\text{parent}}$ = pre-trained *parent* NMT
2. $\boldsymbol{M}$ = linear mapping $\boldsymbol{E}_{\text{child}} \rightarrow \boldsymbol{E}_{\text{parent}}$ (e.g. MUSE)

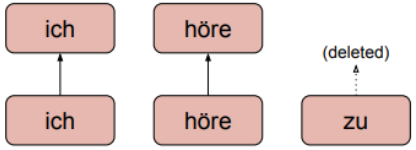
$$\boldsymbol{M}_i = \underset{\boldsymbol{M}'}{\operatorname{argmin}} \sum_{(w,w') \in \boldsymbol{D}_i} \|\boldsymbol{M}'\boldsymbol{E}_{\text{child}}(w) - \boldsymbol{E}_{\text{parent}}(w')\|_2$$

Problem: Word order difference between *parent* / *child* languages
Solution: Pre-train syntax-agnostic *parent* encoder with noisy input

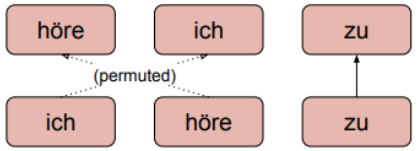
Insertion (10%)



Deletion (10%)



Permutation (100%, max. dist. 3)



- Randomized for each batch
  - *parent* learns various syntax
  - General monotonicity is retained
- Easier to adapt to a new language

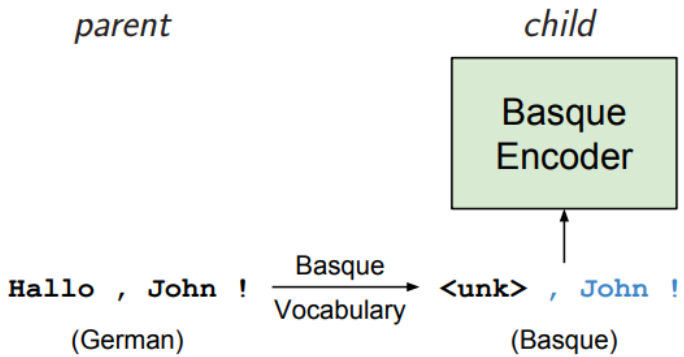
Kim et al., **Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies**, ACL 2019.

Cross-Lingual Transfer Learning

Problem: Back-translation does not work (poor English→xx model)

Solution: Reuse *parent* training data and adjust to *child* vocabulary

- Keep shared tokens and map the rest to <unk>



Results (BLEU [%])

System	eu-en	sl-en	be-en	az-en	tr-en
Transformer baseline (*child* only)	1.7	10.1	3.2	3.1	0.8
Multilingual (*parent* + *child*)	5.1	16.7	4.2	4.5	8.7
Transfer	4.9	19.2	8.9	5.3	7.4
+ Cross-lingual word embedding	7.4	20.6	12.2	7.4	9.4
+ Artificial noises	8.2	21.3	12.8	8.1	10.1
+ Synthetic data from *parent*	**9.7**	**22.1**	**14.0**	**9.0**	**11.3**

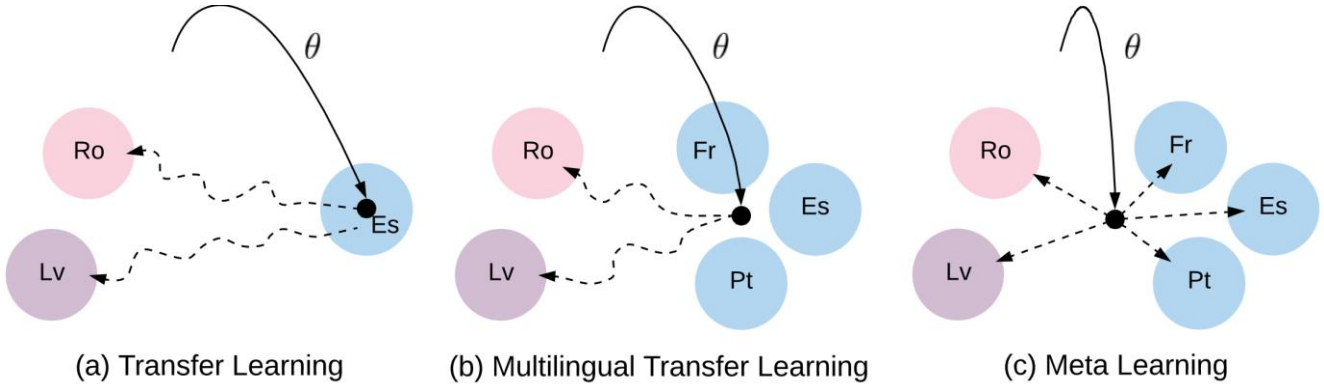
Kim et al., **Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies**, ACL 2019.

Microsoft

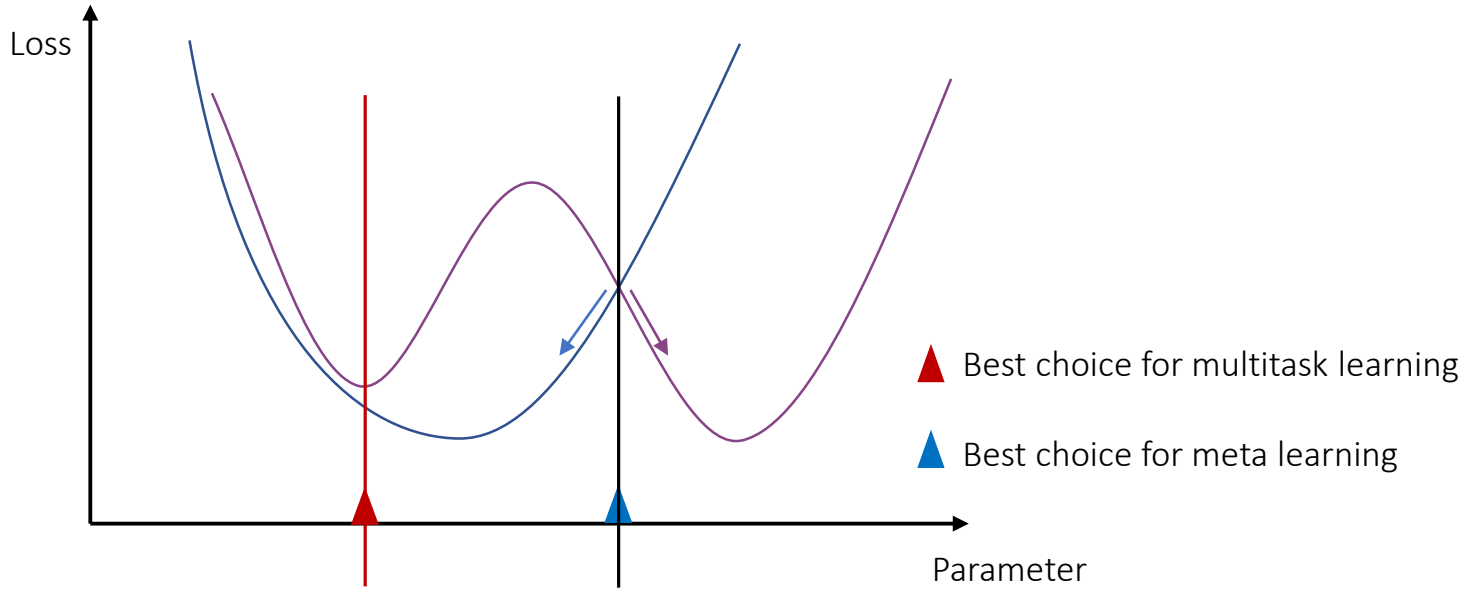
Pre-trained Models and Transfer Learning

- Cross-lingual Language Model Pretraining
- Towards Making the Most of BERT in Neural Machine Translation
- Improving Neural Machine Translation with Pre-trained Representation
- Cross-Lingual Transfer Learning
- **Meta-learning for Low-resource Neural Machine Translation**

