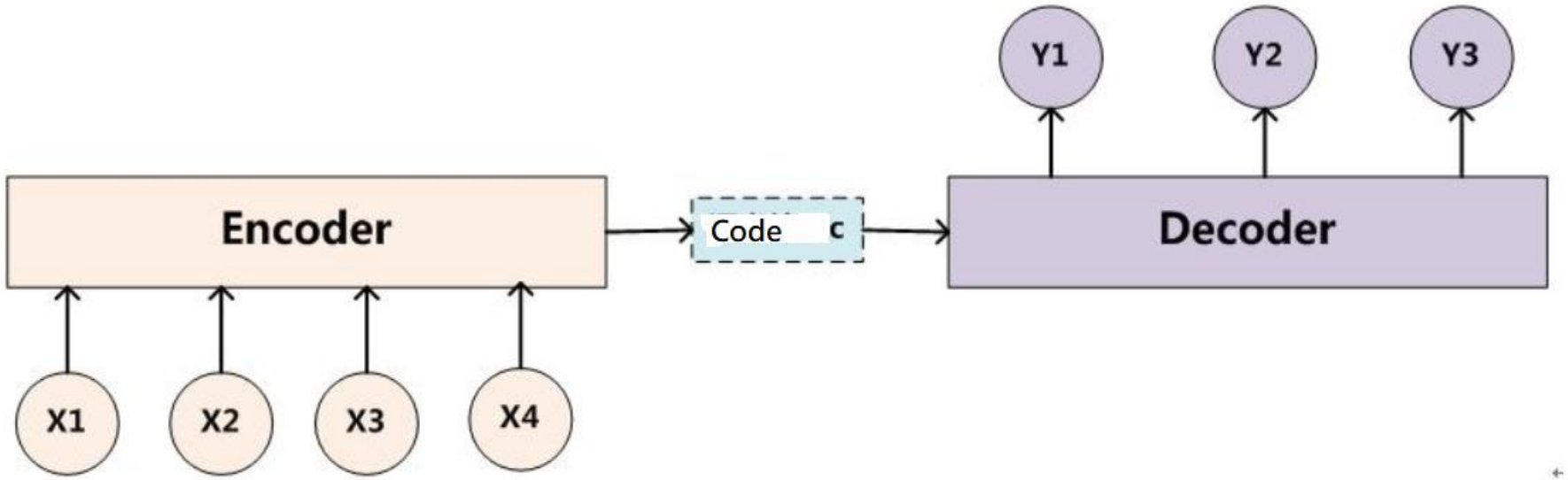
ATTENTION MODEL (AM)

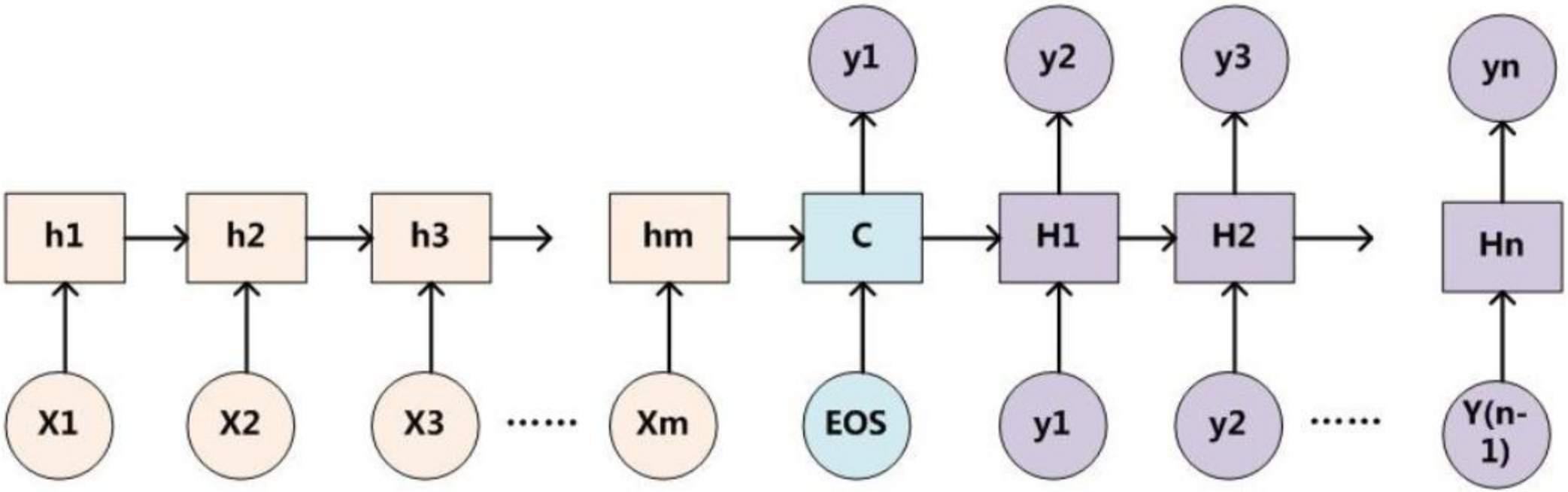
- **Why attention model**
- **Attention Model**
- **Bi-directional RNN**

