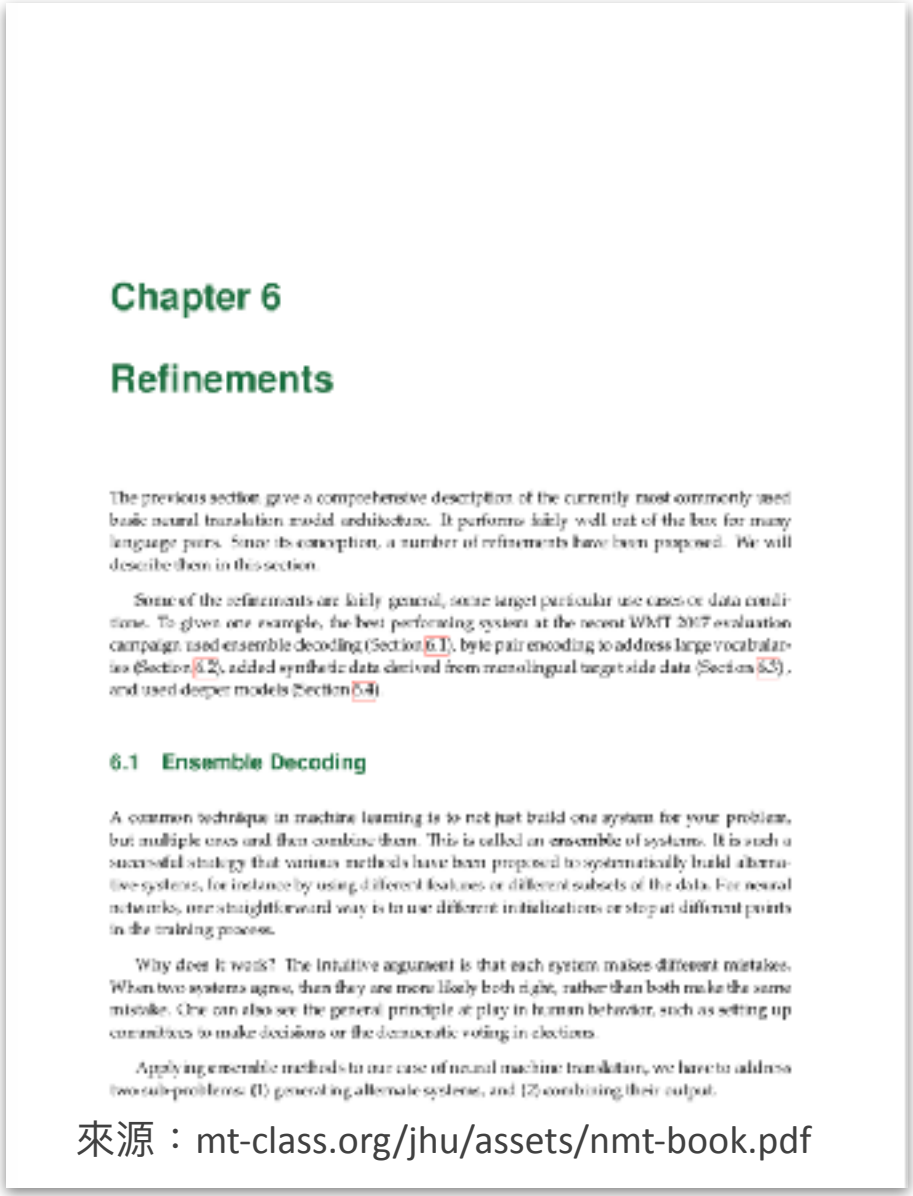
NMT 第 6 章 NMT解碼與各種改進

教科書與課程網站：mt-class.org/jhu/syllabus.html （草稿）

2018 1120

教科書相關章節



Chapter 6

Refinements

The previous section gave a comprehensive description of the currently most commonly used basic neural translation model architecture. It performs fairly well out of the box for many language pairs. Since its conception, a number of refinements have been proposed. We will describe them in this section.

Some of the refinements are fairly general, some target particular use cases or data conditions. To give one example, the best performing system at the recent WMT 2017 evaluation campaign used ensemble decoding (Section 6.1), byte pair encoding to address large vocabularies (Section 6.2), added synthetic data derived from monolingual target side data (Section 6.3), and used deeper models (Section 6.4).

6.1 Ensemble Decoding

A common technique in machine learning is to not just build one system for your problem, but multiple ones and then combine them. This is called an ensemble of systems. It is such a successful strategy that various methods have been proposed to systematically build alternative systems, for instance by using different features or different subsets of the data. For neural networks, one straightforward way is to use different initializations or stop at different points in the training process.

Why does it work? The intuitive argument is that each system makes different mistakes. When two systems agree, then they are more likely both right, rather than both make the same mistake. One can also see the general principle at play in human behavior, such as setting up committees to make decisions or the democratic voting in elections.

Applying ensemble methods to our case of neural machine translation, we have to address two sub-problems: (1) generating alternate systems, and (2) combining their output.