Meta-Learning for Low-Resource NMT

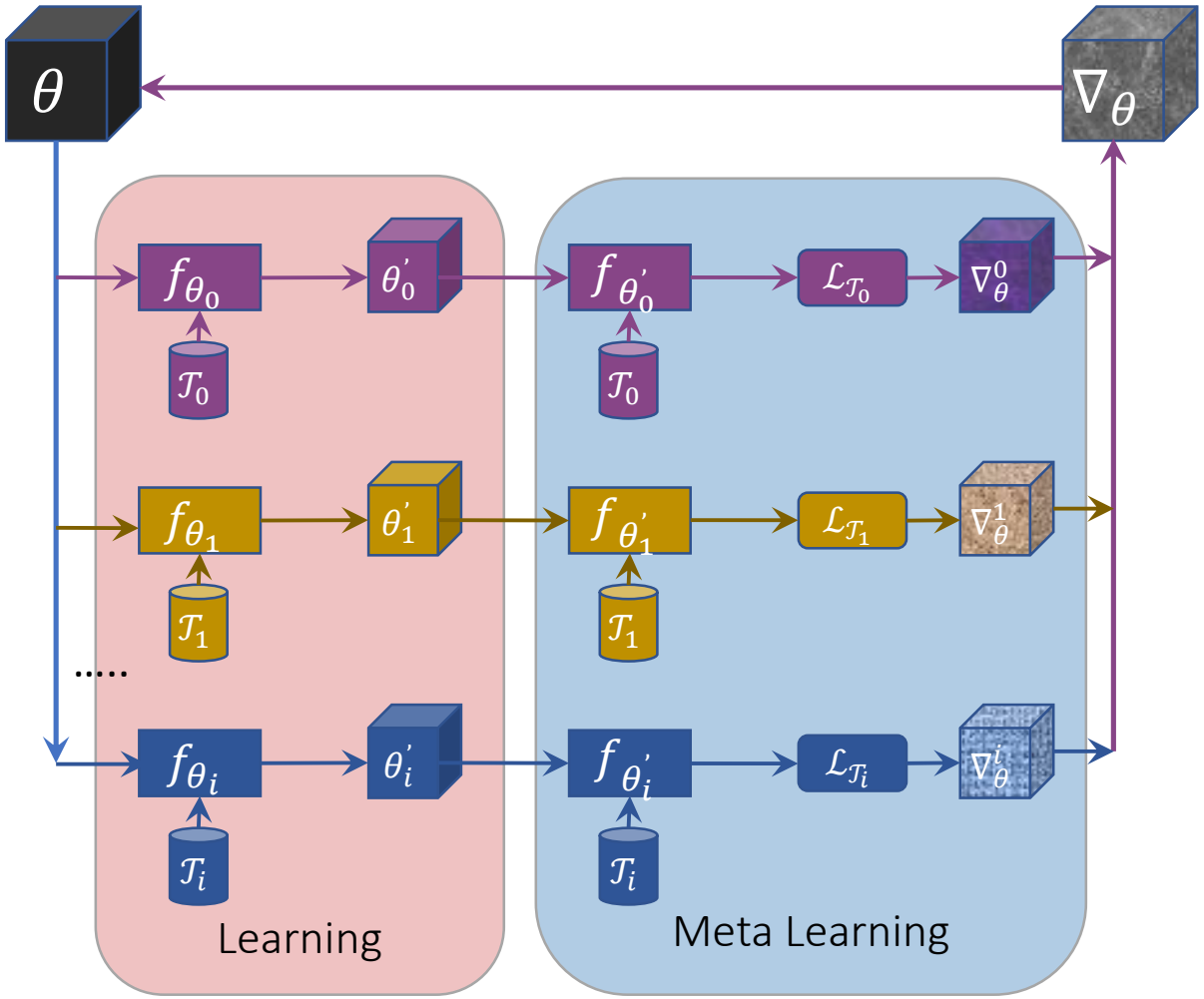


(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning



Gu et al., **Meta-learning for Low-resource Neural Machine Translation** , EMNLP 2018.

Meta-Learning for Low-Resource NMT



Algorithm 2 MAML for Few-Shot Supervised Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Sample K datapoints $\mathcal{D} = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$
6: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ using $\mathcal{D}$ and $\mathcal{L}_{\mathcal{T}_i}$ in Equation (2) or (3)
7: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
8: Sample datapoints $\mathcal{D}_i' = \{\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\}$ from $\mathcal{T}_i$ for the meta-update
9: **end for**
10: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$ using each $\mathcal{D}_i'$ and $\mathcal{L}_{\mathcal{T}_i}$ in Equation 2 or 3
11: **end while**

Gu et al., **Meta-learning for Low-resource Neural Machine Translation** , EMNLP 2018.

Meta-Learning for Low-Resource NMT

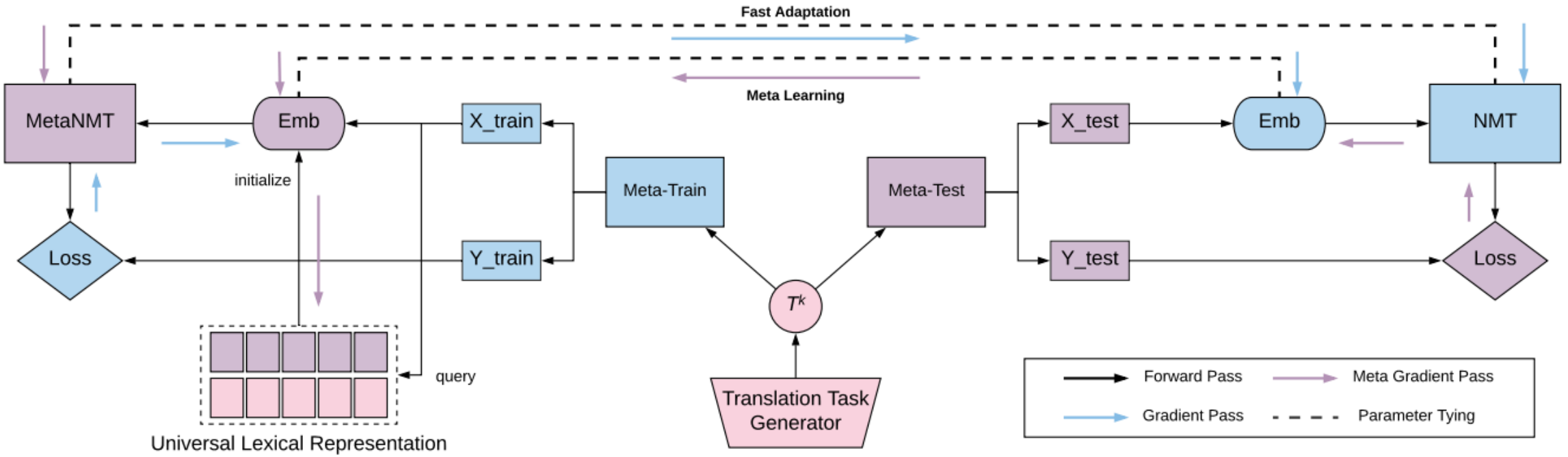
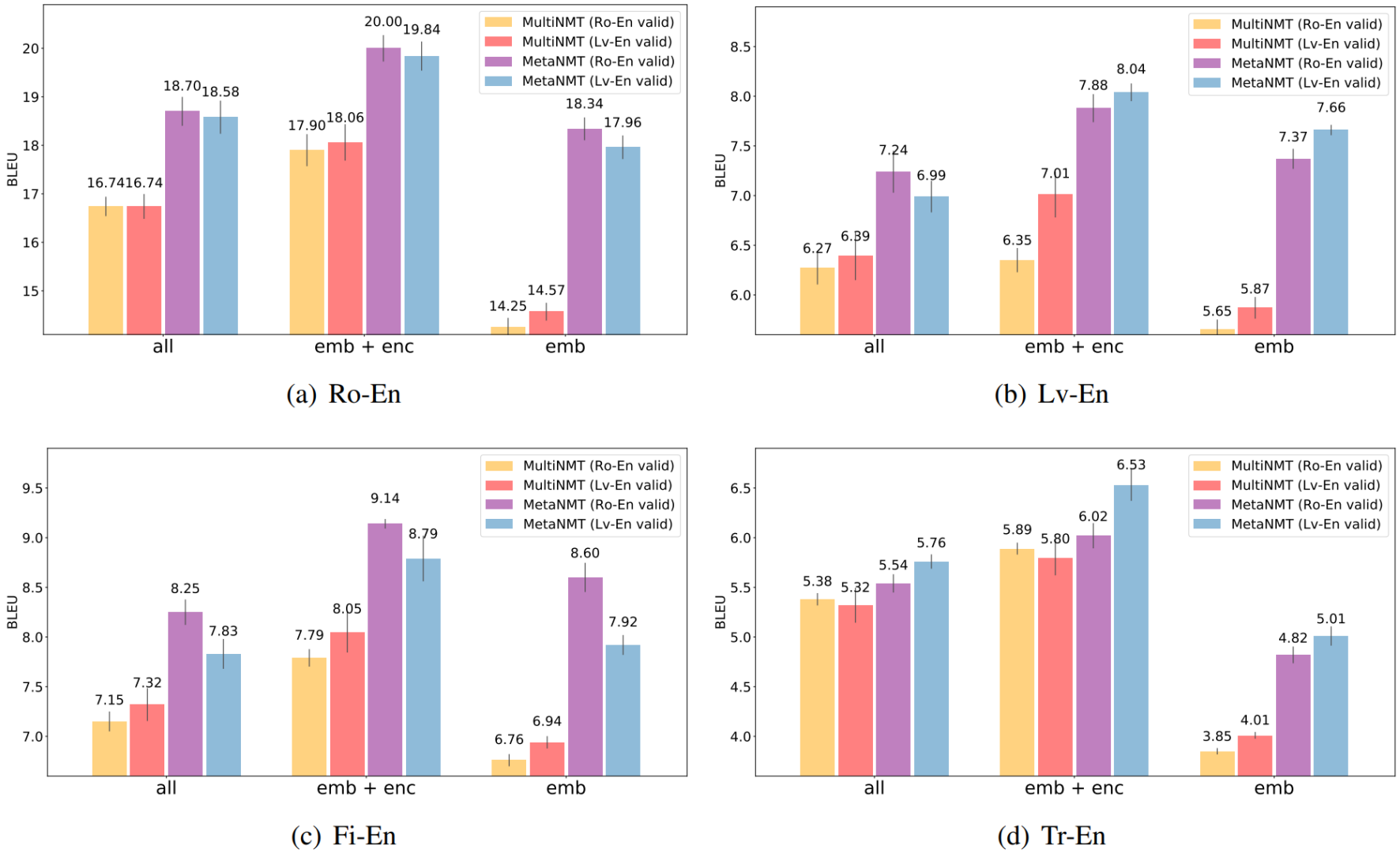


