AI Clerk Platform：資訊擷取 DIY 平台
AI Clerk Platform： Information Extraction DIY Platform

Ru-Yng Chang Wen-Lun Chen Cheng-Ju Kao
AI Clerk International Co., LTD.
13F., No. 502, Sec. 2, Ren'ai Rd., Linkou Dist., New Taipei City 244020 , Taiwan (R.O.C.)
changruyng889@gmail.com; lagame53@yahoo.com.tw; sports.exp@gmail.com

摘要

自然語言處理的一門核心技術資訊擷取（Information Extraction），將非結構化（Unstructured）或是半結構化（Semi-structured）的內容，擷取部分有意義的片語/子句對應到某一個特殊的主題。可說是許多語言技術和應用的核心技術，本論文提出 AI Clerk Platform，旨在加速和提升研發資訊擷取工具的整個流程和便利性，提供友善直覺視覺化的人工標記介面，設定符合欲擷取的語意類別，執行、分配與控管人工標記任務，讓使用者在不用寫程式的情況下，就可完成客製化資訊擷取模組，並提供三種方式瀏覽和使用自建模組與其 API，進而協助其它自然語言處理技術研發與應用服務的衍生。

Abstract

Information extraction is a core technology of natural language processing, which extracts some meaningful phrases/clauses from unstructured or semi-structured content to a particular topic. It can be said to be the core technology of many language technologies and applications. This paper introduces AI Clerk Platform, which aims to accelerate and improve the entire process and convenience of the development of information extraction tools. AI Clerk Platform provides a friendly and intuitive visualized manual labeling interface, sets suitable semantic label in need, and implements, distributes and controls manual labeling tasks, so that users can complete customized information extraction models without programming and view the automatically predict results of models by three method. AI Clerk Platform further assists in the development of other natural language processing technologies and the derivation of application services.

關鍵字：資訊擷取、資訊擷取平台、資訊擷取 API、自然語言處理、DIY、AI Clerk Platform

Keywords: Information Extraction, Information Extraction Platform, Information Extraction API, Natural Language Processing , DIY, AI Clerk Platform

1 背景動機

檔案中的內容常常是用連續性的字元所組成和表達，對電腦而言這樣的呈現和儲存是難以統計、分析、理解與應用的。自然語言處理的一門核心技術資訊擷取（Information Extraction），是將非結構化（Unstructured）或是半結構化（Semi-structured）的內容，擷取部分有意義的片語/子句對應到某一個特殊的主題（Appelt, 1999）。舉例，辨識實體--人、事、時、地、物，如圖 1。透過資訊擷取讓那些非結構化（Unstructured）或是半結構化（Semi-structured）的內容，自動轉化據語意的結構化資訊，所以可以知道圖 1 中“蔡桃貴”和“蔡阿嘎”是人名，“2018 年”是時間，台北市是地，“恐龍玩具”是物，在這串非結構化的內容中，出現兩次人名。也就是透過資訊擷取那些難以統計、分析、應用與理解的非結構（Unstructured）或是半結構化（Semi-structured）資料變成可以統計、分析的。

圖 1. 資訊擷取之釋例

資訊擷取也可說是很多自然語言處理應用服務或是語言技術研發的核心技術（Wilks,, 1997）。資訊擷取最直覺的應用就是幫助達到語意搜尋，可分別指定找出文件中意指水果的“蘋果”或是品牌名稱的“蘋果”。有學者將資訊擷取用於作為文本探勘的基礎找出文本之間的脈絡（Mooney & Bunescu, 2005)、也有學者用來輔助文本生成（Koncel-Kedziorski et al., 2019)（Venkatachalam, 2020）、甚至是輔助作文本摘要（Venkatachalam, 2020）、對話系統（Yoshino et al., 2011）、聊天機器人（Ali, 2020; Jiao, 2020）。為了達到不同應用服務，根據不同應用情境領域、應用技術研發和資料特性，其中資訊擷取和識別的語意類別各有不同，例如：譬如在 Yoshino 等人（2011）的對話系統，是資訊擷取技術獲得的術語論證結構（predicate argument structures）資訊來輔助對話系統技術研發，而 Ali（2020）研發的聊天機器人所應用的資訊擷取資訊是將聊天內容中所有的實體節取出，以“I want to know the taxi rate in Islamabad（我想要知道伊斯蘭堡計程車費率）”這句話為例，“Islamabad”和“taxi”就會被擷取出認為是一個實體值。同樣是聊天機器人技術研發，Jiao（2020）研發的與股票議題相關的聊天機器人所應用的資訊擷取資訊是股票、名稱、數量、上限符號、價格、名稱這類的語意。如果能加速資訊擷取的研發過程，而且是能夠符合不同應用情境和後續技術研發，資訊擷取工具能擷取並辨識不同語意，將大幅加速和衍生其它各式語言技術研發。

