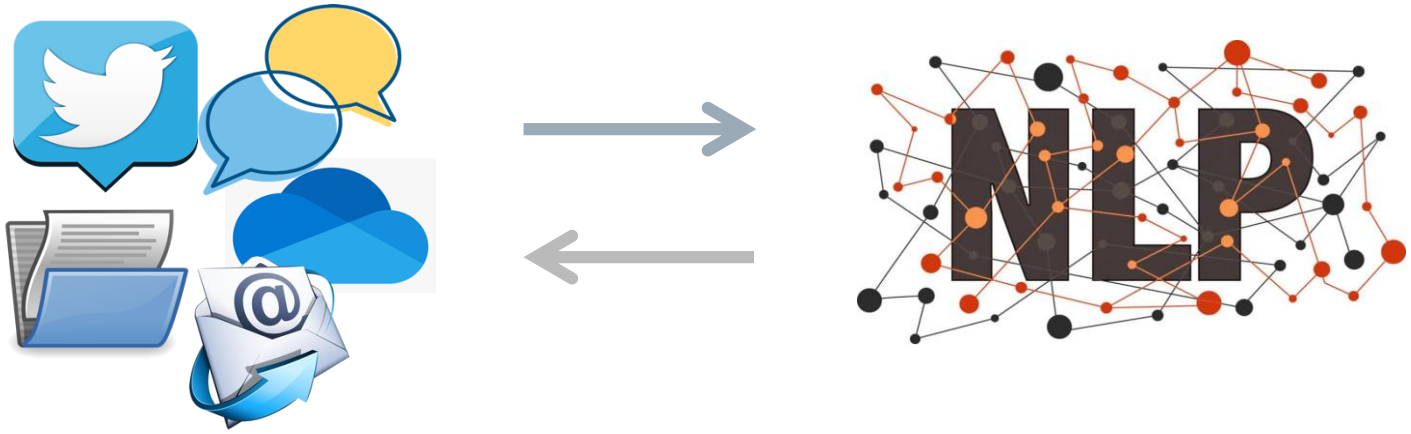
[AI Theory&App] 04 Nature Language Processing

Instructor: Hao-Shang Ma

Department Of Computer Science And Information Engineering, NTUST

什麼是自然語言處理?

- 語言是思維的載體，是人類交流思想、表達情感最自然、最方便的工具。
  - 人類歷史上大部分知識是以語言文字形式記載和流傳的。
- 自然語言指的是人類語言，尤其是指文本符號。
- 自然語言處理（Natural Language Processing，NLP）是用計算機來理解和生成自然語言的各種理論和方法。



自然語言處理的難點與特點

阿明：“你這是什麼意思？”
阿呆：“沒什麼意思，意思意思。”
阿明：“你這就不夠意思了。”
阿呆：“小意思，小意思。”
阿明：“你這人真有意思。”
阿呆：“其實也沒有別的意思。”
阿明：“那我就不好意思了。”
阿呆：“是我不好意思。”

自然語言處理任務層級


應用系統(NLP+)
• 教育，醫療，司法，金融，機器人等
應用任務
資訊擷取，情感分析，機器翻譯，對話系統等
基礎任務
分詞，詞性標註，句法分析，語義分析等
資源建設
• 語言學知識庫建設，語料庫資源建設等

歸結為五個基本問題

- 回歸預測問題
- 文本分類問題
- 文本比對問題
- 序列到序列問題
- 文本生成問題

回歸預測問題

- 將輸入文字映射成為一個連續的數值
- 例如對作文的評分，判斷罰款的金額或預測一個價格等

文本分類問題

- 判斷一個輸入的文字所屬的類別
- 辨識垃圾郵件，即為將文件分類成正常郵件和垃圾郵件
- 情感分析任務，分類一段文字所代表的情感

比對問題

- 判斷兩個輸入文字之間的關係
  - 是否為同義、矛盾或無關
- 辨識兩個輸入文字的相似度(0-1之間)

