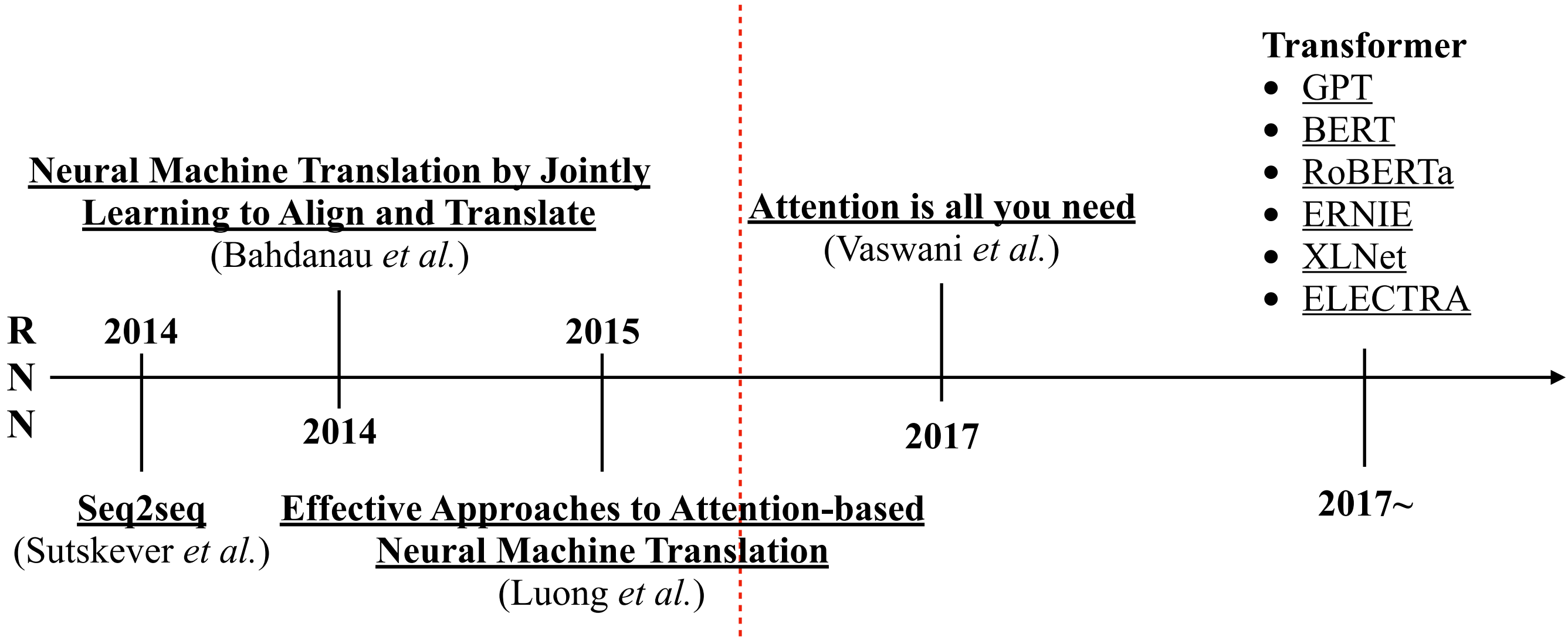
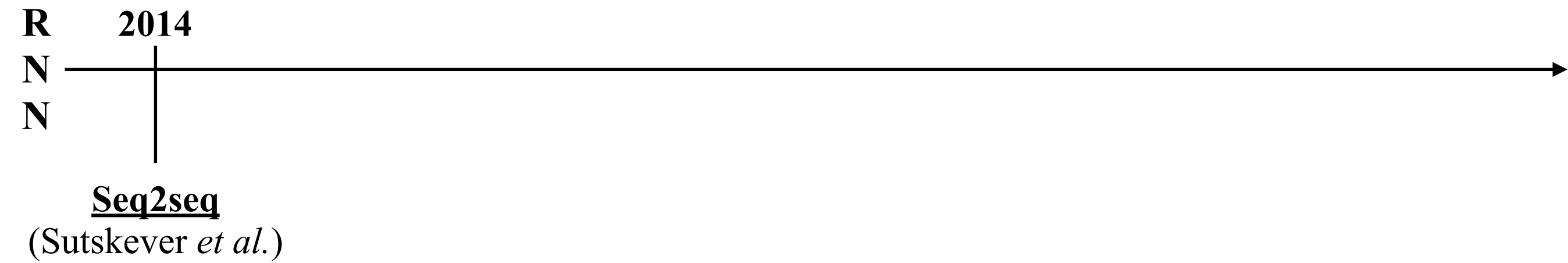


從Seq2seq到Transformer



- Seq2seq：首次將類神經網路應用在翻譯任務(Neural Machine Translation)。
- Attention：設計網路結構時，考慮人類在翻譯時的思考邏輯(Soft-Attention)。
- Transformer：克服遞迴神經網絡(RNN)的缺點，大幅度提升翻譯水準。



- Seq2seq：首次將類神經網路應用在翻譯任務(Neural Machine Translation)。

RNN

RNN series model: Vanilla RNN, LSTM, GRU

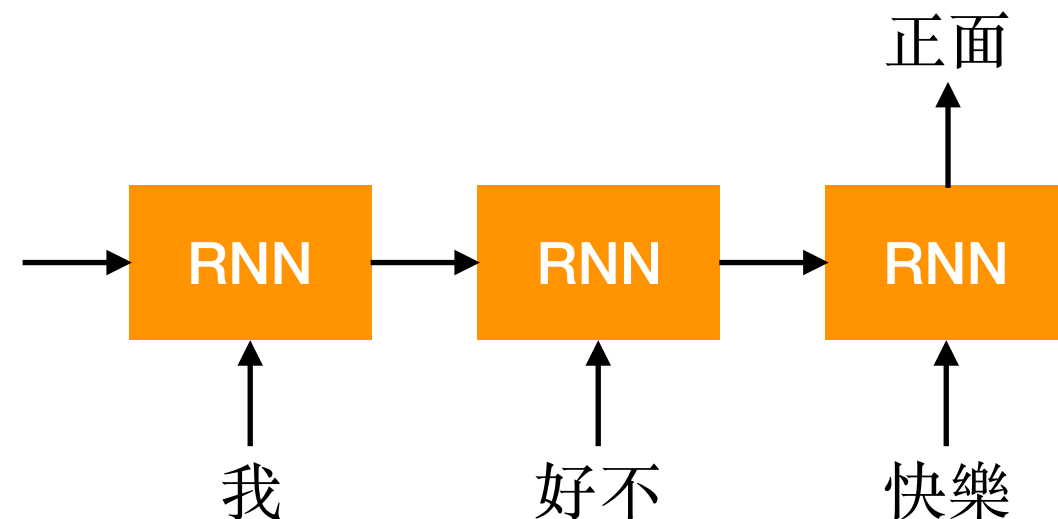
能夠處理輸入與輸出固定長度之序列任務

Many to one :

