May I Have Your Attention Please?

Lecture 5.3: Introduction to Attention Mechanisms

NLP + DL

RECAP

From Last Lecture

Learning Objectives

Be able to answer:

- What is an attention mechanism?
- Difference between soft and hard attention?

- What are some examples of attention?

- How does attention relate to interpretability?

Consider Machine Translation

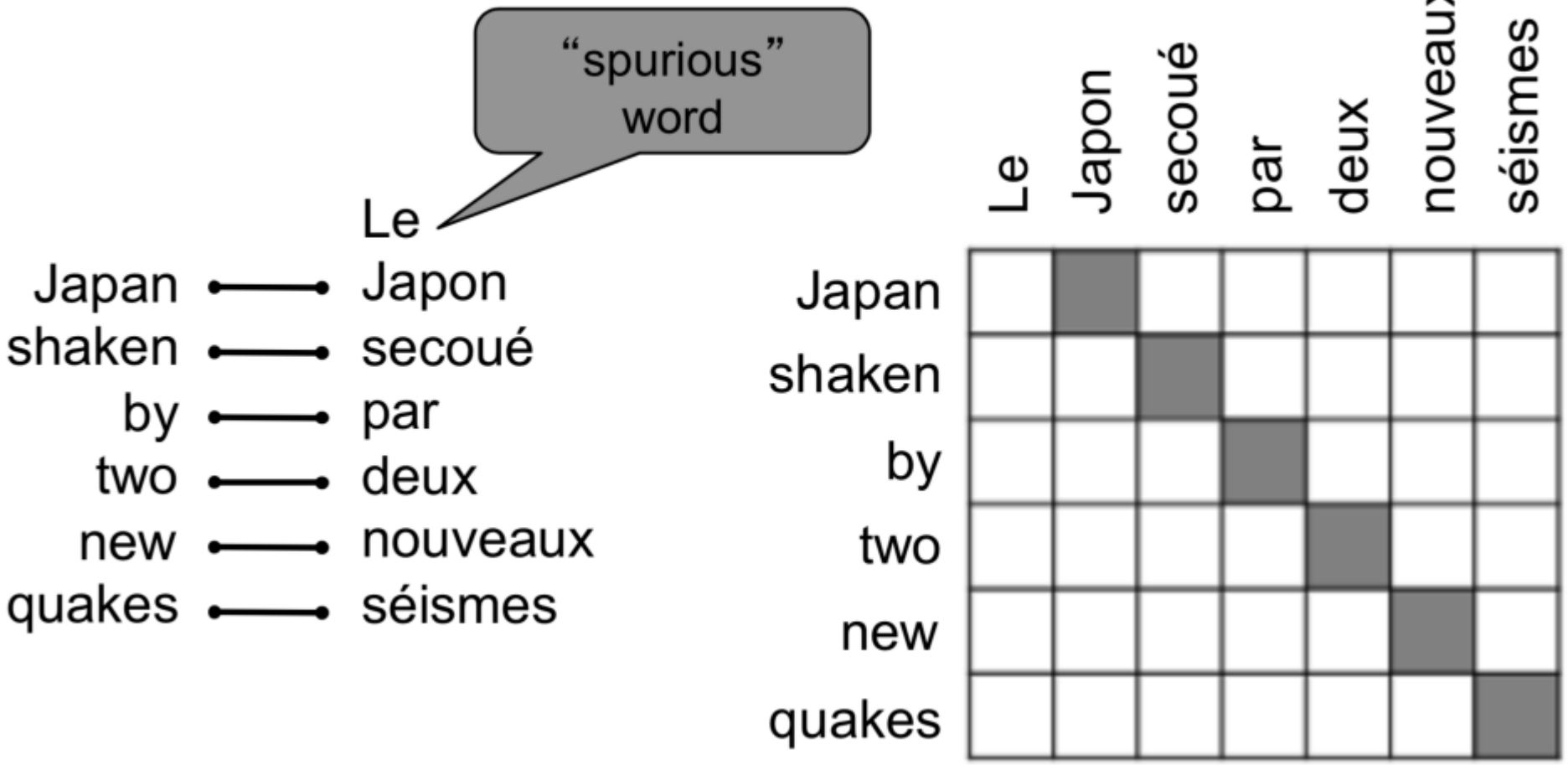
Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes.

Slide adapted from Chris Manning. Example from "The Mathematics of Statistical Machine Translation: Parameter Estimation, 1993.

Consider Machine Translation

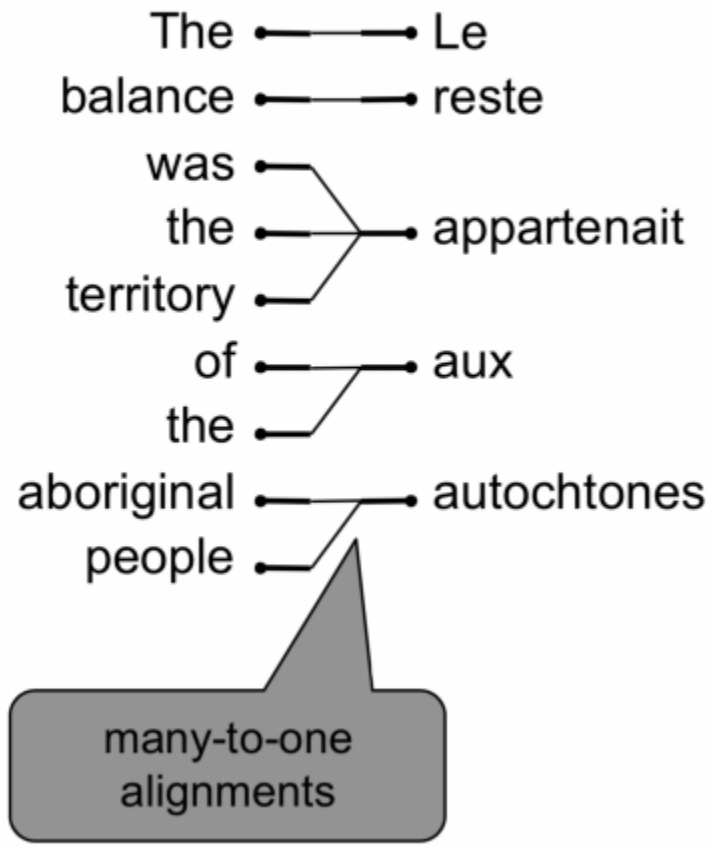
Some words have no alignment in the source language

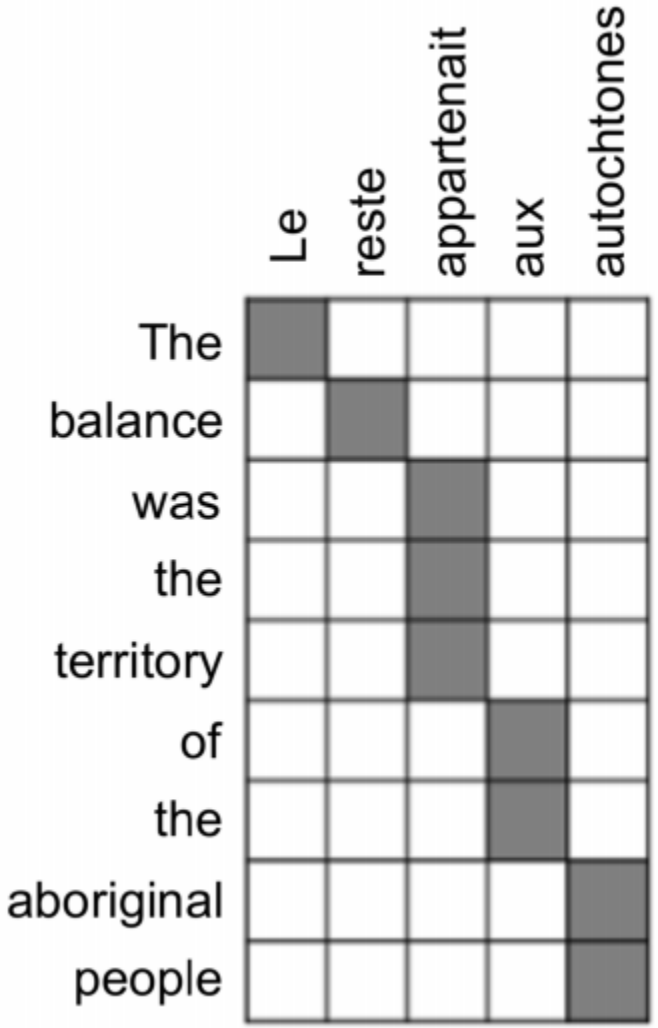


Slide adapted from Chris Manning. Example from "The Mathematics of Statistical Machine Translation: Parameter Estimation, 1993.

Consider Machine Translation

Some alignments are many-to-one

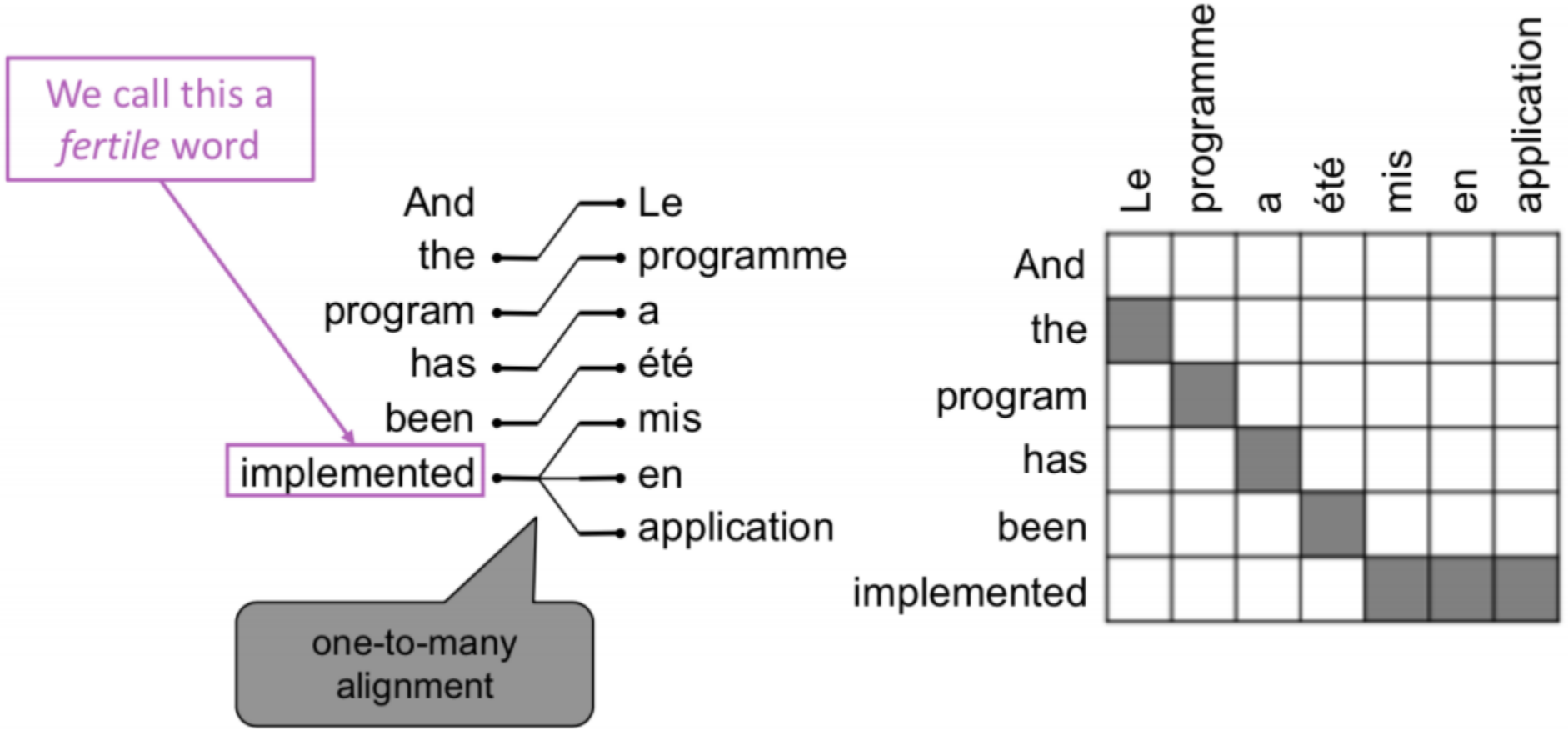




Slide adapted from Chris Manning. Example from "The Mathematics of Statistical Machine Translation: Parameter Estimation, 1993.

