Google Translate
助力自然语言理解

田野
2018/4/21

自然语言和翻译

语言

是用於沟通的一套方式
有其符号与处理规则
会以视觉、声音或者触觉方式来进行传递

- 规则 ～ 文法
- 符号 ～ 文字

自然语言 ～ 人类的语言

文字语言 ～ 300种

声音语言 ～ 7000种

信、达、雅

信 ～ 准确

达 ～ 通顺

雅 ～ 得体

严复

翻译家

意大利传教士翻译家——利玛窦

英国科学家翻译家——李约瑟

唐代佛经翻译家——玄奘

西方美学翻译大师——朱光潜

国学大师翻译家——季羡林

外国文学翻译家——傅雷

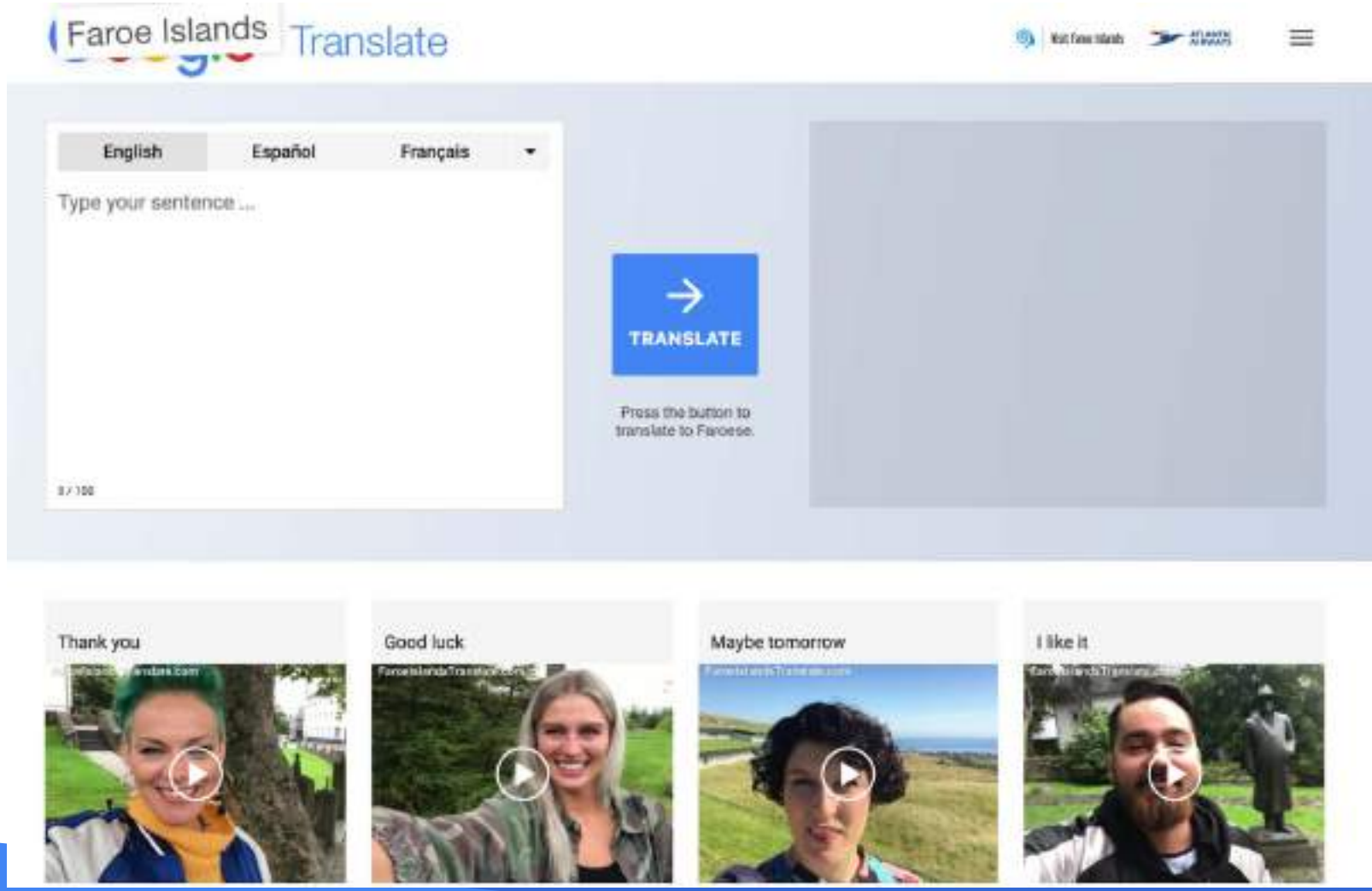
翻译官

“国翻”张璐

亦余心之所善兮，虽九死其尤未悔。

For the ideal that I hold near to my heart,
I'd not regret a thousand times to die.

众翻


Faroe Islands Translate
English
Español
Français
Type your sentence ...
0 / 100
TRANSLATE
Press the button to translate to Faroese.
Thank you
Good luck
Maybe tomorrow
I like it

机器翻译

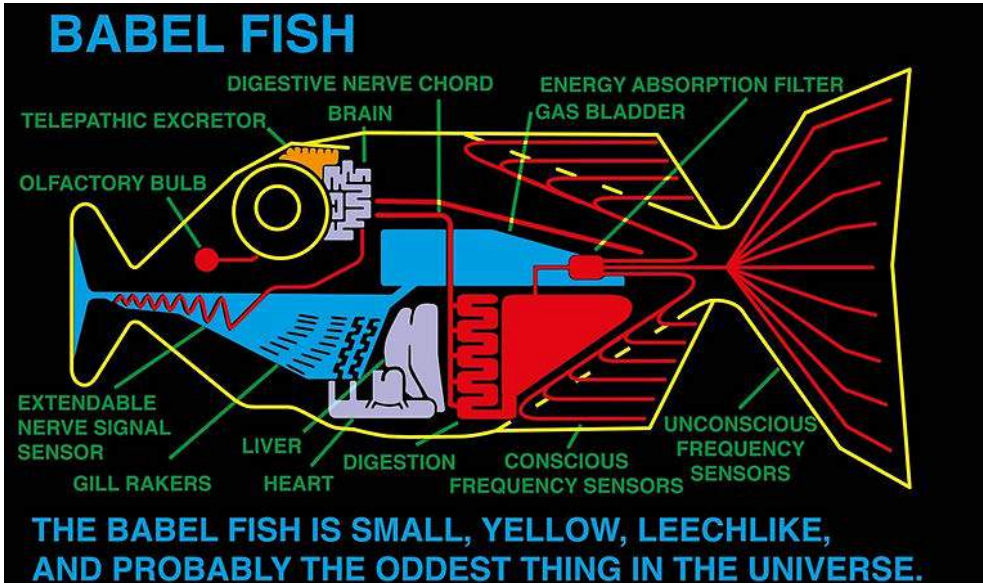
把同样的意思用另外一种语言表达出来。

语义理解

上下文理解

相关知识理解

AI Complete问题



2001 上线

50% 网络内容是英语

20% 全球人口使用过英语

103 种文字语言互译

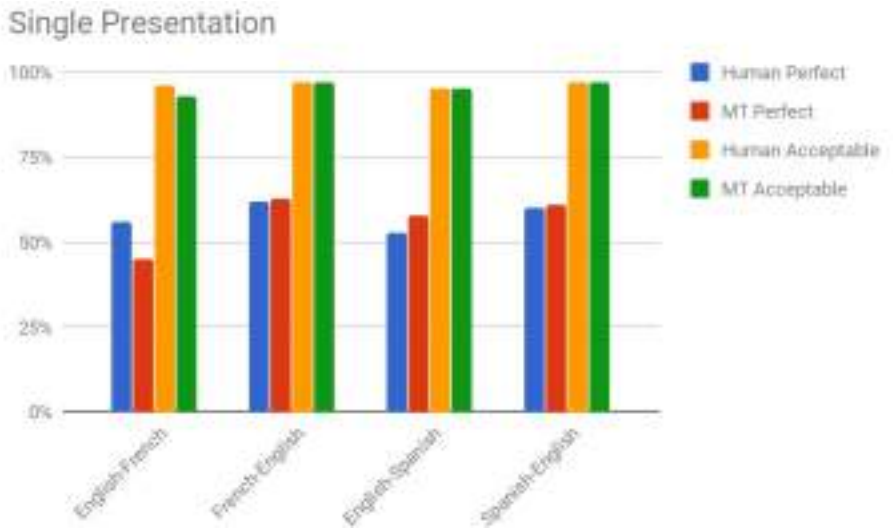
99% 网络语言覆盖

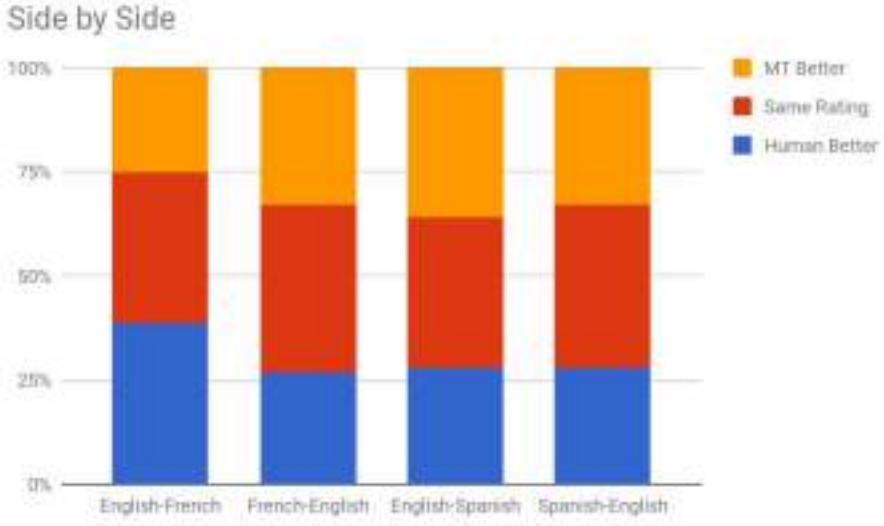
1400亿 词翻译/日

95% 使用来自美国之外

1.2% 搜索流量

5亿 月活






Google
Sign in
Translate
From: English - detected
To: Chinese (Simplified)
Translate
English
Spanish
French
Where is the Shangri La hotel?
English
Chinese (Simplified)
Chinese (Traditional)
香格里拉酒店在哪里?
New! Click the words above to edit and view alternate translations. Dismiss

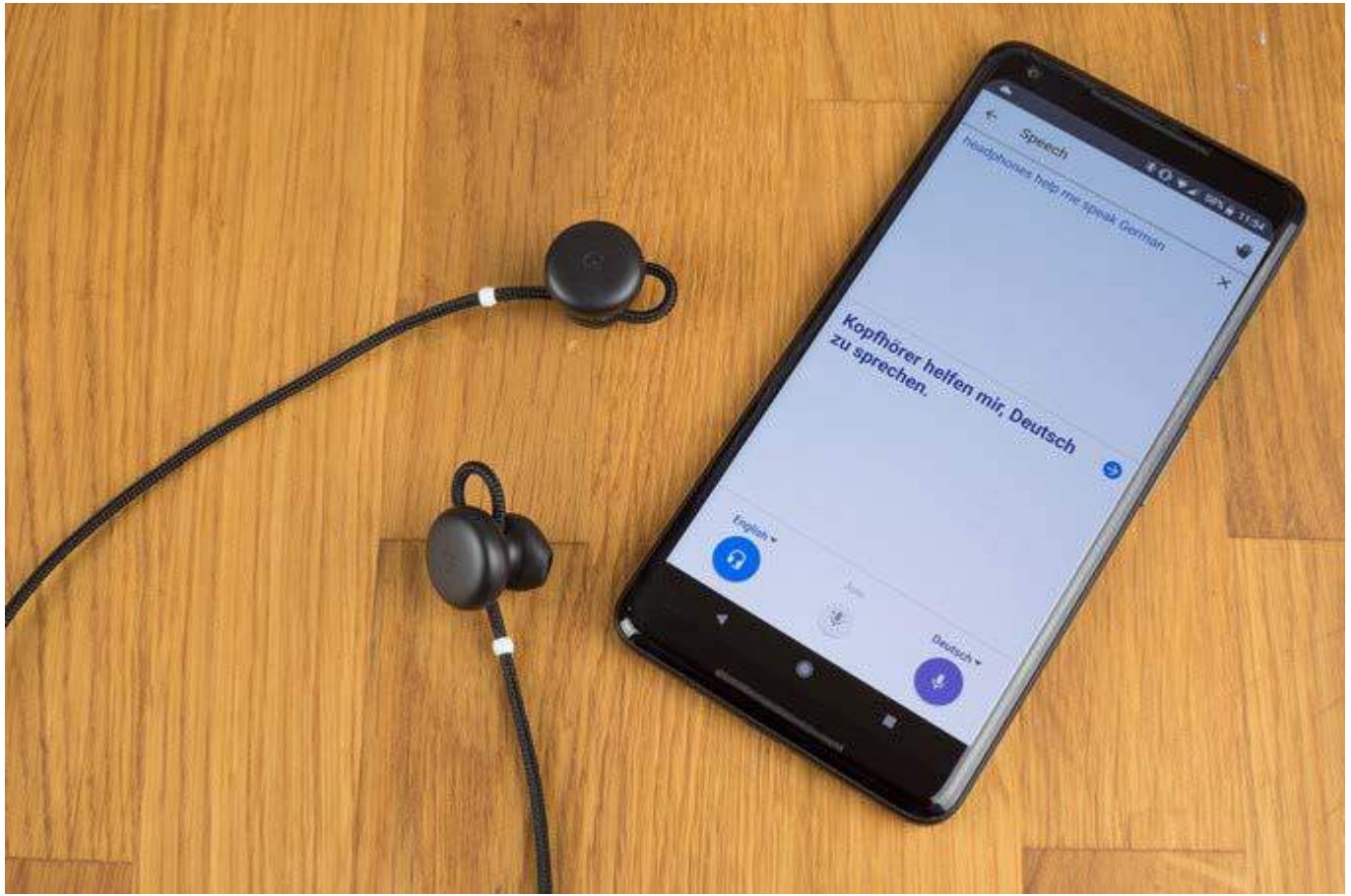

Type text or a website address or translate a document.
饺子
饺子
饺女子
饺交子
汝子
伐子
饭子
or Business:
Translator Toolkit
Website Translato

