

自然语言处理

2022年秋季

黄河燕、鉴萍

北京理工大学 计算机学院

hhy63, pjian@bit.edu.cn

机器翻译

黄河燕、鉴萍

北京理工大学 计算机学院

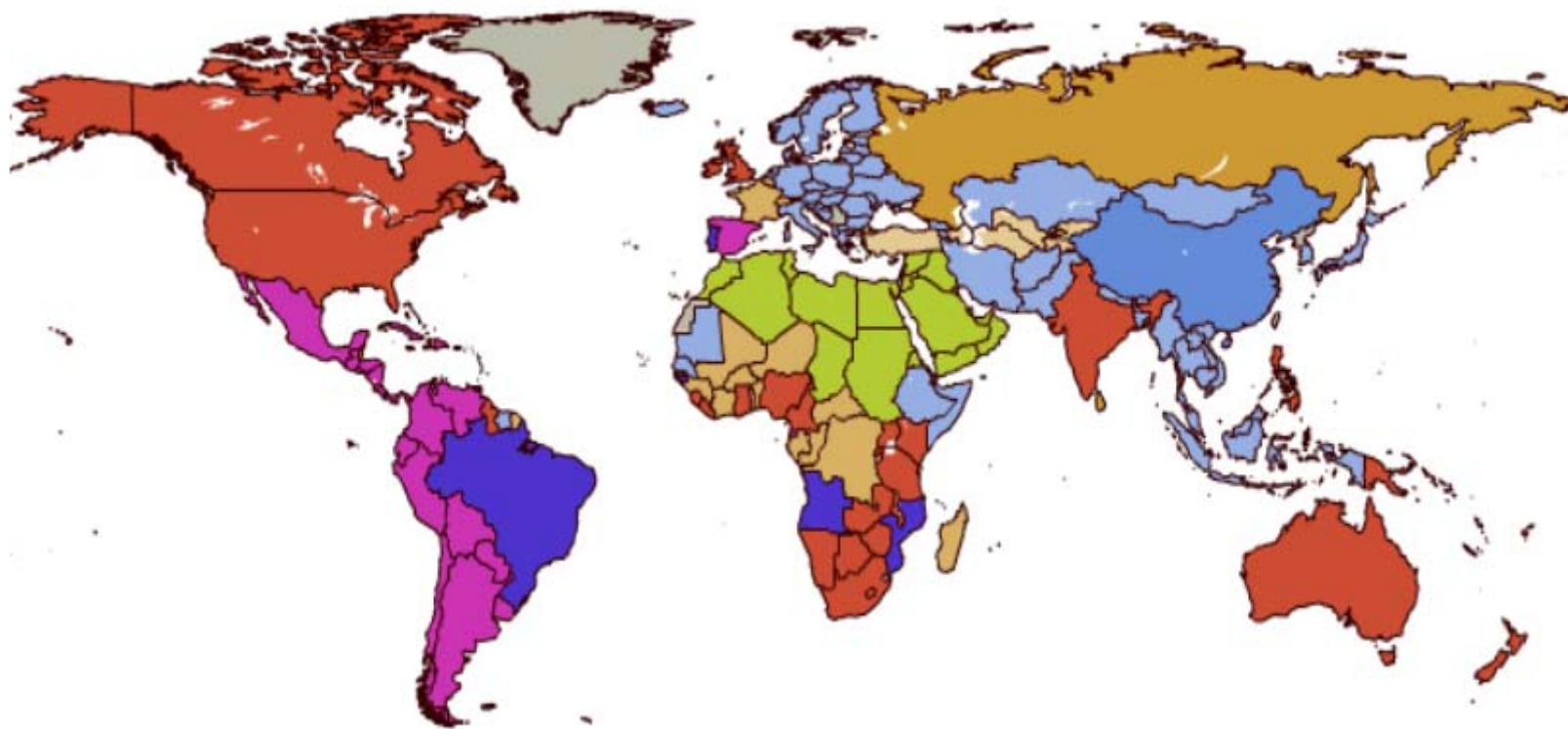
pjian@bit.edu.cn

巴别塔的故事

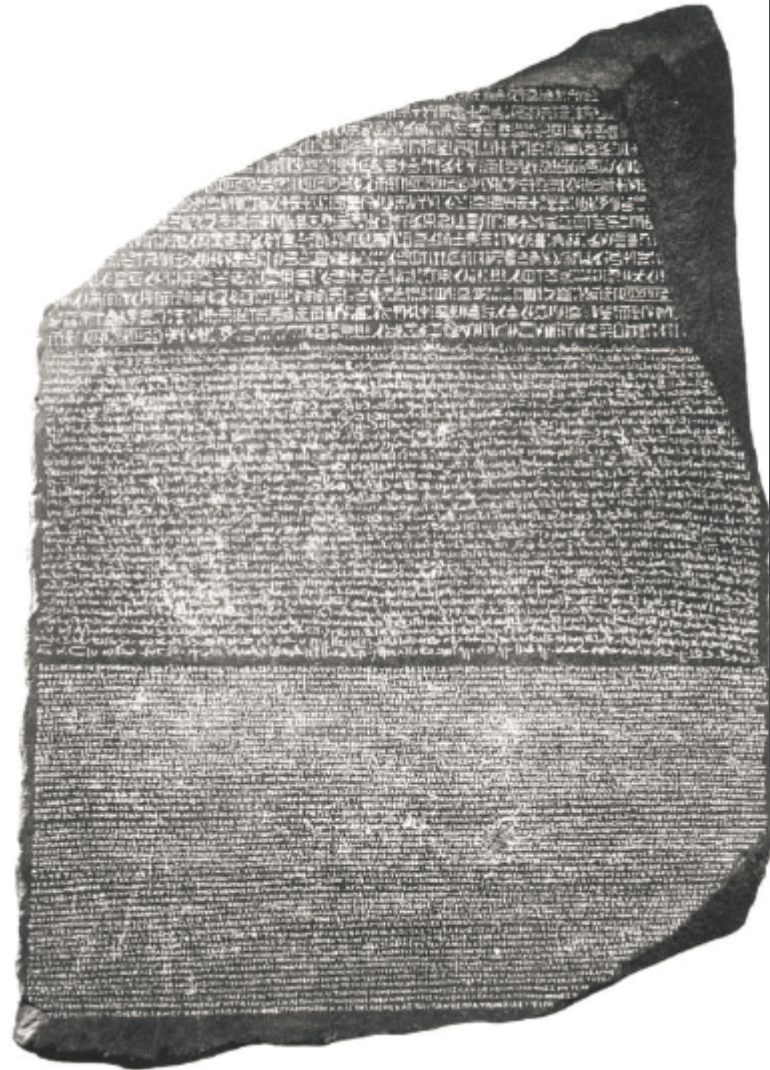
Babel



世界上语言的分布



Rosetta Stone



	a		j		s
	b		k		t
	c		l		u
	d		m		v
	e		n		w
	f		o		x
	g		p		y
	h		q		z
	i		r		

一些翻译的例子

黛玉自在枕上感念宝钗……

又听见窗外竹梢蕉叶之上，雨声淅沥，清寒透幕，不觉又滴下泪来。

As she lay there alone, **Dai-yu**'s thoughts turned to **Bao-chai**...

Then she listened to the insistent rustle of the rain on the **bamboos and plantains** outside **her** window. The coldness **penetrated** the **curtains of her bed**. Almost without noticing it she had begun to cry.

——Translated by David Hawkes

My enemies are many, my equals are none. In the shade of olive trees, they said Italy could never be **conquered**. In the **land** of Pharaohs and kings, they said Egypt could never be **humbled**. In the **realm** of forest and snow, they said Russia could never be **tamed**. Now **they say nothing**. They fear me, like a force of nature, a dealer in thunder and death. I say I am Napoleon, I am emperor..... Burn it.

Translation by Google (SMT):

“我的敌人很多，我等于没有。在橄榄树的树荫下，他们说意大利永远无法征服。在法老和国王的土地，他们说，埃及永远卑微。森林和雪的境界，他们说，俄罗斯可能永远不会被驯服。现在他们什么都不说。他们害怕我，就像大自然的力量，雷鸣和死亡的经销商。我说我是拿破仑，我是皇帝.....燃烧。”

Translation by Google (NMT):

“我的敌人很多，我的平等没有。在橄榄树的阴影中，他们说意大利永远不可征服。他们说，在法老和国王的土地上，埃及永远不会谦卑。他们说，在森林和雪的领域，俄罗斯永远不会被驯服。现在他们什么也没说。他们像自然力量一样恐惧我，是雷电和死亡的经销商。我说我是拿破仑，我是皇帝……燃烧它。”

Translation by human:

“我树敌无数，却从未逢对手。在橄榄树荫下，他们说意大利永远不会被征服。在法老和国王的土地上，他们说埃及永远不会臣服。在森林与暴雪的国度，他们说俄国永远不会屈服。现在他们已无话可说。他们畏惧我，如同畏惧带来雷霆和死亡的自然的力量。我就是拿破仑，我就是皇帝……烧掉它！”

机器翻译

人工翻译为古汉语：

“朕之仇寇多矣，然敌手则未之有也。大秦、大食、罗刹，皆自诩不可胜之，而今寂然。彼畏朕，犹若畏天。朕，天之子也……焚！”

机器翻译

“译事三难：信、达、雅。求其信已大难矣，
顾信矣不达，虽译犹不译也，则达尚焉。”

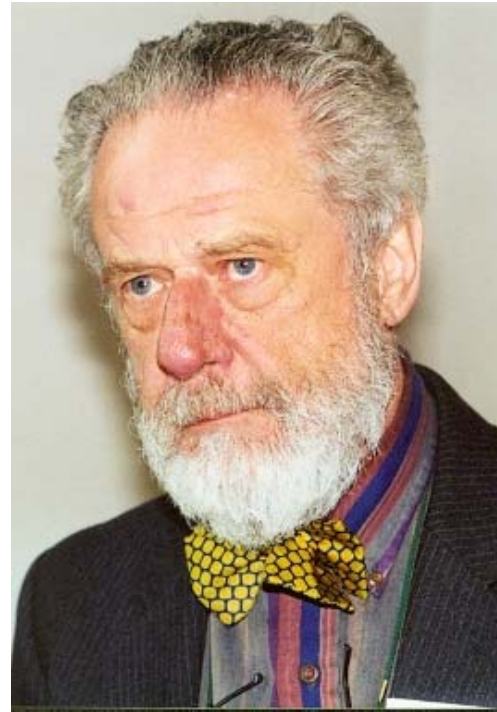
——严复《天演论》

- “信”：意义不背原文，即是译文要准确，不歪曲，不遗漏，也不要随意增减意思；
- “达”：不拘泥于原文形式，译文通顺明白；
- “雅”：译文时选用的词语要得体，追求文章本身的古雅，简明优雅。

机器翻译

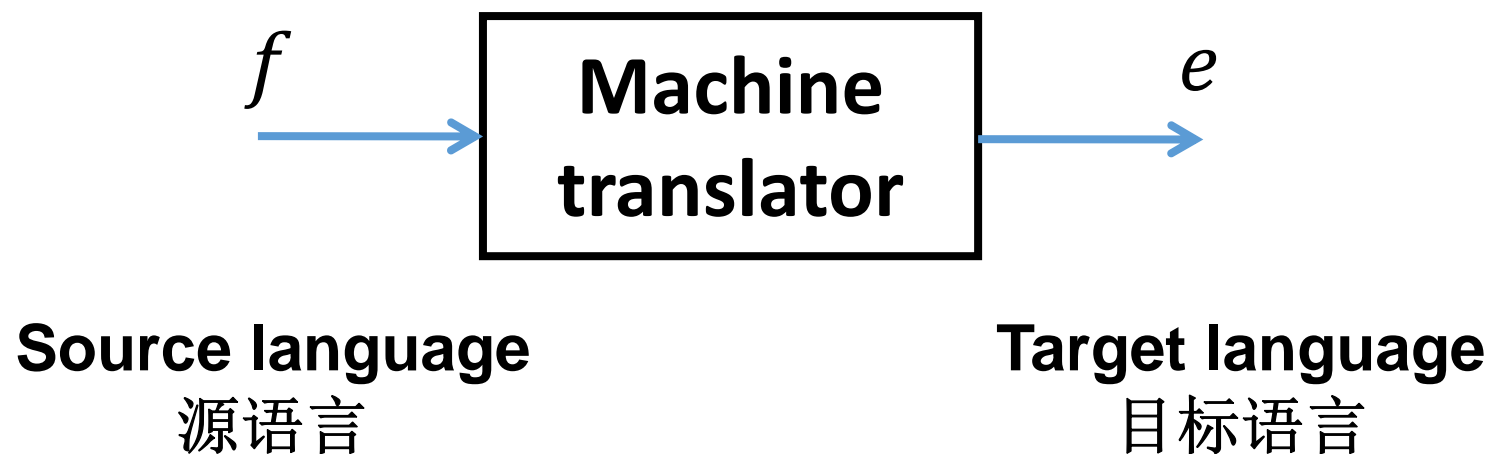
- Translation is a fine and exacting art, but there is much about it that is **mechanical and routine**.

——Martin Kay(1980)



机器翻译

- 机器翻译(machine translation)：使用计算机自动将一种语言翻译成另一种语言的技术



大纲

□ 历史

□ 方法

- RBMT
- EBMT
- SMT
- NMT

历史

□ Before Weaver

- 1930s
 - 机器脑 (G. B. Arsouni, 法国工程师)
 - 翻译的机械方法 (П.П.ТРОЯНСКИЙ, 前苏联发明家)

历史

□ Warren Weaver

- 1946, 酝酿, Booth and Weaver
- 1947, 遭到维纳反对, Weaver and Wiener
- 1949, “Translation”备忘录发表, Weaver

历史



□ When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

——Warren Weaver(1947)

历史

▣ IBM

- 1954年，第一个机器翻译系统, Georgetown University和IBM
 - 系统只有**250**条俄语词汇，**6**条语法规则，将**60**个俄语词组翻译成了英语

IBM-701



历史

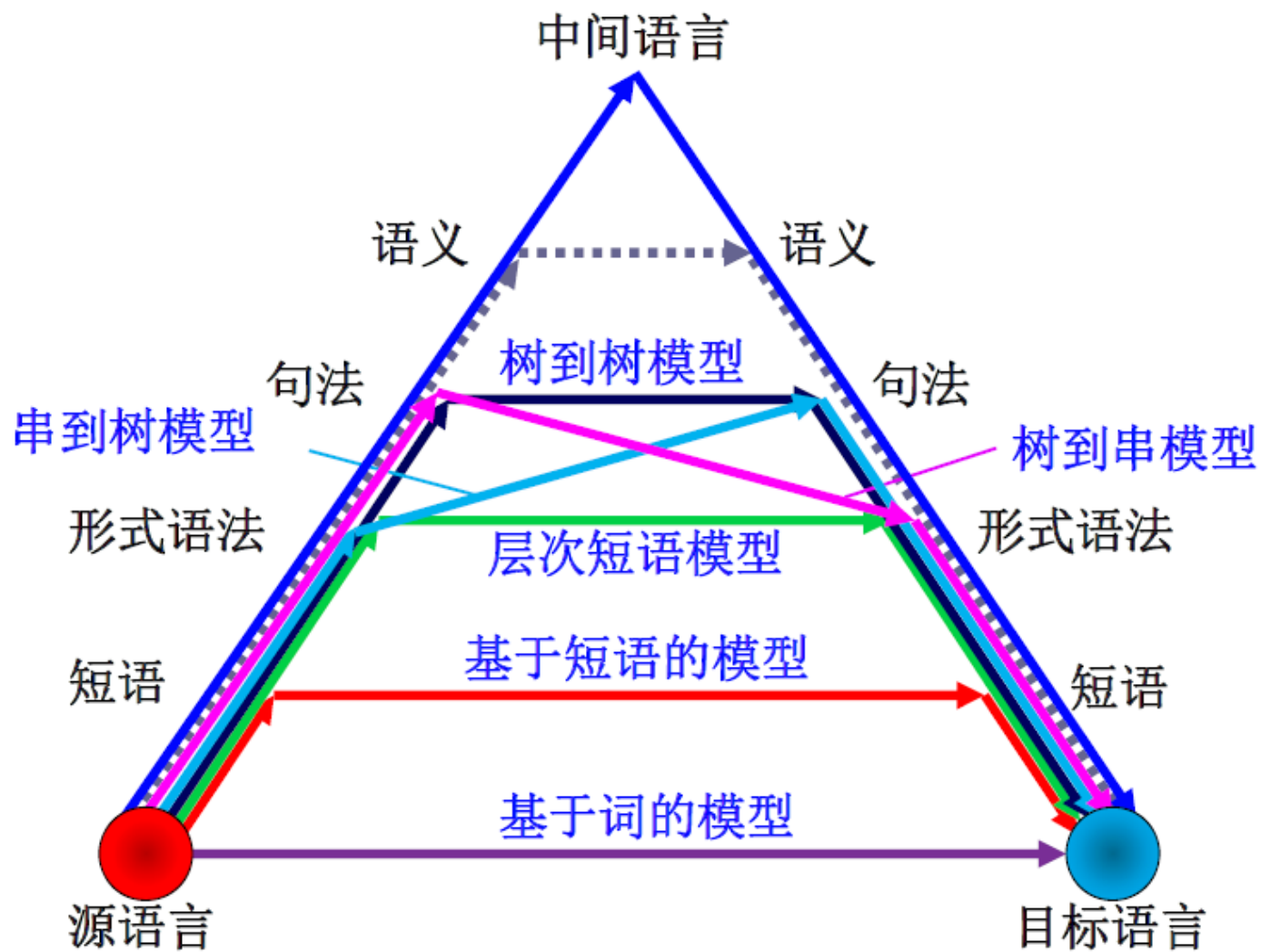
▣ ALPAC报告

- 1964, 语言自动处理委员(Automatic Language Processing Advisory Committee, ALPAC)
- 1966, 语言与机器报告(ALPAC报告)
 - 在目前给机器翻译以大力支持还没有多少理由
 - 机器翻译遇到了难以克服的语义障碍
 - **NLP走向NLU**
 - **AMTCL→ACL**
- 定位错了，模型本身也太简陋

历史

▣ 基于NLU的规则翻译

- 1970s前, word2word→句法转换
- 1970s, 商业机器翻译系统
- 我国从1959年开始机器翻译研究, 俄汉翻译系统
- 1992, 快译通EC-863A



历史

□ IBM

- 1990s, 提出了5个word-based翻译系统, SMT的开端

A Statistical Approach to Machine Translation

Peter F. Brown John Cocke Stephen A. Della Pietra
Vincent J. Della Pietra Fredrick Jelinek John D. Lafferty
Robert L. Mercer Paul S. Roossin

历史

- Some of us started to wonder in the mid of 1980s whether our ASR methods could be applied to new fields. Bob Mercer and I ... came up with two: **machine translation** and stock market modeling.