就如許多自然語言技術研發過程一樣，資訊擷取的技術研發過程，會需要先有人工標記的語料，接著是演算法建立模組，產生執行檔或 API 等型態供使用。傳統若需要針對不同領域、應用情境需要資訊擷取的核心技術，需要在技術研發環境裡開始自建人工標記語料，因為需要一定數量的標記語料所建立的模組，才比較容易達到一定效能，自建人工標記語料庫的過程變成是一非常耗時、又耗人力的過程，尤其資訊擷取是要在一堆文字內容中，找出要標記的字串，並且記錄下需要被標記的字串、被標記的字串位置和所對應的語意，因此，很耗眼力，往往人工標記的任務會是由多人一起分攤與執行。然而，「需標記的字串是哪些？需標記的邊際怎麼界定？」這是最常遇到的問題，往往也跟技術研發未來的應用有關，如果遇到比較需要標記比較專業的內容，譬如：前述的術語論證結構，更非一般人可應付。無論人工標記是哪種內容或挑戰難度，標記時的標準和品質都將影響後續自動化模組效能的表現。而且，多人一起分擔還衍生出標記的品質和進度控管的問題。

因此，若有一個友善直覺視覺化的人工標記介面，能讓標記人員清楚且便利的達成人工標記任務，也可將標記任務分配給不同人，讓眾人分攤並監管整個人工標記品質和進度，標記哪些語意也是根據不同需求而自行設定，讓使用者在整個資訊擷取工具的研發過程不用寫程式，透過簡單的設定和操作步驟便能完成一個資訊檢索的 API 供呼叫引用，也就是加速整個資訊擷取工具的研發流程，相信定可對許多自然語言處理技術的各式應用服務研發有很大的幫助，而其中最大的挑戰便在於怎樣讓各領域的人都可以透過這介面完成符合客製化需求的資訊擷取工具，必須將整個流程標準化而且操作介面易理解。本論文介紹的 AI Clerk Platform 就是基於這些緣由與目標。

2 相關工具

表 1.相關工具整理

競品名稱	用途與特色	效益
Apache cTAKES, MedLee (Friedman et al. 1994,Friedman et al., 1995),等	• toolkit 單機執行工具 • 特定（醫療）領域的資訊擷取工具 • 自動擷取固定的語意字串	• 降低研發人員技術研發過程中的寫程式的工作量，直接呼叫產生更多應用

US National Center SIFR annotator (Tchechmedjiev et al., 2018)	• API, Web Demo 介面 • 特定（醫療）領域的資訊擷取工具 • 自動擷取固定的語意字串	• 降低研發人員技術研發過程中的寫程式的工作量，直接遠端呼叫 API 產生更多應用
Google AuotML NLP	• 雲端 Platform • 匯入含可種自訂各種語意的人工標記語料 • 各領域的資訊擷取模組建置，產生自製客製化 API 供遠端呼叫	• 降低研發人員技術研發過程中的寫程式的工作量，直接呼叫遠端 API 產生更多應用
IBM Watson Knowledge Studio	• 雲端 Platform • 各領域標記語料建置 • 提供友善介面，建置人工標記資料 • 在平台上建立模組，減少建立模組時撰寫程式以進行語料格式轉換和特徵擷取，模組建完，讓其餘語料以自動化完成標記	• 降低研發人員技術研發過程中的資料建置時間
	• API • 特定一般領域的資訊擷取工具 • 自動擷取固定的語意字串（人事時地物等）	• 降低研發人員技術研發過程中的寫程式的工作量，直接呼叫 API 產生更多應用
AI Clerk Platform	• Platform、API • 免寫程式 • 提供友善介面，建置各領域標記資料 • 人工標記任務分配 • 可自訂欲自動擷取的語意 • 在平台上建立各領域模組，完成自動標記，減少建立模組時撰寫程式，產生 API 供遠端呼叫或在此平台直接引用。	• 讓無資訊背景的人都可使用。 • 降低技術研發過程中的資料建置時間 • 降低研發人員技術研發過程中的寫程式的工作量，直接遠端或在平台呼叫自製客製化 API 產生更多應用

