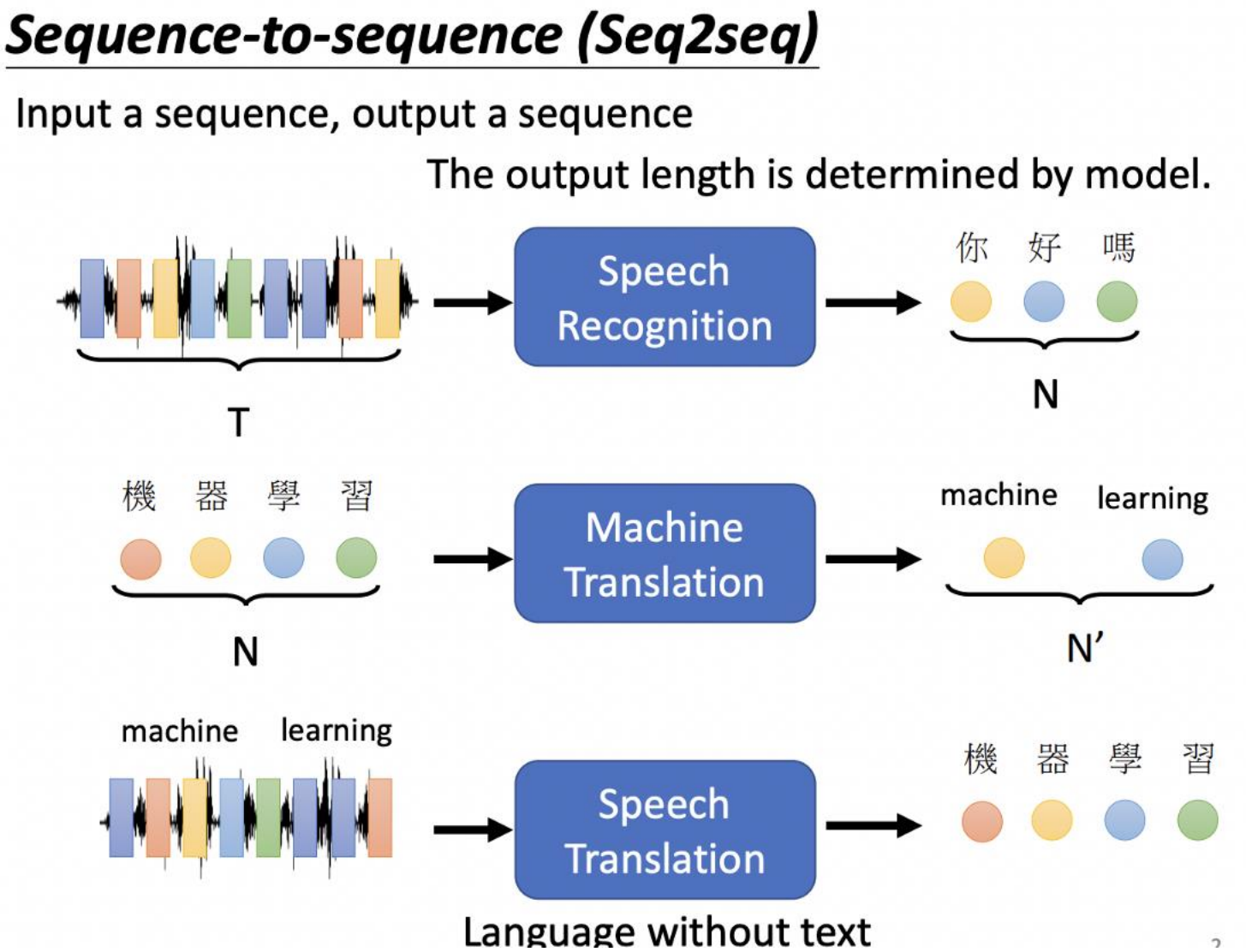
NLP簡介和Tokenizer

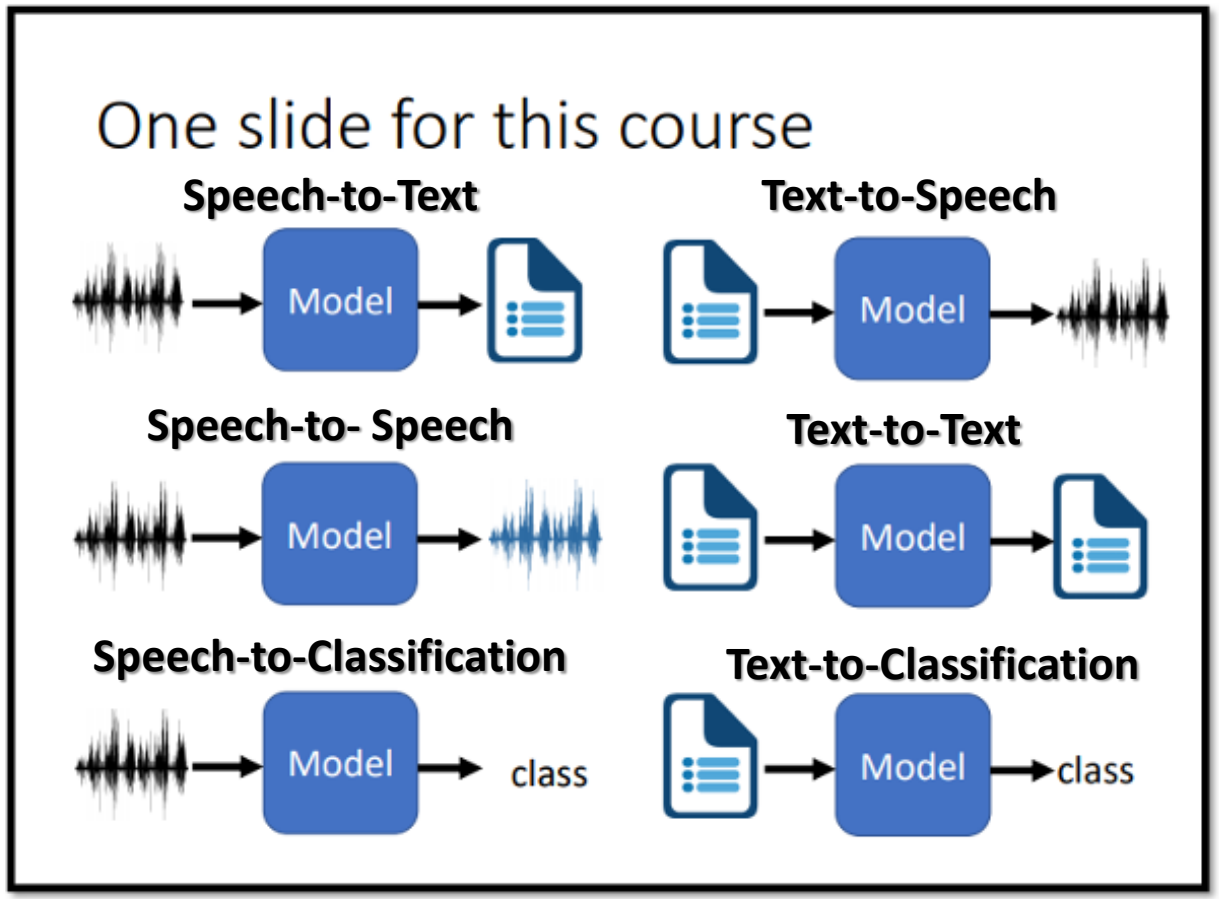
黃志勝（Tommy Huang）
義隆電子 人工智慧研發部
國立陽明交通大學 AI學院 合聘助理教授
國立台北科技大學 電資學院合聘助理教授