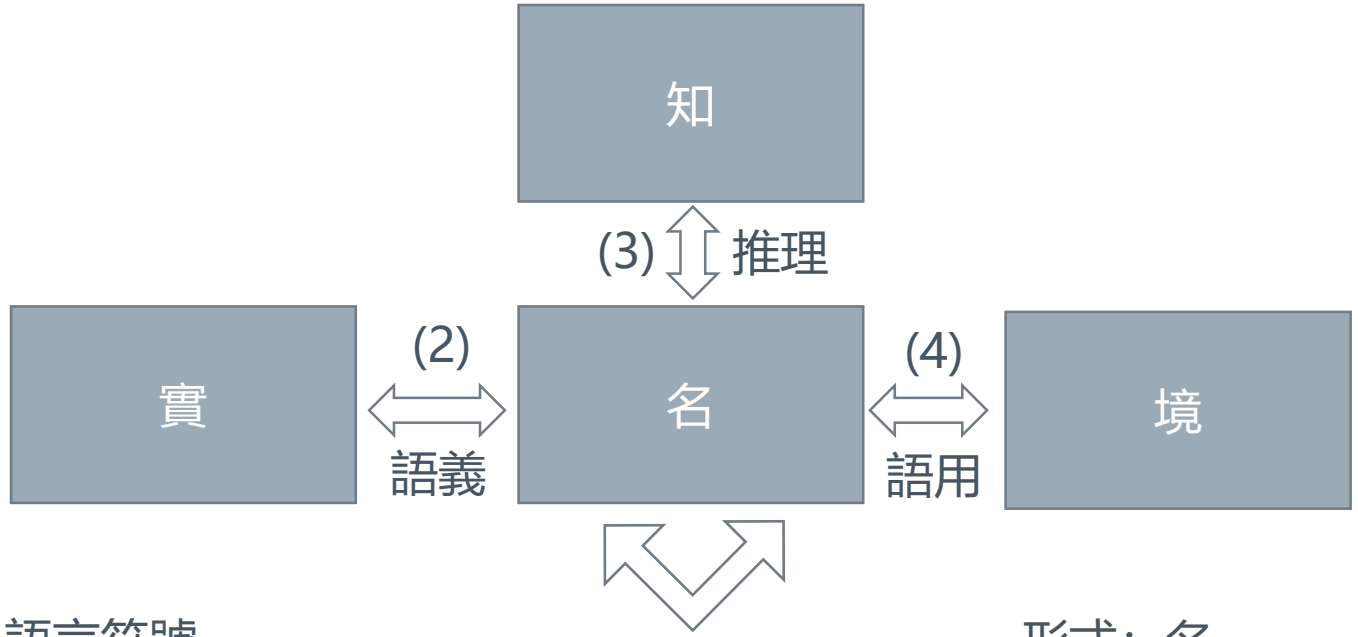
解析問題

- 對文字中的詞語進行標註或辨識詞語之間的關係
  - 詞性標註
  - 句法分析
  - 分詞
  - 命名實體辨識

生成問題

- 根據輸入的內容，生成一段自然語言
- 機器翻譯
- 文字摘要
- 圖形描述生成

研究目標與層次



名：語言符號
實：客觀事實、主觀意見
知：知識
境：語言所處的環境

形式：名
語義：名+實
推理：名+實+知
語用：名+實+知+境

研究目標與層次

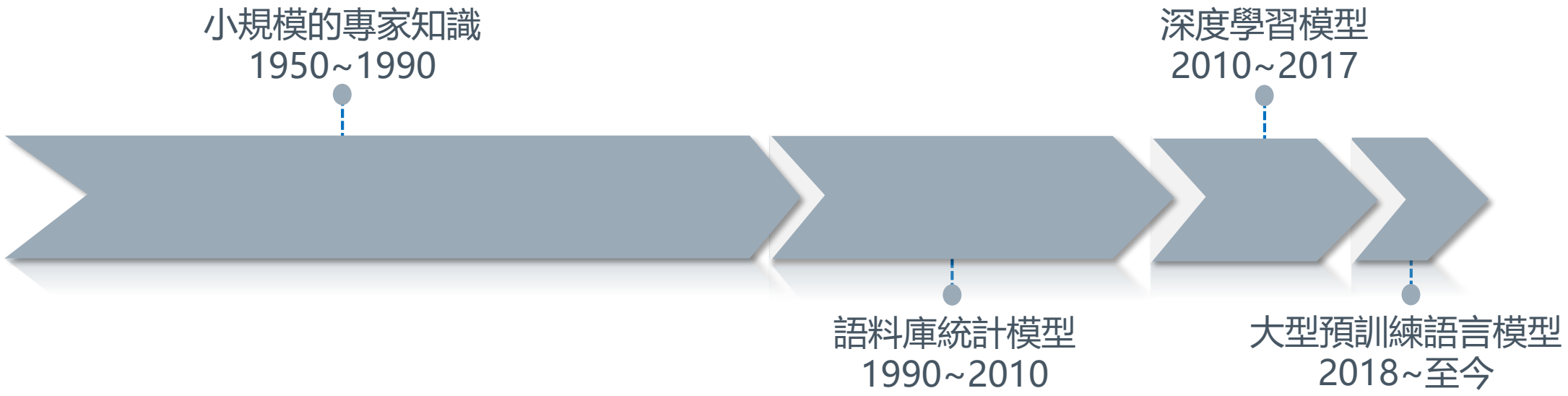
- 形式研究的是名與名之間的關係
  - 計算字詞之間的相似度
- 語義研究的是名與實之間的關係
  - 兩段文字不同卻代表同樣的語義
- 推理研究的是在語義的基礎上進一步引入知識的運用
  - 包括常識、世界知識和領域知識等
- 語用則是還更需要考慮語言使用的環境
  - 同樣一段文字或語言在不同環境可能代表不同意義

研究層次VS任務

	分類	解析	比對	生成
形式	文本分類	詞性標註 句法分析	搜尋	文章摘要
語義	情感分析	命名實體識別 語義角色標註	問答	機器翻譯
推理	隱式情感分析		文本理解	寫故事
語境	反語識別			聊天

