



# Natural Language Processing

Introduction to Data Science  
Spring 1404

Yadollah Yaghoobzadeh

# Agenda

- What is NLP?
- Why NLP is important?
- NLP applications (translation, sentiment analysis, summarization)
- Word embedding (word2vec)
- RNN (sequence processing)
- Attention mechanism, Transformers
- Language modeling (task definition)
- From LMs to general-purpose chatbots

# Goal

Comprehension and generation of **natural language**

# Natural language

- Languages that evolved naturally through human use
    - e.g., Spanish, English, Arabic, Hindi, etc.



# Machine translation

Google Translate

The screenshot shows the Google Translate interface. At the top, there are two tabs: "Text" (selected) and "Documents". Below the tabs, the source language is set to "ENGLISH - DETECTED" and the target language is "PERSIAN". A double-headed arrow icon indicates the bidirectional nature of the translation. The input text in English is "We are starting to learn artificial intelligence." The output text in Persian is "ما در حال یادگیری هوش مصنوعی هستیم." (Ma dar hal yadgir-e hosh makhnou'i hastim). There are icons for microphone, speaker, and edit. The character count is 49 / 5000. At the bottom right, there are "Send feedback" and sharing icons.

We are starting to learn artificial intelligence.

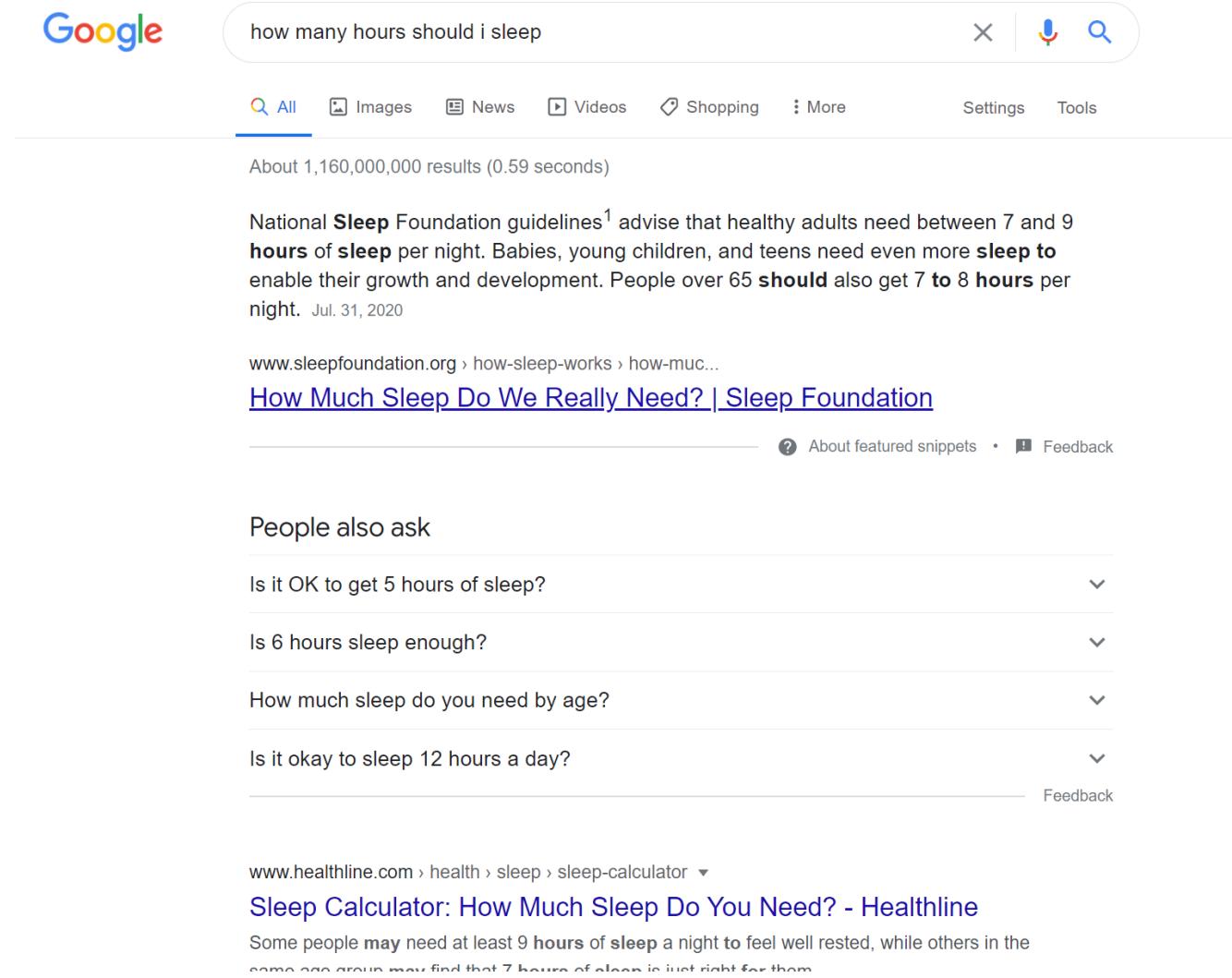
ما در حال یادگیری هوش مصنوعی هستیم.

ENGLISH - DETECTED GERMAN FRENCH PERSIAN PERSIAN ENGLISH SPANISH

49 / 5000

Send feedback

# Search & QA



Google

how many hours should i sleep

All Images News Videos Shopping More Settings Tools

About 1,160,000,000 results (0.59 seconds)

National **Sleep** Foundation guidelines<sup>1</sup> advise that healthy adults need between 7 and 9 hours of **sleep** per night. Babies, young children, and teens need even more **sleep** to enable their growth and development. People over 65 **should** also get 7 to 8 hours per night. Jul. 31, 2020

[www.sleepfoundation.org](http://www.sleepfoundation.org) › how-sleep-works › how-muc...

[How Much Sleep Do We Really Need? | Sleep Foundation](http://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need)

About featured snippets • Feedback

People also ask

Is it OK to get 5 hours of sleep?

Is 6 hours sleep enough?

How much sleep do you need by age?

Is it okay to sleep 12 hours a day?

Feedback

[www.healthline.com/health/sleep/sleep-calculator](http://www.healthline.com/health/sleep/sleep-calculator)

[Sleep Calculator: How Much Sleep Do You Need? - Healthline](http://www.healthline.com/health/sleep/sleep-calculator#calculator)

Some people may need at least 9 hours of **sleep** a night to feel well rested, while others in the same age group may find that 7 hours of **sleep** is just right for them

# Search autocorrect and autocomplete

A screenshot of a Google search bar. The search term "wordls fas" is typed in, with the "o" in "fas" underlined in red, indicating it's a misspelling. Below the search bar, a list of ten search suggestions appears, all starting with "world's fastest" followed by a different subject: car, bike, train, supercomputer, 2020, phone, missile, man, bird, and animal. Each suggestion has a small magnifying glass icon to its left.

- wordls fas
- world's fastest car
- world's fastest bike
- world's fastest train
- world's fastest supercomputer
- world's fastest car 2020
- world's fastest phone
- world's fastest missile
- world's fastest man
- world's fastest bird
- world's fastest animal

# Social media analysis



# Chatbots

10:05

Hello

Hi, what can we do for you?

Event Feedback

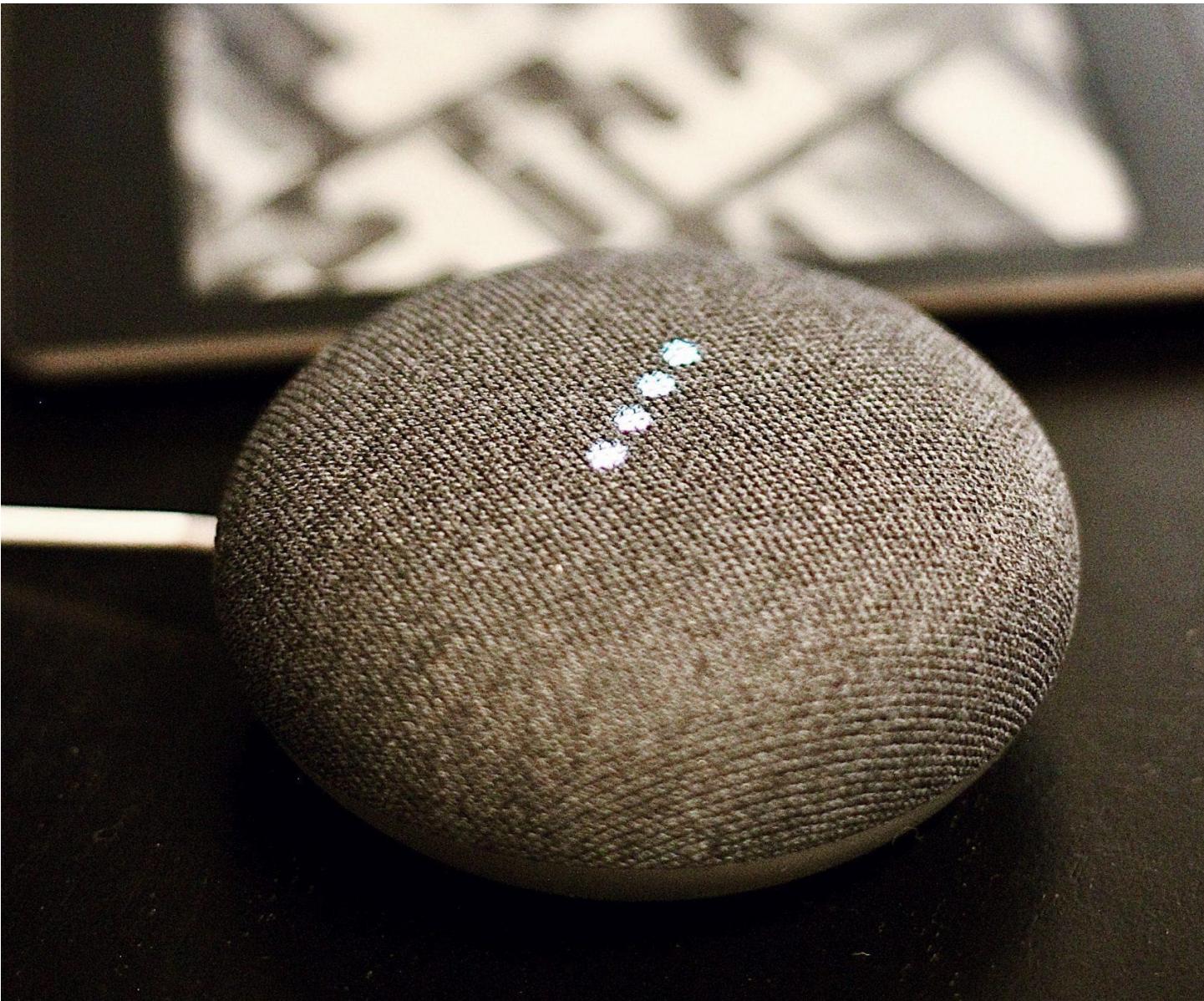
Thanks for coming along today - we'd love to hear what you think about today's event and the presentations. Let's get started.

So, what do you think? Give me your gut feeling!

# Hiring and recruitment



# Voice assistants



# Grammar checkers

The most common type of marketing channel is the wholesale market.

