國立台北大學資訊工程學系專題報告

# 裁判字號引用偵測與分類：結合兩階段標註與資料自動化擴增策略

專題組員：陳冠任、陳恪行、徐少綸、薛浥塵

專題編號：PRJ-NTPUCSIE-114-001
執行期間：2025 年 9 月 至 2026 年 6 月

## 摘要

法院判決間的引用關係反映司法見解的形成與演變，但判決書中裁判字號的出現未必代表實質引用，使引用關係自動辨識具有挑戰性。本專題提出一套兩階段語意理解架構，第一階段辨識判決書中的裁判字號（Judgment Number, JN），第二階段判斷其是否構成實質引用（Semantic Support, SS）；並結合自我訓練（Self-Training）機制，於偽標籤篩選過程導入專家知識，以降低人工標註成本並提升模型效能。實驗結果顯示，我們的方法在少量標註資料下仍可透過迭代學習達成約九成以上的引用偵測準確度，展現其於法律資訊檢索與判決引用網絡建構之應用潛力。

關鍵字：法律科技、判決引用偵測、語意理解、自我訓練

## 1. 緒論

法院判決書判決之間的引用關係，反映法院見解形成、延續與演變的脈絡。一份裁判引用哪些前案，往往代表其法律立場與論理來源；而一份裁判若被後續判決反覆引用，則可反映其在司法實務中的影響力與穩定性，因此，若能有效整理判決引用關係，便能協助使用者快速掌握關鍵前案、理解法律見解的發展方向，並提升法律檢索與研究效率。

然而，據我們所知，目前並沒有相關研究與技術能自動化處理判決間之引用關係。現有法律資料庫仍多以全文檢索為主，使得使用者輸入關鍵字後，仍須自行從大量判決中判斷哪些裁判真正具有參考價值。

判決書引用自動偵測的技術困難點在於：判決書中裁判字號格式常因法院、年代或法官文書習慣而具有高度多樣性，使得裁判字號的出現偵測具有一定的困難度，並且，由於判決書中常同時包含案件事實、程序經過、當事人主張與法院理由，因此，裁判字號的出現並不一定代表法院實質引用該裁判作為論理基礎，使得問題更加難以處理。

為了解此問題，本專題以「從長篇判決書中自動辨識“被實質引用的裁判字號”」為研究目標，提出兩階段語意理解模型：第一階段負責偵測判決書中的裁判字號（Judgment Number, JN），第二階段則根據 JN 附近的語意脈絡，判斷其是否構成實質引用（Semantic Support, SS）；並搭配自我訓練機制（Self-Training），在偽標籤篩選（Pseudo-Label Selection）環節，引入專家知識引導，以有效減低法律專業人力負擔，使方法可務實落地。實驗結果顯示，語意理解模型能有效處理判決書中裁判字號格式多樣與引用語句表達不固定的問題，並且在兩階段學習架構下，我們的方法可僅以少量人工標註資料，迭代訓練達成九成左右以上的判決書引用偵測準確度。

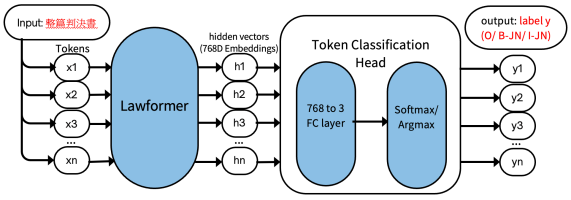
## 2. 研究方法：判決書引用偵測

從長篇判決書中偵測被引用裁判字號主要有兩項挑戰。第一，裁判字號書寫方式可能因法院、年代或法官文書習慣而有所差異，使得其格式具有高度變異，例如：「最高法院刑事大法庭110年度台上大字第5660號裁定」、「112年度偵字第2563、5943號」、「臺灣高等法院臺南分院一一一年度上訴字第一六七一號刑事判決」、「臺灣高等法院暨所屬法院民國105年法律座談會民事類提案第43號」等，因此，若僅依賴正規表達式，規則過嚴容易漏判，規則過寬則容易誤判；第二，裁判字號可能出現在不同語境中，而僅有當法院用以支持其法律見解的前案依據，為實質引用；其餘則僅是為案件事實、程序經過或當事人主張中的單純提及。顯然「引用」與「提及」之法律意義與重要性截然不同，因此，我們仍需依據上下文進行語意判斷。