- Adapted 張俊林

- Encoder-Decoder model for seq2seq translation



- The details—RNN encoder-decoder



- Problem: same code c is used for all yi
- Assume that

$$\mathbf{X} = \langle \mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_m \rangle$$

$$\mathbf{Y} = \langle \mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_n \rangle$$

- Intermediate code C
- depends on xi

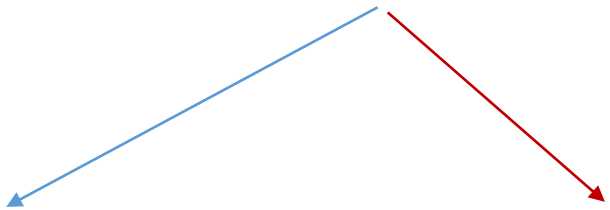
$$\mathbf{C} = \mathcal{F}(\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_m)$$

- To decode yi we use the same code C

$$y_i = G(C, y_1, y_2 \cdots y_{i-1})$$

- Of course with y1, y2, …, yi-1

- No attention in this case:



- To translate
- Tom chase Jerry → “湯姆” “追” “傑瑞”

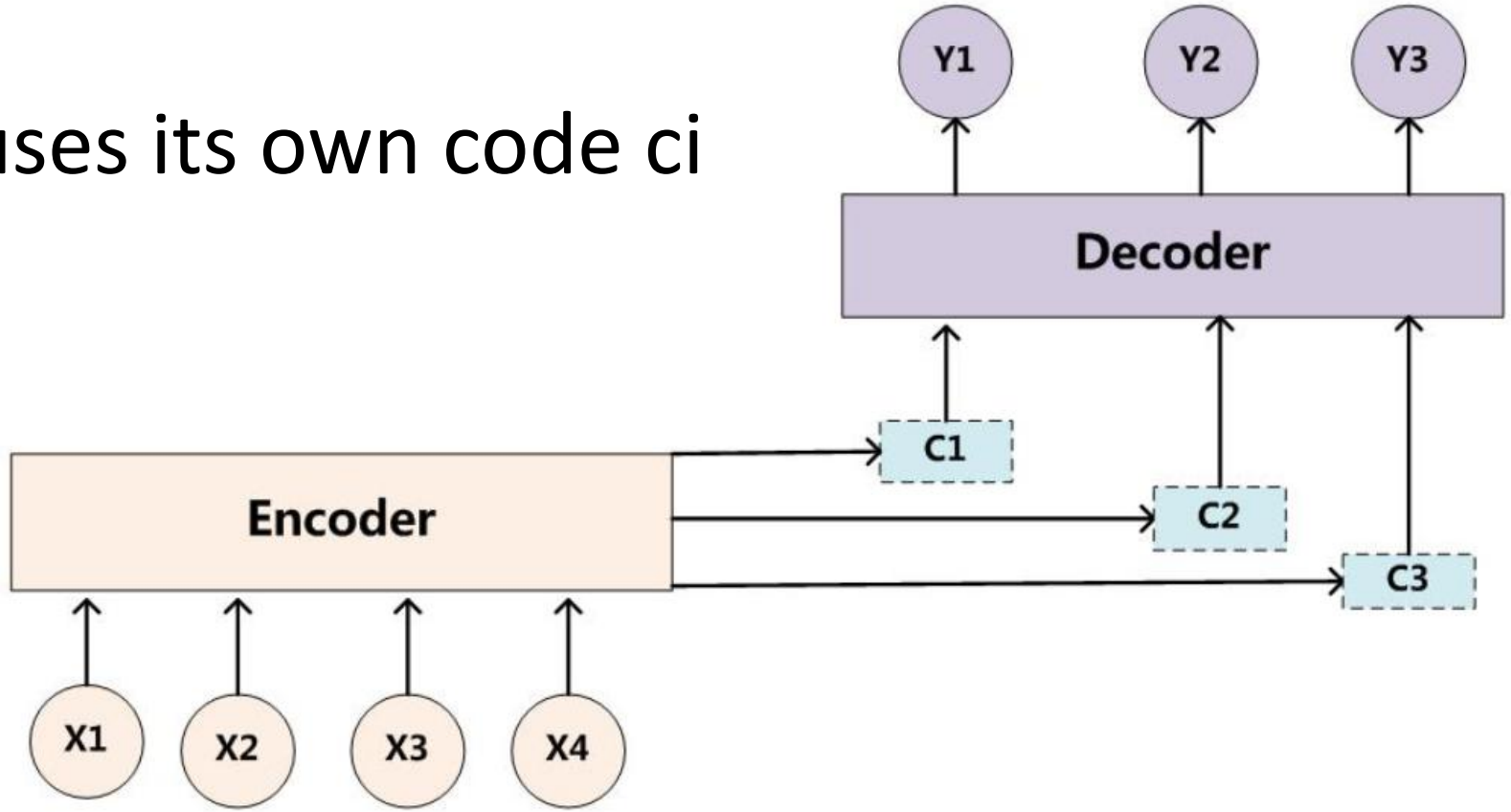
We should focus more on “Jerry” when translate it into “傑瑞“