Varies kinds of **produce** are supplied from different areas are assembled at one place  
and sold thrc  
vegetables s

Replace the word

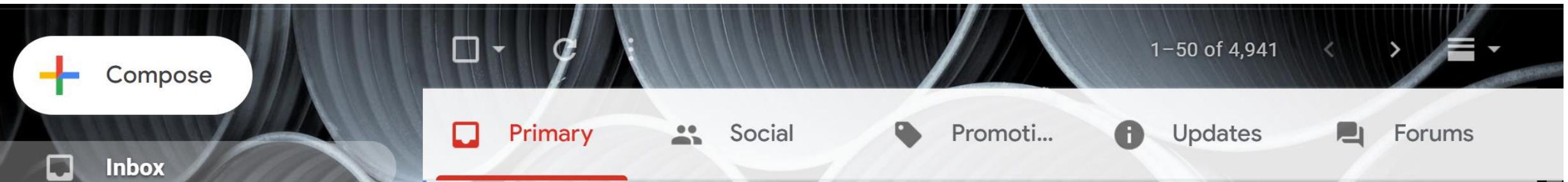
**products**

naller regional markets, etc. Fruits and  
market handling and transport methods.

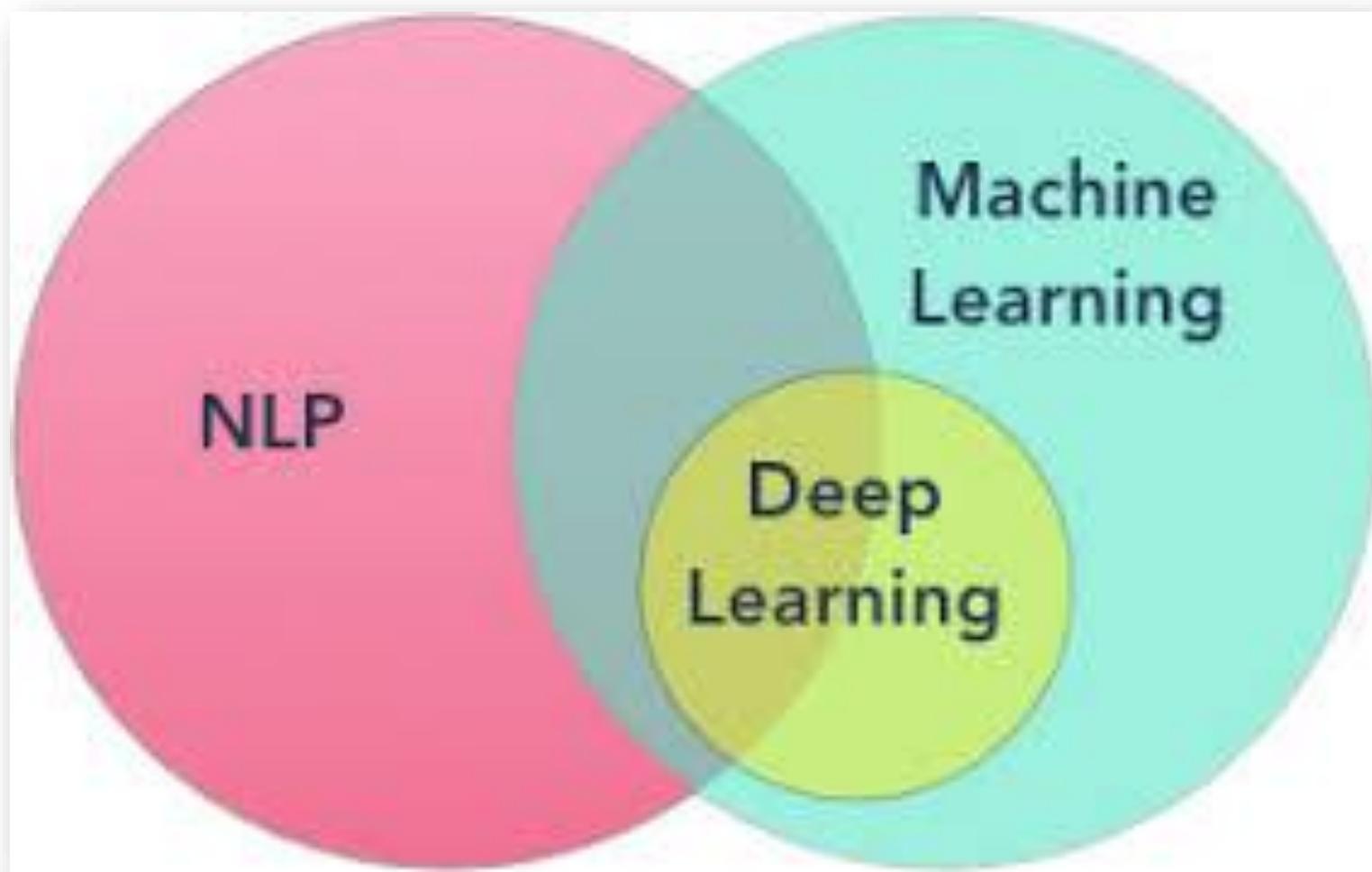
 Dismiss

Suggested by Grammarly

# Email classification



# NLP is not just machine learning



# Levels of linguistic structure

Discourse

Semantics

Syntax: Constituents

Syntax: Part of Speech

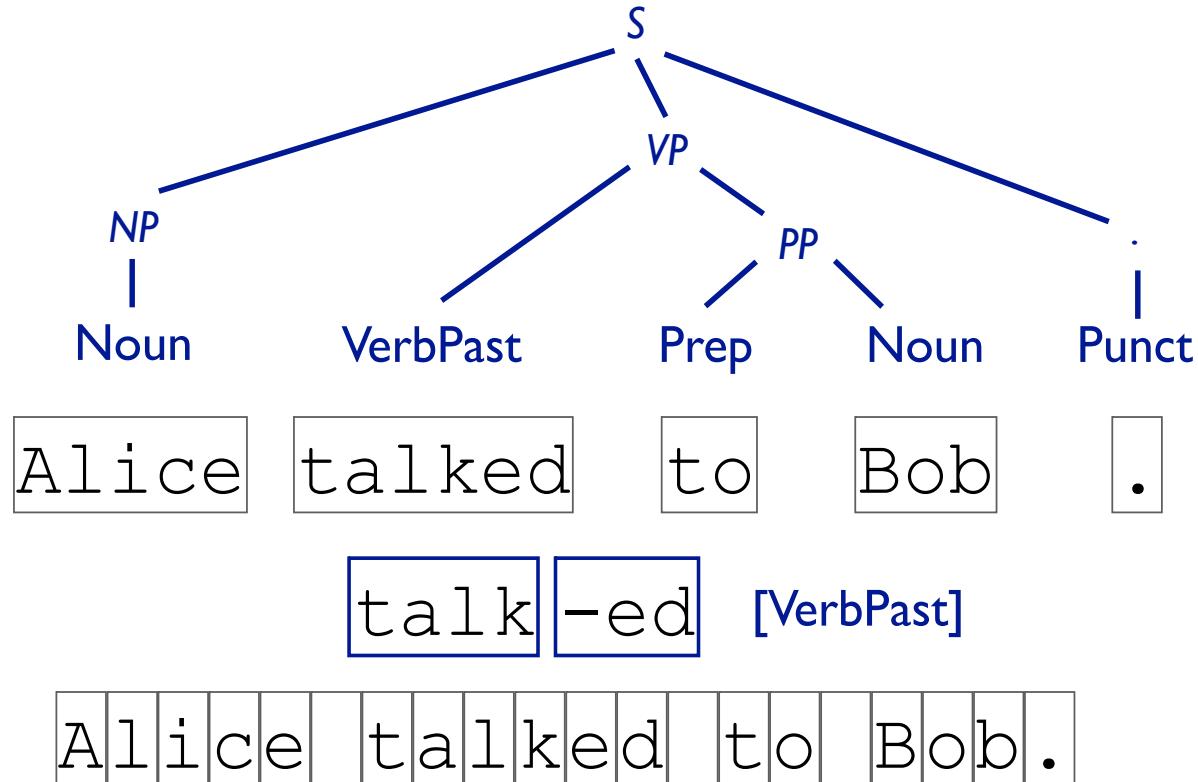
Words

Morphology

Characters

CommunicationEvent(e)  
Agent(e, Alice)  
Recipient(e, Bob)

SpeakerContext(s)  
TemporalBefore(e, s)



# Deep Learning for Text Classification

# Classification

- Output a choice from a fixed set of labels
- For sentiment:
  - Positive/negative
  - Star rating
  - ...

# Some examples of binary sentiment classification

*this movie was great! would watch again*

+

*the movie was gross and overwrought, but I liked it*

+

*this movie was not really very enjoyable*

-

# Building a classifier

- Let's say we have 10k labeled sentences
- We want to learn a function  $f$  that
  - maps an unseen sentence to one of the labels

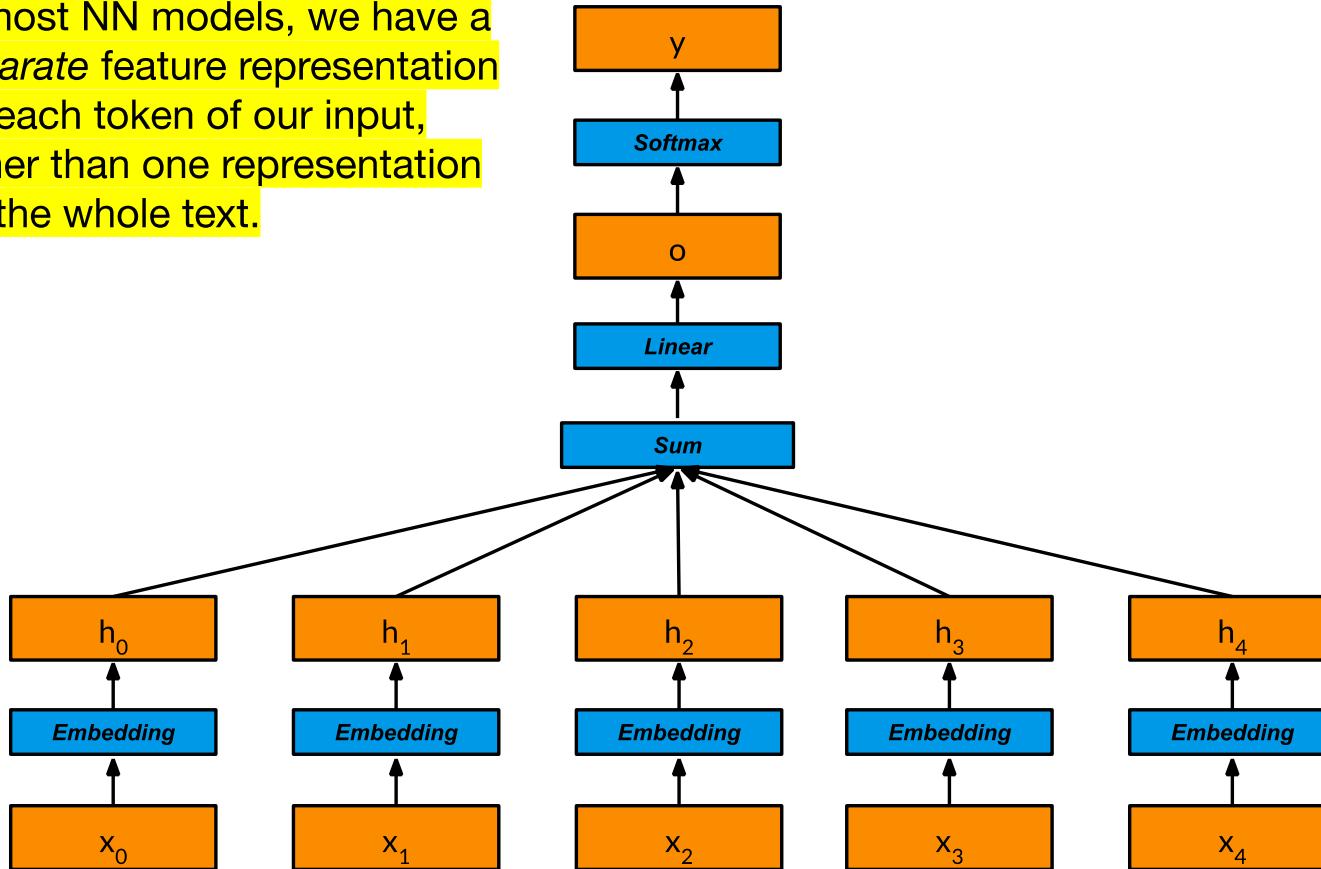
# From strings to words

- I don't like any of Ford's trucks.
- I do n't like any of Ford 's trucks .