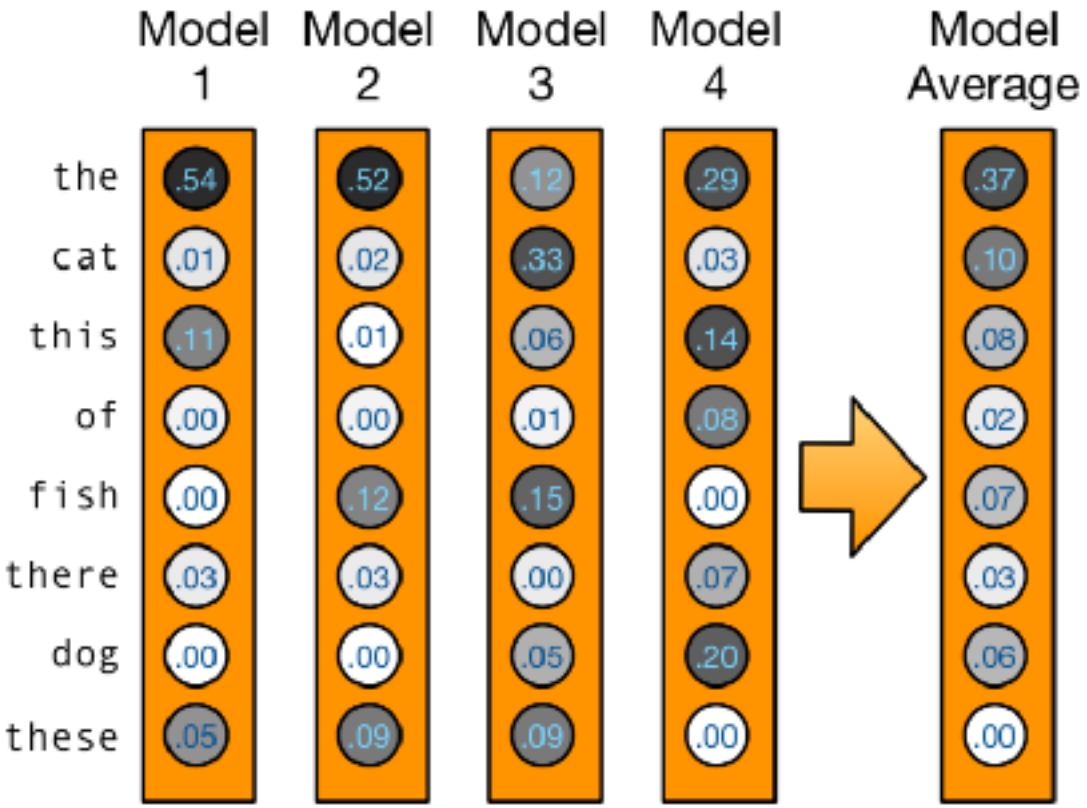
來源：mt-class.org/jhu/assets/nmt-book.pdf

6 類神經機器翻譯的改進—解碼器

- 第5章介紹的模型，可應用在很多語言配對，有不錯的效果
  - 有很多議題仍有改進空間
  - WMT 2017 評估競賽中，提出不少改進的作法
- 第6章大綱
  - 6.1 集合式解碼系統 Ensemble Decoding
  - 6.2 應付大詞彙集：byte pair 輸入編碼
  - 6.3 加入合成資料
  - 6.4 更深的模型
  - 6.5 用詞對應引導訓練
  - 6.6 把翻譯涵蓋範圍納入模型
  - 6.7 調整訓練與執行的差異
  - 6.8 加入語言學註記
  - 6.9 訓練多個語言配對

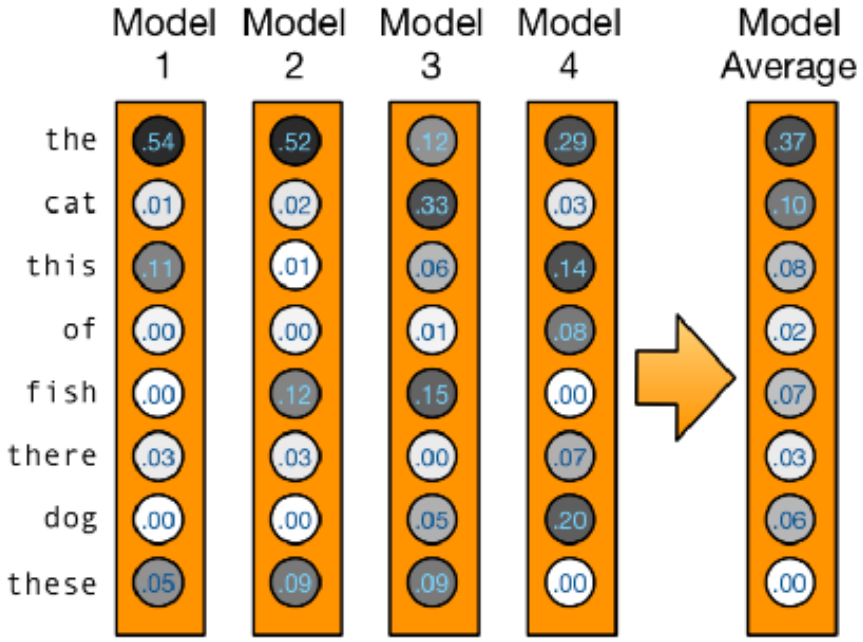
6.1.2 組合系統輸出

- 簡單的作法：取各系統輸出預測機率值
  - 加總取平均值
  - 輸出最佳預測
- 給不同系統不同加權(少見)



6.1.3 系統組合的變形： 解碼中的答案重排序 reranking

- 可以組合正向、反向輸出的系統
  - 不適合 6.1.2 的簡易組合，需要更深的整合
- 重排序的整合方法
  - 用一組由左至右的系統，產生n 個最佳翻譯句
  - 產生所有系統對n 個最佳翻譯句 的評分
  - 結合評分 (簡單平均值) ，以選擇最佳輸出句



6.1 集合式解碼系統 Ensemble Decoding

- 非常有效的策略
  - 不同方法、設定導致不同錯誤，一致結果正確的機率較高
- 引發的次問題
  - (1) 產生不同的系統 (2) 合併所有系統的輸出
- 作法
  - 不同訓練起始條件——效果較佳
    - 權重初始值 (收斂到不同的局部最佳化點)
    - 隨機產生的訓練資料順序
  - 不同訓練結束點 (checkpoint ensemble)——效果有限

結束點集合 Checkpoint ensemble

多次訓練集合 Multi-run ensemble

6.2 應付未知詞：byte pair 輸入編碼

- Zipf 規則指出詞有非常不平均的分佈
- 高頻詞很少，低頻詞非常多
  - 包括公司名，如 eBay, Yahoo, Microsoft
- 類神經網路 (比傳統作法) 更難處理大詞彙集 large vocabularies
  - 使用預先訓練詞內嵌
  - 輸出端需要產生一個很大的矩陣
  - 速度上受影響 (和詞彙集大小成正比)
- 為了速度、為了使用詞內嵌
  - 限制詞彙集
  - 使用 UNK 代表不在詞彙集或未知詞
  - 使用預備詞典處理 UNK 的翻譯vback-off dictionary
- 新作法—用比詞更小的單位，表達未知詞
  - 次詞單位 sub-word units

6.2 Byte Pair 輸入編碼

- 用平行語料庫訓練
  - 由字母開始，最後產生一組最佳的字母ngrams (很多就是高頻詞)
  - 反覆合併最高頻的相鄰 n-字母 (e+r, t+h, c+h)
- 49,500 回合之後， 大部分常見詞不變，
  - 罕見詞：構詞分解 (im@@ pending)
    其他方式 (stra@@ined, Net@ @any@ @ahu)

