複雜系統在語言科技中的應用

（生成語法與形式語意學的整合視角）

學習目標

- 第一部分：複雜系統基礎理論
- 第二部分：生成語法的複雜系統觀點
- 第三部分：形式語意學與複雜性
- 第四部分：語言科技的應用
- 第五部分：批判性反思
- 第六部分：未來展望與結論

第一部分：複雜系統基礎理論

什麼是複雜系統？

- **簡單定義：由大量相互作用的組件構成的系統**
- **複雜定義：從「遊戲理論」的「觀察者亦為受觀察物」到「混沌理論」的「看似隨機的表現其實由簡單規則累積勒成」最終到「複雜系統」的「湧現、自組織、非線性」。**
- **關鍵概念：系統的整體行為無法僅由個別組件預測**

複雜系統的核心特徵

- **多尺度組織（Multi-scale organization）：從微觀到宏觀的層次結構**
- **路徑依賴（Path dependency）：歷史影響現在、前項影響後項、下層影響上層**
- **湧現（Emergence）：整體大於部分之和**
- **自適應性（Self-Adaptivity）：系統隨環境變化而演化**
- **非線性互動（Nonlinear interactions）：小變化可能導致大影響**

複雜系統的數學工具

- **網路理論（Network Theory）：**
  - **描述節點與連接的關係，即「圖論、集合與樹」**
- **動態系統理論（Dynamical Systems）：**
  - **研究系統隨時間的演化，即「 ΔT 必然為系統的因子之一」**
- **資訊理論（Information Theory）：**
  - **量化系統的複雜度與熵，即「 info vs. data 」或「 signal vs. noise 」**
- **計算理論（Computation Theory）：**
  - **分析系統的計算能力，即「什麼是能算的，什麼是不能算的？」**

為何語言是複雜系統？

- **多層次結構：音韻、詞彙、句法、語意、語用**
- **湧現性質：語意不僅是詞彙的簡單組合**
- **社會互動：語言透過使用者互動而演化**
- **適應性：語言持續適應溝通需求**
- **歷史演化：語言變遷展現路徑依賴性**

語言複雜性的層次

- 微觀層次：音素、語素的組合規則
- 中觀層次：詞組結構、句法樹、語意合成
- 宏觀層次：語篇結構、對話系統、語言變遷
- 跨層互動：各層次之間的雙向影響
- 認知基礎：語言處理的神經機制

複雜系統觀點的優勢

- 整合視角：連結不同語言學子領域
- 動態建模：捕捉語言的時間演化
- 預測能力：理解語言變遷與習得
- 跨學科對話：與認知科學、計算機科學整合
- 實用價值：改進語言科技應用

語言系統的自組織性：語言不是約定俗成的

- 詞彙網絡的無尺度特性（scale-free)
- 語法規則的自發湧現：
  - 語法規則是在觀察中發生的，還是它本體（自適應）發生的？
- 語言習得中的模式形成：
  - 雙語環境的幼童自主把 ありがとうございます 說成 ありがとう many thanks
- 社會語言變異的動態平衡
- 語言接觸中的結構轉移

複雜性度量

- **Kolmogorov 複雜度：**
  - **描述語言模式的複雜度。複雜度愈高，愈不可能是語言 / 訊號。**
- **熵與資訊率：量化語言的不確定性**
- **網路複雜度：詞彙與句法網路的拓撲性質**
- **層次深度：句法樹與語意組合的遞迴層次**
- **計算複雜度：語言處理所需的計算資源 (e.g., O(n))**

第一部分小結

- 複雜系統提供理解語言的新框架
- 語言展現典型的複雜系統特徵
- 需要多尺度、動態的分析方法
- 數學工具有助於精確建模
- 為後續討論奠定理論基礎

第二部分：生成語法

（作為複雜系統的句法理論）

生成語法基本概念

- Chomsky 的語言能力（competence）理論
- 普遍語法（Universal Grammar, UG)
- 參數理論（Parametric variation)
- 遞迴性（Recursion）與無限生成
- 內在語言（I-language) vs. 外在語言（E-language)

