Chapter 2

Words, Sentences, Corpora
詞、句子、語料庫

Statistical Machine Translation

Chapter 2 詞、句子、語料庫

- 這一章寫給沒有「自然語言處理」基礎的讀者

statistical machine translation

- 介紹和 SMT 相關的語言學概念
- 從「詞 words」和其語言性質開始
- 接著談「句子 sentences」、句法、語意
- 討論「文本語料庫 text corpora」在發展 SMT 系統的角色
- 描述取得語料庫、處理語料庫的基本方法.

2.1 詞

- 詞 = 語言語意的「原子」
- 以 house (一般的意思) 為例
  - rectangular building with a roof and smoking chimney
  - surrounded by grass and trees
  - inhabited by a happy family
- 'house' 的意義，隨著前後文（文脈）而改變
  - Her grandmother's house 和 White House 不同
  - 雖然各種 house (e.g., publishing house) 意義不同
  - 但是都和基本的意義「rectangular building」有關

詞的標記 marking of Words

- 英語文本中，「詞」的兩側都有空白 space，但英語語音中，我們雖然不在「詞」和「詞」之間停頓，但是母語者似乎可以聽到分隔的「詞」
- 但是若不懂外語就聯成一片，聽不出來詞
- 華語（中文）就沒有空格
- 但是，英文有沒有空格也不完全一致
  - bank vs. riverbank, Joe's, doesn't
- 詞＋詞的意思是否具有組合性 compositional——不一定
  - bird cage = bird + cage + (to keep animal in)
  - toy elephant $\neq$ toy + elephant

2.1.1 分詞作業 Tokenization

- 分詞作業把原始文本切分成「詞」（包括標點）
- 對拉丁字母的語言，如英文比較簡單
- 對於華語，就比較困難 (中文斷詞是重要研究，有問題還未完全解決)
  - 需要一個所有詞彙的辭典（可能嗎？需要嗎？）
  - 需要解決歧義 ambiguity，如「才｜能｜打開」 v.s. 「很｜有｜才能」

英語分詞作業的一些議題

- 有困難的例子:
  - Joe's, doesn't co-operate
  - so-called, high-risk
- 有些複合詞 riverbank, uninvite 可以拆開來比較好
- 可能也需要大小寫正規化 (全小寫 lowercasing 或 正常大小寫 trucasing)
  - 如此可以縮小辭典表 vocabulary
  - 更可靠的估算「詞彙翻譯機率」translation probability

反分詞 Detokenizing 以及 恢復大小寫 Recasing

- 反分詞 = 把標點連到詞後（沒有空格），以及恢復分詞前的狀況
- 恢復大小寫 = 把正規化 (全小寫) 文本，改成適當的大小寫（句首、專有名詞等，如 Mr Fisher)
- 在機器翻譯中，加入分詞工具
  - 分詞＋改小寫 tokenize and lowercase
  - 機器翻譯 (用全小寫)
  - 反分詞＋恢復大小寫

2.1.2 詞的機率分布 Distribution

- 英語的詞彙有什麼現象?
  - 新詞不斷出現或鑄造出來 coined (詞彙如貨幣)
  - 就詞會過時，不再使用
  - 英語詞彙變動激烈 fluid

- 那我們來問英語有多少詞彙（例如在 Europarl 語料庫中）

數字不算在詞彙裡

- Europarl 有 29 million 詞，1 million 句，詞彙表 86,699 詞 (分詞、小寫化後、不計算數字)

一般人詞彙量：兩萬

2.1.3 詞性 Parts of Speech

- 詞有不同的詞性 parts of speech (POS) 就像戲劇中扮演不同的角色 roles
  - 名詞 nouns (實詞、開放集) 指涉世上實體 (house) 或抽象對象 (freedom)
  - 介詞 (虛詞或功能詞、封閉集合) 表示句子中名詞、動詞的關係 (A in B)
- 實詞容易產生新詞，廣受接受，通常新詞是發生在我們周遭的新事物 (e.g., Google, web site or website)
- 名詞容易產生新詞，動詞比較不容易（unfriend）
- 虛詞有封閉性，不容易產生新詞

實詞 Content 虛詞 Function 對 MT 挑戰不同

- 實詞通常太多，而雙語語料庫對於其翻譯的資訊太少 (很多只出現一次)
- 虛詞的問題：不同語言中，虛詞的運作很不同
  - 有時甚至在目標語言（如華語中）沒有適當方式，來對應英語的虛詞
  - 日文的後置詞 (wa) 和英文很不同
  - 翻譯多元，難以掌握，常受前後文影響 context-sensitive

詞性分類

- 名詞指涉世界上抽象、實題對象
  - 普通名詞 (單數: house/NN, 複數: houses/NNS)
  - 專有名詞 nouns (單數: Britain/NP, 複數: Americas/NPS)

- 動詞代表行動 actions 有時態 tense
  - 基本型: go/VB
  - 簡單過去: went/VBD
  - 過去分詞 participle: gone/VBN,
  - 進行式 Gerund: going/VBG
  - 第三人稱單數現在式 3rd person singular present: goes/VBZ
  - 其他 singular present: am/VBP.
  - 特殊狀況: 助詞 can/MD, 動詞＋分詞：switch on/RP (particles)

- 形容詞代表名詞的性質、特徵 properties
  - 一般: green/JJ
  - 表較級 Comparative: greener/JJR 最高級 Superlative: greenest/JJS
- 副詞代表動詞或形容詞的性質
  - 一般: happily/ RB
  - 比較: ran faster/RBR
  - 最高: ran fastest/RBS
  - 疑問副詞 Wh-adverbs: how/WRB fast
- 定詞 (also called 冠詞) 指涉特定名詞
  - 一般: the/DT house
  - 前定詞: all/PDT the houses
  - 疑問定詞: which/WDT.