Consider Machine Translation

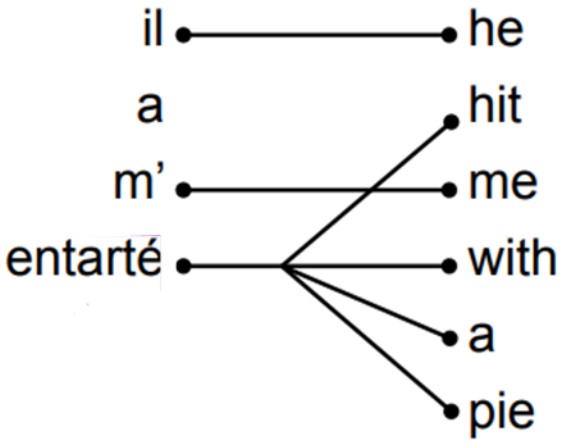
Others are one-to-many

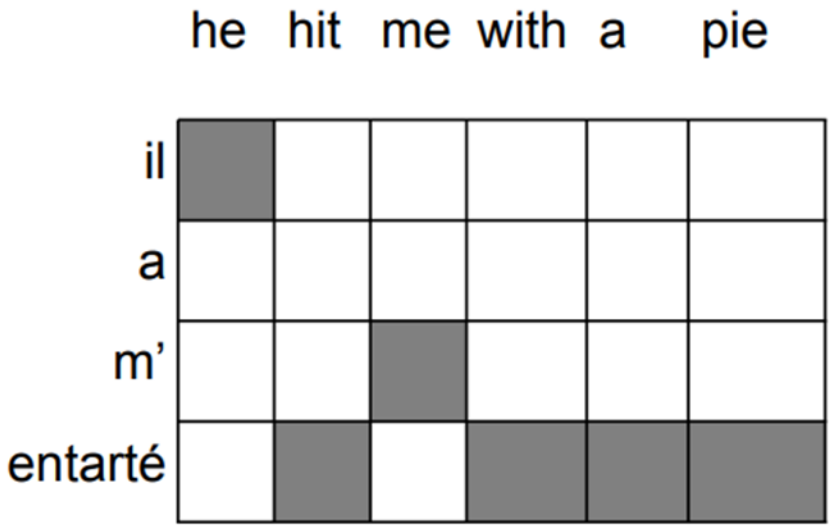


Slide adapted from Chris Manning. Example from "The Mathematics of Statistical Machine Translation: Parameter Estimation, 1993.

Consider Machine Translation

Some words are very fertile – especially words without direct translations

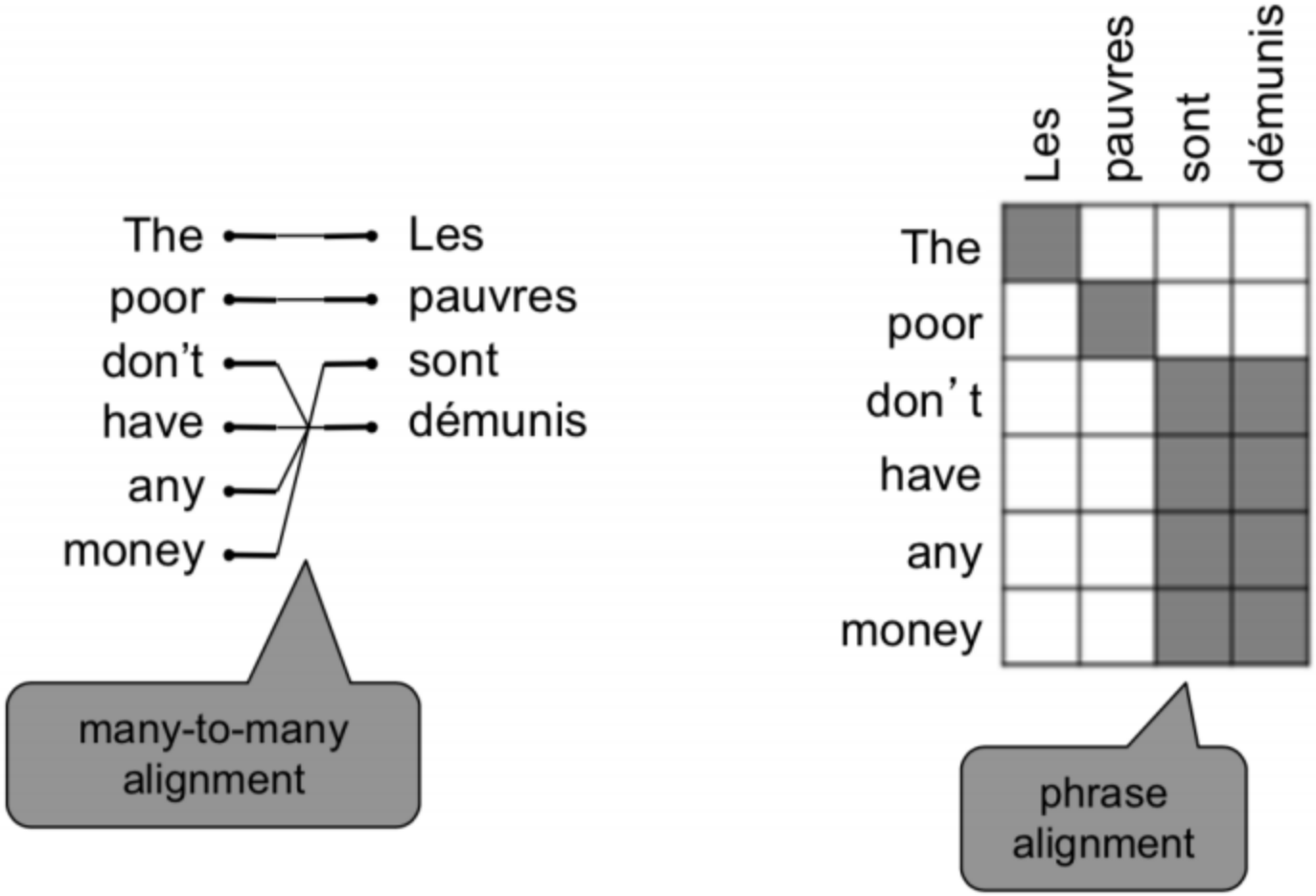




Slide adapted from Chris Manning. Example from "The Mathematics of Statistical Machine Translation: Parameter Estimation, 1993.

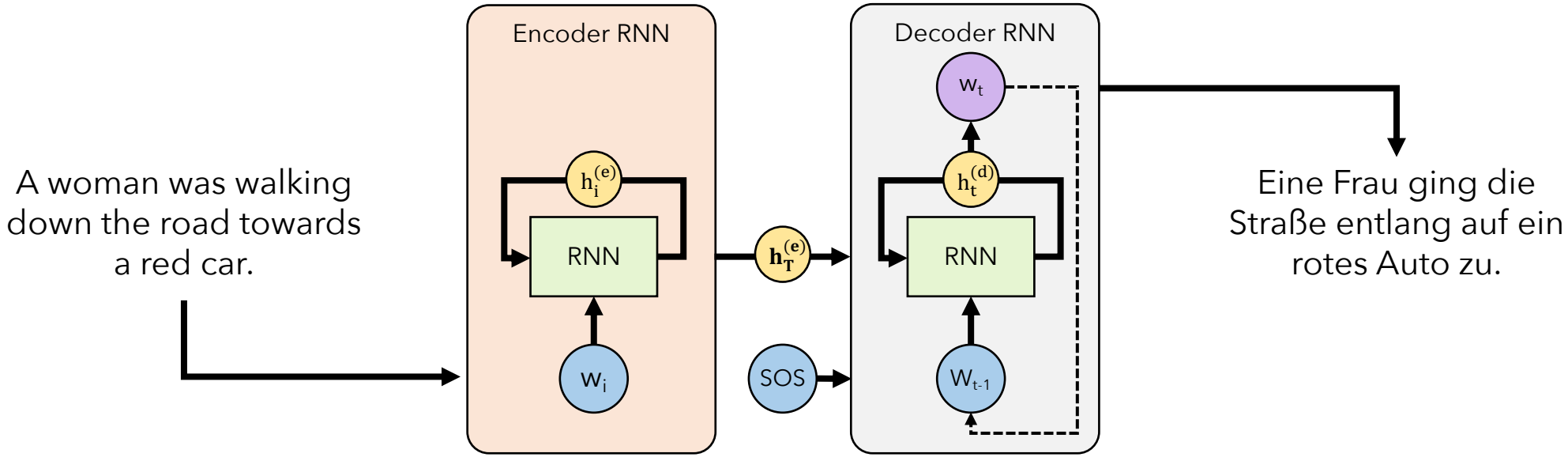
Consider Machine Translation

Phrase translations can be many-to-many



Slide adapted from Chris Manning. Example from "The Mathematics of Statistical Machine Translation: Parameter Estimation, 1993.

Encoder-Decoder for Machine Translation



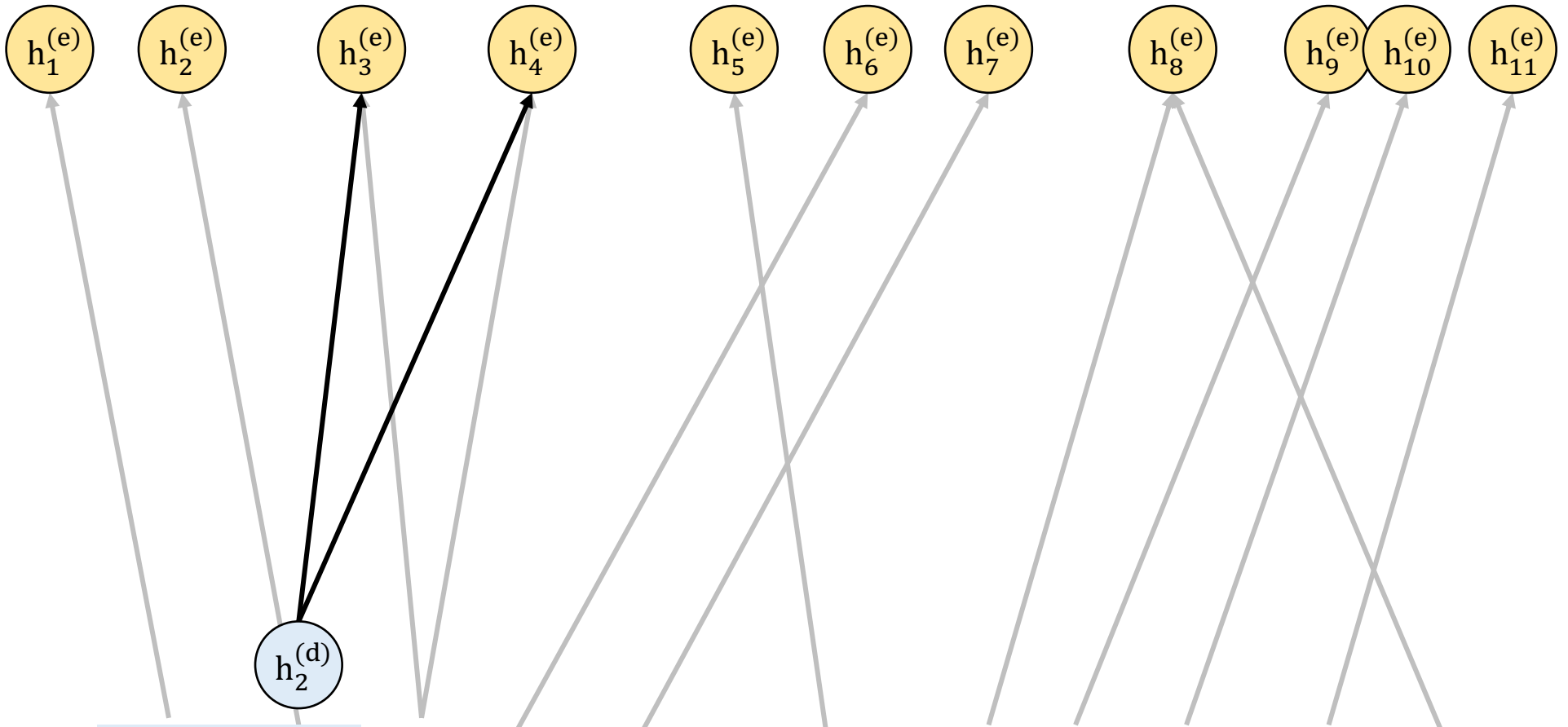
English A woman was walking down the road towards a red car.

German Eine Frau ging die Straße entlang auf ein rotes Auto zu.

Encoder-Decoder for Machine Translation

English A woman was walking down the road towards a red car.

Encoder States

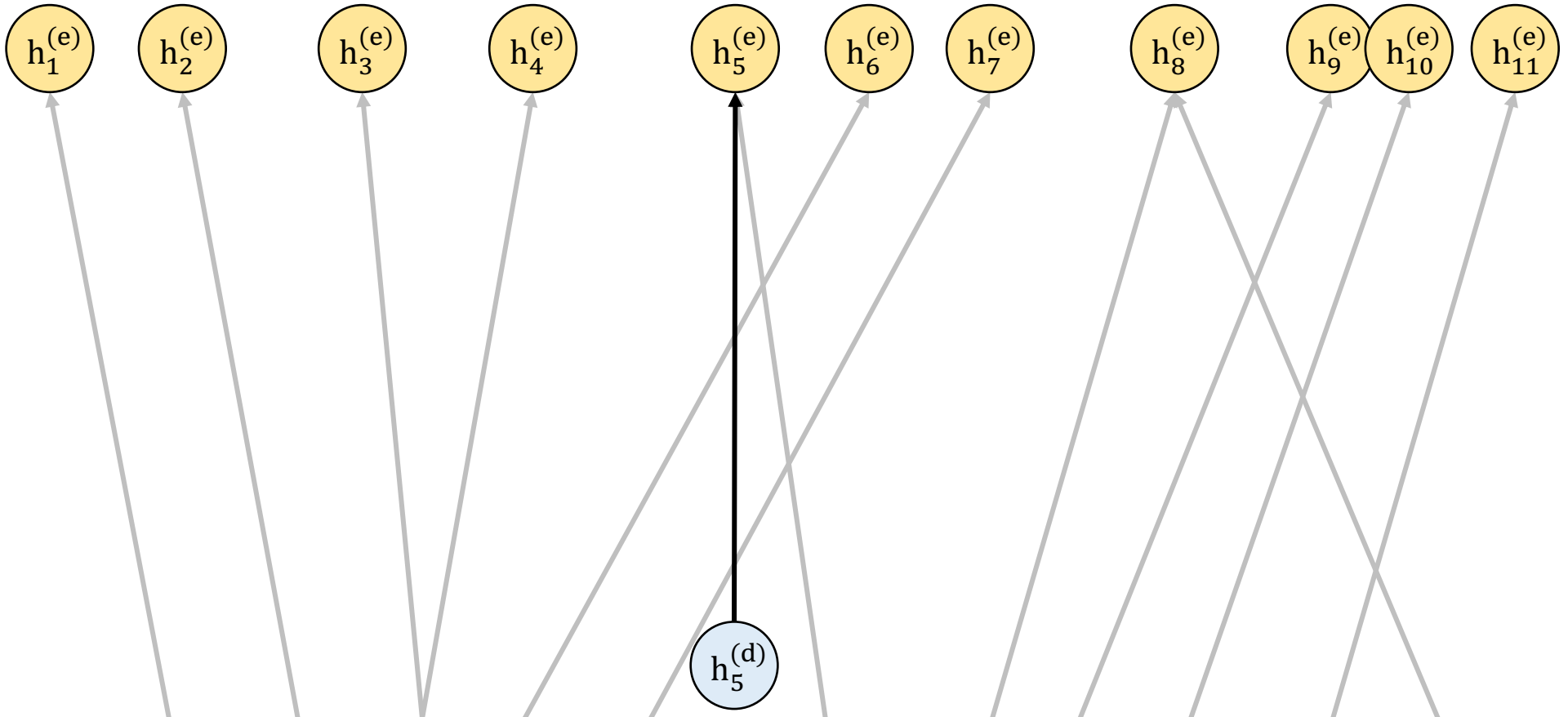


German Eine Frau ging die Straße entlang auf ein rotes Auto zu.

