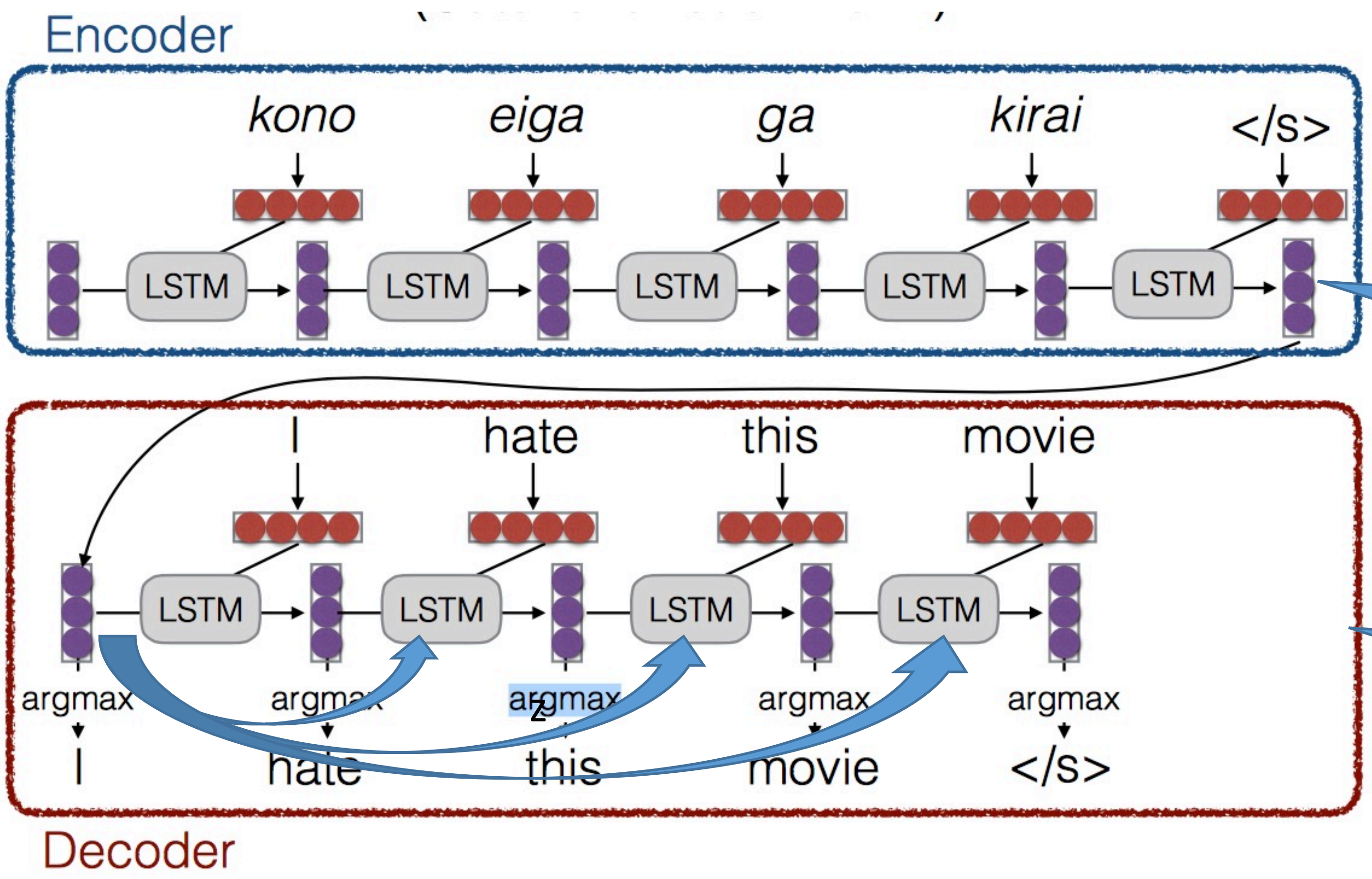
Attention in NLP

Presented by Yan Li

Outline

- Basic seq2seq model & its drawback
- Basic attention and its application in NMT
- Variants of attention

Seq2seq Models



A fixed length representation of the sequence as auxiliary information "X"

