HW3 ZhuYin Decoding

R08922042 鄭光宇

實作 Bigram

觀察一個注音文組成的序列，我們有 Bigram 的 Language Model，可以使用 Viterbi 算法算出機率最高的漢字序列。我們可以使用如下遞迴式並回溯出最佳的漢字序列。

當 t > 1 時，

$$\delta_t(q_i) = \max_{qj} P(q_i \mid q_j)\delta_{t-1}(q_j)$$

當 t=1 時(初始條件)，

$$\delta_1(q_i) = P(q_i)$$

實作 Trigram

同上，但使用 Trigram 的 Language Model。

當 t > 2 時，

$$\delta_t(q_i, q_j) = \max_{q_k} P(q_i \mid q_j, q_k)\delta_{t-1}(q_j, q_k)$$

否則，初始化 δ，

$$\delta_1(q_i, q_j) = P(q_i)$$
$$\delta_2(q_i, q_j) = \max_{qk} P(q_i \mid q_j)\delta_1(q_j, q_k)$$

實作上可以展開 δ_1 放進 δ_2，因為 q_k 在 δ_1 裡沒有意義，

$$\delta_2(q_i, q_j) = P(q_i \mid q_j)P(q_j)$$

回溯時只要求到 t = 2 即可，因為 δ_2 包含 t=1 和 t=2 的結果。

另外，因為 Trigram 外部至少有四層迴圈 (t, i, j, k)，Language Model 內部又有一層迴圈，複雜度非常高，沒有辦法在一分鐘內算完。為了減少複雜度，在儲存 DP 表時，使用的是一個 hash table，並且借鑒 beam search 的想法，每個時間點 (t) 最多只考慮前 200 個候選路徑，降低 (j, k) 迴圈的迭代次數，實際測試起來和正常版本只有些微差異。另外，讀資料時也是一次讀入一個 batch，然後分配給多個 thread 平行計算。雖然有加速，但是第 10 筆測資仍然需要大於一分鐘的時間，也許還有其他加速的方式。

觀察 mydisambig 與 SRILM disambig 的差異

以下都是 trigram 的結果:

mydisambig: <s> 外 來 客 不 見 得 懂 </s>

disambig: <s> 未 來 客 不 見 得 懂 </s>

mydisambig: <s> 但 卻 肯 定 讓 世 會 的 智 慧 財 產 權 保 護 更 多 </s>

disambig: <s> 但 卻 肯 定 讓 世 會 的 智 慧 財 產 權 保 護 更 大 </s>

在額外的 Wiki 語料上建立 Language Model (trigram) 並測試

語料來源:

https://dumps.wikimedia.org/zhwiki/20201201/zhwiki-20201201-pages-articles-multistream1.xml-p1p187712.bz2

共得到 5627427 組 trigram。

輸入:

ㄧ ㄐ 是 ㄕ 學 證 明 ㄓ 很 重 要 且 基 本 的 一 部 份 ㄕ 學 家 ㄒ 望 ㄊ ㄇ 的 定 ㄌ 以 ㄒ 統 化 的 推 理 ㄧ 著 公 理 ㄅ 推 論 下 ㄑ 這 是 ㄨ 了 避 免 依 著 不 可 ㄎ 的 直 觀 ㄦ 推 出 ㄘ ㄨ 的 定 理

mydisambig:

也 就 是 數 學 證 明 中 很 重 要 且 基 本 的 一 部 份 史 學 家 希 望 他 們 的 定 理 以 系 統 化 的 推 理 有 著 公 理 被 推 論 下 去 這 是 為 了 避 免 依 著 不 可 靠 的 直 觀 而 推 出 錯 誤 的 定 理

disambig:

也 就 是 數 學 證 明 中 很 重 要 且 基 本 的 一 部 份 史 學 家 希 望 他 們 的 定 理 以 系 統 化 的 推 理 有 著 公 理 被 推 論 下 去 這 是 為 了 避 免 依 著 不 可 靠 的 直 觀 而 推 出 錯 誤 的 定 理

Ground truth:

嚴 謹 是 數 學 證 明 中 很 重 要 且 基 本 的 一 部 份 數 學 家 希 望 他 們 的 定 理 以 系 統 化 的 推 理 依 著 公 理 被 推 論 下 去 這 是 為 了 避 免 依 著 不 可 靠 的 直 觀 而 推 出 錯 誤 的 定 理

觀察

Viterbi + trigram 似乎沒有抓到整個語句的前後關係，例如：「○○是數學證明中…」，這裡應該是一個名詞但並沒有被抓出來。還有這段文字描述的主題是數學，但 Viterbi 卻選到了「史學家」，似乎也沒有抓到整篇文章的主題。也許可以嘗試看看較新的語言模型像是 BERT，這種用更長的語句做填空題來訓練的語言模型，更可以抓到文章的前後關係吧。