Sequence-to-sequence



圖片：李宏毅老師

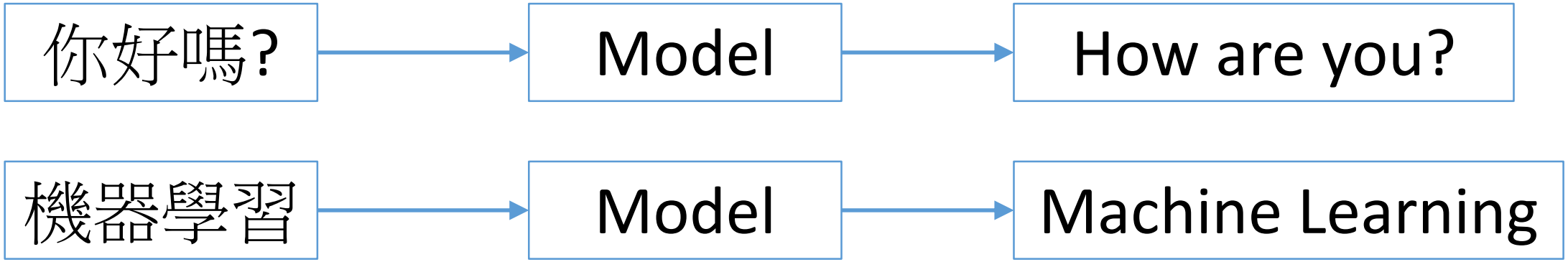
Deep Learning for Human Language Processing
深度學習與人類語言處理



Source webpage: https://speech.ee.ntu.edu.tw/~hylee/dlhlp/2020-spring.html

圖片：李宏毅老師

Sequence-to-sequence
Text-to-Text: 中英文翻譯



電腦怎麼知道

機器學習→Machine Learning

你好嗎? →How are you?

Sequence-to-sequence
Text-to-Text: 中英文翻譯

電腦怎麼知道
機器學習→Machine Learning
你好嗎? →How are you?

給電腦一堆配對好的資料，讓模型去學中文和英文之間的關聯性。

Example/dataset.xlsx

	A	B
1	english	chinese
2	Hi.	嗨。
3	Hi.	你好。
4	Run.	你用跑的。
5	Wait!	等等！
6	Wait!	等一下！
7	Begin.	开始！
8	Begin	开始
9	Hello!	你好。
10	I try.	我试试。
11	I won!	我赢了。
12	Oh no!	不会吧。
13	Cheers!	乾杯!
14	Got it?	你懂了吗？
15	He ran.	他跑了。
16	Hop in.	跳进来。

Tokenization

- Why do we need Tokenization?

你好嗎? → ”你”、”好”、”嗎”、”?”

How are you? → ” How ”、” are ”、” you”、”?”

1. 基於詞的(Word-based):

我是黃志勝

→ “我”、“是”、“黃”、“志”、“勝”

→ “我是”、“黃志勝”

→ “我”、“是”、“黃”、“志勝”

2. 基於字符(Character-based):

How are you? → “H ”、“o”、“w”、“a “、“r”、“e”、“y”、“o”、“u”、“?”

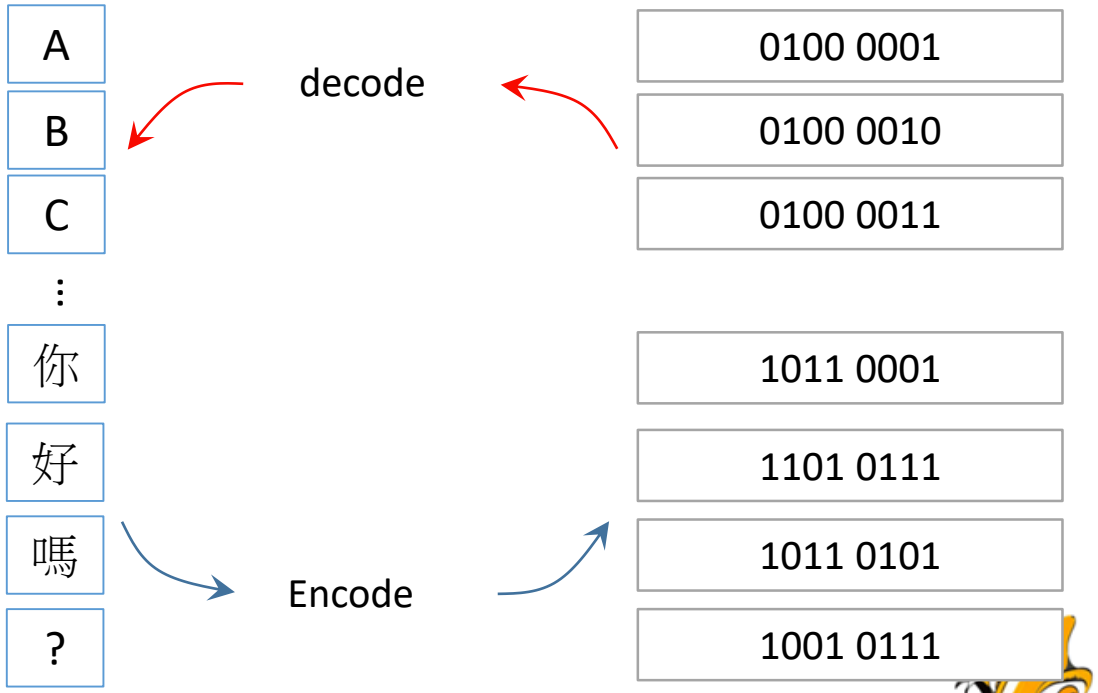
Tokenization

編碼:
ASCII
Big5
Unicode
Base64

電腦看得懂文字(中文/英文/日文/...)?