- Sentence classification

輸入：我好不快樂。

輸出：正面v負面

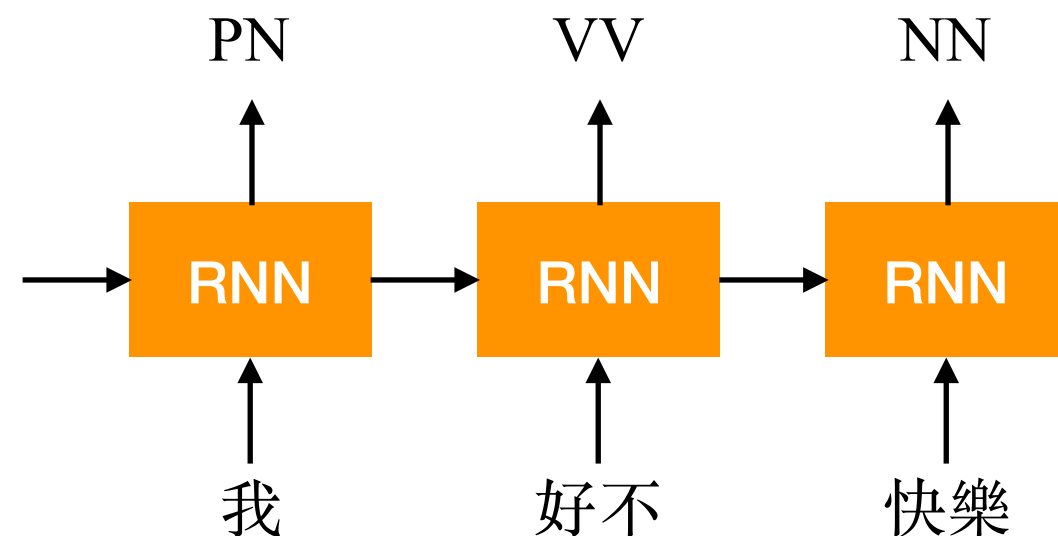


Many to many:

- POS tagging (part-of-speech tagging)

輸入：我好不快樂

輸出：PN/VV/NN (我/好不/快樂)



- Stock price prediction

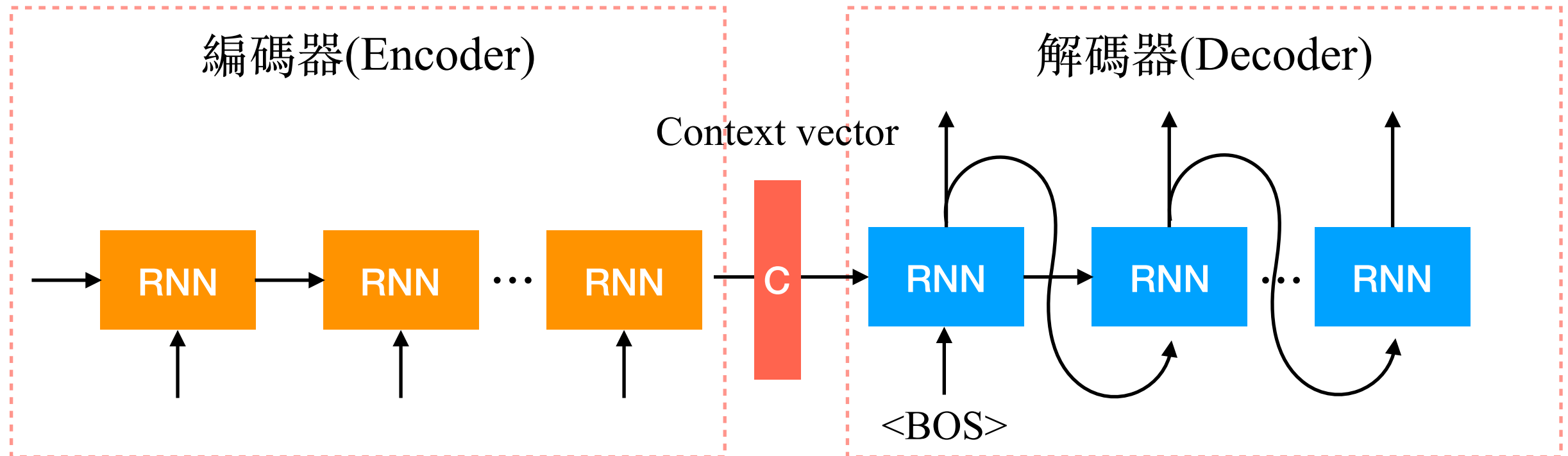
輸入：前三天股價

輸出：後三天股價

Seq2seq(Sequence to Sequence)

Seq2seq model:

能夠處理輸入與輸出不固定長度之序列任務



應用場景：

- Neural Machine Translation

輸入：Taiwan is a beautiful country.

輸出：台灣是一個美麗的國家。

- Chatbot

輸入：台灣是一個美麗的國家。

輸出：當然。

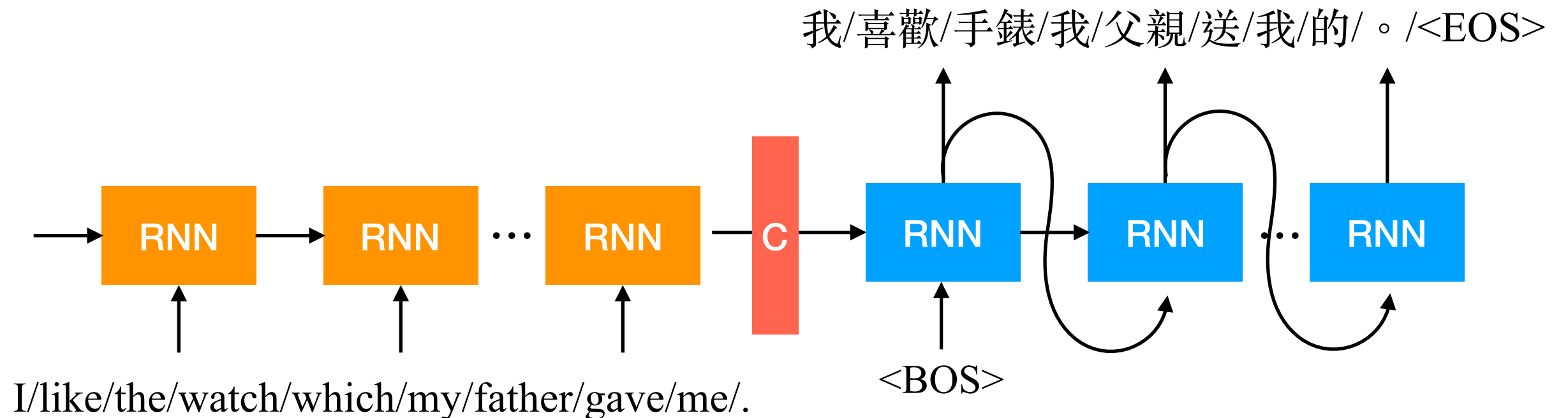
Seq2seq(Sequence to Sequence)

輸入：I like the watch which my father gave me.