Obama receives Net@ @any@ @ahu

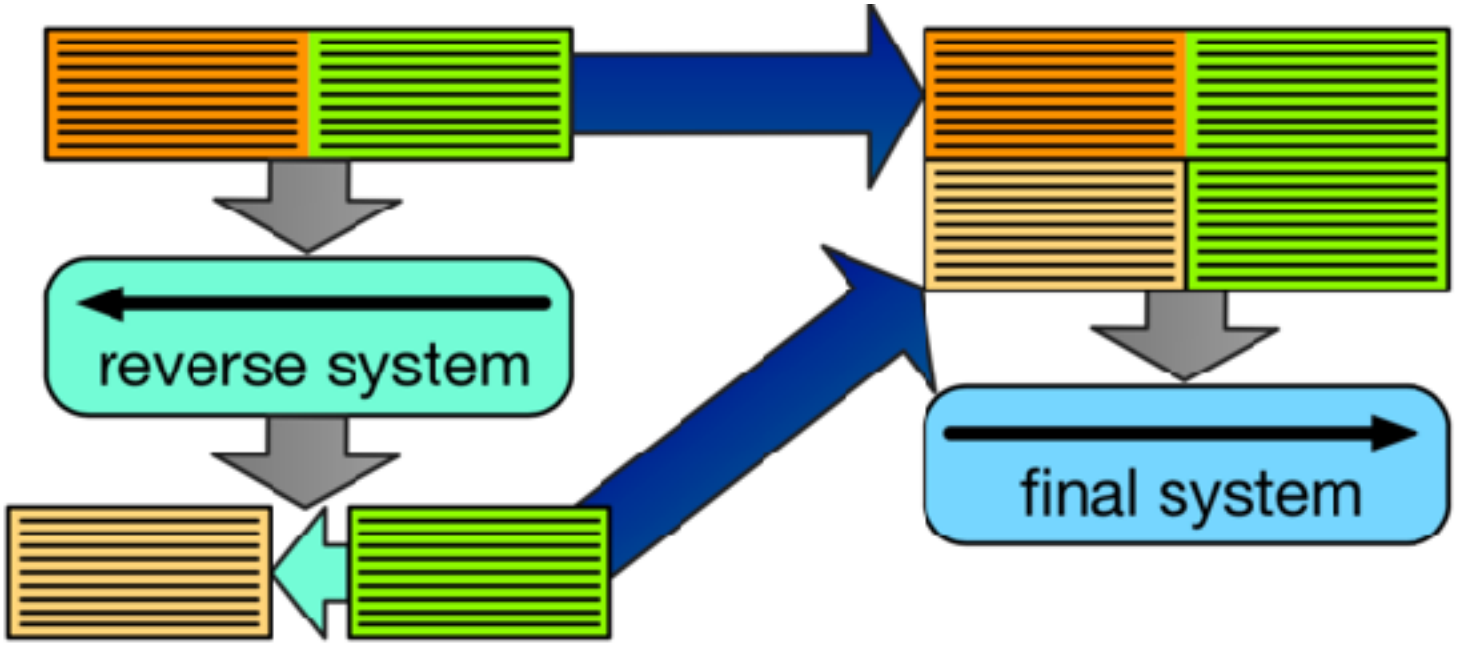
the relationship between Obama and Net@@any@ @ahu is not exactly friendly . the two wanted to talk about the implementation of the international agreement and about Teheran 's destabil@ @ising activities in the Middle East . the meeting was also planned to cover the conflict with the Palestinians and the disputed two state solution . relations between Obama and Net@@any@ @ahu have been stra@@ ined for years . Washington critic@ @ises the continuous building of settlements in Israel and acc@ @uses Net@@ any@@ ahu of a lack of initiative in the peace process . the relationship between the two has further deteriorated because of the deal that Obama negotiated on Iran 's atomic programme . in March , at the invitation of the Republic@ @ans , Net@@any@ @ahu made a controversial speech to the US Congress , which was partly seen as an aff@ @ront to Obama . the speech had not been agreed with Obama , who had rejected a meeting with reference to the election that was at that time im@@pending in Israel .

6.3 加入合成資料

- 統計式機器翻譯特色—大型單語語料庫訓練語言模型
  - 語言模型愈大，翻譯的效果愈好
  - 曾有一兆詞的語言模型用於機器翻譯
- 反而，基本的類神經網路模型沒有利用單語資料
  - 用前隱藏狀態＋前一輸出詞，預測下一詞
  - 和翻譯模型一起，用平行語料庫 (通常比較小) 訓練

6.3.1 反向翻譯 back translation

- 可以用目標語資料產生「合成」的平行資料 synthetic parallel data
  - 用資料訓練一個反向系統
  - 用反向系統翻譯目標語資料成為來源語資料＝合成平行資料
  - 結合原有資料＋合成資料，訓練最終的系統



6.3.2 用單語資料訓練語言模型，加入系統

- 原有 NMT 模型的解碼部分

$$s_i = f(s_{i-1}, Ey_{i-1}, c_i)$$

  - 輸出條件：前一狀態 s_{i-1}
    前一輸出 Ey_{i-1}
    目前文脈 c_i
- 改成

$$e_i = g(c_i, s_i^{\mathrm{TM}}, s_i^{\mathrm{LM}}, e_{i-1})$$

  - 輸出條件：前一狀態 e_{i-1}
    目前文脈 c_i
    語言模型狀態 s_i^{TM} s_i^{LM}

6.3.2 語言模型— 調整 LM 和 TM的相對重要性

- 用閘門元件 $\text{gate}_i^{\text{LM}}$
- 調整 LM 和 TM的相對重要性

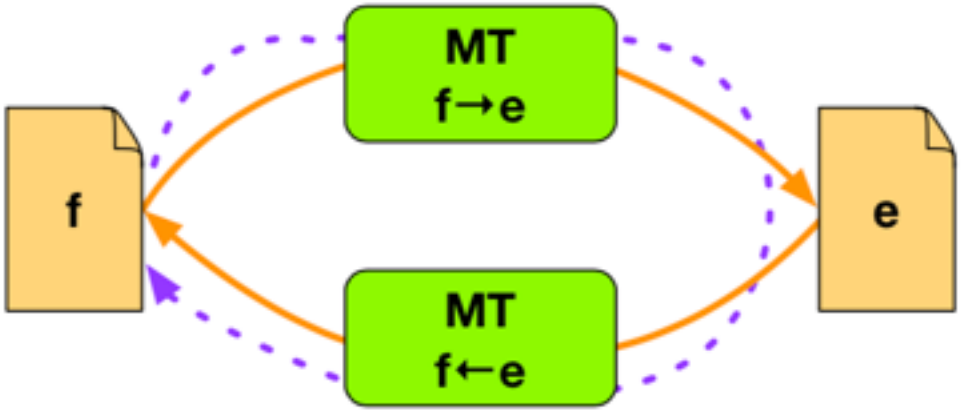
$$\text{gate}_i^{\text{LM}} = f(s_i^{\text{LM}})$$