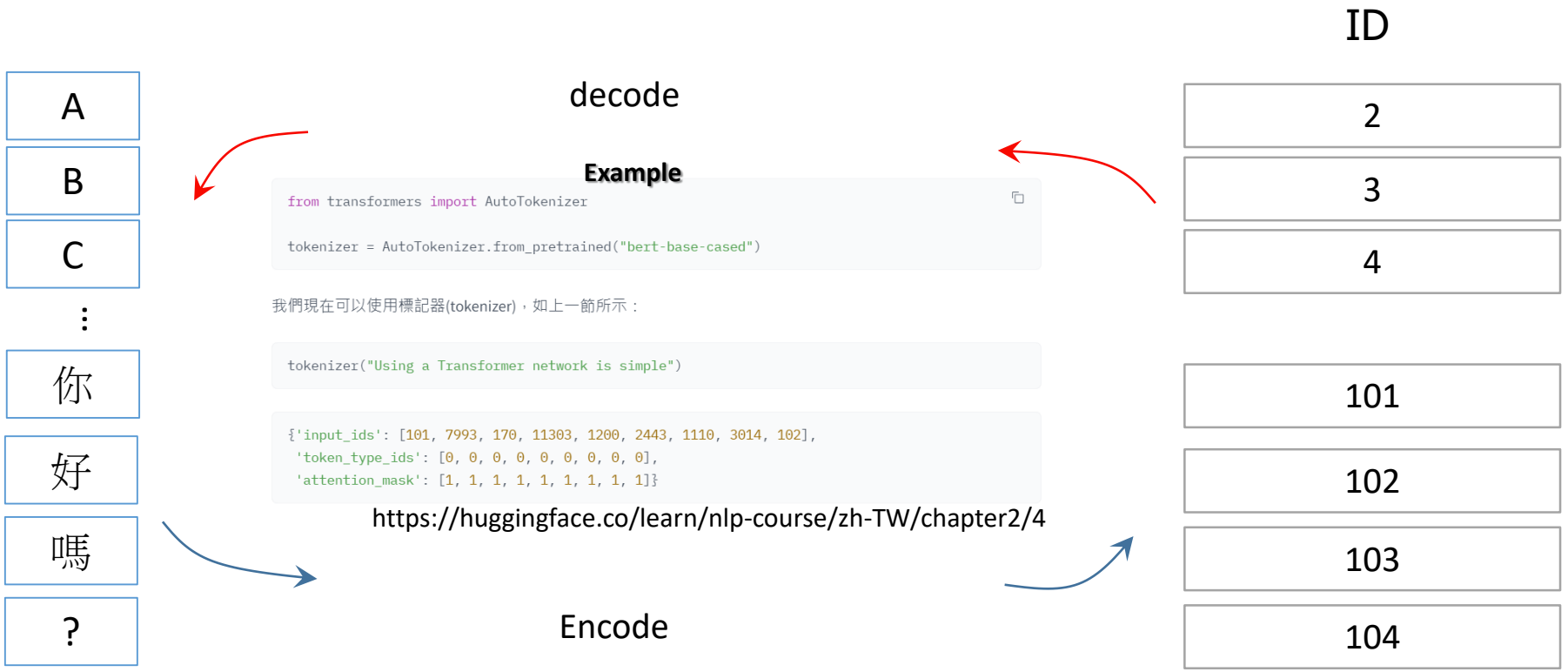
編碼: **ASCII**

二進制	十進制	十六進制	圖形	二進制	十進制	十六進制	圖形	二進制	十進制	十六進制	圖形
0010 0000	32	20	(space)	0100 0000	64	40	@	0110 0000	96	60	`
0010 0001	33	21	!	0100 0001	65	41	A	0110 0001	97	61	a
0010 0010	34	22	"	0100 0010	66	42	B	0110 0010	98	62	b
0010 0011	35	23	#	0100 0011	67	43	C	0110 0011	99	63	c
0010 0100	36	24	$	0100 0100	68	44	D	0110 0100	100	64	d
0010 0101	37	25	%	0100 0101	69	45	E	0110 0101	101	65	e
0010 0110	38	26	&	0100 0110	70	46	F	0110 0110	102	66	f
0010 0111	39	27	'	0100 0111	71	47	G	0110 0111	103	67	g
0010 1000	40	28	(	0100 1000	72	48	H	0110 1000	104	68	h



Note:中文編碼是我亂打的

Tokenization



在NLP很重要

(1)切詞

(2)文字編碼解碼

簡單說就是查字典 → 字典還是辭典

- GPT、BERT都有自己的編碼
- 幾百萬個ID怎麼讓查字典變快

Tokenization

Example/dataset.xlsx

	A	B
1	english	chinese
2	Hi.	嗨。
3	Hi.	你好。
4	Run.	你用跑的。
5	Wait!	等等！
6	Wait!	等一下！
7	Begin.	开始！
8	Begin	开始
9	Hello!	你好。
10	I try.	我试试。
11	I won!	我赢了。
12	Oh no!	不会吧。
13	Cheers!	乾杯!
14	Got it?	你懂了吗？
15	He ran.	他跑了。
16	Hop in.	跳进来。

Word Embedding:
將『字詞/句子/文件』轉換成『向量』形式

英文有自己的字典

Hi → [30, 7]

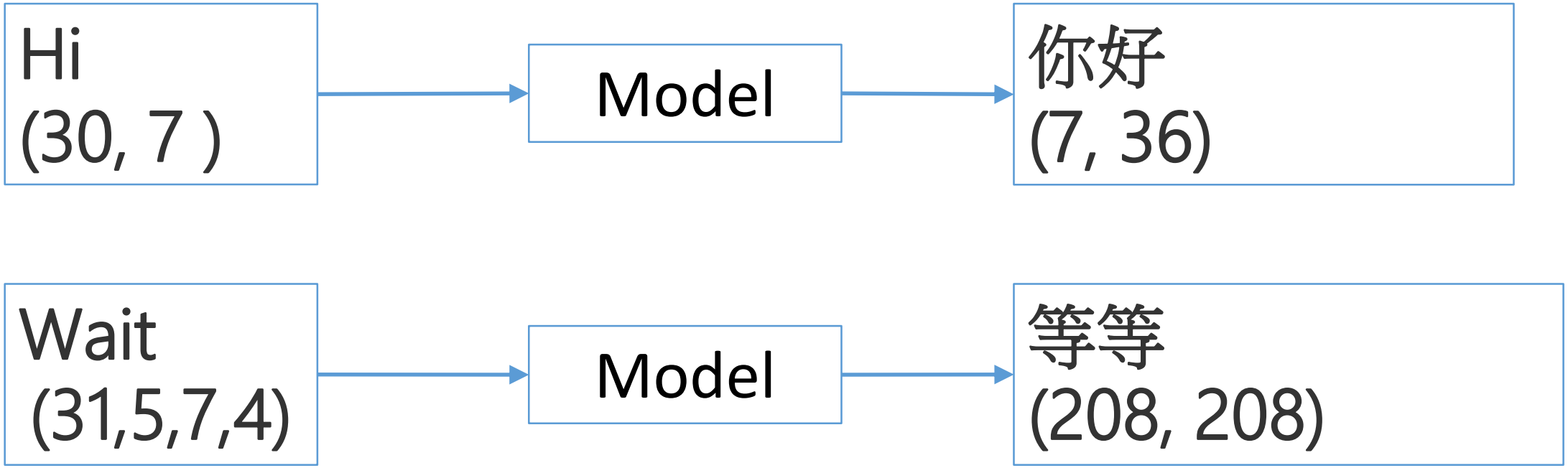
Wait → [31, 5, 7, 4]