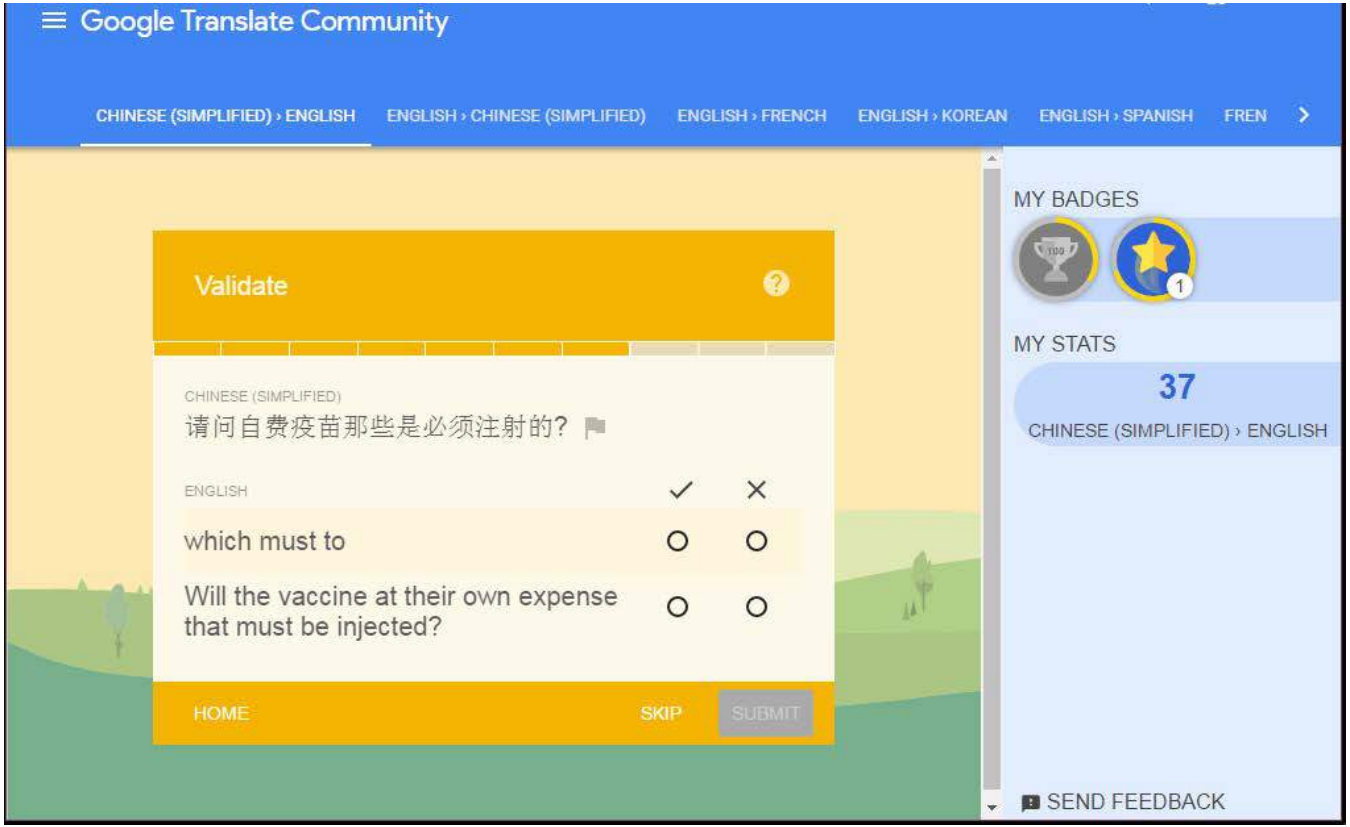

英语
中文(简体)
SORRY
CASH ONLY
很 抱歉
只收 现金


Emergency calls only
Speech
5:19 PM
tell me a story about a dog please
告诉我一个故事关于狗吧。
English
中文(简体)
HUAWEI

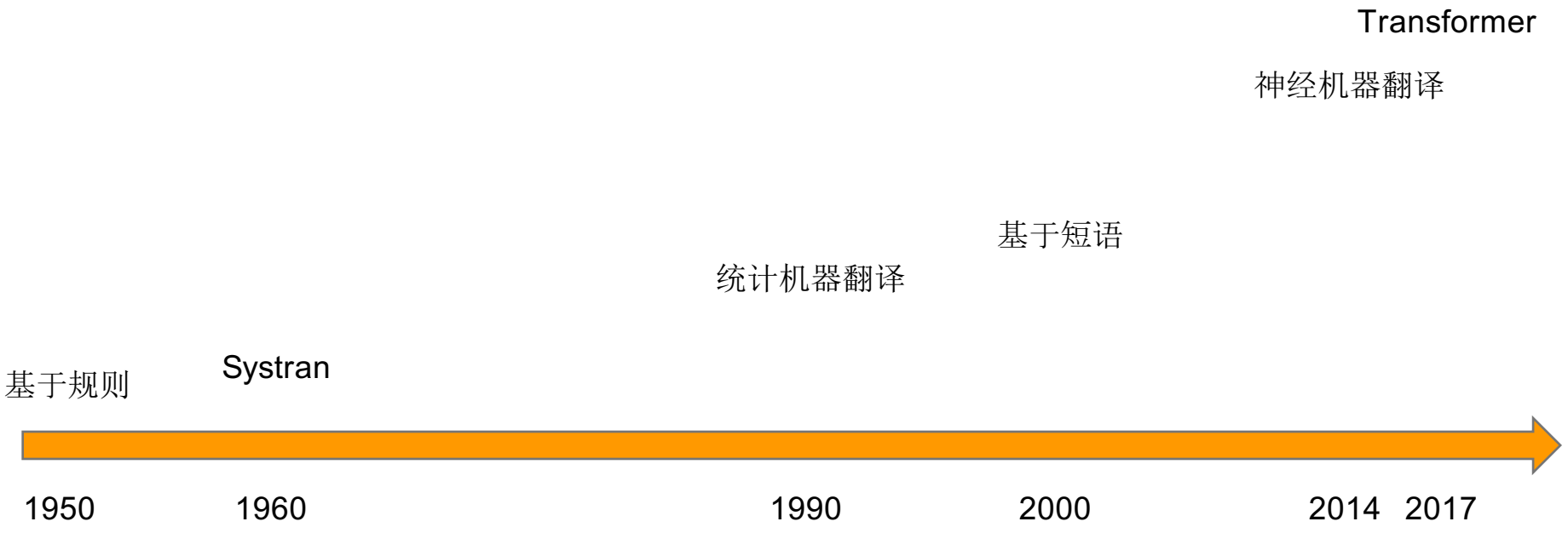

Where is the museum


博物館
在哪裡

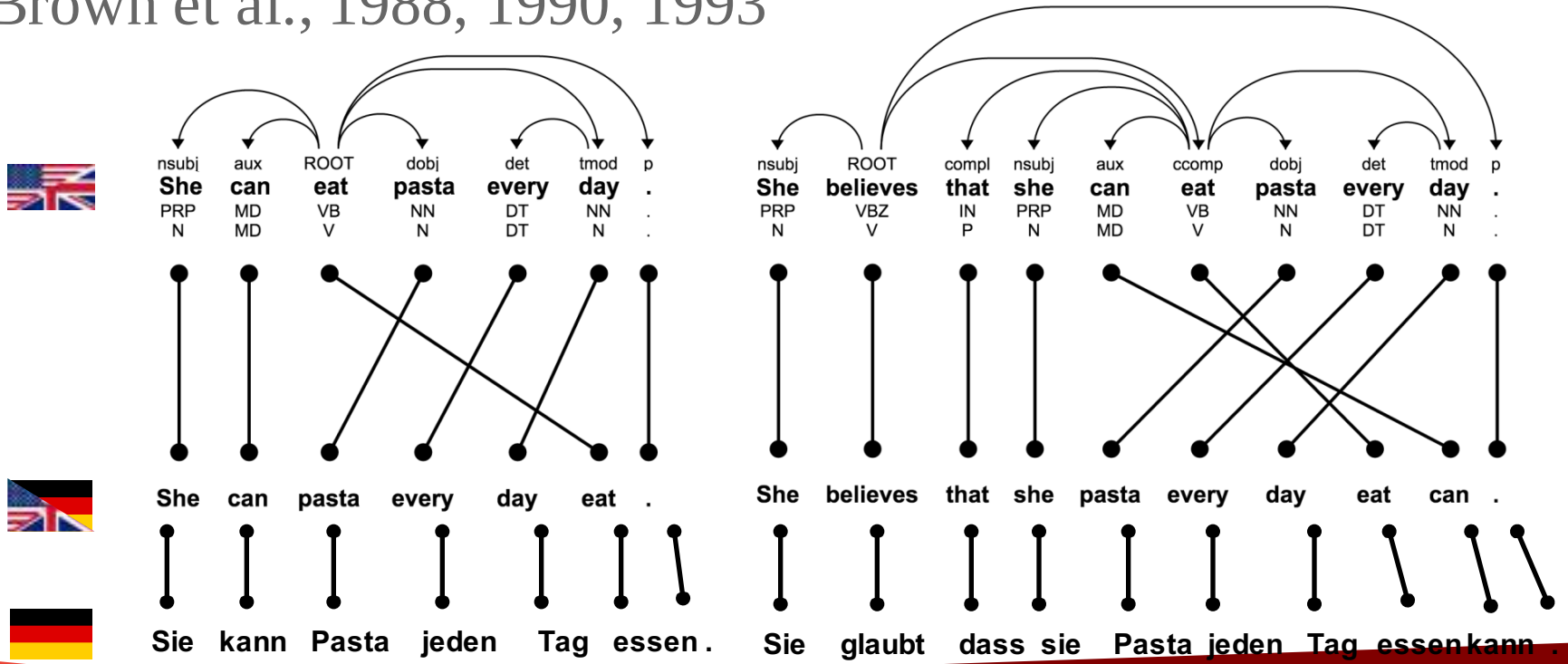

Speech
headphones help me speak German
Kopfhörer helfen mir, Deutsch zu sprechen.
English
Deutsch


Google Translate Community
CHINESE (SIMPLIFIED) › ENGLISH
ENGLISH › CHINESE (SIMPLIFIED)
ENGLISH › FRENCH
ENGLISH › KOREAN
ENGLISH › SPANISH
FREN
Validate
CHINESE (SIMPLIFIED)
请问自费疫苗那些是必须注射的?
ENGLISH
which must to
Will the vaccine at their own expense that must be injected?
HOME
SKIP
SUBMIT
MY BADGES
MY STATS
37
CHINESE (SIMPLIFIED) › ENGLISH
SEND FEEDBACK