Decoder is a RNN based conditional language model

- Read whole sentience once and then translate it. Hard!
- "You can't cram the meaning of a whole sentence into a single vector!"— Ray Mooney

Outline

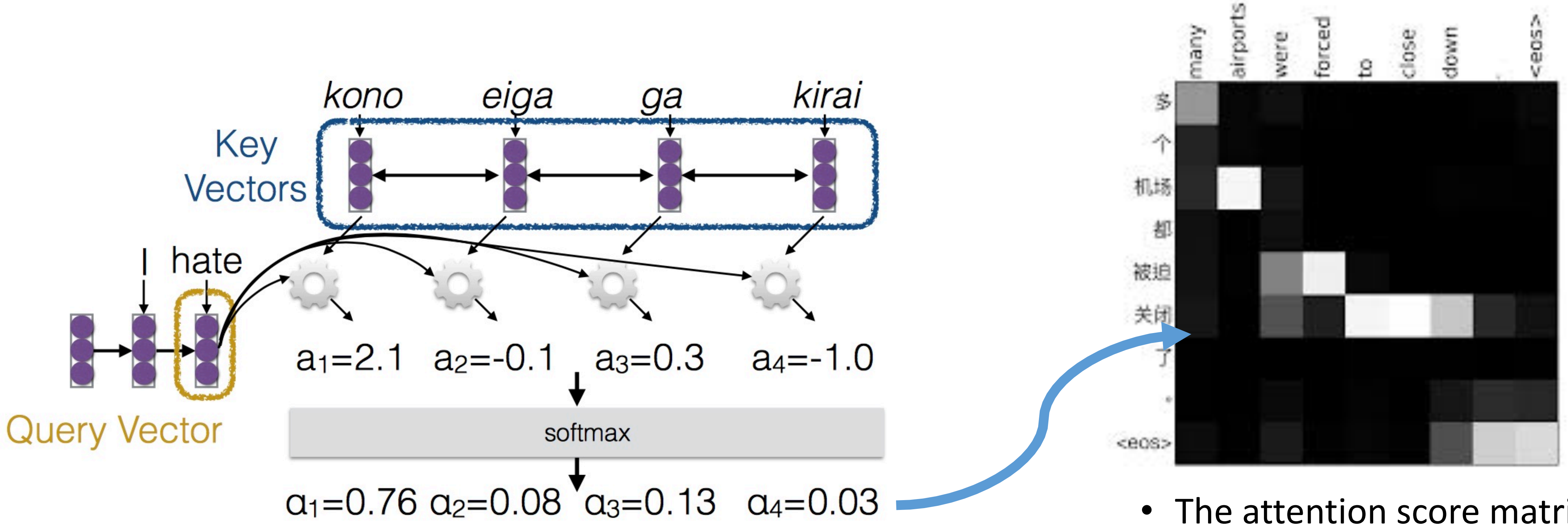
- Basic seq2seq model & its drawback
- Basic attention and its application in NMT
- Variants of attention

Attention Basic Idea

- Encode each word in the sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by “attention weights”
- Use this combination in picking the next word
- Read the source sentence and when you translate you can go back to re-read the source sentence again and again. Relatively easy!

