# 中文詞庫群眾智慧優化計畫

## Make Robot Write In Chinese

第四組

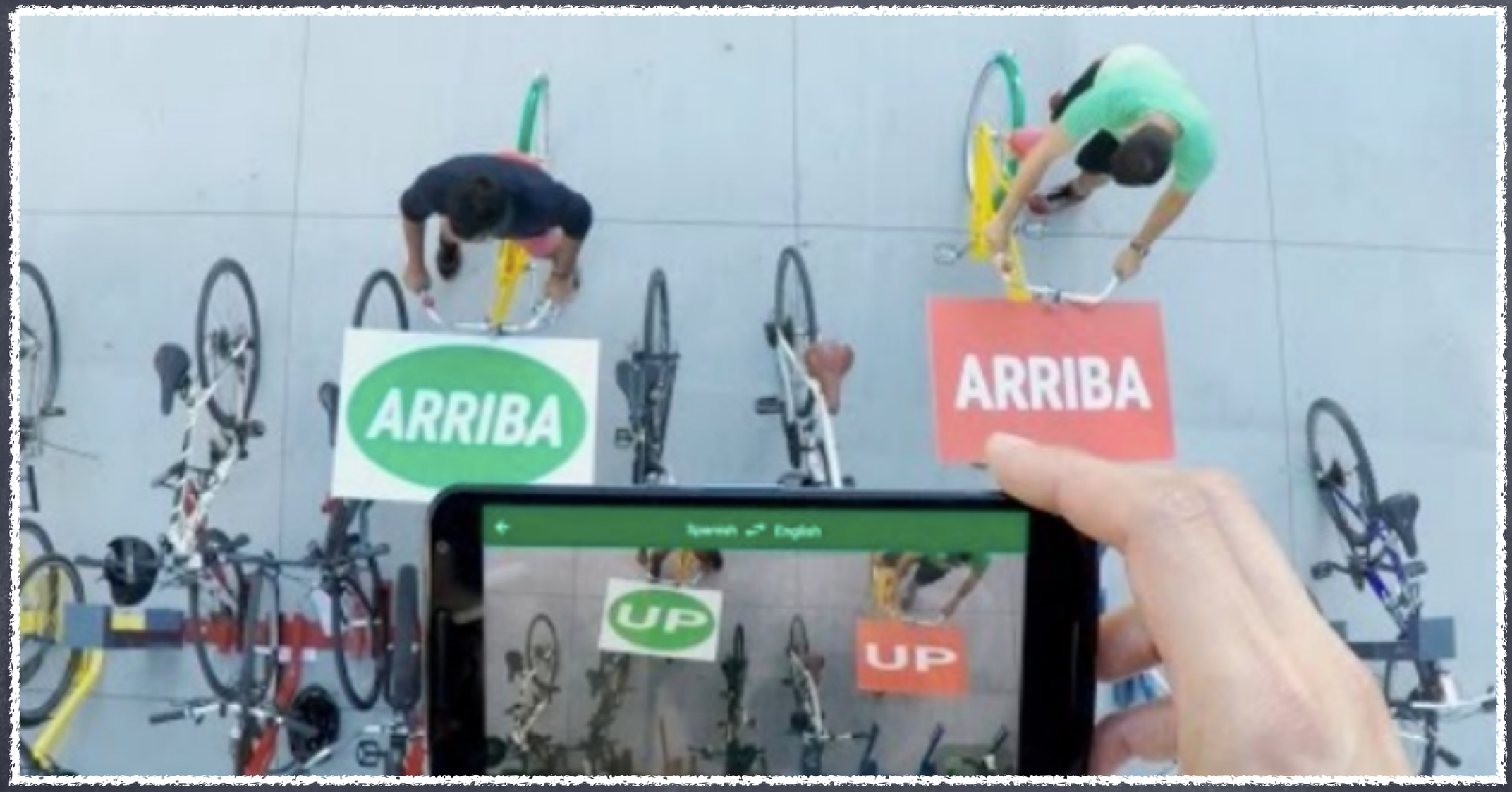
顏照銓 劉其峰 黃兆椿

吳恒毅 鍾郁婕

AP

ai AUTOMATED INSIGHTS

# 機器人撰稿




# GOOGLE 即時照相翻譯

# 不支援中文

斷詞為機器自然語應用最基本的步驟

中文斷詞較英文麻煩

# 英文斷句實例

The rapid advance across Syria and Iraq by militant fighters from Islamic State (IS) in 2014 threw the region into chaos and led to US air strikes against their key positions.