机器翻译极简史


Transformer
神经机器翻译
基于短语
统计机器翻译
基于规则
Systran
1950
1960
1990
2000
2014
2017

统计机器翻译
IBM models
Brown et al., 1988, 1990, 1993



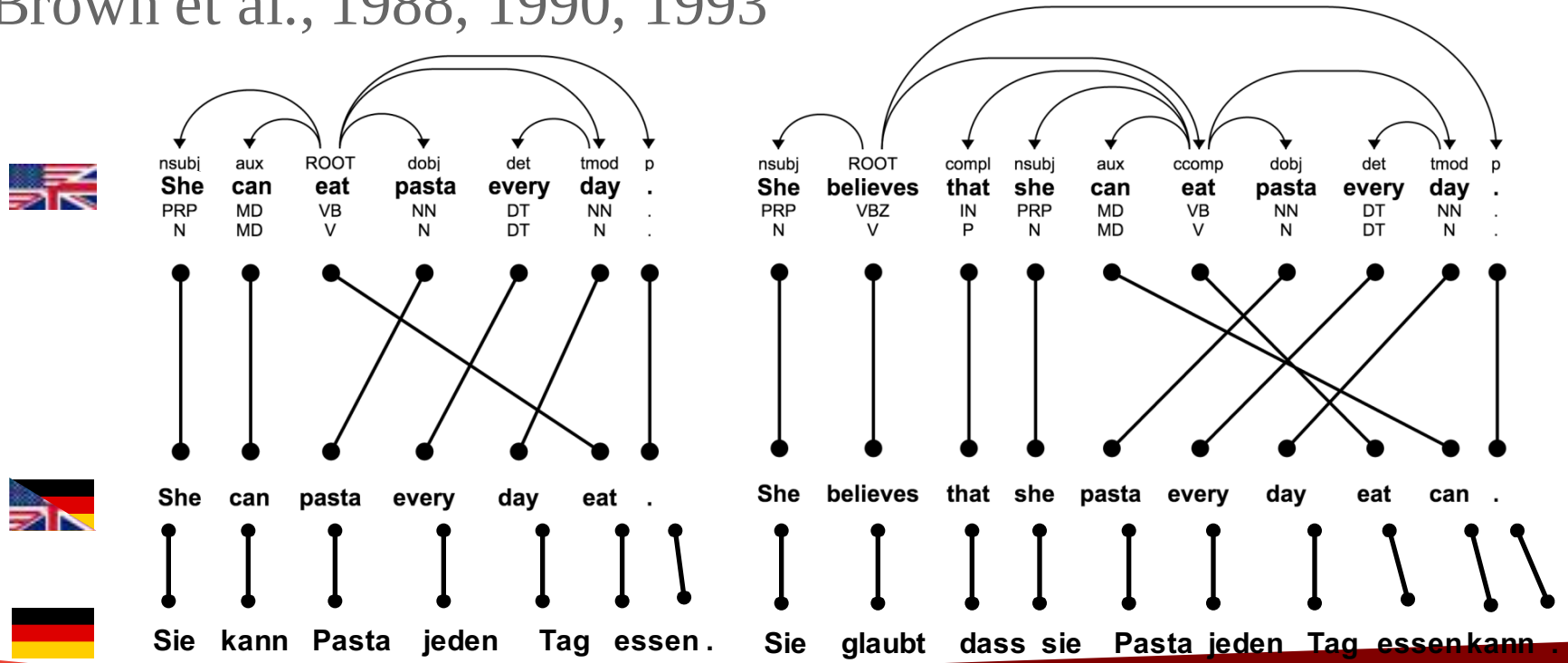
统计机器翻译

IBM models

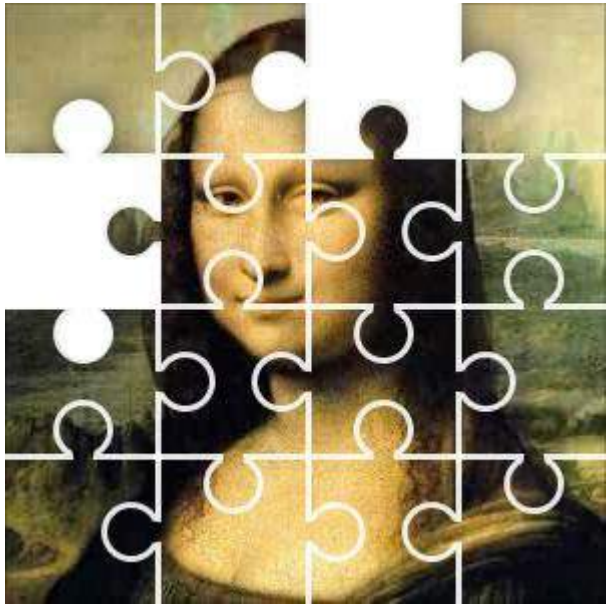
Brown et al., 1988, 1990, 1993

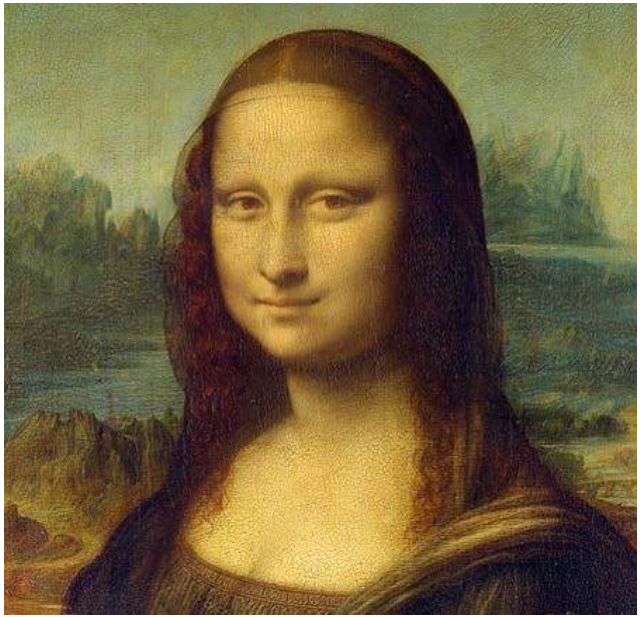
"Every time I fire a linguist, the performance of the speech recognizer goes up".

--Frederick Jelinek



统计方法到深度学习



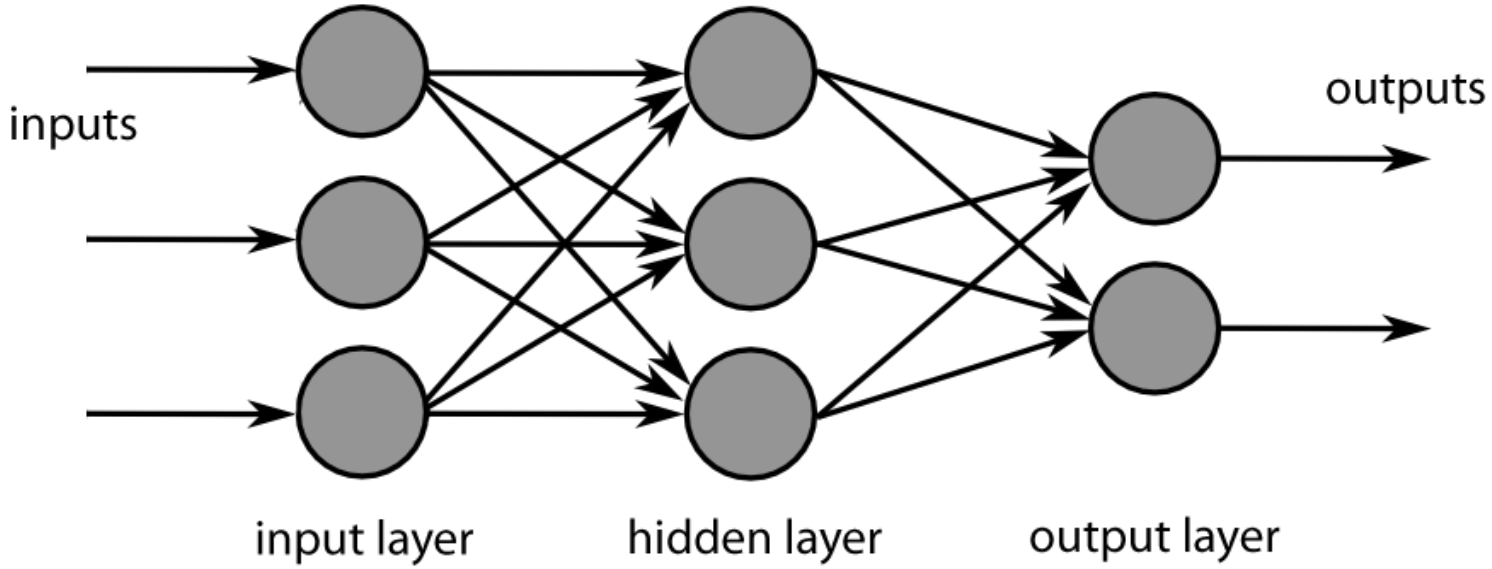


前向传播网络

尺寸固定

无状态

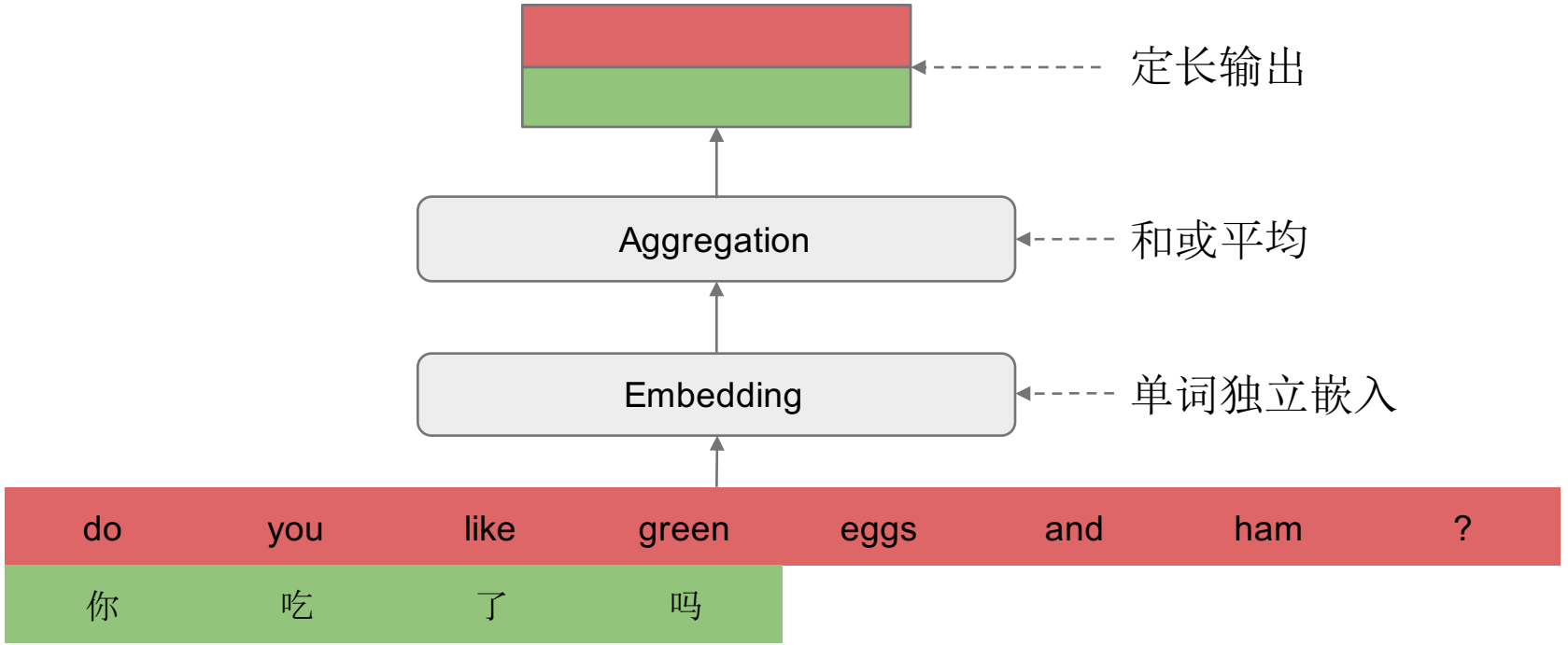
输入无循序



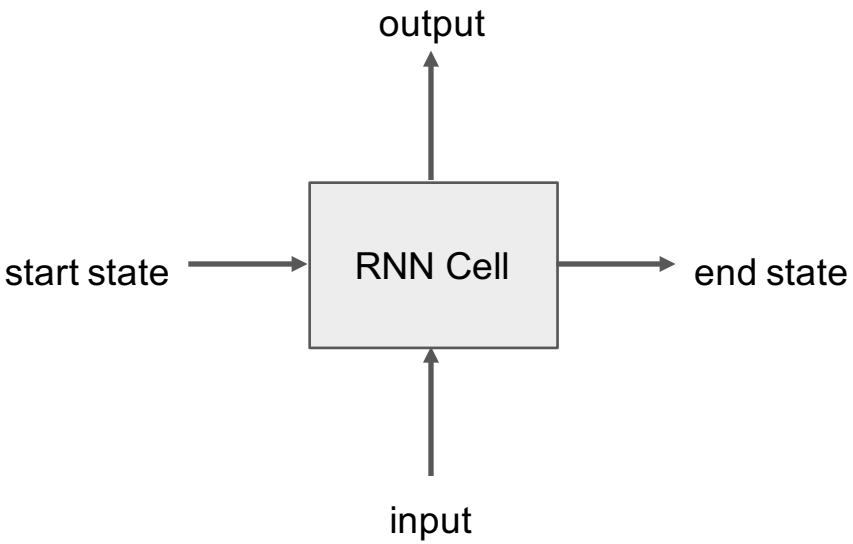
语言长度不固定，有顺序

do	you	like	green	eggs	and	ham	?
你	吃	了	吗				

语言输入在前向传播网络：bag of words



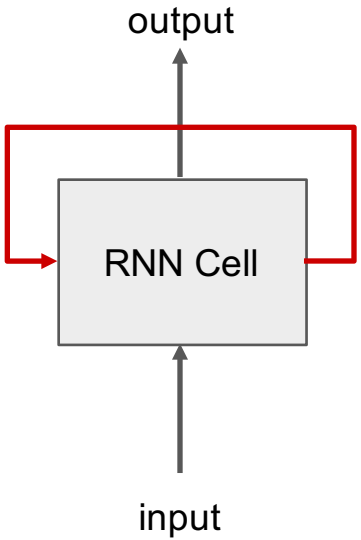
For循环




```
output, end_state =
  rnn_cell(input, start_state)
```

start state和end state尺寸相同

For循环

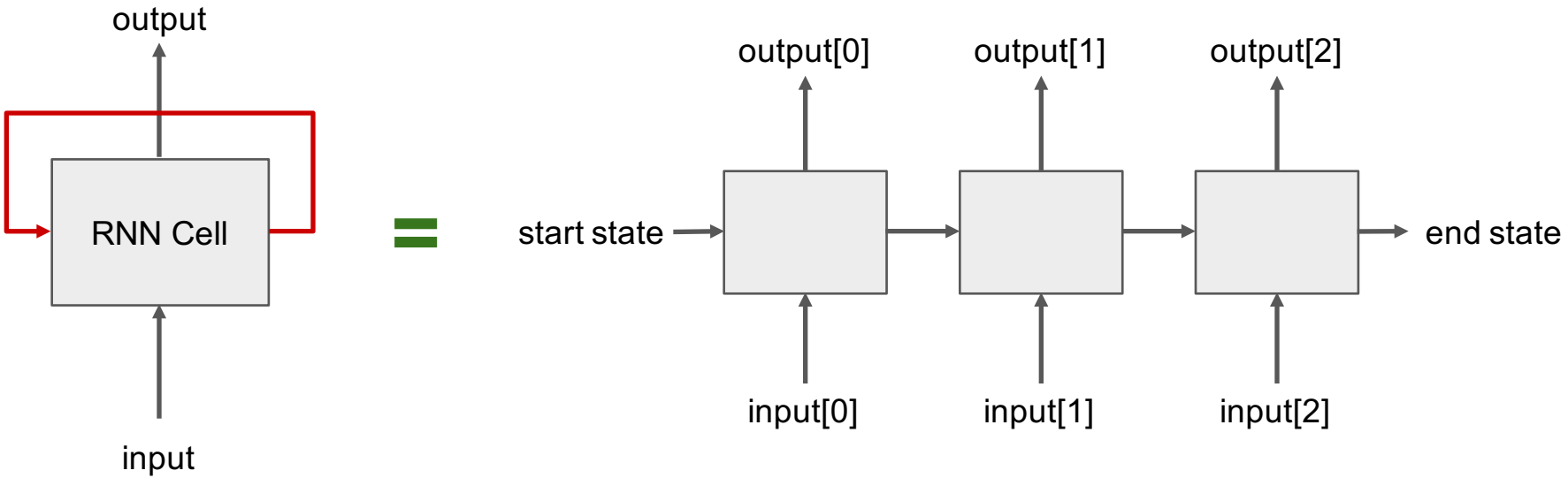



```
for i in range(sequence_length):

  output[i], end_state =
    rnn_cell(input[i], start_state)

  start_state = end_state
```

