醫學電資整合創意專題報告

報告主題：自然語言處理

副標：醫學詞彙縮寫轉換為全稱之 NLP 實作與應用

組員：林聖硯、杜冠勳、李承祐、彭鈺峯、鄧立玨、謝一陽

一、簡介

此次報告目標主要為訓練一個可以將醫學詞彙縮寫轉換為全稱的自然語言處理模型。使用資料集來源為 MeDAL 資料集，並參考先前研究之模型架構進行 NLP Training，以下將依序描述選擇此題目的動機、此模型之可能使用情境、訓練使用的資料集、訓練使用的模型、實驗結果、使用者介面之設計與未來可改善方向。

二、動機

縮寫在拼音文字中常被使用，定義為將根據意義來把全稱縮短為以全稱內特定文字代表的短詞，通常取全稱內各單詞之首字母以大寫形式表示。例如 Parkinson's Disease 取 PD 兩單詞首字母作為縮寫，經過群體內協商縮寫之實質涵義後，縮寫便是個能縮短使用者撰寫文章時間且較方便使用者記憶與使用的工具。但在生活情境中，縮寫亦有其不便之處，主要分為三個面向，第一個面向，若不同群體間不了解對方縮寫之意義，便無法以此縮寫進行溝通，而若其中有一群體使用縮寫時，會使另一群體無法理解詞句想表示的意涵；第二個面向，同一縮寫於不同群體間可能對應至不同意義，於群體間交流時便可能導致詞句理解不同，而無法成功傳達縮寫使用者想要表達的意思；第三個面向，同一群體中可能有相似縮寫，若於書寫時拼寫略為錯誤，或閱讀者意會成另一個相似縮寫所對應的意思，都可能傳達錯誤的資訊，於醫療場域中，此微小差異便可能導致醫事人員採取錯誤的醫療行為，或阻礙醫事人員間，甚至是醫事人員與病患間的溝通。

因此，我們便想利用機器學習，藉由將縮寫的上下文（英文）與縮寫（全大寫英文組合而成）輸入模型，可以自模型中輸出可能的醫學全稱（英文），幫助使用者判讀縮寫於此文章內所對應的全稱，同時會輸出與此醫學名詞相關的 Wikipedia 簡介與 PubMed 相關論文，以供使用者進行進一步查詢。此外，若能將此模型整合到電子文檔之閱讀器中，便可以讓醫生無所顧忌地使用縮寫，提高醫生撰寫文章，尤其是撰寫病歷、論文的效率，且讓使用者閱讀以上文本時，可以正確解讀文本資訊。

三、模型使用情境

以下分三種情境說明模型可供醫事人員與醫事人員教育之使用情境，以供進一步思考模型運用與修正之發想。

(1) 情境一：臨床醫師照會或調閱病歷

不同科別之中對於不同縮寫有不同解釋，例如 PD 於精神科被解讀為

Psychiatric disorder（精神疾患），而於神經科被解讀為 Parkinson's disease（帕金森氏症），且兩科之間由於領域之相似性常需要共同處理病患，於閱讀病歷或溝通時，可能將文本中的縮寫解釋為不同含意而造成決策失誤，故醫院會限定僅有特定詞彙能使用縮寫。但若能將此模型應用至此情境，模型便可依照病歷中的描述判定縮寫之含意，而不會因縮寫造成誤解，如此醫生們便可隨心所欲地使用縮寫，減少醫生撰寫或解讀文本所耗的心力與時間。

(2) 情境二：醫學生之臨床學習

醫學生剛進入臨床端學習時，對於常使用的縮寫詞彙不甚理解，且在各科病房學習時，因接收到過多縮寫卻對縮寫涵義不甚熟悉，可能導致縮寫間含意的混淆，使學生學習效率降低。而此模型便可幫助醫學生篩選於特定情境下縮寫詞彙之可能含意，並提供進一步理解詞義的方向，提高醫學教育的效率。

