第 5 章
片語為本模型

教科書網站：www.statmt.org/book/

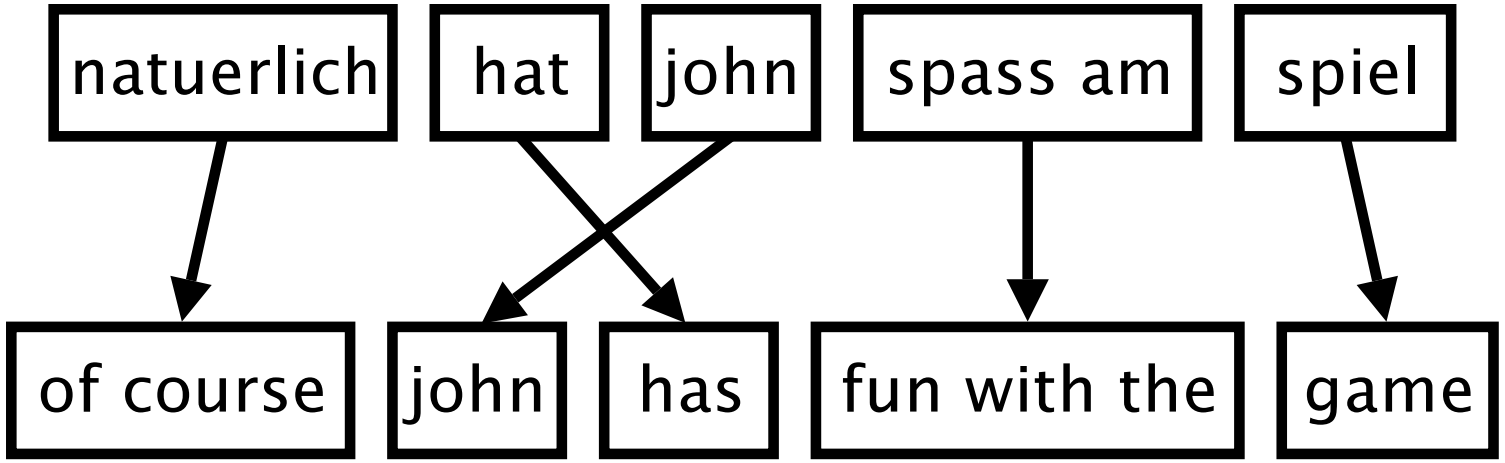
參考課程網站：mt-class.org/jhu/syllabus.html

Oct. 9, 2018

動機：為什麼不用「詞」要用「片語」

- 詞彙為本模型，以「詞」為最小單位（原子）進行翻譯
- 片語為本模型，以「片語」（其實是 n-連詞）為最小單位進行翻譯
- 優點：
  - 可以更合理處理「多到多」的非組合性片語 non-compositional phrases
  - 可以有效運用局部文脈（幫助解歧，虛詞增刪、翻譯選擇）
  - 資料愈多，可以學習愈長、愈有效的片語
  - 研究先驅：Phillis Keohn
  - 後效發展：句法、類神經
- 直到 2016 年，可以說是標準作法 Standard Model
  - Google Translate，Microsoft Bing 以及很多單位

片語為本模型 Phrase-Based Model



- 外語句子，分割 segment 成片段（ n-連詞）
- 把每個外語句的「片語」翻譯成英語（這部份需要學習）
- 把「翻譯片段」的順序重排
- 隨意選擇「片語」（所有排列組合），由左至右輸出翻譯 $\rightarrow$ 順序重排

片語翻譯表 Phrase Translation Table

- 主要知識來源：含片語、翻譯、機率 ($\phi(\bar{e}|\bar{f})$) 的「片語翻譯表」
- 例子: natuerlich 的（片語）翻譯

片語	翻譯	機率
natuerlich	of course	.50
natuerlich	naturally	.30
natuerlich	of course ,	.15
natuerlich	, of course ,	.05

實際發生的例子

- 從 Europarl 語料庫學到的詞彙翻譯 （以德語den Vorschlag為例）

德文	英語	$\phi(\bar{e}\|\bar{f})$	英語	$\phi(\bar{e}\|\bar{f})$
den Vorschlag	the proposal	.62	the suggestions	.11
den Vorschlag	's proposal	.11	the proposed	.11
den Vorschlag	a proposal	.034	the motion	.0091
den Vorschlag	the idea	.025	the idea of	.0091
den Vorschlag	this proposal	.023	the proposal ,	.0068
den Vorschlag	proposal	.021	its proposal	.0068
den Vorschlag	of the proposal	.016	it	.0068
den Vorschlag	the proposals	.016	...	...