For循环展开

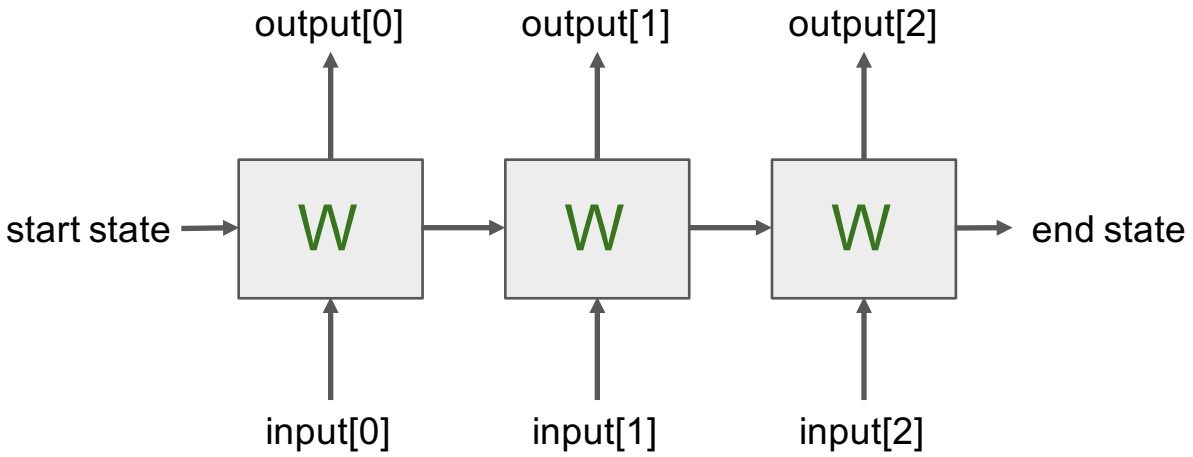

output
RNN Cell
input
=
output[0]
output[1]
output[2]
start state
end state
input[0]
input[1]
input[2]

Recurrent Neural Network

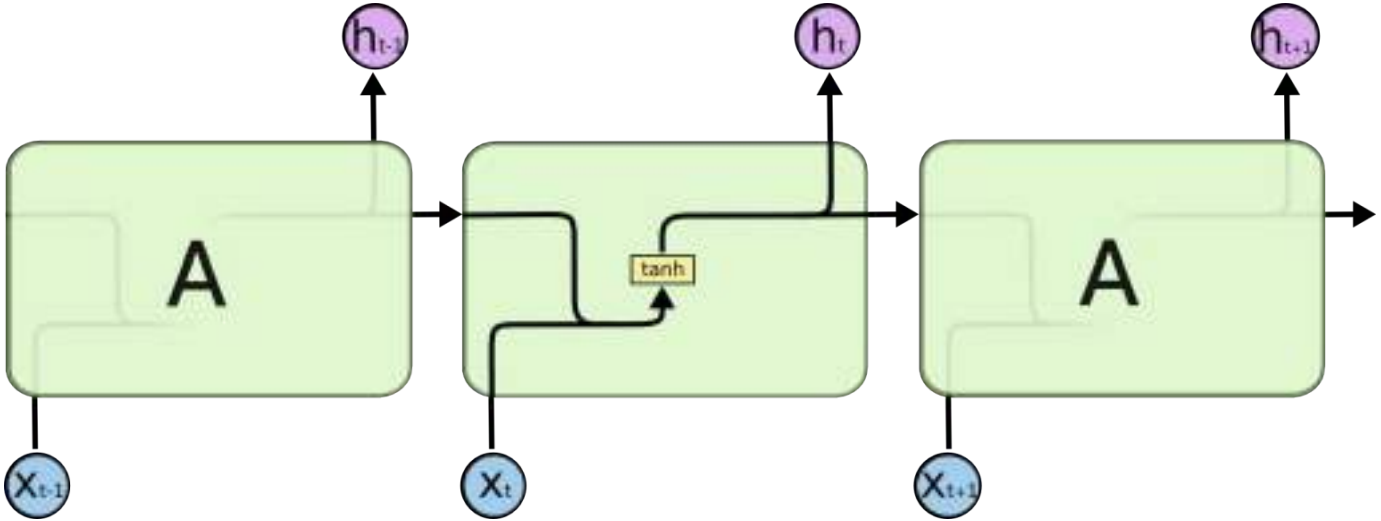
每个单元

结构完全一样

共享相同的权重



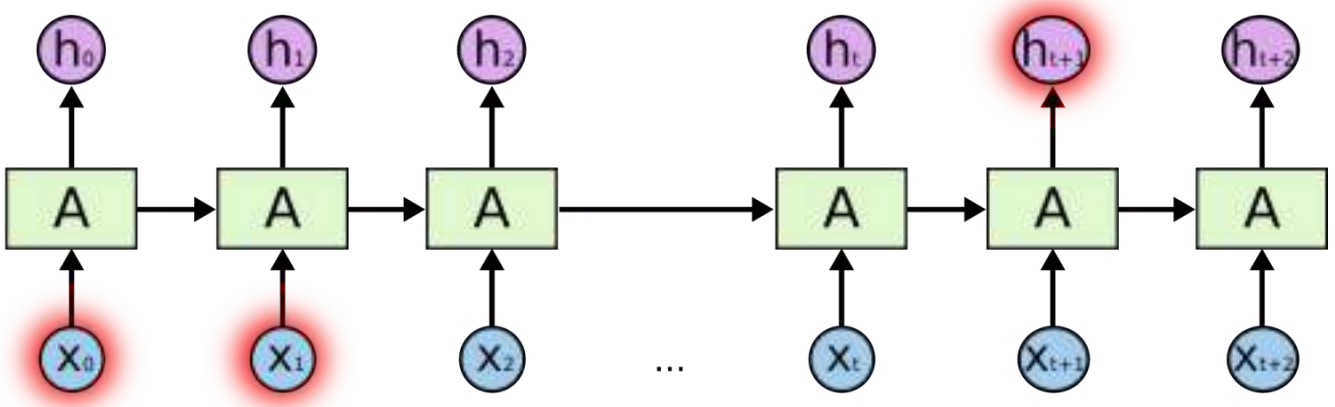
简单RNN单元



$$\mathrm{h}_t = \tanh\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right)$$

