10/11-10/18

WH Observation：

觀察一百句處理過的語料

a. Interrogative

b. Existential

c. Universal

$. 語料有問題

？. 不確定/待討論

我們會一起在 WH Observation上把一百句語料依上表做分類，這邊的分類是以 paper 中的語言學理論做分類。

WH Patterns：

依據 WH Observation 中的分類，在 abc 三類（也就是Loki 中project 「誰」之下的三個 intent）之下去做 pattern 的分類。

比如：

在語料中有「誰碰了我的電腦」，

此句屬於 a intent（Interrogative），

他可以被看作是 「誰+V了+N」，

所以我就會在 Interrogative 這個 intent 中建「誰碰了我的電腦」這個例句，其代表的就是「誰+V了+N」這個 pattern。

爾後就只需對這些 pattern 做編輯和增減即可！這部分就會以paper 中的語言學理論為主，但考量一些 Loki 實作上的因素（Articut 結果）做分類。

10/25 WH 詞進度：

繼續再觀察前一百句，然後再把一些 interrogative 的句子建進去 Loki，但因為有一些 structure 會包含到太多的句子（例1），所以我需要再觀察一下其他兩類的結構之後做比較，這部分應該會在 Loki 先放寬，之後到程式碼的時候再做區分。

例1

「票投給誰只有天知、地知、我知」屬於 interrogative

這句如果建 「V + 誰」會同時對到如「選誰來做都一樣可行」這種 universal 的句子，雖然說直接用 regex 在「誰」後面排除「都」是可行的方案，但除了都以外還會有其他比如「也+ negation」甚至 「也都+negation」等需要排除，這樣 regex 不好寫，所以我想可以在程式端做處理即可，方式大概就會是先看是不是 universal 不是的話再看是不是 interrogative。

這週我們的看法是，因為 interrogative 的語料最多（結構上單純），建很少的 pattern 會吃到很多句子，但這樣會同時包含到其他兩個用法的句子，所以我們接下來一週的做法如下，先把前100句中的 interrogative 都歸納建到 Loki 上，完成之後看看其他兩個是否有被包含進去，有的話就修掉，這樣會比較乾淨簡單。

另外，手動標記的部分我有請孟軒有空就繼續往下標，標到差不多300-500句，待前100句都在 Loki 上做好之後，直接拿後面標過的句子當作 test，這樣 Loki 分出來的句子也可以有正確標記來對照，我們再針對有分錯的進行修改即可，這樣反覆進行，最後就可以慢慢把句子都做完。

針對「誰」的存在語意判斷有些問題。

有些句子沒辦法透過 Tim 給的paper範例核可EPW 語意（e,g, 條件子句、是非問句）判斷。

孟軒說她在標記的時候詢問你，你用將「誰」替換成「某人」之類的方式來做論證，這邊我們想過了之後都是通的，但卻想不出怎麼在 Loki 上不使用替換的方式處理。（只在 pattern 上針對這些句子做篩選）🤣 所以我們在 Google docs 留了六句想詢問 Tim。感謝Tim！

11/15

這週我們已經開始用Loki批次處理了，我們先處理了前400句語料（200句是我們人工分類過的），預計只先驗收b、c的結果，孟軒會再把結果整理出來放到docs上～

bc的部分應該是不太會有混淆的情況但是針對 Tim 說「呀」的不確定性會帶出存在語意，這點我是感覺得到也可以處理的，但語氣一變變成「ㄧㄚ」一聲，是不是會變成疑問語意呢？

我們調整完之後會再往下300句，這次我讓孟軒先調看看，因為前400效果應該不差，我預期有問題的應不超過30句～

我們昨天針對上週跑出來的結果進行修正，有一個我覺得很不錯的 progress。

上週前200句的結果顯示，模型在分類：

b/c 表現佳 100% 無搞混

a/b 表現尚可 搞混一句

a/c 搞混案例較多，約 8 句  
不是你想約誰誰就來的  
這場官司誰勝誰負  
誰也不識誰的廬山真面目

誰又有理由說誰走入邪途呢

誰叫人人都認識我呢

然而誰也不認識誰

並聲稱誰再出聲便處分誰

無論誰當權

無論誰接班都將面對如何處理六四歷史定位

誰做好公民誰就是傻瓜

誰打破了這個平衡誰就輸

不論誰將是台灣第一任民選總統

向來都是誰也不服誰

無論是誰接手寫下一個續集

?誰做了這種這麼缺德的事還誰敢承認

我要看誰跟誰求婚

解決方式: 我們在 Loki 裡面建的 regex 其實能區辨出是否有強疑問詞，也就是說，我們能讓 Loki 知道「在強疑問詞出現的情況下，疑問語意效力大，反之即小」這件事情，這禮拜會試看看在程式上能不能辦到，我預計是可以且不難，這樣就解決了我們前一週卡住的沒有辦法逐一排除「無論、不論」等全稱語意cue 的問題。

也就是說，我們在疑問語意部分的 regex 應該要做的更細，區分不同句法的cue。

a/b/c 語意皆沒有捕捉到的有約 30 句，但其中包含了 ungrammatical 的句子，所以我覺得這是好的現象，我們下週會再針對這邊做補強。

11/22

已經把 Loki 批次處理語料的程式寫好成可以沿用的了，所以之後的模式就會變成：

週三 meeting 討論+修改上一次處理的結果，修改好再丟新的語料下去測試，直到兩千XX句都處理完，這樣「誰」的部分就大致完成了，可能就差與「什麼」的共現問題，以及要跟 Tim 再對一次我們建在 Loki 裡面的邏輯是否足夠貼近句法上的 cue，這邊校對完若需要的話我可以再寫說明的程式即：