因此，我們將判決書引用偵測切分成兩階段任務，並提出兩階段語意理解模型，分別處理裁判字號偵測與引用語意判斷。

### 2.1. 裁判字號偵測（JN Detection）

由於裁判字號（JN）通常是一段具有明確起點與終點的連續文字，為了從完整判決書中找出可能的裁判字號，我們將裁判字號偵測問題化為序列標註任務。標註模型架構如圖一所示 。




圖一、裁判字號標註模型架構。

具體而言，給定一判決書文本 $D$，首先，將其轉換為 token 序列：

$$\mathbf{x} = (x_1, x_2, \ldots, x_N)$$

模型目標為學習一函數 $f(\cdot)$，對每個 token 預測其 BIO 標籤，亦即：

$$f(\mathbf{x}) \to \mathbf{y}, \quad y_i \in \mathcal{Y}$$

其中標籤集合定義為：

$$\mathcal{Y} = \{\text{O}, \text{B-JN}, \text{I-JN}\}$$

分別表示非裁判字號、裁判字號起始 token 與裁判字號內部 token。

為捕捉法律文本中長距依賴與複雜的專業語境，我們採用具有處理長文本法律語料優勢的 Lawformer [1] 作為 tokenizer 與編碼器，將輸入文本序列化後映射至上下文語意空間：

$$\mathbf{H} = \text{Lawformer}(\mathbf{x}) = (h_1, h_2, \ldots, h_N)$$

其中 $h_i \in \mathbb{R}^d$ 表示 token $x_i$ 之上下文語意表示。接著，基於上下文表示，以 token-level classifier 進行標註預測：

$$p(y_i \mid \mathbf{x}) = \text{Softmax}(\mathbf{W}h_i + \mathbf{b})$$

其中 $\mathbf{W}$ 與 $\mathbf{b}$ 分別為分類器之權重矩陣與偏置向量。並以最大後驗機率進行決策：

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} p(y_i = y \mid \mathbf{x})$$

模型訓練採用 cross-entropy loss 作為目標函數：

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k \in \mathcal{Y}} y_{i,k} \log p(y_i = k \mid \mathbf{x})$$

其中 $y_{i,k}$ 為 ground truth label 的 one-hot 表示。模型透過反向傳播（backpropagation）最小化上述損失函數，同時更新 Lawformer 編碼器與分類器參數，使其學習判決書中裁判字號之邊界與語境特徵。

此外，考量法律判決書可能超過模型最大輸入長度（4096 tokens），我們採用 sliding window 機制對長文本進行分段處理：當序列長度超過 4096 個 token 時，將文本切分為多個長度為 4096 token 的區塊，並於相鄰區塊間保留 2048 token 的重疊區域。透過重疊設計，可使跨區段的重要上下文資訊被同時納入相鄰 window 的計算範圍，降低實體邊界因切分而遭破壞的風險。

由於同一 token 可能出現在多個窗口中，我們設計基於 confidence 的整合策略：假設 token $x_i$ 在不同窗口 $w \in \mathcal{W}$ 中之預測分佈為 $p_i^{(w)}$，則最終標籤定義為：

$$\hat{y}_i = \arg\max_{w \in \mathcal{W}} \max_{y \in \mathcal{Y}} p_i^{(w)}(y)$$

此策略透過選擇最高信心預測，以降低窗口邊界造成的不一致性。

最後，在完成 token-level 標註後，將預測結果轉換為實際的裁判字號實體：以標記為 B-JN 的 token 作為實體起點，再向後依序搜尋連續標記為 I-JN 的 token，直到遇到第一個非實體標籤 O 為止，即可重建完整之裁判字號實體。

## 2.2. 引用語意判斷（SS Identification）

判決書中出現裁判字號並不必然代表實質引用，仍需依據上下文判斷其是法律見解上的引用，或只是案件事實與程序中的單純提及。因此，承接第一階段的結果，第二階段根據 JN 附近的語意脈絡，判斷其是否構成實質引用，並辨識對應的引用語意關鍵詞（Semantic Support, SS），如「意旨」、「參照」。

如同第一階段，我們仍選擇使用理解法律語意脈絡具備優勢的Lawformer語言模型，因此，第二階段的引用標註模型架構與第一階段的裁判字號標註模型雷同。惟不同之處在於：（1）輸入資料改為以單一 JN 為中心、保留其周邊語意脈絡的片段（如，前後50個字元）。（2）標籤集合定義為：

$$\mathcal{Y} = \{\text{O}, \text{B-SS}, \text{I-SS}\}$$

然而，依個案情境，判決文書中常有多個字號連續出現之情形，使得一段文字可能同時被用來判斷兩個鄰近的 JN 是否為引用；如此，將造成引用判斷的混亂。為避免此問題，根據就近原則，當兩個 JN 相鄰過近時，我們改以兩個 JN 之間的中間點作為切分邊界，來避免同一段文字同時被歸屬於兩個不同的裁判字號，進而降低 SS 判斷時的語意混淆。