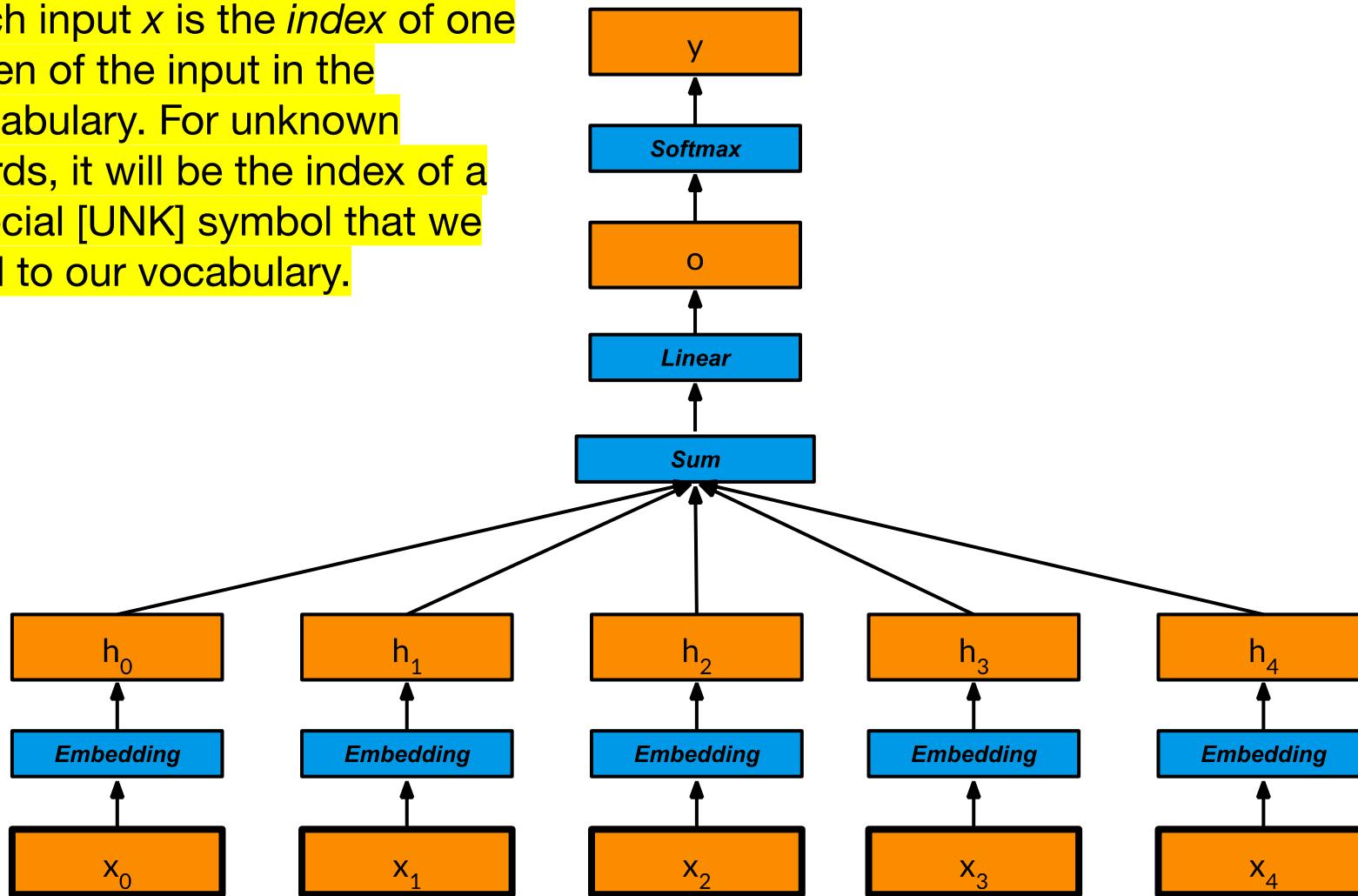
Tokenization: Turning strings into a sequence of symbols (e.g., words, subwords, characters, etc)

# NN sentiment classifier

In most NN models, we have a *separate* feature representation for each token of our input, rather than one representation for the whole text.

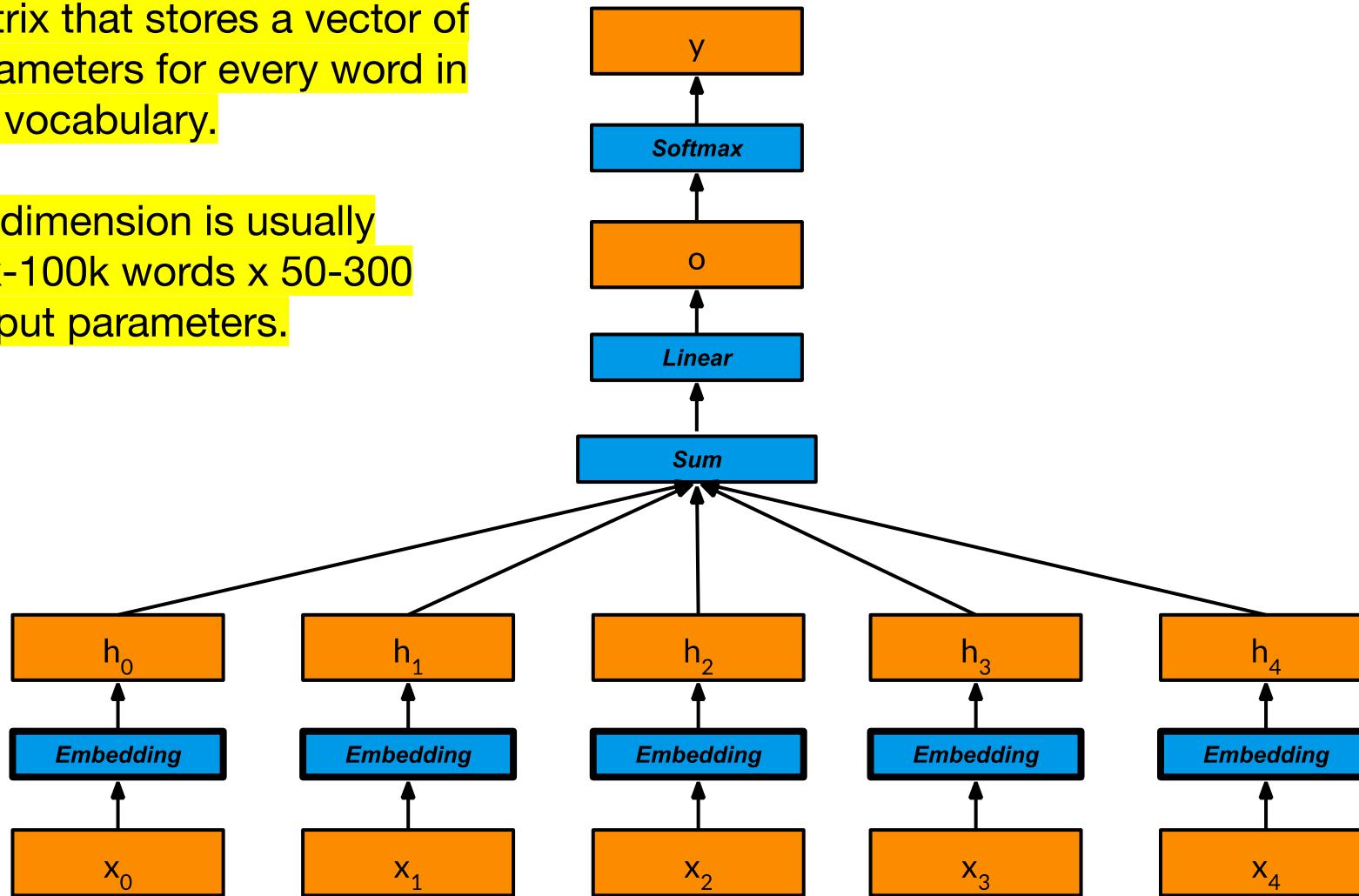


Each input  $x$  is the *index* of one token of the input in the vocabulary. For unknown words, it will be the index of a special [UNK] symbol that we add to our vocabulary.

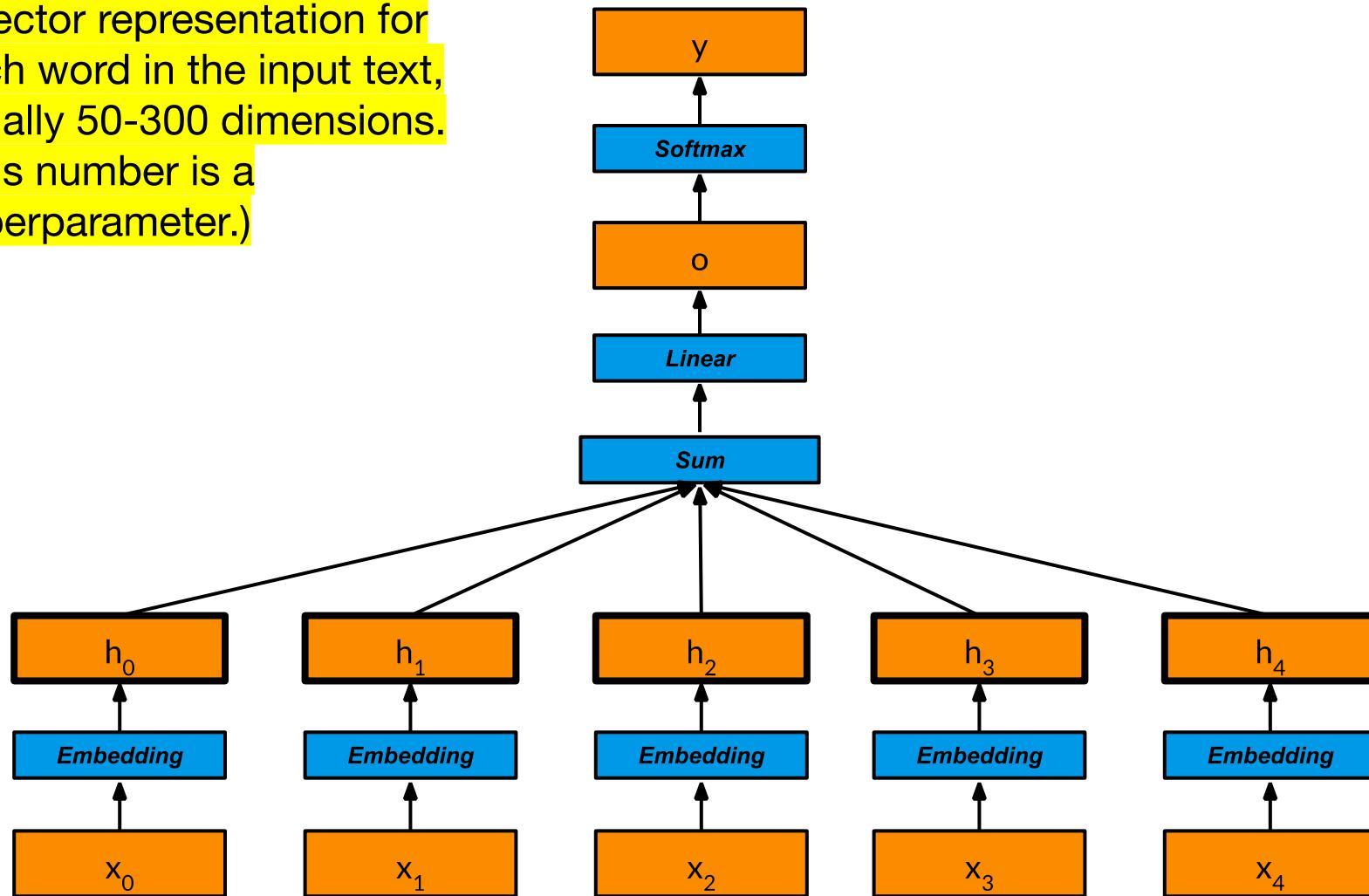


The *embedding layer* is a matrix that stores a vector of parameters for every word in the vocabulary.