——Fred Jelinek (2009)



- 1999, JHU workshop 复现IBM工作, GIZA, Knight
- 优化IBM模型, GIZA++, Och
- 2002, 2003, SMT的对数线性模型和最小错误率训练方法(log-linear models with minimum error rate training), Och
- 2004, Pharaoh, Koehn
- 2005, Moses, Koehn



Kevin Knight



Franz Och



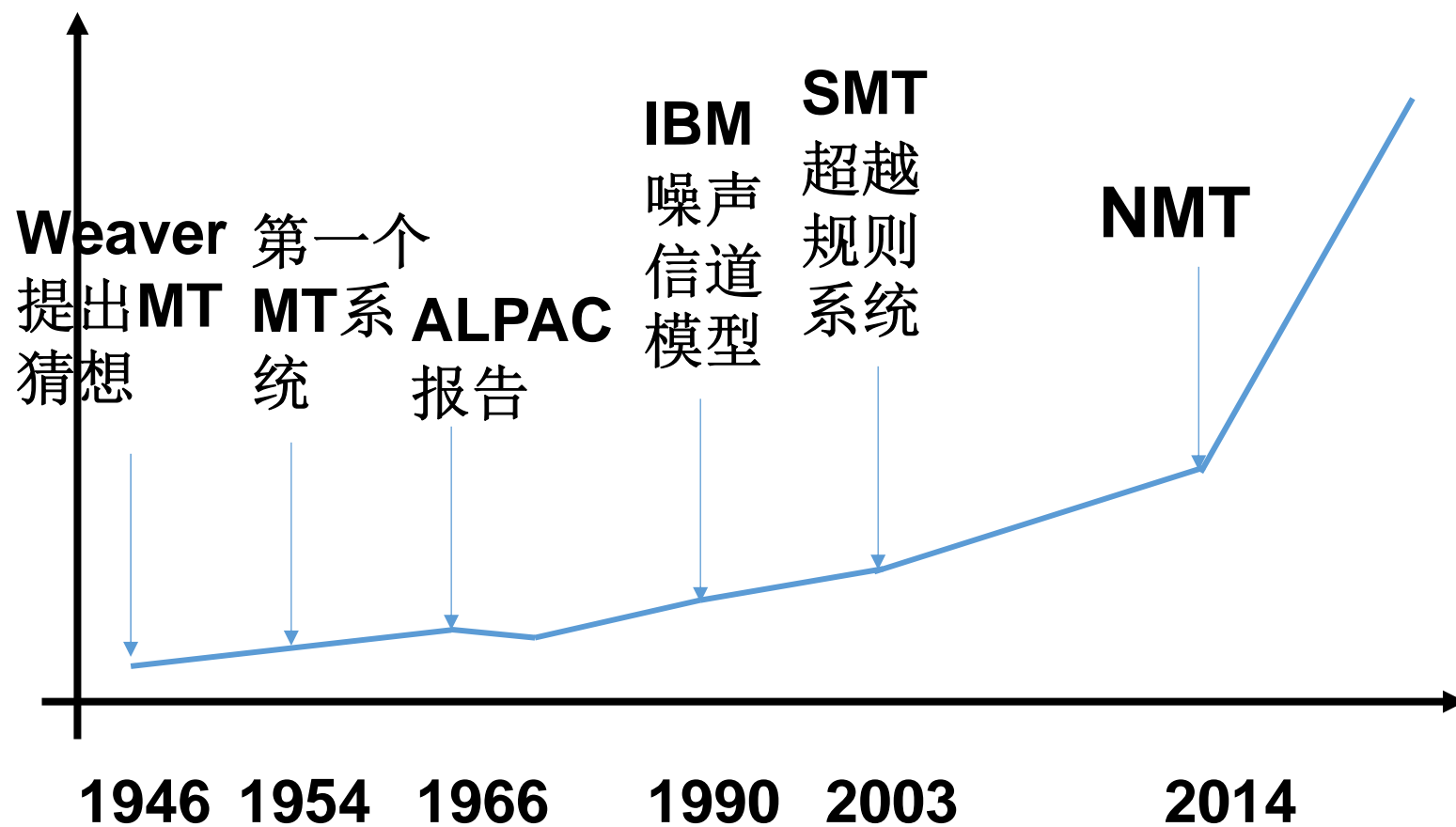
Philipp Koehn

历史

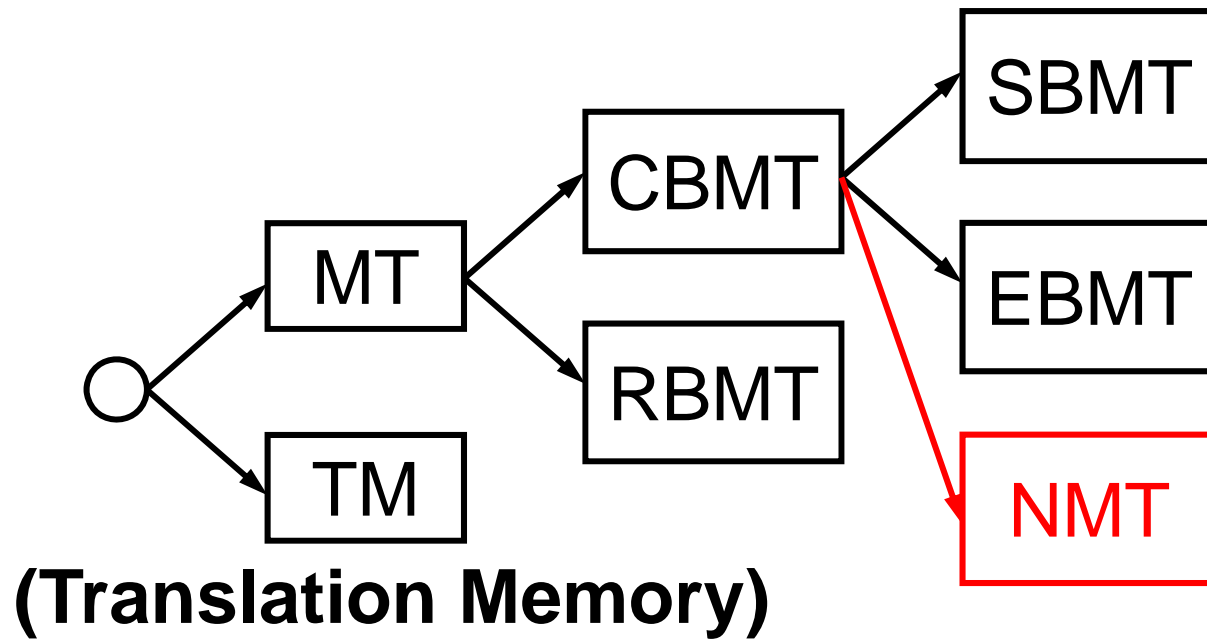
- 在机器翻译研究中实现人机共生(man-machine symbiosis), 人机互助, 比追求全自动的高质量的翻译(Full Automatic High Quality Translation, FAHQT) 更现实、更切合实际。

——Hutchins, 1995

历史



方法



—— Eiichiro Sumita (2002)

大纲

□ 历史

□ 方法

- RBMT
- EBMT
- SMT
- NMT

RBMT

- ▣ 基于规则的翻译(rule based machine translation, RBMT)：使用语言规则或语言知识
 - 直接翻译(使用双语词典)
 - 基于转换的(transfer-based)
 - 中间语言(interlingua)

RBMT

□ 直接翻译

➤ 举例

Chinese: 我 喜爱 运动。

Translate to English:  I  like  sports.

Trans to Japanese: 私 好きだ スポーツ

Reference : 私 は スポーツ が 好きだ

我(助词) 运动 (助词) 喜爱

直接翻译也需考虑目标语言的词法和句法特征，增加规则来改善目标译文的可读性

RBMT

□ 基于转换的

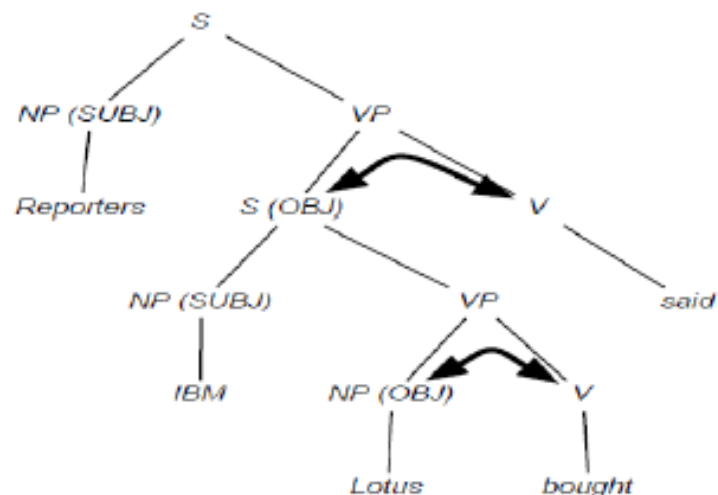
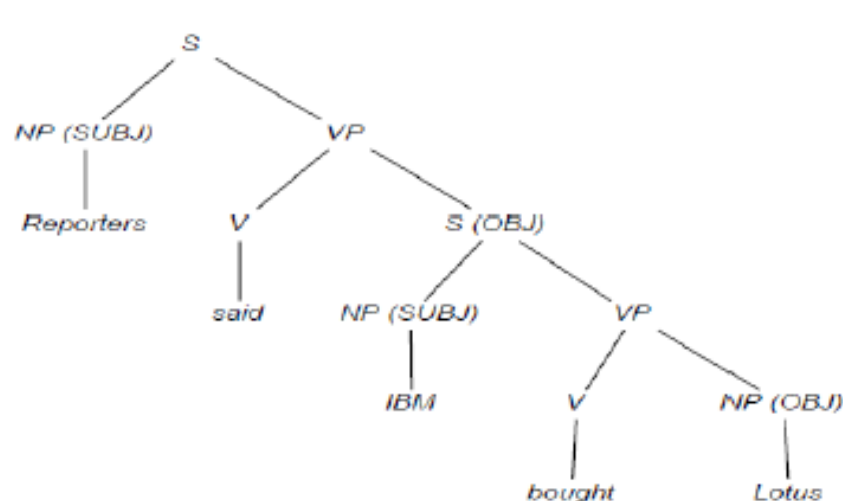
- Example: 一个从西班牙语翻译为英语的转换规则

gustar[SUBJ(ARG2:NP), OBJ1(ARG1:CASE)]

→ **like** [SUBJ(ARG1:NP), OBJ1(ARG2:NP)]

翻译对: **Maria me gustar → I like Mary.**

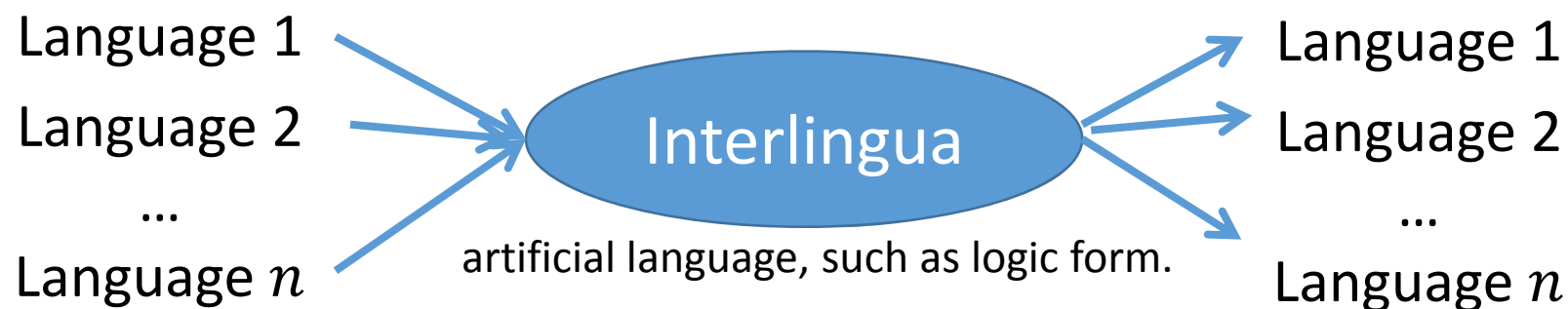
词到词的翻译体现了“信”，而后续的转换规则，则体现了“达”



在句法层面的转换

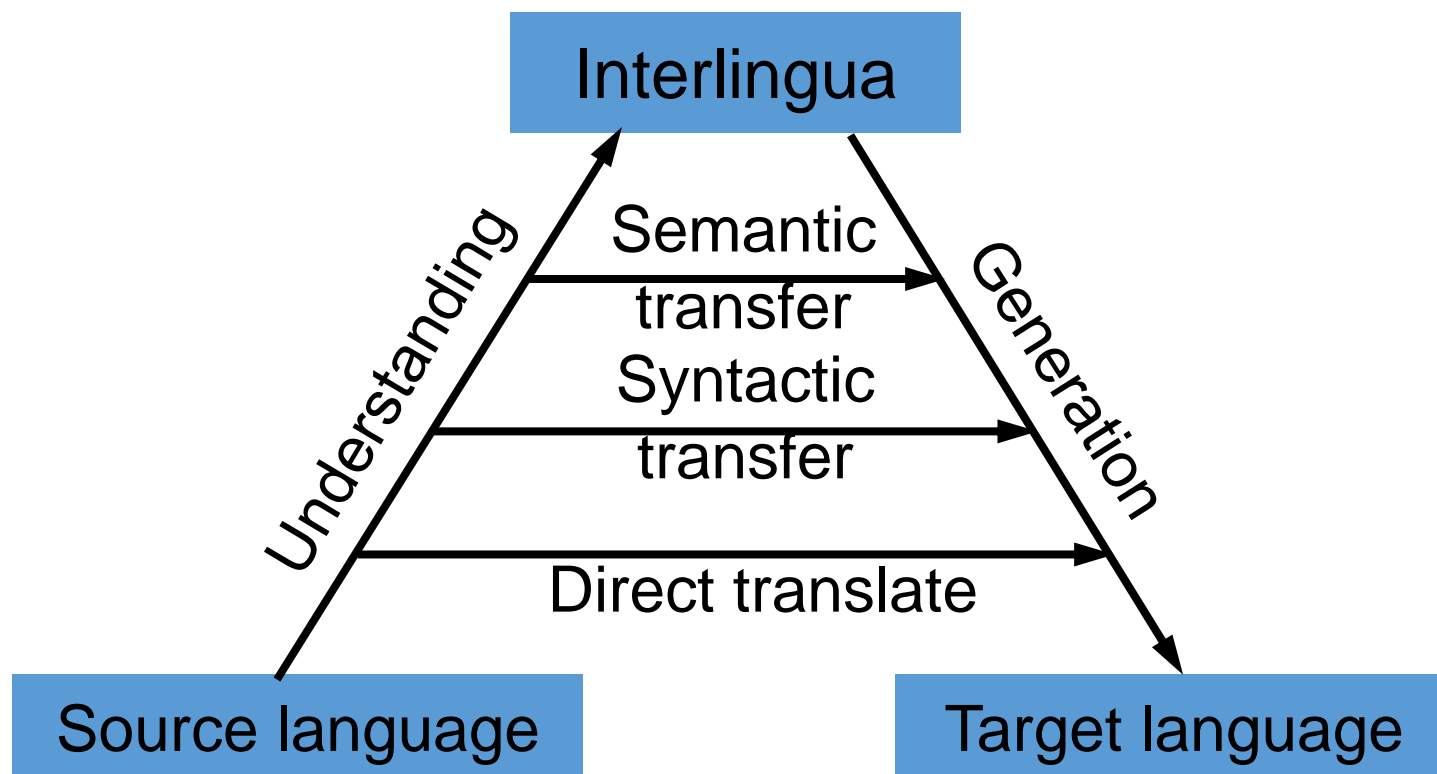
RBMT

□ 基于中间语言



- 中间语言可以是一种人工语言，比如logic form——也相当于一种转换

RBMT



RBMT

乐学上传有一个基于中间语言的规则翻译系统的完整实例，供课下参考

大纲

□ 历史

□ 方法

- RBMT
- EBMT
- SMT
- NMT

EBMT

□ 基于实例的翻译(example based MT, EBMT)

- 人们在翻译简单句子时并没有对语言进行深层分析，而是将句子**分解**为几个片段(短语)，**借助**已有片段的翻译将分解的片段译成目标短语，最后将这些短语**组合**为句子



Makoto Nagao
长尾真

EBMT

English

How much is that **red umbrella**?

How much is that **small camera**?

Japanese

Ano **akai kasa** wa ikura desu ka.

Ano **chiisai kamera** wa ikura desu ka.

从上述双语语料中抽取的模板或实例:

1. **How much is that X ? \Leftrightarrow Ano X wa ikura desu ka**
2. **red umbrella \Leftrightarrow akai kasa**
3. **small camera \Leftrightarrow chiisai kamera**

EBMT

➤ 一个例子

Chinese examples	English examples
<他> _{1c} <把收音机> _{3c} <打开了> _{2c}	<He> _{1e} <turned on> _{2e} <the radio> _{3e}
<钱包> _{4c} <放> _{5c} <在桌上> _{6c}	<The wallet> _{4e} <is put> _{5e} <on the table> _{6e}

Input:

他把花放在桌上

他 把花 放 在桌上



$1c+3c+replace(2c,(5c+6c))$



Output:

He is put on the table flower →



**Translation
generation**

$1e+replace(2e,(5e+6e))+3e$

EBMT

□ 实例泛化(example generalization, 也称实例抽象)

- 将一类语言现象用统一的模式(pattern)或模板(template)来表示

Karl Marx was born in Trier, Germany on May 5, 1818.

卡尔·马克思于1818年5月5日出生在德国特里尔城。



[Person] was born in [City] on [Date]

[Person] 于 [Date] 出生在 [City]

中括号括起的是模板的槽，相当于变量

EBMT

□ 实例泛化

美国国务卿希拉里今天起开始访问日本。

U.S. Secretary of State Hillary Clinton begins a visit to Japan from today.

美国总统奥巴马明日起开始访问中国。

U.S. President Barack Obama begins a visit to China from tomorrow.



[国家名]+[职位]+[姓名]+[时间]+起+开始访问+[地名]

[location]+[position/titile]+[name]+begin a visit to+[location]+from+[time]

EBMT

□ EBMT vs. TM(translation memory)

- EBMT

他把花放在桌上 →

He puts the flower on the table

- TM

他把花放在桌上 →

He turned on the radio.

The wallet is put on the table

TM不能算是机器翻译技术

大纲

□ 历史

□ 方法

- RBMT
- EBMT
- SMT
- NMT

超声检查中，请稍后

Ultrasonic examination is in progressing, please wait

美国总统布什昨天在白宫与以色列总理沙龙就中东局势举行了一个小时的会谈。

US President George W. Bush head an hour-long meeting with Israeli Prime Minister Ariel Sharon on the situation in the Middle East yesterday at the White House.