Calculate Attention

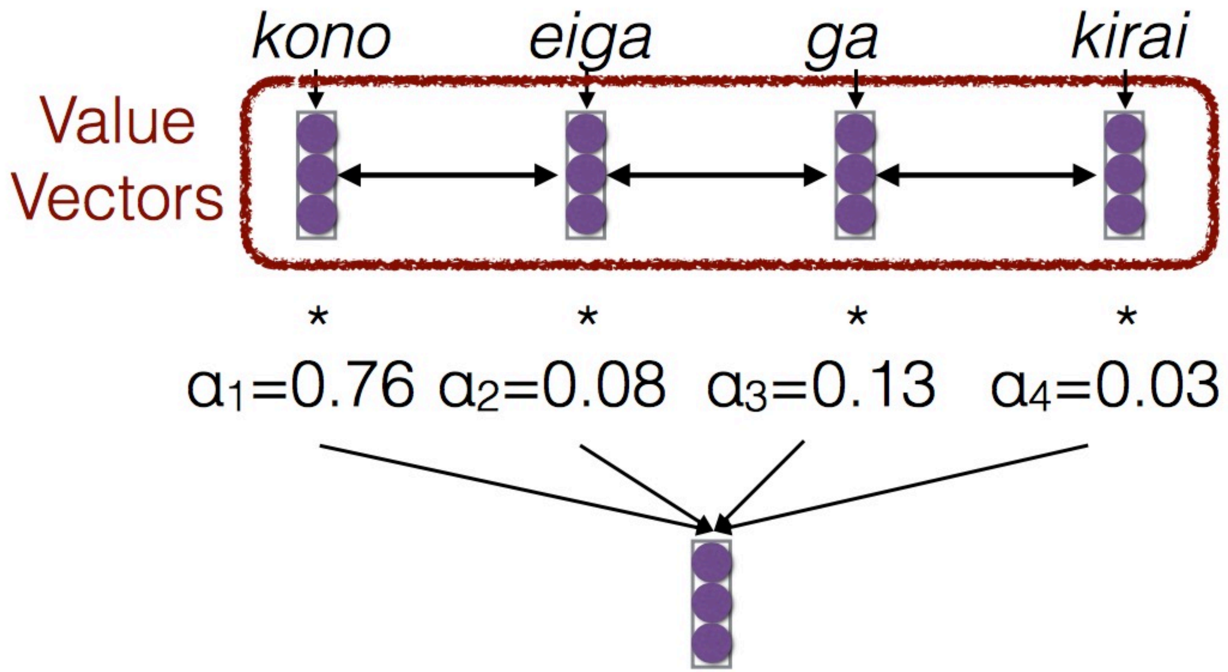
1. Use “query” vector (decoder state) and “key” vectors (all encoder states)
2. For each query-key pair, calculate weight
3. Normalize to add to one using softmax
4. Combine together value vectors (usually encoder states, like key vectors) by taking the weighted sum



