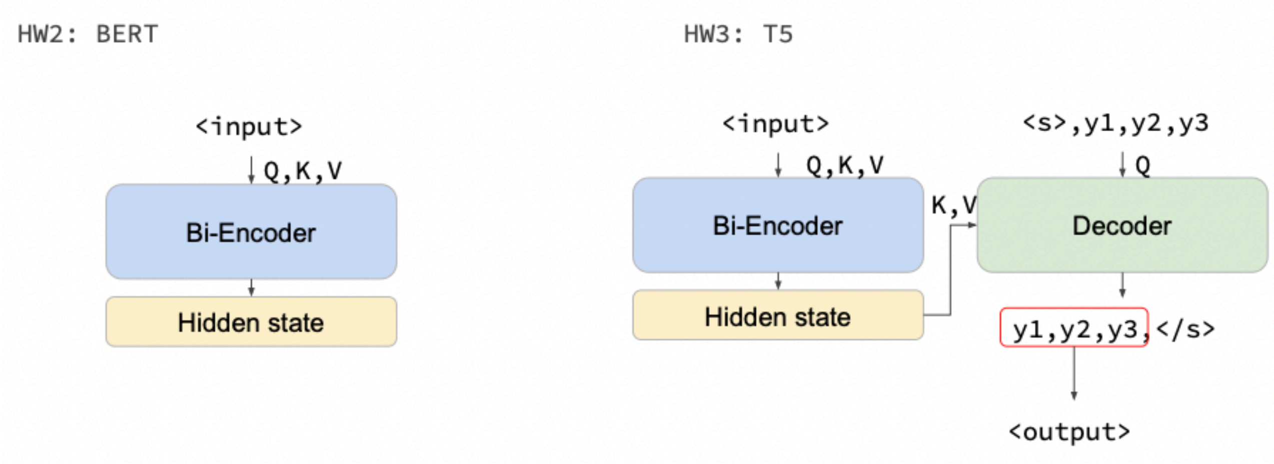
#ADL 2022
#Homework 3
R10946029，陳姵如

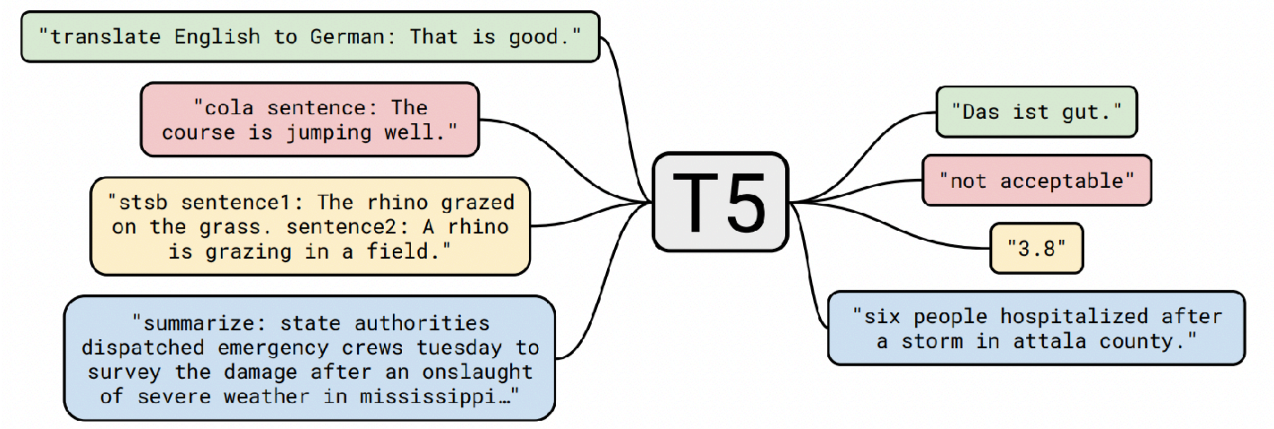
Q1: Model (2%)

Model (1%)

Describe the model architecture and how it works on text summarization.