Encoder-Decoder for Machine Translation

English A woman was walking down the road towards a red car.

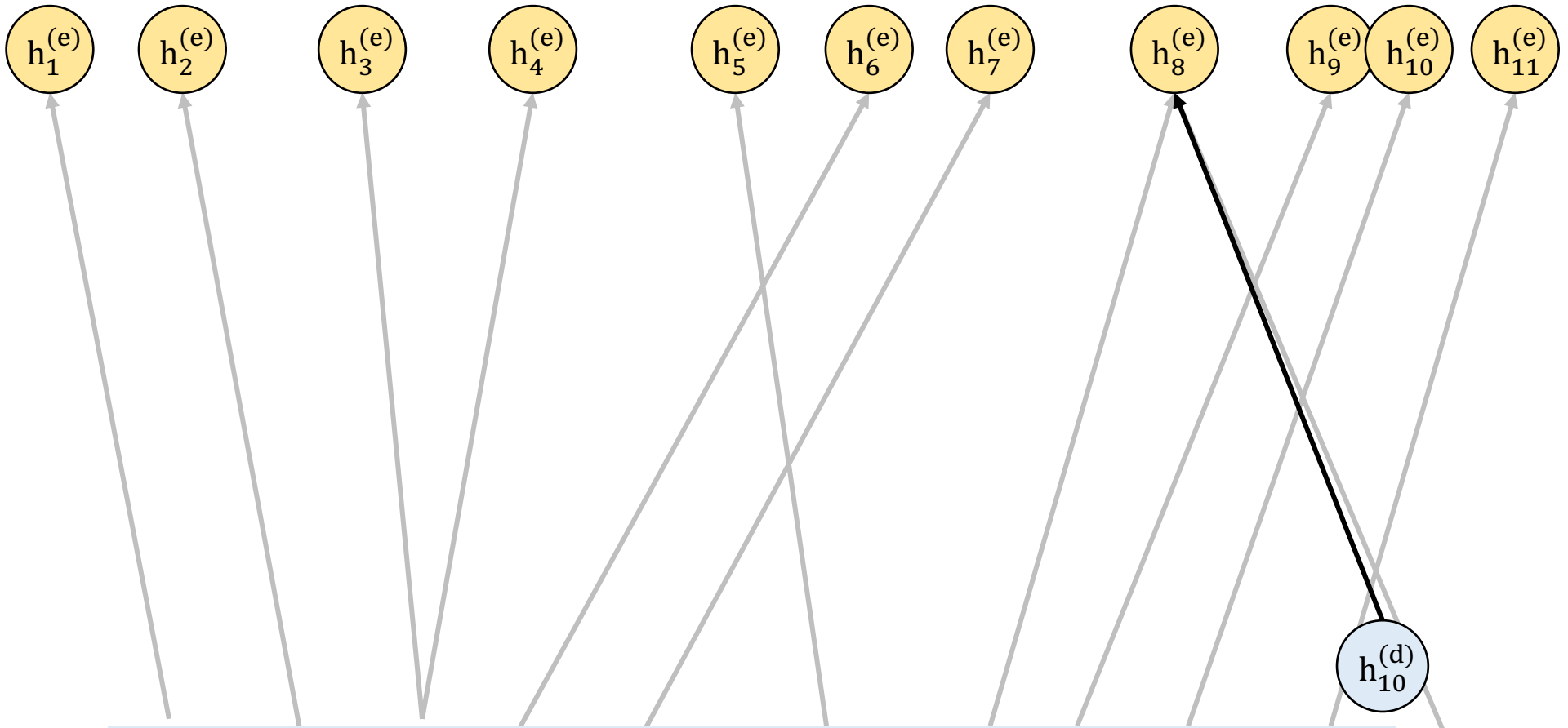
Encoder States



German Eine Frau ging die Straße entlang auf ein rotes Auto zu.

Encoder-Decoder for Machine Translation

English A woman was walking down the road towards a red car.



German Eine Frau ging die Straße entlang auf ein rotes Auto zu.

How can we codify this notion of attention?

Set of Candidates v_1 v_2 v_3 v_4 v_5 $\cdots$ v_k $\boldsymbol{v_i} \in \mathbb{R}^{\boldsymbol{d}}$

Query q $\boldsymbol{q} \in \mathbb{R}^{\boldsymbol{c}}$

Our Goal: Some mechanism to select between k candidates given the query q

Soft-attention – This selection is a convex combination of candidates

Hard-attention – A true selection of only one candidate.

Learning Objectives

Be able to answer:

- ~~What is an attention mechanism?~~
- ~~Difference between soft and hard attention?~~

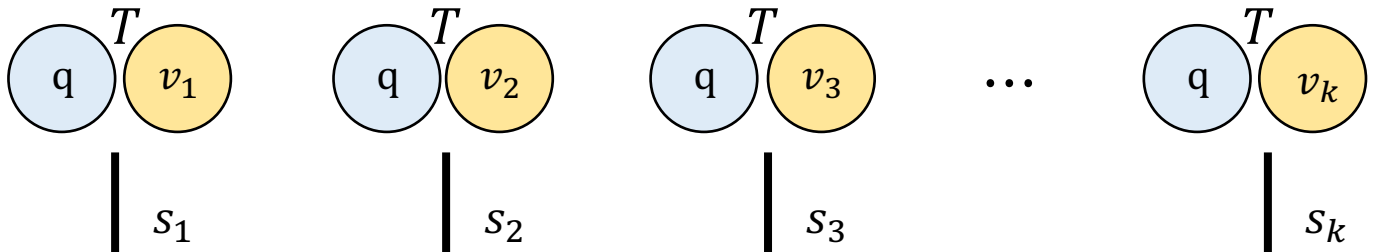
- What are some examples of attention?

- How does attention relate to interpretability?

One Example of A Simple Soft Attention Mechanism

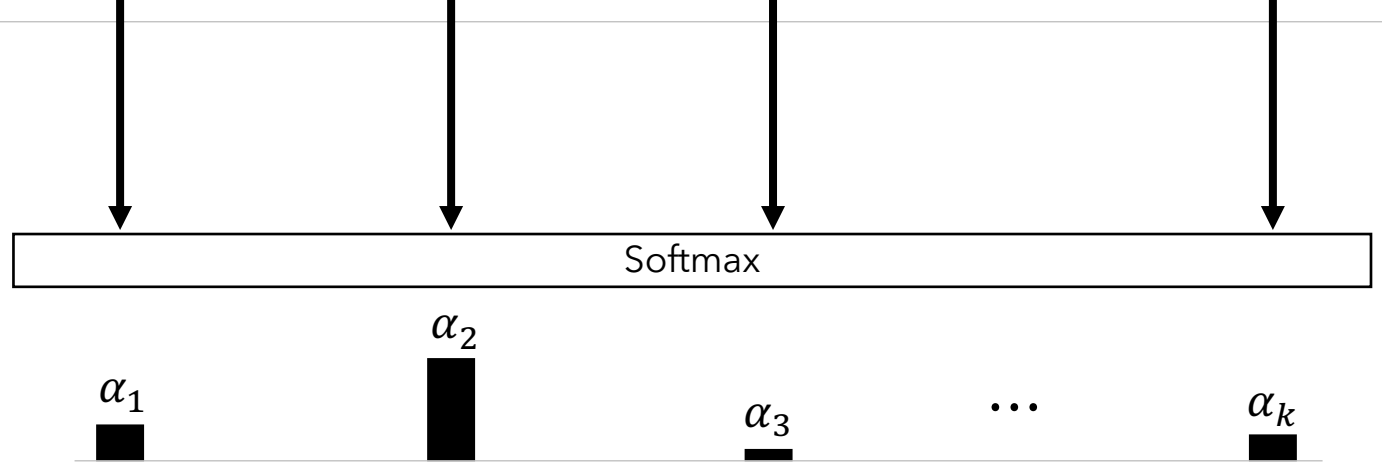
Step 1:
Compute Some
Relevancy Score

$$s_i = q^T v_i$$



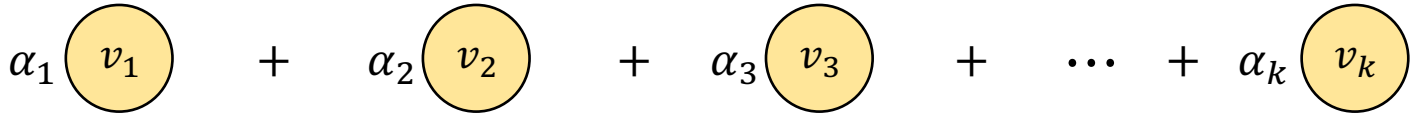
Step 2:
Normalize to
Attention Distribution