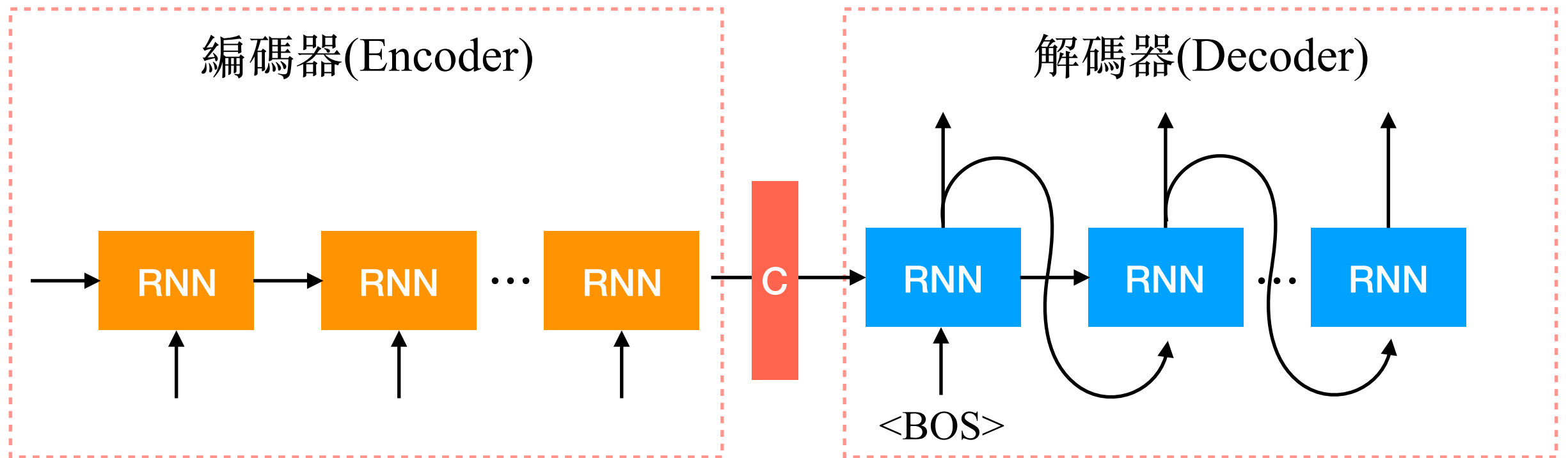
- Hard-Attention：我喜歡手錶我父親送我的。

輸出：

- Soft-Attention：我喜歡我父親送我的手錶。



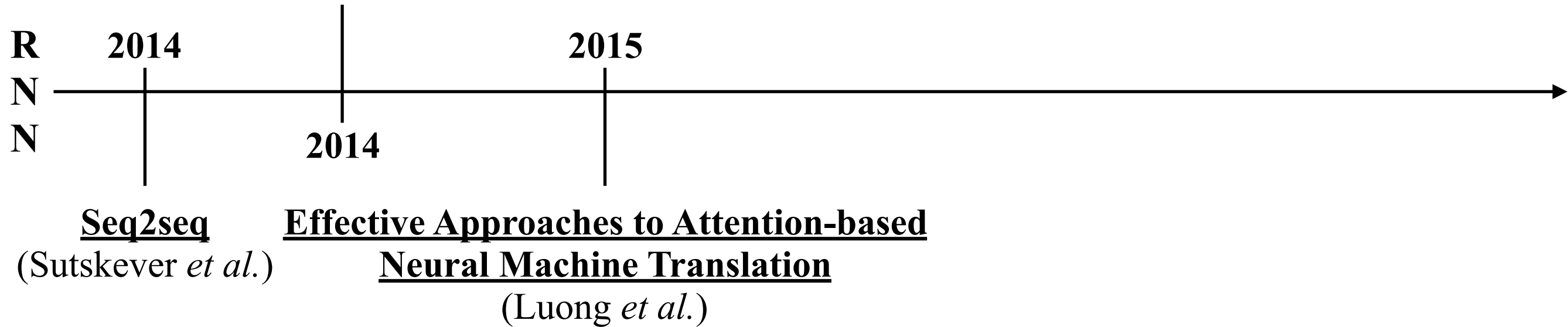
Seq2seq(Sequence to Sequence)



- **長期依賴(Long-term dependency)：**
當序列越來越長的時候，越前後的特徵會被RNN所遺忘。
- **Context vector：**
對於編碼器來說，不管輸入序列有多長，最終都會被一個context vector(向量)來表示。