自然語言的表示

- 語意在電腦內部是如何表示的？
- 根據表示方法的不同，自然語言處理經歷了四個時代變遷

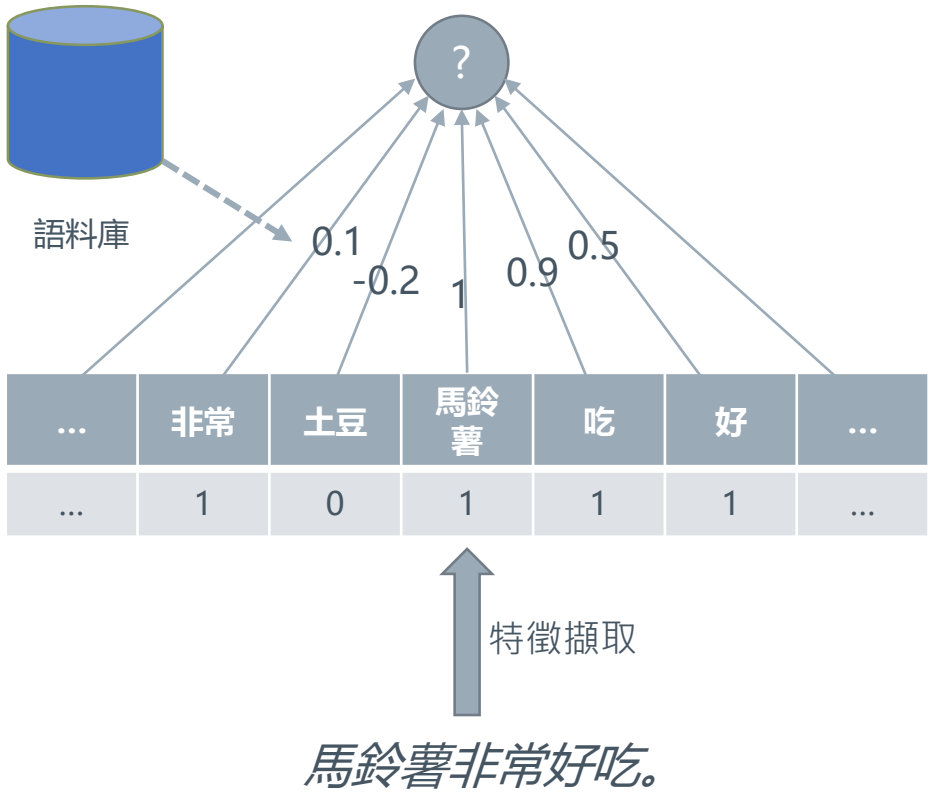


基於符號（字串）表示的專家知識

- 「馬鈴薯非常好吃。」的情感傾向性？
- 如果：出現褒義詞「好」「喜歡」等
  - 結果為正向情緒(褒義)
- 如果：出現"不"
  - 則結果傾向性為負面

- 優點
  - 符合人類的直覺
  - 可解釋、可干預性好
- 缺點
  - 知識不夠完備
  - 需要專家建構與維護
  - 不便於計算

基於向量表示的統計模型



- 獨熱編碼(one-hot encoding)使用高維度、離散、稀疏的向量表示詞
- 維度為詞列表大小，表示法中所有維度只有一位為1，其餘為0
  - 馬鈴薯：[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, …]
  - 好：[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …]
  - 吃：[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …]

缺點

- 嚴重的資料稀疏問題
- 無法處理「多詞一義」的現象

詞的分佈語意假設

- 分佈語意假設（Distributional semantic hypothesis）
- 詞的意思可由其上下文詞的分佈來表示
  - You shall know a word by the company it keeps -- Firth J.R. 1957

詞的分佈（Distributional）表示

› 分佈詞向量

	shinning	bright	trees	dark	look
moon	38	45	2	27	12

he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

語言模型

- 語言模型（Language Model，LM）
  - 描述一段自然語言的概率或給定上文時下一個詞出現的概率
    - $P(w_1, \cdots w_l)$，$P(w_{l+1}|w_1, \cdots w_l)$
  - 以上兩種定義等價

$$P(w_1 w_2 \cdots w_l) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)\cdots P(w_l|w_1 w_2 \cdots w_{l-1})$$

$$= \prod_{i=1}^{l} P(w_i|w_{1:i-1})$$

  - 廣泛應用於多種自然語言處理任務
    - 機器翻譯（詞排序）
      - P(the cat is small) > P(small the is cat)
    - 語音識別（詞選擇）
      - P(there are four cats) > P(there are for cats)

分詞

- 詞（Word）
  - 是最小的能獨立使用的音義結合體
  - 以漢語為代表的漢藏語系，以阿拉伯語為代表的閃-含語系中不包含明顯的詞之間的分隔符
- 中文分詞是將中文字序列切分成一個個單獨的詞
- 分詞的歧義如：把手機關了
  - 把\手機\關了
  - 把手\機關\了
  - 把\手\機關\了

分詞

- 以英語為代表的印歐語系語言，是否需要進行分詞？
- 這些語言詞形變化複雜如：computer、computers、computing等
  - 僅用空格切分的問題
    - 數據稀疏詞表過大，降低處理速度
- 子詞切分
  - 將一個單詞切分為若干連續的片段（子詞）方法眾多，基本原理相似
    - 使用盡量長且頻次高的子詞對單詞進行切分

斷詞/分詞演算法

- **基於設定好的單位切分**：將整個字符串以固定單位來切分，例如N-Gram
- **基於詞典的分詞法**：將待匹配的字符串和一個已建立好的詞典中的詞進行匹配，通常會採用雙向匹配的方法，但這方法的能力有限，例如像是新發明的詞就無法進行匹配
- **統計的機器學習算法**：如HMM，CRF (Conditional Random Field)，常見中文斷詞Jieba套件，對於不存在於字典的字詞就是用統計的方法來處理的

N-Gram 斷詞模型

- N-gram 模型是一種基於統計機率的自然語言處理模型，用於對文本進行建模和預測。它基於一個簡單的假設，即在一個句子或文本中，下一個詞的出現只與前面的N-1 個詞有關，與整個文本的上下文無關。
- N-gram 模型將文本拆分為一系列的N 個詞的序列，這些序列被稱為N-gram。
- 假設N=2，文本為“ChatGPT is a language model”，則會被拆分成:(ChatGPT, is) (is,a) (a,language) (language,model)