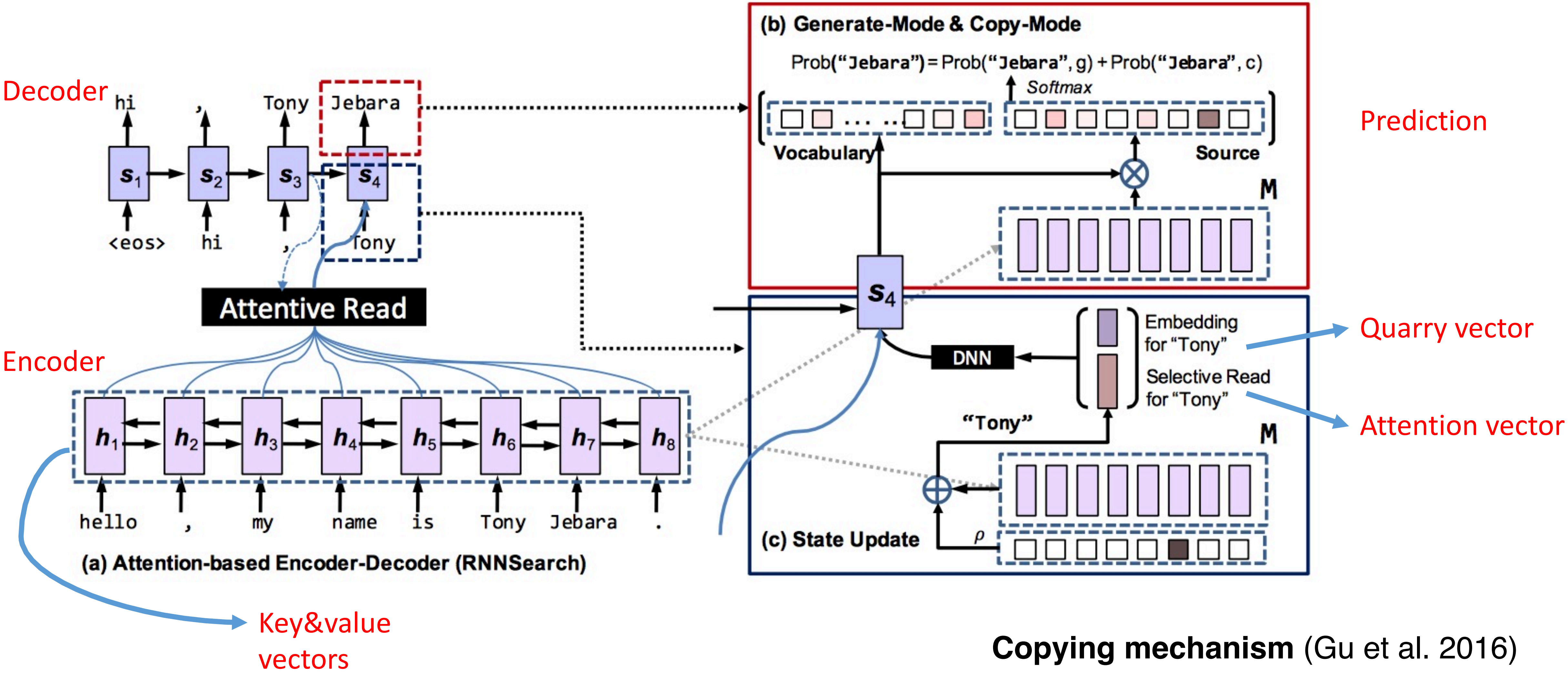
- The attention score matrix

Calculating Attention (cont.)

- Combine together value vectors (usually encoder states, like key vectors) by taking the weighted sum

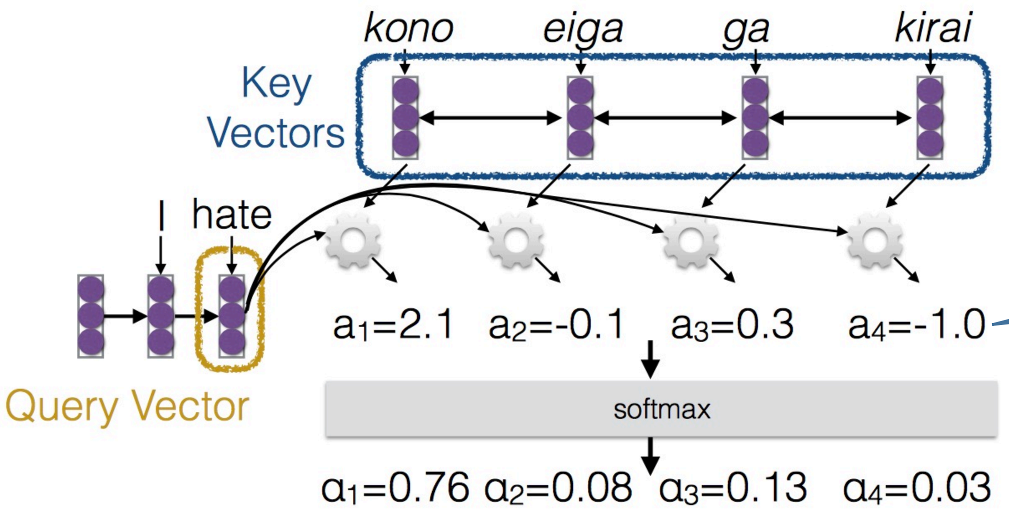


Attend to the source language



Copying mechanism (Gu et al. 2016)

Different Attention Score Functions



How to design the computational mechanism to compute attention score?

q is the query and ***k*** is the key

- **Multi-layer Perceptron:** $a(\boldsymbol{q}, \boldsymbol{k}) = \boldsymbol{w}_2^\intercal \tanh(W_1[\boldsymbol{q}; \boldsymbol{k}])$ (Flexible, often very good with large data)
- **Bilinear:** $a(\boldsymbol{q}, \boldsymbol{k}) = \boldsymbol{q}^\intercal W \boldsymbol{k}$
- **Dot Product:** $a(\boldsymbol{q}, \boldsymbol{k}) = \boldsymbol{q}^\intercal \boldsymbol{k}$ (No parameters, But requires sizes to be the same.)
- **Scaled Dot Product:** $a(\boldsymbol{q}, \boldsymbol{k}) = \frac{\boldsymbol{q}^\intercal \boldsymbol{k}}{\sqrt{|\boldsymbol{k}|}}$ (scale by size of the vector, this is because dot product increases as dimensions get larger)