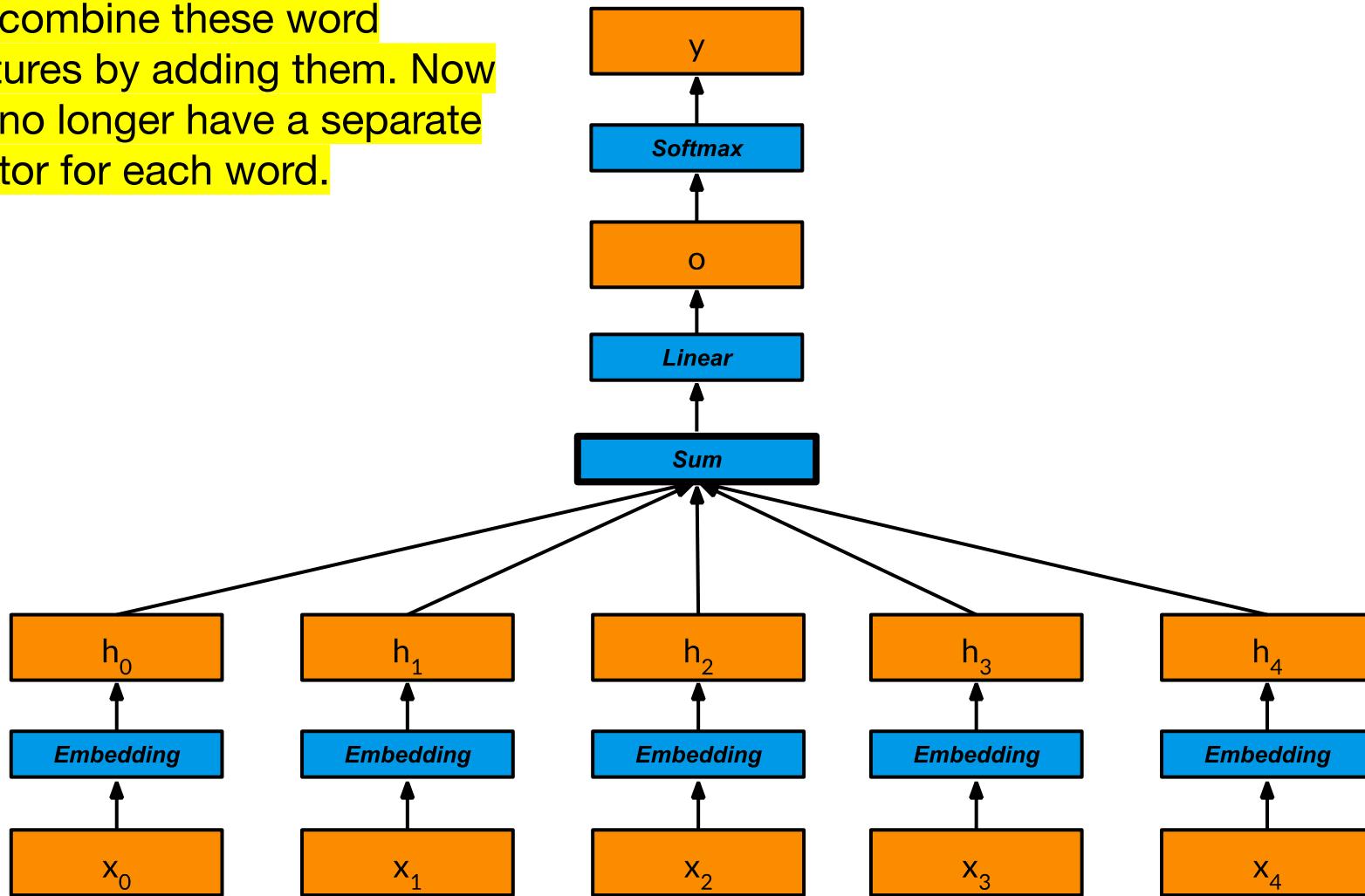
It's dimension is usually 10k-100k words x 50-300 output parameters.



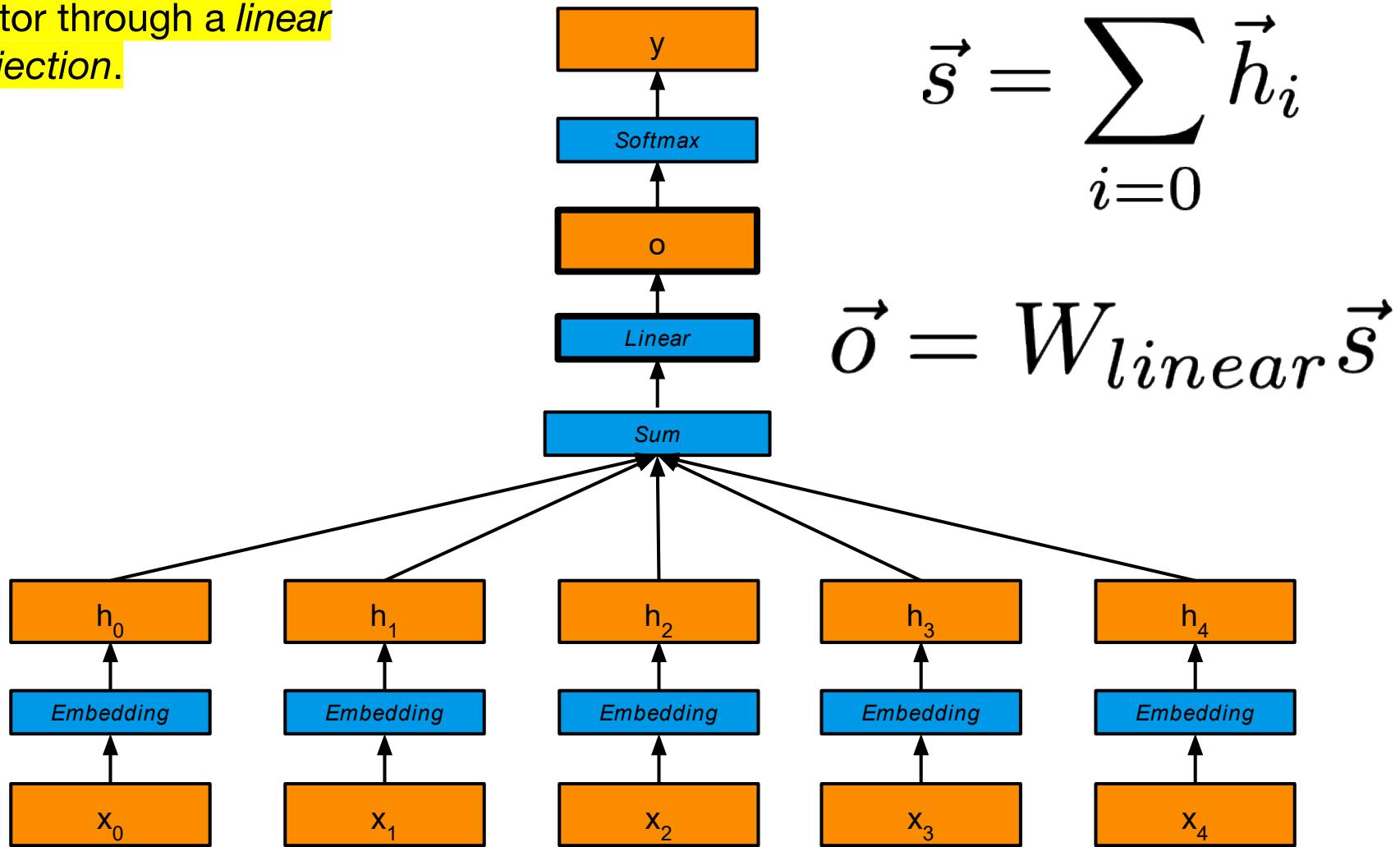
The embedding layer produces a vector representation for each word in the input text, usually 50-300 dimensions.  
(This number is a hyperparameter.)



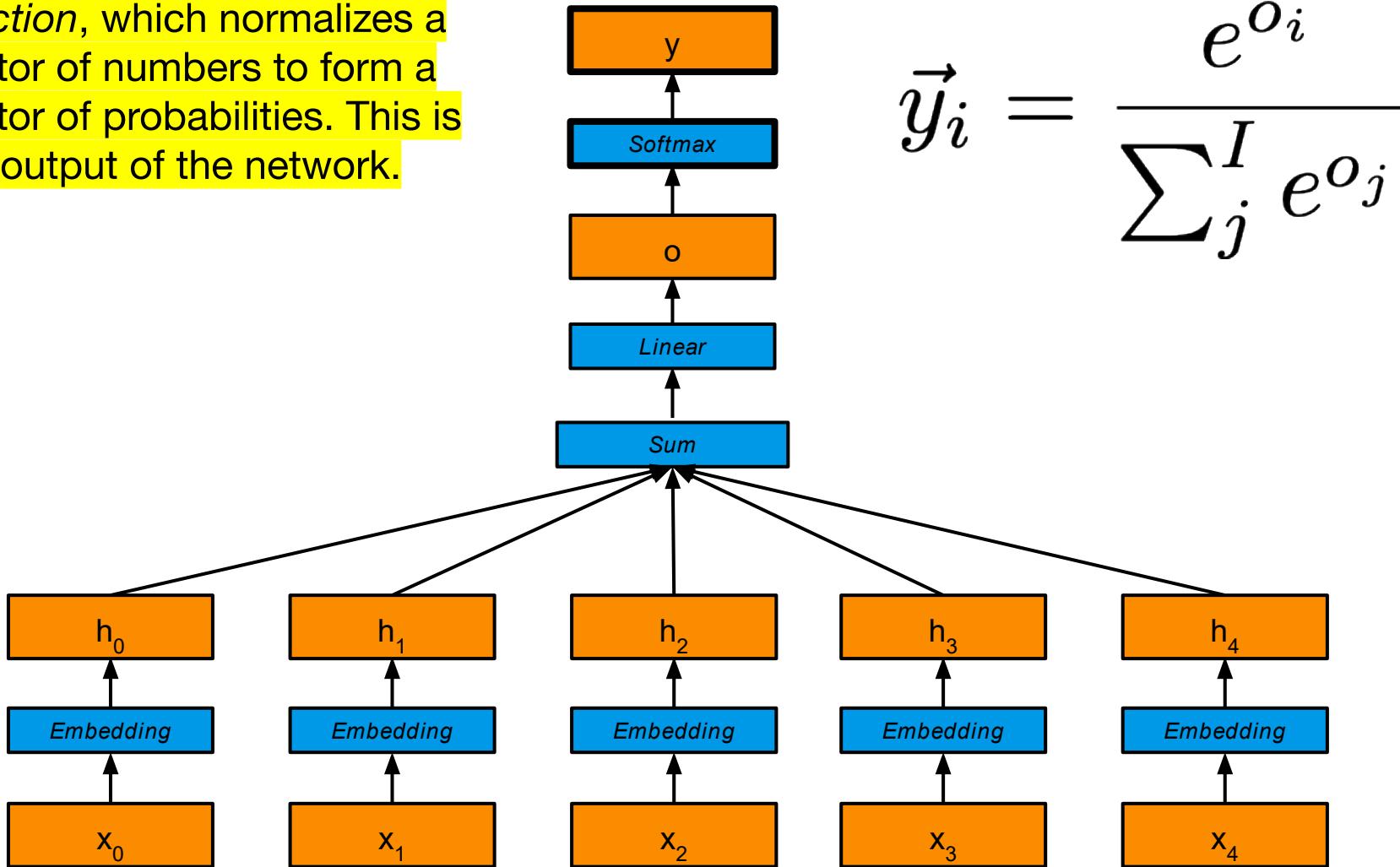
In this simple neural network, we combine these word features by adding them. Now we no longer have a separate vector for each word.



