

翻譯時用手指著原文

本課講 2014 年的一篇文章。Bahdanau、Cho 和 Bengio 三位科學家教翻譯機器邊寫邊回頭看原句。這就是「注意力」的起點。

適合 10 歲左右

約 40 分鐘

需要：3 種顏色的彩筆、幾張小卡片、幾位同伴

1 熱身：一張只能寫 6 個字的卡片

讀下面兩句話。每讀完一句，就在卡片上寫**最多 6 個字**，幫自己記住它。

A. 小貓在睡覺。 卡片：□□□□□□

B. 星期六早上，小明和妹妹帶著一個紅色的風箏、兩瓶橙汁和三個肉包，坐 8 路公交車去河邊的公園。

卡片：□□□□□□

現在用紙把句子蓋住，**只看卡片**回答：

A 句裡，小貓在幹什麼？ _____

B 句裡，風箏是什麼顏色？ _____

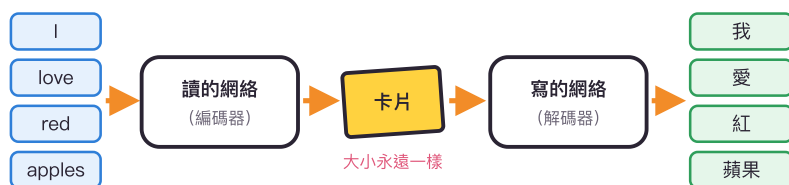
B 句裡，帶了幾個肉包？ _____

B 句裡，坐幾路車？ _____

A 句很簡單，6 個字全裝得下。B 句就難了，卡片太小，總有東西掉出去。2014 年以前，翻譯機器也被這樣一張卡片困住了。

2 翻譯機器的老辦法：讀完整句，寫一張卡片

那時候，最新的翻譯機器由 2 個電腦程序組成：

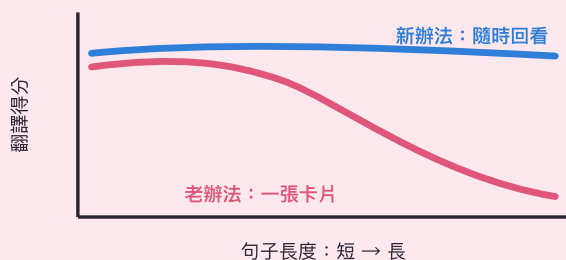


第 1 課講過電腦的「小本子」：讀一個字，記一筆。讀的網絡也帶著小本子，一個詞一個詞地讀原句。讀完後，整本筆記被壓成一張卡片。寫的網絡只看這張卡片，寫出譯文。科學家把讀的網絡叫**編碼器**，把寫的網絡叫**解碼器**。論文裡做的是英文翻法文，上圖換成了中文，道理一樣。

問題出在壓卡片這一步。這張**總結卡片**的大小是固定的。原句有 5 個詞，卡片這麼大。原句有 50 個詞，卡片還是這麼大。

3 句子一長，卡片裝不下

句子越長，卡片越擠。科學家按句子長短給翻譯打分，也發現了這件事：



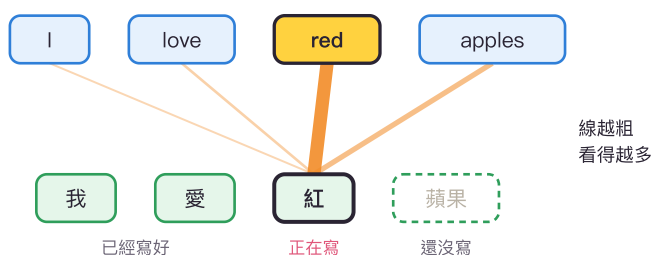
示意圖，形狀參照論文裡的結果圖。粉色線是老辦法：句子一長，分數往下掉。藍色線是這篇文章的新辦法：長句子也幾乎不掉分。

