

## NLP Lab 第 13 組

### Evaluate Student Writing by using longformers

謝沐昕111065551

李昀浩107062373

吳晉廷111062664

#### Abstract:

在台灣，英文作文是學測中必考的科目。然而根據研究統計，偏鄉學生的英文作文能力遠遠不及都市學生，其中原因不乏為無法獲得充分的教學資源，英文作文沒有人可以批改，並提出建議。

而市面上目前有許多自動寫作反饋工具，但都有局限性，絕大多都是Grammar check或是Spelling check。但更多人的問題是無法擬定議論文中的寫作結構，例如議論陳述和起承轉合，抑或是沒有足夠多的論點支持觀點。更重要的是，由於成本高，教育工作者無法使用其中許多書寫工具。對於教育資源落後的學校來說，這些工具形同虛設，簡而言之，自動寫作反饋領域的創新時機已經成熟，可以幫助偏鄉學童的英文作文能力。

此專案運用本學期教學的元素(Transformer, NER, Spacy, Streamlit)實作了:自動對 6 至 12 年級學生撰寫的論文中的文本進行分段，並對議論文類型進行分類。我們將實作出來的結果製作成使用者可自行上傳資料並分類的網頁，希望藉此降低操作者的上手難度。

#### Introduction:

我們使用Georgia University, Learning Agency Lab 所提供的15600筆美國 6 至 12 年級學生撰寫的議論文，區間為2010-2020年。這些文章經由多名專業語言學家標記每個段落的類型、開始與結束及單詞。每篇字數不等(從100字至4500字)、主題不拘。

因為資料屬於NLP文字分類範疇，我們使用Transformer來訓練模型，在method中，可以清楚地看到比較兩種不同Transformer的結果。資料前處理的部分，因為同一個段落文字長，參考了NER、Spacy，我們使用BIOES-style方法，將一段句子同個class拆分成Begin、Inside 這兩種label，方便模型讀取，最後再將預測結果的Begin、Inside合併成同一個Label。

而模型輸出的結果是一連串由discourse\_num、Label組成的數據。為了方便使用者查看結果，我們將其視覺化。我們將這些輸出的資料在原本的文檔根據7種不同的類別塗上不同的顏色，可以一眼就在文檔上看出預測的類別。最後，再將所有程式運用streamlit部署於網路上，使用者只需要在網頁表單裡輸入文章，即可得到一份精美的

預測成果。

## Method:

我們會input一篇文章，一篇文章裡會有很多個段落，我們稱之為discourse，如同中文的起承轉合一般。

我們參考NER BIO Tag 作為Label的Method，BIO Tag分別代表(Beginning Inside Outside)，例如:United(B\_Country) States(I\_Country) of(I\_Country) America(E\_Country)。由於我們為了簡化標籤數，我們只考慮Beginning以及Inside。所以我們總共標籤數為 $7 \times 2 + 1(\text{other})$ 共15個標籤，我們把將label後的文章丟進Model後predict，最後在視覺化前合併為原先的7個label。

我們使用了Longformer的tokenizer裡的offset\_mapping欄位，當我們input文章的時候，會對每個token給予(start,end)的一個tuple。例如: better choices，and 我們分別給予 (70,76),(77,84),(84,85),(86,89) [以字母計算]

如此一來我們可以把資料集裡每個discourse的start end與offset\_mapping做比對後，為我們的token標註上我們想要的label。最後，丟進Model裡預測，模型會預測每個字最有可能出現在哪個class，達到分類效果，並使用spacy工具視覺化它。

由於可能會有預測錯誤，如少數兩個字或三個字被當作一個discourse，我們做法就是把少於6個字的discourse捨棄。

## Result:

The question of whether in this age of intensive media coverage, it is no longer possible for a society to regard any living man or woman as a hero is one that has been grappled with many people and one that deserves serious consideration. However, the argument is based on generalization. The statement simply assumes that hero must be impeccable, but in reality, there are many ways to define a person as a hero. Moreover, it is presumptuous to suggest that media only plays a role of disparaging a person. Therefore, even though the number of flawless hero decreases, there is no guarantee that it is impossible for a living man or woman as a hero.

<span style="font-size: 0.8em; font-weight: bold; line-height: 1; border-radius: 0

Admittedly, the speaker's claim might appear to have considerable merit. After all, hero that doesn't make any mistakes decreases **Counterclaim**. Yet, the speaker fails to provide a clear litmus test for measuring which kind of hero the argument is discussing. **Rebuttal** When we speak of "hero", the following come to mind: hero that protects the world, hero that has great success in a certain field, and hero that is the person you admire most. Obviously, the argument is discussing the ideal hero that doesn't make any kind of mistake. Thus, in order to determine the prerequisites of regarding a person as a hero, it is necessary to examine the different sorts of heroes.