$$\bar{s}_i^{\text{LM}} = \text{gate}_i^{\text{LM}} \times s_i^{\text{LM}}$$

$$e_i = g(c_i, s_i^{\text{TM}}, \bar{s}_i^{\text{LM}}, e_{i-1})$$

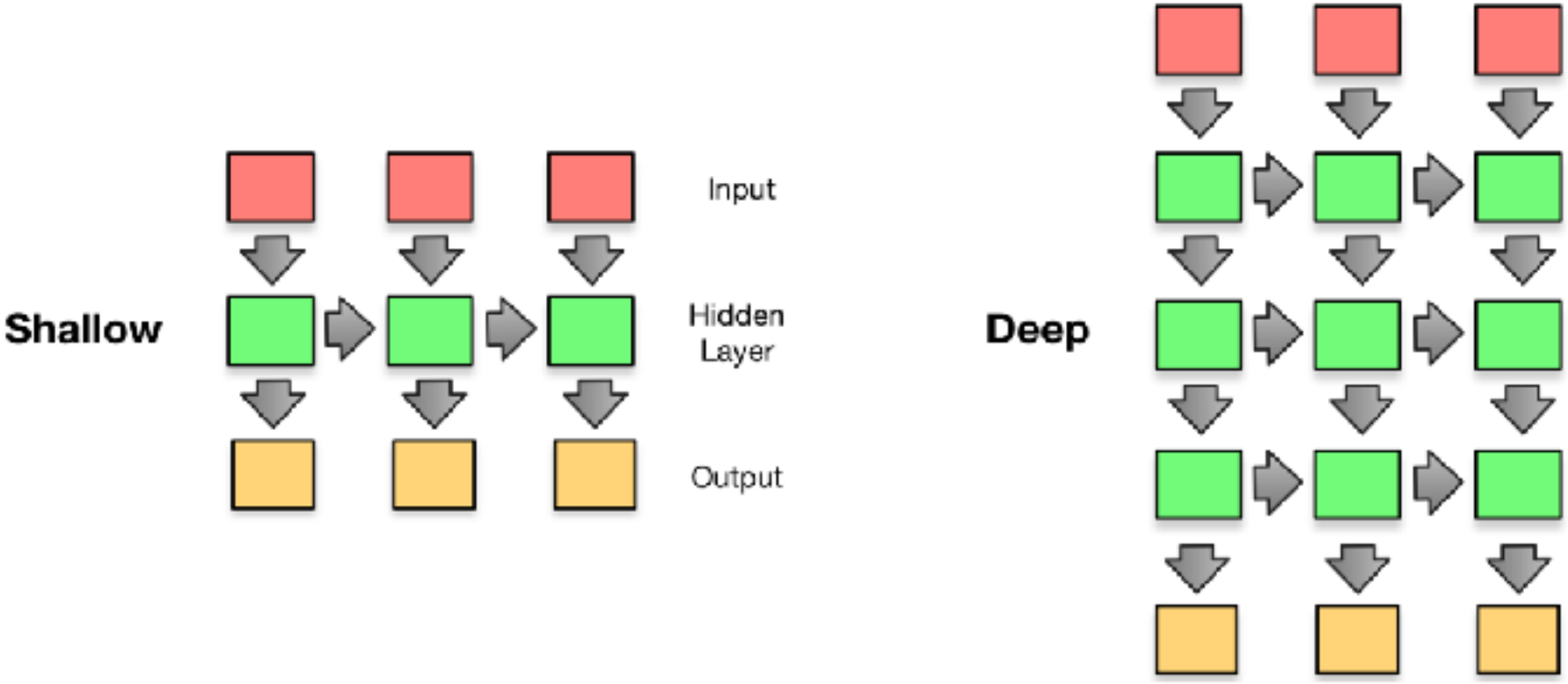
6.3.3 來回訓練 round trip training

- 如果用反向翻譯產生新的資料
  - 應該讓輸出句反向翻譯時，也能保持其「意義」
- 可以同時訓練不同方向的兩個模型
  - 給予句子 f 的 翻譯 e’ ，必須有高的語言模型機率 LM(e’).
  - 把翻譯 e’ 反向轉回 f 應該有高的翻譯模型機率MTe→f (f|e’)
  - 這兩個目標可以用來更新模型參數 MTf→e andMTe→f.



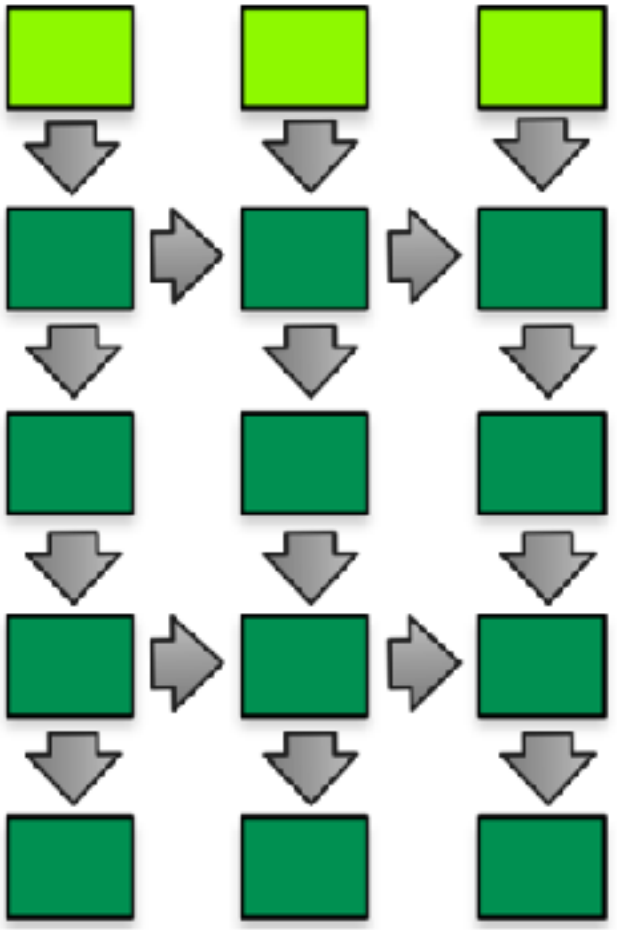
6.4 深度解碼器 (1)