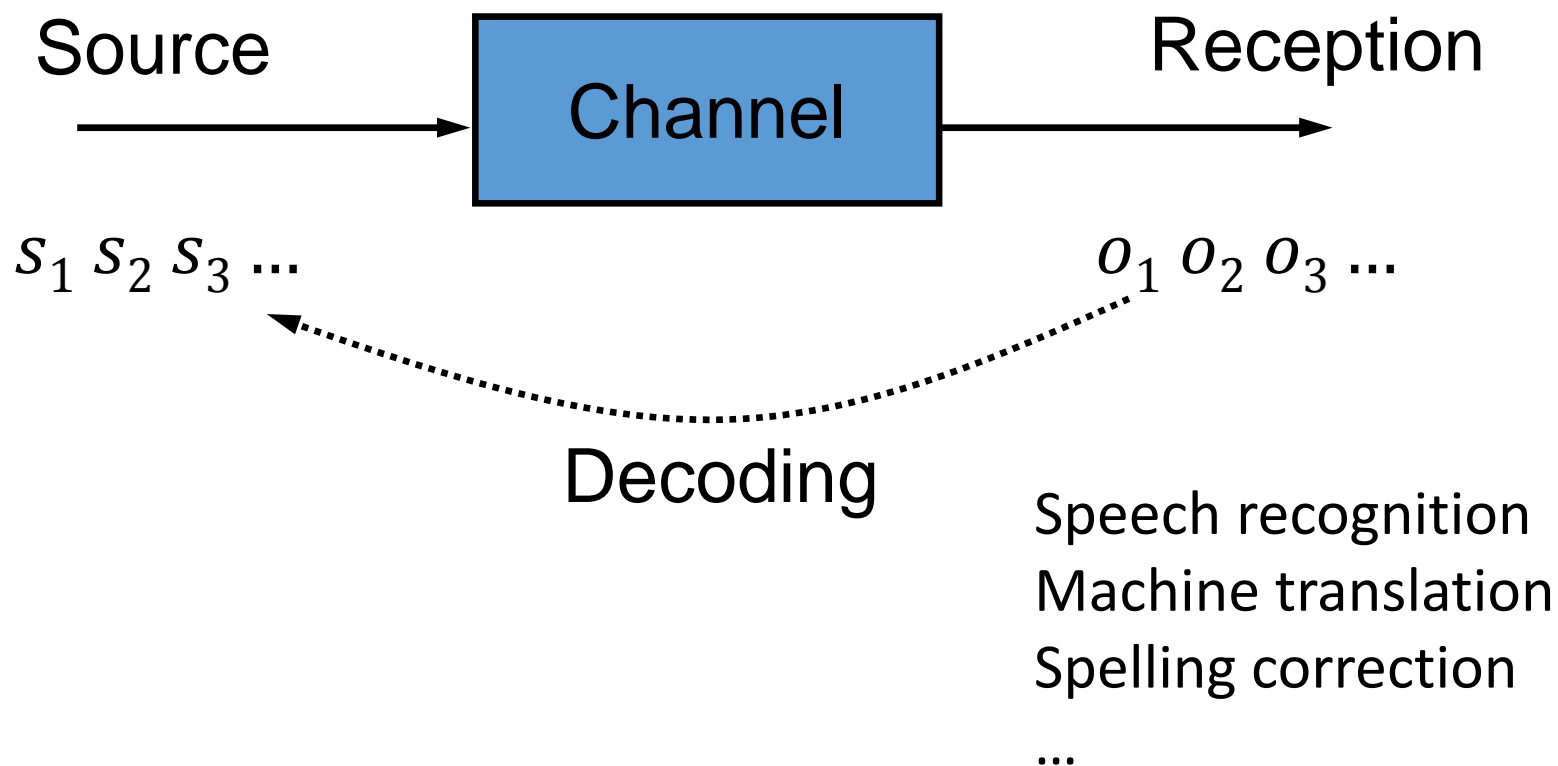
SMT

□ 统计机器翻译(statistical MT, SMT)

- Word based
- Phrase based
- Syntax based

Noisy channel model

噪声信道模型



SMT-word based

- 回顾Bayes法则及概率图模型

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$$S^* = \operatorname{argmax}_S P(S|O)$$

$$= \operatorname{argmax}_S P(O|S)P(S)$$

$$= \operatorname{argmax}_S P(o_1 o_2 o_3 \dots | s_1 s_2 s_3 \dots) P(s_1 s_2 s_3 \dots)$$

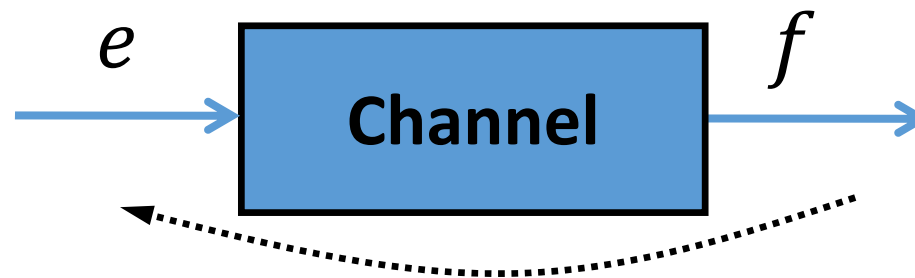
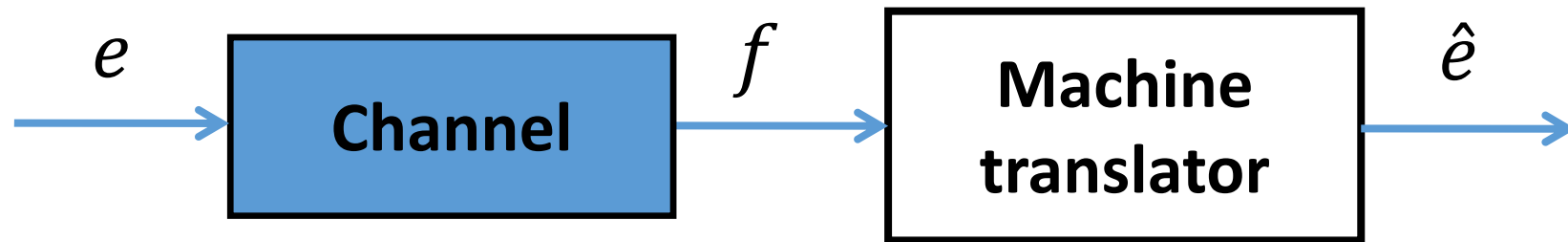
Acoustic Model
Translation model
Correction model
...

Language model

HMM

$$\approx \operatorname{argmax}_S P(o_1|s_1)P(o_2|s_2)P(o_3|s_3) \dots P(s_1)P(s_2|s_1)P(s_3|s_2) \dots$$

机器翻译可以像**HMM**这样建模吗？



Translation=Decoding

SMT-word based

$$e^* = \arg \max_e P(e|f) = \arg \max_e \frac{P(f|e)P(e)}{P(f)}$$
$$= \arg \max_e P(f|e)P(e)$$

翻译模型

语言模型

人工评价指标:

adequacy

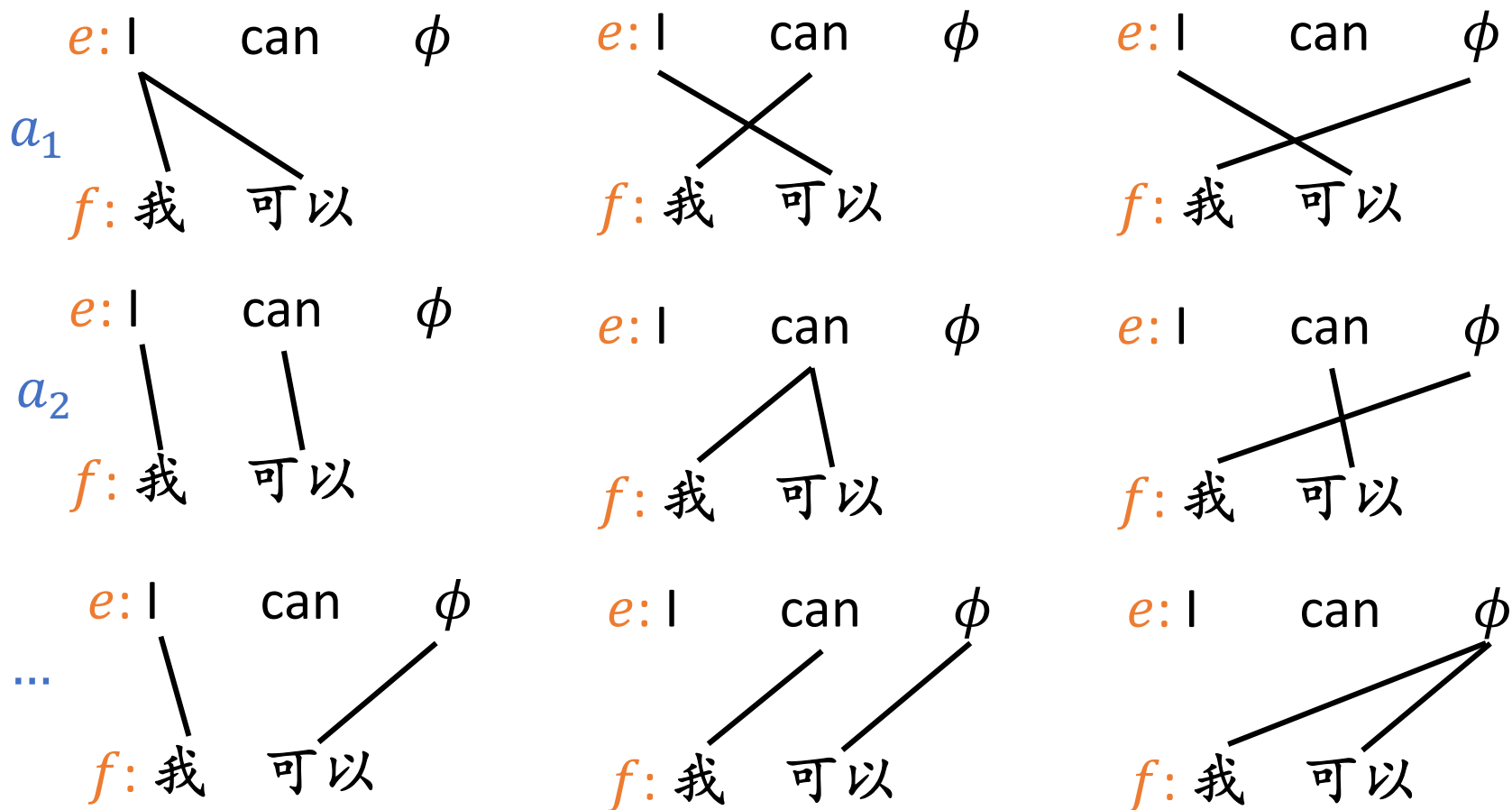
fluency

SMT-word based

□ IBM model1 (词汇翻译)

$$P(f|e) = \sum_a P(f, a|e)$$
$$P(f, a|e) = P(a|e) \underbrace{P(f|e, a)}_{\text{因为基于了对齐 } a, \text{ 翻译概率基于条件独立性假设写成条件概率因子的连乘}} = \frac{\varepsilon}{(l_e + 1)^{l_f}} \prod_{j=1}^{l_f} P(f_j | e_{a_j})$$

$P(a|e)$: 长度为 l_e 的句子 e 生成长度为 l_f 的句子 f , 对齐 a 的概率; IBM model1 假设每个 alignment 所发生的概率相同, 分母是多少种对齐方式, 因为引入了空标记所以 +1



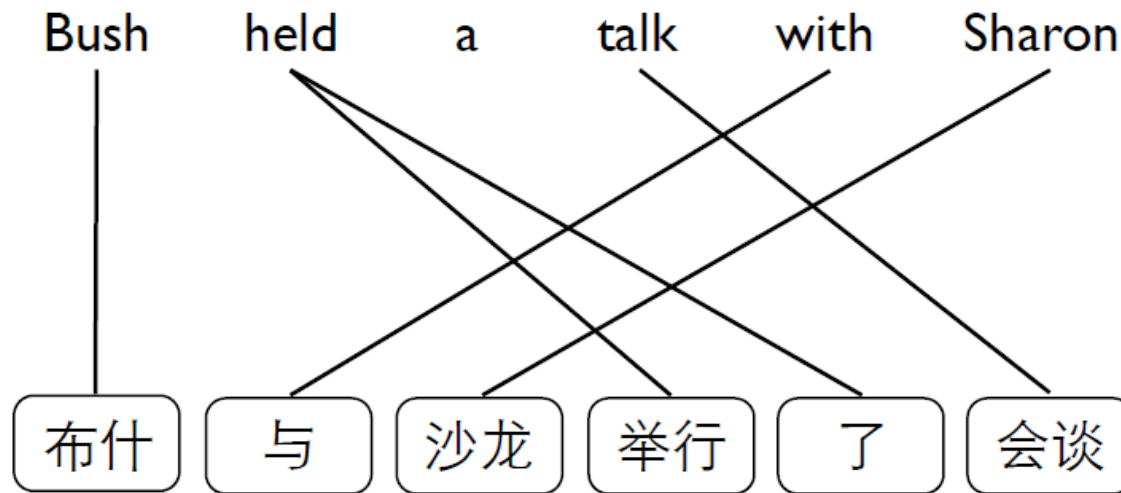
对 $I \text{ can} \rightarrow \text{我 可以}$, 一共 $(l_e + 1)^{l_f} = (2 + 1)^2 = 9$ 种对齐方式

$$P(\text{我可以}, a_1 | I \text{ can}) = \frac{\varepsilon}{9} P(\text{我} | I) P(\text{可以} | I)$$

$$P(\text{我可以}, a_2 | I \text{ can}) = \frac{\varepsilon}{9} P(\text{我} | I) P(\text{可以} | \text{can})$$

SMT-word based

□ IBM model1



$$P(f, a|e) = \frac{\varepsilon}{(6 + 1)^6} P(\text{布什}|\text{bush})P(\text{与}|\text{with}) \dots P(\text{会谈}|\text{talk})$$

如何learning?

SMT-word based

□ Learning

- MLE
- EM

□ 副产品: word alignment (词对齐)

GIZA, GIZA++

训练语料:

I can

He can

EM算法推导
IBM model1
参数

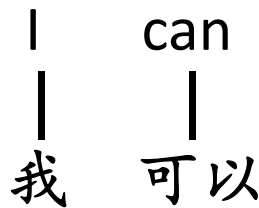
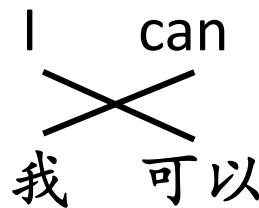
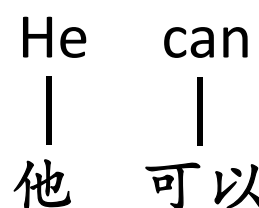
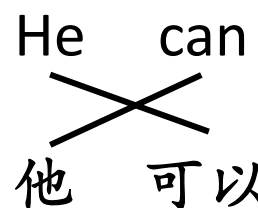
我 可以

他 可以

初始时词翻译
概率:

	我	他	可以
I	1/3	1/3	1/3
he	1/3	1/3	1/3
can	1/3	1/3	1/3

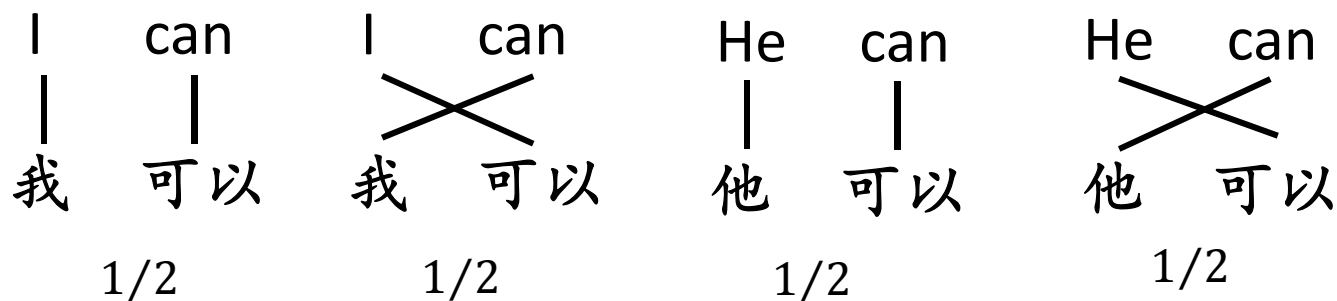
计算 $P(f, a|e)$

				这里简 化了对 齐形式
$1/3 \times 1/3 = 1/9$	$1/3 \times 1/3 = 1/9$	$1/3 \times 1/3 = 1/9$	$1/3 \times 1/3 = 1/9$	

归一化得

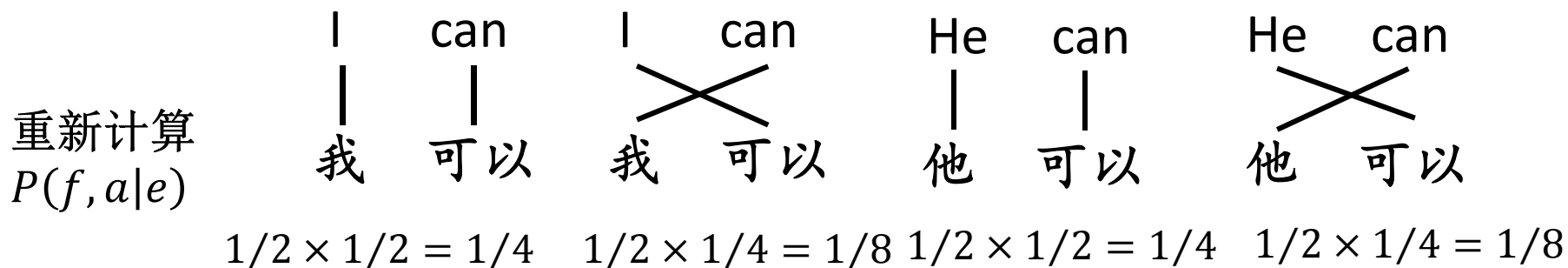
$\frac{1/9}{2/9} = 1/2$	$\frac{1/9}{2/9} = 1/2$	$\frac{1/9}{2/9} = 1/2$	$\frac{1/9}{2/9} = 1/2$
-------------------------	-------------------------	-------------------------	-------------------------

$$P(a|f, e) = \frac{P(f, a|e)}{\sum_a P(f, a|e)}$$



更新词翻译概率:

	我	他	可以
I	1/2	0	1/2
he	0	1/2	1/2
can	1/4	1/4	1/2



归一化得 $P(a|f, e)$

$$\frac{1/4}{3/8} = 2/3 \quad \frac{1/8}{3/8} = 1/3 \quad \frac{1/4}{3/8} = 2/3 \quad \frac{1/8}{3/8} = 1/3$$

重复上述E和M过程，直到词翻译概率不再发生变化

SMT

□ 统计机器翻译(statistical MT, SMT)

- Word based
- Phrase based
- Syntax based

SMT-phrase based

$$e^* = \arg \max_e P(e|f) = \arg \max_e \frac{P(f|e)P(e)}{P(f)}$$

基于词的模型： $= \arg \max_e P(f|e)P(e)$

没法使用上下文，有时短语的意义并不是词的意义加起来

□ 基于短语的模型

● Modeling

$$P(f_1^I | e_1^I) = \prod_{i=1}^I \phi(f_i | e_i) \underline{d(\text{start}_i - \text{end}_{i-1} - 1)}$$