## 2.3. 判決書引用偵測

基於階段一 JN 與階段二 SS 之標註結果，排除未構成引用的 JN 後，即為本專題研究方法偵測得到的「被實質引用的裁判字號」。

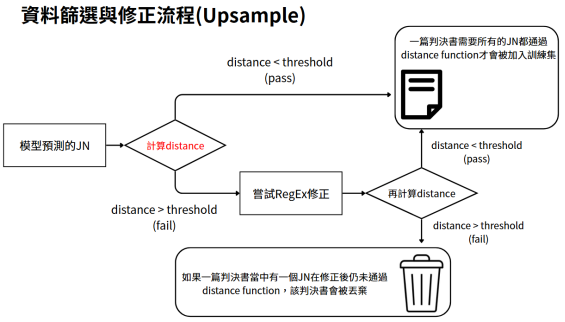
## 3. 研究方法：自我訓練機制設計

在判決書引用偵測的研究方法中，要訓練裁判字號與引用標註模型，在實務上都必然遭遇對於法律專業人力質與量的高度需求。為有效降低對人工標註資源的依賴，本研究進一步探討自我訓練（Self-Training）機制於本任務中的應用可行性。

考量偽標籤之生成與篩選品質將直接影響自我訓練之學習成效，我們從以下四個面向設計品質控制機制，以提升偽標籤可靠性並降低錯誤標註累積之風險：（1）新增資料標註策略（Pseudo-Labeling Strategy），（2）偽標註修正與篩選（Pseudo-Label Refinement and Selection），（3）自我訓練迭代對象（Self-Training Target），（4）標註之最終推論整合。

（1）新增資料標註策略決定由哪個模型產生偽標籤：當獲取未經人工標註之新資料時，可利用歷次迭代過程中所產生之模型集成（model ensemble）進行偽標籤標註，以整合不同階段模型之知識並提升標註品質，我們將之稱為P-all ensemble。

（2）偽標註修正與篩選，稱之為Upsample：為降低錯誤偽標註造成之誤差累積，本研究提出基於 Distance Function 的標註修正與篩選機制。對於每筆偽標註結果，先計算其信心分數並進行必要修正，再以門檻值篩選樣本，僅將高信心偽標註納入擴充訓練集，作為後續自我訓練之資料來源。流程如圖二所示。




圖二、偽標註修正與篩選流程。

為實作上述信心評估機制，考量判決字號之引用撰寫雖具有一定程度非約定成俗之習慣格式，但實務上仍呈現高度多樣性，且其是否構成實質且正確之引用，需同時兼顧「指涉正確性」、「邊界正確性」及「語義正確性」。上述三種判準分別涉及語意層級與表層結構之不同面向，單一評估標準難以同時涵蓋，因此本研究設計 Distance Function 作為偽標註信心分數之計算方式，由語意資訊與文書脈絡特徵兩部分組成。

具體而言，基於訓練集中所有由人工標註之裁判字號、引用語意關鍵詞之片段集合$G = \{g_1, g_2, \ldots, g_N\}$，對於一偽標註樣本 $x'$，其信心值由語意相似度分數 $D_{\mathrm{E}}(G, x')$ 與文書脈絡特徵匹配分數 $D_{\mathrm{R}}(x')$ 共同決定：

$$D_{\mathrm{Total}}(G, x') = \alpha D_{\mathrm{E}}(G, x') + (1 - \alpha) D_{\mathrm{R}}(x')$$

語意相似度分數 $D_{\mathrm{E}}(G, x')$ 衡量偽標註樣本與既有正確標註樣本於語意空間中的相似程度，用以反映指涉語義是否一致：

$$D_E(G, x') = 1 - \frac{\mathbf{e}_{x'} \cdot \mu_G}{\|\mathbf{e}_{x'}\|_2 \|\mu_G\|_2}$$

其中，$\mu_G = \frac{1}{N} \sum_{i=1}^{N} \mathbf{e}_i$ 為 $G$ 之語意中心（semantic centroid），$\mathbf{e}_i = \text{BGE-M3}(s_i) \in \mathbb{R}^d$ 為利用預訓練語意模型 BGE-M3 [2] 將每一個裁判字號、引用語意關鍵詞片段映射至高維語意空間之 embedding 表示。

