EE-608: Deep Learning For Natural Language Processing: Machine Translation, Sequence2Sequence, Attention

James Henderson

Idiap Research Institute

DLNLP, Lecture 3

Outline

Outline

Machine Translation

Machine Translation (MT) is the task of translating a sentence *x* from one language (the source language) to a sentence *y* in another language (the target language).

x: *L'homme est né libre, et partout il est dans les fers*

$$\downarrow$$

y: *Man is born free, but everywhere he is in chains*

– Rousseau

Slide from Christopher Manning

1990s-2010s: Statistical Machine Translation

- <u>Core idea</u>: Learn a probabilistic model from data
- Suppose we're translating French → English.
- We want to find best English sentence *y,* given French sentence *x*

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

- Use Bayes Rule to break this down into two components to be learned separately:

$$= \text{argmax}_y \underbrace{P(x|y)}_{\text{Translation Model}} \underbrace{P(y)}_{\text{Language Model}}$$

Translation Model

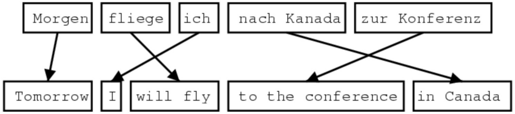
Models how words and phrases should be translated (*fidelity*). Learned from parallel data.

Language Model

Models how to write good English (*fluency*). Learned from monolingual data.

Slide from Christopher Manning

What happens in translation isn't trivial to model!



1519年600名西班牙人在墨西哥登陆，去征服几百万人口的阿兹特克帝国，初次交锋他们损兵三分之二。

In 1519, six hundred Spaniards landed in Mexico to conquer the Aztec Empire with a population of a few million. They lost two thirds of their soldiers in the first clash.

translate.google.com (2009): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of soldiers against their loss.

translate.google.com (2013): 1519 600 Spaniards landed in Mexico to conquer the Aztec empire, hundreds of millions of people, the initial confrontation loss of soldiers two-thirds.

translate.google.com (2015): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of the loss of soldiers they clash.

Slide from Christopher Manning

1990s–2010s: Statistical Machine Translation

- SMT was a huge research field
- The best systems were extremely complex
  - Hundreds of important details
- Systems had many separately-designed subcomponents
  - Lots of feature engineering
    - Need to design features to capture particular language phenomena
  - Required compiling and maintaining extra resources
    - Like tables of equivalent phrases
  - Lots of human effort to maintain
    - Repeated effort for each language pair!

Slide from Christopher Manning

Outline

Neural Machine Translation: Idea

- Like many tasks, MT is a **sequence-to-sequence** problem.
- We know how to **encode** sequences of words with recurrent neural networks.
- We know how to conditionally **generate** sequences of words with recurrent neural networks.
- Why not simply encode the source sentence and condition on that to generate the target sentence?

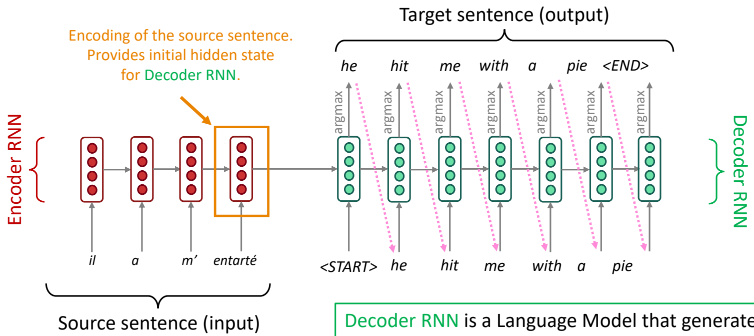
What is Neural Machine Translation?

- Neural Machine Translation (NMT) is a way to do Machine Translation with a *single end-to-end neural network*
- The neural network architecture is called a sequence-to-sequence model (aka seq2seq) and it involves *two* RNNs

Slide from Christopher Manning

Neural Machine Translation (NMT)

The sequence-to-sequence model



Target sentence (output)
Encoding of the source sentence. Provides initial hidden state for Decoder RNN.
he hit me with a pie <END>
argmax
Encoder RNN
Decoder RNN
il a m' entarté
<START> he hit me with a pie
Source sentence (input)

Decoder RNN is a Language Model that generates target sentence, *conditioned on encoding*.

Note: This diagram shows **test time** behavior: decoder output is fed in ·······▸ as next step's input

Encoder RNN produces an encoding of the source sentence.

Slide from Christopher Manning

Sequence-to-sequence is versatile!

- The general notion here is an encoder-decoder model
  - One neural network takes input and produces a neural representation
  - Another network produces output based on that neural representation
  - If the input and output are sequences, we call it a seq2seq model

- Sequence-to-sequence is useful for *more than just MT*
- Many NLP tasks can be phrased as sequence-to-sequence:
  - Summarization (long text → short text)
  - Dialogue (previous utterances → next utterance)
  - Parsing (input text → output parse as sequence)
  - Code generation (natural language → Python code)

Slide from Christopher Manning

Neural Machine Translation (NMT)

- The sequence-to-sequence model is an example of a **Conditional Language Model**
    - **Language Model** because the decoder is predicting the next word of the target sentence *y*
    - **Conditional** because its predictions are *also* conditioned on the source sentence *x*

- NMT directly calculates $P(y|x)$:

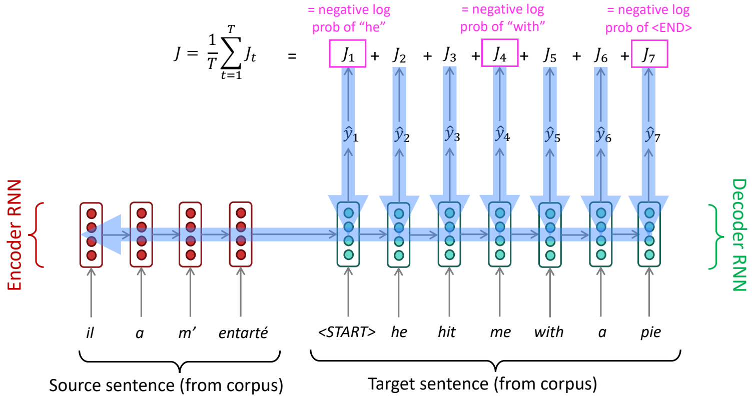
$$P(y|x) = P(y_1|x)\,P(y_2|y_1,x)\,P(y_3|y_1,y_2,x)\ldots\underbrace{P(y_T|y_1,\ldots,y_{T-1},x)}_{\text{Probability of next target word, given target words so far and source sentence } x}$$

- **Question:** How to train an NMT system?
- **(Easy) Answer:** Get a big parallel corpus…
    - But there is now exciting work on "unsupervised NMT", data augmentation, etc.

Slide from Christopher Manning

Training a Neural Machine Translation system

$$J = \frac{1}{T}\sum_{t=1}^{T} J_t \quad = \quad \boxed{J_1} + J_2 + J_3 + \boxed{J_4} + J_5 + J_6 + \boxed{J_7}$$


= negative log prob of “he”
= negative log prob of “with”
= negative log prob of <END>
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$ $\hat{y}_4$ $\hat{y}_5$ $\hat{y}_6$ $\hat{y}_7$
Encoder RNN
Decoder RNN
il a m’ entarté
<START> he hit me with a pie
Source sentence (from corpus)
Target sentence (from corpus)

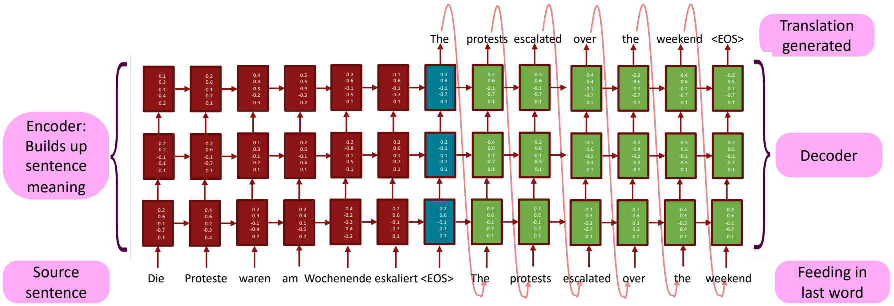
Seq2seq is optimized as a **single system.** Backpropagation operates *“end-to-end”*.

Slide from Christopher Manning

Multi-layer deep encoder-decoder machine translation net

[Sutskever et al. 2014; Luong et al. 2015]

The hidden states from RNN layer *i* are the inputs to RNN layer *i*+1

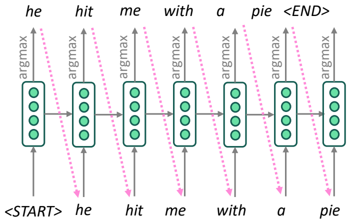


Conditioning = Bottleneck

Slide from Christopher Manning

Decoding: Greedy decoding

- We saw how to generate (or “decode”) the target sentence by taking argmax on each step of the decoder



- This is greedy decoding (take most probable word on each step)

Slide from Christopher Manning

Problems with greedy decoding

- Greedy decoding has no way to undo decisions!
  - Input: *il a m’entarté* *(he hit me with a pie)*
  - → *he* ____
  - → *he hit* ____
  - → *he hit a* ____ *(whoops! no going back now…)*

- How to fix this?

