Lab3

額外補充

大綱

- Word Embedding
- Keras

WordNet

WordNet 包含有同義詞和上位詞的英語字典。

同義詞的範例

```
from nltk.corpus import wordnet as wn
```

```
for synset in wn.synsets("chopper"):
    print("(%s)" % synset.pos(),
    ", ".join([l.name() for l in synset.lemmas()]))
```

```
(n) chop, chopper
(n) chopper, pearly
(n) helicopter, chopper, whirlybird, eggbeater
(n) cleaver, meat_cleaver, chopper
```

上位詞的範例

```
from nltk.corpus import wordnet as wn
panda = wn.synset("panda.n.01")
hyper = lambda s: s.hypernyms()
list(panda.closure(hyper))
```

```
[Synset('procyonid.n.01'),
 Synset('carnivore.n.01'),
 Synset('placental.n.01'),
 Synset('mammal.n.01'),
 Synset('vertebrate.n.01'),
 Synset('chordate.n.01'),
```

WordNet 的問題

- 需要人工去標註
- 主觀性
- 難以數值化，不利進行機器學習訓練

用R[1]來做 Word Vector

以這些字為例: apple, ball, cat, dog, elephant …

以ID表示

相鄰ID沒有特別的關係

apple	0
ball	1
cat	2
dog	3
Elephant	4

用R^v來做 Word Vector

V = 單字數量 : One-hot encoding

V < 單字數量 : Word embedding

One-hot Encoding

以這些字為例: apple, ball, cat, dog, elephant …

該字ID的維度值為1，其餘皆為零。

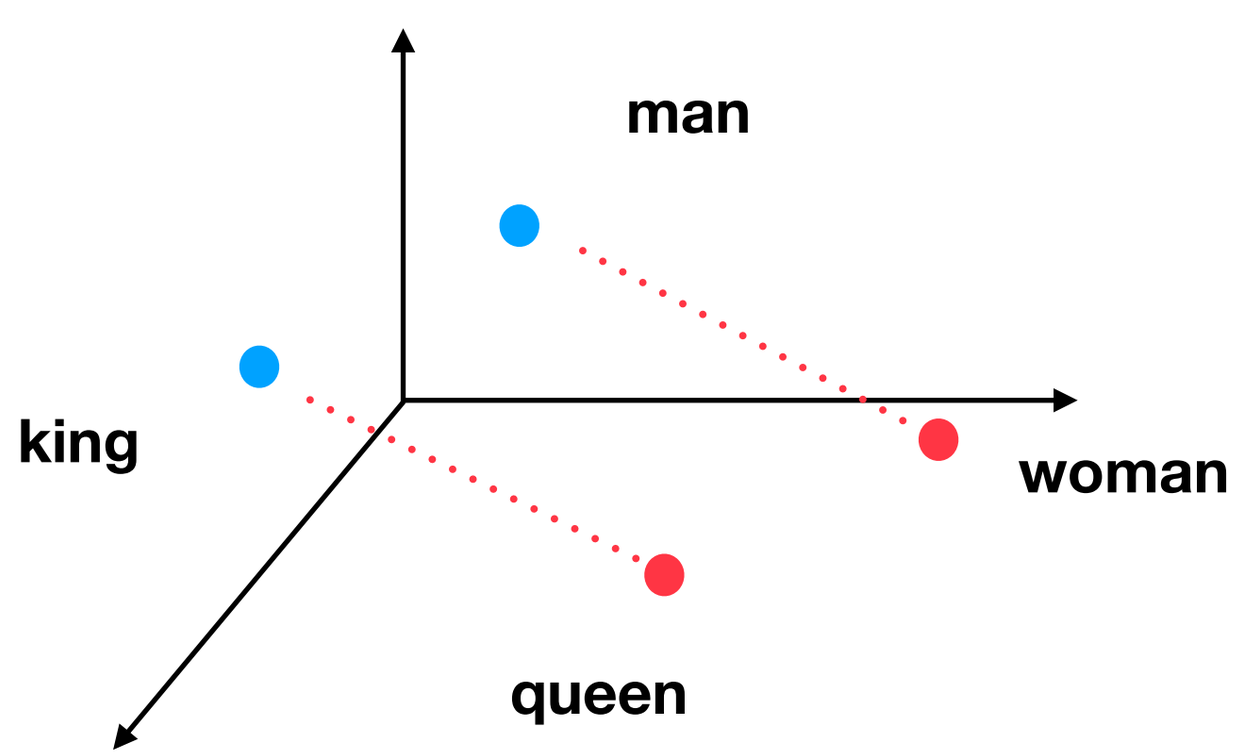
屬於離散，每個字的距離一樣。

apple	(1, 0, 0, 0, 0)
ball	(0, 1, 0, 0, 0)
cat	(0, 0, 1, 0, 0)
dog	(0, 0, 0, 1, 0)
elephant	(0, 0, 0, 0, 1)

Word Embedding

相似的字要在附近

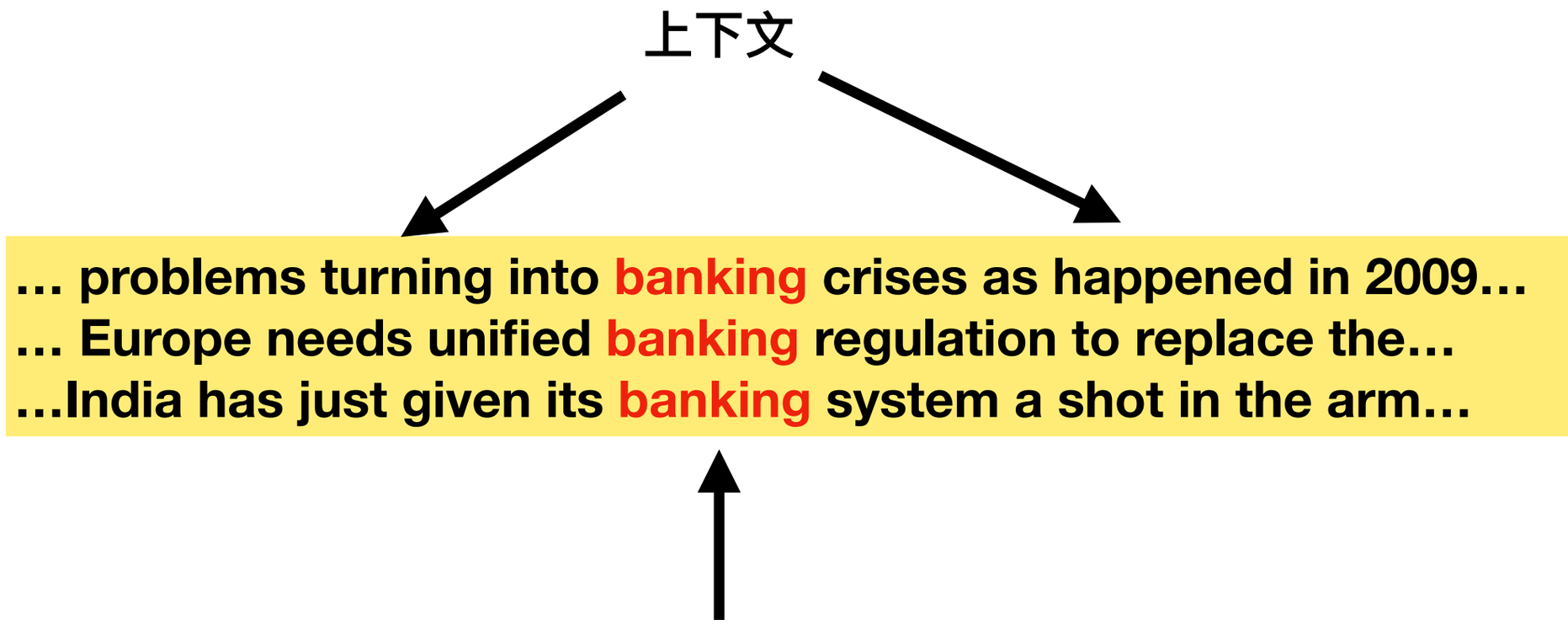
$V_{king} - V_{queen}$ 要相近於 $V_{man} - V_{woman}$