此句中 「誰」為XX語意，此語意來自句中XXX。

的輸出結果，這樣應該就是比較理想的一個語言學模型了。

前幾週因為要討論的地方都還太多，不太好有規劃的處理問題，現在感覺到問題慢慢在縮限，可以好好想方法處理之後我覺得處理起來就很貼近論文中的內容，越來越好做了。

11/29

針對疑問和全稱語意混淆的問題做了一些程式碼中的調整，其中包含捕捉強疑問詞的 cue，誰…誰就…的 cue 等等，進行效力的比較之後選擇答案，這邊效果不錯，也一定程度上解決太容易出現疑問語意的情況，因為在大多數情況下疑問和全稱不容易同時出現在沒有 embedding 的句子當中。

調整後我們實測了400句語料：

正確率大概在87%左右

疑問存在都有且是錯的：9

(有五個是”呀”相關的問題)

Missed ：48

（其中1/4是不合法句）

存在全稱都有且是錯的：2

（這邊我會再看看是怎麼回事）

那誰坐哪的話教室裡那不都是先生指定的嗎

比如誰誰誰挨批評

目前看起來我們三種語意總共才建20個左右的 pattern而已，卻已經漸漸不用再增加，開始在處理效力大小問題了。

下週再做多一點測試，讓測試總量落在大約700-1000之間，正確率夠高之後，我會再來試試看透過embedding 來區分疑問/存在語意的問題。

12/6

針對「誰」的部分已經差不多做完句型的建置了，2696句語料已經大多能捕捉（25句 missed但其中過半數是句子不完整），約剩下個位數尚未處理，目前我們只建了41個 pattern。

剩餘待處理的部分僅剩：

1.一些句子在被embed的情況下語意會發生變化

2.有些成語或是文言文用法

3.少部分Articut 斷詞有誤

接下來針對「誰」需要做的：

1. 把平衡語料庫的句子判斷做到對，要處理以上三點。

2. 找新的語料來測試，做出一個 accuracy 的 %數。

3.若有需要再撰寫判斷說明。

剩下我目前想到的有：

1. 把「什麼」做完 一樣要給出一個 accuracy。

2. 處理共現問題。（這點尚無法估計所需時間）(兩個以上語義)

12/15

1. 建pattern 這個部分我和陳峻田是分工的，我會和他確認他是否有將他在建的所有pattern都同步放在Google docx 上。若是都有同步的話，我們三種語意的所有pattern都會在pattern 的Google docx 裡。

2、3 我們這週三對程式和pattern有一些改動，最新的輸出結果我預計今天以前會檢查完，再回報老師。

今天以前會補上回覆老師的前三項問題。

4.我認為學期結束前是可以做得完的。

5.了解！我先把重心放在完善「誰」的部分

剛才檢查完檔案已經整理出前三項問題的回覆，再請老師看看

1. 已經將目前所建的所有Pattern同步更新到Pattern的google docx

2.

疑問語意:98%

存在語意:99%

全稱語義:99%

(不包含ungrammatical但對到東西:1%)

3.

(1)

句尾有"呀"的句子，尚未精準判斷是疑問語意還是存在語意，需要使程式能夠判斷是否是embedded clause

(2)

有兩個"誰"的類成語會是疑問語意，但有些這種類成語，檢查器會判斷成全稱語義

如:誰贏誰輸、誰勝誰負

需討論是否一個一個加進utterance，或是有能夠一次解決類似情況的辦法

(3)

有些ungrammatical的utterance仍對到某個語意

如:

94. $ 知道誰--> 檢查器輸出疑問語義

134 $ 好像誰上壘--> 檢查器輸出疑問語義

(4)

island外的"都"應無法作為island裡的"誰"的核可詞，但會被檢查機器人算成全稱語義

如:

51. a 誰的辦法能讓網球界的教練和選手都服氣--> 檢查器輸出全稱語義

(5)

古文中的"誰"會抓不到語意

如:捨他其誰、人生自古誰無死...等，顯示Missed，可能再建一個intent另外處理

另外，有一些我認為待討論的句子，不知老師認為以下輸出結果是否合理?

166 ? 誰做了這種這麼缺德的事還誰敢承認-->檢查器輸出全稱語義

430 ? 我說了我們在笑誰-->檢查器輸出疑問語義

437 ? 但是如果整個兒這個醫藥為誰服務的這個想法不變的話-->檢查器輸出疑問語義

443 ? 大人講說你要學習誰你要學習誰-->檢查器輸出疑問語義

453 ? 總是免不了誰家的小孩子會讀書-->檢查器輸出疑問語義

463 ? 誰入地獄的精神為我們服務-->檢查器輸出疑問語義

12/21

以下是三種語意的正確率(該語意的句子數量為分母/全部句子的數量為分母)

疑問語意(錯56句): 97.3%/97.9%

存在語意(錯38句): 26.9%/98.5%

全稱語義(錯37句): 92%/ 98.6%

待討論(有爭議)句子(22句): 0.008%

不合文法但輸出語意的句子(18句): 0.006%

老師，目前存在語意的問題比較大，但因為存在語意句子原本就較少，以全部句子的數量為分母的話占比很少。大部分是因為在程式沒有排掉含有弱疑問詞的疑問語意，使檢查器輸出結果會同時有疑問語意和存在語意。目前還在尋找問題。而全稱語義的檢查器在同時對到全稱語義和含有弱疑問詞的疑問語意時，排掉疑問語意的方法和邏輯和在存在語意的檢查器是一樣的，但不知為何存在語意的程式很多都沒有排掉含有弱疑問詞的疑問語意，全稱語義的檢查器就沒有這個問題。這部分可能需要進一步修正，但解決了這個問題之後存在語意檢查器的結果就會好很多。