目錄

一、標註問題定義說明

二、標註規定與費用

三、標註工具介紹與範例

附錄：10篇論文摘要範例之標籤答案

結尾：聯絡方式

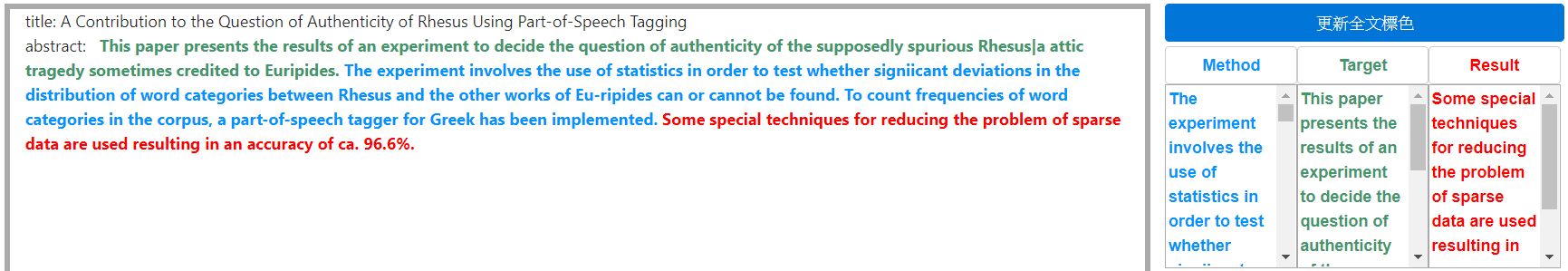
**一、標註問題定義說明**

本計畫預計以人工方式獲得論文中摘要段落的標籤標註。其中，我們預計以人工方式請專人標註出一篇摘要中，是否包含以下三種段落之定義：

1. Method(包含：演算法、模型、框架等)
2. Target(包含：欲解決之問題、欲達成的目標等)
3. Result(包含：論文最終解決Target問題的質化或量化成效、論文的創新發現或發明等)

我們的資料集採取arXiv.org網站上所有包含cs標籤以及其下各子標籤的論文，預期能將廣泛的電腦科學類的論文摘要做標註。

範例標註：



更多高達10篇的範例摘要將條列在附錄。

**二、標註規定與費用**

**標註規則：**

1. 每次以100篇論文摘要為單位，亦即有100個.txt檔案。
2. 每篇.txt檔案標註完成後的命名規則：[原名稱]\_marked.txt
   1. 範例: 輸入檔案：cs.AI\_0.txt ，輸出檔案：cs.AI\_0\_marked.txt
3. 各標籤包含的單字序列不能有部分重疊。
4. 同一標籤之內容於一篇摘要內只能恰好包含一個不限長度之連續單字序列，不可以斷續的方式標註。 EX: “A winning maker strategy for the (a, b) game can be built from wining strategies for games involving fewer marks for the maker and the breaker.” 恰可被標籤為Target.

**計費方式：**

1. 每篇摘要報酬之實際計算方式請參考下方規則。
2. 每人每批(100篇論文摘要)之任務如符合以下評量標準，原則上可獲得新台幣一千六百元之報酬
3. 本實驗室保留調整計酬之計費方式。

**標籤質量的衡量規範：**

由於缺乏正確答案，我們將會採取以下的政策來檢驗以及評估標籤的良好程度。

1. 每人每批的100篇論文將以不公布編號的方式，被區分成:
   1. 10篇包含在本實驗室具有正確答案之要完成的摘要集
   2. 90篇該標籤者取得之要完成的摘要集
2. 檢驗100篇摘要中的部分若干篇之標註結果與本實驗室事先準備的正確答案標註中的作比對驗證
3. 標註摘要的比對與衡量方式：
   1. 檢查其中10篇有正確答案之論文摘要，如有超過四篇(含)以上與正確標籤差異過大者，會請退回並重新替100篇標註。
   2. 與正確答案比對檢查時(如有正確答案)，容許標籤有前後的字數offset誤差，只要大致涵蓋的字句相同即可。
   3. 所有摘要將以一篇單價16元新台幣作為最終計酬。
4. 所有數字與比例分配將於實際進行後視實際情況保留調整空間。後期如退回重標的次數過多者，經研議者，可能不再給予任務分配。

**三、標註工具介紹與範例**

**資料夾結構**

./marker/

於marker資料夾內：

./static/                               # require files of index.html

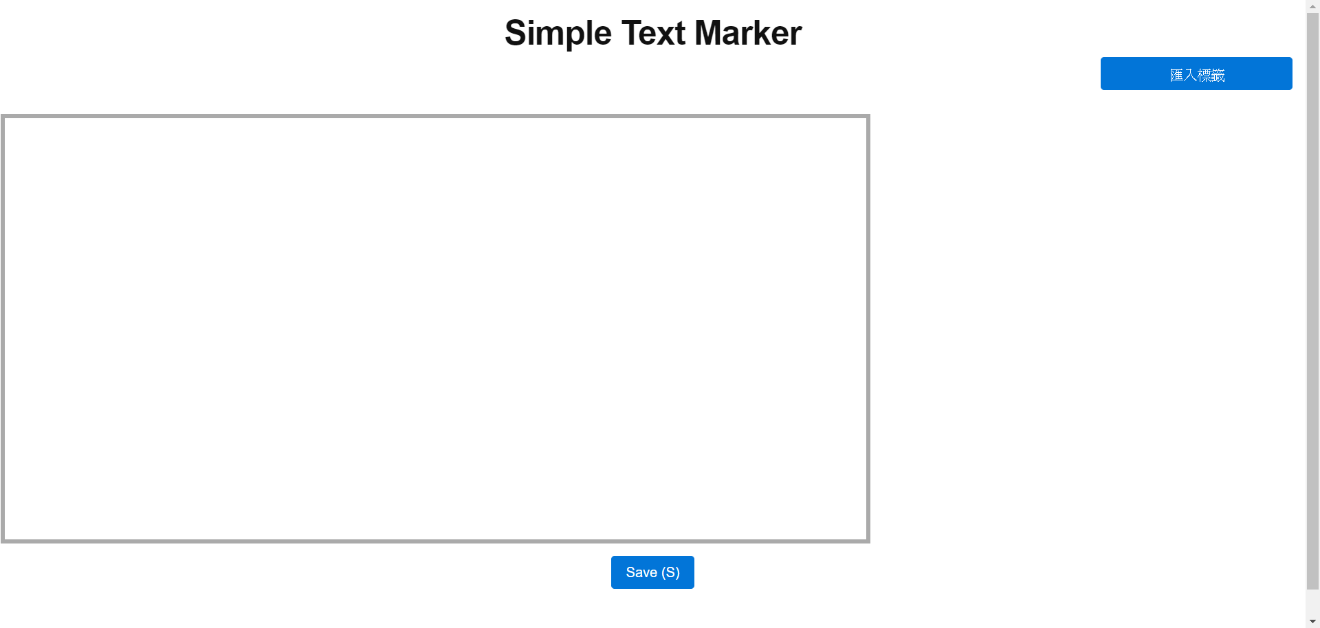
./abstract\_labels.txt           # self-defined labels list

