AI CUP 2022 FALL RNER

Team: TEAM_2218

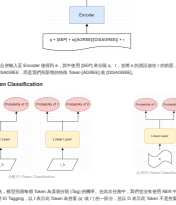
GitHub: https://github.com/Shmuhammad/AI-CUP-2022-Fall-NLP

模型架構

在這次的比賽中，我們做一個 **Token Classification** 的任務，目標是從給定的文章中，抓出特定的資訊。我們使用的資料集為 **DeBERTaV3**，其預訓練的 **SQuAD2.0** 資料集上，我們將其作為 Encoder 從後接輸入文章的 Contextualized Embedding，之後使用不同的架構處理這些 token embedding，最終輸出 token 的類別。以下為 Encoder 後面的幾種架構：

而我們的輸入了多種類型資訊，包含了 token 的文字、token 的 NER 類別、token 的 POS 類別、token 的 dep 4 種

輸入格式 - 分類式



其中，x 表示 token 的 embedding 及 Encoder 後的特徵，其中使用 [SHP] 來分隔 x，x，並將 x 的開頭放在 x 的開頭，而這樣 x 並非原 token 的 NORMS 或 DISAGNOSE，而是我們預定義過的特徵 Token [NORMS] 或 [DISAGNOSE]。

輸出格式 - Token Classification

分別從 NER 的任務，模型預測每個 Token 為是否屬於類型 (Tag) 的機率，在此次比賽中，我們定義每個類別 NER-可能的類型的 BIO Tagging，用意是將 O Tag，(這 1 表示沒 Token 為類型 or 這 token 的 1 的一部分，型以 O 表示沒 1 不是類型的一部分 B) 分開的原因，是因為 B 後面的 token 大多會是 Linear Layer 之後的 1 表示 Start token 之後的 1 的機率，其中會不太可能會是 Linear Layer 是先預測到 B 的原因。

在這邊的資訊來源，則是將 x 輸入 Linear Layer 來獲得 1 到 O 的機率

輸出格式 - Span Prediction

分別從 Extractive QA 的方式，模型預測每一個 Token 為是否開始的開始和結束的機率

在這邊的資訊來源，則是將 x 輸入 Linear Layer 分別獲得 x 開 / 的結束和開始機率，而最中間 Linear Layer 是一個共用模組的

在這邊的資訊來源，則是將 x 輸入 Linear Layer 一次得到 x 開 / 的結束和開始機率

預測流程

1. 前處理

- 分句斷詞，分別將 x 和 y 做 tokenize 並轉換為數值以及 Span
- 合併成 [開始 x, x, x,