相關工具的功能特色與效益，整理如表 1，可以發現相關工具多數著重在協助標記語料建置、協助模組建置或是提供既有資訊擷取工具的單一面向，提供既有資訊擷取工具提供固定擷取的語意，沒法滿足研發各種自然語言處理過程中針對不同應用情境可能會需要擷取不同語意的各種不同客製化資訊擷取，譬如：即便都是研發智能客服系統，但研發零售業智能客服系統和政府單位智能客服系統其中所需要的資訊擷取技術所要擷取的語意就不會相同。擷取人名、組織名等的資訊擷取結果，對零售業的智能客服的技術研發效益不大。

Google AuotML NLP 雖然可以協助各領域的資訊擷取模組建置，產生自製客製化 API 供遠端呼叫，但它是以對自然語言處理或是機器學習知識相當熟悉的技術人員所使用，必須書打很多指令，人工標記的過程則非 Google

AuotML NLP 想要便民處，連匯入的人工標記資料是經由工程師轉換的格式，如圖 2。

Google AutoML：匯入人工標記資料


```
{
  "annotations": [
    {
      "text_extraction": {
        "text_segment": {
          "end_offset": 8, "start_offset": 1 }
      },
      "display_name": 疾病名稱
    },
    {
      "text_extraction": {
        "text_segment": {
          "end_offset": 68, "start_offset": 62 }
      },
      "display_name": 器官
    },
    ],
  "text_snippet":
    {"content": Fibroids are abnormal growths that develop in or on a woman uterus. }
}
```


圖 2. Google AuotML NL 人工標記資料匯入

IBM 有提供協助人工標記和任務分配的功能，但建立好的模組僅限於在平台上自動標記其餘尚未標記的語料，所以目標著重在協助建立人工標記語料。

本論文提出的 AI Clerk Platform 則包含提供友善介面，將人工標記人物分配，協助人工標記語料建置，並且可自訂語意，在平台上建立各領域模組，完成自動標記，減少建立模組時撰寫程式，產生 API 供遠端呼叫或在此平台直接引用，所以整合並簡化了資料擷取技術研發過程中的各階段步驟，整個過程免寫程式，讓即便是非資訊人員的人，也可輕易的完成資訊擷取工具。

3 AI Clerk Platform 功能

● 協力建置領域標記語料

建立人工標記語料是自然語言處理領域最耗時的地方，改善人工標記是必要的，為了解決此情形，AI Clerk Platform 可以讓使用者根據自己的領域，自訂該領域的語意標籤，如圖 3，並且提供友善化人工標記介面，如圖 4，使用者只需要選取文字，再選擇語意標籤即可，介面會以顏色區隔標記文字，使用者可以更容易查看標記的文本以及輕鬆完成標記，不再需要建立 Excel 檔或自行撰寫標記介面。

圖 3. 自訂欲自動擷取的語意標籤

圖 4. 友善的人工標記介面

圖 5. 標記員任務管理

除此之外，AI Clerk Platform 提供任務管理機制，如圖 5，讓使用者將人工標記任務分割並分配給不同標記者，並且查看標記進度，與進行針對標記狀況進行審核，以利標記語料建置成果維持一定品質。

● 自建領域資訊擷取模組和 API

AI Clerk Platform 可以讓使用者建立資訊擷取模組和 API，如圖 6，使用者不需撰寫程式，只需點選按鈕即可進行模組訓練，平台提供包含常見的「訓練全部語料」、「80%訓練語料 20% 預測語料」、「5 Fold Cross Validation」三種訓練模式，滿足各種實驗設定，並降低了技術門檻和實踐過程，人人都可自建模組。