文書脈絡特徵則衡量偽標註樣本是否符合法律文書中裁判字號之常見引用語境、書寫習慣及結構模式，用以補足邊界與格式層面的約束。此部分參考法學專家之建議，依照裁判字號之書寫形式與語境性引用特徵，設計具體文書脈絡匹配分數 $D_{\mathrm{R}}(x')$。

（3）自我訓練迭代對象決定基於哪個模型進行下一輪訓練：不同於每輪皆從原始預訓練模型重新開始，我們以前一輪訓練所得之模型作為後續訓練的初始模型，稱之為 P-model（Previous model）。此策略使模型參數能持續承接歷次訓練所學習之知識，逐步整合不同迭代階段所使用的資料資訊。隨著自我訓練輪次增加，模型得以在保留既有知識的同時吸收新資料中的潛在模式，形成更完整的資料分布理解。

（4）標註之最終推論整合決定最終如何產生預測結果：在自我訓練（Self-Training）架構中，未標註資料所產生的偽標籤將被納入後續訓練資料，因此偽標籤品質將直接影響後續模型學習結果。由於錯誤標註可能在多輪迭代過程中持續累積並被模型反覆學習，進而造成誤差擴散（error propagation），因此如何提升偽標籤之可靠性與穩定性為自我訓練中的重要課題。

為降低單一模型預測偏差所帶來的影響，本研究於偽標籤產生階段採用模型

集成策略（model ensemble strategy）機制。具體而言，對於同一份未標註判決書，分別利用多個既有標註模型進行裁判字號辨識，各模型皆可輸出每個 token 屬於 BIO 標籤之機率分佈。接著，採用軟投票法（soft voting）整合各模型之預測結果，將同一 token 於各類別上的預測機率進行平均：

$$p_{sv}(k \mid \mathbf{x}, i) = \frac{1}{M} \sum_{m=1}^{M} p_m(y_i = k \mid \mathbf{x})$$

其中，$M$為參與集成之模型數量，$p_m(y_i = k \mid \mathbf{x})$表示第 $m$ 個模型對輸入文本 $\mathbf{x}$ 中第 $i$ 個 token 預測為類別 $k$ 之機率。最終每個 token 的標註結果皆由整合後機率最大的類別決定：

$$\hat{y}_i = \arg\max_{k \in \mathcal{Y}} p_{sv}(k \mid \mathbf{x}, i)$$

透過整合多個模型所學習到的不同決策邊界與語意特徵，Soft Voting 能降低個別模型誤判對偽標籤品質之影響，提升標註結果的一致性與穩健性。

綜合上述，我們將提出具備自我訓練機制之 PUPE（P-all ensemble Upsample P-model Ensemble）標註模型。

## 4. 實驗

### 4.1. 實驗設定

本研究以司法院資料開放平臺公開之2024年裁判書為資料來源，採隨機抽樣方式取得1,068篇裁判書作為研究語料。為建立監督式學習所需之標註資料，由法律專業人士針對裁判書內之判決字號（Judgment Number, JN）及裁判引用關鍵詞片段（Semantic Support, SS）進行人工標註，作為模型訓練與最終評估之依據。

評估指標方面，所有實驗均以精確率（Precision）、召回率（Recall）與 F1 分數（F1-score）作為主要衡量標準，並採用 chunk-level exact match 評估方式，亦即模型預測之 JN 片段必須與人工標註之 ground truth 在起訖位置上完全一致才計為正確，若僅擷取部分字號或額外包含前後文字，即使大致指向同一引用，仍視為錯誤。為降低資料切分所造成之評估偏差，所有實驗皆採用 cross-validation 且套用相同的超參數設定，最終效能以交叉驗證各次結果之平均值表示。

### 4.2. 裁判字號偵測（JN Detection）評估

判決字號偵測之目標為從完整判決書中標記出所有可能之判決字號。為驗證語意理解模型相較於規則式方法之優勢，本研究將階段一模型，於充足之人工標註資料下訓練後，與兩種正規表達式基準進行比較：

- RegEx：依判決字號常見組成結構進行匹配，即（法院名稱）X年（度）Y字第Z號（裁判類型）。
- RegEx-Strict：於 RegEx 的基礎上進一步要求字號必須同時包含法院名稱與裁判類型，以建立高精確率之嚴格基準。