Figure 1: The graphical illustration of the training process of the proposed MetaNMT. For each episode, one task (language pair) is sampled for meta-learning. The boxes and arrows in blue are mainly involved in language-specific learning (§3.1), and those in purple in meta-learning (§3.2).

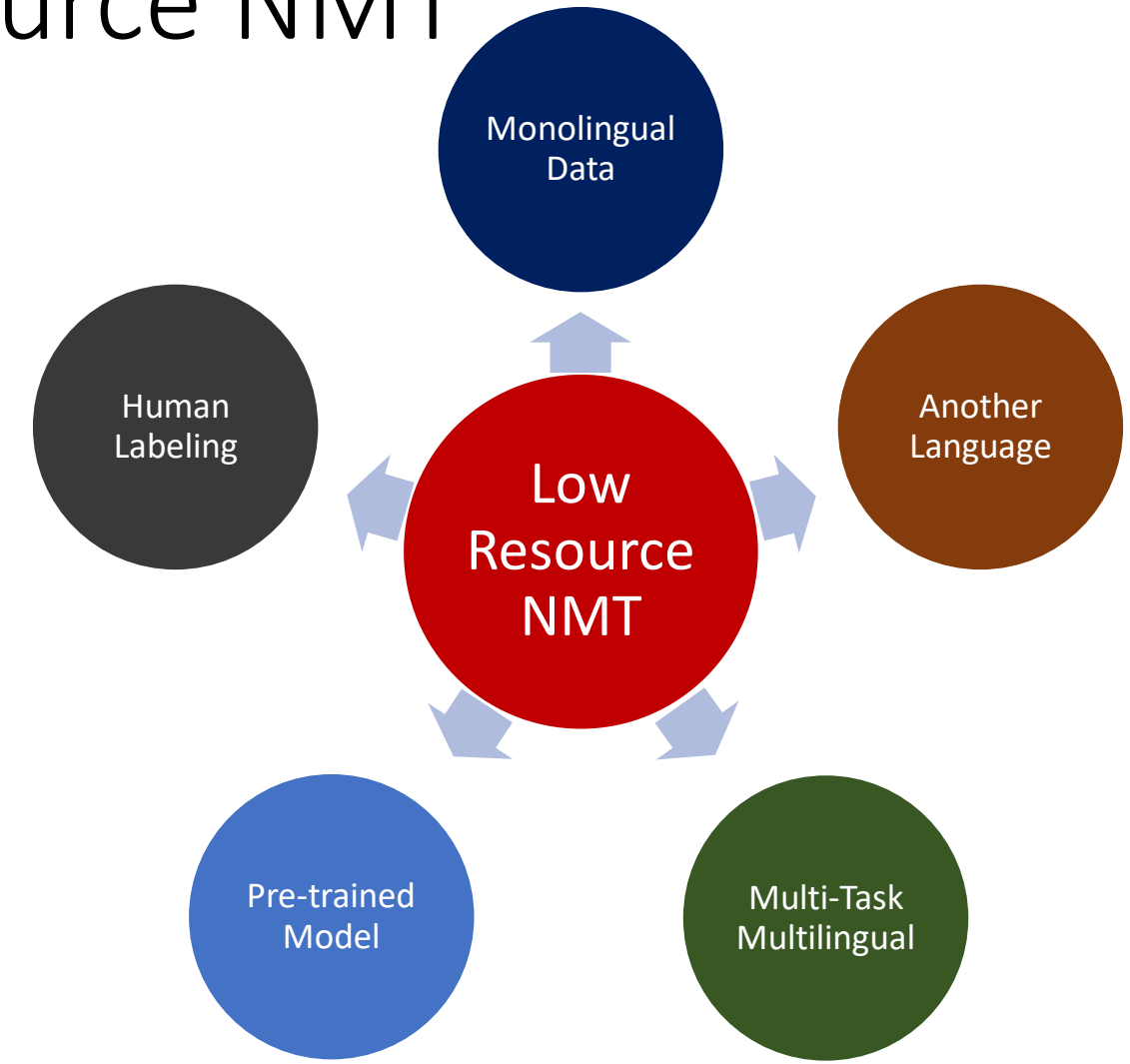
Gu et al., **Meta-learning for Low-resource Neural Machine Translation** , EMNLP 2018.

Meta-Learning for Low-Resource NMT

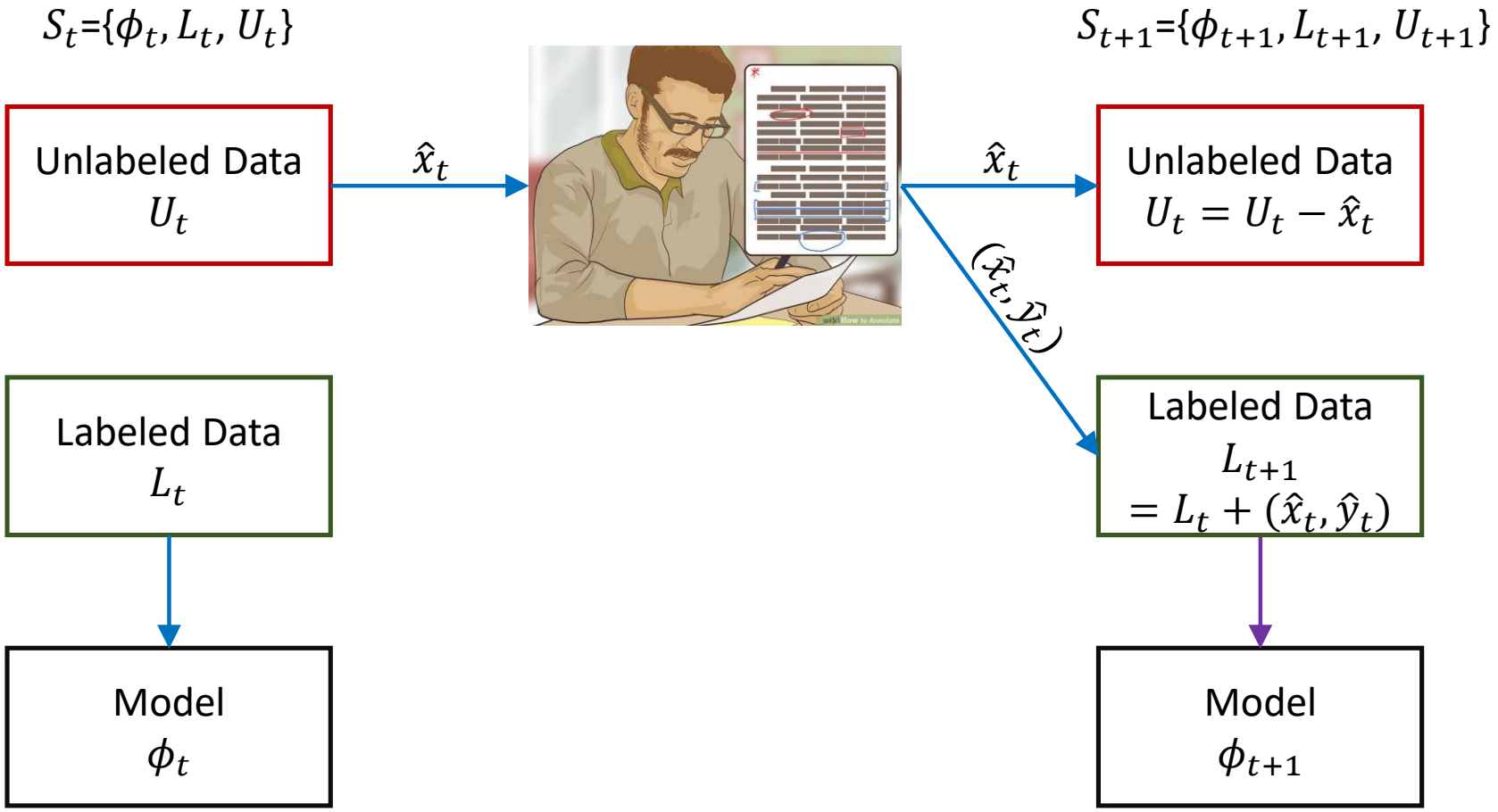


Gu et al., **Meta-learning for Low-resource Neural Machine Translation** , EMNLP 2018.

