Statistical Machine Translation (SMT)

Core idea: learn a probabilistic model from data.

French → English

x y

$$\underset{y}{\text{argmax}}\ P(y|x)$$

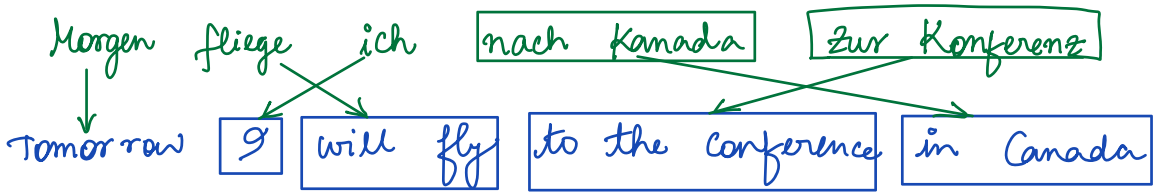
Use Bayes rule to break this down into:

$$\underset{y}{\text{argmax}}\ P(x|y)\,P(y)$$

$P(x|y)$: Translation Model

$P(y)$: Language Model

Alignment



Spurous words

	Le
Japan	Japon
Shaken	Secoue
by	par
two	deux
new	nouveaux

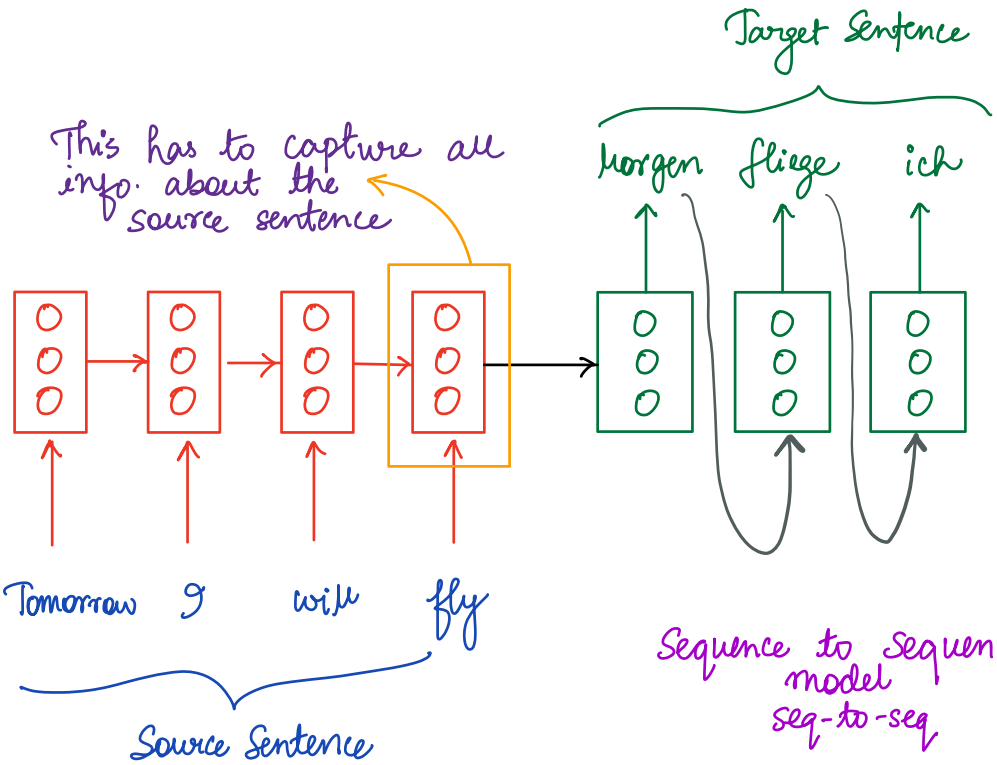
One to many

And	Le
the	programme
program	a
has	ete
been	mis

quakes seismes implemented ← en application