  - 翻譯的詞彙變化 lexical variation (proposal vs suggestions)、 構詞變化 (proposal vs proposals)、虛詞有無 (the, a, ...)、雜訊 (it)

這是語言學所稱的「片語」 phrases?

- 模型考慮任何 n-連詞，而不受限於語言學勝認定的片語 (名詞片語、動詞片語、介詞片語等)
- 非語言學片語的例子：

  spass am → fun with the （不完整的名詞＋介詞片語）

- 介詞前的名詞（spass → fun）常常有助於介詞的翻譯（am → with）
- 實驗顯示，加上「語言學片語」的限制，會讓機器翻譯品質降低

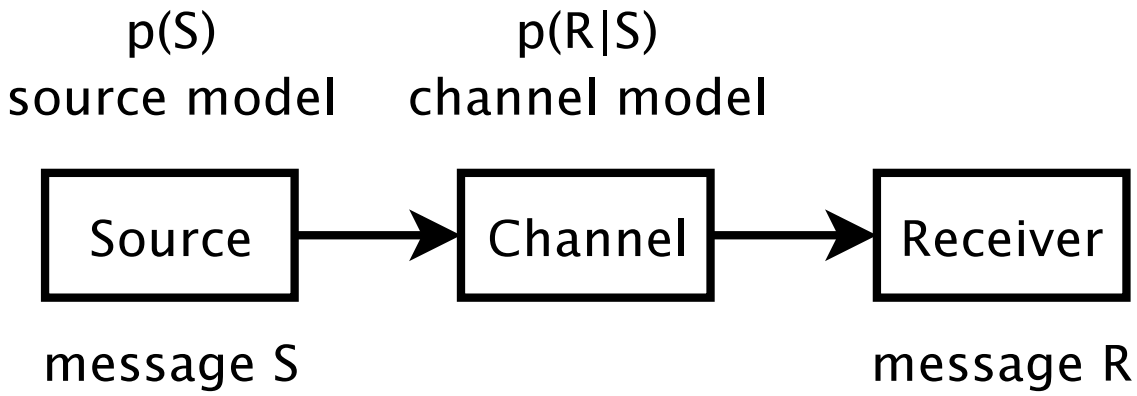
雜訊通道模型 Noisy Channel Model

- 把語言模型整合到機器翻譯系統
- 如何：貝氏定理 Bayes rule

$$\begin{aligned} \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) &= \text{argmax}_{\mathbf{e}} \frac{p(\mathbf{f}|\mathbf{e})\, p(\mathbf{e})}{p(\mathbf{f})} \\ &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{f}|\mathbf{e})\, p(\mathbf{e}) \end{aligned}$$

- 因為 p(**f**) 是輸入，數值不變，所以可以省略

雜訊通道模型 Noisy Channel Model



- 運用貝氏定理，得到 noisy channel model
  - 看到有雜訊的訊息 R (外語 **f**)
  - 模型描述如何產生這樣的雜訊 (翻譯模型)
  - 模型描述正確的訊息 S 如何產生 (語言模型)
  - 目的：如何把 R 轉換（解碼）為 S (英語句 **e**)

雜訊通道模型細節：機率模型

- 貝氏定理

$$\begin{aligned}\mathbf{e}_{\text{best}} &= \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}}\ p(\mathbf{f}|\mathbf{e})\ p_{\text{LM}}(\mathbf{e})\end{aligned}$$

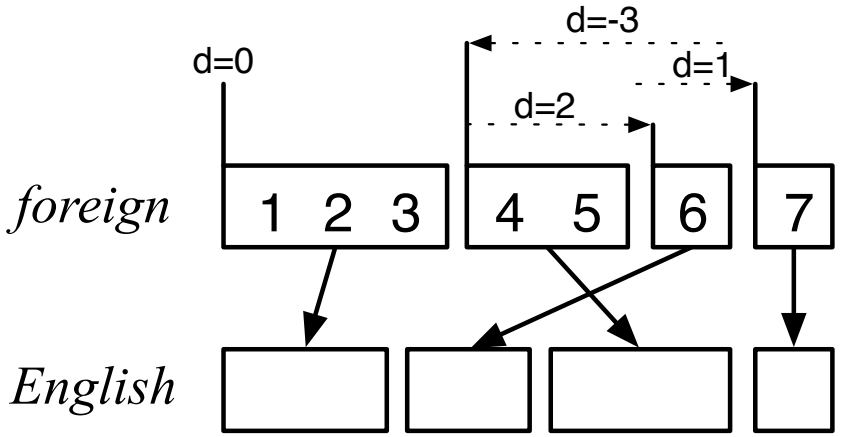
  - 翻譯模型 $p(\mathbf{e}|\mathbf{f})$
  - 語言模型 $p_{\text{LM}}(\mathbf{e})$