Low Resource NMT

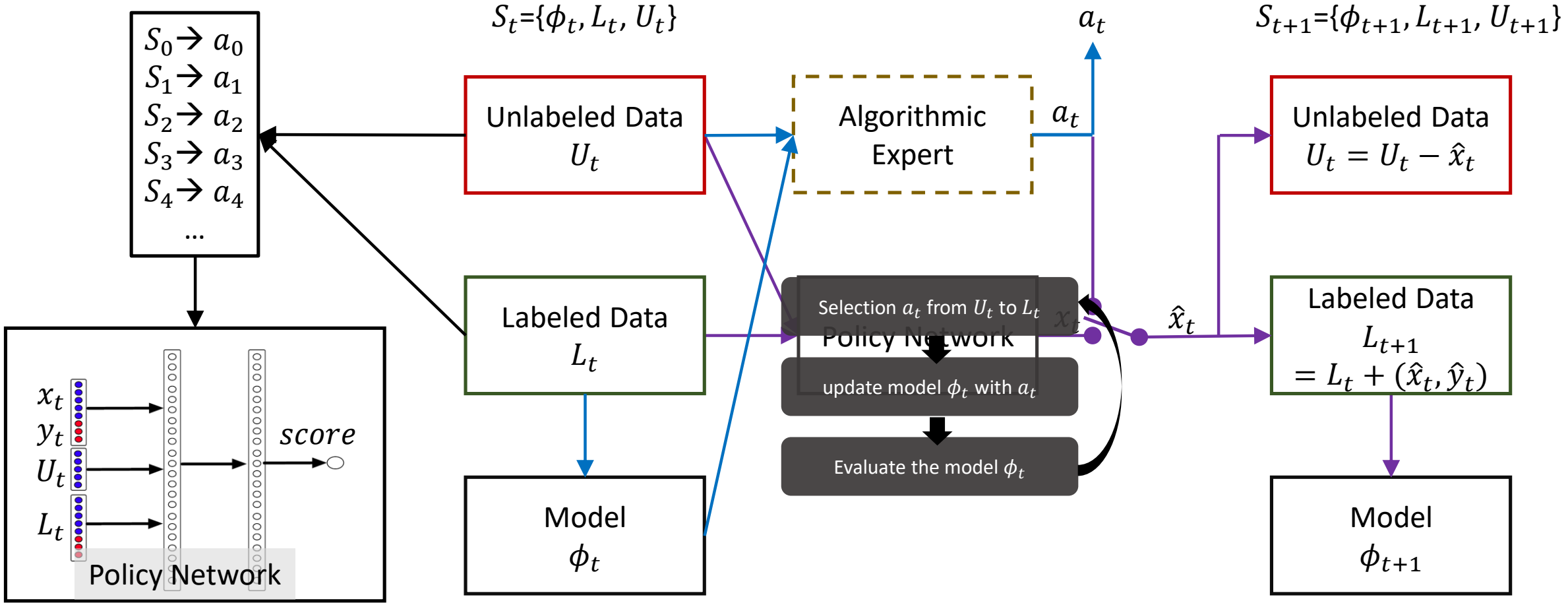


Learning to Actively Learn Neural Machine Translation



Liu et al., **Learning to Actively Learn Neural Machine Translation**, CoNLL, 2018

Learning to Actively Learn Neural Machine Translation



Liu et al., **Learning to Actively Learn Neural Machine Translation**, CoNLL, 2018

Microsoft

Learning to Actively Learn Neural Machine Translation

System	EN→DE	EN→FI	EN→DE	EN→CS	EN→CS	EN→FI
Base NMT (100K)	13.2	10.3	13.2	8.1	8.1	10.3
AL with 135K token budget						
Random	13.9	11.2	13.9	8.3	8.3	11.2
Shortest	14.5	11.5	14.5	**8.6**	8.6	11.5
Longest	14.1	11.3	14.1	8.2	8.2	11.3
TTE	14.2	11.3	14.2	8.5	8.5	11.3
Token Policy	**15.5**	**12.8**	**14.8**	8.5	**9.0**	**12.5**
AL with 677K token budget						
Random	15.9	13.5	15.9	**9.2**	9.2	13.5
Shortest	15.8	13.7	15.8	8.9	8.9	13.7
Longest	15.6	13.5	15.6	8.5	8.5	13.5
TTE	15.6	13.7	15.6	8.6	8.6	13.7
Token Policy	**16.6**	**14.1**	**16.3**	9.2	**10.3**	**13.9**
FULL bitext (500K)	20.5	18.3	20.5	12.1	12.1	18.3

Table 1: BLEU scores on tests sets with different selection strategies, the budget is at token level with annotation for 135.45k tokens and 677.25k tokens respectively.

Liu et al., **Learning to Actively Learn Neural Machine Translation**, CoNLL, 2018

Microsoft

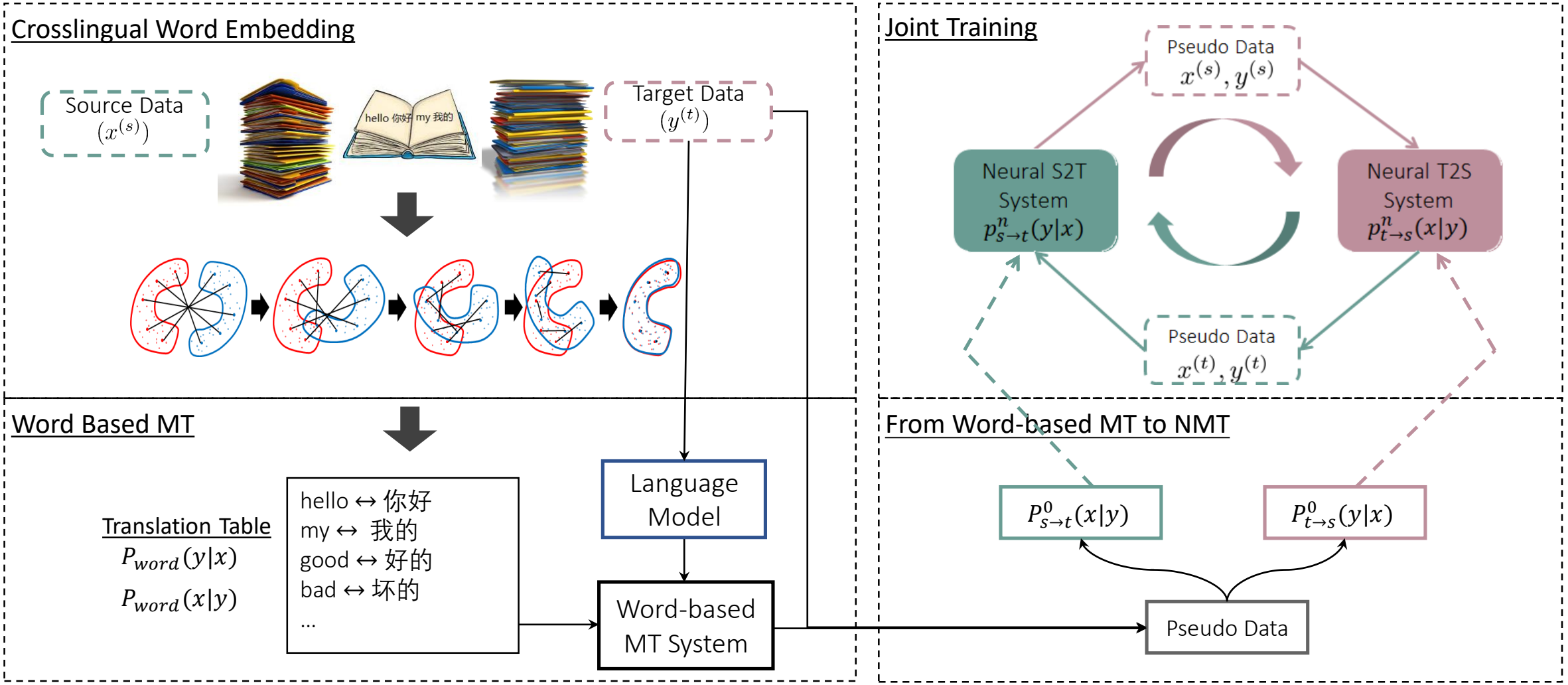
Unsupervised Neural Machine Translation

Microsoft

Unsupervised Neural Machine Translation

- **Cross-Lingual Word Embedding**
- Joint Training for Unsupervised NMT
- Posterior Regularization with SMT
- Explicit Pre-training for Unsupervised NMT