./index.html                        # for labeling .html file

操作步驟

**步驟一、開啟程式**

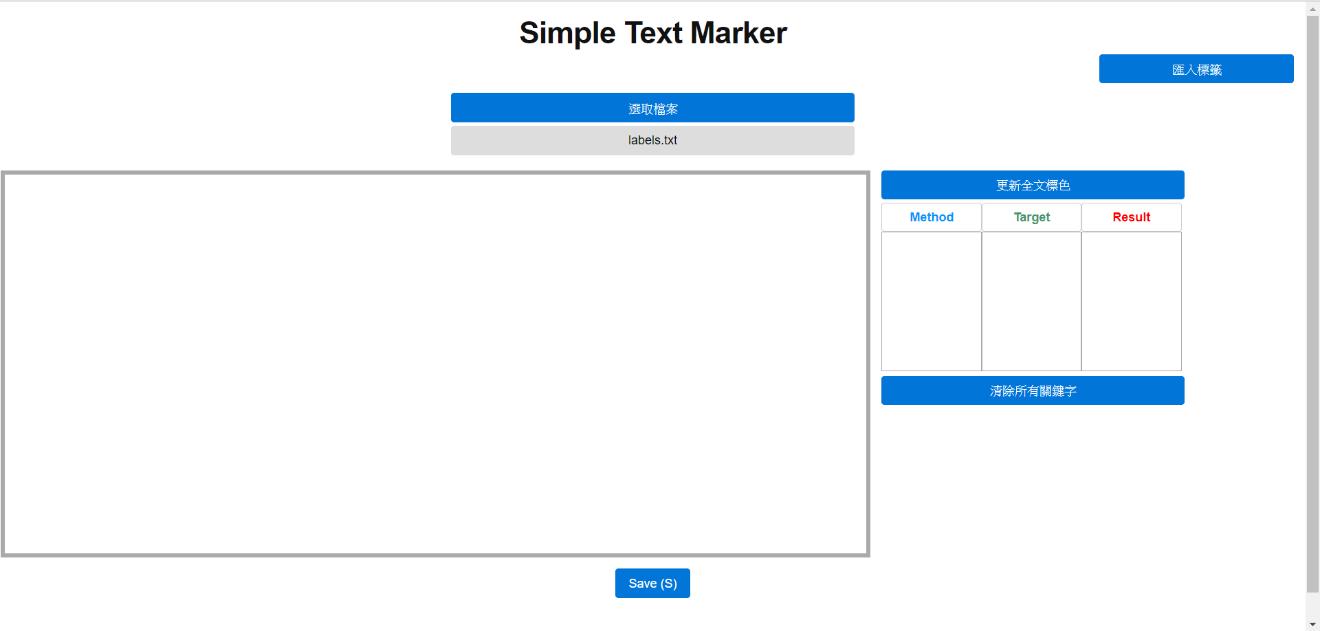
在當地資料夾使用瀏覽器打開index.html即可開啟標註程式(以網頁呈現)，如下圖所示:

**步驟二、匯入標籤**

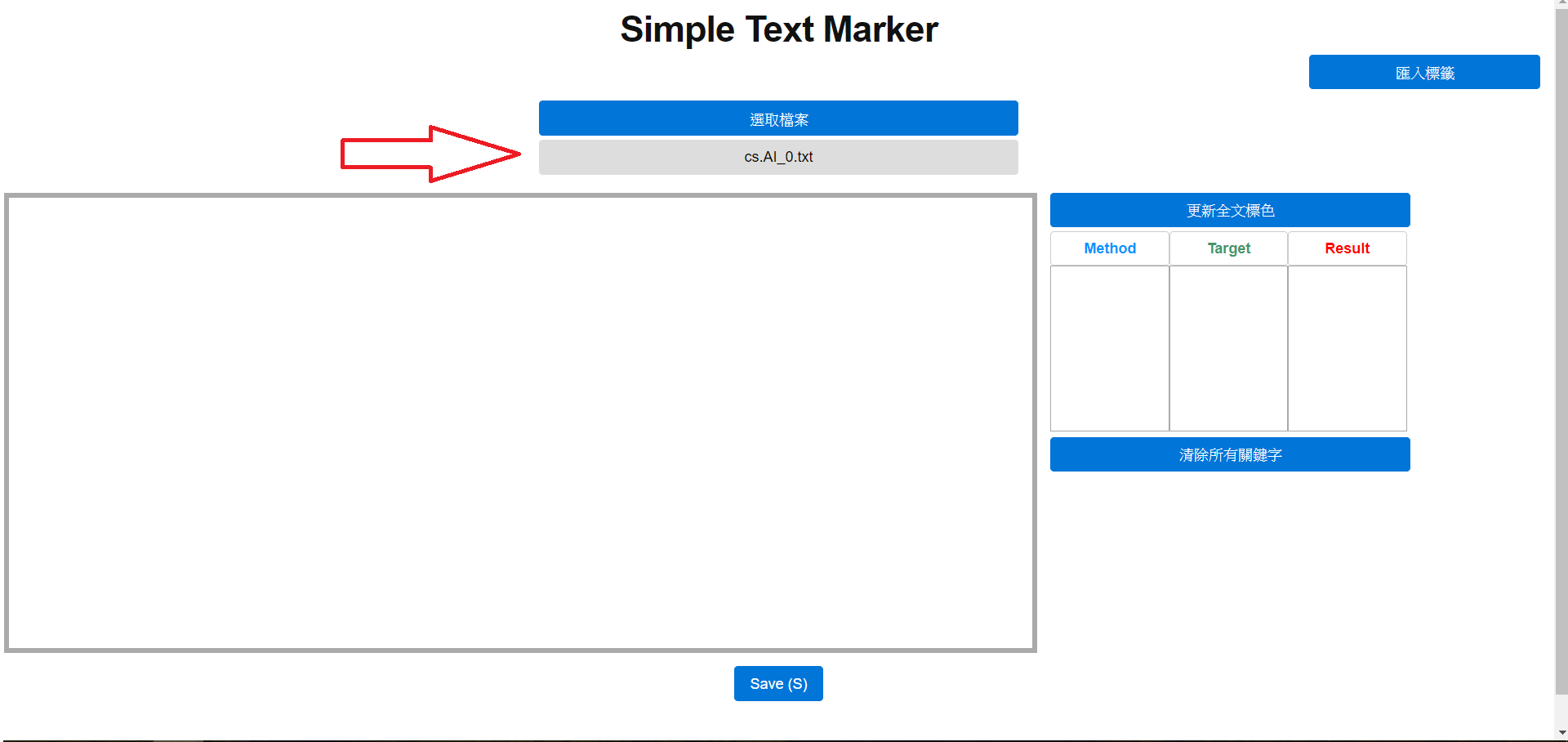
按下「匯入標籤」的按鈕，並選取marker資料夾內的abstract\_labels.txt 檔案。該檔案為符合本實驗需要而定義之標籤。然為統一需要，勿自行更改標籤名稱與顏色編碼。





