机器翻译

玖强

OUTLINE

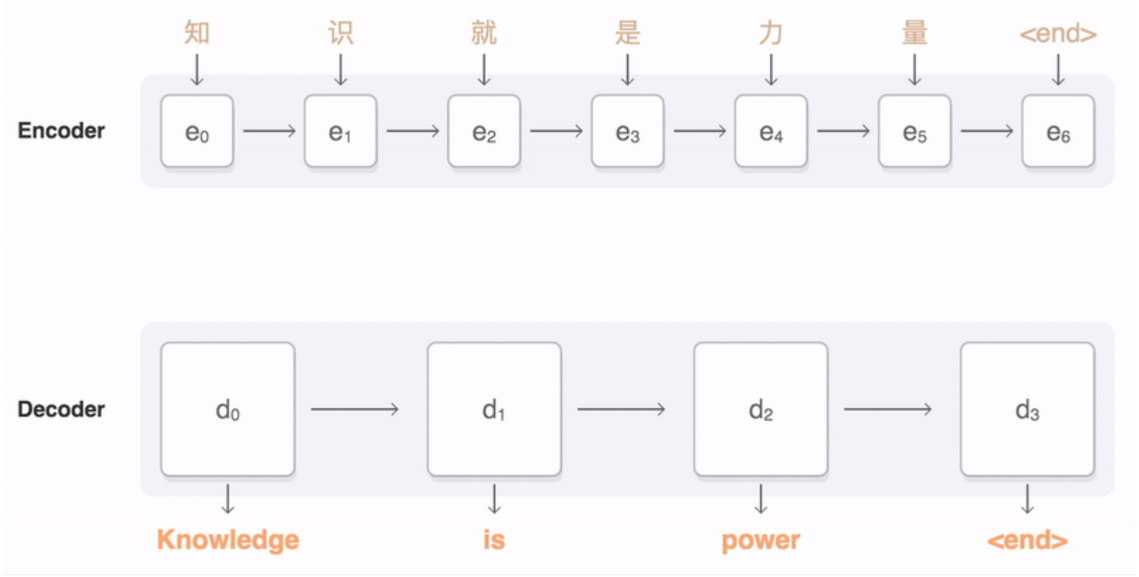
- 机器翻译概述
- 基于统计的机器翻译

机器翻译概述

机器翻译概述

- **Machine Translation （MT） 机器翻译**

用计算机实现从一种自然语言文本（源语言/source language）到另一种自然语言文本（目标语言/target language）的翻译

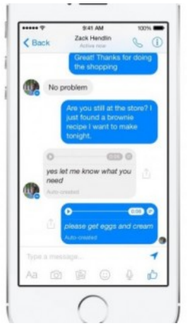


按需求分类

- 传播信息（dissemination）➔出版/信息发布

- 浏览信息（assimliation） ➔ 网页翻译

- 交流信息（interchange） ➔ 实时/多语聊天室



- 查询信息（information access） ➔ 跨语言信息检索

按技术方法分类

理性主义/基于规则的MT方法（RBMT）

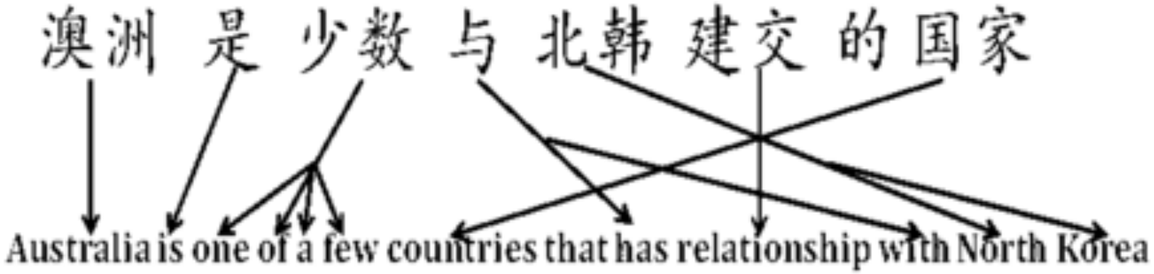
- 直接翻译法
- 转换法
- 中间语言法

经验主义/基于语料库、基于统计的MT方法

- EBMT
- Translation Memory
- Pattern-based MT
- Statistical approach to MT

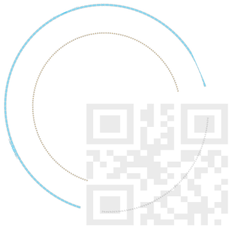
词语对齐

- 词对齐（Word Alignment），所谓词对齐，简而言之就是知道源语言句子中某个词是由目标语言中哪个词翻译而来的



词语对齐

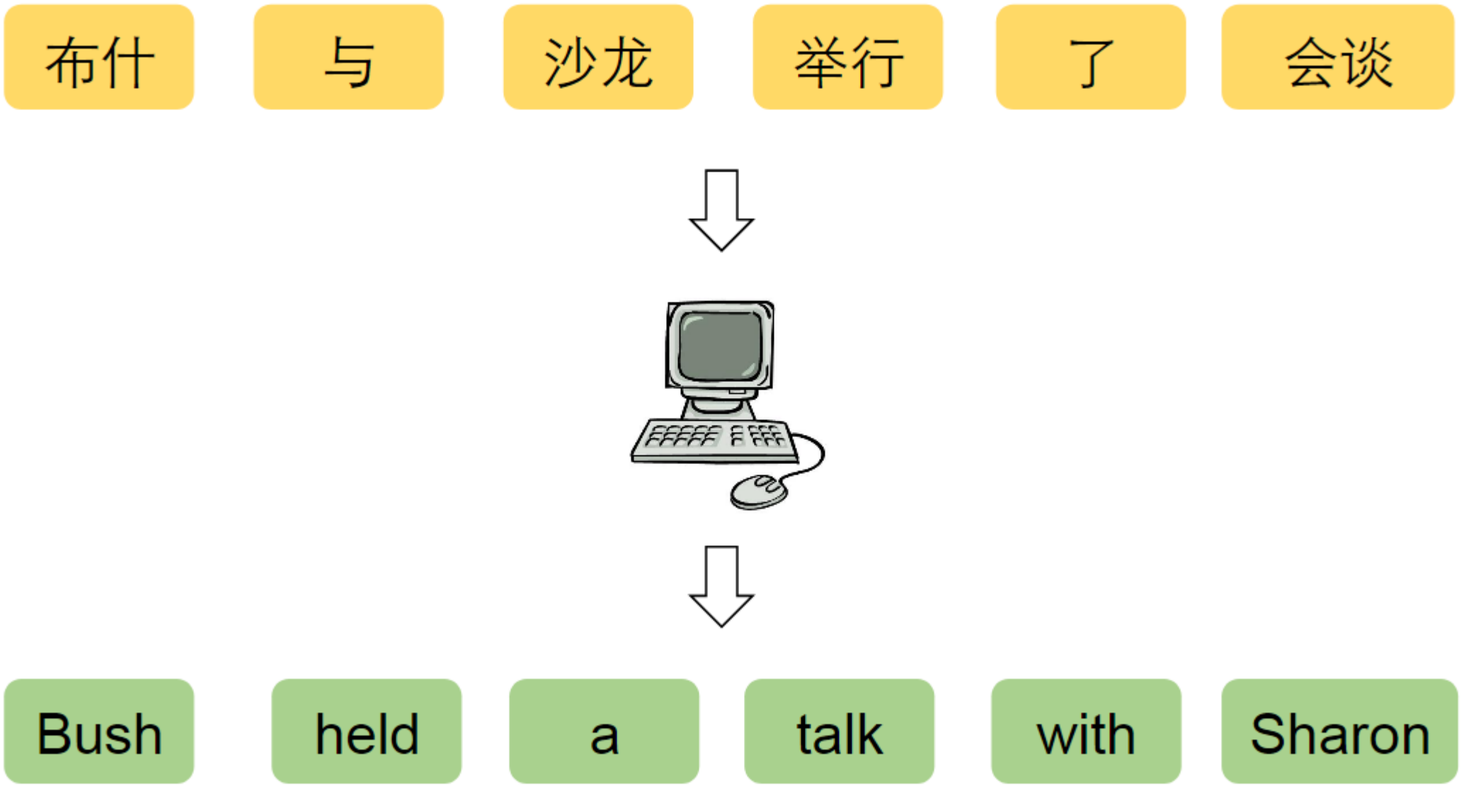
- 自动词语对齐技术在统计机器翻译领域中起了很大的作用
- 词语对齐是自然语言处理领域的一个基本的问题，许多基于双语语料库的应用（如统计机器翻译（SMT）、基于实例的机器翻译（EBMT）、词义消歧（WSD）、词典编撰等）都需要词汇级别的对齐
- 一般来讲，对齐有篇章(section)、段落(paragraph)、句子(sentence)、短语(phrase)、词语(word)等不同级别的对齐，其目的就是从双语互译的文本中找出互译的片段
- 其中篇章、段落、句子的对齐技术主要用于语料库的整理，而短语和词语对齐，就是要找出相互翻译的文本中对应的词与词、词与短语、短语和短语之间的相互翻译对。



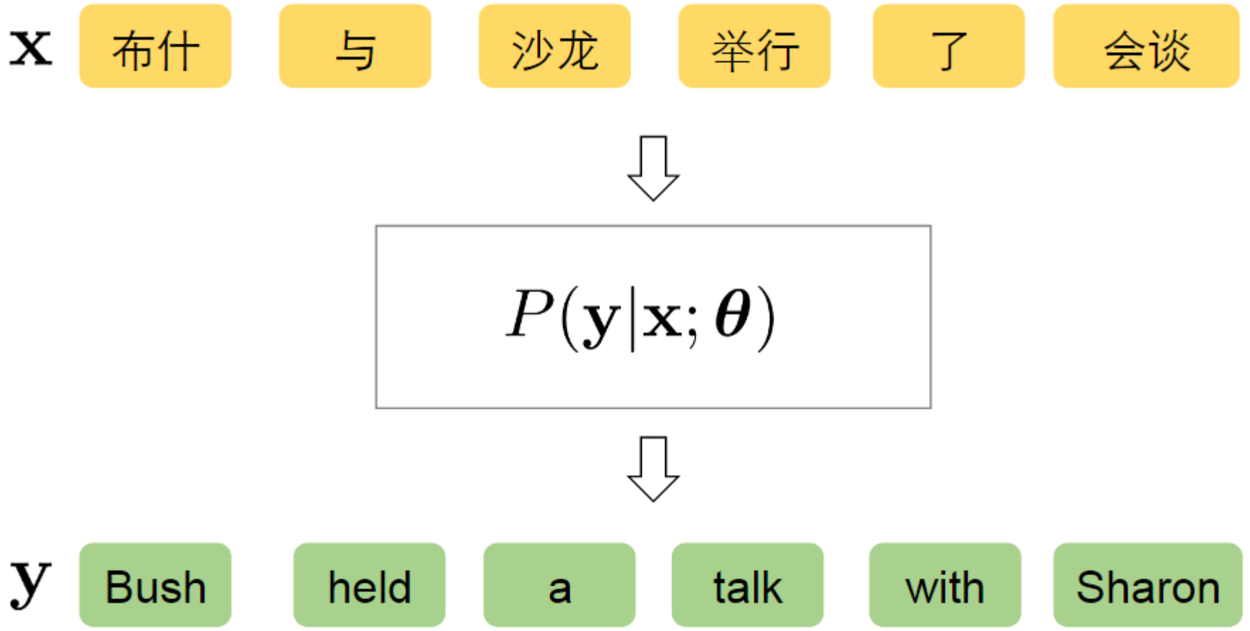
从基本seq2seq模型到最新进展

Encoder-decoder 框架

- End-to-end NMT的基本思想是通过neural network实现自然语言之间的自动翻译
- 为此，NMT通常采用encoder-decoder实现sequence到sequence的转换
- 举例，