(3) 情境三：閱讀論文或醫學文章

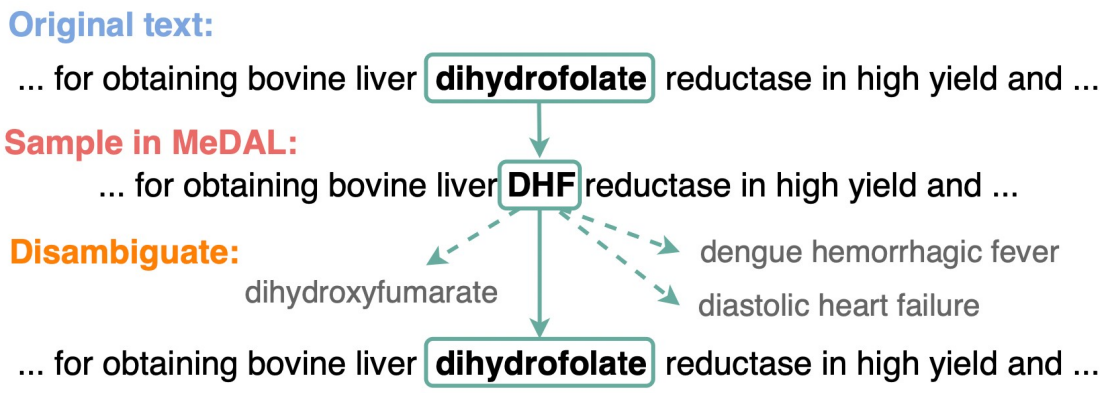
閱讀論文或醫學文章時，雖然文本之開頭皆會提示文本中將以哪一個縮寫來代替特定全稱，但若需回顧眾多文章時，可能無法記憶特定文章中縮寫的意義，導致進行統整時需要不斷查詢該文章中縮寫對應的含意。此模型便可以藉由上下文，找出文章中縮寫的涵義，而無須再次翻回文章開頭尋找縮寫所對應的全稱，避免執行此動作所消耗的時間與心力。

四、任務介紹

(1) 任務定義：Abbreviation Disambiguation (AD)

我們這次要完成的任務在機器學習的領域中被稱為岐異消除（Abbreviation Disambiguation，簡稱 AD），這個任務是指在特定情境中識別並解析縮寫意義的過程。由於相同的縮寫可能在不同情境中用於不同的事物，因此是一項具有挑戰性的任務。在這個任務中，由於 output space 裡面涵蓋很多 expansions，而一個縮寫只會對應到一個 expansion，故我們也可以把這個任務視為一個多分類任務（Multi-class classification）。

圖一中，在不同的情境下，縮寫"DHF"可以代表"dengue hemorrhagic fever"（登革熱）、"dihydroxyfumarate"（草醯羥乙酸還原酶）或"diastolic heart failure"（舒張性心衰竭），但在此文本這個特定情境中，DHF 指的是 dihydroxyfumarate。為了確定縮寫的正確意義，模型必須透過查看周圍的文本來考慮其使用的情境。所以岐異消除是自然語言處理（NLP）領域中的一項重要任務，它可以提高文本分析和搜索結果的準確性和效率。透過正確識別和解析縮寫的意義，人們可以更好地理解文檔或溝通的內容和情境。