Slide from Christopher Manning

Exhaustive search decoding

- Ideally, we want to find a (length *T*) translation *y* that maximizes

$$P(y|x) = P(y_1|x)\, P(y_2|y_1, x)\, P(y_3|y_1, y_2, x) \ldots, P(y_T|y_1, \ldots, y_{T-1}, x)$$
$$= \prod_{t=1}^{T} P(y_t|y_1, \ldots, y_{t-1}, x)$$

- We could try computing all possible sequences *y*
  - This means that on each step *t* of the decoder, we're tracking V^t possible partial translations, where *V* is vocab size
  - This $O(V^T)$ complexity is far too expensive!

Slide from Christopher Manning

Beam search decoding

- Core idea: On each step of decoder, keep track of the *k* most probable partial translations (which we call *hypotheses*)
  - *k* is the beam size (in practice around 5 to 10, in NMT)

- A hypothesis $y_1, \ldots, y_t$ has a score which is its log probability:

$$\text{score}(y_1, \ldots, y_t) = \log P_{\text{LM}}(y_1, \ldots, y_t | x) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$$

  - Scores are all negative, and higher score is better
  - We search for high-scoring hypotheses, tracking top *k* on each step

- Beam search is not guaranteed to find optimal solution
- But much more efficient than exhaustive search!

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

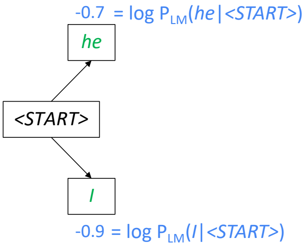
<START>

Calculate prob dist of next word

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$



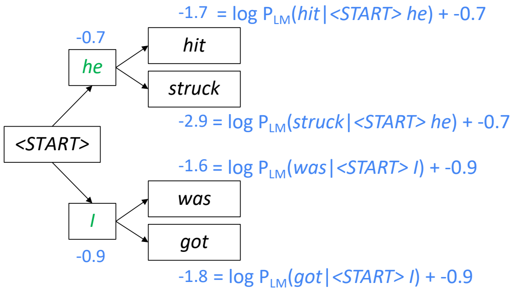
-0.7 = log P_{LM}(*he*|*<START>*)
he
<START>
I
-0.9 = log P_{LM}(*I*|*<START>*)

Take top *k* words and compute scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$



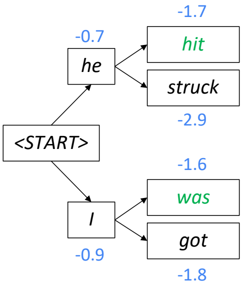
-1.7 = log P_{LM}(*hit*|*<START> he*) + -0.7
-0.7
hit
he
struck
<START>
-2.9 = log P_{LM}(*struck*|*<START> he*) + -0.7
-1.6 = log P_{LM}(*was*|*<START> I*) + -0.9
was
I
got
-0.9
-1.8 = log P_{LM}(*got*|*<START> I*) + -0.9

For each of the *k* hypotheses, find top *k* next words and calculate scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

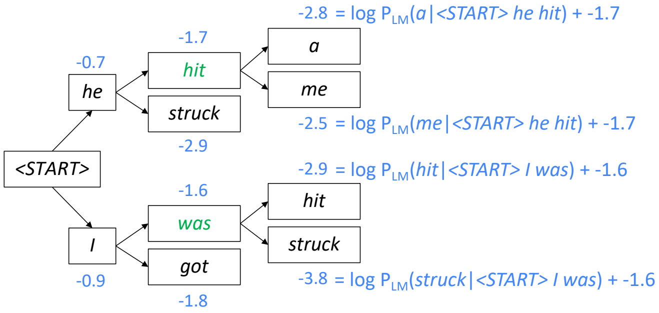


Of these k^2 hypotheses, just keep *k* with highest scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$



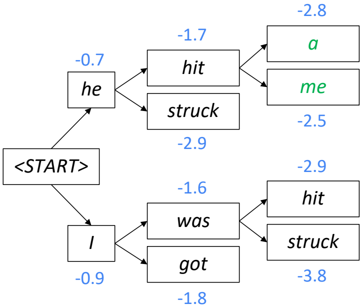
-2.8 = log P_LM(a|<START> he hit) + -1.7
-1.7
a
-0.7
hit
he
me
struck
-2.5 = log P_LM(me|<START> he hit) + -1.7
<START>
-2.9
-2.9 = log P_LM(hit|<START> I was) + -1.6
-1.6
hit
was
I
struck
got
-0.9
-3.8 = log P_LM(struck|<START> I was) + -1.6
-1.8

For each of the *k* hypotheses, find top *k* next words and calculate scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

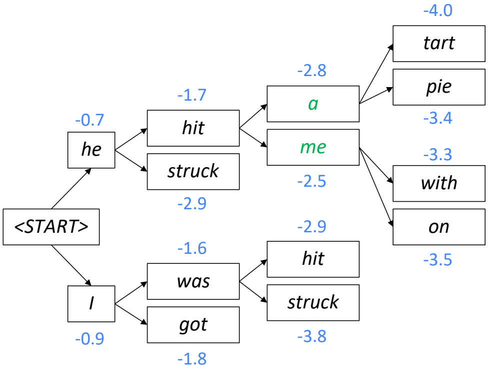


Of these k^2 hypotheses, just keep *k* with highest scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

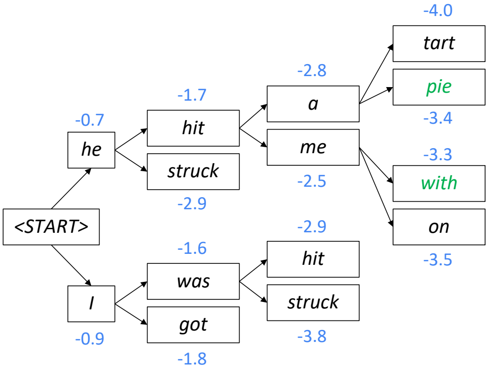


For each of the *k* hypotheses, find top *k* next words and calculate scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

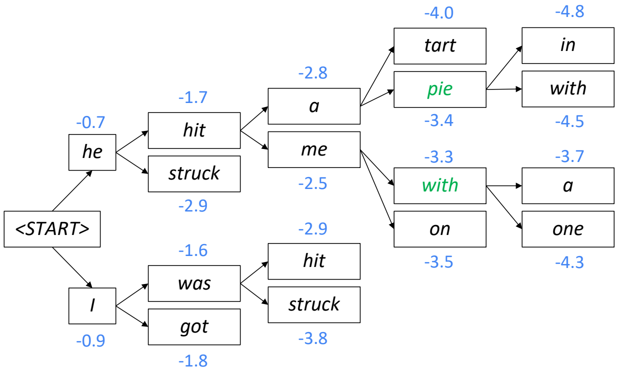


Of these k^2 hypotheses, just keep k with highest scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

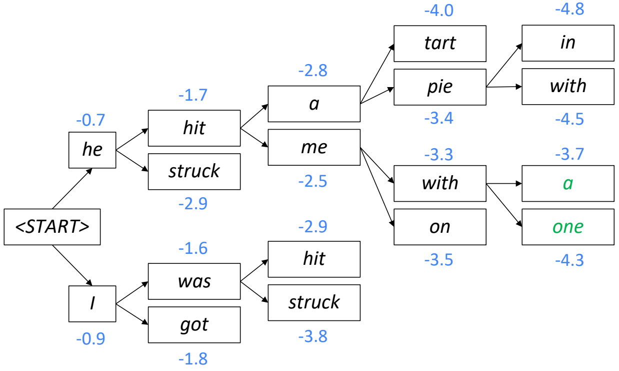


For each of the *k* hypotheses, find top *k* next words and calculate scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

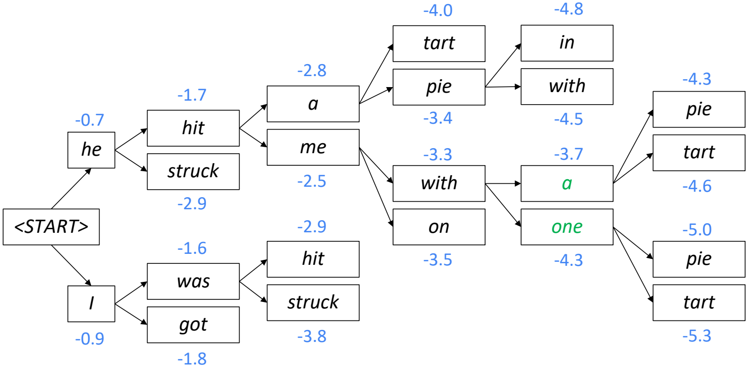


Of these k^2 hypotheses, just keep *k* with highest scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