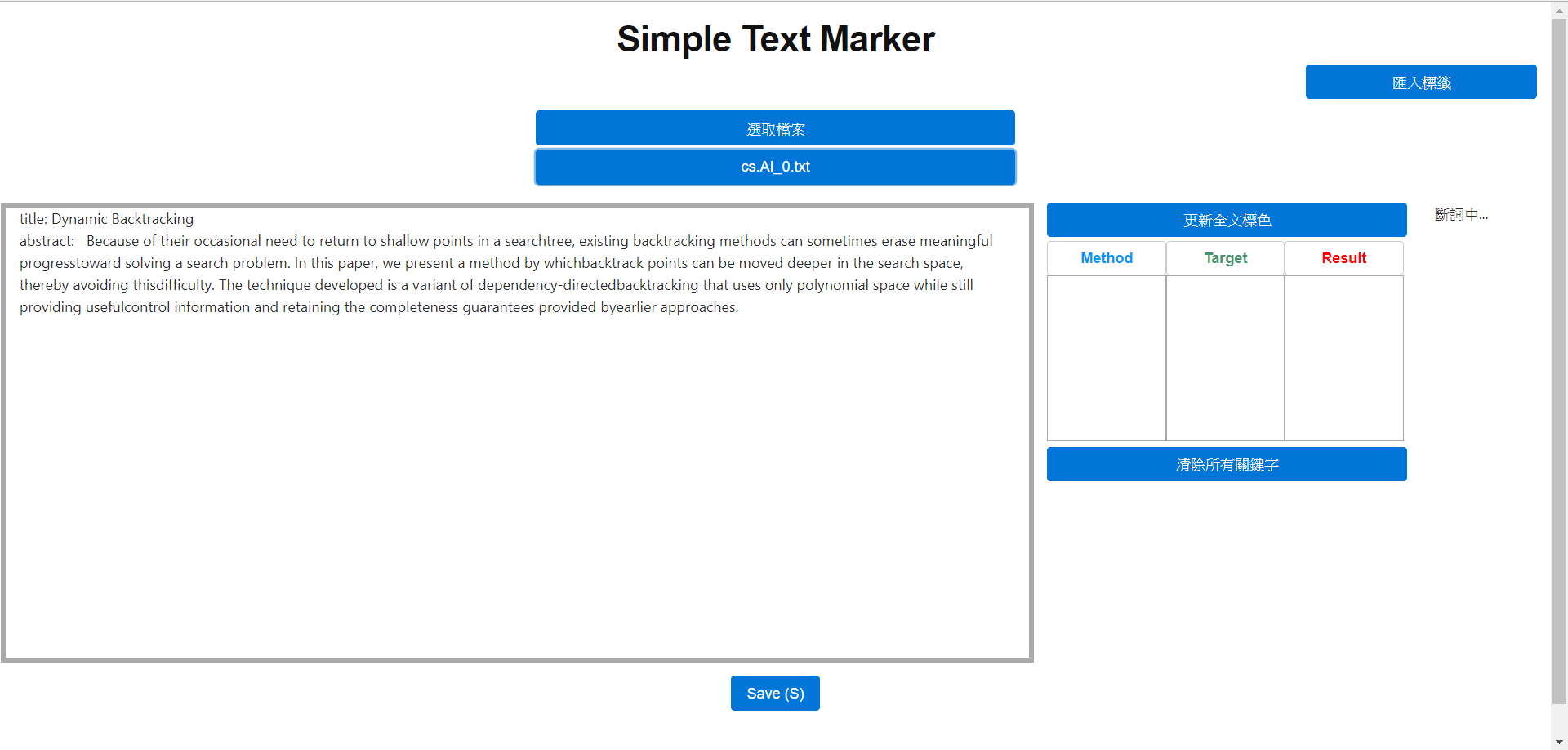
**步驟三、匯入單一未標註文檔(.txt file)**

選擇網頁中「選取檔案」按鈕，選擇目標.txt檔案匯入，完成後如下圖所示。



**步驟四、選取該文檔**

匯入指定文檔之後，請點選網頁中該文檔名稱之按鈕，如下圖所是即為「cs.AI\_0.txt」。選取成功後會出現如下圖所示之畫面。其中論文摘要則出現在網頁左方空白格中。