General Framework of UNMT



Cross Lingual Word Embedding

Seed Dictionary

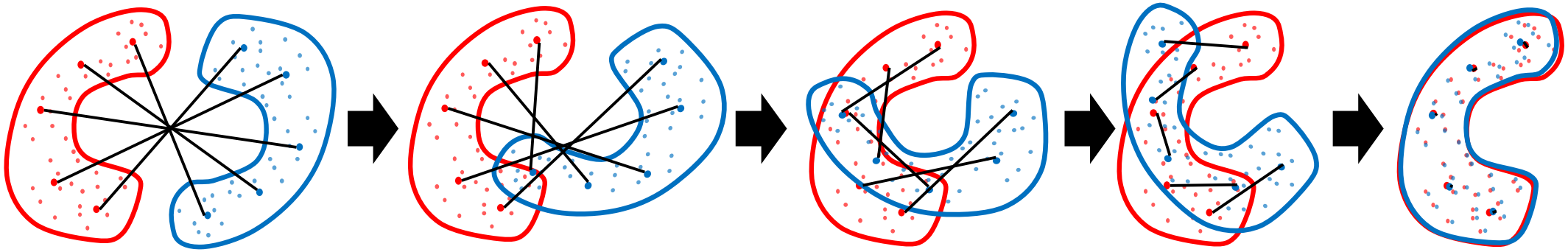
hello	你好
my	我的
good	好
…	…
are	是

Monolingual Embedding

hello	$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix}$	$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix}$ 你好
my		我的
good		好
…		…
welcome		欢迎

Cross Lingual Embedding: $\underset{W}{\operatorname{argmin}} \sum_i \|x_i W - y_i\|^2$

hello	$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix}$	$[W] \approx$	$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix}$ 你好
my			我的
good			好
…			…
welcome			欢迎



Artetxe et al., **Learning bilingual word embeddings with (almost) no bilingual data**, ACL, 2017

Cross Lingual Word Embedding

Algorithm 1 Traditional framework

Input: X (source embeddings)
Input: Z (target embeddings)
Input: D (seed dictionary)

1: $W \leftarrow$ LEARN_MAPPING(X, Z, D)
2: $D \leftarrow$ LEARN_DICTIONARY(X, Z, W)
3: EVALUATE_DICTIONARY(D)

Algorithm 2 Proposed self-learning framework

Input: X (source embeddings)
Input: Z (target embeddings)
Input: D (seed dictionary)

1: **repeat**
2: $W \leftarrow$ LEARN_MAPPING(X, Z, D)
3: $D \leftarrow$ LEARN_DICTIONARY(X, Z, W)
4: **until** convergence criterion
5: EVALUATE_DICTIONARY(D)

Artetxe et al., **Learning bilingual word embeddings with (almost) no bilingual data**, ACL, 2017

Cross Lingual Word Embedding

	English-Italian			English-German			English-Finnish		
	5,000	**25**	**num.**	**5,000**	**25**	**num.**	**5,000**	**25**	**num.**
Mikolov et al. (2013a)	34.93	0.00	0.00	35.00	0.00	0.07	25.91	0.00	0.00
Xing et al. (2015)	36.87	0.00	0.13	41.27	0.07	0.53	28.23	0.07	0.56
Zhang et al. (2016)	36.73	0.07	0.27	40.80	0.13	0.87	28.16	0.14	0.42
Artetxe et al. (2016)	39.27	0.07	0.40	**41.87**	0.13	0.73	**30.62**	0.21	0.77
Our method	**39.67**	**37.27**	**39.40**	40.87	**39.60**	**40.27**	28.72	**28.16**	**26.47**

Table 1: Accuracy (%) on bilingual lexicon induction for different seed dictionaries

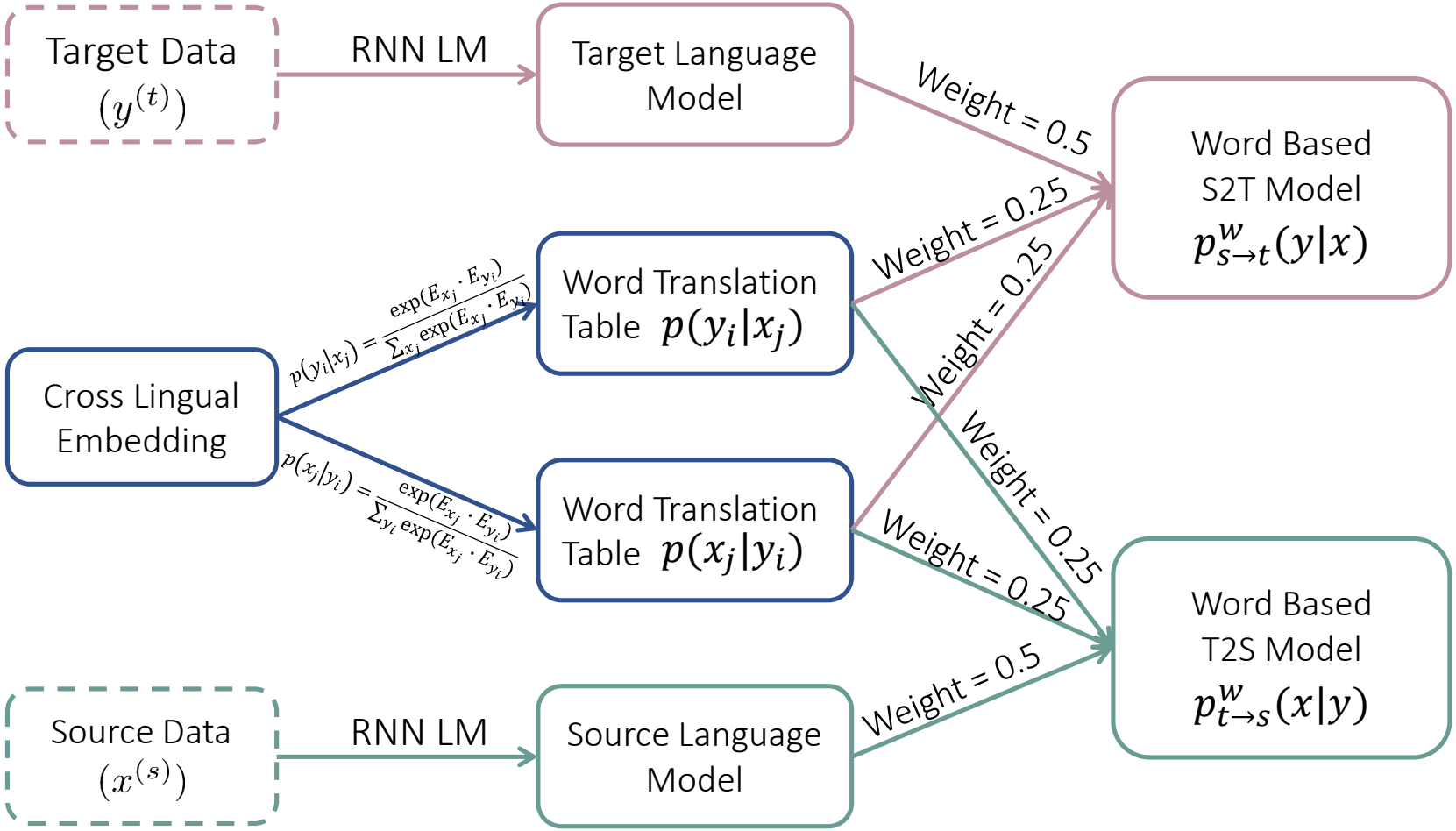
Artetxe et al., **Learning bilingual word embeddings with (almost) no bilingual data**, ACL, 2017