對於解碼器來說，只有第一個RNN會接收到編碼器傳遞過來的資訊(context vector)，之後會漸漸的遺忘。

**Neural Machine Translation by Jointly
Learning to Align and Translate**
(Bahdanau *et al.*)

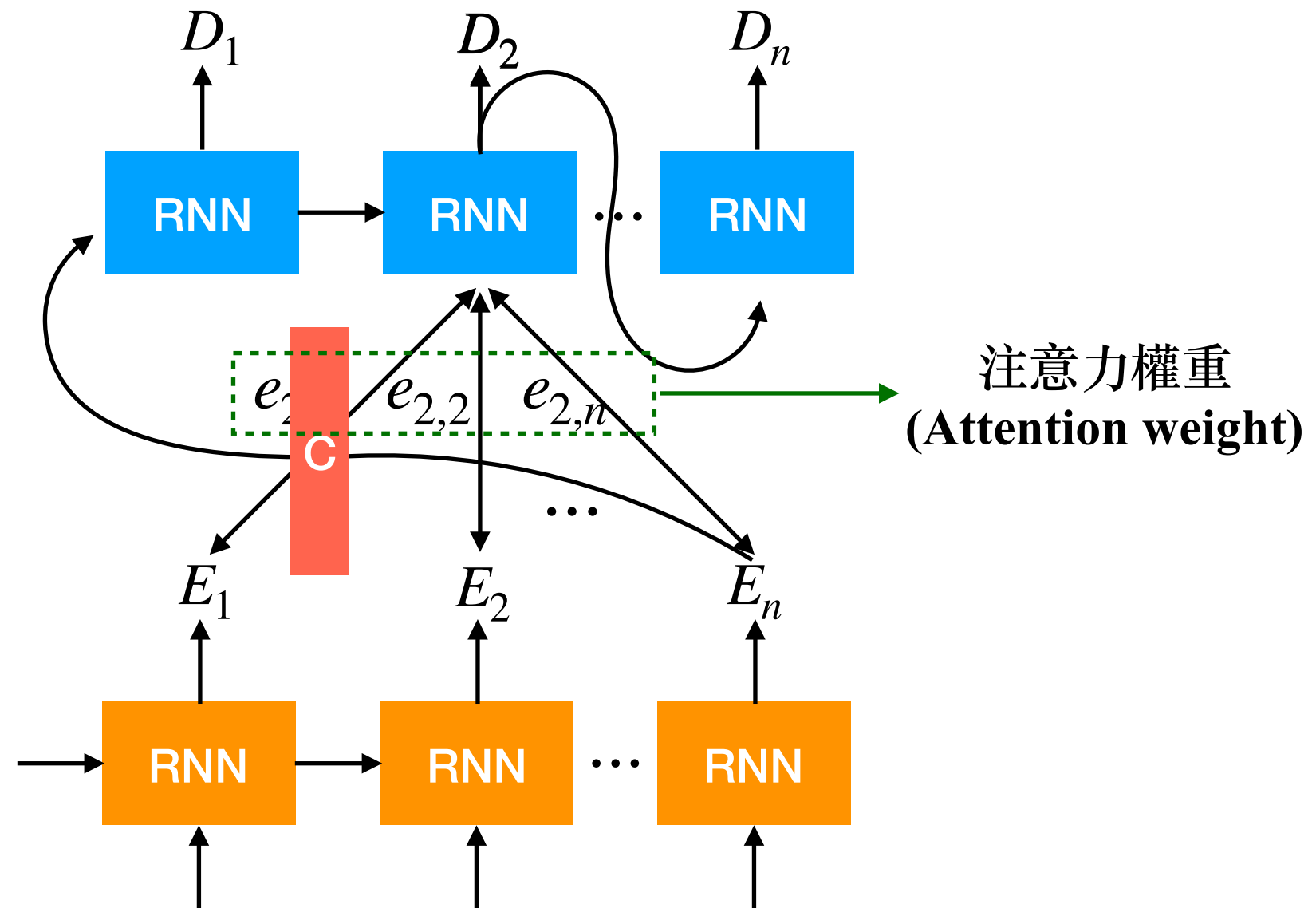


- Seq2seq：首次將類神經網路應用在翻譯任務(Neural Machine Translation)。
- Attention：設計網路結構時，考慮人類在翻譯時的思考邏輯(Soft-Attention)。

Seq2seq with Attention mechanism

解碼器 :

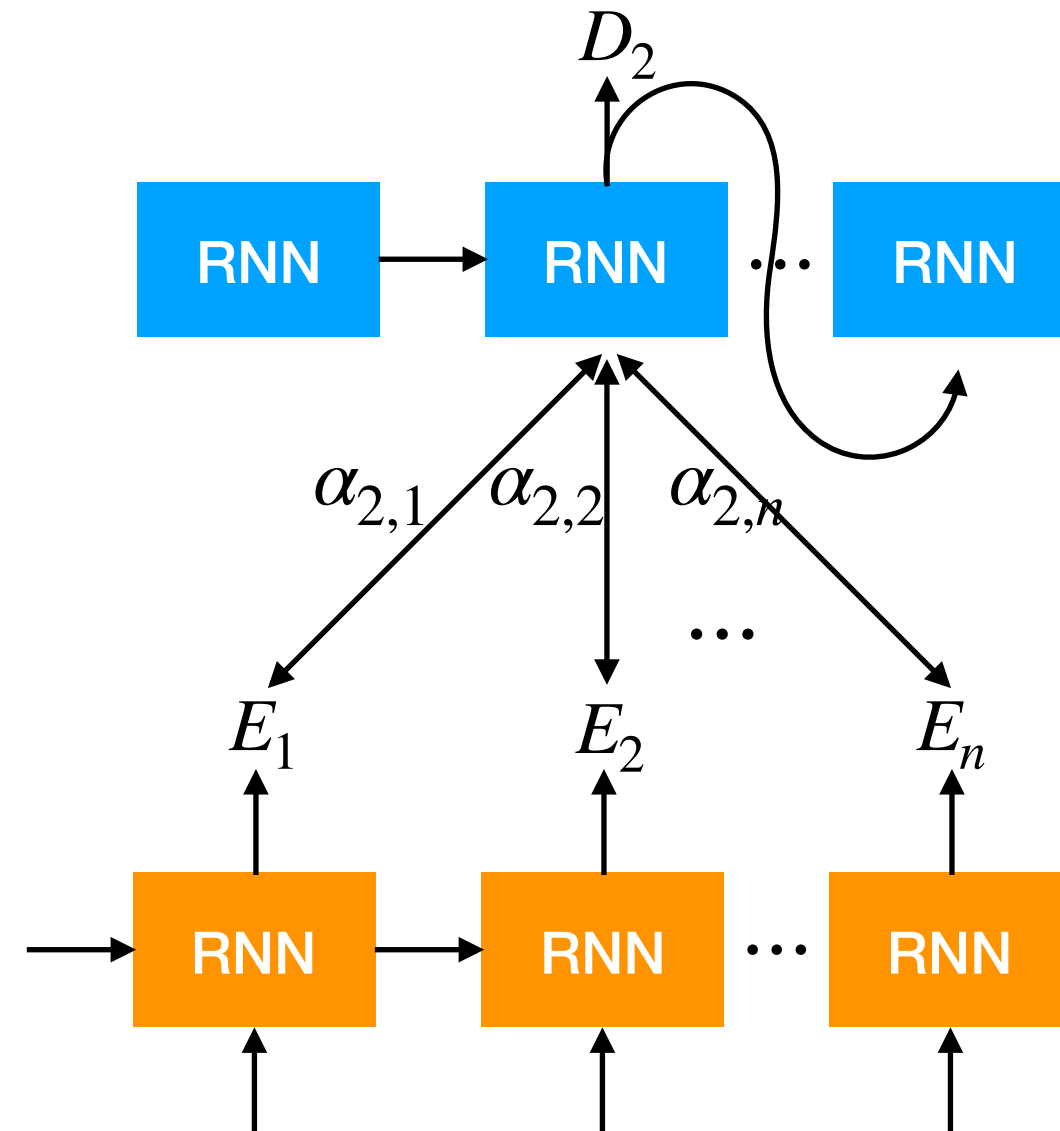
編碼器 :



$$e_{i,j} = f(E_i, D_j) = \begin{cases} E_i D_j^T, & \text{dot} \\ E_i W D_j, & \text{general} \\ \tanh(w^T (E_i W_E + D_j W_D)), & \text{feedforward} \end{cases}$$