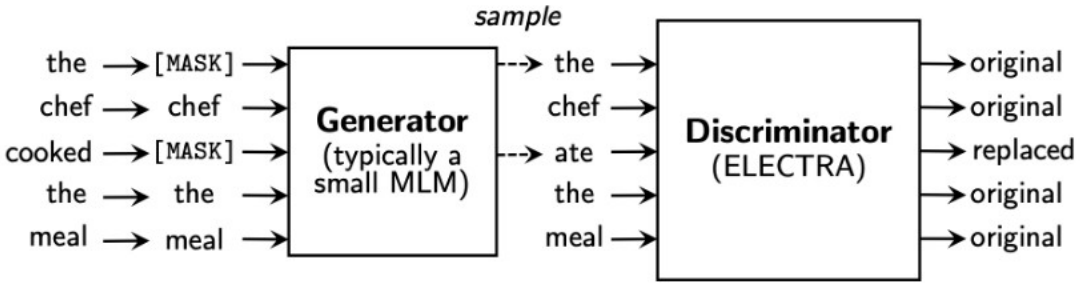
圖一

(2) 資料集：MeDAL

我們這次任務訓練所使用之資料集為 The Medical Dataset for Abbreviation Disambiguation for Natural Language Understanding (MeDAL) [1]，此為一大型醫學文本的資料集，專門為將縮寫消除歧義相關任務所設計的資料庫。資料庫之資料來源為 PubMed，總共包含 14,393,619 份論文，而每篇論文平均約有三個縮寫，其中縮寫與全稱的配對通常為單一縮寫對應到多個全稱，整個資料集包含 5,886 個縮寫，共有 24,005 組縮寫與全稱的配對。資料格式包含四欄，分別為 abstract_id、text、location、label；abstract_id 為 PubMed 論文之代號；text 為取自 PubMed 中 abstract_id 對應之論文中含有縮寫的段落，以字串儲存，location 為 text 段落中縮寫所對應的字元位置，數字表示的是第幾個單詞為縮寫，以序列儲存，label 為 location 代表字元之縮寫對應的英文全稱，同樣以序列儲存。

(3) Benchmark model: ELECTRA

Wen. et al. (2020) [1] 在提出 MeDAL 這個資料集時就同時使用 ELECTRA 架構訓練出一個解決這個任務的 Benchmark model。ELECTRA [3]是由 OpenAI 研發的自然語言模型，這個模型的架構（圖二）包含一個生成器和一個識別器的生成對抗網絡（Generative Adversarial Network，GAN）。生成器的目的是生成虛假的文字，而識別器的目的是將真實文字與虛假文字區分開來。通過不斷同時訓練這兩個網絡，ELECTRA 模型可以學習文本中常見的語言模式，並使用這些模式來預測文本中的單詞或字符是否被改變。



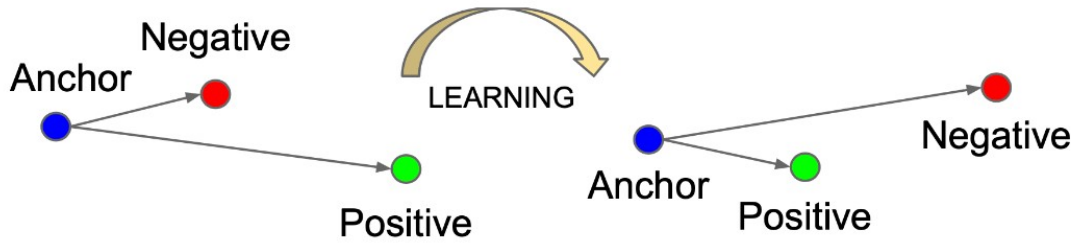
圖二

在這篇論文內，Wen. et al. (2020) [1] 上述提及的 MeDAL 資料集預訓練整個 ELECTRA-GAN 的模型架構，並且最後使用了 discriminator 的部分當作 AD 任務的 Encoder，最後接上 Multi-Layer Perceptron 來完成分類任務。

(4) Proposed method: Contrastive learning – triplet network

ELECTRA 使用的 MeDAL 資料集裡面共有三百萬筆的 training data，但 output space 的 expansion 數量卻有約兩萬四千筆，等於每一個 expansion 對應到的權重只會被更新到一百五十次，這就讓我們想到了深度學習中一項能夠讓非正確 expansion 的權重也被更新的技巧──對比學習（Contrastive learning）。

對比學習是一種機器學習方法，用於學習對於特定任務有用的 representation，這個方法的精髓在於讓模型學習資料點的不同並試圖透過並透過距離損失函數計算模型的 loss 來反向傳播並學習他們之間的差異。而我們這次使用 contrastive learning 中的三元網路（triplet network）來套用在我們這個 AD 的任務上。Triplet network（圖三）是基於三元損失（Triplet loss function）的概念來設計的網絡，在這個網路中我們要讓模型吃進三個不同的資料點，分別為錨點（Anchor）、正樣本（Positive）和負樣本（Negative）。在三元網絡中，模型透過三元損失被訓練為想要「拉近」正樣本與錨點的距離並且「推遠」負樣本與錨點的距離。



圖三

在我們的任務中，我們的錨點會放入尚未將縮寫還原為 expansion 的句子、正樣本會放入將縮寫還原為「正確」expansion 的句子、負樣本會放入將縮寫還原

為「錯誤」expansion 的句子，範例資料如下：

- Anchor: Inotropic reserve identified by dobutamine or dipyridamole SE is associated with a better outcome in patients with idiopathic dilated cardiomyopathy dcm …
- Positive: Inotropic reserve identified by dobutamine or dipyridamole stress echocardiography is associated with a better outcome in patients with idiopathic dilated cardiomyopathy dcm …
- Negative: Inotropic reserve identified by dobutamine or dipyridamole state entropy is associated with a better outcome in patients with idiopathic dilated cardiomyopathy dcm …