圖 6. 透過點擊與設定啟動模型之訓練

提供三種呼叫和瀏覽模組預測結果，如圖 7。第一種是線上預覽，使用者可以在平台上直接輸入文字，平台即時預測並顯示在介面上；第二種是遠端呼叫 API，系統會告知使用者呼叫方式，使用者就可以自行撰寫程式來進行大量呼叫使用；第三種是匯入 Excel 並執行自動預測，使用者將預測文本以 Excel 上傳至平台，同時預測多筆文本，並可下載包含預測結果的 Excel 檔案，如圖 8 為匯入 Excel 自動預測結果之釋例，Excel 是最普及且親民的文書處理軟體，使得人人都可以享受資訊擷取的強大效果。

圖 7. 提供三種方式呼叫和瀏覽模組預測結果

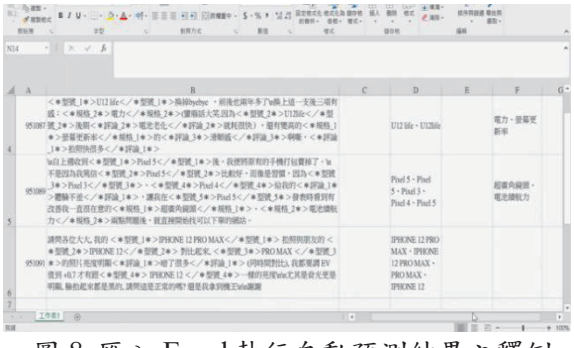

圖 8. 匯入 Excel 執行自動預測結果之釋例

- 提供資訊擷取特殊領域 API

建立特殊領域資訊擷取模組除了考驗人工標記人力，也考驗特殊領域的人力，標記人力也需要具備特殊領域的知識才可以標記，教育時間也更為曠日廢時。

因此，AI Clerk Platform 提供一些訓練完成的資訊擷取特殊領域 API，讓使用者可以直接呼叫使用，目前已經有 3C 產品（如圖 9）以及保險商品領域可以使用，3C 產品 API 針對手機產品擷取效果最好，保險商品 API 可處理常見意外險、醫療險、壽險、罐頭保單的相關文本內容。

圖 9. 3C 產品資訊擷取 API 預測結果之釋例

4 AI Clerk Platform 平台效益

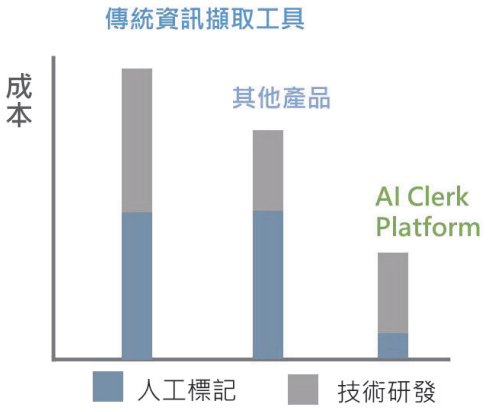

圖 10. 人力成本相對性比較示意圖

- 大幅縮減人工處理資料成本

AI Clerk Platform 藉由友善人工標記介面、任務管理並輔以後端演算法機制，大幅下降了人工標記人力，其餘競品都仍需要大量人力投入。如圖 10，AI Clerk Platform 與相關工具所提及的產品做比較，做了成本相對性比較示意圖，表達成本相對高低的概念，AI Clerk Platform 可以大量下降人力成本。根據本團隊針對建立 3C 產品領域模組為實驗，以手機類文章和其它相機、電腦類文章相比，手機類模組訓練全部此採用人工標記資料建置，而相機與電腦類文章僅有部分訓練語料是由人工標記，同樣是在論壇、業配文中找出商品種類、型號、規格、功能、描述、評論、價格的語意概念，透過後端演算法機制，當電

腦類和相機類人工標記文章數量為手機類的20%時，就可達到和手機類文章同等級效益。因此，推估約可以節省 80%的人工成本。因此可加速在更多特殊領域的技術研發和應用是可期待的。

- 不用寫程式，完成客製化資訊擷取 API

不用寫程式是 AI Clerk Platform 重要特色，使用者透過人工標記介面，可以用設定、選取的方式完成標記，模組與 API 建立只需點選按鈕進行，使用模組也可以透過匯入 Excel 執行和瀏覽自動預測結果，這些功能特色除了對研發更為便利，也更有助於連非資訊背景的人都可使用。