最簡方案（Minimalist Program)

- 核心操作： Merge （融接）、 Move （移位）
- 介面條件： PF （語音形式）與 LF （邏輯形式）
- 經濟原則：最小化計算負擔
- 特徵檢驗（Feature checking)
- 階段理論（Phase theory)

生成語法的複雜系統性質

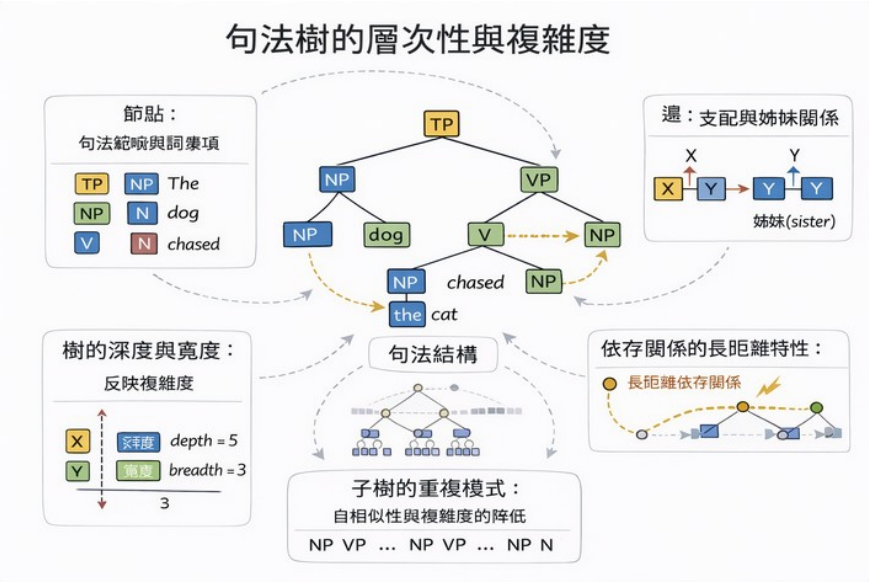
- 遞迴合併產生離散無限性
- 少數原則生成豐富結構（湧現）
- 參數設定的級聯效應（非線性）
- 句法樹的階層組織（多尺度）
- 跨語言變異的系統性模式

Merge 操作的複雜性

- 二元合併：最簡單的結構建構 （為什麼 2 最簡單？）
- 遞迴應用：產生任意複雜的結構
- 標籤演算 (Labeling algorithm)
- 內部 Merge (Internal Merge)：產生位移
- 計算複雜度： context-free 到 mildly context-sensitive

