Articut/Loki (Wang et al. 2019) 為根據生成語言學原理建構的 NLP/NLU 系統， 能在保有自然語言結構本質的前提下完成運算。Articut 不同於現行以統計機率計算詞頻或語料庫字典驅動的斷詞系統 (Jieba、CKIP tagger)，其以X-Bar Theory 為基礎運作，在完整保留功能詞 (functional words) 的前提之下，以其為核心語 (functional heads)，將句子套用到不同句法樹進行特徵檢查 (feature checking)，並在執行中文斷詞 (Chinese Word Segment，以下簡稱 CWS) 的同時依節點加註詞性標記 (Part of Speech, 以下簡稱 POS)，其過程中考量到漢語中雙音詞及非核心語重音等構詞上的特性。而 Loki 將經 Articut 標記斷詞的語句以正規表示式 (Regular Expression, 以下稱 Regex) 輸出，以句法及語義結構做為運算操作符號，在不同的語境和意圖之下，比對語料結構，使用者可於相關程式區塊中撰寫程式碼，以針對不同結構進行詞彙參數抽取，設計回應等操作。

二、Articut 斷詞工具與 Loki 意圖模型工具

（一）Articut 斷詞與詞性標記工具

Articut 是一個商用中文斷詞及詞性、命名實體標記工具。Articut 依 Chomsky 對句子內部結構的觀點，定義了 X-bar 語言樹狀結構框架，再將輸入的中文句子由下而上地，透過多組中文詞組構詞原則決定詞的邊界。以能最接近句法樹的最高點為輸出結果。 依句法樹的運作原則，一個詞彙被定位在句法樹上的某個節點時，其詞性亦固定下來。 因此節點之間的分界，就是詞組的分界；而節點的位置，就標示了詞性的推算結果。此外，Articut 在設計上另外收集了一組用以表示台灣地址、台灣道路名稱、法條索引、網址以及金錢的字串模式 (string pattern) 以及可供使用者動態選用的 WikiData 詞條和 政府公開資料中的景點名稱資訊做為外部字典以完成命名實體辨識的工作。綜合以上流程，Articut 的斷詞、詞性標記以及命名實體辨識是同步完成而無法分割的。 其輸出結果除了詞彙邊緣外，亦已隱含句法、句型資訊在內。

（二）Loki 設計原理

Loki為 Linguistic Oriented Keyword Interface 之字首組合詞。其設計之目的在 於透過句型比對，以及挑選語義計算時所需的詞彙做為參數，以便在保有「句型-語義」 的語言表現關係之餘，也能擷取出關鍵詞彙做為計算介面之所需。 Loki 意圖分析工具的架構，是依 Sinclair, J. & Coulthard, R.M. 1975 年在 Toward an Analysis of Discourse. 書中所提，注意到的課堂上教師與學生之間的言語互動模式而 設計。在該研究中提出的三個層次分別為： 效果 (Act)：例如「教學中與學生互動」、「考試時令學生安靜」…等效果。 功能 (Function) ：例如「問與答的互動」、「點名與答有的互動」 實例 (Example)：例如「教師：『ㄎㄨㄞ 四聲？』學生：『快』」 這三個層次在 David Nunan (1993). Introducing Discourse Analysis. Penguin Group 一書中，被擴充為：場景 (Discourse) 對應效果 (Act) 對話的語境 (Context) 對應功能 (Function) 對話的實際內容 (Utterance) 對應實體 (Example) 本工具在設計時，使用 NLP 領域較熟悉的詞彙，將對話分成三層，分別是： 專案名稱 (Project)：某組意圖適用的場景。例如在便利商店的場景，具備繳費意 圖、購票意圖。但不具備住宿意圖。對應 Nunan 的「場景」。 意圖名稱 (Intent)：某一種意圖。例如在便利商店場中的繳費意圖。對應 Nunan 的語境。 語言表達 (Utterance)：一組可以用來表達某一場景下，某一意圖的語言表達。可以 是完整的句子或是不完整的句子。對應 Nunan 的對話的實際內容。 首先，本研究利用 Loki 建立「中文wh疑問詞：誰」做為專案名稱 (Project name)，說明這一組意圖將適用於含有中文wh疑問詞「誰」的語言場景 (Discourse)。接著建立三種「語義」的意圖名稱 (Intent name)。在這意圖下，所有的句子都是為了描述這三種「語義」的語言表達 (Utterance)。 例如「你寫給誰呢」、「沒有誰注意我的存在」和「無論誰當權」…等。 透過 Loki 呼叫 Articut 進行斷詞、詞性標記與命名實體辨識處理後，所有的句子都將轉化為只保留標記做為句式辨識用的正規表式示，並透過保留語言表達 (Utterance) 句型的方式來區分語義意圖，獲得最大的兼容性。