- 衍生各式資訊擷取或自然語言處理應用服務，減低技術門檻和成本

一般的競品需耗費大量的人力成本來建置資訊擷取模型，而且都需要仰賴工程師來建立模組，也意味著需耗費更多的時間成本來完成智慧應用服務。

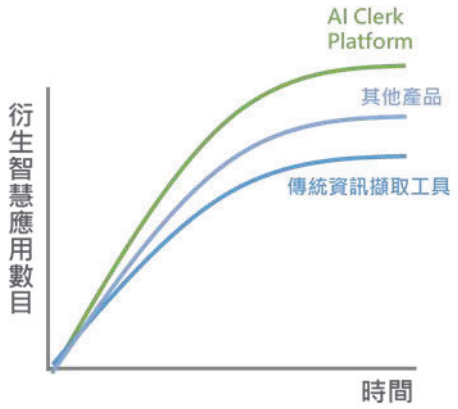

圖 11. 智慧應用相對性比較示意圖

AI Clerk Platform 搭配人工標記介面、任務管理和後端演算法機制降低人力成本與耗費時間，相對於其它競品來說 AI Clerk Platform 的功能可以大幅簡化了資料擷取技術研發過程，節省了人力成本等於也加快衍生各項以資訊擷取技術為基底的各類智慧應用服務之研發，如圖 11，與相關工具所提及的產品做比較，做了智慧應用數目相對性比較示意圖，以曲線相對高低來呈現數目相對多寡。

5 結論

本論文提出一個結合「協力建置領域人工標記語料」、「自建領域資訊擷取模組和 API」、「提供資訊擷取特殊領域 API」特色的 AI Clerk Platform。有別於現存的資訊擷取工具，AI Clerk Platform 可以讓使用者自訂語意標籤滿足客製化需求，透過任務管理機制和友善的人工標記介面，讓使用者可以輕鬆建置領域標記語料，並且可以在無需撰寫程式的前提下，自行建立領域資訊擷取模組和 API，並提供線上預覽、遠端呼叫以及匯入 Excel 執行和瀏覽自動預測結果，也提供特殊領域的資訊擷取 API，有 3C 產品以及保險商品。AI Clerk Platform 可以大幅縮減人工處理資料成本，並且快速衍生各種自然語言處理應用服務，相信 AI Clerk Platform 可以協助學界提升產能與效率。

參考文獻

Appelt, D. E. (1999). Introduction to information extraction. Ai Communications, 12(3), 161-172.

Wilks, Y. (1997, July). Information extraction as a core language technology. In International Summer School on Information Extraction (pp. 1-9). Springer, Berlin, Heidelberg.

Mooney, R. J., & Bunescu, R. (2005). Mining knowledge from text using information extraction. ACM SIGKDD explorations newsletter, 7(1), 3-10.

Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., & Hajishirzi, H. (2019). Text generation from knowledge graphs with graph transformers. arXiv preprint arXiv:1904.02342.

Venkatachalam, S., Subbiah, L. P., Rajendiran, R., & Venkatachalam, N. (2020). An ontology-based information extraction and summarization of multiple news articles. International Journal of Information Technology, 12(2), 547-557.

Yoshino, K., Mori, S., & Kawahara, T. (2011, June). Spoken dialogue system based on information extraction using similarity of predicate argument structures. In Proceedings of the SIGDIAL 2011 Conference (pp. 59-66).

Ali, N. (2020). Chatbot: A Conversational Agent employed with Named Entity Recognition Model using Artificial Neural Network. arXiv preprint arXiv:2007.04248.

Jiao, A. (2020, March). An intelligent chatbot system based on entity extraction using RASA NLU and

neural network. In Journal of Physics: Conference Series (Vol. 1487, No. 1, p. 012014). IOP Publishing.

Apache cTAKES, https://ctakes.apache.org/

Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. J Am Med Inform Assoc. 1994; 1:161 – 74. [PMC free article] [PubMed] [Google Scholar]

Friedman C, Starren J, Johnson SB. Architectural requirements for a multipurpose natural language processor in the clinical environment. In: Gardner RM (ed). Proceedings of SCAMC 1995. Philadelphia: Hanley & Belfus, 1995:347 – 51.

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Zevio, S., & Jonquet, C. (2018). SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. BMC bioinformatics, 19(1), 1-26.

Google AuotML NLP, https://cloud.google.com/natural-language/

IBM Watson Knowledge Studio, https://www.ibm.com/tw-zh/cloud/watson-knowledge-studio