這樣的假設是我們認為模型有辦法辨別出，一個字的改變造成整句話意思改變的細微差別，而我們希望透過 triplet network 的架構可以讓網路學習到能辨別出正樣本、負樣本的 backbone。在 Inference 的階段，若模型發現正樣本相較於負樣本與錨點的距離較近，代表模型將樣本分類正確。

五、實驗

(1) 實驗設定

由於時間上的限制，我們只使用了 MeDAL 原論文資料集 5%的資料點，也就是 training、validation 以及 testing 的資料點數分別為 150,000、50,000 以及 50,000。資料採樣的過程中，我們使用了 Stratified sampling 來讓每一個 class 的資料平均分佈在每一份資料集內。我們參考 Seneviratne et al. (2022) [4]的作法，使用 triplet network 的架構，採用預訓練在 Biomedical 任務上面的 BioBERT(Lee et al., 2020) [5] 來訓練整個網路，並在 BioBERT 之後加入了一層將 768 維度 features 映射到 64 維度的 MLP，再對結果算 triplet loss。在超參數的部分，也以 Seneviratne et al. (2022)的作法為基準，使用 Adam optimizer、並調整了 batch size、learning rate、dropout rate 三個超參數來優化模型。而在 ELECTRA 的部分，我們使用了論文 github[2]所提供的 ELECTRA 預訓練好的模型，並將此模型 testing 在我們新採樣出來的 testing 資料集上。

我們使用的 metrics 為 accuracy，模型的訓練結果如下：

模型	Testing accuracy
BioBERT with triplet network[a]	67.032%
ELECTRA	83.184%

a. 超參數：dropout = 0.05, batch size = 2, learning rate = 1e-5

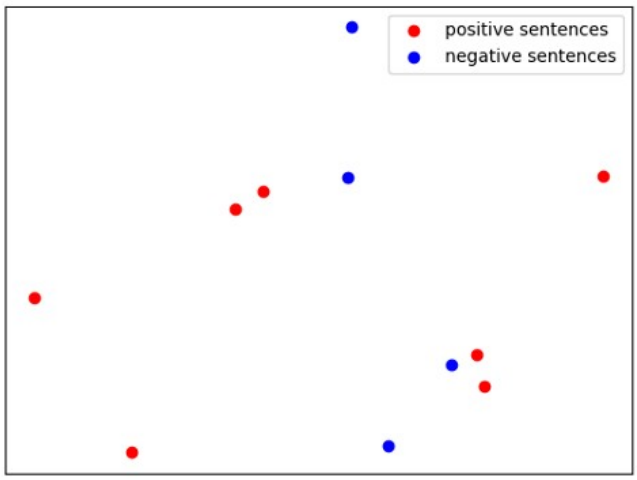
(2) 實驗結果

在訓練的過程中，我們發現 dropout rate 對模型的訓練結果有很大的影響，而其他兩個參數的影響甚小。由於在最後一層 MLP 之後，我們依照原本的 Seneviratne et al. (2022)的作法會接上一層 dropout layer，但這樣會導致損失函數的結果大大受到 dropout layer 的影響，直到將 dropout rate 的比率逐漸調低，loss 以及 accuracy 才慢慢回復正常。而 batch size 因為硬體的限制，無法像 Seneviratne et al. (2022)設置為 32，我們最高只能上升到 4，但最後發現 batch size 為 2 的結果稍佳。

最後在盡力嘗試之下，accuracy 仍無法超過原本 ELECTRA 的表現，由於在 triplet network 網路中影響最大的是最後一層 MLP 的 feature，我們便將最後一層的 feature 降維視覺化後，來觀察網路有沒有將正樣本與負樣本分開。這代表我們的假說可能有問題，模型並不會因為一個字的改變就學習整個句子的語意變化。

(3) 資料視覺化

我們透過 Principal component analysis (PCA)將縮寫涵蓋"DHF"的資料點的正樣本以及負樣本的 feature 的維度降為 2 之後做視覺化，結果如下：