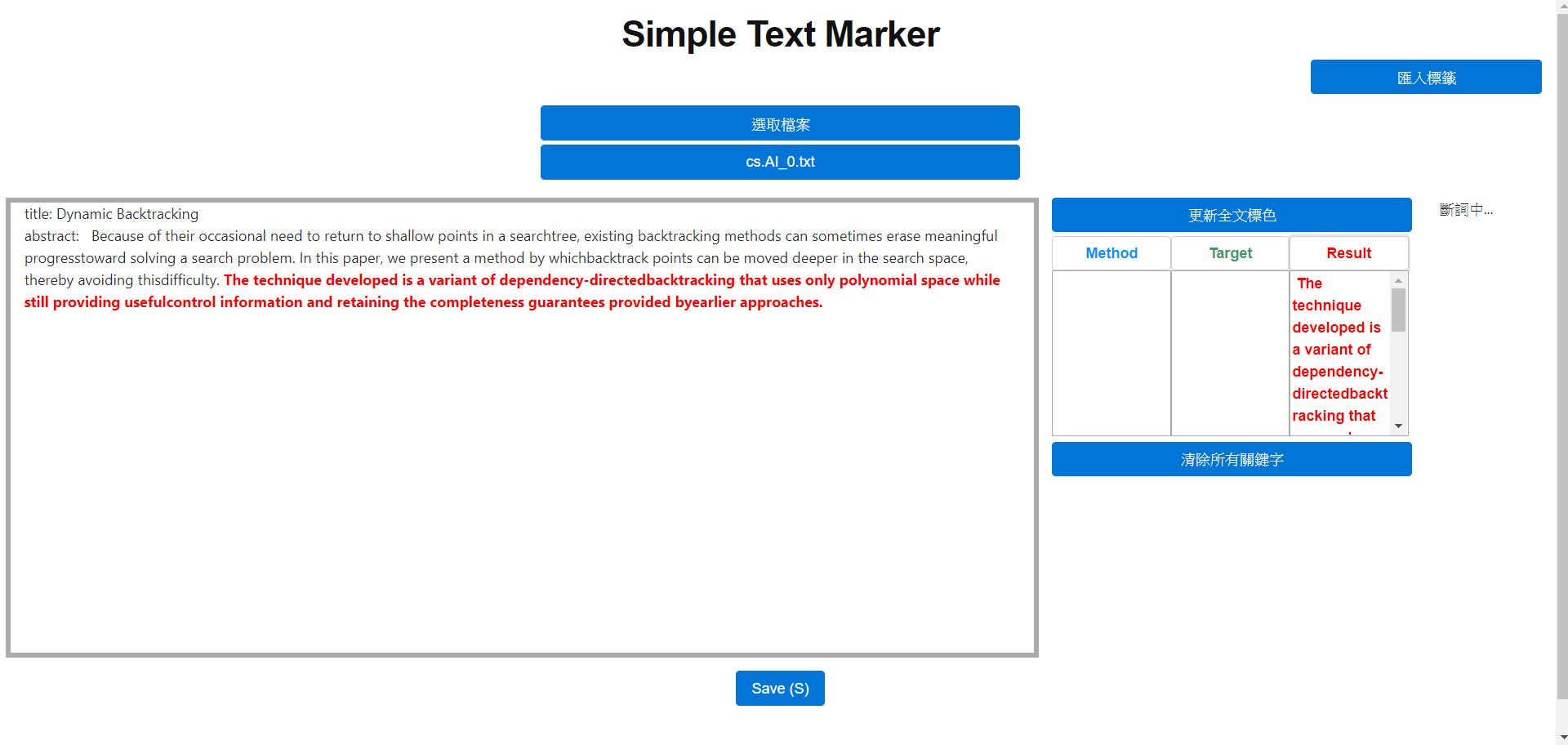


**步驟五、開始標註**

個別標籤的標註方式為：

1) 反白指定段落 

2) 點選標籤，如按下紅色之「Result」按鈕後，該段落即以紅色呈現，並在標籤下方欄位出現對應原文。



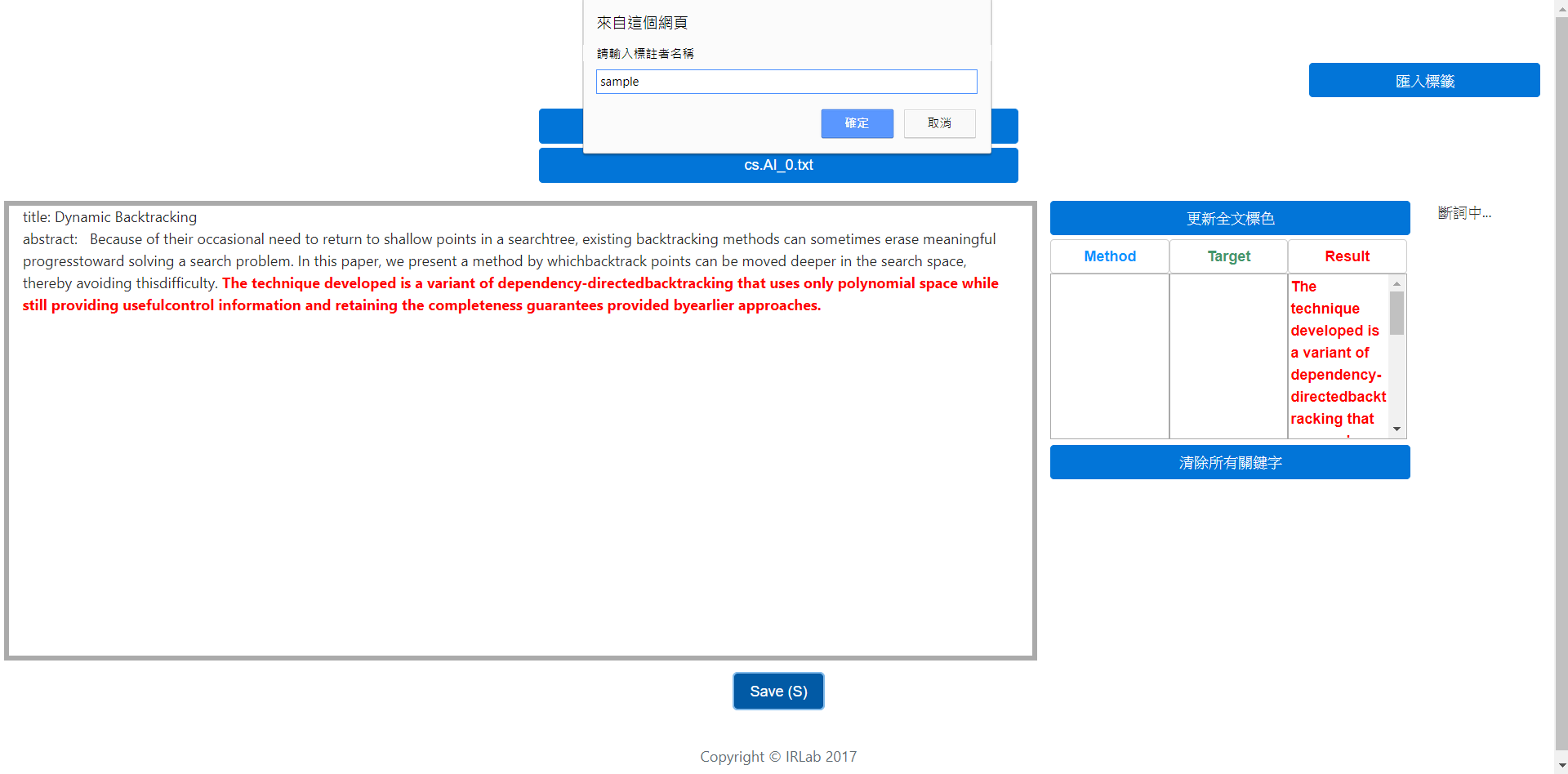
※註：如選取之段落在整篇摘要中有其他重複，則會同時將各重複段落一起選起，不須再個別選起。然而此為極少之情況，在此僅做說明。如發生，則為我們想要的結果。我們要的是重複段落都需標註。

其他操作說明：

如需清除暫時標註之結果，請按下網頁中「清除所有關鍵字」按鈕。如標註至一半欲暫存以便下次繼續標註，請先儲存檔案，如下一步驟說明。往後可載入標註半成品檔案繼續標註。

**步驟六、儲存檔案**

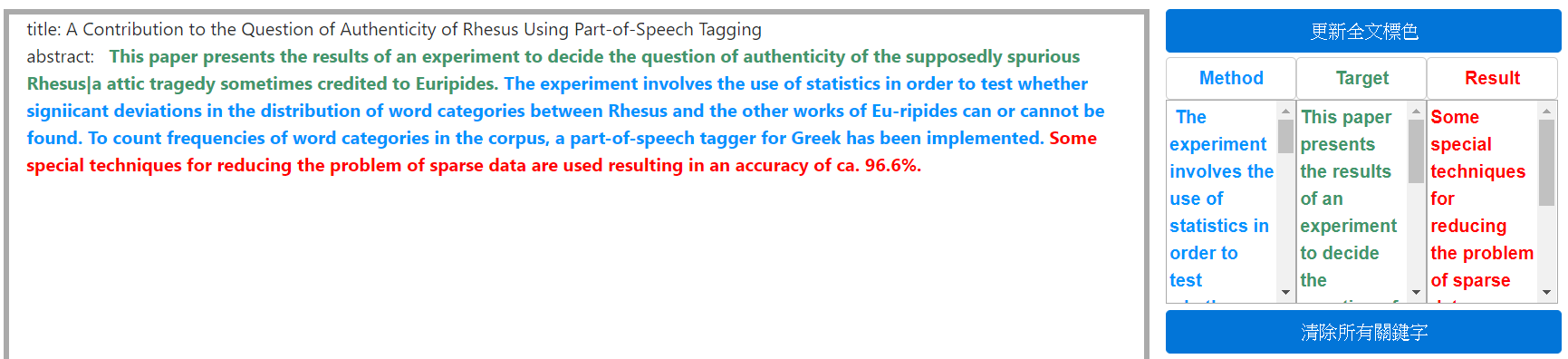
欲儲存檔案，請按下網頁中下方「save(s)」按鈕，會出現一彈出式視窗。請在該視窗中鍵入欲命名之檔案名稱。請注意，命名規則請遵守本計畫統一規定。下圖僅為示範。另外，下一步驟說明完成的一批摘要任務如何繳交。