N-Gram 斷詞模型

- **我的興趣是看電影和讀書**
- Uni-Gram
  - 「我\的\興\趣\是\看\電\影\和\讀\書」
- Bi-Gram
  - 「我的\的興\興趣\趣是\是看\看電\電影\影和\和讀\讀書」
- Tri-Gram
  - 「我的興\的興趣\興趣是\趣是看\是看電\看電影\電影和\影和讀\和讀書」

N-Gram 斷詞模型

› Bi-Gram

$$P(w_n|w_{n-1}) = \frac{count(w_{n-1}w_n)}{count(w_{n-1})}$$

› Tri-Gram

$$P(w_n|w_{n-2}, w_{n-1}) = \frac{count(w_{n-2}w_{n-1}w_n)}{count(w_{n-2}w_{n-1})}$$

基於詞典的分詞法

- **我的興趣是看電影和讀書**
  - 「我\的\興趣\是\看電影\和\讀\書」

詞典
我
興趣
電影
讀
看
書
看電影

中文斷詞

- 常遇到的難題
  - 歧異詞
    - 「我們 / 在野 / 生動 / 物 / 園 / 玩」
    - 「我們 / 在 / 野生 / 動物園 / 玩」
  - 新詞識別
    - 特有名詞(ptt梗、溫拿)
    - 人名地名

中文斷詞

- 正向最大批配法
  - 會有一個詞典，將句子在詞典中由前向後比對，一一比對最長詞的匹配結果
  - 「我們 / 在野 / 生動 / 物 / 園 / 玩」
- 逆向最大批配法
  - 會有一個詞典，將句子在詞典中由後向前比對，一一比對最長詞的匹配結果
  - 「我們 / 在 / 野生 / 動物園 / 玩」
- 正反都做完之後取最長詞
  - 「我們 / 在 / 野生 / 動物園 / 玩」
  - 「我們 / 在野 / 生 / 動物園 / 玩」

中文斷詞

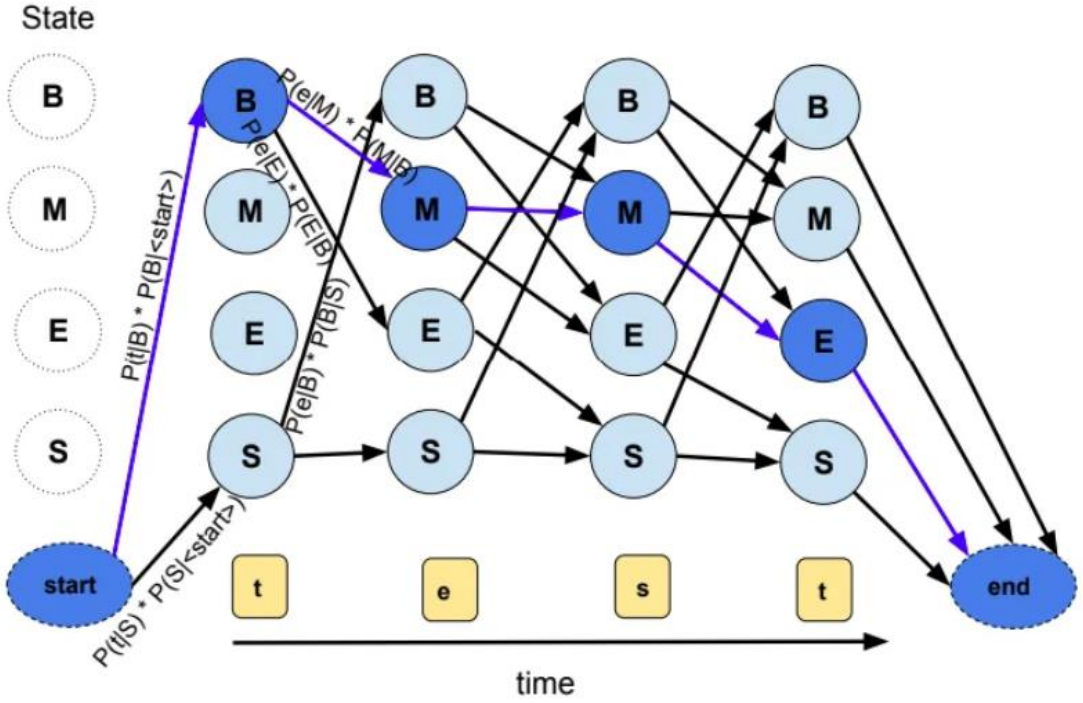
- 基於詞典的斷詞方式，只要詞典中沒有收錄句子中的詞，那可能效果會非常差
- 大部份比較好的斷詞系統都是使用全切分方法，切分出與詞庫匹配的所有可能，然後再運用統計模型決定最好的切分結果

統計的機器學習算法

› HMM(隱馬可夫模型Hidden Markov Model)
  – 運用大量文本去統計"詞與詞"
    之間的關聯性來模擬機率

B→Begin(詞開頭)
M→Middle(詞中間)
E→End(詞結尾)
S→Single(單獨)