表一、判決字號偵測之效能比較

| 方法 | Precision | Recall | F1-score |
|---|---|---|---|
| RegEx | 0.7505 | 0.807 | 0.7776 |
| RegEx -Strict | 0.9565 | 0.1463 | 0.2535 |
| JN-BIO | 0.8986 | 0.926 | 0.9123 |

表一呈現各方法於判決字號偵測任務之實驗結果。結果顯示，JN-BIO標註模型於三項指標上皆維持穩定表現，其中 F1 score 為所有方法中最佳。相較之下，RegEx-Strict雖因採用較嚴格之比對條件而取得最高精確率，但召回率僅 0.1463，導致 F1 score 大幅下滑。此結果反映規則式方法在判決字號偵測任務中所面臨之限制：當規則設計較為寬鬆（RegEx）時，雖能提升召回率，卻容易引入誤判；而當規則設計較嚴格（RegEx-Strict）時，則可能遺漏大量有效目標，造成召回率明顯下降。因此，規則式方法往往難以同時兼顧兩者平衡。相較之下，語意模型能夠利用上下文語意資訊，正確辨識如「臺灣高等法院暨所屬法院民國 105 年法律座談會民事類提案第 43 號」此類非典型格式之判決字號。

此結果顯示，語意模型相較於傳統規則式方法具有更佳的泛化能力，驗證語意理解方法在判決字號偵測任務上的有效性。

### 4.3. 引用語意判斷（SS Identification）評估

為驗證階段二中判決字號前後語句脈絡對是否構成實質引用的影響，本研究將階段二提出之方法與下列兩者進行比較：

- Keyword：以人工整理之引用關鍵詞（如：意旨參照、理由書揭示、揭明在案、…等共27個關鍵詞）進行規則式判斷。
- Classifier：將判決字號前後所截出之片段視為單一輸入，以 Lawformer 進行二元分類，預測整段是否構成引用。

表二呈現三種方法於引用語意判斷之

效能表現。實驗結果顯示，三種方法皆能有效判斷引用語意，其中 SS-BIO 標註模型表現最佳。對比 Classifier 方法顯示， SS-BIO 透過序列標註直接定位 SS 片段，能較精確掌握引用語句之位置與邊界，進一步提高整體效能。最後 Keyword 方法雖能涵蓋多數引用情形，但其較低的精確率反映出單純依關鍵詞出現與否進行判斷，容易忽略語意脈絡，因而產生較多誤判。整體而言，語意模型之表現優於關鍵詞比對規則，其中標註模型又比分類模型更為優秀。

表二、引用語意判斷之效能比較

| 方法 | Precision | Recall | F1-score |
|---|---|---|---|
| Keyword | 0.8868 | 0.9739 | 0.9282 |
| Classifier | 0.9566 | 0.9718 | 0.9641 |
| SS-BIO | 0.9777 | 0.9787 | 0.9781 |

### 4.4. 判決書引用偵測（JN+SS）評估

為驗證本研究所提出之兩階段方法於原始判決書中標記「被引用判決字號」之有效性，與下列兩種方法進行對比：

- RegEx + Keyword：以正規表達式擷取判決字號後再以關鍵詞判斷是否具實質引用之兩階段規則式方法。
- 1-Phase BIO：以單一語言標註模型自完整判決書中直接標註出具實質引用之判決字號之序列標注方法。

表三呈現所有方法於判決字號引用偵測任務上的效能表現。其中本研究所提出之兩階段標註（JN&SS-BIO）方法整體表現最佳，不僅優於一次性標註（1-Phase BIO）方法，更大幅領先RegEx+Keyword。此結果顯示將任務拆解為「先辨識判決字號、再判斷其是否構成實質引用」之策略的有效性。此外RegEx+Keyword雖具備規則簡單、可解釋性高之優點，但其較低的召回率顯示依賴字串格式與關鍵詞比對容易受格式變化與引用語句表達多樣性之影響；相較之下，語意模型能學習上下文資訊與語意關係，於三項指標上皆有明顯提升。

表三、引用語意判斷之效能比較

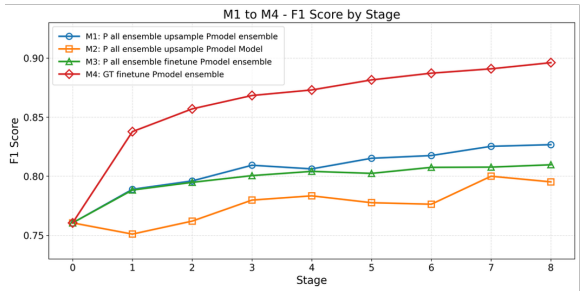
| 方法 | Precision | Recall | F1-score |
|---|---|---|---|
| RegEx+ Keyword | 0.755 | 0.58 | 0.655 |
| 1-Phase BIO | 0.8726 | 0.9029 | 0.8874 |
| JN&SS-BIO | 0.9256 | 0.9193 | 0.9224 |