透過上下文來表示一個字

核心的想法：一個字的意思跟他鄰近的字有相關。

盡可能地去收集越多得上下文，只看前後n個字即可。



Word Vectors

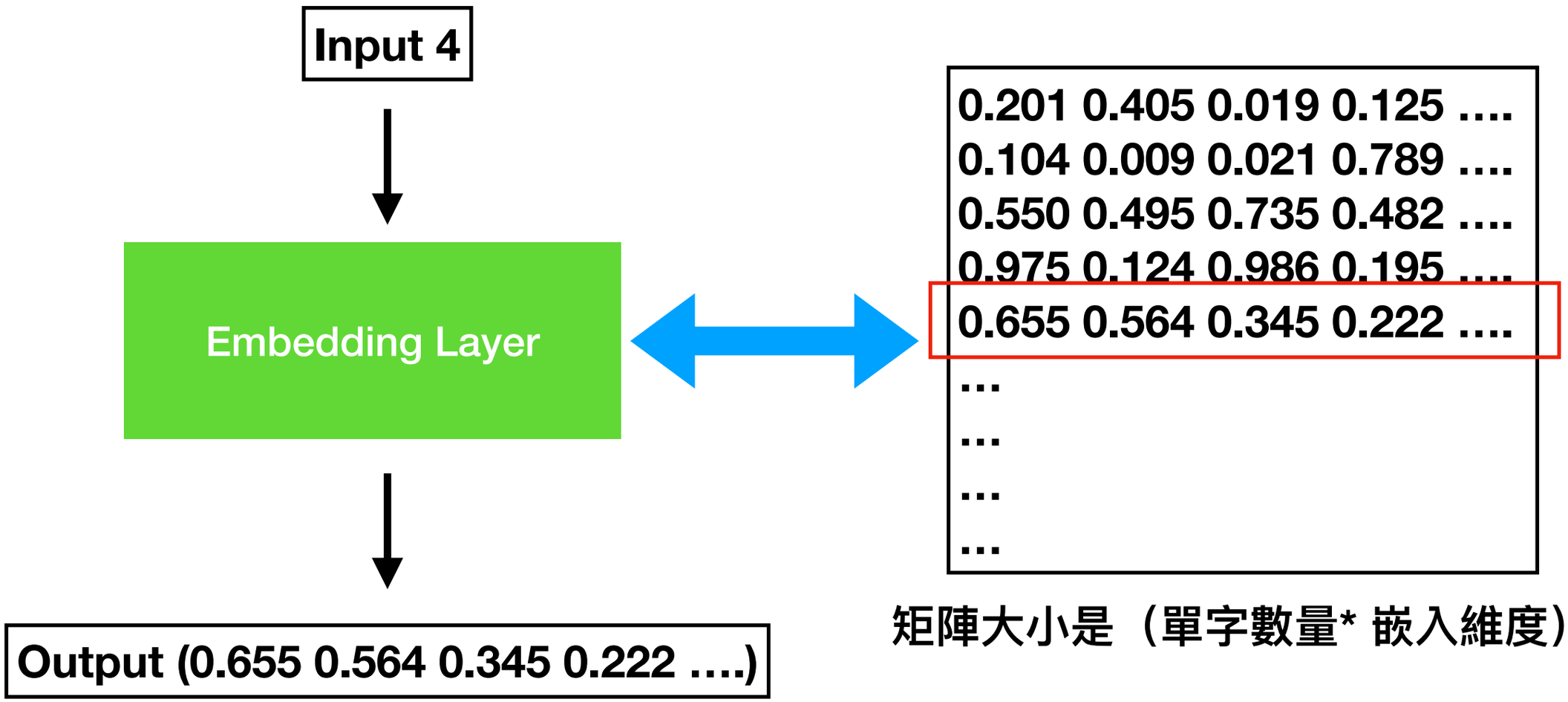
- 每個字代表一個向量，我們利用上下文來決定向量的數值。

$$\textit{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Word Embeddings

1. Word2vec
2. GloVe
3. fastText

Word Embeddings



Keras

1. Build model
2. Data preprocess
3. Compilation
4. Training

Build Model

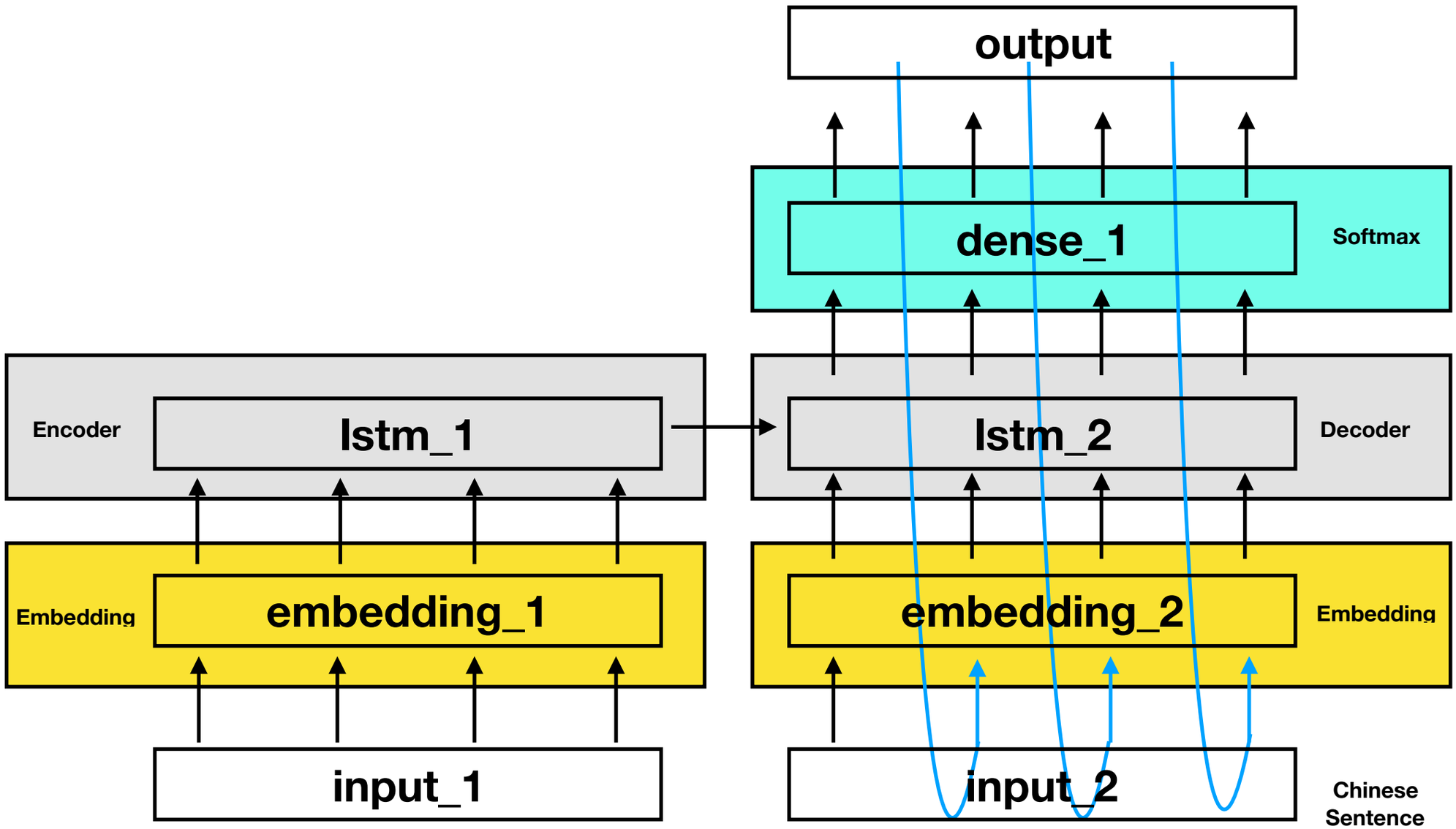
- 决定你的input shape和output shape
- Embedding Layer
  - Input shape: (batch_size, sequence_length)
  - Output shape: (batch_size, sequence_length, output_dim)

```
encoder_inputs = Input(shape=(None, ))
encoder_embedding = Embedding(input_dim=1000, output_dim=256)
en_x=  encoder_embedding(encoder_inputs)
print(encoder_embedding.input_shape, encoder_embedding.output_shape)
```

```
(None, None) (None, None, 256)
```