在野生動物園→SBEBME

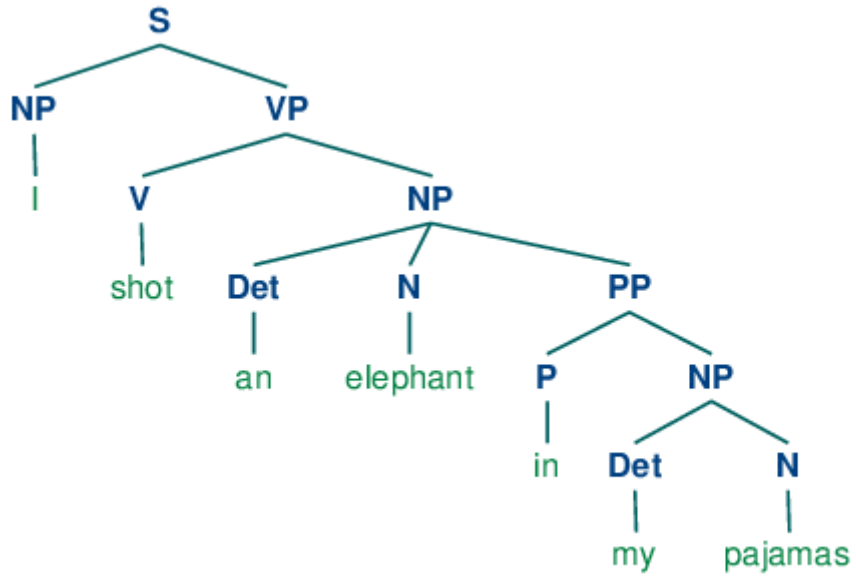


句法分析

- 分析句子的句法成分
  - 以英文為例，構成一個句子可能需包含名詞片語（NP）、動詞片語（VP）和可有可無的 preposition phrase (PP)
- 將詞序列表示的句子轉換成樹狀結構

I shot an elephant in my pajamas

句法樹



情感分析（Sentiment Analysis）

- 分析文本中隱含的對外界事物的態度、觀點或傾向性
  - 正面、負面等
  - 人自身的情緒（Emotion），如喜怒哀懼等
- 輸入
  - 這款手機的螢幕很不錯，性能也還可以。

情感分析子任務	分析結果
情感分類	正面
情感資訊擷取	評價用的詞:很不錯; 還可以 評價對象: 手機螢幕; 性能

資訊萃取（Information Extraction）

› 從非結構化的文本中自動擷取出結構化資訊
  – 命名實體識別
  – 關係擷取
  – 時間表達式擷取
  – 事件擷取

輸入:
10月28日，AMD宣布斥資350億美元收購FPGA晶片巨頭賽靈思。這兩家傳了多年緋聞的晶片公司終於走到了一起。

- 命名實體識別 → 公司名:AMD、賽靈思
- 實體關係擷取 → 賽靈思屬於AMD
- 時間表達式擷取→10月28日
- 事件擷取→收購

問答系統（Question Answering，QA）

- 問答系統可以分為4種主要的類型
  - 檢索式問答系統→答案來源於固定的文本語料庫或互聯網，系統通過查找相關文檔並選擇答案完成問答
  - 知識庫問答系統→回答問題所需的知識以資料庫等結構化形式儲存，問答系統首先將問題解析為結構化的查詢語句，通過查詢相關知識點，並結合知識推理獲取答案
  - 常問問題集問答系統→通過對歷史積累的常問問題集合進行檢索，回答使用者提出的類似問題
  - 閱讀理解式問答系統→通過擷取給定文檔中的文本片段或生成一段答案來回答使用者提出的問題

更多自然語言任務與應用

- 機器翻譯（Machine Translation，MT）
  - 中翻英、各國語言翻譯
- 對話系統（Dialogue System）
  - 任務型系統: Siri、Cortana、Google Assistant
  - 聊天型系統: 小冰
  - 問答型系統: 線上客服

自然語言處理應用

機器翻譯(Machine Translation)

- 機器翻譯是指運用機器，透過特定的電腦程式，將一種文字或聲音形式的自然語言，翻譯成另一種文字或聲音形式的自然語言。
- 透過計算機語言學、人工智慧和數理邏輯來教會機器理解人類的語言，機器翻譯是先把複雜的語言進行編碼，並轉換成電腦理解可計算的公式、模型和數字，再解碼成另一種語言。

機器翻譯(Machine Translation)

若要將中文句子翻譯成英文句子，如「猴子吃香蕉」

我們會先將句子進行斷詞，讓機器容易了解，即「猴子 / 吃 / 香蕉」再經過編碼器分析句子，包含語法及語義的自動解析，並透過翻譯及調序模型將句子完整翻譯，即「猴子為 monkey / 吃為 eat / 香蕉為 banana」，最後利用解碼器與後處理，轉換為人類理解的英文句子。