- 代名詞代表前面題過的名詞
  - 人稱代名詞 Personal pronoun: she/PP
  - 所有格代名詞: her/PP$
  - 疑問代名詞: who/WP
  - 疑問代名詞所有格: whose/WP$

  .

- 介系詞（在名詞、子句前）表示在句子的角色（工具、共同）: from/IN here
  - 特殊詞性： to/TO.
- 對等連接詞 Coordinating conjunctions: and/CC.
- 數字: 17/CD

- 所有格符號：Joe 's/POS
- 條列符號 List item markers: A./LS
- 特殊符號: $/SYM
- 外來語（拉丁、法文）：de/FW facto/FW
- 感嘆詞 Interjections: oh/UH

2.1.4 構詞學 Morphology

- 讓意義更精確的兩種作法
  - 修飾語 Modifier: 加形容詞 around it (e.g., add small to house—-small house)
  - 構詞變化：做後綴的小變形 (e.g., add s to house–houses)

- 英文詞常常因為下列原因，而有構詞變形
  - 數量 count (noun singular or plural)
  - 時態 tense (present or past)
  - 人稱＋數量＋時態 (3rd person, singular, present)

時態 Tense 助詞 auxiliary verbs

- 有時候還要加上數量詞（例如：one, many）、助詞
  - 簡單現在（或無時間性）：walk
  - 現在進行：is walking
  - 過去：walked
  - 現在完成：have walked
  - 過去完成：had walked
  - 未來：will walk
  - 未來完成：will have walked

格位 **Case** 代名詞 **prepositions**

- Many languages use morphology to express additional information —- case and gender
- E.g., given the bag of words eat, lion, zebra, we need to indicate who is eating whom
- English indicate the relation by sentence order
- Other (free-order) languages indicate the relation by case

德文的格位很複雜

- German cases: subject and object
  - 1. Der Löwe frißt das Zebra.
  - 2. Den Löwen frißt das Zebra.
  - 3. Das Zebra frißt der Löwe.
  - 4. Das Zebra frißt den Löwen.
- In sentences 1 and 3 the lion is doing the eating (using subject case = der X)
- In sentences 2 and 4 the zebra is doing the eating (using object case = den X)
- Two more German cases
  - Genitive: indicating ownership relationships (i.e., English possessive: lion's food)

其他語言的格位

- Latin has two further cases
  - vocative (for the entity that is addressed as in John, I'm eating dinner!)
  - ablative (used for instance to indicate location)
- Finnish has 15 cases.
- English nouns do not have case markers, but the pronouns have (he, him himself)
- English is in the process of losing morphological case

續

- English uses *sentence order* and *preposition* to indicate the role of nouns
  - I go *to* the house.
  - I go *from* the house.
  - I go *in* the house.
  - I go *on* the house.
  - I go *by* the house.
  - I go *through* the house.
- But, prepositions are ambiguous (movement, location, and time meaning), and different propositions could mean the same thing (time)
  - Let's meet *at* 9am. vs. Let's meet *on* Sunday. vs. Let's meet *in* December.
- Prepositions are difficult to translate, because of mismatch between languages

性別 Gender

- Gender shows up only in pronouns (she, he, her, his) in English

Case	Singular			Plural		
	male	fem.	neu.	male	fem.	neu.
nominative (subject)	der	die	das	die	die	die
genitive (possessive)	des	der	des	der	der	der
dative (indirect object)	dem	der	dem	den	den	den
accusative (direct object)	den	die	das	die	die	die

,

- Figure 2.6 德文定冠詞的構詞變化——數量、格位、性別。造成很多自期異性高（如 der）–(1) singular nominative (2) female singular genitive/dative, (3) plural genitive for any gender.

構詞與機器翻譯

- 複雜的構詞，造成機器翻譯上的困難
- 有些語言透過位置、虛詞來表達詞和詞的關係
- 有些語言卻過過構詞變化來表達詞和詞的關係
- 翻譯這種構詞變化常常很困難
- 構詞變化也會增加詞彙的數量
- 構詞豐富的語言（如德文）翻譯時，就相對比較困難
- 需要額外的資訊 (例如，從輸入中常常無法知道性別的資訊）

2.1.5 詞彙語意學 Lexical Semantics

- 翻譯 = 把一個語言的詞彙、句子，轉換成「意義相等」的另外一個語言的詞彙與句子
- 然而，「意義」是非常難處理的
- 像 $house$ 就不一定是「家、房子」的意義，如$House\ of\ Windsor$, 就是只家族、組織，而非實體的「房子」
  - (Wikipedia) $House\ of\ Windsor$ is the royal house of the United Kingdom ... most prominent member of the $House\ of\ Windsor$ is Queen Elizabeth II
  - (WordNet) the British royal $family$ since 1917

詞彙語意 Word Senses

- 詞常常有不只一個意義 (多義詞 polysemy、同音異義詞homonymy)

plant, can

- 同音異義詞是兩個詞有完全不相關的意思，恰巧同音同形，如 (e.g., *can* 是助詞、名詞罐頭)

- 多義詞是指詞有一組相關的意義，如 house 是家、房子、（苦茶之）家
  - interest: 興趣 (對足球的 interest)
  - intesest: 股份（5% Google 股票)
  - intesest: 利息（interest rate of 4.9%)