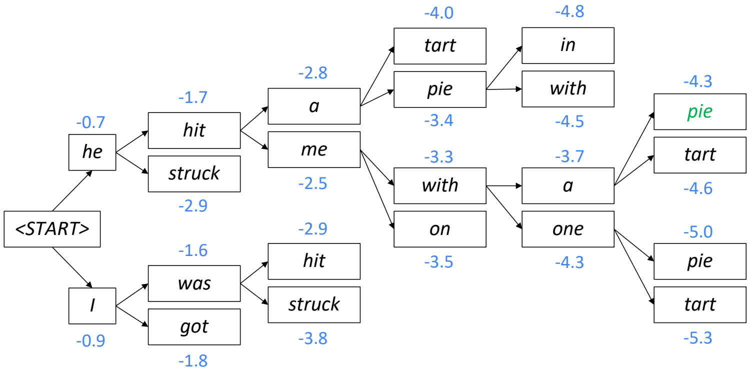


For each of the *k* hypotheses, find top *k* next words and calculate scores

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \ldots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$

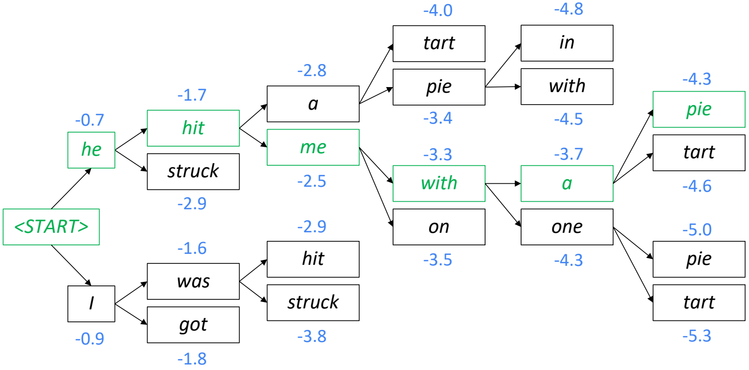


This is the top-scoring hypothesis!

Slide from Christopher Manning

Beam search decoding: example

Beam size = k = 2. Blue numbers = $\text{score}(y_1, \dots, y_t) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \dots, y_{i-1}, x)$



Backtrack to obtain the full hypothesis

Slide from Christopher Manning

Beam search decoding: stopping criterion

- In greedy decoding, usually we decode until the model produces an <END> token
  - **For example:** *<START> he hit me with a pie <END>*

- In beam search decoding, different hypotheses may produce <END> tokens on different timesteps
  - When a hypothesis produces <END>, that hypothesis is complete.
  - Place it aside and continue exploring other hypotheses via beam search.

- Usually we continue beam search until:
  - We reach timestep *T* (where *T* is some pre-defined cutoff), or
  - We have at least *n* completed hypotheses (where *n* is pre-defined cutoff)

Slide from Christopher Manning

Beam search decoding: finishing up

- We have our list of completed hypotheses.
- How to select top one?

- Each hypothesis $y_1, \ldots, y_t$ on our list has a score

$$\text{score}(y_1, \ldots, y_t) = \log P_{\text{LM}}(y_1, \ldots, y_t | x) = \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$$

- **Problem with this:** longer hypotheses have lower scores

- **Fix:** Normalize by length. Use this to select top one instead:

$$\frac{1}{t} \sum_{i=1}^{t} \log P_{\text{LM}}(y_i | y_1, \ldots, y_{i-1}, x)$$

See also discussion of sampling-based decoding in the NLG lecture

Slide from Christopher Manning

How do we evaluate Machine Translation?

BLEU (**Bil**ingual **E**valuation **U**nderstudy)

You'll see BLEU in detail in Assignment 4!

- BLEU compares the machine-written translation to one or several human-written translation(s), and computes a similarity score based on:
  - Geometric mean of *n*-gram precision (usually for 1, 2, 3 and 4-grams)
  - Plus a penalty for too-short system translations

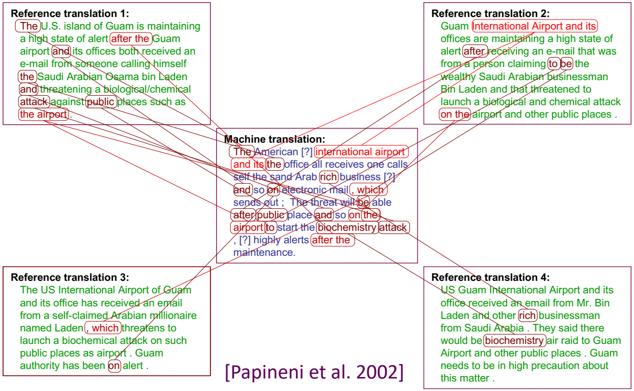
- BLEU is useful but imperfect
  - There are many valid ways to translate a sentence
  - So a good translation can get a poor BLEU score because it has low *n*-gram overlap with the human translation ☹

See discussion of evaluation in NLG lecture

Source: "BLEU: a Method for Automatic Evaluation of Machine Translation", Papineni et al, 2002. http://aclweb.org/anthology/P02-1040

Slide from Christopher Manning

BLEU score against 4 reference translations



Reference translation 1:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

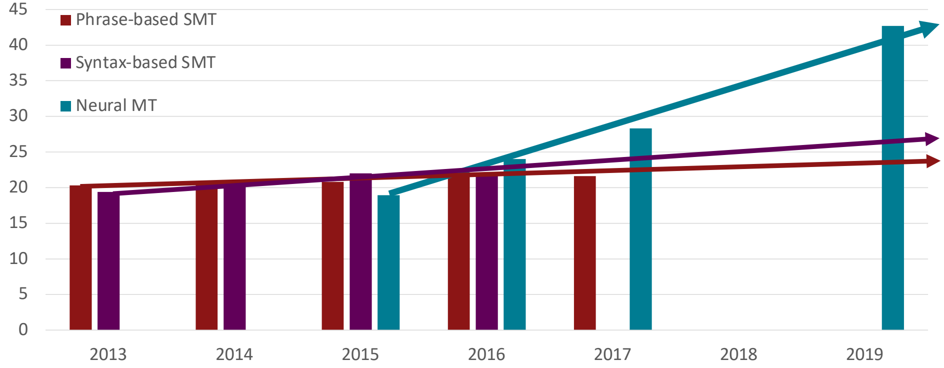
Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

[Papineni et al. 2002]

Slide from Christopher Manning

MT progress over time

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal; NMT 2019 FAIR on newstest2019]



Sources: http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf & http://matrix.statmt.org/

Slide from Christopher Manning

Advantages of NMT

Compared to SMT, NMT has many advantages:

- Better performance
  - More fluent
  - Better use of context
  - Better use of phrase similarities
- A single neural network to be optimized end-to-end
  - No subcomponents to be individually optimized
- Requires much less human engineering effort
  - No feature engineering
  - Same method for all language pairs

Slide from Christopher Manning

Disadvantages of NMT?

Compared to SMT:

- NMT is less interpretable
  - Hard to debug

- NMT is difficult to control
  - For example, can't easily specify rules or guidelines for translation
  - Safety concerns!
    - Invention of content not in source
    - Systematic gender biases

Slide from Christopher Manning

NMT: the first big success story of NLP Deep Learning

Neural Machine Translation went from a fringe research attempt in **2014** to the leading standard method in **2016**

- **2014**: First seq2seq paper published [Sutskever et al. 2014]
- **2016**: Google Translate switches from SMT to NMT – and by 2018 everyone has

Microsoft SYSTRAN beyond language Google facebook

Bai du 百度 網易 NetEase www.163.com Tencent 腾讯 搜狗搜索

- This is amazing!
  - **SMT** systems, built by hundreds of engineers over many years, outperformed by NMT systems trained by small groups of engineers in a few months

Slide from Christopher Manning

Summary of Sequence 2 Sequence NMT

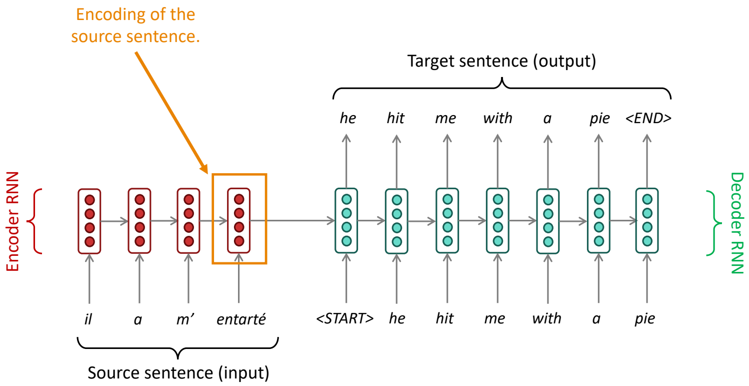
- Seq2seq NMT models model the probability of the target sentence conditioned on the source sentence
- Seq2seq models have two components
  - an encoder model converts the input sequence into a vector (or vectors)
  - a decoder model generates the output sequence conditioned on the encoding vector (or vectors)
- Searching the space of output sequences (also called “decoding”) can be done with beam search

Outline

Attention in NMT: Idea

- For long sentences, a fixed-length vector encoding introduces a bottleneck.
- Even for shorter sentences, conditioning on the entire sentence is hard.
- Solution: reintroduce a model of latent **alignment**, as in SMT.
- **Attention** is a soft latent alignment.

