智能記帳分類模型

一、前言與目的

Chat&Track希望建立一個模型，透過檢視一個至數個名詞以分類日常消費的用品。**但文字是被發明的符號，並不像圖像、聲波具備物理上的意義**。傳統電腦學習字義的辦法為記住每一個字，但此法無法學習字與字間的關係。因此現行多採用Word Embedding[1]的概念，**透過分析語法及文字相鄰的頻率來教導電腦學會多維度的字義**。

Word Embedding除了具備賦予詞義的功能外，它的訓練資料並不需要標記，僅需大量文本資料即可訓練，**節省標記訓練資料的時間成本**。

二、初期想法與實作

我們以Word2vec[2]訓練維基百科釋出的開源文件[3]，在訓練過後得到所有詞的數值矩陣A。每個類別**抽樣**上百個相關名詞[4]，各個抽樣名詞與目標名詞B基於A**計算相似度數值**，最後**加總平均**得到B被分到此類別的權重分數。

重複上述步驟後可得到不同類別的權重分數，**取其最大值**C；若C有**高於人工訂定的分數門檻**則輸出其相對應的類別，若無則輸出其他/無法分類。

三、實際困難

困難依處理順序、範圍，分為以下四種：切詞、抽樣名詞、分數門檻、分類衝突。

1. 切詞：**中文無法如同英文用空白切詞**，在複合名詞、人名、從未得知的專有名詞的切詞判斷，是相當困難的一個議題。

2. 抽樣(sampling)名詞

(1) 判斷各類別**最具鑑別力**的名詞。如：飯、麵之於飲食類別。

(2) 抽樣名詞**分佈是否符合現實**：若飲食類別的抽樣名詞都是跟"麵"有關的食物，那麼輸入"飯"時則有可能就會因為相關性被稀釋(加總平均)而使得模型權重分數數值不高。

3. 分數門檻：若分數門檻太高，則模型一遇到稍不確定的便放棄判斷，若太低則分類精準度將下降。

4. 分類衝突：使用者輸入複合詞時，**如何賦予不同切詞分出分類權重**也是個問題，如：輸入"台大小木屋鬆餅"，可能會切成"台大/小/木屋/鬆餅"，分別輸入分類模型後得到"育/無法分類/住/飲食"，因類別並非一致，因此如何決定要選一種類別便是個問題[5]。

四、衡量與改善

1. 初步衡量

---

[1] 每個文字就像是被嵌上矩陣(matrix)般，賦予其自身一個向量(vector)數值

[2] Google 在 2013 年實作 Word Embedding 的開源工具，有很好的效率及表現

[3] 資料網址：https://dumps.wikimedia.org/zhwiki/

[4] 為什麼不直接對食、衣…等名詞進行相似度衡量？因為這些詞彙在 word vector 中僅是一個名詞，不具備"範圍"的概念。

[5] 根據詞性、詞出現先後，決定其所占的權重，如：阿明麵店，或許可給予後面的名詞"麵店"較大的權重

透過人工搜尋及程式標記的方式蒐集近4000筆關於飲食、交通的測試資料[6]，衡量指標：(1) accuracy[7] (2) precision[8]

我們在兩種分類的測試資料上有著近62% accuracy及90% precision的表現。

2. 改善

我們發現了以下數個問題，並依此進行改善：

(1) 分類詞中出現**特殊符號及無意義詞彙**：Normalization (去除特殊字元及StopWord[9])。

(2) word vector太過**發散**[10]：蒐集了約24萬篇與各個記帳類別關係密切的新聞[11](食記、旅遊..等)加入訓練資料。

(3) 類別**抽樣有主觀**問題：依Kappa statistic[12]判斷抽樣是否客觀。

(4) **分數門檻數值難以取得平衡**：改為KNN投票制度[13]。

更動後accuracy上升至約82%、precision維持不變。

五、結論與未來方向

現行透過抽樣來取代分類邊界的切分在表現上仍有很大的改善空間。但**若要再進步勢必建立更完整大量、且分佈貼合現實分佈的抽樣名詞列表**，但根據Zipf's law[14]，上述改善方法將付出昂貴成本。

我們認為現行**透過人工索引的抽樣無法精確定義記帳類別間的分類邊界**，且詞彙有著**多維的構面，無法歸類單一類別**。

因此我們之後計畫在Word Matrix(經過Word Embedding處理後的矩陣)後疊加類神經網路研究分類間的分類邊界。

References

1. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality
2. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space
3. Omer Levy, Yoav Goldberg, and Israel RamatGan. 2014. Linguistic regularities in sparse and explicit word representations

---

[6] 此測試資料的多元性及數量都不足以代表模型在現實中真實表現，僅能作為改良的參考指標之一

[7] 在所有可能的分類中，正確將名詞分類至類別的比例

[8] 當模型將名詞分類至某類別時，其分類判斷正確無誤的比例

[9] 在文本中無鑑別力的詞彙，在此模型中通常為無意義的介系詞：之、的…等

[10] 像是"電動"一詞，若無大量關於交通(電動車)、娛樂(電動玩具)，則在分類上會出現誤差，甚至無法分類

[11] 資料來源為台大租訂的 WiseSearch 資料庫

[12] 假設決議個體為獨立，基於其提出相同/不同意見機率，計算而得的可靠度[-1 - 1]

[13] 當切出的名詞分別符合類別抽樣的詞彙後即直接增加 1 票，最後比各個類別的票數，若一致則輸出無法分類

[14] 表示詞的稀疏性，指在分佈當中常用的字非常少，而會有許多僅出現 1 次、不出現的字