- 詞彙語意難以定義，定義難以周全（有人説語意有無限多）

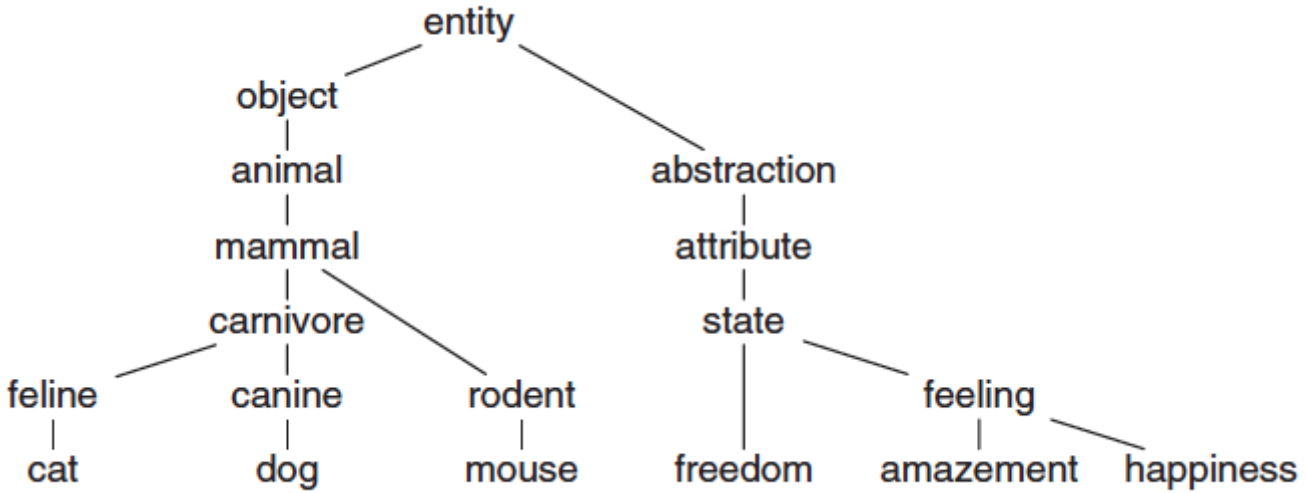
- Is *national interest* 就和已上三個意義都不同——共同的目標 common goal

WordNet記錄的詞彙語意常用於研究

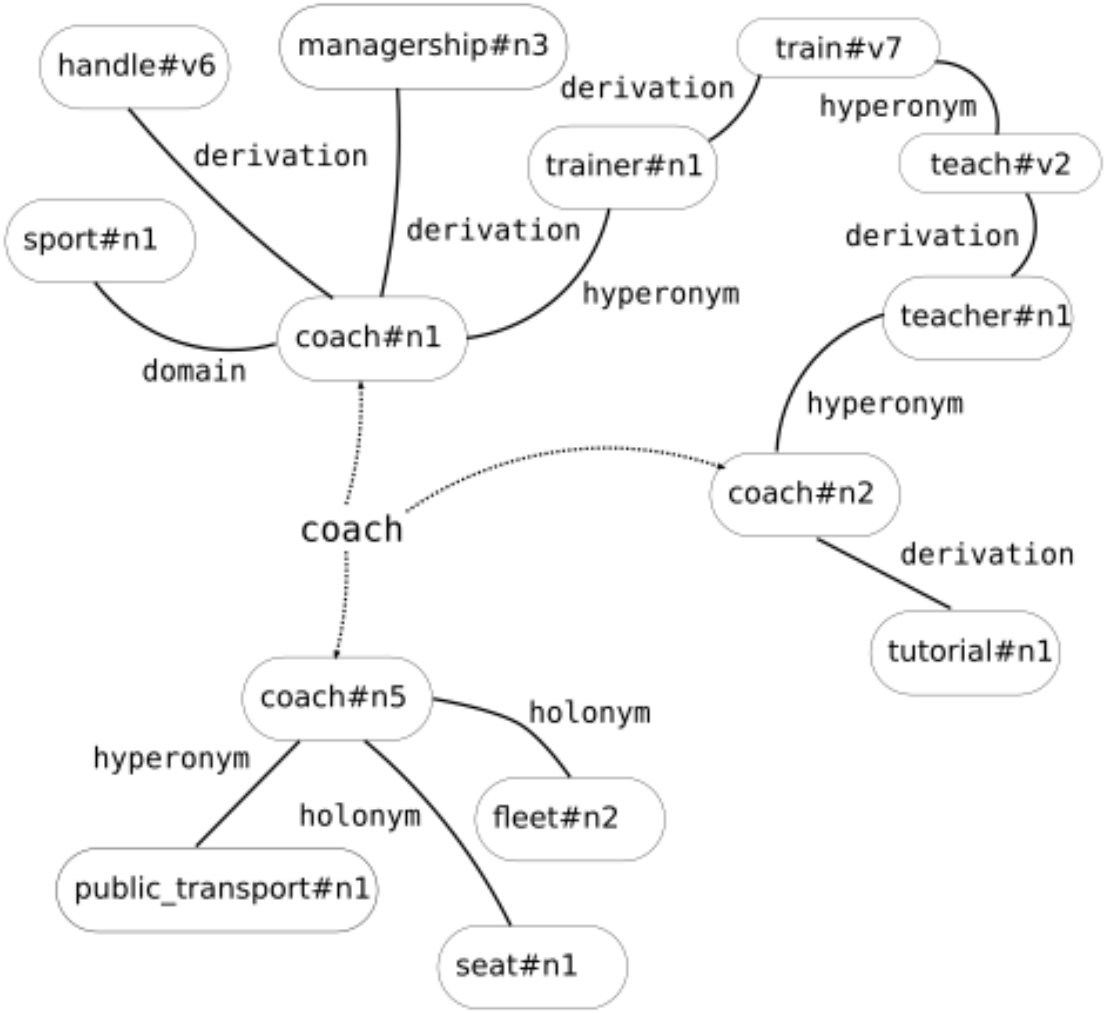
- WordNet 為廣為使用的詞彙語意資料庫 database for lexical semantics
- 根據WordNet *interest* 有 7 個不同的意義
- 或許 7 個有點太多，因為有幾個其實翻譯是一樣的
- 根據翻譯來定義語意比較清楚
- 例如，三個意義有不同的德文翻譯：
  - Interesse (curiosity sense) 興趣
  - Anteil (stake sense) 股份
  - Zins (money sense) 利息