中文有自己的字典

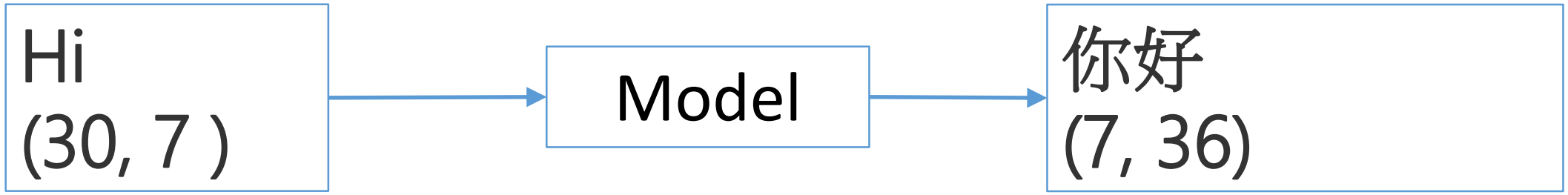
你好 → [7, 36]

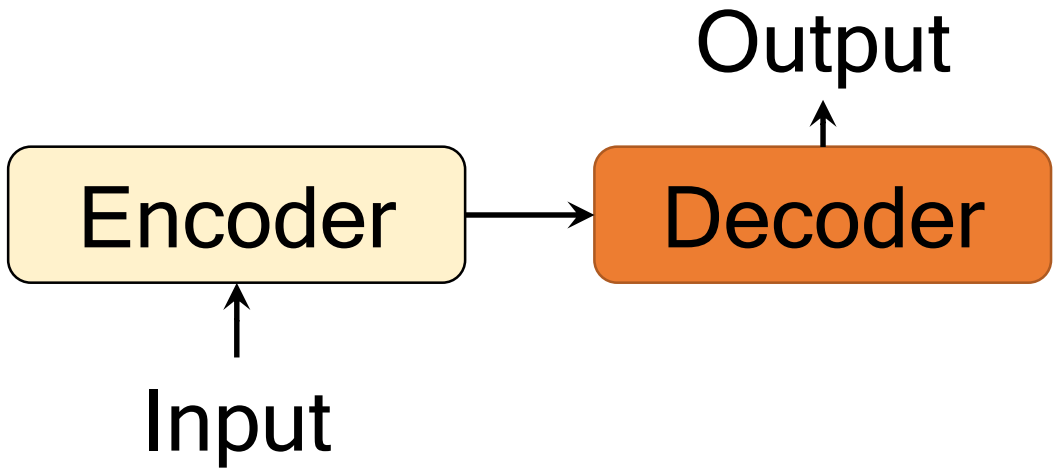
等等 → [208, 208]

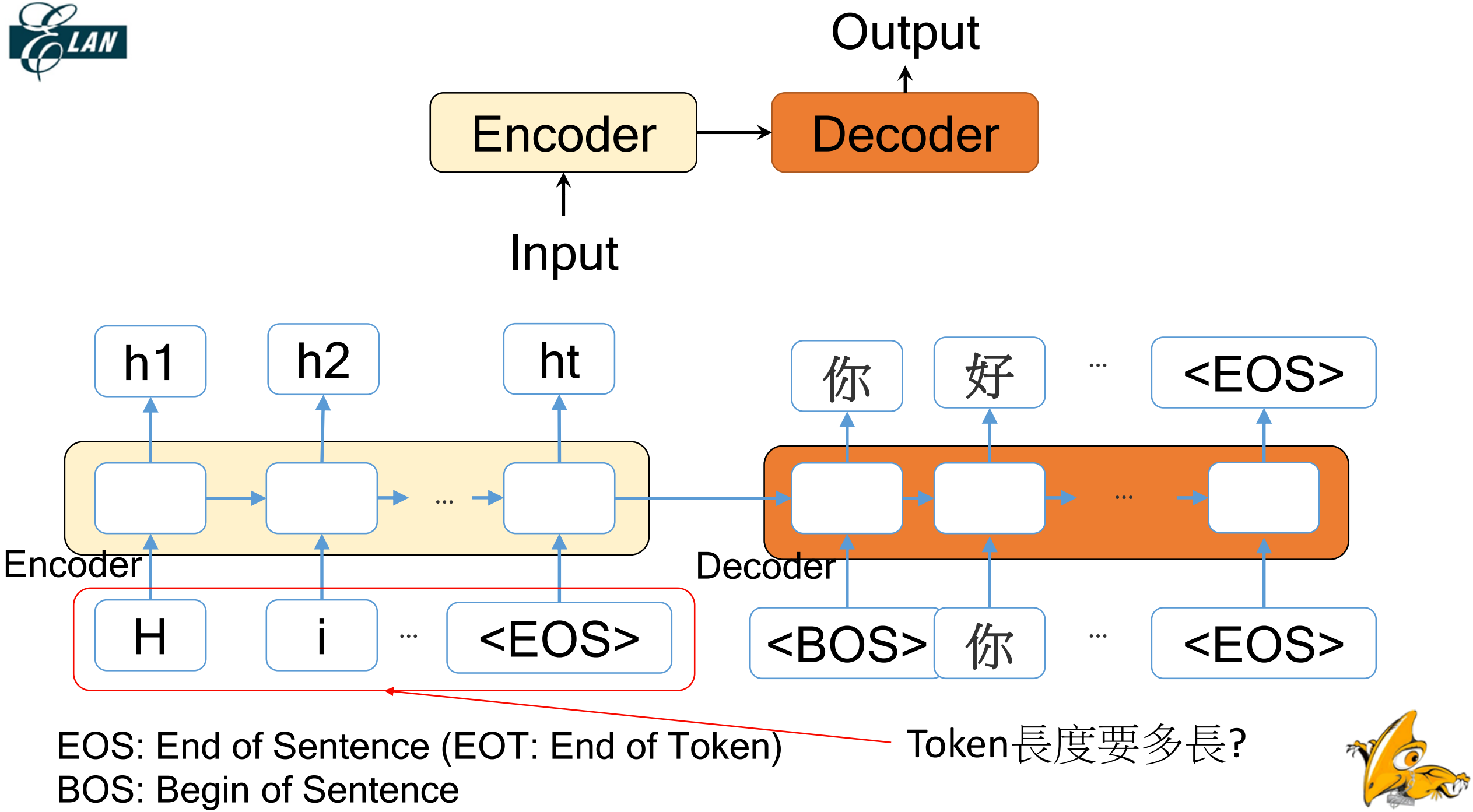
Sequence-to-sequence
Text-to-Text: 中英文翻譯

