

翻译时用手指着原文

本课讲 2014 年的一篇文章。Bahdanau、Cho 和 Bengio 三位科学家教翻译机器边写边回头看原句。这就是“注意力”的起点。

适合 10 岁左右

约 40 分钟

需要：3 种颜色的彩笔、几张小卡片、几位同伴

1 热身：一张只能写 6 个字的卡片

读下面两句话。每读完一句，就在卡片上写**最多 6 个字**，帮自己记住它。

A. 小猫在睡觉。 卡片：□□□□□□

B. 星期六早上，小明和妹妹带着一个红色的风筝、两瓶橙汁和三个肉包，坐 8 路公交车去河边的公园。

卡片：□□□□□□

现在用纸把句子盖住，**只看卡片**回答：

A 句里，小猫在干什么？ _____

B 句里，风筝是什么颜色？ _____

B 句里，带了几个肉包？ _____

B 句里，坐几路车？ _____

A 句很简单，6 个字全装得下。B 句就难了，卡片太小，总有东西掉出去。2014 年以前，翻译机器也被这样一张卡片困住了。

2 翻译机器的老办法：读完整句，写一张卡片

那时候，最新的翻译机器由 2 个电脑程序组成：

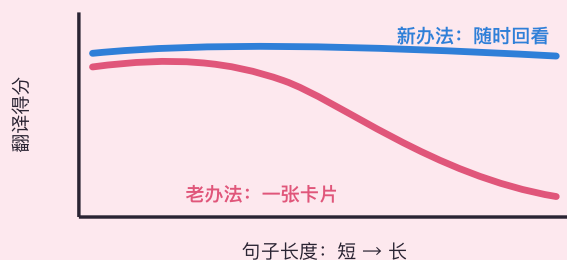


第 1 课讲过电脑的“小本子”：读一个字，记一笔。读的网络也带着小本子，一个词一个词地读原句。读完后，整本笔记被压成一张卡片。写的网络只看这张卡片，写出译文。科学家把读的网络叫**编码器**，把写的网络叫**解码器**。论文里做的是英文翻法文，上图换成了中文，道理一样。

问题出在压卡片这一步。这张**总结卡片**的大小是固定的。原句有 5 个词，卡片这么大。原句有 50 个词，卡片还是这么大。

3 句子一长，卡片装不下

句子越长，卡片越挤。科学家按句子长短给翻译打分，也发现了这件事：



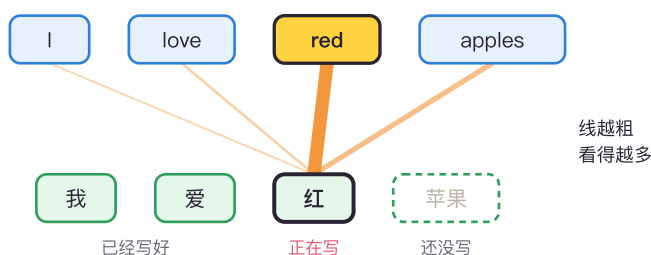
示意图，形状参照论文里的结果图。粉色线是老办法：句子一长，分数往下掉。蓝色线是这篇文章的新办法：长句子也几乎不掉分。

要怎么办？把卡片做大一点吗？这篇文章想了一个更聪明的办法。

4 新办法：每写一个词，就回头看一眼

你自己翻译句子时，不会先背下整句再闭着眼睛写。你会用手指指着原文，写到哪个词，就看看原文对应的地方。

这篇文章让写的网络也这样做。每写一个词之前，它都回头把原句的每个词看一遍，再决定**这一次主要看哪几个**。这就叫**注意力**：注意力多给几个词，少给其他词。



写“红”的时候，它主要看 red，也稍微看一眼 apples，因为“红”后面要接“苹果”。写下一个词时，它会**重新分配一次注意力**。

老办法：整句压成 1 张总结卡片，写每个词都看这一张。

新办法：原句每个词旁边都留一张小便签。写每个词时，回头看所有便签，现做一张新卡片。句子再长，便签也一直都在。

小便签上记着这个词和它左右邻居的意思。为了写好便签，读的网络读了两遍：一遍从左往右，一遍从右往左。

5 该看哪里，机器自己学会了

找出“译文的这个词对着原文的哪个词”，叫**对齐**。最神奇的是：**没有人教机器怎么对齐**。练习材料里只有很多“英文句子 + 人翻好的法文句子”。没有人画线告诉它哪个词对哪个词。为了翻得准，它自己学会了该看哪里。

科学家把机器“看哪里”画成了一张格子图，叫**热力图**。格子越深，表示写这个词时看那个词越多。下面这张是照论文里的一个例子画的示意图：

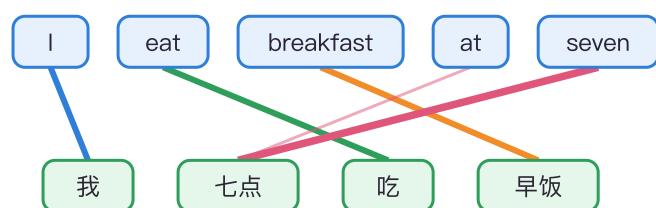
	the	European 欧洲	Economic 经济	Area 区
la				
zone (区)				
économique (经济)				
européenne (欧洲)				