- 分解翻譯模型

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\ d(start_i - end_{i-1} - 1)$$

  - 片語翻譯機率 ϕ
  - 詞序重組機率 d

移動量為本的重組 Distance-Based Reordering



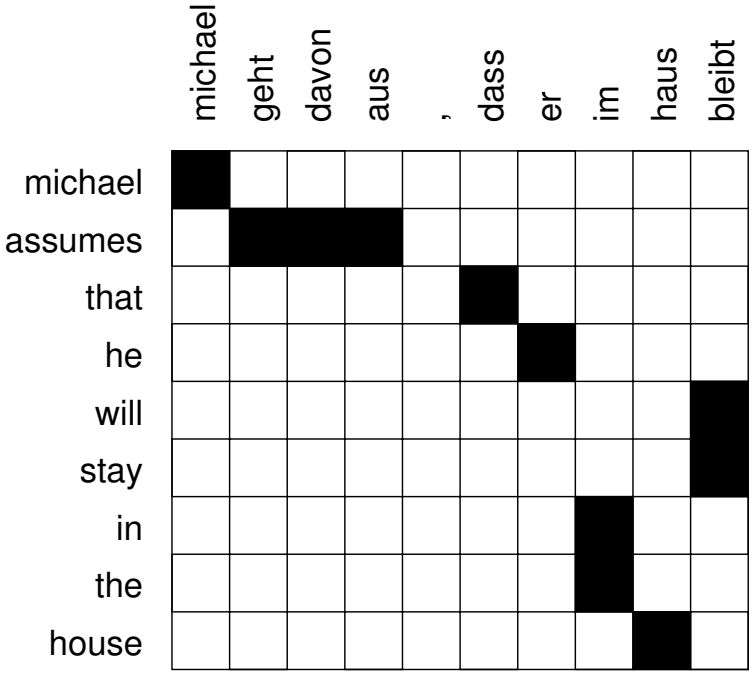
片語	翻譯	移動	距離
1	1–3	從最左邊 1 開始	0
2	6	往前跳過 4–5	+2
3	4–5	往後跳回 4–5	-3
4	7	往前跳過 6	+1

- 評分函數 scoring function：$d(x) = \alpha^{|x|}$ – 距離的對數函數

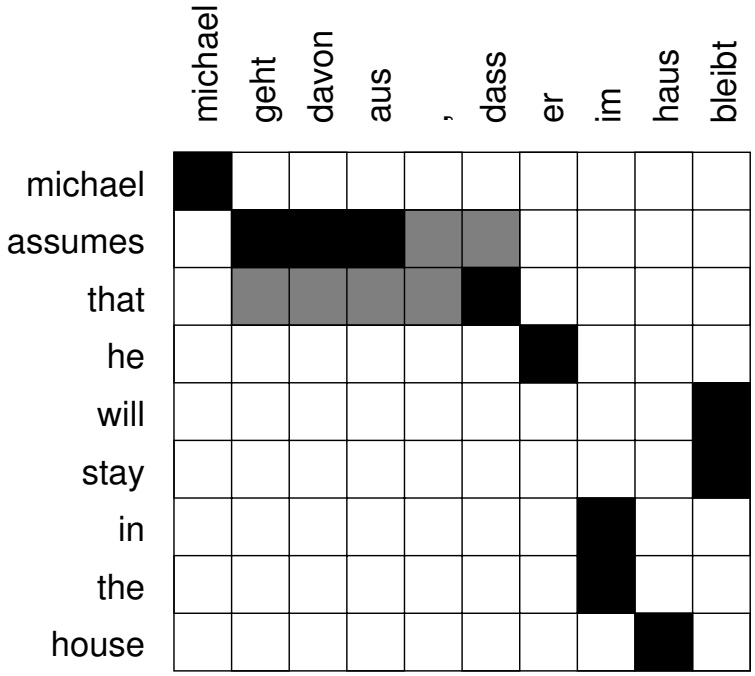
習得片語翻譯表 Learning PTT

- 任務：由給予的平行語料庫，學習得到片語翻譯模型
- 三階段：
  - 詞彙對齊：使用 IBM 幾個模型（或其他方法）
  - 擷取片語到片語的配對
  - 賦予每個片語配對，一個分數