- 不同的模型變形
  - 前饋類神經網路 feed-forward neural network



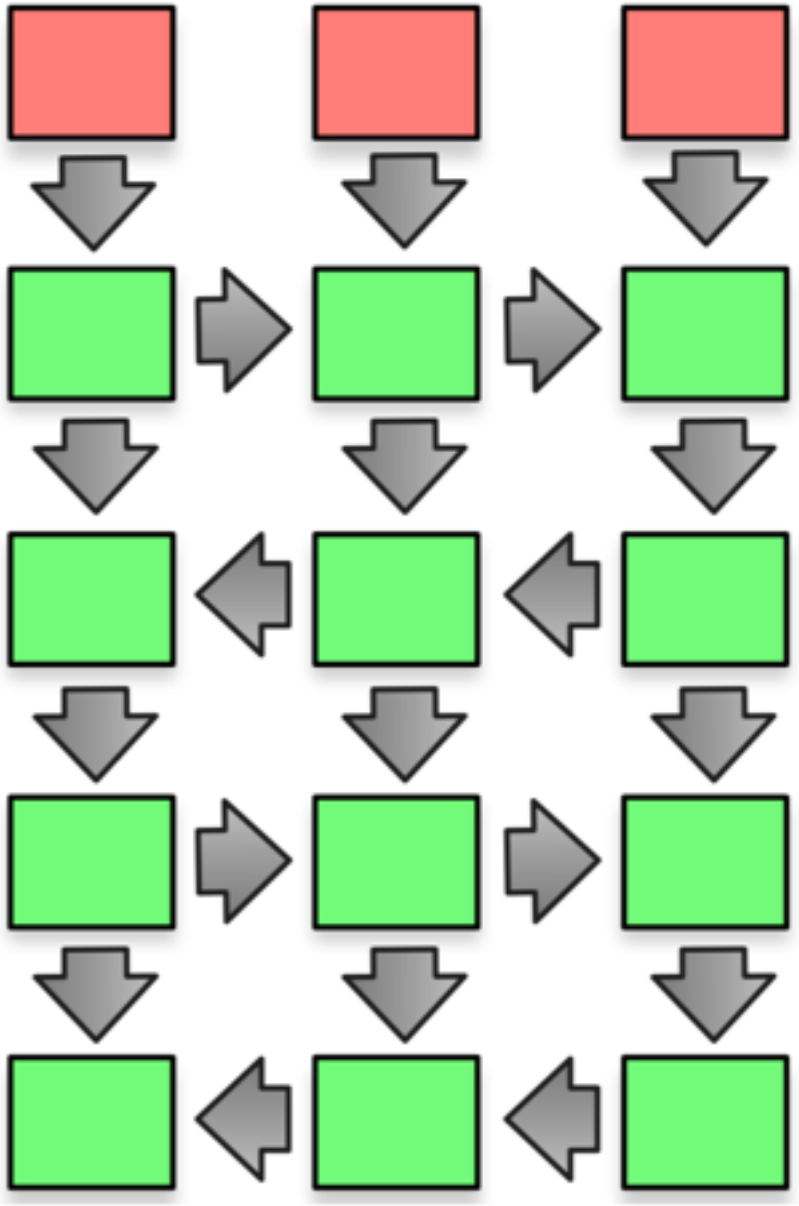
6.4 深度解碼器 (2)

- 不同的模型變形
  - 前饋類神經網路 feed-forward neural network
  - 遞迴類神經網路 recurrent neural network
  - 長短期記憶類神經網路 long short term memory neural network
- 可能包括輸入的文脈



6.4 深度解碼器 (3)

- 不同的模型變形
  - 前饋類神經網路 feed-forward neural network
  - 遞迴類神經網路 recurrent neural network
  - 長短期記憶類神經網路 long short term memory neural network
- 可能包括輸入的文脈



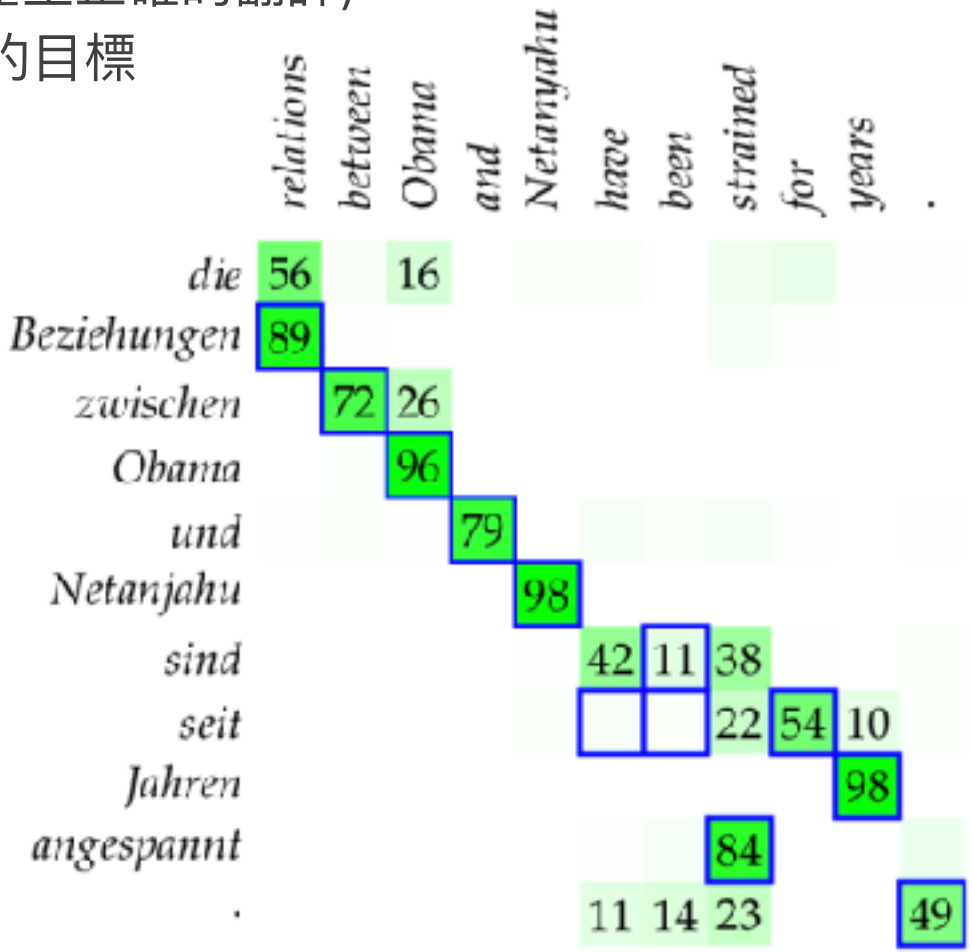
6.5 用詞對應引導訓練 Guided Alignment Training

- 用傳統詞彙對應 (方框)、注意模型大致相符 (百分比)
- 把詞彙對應也視為訓練資料
  - 改變訓練的目標 (通常是產生正確的翻譯)
  - 加 α_{ij} 和詞彙對應 A_{ij} 相符的目標
- Cross Entropy (CE)

$$\text{cost}_{\text{CE}} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} A_{ij} \log \alpha_{ij}$$