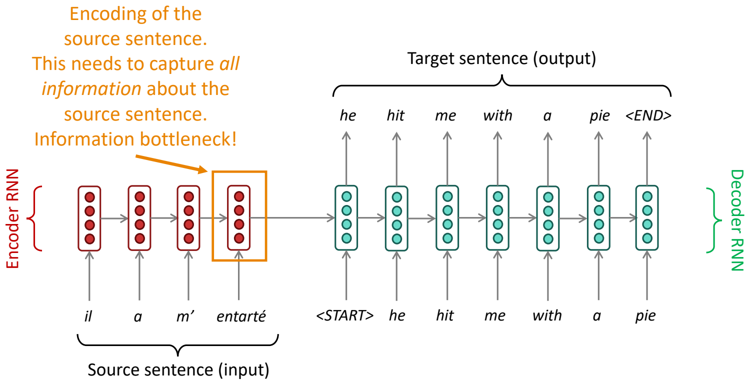
2. Why attention? Sequence-to-sequence: the bottleneck problem



Problems with this architecture?

Slide from Christopher Manning

1. Why attention? Sequence-to-sequence: the bottleneck problem



Encoding of the source sentence. This needs to capture *all information* about the source sentence. Information bottleneck!
Target sentence (output)
he
hit
me
with
a
pie
<END>
Encoder RNN
Decoder RNN
il
a
m'
entarté
<START>
he
hit
me
with
a
pie
Source sentence (input)

Slide from Christopher Manning

Attention

- **Attention** provides a solution to the bottleneck problem.

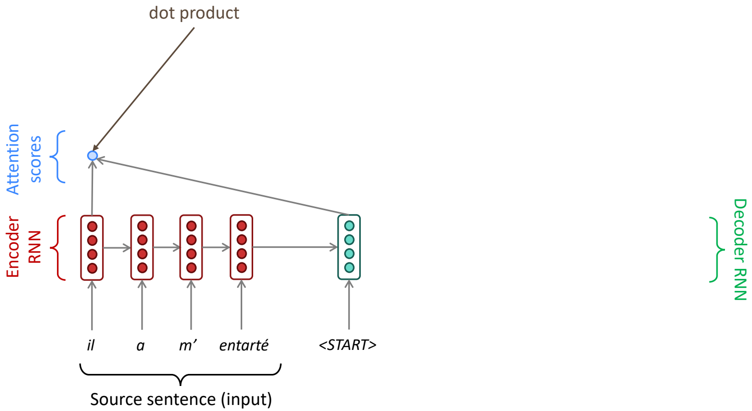
- **Core idea**: on each step of the decoder, *use direct connection to the encoder* to *focus on a particular part* of the source sequence

- First, we will show via diagram (no equations), then we will show with equations

Slide from Christopher Manning

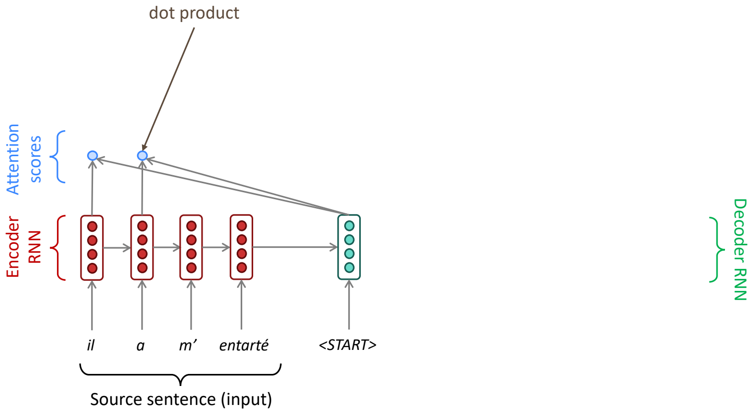
Sequence-to-sequence with attention

Core idea: on each step of the decoder, *use direct connection to the encoder* to *focus on a particular part* of the source sequence



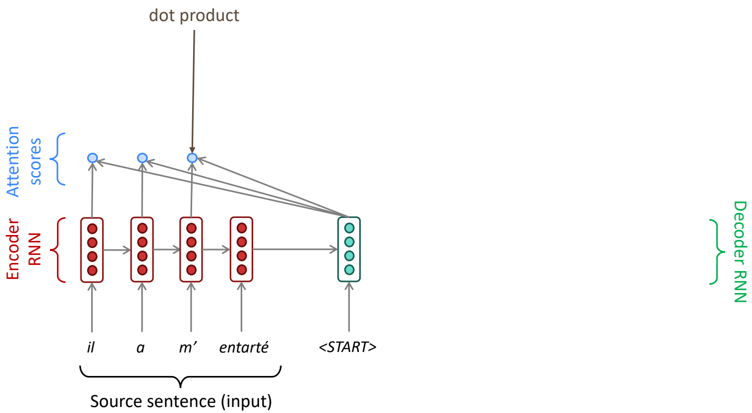
Slide from Christopher Manning

Sequence-to-sequence with attention



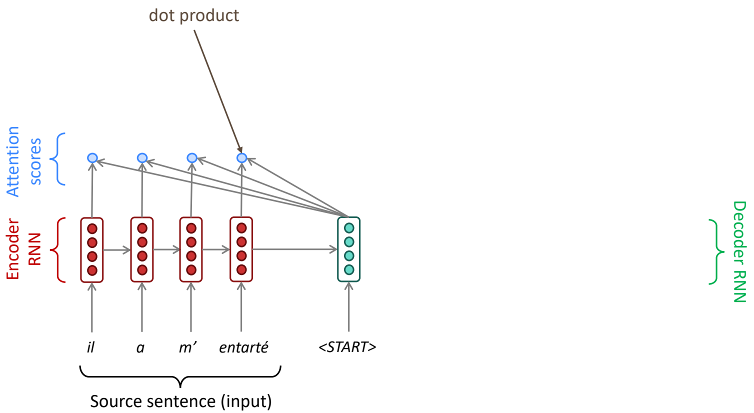
Slide from Christopher Manning

Sequence-to-sequence with attention



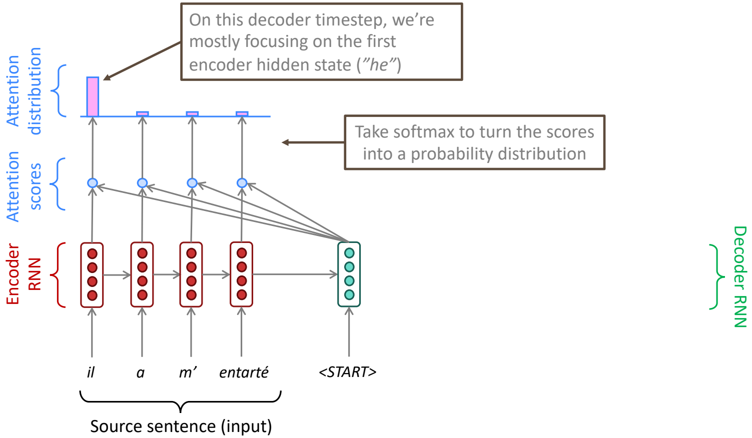
Slide from Christopher Manning

Sequence-to-sequence with attention



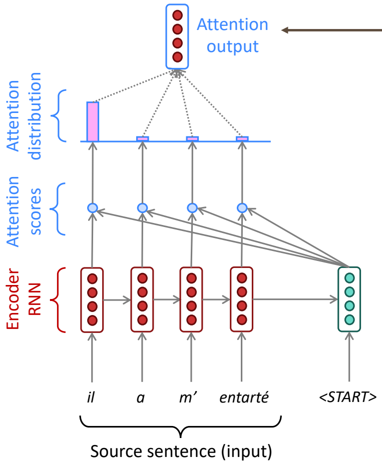
Slide from Christopher Manning

Sequence-to-sequence with attention


On this decoder timestep, we're mostly focusing on the first encoder hidden state ("he")
Take softmax to turn the scores into a probability distribution
Attention distribution
Attention scores
Encoder RNN
Decoder RNN
il
a
m'
entarté
<START>
Source sentence (input)

Slide from Christopher Manning

Sequence-to-sequence with attention

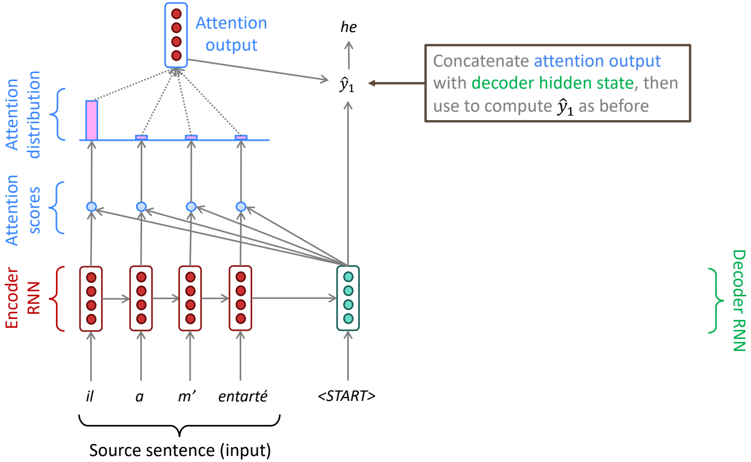


Use the attention distribution to take a **weighted sum** of the encoder hidden states.