$$\alpha_i = softmax_i(s_i)$$

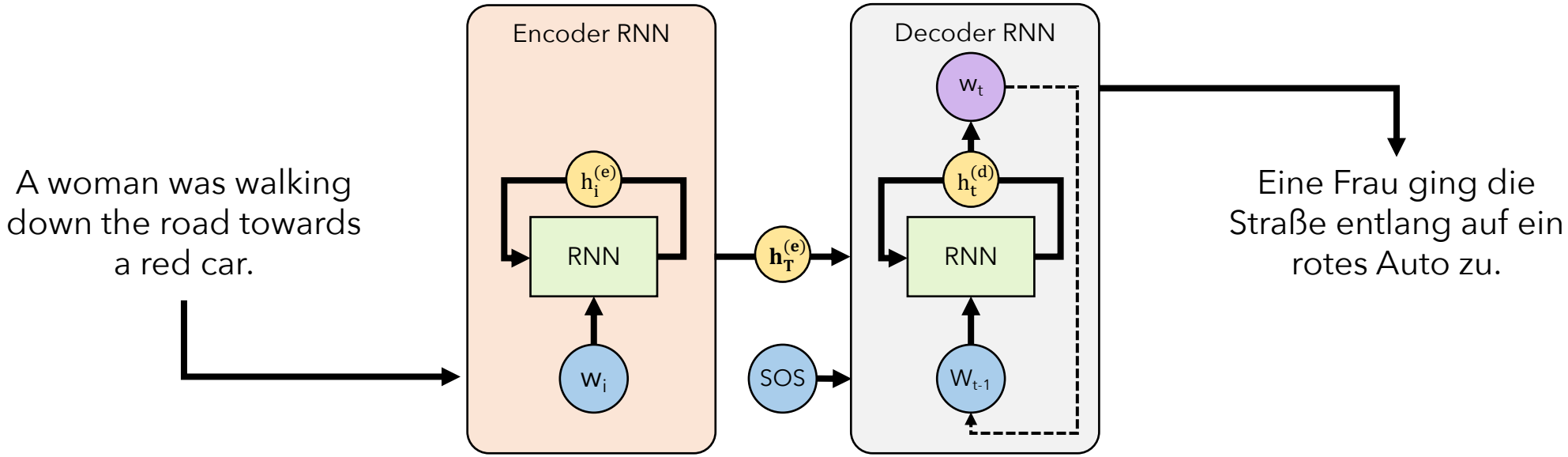


Step 3:
Produce Attended
Feature

$$v = \sum_i \alpha_i v_i$$



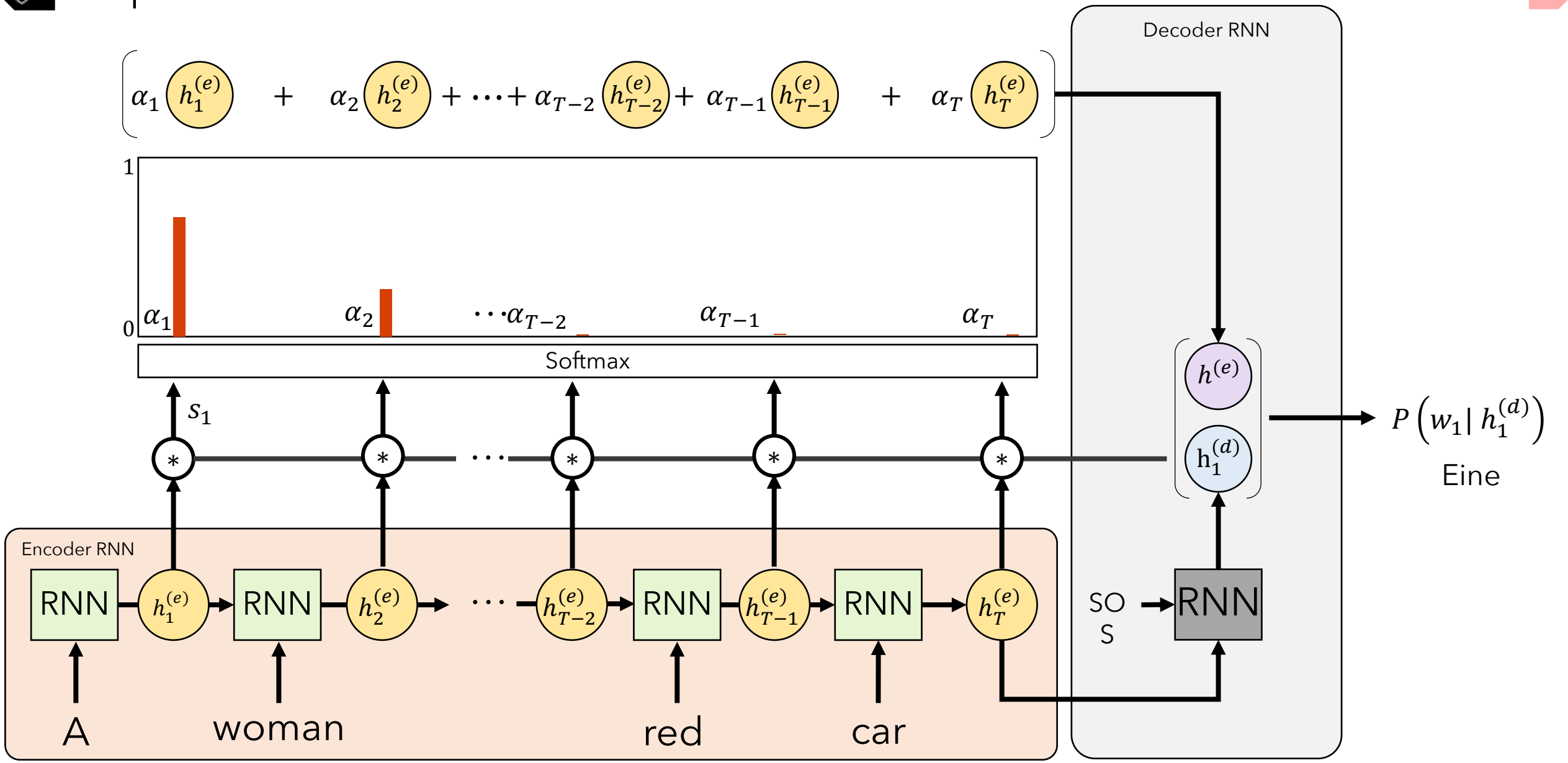
Machine Translation



English A woman was walking down the road towards a red car.

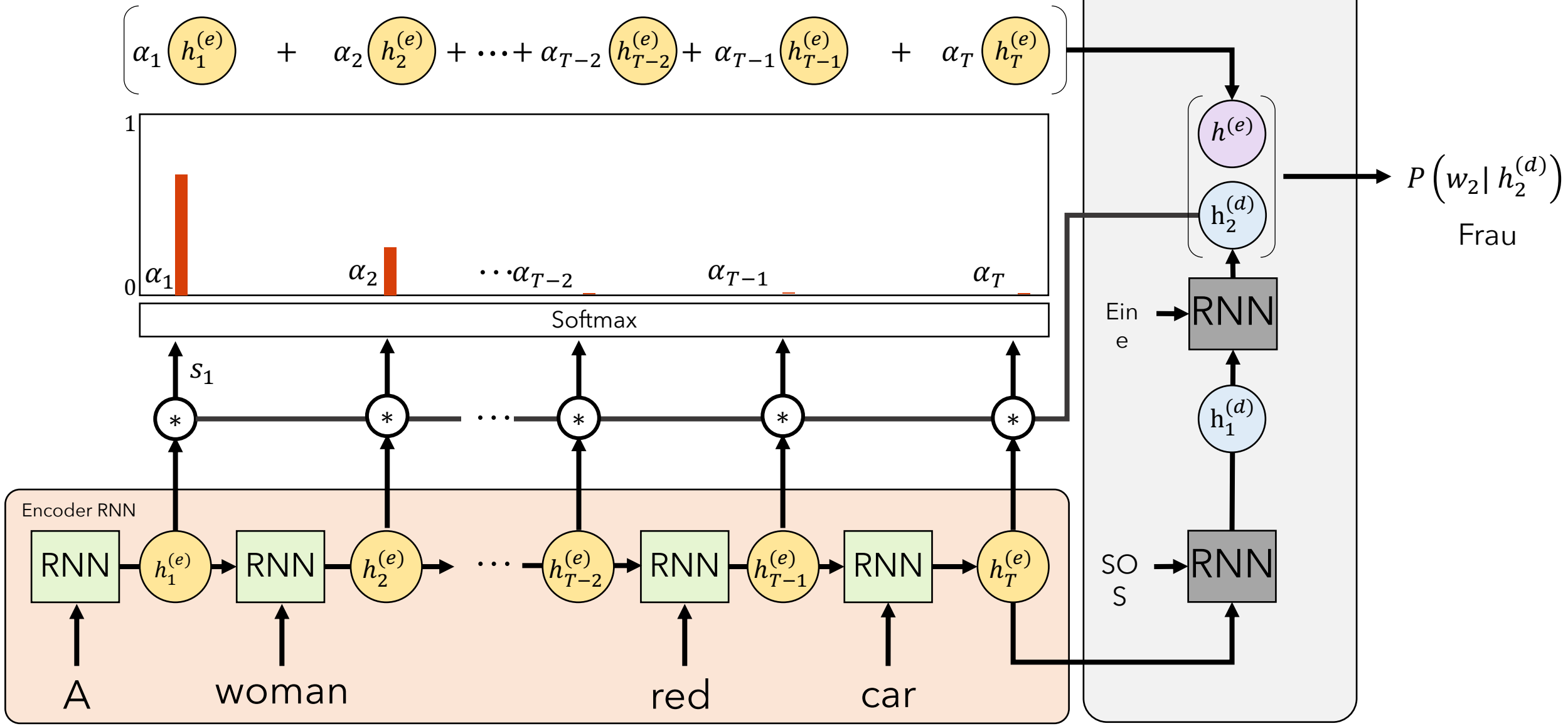
German Eine Frau ging die Straße entlang auf ein rotes Auto zu.

Simple Soft Attention for MT In Pictures



A woman was walking down the road towards a red car

Simple Soft Attention for MT In Pictures



A woman was walking down the road towards a red car

Simple Soft Attention for MT In Equations

Encode source sentence

$$h_1^{(e)}, \ldots, h_T^{(e)} = BiRNN_{(e)}(s_1, \ldots, s_T)$$

For Each Step of Decoder:

$$h_t^{(d)} = RNN_{(d)}\left(w_{t-1}, h_{t-1}^{(d)}\right)$$ *Update hidden state*

$$s_i = {h_{t-1}^{(d)}}^T h_i^{(e)} \quad \forall i$$ *Compute relevancy scores*

$$\alpha_i = \mathrm{softmax_i}(s_1, \ldots, s_T) \quad \forall i$$ *Normalize to attention distribution*

$$h^{(e)} = \sum_i \alpha_i h_i^{(e)}$$ *Calculate attended state*

$$P(w_t \mid w_{<t}, s_1, \ldots, s_T) = softmax\left(W_V \left[\boldsymbol{h}_t^{(d)}, \boldsymbol{h}^{(e)}\right]\right)$$ *Predict next word*

A More General Notion Of An Attention Mechanism

Set of Candidates v_1 v_2 v_3 v_4 v_5 $\cdots$ v_k

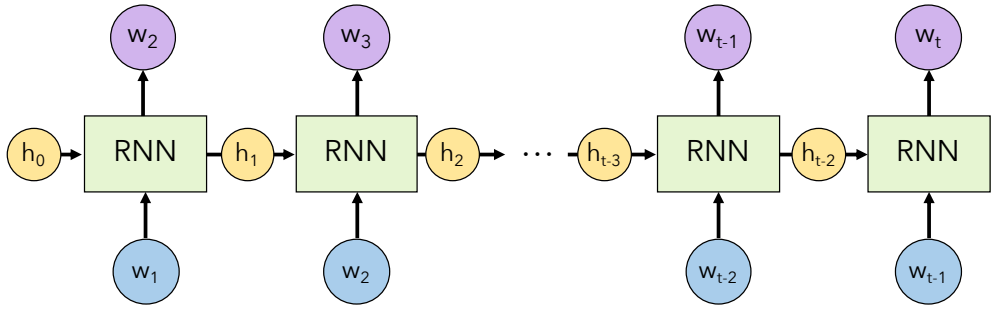
Query q

Attention Function $\alpha_i = f(\mathrm{q}, \mathrm{v_i}) \quad \text{s.t.} \quad \sum_i a_i = 1 \quad \forall i, a_i \geq 0$

Attended Feature $c = \sum_i f(\mathrm{q}, \mathrm{v_i}) * \mathrm{c_i}$

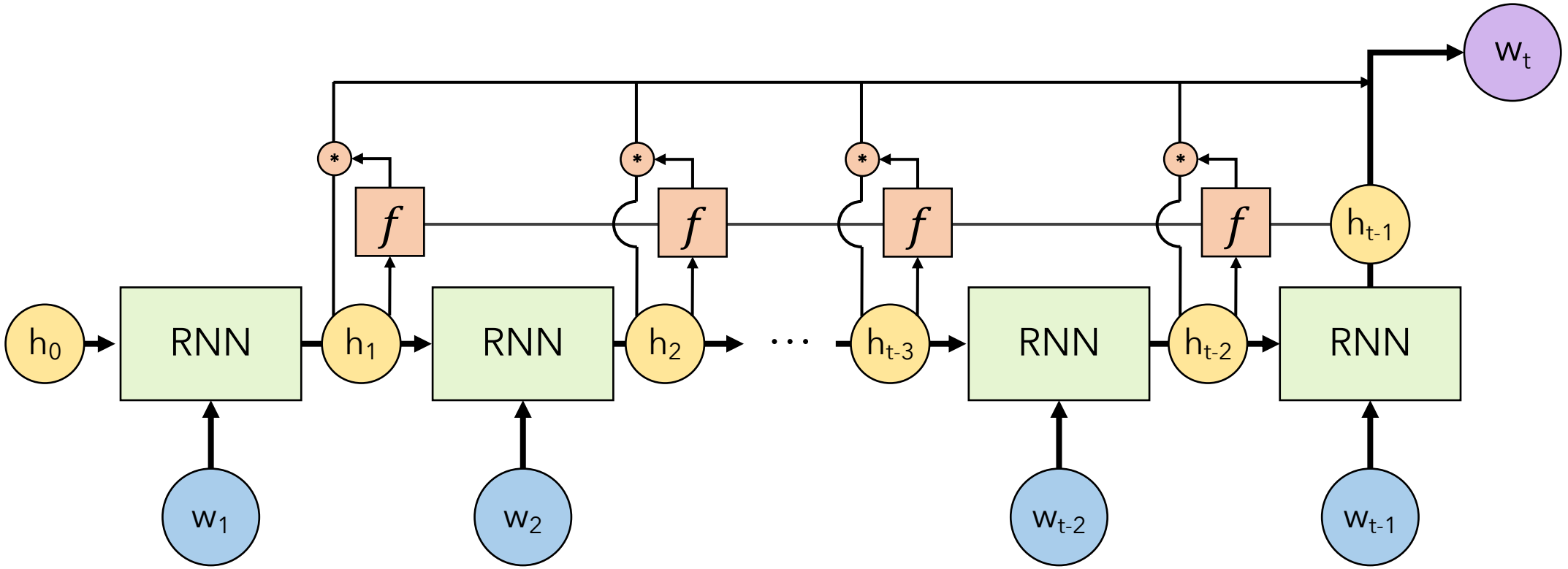
Questions Break!

Say I'm just doing plain language modeling like before, how might I use attention?



Questions Break!

Say I'm just doing plain language modeling like before, how might I use attention?



Example Attention Mechanisms

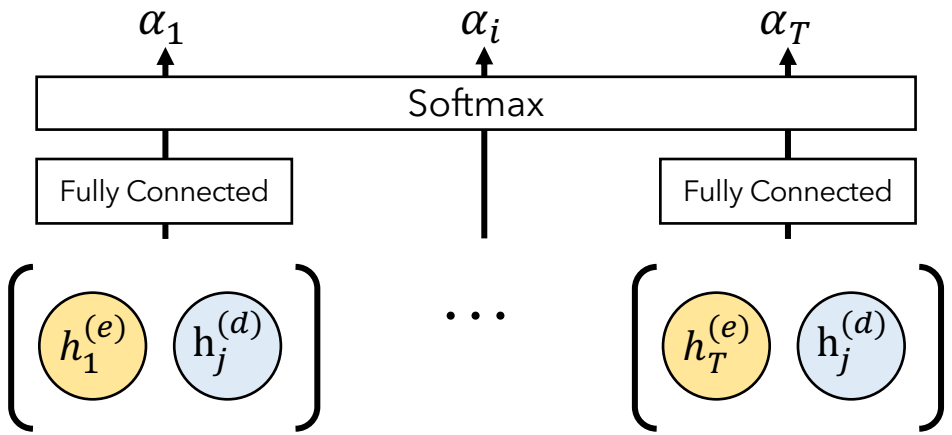
Example: Neural Machine Translation by Jointly Learning to Align and Translate, 2016.