- Mean Square Error (MSE)

$$\text{cost}_{\text{MSE}} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} (A_{ij} - \alpha_{ij})^2$$



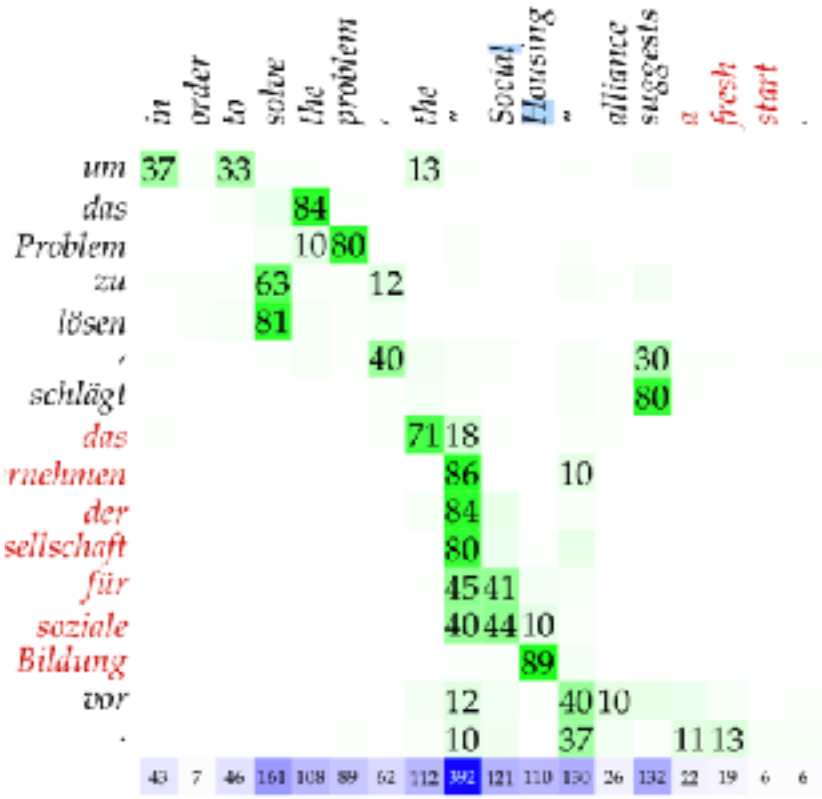
6.6 把翻譯涵蓋範圍納入模型 Modeling Coverage

- 類神經機器翻譯，可很有效地翻譯、重排整個句子
- 但是，偶而模型會出問題
  - 一個輸入詞翻譯了好幾次 、輸入詞未翻譯出
- 圖 6.9 的例子有兩個問題
  - 開始的 Social Housing 翻譯成太多冗餘詞：
  - ***das Unternehmen der Gesellschaft für soziale Bildung*** *(the company of the society for social education)*
  - 結束的 ***a fresh start***
    - 沒有得到「注意」因此未譯出
- 使用下列的評分函數

$$\text{coverage}(j) = \sum_i \sum_k \alpha_{i,k}$$

$$\text{over-generation} = \max\left(0, \sum_j \text{coverage}(j) - 1\right)$$

$$\text{under-generation} = \min\left(1, \sum_j \text{coverage}(j)\right)$$

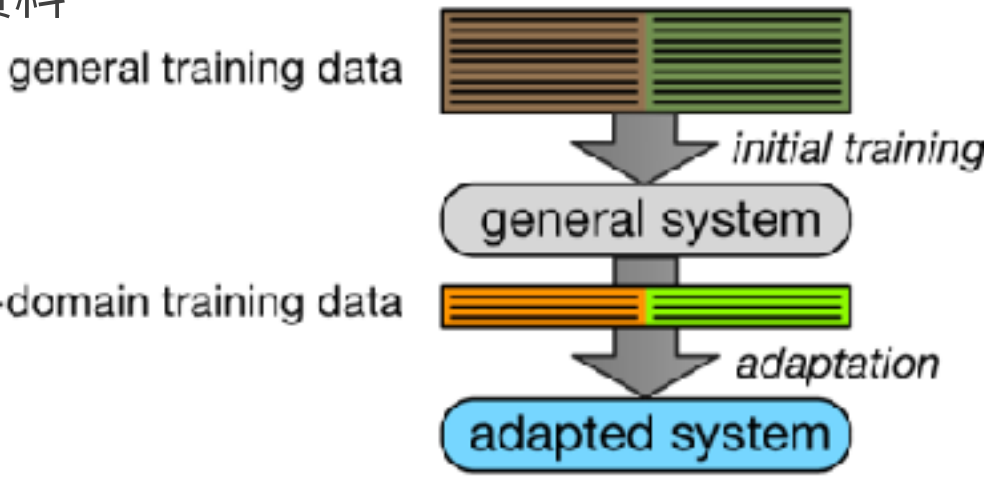


6.7 調整適用領域 Adaptation

- 先用一般資料訓練模型，再用領域內資料，調整模型 (如圖)
- 在大資料集中，取樣「領域內」資料
  - 使用領域分類器
  - 在大型「領域外」資料集中擷取「領域內、資料
  - 用語言模型，決定領域相關性

$$\text{relevance}_{e,f} = \left(\text{LM}_e^{\text{in}}(e) - \text{LM}_e^{\text{out}}(e)\right) + \left(\text{LM}_f^{\text{in}}(f) - \text{LM}_f^{\text{out}}(f)\right)$$

- 使用「領域內」單語資料
  - 使用單語資料 (來源、目標皆可) 在平行語料庫中取樣
  - 反向翻譯「領域內」單語資料
- 多重領域
  - 發展多個領域的翻譯模型
  - 先預測輸入句的領域
  - 用特定領域的模型翻譯