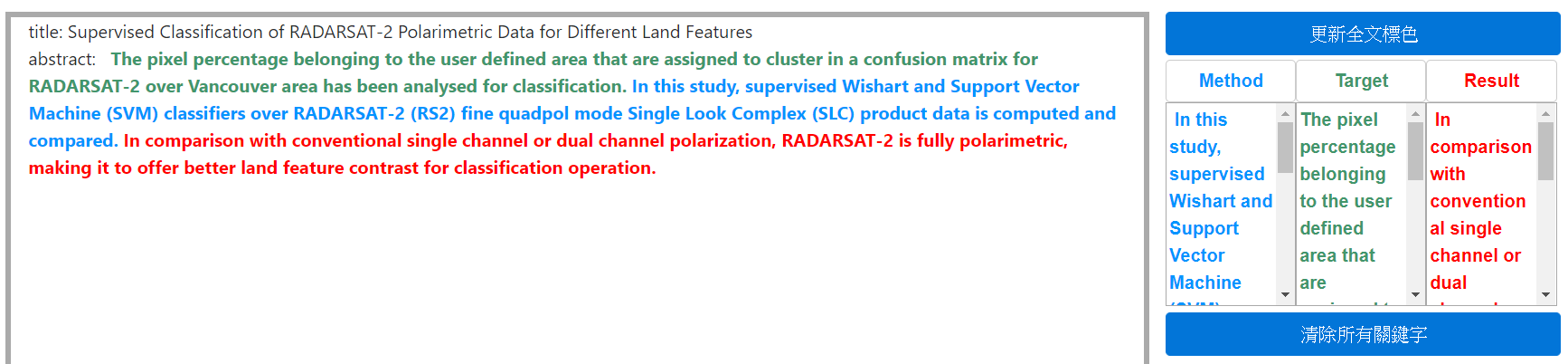


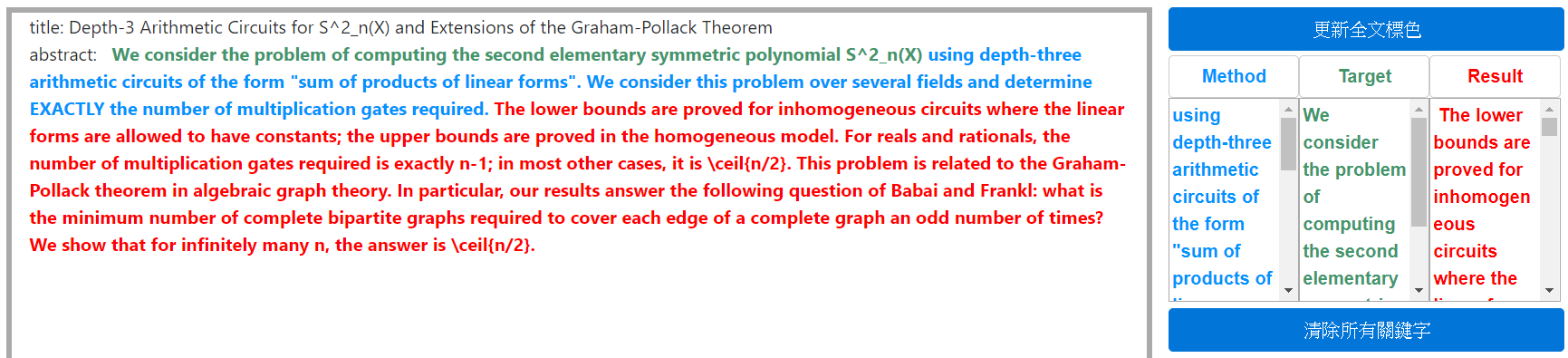
**步驟七、繳交任務**

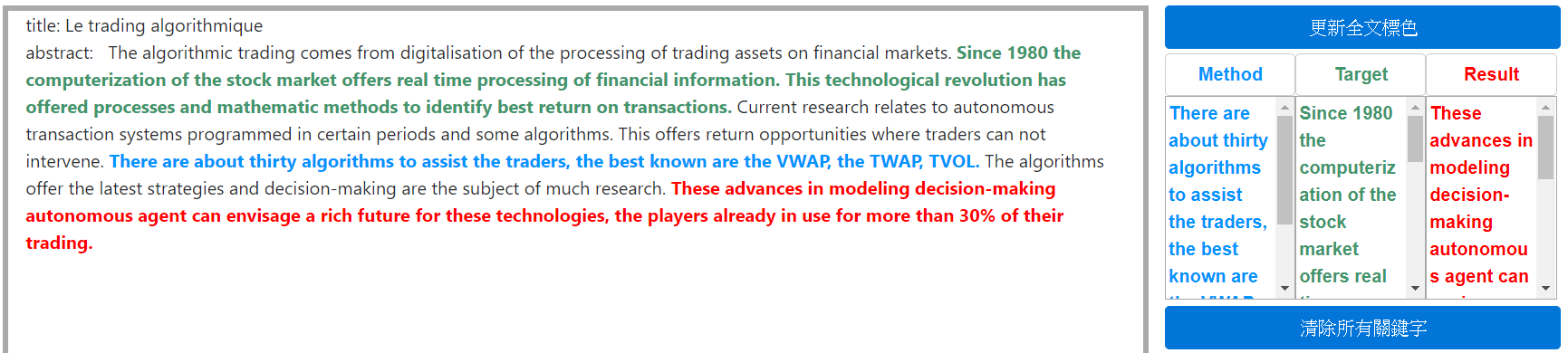
當完成一批100篇論文摘要的標註任務後，我們要求標註人員將該批.txt檔案上傳至指定雲端空間，使用dropbox帳號來完成。您需要一個dropbox帳號，並請來信聯絡取得上傳網址。每人每批的100篇完成之.txt檔案須放在同一資料夾內，並以個人姓名或帳號代稱命名以利辨別。實驗室團隊這邊將會執行程式檢驗您所上傳的檔案並自動計算分數後告知結果。