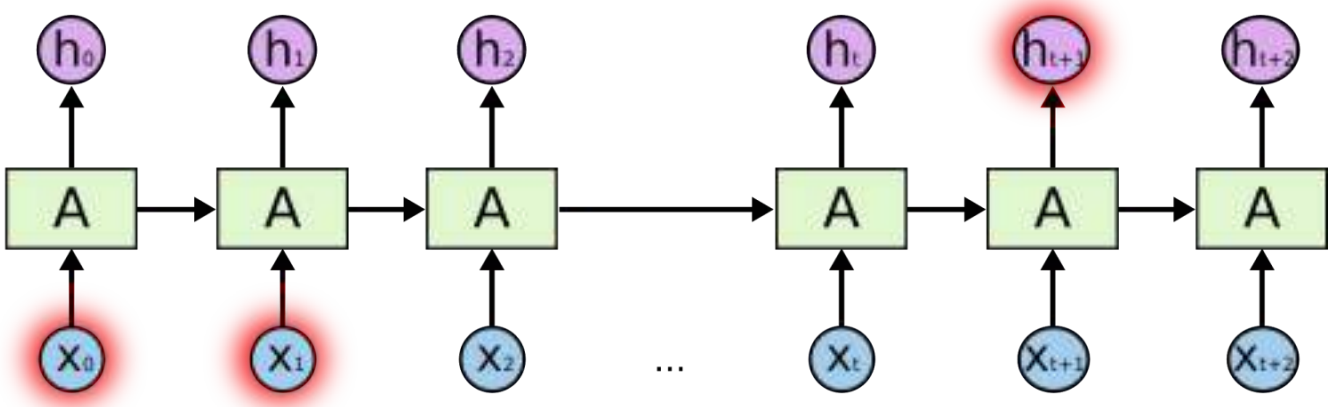
colah's blog

简单RNN单元



colah's blog

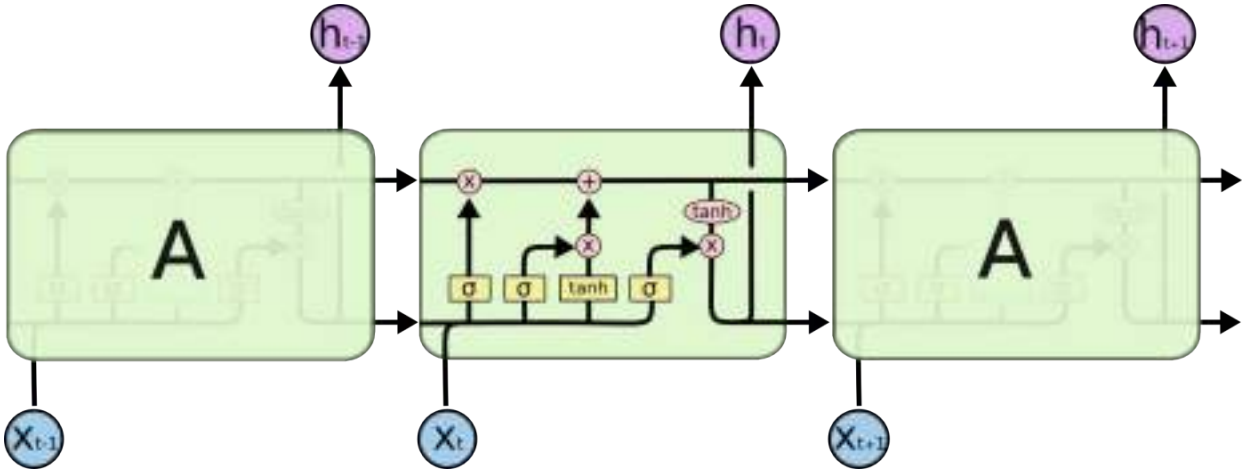
梯度消失/梯度爆炸



长距离节点的梯度会指数减小/增大

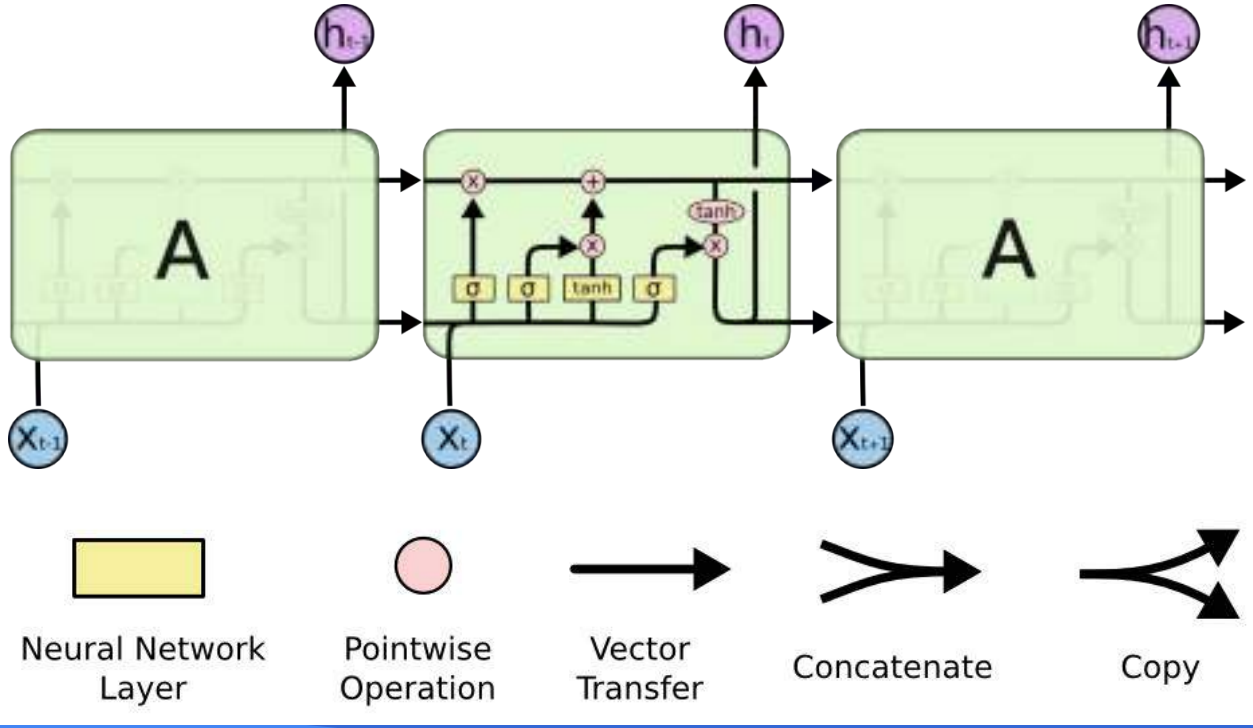
colah's blog

Long Short Term Memory单元



colah's blog

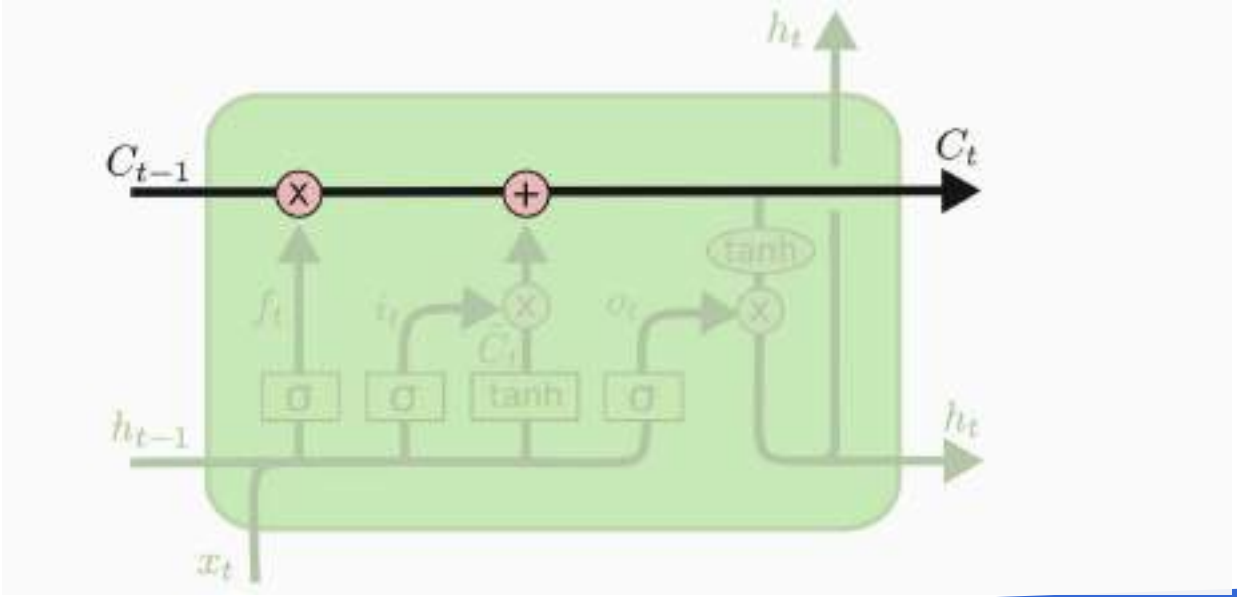
Long Short Term Memory单元



colah's blog

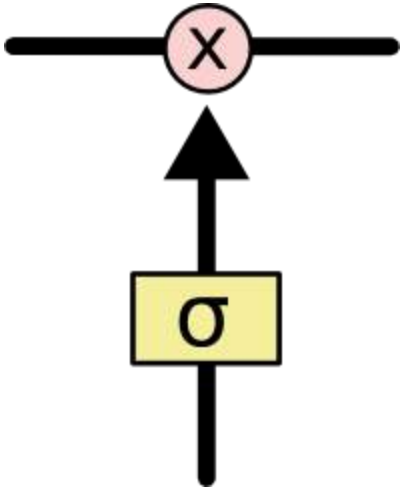
LSTM单元状态

传送带
在传送带上增减信息



colah's blog

门：信息开关

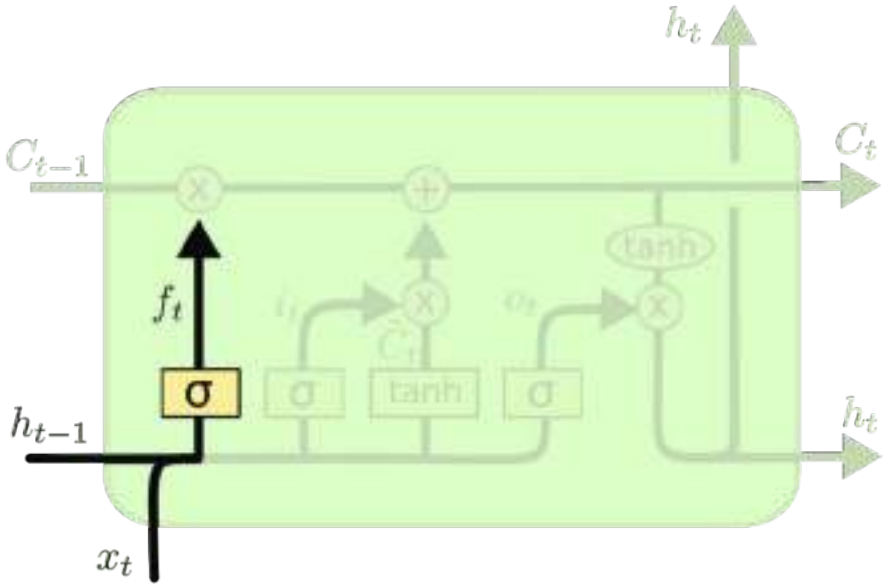


- Element-wise sigmoid
- Element-wise multiplication

$\sigma(\cdot) \in [0, 1]$

colah's blog

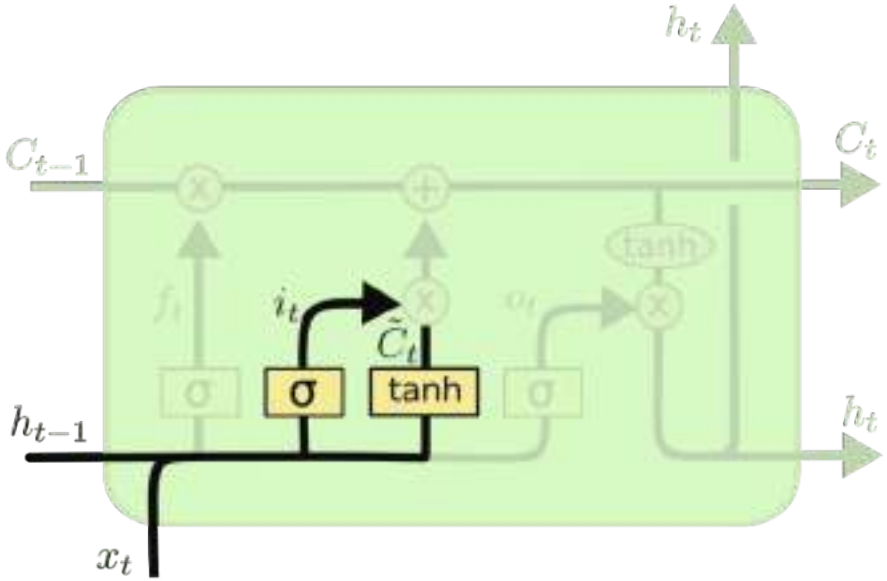
遗忘门



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

colah's blog

输入门/候选单元状态

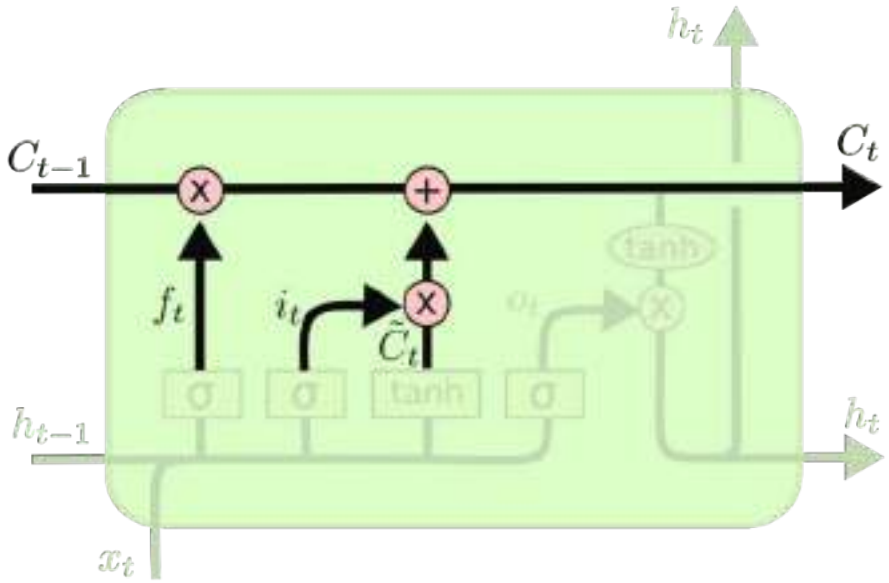


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

colah's blog

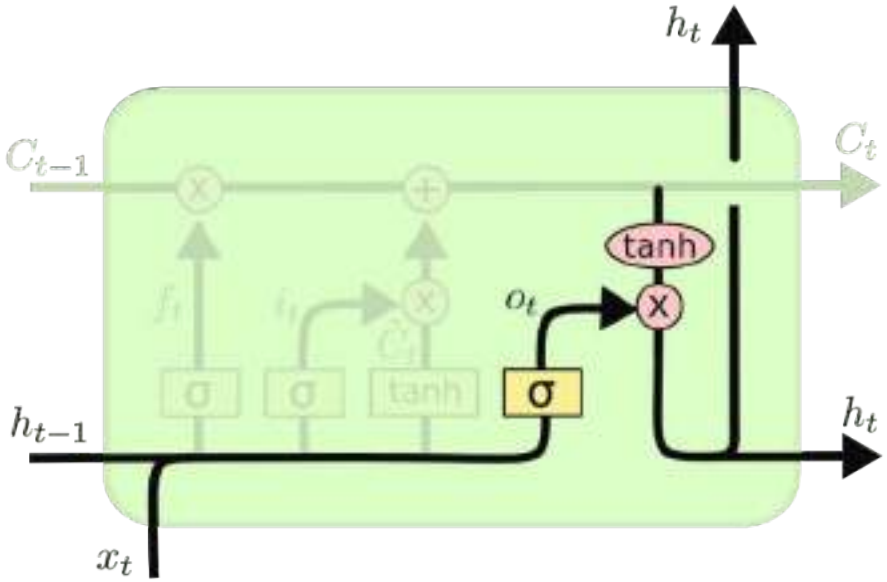