Outline

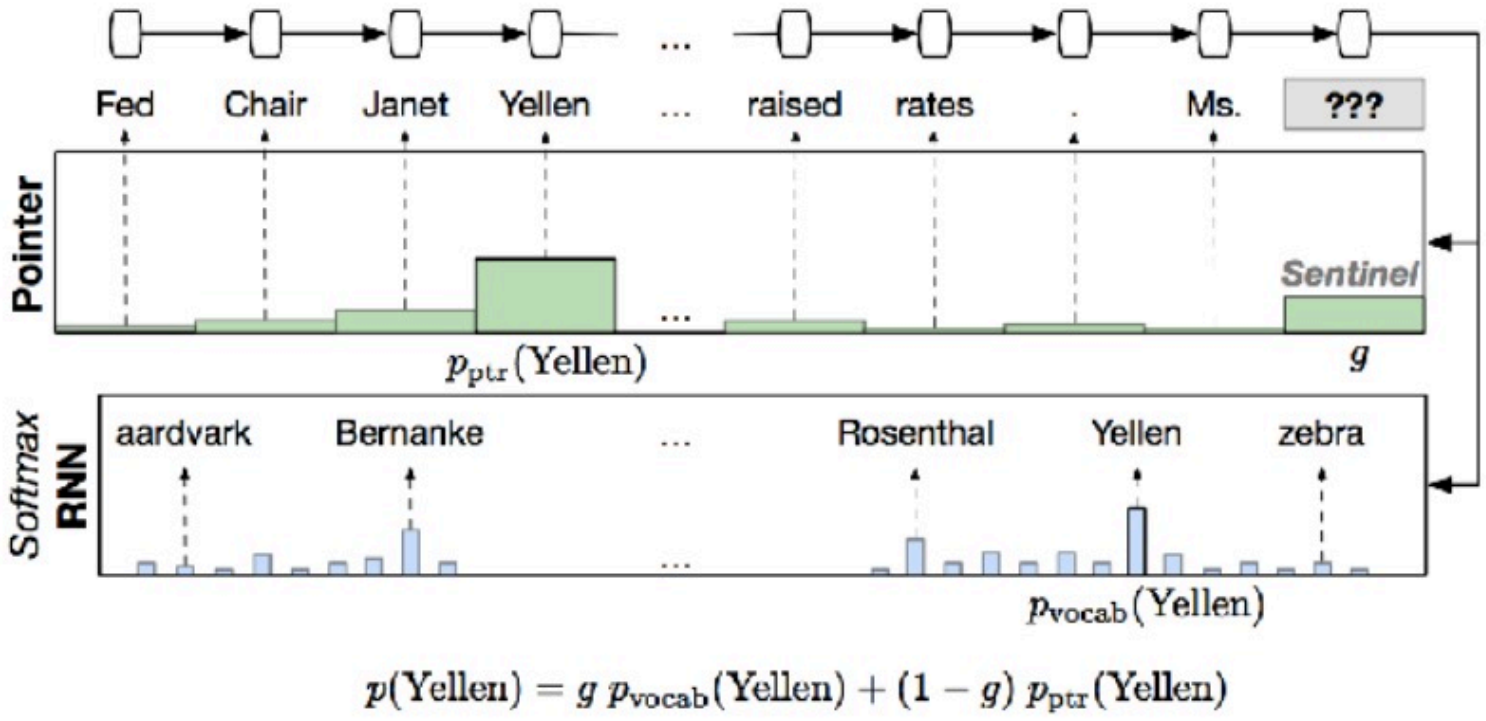
- Basic seq2seq model & its drawback
- Basic attention and its application in NMT
- Variants of attention

Attend to the previously generated things

- Key point: Read the previously generated words again and again while generate the new words.
- Reason: For example, in a long paragraph the key words might need many times.

