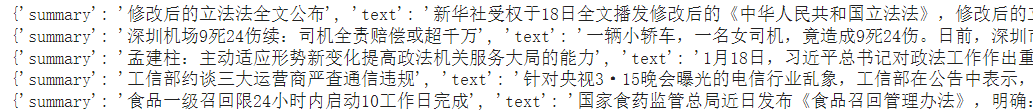
Dataset

huggingface已經把train, validation的dataset弄好了，我只要整理一下就好，我把dataset整理成和英文文本一樣的dataframe(有text和summary)，我取的資料是前2000筆資料。T5和gpt2都是用同樣的資料。



Model

T5 模型的tokenizer沒辦法處理中文，所以我換成其他tokenizer，但是結果還是沒有很好。雖然t5 tokenizer的分數看起來最高，但是實際上他根本看不懂，把輸出印出來看就是一堆Unknown。

這是t5 模型用各種tokenizer的evaluation

|  |  |
| --- | --- |
| [t5 tokenizer] | [bert\_fast tokenizer] |
| Rouge-L-P': 0.23259874230344826,  'Rouge-L-R': 0.31306666666666666,  'Rouge-L-F': 0.31306666666666666,  'Rouge-2-P': 0.0010158730158730158,  'Rouge-2-R': 0.0026666666666666666,  'Rouge-2-F': 0.0026666666666666666 | Rouge-L-P': 0.020760127685004695,  'Rouge-L-R': 0.044308225295973246,  'Rouge-L-F': 0.044308225295973246,  'Rouge-2-P': 0.003104483688737341,  'Rouge-2-R': 0.0068518832392900075,  'Rouge-2-F': 0.0068518832392900075} |
| [bert tokenizer] | [Roberta tokenizer] |
| Rouge-L-P': 0.023500622810601363,  'Rouge-L-R': 0.04557549238840649,  'Rouge-L-F': 0.04557549238840649,  'Rouge-2-P': 0.003721591667025889,  'Rouge-2-R': 0.007925213132299378,  'Rouge-2-F': 0.007925213132299378} | 'Rouge-L-P': 0.005650396093632591,  'Rouge-L-R': 0.009910578700485482,  'Rouge-L-F': 0.009910578700485482,  'Rouge-2-P': 0.00028707135027924215,  'Rouge-2-R': 0.0005839727195225917,  'Rouge-2-F': 0.0005839727195225917 |

T5預訓練的資料有蠻多種的， 有t5-small, t5-base, t5-large, t5-3b, t5-11b，但後面兩個太大了跑不動，所以我就用前面3個跑跑看

|  |  |
| --- | --- |
| T5-small | T5-large |
| Rouge-L-P': 0.08999413719466974,  'Rouge-L-R': 0.18579441575105168,  'Rouge-L-F': 0.18579441575105168,  'Rouge-2-P': 0.04394062043989692,  'Rouge-2-R': 0.09282632838589598,  'Rouge-2-F': 0.09282632838589598 | Rouge-L-P': 0.09628490606639933,  'Rouge-L-R': 0.16936987135680664,  'Rouge-L-F': 0.16936987135680664,  'Rouge-2-P': 0.06994522110493928,  'Rouge-2-R': 0.13168205444935036,  'Rouge-2-F': 0.13168205444935036 |

Train

Gpt2模型我就直接用gpt2自己的tokenizer，可是在訓練時，input batch size和target batch size都會不合，所以我在target 後面補零當作padding，讓他們大小符合可以計算loss，我還有加attention mask，讓模型在訓練時不要過於注意填充的部分。

這是gpt2的分數

Average ROUGE-2 score: 0.018863725936945745

T5和GPT2

T5和gpt2在架構上的差別是t5有encoder和decoder，但gpt2只有decoder。T5的概念是文本到文本的轉移任務，輸入是文本，輸出也是文本；

gpt2的輸出會由前面的一部分輸出來決定，這種機制叫做自回歸。