a. 就只有使用助教所提供的 data 的相關資料

b. 有把 source.csv 跟 target.csv 用 write\_csv.py 串成 text.csv，並用t5-small的架構做了tokenize，然後有對 target 做 padding，設 max\_input\_length = 60 和 max\_target\_length = 30

c. 我是使用 t5-small 的架構

model 的 configuration 如下

T5Config {  
  "\_name\_or\_path": "t5-small",  
  "architectures": [  
    "T5WithLMHeadModel"  
  ],  
  "d\_ff": 2048,  
  "d\_kv": 64,  
  "d\_model": 512,  
  "decoder\_start\_token\_id": 0,  
  "dropout\_rate": 0.1,  
  "eos\_token\_id": 1,  
  "feed\_forward\_proj": "relu",  
  "initializer\_factor": 1.0,  
  "is\_encoder\_decoder": true,  
  "layer\_norm\_epsilon": 1e-06,  
  "model\_type": "t5",  
  "n\_positions": 512,  
  "num\_decoder\_layers": 6,  
  "num\_heads": 8,  
  "num\_layers": 6,  
  "output\_past": true,  
  "pad\_token\_id": 0,  
  "relative\_attention\_max\_distance": 128,  
  "relative\_attention\_num\_buckets": 32,  
  "task\_specific\_params": {  
    "summarization": {  
      "early\_stopping": true,  
      "length\_penalty": 2.0,  
      "max\_length": 200,  
      "min\_length": 30,  
      "no\_repeat\_ngram\_size": 3,  
      "num\_beams": 4,  
      "prefix": "summarize: "  
    },  
    "translation\_en\_to\_de": {  
      "early\_stopping": true,  
      "max\_length": 300,  
      "num\_beams": 4,  
      "prefix": "translate English to German: "  
    },  
    "translation\_en\_to\_fr": {  
      "early\_stopping": true,  
      "max\_length": 300,  
      "num\_beams": 4,  
      "prefix": "translate English to French: "  
    },  
    "translation\_en\_to\_ro": {  
      "early\_stopping": true,  
      "max\_length": 300,  
      "num\_beams": 4,  
      "prefix": "translate English to Romanian: "  
    }  
  },  
  "transformers\_version": "4.18.0",  
  "use\_cache": true,  
  "vocab\_size": 32128  
}

問題2

a.

in-domain sacrebleu : {'bleu': 4.715995127409364}

out-of-domain sacrebleu : {'bleu': 5.328754326230293}

in-domain perplexity score : 32.63237380981445

out-of-domain perplexity score : 31.199960708618164

b.

in\_domain:

19:

i like babies. I like to make babies laugh. i enjoy making other people laugh.

112

i enjoy swimming. I like to swim. I travel to new places. i travel to new places.

out\_of\_domain :

578

i like bears. I like bears and surfing. i enjoy surfing.

775

we own several dogs. I have a dog. my mother works in healthcare.

846

i like songs. I like to sing songs in the mountains. i like to be surrounded by nature.

 c.

（1）從隨便取的5個例子來看，in domain 的第19跟out-of-domain的578跟846感覺蠻自然的，就蠻像是一般講話的內容，然而in domain的112就有點硬轉的感覺（從喜歡游泳突然說我去過很多地方），而out of domain的775更是完全不相關，我有一隻狗跟我媽媽在healthcare工作沒什麼關係。

  (2)   in domain 的第19 輸入baby就可以在concept net 上找到 laugh (laugh at its mother)，in domain 的 swim 雖然可以在concept net上面找到 travel，但是以一般人的理解來說這兩個詞彙的連結是有點牽強的，像是out domain的775 dog 在 concept net 上面就找不到 health care，也很明顯這兩個字也沒有直接的關係，在out domain的846 mountains的concept net 上可以找到群山，群山代表著被大自然環繞，也可以在nature 的concept net 上找到 mountain，也代表彼此蠻有關聯的，在彼此之間相差太遠的在concept net 的 hidden concepts 上可能也會找不到。

（3）有好有壞， in\_domain的19就蠻好的，但像112就有點牽強硬轉，775只是說自己也有一隻狗，跟想導到的主題毫無關聯，但產生出的句子至少都是可閱讀的。