翻译成第*i*个目标语言短语
的源语言短语的起始位置

思考这里为什么都是*I*长，而且没有对齐模型

SMT-phrase based

- ✓ 相比基于词的翻译模型，这里多学到了一个短语翻译表
 - 德语的*natuerlich*的phrase translation table

Translation	Probability $p(e f)$
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

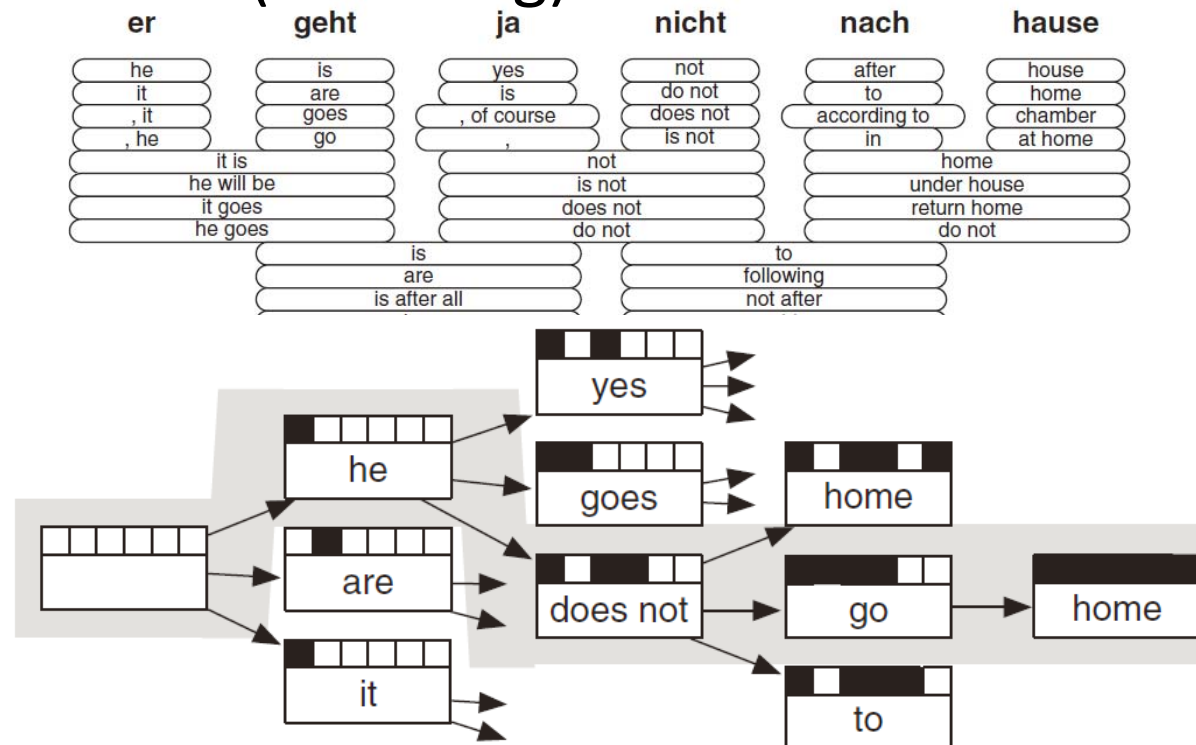
不是语言学意义上的短语；统计出来的反而效果更好。

SMT-phrase based

- Learning: 即估计短语翻译的概率

$$\phi(f|e) = \frac{\text{count}(e, f)}{\sum_{f_i} \text{count}(e, f_i)}$$

- Inference (decoding): Beam search



SMT-phrase based

▣ ME based SMT

- 不再基于噪声信道，直接对后验概率建模

IBM噪声信道模型：

$$\begin{aligned} e^* &= \arg \max_e P(e|f) = \arg \max_e \frac{P(f|e)P(e)}{P(f)} \\ &= \arg \max_e P(f|e)P(e) \end{aligned}$$

最大熵对数线性SMT模型：

$$e^* = \arg \max_e P(e|f) = \arg \max_e \sum_{m=1}^M \lambda_m h_m(e, f)$$

这是对后世影响最大的应用最广泛的**SMT**模型：
对数线性模型+**beam search**解码+最小错误率训练

SMT

□ 机器翻译的评价指标BLEU (BiLingual Evaluation Understudy)

$$\text{BLEU} = BP \cdot \exp \left(\sum_{n=1}^N w_n \log P_n \right)$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

句长惩罚
 c : candidate长度
 r : reference长度

P_n : Modified n -gram precision

2002, BLEU: a Method for Automatic Evaluation of Machine Translation

SMT

Candidate: It is a nice day today↵

1-gram匹配5/6

Reference: Today is a nice day↵

Candidate: It is a nice day today↵

3-gram匹配2/4

Reference: Today is a nice day↵

SMT

□ GIZA , Pharaoh , Moses , 丝路 (骆驼、商队、绿洲), ELMo, BERT, ERNIE



ROUGE: 2004, Recall-oriented understudy for gisting evaluation

SMT

□ 通常，BLEU打分机制和常见的解码优化目标不一致

——模型优化目标和机器翻译评价的目标不一致

- 最小错误率训练MERT
- 最小风险训练MRT
- 强化学习
 - Actor-critic

SMT

□ 统计机器翻译(statistical MT, SMT)

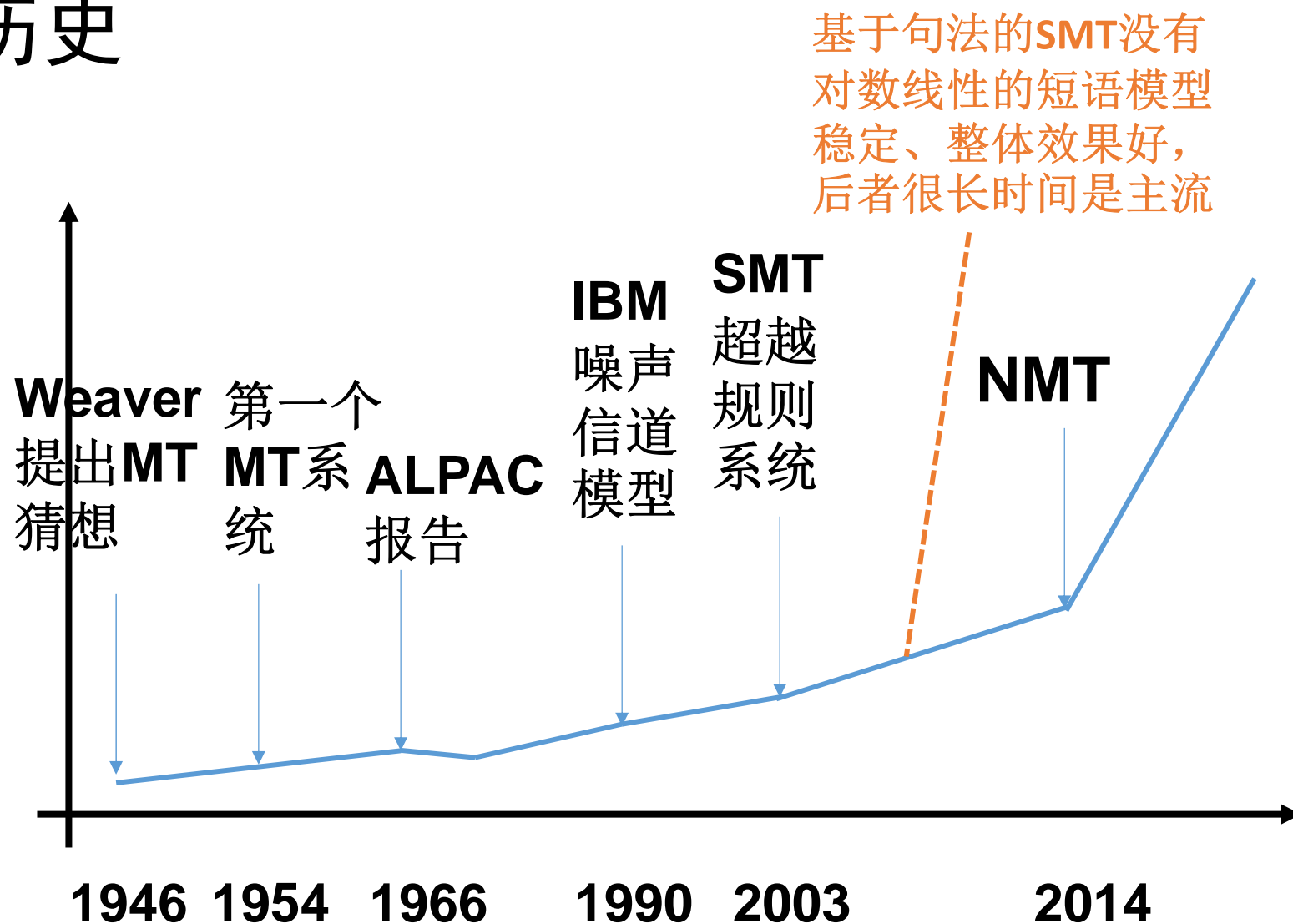
- Word based
- Phrase based
- Syntax based

SMT-syntax based

□ 出发点：没有融合句法信息，生成的目标句子可能不合语法

- 形式句法方法(formally syntax based)：采用的短语串只是形式上的，不符合语言学约束
Hierarchical phrase-based model (2005)
- 语言学句法方法(linguistically syntax based)
 - Tree to string
 - String to tree
 - Tree to tree

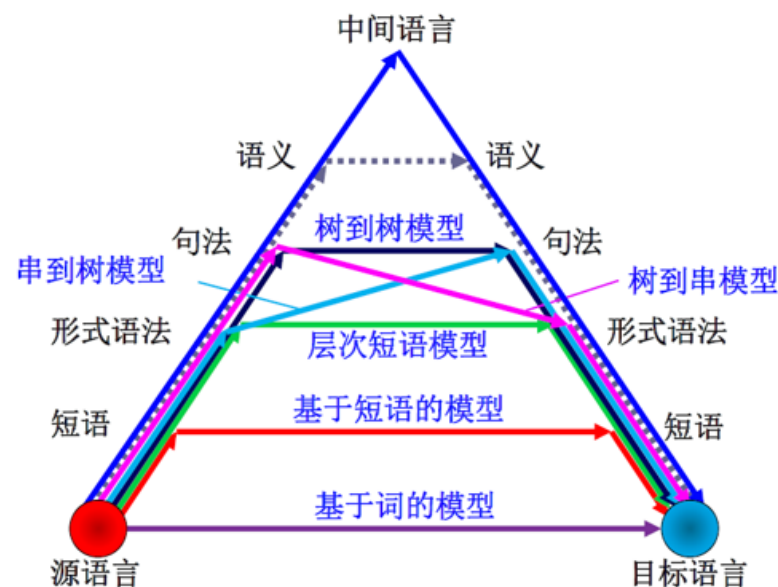
历史

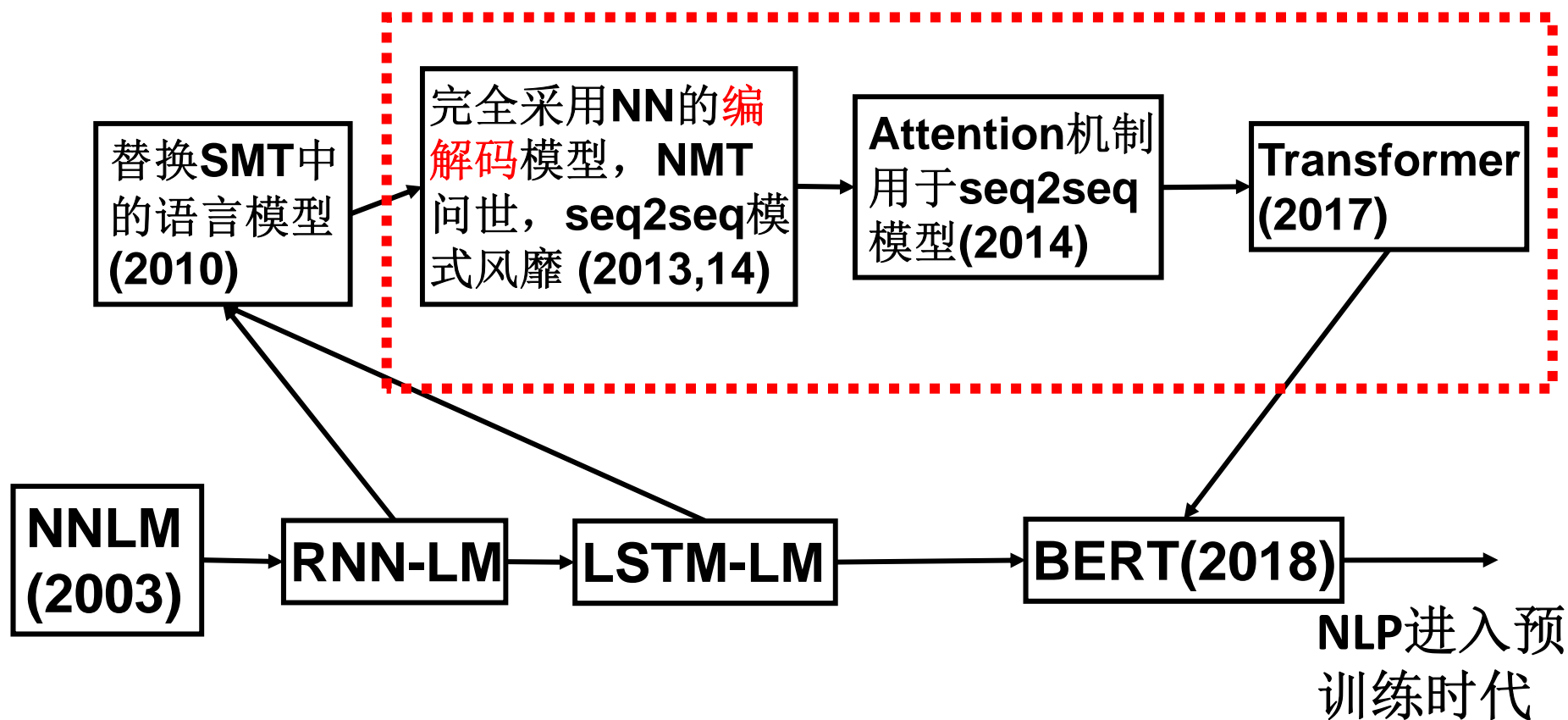


SMT

□ SMT到达瓶颈后

- 深度上
 - 基于语义的翻译
 - 篇章级翻译
 - DNN引入SMT
- 广度上
 - 半监督、无监督的MT
 - 互联网环境下的MT





大纲

□ 历史

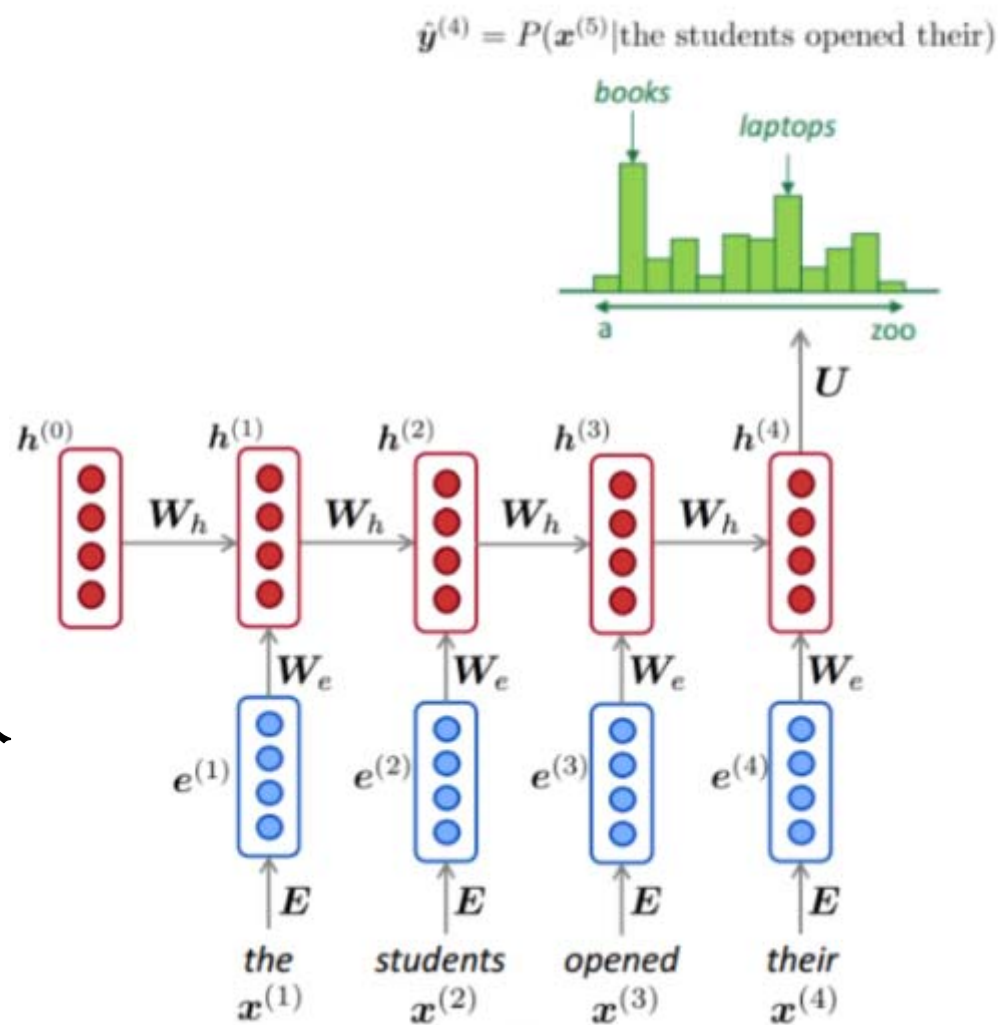
□ 方法

- RBMT
- EBMT
- SMT
- NMT

NMT

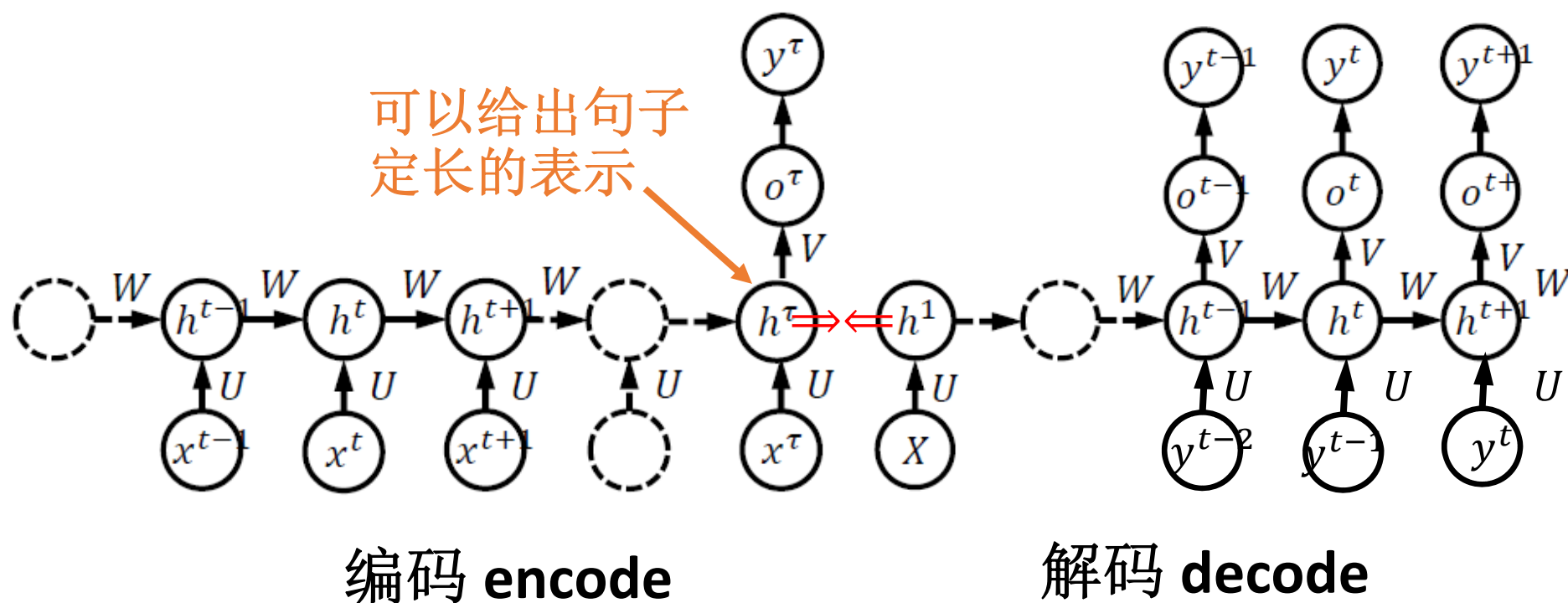
- ▣ RNN-NMT
- ▣ Attention RNN-NMT
- ▣ Transformer