Next, we pass the resulting vector through a *linear projection*.

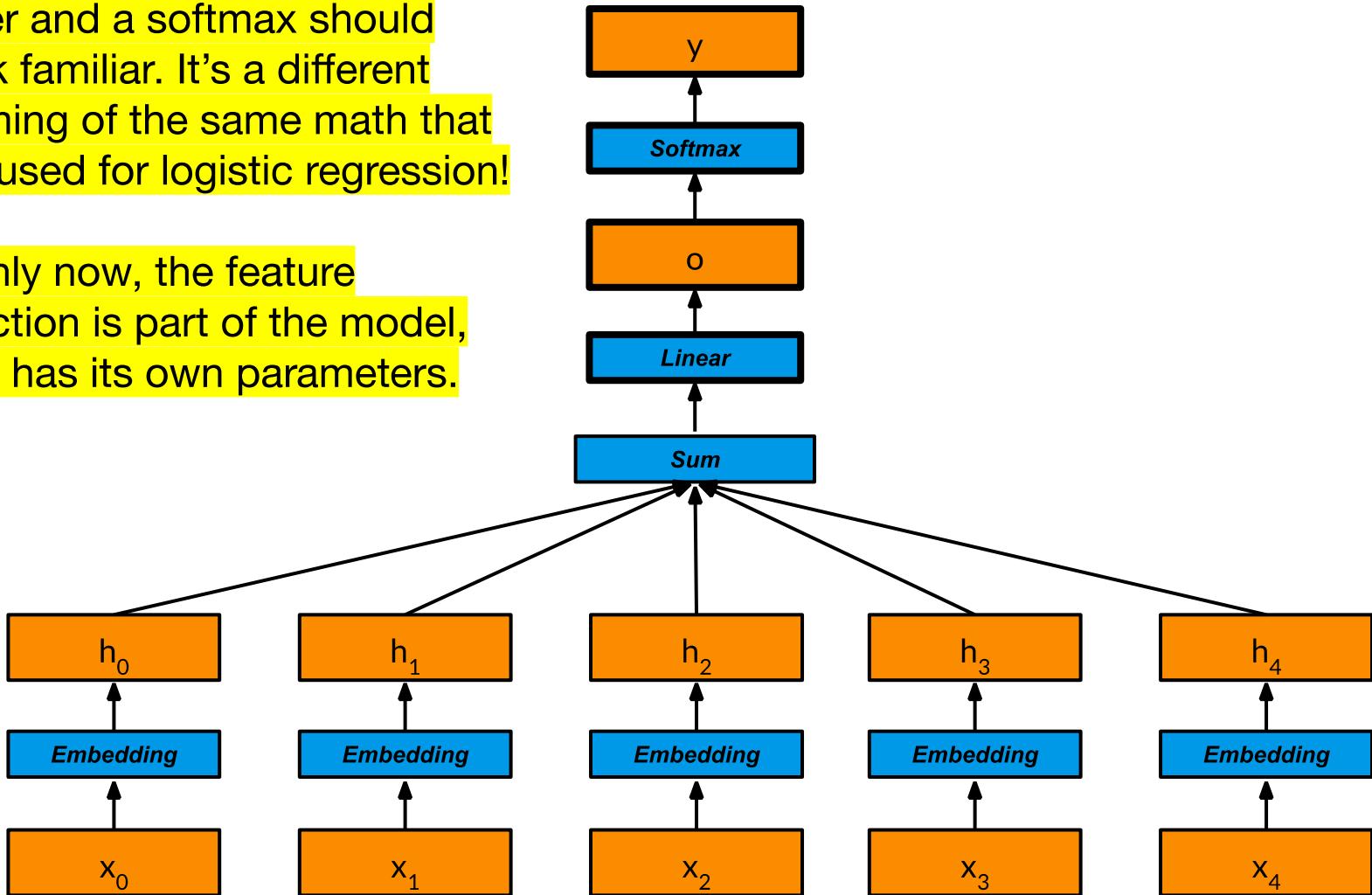


After that, we use the softmax function, which normalizes a vector of numbers to form a vector of probabilities. This is the output of the network.

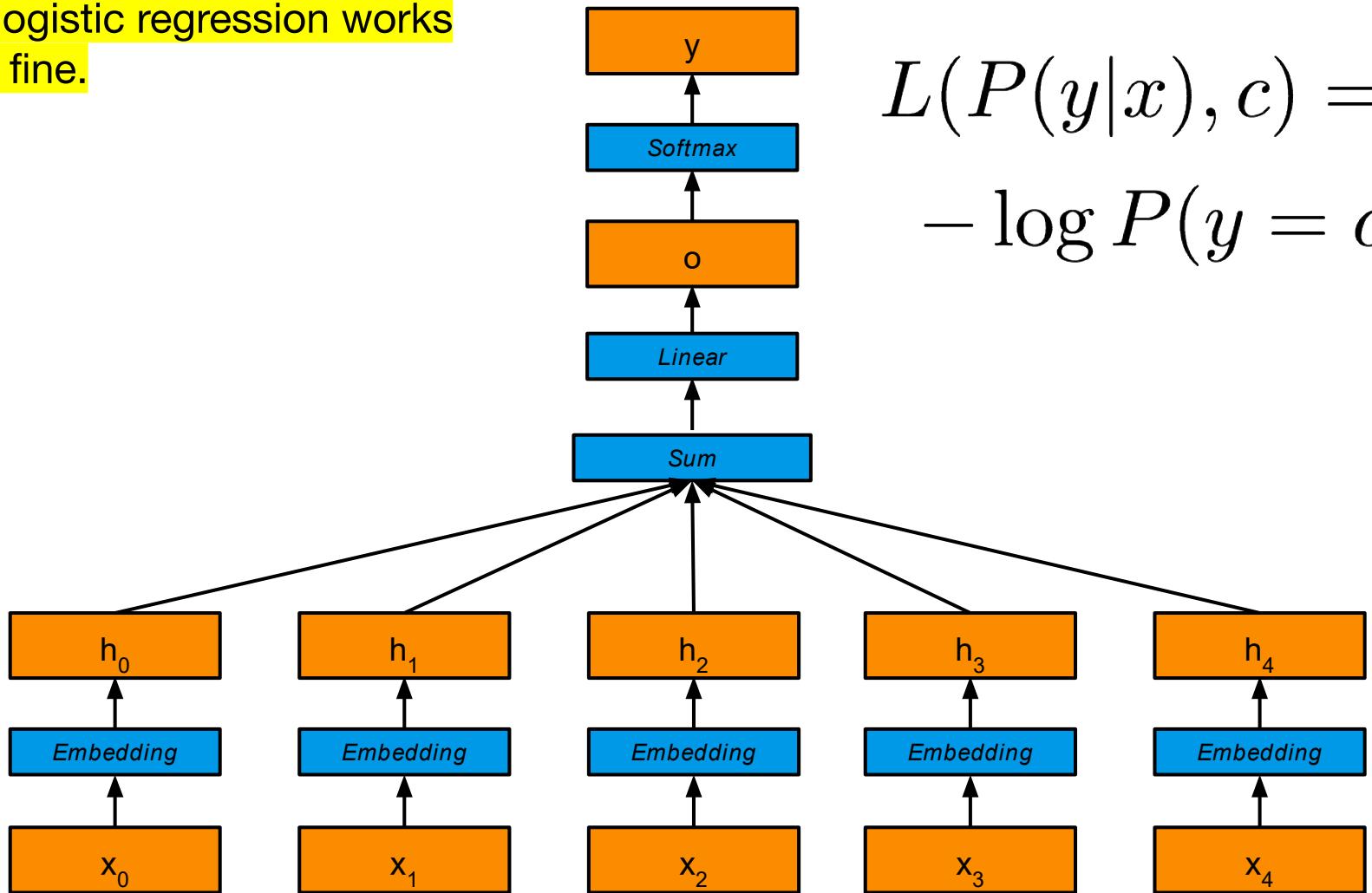


This combination of a linear layer and a softmax should look familiar. It's a different framing of the same math that we used for logistic regression!

...only now, the feature function is part of the model, and has its own parameters.



The *loss function* that we used  
for logistic regression works  
just fine.