句法樹作為網路結構

- 節點：句法範疇與詞彙項
- 邊：支配與姊妹關係
- 樹的深度與寬度：反映複雜度
- 子樹的重複模式：自相似性與複雜度的降低
- 依存關係的長距離特性



參數理論與複雜性

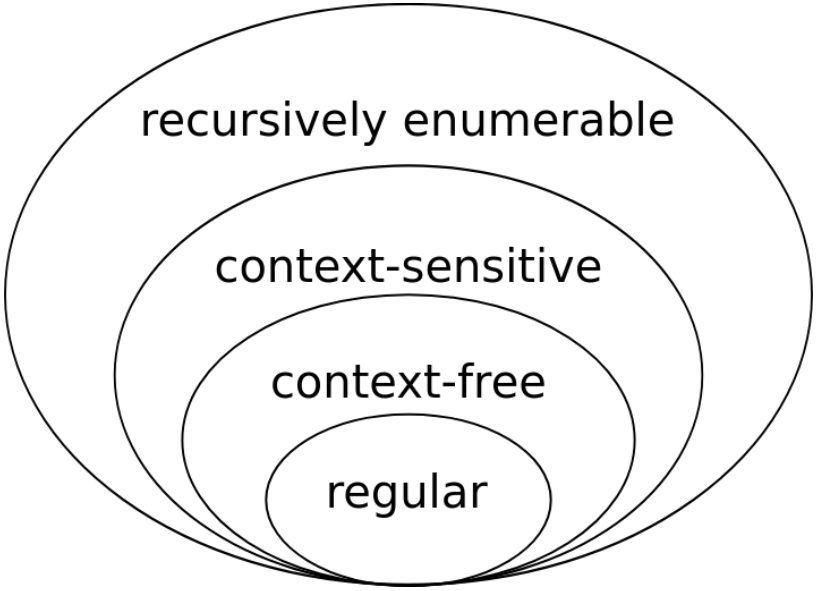
- 參數：語言變異的開關
- 參數空間：可能語法的組合爆炸
- 參數設定的觸發：語言習得
- 參數級聯：一個參數影響多個現象
- 宏參數 vs. 微參數的辯論

移位與依存關係

- wh- 移位、主題化、焦點移位
- 長距離依存的複雜度
- 島嶼限制（Island constraints）
- 痕跡理論與複本理論（Copy theory）
- 移位的認知成本

生成語法的計算模型

- 形式文法層級： Chomsky hierarchy
- 自然語言的溫和語境敏感性（mild-context sensitivity)
- Minimalist Grammars (MG) 的形式化
- 可分析性與可學習性



句法處理的神經基礎

- Broca 區與句法處理
- 遞迴結構的神經實現
- 左半球優勢與句法網絡
- 句法與工作記憶的關係

跨語言的句法變異

- 詞序參數： SVO, SOV, VSO 等
- Pro-drop 參數
- wh- 移位 vs. wh-in-situ
- 名詞組的結構變異
- 變異背後的普遍原則

生成語法與語言科技

- 句法分析器（Parsers）的理論基礎
- 樹庫（Treebanks）的構建（obsolete...maybe?）
- 機器翻譯中的句法轉換
- 語法錯誤檢測
- 神經網路能否學會句法？ (we know it can't now by 2024.)

第二部分小結

- 生成語法描述離散的結構系統
- 遞迴與參數產生複雜性
- 句法是多層次的組織
- 湧現性質來自簡單操作
- 為語言科技提供理論指導

第三部分：形式語意學

- **意義的非複雜系統觀：形式語意學的計算「並不是」複雜系統！**
- **它是因果邏輯的計算過程。**

形式語意學基礎

- 真值條件（Truth conditions）
- 指稱（Reference）與謂述（Predication）
- 組合性原則（Compositionality）：Frege 原則
- 邏輯形式（Logical Form, LF）
- 模型論語意學（Model-theoretic semantics）：
  - 把前四者結合起來，也可以成為一個語意模型
  - 模型「並不是」只有機率模型一種

Montague 語意學

- 類型論（Type theory）：e（個體）、t（真值）
- λ- 演算（Lambda calculus）
- 內涵邏輯（Intensional logic）
- 可能世界語意學（Possible worlds）
- 自然語言作為形式語言

語意組合的複雜性

- 函數應用（Function application）
- 類型提升（Type raising）
- 組合子邏輯（Combinatory logic）
- 量化詞的範疇提升
- It's complicated, but it's not complex!