Figure 2.7 Semantic Relations in WordNet

- WordNet （名詞）的主要階層結構有三：is-a (上位詞hypernym), part-of （部分整體）, and member-of （成員群體）



Couch 這個詞的 WordNet 語意關係



2.2 句子

- 句子把詞表達的概念組合成，來陳述事實、事件 event 或命令或疑問
  - 2.2.1 句子結構
  - 2.2.2 文法理論
  - 2.2.3 翻譯句子的結構
  - 2.2.4 句子之間的文脈關係 discourse

2.2.1 句子結構

- 例子 $Jane\ bought\ the\ house.$
- 中心成份 central element 動詞 $bought$
- 主詞 **subject** $= Jane$ (buyer)
- 受詞 **object** $= the\ house$ (object to be bought)
- 有些動詞不需要受詞 (不及物 intransitive), 例如： $Jane\ swims.$
- 有些動詞需要2個受詞，例如：$Jane\ gave\ Joe\ the\ book.$
- 價位 **valency** = 動詞需要的參數數量

續

- 可有可無附加資訊，叫做修飾語 *adjuncts*
- 如：*Jane bought the house from Jim cheaply, without hesitation, yesterday.*
- 修飾語：from Jim, without hesitation (pp), cheaply, yesterday (adv)
- 同一動詞的不同意義 *meanings* 可能會有不同的價位 *valency*（參數個數以及文法規則）
- 參數和修飾語可以愈加愈複雜
- 例如：*Jane bought the house in the posh neighborhood.*
- 例如：*Jane bought the house in the posh neighborhood across the river.*

續

- 遞迴 *recursion* = 人類的語言能力，可以創造層層包孕的結構
- 例如：*[Jane [who recently won the lottery] bought the house] [that was just put on the market].*
- 新定義：句子可以含有一到多個子句，而每個子句各自有動詞、參數、修飾語
- 而參數、修飾語也可以是子句，例如：*I proposed to go swimming.*

續

- 句子的結構常常有歧義，因此難處理，例如：

  1 *Joe eats steak with a knife.*
  2 *Jim eats steak with ketchup.*
  3 *Jane watches the man with the telescope.*

- (1) 句的 with 介詞片語，結構上聯繫到動詞 *eat*

- (1) 句的 with 介詞片語，結構上聯繫到名詞 *steak*

- a(1) 句的 with 介詞片語，結構上可以聯繫到名詞或動詞—-歧義*ambiguity*

續

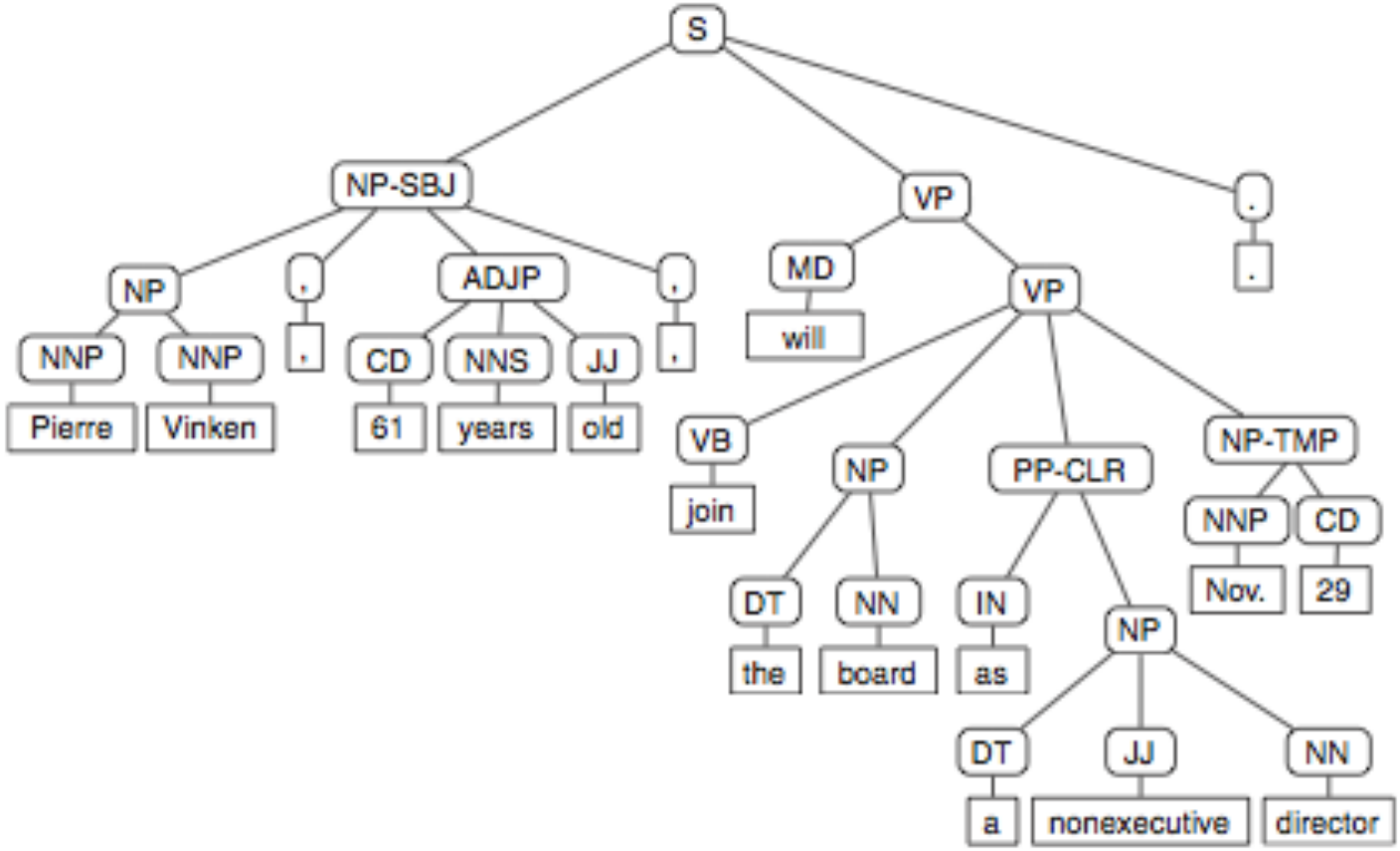
- 連接詞的範圍 *scope* 不清楚時，連接詞 connectives 就引發歧義
- 看看這個例句： *Jim washes the dishes and watches television with Jane.*
- 連結詞 and 的範圍是什麼：
  - (1) [*washes the dishes*] *and* [*watches television with Jane*]
  - (2) [*washes the dishes*] *and* [*watches television*] *with Jane*
- (1) Jane helping with the dishes
- (2) Jane helping with the dishes and joining for television?