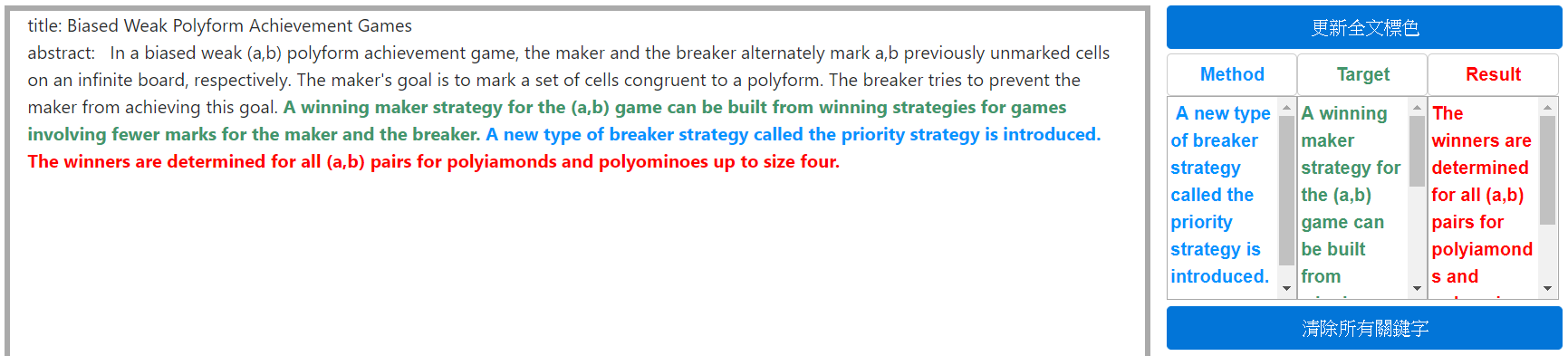
附錄:

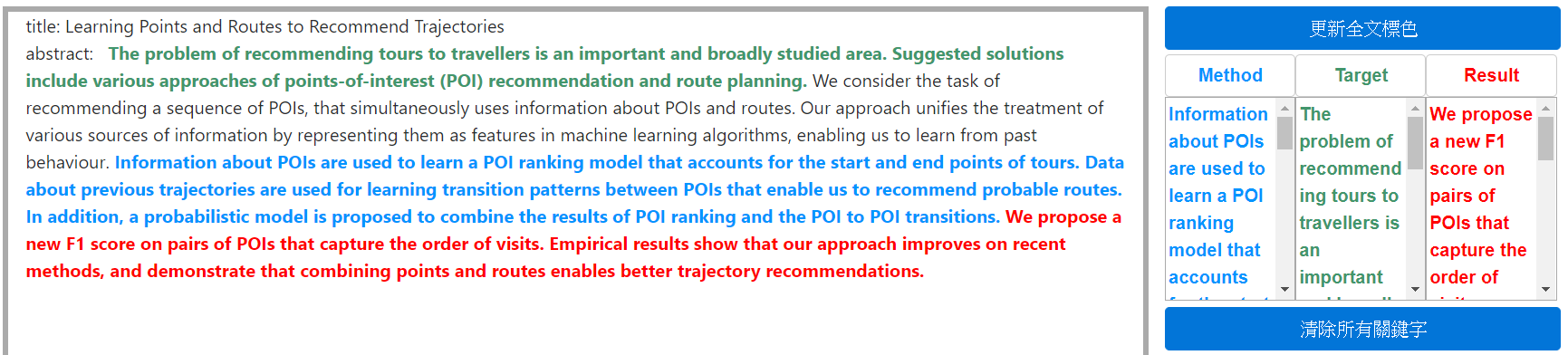


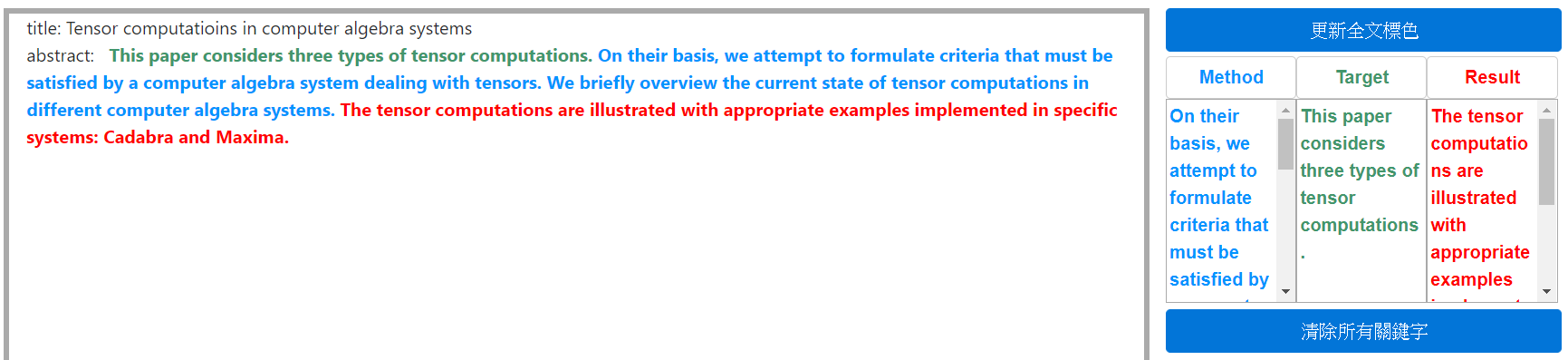


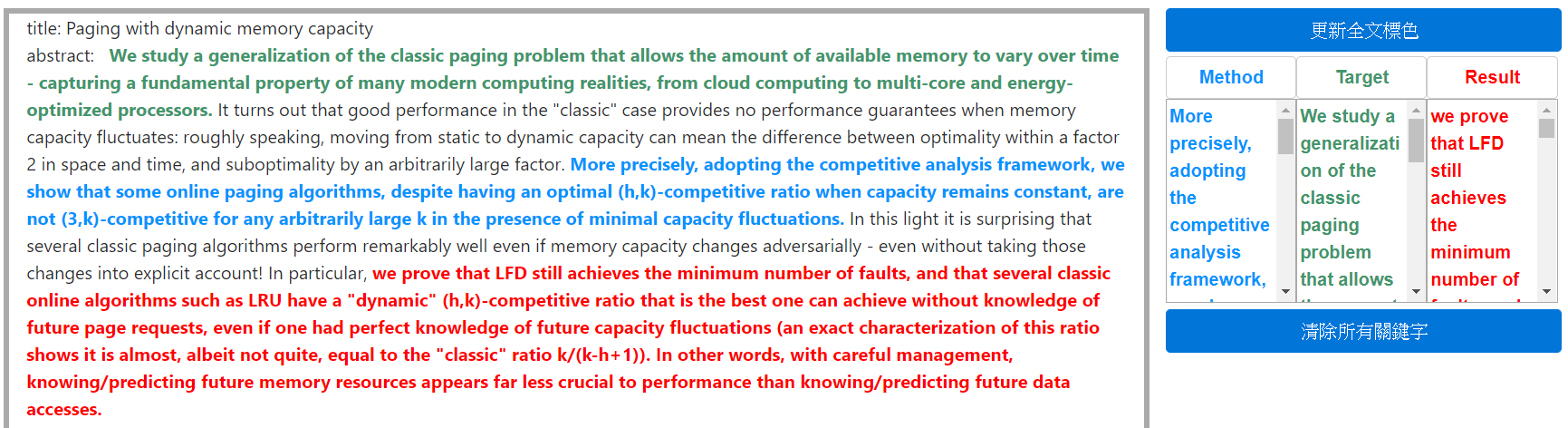


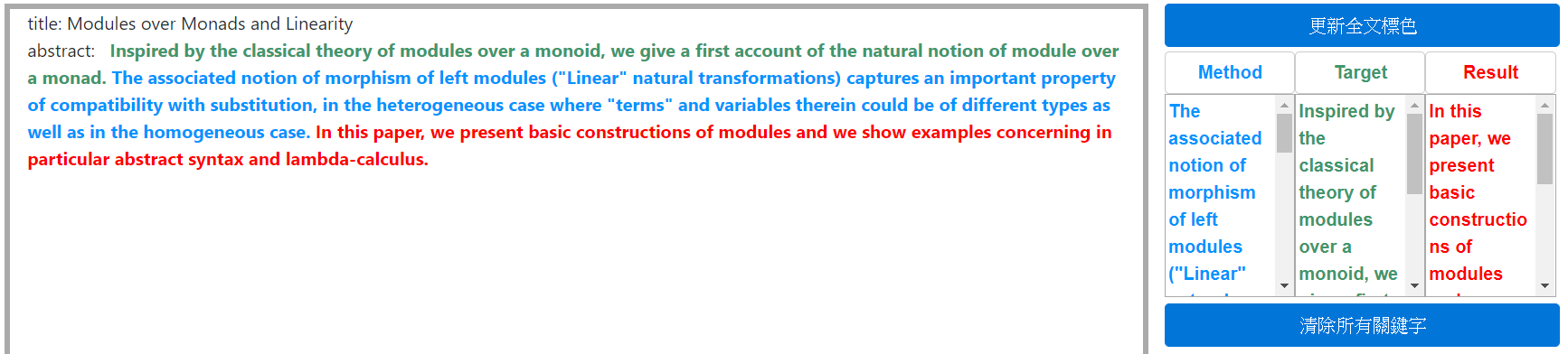


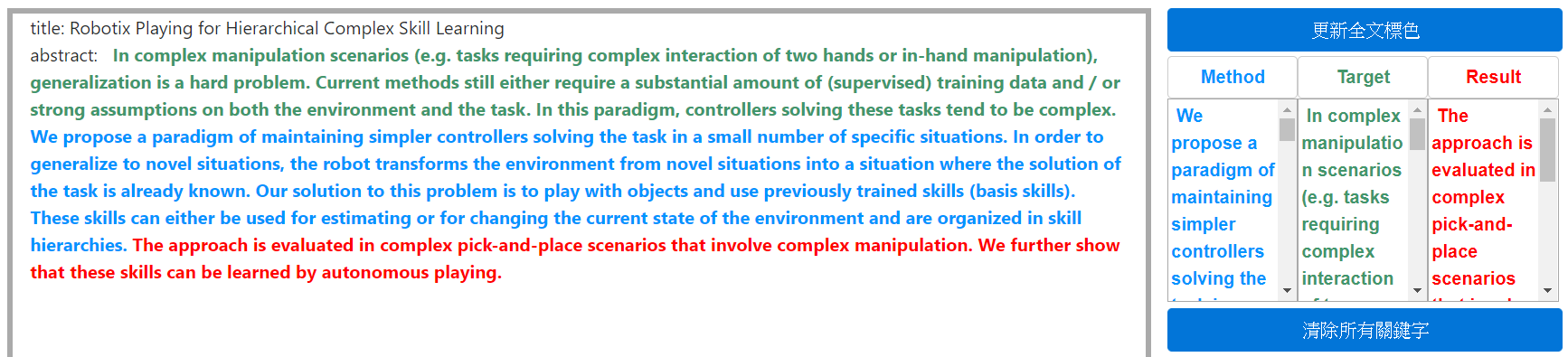












如有任何問題，敬請來信聯絡：

蔡仕竑 [r05921030@ntu.edu.tw](mailto:r05921030@ntu.edu.tw)

許冠傑 [jk78346@gmail.com](mailto:jk78346@gmail.com)

台灣大學資訊工程學系 林守德教授 機器發明與社群網路探勘實驗室