The attention output mostly contains information from the hidden states that received high attention.

Slide from Christopher Manning

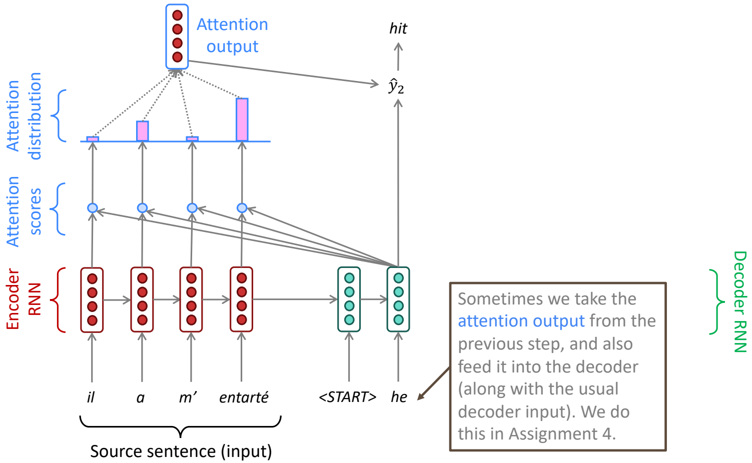
Sequence-to-sequence with attention



Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

Slide from Christopher Manning

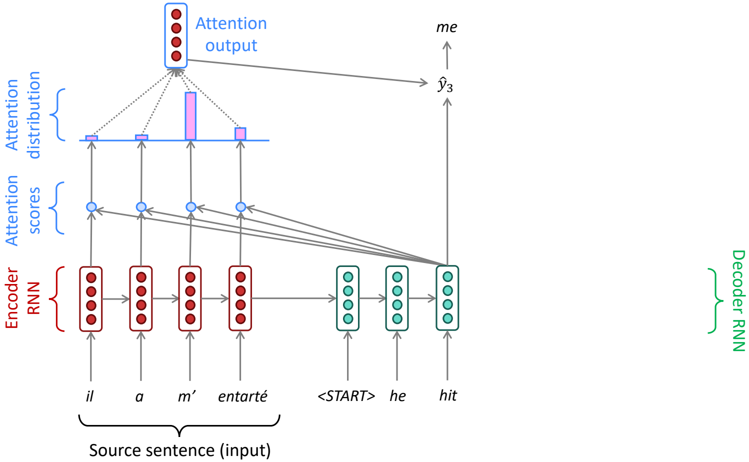
Sequence-to-sequence with attention



Sometimes we take the attention output from the previous step, and also feed it into the decoder (along with the usual decoder input). We do this in Assignment 4.

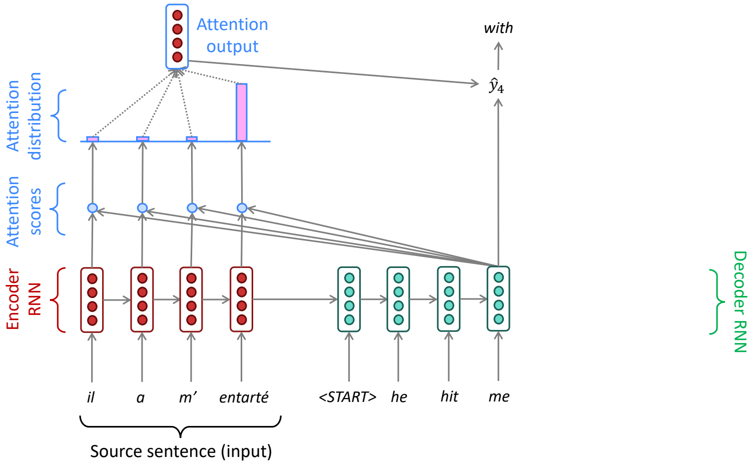
Slide from Christopher Manning

Sequence-to-sequence with attention



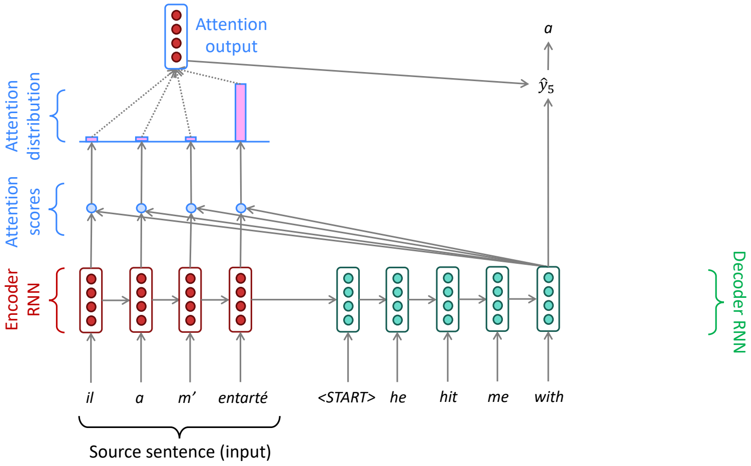
Slide from Christopher Manning

Sequence-to-sequence with attention



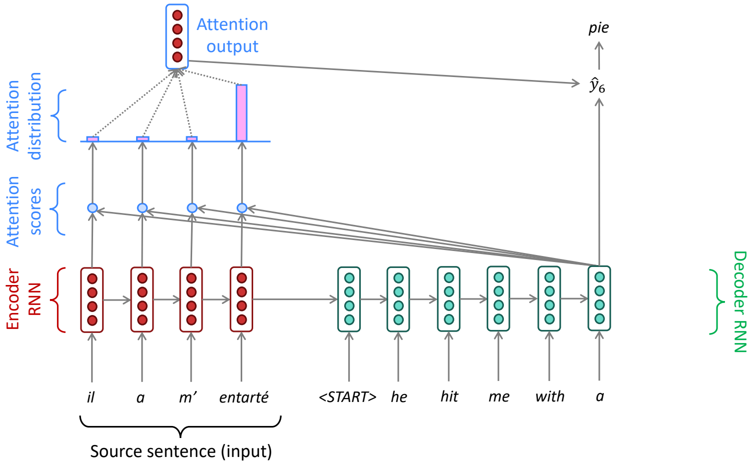
Slide from Christopher Manning

Sequence-to-sequence with attention



Slide from Christopher Manning

Sequence-to-sequence with attention



Slide from Christopher Manning

Attention: in equations

- We have encoder hidden states $h_1, \ldots, h_N \in \mathbb{R}^h$
- On timestep *t*, we have decoder hidden state $s_t \in \mathbb{R}^h$
- We get the attention scores $\boldsymbol{e}^t$ for this step:

$$\boldsymbol{e}^t = [\boldsymbol{s}_t^T \boldsymbol{h}_1, \ldots, \boldsymbol{s}_t^T \boldsymbol{h}_N] \in \mathbb{R}^N$$

- We take softmax to get the attention distribution α^t for this step (this is a probability distribution and sums to 1)

$$\alpha^t = \mathrm{softmax}(\boldsymbol{e}^t) \in \mathbb{R}^N$$

- We use α^t to take a weighted sum of the encoder hidden states to get the attention output $\boldsymbol{a}_t$

$$\boldsymbol{a}_t = \sum_{i=1}^{N} \alpha_i^t \boldsymbol{h}_i \in \mathbb{R}^h$$

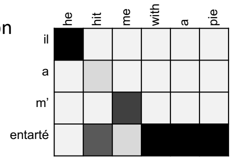
- Finally we concatenate the attention output $\boldsymbol{a}_t$ with the decoder hidden state s_t and proceed as in the non-attention seq2seq model

$$[\boldsymbol{a}_t; \boldsymbol{s}_t] \in \mathbb{R}^{2h}$$

Slide from Christopher Manning

Attention is great!