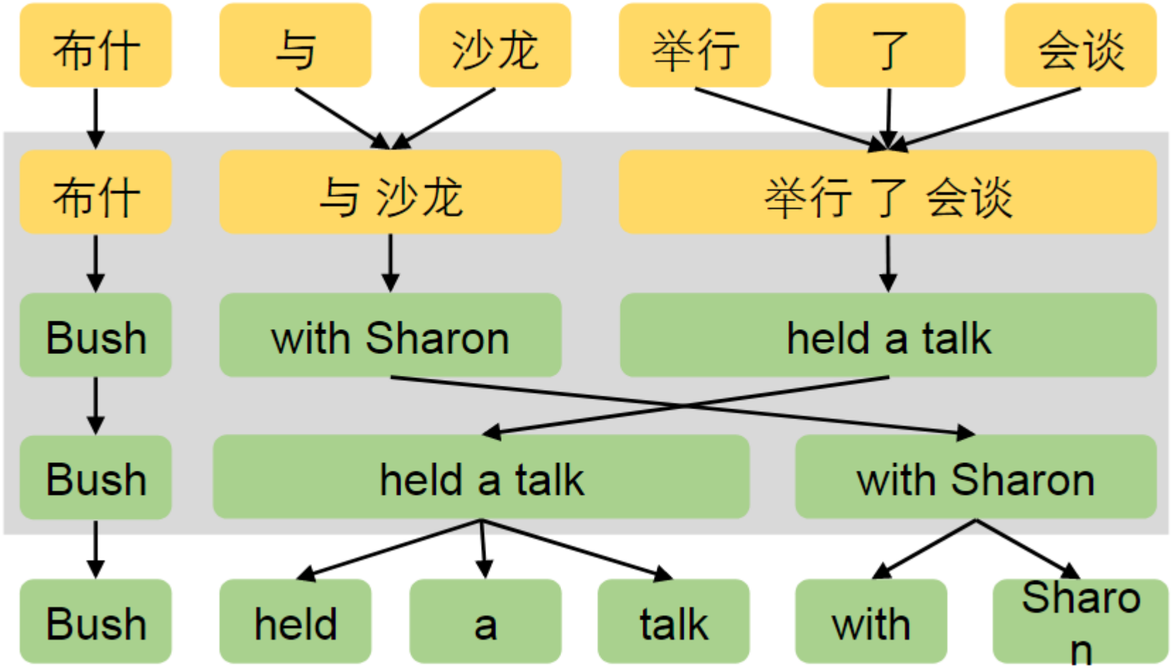


如何为翻译过程建立概率模型？



基于短语的统计机器翻译

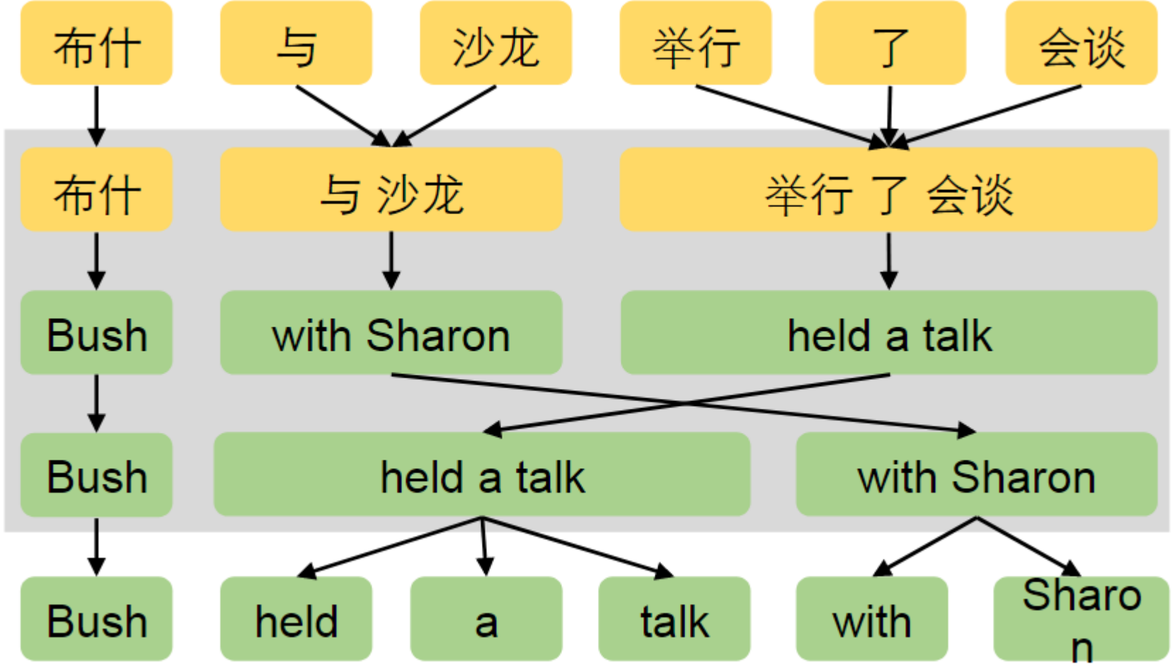
- 基于短语的翻译模型
  - 以隐结构短语为基本翻译单元
  - 选词+调序



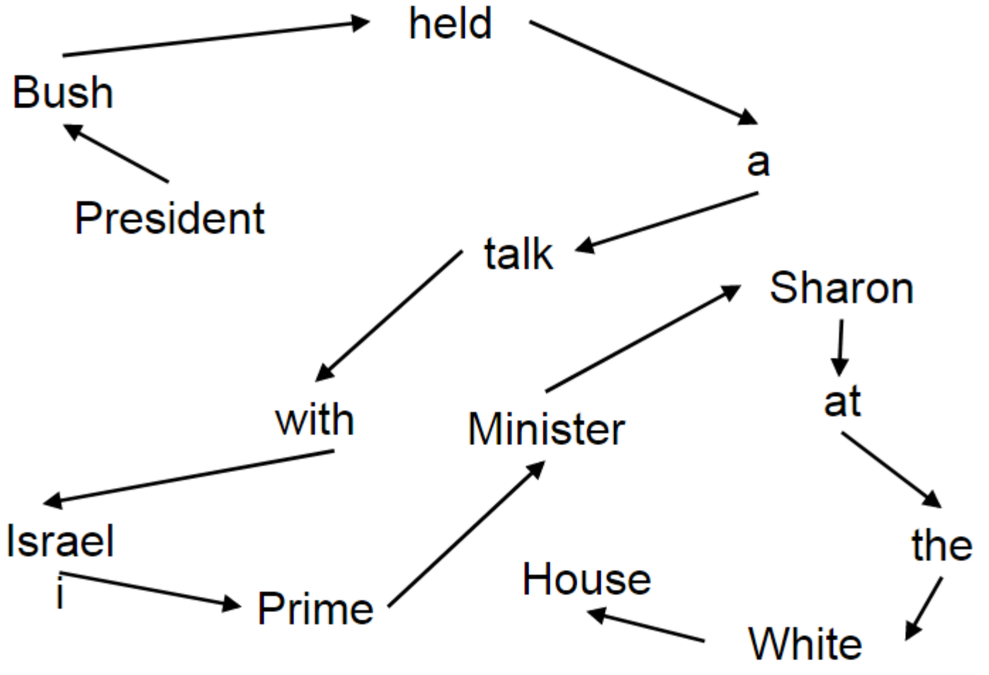
(Koehn et al., 2003)

基于短语的统计机器翻译

- 难点：长距离调序
  - 对于解码而言，层次短语模型使用最基本的CYK算法。自底向上逐渐填充每一个span。对于填充某一个span时，首先枚举该span可能包含的所有规则，对于每一个规则产生一个cube候选，然后对于所有的cube候选进行cube pruning，然后添加到chart图中。 https://en.wikipedia.org/wiki/CYK_algorithm
  - 短语模型依靠词汇化调序模型以及distortion调序模型进行随机调序，产生比较大的歧义，很难解决长距离调序问题



(Koehn et al., 2003)

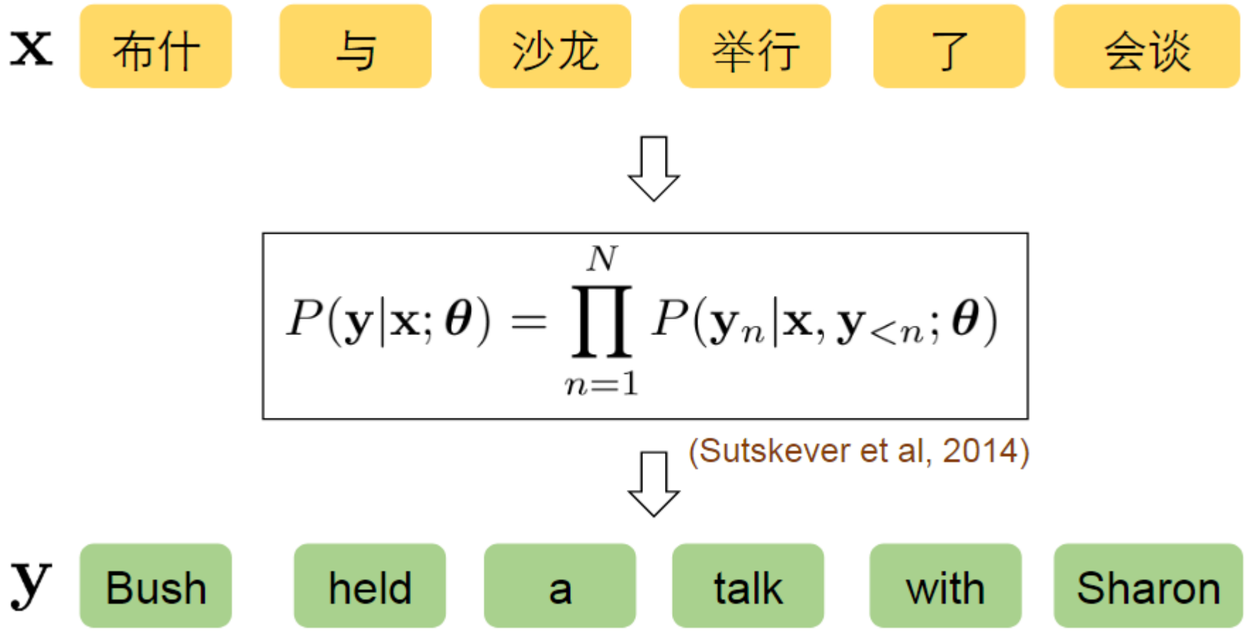


如何用上述词语拼成合理的译文?