The/rapid/advance/across/Syria/and/Iraq/by/militant/fighters/from/Islamic/State/(IS)/in/2014/threw/the/region/into/chaos/and/led/to/US/air/strikes/against/their/key/positions.

# 中文斷句實例

臣亮言：先帝創業未半，而中道崩殂；今天下三分，益州疲敝，此誠危急存亡之秋也。然侍衛之臣，不懈于內；忠志之士，忘身于外者

臣亮言/先帝/創業/未/半/而/中道/崩殂/今天/下/三/分/益/州*/疲敝/此/誠/危急存亡/之秋/也/然/侍衛/之/臣/不懈/于/內/忠志/之/士/忘/身/于/外/者

* 「益州」斷句失敗，應為一詞

中文無法以空白斷句

必須逐字分析後套用統計方法

# 現有的中文斷詞資料庫


## CKIP 中文詞知識庫小組

- 字典檔年代久遠，長時間未做更新
- 每天上午六點要進行系統維護
- 不能一次送出大量資料或密集傳送資料

# 現有的中文斷詞資料庫


## JIEBA 結巴中文分詞

- 對繁體中文支援不足
- 缺乏網路的語料
- 2012年後字典即未更新

# 中文斷詞庫面臨瓶頸

- 未知詞
- 字典檔長時間未做更新
- 缺乏網路與新興用詞的語料

# 資料庫更新的速度跟不上
# 語言演進的速度

共同問題

語言是活的

會不斷演化、改變

我們相信

問題來源：中文語句的斷詞不夠精確

JIEBA +

T-JIEBA

# T-JIEBA 流程

新聞、ptt

JIEBA

CKIP

未知詞

排列
組合

推撥題目

yes

no

skip

收集回應

yes/total > 0.8

建立新詞

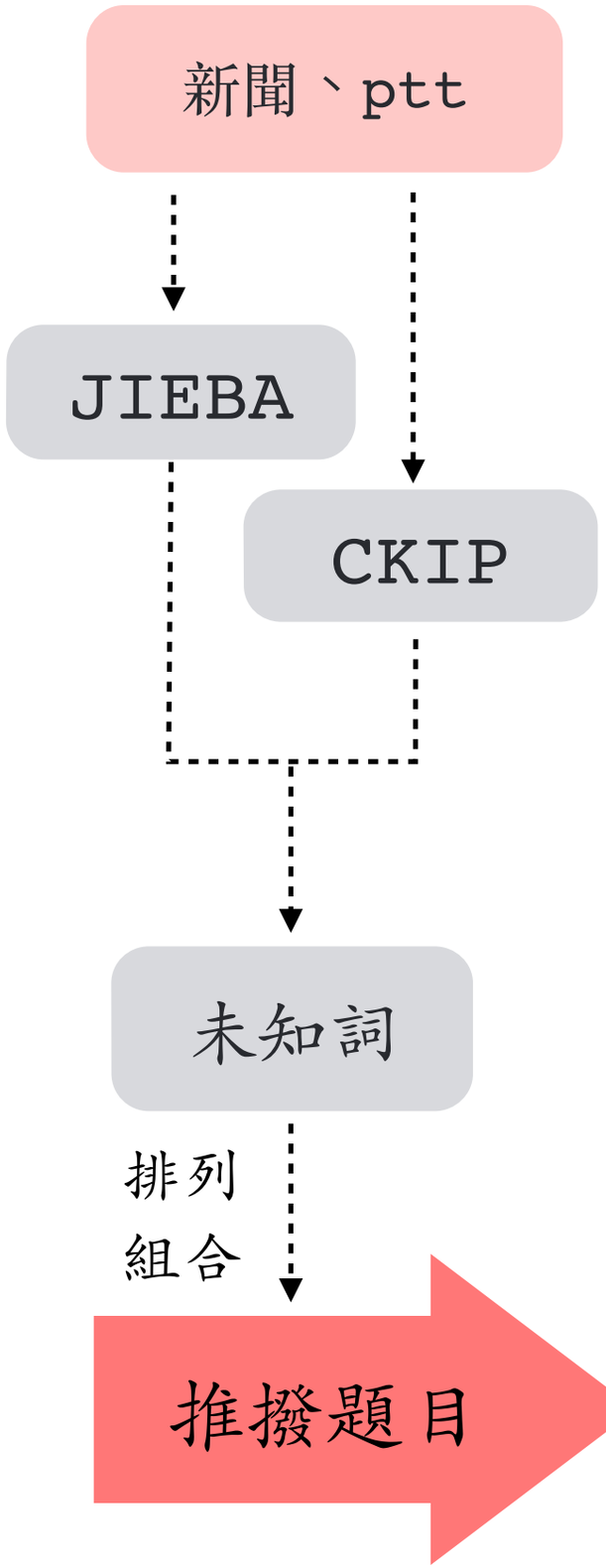
skip 次數最多

二次推撥

分析處理

優化字詞庫

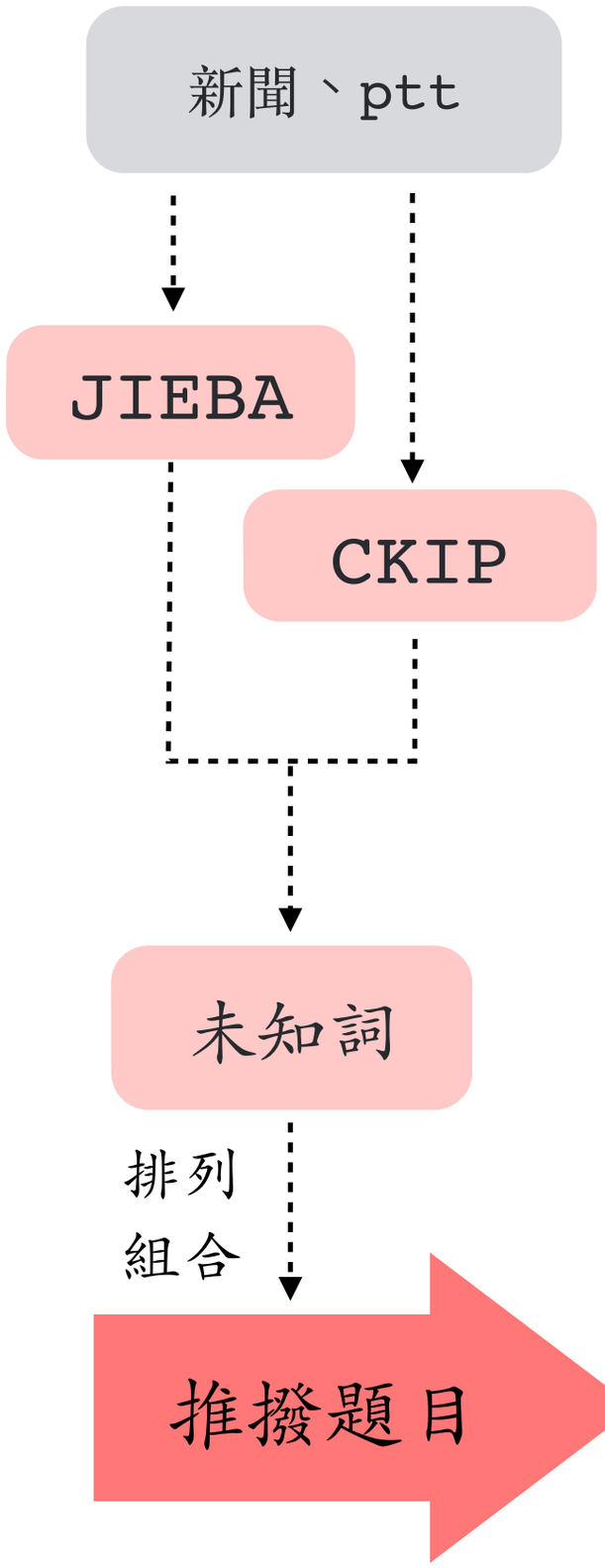
# T-JIEBA 流程




1. 資料來源

- 語料來源：東森ETtoday、PTT
- 抓取方法：爬蟲

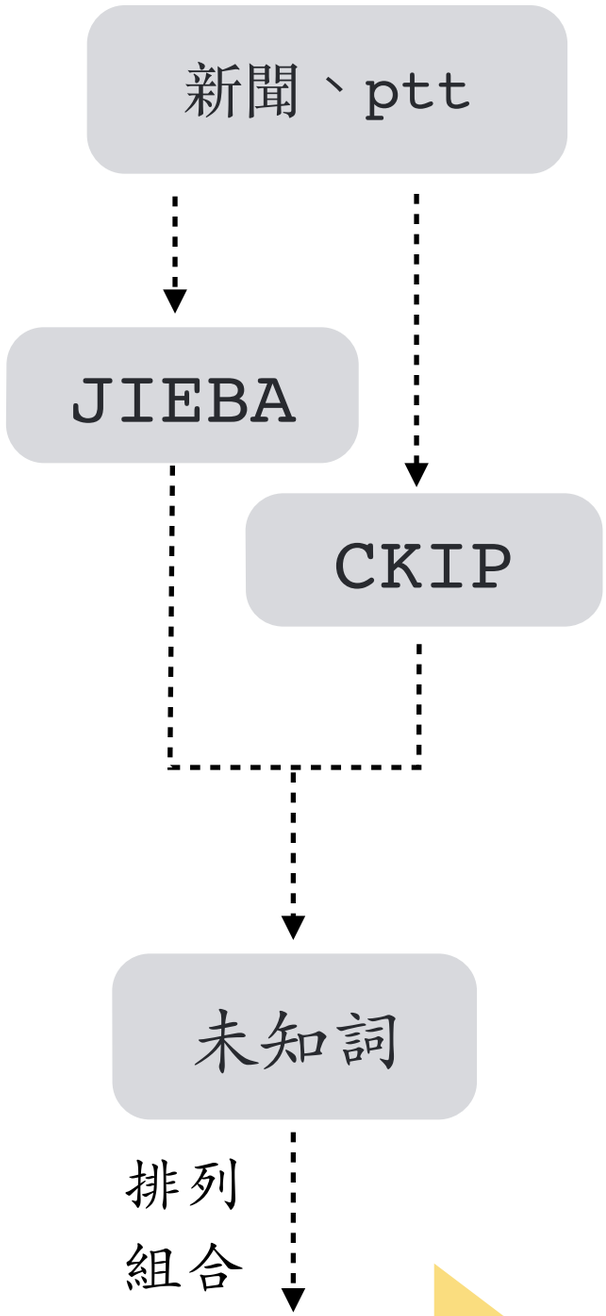