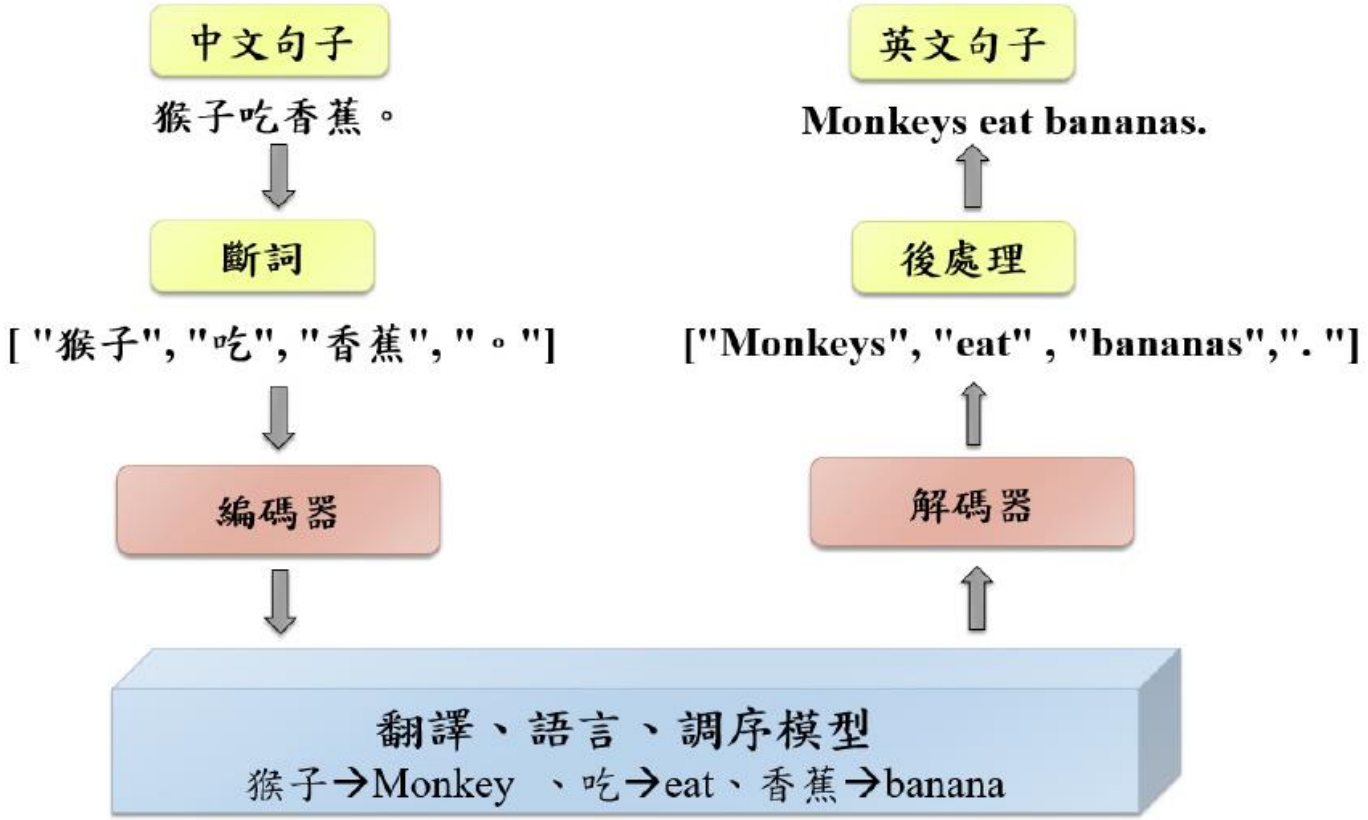

中文句子
猴子吃香蕉。
斷詞
["猴子", "吃", "香蕉", "。"]
編碼器
英文句子
Monkeys eat bananas.
後處理
["Monkeys", "eat" , "bananas",". "]
解碼器
翻譯、語言、調序模型
猴子→Monkey 、吃→eat、香蕉→banana

圖 2-6　翻譯猴子吃香蕉

機器翻譯可分成三類

- 文本翻譯
- 語音翻譯
- 圖像翻譯

圖 2-7　機器翻譯分成三類

文本翻譯

目前最為主流的應用仍然是以傳統的統計機器翻譯和神經網絡翻譯為主，如Google、微軟與百度等公司，都為使用者，提供了免費的在線多語言翻譯系統。

語音翻譯

- 即時翻譯技術是語音翻譯最廣泛的應用，最常出現在會議場所，演講者的語音能即時轉換成文本，並且進行同步與低延遲的翻譯，能夠取代口譯員的工作，實現不同語言的交流。

圖像翻譯

- 人們習慣透過Google 翻譯來查詢看不懂的外文字，例如餐廳裡的菜單、街道看板等等，但要把它輸入到手機翻譯很浪費時間，而且某些看不懂語言也無法輸入，這時候只要打開手機相機鏡頭，對準擬翻譯的文字，就能即時將它轉譯為我們熟悉的語言，能夠即時翻譯招牌、指示等照片上的文字。

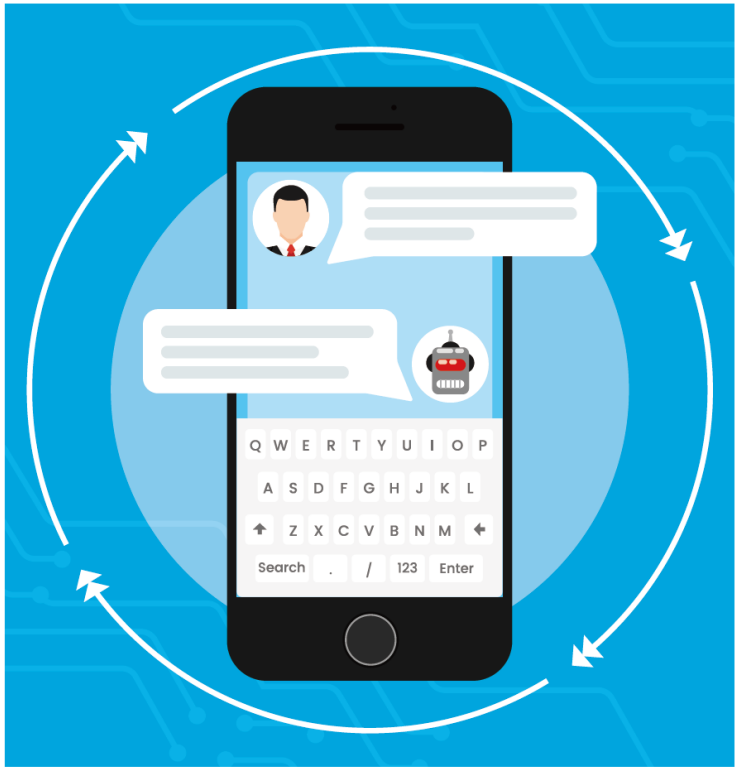
聊天機器人(Chatbot)