6.8 加入語言學註記 Linguistic Annotation

- 目前的研究關鍵
  - 發展一般性機器學習方法 (「隱含」的語言特徵)
  - 用語言直覺增強資料與模型 (使用「外顯」的語言特徵)
- 用語言學啟發的模型有效果
- 目前最佳的中英、德英翻譯系統不是類神經，而是句法為本
- 也有研究探索深度的語意的機器翻譯系統
- 不同的作法
  - 輸入句的語言學註記
  - 輸出句的語言學註記
  - 結構化語言學模型

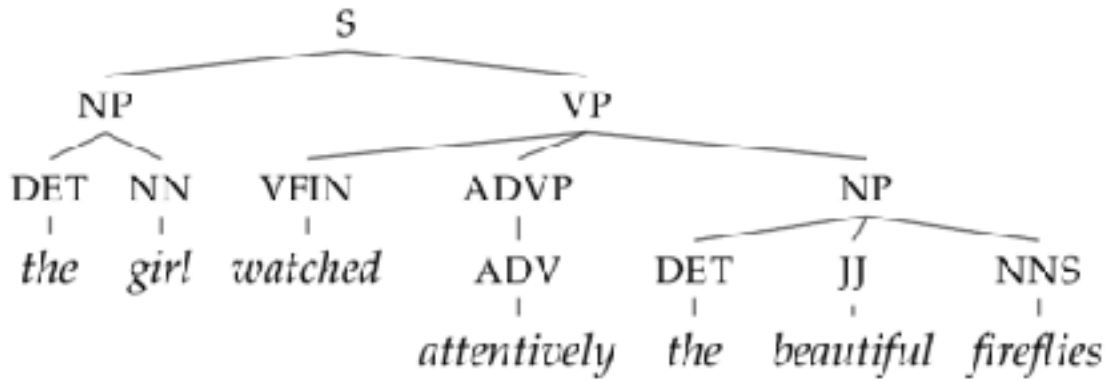
6.8.1 輸入句的語言學註記

- 在類神經網路加入更多資訊很簡單
- 通常加入的層面： 詞性、原形、構詞、句法結構、相依關係、詞彙語意
- 每一個層面都轉換成內嵌向量
- 最後每一詞的表達方式
  - 所有層面的內嵌的總合

Words	*the*	*girl*	*watched*	*attentively*	*the*	*beautiful*	*fireflies*
詞性	DET	NN	ADV	VFIN	DET	JJ	NNS
原形化	*the*	*girl*	*watch*	*attentive*	*the*	*beautiful*	*firefly*
構詞	-	SING.	PAST	-	-	PLURAL	
名詞組	BEGIN	CONT	OTHER	OTHER	BEGIN	CONT	CONT
動詞組	OTHER	OTHER	BEGIN	CONT	CONT	CONT	CONT
句法相依	*girl*	*watched*	-	*watched*	*fireflies*	*fireflies*	*watched*
相依關係	DET	SUBJ	-	ADV	DET	ADJ	OBJ
語意	-	ACTOR	-	MANNER	-	MOD	PATIENT
	-	HUMAN	VIEW	-	-	-	ANIMATE

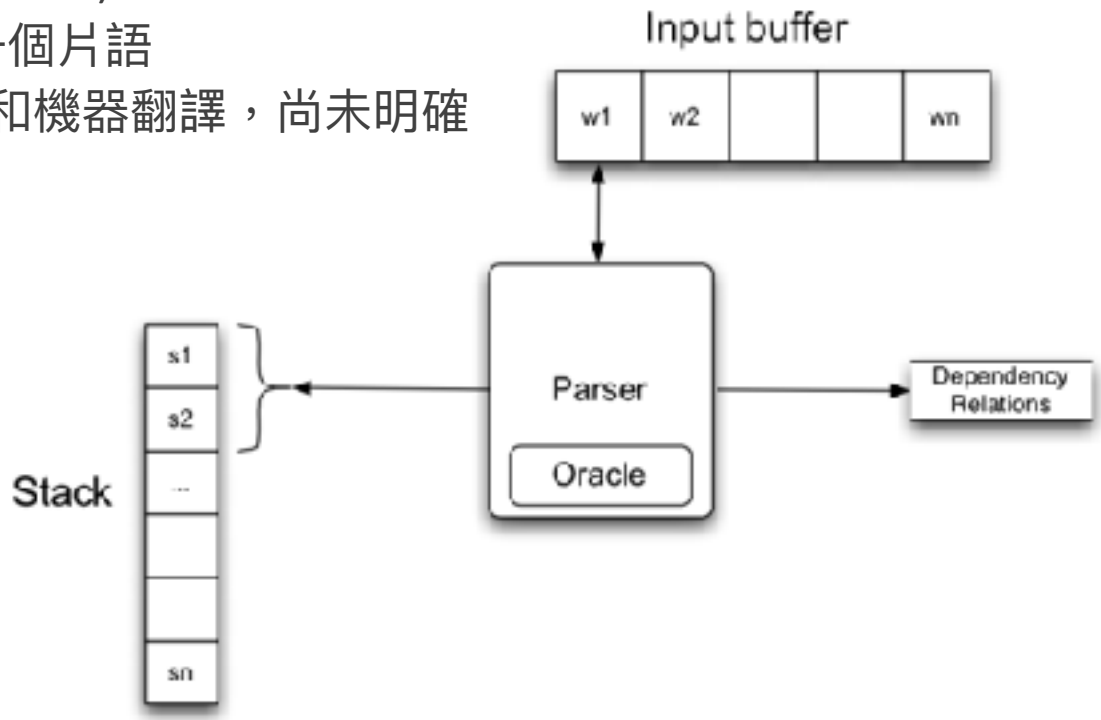
6.8.2 輸出句的語言學註記

- 在輸出端加入句法的註記
  - 可保證輸出句的整體合法性
  - 通常用句法剖析束代表句法 (但有困難，所以用線性化)
- 通常線性化剖析樹—用 “(NP” 和 “)" 代表片語開始和結束
  - 強迫序列到序列模型產生合乎句法的句子 (和註記)

Sentence	*the girl watched attentively the beautiful fireflies*
Syntax tree	
Linearized	(S (NP (DET *the*) (NN *girl*)) (VP (VFIN *watched*) (ADVP (ADV *attentively*)) (NP (DET *the*) (JJ *beautiful*) (NNS *fireflies*))))