### 4.5. PUPE 自我訓練機制評估

有鑒於法律文書的處理一直面臨巨大的專業人力缺口，為驗證 PUPE 自我訓練機制於資料不足情境下的有效性，我們以裁判字號偵測（JN Detection）具體實踐評估。

為此，我們將訓練資料切分為九份並進行九個階段之迭代訓練，以模擬、評估資料不足時進行自我訓練之情境；並參考法學專家之建議，根據裁判字號高度結構化「（法院名稱）X 年（度）Y 字（第） Z 號（裁判類型）」的特性，我們依照其格式的重要性，設計三項懲罰因子：「年、字、號」必要結構缺失、「法院名稱、度、第、裁判類型」次要結構缺失、「X、Y、Z」長度異常，分別以 $P_{\text{ess}}(x')$、$P_{\text{ct}}(x')$、$P_{\text{len}}(x')$ 表示。

$$D_R(x') = 0.7P_{\text{ess}}(x') + 0.2P_{\text{ct}}(x') + 0.1P_{\text{len}}(x')$$




圖三、PUPE之重要消融性比較。

圖三呈現PUPE之重要消融性比較結果：M1為PUPE，M2不做標註之最終推論整合（僅以最新模型進行標注），M3不做偽標註修正與篩選（直接將偽標註為 JN 之資料全數納入擴充訓練集，進行後續自我訓練），M4為分階段使用專業人力標註資料並進行後續模型迭代訓練之方法，屬於理想上界。

根據實驗結果：（1）M1 於各階段表現皆優於未在推論階段使用 ensemble 機制的 M2，且各階段表現震盪較不明顯；M3 比之 M2 也呈現相同趨勢；這顯示整合多階段模型之ensemble 機制能有效抹平單一模型於迭代過程中之不穩定波動，對整體效能與穩定性皆有正向貢獻。（2）對比 M1 與 M3，M3 因保留全部新增的偽標籤資料未進行篩選而引入較多雜訊，使其於後續迭代階段成長趨緩，顯示信心值篩選與修正流程能有效過濾雜訊樣本，維持效能穩定成長。（3）理想上界 M4 與 M1 之差距反映了偽標籤雜訊相較於完全正確標註所付出之代價，然而 M1 能在不依賴額外專業人力標註之前提下逐步逼近此上界，顯示本研究所提出之自我迭代流程於實務上

具有可行性。

綜合以上實驗結果，語意模型於判決字號偵測與引用語意判斷兩項子任務上均明顯優於規則式方法，而將二者整合之分段式標注策略亦優於一次性標註；同時，PUPE 自我學習機制能在有限人工標註之前提下，透過 ensemble 與信心值篩選兼顧效能之穩定性與成長性，驗證了「語意模型結合分階段任務設計」對於判決字號引用偵測任務之有效性。

## 5. 結論

本專題針對法院判決書中裁判字號引用關係自動辨識問題，提出結合語意模型與自我學習機制之兩階段架構，分別處理裁判字號辨識（JN）與實質引用判斷（SS）任務。實驗結果顯示，語意模型於兩項子任務上均明顯優於規則式方法，而分階段標註策略亦優於一次性標註方式，驗證了任務拆解設計的有效性。

此外，PUPE 自我學習機制透過 ensemble 與信心值篩選，有效降低偽標籤雜訊對模型的影響，使模型在有限人工標註資料下仍能穩定提升效能。雖然偽標籤品質與完全正確標註仍存在差距，但實驗結果顯示所提出方法能逐步逼近理想上界，兼顧效能與標註成本。

綜合而言，本研究證明「語意模型結合分階段任務設計與自我學習機制」能有效提升判決字號引用偵測效能，並在有限標註資源下達成約九成以上準確度，具備應用於法律資訊檢索與判決引用網絡建構之潛力。

## 6. 參考文獻


1. Xiao, Chaojun, et al. "Lawformer: A pre-trained language model for chinese legal long documents." AI Open 2 (2021): 79-84.
2. Chen, Jianlv, et al. "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation." arXiv preprint arXiv:2402.03216 4.5 (2024).