Candidates Hidden states from a bidirectional LSTM encoder.

Query Hidden state from the LSTM decoder at current step.

Attention Function

f:

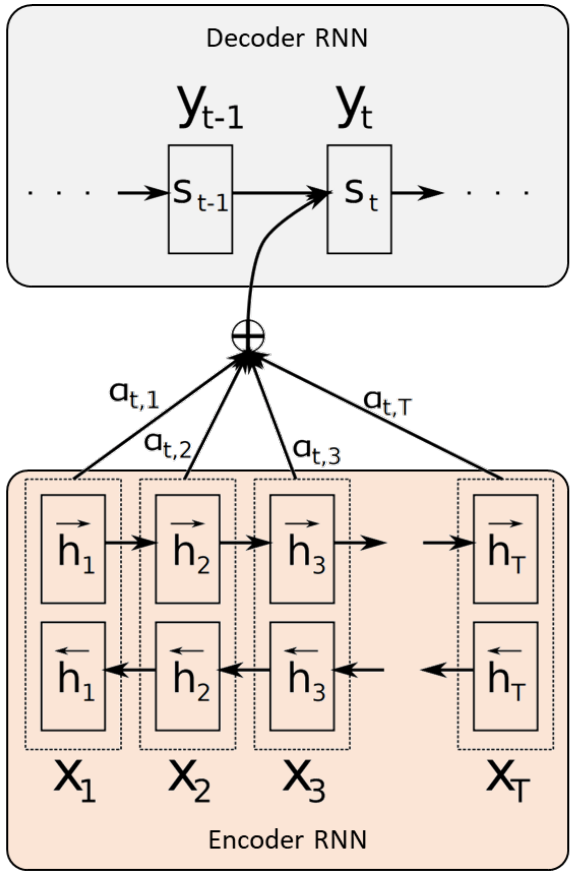


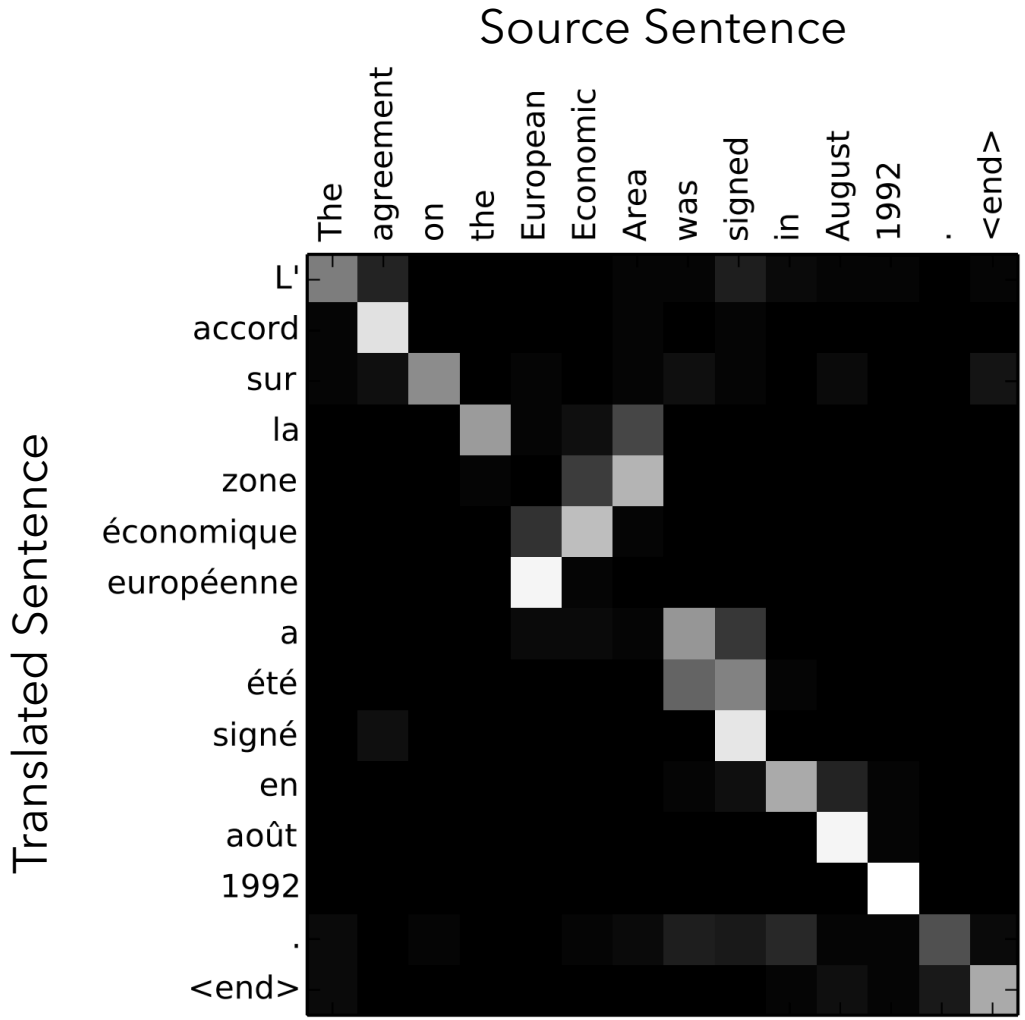
Attended Feature

$$c_j = \sum_i \alpha_i h_i^{(e)}$$

Example Attention Mechanisms

Example: Neural Machine Translation by Jointly Learning to Align and Translate, 2015.

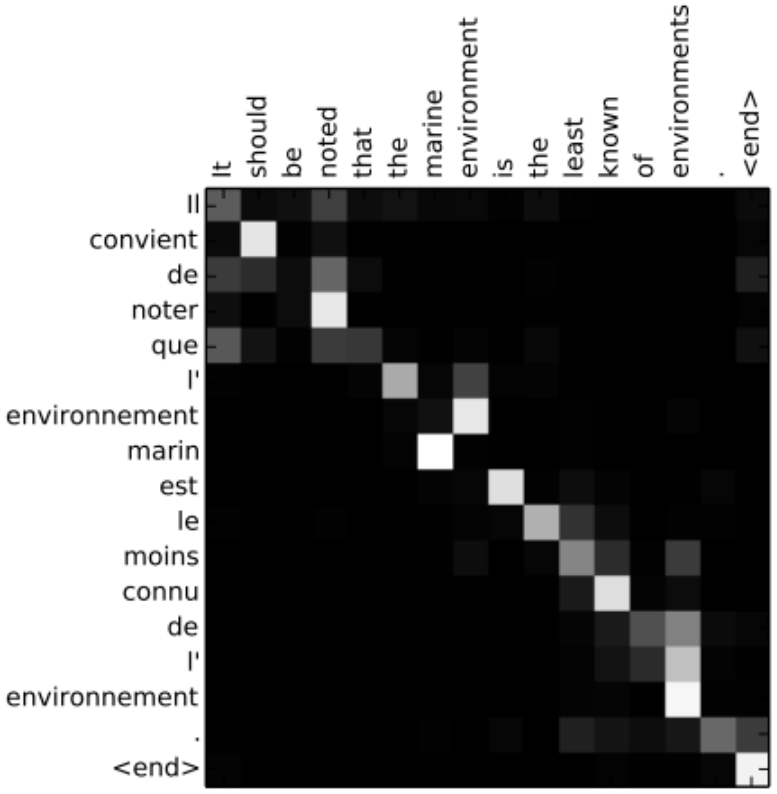


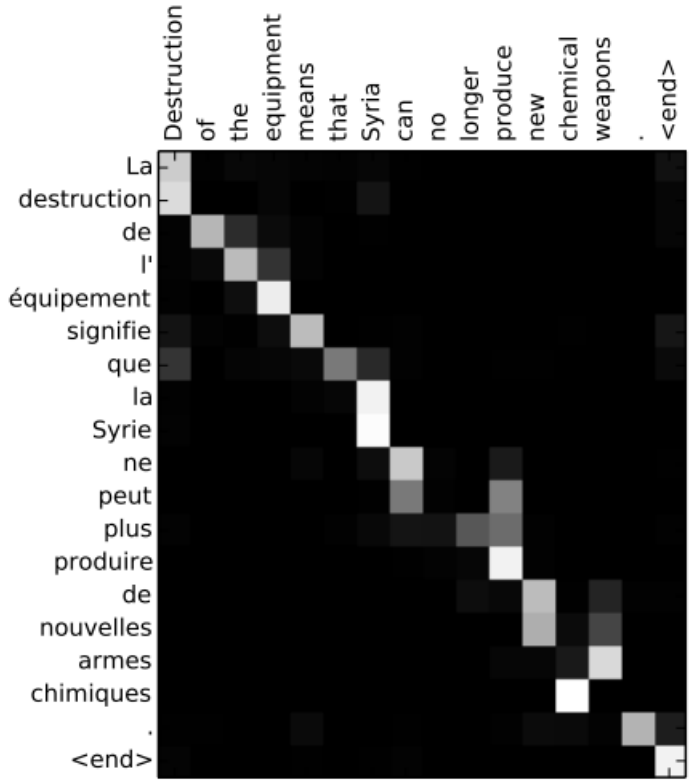


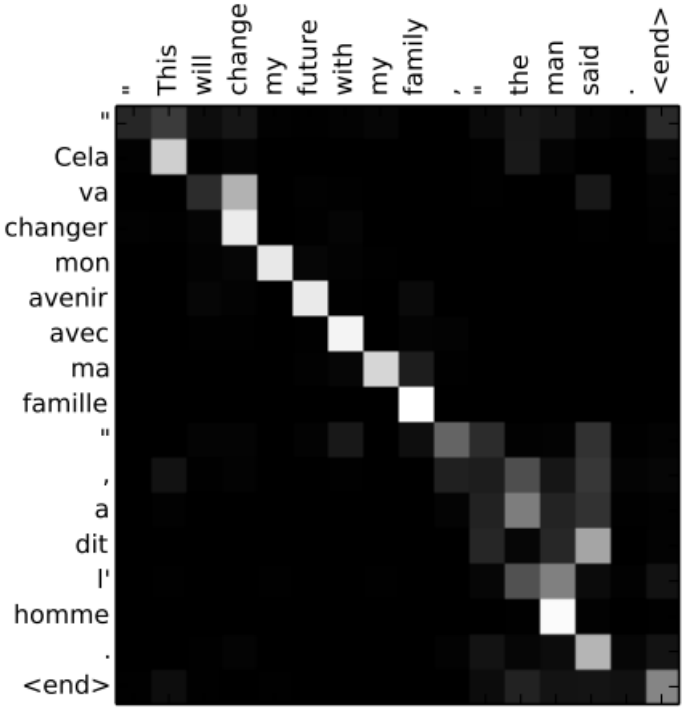
Neural Machine Translation by Jointly Learning to Align and Translate, 2016.

Example Attention Mechanisms

Example: Neural Machine Translation by Jointly Learning to Align and Translate, 2015.





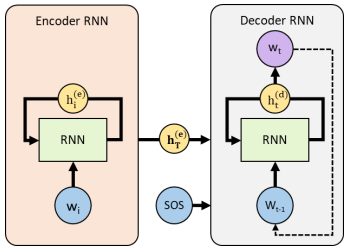


Neural Machine Translation by Jointly Learning to Align and Translate, 2016.