神经机器翻译

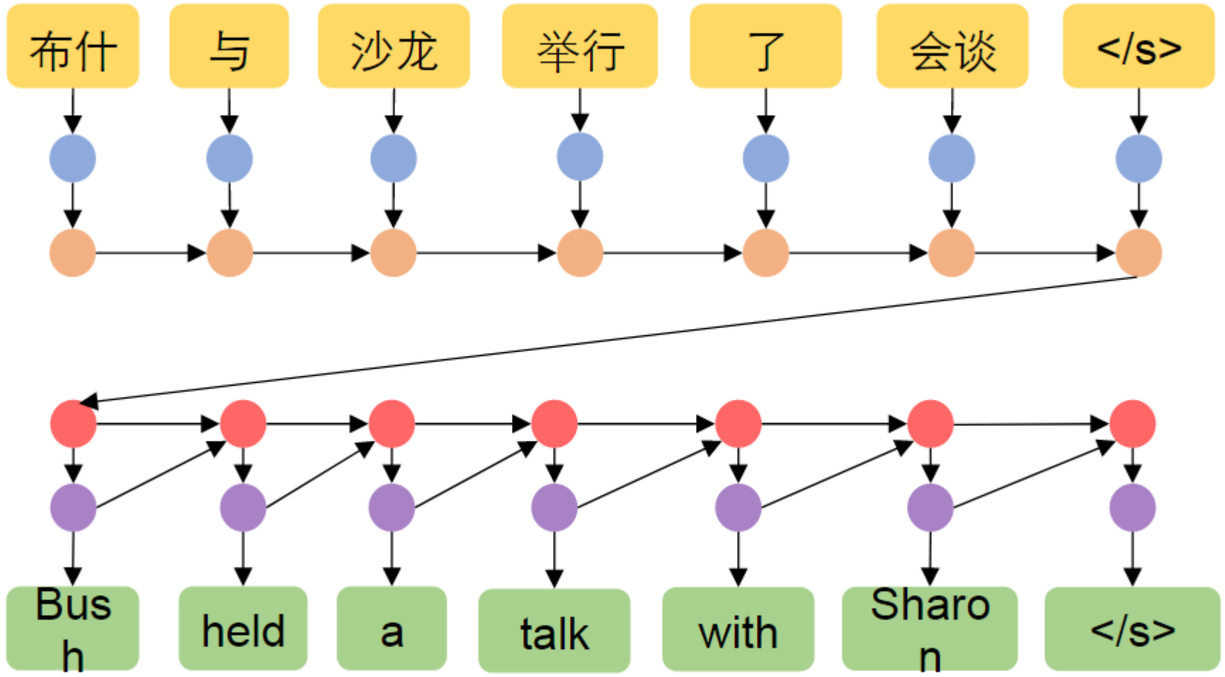
利用神经网络实现自然语言的映射

- X：源文本
- Y：目标文本
- P：联合概率

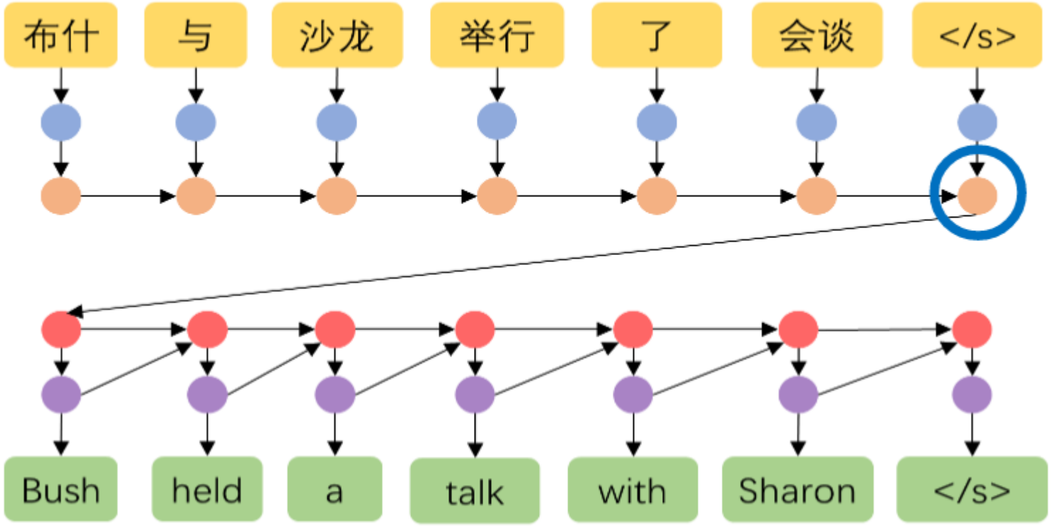


编码器-解码器框架

- 利用RNN实现源Sentence的编码和目标Sentence的解码
- Encoder的hidden输出作为Decoder的输入
- 截止标志位</s>

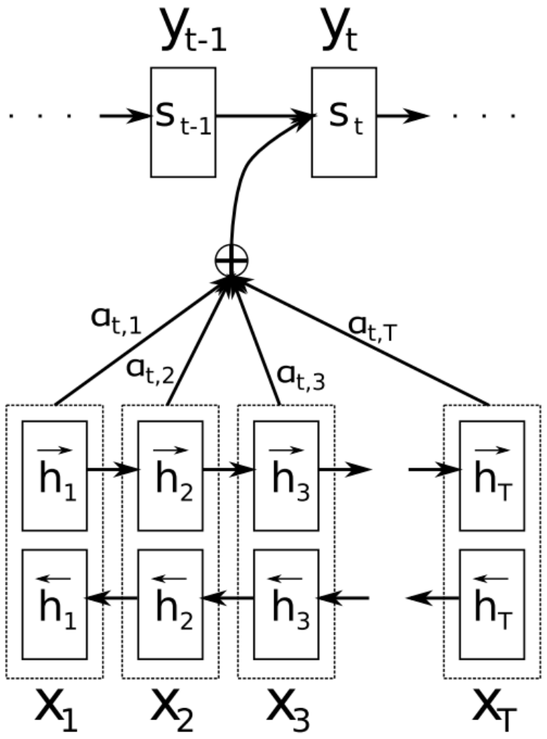


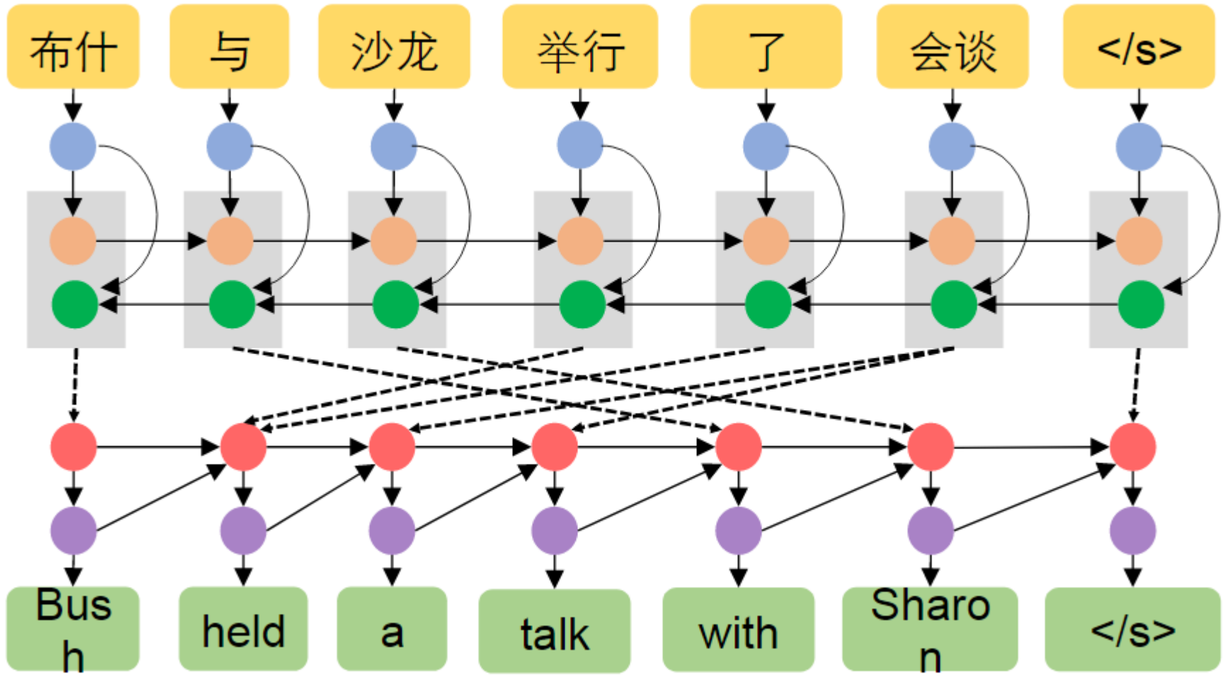
Encoder-Decoder的优缺点



- 优点：利用LSTM处理Long-term dependency
- 缺点：任意长度的句子都编码为固定维度的向量

基于Attention的NMT



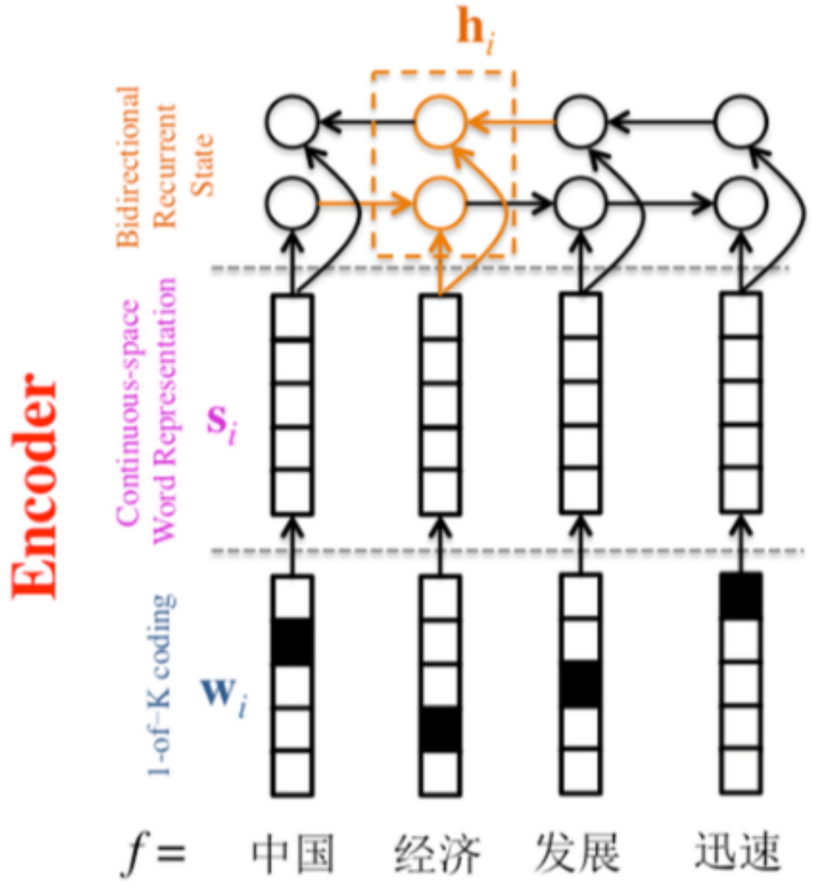


- 利用注意力机制动态计算源语言端相关上下文

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Encoder的改进

- 双向GRU包含一个前向和一个后向的GRU
- 前向按照词序列的顺序依次压缩源语言端词 $(\overrightarrow{h}_1, \overrightarrow{h}_2, ..., \overrightarrow{h}_T)$
- 后向安装逆序压缩源语言端词 $(\overleftarrow{h}_1, \overleftarrow{h}_2, ..., \overleftarrow{h}_T)$
- 对于词i，它的隐藏状态通过连接2个RNN的结果得到，$h_i = [\overrightarrow{h}_i^{\mathrm{T}}; \overleftarrow{h}_i^{\mathrm{T}}]^{\mathrm{T}}$
- 新的hidden state压缩了前向和后向的表示，并且更加关注于词i的周围词，使得RNN更好的表达当前的输入。



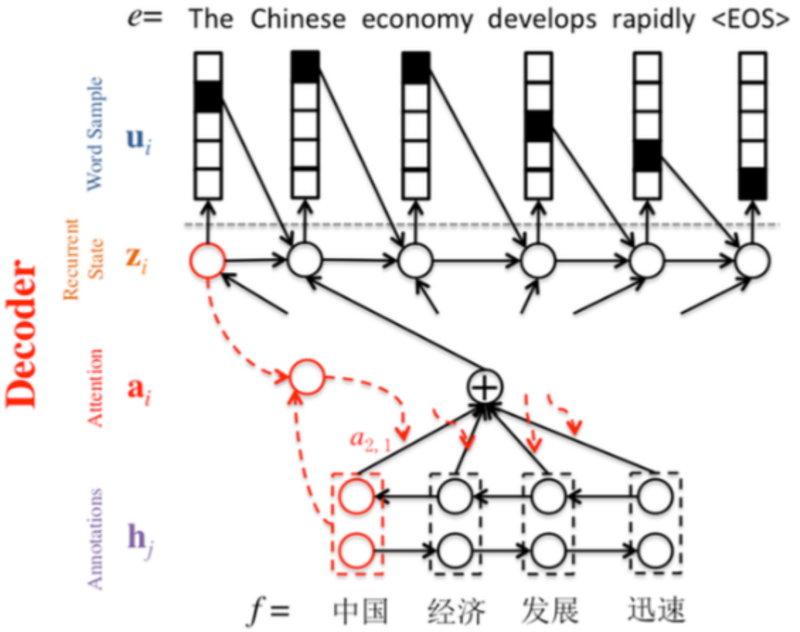
Decoder的改进

- Attention的加权得到 $c_i = \sum_{j=1}^{T} \alpha_{ij} h_j$

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

- 权重计算为

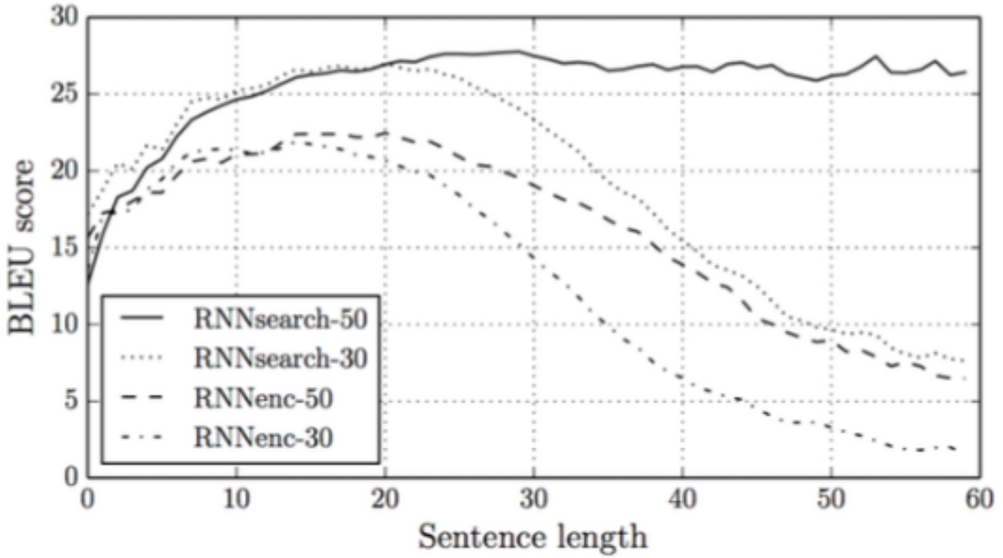
$$\begin{aligned} e_{ij} &= a(z_{i-1}, h_j) \\ &= v_a^{\mathrm{T}} \tanh(W_a z_{i-1} + U_a h_j) \end{aligned}$$