Seq2seq with Attention mechanism

解碼器 :



編碼器 :

$$1. \alpha_{ij} = \text{softmax}(e_{ij} | e) = \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})}$$
$$2. c_i = \sum_{j=1}^n a_{ij} E_i$$

Seq2seq with Attention mechanism

Seq2seq:

- **Context vector :**

對於編碼器來說，不管輸入序列有多長，最終都會被一個context vector(向量)來表示。

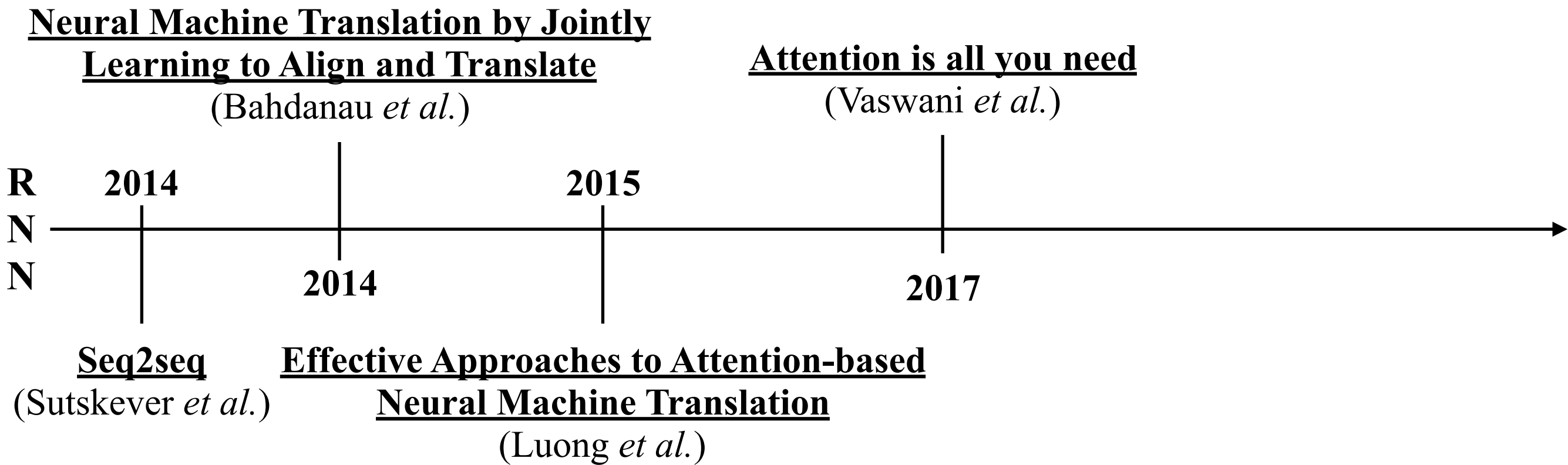
對於解碼器來說，只有第一個RNN會接收到編碼器傳遞過來的資訊(context vector)，之後會漸漸的遺忘。

Seq2seq with Attention mechanism:

- **Context vector :**

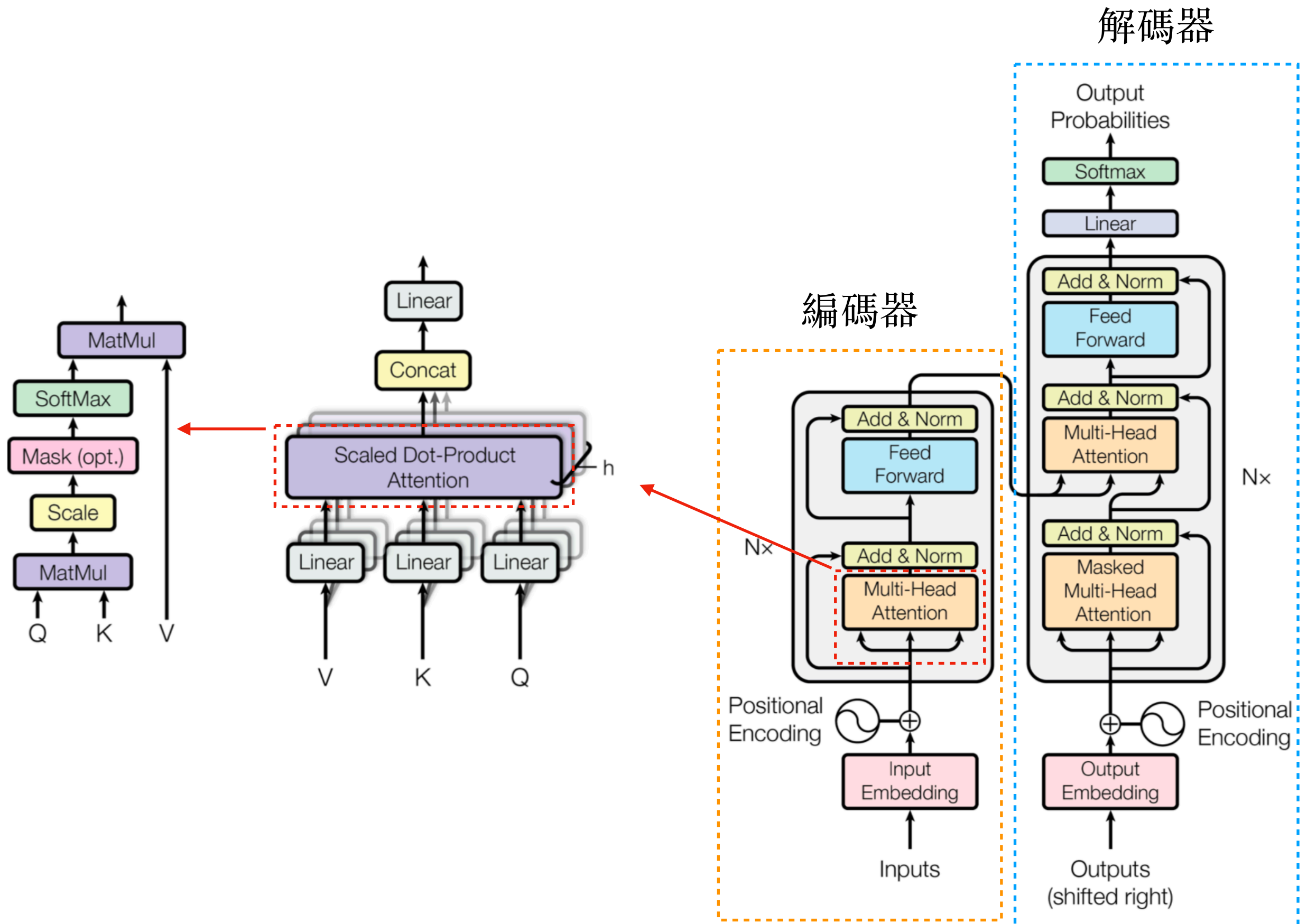
對於編碼器來說，每個時間點的資訊特徵都會被考慮到。

對於解碼器來說，每個時間點都能夠去選擇哪些資訊重要哪些資訊不重要。

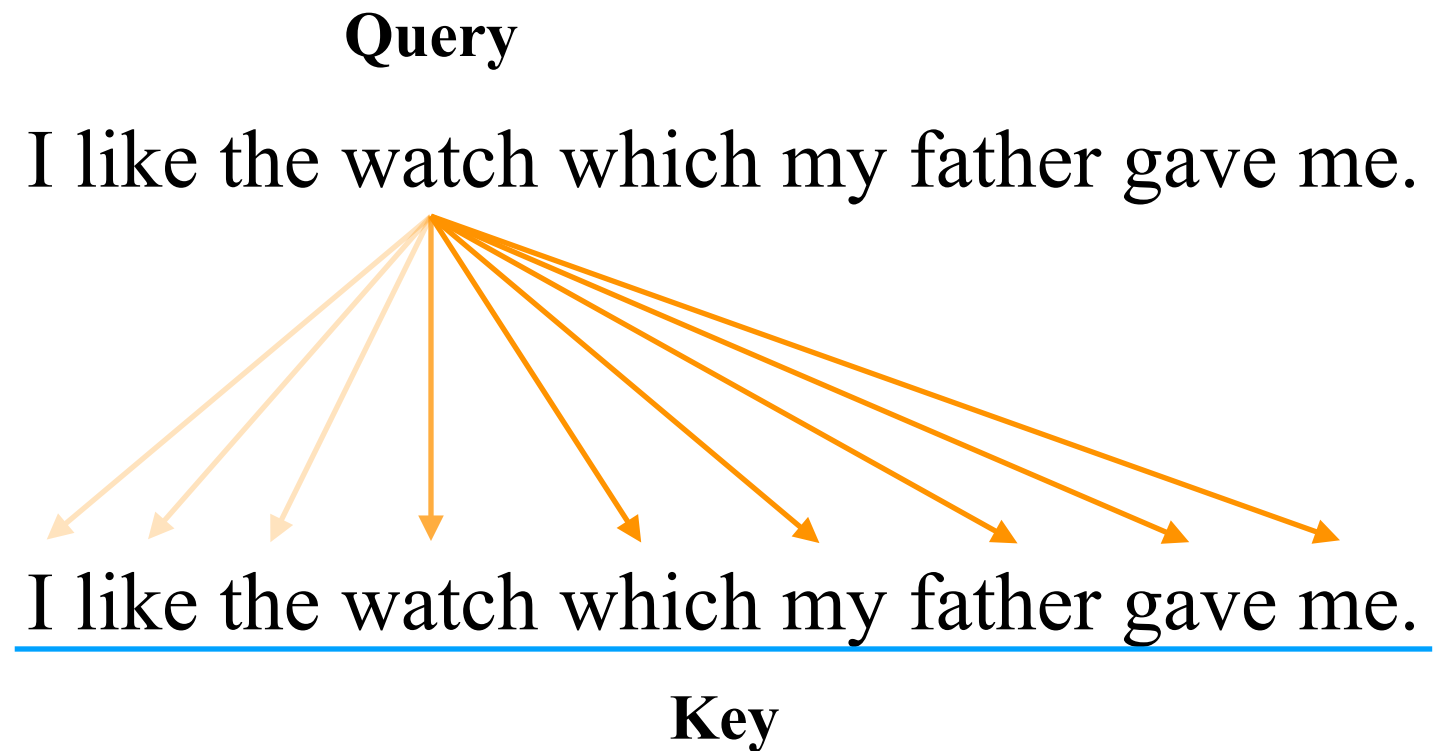
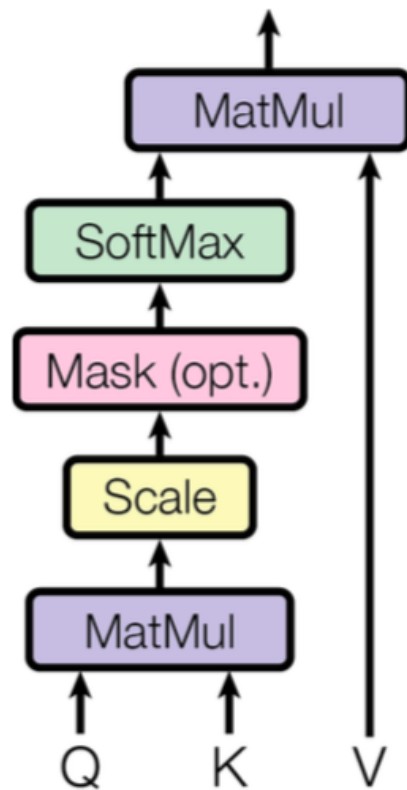


- Seq2seq：首次將類神經網路應用在翻譯任務(Neural Machine Translation)。
- Attention：設計網路結構時，考慮人類在翻譯時的思考邏輯(Soft-Attention)。
- Transformer：克服遞迴神經網絡(RNN)的缺點，大幅度提升翻譯水準。

Transformer



Transformer

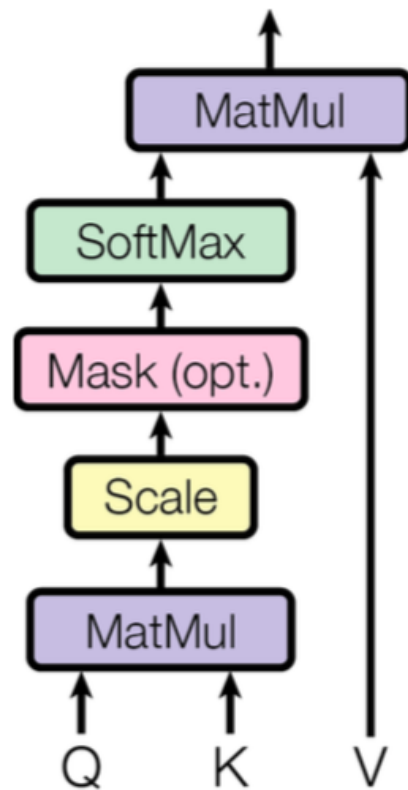


- **Q(Query)**：要搜索重要性的字(watch)。
- **K(Key)**：待搜索重要性的字，對於每個字都有一個重要性 α 。
- **V(Value)**：搜索完之後進行整理的詞嵌入矩陣。