Key vectors from previously generated words

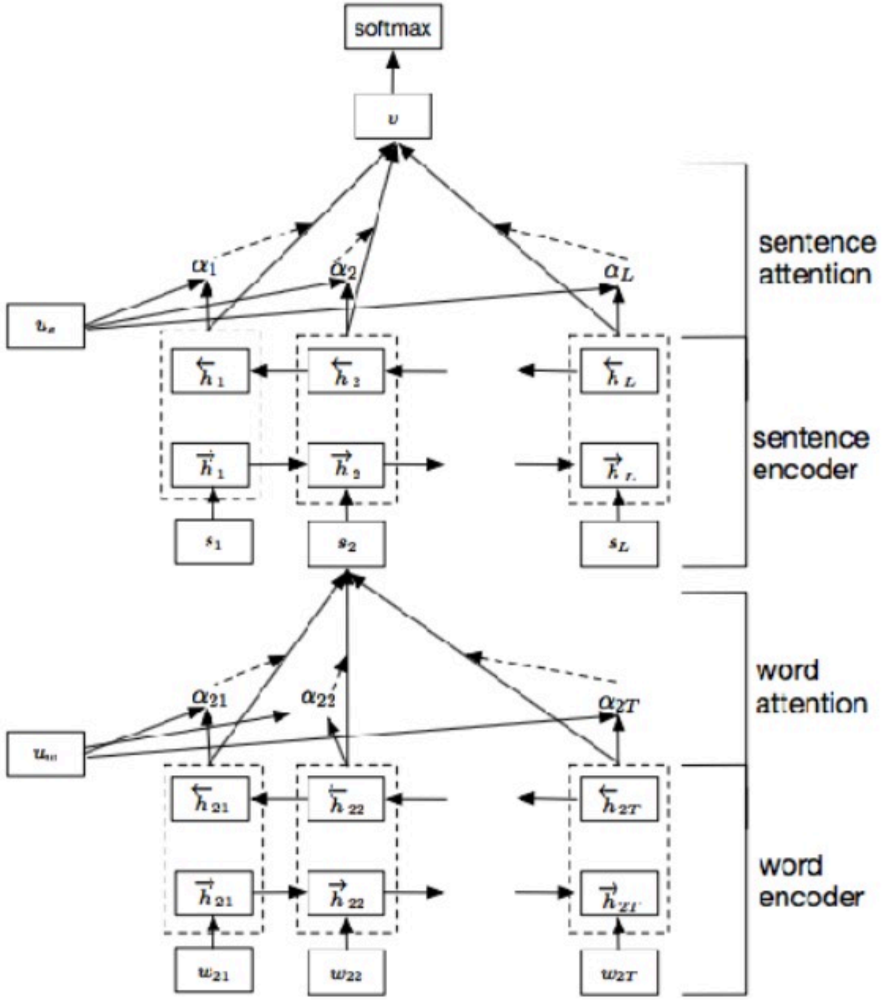
Key vectors from vocabulary



(Merity et al. 2016)

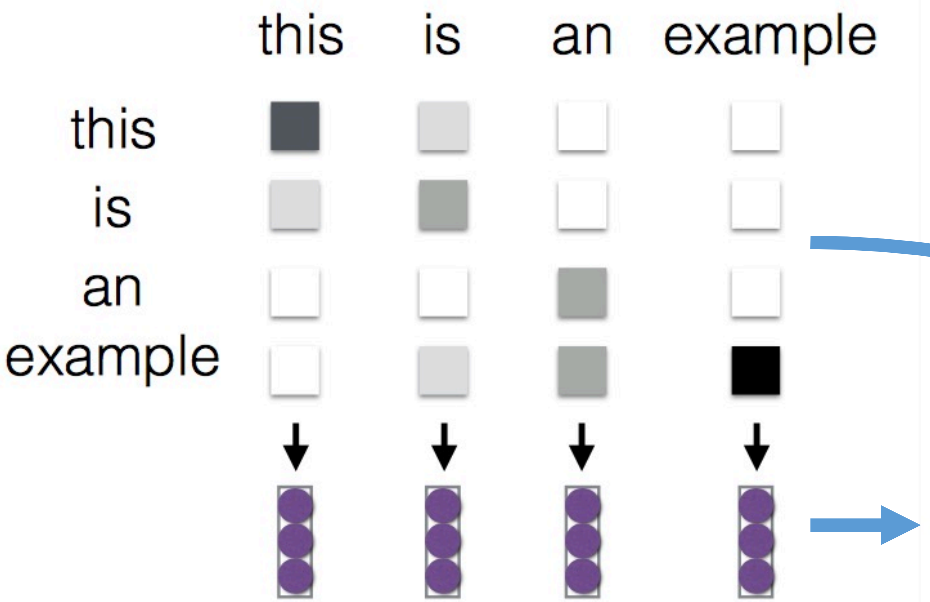
Hierarchical Structures

- Encode with attention over each sentence, then attention over each sentence in the document. (Yang et al. 2016)