RNNLM，根据输入
预测下一个词：



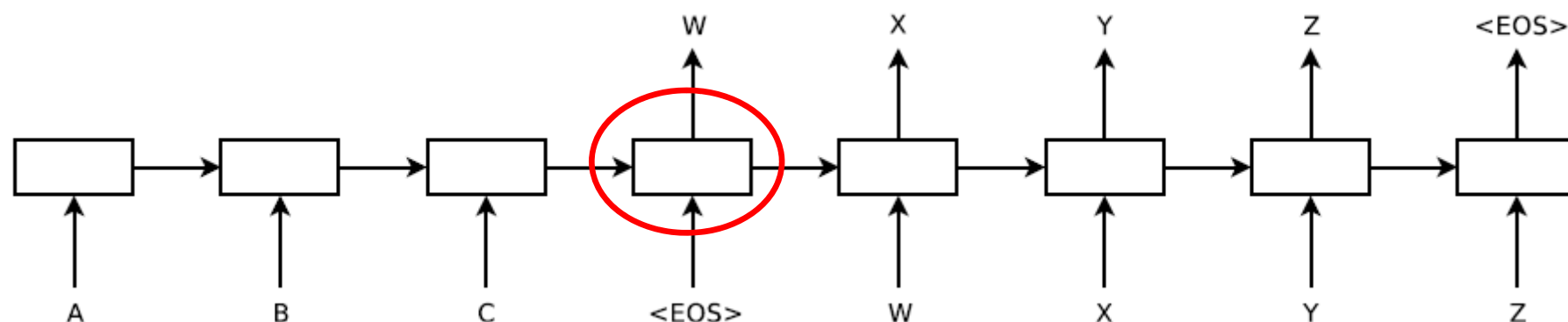
RNN-NMT

□ 回顾 $n \rightarrow 1$ 的RNN和 $1 \rightarrow n$ 的RNN



RNN-NMT

- 用这个定长的encoding出来的向量表示，来逐词生成目标语言句子



编码 **encode**

解码 **decode**

- 基于RNN的NMT，一个seq2seq的编解码模型 (encoder-decoder)

最终成为一种框架，主要为**NLP**的序列转换问题建模

RNN-NMT

$$c = h_T = \tanh(b + Wh_{T-1} + Ux_T)$$

$$s_t = f(s_{t-1}, y_{t-1}, c)$$

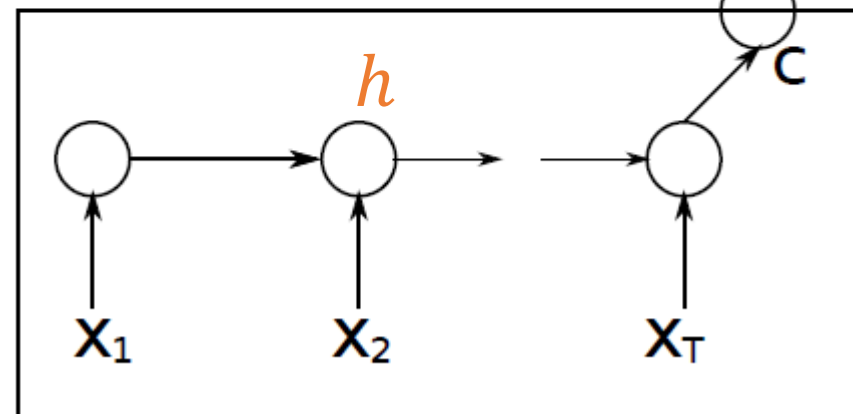
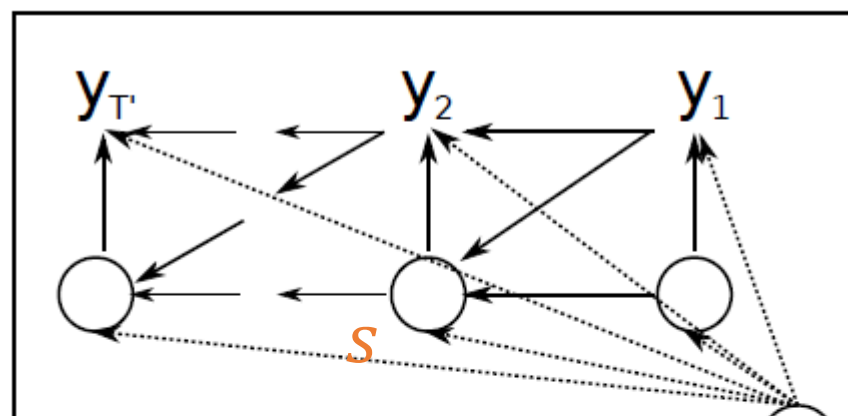
$$P(y_t | y_{t-1}, y_{t-2}, \dots, y_1, c) \\ = g(s_t, y_{t-1}, c)$$

g 通常是softmax

优化目标:

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^N \log P_{\theta}(y_n | x_n)$$

Decoder



Encoder

RNN-NMT

□ 存在的问题

- 一个实数向量无法表示源语言句子的完整语义

□ 缓解

- 解码时引入Attention机制——解码器在解码(生成目标语言词语)时更多关注输入中的相关部分

NMT

- ▣ RNN-NMT
- ▣ Attention RNN-NMT
- ▣ Transformer

Attention RNN-NMT

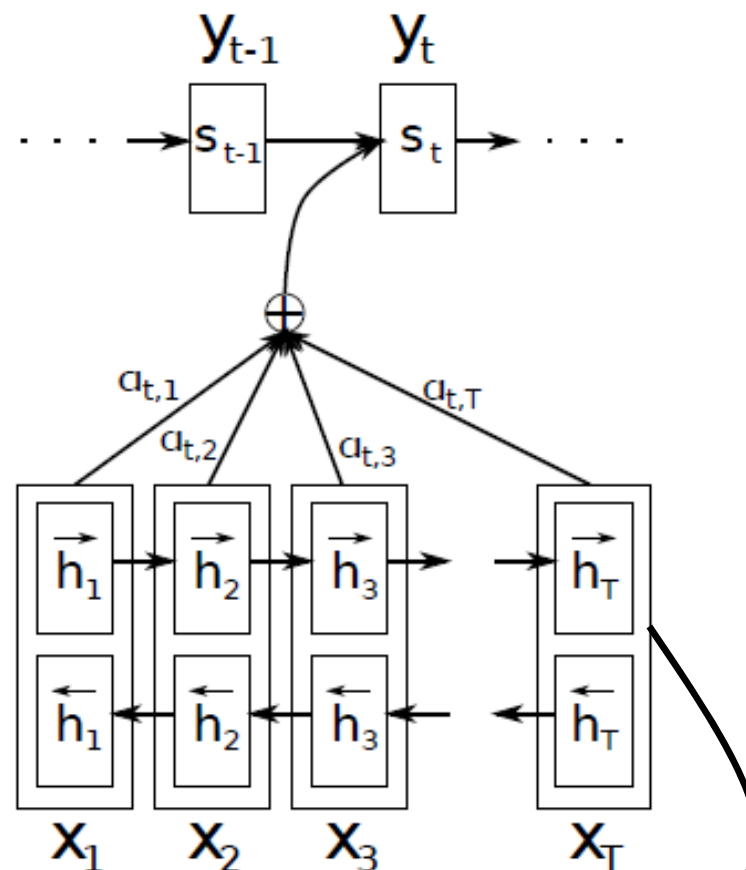
$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

$$c_t = \sum_{j=1}^T \alpha_{tj} h_j$$

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^T \exp(e_{tk})}$$

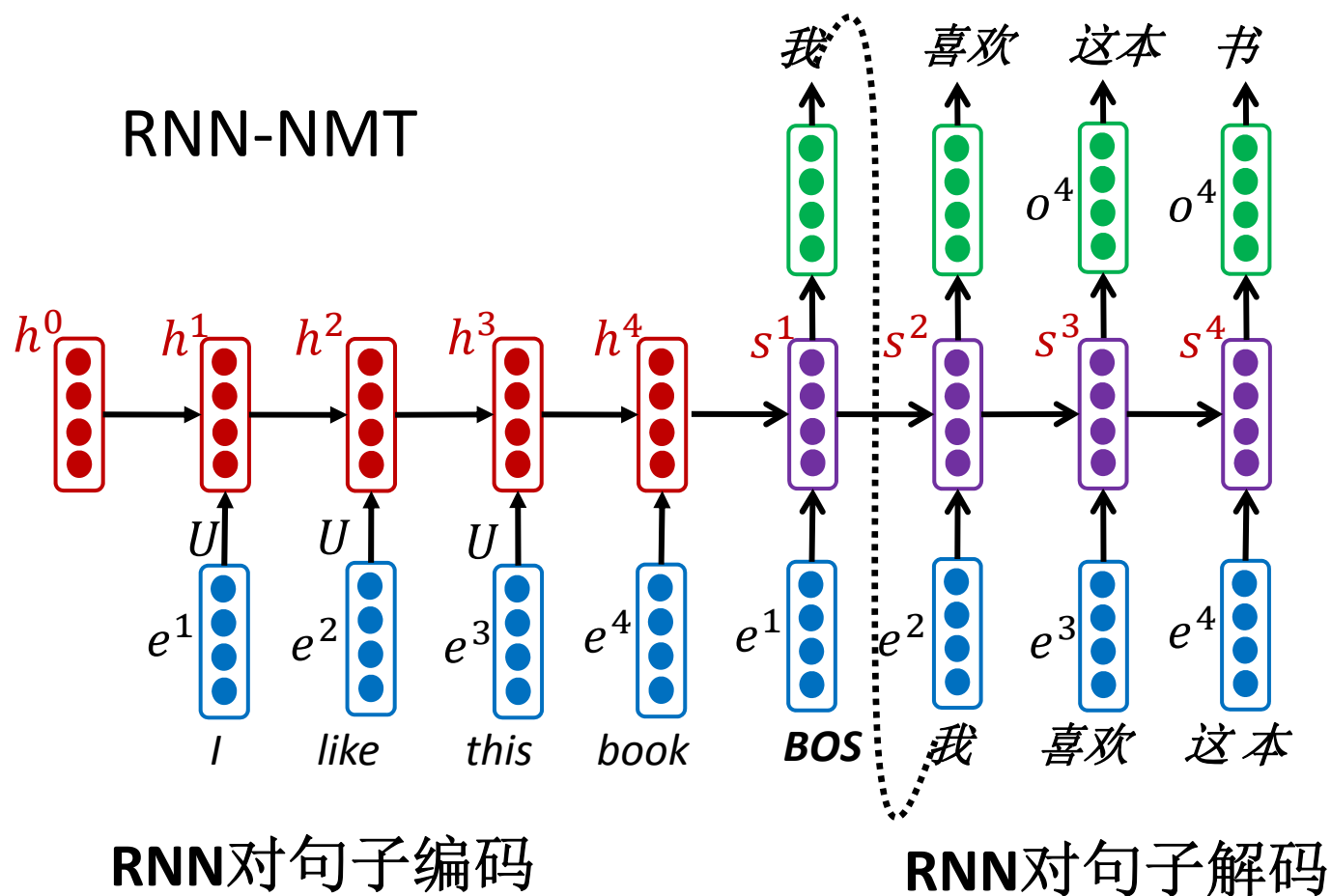
$$e_{tj} = a(s_{t-1}, h_j)$$

对齐网络 a 是一个前馈神经网络 $v_a^T \tanh(W_a s_{t-1} + U_a h_j)$ 这里已经使用的双向LSTM了 $h_t = [\vec{h}_t^T; \overleftarrow{h}_t^T]^T$



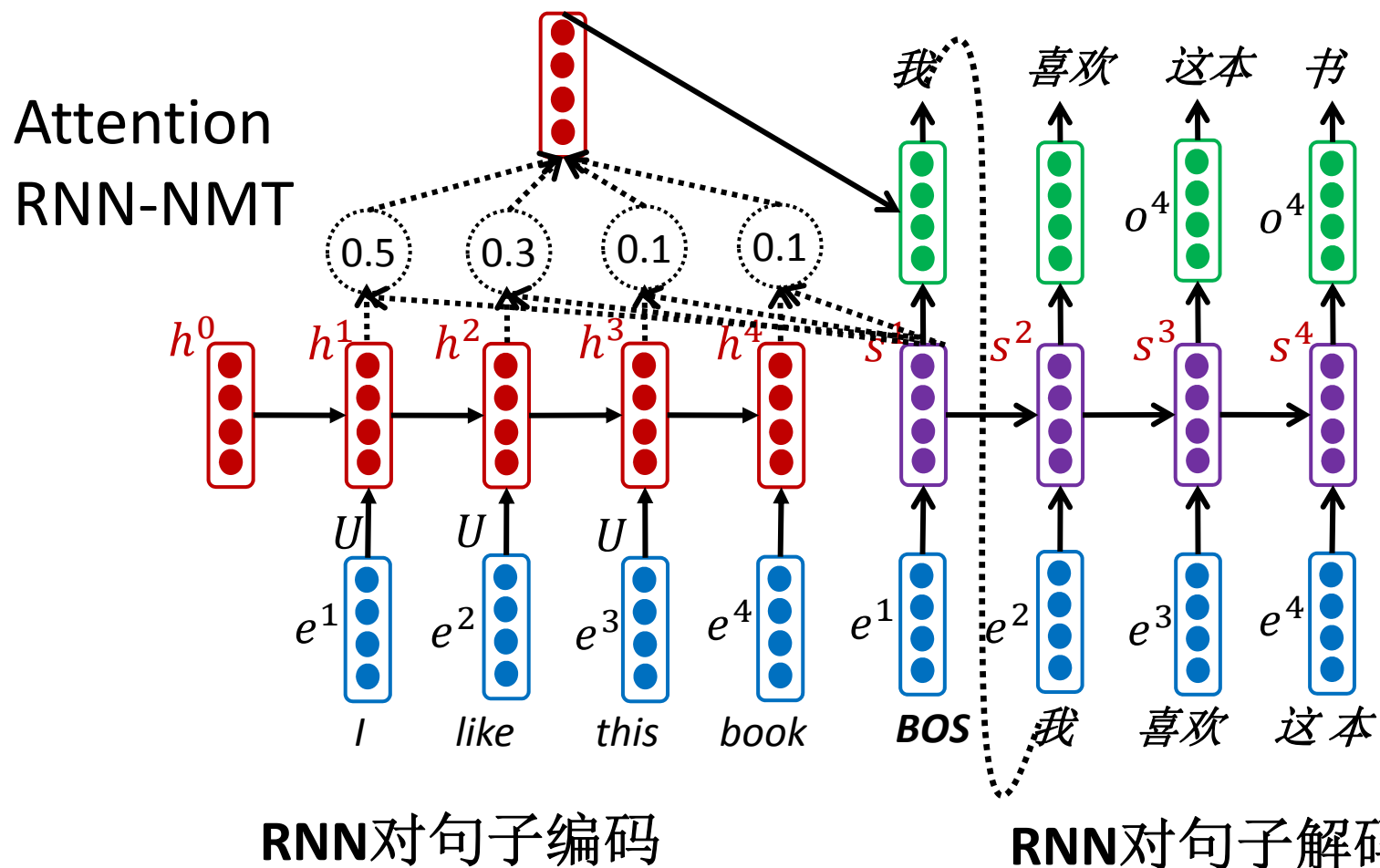
解码器重点注意了与生成目标语言词最相关编码的源语言向量的权重分布

比较RNN-NMT和Attention RNN-NMT



这个模型加入了解码端对编码端的**Attention**以后，才有较好的效果

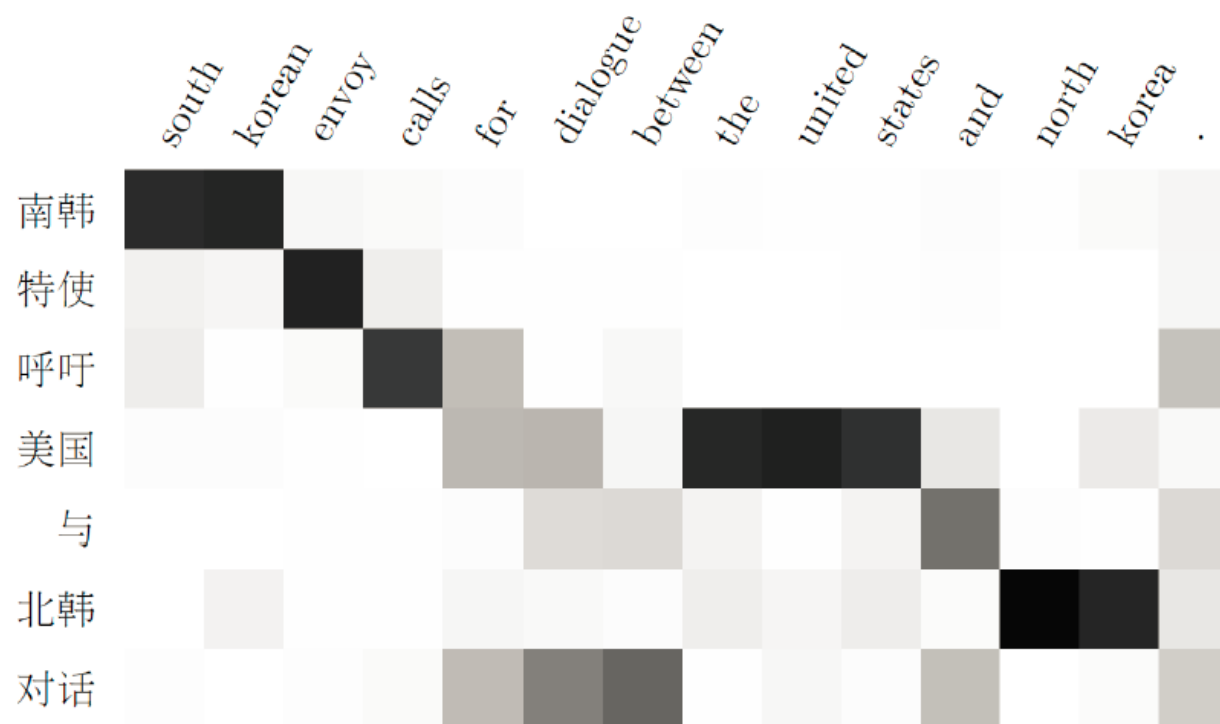
解码端对编码端的**Attention**——当前解码的隐状态对输入的每个位置有多亲密：