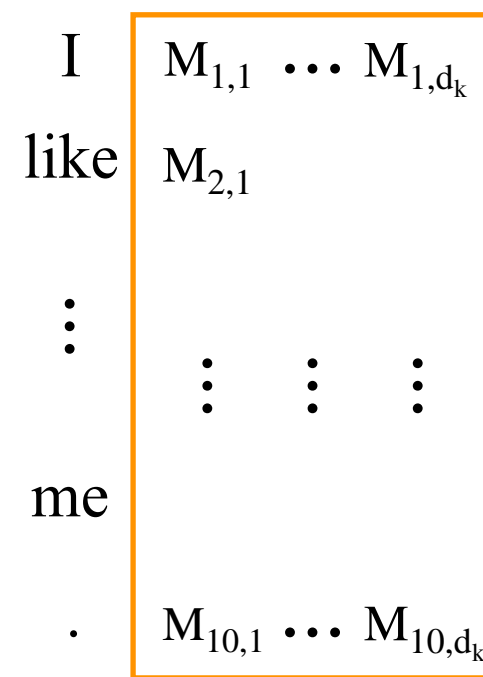
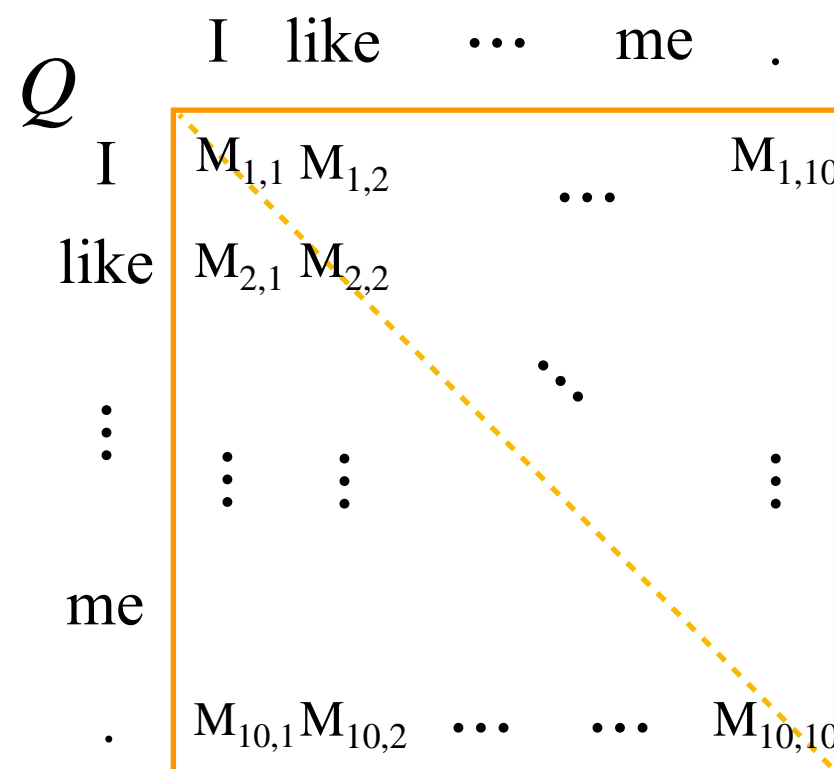
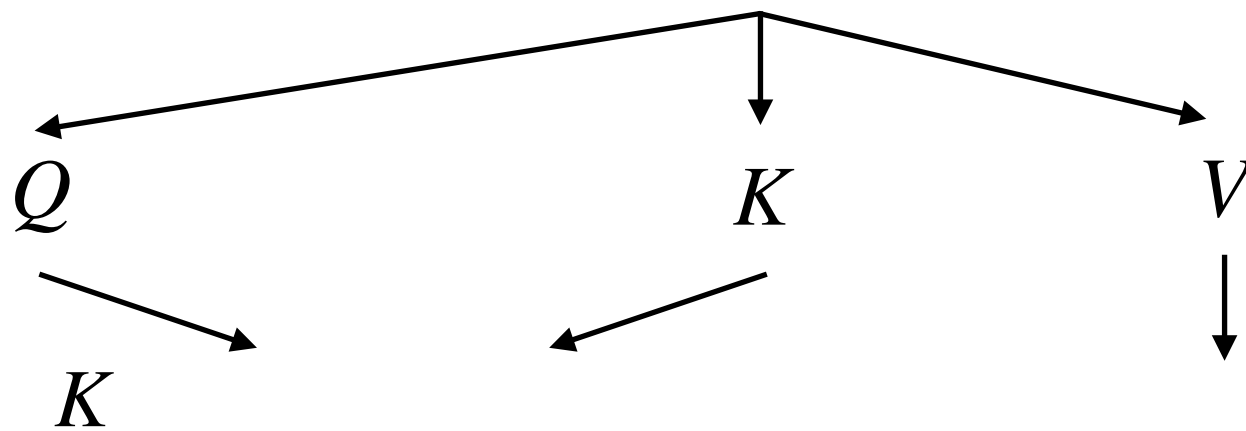
Transformer

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

I/like/the/watch/which/my/father/gave/me/.