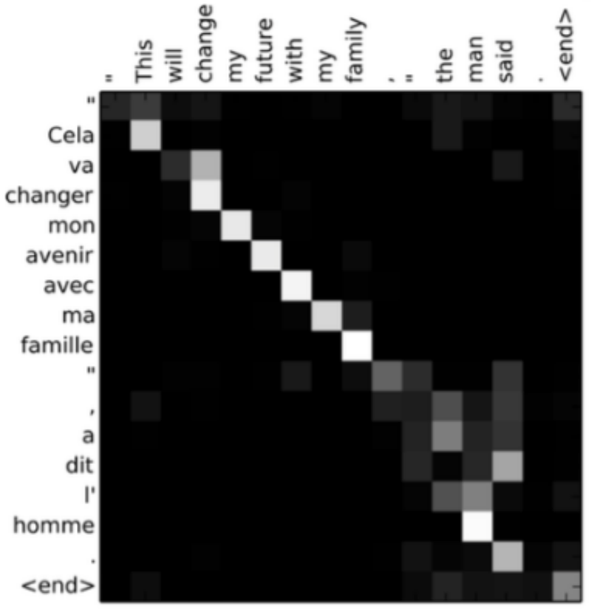
- Decoder中的a可以看做是一个对齐模型，用来衡量第j个源端词与目标端第i个词的匹配程度
- 不同于传统的对齐模型：源语言端每个词明确对起到目标语言端的一个或多个词，而这种方法计算得到的是一种soft alignment，可以融入整个nmt框架，通过反向算法求梯度以及更新参数。
- Attention机制引入NMT中，可以使得Decoder更多的关注于源语言端部分词，从而缓解Encoder-Decoder框架中将源语言压缩成固定向量带来的问题。

结果分析WMT'14 English/French

- RNNsearch结果优于传统基于短语的统计机器翻译系统
- 引入attention是为了解决不同长度的源语言句子都用相同且固定维度的压缩向量表示所带来的性能瓶颈。
- 右图比较了不同句子长度的模型结果，可以看出RNNsearch的鲁棒性比RNNEnc更好

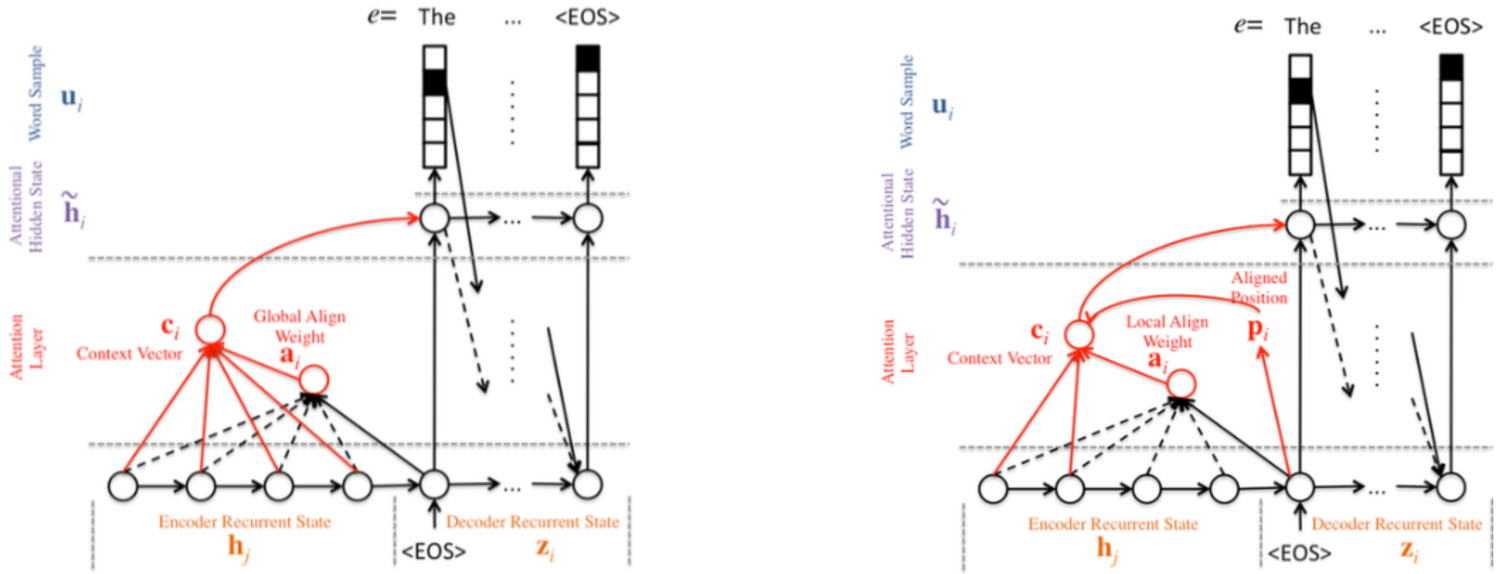


- 对齐模型分析：所选的例子是随机从句长10-20且不包含UNK的结果中采样得到的。横轴对应源语言，纵轴对应目标语言，对齐权重由灰度值表示。
- 可以看出，颜色较白的块大致分布于对角线附近，也有一定效果
- Soft-attention: 为t时刻的输出y_t求上下文向量c_t时，为输入句子的每一个单词都给出一个注意力概率（也可以看做一个对齐模型），得到的结果是一个概率分布
- Hard-attention：对于图像有用，但是对于nlp，对齐要求太高，可能会导致负面影响。同时，hard-attention模型不可微，需要复杂的优化方法



Attention优缺点

- Luong 等人在NMT中引入了Global + Local的方式
- Global的方式和之前的soft-attention类似
- Local可以看做soft-attention和hard-attention的一种折中，计算代价比全局的方式或者soft-attention的方式小。且对于hard-attention，soft-attention可微且便于训练。



Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

https://rawgit.com/gujiuxiang/PaperNotes/master/post/Recurrent_Neural_Network/20150804_Effective_approaches_to_attention-based_neural_machine_translation.md.html

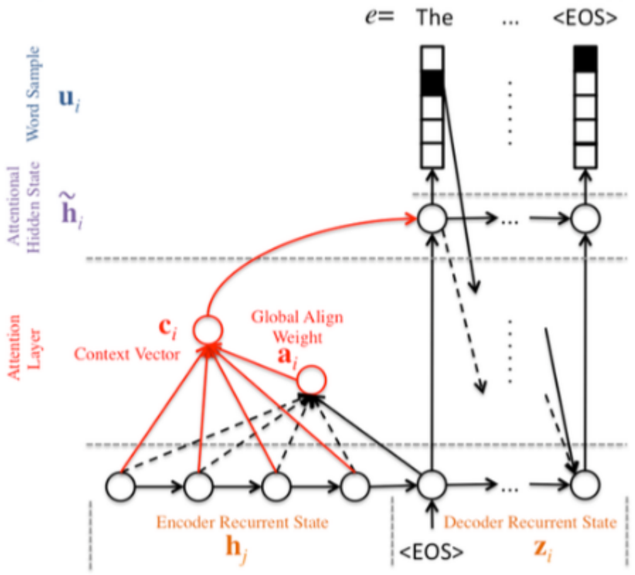
Global Attention

- Global attention的思想是计算source端上下文向量c_i时，考虑Encoder的所有隐藏状态（h_1,…,h_T）。其间的对应权重（align probability）通过比较当前目标端隐层状态z_i和source端的hidden state h_j得到：

$$a_{i,j} = \text{align}(z_i, h_j) = \frac{\exp(\text{score}(z_i, h_j))}{\sum_{j'} \exp(\text{score}(z_i, h_{j'}))}$$

- 这里的score函数，他们定义了3种方式：

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_a \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_a [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

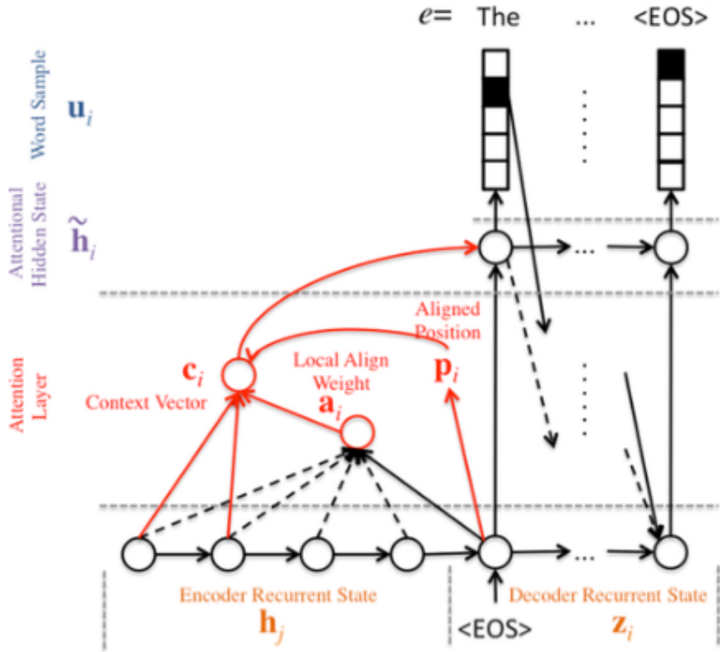


$$p(y_i | y_{<i}, \mathbf{x}) = \text{softmax}(\mathbf{W}_s \tilde{h}_i)$$

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

Local Attention

- Global attention的思想是计算每个目标端词和每个源语言词端的对齐概率，这也许会称为一种缺点，尤其针对长句子，这种方法的代价很大。
- 因此，有了一种折中的方法，来自xu etal中soft+hard
- 局部attentio对i时刻的输出生成一个它在源语言端的对齐位置p_i，接着在源语言端取窗口[p_i-D,p_i+D]，上下文向量c_i则通过计算窗口内的hidden state的加权平均得到
- 至于这个对齐位置p_i如何确定，他们定义了2种：local-m和local-p。



$$p(y_i|y_{<i}, \mathbf{x}) = \text{softmax}(\mathbf{W}_s\tilde{h}_i)$$

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

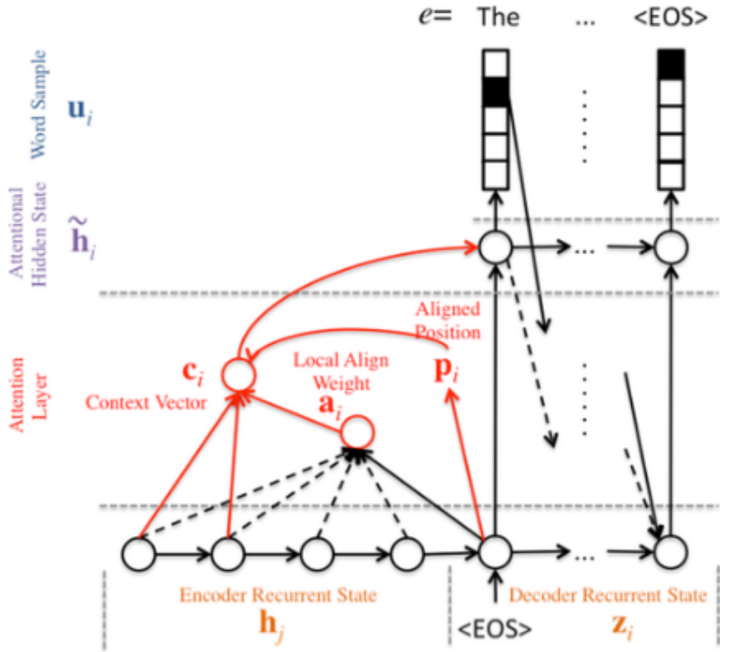
Local Attention

- Monotonic alignment (local-m):就是简单设置p_i=I
- Predictive alignment (local-p): 针对每个目标端输出，预测它在源语言端的对齐位置，计算公式为：

$$p_i = T \times \text{sigmoid}\left(v_p^{\mathrm{T}} \tanh\left(\mathbf{W}_p z_i\right)\right)$$

- 其中，W_p和v_p都是模型参数，T是源语言端句子长度。最后文章引入一个服从于N(p_i, D/2)的高斯分布来设置对齐权重，因为直觉上，离对齐位置p_i距离越近，对后续决策的影响越大。那么目标端位置i与源语言端位置j（在窗口内）的对齐概率计算如下：

$$a_{i,j} = \text{align}(z_i, h_j) \exp\left(-\frac{(j - p_i)^2}{2\sigma^2}\right)$$



$$p(y_i|y_{<i}, \mathbf{x}) = \text{softmax}(\mathbf{W}_s \tilde{h}_i)$$

- Align函数的定义与softmax定义类似

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." *arXiv preprint arXiv:1508.04025* (2015).