<span style="font-size: 0.8em; font-weight: bold; line-height: 1; border-radius: 0

It is thus justifiable to take hero that is faultless and hero that has great success in a particular field into account, and we will find the statement problematic. **Claim** For example, many people regard Michael Jordan, the greatest player in the NBA of all time, as a hero and even want him to play for extra season during the 911 even though he retired for several years. People all know that he has a bad habit of gambling, but it doesn't influence his place in NBA history. Therefore, the complicated big picture, composed of hero in different definitions, need to be seen.

<span style="font-size: 0.8em; font-weight: bold; line-height: 1; border-radius: 0

Moreover, the speaker unfairly assumes that by intensive media converging, his or her fame will be deprecated **Counterclaim**. However, the role of media is to play a objective role and describe what happens in a neutral situation. **Rebuttal** For example, many newspaper editors have a strong orientation of partisan and always blame the other one even though they do nothing wrong. In short, the media should play a role in delivering the message not giving the tendency.

<span style="font-size: 0.8em; font-weight: bold; line-height: 1; border-radius: 0

In summary, this claim is not as persuasive as it stands **Concluding Statement**. Although the number of some hero declines, not all hero is disappearing. If the media can't play a neutral role, no matter how perfect the person is, it is no longer possible to regard a living man or woman as a hero.

## Discussion & Conclusion

比較bert模型與longformer模型，最顯著的差距是tokenizer 的 max\_length。如圖一所述，bert的token超過512之後就無法繼續使用了，若是分割一篇文章至兩個部分分別使用bert預測，因為文章有其先後性、結構性，因此不會有任何效果。相對的，與longformer至多可以tokenize 4096個token，對議輪文的範疇相當足夠。再者，觀察bert與longformer的準確率，可以觀察到longformer比bert高約5%。

我們針對bert與longformer進行比較。

### 傳統 BERT 長文本輸入處理方式

1. Truncation methods(截斷法)：直觀的取其中的 512 token 符輸入，也是已知最常用的暴力解，但在這個資料集不適用
2. Pooling 法：把長文本分成多個 512 字以內的 segments (段落)，對每個段落都轉換為 BERT Embedding 再對 [CLS] 進行 max-pooling，拆分文本的方式會考慮不到長距離字符之間的上下文關聯，因為他們被拆開到不同 segment 沒辦法進行 attention
3. 壓縮法：按照句子分成多個 segments，接著用規則或單獨訓練一個模型的方式來排除一些權重較低的 segments 和 Pooling 法一樣有無法考慮長距離關聯的問題
4. BERT 之所以難以處理長文本，是因為傳統 Transformer-based 模型採用「全連接」的 Attention 機制，每一 token 都要和另一 token 組合，attention 的時間&空間複雜度高達  $O(n^2)$ 。

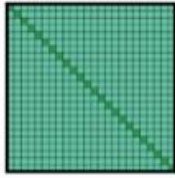
### Longformer 算法介紹

Longformer 改進 Transformer 傳統 attention 機制：對每一 token 「只對固定窗口大小附近的 token」計算 local attention，再結合下游任務計算少量 global attention，將原始 Transformer 的複雜度降至  $O(n)$

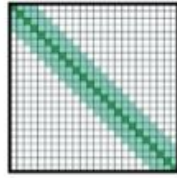
滑窗機制 (Sliding window attention)：對每一 token 只對附近  $w$  個 token 計算 attention， $w$  大小隨任務不同調整，複雜度  $O(w*n)$

膨脹滑窗機制 (Dilated sliding window)：考量更全面的上下文資訊。在滑動窗格中，被 attention 的 2 token 之間存在大小為  $d$  的空隙，可展開比普通滑窗機制更廣的 attention 範圍、又不增加計算負荷。論文中實驗證明這種方法表現得比第一種好。

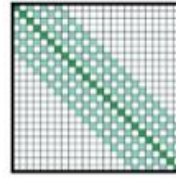
融合全局訊息的滑窗機制 (Global + sliding window)：根據下游任務的不同添加少量 global attention（例如分類任務的 [CLS] 處會有一個 global attention）。也就是部分 token 會被添加 global attention



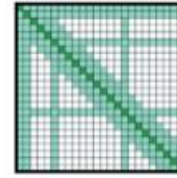
(a) Full  $n^2$  attention



(b) Sliding window attention



(c) Dilated sliding window



(d) Global+sliding window