T5 不同於 BERT 是 encoder-decoder 架構 (如上方右圖)，特色是 pre-train 的時候把很多本來不是 seq-to-seq 的 task 也用 seq-to-seq 來做，如下圖範例有數值評分預測、分類問題等等。



#ADL 2022
#Homework 3
R10946029，陳姵如

Preprocessing (1%)

Describe your preprocessing (e.g. tokenization, data cleaning and etc.)

```
tokenizer = AutoTokenizer.from_pretrained("google/mt5-small")

def preprocess(articles):
    encode_articles = {}
    maintext = [article["maintext"] for article in articles]

    # Tokenize
    encode_articles = tokenizer(maintext,
                                max_length=max_length,
                                truncation=True,
                                padding=True)

    if 'title' in examples[0].keys():
        titles = [article["title"] for article in articles]
        encode_articles['title'] = tokenizer(titles,
                                             max_length=56,
                                             truncation=False,
                                             padding=True)
    return encode_articles
```

我僅使用文章前 384 字、title 的部分則取 56 字（原因是 training data 中最長 title 是 56 字）。

Q2: Training (2%)

Hyperparameter (1%)

Describe your hyperparameter you use and how you decide it.

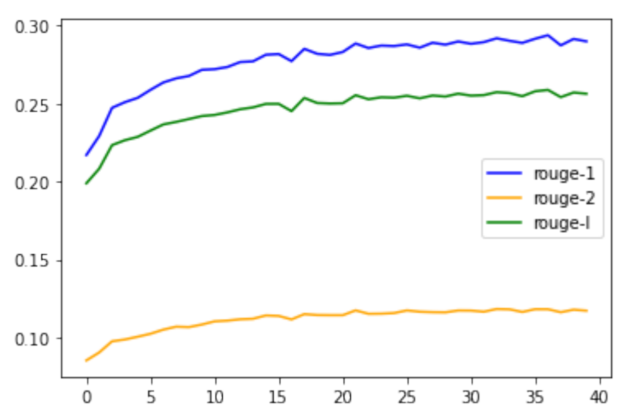
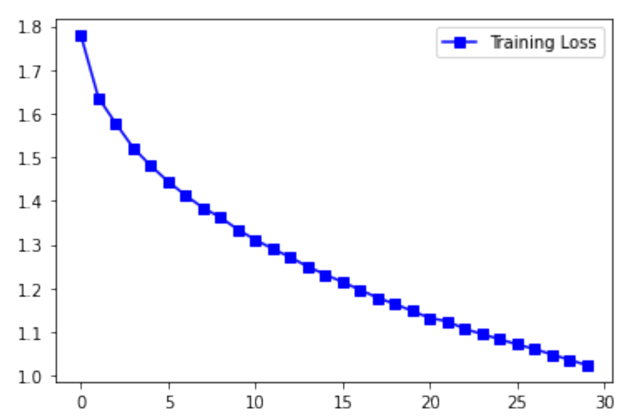
Batch size: 32

Optimizer: AdamW

Learning rate: 1e-4

Learning Curves (1%)

Plot the learning curves (ROUGE versus training steps)

Rouge score (Beam Search=5, Max Length=64)	Training loss
	

#ADL 2022
#Homework 3
R10946029，陳姵如

Q3: Generation Strategies(6%)

Stratgies (2%)

Describe the detail of the following generation strategies:

	Description
Greedy	 每次選最高機率的字。
Beam Search	每次選機率最大的字，不一定會是最佳路徑。但因為計算所有可能的 word 組合，time complexity 會很高，因此會決定要觀察 k 條 path，取最終機率最大的那條。也就是說當 k=1 時，等同於 Greedy。
Top-k Sampling	從大到小排序前 k 的字 sample
Top-p Sampling	從大到小排序機率加起來小於 p 的字 sample
Temperature	$$P(w_t) = \frac{e^{s_w/\tau}}{\sum_{w' \in V} e^{s_{w'}/\tau}}$$ 字在算 softmax 的時候指數多除上一個參數 temperature hyperparameter τ，所以當 τ 大，機率分佈會比較請向於 uniform 分佈，反之則會集中於某幾個字。

#ADL 2022
#Homework 3
R10946029，陳姵如

Hyperparameters (4%)

Try at least 2 settings of each strategies and compare the result.
What is your final generation strategy? (you can combine any of them)

	Greedy	Beam Search (K=5)
Rouge-1	0.280191408252007	0.289943029055722
Rouge-2	0.103900554513634	0.117273801546114
Rouge-l	0.245942529959404	0.256366734128703
Elapse (secs.)	178.7218	314.4461

從上面數值可以看到使用 beam search 在各項 rouge score，然而在時間上的確也需約一倍的時間。