- Attention significantly improves NMT performance
  - It's very useful to allow decoder to focus on certain parts of the source
- Attention provides a more "human-like" model of the MT process
  - You can look back at the source sentence while translating, rather than needing to remember it all
- Attention solves the bottleneck problem
  - Attention allows decoder to look directly at source; bypass bottleneck
- Attention helps with the vanishing gradient problem
  - Provides shortcut to faraway states
- Attention provides some interpretability
  - By inspecting attention distribution, we see what the decoder was focusing on
  - We get (soft) alignment for free!
  - This is cool because we never explicitly trained an alignment system
  - The network just learned alignment by itself



Slide from Christopher Manning

There are *several* attention variants

- We have some *values* $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N \in \mathbb{R}^{d_1}$ and a *query* $\boldsymbol{s} \in \mathbb{R}^{d_2}$

- Attention always involves:
  1. Computing the *attention scores* $\boldsymbol{e} \in \mathbb{R}^N$ ← There are multiple ways to do this
  2. Taking softmax to get *attention distribution* α:

$$\alpha = \mathrm{softmax}(\boldsymbol{e}) \in \mathbb{R}^N$$

  3. Using attention distribution to take weighted sum of values:

$$\boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \boldsymbol{h}_i \in \mathbb{R}^{d_1}$$

thus obtaining the *attention output* ***a*** (sometimes called the *context vector*)

Slide from Christopher Manning

Attention variants

You'll think about the relative advantages/disadvantages of these in Assignment 4!

There are several ways you can compute $\boldsymbol{e} \in \mathbb{R}^N$ from $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N \in \mathbb{R}^{d_1}$ and $\boldsymbol{s} \in \mathbb{R}^{d_2}$:

- Basic dot-product attention: $\boldsymbol{e}_i = \boldsymbol{s}^T \boldsymbol{h}_i \in \mathbb{R}$
  - Note: this assumes $d_1 = d_2$. This is the version we saw earlier.
- Multiplicative attention: $\boldsymbol{e}_i = \boldsymbol{s}^T \boldsymbol{W} \boldsymbol{h}_i \in \mathbb{R}$ [Luong, Pham, and Manning 2015]
  - Where $\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}$ is a weight matrix. Perhaps better called "bilinear attention"
- Reduced-rank multiplicative attention: $e_i = s^T(\boldsymbol{U}^T\boldsymbol{V})h_i = (\boldsymbol{U}s)^T(\boldsymbol{V}h_i)$ $\leftarrow$ Remember this when we look at Transformers next week!
  - For low rank matrices $\boldsymbol{U} \in \mathbb{R}^{k \times d_2}$, $\boldsymbol{V} \in \mathbb{R}^{k \times d_1}$, $k \ll d_1, d_2$
- Additive attention: $\boldsymbol{e}_i = \boldsymbol{v}^T \tanh(\boldsymbol{W}_1 \boldsymbol{h}_i + \boldsymbol{W}_2 \boldsymbol{s}) \in \mathbb{R}$ [Bahdanau, Cho, and Bengio 2014]
  - Where $\boldsymbol{W}_1 \in \mathbb{R}^{d_3 \times d_1}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d_3 \times d_2}$ are weight matrices and $\boldsymbol{v} \in \mathbb{R}^{d_3}$ is a weight vector.
  - d_3 (the attention dimensionality) is a hyperparameter
  - "Additive" is a weird/bad name. It's really using a feed-forward neural net layer.

More information: "Deep Learning for NLP Best Practices", Ruder, 2017. http://ruder.io/deep-learning-nlp-best-practices/index.html#attention
"Massive Exploration of Neural Machine Translation Architectures", Britz et al, 2017, https://arxiv.org/pdf/1703.03906.pdf

Slide from Christopher Manning

Query-Key-Value Attention

- Given a sequence-of-vectors $\langle h_1, \ldots, h_N \rangle$ and a state vector s_t,
- and three parameter matrices W^q, W^k, W^v,
- $$\begin{aligned} e_i^t &= (W^q s_t)^T W^k h_i \\ \alpha^t &= \mathit{softmax}(e^t) \\ a_t &= \sum_{i=1}^{N} \alpha_i^t W^v h_i \end{aligned}$$

Attention function:

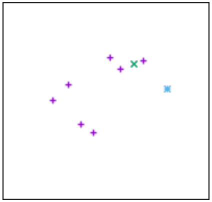
- permutation invariant, so $\langle h_1, \ldots, h_N \rangle$ is a set
- size invariant, so $\langle h_1, \ldots, h_N \rangle$ is unbounded
- normalised weighting, so $\langle \alpha_1^t, \ldots, \alpha_N^t \rangle$ is a distribution

Understanding Attention

- Attention function is **permutation invariant** in the vectors

$$Attn(\boldsymbol{u}, \boldsymbol{Z}) = \sum_{i=1}^{n} a_i \boldsymbol{z}_i$$

$$a_i = \frac{\exp(\frac{1}{\sqrt{d}} \boldsymbol{u}\boldsymbol{z}_i)}{\sum_{i=1}^{n} \exp(\frac{1}{\sqrt{d}} \boldsymbol{u}\boldsymbol{z}_i)}$$

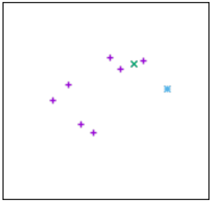


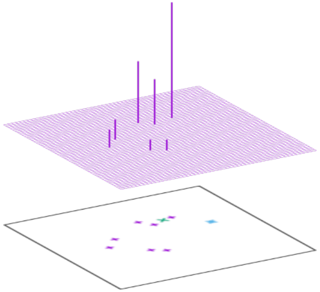
Understanding Attention

- Attention function is **permutation invariant** in the vectors
- Attention imposes a **normalised weighting** over vectors

$$Attn(\mathbf{u}, \mathbf{Z}) = \sum_{i=1}^{n} a_i \mathbf{z}_i$$

$$a_i = \frac{\exp(\frac{1}{\sqrt{d}} \mathbf{u}\mathbf{z}_i)}{\sum_{i=1}^{n} \exp(\frac{1}{\sqrt{d}} \mathbf{u}\mathbf{z}_i)}$$



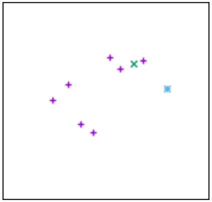


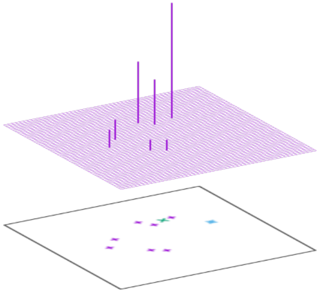
Understanding Attention

- Attention function is **permutation invariant** in the vectors
- Attention imposes a **normalised weighting** over vectors
- Attention supports a **variable number** of vectors

$$Attn(\mathbf{u}, \mathbf{Z}) = \sum_{i=1}^{n} a_i \mathbf{z}_i$$

$$a_i = \frac{\exp(\frac{1}{\sqrt{d}} \mathbf{u}\mathbf{z}_i)}{\sum_{i=1}^{n} \exp(\frac{1}{\sqrt{d}} \mathbf{u}\mathbf{z}_i)}$$





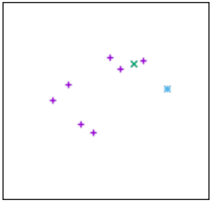
Understanding Attention

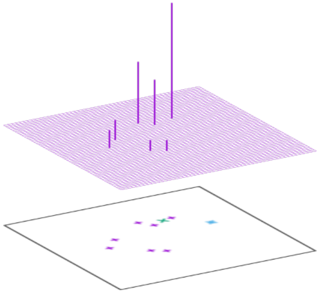
- Attention function is **permutation invariant** in the vectors
- Attention imposes a **normalised weighting** over vectors
- Attention supports a **variable number** of vectors

Like a **nonparametric mixture of impulse distributions**

$$Attn(\boldsymbol{u}, \boldsymbol{Z}) = \sum_{i=1}^{n} a_i \boldsymbol{z}_i$$

$$a_i = \frac{\exp(\frac{1}{\sqrt{d}} \boldsymbol{u}\boldsymbol{z}_i)}{\sum_{i=1}^{n} \exp(\frac{1}{\sqrt{d}} \boldsymbol{u}\boldsymbol{z}_i)}$$





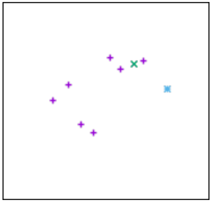
Understanding Attention

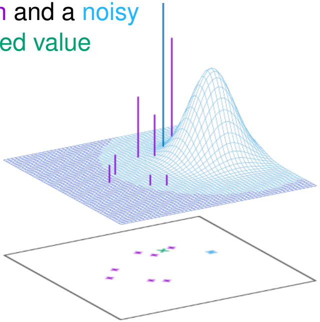
The attention function is **query denoising** with a **nonparametric mixture of impulse distributions**

- Attention takes a sequence of vectors and a query vector and returns an attention vector
- Denoising takes a prior distribution and a noisy observation and returns its expected value

$$Attn(\boldsymbol{u}, \boldsymbol{Z}) = \sum_{i=1}^{n} a_i \boldsymbol{z}_i$$

$$a_i = \frac{\exp(\frac{1}{\sqrt{d}} \boldsymbol{u}\boldsymbol{z}_i)}{\sum_{i=1}^{n} \exp(\frac{1}{\sqrt{d}} \boldsymbol{u}\boldsymbol{z}_i)}$$





Attention is a *general* Deep Learning technique

- We've seen that attention is a great way to improve the sequence-to-sequence model for Machine Translation.
- However: You can use attention in many architectures (not just seq2seq) and many tasks (not just MT)

- **More general definition of attention**:
  - Given a set of vector *values*, and a vector *query*, **attention** is a technique to compute a weighted sum of the values, dependent on the query.