詞彙對齊 Word Alignment

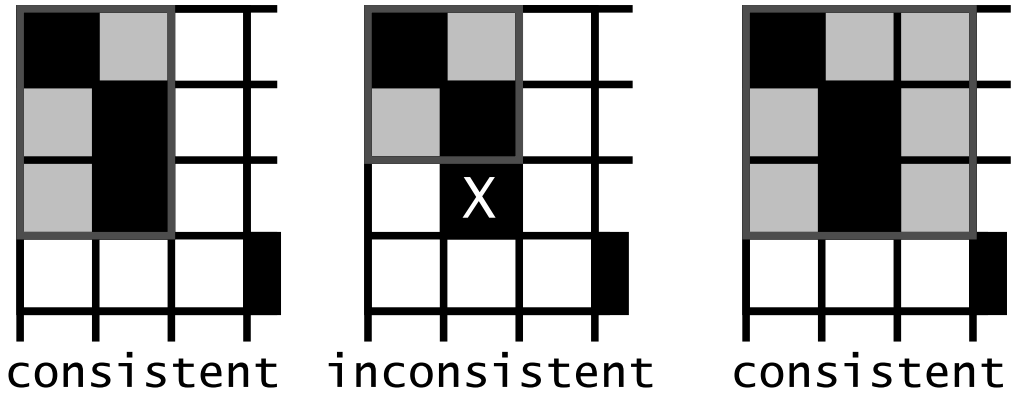


擷取片語到片語的配對



- 擷取和詞彙對齊不衝突的片語： assumes that ||| geht davon aus , dass

一致性：片語對齊和詞彙對齊有沒有衝突



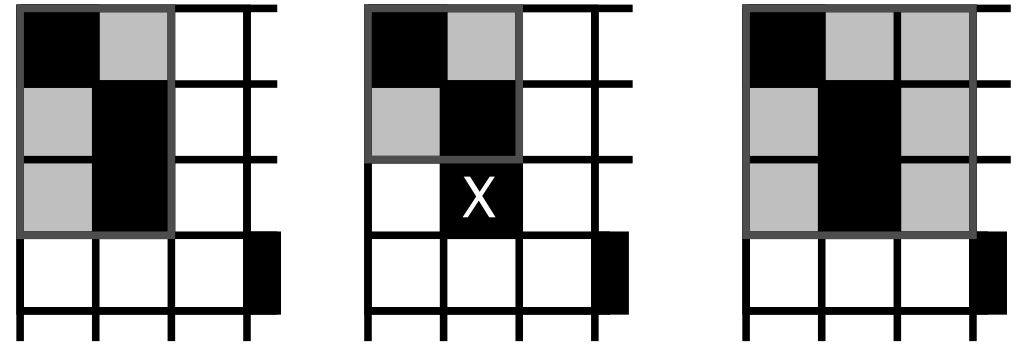
沒問題

違反了
外部有詞
沒對應

沒問題
0-1 沒對齊
沒關係

- 所有涉及的詞都有對應，沒有對應落到片語的外部

一致性 consistency 的正式定義



- 若 $\bar{f}$ 的詞 $f_1, ..., f_n$ 透過 A 的對應都在 $\bar{e}$ 的詞 $e_1, ..., e_n$ 內，反之亦然
- 則片語配對 $(\bar{e}, \bar{f})$ （稱為一致矩型）和對齊 A 是一致的 (不違反、不衝突)

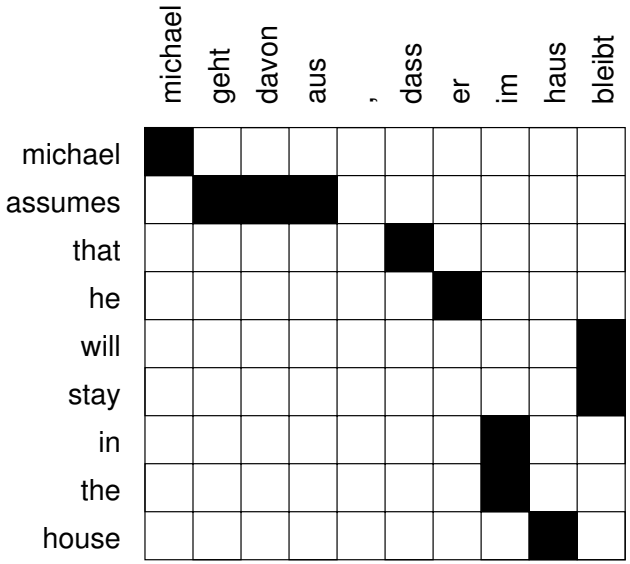
$$(\bar{e}, \bar{f}) \text{ consistent with } A \Leftrightarrow$$