Seq2seq+Attention(典型实现: RNNSearch)

Attention RNN-NMT

□ Attention的可视化



NMT

- ▣ RNN-NMT
- ▣ Attention RNN-NMT
- ▣ Transformer

Transformer

Attention Is All You Need

Google 2017

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

- 不再用RNN等序列结构
- 前向全连接网络 + 多头注意力机制
- 目的是
 - 并行能力
 - 长期依赖

Transformer

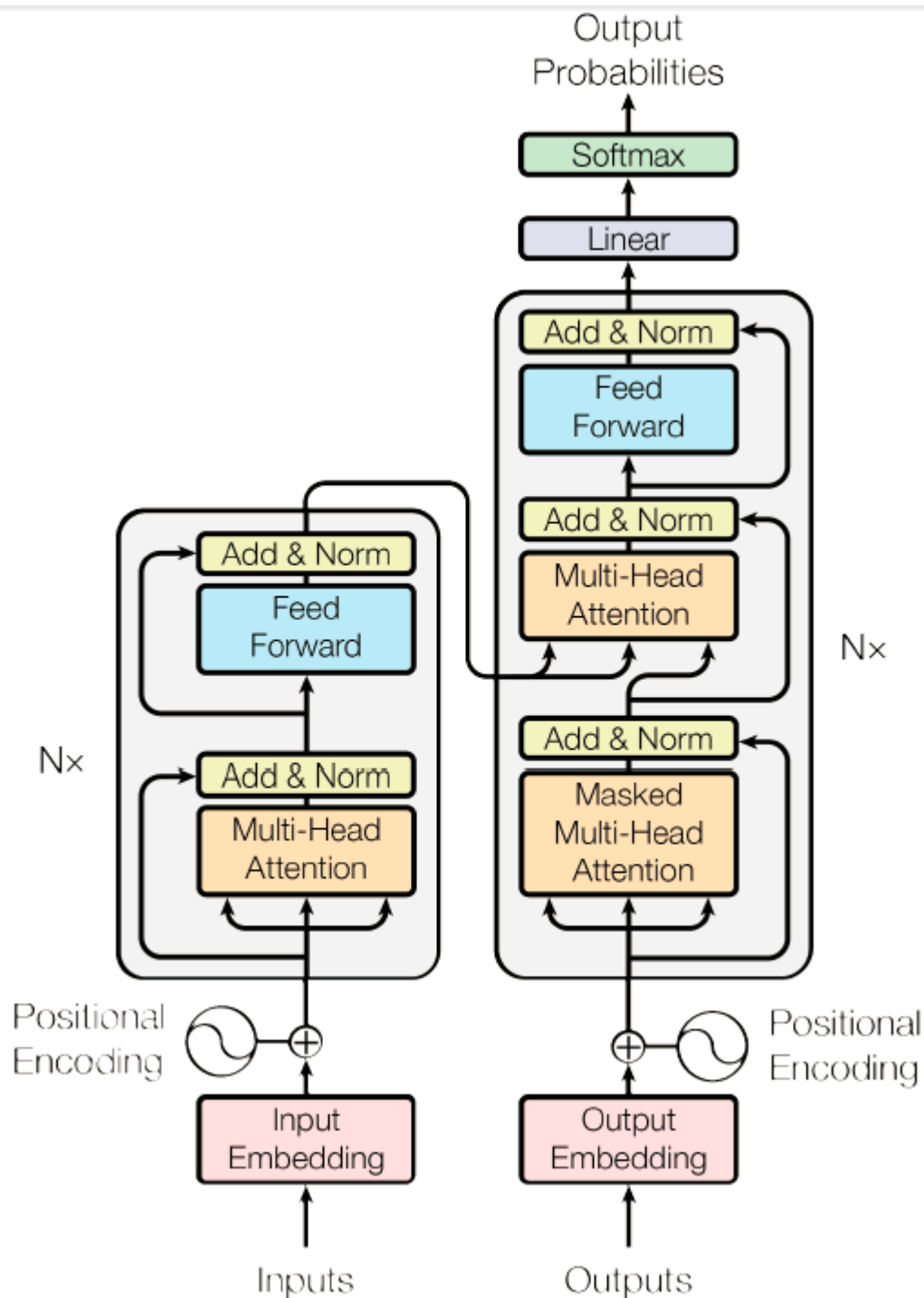
Attention Is All You Need

Google

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer
Google Brain
noam@google.com

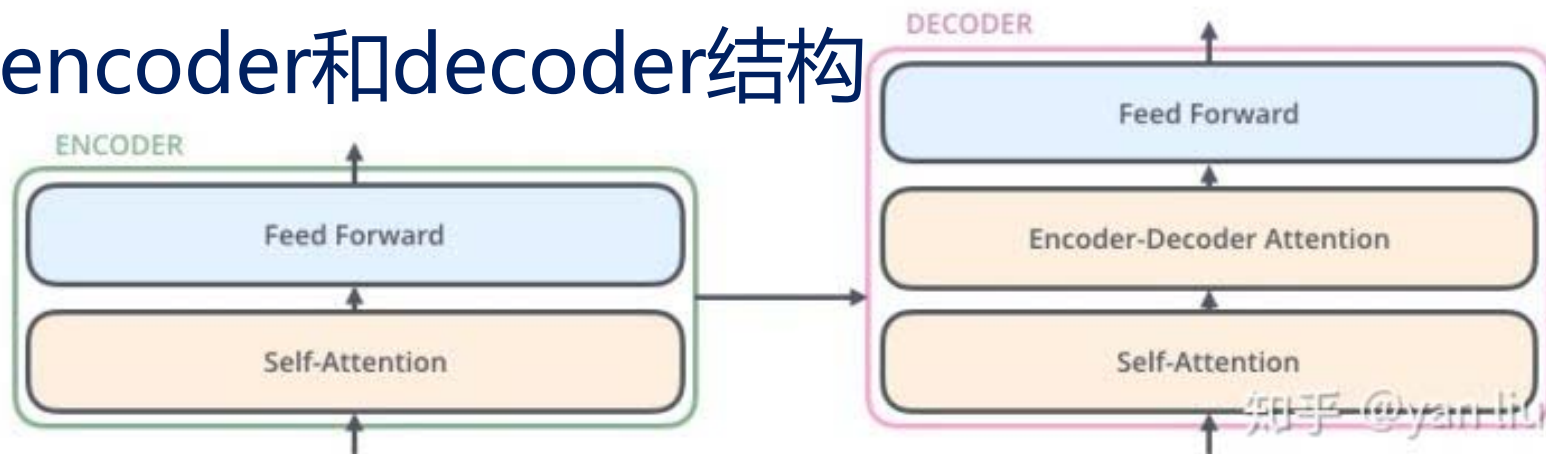
- 不再用RNN等序列模型
- 前向全连接网络 + 多头注意力机制
- 目的是
 - 并行能力
 - 长期依赖



Transformer



□ encoder和decoder结构

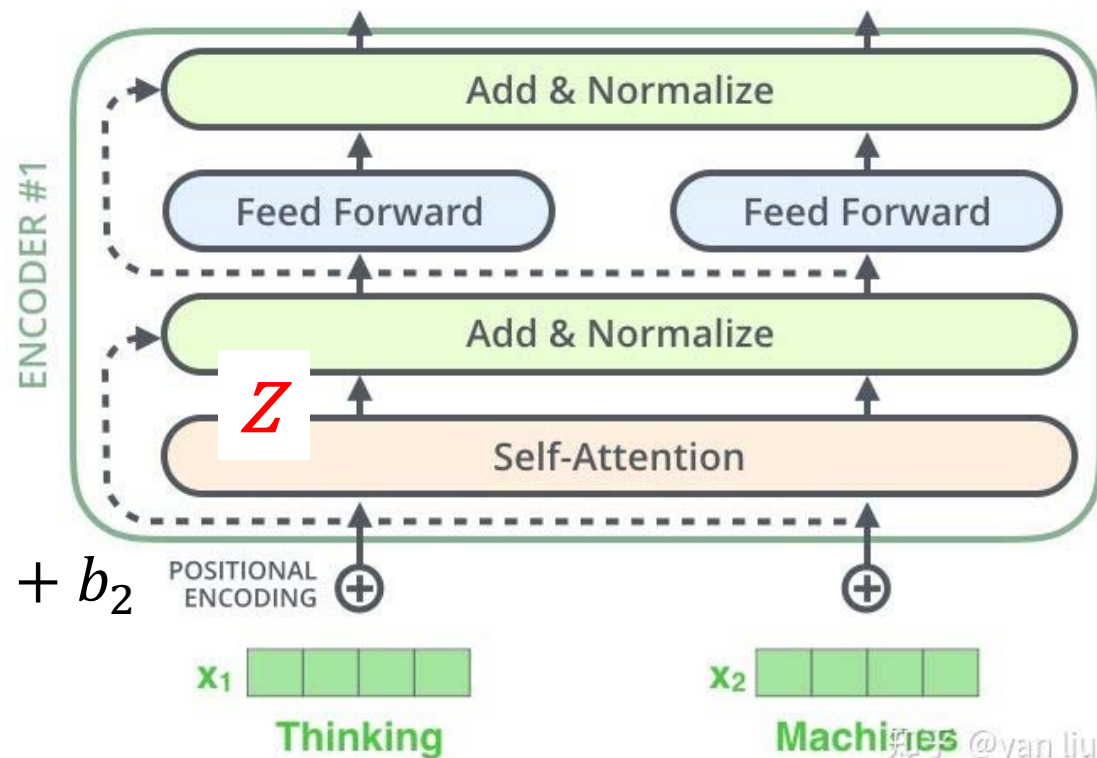


$$Z = \text{Att}(Q, K, V)$$

$$= \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

$$\text{FFN}(Z)$$

$$= \max(0, ZW_1 + b_1)W_2 + b_2$$



Transformer

□ Self-attention

- 每个输入的token都有“学”出来的三个向量，每个输入句子 x 有学出来的三个矩阵：

- **Query:** $Q = xW^Q$

- **Key:** $K = xW^K$

- **Value:** $V = xW^V$

- Self-attention

$$Z = \text{softmax} \left(\frac{QK^T}{\sqrt{d_K}} \right) V$$

$$Z_i = \sum_j \left(\frac{\exp \left(\frac{q_i^T k_j}{\sqrt{d_K}} \right)}{\sum_k \exp \left(\frac{q_i^T k_k}{\sqrt{d_K}} \right)} v_j \right)$$

为了梯度的稳定，使用了score的归一化， d_K 是 K 的维度

Transformer

□ Self-attention

- 每个输入的token都有 “权重”，
每个输入句子 x 有学出来的

- **Query:** $Q = xW^Q$

- **Key:** $K = xW^K$

- **Value:** $V = xW^V$

- Self-attention

$$Z = \text{softmax} \left(\frac{QK^T}{\sqrt{d_K}} \right) V$$

$$\begin{aligned} c_t &= \sum_{j=1}^T \alpha_{tj} h_j \\ \alpha_{tj} &= \frac{\exp(e_{tj})}{\sum_{k=1}^T \exp(e_{tk})} \\ e_{tj} &= a(s_{t-1}, h_j) \end{aligned}$$

比较一下传统的attention

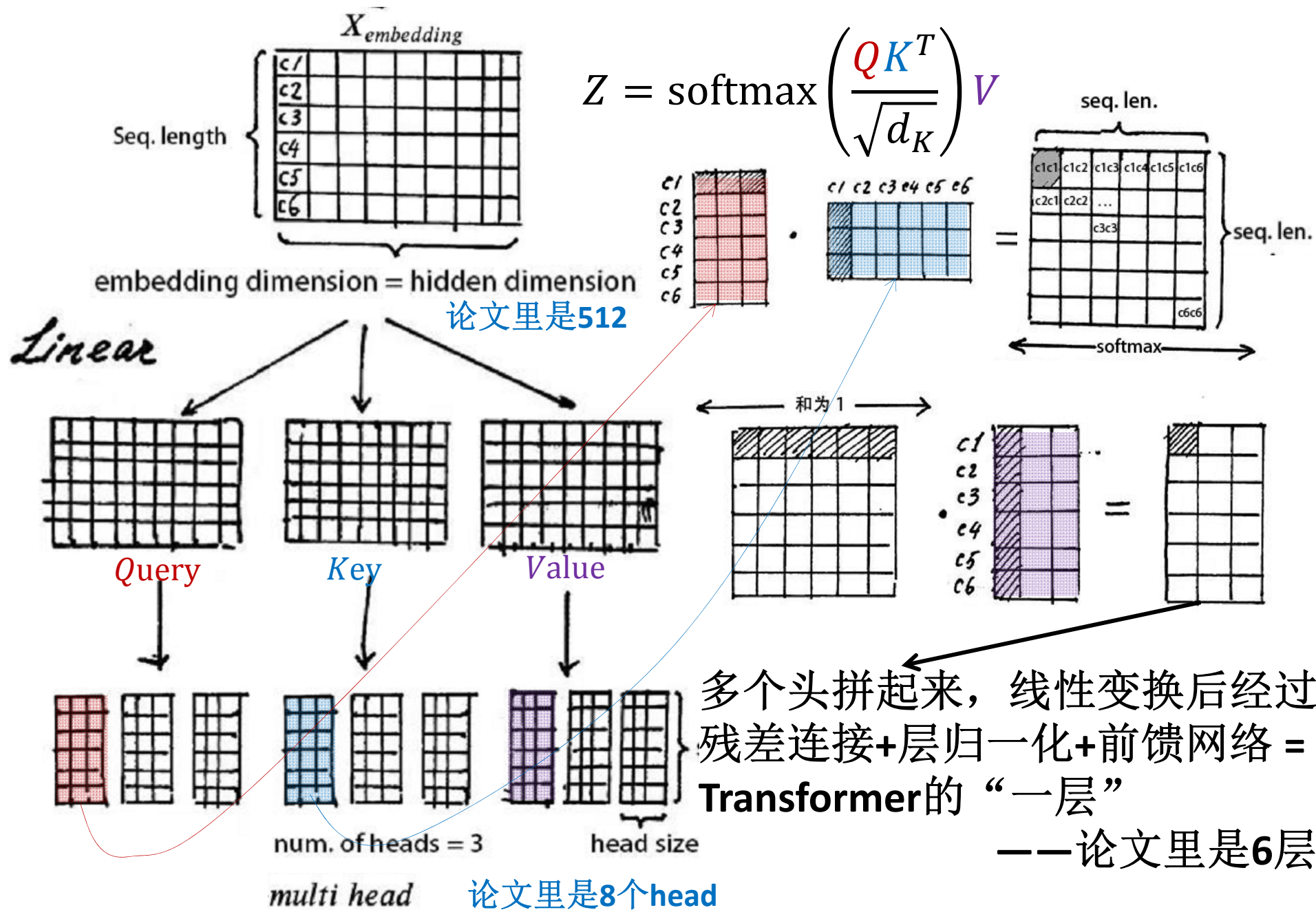
$$Z_i = \sum_j \left(\frac{\exp \left(\frac{q_i^T k_j}{\sqrt{d_K}} \right)}{\sum_k \exp \left(\frac{q_i^T k_k}{\sqrt{d_K}} \right)} v_j \right)$$

为了梯度的稳定，使用了score的归一化， d_K 是 K 的维度

Transformer

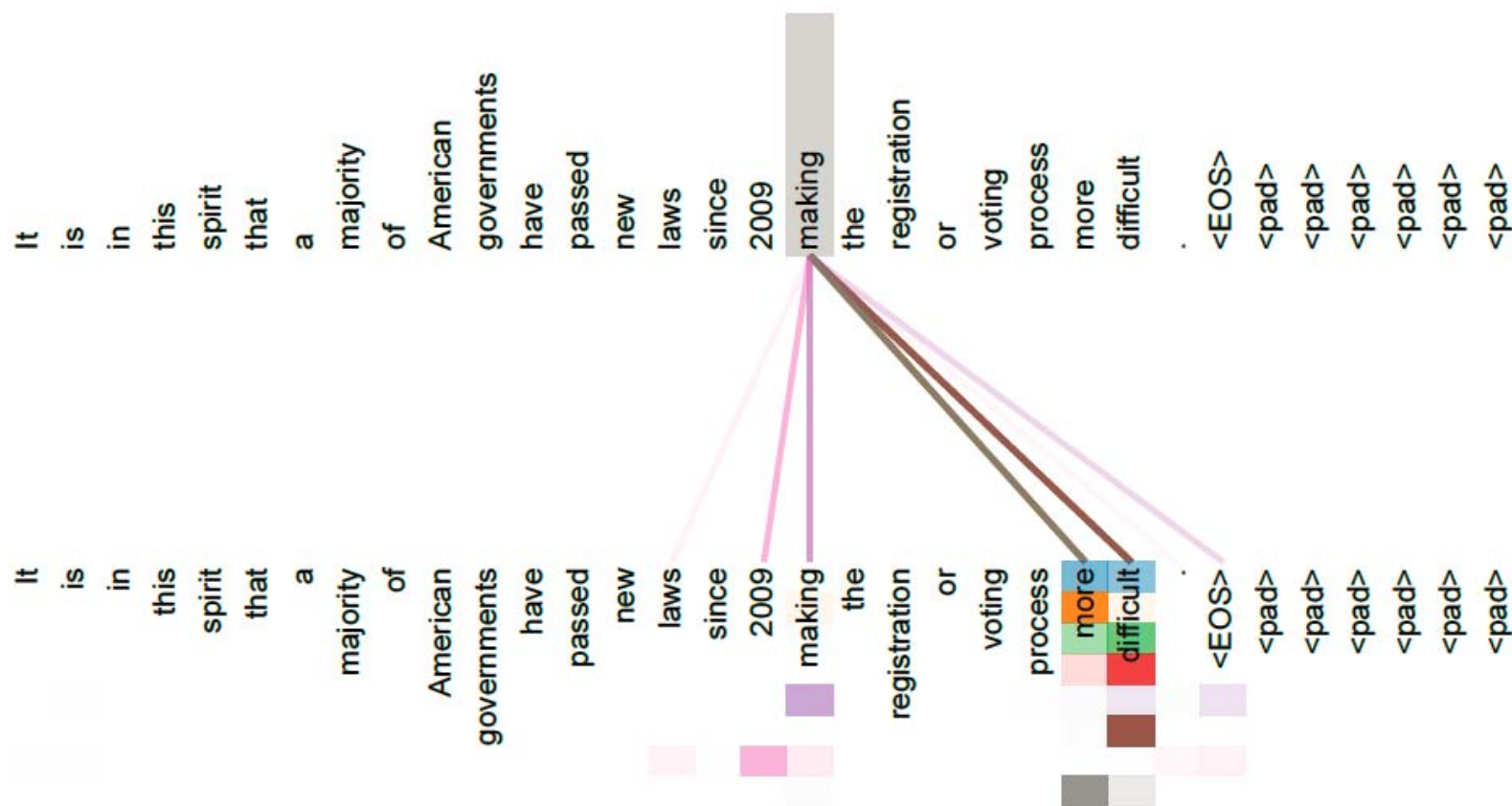
□ Multi-head

- 相当于 h 个不同的self-attention的集成，以 $h = 8$ 为例：
 - 将 x 分别输入到8个**self-attention**中，得到8个加权后的特征矩阵 Z_i
 - 将8个 Z_i 按列拼成一个大的特征矩阵
 - 特征矩阵经过一层全连接后得到输出 Z