更新单元状态



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

colah's blog

输出门



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

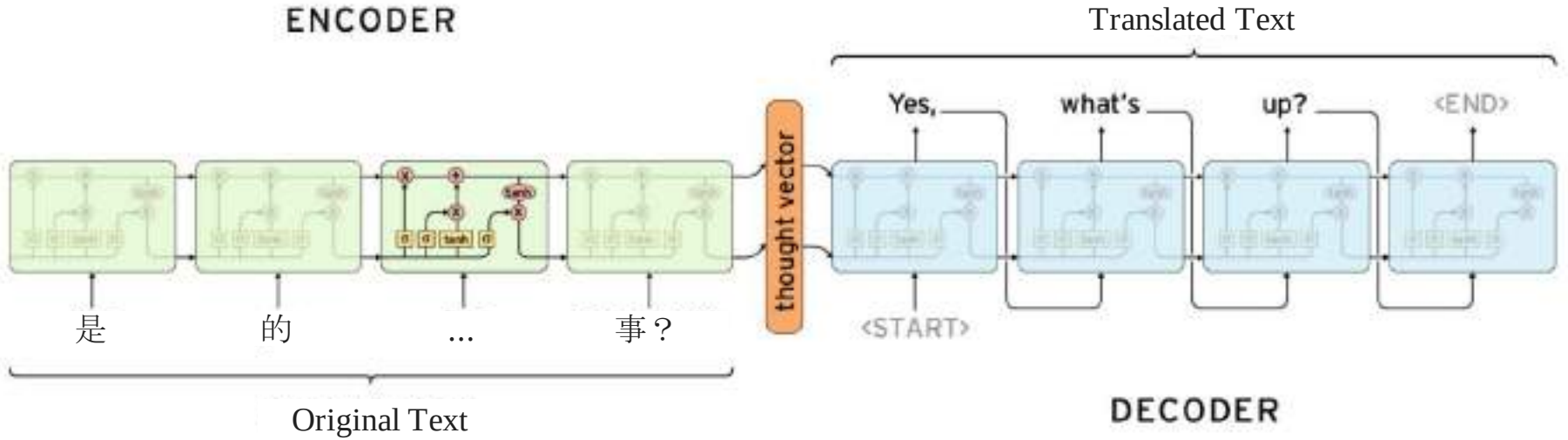
$$h_t = o_t * \tanh\left(C_t\right)$$

- Output filtered version of cell state

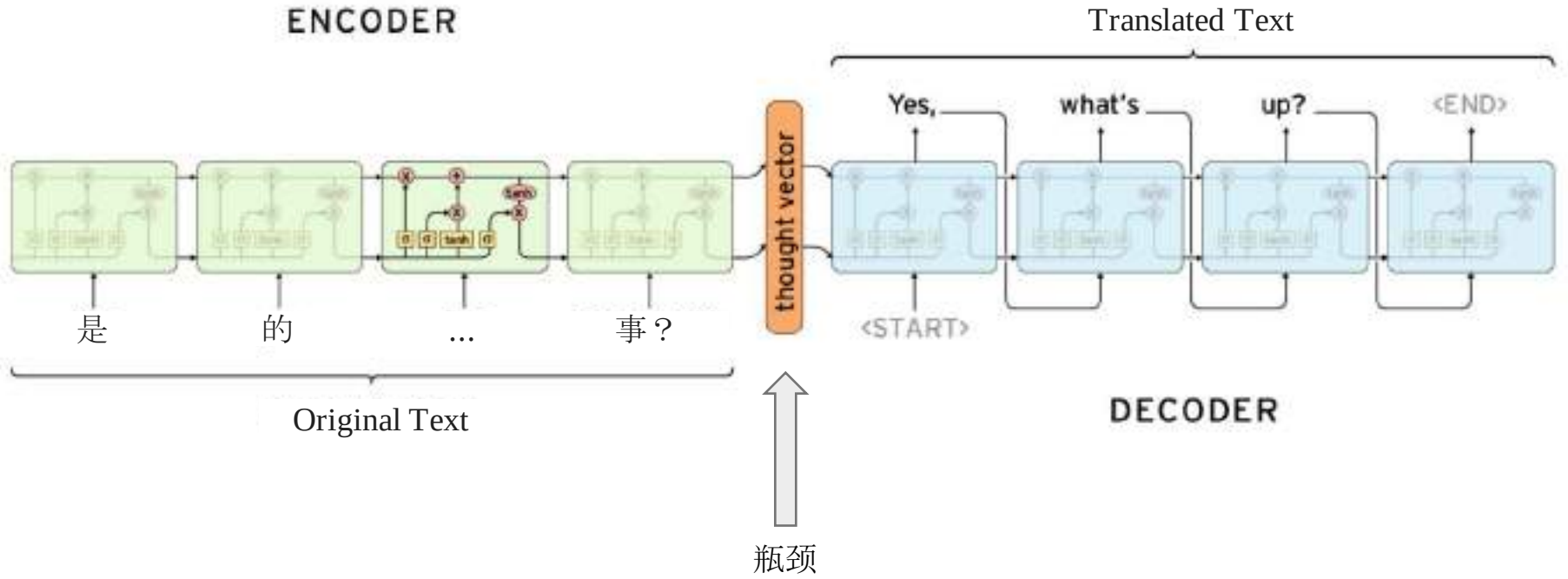
$\tanh(\cdot) \in [-1, 1]$

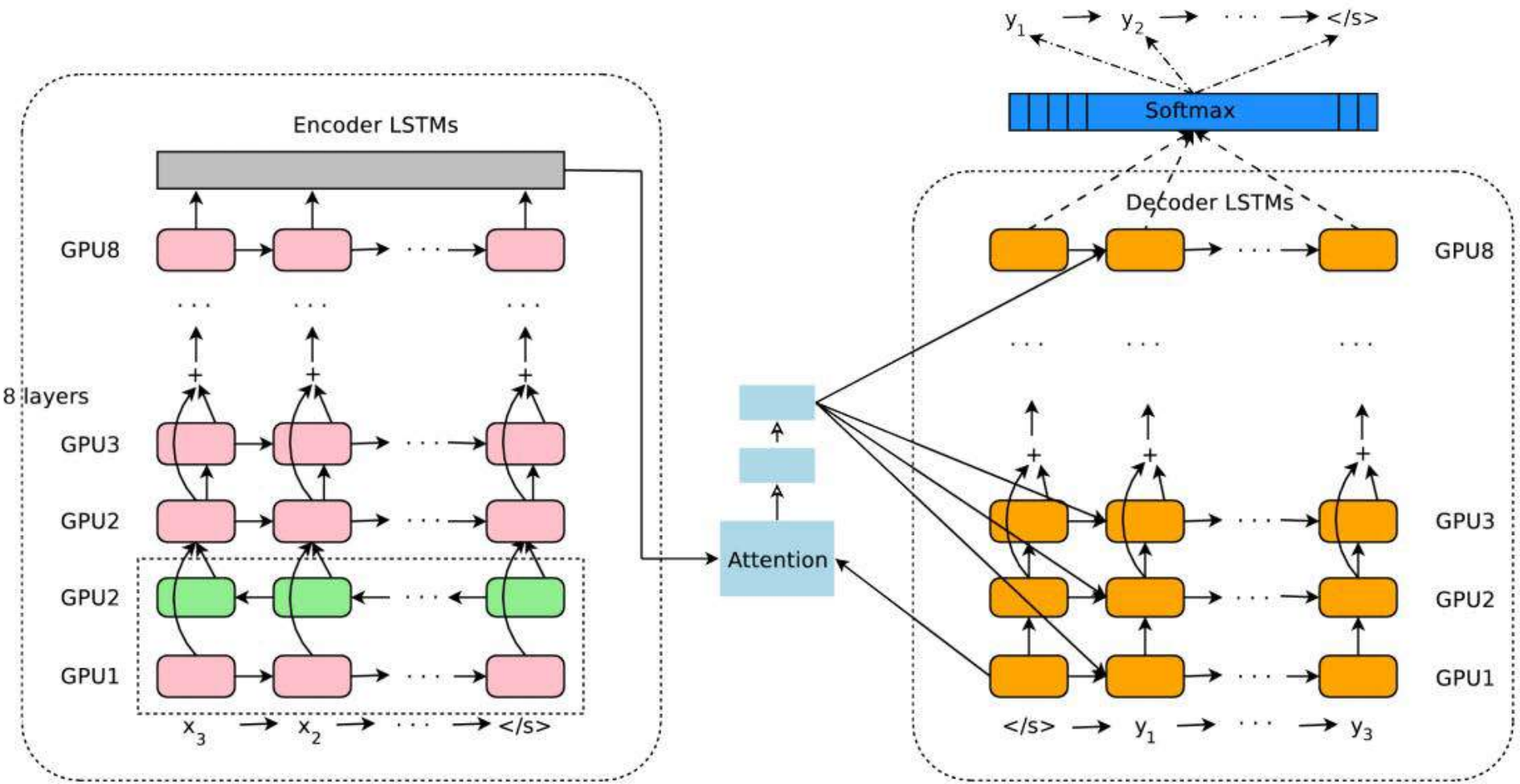
colah's blog

LSTM翻译原型

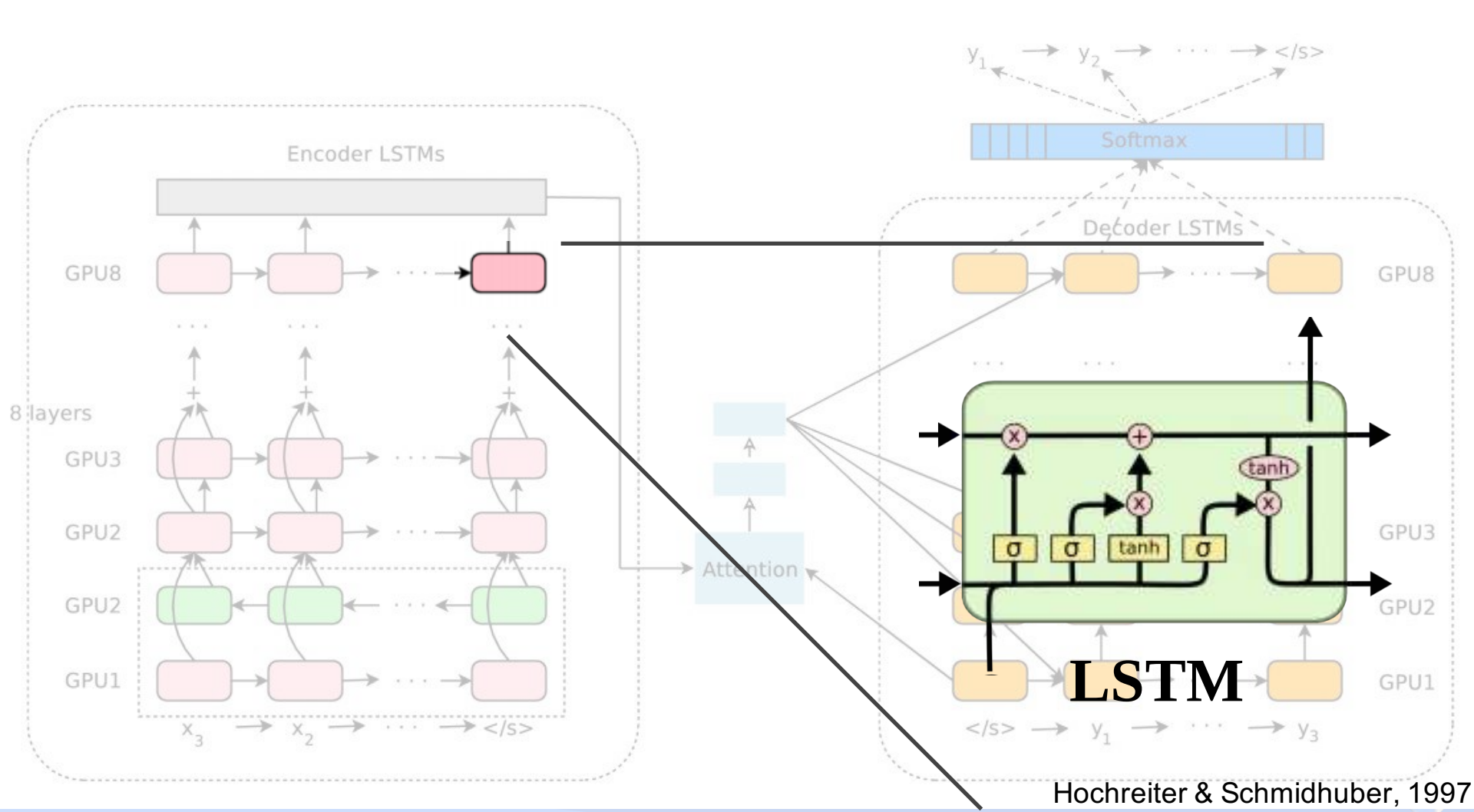

ENCODER
Translated Text
Yes,
what's
up?
<END>
thought vector
是
的
...
事？
<START>
Original Text
DECODER

LSTM翻译原型

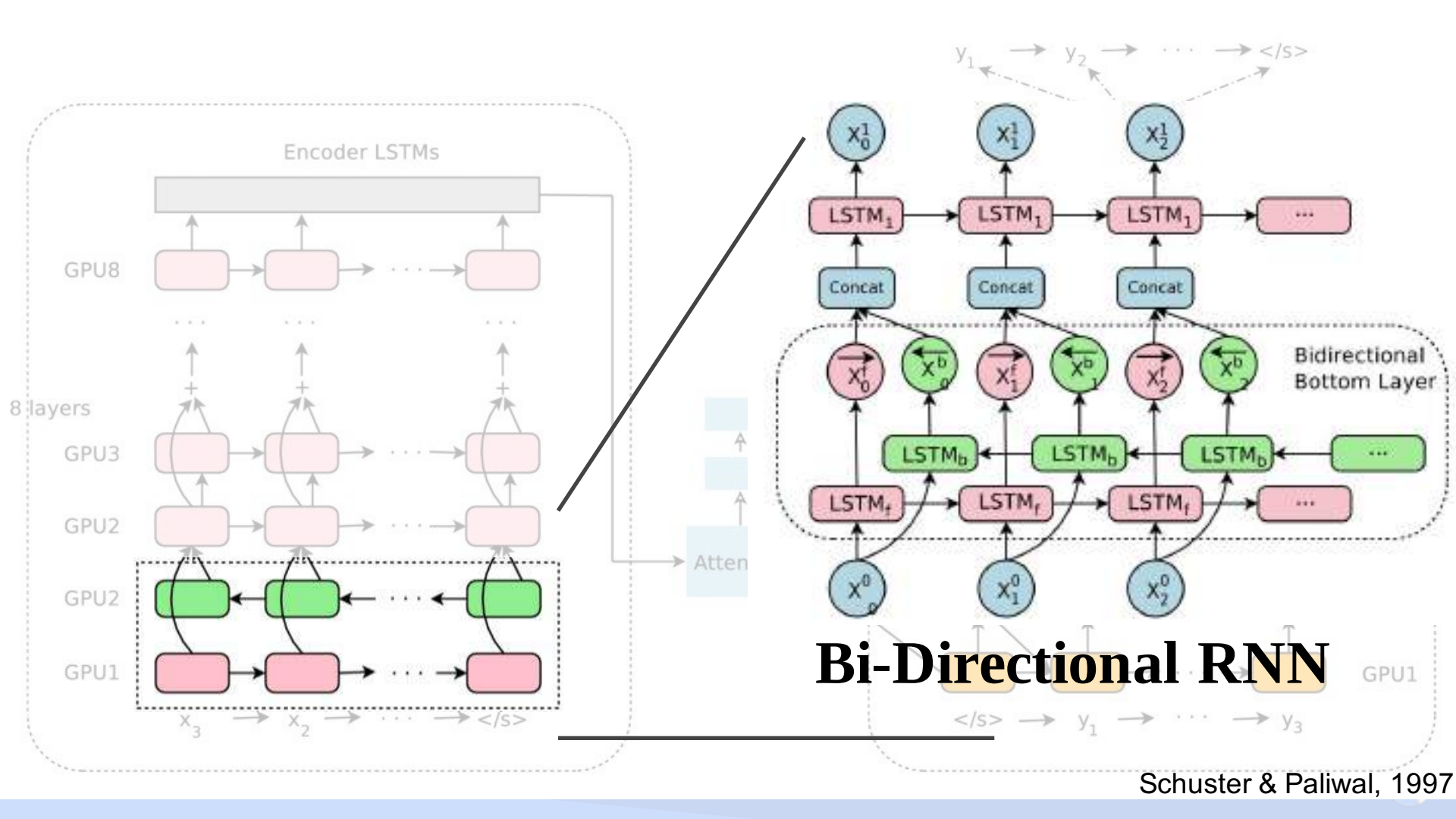

ENCODER
Translated Text
Yes,
what's
up?
<END>
thought vector
是
的
...
事？
<START>
Original Text
DECODER
瓶颈


Encoder LSTMs
GPU8
8 layers
GPU3
GPU2
GPU2
GPU1
x3
x2
</s>
Attention
y1
y2
</s>
Softmax
Decoder LSTMs
GPU8
GPU3
GPU2
GPU1
</s>
y1
y3