圖四

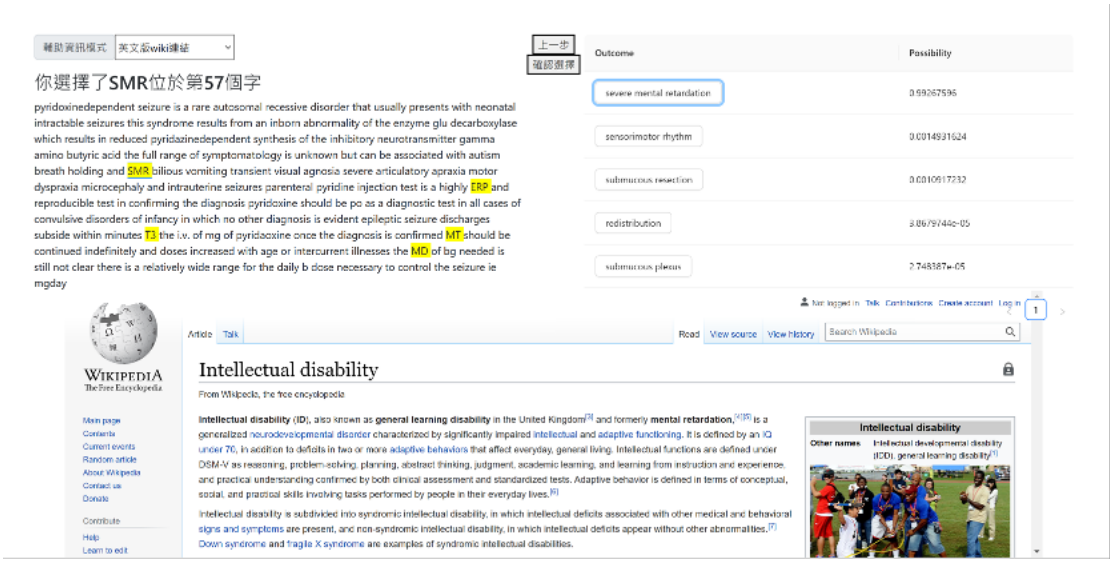
從圖四的結果我們可以發現，模型能有效將一些負樣本與正樣本的資料點分開，左邊的四個正樣本距離較近，而中間的負樣本距離也較接近，但右邊仍有一些正樣本與負樣本的距離較接近，顯示模型並不能完全有效地將所有正樣本與負樣本的距離拉開，也可能是導致模型表現不好的原因。

六、使用者系統介紹

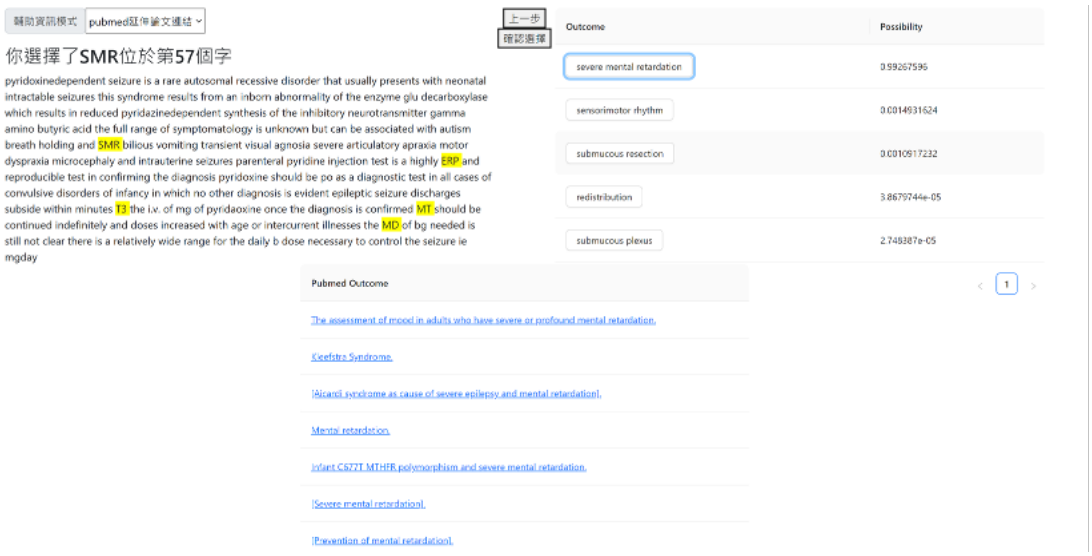
我們這次用 react 這個框架架設一個前端網站 demo，將選取的文字和位置以 API 打包後，傳到後端的深度學習模型，並回傳我們預測縮寫的結果。由於後端深度學習使用的語言為 Python，我們使用 child_process 的模組來執行 shell commands 來呼叫模型。