量化與範疇

- 廣義量詞理論（Generalized quantifiers)
- 量化詞的單調性（Monotonicity)
- 範疇的包含關係
- 多重量化的歧義
- 約束變項的語意解釋

事件語意學

- 事件結構與論元結構
- 體貌（Aspect）的形式化
- 因果關係的表徵

時態與模態

- 時間邏輯（Temporal logic）
- 過去、現在、未來的表徵
- 模態算子：必然、可能
- 時態與模態的互動

認識模態 vs. 道義模態

- 認識模態 (Alethic Modality)
- 定義：描述事物真實性或可能性狀態的模態。
- 例子：
  - 「1+1=2」是一個必然為真的命題。
  - 「我現在沒有吃飽」這個命題是偶然的，可能為真也可能不為真。
  - 「火星上有生命」是可能的命題，但尚未被證明是必然或偶然的。
  - 核心概念：必然性、偶然性、可能性。

- 道義模態 (Deontic Modality)
- 定義：描述行為是否被允許、禁止或是否為義務的模態。
- 例子：
  - 「你必須遵守交通規則」（義務）。
  - 「你不能在圖書館大聲喧嘩」（禁止）。
  - 「你可以選擇是否要參加派對」（允許）。
  - 核心概念：義務、許可、禁止。

指稱與回指

- 指示詞（Indexicals）的語意
- 定描述詞（Definite descriptions）： Russell vs. Strawson
- 代詞的約束理論
- 動態語意學（Dynamic semantics)
- 話語表徵理論（DRT)

預設與蘊涵

- 語意預設（Presupposition）
- 預設投射問題（Projection problem）
- 蘊涵（Entailment ；必然性）與推論（Implication ；可能性）
- 會話蘊涵（Conversational implicature）
- Grice 的合作四原則

語意學的非複雜系統特徵

- 組合性：
  - 局部規則產生全局意義
  - 但全局即為「局部的總合」，沒有超過！
- 上下文依賴：語境影響解釋的方式可用同一套邏輯推算
- 多義性：一個表層形式多種底層意義可用同一套邏輯推算
- 推理網絡：意義之間的蘊涵與推論關係
- 語用與語意的介面

向量空間語意學？

- 詞向量（Word embeddings）： Word2Vec, GloVe
- 分佈假說：相似上下文產生相似意義（嗎？）
- 向量的算術運算： king - man + woman ≈ queen
- 高維空間的複雜結構
- 與形式語意學的整合

組合向量語意學

- 張量積與組合函數
- 遞迴神經網絡（RNN）的語意組合
- Tree-LSTM 與句法引導組合
- 注意力機制與語意對齊
- Transformer 的語意表徵

多義性與歧義消解

- 字典：
  - 詞彙多義：一詞多義 vs. 同形異義
- 句法：
  - 結構歧義：附加歧義、範疇歧義
- 語意：
  - 語用歧義：指稱歧義
- 語意 + 語用：
  - 上下文線索與消解策略
- 資料分佈：
  - 貝葉斯推論模型

第三部分小結

- 形式語意學提供精確的意義理論
- 組合性是語意複雜性的核心
- 語意網絡展現複雜的結構
- 上下文與推理增加複雜度
- 形式與分佈方法的互補
- 然而，它並不是複雜系統。 Why?

第四部分：語言科技應用

複雜系統觀點的實踐

自然語言處理概覽

- 核心任務：分詞、詞性標註、句法分析、語意分析
- 傳統方法：基於規則、統計模型
- 深度學習革命：神經網絡方法
- 大型語言模型（LLM）的興起
- 評估標準與挑戰

句法分析技術

- **結構句法分析（Constituency parsing)**
- **依存句法分析（Dependency parsing)** ***(deprecated now)***
- **神經句法分析器** ***(never actually works...)***

語意角色標註（效果不如預期的那些方法）

- PropBank 與 FrameNet
- 論元結構的自動識別
- 序列標註模型：條件隨機場 (Conditional Random Field, CRF), BiLSTM
- 預訓練模型的語意理解
- 跨語言語意角色標註

機器翻譯

- 統計機器翻譯（SMT）：短語基礎
- 神經機器翻譯（NMT）： Seq2Seq 模型
- 注意力機制的引入
- Transformer 架構： BERT, GPT
- 多語言模型與零樣本翻譯

