Schema Matching

資工三 楊佳峻

1. Introduction

分組資料並train Bert模型，使用模型時相似度高的較有可能為相同項，並以此做matching，不存在不進行配對的策略

1. Related work

常用方法: 根據維基百科

Pre-integration – 對當前模板進行分析，決定整合策略，策略可以是全部或是部分、先後順序、及模式的選擇

Comparison of the Schemas – 跟去兩模板的內容去判斷屬性，可能會發現衝突點，也會有配對成功之處

Conforming the Schemas – 試圖解決衝突點，努力使兩者適應

Merging and Restructuring – 依據上面方法，這時候應該可找到一些方法使兩模板合併，有時還會進行重組

1. Method
   * Data Pre-processing

將助教配對好的歸為一組，總共就會有70組，這些資料將會丟入Bert模型做表層weight的training。

* + Matching Algorithm

Step1: 求得相似度

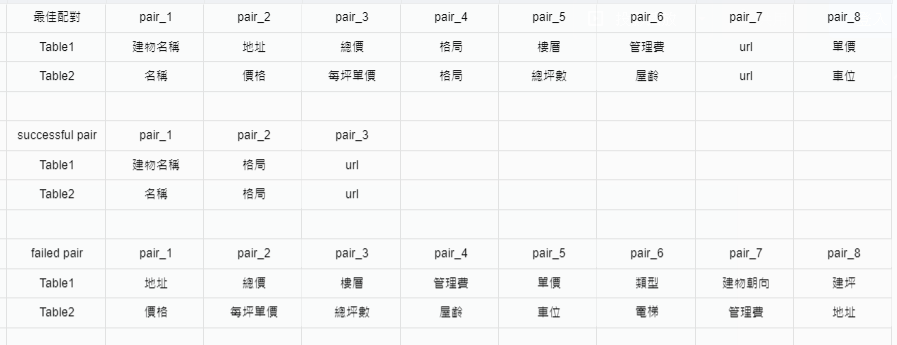
Table1.csv與Table2.csv每個column都取前200項(更好是用random取200項)，內容帶入train好的Bert模型，每個column會得到一組向量，最後兩向量做內積即為兩column相似的程度。

Step2: 進行配對

將得出來的分數以column和row2對應做成表格，從最大值由上而下、由左而右，依序查找，且查過的column及row不重複，必配對出min(len(column), len(row))個數量，因次不存在不進行配對這種策略。

1. Result

以pair\_1為例，效果頗低且分數接近





1. Conclusion
   * Problem encountered & Solving

Problem 1: 使用word2vec但詞語庫不夠及缺乏彈性字詞

本來想進行word2vec的tag grouping，讓相似的tag能夠做為同一類進行train，不過實際分類仍分成70類，也就是完全沒有效果，於是便不使用。

Problem 2: Bert得出來的分數極低，且差異不大，並無法準確判斷

如題，有試圖提高岔具，但想不出有什麼方法，所以效益仍是不大，無法對分數判別起大作用‧

* + In conclusion

Schema Matching需要花不少時間在tagging與研究版型上，我認為使用ML利用關鍵字學習類型，可以更快速達到這個效果，不過還是需要做人工教調等的後處理。目前最不可能達成的原因可能會出在訓練時間及語言資料庫上，我自己這部分是沒有做的很好。

1. Reference

My colab: <https://www.kaggle.com/code/staler2019/wimu-hw2b/settings>

My github: <https://github.com/Staler2019/NCU-Projects/tree/master/Web-Intelligence-And-Message-Understanding/Schema%20matching>