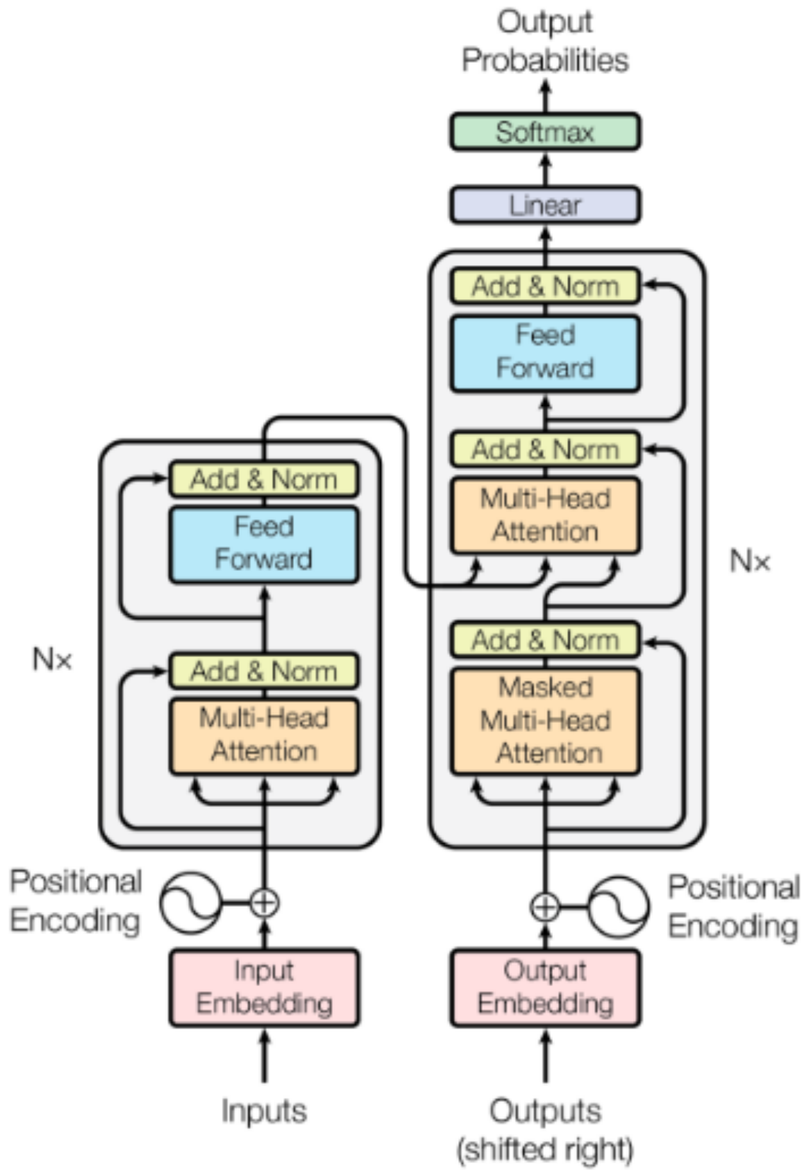
Attention is All You Need

- 卷积序列到序列学习
- RNNs的缺点：依赖于全部历史信息，难以并行化
- CNNs的优点：不依赖于全部历史信息，高度并行化
- Transformer
- 同样试图解决RNNs难以并行化的缺点，但既不使用RNNs也不使用CNNs，只使用注意力机制
- 在这篇文章当中有一个对Attention很好的描述，即attention机制实际上来讲是一个由诸多Query和Key-value pair组成的映射函数

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

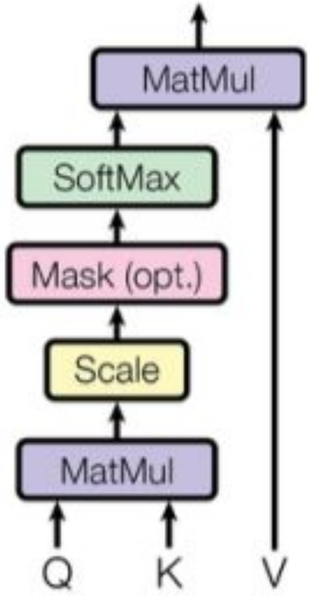
- 上式被作者们称为Scaled Dot-Product Attention. 在这里面 \sqrt{d_k} 是用来约束点积大小的。作者认为当query和key的维度很大时，点击倾向于变得比较大，因而上述因子做约束

https://rawgit.com/gujiuxiang/PaperNotes/master/post/Recurrent_Neural_Network/20170612_Attention_is_All_You_Need.md.html

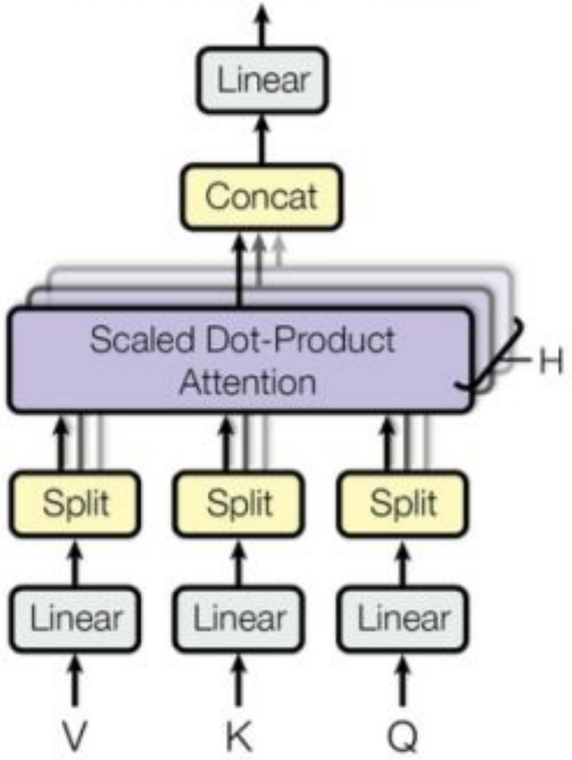


Attention is All You Need

Scaled Dot-Product Attention



Multi-Head Attention



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

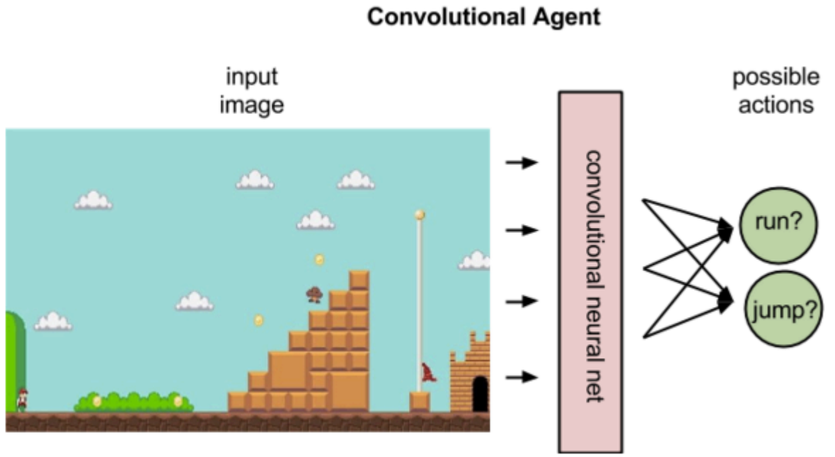
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Training with Reinforcement Learning

- 最大奖励或最小惩罚：针对评价指标训练神经网络

训练数据: $\left\{\left\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \right\rangle\right\}_{s=1}^{S}$

训练目标:
$$\mathcal{R}(\boldsymbol{\theta}) = \sum_{s=1}^{S} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(s)};\boldsymbol{\theta}} \left[\Delta(\mathbf{y}, \mathbf{y}^{(s)})\right]$$
$$= \sum_{s=1}^{S} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(s)})} P(\mathbf{y}|\mathbf{x}^{(s)};\boldsymbol{\theta}) \Delta(\mathbf{y}, \mathbf{y}^{(s)})$$



优化: $\hat{\boldsymbol{\theta}}_{\text{MRT}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left\{\mathcal{R}(\boldsymbol{\theta})\right\}$

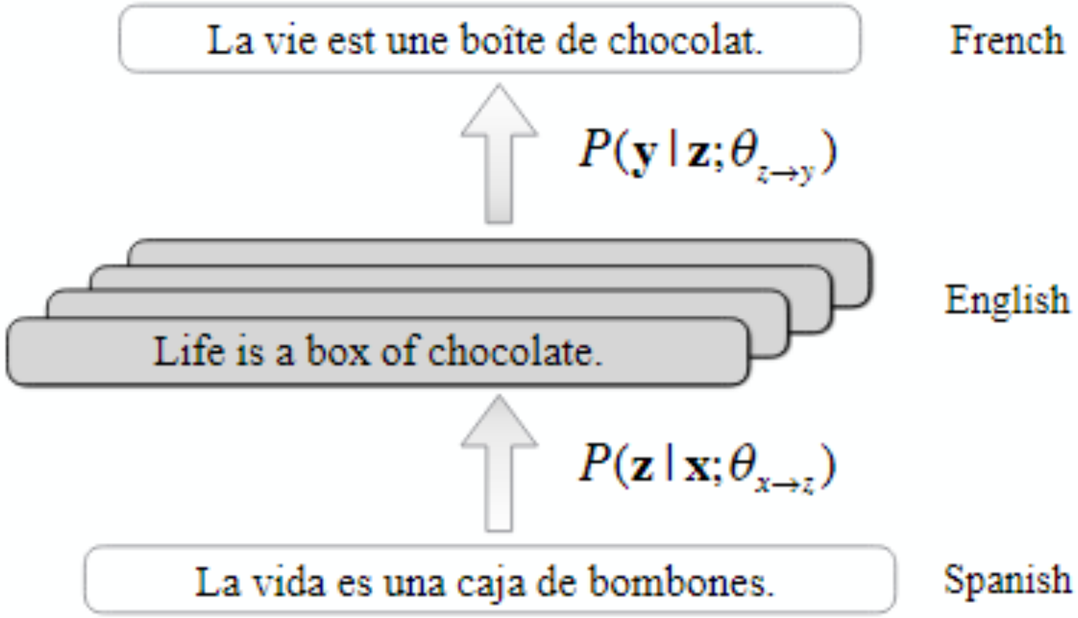
通用性：适用用于任意架构和任意损失函数

零资源语言翻译

- 基于Pivot language的联合训练

Cheng, Yong, et al. "Joint training for pivot-based neural machine translation." Proceedings of IJCAI. 2017.