Wu et al, 2016

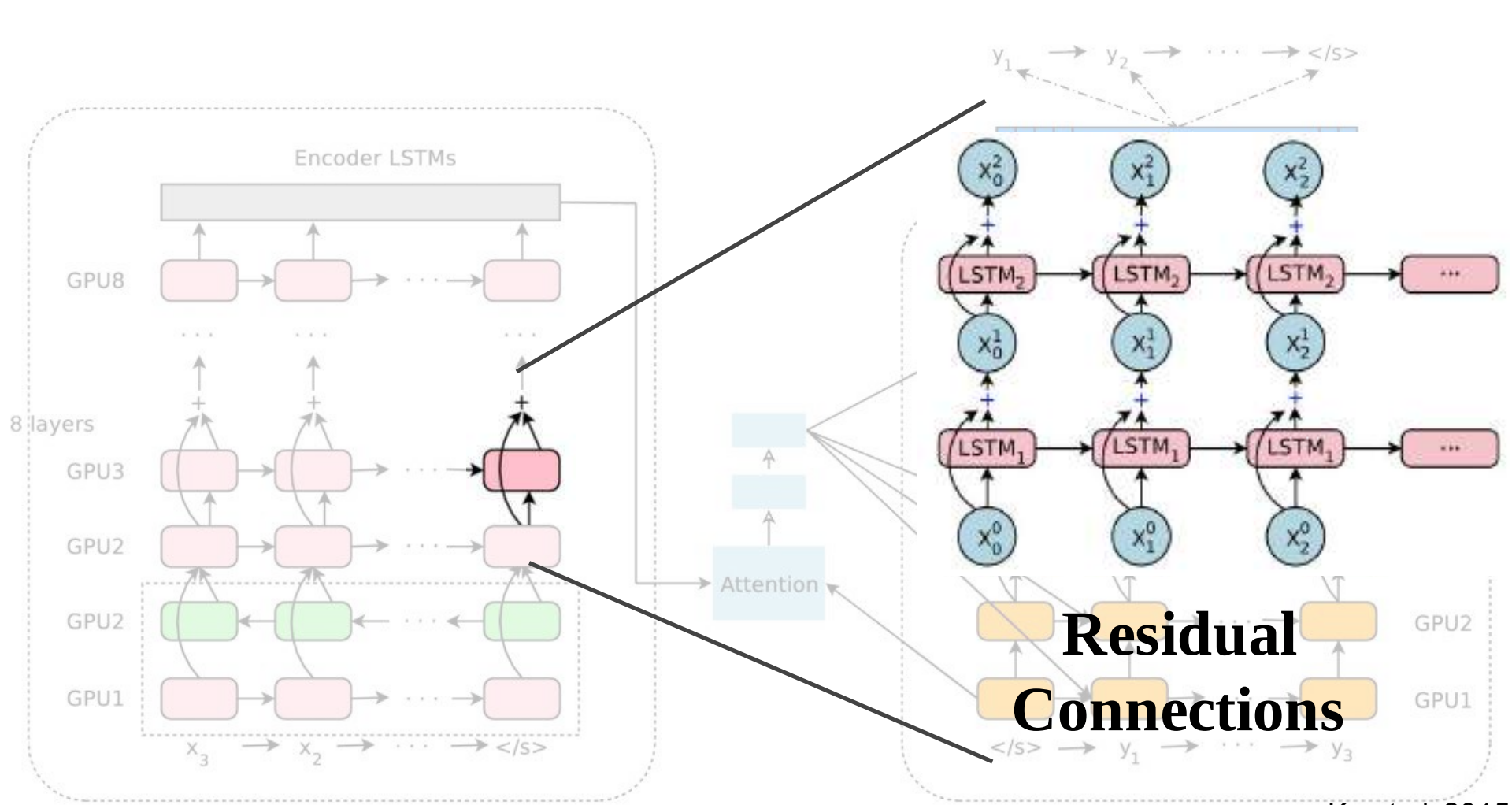


Hochreiter & Schmidhuber, 1997



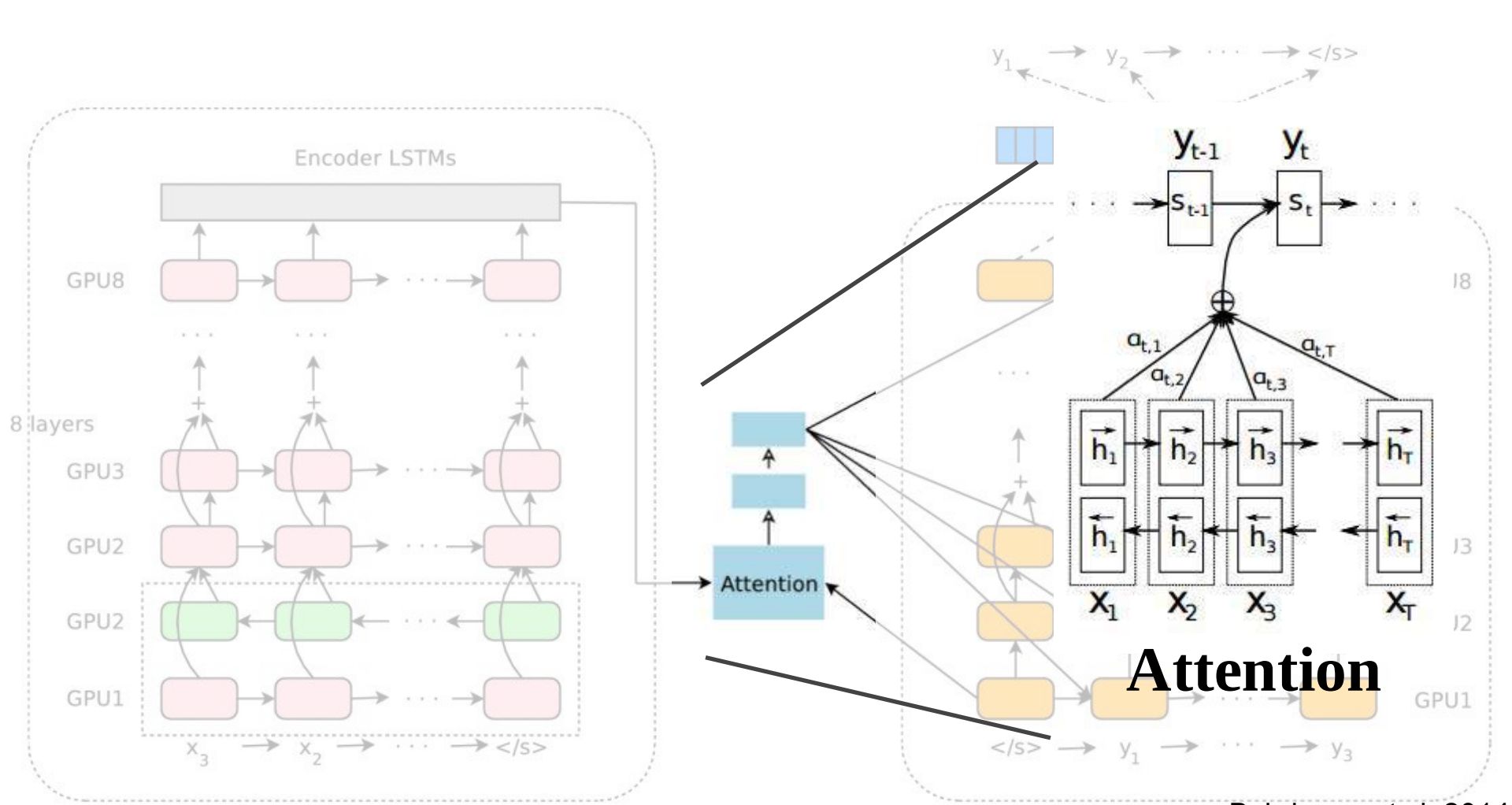
Bi-Directional RNN

Schuster & Paliwal, 1997



Residual Connections

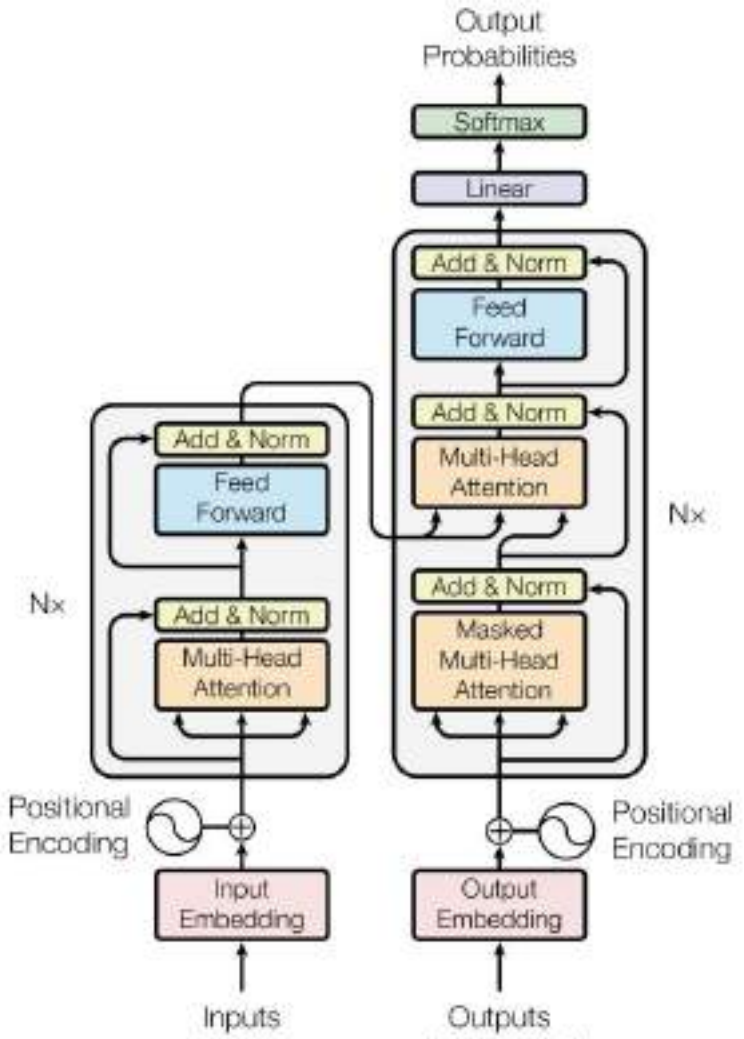
Ke et al, 2015



Bahdanau et al, 2014

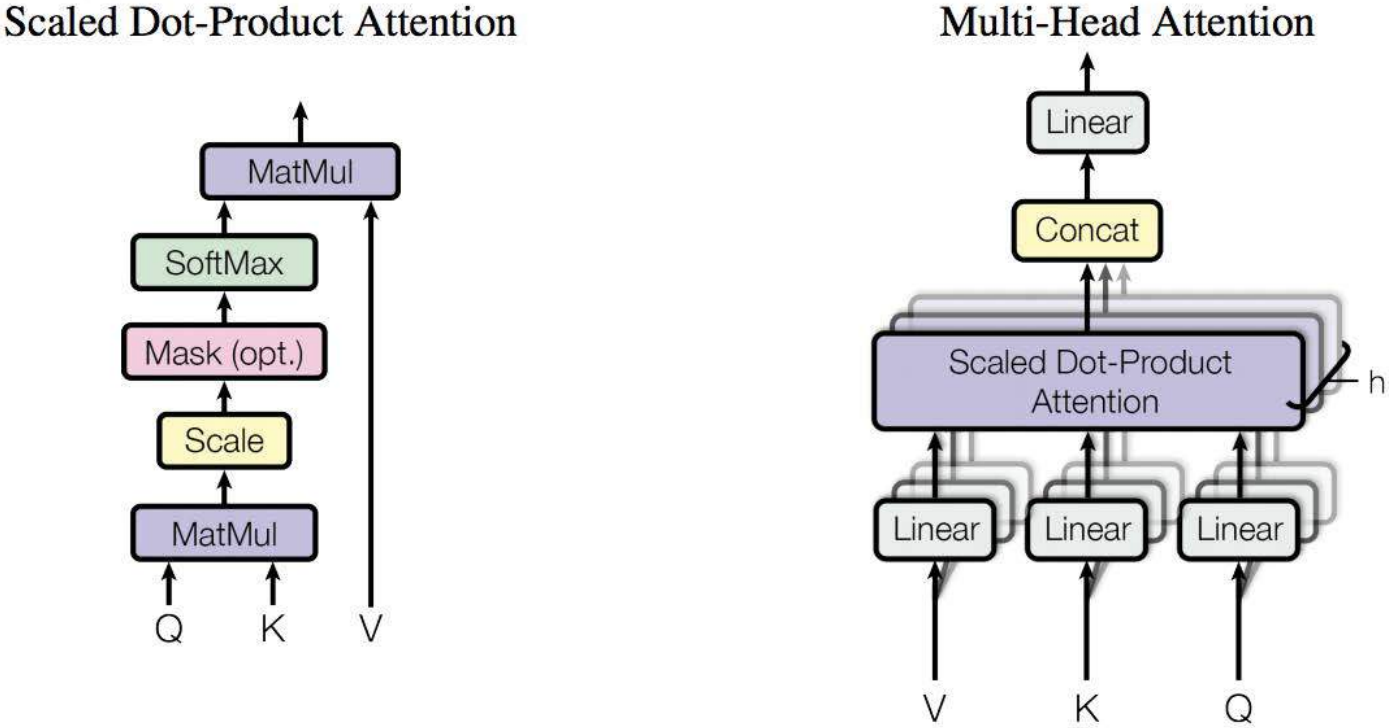
序列就一定需要RNN吗

- 诸多改进:
  - Self-Attention
  - Multi-headed attention
  - Normalized Dot-product Attention
  - Positional Encoding
  - Label Smoothing
  - Layer normalization
  - Residual Layers
  - Learning rate schedule
- 开源Tensor2Tensor