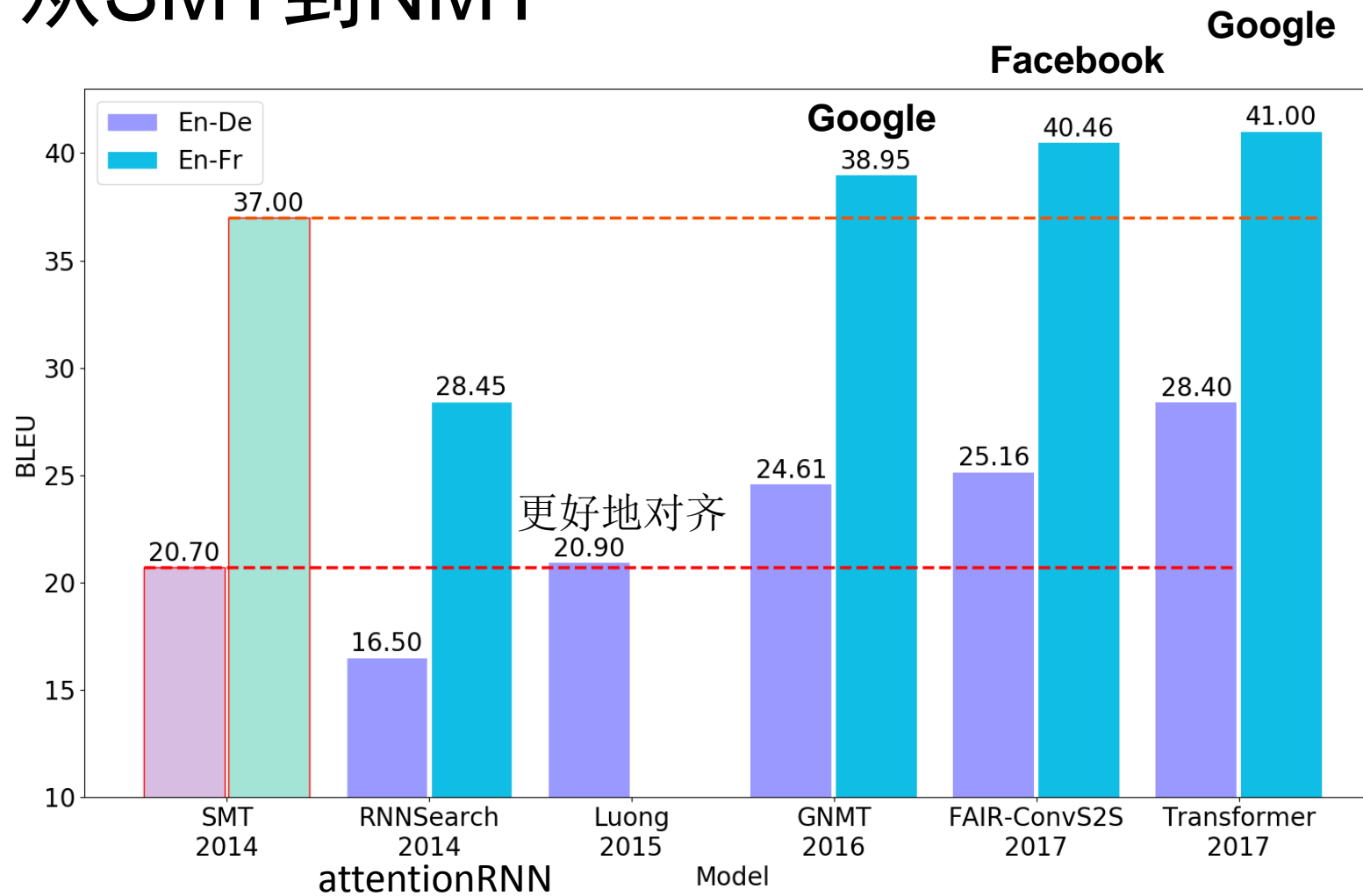


Transformer

□ Attention可视化——长距离依赖



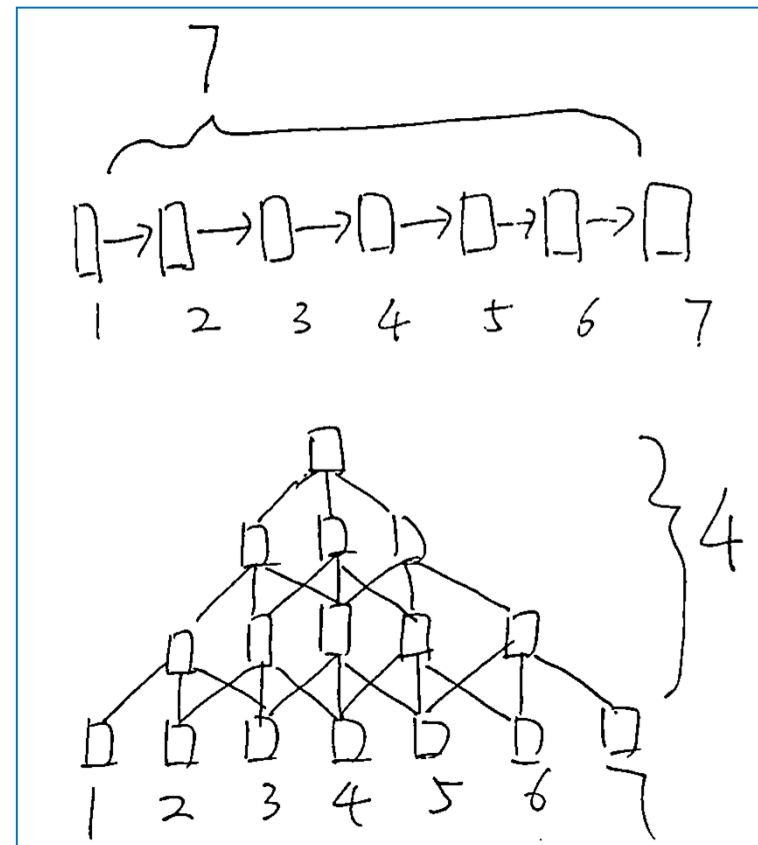
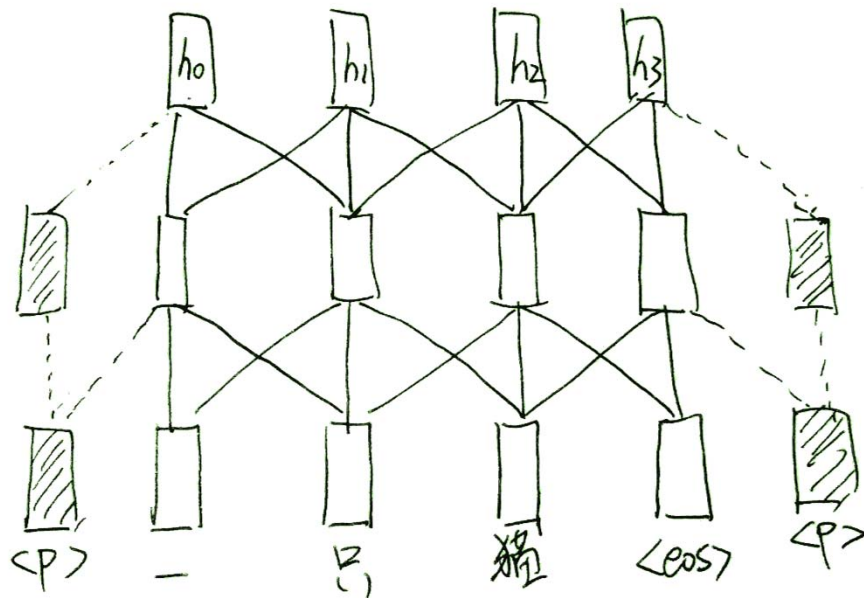
从SMT到NMT



NMT

▣ Convolutional Seq2Seq (Gehring et al., 2017)

- Not so deep
- Parallelization



NMT

□ 一些NMT翻译的例子

Source : you 're beginning to sound like a bloody parrot . go to bed and get some rest .

SMT : 你开始听起来像个该死的鹦鹉。上床去休息一下。

NMT : 你说话开始像只该死的鹦鹉了。去睡觉休息一下。

Google

翻译

英语 中文 德语 检测语言



中文(简体) 英语 日语

翻译

他要不要吃饭
我要不要吃饭
小王要不要吃饭
小王要吃饭
小王不想吃饭
小王想吃饭
小王想不吃饭




47/5000

Does he want to eat?
I want to eat
Xiao Wang wants to eat
Xiao Wang wants to eat
Xiao Wang does not want to eat
Xiao Wang wants to eat
Xiao Wang wants to not eat



NMT



翻译

关闭即时翻

英语 中文 德语 检测语言

↔ 中文(简体) 英语 日语


翻译

这家酒店的性价比

8/5000

The price/performance ratio of this hotel

提出修改建议



翻译

关闭即时翻译

英语 中文 德语 检测语言

↔ 中文(简体) 英语 日语

翻译

这家酒店的性价比相当高

11/5000

The price of this hotel is quite high

提出修改建议

NMT

with the release and spread 随着释放和扩散

with the release and spread of 随着

这件家具才是我想要的 This furniture is what I want

这件衣服才是我想要的 This is what I want

抱歉之前扭曲了文章的用意

Sorry for distorting the meaning of the article

抱歉之前歪曲了文章的用意

I'm sorry to have distorted the meaning of the article

Your school is pku.

你们学校在北大。

✓ 微信翻译



Your school is cumt.

你的学校很糟糕。

✓ 微信翻译



Your school is bit.

你的学校被咬了。

✓ 微信翻译

Your school is thu.

你的学校太烂了。

✓ 微信翻译



NMT

□ Challenges

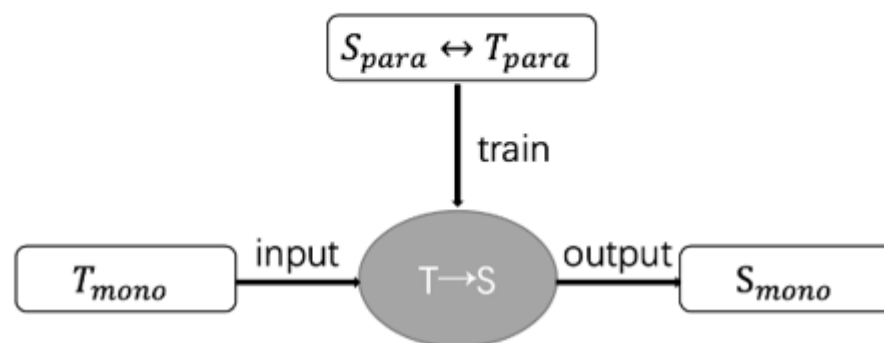
- Vocabularies
 - 主要是UNK (unknown words)问题——使用subwords
- Knowledge
 - 多源(multi-source), 多模态(multi-modal), 篇章上下文(document level NMT)
- Low resources 研究重点
 - Semisupervised, unsupervised, domain adaptation
- Robustness 研究重点
 - 使用数据增强

NMT

▣ Low resources

ACL2016. Improving neural machine translation models with monolingual data. Step1:

最简单有效的方法：
back-translation



Step2:



NMT

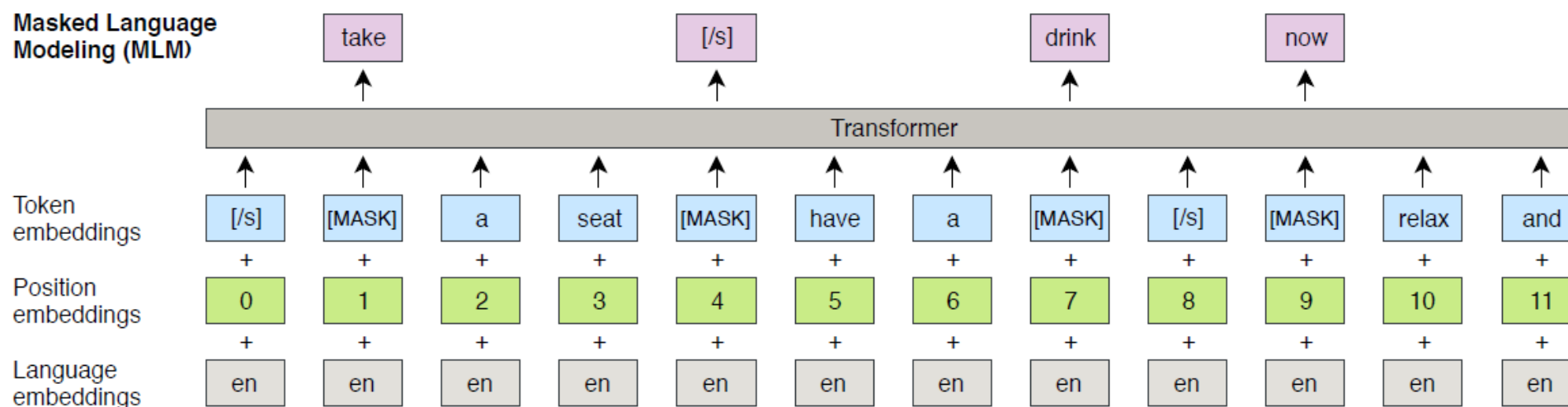
□ 多语言预训练模型

- XLM(Cross-lingual Language Model Pretraining)
 - CLM (causal)
 - MLM (masked)
 - TLM (translation)

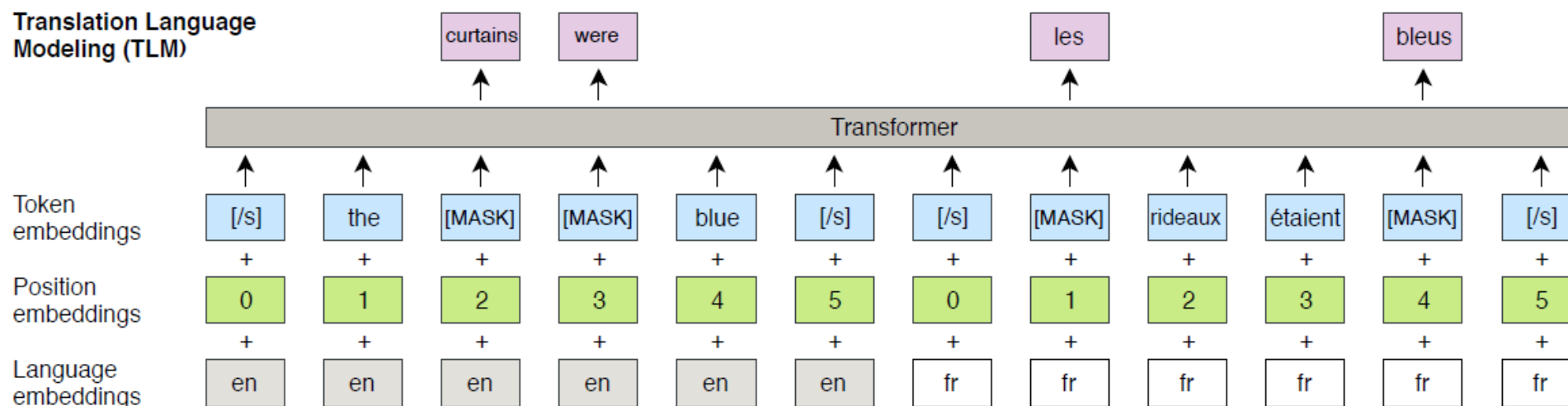
NMT

□ 多语言预训练模型

Masked Language Modeling (MLM)



Translation Language Modeling (TLM)

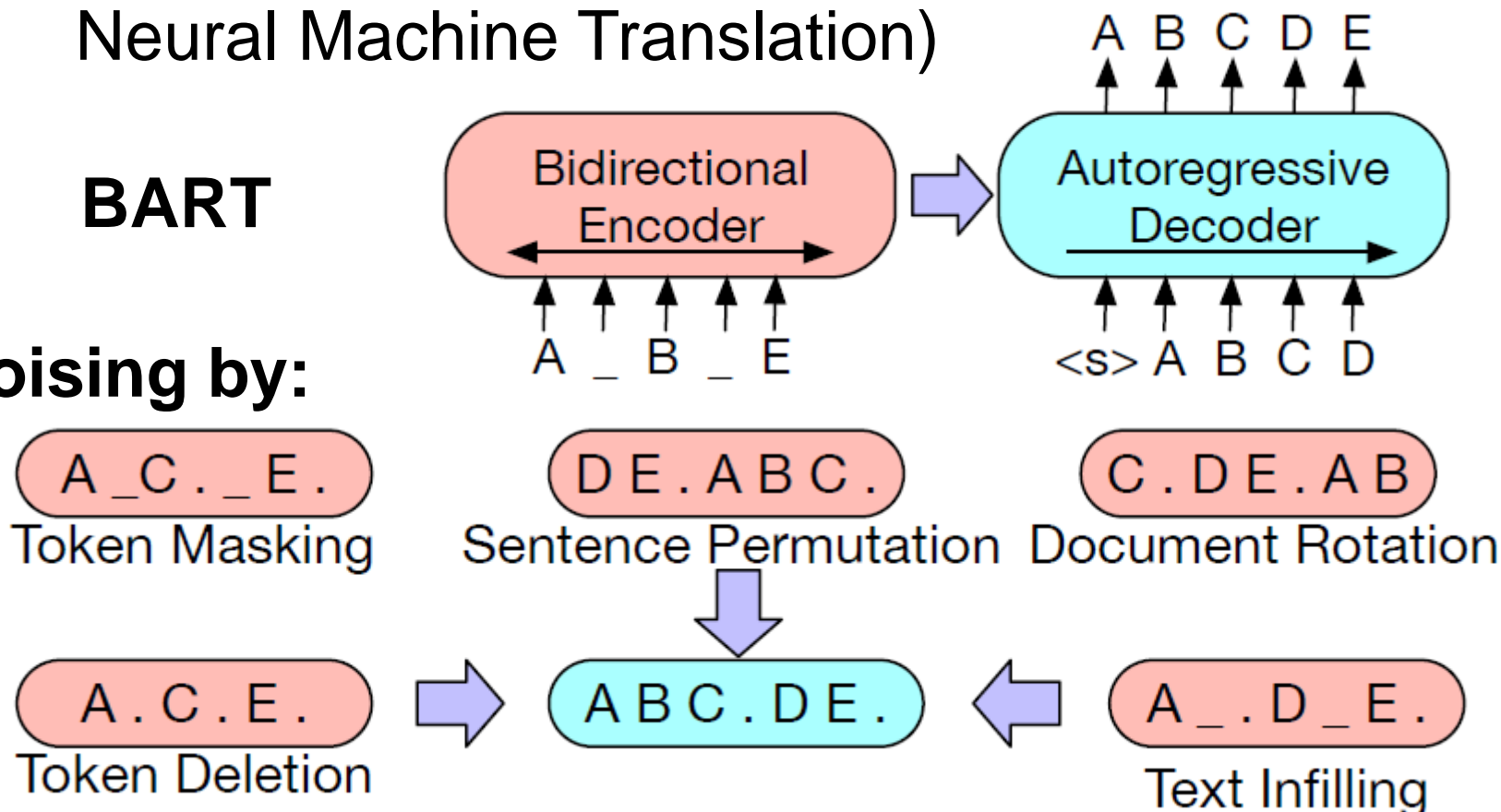


NMT

□ 多语言预训练模型

- mBART (Multilingual Denoising Pre-training for Neural Machine Translation)