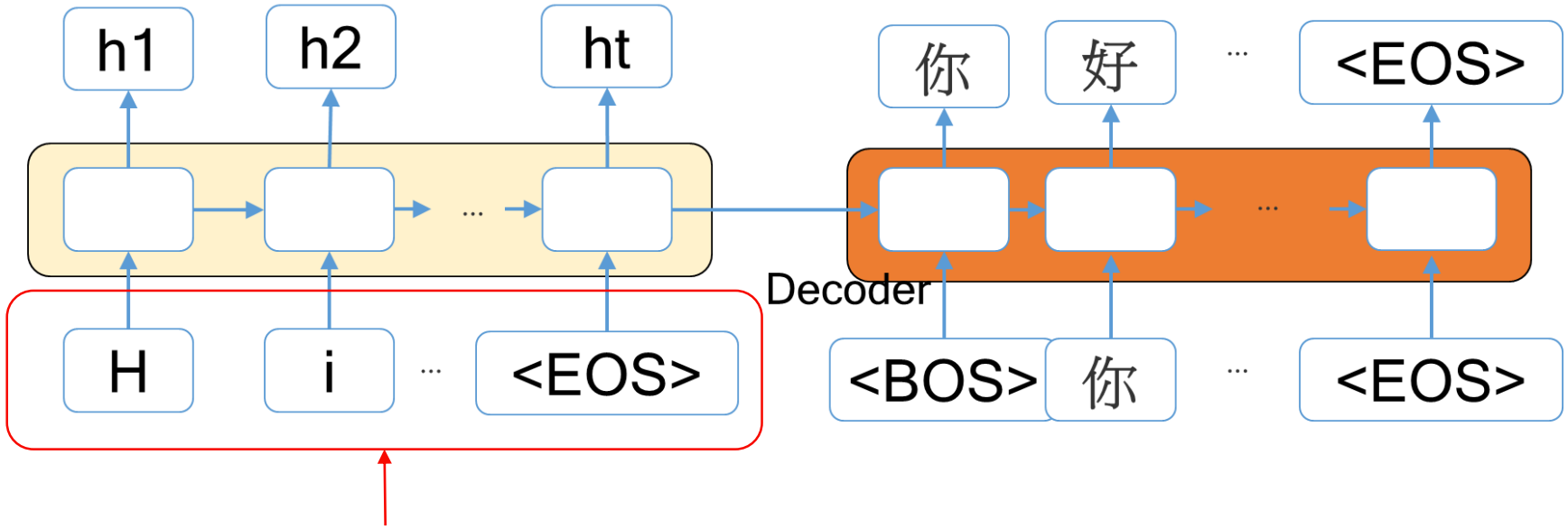


Sequence-to-sequence






Output
Encoder
Decoder
Input
h1
h2
ht
你
好
...
<EOS>
Encoder
Decoder
H
i
...
<EOS>
<BOS>
你
...
<EOS>
EOS: End of Sentence (EOT: End of Token)
BOS: Begin of Sentence
Token長度要多長?



Token長度要多長?

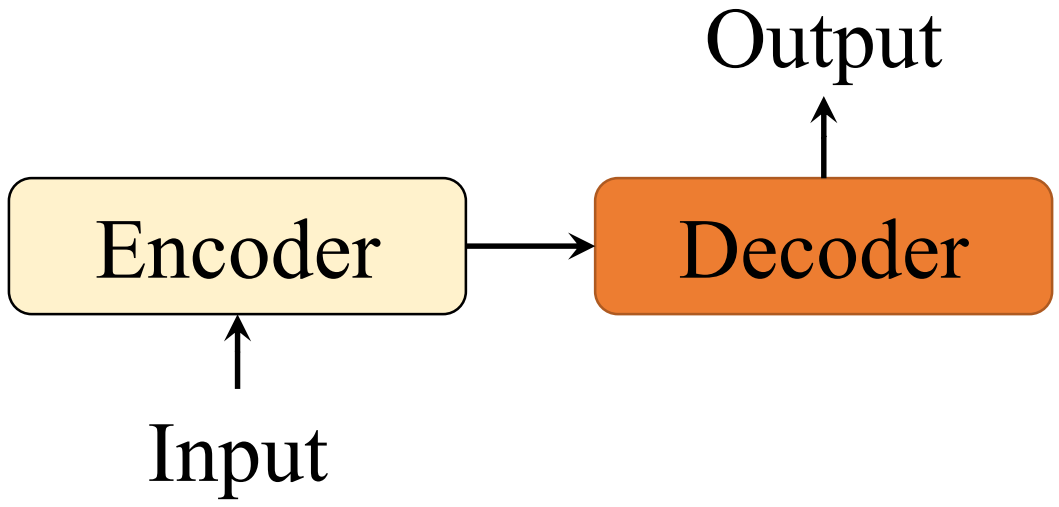
The previous example runs until you hit the model's token limit. With each question asked, and answer received, the `messages` list grows in size. The token limit fo `gpt-35-turbo` is 4,096 tokens. The token limits for `gpt-4` and `gpt-4-32k` are 8,192 and 32,768, respectively. These limits include th token count from both the message list sent and the model response. The number of tokens in the messages list combined with the value of the `max_tokens` parameter must stay under these limits or you receive an error.

https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/chatgpt?tabs=python-new

Tokenization

句子	切字	處理後的編碼	中文	切字	處理後的編碼
I am Tommy	'I', 'am', 'Tommy'	[3, 4, 5, 1, 0, 0, 0]	我是湯米	'我', '是', '湯米'	[10, 9, 8, 1, 0, 0, 0]
I am good.	'I', 'am', 'good', '.'	[3, 4, 6, 10, 1, 0, 0]	我很好	'我', '很好'	[10, 7, 1, 0, 0, 0, 0]
good to see you.	'good', 'to', 'see', 'you', '.'	[6, 7, 8, 9, 10, 1, 0]	很高興見到你	'很', '高興', '見到', '你'	[6, 5, 4, 3, 1, 0, 0]