Self Attention

- Each element in the sentence attends to other elements → context sensitive encodings! Incorporate the surrounding words information into the word embedding



$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V)$$

x_i is the original embedded matrix of i-th word

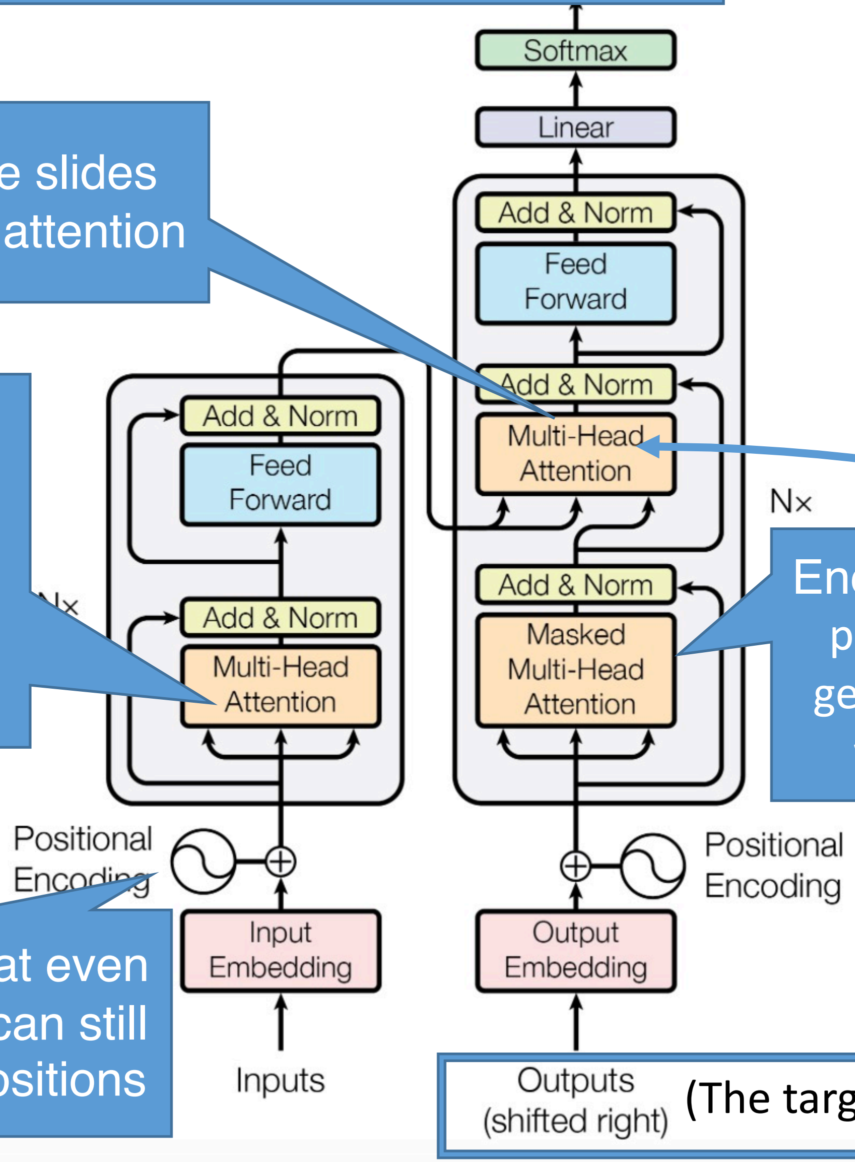
$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}}$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}}$$

- Comparing with Bi-RNN:
  - Much faster than Bi-RNN.
  - More direct than Bi-RNN

Each output element, z_i, is computed as weighted sum of a linearly transformed input elements

Attention Is All You Need (Vaswani et al. 2017)