2.2.2 文法理論

- 有幾種互相競爭的文法理論（形式結構）
  - 文脈無關文法 Context-free grammars
  - 相依文法 Dependency grammars
  - 詞彙功能文法 Lexical functional grammars
  - 組合份類文法 Combinatory categorical grammar
- 學者發展出各種方法來處理文法
- 這些方法需要有註記文法茲鱒的資料集（如 $Penn\ tree\ bank$（大約100萬詞），就是所謂的剖析樹 parse trees.

Figure 2.8 賓州大學樹庫 Penn tree bank 剖析樹

- *Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 2*



文脈無關文法 Context-free grammars

- CFG =
  - 非終端符號 nonterminals (詞性標籤和片語類別)
  - 終端符號 terminal (詞彙)
  - 文法規則：非終端符號 → 一個（以上）的（非）終端符號
  -
  - S → NP VP
  - NP → NNP ADJP
  - NP → NNP NNP
  - VP → VB NP PP NP
  - VB → join

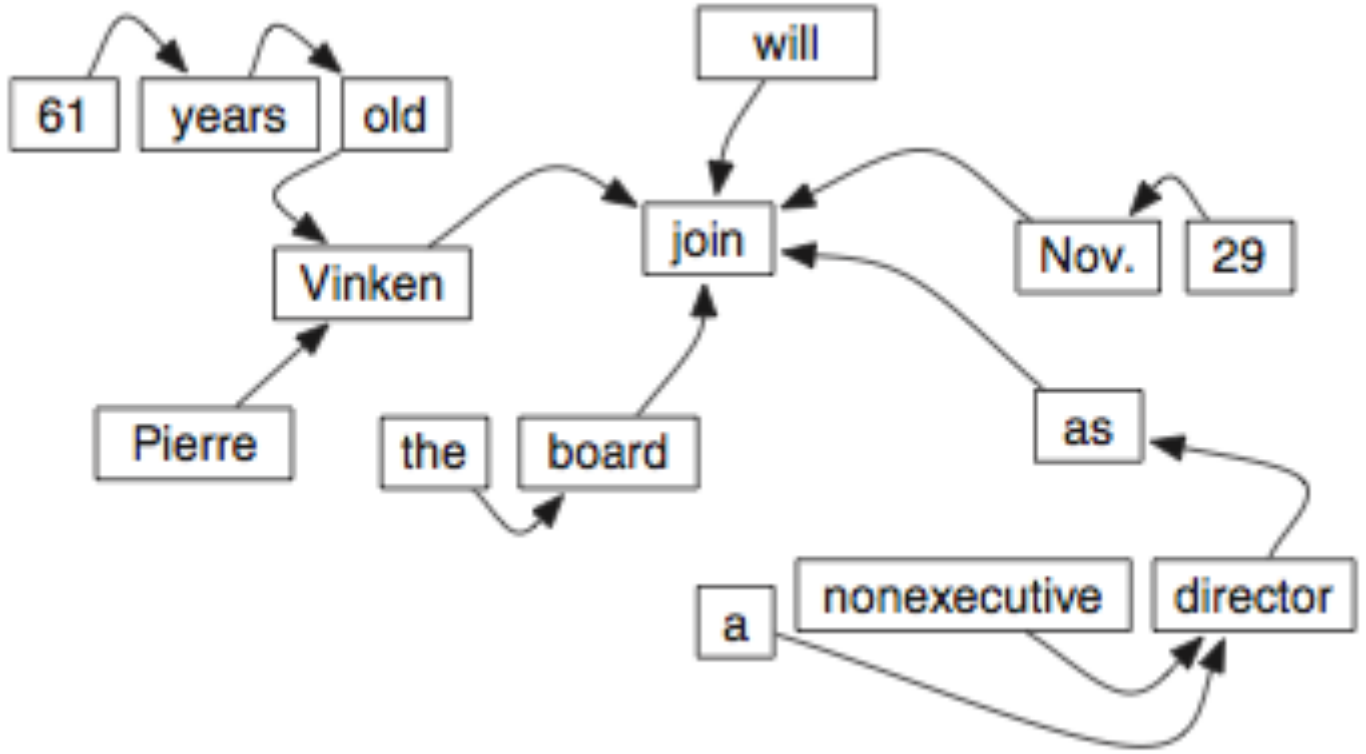
- 規則反應剖析樹節點如何展開 branching

- 一個非終端符號 nonterminal，可能有許多規則，描述不同的展開

- CFG 可以人工撰寫，也可以有樹庫（如 Penn tree bank）來學習
- 規則一多，容易造成繁多的剖析樹 (結構歧義 structural ambiguity)
- CFG 的延伸：加上機率值 → 機率釋文脈無關文法 *probabilistic context-free grammars* (PCFG)