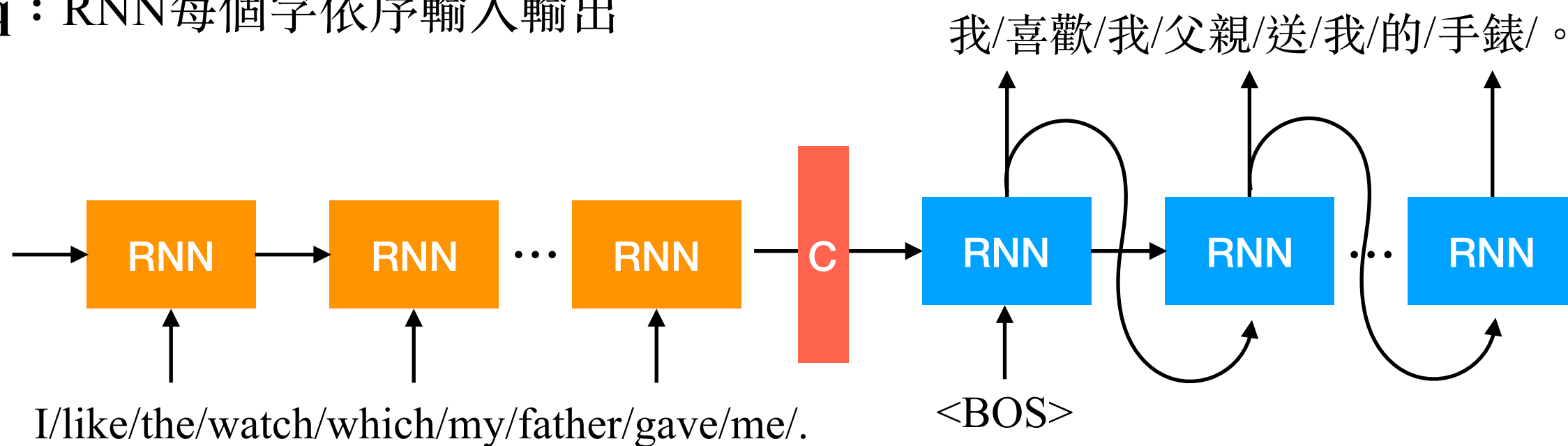


$$X = \{x_1, x_2, \dots, x_{10}\}, x_i \in R^{d_k}$$



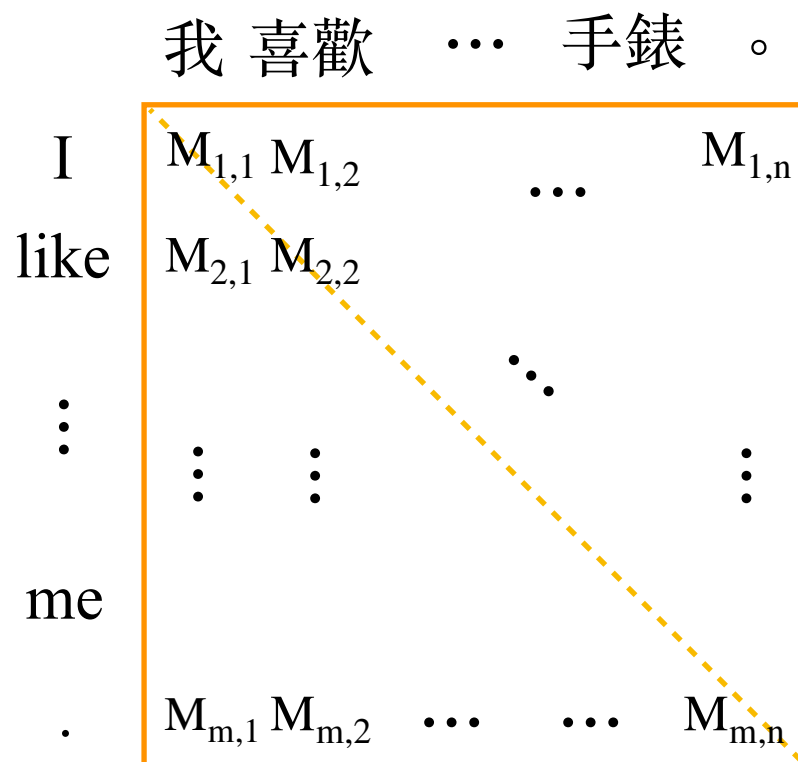
- Transformer：克服遞迴神經網絡(RNN)的缺點，大幅度提升翻譯水準。

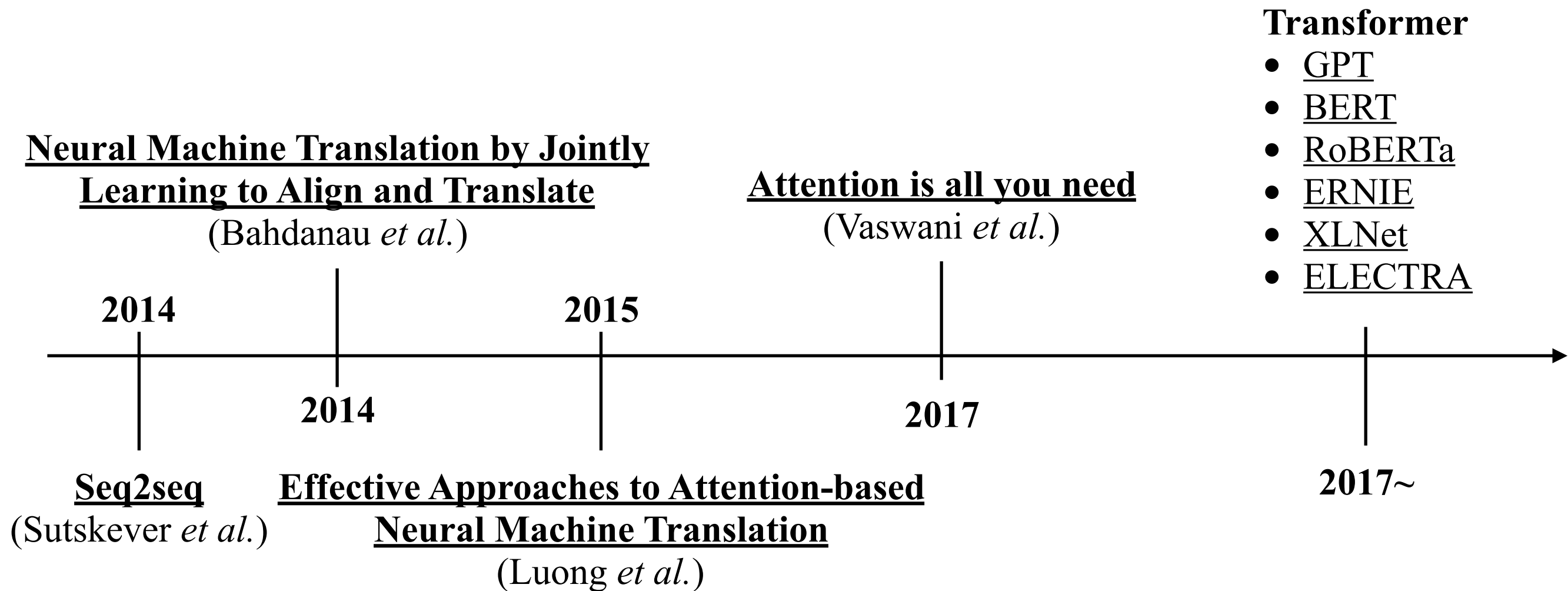
Seq2seq：RNN每個字依序輸入輸出



Transformer：

一次輸入所有的字進行Attention
(但沒有考慮到相鄰字之間的關係)





- Seq2seq：首次將類神經網路應用在翻譯任務(Neural Machine Translation)。
- Attention：設計網路結構時，考慮人類在翻譯時的思考邏輯(Soft-Attention)。
- Transformer：克服遞迴神經網絡(RNN)的缺點，大幅度提升翻譯水準。

- **GPT** : AutoRegressive
- **BERT** : Masked Language Model + Next Sentence Prediction
- **RoBERTA** : 在更大量資料集上訓練
- **ERNIE** : 更改中文斷詞方式，大幅度提升中文任務表現
- **XLNet** : Permutation Modeling
- **ELECTRA** : Transformer + GAN