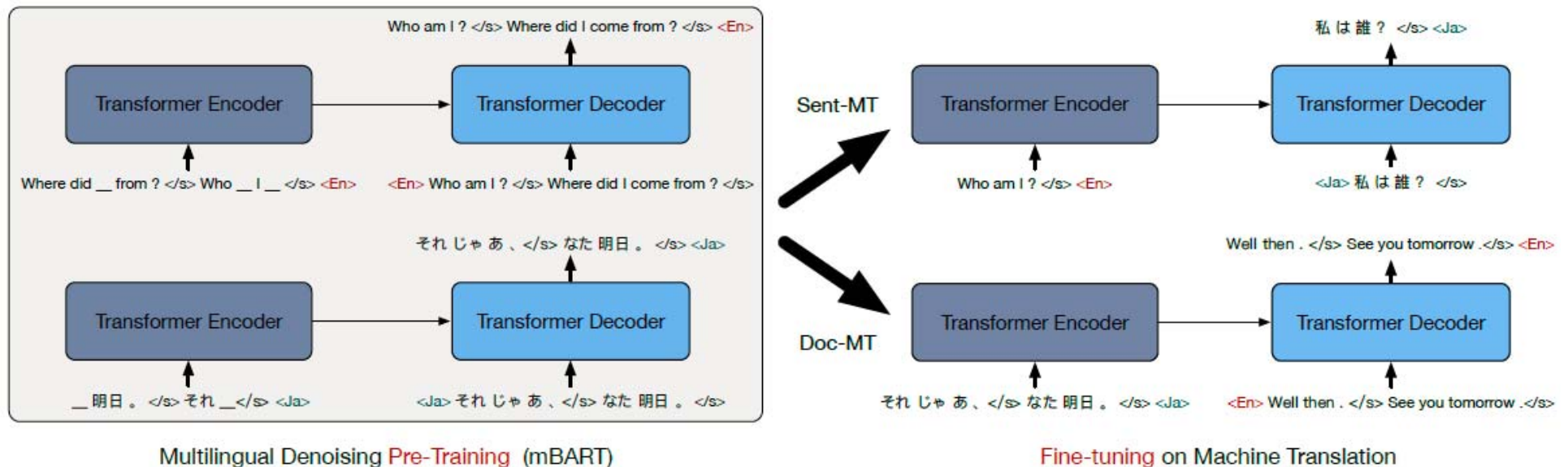
BART
Noising by:



NMT

□ 多语言预训练模型

- mBART (Multilingual Denoising Pre-training for Neural Machine Translation)



mBART

NMT

▣ Robustness

- NMT模型对输入的微小干扰很敏感，从而导致各种不同的错误

这架飞机没有撞上住家或医院，实在是奇迹

It was indeed a miracle that the plane did not touch down at home or hospital.

这架飞机没有撞上住家及医院，实在是奇迹

It was a miracle that the plane landed at home and hospital.

- 在图像中，对像素的微小扰动是无法察觉的，而自然语言中的一个单词的变化都可以被感知到
- 相比SMT，NMT没有明显的翻译、对齐模型，更不可解释

NMT

▣ Robustness

- NMT模型对输入的微小干扰很敏感，从而导致各种不同的错误
 - 如何增强模型的鲁棒性？
 - 人为生成意思相近的输入，用新的输入去“攻击”翻译模型，进而增强模型的鲁棒性。

We have got to follow the rules.

We have got to obey the rules.

NMT

▣ Robustness

- NMT模型对输入的微小干扰很敏感，从而导致各种不同的错误
 - 如何增强模型的鲁棒性？
 - 人为生成意思相近的输入，用新的输入去“攻击”翻译模型，进而增强模型的鲁棒性。



Paraphrase via thesauruses

NMT

▣ Robustness

- NMT模型对输入的微小干扰很敏感，从而导致各种不同的错误
 - 通过对单词微小的改变，使得模型最终的翻译有明显的改变或错误——这种改变最适合作为对抗样本

$$\left\{ \mathbf{x}' \mid \mathcal{R}(\mathbf{x}', \mathbf{x}) \leq \epsilon, \operatorname{argmax}_{\mathbf{x}'} -\log P(\mathbf{y}|\mathbf{x}'; \boldsymbol{\theta}_{mt}) \right\}$$

有目的地设计对抗样本

2019ACL. Robust Neural Machine Translation with Doubly Adversarial Inputs

NMT

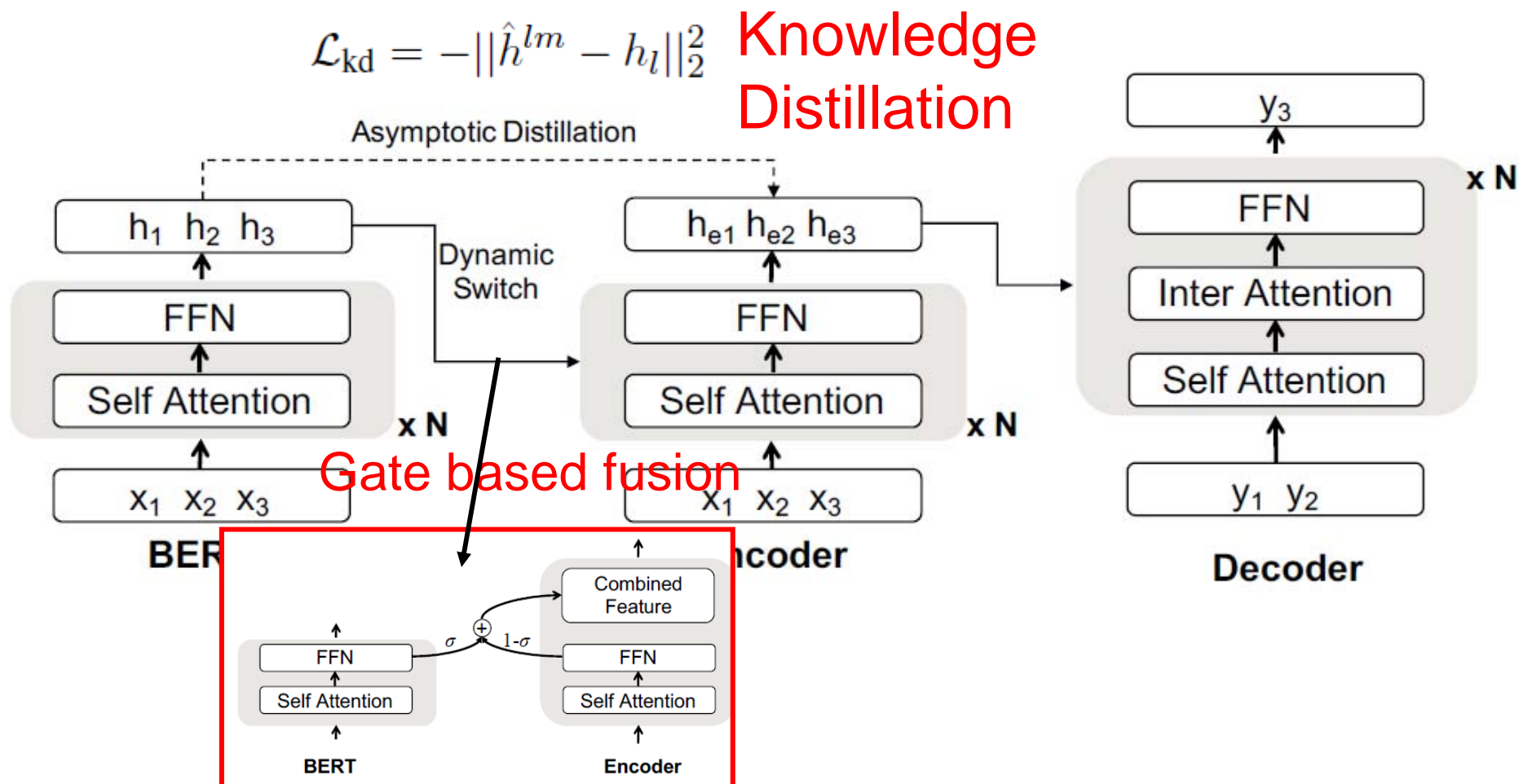
□ 机器翻译是预训练模型应用最不“顺畅”的领域之一

——预训练模型的加入并不能为NMT带来非常显著的提升

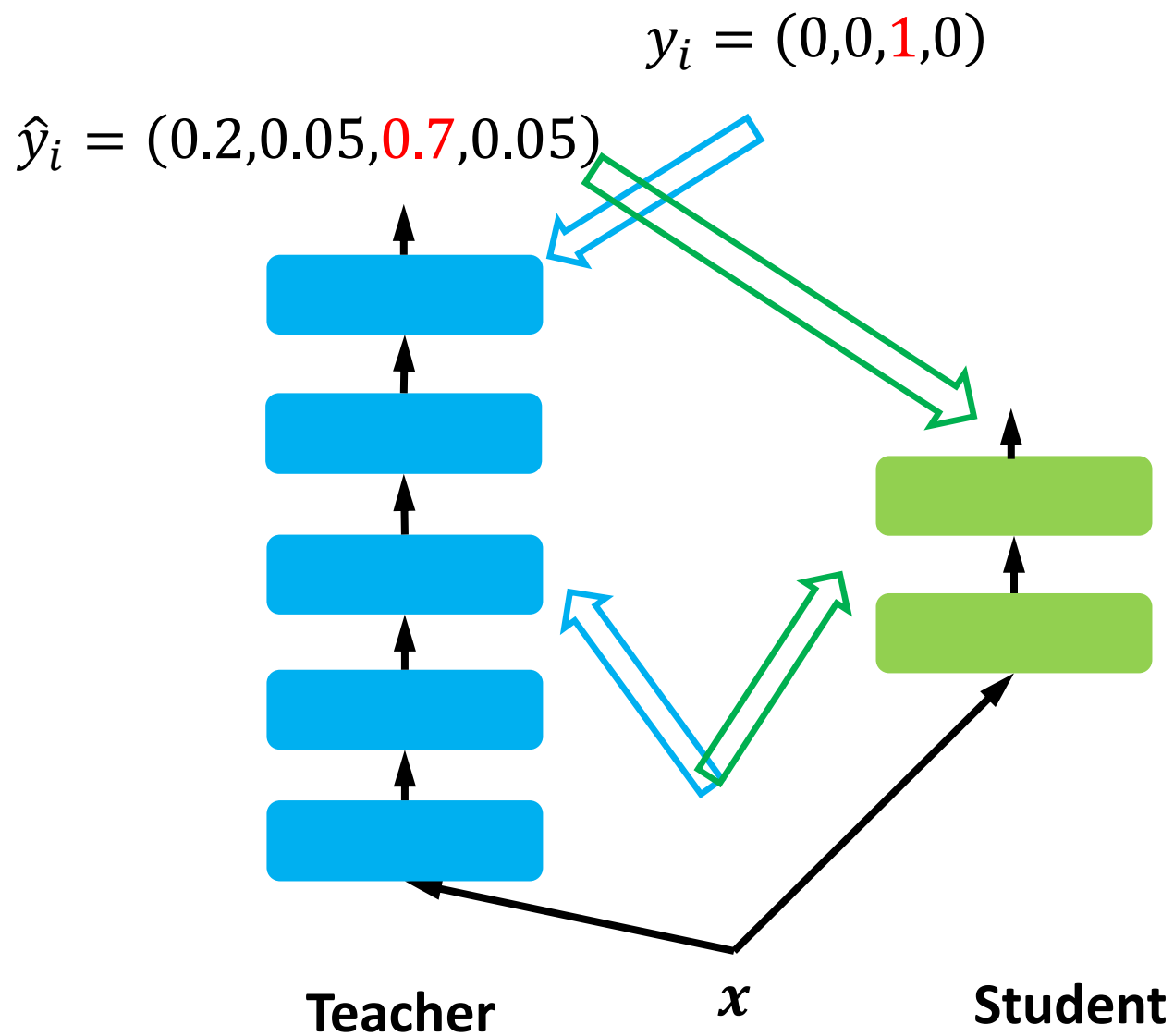
- 任务本身特点
- 语料规模

□ 知识蒸馏

知识蒸馏帮助MT编码端融入预训练知识



□ 知识蒸馏



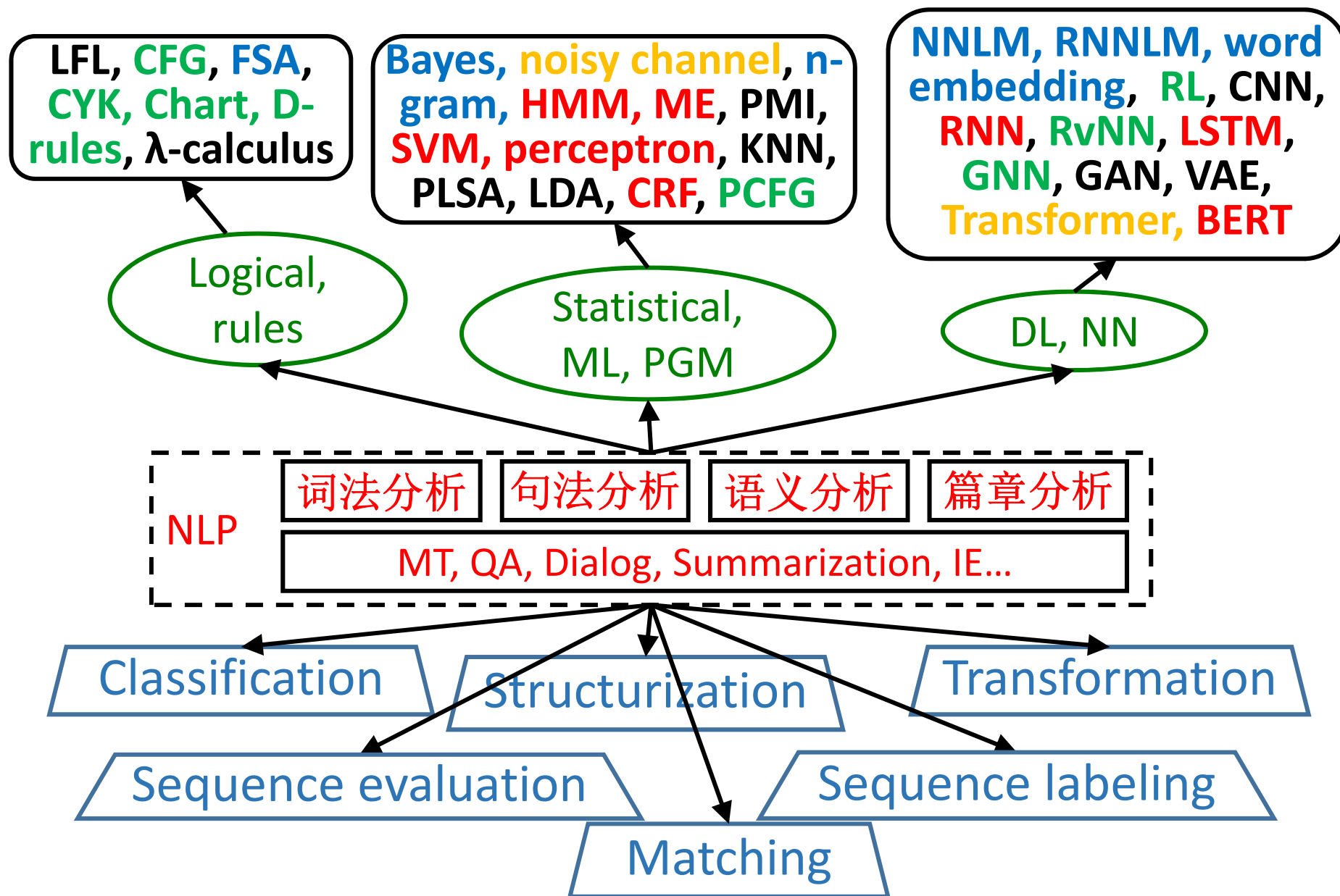
MT是典型的序列转换问题



**Dialog generation, spelling correction,
summarization, paraphrase, addanending**

要求

- 理解噪声信道的word-based SMT、最大熵对数线性的phrase-based SMT原理
- 掌握RNN NMT和Attention NMT原理
- 掌握Attention机制原理和核心计算方法
- 掌握Transformer原理
- 能够利用RNN NMT和Transformer实现NMT
- 能够利用Transformer编码自然语言句子并完成NLP任务



Tasks, problems, methodologies and models in NLP

□ References

- Peter E Brown, Stephen A. DellaPietra, Vincent J. DellaPietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263-311.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133.
- Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 295–302.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404.
- David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of 43rd Annual Meeting of the ACL*, pages 263-270.
- Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insert grammars. In *Proceedings of 43rd Annual Meeting of the ACL*, pages 541-548.
- Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of ACL*, pages 271-279.

□ References

- Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas, pages 115-124.
- Yang Liu , Qun Liu , and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616.
- Philipp Koehn. Statistical Machine Translation. New York, NY: Cambridge University Press. 2009. 宗成庆, 张霄军译. 统计机器翻译. 北京: 电子工业出版社. 2012.
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. 2003. A neural probabilistic language model. JMLR.
- Mikolov, T., Karafiat, S., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language models. In Proceedings of INTERSPEECH 2010.
- Devlin et al. 2014. Fast and robust neural network joint models for statistical machine translation. In Proceedings of ACL.

□ References

- Cho, K., Van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translations. In Proceedings of EMNLP, number 1406.1078 in cs.CL.
- Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. In NIPS.
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. ICML.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. ACL.
- Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, and Guodong Zhou. 2017. Modeling source syntax for neural machine translation. ACL.
- Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term. ACL.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. NIPS.

□ References

- Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. ACL.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. TACL.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. ICLR.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. ICLR.
- Guillaume Lample, Myle Ott, Alexis Conneau, et al. Phrase-Based & Neural Unsupervised Machine Translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.
- Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. ACL.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Masked sequence to sequence pre-training for language generation. ICML