Microsoft

Unsupervised Neural Machine Translation

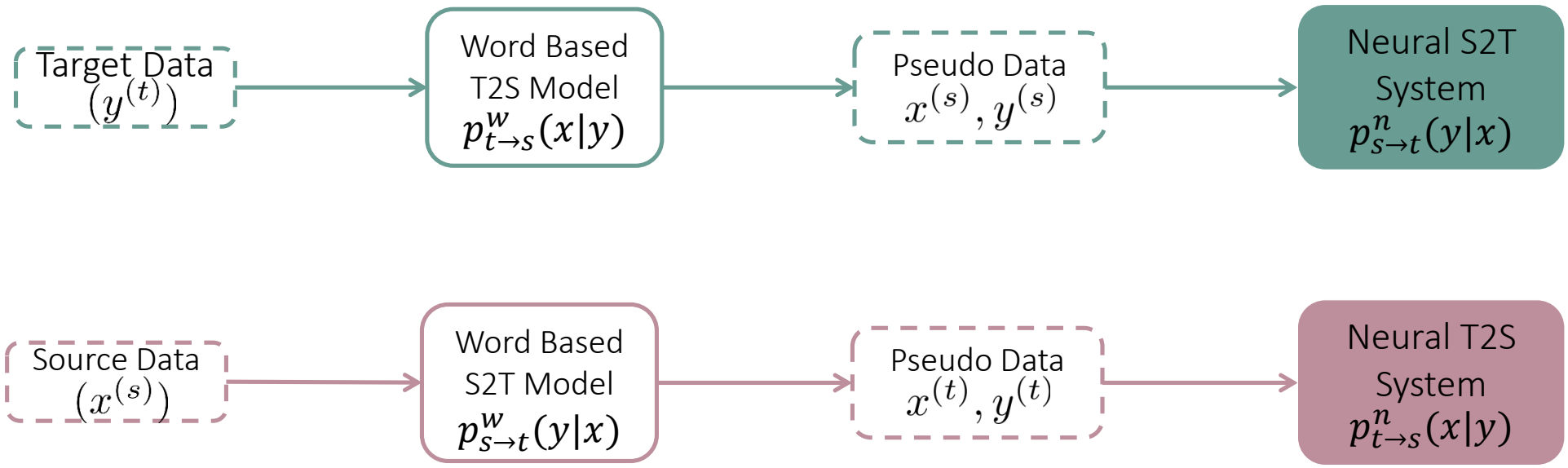
- Cross-Lingual Word Embedding
- **Joint Training for Unsupervised NMT**
- Posterior Regularization with SMT
- Explicit Pre-training for Unsupervised NMT

Word-based Machine Translation



Lample et al., **Phrase-Based & Neural Unsupervised Machine Translation**. EMNLP, 2018

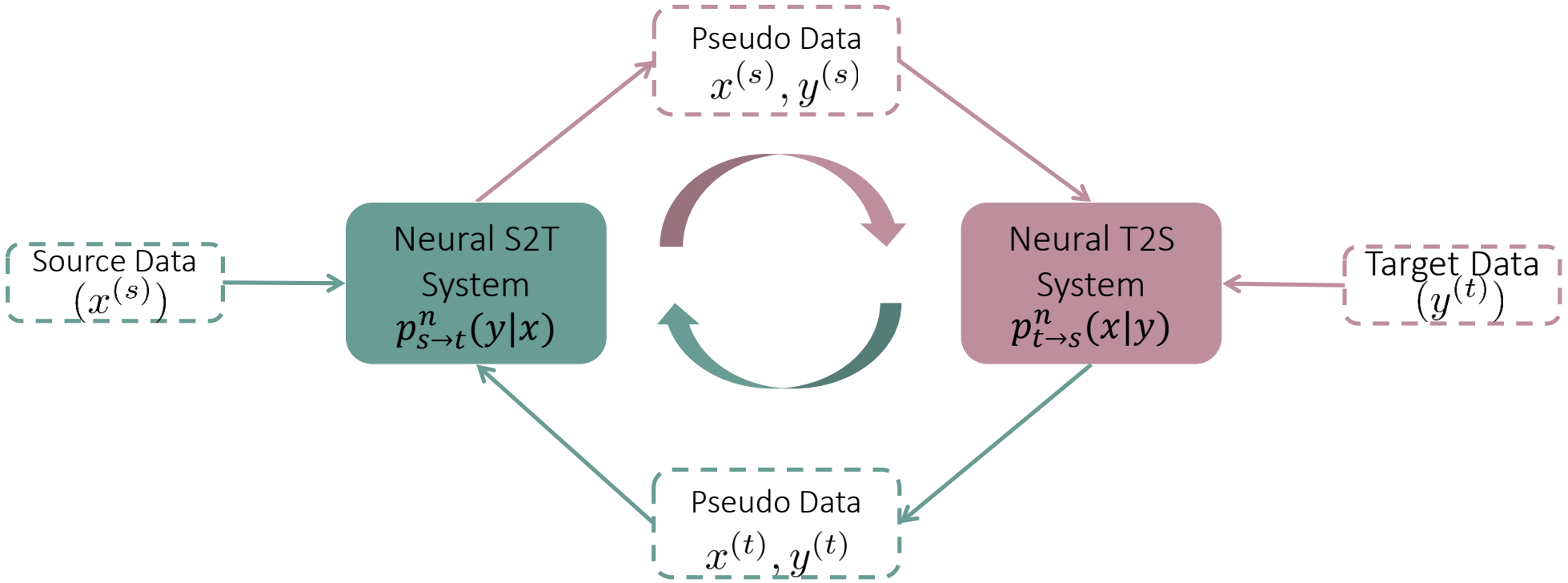
From Word-based MT to NMT



*Back Translation is used

Lample et al., **Phrase-Based & Neural Unsupervised Machine Translation**. EMNLP, 2018

Joint Training for Unsupervised NMT



Lample et al., **Phrase-Based & Neural Unsupervised Machine Translation**. EMNLP, 2018

Microsoft

Joint Training for Unsupervised NMT

Model	en-fr	fr-en	en-de	de-en
(Artetxe et al., 2018)	15.1	15.6	-	-
(Lample et al., 2018)	15.0	14.3	9.6	13.3
(Yang et al., 2018)	17.0	15.6	10.9	14.6
NMT (LSTM)	24.5	23.7	14.7	19.6
NMT (Transformer)	25.1	24.2	17.2	21.0
PBSMT (Iter. 0)	16.2	17.5	11.0	15.6
PBSMT (Iter. n)	**28.1**	27.2	17.9	22.9
NMT + PBSMT	27.1	26.3	17.5	22.1
PBSMT + NMT	27.6	**27.7**	**20.2**	**25.2**

Lample et al., **Phrase-Based & Neural Unsupervised Machine Translation**. EMNLP, 2018

Microsoft

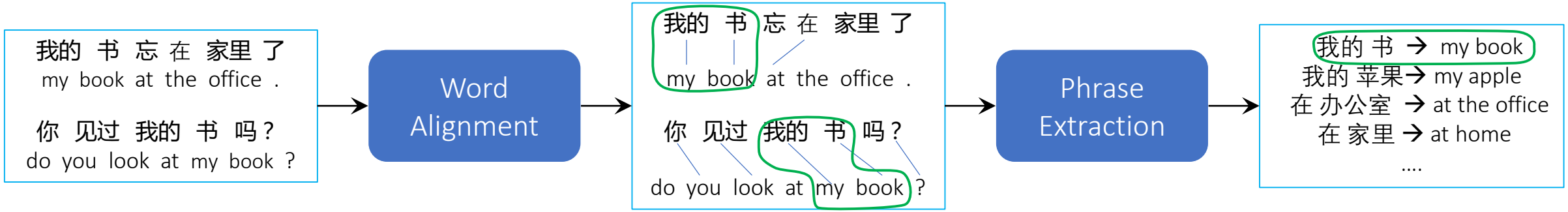
Unsupervised Neural Machine Translation

- Cross-Lingual Word Embedding
- Joint Training for Unsupervised NMT
- **Posterior Regularization with SMT**
- Explicit Pre-training for Unsupervised NMT

Posterior Regularization with SMT

- Pseudo data generated may contain errors.
  - Leading training process to be slow or stacked in a bad local minimal.

- Extract good phrase pairs to guide NMT training.