$$\forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f}$$

$$\text{AND } \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e}$$

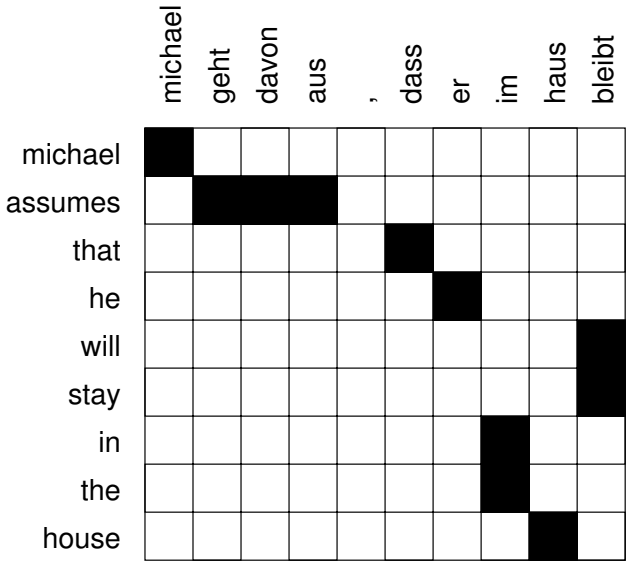
$$\text{AND } \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i, f_j) \in A$$

擷取片語配對



- 最小的片語配對：
  - michael ||| michael　　assumes ||| geht davon aus / geht davon aus ,
  - he ||| er　　that ||| dass / , dass　　he ||| er
  - will stay ||| bleibt　　in the ||| im　　house ||| haus
- 只要有一個詞（如德語的逗號) 沒有對應，就會導致很多配對

更長的片語配對



michael assumes ||| michael geht davon aus / michael geht davon aus ,
assumes that ||| geht davon aus , dass ; assumes that he ||| geht davon aus , dass er
that he ||| dass er / , dass er ; in the house ||| im haus
michael assumes that ||| michael geht davon aus , dass
michael assumes that he ||| michael geht davon aus , dass er
michael assumes that he will stay in the house ||| michael geht davon aus , dass er im haus bleibt
assumes that he will stay in the house ||| geht davon aus , dass er im haus bleibt
that he will stay in the house ||| dass er im haus bleibt ; dass er im haus bleibt ,
he will stay in the house ||| er im haus bleibt ; will stay in the house ||| im haus bleibt

片語翻譯的評分

- 擷取片語配對：從語料庫收集所有的片語配對
- 片語配對評分：計算片語配對的機率值（評分）
- 以相對次數來評分：

$$\phi(\bar{f}|\bar{e}) = \frac{\mathsf{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \mathsf{count}(\bar{e}, \bar{f}_i)}$$

- 會有很多配對次數很低 (1次) ——不可靠

片語表的大小

- 片語翻譯表通常比語料庫大
  - 雖然，通常限制片語長度（例如，最多 7 詞）
  - 會太大，超過記憶體的容量
- 擷取時：如果太大
  - 擷取後，存入硬碟。然後按片語排序，依序計算片語翻譯之條件機率
  - 常用 suffix arrays 儲存，建立快速搜尋的索引
- 解碼時
  - 利用索引，快速查詢句中所有適用片語
  - 取回翻譯、機率

有權重的模型 Weighted Model

- 標準模型中有三子模型
  - 片語翻譯模型 phrase translation model $\phi(\bar{f}|\bar{e})$
  - 重組模型 reordering model d
  - 語言模型 language model $p_{LM}(e)$

$$e_{\text{best}} = \text{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\ d(start_i - end_{i-1} - 1) \prod_{i=1}^{|\mathbf{e}|} p_{LM}(e_i|e_1...e_{i-1})$$

- 三個子模型，重要性不同

- 用三個權重代表重要性 weights λ_ϕ, λ_d, λ_{LM}

$$e_{\text{best}} = \text{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)^{\lambda_\phi}\ d(start_i - end_{i-1} - 1)^{\lambda_d} \prod_{i=1}^{|\mathbf{e}|} p_{LM}(e_i|e_1...e_{i-1})^{\lambda_{LM}}$$

線性對數模型 Log-Linear Model

- 線性對數模型就是權重模型其中的一種：

$$p(x) = \exp \sum_{i=1}^{n} \lambda_i h_i(x)$$

- 其中有特徵函數 feature functions（即機率函數）各有對應的權重
  - 特徵函數數量 $n = 3$
  - 隨機變數 variable $x = (e, f, start, end)$
  - 特徵函數-1 $h_1 = \log \phi$
  - 特徵函數-2 $h_2 = \log d$
  - 特徵函數-3 $h_3 = \log p_{\mathsf{LM}}$