要怎麼辦？把卡片做大一點嗎？這篇文章想了一個更聰明的辦法。

4 新辦法：每寫一個詞，就回頭看一眼

你自己翻譯句子時，不會先背下整句再閉著眼睛寫。你會**用手指指著原文**，寫到哪個詞，就看看原文對應的地方。

這篇文章讓寫的網絡也這樣做。每寫一個詞之前，它都回頭把原句的每個詞看一遍，再決定**這一次主要看哪幾個**。這就叫**注意力**：注意力多給幾個詞，少給其他詞。



寫「紅」的時候，它主要看 red，也稍微看一眼 apples，因為「紅」後面要接「蘋果」。寫下一個詞時，它會**重新分配一次注意力**。

老辦法：整句壓成 1 張總結卡片，寫每個詞都看這同一張。

新辦法：原句每個詞旁邊都留一張小便籤。寫每個詞時，回頭看所有便籤，現做一張新卡片。句子再長，便籤也一直都在。

小便籤上記著這個詞和它左右鄰居的意思。為了寫好便籤，讀的網絡讀了兩遍：一遍從左往右，一遍從右往左。

5 該看哪裡，機器自己學會了

找出「譯文的這個詞對著原文的哪個詞」，叫**對齊**。最神奇的是：**沒有人教機器怎麼對齊**。練習材料裡只有很多「英文句子 + 人翻好的法文句子」。沒有人畫線告訴它哪個詞對哪個詞。為了翻得準，它自己學會了該看哪裡。

科學家把機器「看哪裡」畫成了一張格子圖，叫**熱力圖**。格子越深，表示寫這個詞時看那個詞越多。下面這張是照論文裡的一個例子畫的示意圖：

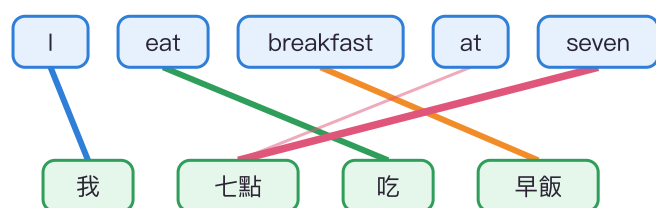
	the	European 歐洲	Economic 經濟	Area 區
la				
zone (區)				
économique (經濟)				
européenne (歐洲)				

上面一排是英文，左邊一列是法文。

英文說「歐洲 經濟 區」，法文說「區 經濟 歐洲」，**順序正好反過來**。

你看，深色格子也跟著拐了個彎，從右上斜到左下。

在整句話的熱力圖裡，大多數深色格子沿著**對角線**排開。因為英文和法文的詞序差不多，按順序一個對一個就行。遇到詞序不同的地方，深色格子就**交叉**了。中文和英文的詞序差別更大，看這個我們編的例子：



英文把「七點」放在最後，中文放在「吃」前面。所以 seven 的線要斜著穿過其他線。「七點」用到了 at 和 seven，所以有 2 條線。

6 結果：長句子翻得好多了

科學家用 **BLEU** 給譯文打分：把機器譯文和人翻的譯文對比，一樣的詞和詞組越多，分數越高。

比什麼	老辦法（一張卡片）	新辦法（隨時回看）
所有測試句子的得分	約 18 分	約 27 分
50 個詞以上的長句子	分數掉得很厲害	幾乎不掉分

兩種機器用的練習材料一樣，只差在「能不能回頭看」。還有一個驚喜：在沒有生僻詞的句子，多練了一段時間的新機器，和當時最好的老式翻譯系統打成平手，還略高一點。

7 這和今天有什麼關係

這篇文章裡，「回頭看」只是加在翻譯機器上的一個小零件。第 13 課講過的 Transformer，把這個零件變成了**整臺機器的主要本事**：每個詞都回頭看其他所有詞。

今天的聊天 AI 就是用 Transformer 造的。它寫回答的每一個字時，都會回頭看你的問題和它已經寫過的話，再決定主要看哪幾處。所以，你今天用的 AI 裡，也有這篇文章的影子。