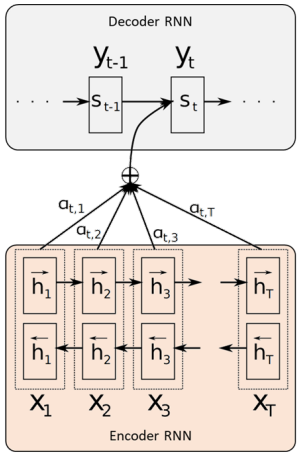
Example Attention Mechanisms

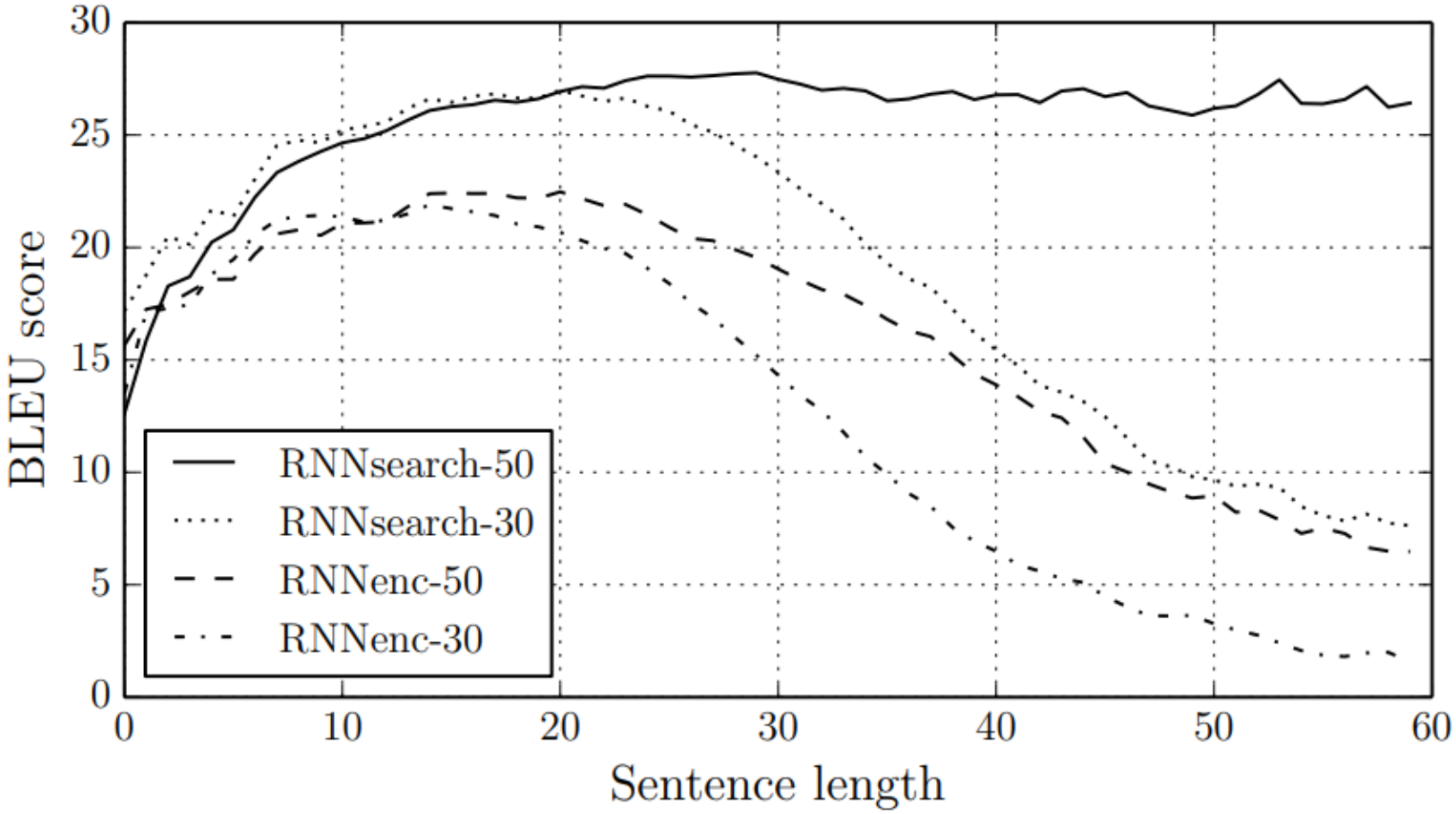
Example: Neural Machine Translation by Jointly Learning to Align and Translate, 2015.

RNNenc



RNNsearch





Packing a whole sentence into a single vector seems hard. Attention helps

Neural Machine Translation by Jointly Learning to Align and Translate, 2016.

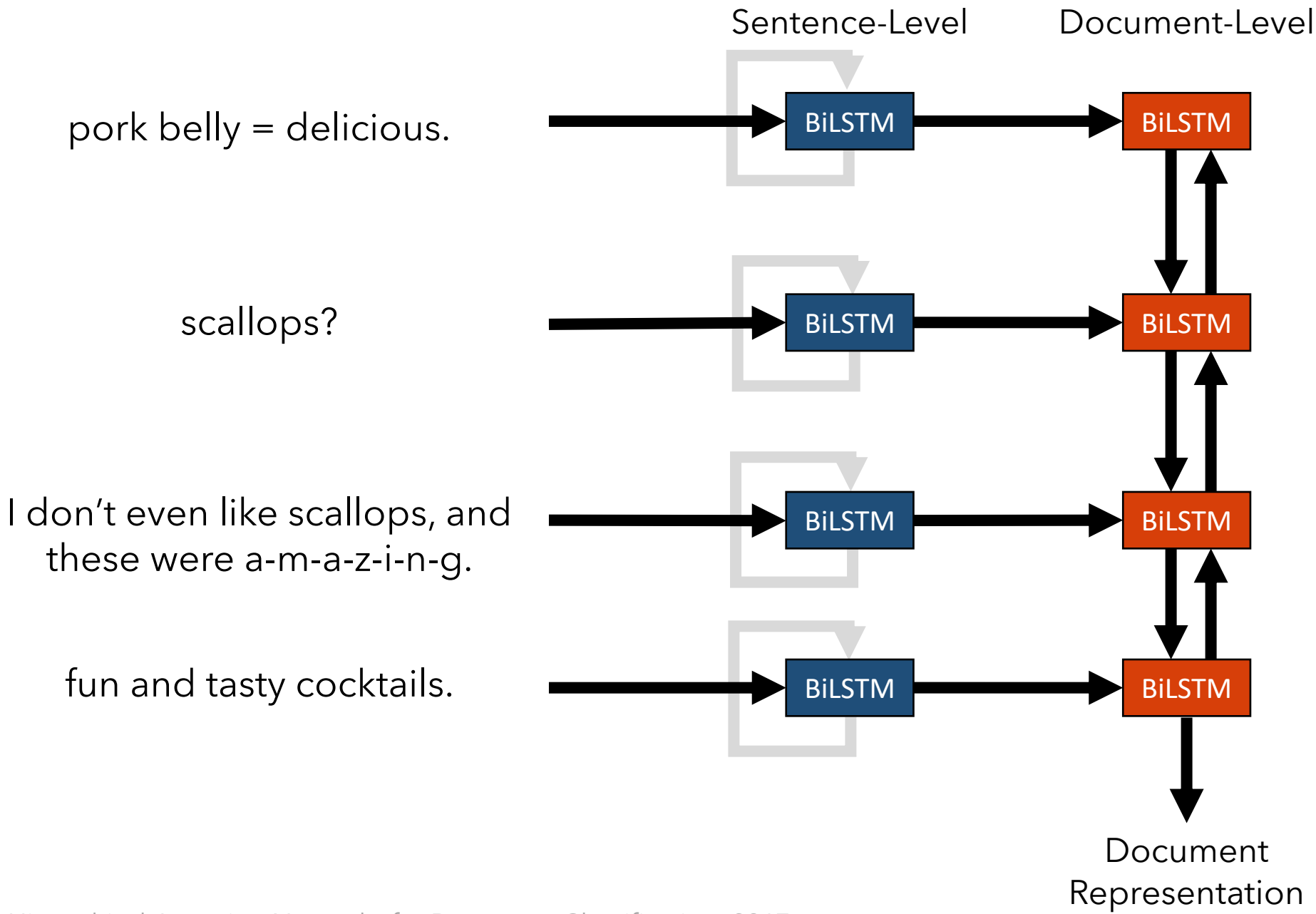
Example Attention Mechanisms

Example: Neural Machine Translation by Jointly Learning to Align and Translate, 2015.

> Our approach, on the other hand, requires computing the annotation weight of every word in the source sentence for each word in the translation. This drawback is not severe with the task of translation in which most of input and output sentences are only 15–40 words. However, this may limit the applicability of the proposed scheme to other tasks.

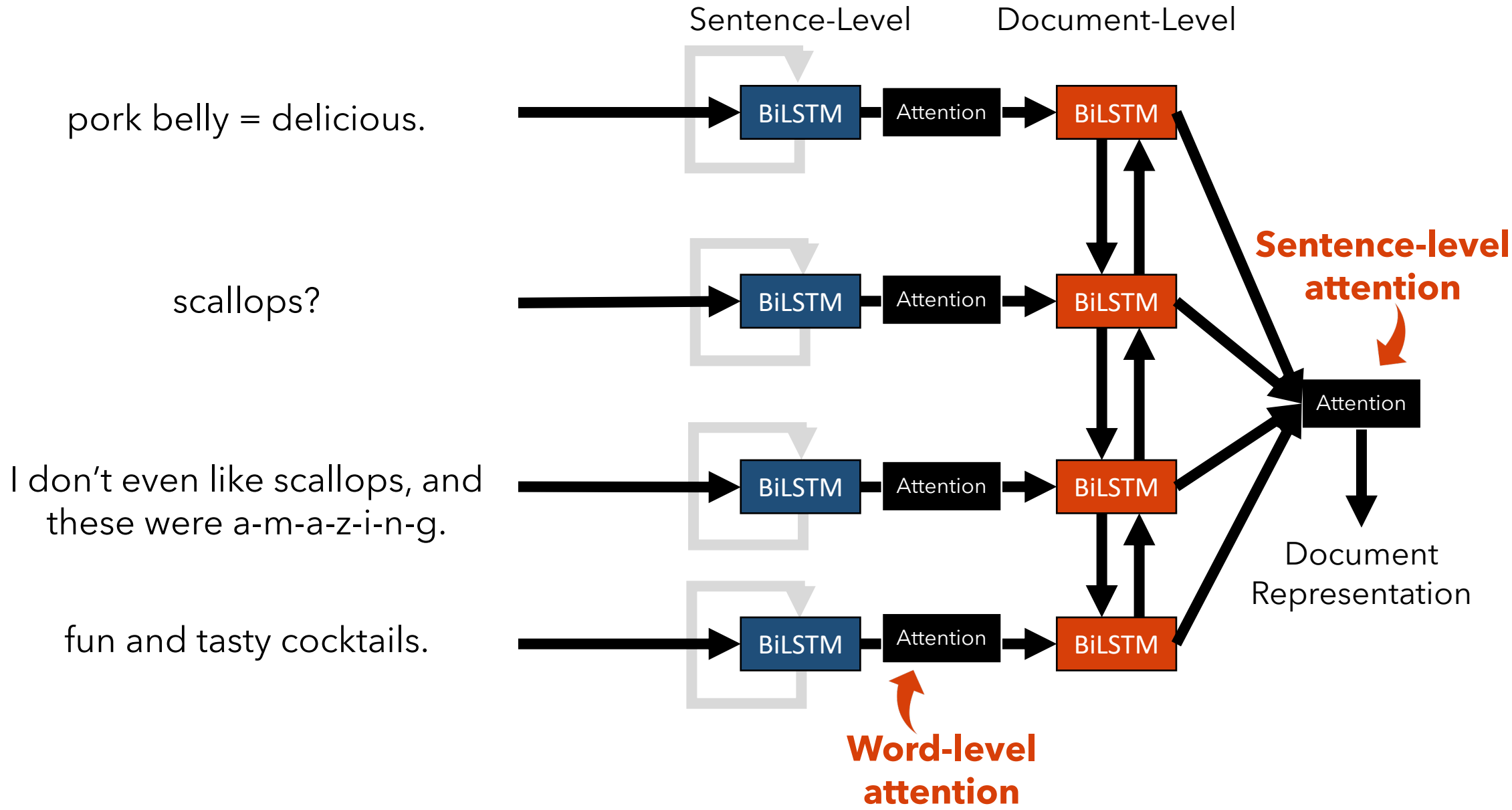
Neural Machine Translation by Jointly Learning to Align and Translate, 2016.

Consider Sentiment Classification with Hierarchical LSTM



Hierarchical Attention Networks for Document Classification, 2017.

Consider Sentiment Classification with Hierarchical LSTM



Hierarchical Attention Networks for Document Classification, 2017.

Example Attention Mechanisms

Example: Hierarchical Attention Networks for Document Classification, 2017.

Word-level attention

$$h_{i0}^{(w)}, \ldots, h_{iT_i}^{(w)} = BiLSTM(w_{i0}, \ldots, w_{iT_i})$$

$$\alpha_{it} = softmax_t\left(\tanh\left(W_w h_{it}^{(w)} + b_w\right)^T u_w \right)$$

$$c_i = \sum_i \alpha_{it} h_{it}^{(w)}$$

Sentence-level attention

$$h_0^{(s)}, \ldots h_I^{(s)} = BiLSTM(c_0, \ldots, c_I)$$

$$a_i = softmax_i\left(\tanh\left(W_s h_i^{(s)} + b_s\right)^T u_s \right)$$

$$d = \sum_i a_i h_i^{(s)}$$

$$\boldsymbol{y} = softmax(Wd + b)$$

Example Attention Mechanisms

Example: Hierarchical Attention Networks for Document Classification, 2017.