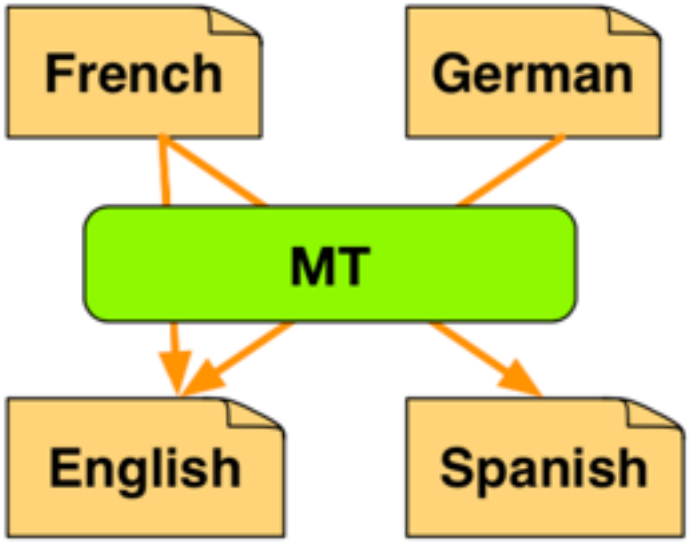
6.8.3 結構化語言學模型

- 最佳的類神經的句法剖析，並非 sequence to sequence 的模型
- 而是，傳統的 left-to-right push-down automata
  - 維持一個未完成片語的堆疊
  - 處理下一詞
    - 放入堆疊，加入未完成片語 (SHIFT)
    - 完成一個片語 (REDUCE)
    - 在堆疊上，開始一個片語
- 新作法：結合句法剖析和機器翻譯，尚未明確
- 未來的研究方向



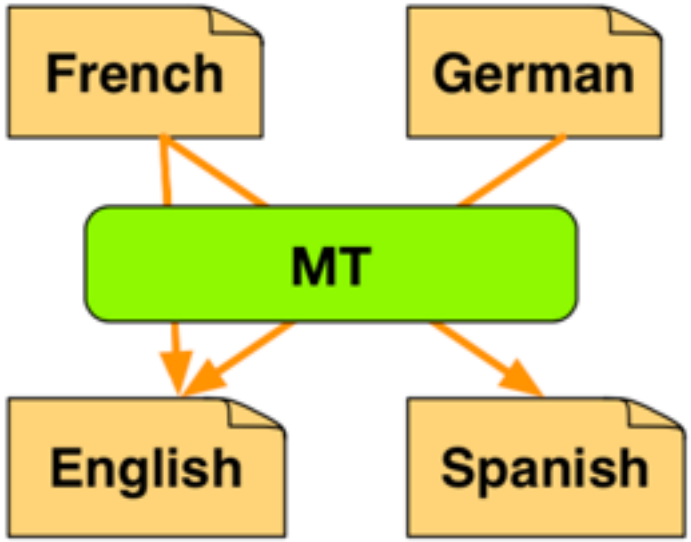
6.9 訓練多個語言配對 Language Pairs

- 多語言機器翻譯系統，可以一次一個配對訓練，例如
  - 先訓練法英系統
  - 再訓練法西系統
  - 最後訓練德英系統
- 產生一個系統，可以同時翻譯多個語言配對
- 甚至，在沒有資料的狀況下，可以將德語翻譯成西班牙語



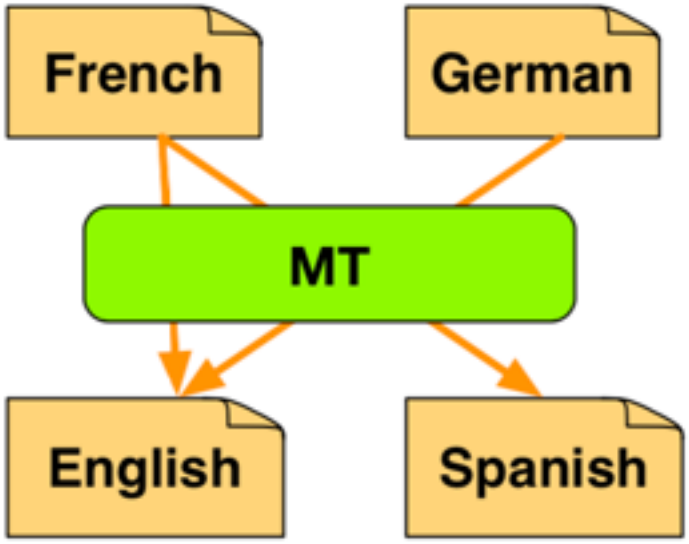
6.9.1 多重輸入語言

- 訓練類神經機器翻譯模型很簡單
  - 直接把兩個平行語料庫接起來 (例如德英＋法英)
  - 輸入的詞彙含有德語、法語
- 比單獨訓練兩個模型的優點？
  - 受惠於兩個平行語料庫中較多英語資料—學到比較好的語言模型
  - 兩個平行語料庫的資料比較多元，語言模型也比較強韌



6.9.2 多重輸出語言

- 和多重輸入語言一樣，把法—英和法—西資料接起來同時訓練
- 執行時加入標籤 (如 [SPANISH] 在輸入句前) 表示翻譯目標語言
- 訓練三組模型 (德英，法英，法西) 可翻譯新配對 (德西)
- 目標語言記號，如 [SPANISH]也可改來代表
  - 領域
  - 語氣 (禮貌)



6.9.3 共享元件

- 可共享元件
  - 如輸入語言相同，可共享編碼器
  - 如輸出語言相同，可共享解碼器
  - 注意機制，可共享於所有語言配對
- 共享元件也就是使用相同的參數值、權重值
- 編碼器可用單語 (來源語) 的資料集訓練
  - 需加新訓練目標 (例如，語言模型的 cross-entropy)
- 解碼器可用單語 (目標語) 資料集，獨立訓練
  - 沒有編碼器傳過來的文脈狀態 (空白掉)
  - 學到：忽視輸入句子，而只是當做目標語語言模型
  - (可能作為起始權重值，最後還是需要用平行資料訓練)

資源

- Keras 範例
  - https://github.com/keras-team/keras
  - https://github.com/awslabs/keras-apache-mxnet/blob/master/examples/lstm_seq2seq.py
- TensorFlow 範例
  - nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf
  - sites.google.com/site/acl16nmt/home/resources
  - www.tensorflow.org/tutorials/
    - github.com/tensorflow/nmt
- PyTorch 範例
  - https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html