- 聊天機器人是指透過人工智慧、電腦程式模擬與使用者互動的對話，利用計算機自動回答使用者所提出的問題，以滿足使用者需求的任務。
- 賦予機器感知、認識、能聽、能看和能與人交流的能力，這樣的溝通方式，不僅符合人類的習慣，也省去編程和輸入的繁瑣，一個語音指令就能達成目的。

相關影片

「聊天機器人」夯！
人工智慧讓服務變科技

現今聊天機器人分成許多類型，最常見的是回答問題、聊天、下訂單、檢索等等。



客服回答問題

› 臉書推出了「Facebook Messenger Platform」，能夠串接 Facebook 粉絲專頁，透過粉絲專頁，直接點選聊天按鈕，讓使用者能夠更加直接與企業粉絲專頁聯絡。

› 而企業常用的方式，就是建立一個匯入常見問題(Frequently Asked Questions, FAQs)系統，當機器人看到關鍵字，機器人將複製 FAQs 裡面的重點，然後採用有禮貌的語氣回答設定好的答案，自動回應形成一對一的對話。

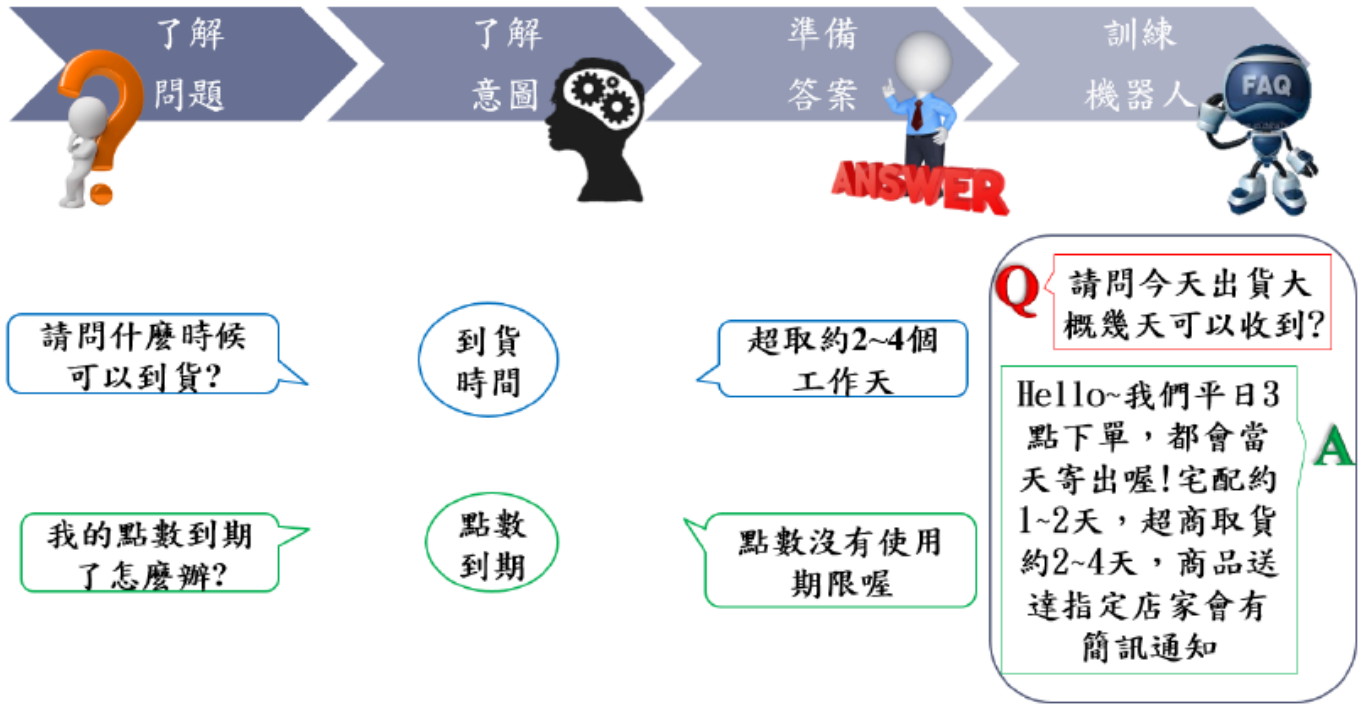

了解
問題
了解
意圖
準備
答案
訓練
機器人
FAQ
ANSWER
請問什麼時候可以到貨?
到貨時間
超取約2~4個工作天
我的點數到期了怎麼辦?
點數到期
點數沒有使用期限喔
Q
請問今天出貨大概幾天可以收到?
Hello~我們平日3點下單，都會當天寄出喔!宅配約1~2天，超商取貨約2~4天，商品送達指定店家會有簡訊通知
A

圖 2-8　FAQs 系統

圖 2-9　FAQs 下訂單的例子

購物助理

› 未來除了在消費購物的習慣上有所改變，在付費的流程上也可能因機器人的盛行而加以改變，比起電話或是網頁，直接使用社群軟體，如 LINE、微信等，可讓消費者能迅速的完成訂單，再用 LINEPay、微信支付進行輔助更能打造出快速付款的購物環境。

檢索

› LINE 推出「國語小幫手」的聊天機器人，能查詢國字的注音、部首或筆畫，也包括造詞、造句和成語查詢的功能，人們透過機器人來進行檢索，能省去查詢的時間，並透過簡單的指令即能找出所困惑的字詞，使在找字的同時亦能學習到不同層面的知識。