上面一排是英文，左边一列是法文。

英文说“欧洲 经济 区”，法文说“区 经济 欧洲”，**顺序正好反过来**。

你看，深色格子也跟着拐了个弯，从右上斜到左下。

在整句话的热力图里，大多数深色格子沿着**对角线**排开。因为英文和法文的词序差不多，按顺序一个对一个就行。遇到词序不同的地方，深色格子就**交叉**了。中文和英文的词序差别更大，看这个我们编的例子：



英文把“七点”放在最后，中文放在“吃”前面。所以 seven 的线要斜着穿过其他线。“七点”用到了 at 和 seven，所以有 2 条线。

6 结果：长句子翻得好多了

科学家用 **BLEU** 给译文打分：把机器译文和人翻的译文对比，一样的词和词组越多，分数越高。

比什么	老办法（一张卡片）	新办法（随时回看）
所有测试句子的得分	约 18 分	约 27 分
50 个词以上的长句子	分数掉得很厉害	几乎不掉分

两种机器用的练习材料一样，只差在“能不能回头看”。还有一个惊喜：在没有生僻词的句子，多练了一段时间的新机器，和当时最好的老式翻译系统打成平手，还略高一点。

7 这和今天有什么关系

这篇文章里，“回头看”只是加在翻译机器上的一个小零件。第 13 课讲过的 Transformer，把这个零件变成了**整台机器的主要本事**：每个词都回头看其他所有词。

今天的聊天 AI 就是用 Transformer 造的。它写回答的每一个字时，都会回头看你的问题和它已经写过的话，再决定主要看哪几处。所以，你今天用的 AI 里，也有这篇文章的影子。

要诚实地说一句：热力图能告诉我们机器“在看哪里”，这是很好的线索。可是“看着哪里”不等于“想明白了为什么”。

8 动手：手指翻译，画出你的对齐图

玩法（1 到 2 个人，约 10 分钟）

1. 读句 1（小字是意思提示）。读中文时，用手指指着它对着的英文词，再用彩笔连线。

I like red apples
我 喜欢 红的 苹果

我 喜欢 红 苹果

2. 句 2 是“她在图书馆看书”，直接画成热力图：写每个中文词时主要看哪个英文词，就涂深那一格。

	She 她	reads 读	books 书	in 在里面	the 那个	library 图书馆
她						
在						
图书馆						
看						
书						

你会发现：句 1 的线不交叉，像对角线。句 2 的格子拐了弯，因为中文把“在图书馆”提到了前面。

9 动手：一张卡片 vs 随时回看

玩法（4 到 6 个人，约 15 分钟）

1. 分成两组。**卡片组**每人一张小卡片，最多写 8 个字。**回看组**可以随时翻看原文。

2. 老师先读短故事，再读长故事。每轮卡片组写一张新卡片，只看卡片答题。

短故事：小兔子早上去河边，钓到了一条鱼。 **问题：**① 谁去钓鱼？ ② 钓到了什么？

长故事：星期天早上 7 点，小熊背着绿色的书包出门。它在面包店买了 3 个菠萝包，在花店买了一束向日葵。下雨了，松鼠借给它一把蓝色的雨伞。中午 12 点，小熊到了奶奶家。

问题：① 几点出门？ ② 书包什么颜色？ ③ 买了几个菠萝包？ ④ 雨伞什么颜色？

	短故事：答对几题（共 2 题）	长故事：答对几题（共 4 题）
卡片组		
回看组		

你会发现：短故事两组差不多。长故事卡片组丢了很多细节，回看组几乎全对。

10 词语小卡片

总结卡片

老翻译机器把整句压成的一串数字。大小固定，句子一长就装不下。

注意力

每写一个词，都回头看原句，决定这一次主要看哪几个词。

对齐

找出译文的每个词对着原文的哪个词。机器自己学会了。

热力图

用深浅不同的格子，画出机器写每个词时在看哪里。

11 想一想

1. 为什么一张大小固定的卡片，装短句子够用，装长句子就不够？用你自己的话说一说。

2. 再想一句中文和英文词序不同的话，写下来，画出它的连线。线在哪里交叉了？

3. 热力图能看出机器“在看哪里”。你觉得“看着哪里”和“想明白了”是一回事吗？为什么？

一句话记住：边写边回头看原句，自己学会看哪里，这就是“注意力”。

给家长和老师：本课讲的是 Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio 的论文 [Neural Machine Translation by Jointly Learning to Align and Translate](#) (arXiv 2014, ICLR 2015)。论文指出编码器-解码器把整句压成固定长度向量，导致长句变差；RNNsearch 用双向 RNN 生成注解，解码器每步对所有注解打分加权，得到专属上下文向量（软对齐，后称加性注意力），无需人工对齐标注。WMT 2014 英法任务上，RNNsearch-50 的 BLEU 为 26.75，RNNencdec-50 为 17.82；无未登录词时，RNNsearch-50* 为 36.15，与 Moses 的 35.63 相当。

讲义中的“总结卡片”“小便签”是比喻；结构图、“I love red apples”连线、得分曲线、热力图都是示意（热力图照论文 European Economic Area 一例画，深浅非论文数据）；“I eat breakfast at seven”和活动里的句子、故事是编的；BLEU 取整。详见 [第 15 天中文导读](#)，代码见开源项目 [30u30 第 15 天目录](#)（英文）。