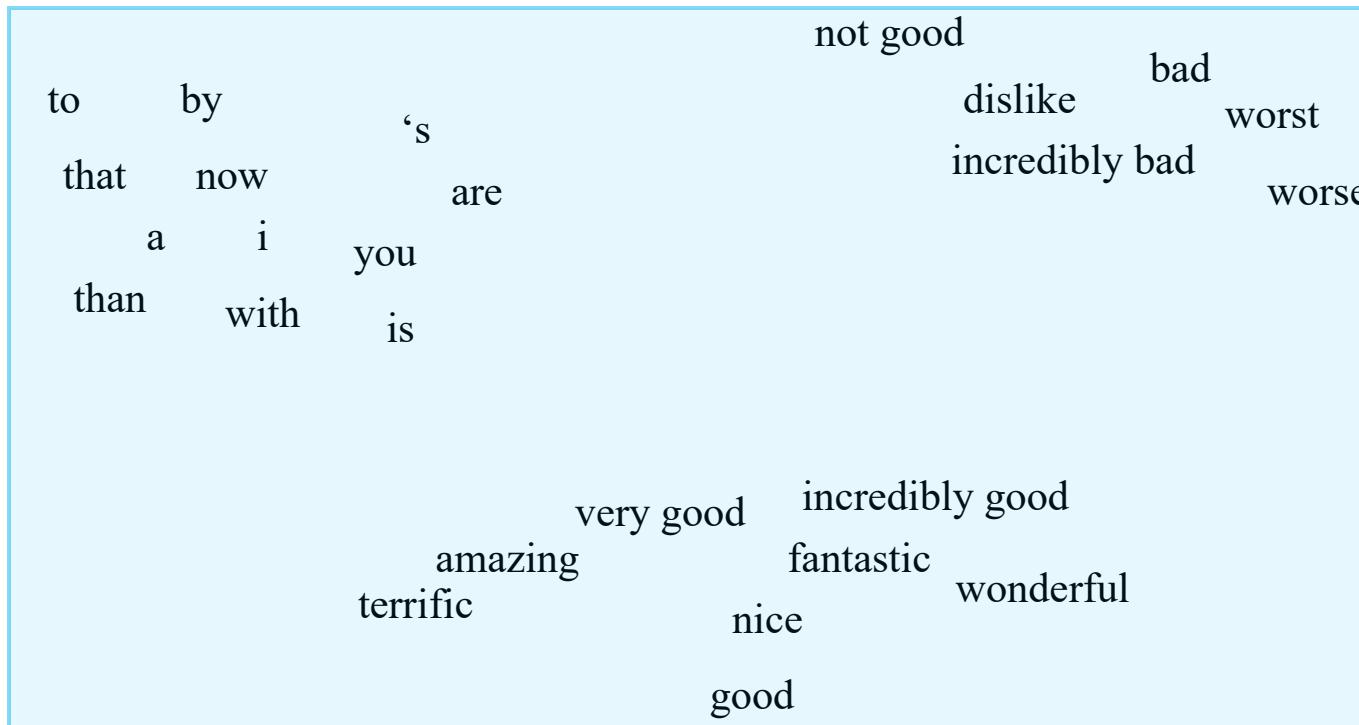


$$L(P(y|x), c) = -\log P(y = c|x)$$

# Word embedding

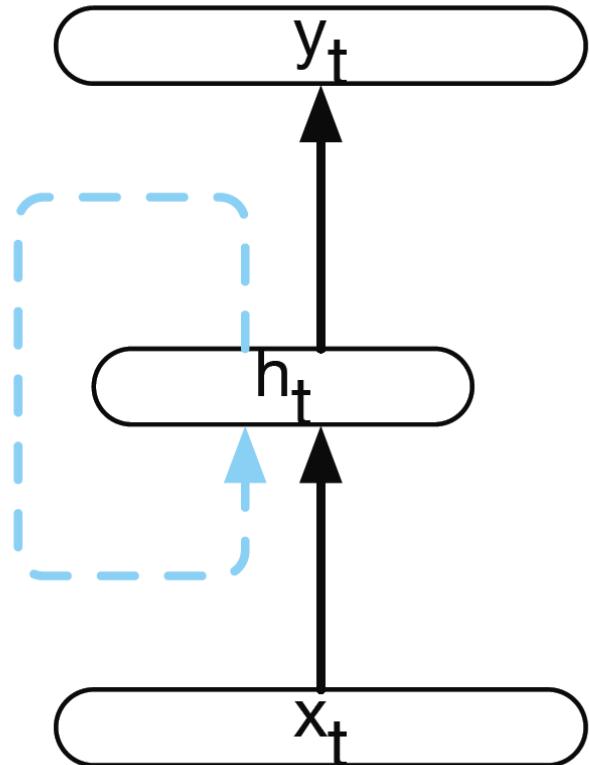
# Word embedding

- Each word = a vector
- Similar words are "nearby in space"



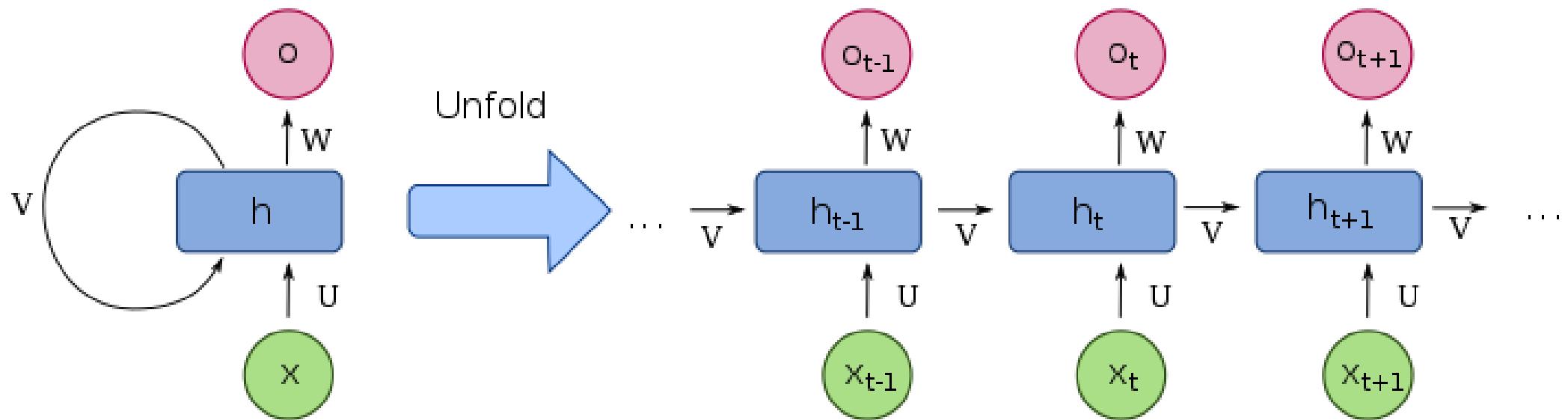
# RNNs, attention, transformers

# Recurrent Neural Networks (RNNs)

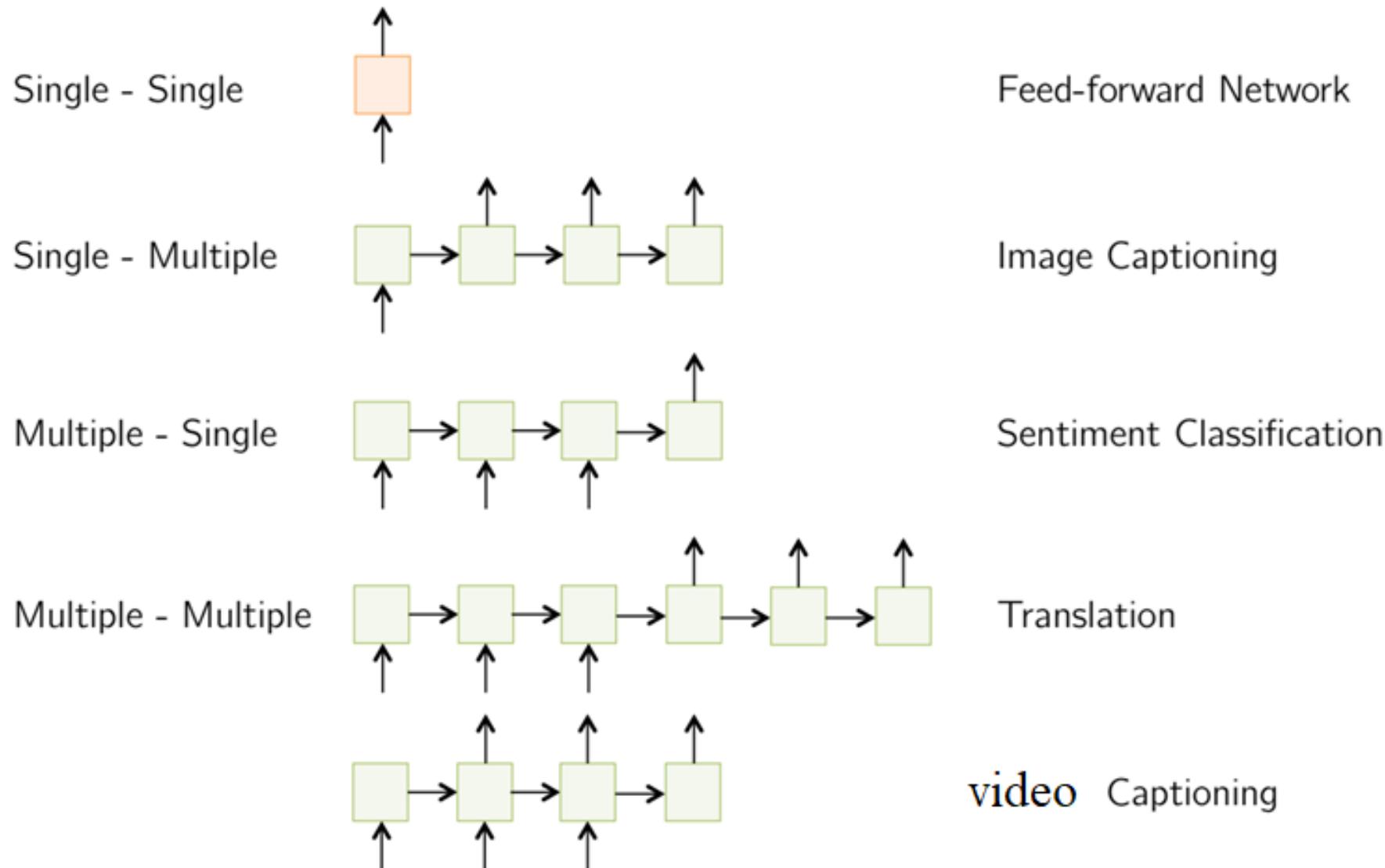


- RNNs take the previous output or hidden states as inputs! The composite input at time  $t$  has some historical information about the happenings at time  $T < t$
- RNNs are useful as their intermediate values (state) can store information about past inputs for a time that is not fixed a priori