Word-level attention

$$h_{i0}^{(w)}, \ldots, h_{iT_i}^{(w)} = BiLSTM(w_{i0}, \ldots, w_{iT_i})$$

$$\alpha_{it} = softmax_t\left(\tanh\left(W_w h_{it}^{(w)} + b_w\right)^T \boldsymbol{u_w} \right)$$

$$c_i = \sum_i \alpha_{it} h_{it}^{(w)}$$

Sentence-level attention

$$h_0^{(s)}, \ldots h_I^{(s)} = BiLSTM(c_0, \ldots, c_I)$$

$$a_i = softmax_i\left(\tanh\left(W_s h_i^{(s)} + b_s\right)^T \boldsymbol{u_s} \right)$$

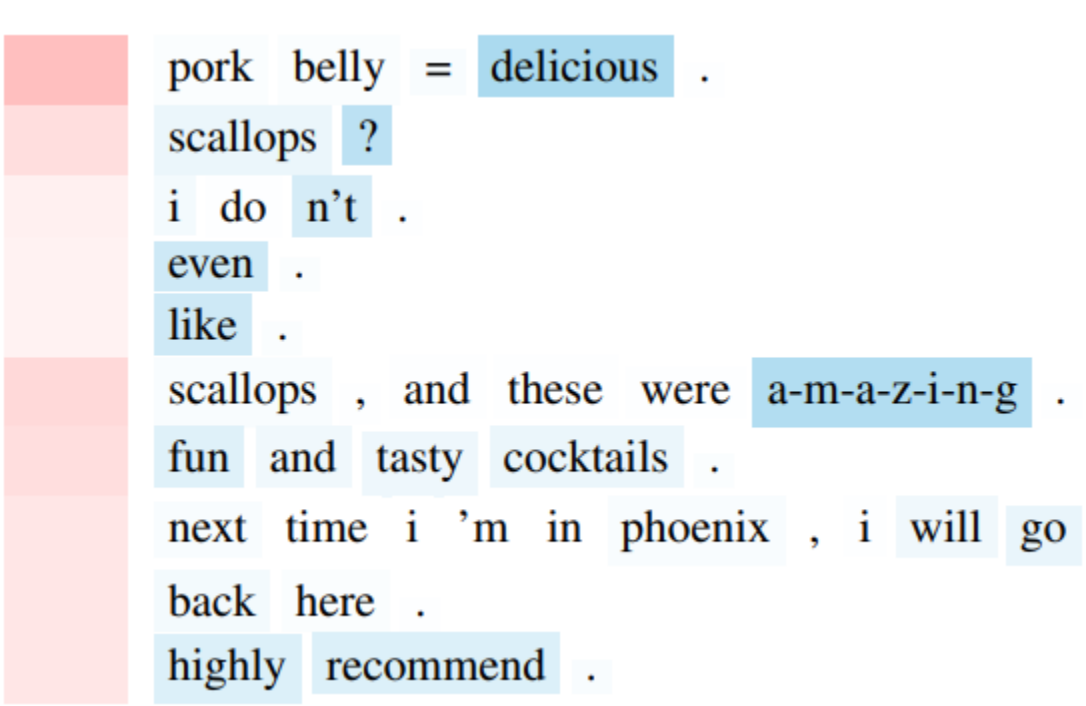
$$d = \sum_i a_i h_i^{(s)}$$

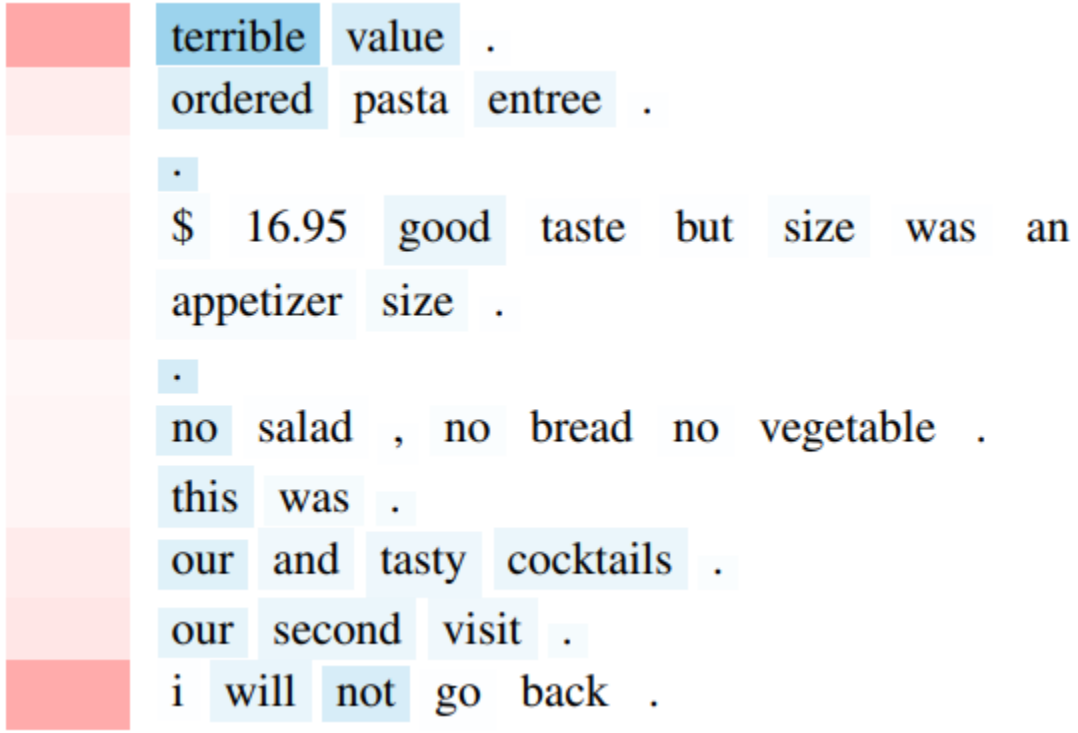
$$\boldsymbol{y} = softmax(Wd + b)$$

Learned vectors denoting "importance" broadly. Not another input acting as query.

Example Attention Mechanisms

$$d = \sum_i a_i h_i^{(s)} \qquad c_i = \sum_i \alpha_{it} h_{it}^{(w)}$$

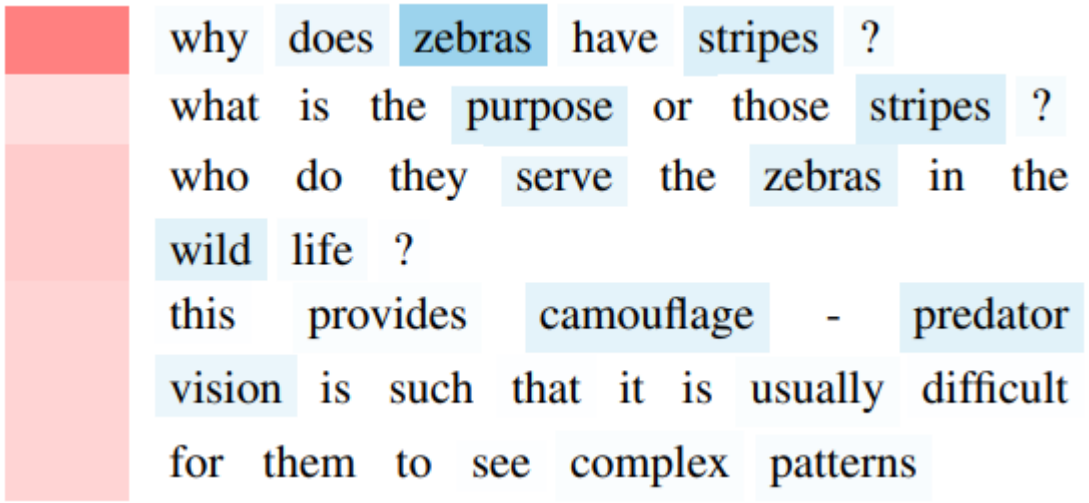


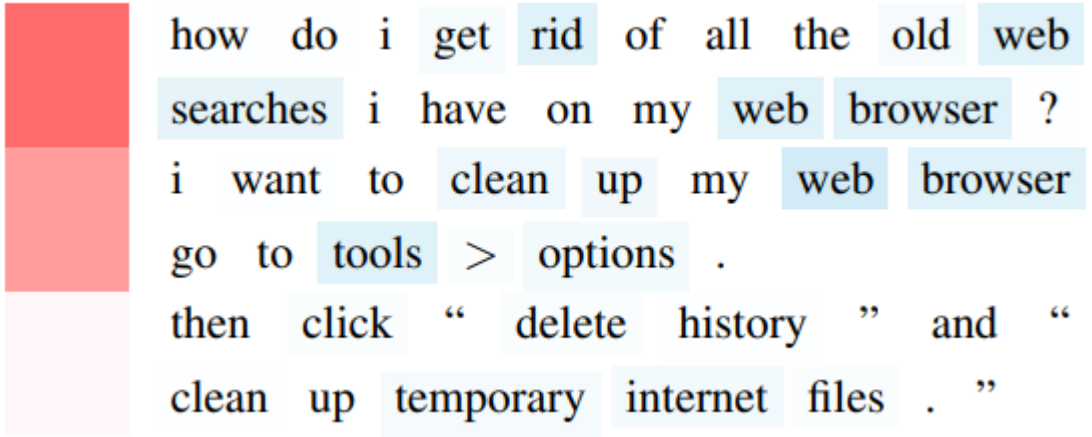


Hierarchical Attention Networks for Document Classification, 2017.

Example Attention Mechanisms

$$d = \sum_i a_i h_i^{(s)} \qquad c_i = \sum_i \alpha_{it} h_{it}^{(w)}$$





Hierarchical Attention Networks for Document Classification, 2017.

Example Attention Mechanisms

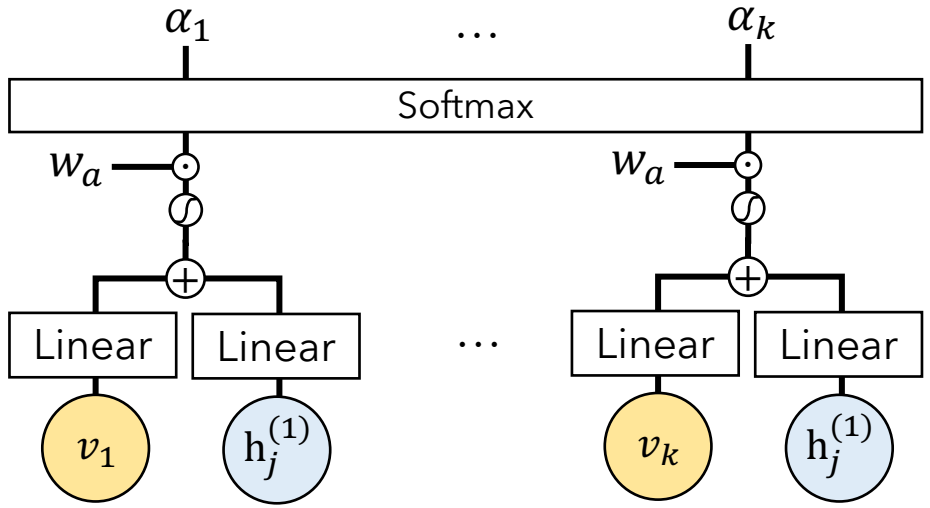
Example: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, 2017.

Candidates Image regions and associated features from Faster RCNN detection model.

Query Hidden LSTM decoder state, previous generated word, and average image feature.

Attention Function

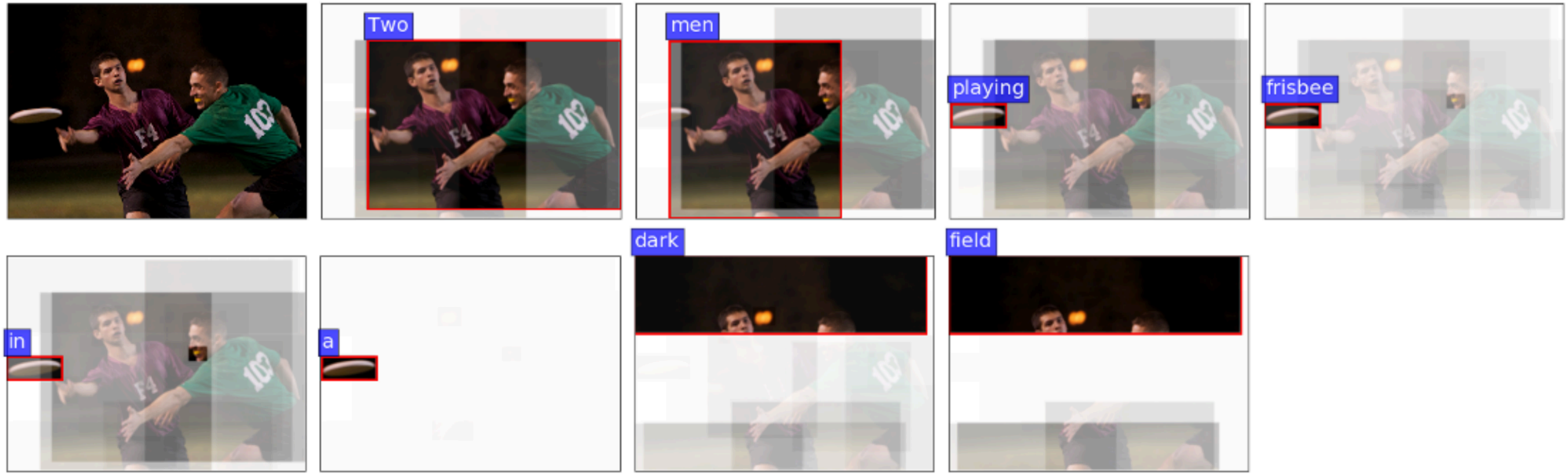
f:



Attended Feature $c_j = \sum_i \alpha_i v_i$

Example Attention Mechanisms

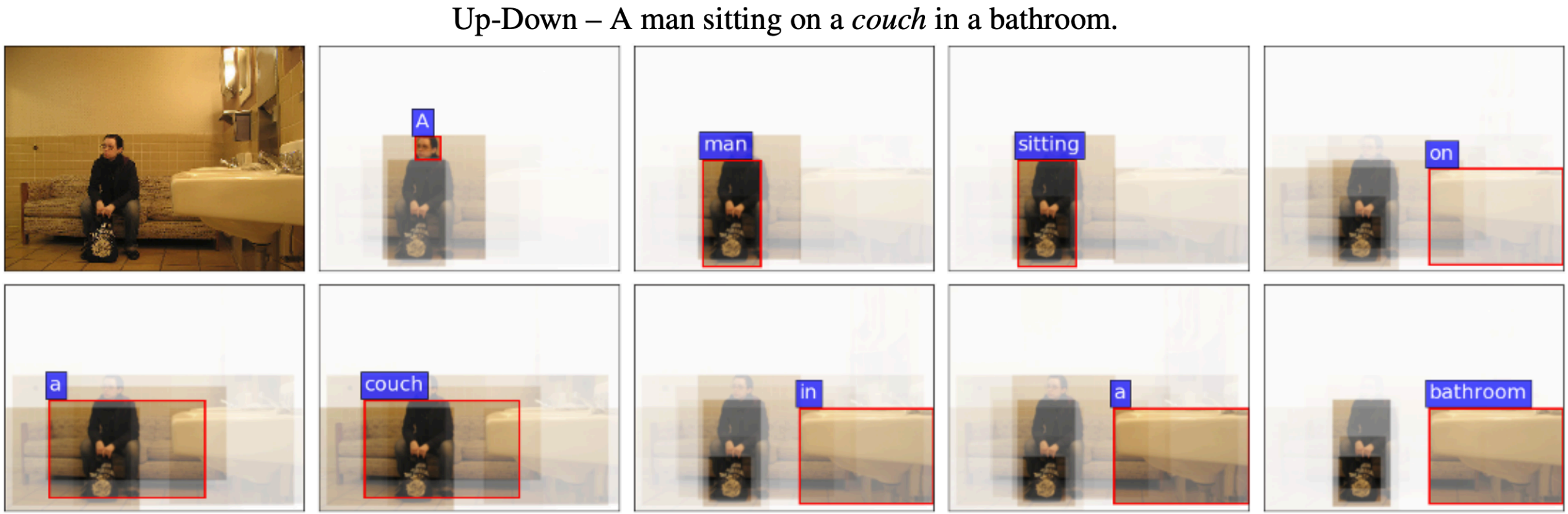
Example: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, 2017.