- 例如，有各種的名詞片語 noun phrase (NP)，其中以 DET NOUN 機率最高

Figure 2.9 相依結構 Dependency structure

- *Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.*



相依結構 Dependency structure

- 除了以上的順序、組成關係，我們也可以用相依關係 dependency structure 來表達結構—只表達關係，不管片語、詞序 (word order)
- 有更多資訊：明顯的指示中心語 $head$
- 例如，動詞 $join$ 有五個相依關係：主詞、助詞、受詞、修飾語

Extending tree and dependency structure

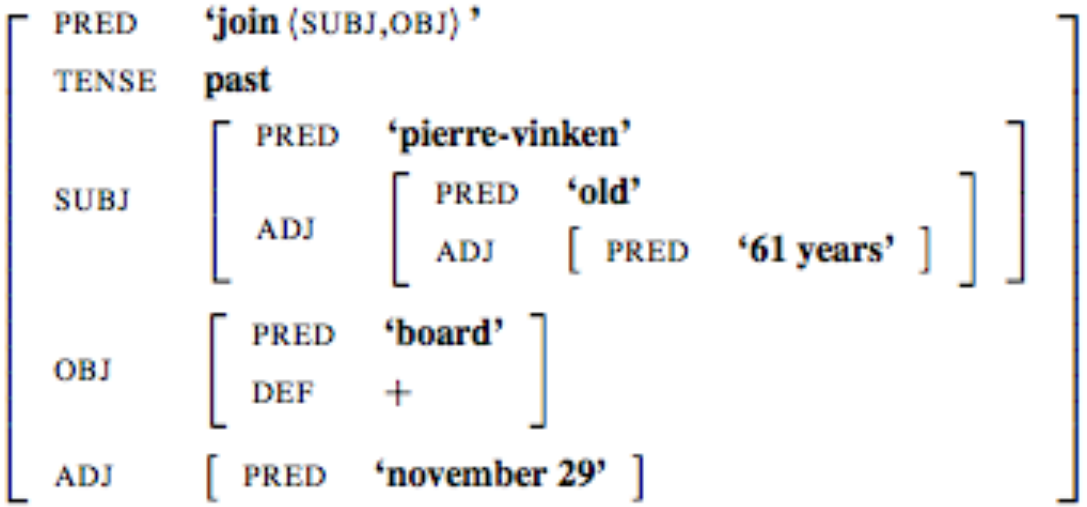
- 剖析樹、相依結構都可以再附加資訊
- 剖析樹可以加註「中心語」
- 相依結構可以標示關係，如主詞 subject 受詞 object 修飾語, adjunct

詞彙功能文法 **Lexical functional grammar**

- LFG 分開表達句子的表面結構 surface structure 和深層結構 deep structure
  - 成分結構 constituent structure 或 c-structure
  - 功能結構 functional structure 或 f-structure
- LFG 用句法性 f-structure 作為基礎表達句子的語意
- 語意也可以用邏輯學的「述詞邏輯」predicate logic/calculus 來表達
- (如$join$(Pierre Vinken, board) 其實 join 和 board 未定義（尚未用於 MT）