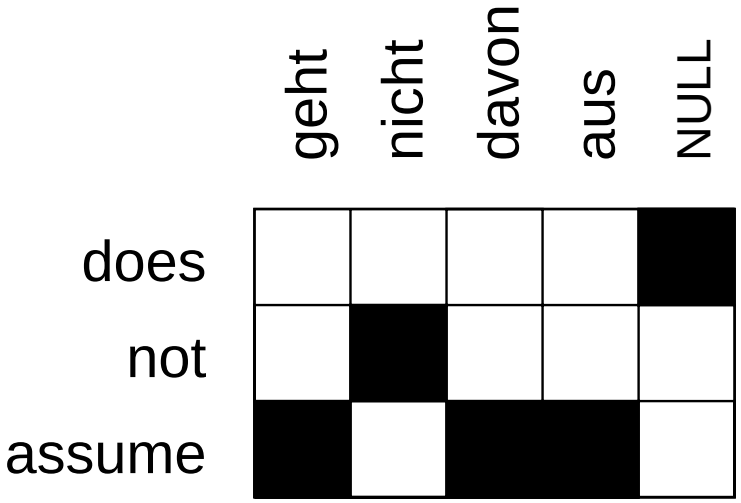
線性對數模型的公式

- 加總加權的特徵函數值
- 取總數的指數

$$p(e,a|f) = \exp(\lambda_\phi \sum_{i=1}^{I} \log \phi(\bar{f}_i|\bar{e}_i) +$$
$$\lambda_d \sum_{i=1}^{I} \log d(a_i - b_{i-1} - 1) +$$
$$\lambda_{LM} \sum_{i=1}^{|\mathrm{e}|} \log p_{LM}(e_i|e_1 ... e_{i-1}))$$

用更多特徵函數，效果更好

- 雙向片語翻譯條件機率：$\phi(\bar{e}|\bar{f})$ and $\phi(\bar{f}|\bar{e})$
- 因為低頻片語配對的機率不可靠 $\rightarrow$ 加入「詞翻譯機率」的特徵函數



$$\text{lex}(\bar{e}|\bar{f},a) = \prod_{i=1}^{\text{length}(\bar{e})} \frac{1}{|\{j|(i,j)\in a\}|} \sum_{\forall(i,j)\in a} w(e_i|f_j)$$

還可以再加特徵函數

- 因為語言模型有偏好短的翻譯的偏差

  $\rightarrow$ 加入詞數的特徵函數：$\mathsf{wc}(e) = \log |\mathbf{e}|^{\omega}$

- 因為片語的長短（少多）對翻譯品質有影響

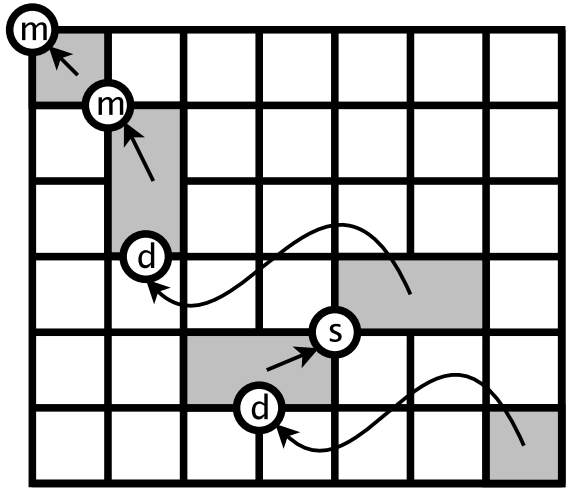
  $\rightarrow$ 加入片語數的特徵函數：$\mathsf{pc}(e) = \log |I|^{\rho}$

- 多種語言模型（有各自的權重）

- 多種翻譯模型（有各自的權重）

- 其他知識來源

詞彙化重組模型 Lexicalized Reordering

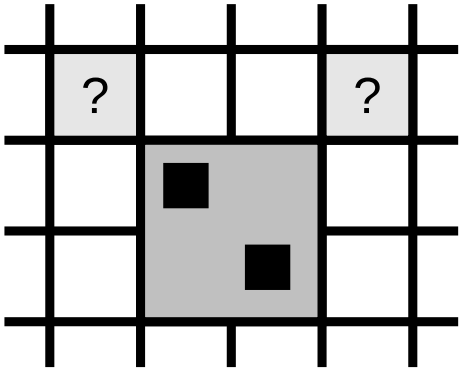


- 距離（移動量）為本的重組模型很「弱」（預測能力查）→ 加入片語為條件，比較可以預測重組的偏好 （移動量的機率）

- 考慮 (m) 單調 monotone, (s) 交換 swap, (d) 不連續 discontinuous 等三種

$$p_o(\text{orientation}|\bar{f},\bar{e}) \quad where \quad \text{orientation} \in \{m,s,d\}$$