圖五

我們設計的介面如圖五，可以見到左方可輸入要分析的文字，右方會顯示我們模型預測的縮寫，並且我們根據前面所提到三個使用情境，再提供三個輔助資訊模式：英文版 wikipedia 查詢、中文版 wikipedia 查詢和 PubMed 的論文連結。前兩者是讓使用者能概略瞭解該詞的意思，直接提供最相關的 wikipedia 頁面；PubMed 的論文連結則是給予需要更深度資訊的使用者，提供相關的 PubMed 連結。圖六及圖七為輔助資訊模式示意：



圖六



圖七

七、結論及未來展望

在這次專題中，我們使用 MeDAL 資料集來進行 Abbreviation Disambiguation 的任務。我們嘗試使用 contrastive learning 來解決每一個類別權重更新過少的問題，但很不幸地模型的結果並沒有超越原本 MeDAL 論文提出的結果。

在技術方面，若沒有硬體及時間上的限制，我們會使用 m-networks，透過多組 negative points 而非一組 negative poin 藉以改善 contrastive learning 的結果。我們也會嘗試使用 gradient accumulation 來加大 batch size，使得模型在更新權重時學習到更準確的梯度。

而在使用者介面方面，我們認為有以下三個面向能夠改善：第一個能改善的方向為程式本身效率，目前從確認輸入到得到結果大概需 3 至 4 秒，尚有一大段進步的空間。這點可以藉優化 child process，後端程式直接用 Python 撰寫，或使用如 flask 等現有框架來增加效率，提升使用者體驗。第二個能著手的方向是增添功能，如讓使用者輸入一段含有縮寫的文章，程式會回傳將所有縮寫都變為全稱的結果。醫事人員為了講求效率常常會使用縮寫，但也會造成溝通之間的問題，若能開發此工具讓使用者能用縮寫完成文章，再放入此工具來將縮寫轉成全稱，便能既有效率的撰寫也避免溝通問題的產生。第三個能改進的方向是增加可以應用的平台，如提供 chrome 外掛等，避免轉貼到網站的麻煩步驟，並使醫事人員可以更輕鬆的瀏覽論文、教科書，提升閱讀的效率。

我們這次專題的貢獻有以下兩點：第一，我們設計了一個專屬於醫學文章 abbreviation disambiguation 的網站，對應不同情境的功能使得醫學人員能更方便取得需要的資訊。第二，我們提出的模型能套用在其他真實的資料集上。若之後能拿到醫生手寫的病人診斷，而非 PubMed 上的期刊文章，就能再次訓練模型並做出一個貼近真實場景的系統，模型判斷的也會更準確。我們期望這次專題的結果能促進相關領域以及縮寫轉換系統的發展。

八、參考資料

[1] Wen, Z., Lu, X. H., & Reddy, S. (2020). MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining

[2] MeDAL github: https://github.com/McGill-NLP/medal

[3] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators

[4] Seneviratne, S., Daskalaki, E., Lenskiy, A., & Suominen, H. (2022, July). m-Networks: Adapting the Triplet Networks for Acronym Disambiguation-Networks: Adapting the Triplet Networks for Acronym Disambiguation.

[5] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

九、組員貢獻

組員	程式	報告
林聖硯	資料處理、模型改善	上台報告資料集、實驗結果和使用模型 書面報告任務介紹、實驗結果、未來改善方向
杜冠勳	使用者介面製作	上台和書面報告使用者介面、未來改善方向
李承祐	使用者介面製作	上台和書面報告使用者介面、未來改善方向
彭鈺峯	N/A	投影片美編、書面報告美編
鄧立珏	N/A	書面報告簡介、動機、使用情境、資料集介紹
謝一陽	N/A	上台報告簡介、動機、使用情境