Figure 2.10 F 結構 F-structure

- *Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.*



組合類別文法 Combinatory categorical grammar

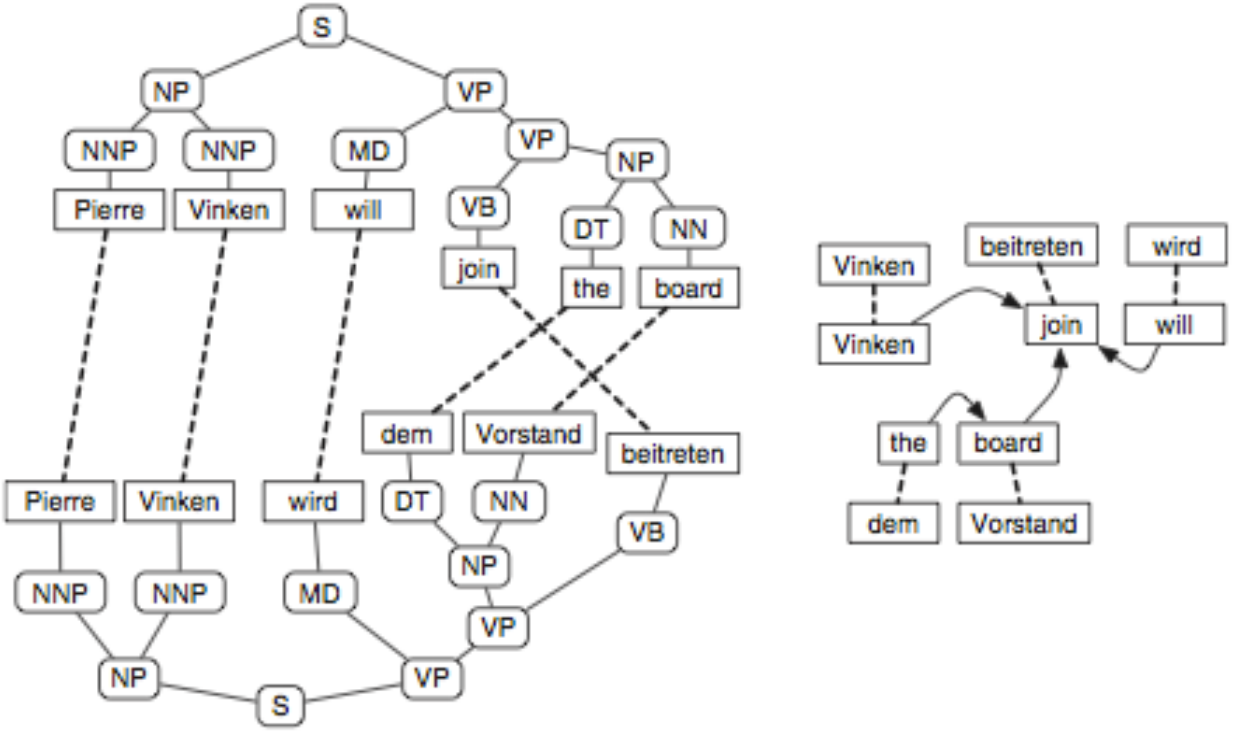
- *Combinatory categorical grammar* (CCG) 依賴詞彙來引導句子的剖析作業
- 動詞註記受詞的類別：
  - swim: $S\backslash NP$ — 左邊需要一個 NP
  - eat: $(S\backslash NP)/NP$ — 右邊還需要一個 NP
  - give: $((S\backslash NP)/NP)/NP)$ — 右邊需要兩個 NP
- 有了這些文法行為的資訊，就可減少結構的歧義

2.2.3 翻譯句子的結構

- 機器翻譯最困難的部份：在兩種語言之家，翻譯轉換句子結構
  - 詞序重組 word reorder
  - 插入、刪除虛詞 function words
- 有時候，兩邊的結構差距很大，就不是這麼簡單
- Figure 2.11 顯示一個簡單的例子（兩邊結構很類似）
- 用剖析樹來轉換句子的結構比較理想，但是剖析樹也非100% 正確，最近作法，潛力很大
  - 文法結構為本的機器翻譯
  - 類神經網路的機器翻譯（分析轉換後得到整句轉換後的資訊，再整體考慮重新產生目標語句子）

Figure 2.11 德文和英文有不同的句法結構

- 再成分結構 constituency structure 中主動詞需要移到最後。但相依結構 Dependency structure 不描述順序，就沒有不同（之後再決定詞序重組）



2.2.4 文脈 Discourse (段落/文件的層次)

- 目前的 SMT 逐句翻譯，不管文脈的現象
- 但是，這會導致有些翻譯的問題，在句子的層次，無法解決
  - 共同指涉 co-references: President George Bush, ... he or the president
  - 向前指涉 anaphora resolution: resolving who is 'the president' male or female
  - 主題與語意 topic 有時候可以幫助語意解歧：運動或生態 bat (蝙糊或球棒)

2.3 語料庫 Corpora

- 統計式機器翻譯需要用有翻譯的文本來訓練系統
- 接著，仔細描述文本與翻譯（Section 2.3.1）、語料庫資料集 (Section2.3.2) 機器翻譯的準備作業 (Section 2.3.3)
- Content
  - 2.3.1 文本分類
  - 2.3.2 取得平行語料庫
  - 2.3.3 句子對應

2.3.1 文本類型

- MT 系統對於特殊文本，效果最好 (新聞、科技論文)
- 詞彙語意隨著主題、領域、文體 modality (書寫、口說)
- 目前大部分的資料來自立法或國際組織 (UN, EU)
- 過去的機器翻譯系統都是為有限領域（氣象報告、旅遊資訊、技術手冊）
- 限制領域後，問題很神奇地簡化、減少

重要的評估比賽 Evaluation Campaigns

- 比賽通常針對特定領域，特定語言配對
- NIST 比賽重點：阿拉伯到英語 Arabic–English 華語到英語、新聞
- IWSLT 任務，主要是亞洲語言到英語，旅遊對話領域
- TC-STAR 以及 ACL WMT 任務，提供歐洲議會會議記錄進行訓練、測試