针对低资源语言的神经机器翻译提出了源语言-桥接语言和桥接语言-目标语言翻译模型的联合训练算法，增强两个模型在参数估计中的关联性

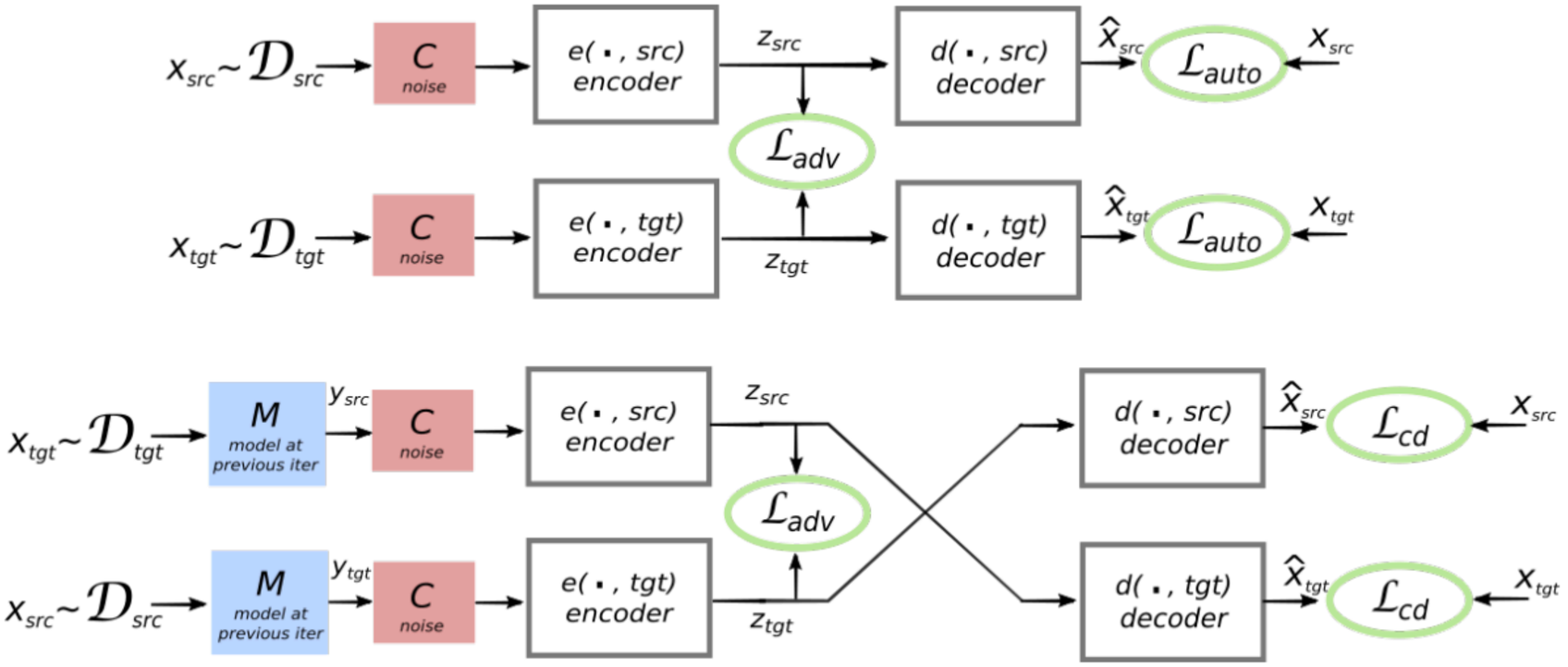


https://rawgit.com/gujiuxiang/PaperNotes/master/post/Recurrent_Neural_Network/20161115_Joint_training_for_pivot-based_neural_machine_translation.md.html

更多的改进。。。

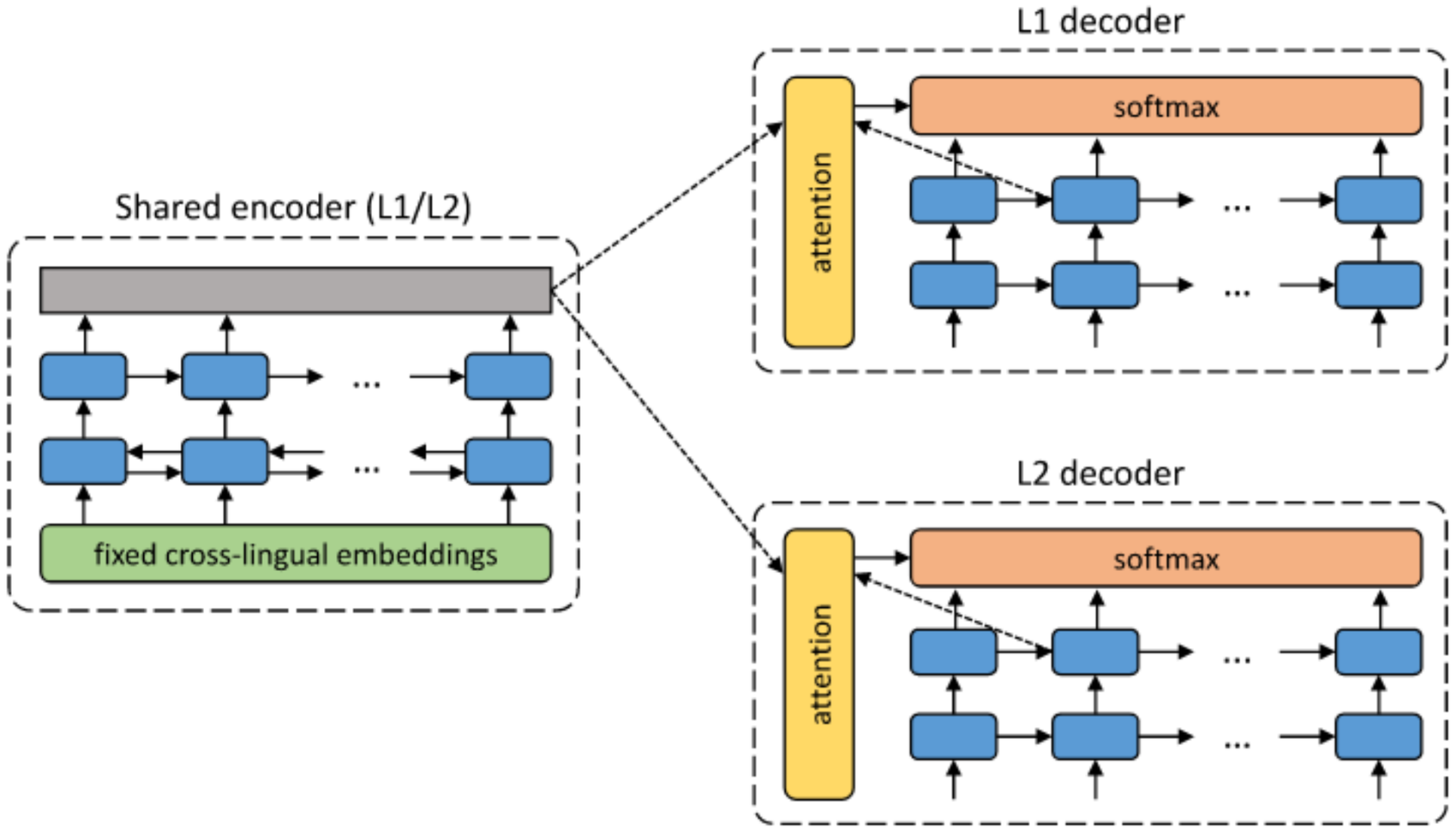
Unsupervised Machine Translation Using Monolingual Corpora Only

Lample, Guillaume, Ludovic Denoyer, and Marc'Aurelio Ranzato. "Unsupervised Machine Translation Using Monolingual Corpora Only." arXiv preprint arXiv:1711.00043 (2017).



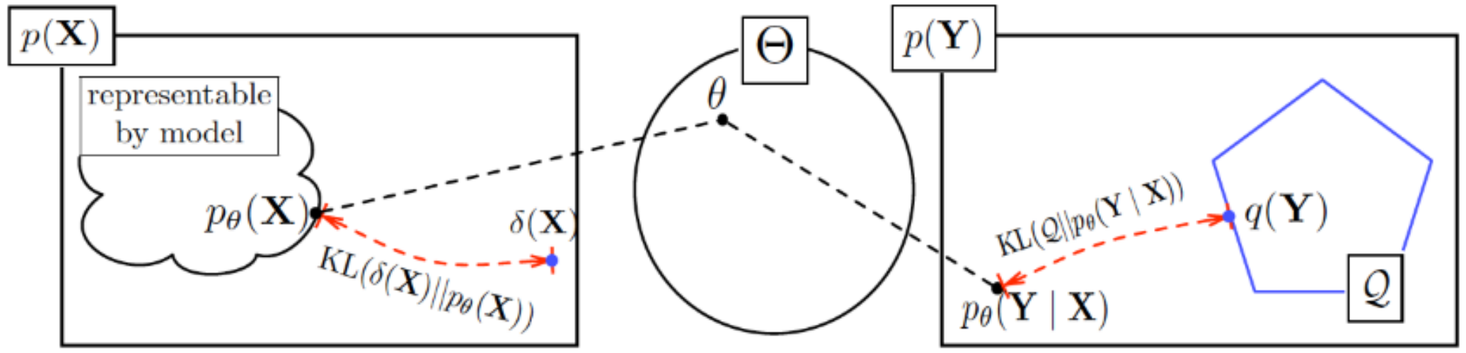
Unsupervised Neural Machine Translation

Artetxe, Mikel, et al. "Unsupervised neural machine translation." arXiv preprint arXiv:1710.11041 (2017).



先验约束知识融合

- 基于后验正则化加入离散特征



Ganchev, Kuzman, Jennifer Gillenwater, and Ben Taskar. "Posterior regularization for structured latent variable models." Journal of Machine Learning Research 11.Jul (2010): 2001-2049.

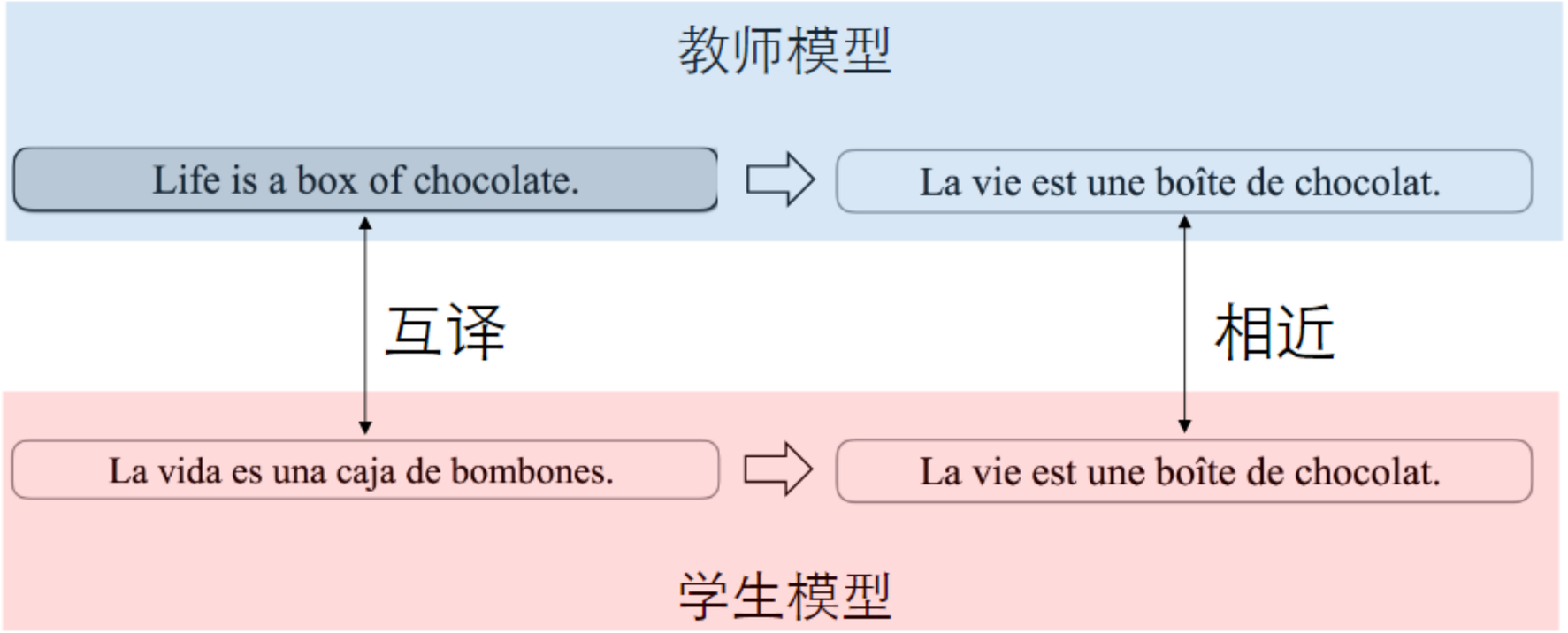
- 训练目标 $$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \lambda_1 \mathcal{L}(\boldsymbol{\theta}) - \lambda_2 \sum_{n=1}^{N} \mathrm{KL}\Big(Q(\mathbf{y}|\mathbf{x}^{(n)}; \boldsymbol{\gamma}) \Big|\Big| P(\mathbf{y}|\mathbf{x}^{(n)}; \boldsymbol{\theta})\Big)$$

- 先验知识 $$Q(\mathbf{y}|\mathbf{x}; \boldsymbol{\gamma}) = \frac{\exp\Big(\boldsymbol{\gamma} \cdot \boldsymbol{\phi}(\mathbf{x}, \mathbf{y})\Big)}{\sum_{\mathbf{y}'} \exp\Big(\boldsymbol{\gamma} \cdot \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}')\Big)}$$

零资源语言翻译

- 面面向零资源翻译的 “教师-学生生” 框架

Chen, Yun, et al. "A Teacher-Student Framework for Zero-Resource Neural Machine Translation." arXiv preprint arXiv:1705.00753 (2017).



$$\mathcal{J}_{\text{SENT}}(\boldsymbol{\theta}_{x\to y}) = \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \text{KL}\Big(P(\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to y}) \Big|\Big| P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big)$$

Interactive Attention for Neural Machine Translation

Meng, Fandong, et al. "Interactive attention for neural machine translation." arXiv preprint arXiv:1610.05011 (2016).