- We sometimes say that the query *attends to* the values.
- For example, in the seq2seq + attention model, each decoder hidden state (query) *attends to* all the encoder hidden states (values).

Slide from Christopher Manning

Attention is a *general* Deep Learning technique

- **More general definition of attention**:
  - Given a set of vector *values*, and a vector *query*, **attention** is a technique to compute a weighted sum of the values, dependent on the query.

Intuition:

- The weighted sum is a *selective summary* of the information contained in the values, where the query determines which values to focus on.
- Attention is a way to obtain a *fixed-size representation of an arbitrary set of representations* (the values), dependent on some other representation (the query).

Upshot:

- Attention has become the powerful, flexible, general way pointer and memory manipulation in all deep learning models. A new idea from after 2010! From NMT!

Slide from Christopher Manning

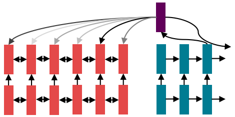
Summary of Attention in NMT

- Attention in NMT learns a soft alignment between output and input tokens
- Attention uses a (non-parametric) **set-of-vector** representation, instead of a (parametric) vector representation, which is more appropriate for representing language
- Attention accesses vectors in the set based only on their content
- Attention is very effective whenever conditioning on (arbitrarily long) text

Outline

Today: Same goals, different building blocks

- Last week, we learned about sequence-to-sequence problems and encoder-decoder models.
- Today, we're **not** trying to motivate entirely new ways of looking at problems (like Machine Translation)
- Instead, we're trying to find the best **building blocks** to plug into our models and enable broad progress.



$$\xrightarrow[\text{and error}]{\text{Lots of trial}}$$

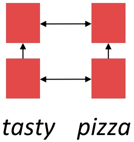
2014-2017ish
Recurrence

2021
??????

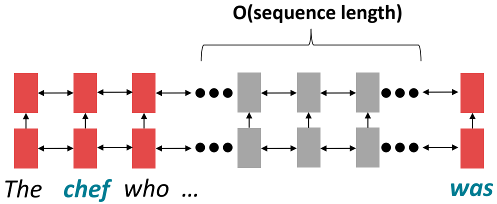
Slide from John Hewitt

Issues with recurrent models: **Linear interaction distance**

- RNNs are unrolled "left-to-right".
- This encodes linear locality: a useful heuristic!
  - Nearby words often affect each other's meanings



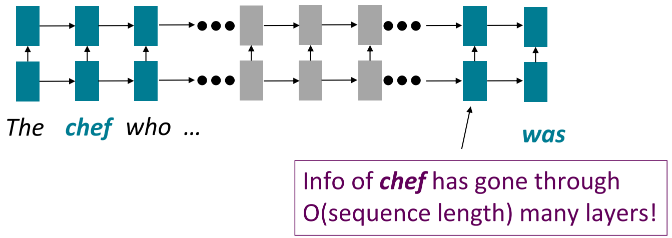
- **Problem:** RNNs take **O(sequence length)** steps for distant word pairs to interact.



Slide from John Hewitt

Issues with recurrent models: **Linear interaction distance**

- **O(sequence length)** steps for distant word pairs to interact means:
  - Hard to learn long-distance dependencies (because gradient problems!)
  - Linear order of words is "baked in"; we already know linear order isn't the right way to think about sentences…

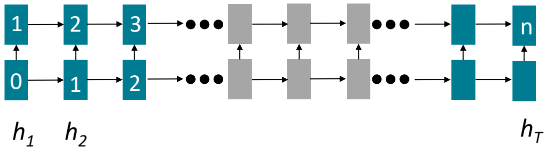


The
chef
who
…
was
Info of ***chef*** has gone through O(sequence length) many layers!

Slide from John Hewitt

Issues with recurrent models: **Lack of parallelizability**

- Forward and backward passes have **O(sequence length)** unparallelizable operations
  - GPUs can perform a bunch of independent computations at once!
  - But future RNN hidden states can't be computed in full before past RNN hidden states have been computed
  - Inhibits training on very large datasets!

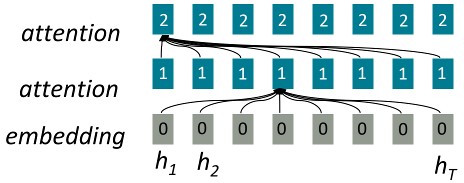


Numbers indicate min # of steps before a state can be computed

Slide from John Hewitt

If not recurrence, then what? **How about attention?**

- **Attention** treats each word's representation as a **query** to access and incorporate information from **a set of values.**
  - We saw attention from the **decoder** to the **encoder**; today we'll think about attention **within a single sentence**.
- Number of unparallelizable operations does not increase with sequence length.
- Maximum interaction distance: O(1), since all words interact at every layer!



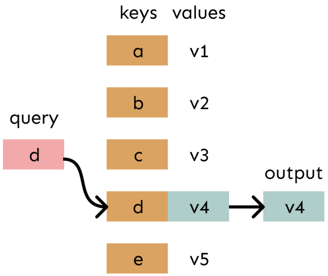
All words attend to all words in previous layer; most arrows here are omitted

Slide from John Hewitt

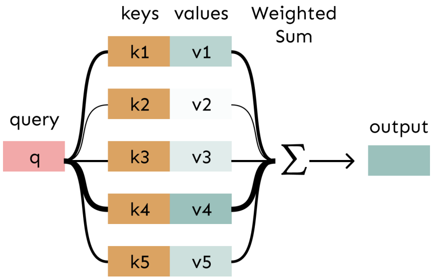
Attention as a soft, averaging lookup table

We can think of **attention** as performing fuzzy lookup in a key-value store.

In a **lookup table**, we have a table of **keys** that map to **values**. The **query** matches one of the keys, returning its value.

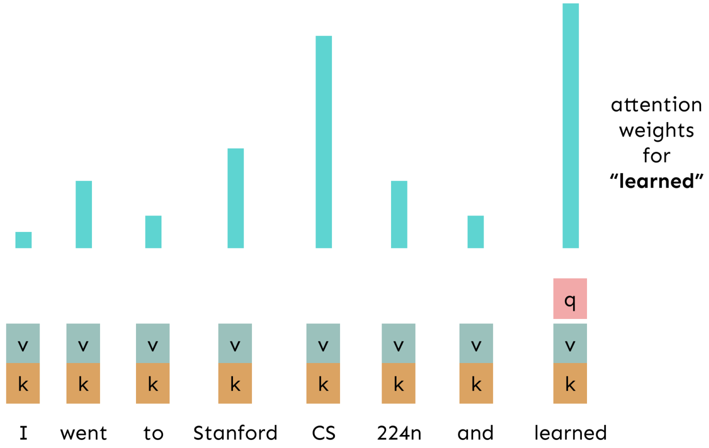


In **attention**, the **query** matches all **keys** *softly*, to a weight between 0 and 1. The keys' **values** are multiplied by the weights and summed.



Slide from John Hewitt

Self-Attention Hypothetical Example



Slide from John Hewitt

Self-Attention: keys, queries, values from the same sequence

Let $\boldsymbol{w}_{1:n}$ be a sequence of words in vocabulary V, like *Zuko made his uncle tea*.

For each $\boldsymbol{w}_i$, let $\boldsymbol{x}_i = E\boldsymbol{w}_{\boldsymbol{i}}$, where $E \in \mathbb{R}^{d\times|V|}$ is an embedding matrix.