英文字典	編碼
I	3
Am	4
Tommy	5
Good	6
To	7
See	8
You	9
.	10

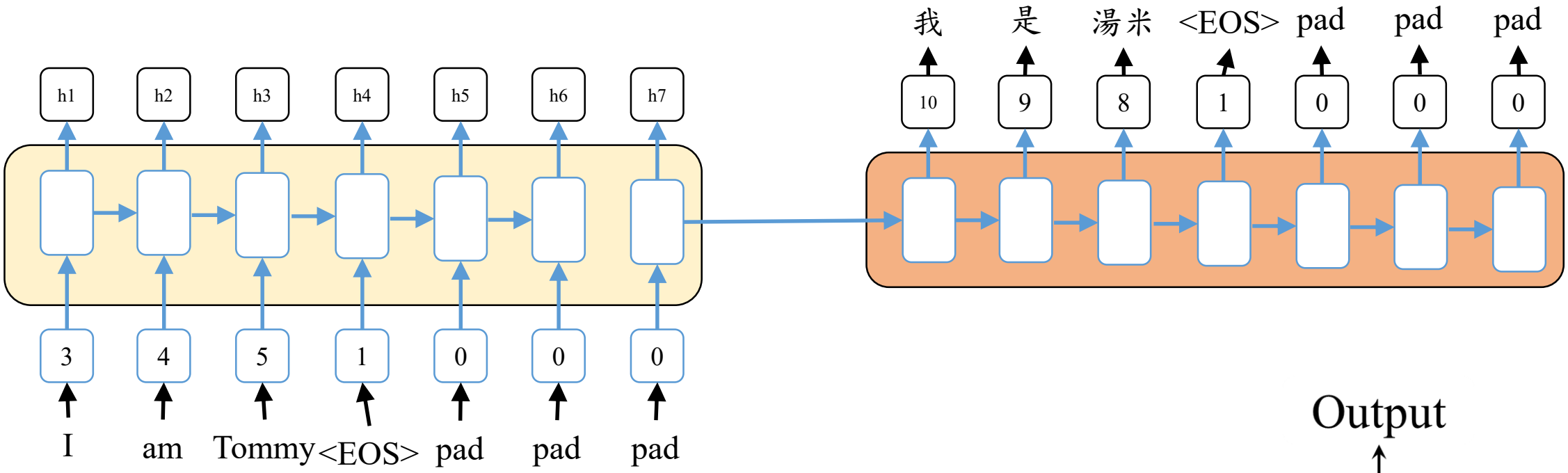


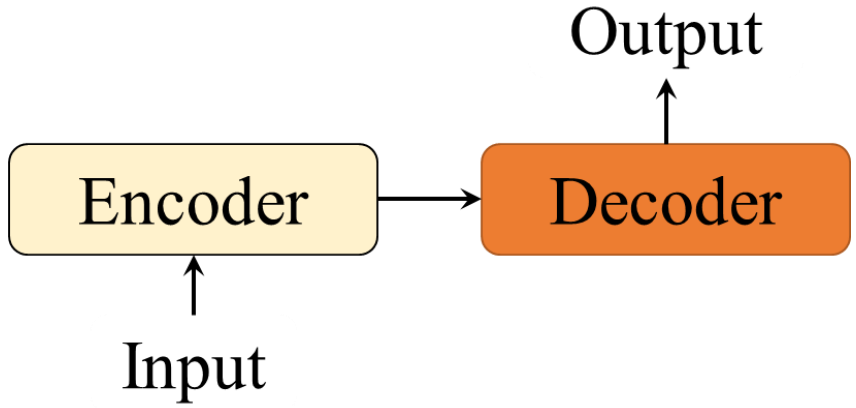
<EOS> : 1
Pad : 0
<BOS> : 2

中文字典	編碼
我	10
是	9
湯米	8
很好	7
很	6
高興	5
見到	4
你	3

Tokenization: Word Embedding

I am Tommy	'I', 'am', 'Tommy'	[3, 4, 5, 1, 0, 0, 0]	我是湯米	'我', '是', '湯米'	[10, 9, 8, 1, 0, 0, 0]



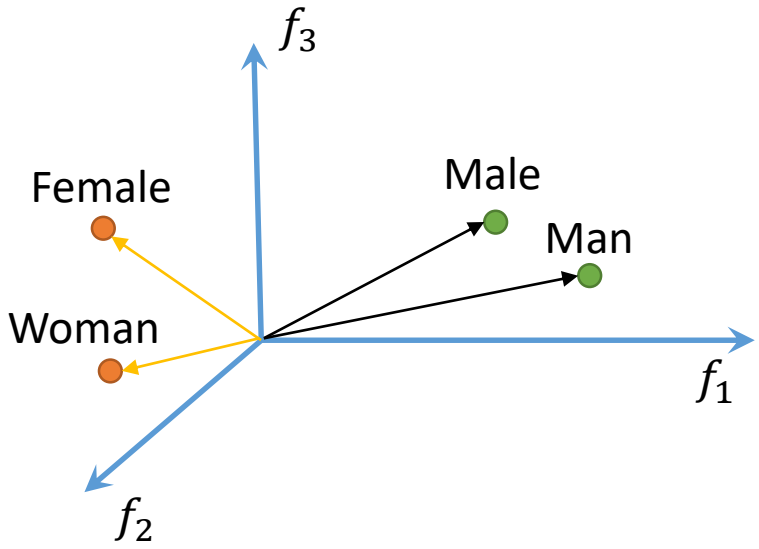


3,4,5,6→數字本身沒有意義

Tokenization: Word Embedding

	ID	Word Embedding
Man	4	$[f_1, f_2, f_3, \dots, f_d]$
Woman	5	$[f_1, f_2, f_3, \dots, f_d]$
Female	6	$[f_1, f_2, f_3, \dots, f_d]$
Male	7	$[f_1, f_2, f_3, \dots, f_d]$