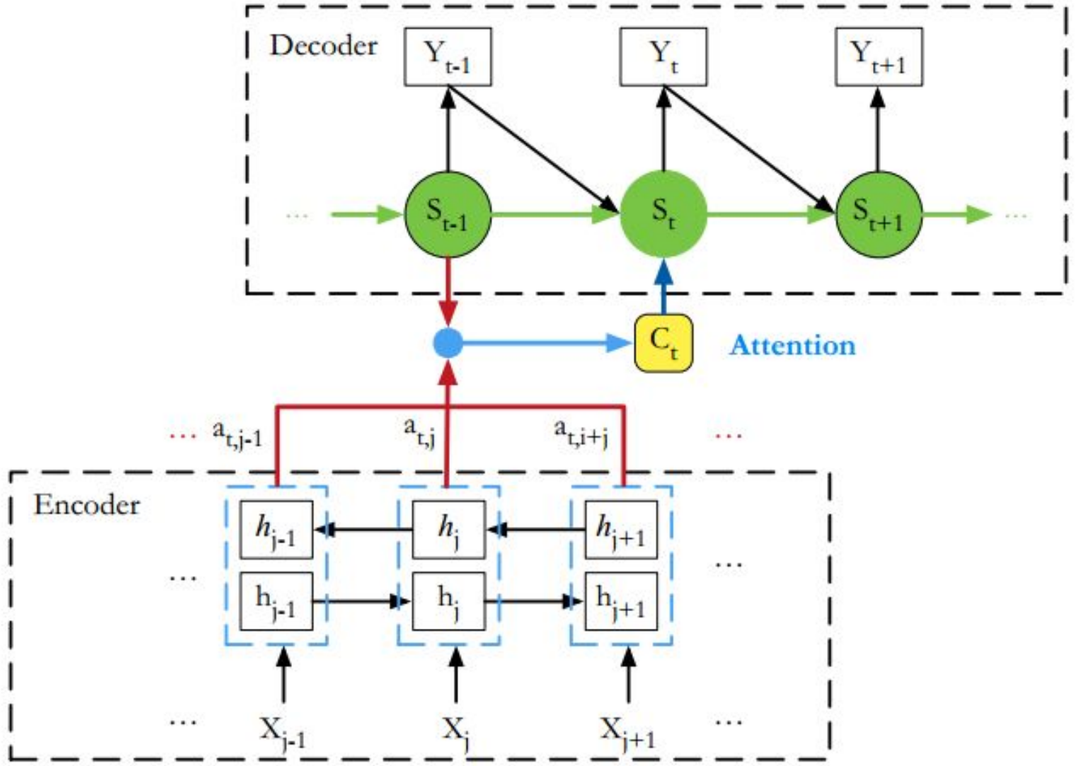


Figure 1: Illustration for attention-based NMT.

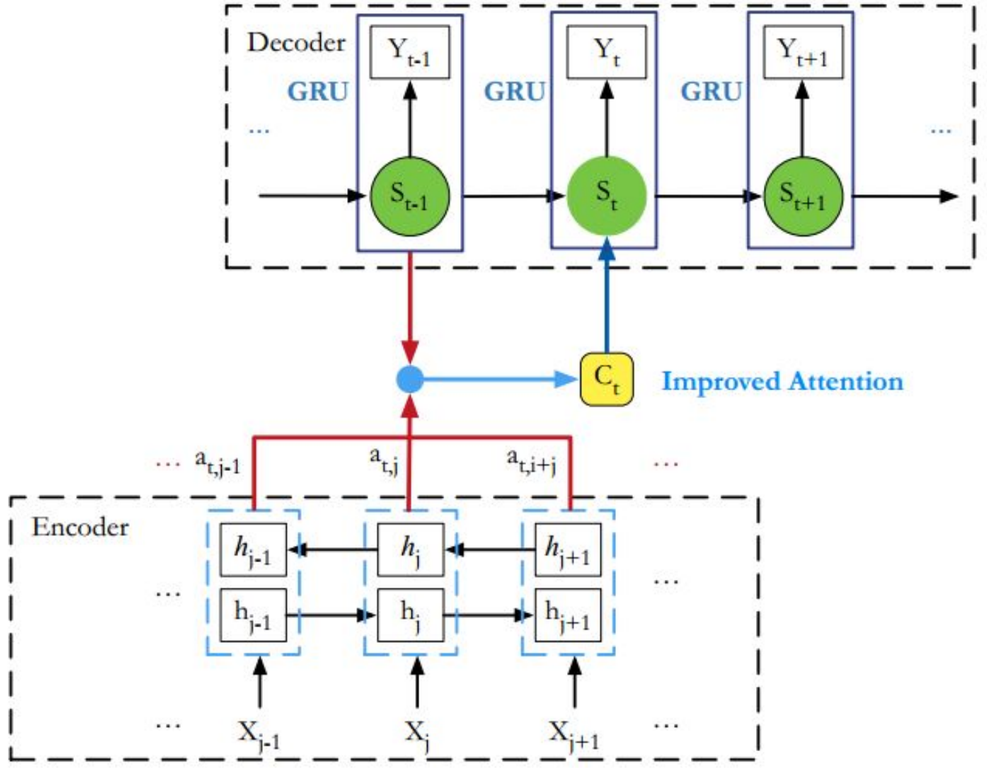
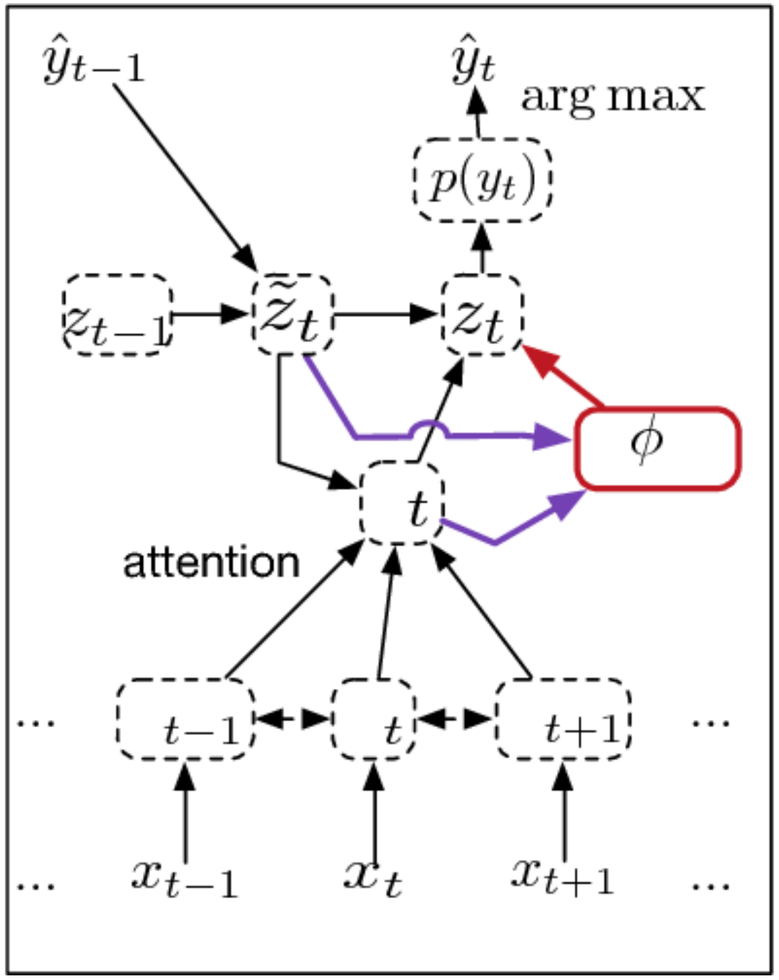
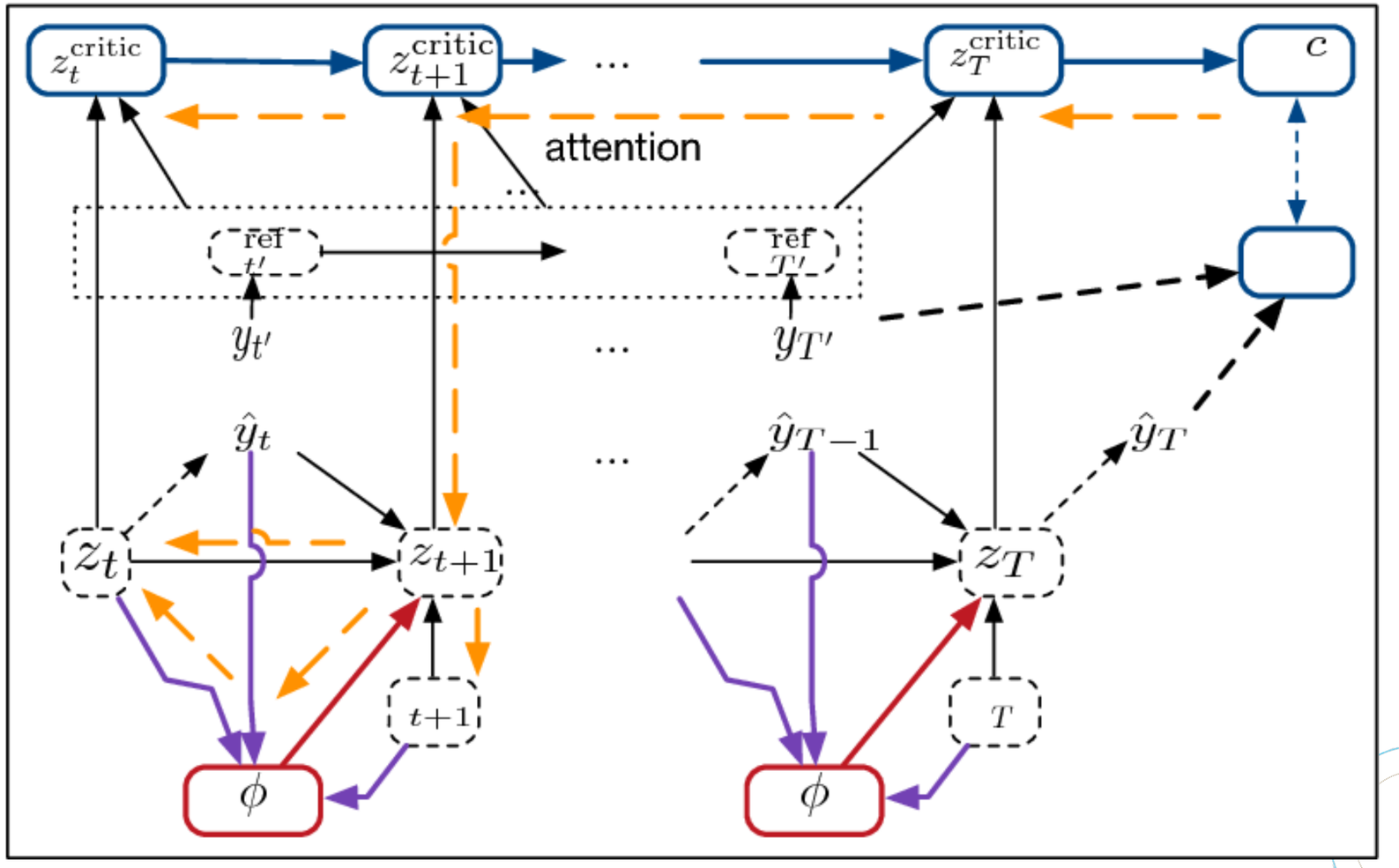


Figure 2: Illustration for improved attention model of NMT.