習得詞彙化重組模型



- 在片語擷取過程中，收集方向的資訊
  - 如果左上角有詞到詞的對應 → **monotone**
  - 如果右上角有詞到詞的對應 → **swap**
  - i如果左、右上角都沒有詞到詞的對應 → **discontinuous**

習得詞彙化重組機率 Lexicalized Reordering

- 用相對次數，計算機率

$$p_o(\text{orientation}) = \frac{\sum_{\bar{f}} \sum_{\bar{e}} count(\text{orientation}, \bar{e}, \bar{f})}{\sum_o \sum_{\bar{f}} \sum_{\bar{e}} count(o, \bar{e}, \bar{f})}$$

- 用「非詞彙化」方向模型，來平滑化 smoothing 模型機率 $p(\text{orientation})$

- 避免 0 次數 0 機率的未出現方向

$$p_o(\text{orientation}|\bar{f}, \bar{e}) = \frac{\sigma\ p(\text{orientation}) + count(\text{orientation}, \bar{e}, \bar{f})}{\sigma + \sum_o count(o, \bar{e}, \bar{f})}$$

對 PBSMT 批評：片語的分段似乎是任意的

- 如果可以好幾種「片語分段」，為何選其中一種？

spass am	piel	vs.	spass	am spiel

- 不知道何時應該選長詞還是幾個短詞?