- 用更多維度來表示每一個字
- 也可以在訓練模型的時候順便學習



Google: Word2vec

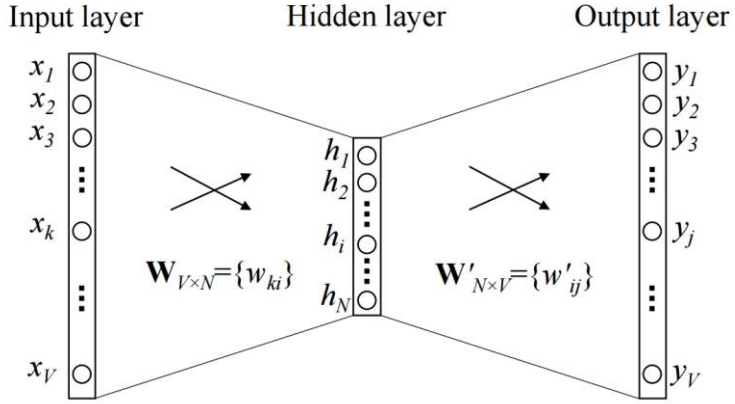
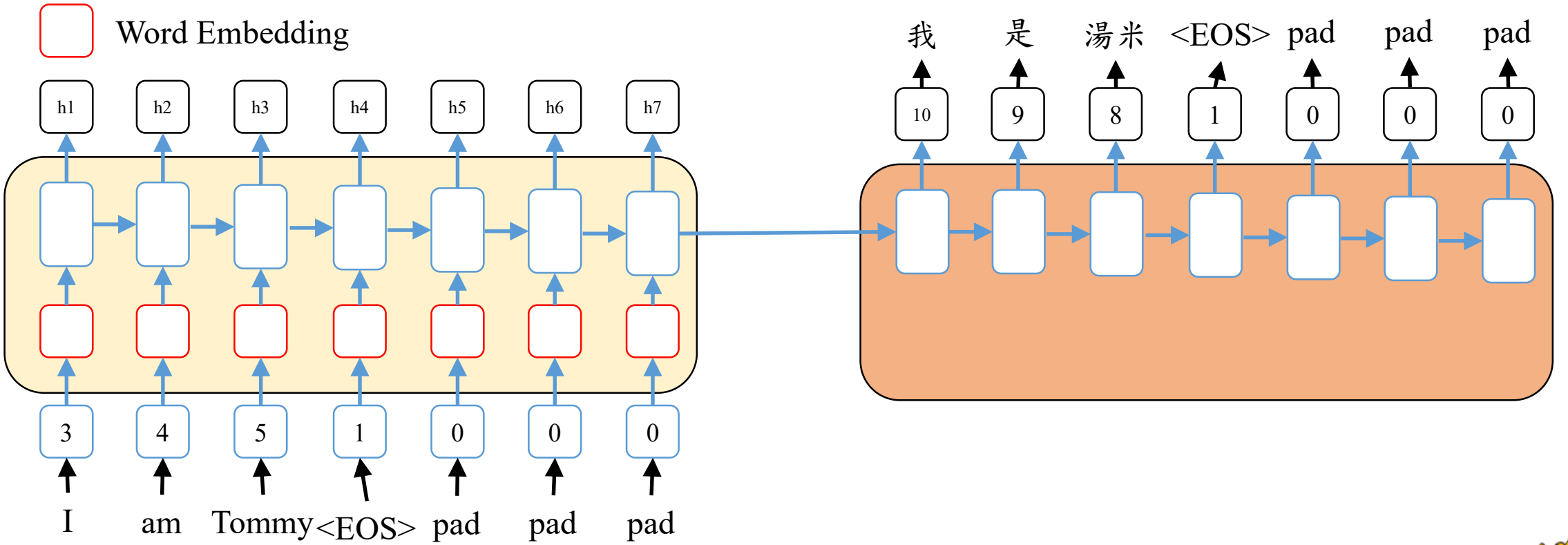


Figure 1: A simple CBOW model with only one word in the context

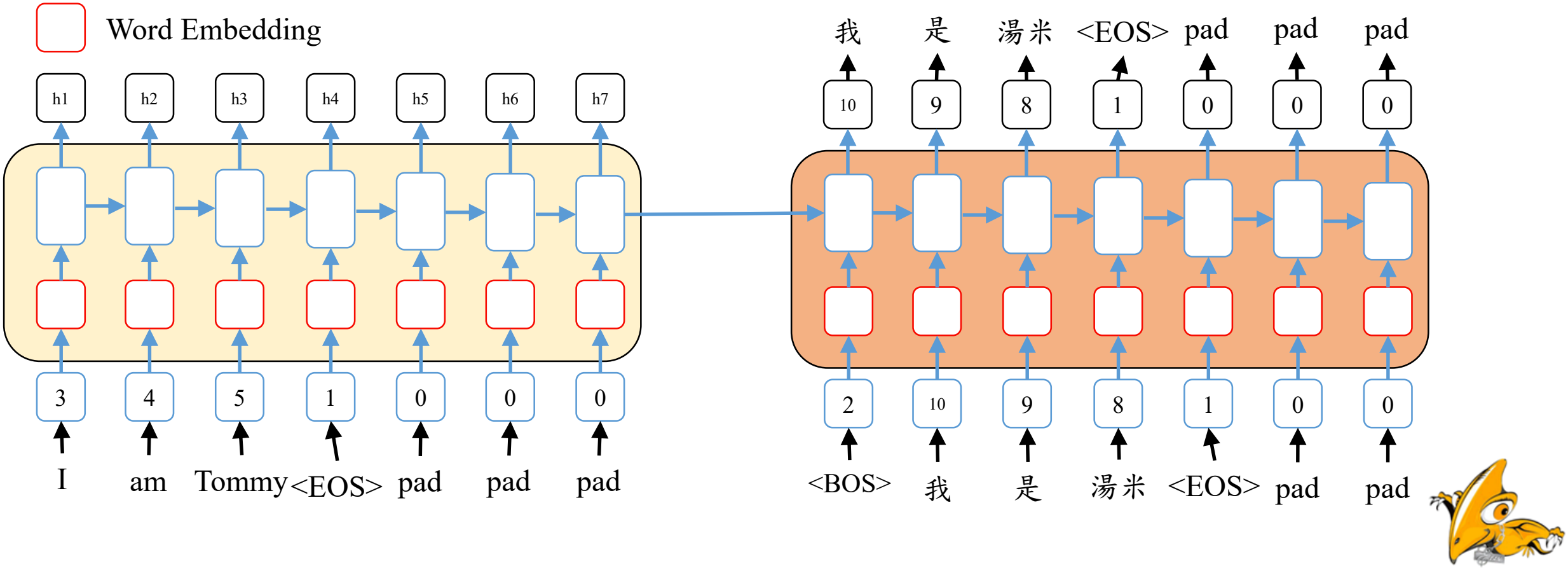
Tokenization: Word Embedding

I am Tommy	'I', 'am', 'Tommy'	[3, 4, 5, 1, 0, 0, 0]	我是湯米	'我', '是', '湯米'	[10, 9, 8, 1, 0, 0, 0]