So we want a probability distribution of importance such as

(Tom,0.3) (Chase,0.2)(Jerry,0.5)

Attention Model (AM)

Each yi uses its own code ci



That is,

$$y_1 = f1(C_1)$$
$$y_2 = f1(C_2, y_1)$$
$$y_3 = f1(C_3, y_1, y_2)$$

- How to find ci: weighted-sum of the component hidden state

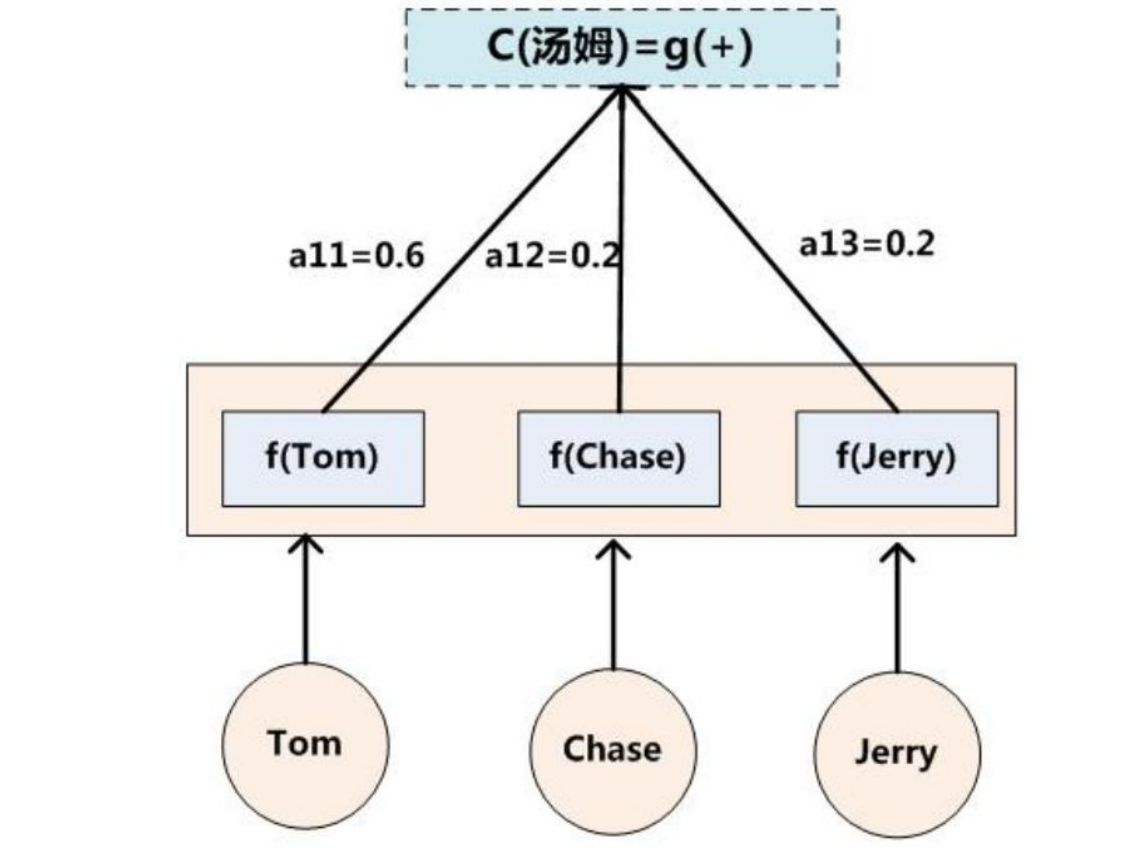
$$C_{汤姆} = g(0.6 * f2("Tom"), 0.2 * f2(Chase), 0.2 * f2("Jerry"))$$