spass am	piel	vs.	spass	am	spiel	vs.	spass am spiel

- 以上的問題都沒有解答

對 PBSMT 批評：片語翻譯互相獨立的假設

- 詞彙的文脈，只在片語內考慮

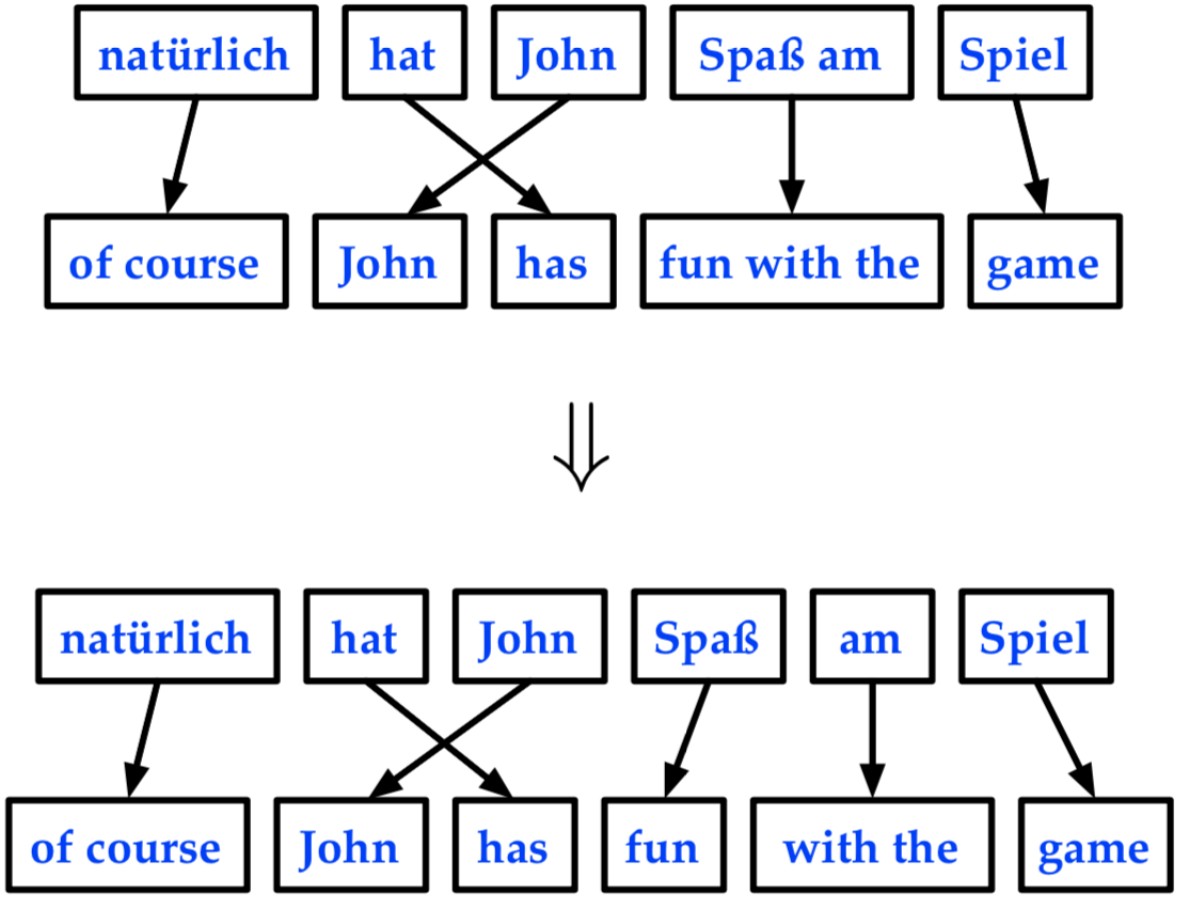
  spass am → fun with

- 片語和片語之間，並無文脈的考慮

  ? spass am ? → ? fun with ?

- 在「詞彙化重組模型」考慮到片語本身，但是鄰近的片語卻沒有納入考慮

如何分段？最少片語配對


natürlich
hat
John
Spaß am
Spiel
of course
John
has
fun with the
game
natürlich
hat
John
Spaß
am
Spiel
of course
John
has
fun
with the
game

獨立性？考慮以下的操作動作序列

o_1	Generate(natürlich, of course)	natürlich ↓ of course
o_2 o_3	Insert Gap Generate (John, John)	natürlich ↓ ☐ John of course John
o_4 o_5	Jump Back (1) Generate (hat, has)	natürlich hat ↓ John of course John has
o_6	Jump Forward	natürlich hat John ↓ of course John has
o_7	Generate(natürlich, of course)	natürlich hat John Spaß ↓ of course John has fun
o_8 o_9	Generate(am, with) GenerateTargetOnly(the)	natürlich hat John Spaß am ↓ of course John has fun with the
o_{10}	Generate(Spiel, game)	natürlich hat John Spaß am Spiel ↓ of course John has fun with the game

動作序列模型 Operation Sequence Mode

- 動作序列模型
  - 產生 (片語翻譯)
  - 只產生目標 target
  - 只產生來源 source
  - 插入空隙 insert gap
  - 往後跳 jump back
  - 往前跳 jump forward

- N-gram 方向系列模型, 例如，5-連 模型：

$$p(o_1) \;\; p(o_2|o_1) \;\; p(o_3|o_1,o_2) \;\; \ldots \;\; p(o_{10}|o_6,o_7,o_8,o_9)$$

實務的考慮

- 動作系列模型 Operation Sequence Model (OSM) 是附加的特徵函數
- 和 PBSMT 相比，PBSMT+OSM 顯著地改進翻譯品質
- 所以，最新的系統都包含了 OSM 模型

用 EM 演算法來訓練片語為本模型

- 以上，我們描述了的「」方式，來建立片語翻譯表 presented a heuristic set-up to build phrase translation table
(雙向詞彙對應、對稱化、抽取片語翻譯、片語翻譯評分)

- 替代方案：用 EM 演算法，直接對應片語
  - 初始化：平均分布模型，所有的片語配對 $\phi(\bar{e}, \bar{f})$ 的機率都一樣
  - 期望值步驟：
    * 在每個句子配對中，估計所有片語配對的機率值
  - 最大似然步驟：
    * 收集片語配對 $(\bar{e}, \bar{f})$ 的加權次數（期望值），權重＝對應機率
    * 用期望值估算、更新聯合詞彙翻譯機率 $p(\bar{e}, \bar{f})$ （以及條件機率）

- 注意：這個方法很容易過度適應（overfits）學到太多片語配對，布滿句子

結語

- 片語為本統計式機率翻譯模型
- 如何訓練模型
  - 詞彙對齊 word alignment
  - 抽取片語配對 phrase pair extraction
  - 片語配對評分 phrase pair scoring
- 線性對數模型 Log linear model
  - 把子模型 sub-models 當做特徵函數 feature functions
  - 加上 詞彙翻譯特徵函數和權重
  - 詞、片語的長度，都可當做附加的特徵函數
- 詞彙化重組模型 Lexicalized reordering model
- 可以用 EM 演算法習得片語模型
- 動作序列模型 Operation sequence model 顯著改進翻譯品質