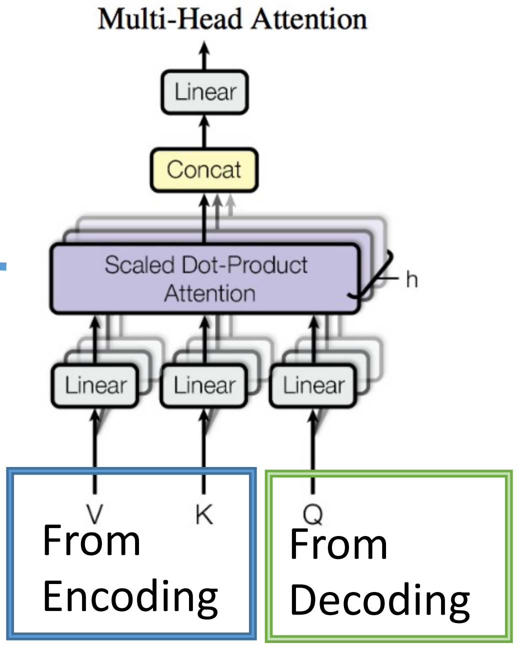
The above slides mentioned attention

Self-Attention over input sentence (encode source)

Encode the previous generated words

Make sure that even if no RNN, it can still distinguish positions

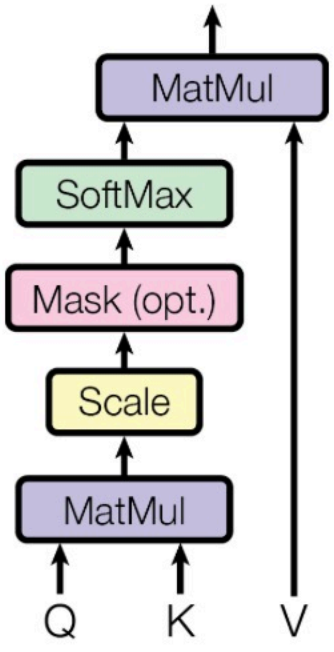
Every step is one training sample, in RNN the whole sentence is one sample, you have to backprob all sentence. Here each token is a sample, one don't need to backprob whole sentence.



This structure is used several times. In the arrowhead one the input from different part

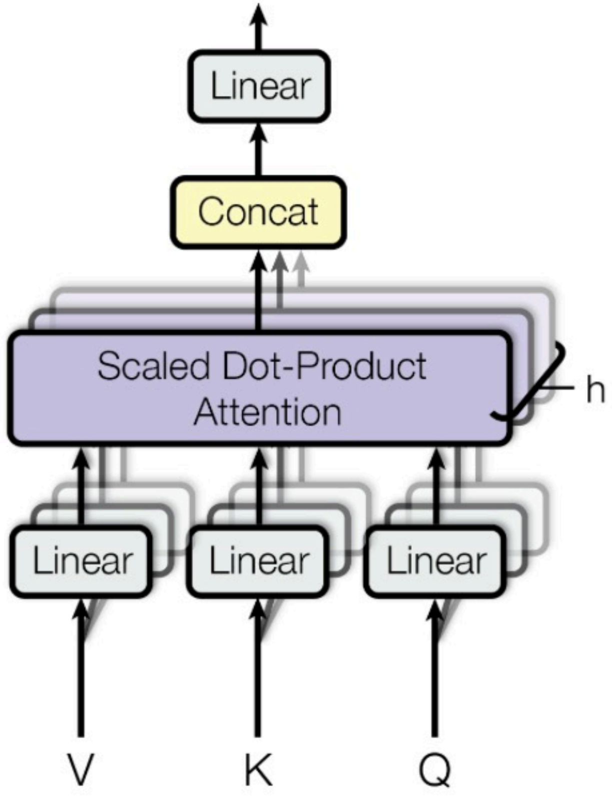
Important components of Attention Is All You Need

Scaled Dot-Product Attention



constitute

Multi-Head Attention



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

Q: query vector
K: Key vector
V: value vector

Idea: multiple attention "heads" focus on different parts of the sentence

- Positional Encoding

use sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

where pos is the position and i is the dimension.

感觉和傅立叶变换有相似的地方，把绝对位置表示为相对位置。