# Recurrent Neural Networks (RNNs)

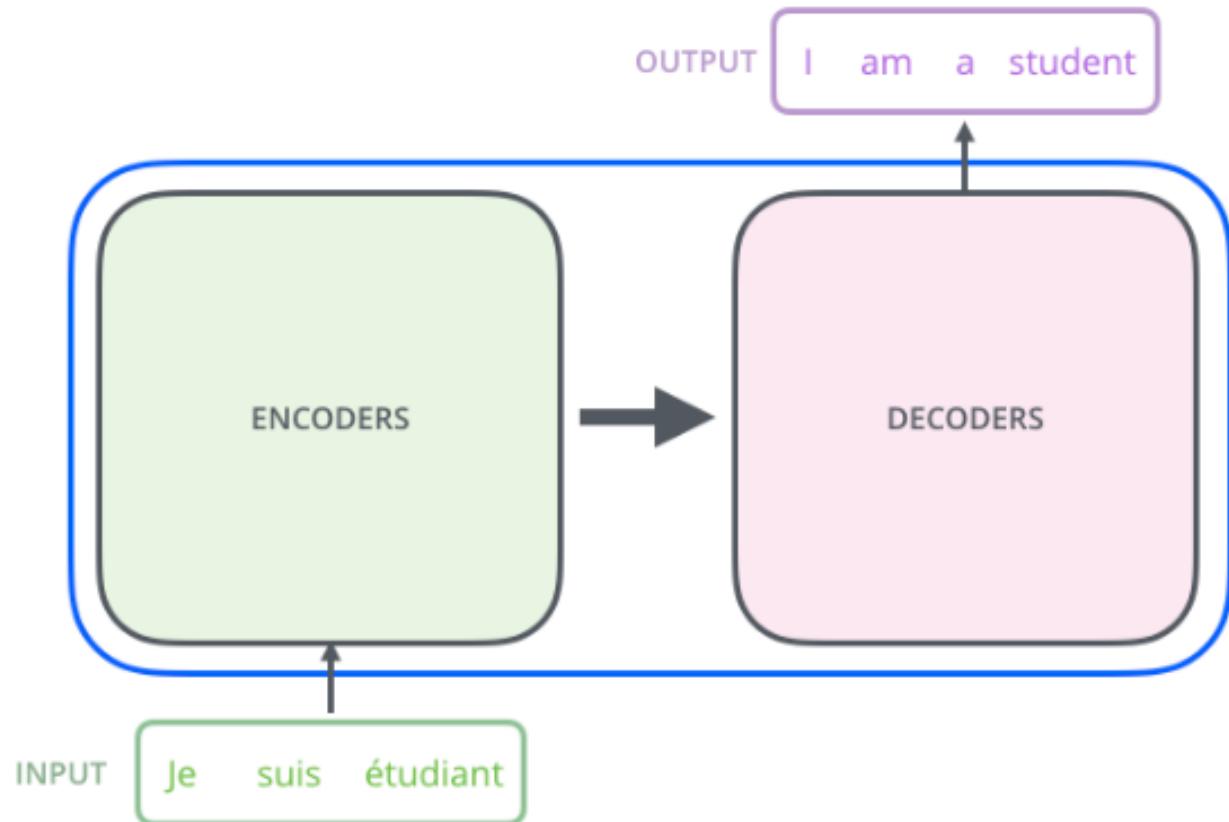


# RNNs applications

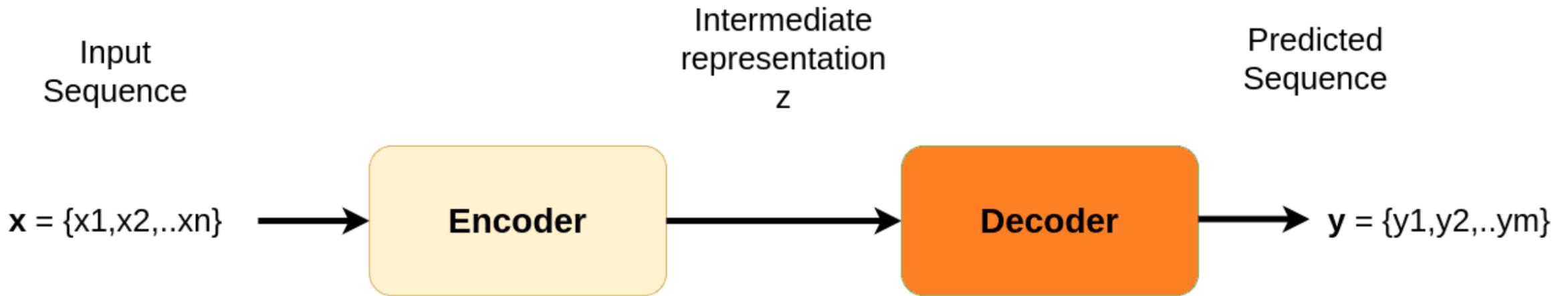


# Encoder-decoder networks

- Used in a wide range of applications including machine translation, summarization, question answering, and dialogue modeling.
- RNNs were the most widely-used and successful architecture for both the encoder and decoder.

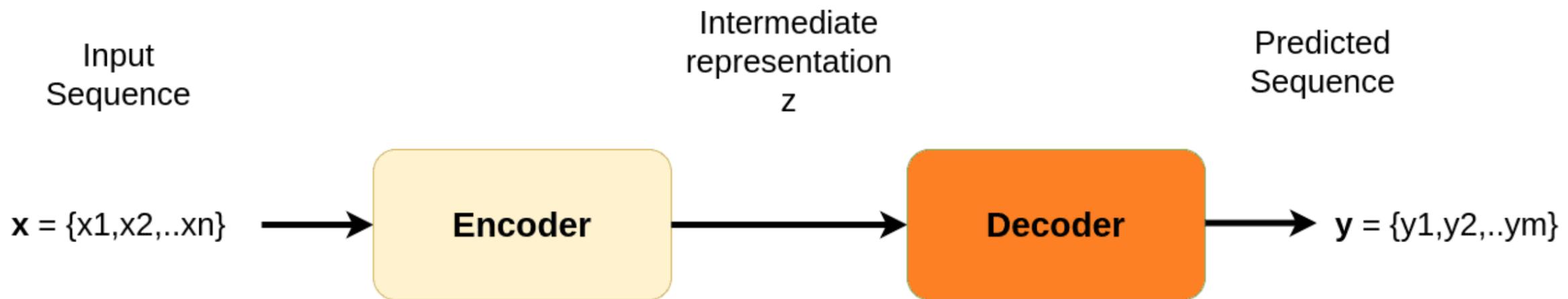


# Encoder-decoder: seq2seq

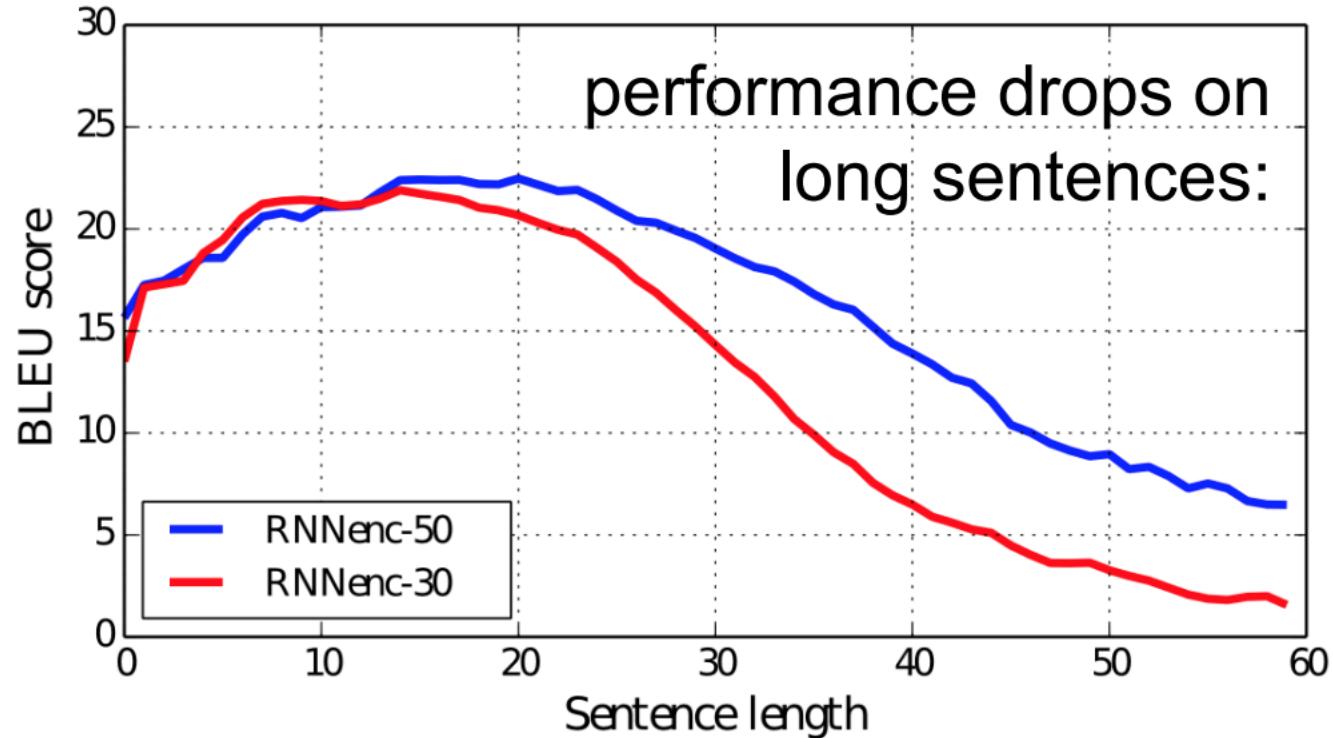


<https://theaisummer.com/attention/>

## What if sequence length is high (say > 30)?



The vector  $z$  needs to capture all the information about the source sentence.

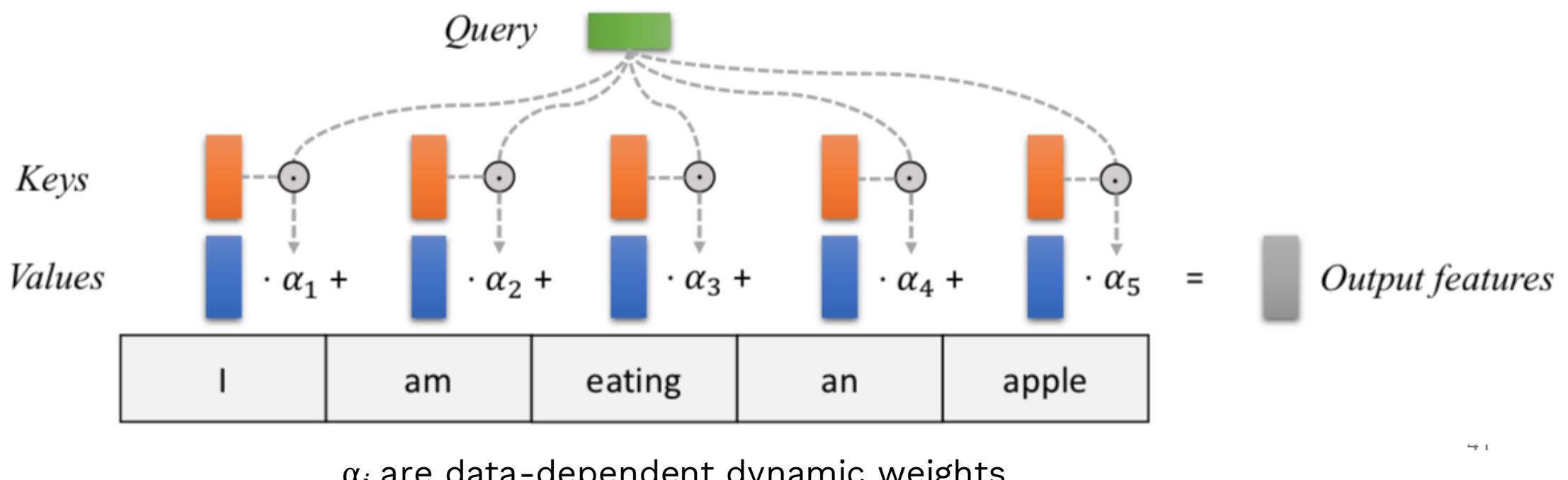


fixed size representation can be the bottleneck

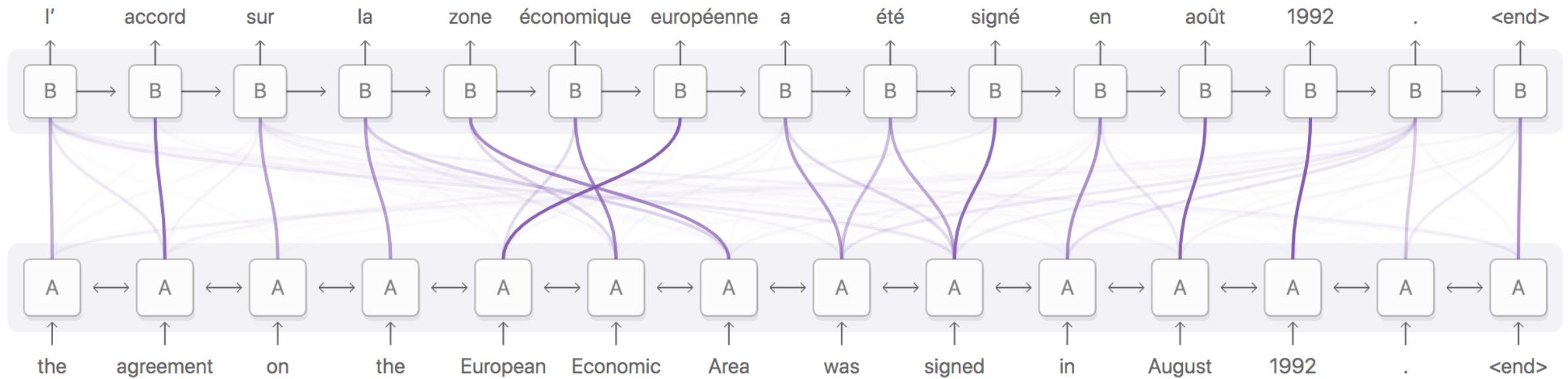
The core idea of **attention** is that the context vector  $z$  should have access to **all** parts of the input sequence instead of just the last one.