$$C_{追逐} = g(0.2 * f2("Tom"), 0.7 * f2(Chase), 0.1 * f2("Jerry"))$$

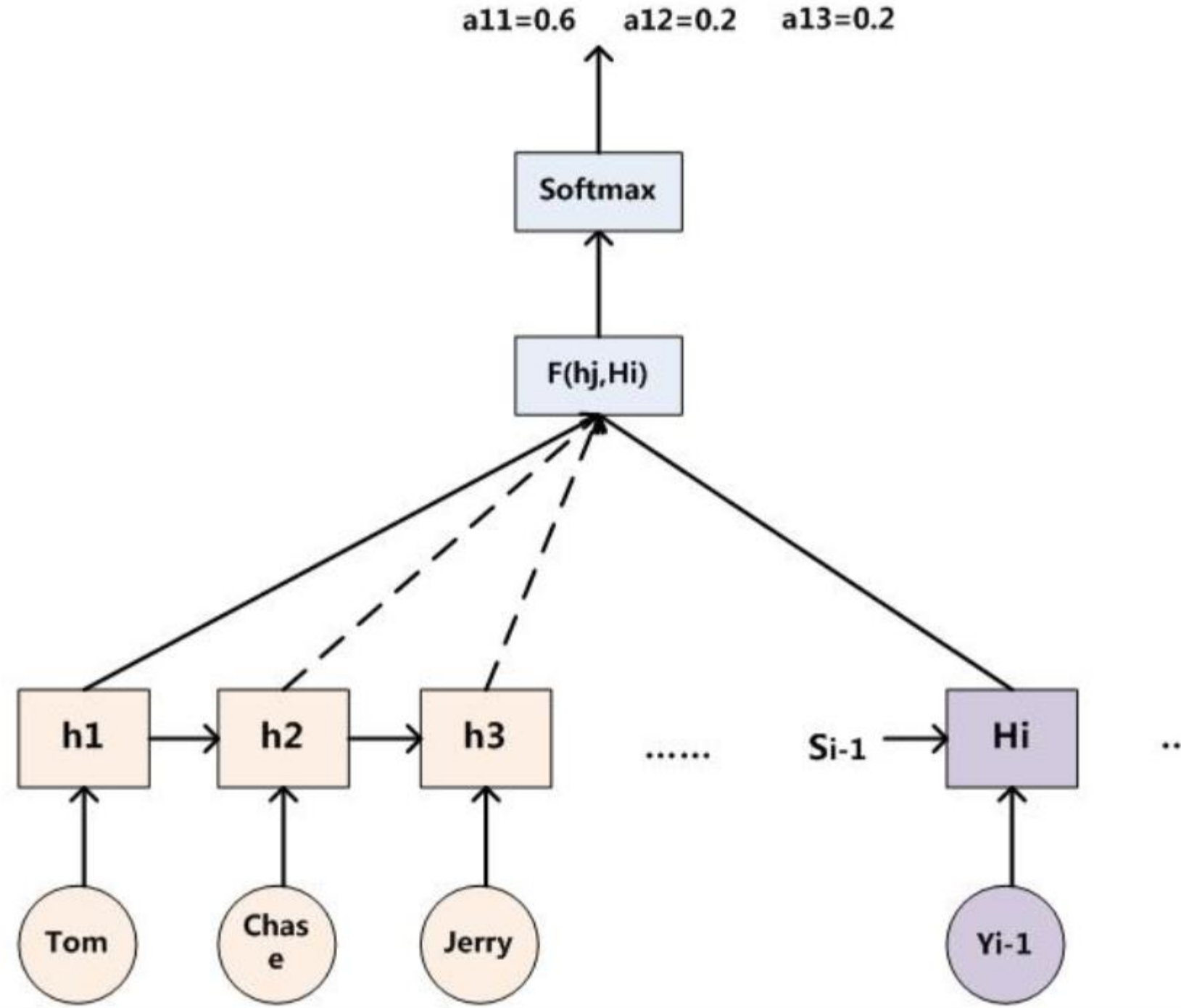
$$C_{杰瑞} = g(0.3 * f2("Tom"), 0.2 * f2(Chase), 0.5 * f2("Jerry"))$$