問答系統

- **抽取式 vs. 生成式問答**
- **知識庫問答（KBQA)**
- **閱讀理解： SQuAD, RACE**
- **檢索增強生成（RAG)**
- **複雜推理與多跳問答**

對話系統

- 任務導向 vs. 開放域對話
- 對話狀態追蹤
- 回應生成策略
- 對話管理的強化學習
- ChatGPT 與大型對話模型

語言生成

- 文本摘要：抽取式與生成式
- 數據到文本生成
- 故事生成與創意寫作
- 可控生成：風格、情感、主題
- 評估指標： BLEU, ROUGE, BERTScore

大型語言模型（LLM）

- GPT 系列：自回歸語言模型
- BERT 系列：掩碼語言模型
- 預訓練與微調範式
- 提示工程（Prompt engineering）
- 觀察到的湧現能力：推理、規劃、工具使用

第四部分小結

- 語言科技整合複雜系統觀點
- 神經方法學習複雜語言模式
- 理論語言學提供指導與評估
- 挑戰來自語言的本質複雜性
- 但需要批判性反思 LLM 的性質

第五部分：批判性反思

兩種湧現的區分

- 弱湧現（Weak Emergence）
  - 原則上可還原，但實務上困難
  - 是認識論的複雜性
  - 來自觀察者的描述視角
- 強湧現（Strong Emergence）
  - 原則上不可還原
  - 新的因果力量
  - 是本體論的複雜性

區別（Diff.）
認識論：牆上不可能有狗臉，只是因為某某原因，人類有此認知或詮釋
本體論：牆上有一張狗臉

LLM 真的是複雜系統嗎？

- 質疑的核心論點
- LLM 的”湧現”可能只是觀察者的描述：
  - 系統本身只做確定性矩陣運算
  - 每一步計算都是可預測的
  - “湧現能力”是我們無法追蹤細節（因實務上的困難）的結果
  - 不是系統固有的本體論特性

還原論視角

- LLM 的運作流程：
- 輸入文本

  → Token 化

  → 嵌入向量

  → 多層 Transformer （矩陣乘法 + 激活函數）

  → 輸出概率分布

  → 生成文本

- 關鍵洞察：
  - 每一步都是確定性計算
  - 沒有真正的自主性或隨機性
  - 原則上完全可預測（給定參數）

Scaling Laws 的啟示

- 平滑的擴展律
- GPT 系列的能力可用平滑曲線預測
- 沒有明顯的”相變點”
- 所謂”湧現”可能是：
  - 人類主觀劃定的能力閾值
  - 漸進量變，而非質變
  - 測量粒度不足造成的錯覺
- 結論： LLM 的湧現可能是認識論的，非本體論的

與真實複雜系統的對比

• 特徵：	自然複雜系統	vs.	LLM
自主性：	個體有自主行為	vs.	無自主性
隨機性：	真實隨機過程	vs.	確定性計算
開放性：	與環境持續互動	vs.	封閉的計算系統
演化：	真實的適應與演化	vs.	參數固定（訓練後）
因果力：	湧現層次有因果作用	vs.	只是底層計算的結果

人類語言 vs. LLM

- 人類語言系統的獨特性
- 真正的複雜系統特徵：
  - 社會互動：多個自主主體的真實互動
  - 歷史演化：語言隨時間真實變遷
  - 認知基礎：神經生物學的隨機性與可塑性
  - 創造性：真正的新穎表達生成
- LLM 的局限：
  - 只是靜態參數的函數映射
  - 沒有真實的社會或認知基礎
  - " 創造性 " 只是訓練數據的重組

對語言學研究的啟示

- ⚠ 重要警告
- 不能直接用 LLM 推論人類語言能力：
- LLM 學到的”語法” ≠ 生成語法的心理現實
- LLM 的”語意理解” ≠ 人類的意義建構
- 統計相關性 ≠ 結構性語言知識
- 函數擬合 ≠ 語言能力（competence）
- Chomsky 的批評是有道理的！