要誠實地說一句：熱力圖能告訴我們機器「在看哪裡」，這是很好的線索。可是「看著哪裡」不等於「想明白了為什麼」。

8 動手：手指翻譯，畫出你的對齊圖

玩法（1 到 2 個人，約 10 分鐘）

1. 讀句 1（小字是意思提示）。讀中文時，用手指指著它對著的英文詞，再用彩筆連線。

I like red apples

我 喜歡 紅的 蘋果

我 喜歡 紅 蘋果

2. 句 2 是「她在圖書館看書」，直接畫成熱力圖：寫每個中文詞時主要看哪個英文詞，就塗深那一格。

	She 她	reads 讀	books 書	in 在裡面	the 那個	library 圖書館
她						
在						
圖書館						
看						
書						

你會發現：句 1 的線不交叉，像對角線。句 2 的格子拐了彎，因為中文把「在圖書館」提到了前面。

9 動手：一張卡片 vs 隨時回看

玩法（4 到 6 個人，約 15 分鐘）

- 分成兩組。**卡片組**每人一張小卡片，最多寫 8 個字。**回看組**可以隨時翻看原文。
- 老師先讀短故事，再讀長故事。每輪卡片組寫一張新卡片，只看卡片答題。

短故事：小兔子早上去河邊，釣到了一條魚。 **問題：**① 誰去釣魚？② 釣到了什麼？

長故事：星期天早上 7 點，小熊揹著綠色的書包出門。它在麵包店買了 3 個菠蘿包，在花店買了一束向日葵。下雨了，松鼠借給它一把藍色的雨傘。中午 12 點，小熊到了奶奶家。

問題：① 幾點出門？② 書包什麼顏色？③ 買了幾個菠蘿包？④ 雨傘什麼顏色？

	短故事：答對幾題（共 2 題）	長故事：答對幾題（共 4 題）
卡片組		
回看組		

你會發現：短故事兩組差不多。長故事卡片組丟了很多細節，回看組幾乎全對。

10 詞語小卡片

總結卡片

老翻譯機器把整句壓成的一串數字。大小固定，句子一長就裝不下。

注意力

每寫一個詞，都回頭看原句，決定這一次主要看哪幾個詞。

對齊

找出譯文的每個詞對著原文的哪個詞。機器自己學會了。

熱力圖

用深淺不同的格子，畫出機器寫每個詞時在看哪裡。

11 想一想

1. 為什麼一張大小固定的卡片，裝短句子夠用，裝長句子就不夠？用你自己的話說一說。

2. 再想一句中文和英文詞序不同的話，寫下來，畫出它的連線。線在哪裡交叉了？

3. 熱力圖能看出機器「在看哪裡」。你覺得「看著哪裡」和「想明白了」是一回事嗎？為什麼？

一句話記住：邊寫邊回頭看原句，自己學會看哪裡，這就是「注意力」。

給家長和老師：本課講的是 Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio 的論文 [Neural Machine Translation by Jointly Learning to Align and Translate](#) (arXiv 2014, ICLR 2015)。論文指出編碼器-解碼器把整句壓成固定長度向量，導致長句變差；RNNsearch 用雙向 RNN 生成註解，解碼器每步對所有註解打分加權，得到專屬上下文向量（軟對齊，後稱加性注意力），無需人工對齊標註。WMT 2014 英法任務上，RNNsearch-50 的 BLEU 為 26.75，RNNencdec-50 為 17.82；無未登錄詞時，RNNsearch-50* 為 36.15，與 Moses 的 35.63 相當。

講義中的「總結卡片」「小便籤」是比喻；結構圖、「I love red apples」連線、得分曲線、熱力圖都是示意（熱力圖照論文 European Economic Area 一例畫，深淺非論文數據）；「I eat breakfast at seven」和活動裡的句子、故事是編的；BLEU 取整。詳見 [第 15 天中文導讀](#)，代碼見開源項目 [30u30 第 15 天目錄](#)（英文）。