2.3.2 如何取得平行語料庫

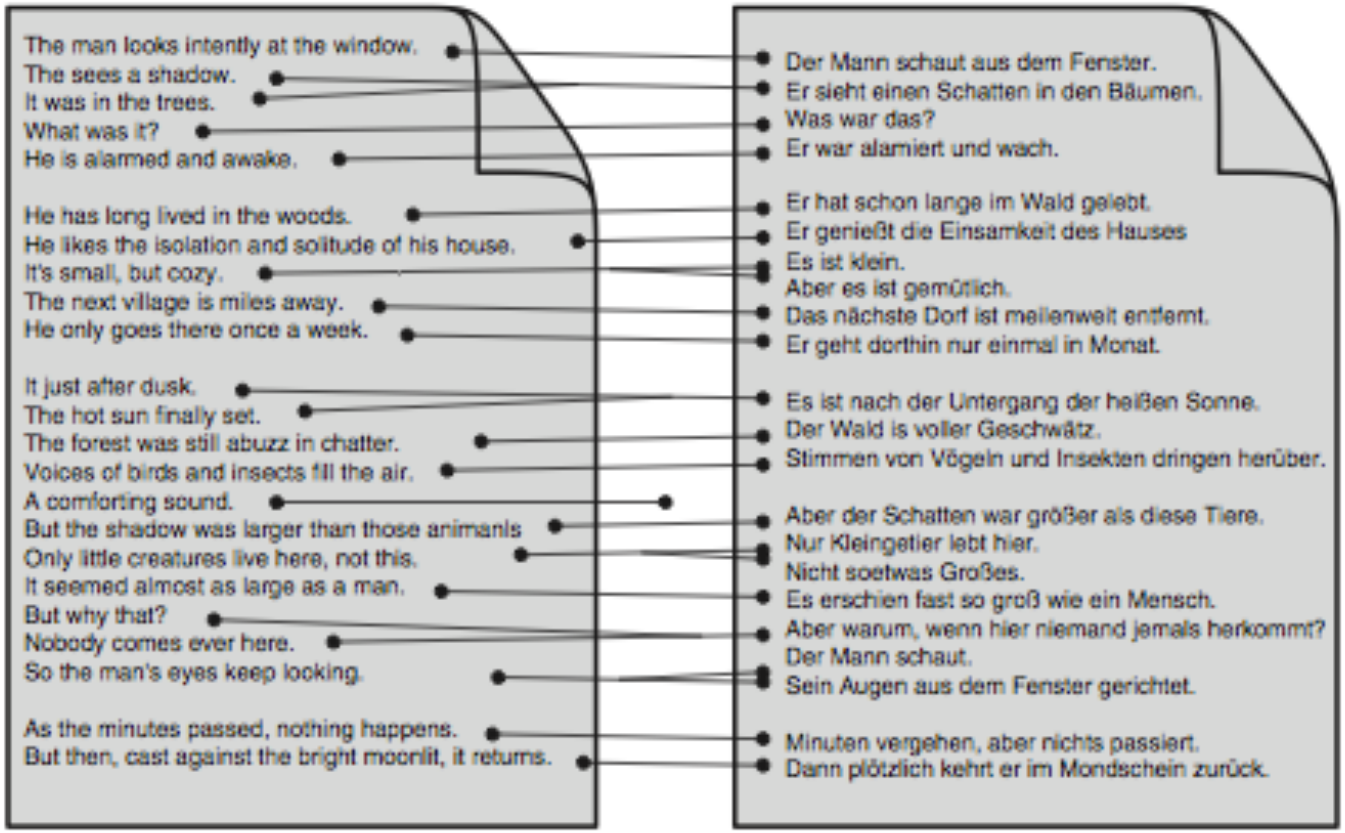
- 平行語料庫 parallel corpus 由一組文本構成，每篇文章都有搭配翻譯
- 數量：10-100 百萬詞
- 來源：網路、LDC
- 寫爬網程式，下載 HTML 網頁，轉成文字檔，對應到翻譯網頁 (文件對應)
- 另外，也可取得相容（但是不平行）語料庫

2.3.3 句子對應 Sentence Alignment

- 有了平行語料庫，我們還需要對齊句子 align sentences
- 大部分翻譯都是一句翻譯成一句
- 長句或許會拆開來翻譯成幾個小句
- 通常會有句號結束的符號（但是：泰語沒有i)

Figure 2.12 句子對應

- 給予一個文件，以及翻譯,，找出互為翻譯的句子



句子對齊的正式定義

- 問題陳述
  - 給予一組句子的文本 $f_1, ..., f_{n_f}$ （外文）
  - 以及對應的翻譯句 $e_1, ..., e_{n_e}$ （英語)
  - 找出 句子對齊資訊 $S = (s_1, ..., s_n)$
  - where $s_i = (f_start - f(i), ..., f_end - f(i), e_start - e(i), ..., e_end - e(i))$

- 假設對應不回頭、不交叉 monotone alignment
  - 開始-f(i) = end-f(i 1) + 1
  - 開始-e(i) = end-e(i 1) + 1

- 其他簡單的限制
  - 開始 f(1) = 1 and start-e(1) = 1

  - 結束 f(n) = n_f and end-e(n) = n_e
  - 開始 f(i) $\leq$ end-f(i), and start-e(i) $\leq$ end-e(i)

- 對應型態有限制：
  - type(s_i) = end-f(i) − start-f(i) + 1 — end-e(i) − start-e(i) + 1
  - e.g., type = 1-1, 1-0, 0-1, 2-2, ...

- 最佳化 $score(S) = \prod_{i}^{n} P(s_i)$

- 其中

$$P(s_i) = \begin{cases} 0.89 & \text{if type}(s_i) = \text{1-1} \\ 0.01 & \text{if type}(s_i) = \text{1-0 or 0-1} \\ 0.09 & \text{if type}(s_i) = \text{2-1 or 1-2} \\ 0.01 & \text{if type}(s_i) = \text{2-2} \end{cases}$$

2.4 摘要

- 因為過讀吹噓，MT 起起伏伏，遭到冷眼 (ALPAC Report)、中斷研究經費
- MT 系統的開發史，就是政府補助機構（如DARPA）的介入的歷史
- 早期的作法包括「直接」、「轉換」、「中介語」
- 70, 80 年代開發的早期系統（如 Météo, Systran, Logos, METAL）如今尚存
- 機器翻譯的主要用途：吸收資訊 assimilation 傳播 dissemination 溝通
- 而需求也有所不同——即使不怎樣的系統，也有用途（網頁、臉書翻譯）
- 機器翻譯可以透過限制輸入：領域限制、控制技術手冊的所使用語言（次語言 sublanguage）——讓讀者易讀，也讓機器翻譯容易處理