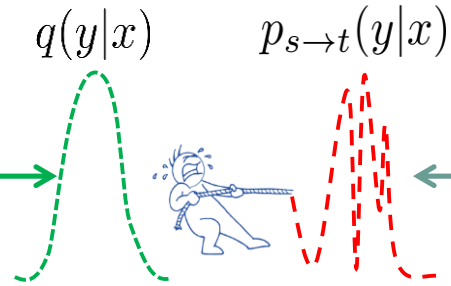


Ren et al., **Unsupervised Neural Machine Translation with SMT as Posterior Regularization**. AAAI, 2019

Posterior Regularization with SMT

我的 书 → my book
我的 苹果→ my apple
在 办公室 → at the office
在 家里 → at home
....

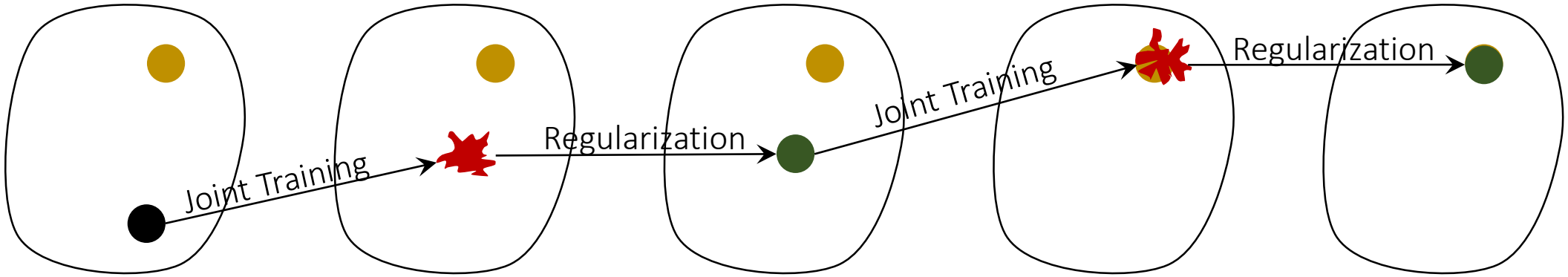
$$q(y|x) = \frac{exp(\sum_i(\lambda_i f_i(x,y)))}{\sum_y exp(\sum_i(\lambda_i f_i(x,y)))}$$



Neural S2T System $p^n_{s\to t}(y|x)$

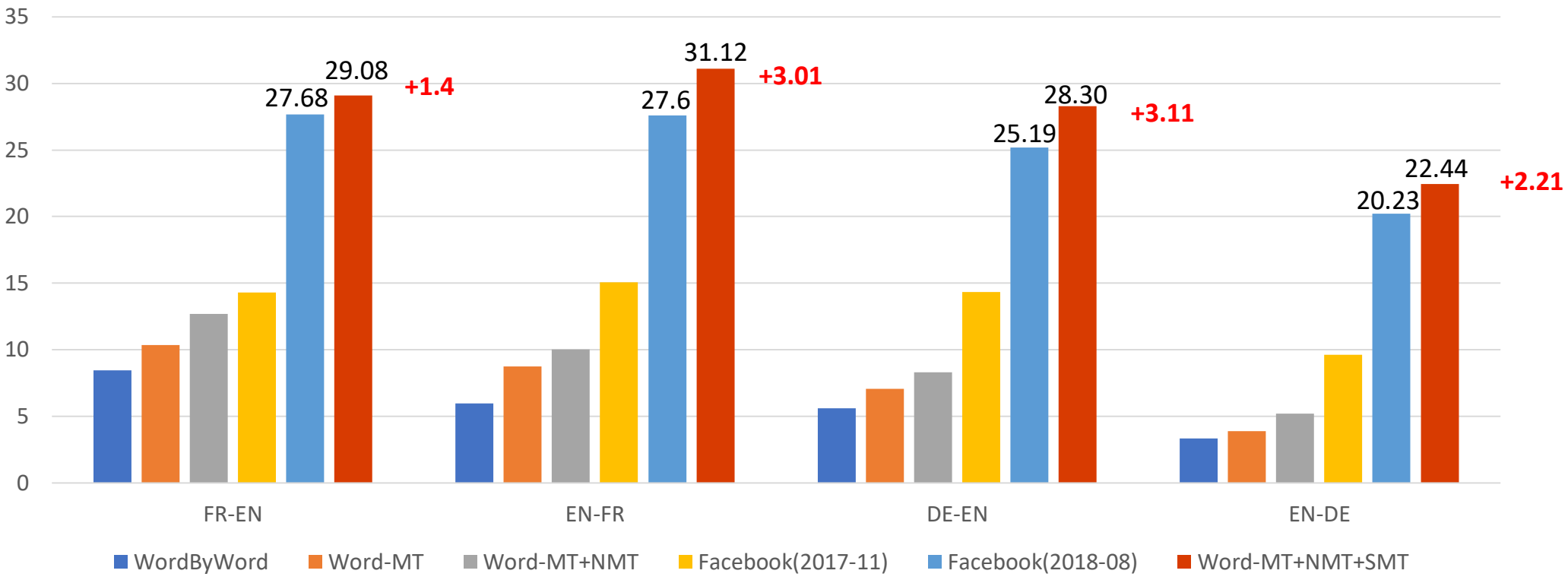
Pseudo Data $x^{(s)}, y^{(s)}$

$$L(\theta_{s\to t}) = \lambda \mathrm{KL}\left(q(y|x) \| p_{s\to t}(y|x)\right) - \log p_{s\to t}(y|x)$$



Ren et al., **Unsupervised Neural Machine Translation with SMT as Posterior Regularization**. AAAI, 2019

Posterior Regularization with SMT



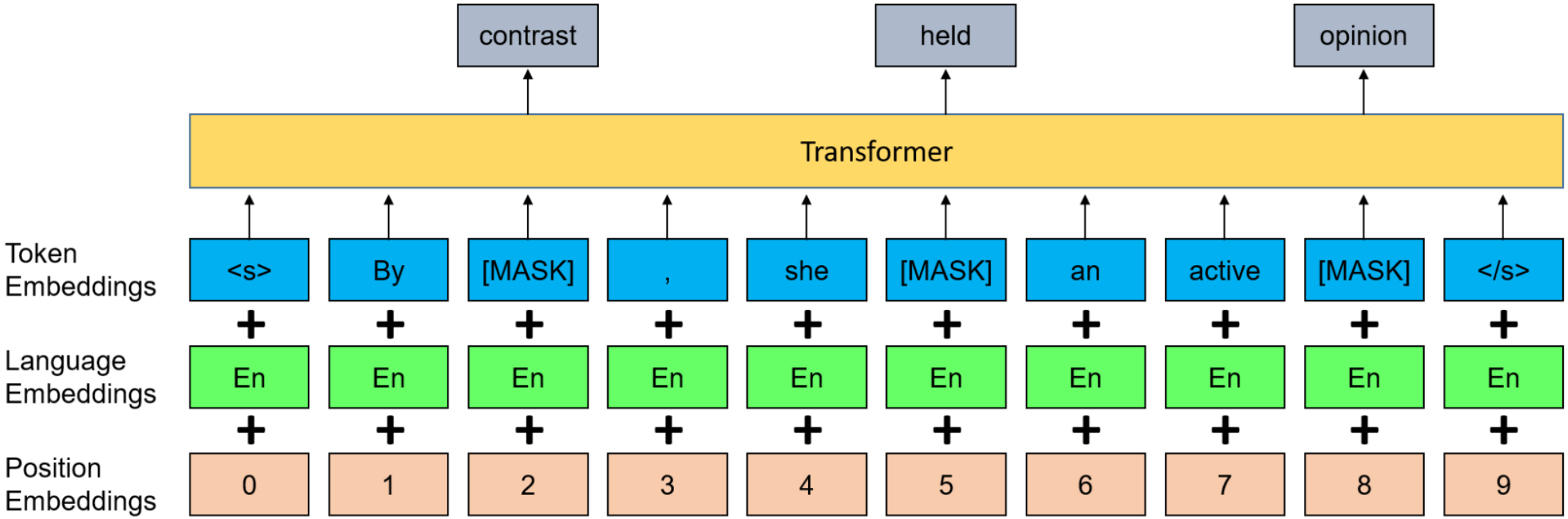
Ren et al., **Unsupervised Neural Machine Translation with SMT as Posterior Regularization**. AAAI, 2019

Microsoft

Unsupervised Neural Machine Translation

- Cross-Lingual Word Embedding
- Joint Training for Unsupervised NMT
- Posterior Regularization with SMT
- **Explicit Pre-training for Unsupervised NMT**