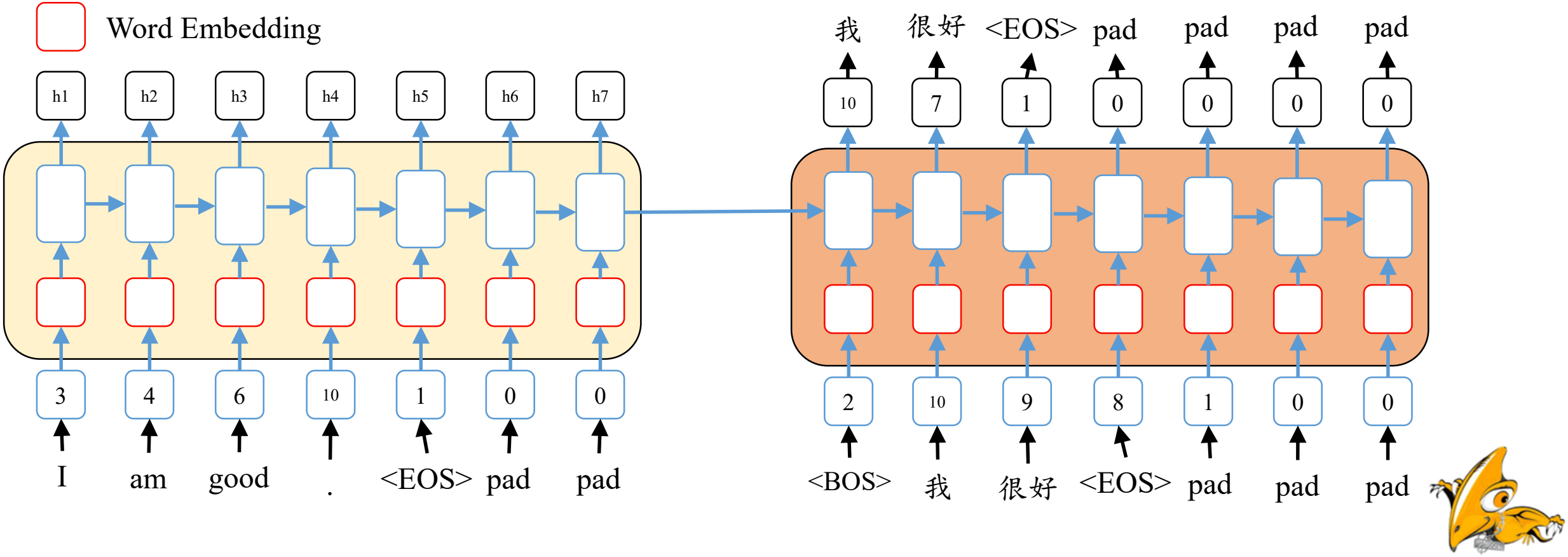
範例-I am Tommy

句子	切字	處理後的編碼	中文	切字	處理後的編碼
I am Tommy	'I', 'am', 'Tommy'	[3, 4, 5, 1, 0, 0, 0]	我是湯米	'我', '是', '湯米'	[10, 9, 8, 1, 0, 0, 0]
I am good.	'I', 'am', 'good', '.'	[3, 4, 6, 10, 1, 0, 0]	我很好	'我', '很好'	[10, 7, 1, 0, 0, 0, 0]
good to see you.	'good', 'to', 'see', 'you', '.'	[6, 7, 8, 9, 10, 1, 0]	很高興見到你	'很', '高興', '見到', '你'	[6, 5, 4, 3, 1, 0, 0]



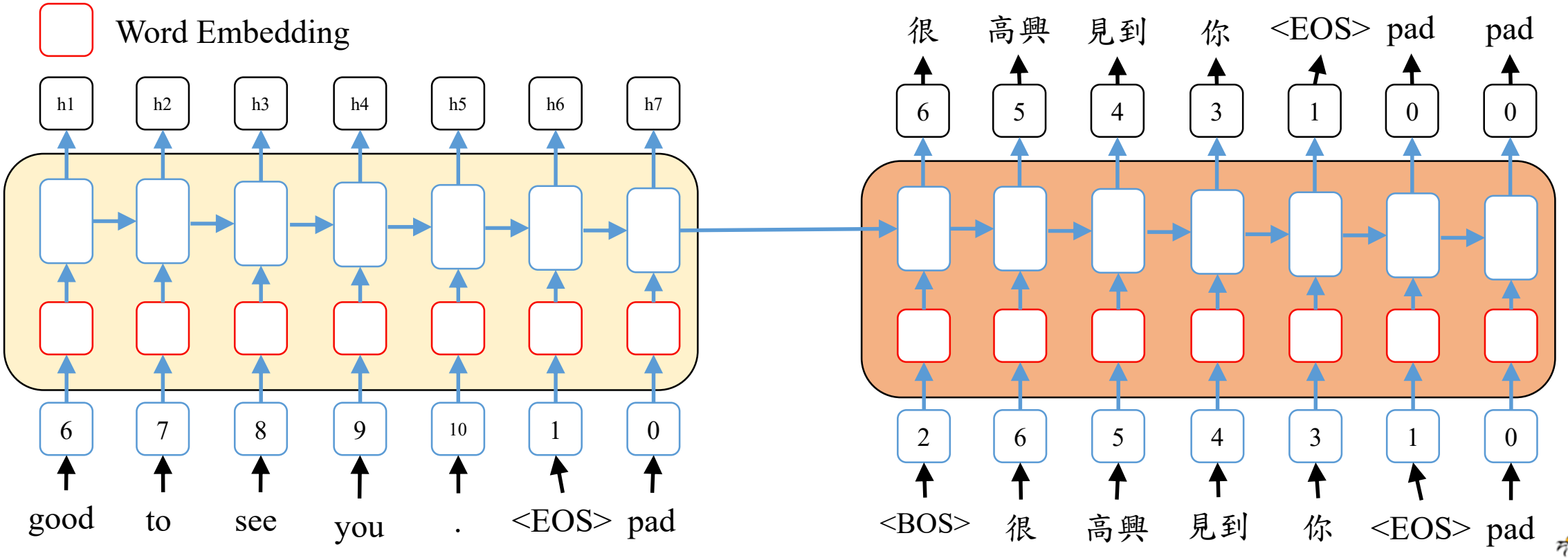
範例-I am good.

句子	切字	處理後的編碼	中文	切字	處理後的編碼
I am Tommy	'I', 'am', 'Tommy'	[3, 4, 5, 1, 0, 0, 0]	我是湯米	‘我’, ‘是’, ‘湯米’	[10, 9, 8, 1, 0, 0, 0]
I am good.	'I', 'am', 'good', '.'	[3, 4, 6, 10, 1, 0, 0]	我很好	‘我’, ‘很好’	[10, 7, 1, 0, 0, 0, 0]
good to see you.	'good', 'to', 'see', 'you', '.'	[6, 7, 8, 9, 10, 1, 0]	很高興見到你	‘很’, ‘高興’, ‘見到’, ‘你’	[6, 5, 4, 3, 1, 0, 0]



範例-good to see you.

句子	切字	處理後的編碼	中文	切字	處理後的編碼
I am Tommy	'I', 'am', 'Tommy'	[3, 4, 5, 1, 0, 0, 0]	我是湯米	‘我’, ‘是’, ‘湯米’	[10, 9, 8, 1, 0, 0, 0]
I am good.	'I', 'am', 'good', '.'	[3, 4, 6, 10, 1, 0, 0]	我很好	‘我’, ‘很好’	[10, 7, 1, 0, 0, 0, 0]
good to see you.	'good', 'to', 'see', 'you', '.'	[6, 7, 8, 9, 10, 1, 0]	很高興見到你	‘很’, ‘高興’, ‘見到’, ‘你’	[6, 5, 4, 3, 1, 0, 0]



Code Example Q&A