修正後的立場

- LLM 是什麼？
- ✅ 計算複雜的系統
  - 大量參數的非線性互動
  - 實務上難以完全理解
- ✅ 認識論意義的複雜性
  - 從觀察者角度展現複雜行為
  - 需要複雜系統工具分析
- ❌ 本體論意義的複雜系統
  - 沒有真正的湧現
  - 沒有自主性和適應性
  - 是確定性的計算裝置

正確的研究態度

- 對 LLM 的合理定位
- 應該做：
  - 研究 LLM 作為語言模式的統計模型
  - 分析其計算複雜性與表現
  - 用作探索語言現象的工具
  - 比較 LLM 與人類的行為差異
- 不應該做：
  - 認為 LLM 複製了人類語言能力
  - 用 LLM 的行為推翻語言學理論（你不會拿香蕉來証明芭樂不夠長，對吧？）
  - 忽視 LLM 與真實語言系統的本質差異

第五部分小結

- 湧現有弱 / 強之分：認識論 vs. 本體論
- LLM 的湧現是觀察者視角的產物
- LLM 是計算複雜但非本體論複雜系統
- 不能等同 LLM 與人類語言能力
- 需要批判性地應用複雜系統框架

第六部分：未來展望

複雜系統、語言學與科技的融合

語言科技的未來趨勢

- 更大規模的語言模型（仍需反思其本質）
- 多模態與具身智能
- 少樣本與零樣本學習
- 可解釋與可信賴的 AI
- 特定領域與低資源語言

複雜系統方法的深化

- 謹慎應用於不同對象
- 自然語言系統（真正的複雜系統）：
  - 動態網絡分析語言演化
  - 多主體模擬語言變遷
  - 跨尺度建模：從神經元到社會
- 人工語言模型（認識論複雜）：
  - 計算複雜度分析
  - 資訊論量化
  - 但不過度詮釋為本體論湧現

生成語法與神經網路

- 關鍵問題
  - 神經網絡能否實現生成語法？
  - 可能學到表面模式
  - 但不等於內化了生成語法規則
- 未來方向：
  - 結構歸納偏置（Inductive biases)
  - 神經符號整合（Neurosymbolic AI)
  - 可微分的句法結構
  - 明確整合語言學知識

跨學科研究機會

- 真正的整合
  - 語言學、計算機科學、認知科學
  - 複雜系統科學（謹慎應用）
  - 神經科學與腦影像研究
  - 社會科學與語言演化
  - 哲學：湧現、意識、意義的本質
- 關鍵：保持各學科的嚴謹性，不過度類比

開放問題（潛在的期中 / 未考題）

- 基本理論問題
  - 語言能力的本質是什麼？
  - 生成規則？統計模式？還是兩者皆有？
  - 組合性如何在神經網絡中實現？
  - 真的實現了嗎？還是只是近似？
  - 語言習得的關鍵機制？
  - 先天 UG ？統計學習？社會互動？
  - 語言與思維的關係？
  - Sapir-Whorf 假說的現代版本

開放問題（續）

- 人工智能問題
  - 通用人工智能需要何種語言能力？
  - 統計模式夠嗎？還是需要結構知識？
  - LLM 的”理解”是什麼？
  - 真正的理解？還是複雜的模式匹配？
  - 如何整合符號與連結主義？
  - 神經符號 AI 的挑戰
  - 可解釋性與複雜性的權衡？
  - 越複雜越難解釋，但越有效

總結：複雜系統觀的核心洞察

- 對自然語言（真正的複雜系統）
  - 語言是多層次的複雜適應系統
  - 簡單規則透過遞迴產生無限複雜性
  - 社會互動產生真實的湧現性質
  - 理論需要複雜系統思維
- 對 LLM （認識論複雜）
  - 計算複雜但本質確定
  - 展現模式但非真正理解
  - 有用但不等於人類能力
  - 需謹慎詮釋