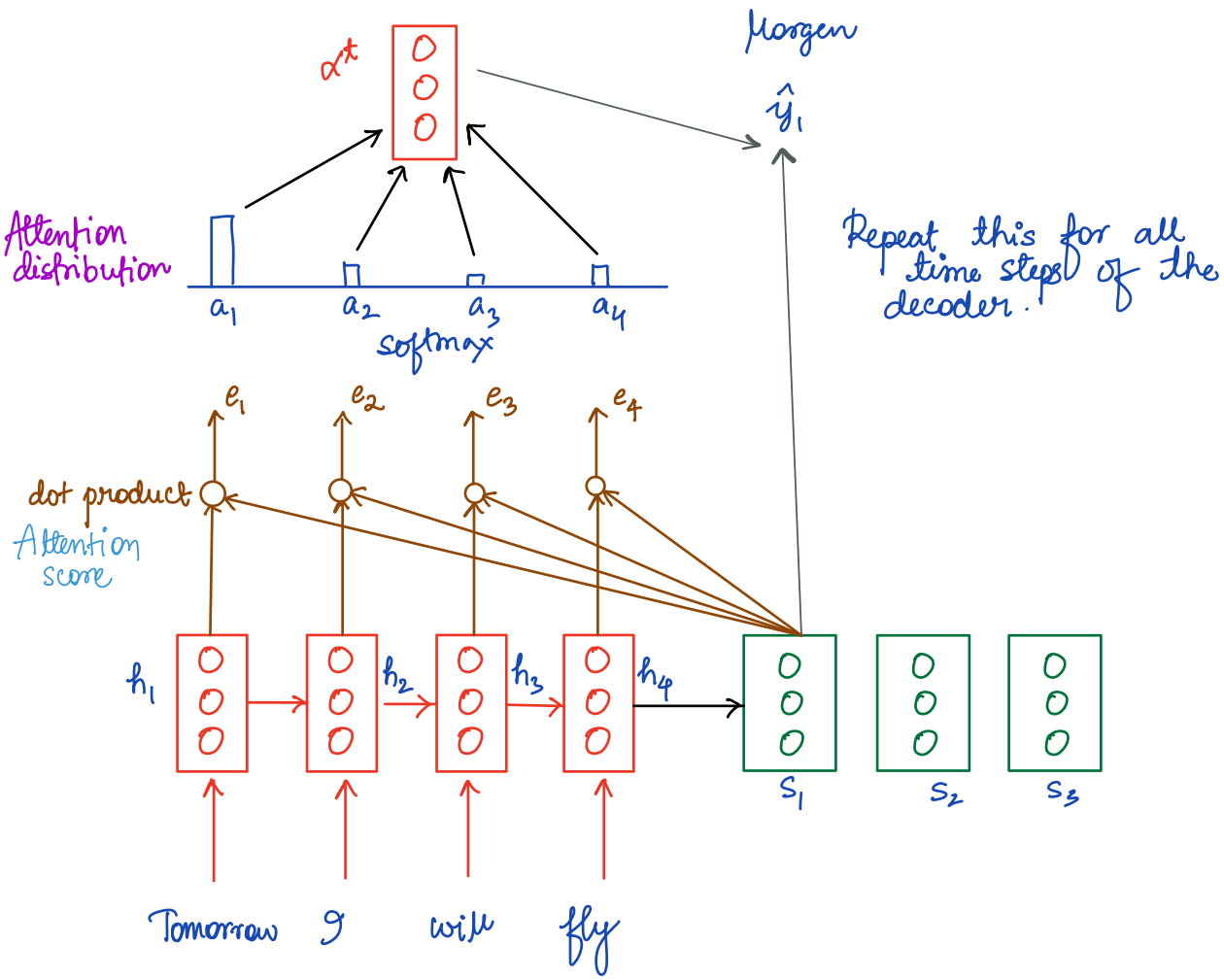
Alignment is complex.

Neural Machine Translation (NMT)



Attention provides a solution to this problem.

Idea: use direct connections to the encoder to focus on a specific part of the input.



at time step (decoder) t.

$$\left[s_t^T h_1 \;,\; s_t^T h_2 \;,\; \ldots \;,\; s_t^T h_N \right]$$

$$e^t = \left[e_1^t \quad e_2^t \;,\; \ldots\ldots \; e_N^t \right]$$

$$a^t = \text{softmax}(e^t)$$

$$a^t = \left[a_1^t \;,\; a_2^t \;,\; \ldots\ldots \;,\; a_N^t \right]$$

$$c^t = \sum_{t=1}^{N} a_t h_t$$

$$[c^t; s^t] \longrightarrow \hat{y}^t$$