# Attention

- A weighted average of (sequence) elements with the weights depending on an input query.
- The idea and the name taken from the intuition that humans attend to certain things at each time.

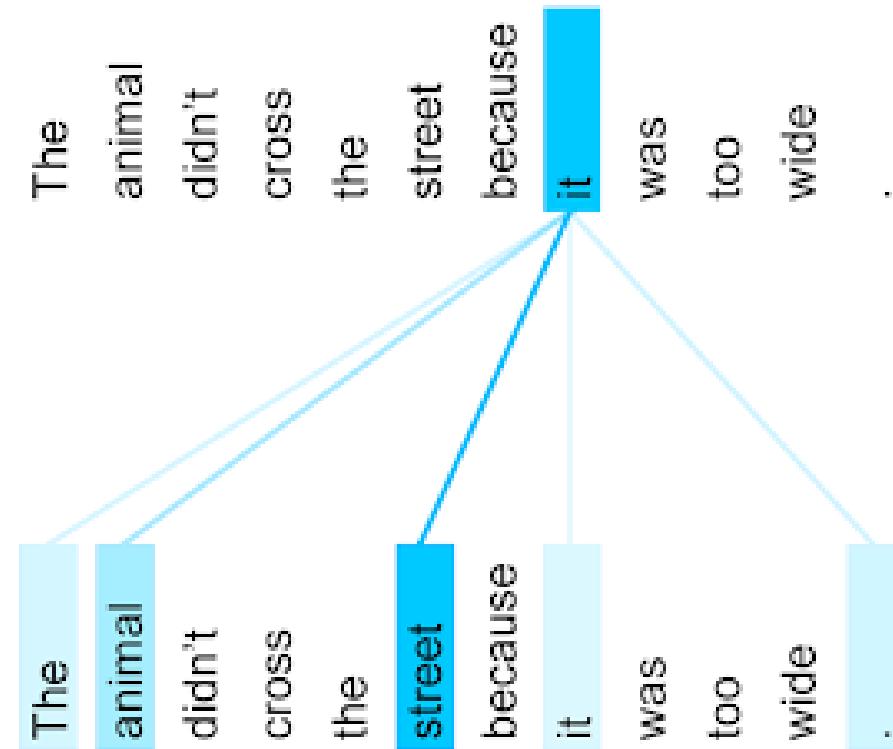
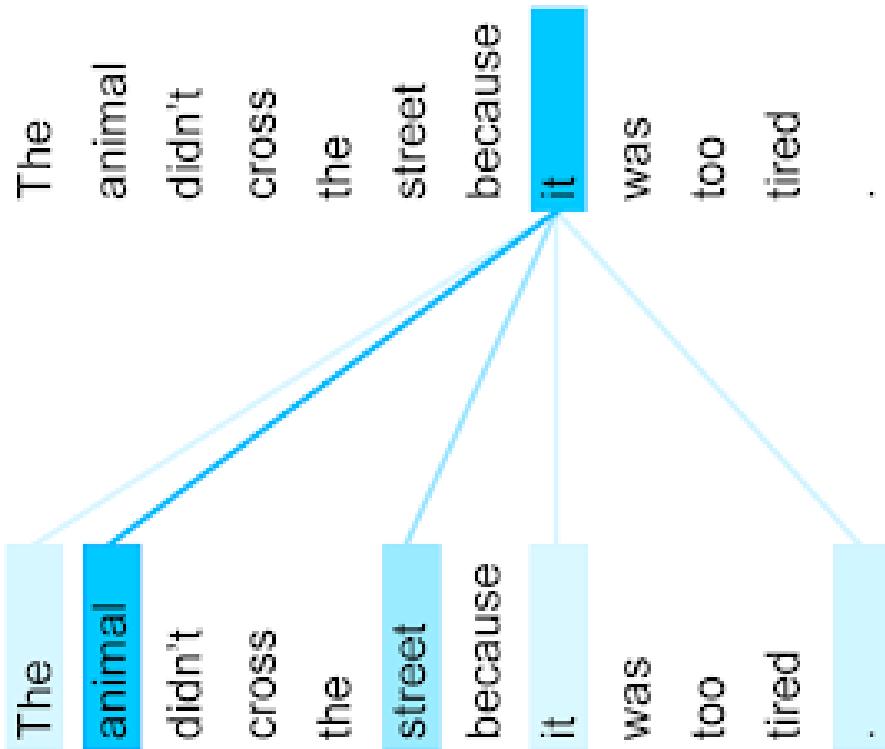


# MT



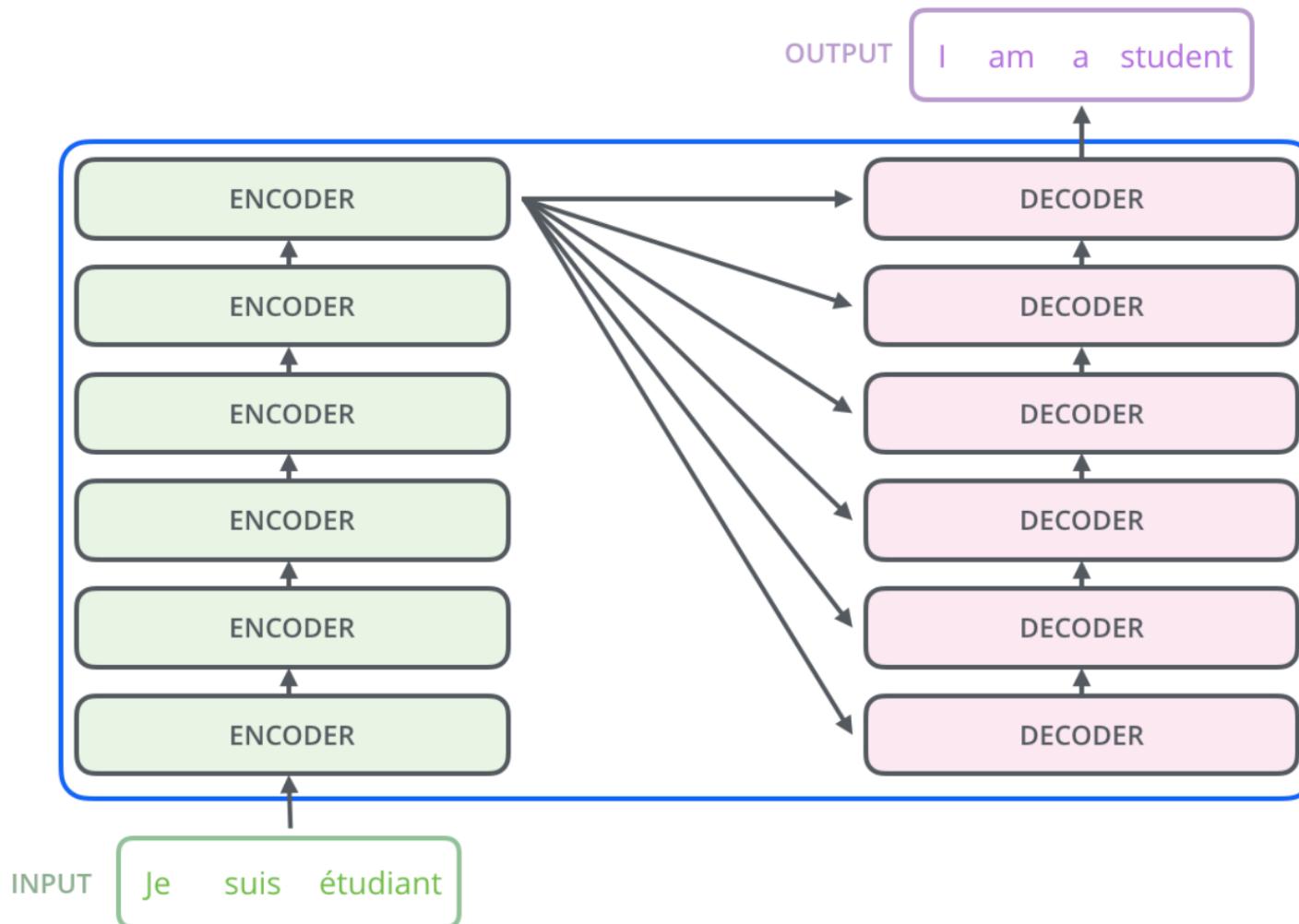
# Self-Attention

- For each word, self-attention allows the model to look at other positions in the input for a better encoding for this word.

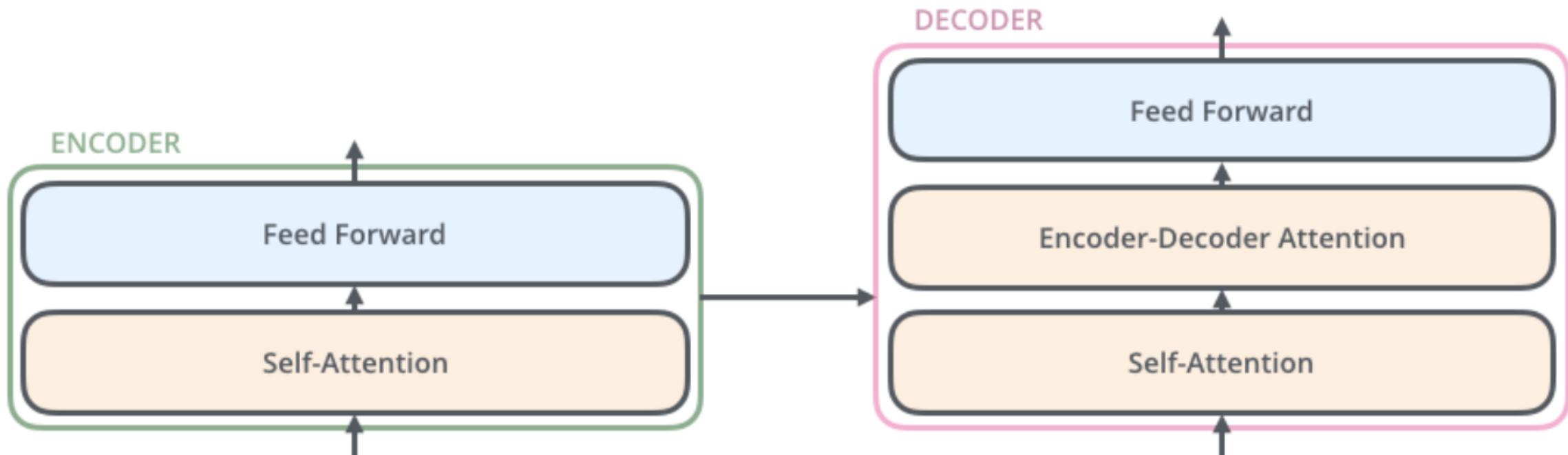


# Transformer

# Transformer Model



# Transformer Model



<http://jalammar.github.io/illustrated-transformer/>

# Transformer model

- Non-recurrent sequence to sequence encoder-decoder model
  - Eliminate recurrence, allows for significantly more parallelization
- Three kinds of attentions:
  - The input and output tokens (solved by traditional attention mechanism)
  - The input tokens themselves
  - The output tokens themselves
- Extend the (self-)attention mechanism to processing input and output sentences as well.