1. Transform each word embedding with weight matrices Q, K, V , each in $\mathbb{R}^{d\times d}$

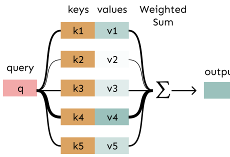
$$\boldsymbol{q}_i = Q\boldsymbol{x}_{\boldsymbol{i}} \text{ (queries)} \qquad \boldsymbol{k}_i = K\boldsymbol{x}_{\boldsymbol{i}} \text{ (keys)} \qquad \boldsymbol{v}_i = V\boldsymbol{x}_{\boldsymbol{i}} \text{ (values)}$$

2. Compute pairwise similarities between keys and queries; normalize with softmax

$$\boldsymbol{e}_{ij} = \boldsymbol{q}_{\boldsymbol{i}}^{\top}\boldsymbol{k}_{\boldsymbol{j}} \qquad \boldsymbol{\alpha}_{ij} = \frac{\exp(\boldsymbol{e}_{ij})}{\sum_{j'}\exp(\boldsymbol{e}_{ij'})}$$

3. Compute output for each word as weighted sum of values

$$\boldsymbol{o}_i = \sum_{\boldsymbol{j}} \boldsymbol{\alpha}_{ij}\,\boldsymbol{v}_i$$



Slide from John Hewitt

Barriers and solutions for Self-Attention as a building block

Barriers

- Doesn't have an inherent notion of order! $\longrightarrow$

Solutions

Slide from John Hewitt

Fixing the first self-attention problem: **sequence order**

- Since self-attention doesn't build in order information, we need to encode the order of the sentence in our keys, queries, and values.
- Consider representing each **sequence index** as a **vector**

$$\boldsymbol{p}_i \in \mathbb{R}^d, \text{ for } i \in \{1,2,\ldots,n\} \text{ are position vectors}$$

- Don't worry about what the p_i are made of yet!
- Easy to incorporate this info into our self-attention block: just add the $\boldsymbol{p}_i$ to our inputs!
- Recall that $\boldsymbol{x}_i$ is the embedding of the word at index i. The positioned embedding is:

$$\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \boldsymbol{p}_i$$

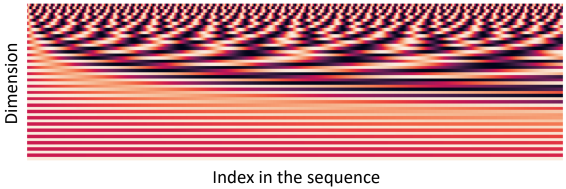
In deep self-attention networks, we do this at the first layer! You could concatenate them as well, but people mostly just add…

Slide from John Hewitt

Position representation vectors through sinusoids

- **Sinusoidal position representations:** concatenate sinusoidal functions of varying periods:

$$\boldsymbol{p}_i = \begin{pmatrix} \sin(i/10000^{2*1/d}) \\ \cos(i/10000^{2*1/d}) \\ \vdots \\ \sin(i/10000^{2*\frac{d}{2}/d}) \\ \cos(i/10000^{2*\frac{d}{2}/d}) \end{pmatrix}$$



- Pros:
  - Periodicity indicates that maybe "absolute position" isn't as important
  - Maybe can extrapolate to longer sequences as periods restart!
- Cons:
  - Not learnable; also the extrapolation doesn't really work!

Image: https://timodenk.com/blog/linear-relationships-in-the-transformers-positional-encodin

Slide from John Hewitt

Position representation vectors learned from scratch

- **Learned absolute position representations:** Let all p_i be learnable parameters!
  Learn a matrix $\boldsymbol{p} \in \mathbb{R}^{d \times n}$*, and let each* $\boldsymbol{p}_i$ *be a column of that matrix!*

- Pros:
  - Flexibility: each position gets to be learned to fit the data
- Cons:
  - Definitely can't extrapolate to indices outside $1, \dots, n$.
- Most systems use this!

- Sometimes people try more flexible representations of position:
  - Relative linear position attention [Shaw et al., 2018]
  - Dependency syntax-based position [Wang et al., 2019]

Slide from John Hewitt

Barriers and solutions for Self-Attention as a building block

Barriers

- Doesn't have an inherent notion of order! $\longrightarrow$
- No nonlinearities for deep learning! It's all just weighted averages $\longrightarrow$

Solutions

- Add position representations to the inputs

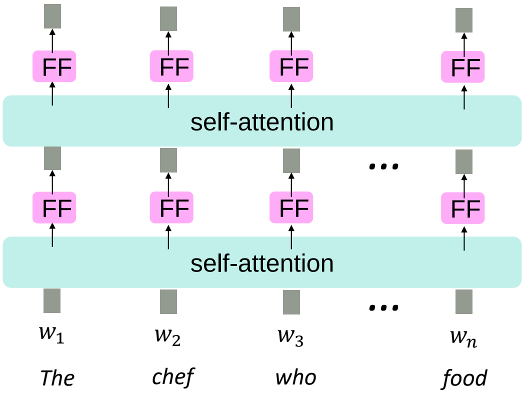
Slide from John Hewitt

Adding nonlinearities in self-attention

- Note that there are no elementwise nonlinearities in self-attention; stacking more self-attention layers just re-averages **value** vectors (Why? Look at the notes!)

- Easy fix: add a **feed-forward network** to post-process each output vector.

$$\begin{aligned} m_i &= MLP(\text{output}_i) \\ &= W_2 * \text{ReLU}(W_1 \text{ output}_i + b_1) + b_2 \end{aligned}$$



Intuition: the FF network processes the result of attention

Slide from John Hewitt

Barriers and solutions for Self-Attention as a building block

Barriers		Solutions
• Doesn't have an inherent notion of order!	$\longrightarrow$	• Add position representations to the inputs
• No nonlinearities for deep learning magic! It's all just weighted averages	$\longrightarrow$	• Easy fix: apply the same feedforward network to each self-attention output.
• Need to ensure we don't "look at the future" when predicting a sequence • Like in machine translation • Or language modeling	$\longrightarrow$	

Slide from John Hewitt

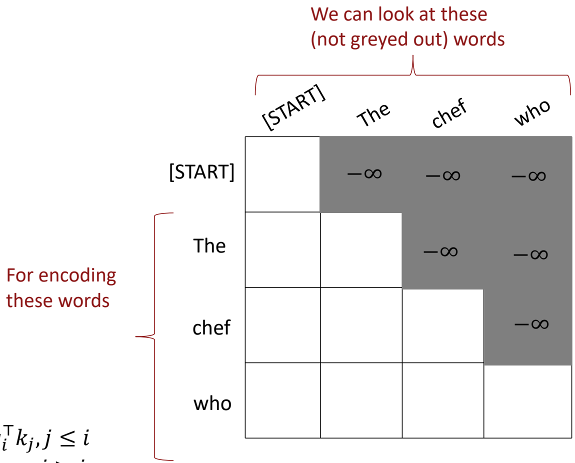
Masking the future in self-attention

- To use self-attention in **decoders**, we need to ensure we can't peek at the future.

- At every timestep, we could change the set of **keys and queries** to include only past words. (Inefficient!)

- To enable parallelization, we **mask out attention** to future words by setting attention scores to $-\infty$.

$$e_{ij} = \begin{cases} q_i^\top k_j, j \le i \\ -\infty, j > i \end{cases}$$



$$\begin{array}{c|cccc}
 & \text{[START]} & \text{The} & \text{chef} & \text{who} \\
\hline
\text{[START]} & & -\infty & -\infty & -\infty \\
\text{The} & & & -\infty & -\infty \\
\text{chef} & & & & -\infty \\
\text{who} & & & &
\end{array}$$

Slide from John Hewitt

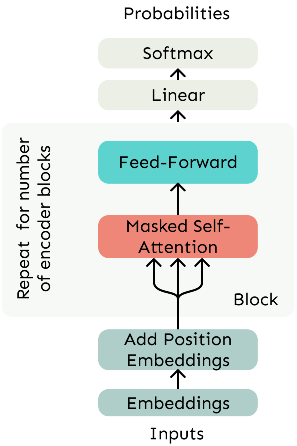
Barriers and solutions for Self-Attention as a building block

Barriers		Solutions
• Doesn't have an inherent notion of order!	$\longrightarrow$	• Add position representations to the inputs
• No nonlinearities for deep learning magic! It's all just weighted averages	$\longrightarrow$	• Easy fix: apply the same feedforward network to each self-attention output.
• Need to ensure we don't "look at the future" when predicting a sequence • Like in machine translation • Or language modeling	$\longrightarrow$	• Mask out the future by artificially setting attention weights to 0!

Slide from John Hewitt

Necessities for a self-attention building block:

- **Self-attention**:
  - the basis of the method.
- **Position representations**:
  - Specify the sequence order, since self-attention is an unordered function of its inputs.
- **Nonlinearities**:
  - At the output of the self-attention block
  - Frequently implemented as a simple feed-forward network.
- **Masking**:
  - In order to parallelize operations while not looking at the future.
  - Keeps information about the future from "leaking" to the past.



Slide from John Hewitt

Summary of Attention instead of Recurrence

- Attention is all you need
- Plus a represention of sequence order, with absolute (or relative) positions
- Plus layers of nonlinearity, for a fixed number of layers
- Plus causal masking, to similate running multiple models on the same computation graph