- f2 for RNN will encode xi into its hidden state hi; every ci has its own α_i

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

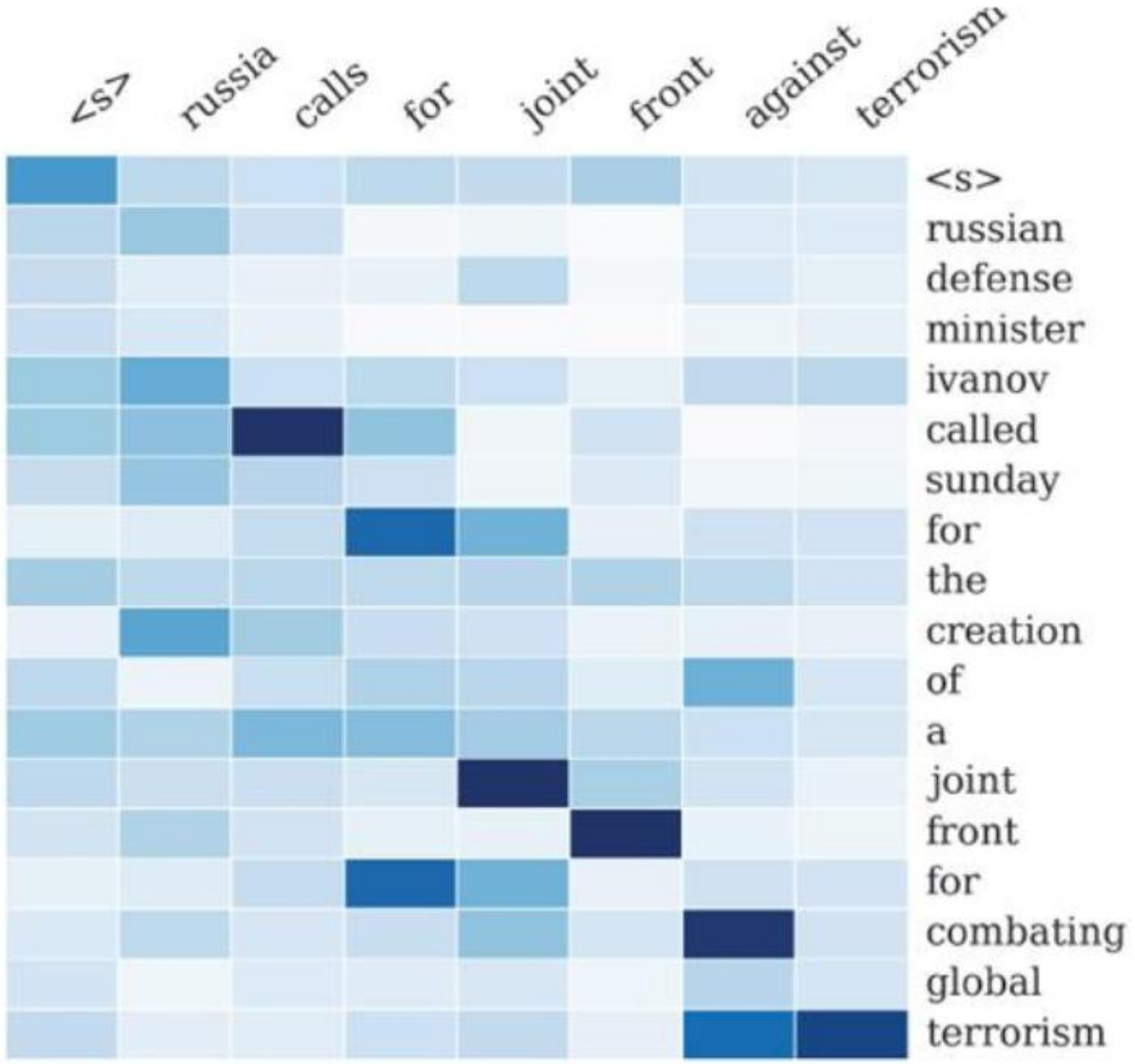

C(汤姆)=g(+)
a11=0.6
a12=0.2
a13=0.2
f(Tom)
f(Chase)
f(Jerry)
Tom
Chase
Jerry

- How to find attent probabilities?
- Use pair of hj and Hi and finc F
- F depends on different applications



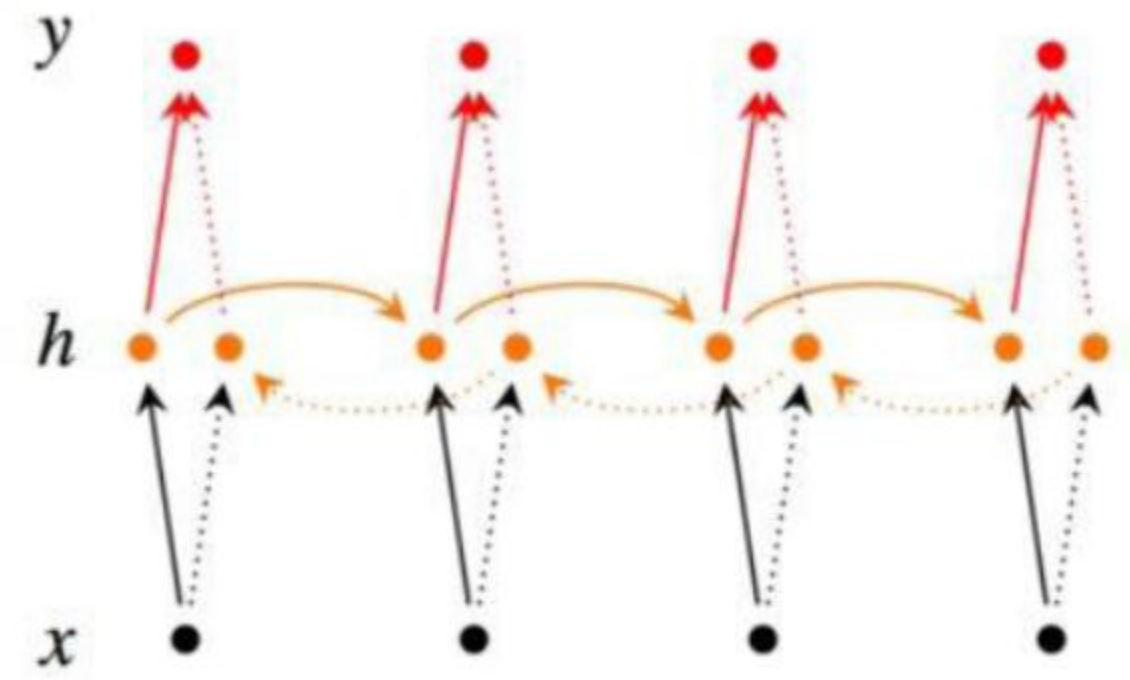
- Attention also called alignment (aligned on focus point)
- Attention link

Attention for summarization



Bi-directional RNN

Output depends on both prefix and suffix



- h^1 forward, h^2 backward

$$h_i^1 = f(U^1 * X_i + W^1 * h_{i-1})$$

$$h_i^2 = f(U^2 * X_i + W^2 * h_{i-1})$$

$$y_i = \text{softmax}(V * [h_i^1; h_i^2])$$

- From beginners to experts in deep learning