None 表示任意大小

Build Model


output
dense_1
Softmax
Encoder
lstm_1
lstm_2
Decoder
Embedding
embedding_1
embedding_2
Embedding
input_1
input_2
Chinese
Sentence

Date Preprocess

- 準備numpy array資料
- 例子: lstm_seq2seq.py 第103~110行

```
encoder_input_data = np.zeros(
    (len(input_texts), max_encoder_seq_length, num_encoder_tokens),
    dtype='float32')
decoder_input_data = np.zeros(
    (len(input_texts), max_decoder_seq_length, num_decoder_tokens),
    dtype='float32')
decoder_target_data = np.zeros(
    (len(input_texts), max_decoder_seq_length, num_decoder_tokens),
    dtype='float32')
```

Compilation

- optimizer : 優化器，調整模型參數的方法。
- loss : 計算損失的方式
  - categorical_crossentropy
  - sparse_categorical_crossentropy

categorical_crossentropy & sparse_categorical_crossentropy

```
decoder_target_data = np.zeros((len(input_texts),
max_decoder_seq_length, num_decoder_tokens), dtype='float32')
```

categorical_crossentropy 使用one-hot encoding

```
decoder_target_data = np.zeros((len(input_texts),
max_decoder_seq_length, 1), dtype='float32')
```

sparse_categorical_crossentropy 只需要 index 就行

Training

- Epochs : 一個epoch 等於訓練整份資料一次。
- Batches : 一次使用幾筆資料作為訓練。

作業繳交

- lab3_<學號>.ipynb, output.txt
- 使用全部資料，建議train 50個epoch以上
- 依順序輸出第4077, 2122, 3335, 1464, 8956, 7168, 3490, 4495, 5100, 119行到output.txt，第4077例子如下。

  Input sentence<tab>He is afraid of snakes.
  Decoded sentence<tab>他害怕蛇。<end>

- 翻譯的結果必須有一定成效，如果結果太差或自行修改output.txt，以零分算。

TA’s output

```
-
Input sentence: he is afraid of snakes .
Decoded sentence:  他 害怕 蛇 。 <end>
-
Input sentence: i miss you so much .
Decoded sentence:  我 如此 想念 你 。 <end>
-
Input sentence: we 're going by train .
Decoded sentence:  我们 要 乘火车 去 。 <end>
-
Input sentence: the sky is clear .
Decoded sentence:  天空 很 晴朗 。 <end>
-
Input sentence: wearing a suit , he stood out .
Decoded sentence:  他 穿著 西 裝 站 了 出 來 。 <end>
-
Input sentence: she made a serious mistake .
Decoded sentence:  她 犯 了 一 個 嚴 重 的 錯 <end>
-
Input sentence: have you eaten dinner ?
Decoded sentence:  你 吃 晚 飯 了 嗎 ? <end>
-
Input sentence: what do you want to be ?
Decoded sentence:  你 想 成为 什么 ? <end>
-
Input sentence: tom is going to help us .
Decoded sentence:  汤姆 要 帮助 我们 。 <end>
-
Input sentence: he 's lazy .
Decoded sentence:  他 很 懶 。 <end>
```