Trainable Greedy Decoding for Neural Machine Translation

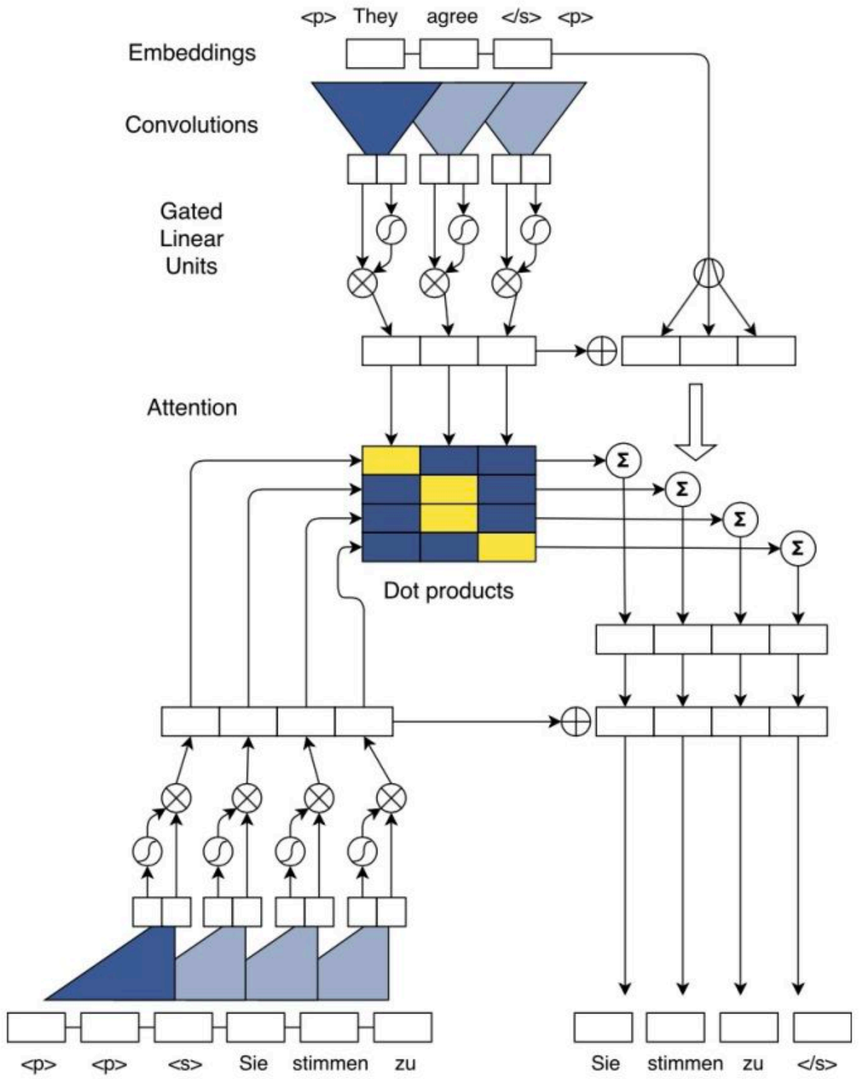
Gu, Jiatao, Kyunghyun Cho, and Victor OK Li. "Trainable greedy decoding for neural machine translation." arXiv preprint arXiv:1702.02429 (2017).





Convolutional Sequence to Sequence Learning

Gehring, Jonas, et al. "Convolutional sequence to sequence learning." arXiv preprint arXiv:1705.03122 (2017).

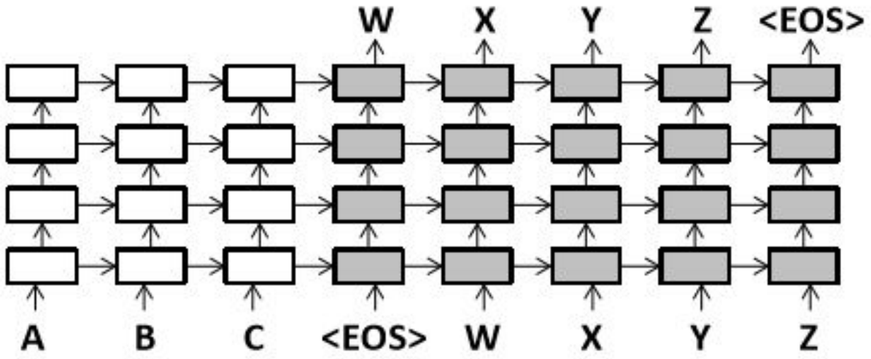


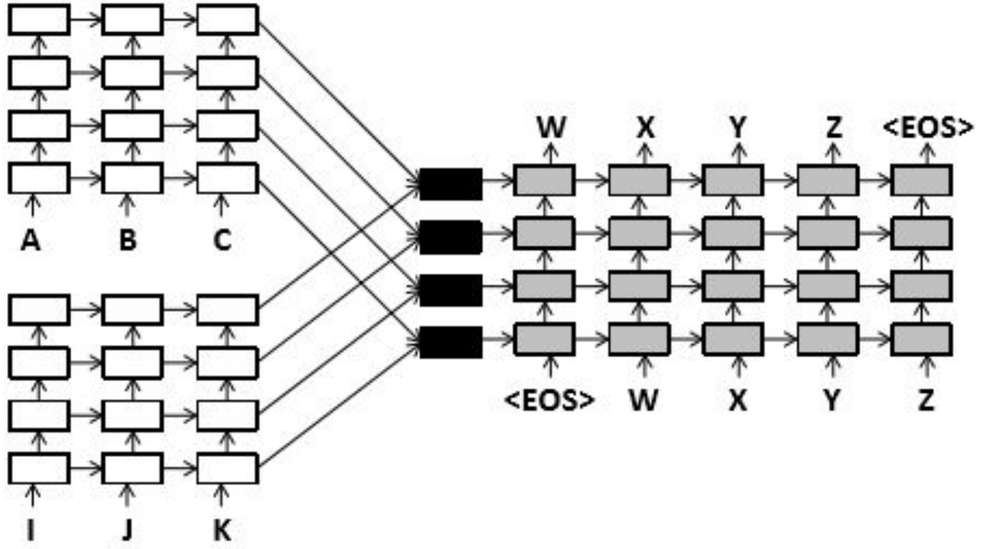
Prior Knowledge Integration for Neural Machine Translation using Posterior Regularization

Zhang, Jiacheng, et al. "Prior knowledge integration for neural machine translation using posterior regularization." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2017.

Multi-Source Neural Translation

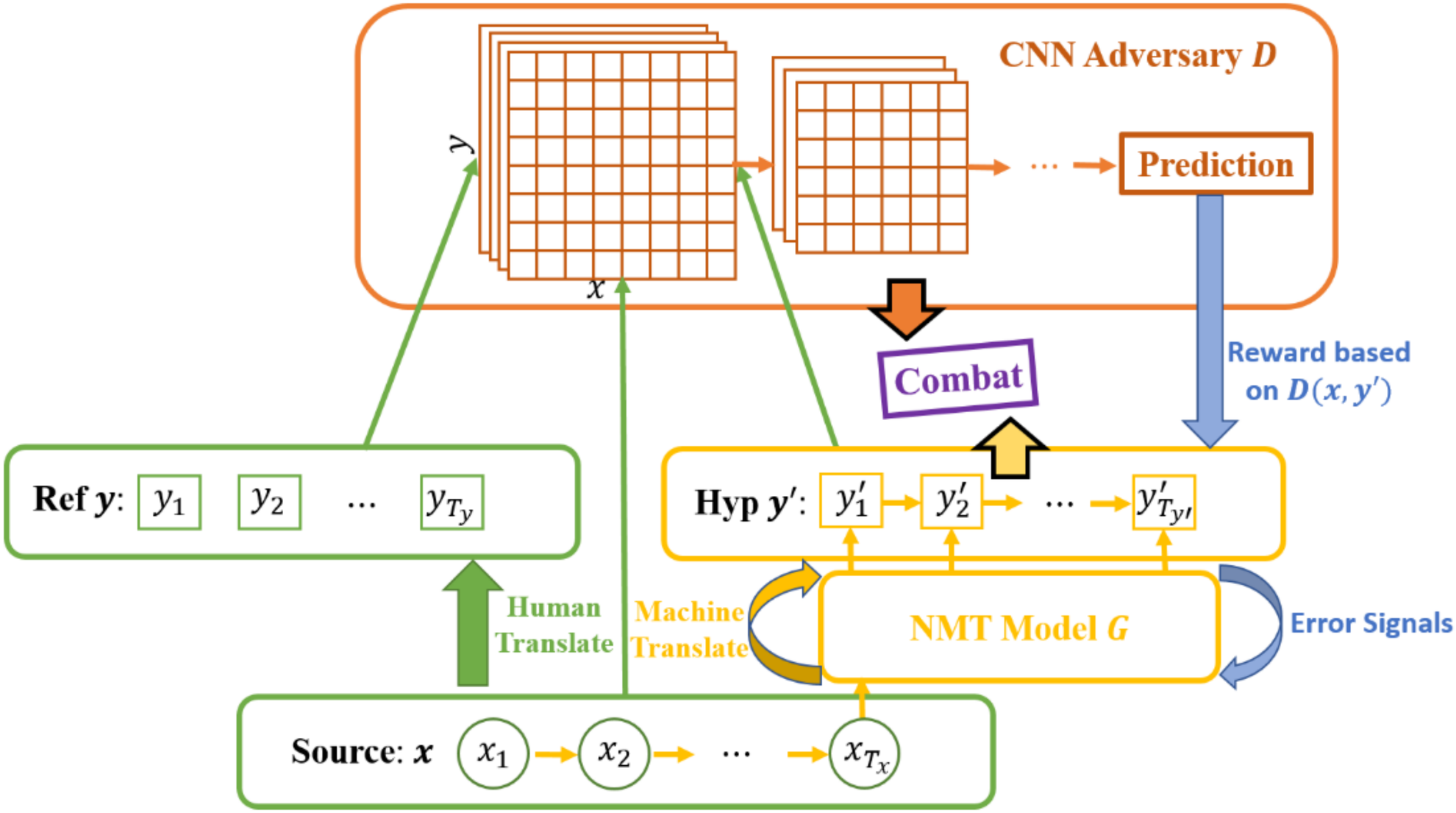
Zoph, Barret, and Kevin Knight. "Multi-source neural translation." arXiv preprint arXiv:1601.00710 (2016).





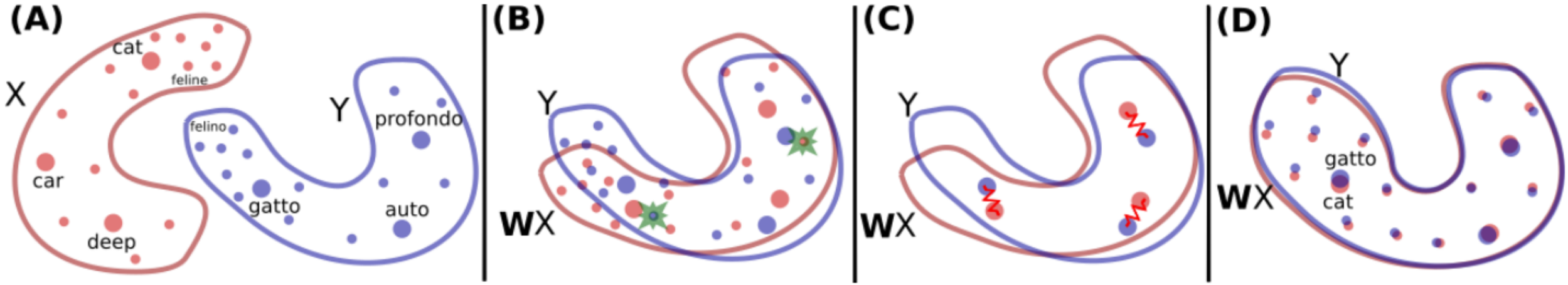
Adversarial Neural Machine Translation

Wu, Lijun, et al. "Adversarial neural machine translation." arXiv preprint arXiv:1704.06933 (2017).



Word Translation Without Parallel Data

Conneau, Alexis, et al. "Word translation without parallel data." arXiv preprint arXiv:1710.04087 (2017).



END