› 聊天機器人在商業的應用越來越蓬勃，其主要有四大原因:

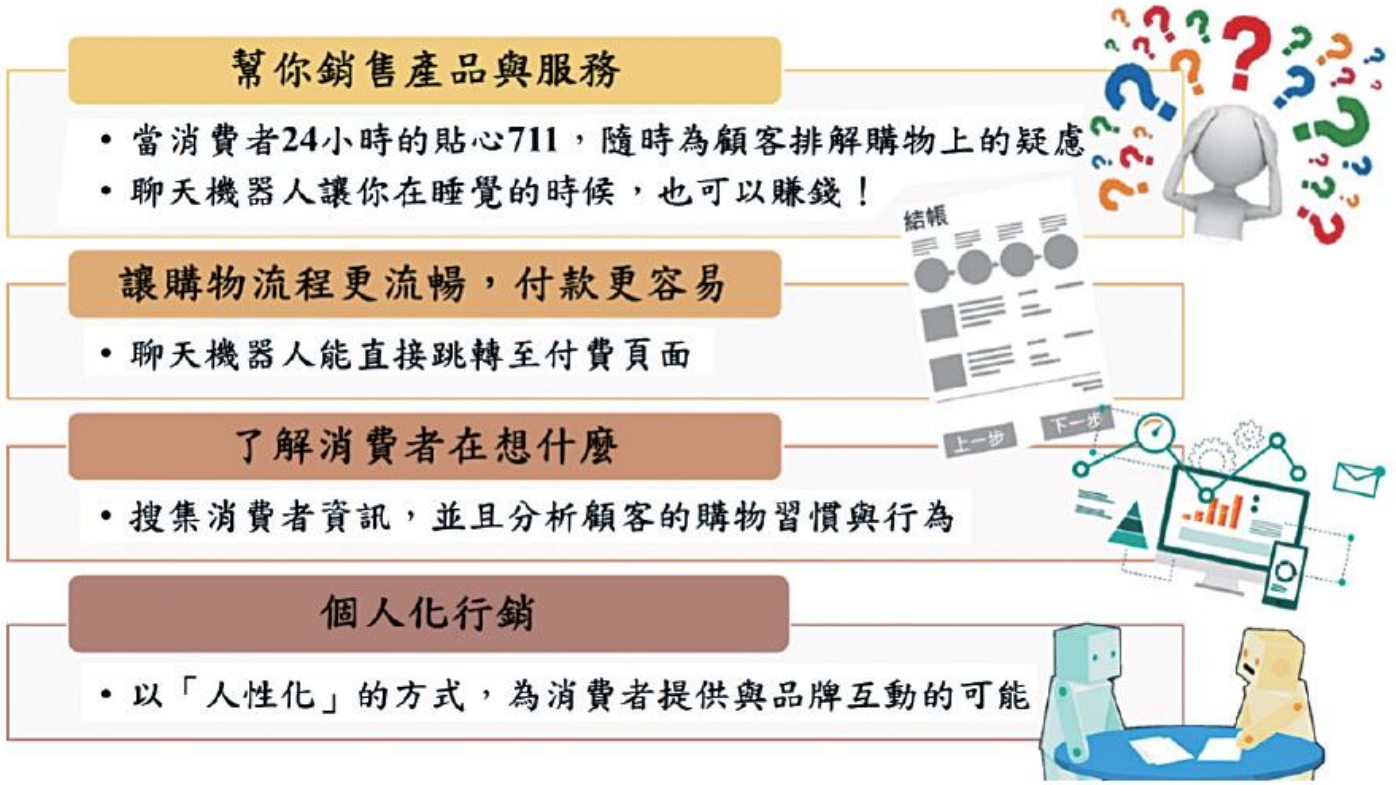


圖 2-10　聊天機器人在商業的發展

關鍵字

- 透過關鍵字能了解不同身分與當前社會關注的熱門議題，如：大學生群體，常出現的關鍵字可能為出國、留學、就業等。
- 以個人利益來說，利用關鍵字能提高搜索效率，在最短的時間找到你所需要的相關訊息。
- 而商家則可透過關鍵字廣告，發掘巨大的商業價值，這個價值可能體現在短期的銷售增長上，更可能是長期企業品牌形象的提升上。

輸入法選字

› 當你在手機上打字時，你常會看到選字建議，利用自動學習，使機器能自動學習各個使用者過去輸入的詞彙與內容，成為他們常用的關鍵字並進行建議的行為。

圖 2-11　輸入法選字建議

檢測垃圾電子郵件(Spam Email)

- 垃圾郵件指未經請求而發送的電子郵件，例如未經發件人請求或允許而發送的各種宣傳廣告或具有破壞性附有病毒的電子郵件。
- 常見內容包括賺錢信息、成人廣告、商業或個人網站廣告、電子雜誌和連環信等。

圖 2-12　檢測垃圾電子郵件

文本情感分析(Sentiment Analysis)

- 文本情感分析，也稱為意見挖掘，是指用自然語言處理、文本挖掘以及計算機語言學等方法，來對帶有情感色彩的主觀性文本進行分析、處理、歸納和推理的過程。
- 情感分析的商業價值，除了可以提早了解顧客對於產品或公司的觀感，進而調整營運策略方向。

個人助理-Siri

- Siri最早內建在iPhone 4，使用者可以使用自然的對話與手機進行互動，完成搜尋資料、查詢天氣、設定手機日曆、設定鬧鈴、及對話聊天等服務。

相關影片

找資料又會搞笑！語音助理 Siri、OK Google 大 PK

圖 2-14　Siri 的功能