三、研究方法與流程

本研究將 Loki 意圖分類工具操作流程分成四部份來達成辨析中文wh疑問詞「誰」之語義的目標：

1. 中文wh疑問詞的斷詞及詞性標記預處理（Articut）

2. 完成預處理後，建立 Loki 意圖模型 shei\_sinica

3. 設計數學運算的函式（本研究以「加法事件」和「減法事件」來處理加減法） 4. 使用 shei\_sinica 進行中文wh疑問詞語義辨析與檢查，其中流程 1 的預處理，在 Loki 意圖分析工具中會自行於後台處理。

（一）建立 Loki 意圖模型 shei\_sinica

Loki 的架構分成專案名稱 (Project)、意圖名稱 (Intent) 和語言表達 (Utterance) 三層。 本研究建立一專案名為shei\_sinica，接著根據語言學研究歸納出的三種中文wh疑問詞的語義定義為三個主要意圖，分別為： a. 表達「疑問語義(interrogative\_use)」的意圖 b. 表達「存在語義 (existential\_use)」的意圖 c. 表達「全稱語義 (universal\_use)」的意圖

以句子「你寫給誰呢」、「沒有誰注意我的存在」和「無論誰當權」來說明，疑問語義是「你寫給誰呢」，存在語義是「沒有誰注意我的存在」，全稱語義是「無論誰當權」

本研究利用中央研究院平衡語料庫，蒐集含有疑問詞「誰」的所有句子。將平衡語料庫裡，可以代表以表示疑問語義的句子，例如「你寫給誰呢」或「還有誰會來買本國的產品」，輸入疑問語義 (interrogative\_use) 中建立相關意圖 (Intent) 模型；再將平衡語料庫裡用來表示存在語義的句子，例如「沒有誰注意我的存在」或「彷彿生著誰的氣」…等句子，輸入存在語義 (existential\_use)的意圖中；最後將表達全稱語義的句子， 例如「無論誰當權」或「誰也不敢動」輸入全稱語義(universal\_use) 中 建立 shei\_sinica 模型。

有了 Loki 意圖模型，符合疑問語義句型的文字，就會被 Loki 分類為 interrogative\_use意圖。以此類推，「沒有誰注意我的存在」就會被分類為 existential\_use 意圖，而「無論誰當權」則是 universal\_use意圖。

歸到Pattern:

題目文字轉成 Definition, Calculation, Question 三種意圖的資料結構如下圖一、二、三： 圖一、Definition 意圖資料 結構 圖二、Calculation 意圖資 料結構 圖三、Question 意圖資料 結構 在定義語境場景 (Definition) 的意圖中，依句型取出其實體名稱 (Entity)、物體持有人 (Possessor)、分類/量詞 (Classifier) 以及其數量 (Quantity)。以若無，則儲存空字串。例 題中的「蘋果」是 Entity，「桌上有」是 Possessor，「顆」是 Classifier，而「一」是 Quantity。 若在同一題目中有多組定義，例如「姐姐有三張黑紙，妹妹有五張白紙」，則儲存多組 定義。 在計算過程 (Calculation) 的意圖中，依句型取出其實體名稱 (Entity)、物體持有人 (Possessor)、分類/量詞 (Classifier) 以及其數量 (Quantity)。若在同一題目中有多組計算 過程，例如「姐姐早上搞丟三枝筆，下午去買了五枝」，則儲存多組過程以記錄發生順 序。 在求解目標 (Question)的意圖中，依句型取出其實體名稱 (Entity)、物體持有人 (Possessor)、分類/量詞 (Classifier) 以及其數量 (Quantity)。若在同一題目中有多組計算 The 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020) Taipei, Taiwan, September 24–26, 2020. The Association for Computational Linguistics and Chinese Language Processing 214 過程，例如「姐姐總共吃了幾塊蛋糕，妹妹剩下多少蛋糕」，則儲存多組求解目標以待 稍後依序求解時使用。 模型建立後，Loki 系統會將產生的 Loki\_Math.atm 儲存於雲端，同時產生用以表示每 一個「句型 - 意義」連結關係以及自動取出該句型中可供計算的「詞彙元素 (e.g., 兩顆，蘋果)」的程式碼。