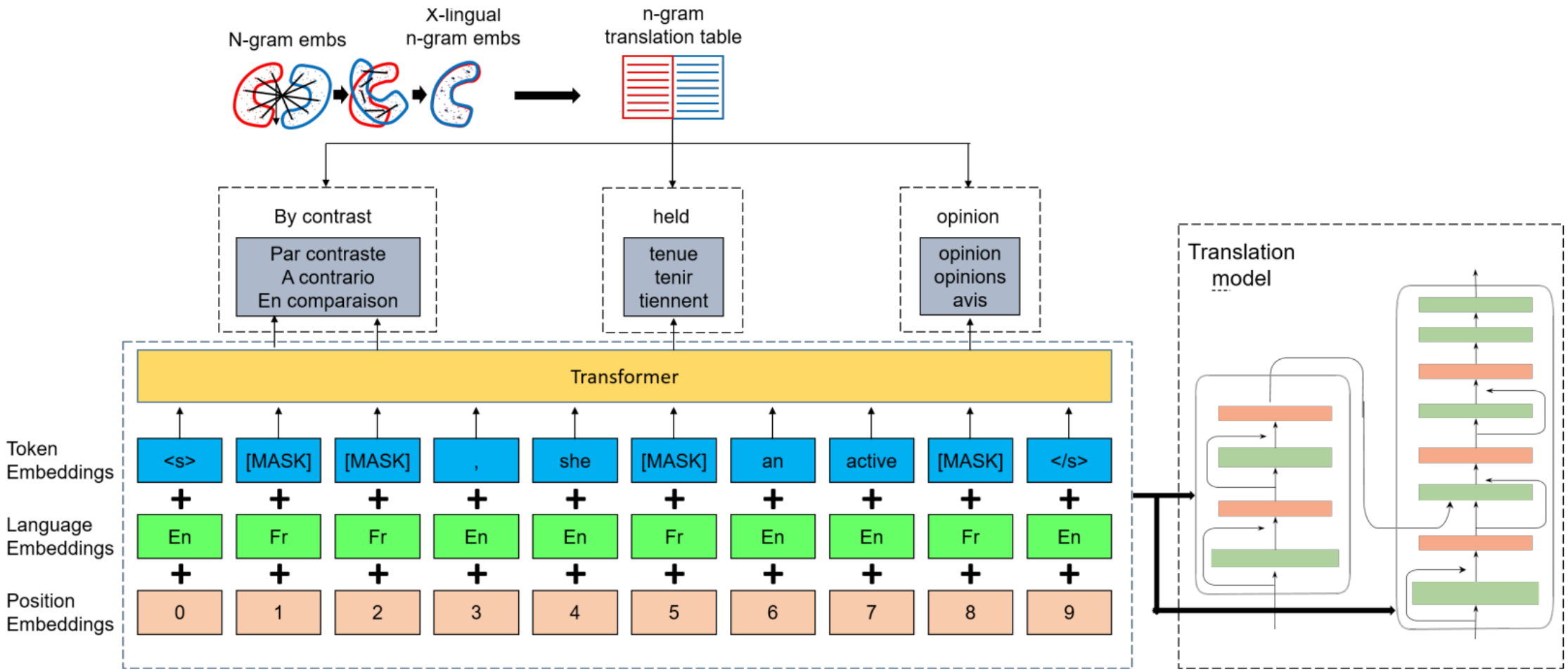
Explicit Pre-training for Unsupervised NMT



MLM: Masked Language Model

Ren et al., **Explicit Cross-lingual Pre-training for Unsupervised Machine Translation**. EMNLP, 2019

Explicit Pre-training for Unsupervised NMT



CMLM: Crosslingual Masked Language Model

Pre-trained NMT Model

Ren et al., **Explicit Cross-lingual Pre-training for Unsupervised Machine Translation**. EMNLP, 2019

Explicit Pre-training for Unsupervised NMT

$$\mathcal{L}_{pre} = \mathcal{L}_{cmlm} + \mathcal{L}_{mlm}$$

$$\mathcal{L}_{cmlm} = -\log \mathbf{Pr}(y_1^m|x_1^l) = -\log \epsilon - \sum_{j=1}^{m} \log \sum_{i=0}^{l} a(i|j,l,m)p(y_j|x_i)$$

$$\mathbf{Pr}(y_1^m|x_1^l) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} a(i|j,l,m)p(y_j|x_i)$$

$$\nabla_\theta \mathcal{L}_{cmlm} = -\sum_{j=1}^{m} \frac{a(i|j,l,m)p(y_j|x_i)}{\sum_{i=0}^{l} a(i|j,l,m)p(y_j|x_i)} \nabla_\theta \log p(y_j|x_i)$$

Ren et al., **Explicit Cross-lingual Pre-training for Unsupervised Machine Translation**. EMNLP, 2019

Explicit Pre-training for Unsupervised NMT

	Method	fr2en	en2fr	de2en	en2de	ro2en	en2ro
Baselines	(Artetxe et al., 2017)	15.6	15.1	-	-	-	-
	(Lample et al., 2017)	14.3	15.1	13.3	9.6	-	-
	(Artetxe et al., 2018b)	25.9	26.2	23.1	18.2	-	-
	(Lample et al., 2018)	27.7	28.1	25.2	20.2	23.9	25.1
	(Ren et al., 2019)	28.9	29.5	26.3	21.7	-	-
	(Lample and Conneau, 2019)	33.3	33.4	34.3	26.4	31.8	33.3
CMLM	Iter 1	34.8	34.9	35.5	**27.9**	33.6	34.7
	Iter 2	**34.9**	**35.4**	**35.6**	27.7	**34.1**	**34.9**

Ren et al., **Explicit Cross-lingual Pre-training for Unsupervised Machine Translation**. EMNLP, 2019

Microsoft

Summary

- Introduction to Machine Translation
  - Background, Methods, Evaluation
- Neural Machine Translation
  - RNN, RNN-based NMT, Transformer, NMT Training
- Low Resource Neural Machine Translation
  - Monolingual Data, Other Languages, Multi-Task Learning, Pre-trained Model and Transfer Learning, Active Learning
- Unsupervised Neural Machine Translation
  - Cross-lingual Word Embedding, Joint Training for UNMT, SMT as Posterior Regularization, Explicit Pre-training

Microsoft

Thank You for Your Attention!

Reference

- Papineni, et al. BLEU: a Method for Automatic Evaluation of Machine Translation. ACL, 2002
- Sutskever, et al. Sequence to sequence learning with neural networks. NIPS, 2014
- Bahdanau, et al. Neural Machine Translation by Jointly Learning to Align and Translate.ICLR, 2015
- Vaswani, et al. Attention is all you need. NeurIPS, 2017
- Gulcehre et al., On Using Monolingual Corpora in Neural Machine Translation. arxiv, 2015.
- Cheng et al., Semi-Supervised Learning for Neural Machine Translation. ACL, 2016.
- He et al., Dual Learning for Machine Translation. NIPS, 2016.
- Sennrich et al., Improving Neural Machine Translation Models with Monolingual Data. ACL, 2016.
- Zhang et al., Joint Training for Neural Machine Translation Models with Monolingual Data. AAAI, 2018.

Reference

- Chen et al., A Teacher-Student Framework for Zero-Resource Neural Machine Translation. ACL, 2017.
- Ren et al., Triangular Architecture for Rare Language Translation. ACL, 2018.
- Dong et al., Multi-Task Learning for Multiple Language Translation. ACL, 2015.
- Johnson et al., Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. TACL, 2017.
- Zhang et al., Regularizing Neural Machine Translation by Target-Bidirectional Agreement. AAAI, 2019.
- Lample et al., Cross-lingual Language Model Pretraining, arxiv, 2019.
- Yang et al., Towards Making the Most of BERT in Neural Machine Translation, arxiv, 2019.
- Weng et al., Improving Neural Machine Translation with Pre-trained Representation, arxiv, 2019.

Reference

- Kim et al., Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies, ACL 2019.
- Liu et al., Learning to Actively Learn Neural Machine Translation, CoNLL, 2018
- Artetxe et al., Learning bilingual word embeddings with (almost) no bilingual data, ACL, 2017
- Lample et al., Phrase-Based & Neural Unsupervised Machine Translation. EMNLP, 2018
- Ren et al., Unsupervised Neural Machine Translation with SMT as Posterior Regularization. AAAI, 2019
- Ren et al., Explicit Cross-lingual Pre-training for Unsupervised Machine Translation. EMNLP, 2019
- Gu et al., Meta-learning for Low-resource Neural Machine Translation , EMNLP 2018.
- Gu et al., Universal Neural Machine Translation for Extremely Low Resource Languages. NAACL-HLT, 2018.

Reference

- Al-Shedivat and P. Parikh, Consistency by Agreement in Zero-shot Neural Machine Translation . NAACL, 2019.
- Xia et al., Generalized Data Augmentation for Low-resource Translation . ACL, 2019.