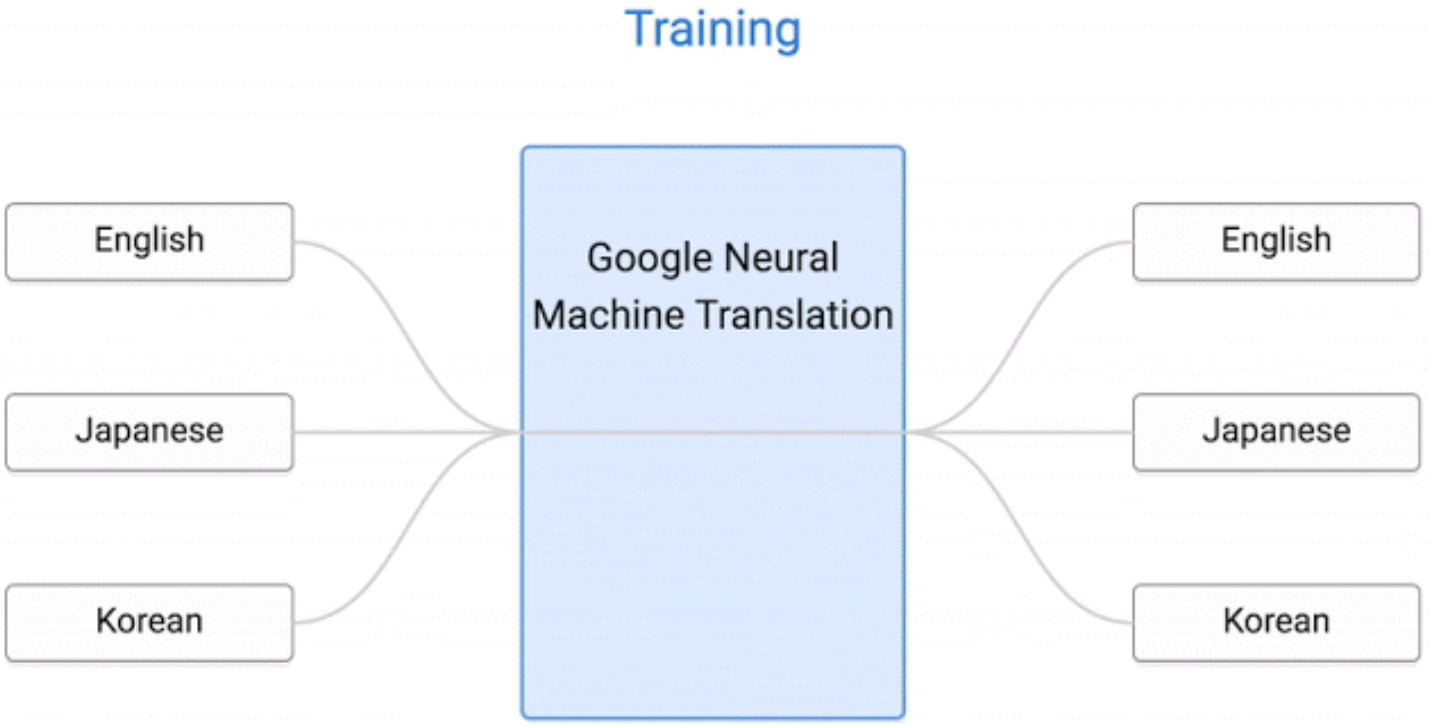
Vaswani et al, 2017

注意力



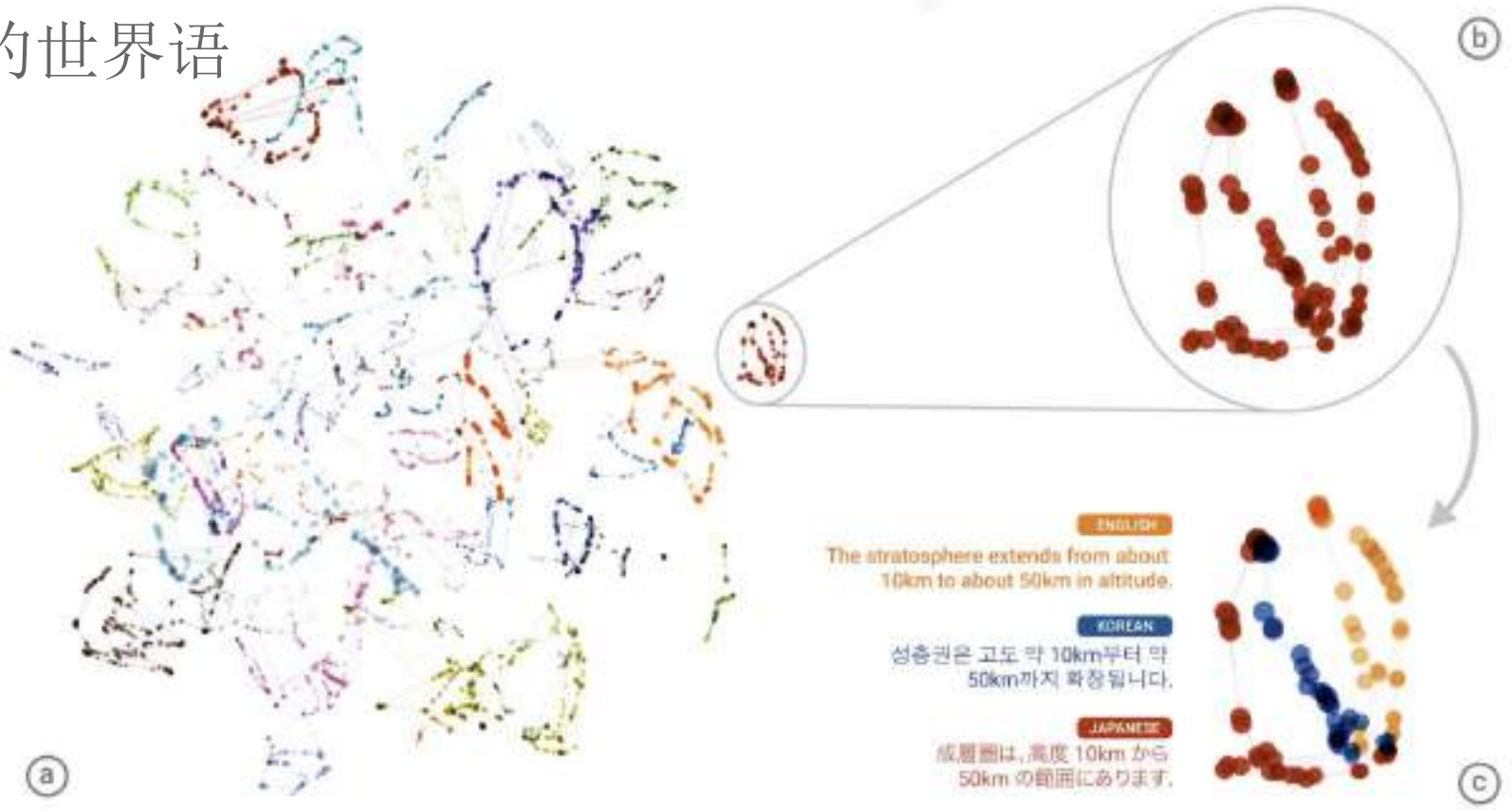
Vaswani et al, 2017

多语言共享模型

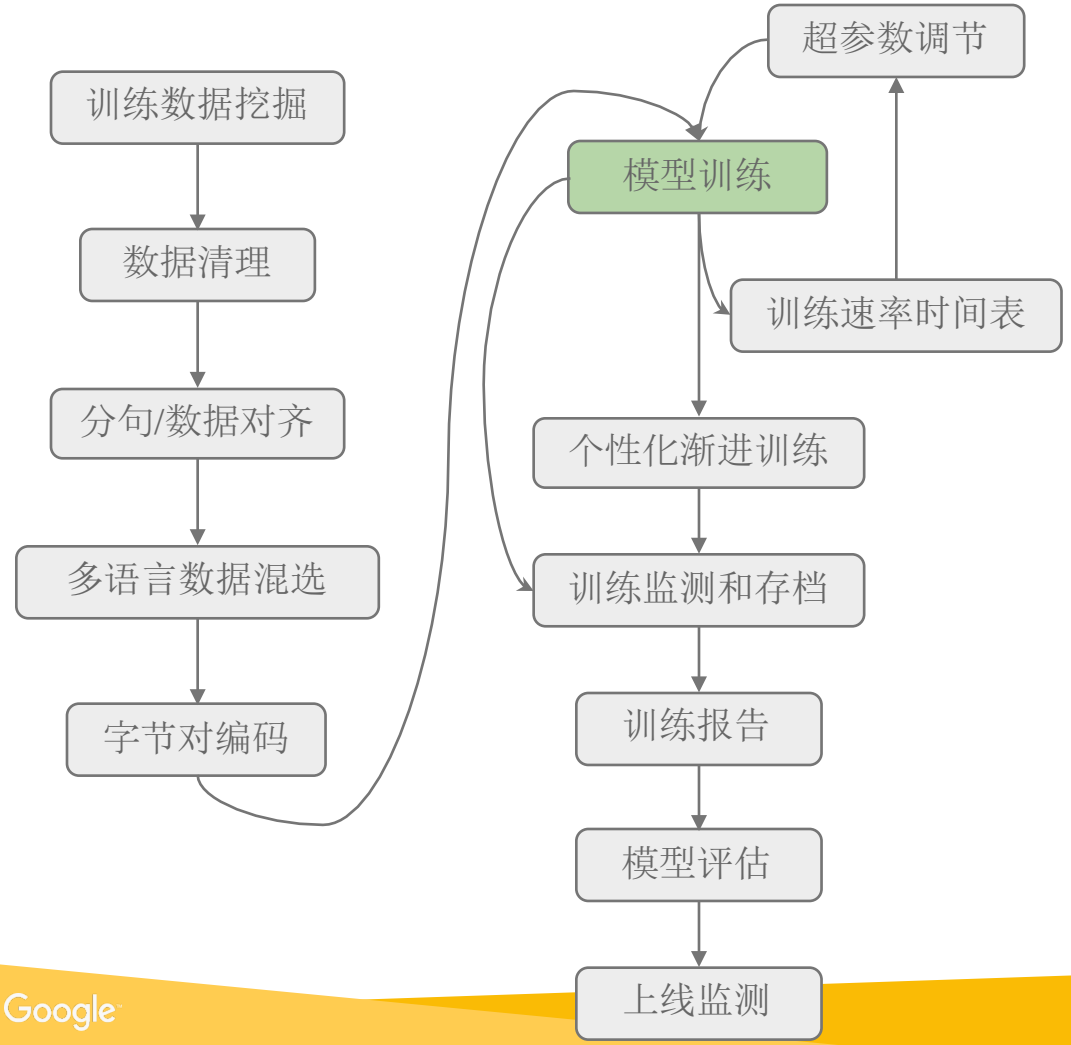


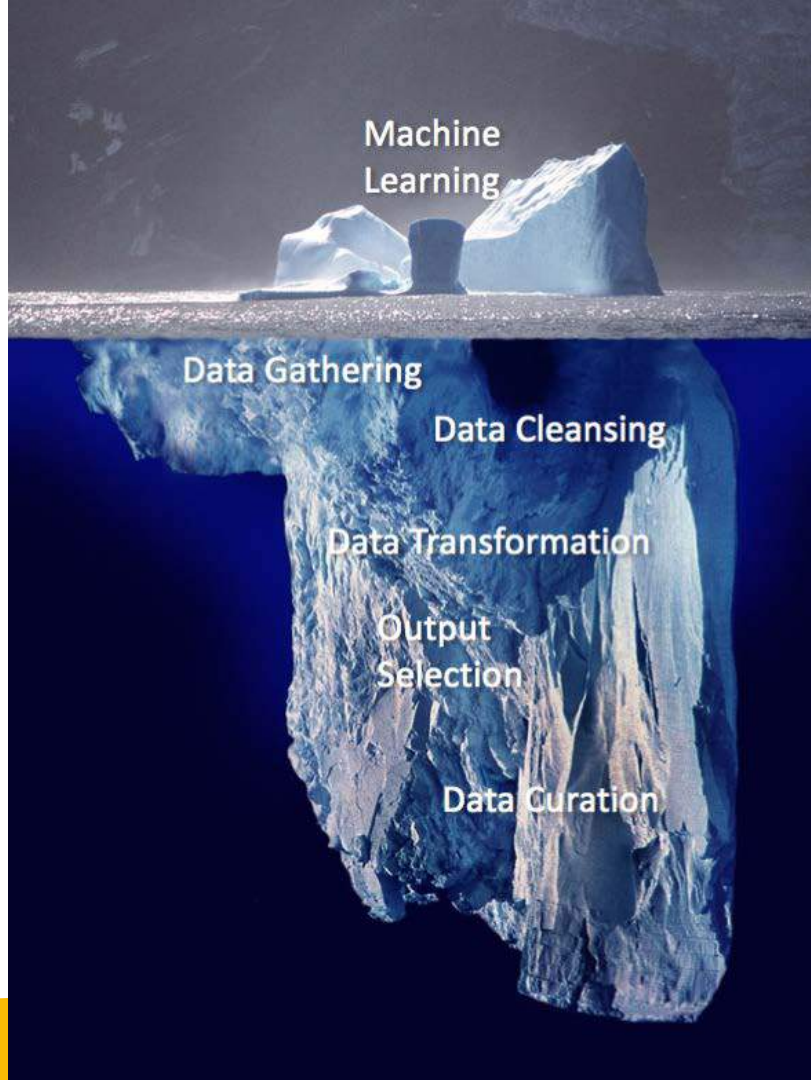
Johnson et al, 2017

隐藏的世界语



Johnson et al, 2017


训练数据挖掘
数据清理
分句/数据对齐
多语言数据混选
字节对编码
超参数调节
模型训练
训练速率时间表
个性化渐进训练
训练监测和存档
训练报告
模型评估
上线监测


Machine Learning
Data Gathering
Data Cleansing
Data Transformation
Output Selection
Data Curation

个性化训练

新闻领域模型

翻译原文： China encourages telecommunication cos. to increase spending on network security and submit trial projects for govt funding, according to statement on Ministry of Industry and Information Technology's website.

初始模型： 中国鼓励电信公司增加网络安全支出，并提交政府融资试点项目，根据工业和信息化部网站的声明。

个性化模型：工信部网站公告称，中国鼓励电信公司增加网络安全支出，并提交政府投资试点项目。

个性化训练

旅游领域模型

翻译原文： A tourist information package is available in the room and the English - speaking host is on - calls 24 hours.

初始模型： 客房提供旅游信息包，讲英语的主机24小时开放。

个性化模型 :房间内有旅游信息包，讲英语的房东24小时随叫随到。

Google Cloud Platform

Translation API

一站式服务

语言检测

高性价比

REST API

未来发展方向

- 文档翻译

  例如：人称代词的一致性

- 语境翻译

  例如：是正话还是反话？

- 多模态翻译

  例如：结合图像的字幕翻译

- 语音到语音直接翻译

  例如：真正的同声传译，离线？

- 翻译的公平性

  例如：性别的平等性

https://translate.google.cn/

谢谢