# T-JIEBA 流程

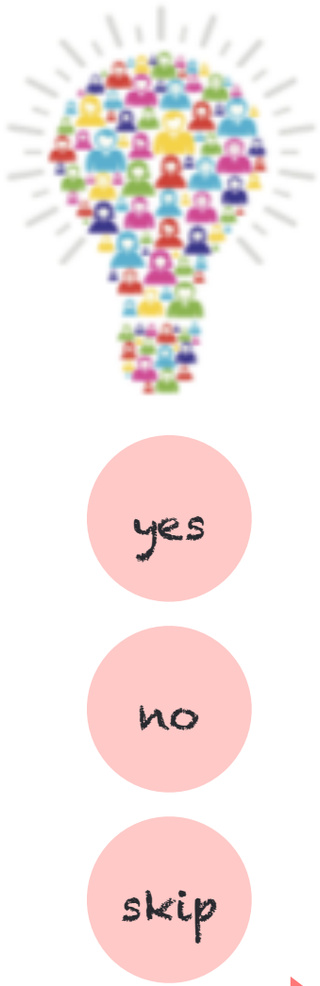



2. 前置處理

- 使用Jieba與CKIP斷詞處理
  - 未知詞
  - 相異處
- 未知詞的處理
  - 單字不送
  - 兩個詞庫的未知詞

# T-JIEBA 流程




3. 使用者回饋

- 是、不是、跳過
- 回答總數>100
- 是/回答總數 >0.8 → 詞
- 跳過的回答次數最多

→ 加入前後詞進行二次推撥

收集回應 → 分析處理 → 優化字詞庫

# T-JIEBA 流程

新聞、ptt

JIEBA

CKIP

未知詞

排列
組合

推撥題目

yes

no

skip

收集回應

yes/total > 0.8

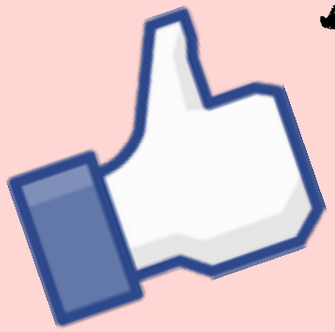
建立新詞

skip 次數最多

二次推撥

分析處理

優化字詞庫



4. 貢獻

新聞、網路用語

可以更精確斷詞

# 為什麼不考慮機器學習

- 機器學習能力有限
- 群眾智慧是現階段最直接、有效的解決方案

# THANKS FOR LISTENING