Two men playing frisbee in a dark field.

Example Attention Mechanisms

Example: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, 2017.



Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, 2017.

Learning Objectives

Be able to answer:

- ~~What is an attention mechanism?~~
- ~~Difference between soft and hard attention?~~

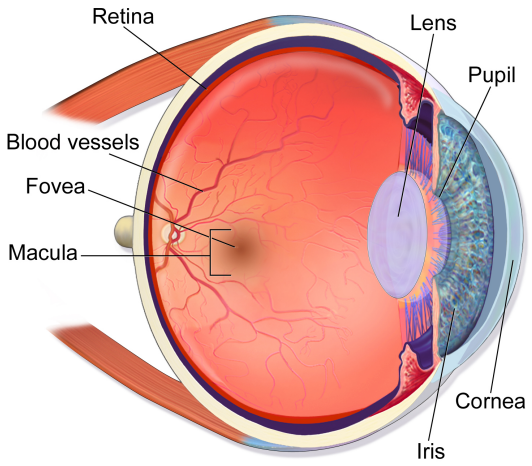
- ~~What are some examples of attention?~~

- How does attention relate to interpretability?

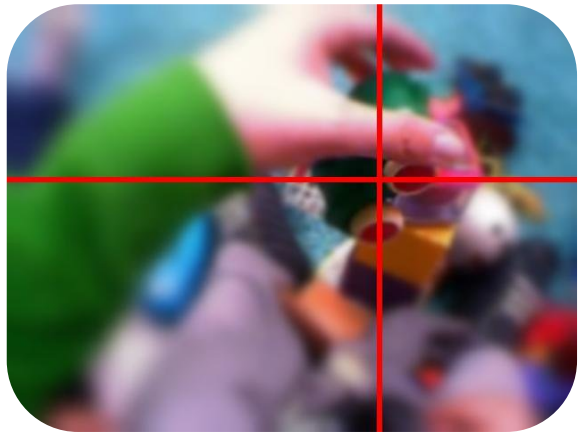
What motivations have been proposed for attention?

- **Pragmatic.** It is a useful computational bias. Let's us access far-away elements select from a large set of possible information.

- **Biologic.** Similar bias present in human visual system (foveated vision + sacca



Eye Anatomy

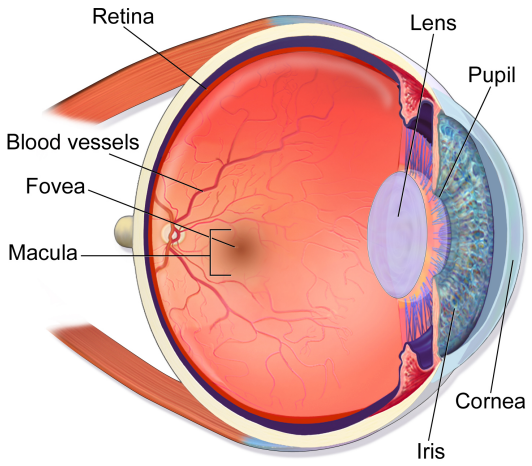


- **Interpretability.** Looking at attention maps explains some part of the model's dependence on the input.

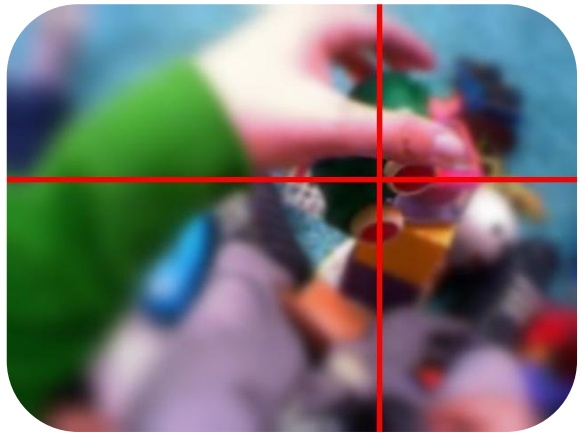
What motivations have been proposed for attention?

- **Pragmatic.** It is a useful computational bias. Let's us access far-away elements select from a large set of possible information.

- **Biologic.** Similar bias present in human visual system (foveated vision + sacca



Eye Anatomy



- **Interpretability.** Looking at attention maps explains some part of the model's dependence on the input.

Is Attention Explanation?

Attention is not explanation

S Jain, BC Wallace - arXiv preprint arXiv:1902.10186, 2019 - arxiv.org

Attention mechanisms have seen wide adoption in neural NLP models. In addition to improving predictive performance, these are often touted as affording transparency: models equipped with attention provide a distribution over attended-to input units, and this is often …

☆ 99 Cited by 66 Related articles All 5 versions »

Attention is not not Explanation

S Wiegreffe, Y Pinter - arXiv preprint arXiv:1908.04626, 2019 - arxiv.org

Attention mechanisms play a central role in NLP systems, especially within recurrent neural network (RNN) models. Recently, there has been increasing interest in whether or **not** the intermediate representations offered by these modules may be used to explain the …

☆ 99 Cited by 12 Related articles All 4 versions »

Attention is not explanation, 2019.

Is Attention Explanation?

Asks two questions:

1. To what extent do induced attention weights correlate with measures of feature importance (gradients and leave-one-out (LOO) methods)?

2. Would alternative attention weights necessarily yield different predictions?

Attention is not explanation, 2019.

Does attention correlate with feature importance?

Step 1: Train models with attention on a bunch of language tasks.

Dataset	$\|V\|$	*Avg. length*	*Train size*	*Test size*	*Test performance (LSTM)*
SST	16175	19	3034 / 3321	863 / 862	0.81
IMDB	13916	179	12500 / 12500	2184 / 2172	0.88
ADR Tweets	8686	20	14446 / 1939	3636 / 487	0.61
20 Newsgroups	8853	115	716 / 710	151 / 183	0.94
AG News	14752	36	30000 / 30000	1900 / 1900	0.96
Diabetes (MIMIC)	22316	1858	6381 / 1353	1295 / 319	0.79
Anemia (MIMIC)	19743	2188	1847 / 3251	460 / 802	0.92
CNN	74790	761	380298	3198	0.64
bAbI (Task 1 / 2 / 3)	40	8 / 67 / 421	10000	1000	1.0 / 0.48 / 0.62
SNLI	20982	14	182764 / 183187 / 183416	3219 / 3237 / 3368	0.78

Attention is not explanation, 2019.

Does attention correlate with feature importance?

Step 2: Run the models and store the attention values for each instance.

Original: reggio falls victim to relying on the very digital technology that he fervently scorns creating a meandering inarticulate and ultimately disappointing film

Original:general motors and daimlerchrysler say they # qqq teaming up to develop hybrid technology for use in their vehicles . the two giant automakers say they have signed a memorandum of understanding

Attention is not explanation, 2019.

Does attention correlate with feature importance?

Step 3: Compute the gradients with respect to the candidates and compare.

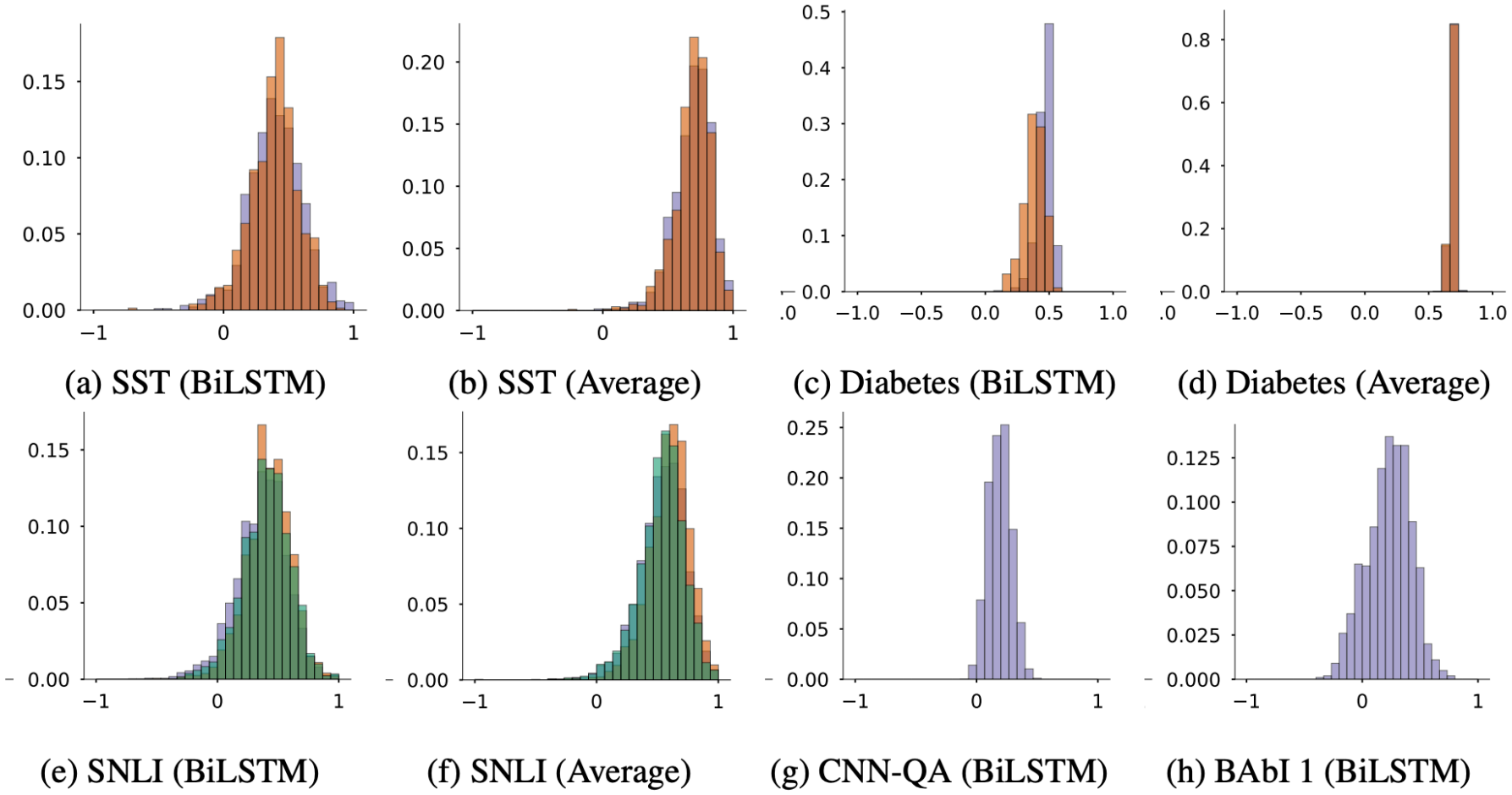


Figure 2: Histogram of **Kendall** τ between attention and gradients. Encoder variants are denoted parenthetically; colors indicate predicted classes. Exhaustive results are available for perusal online.

Attention is not explanation, 2019.

Is Attention Explanation?

Asks two questions:

1. To what extent do induced attention weights correlate with measures of feature importance (gradients and leave-one-out (LOO) methods)?

Attention weakly (and variably) correlates with gradients for many task – what is attended is not always what is most influential.

2. Would alternative attention weights necessarily yield different predictions?

Attention is not explanation, 2019.

Step 4: Run an optimization procedure to find new attentions that maximize difference (JS-Divergence) with original set, but doesn't change output.

Algorithm 3 Finding adversarial attention weights

$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x}), \hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \text{Dec}(\mathbf{h}, \hat{\alpha})$
$\alpha^{(1)}, ..., \alpha^{(k)} \leftarrow$ Optimize Eq 1
for $i \leftarrow 1$ to k **do**
  $\hat{y}^{(i)} \leftarrow \text{Dec}(\mathbf{h}, \alpha^{(i)})$ ▷ $\mathbf{h}$ is not changed
  $\Delta\hat{y}^{(i)} \leftarrow \text{TVD}[\hat{y}, \hat{y}^{(i)}]$
  $\Delta\alpha^{(i)} \leftarrow \text{JSD}[\hat{\alpha}, \alpha^{(i)}]$
end for
ϵ-max JSD $\leftarrow \max_i \mathbb{1}[\Delta\hat{y}^{(i)} \leq \epsilon]\Delta\alpha^{(i)}$

Do there exist adversarial attentions?

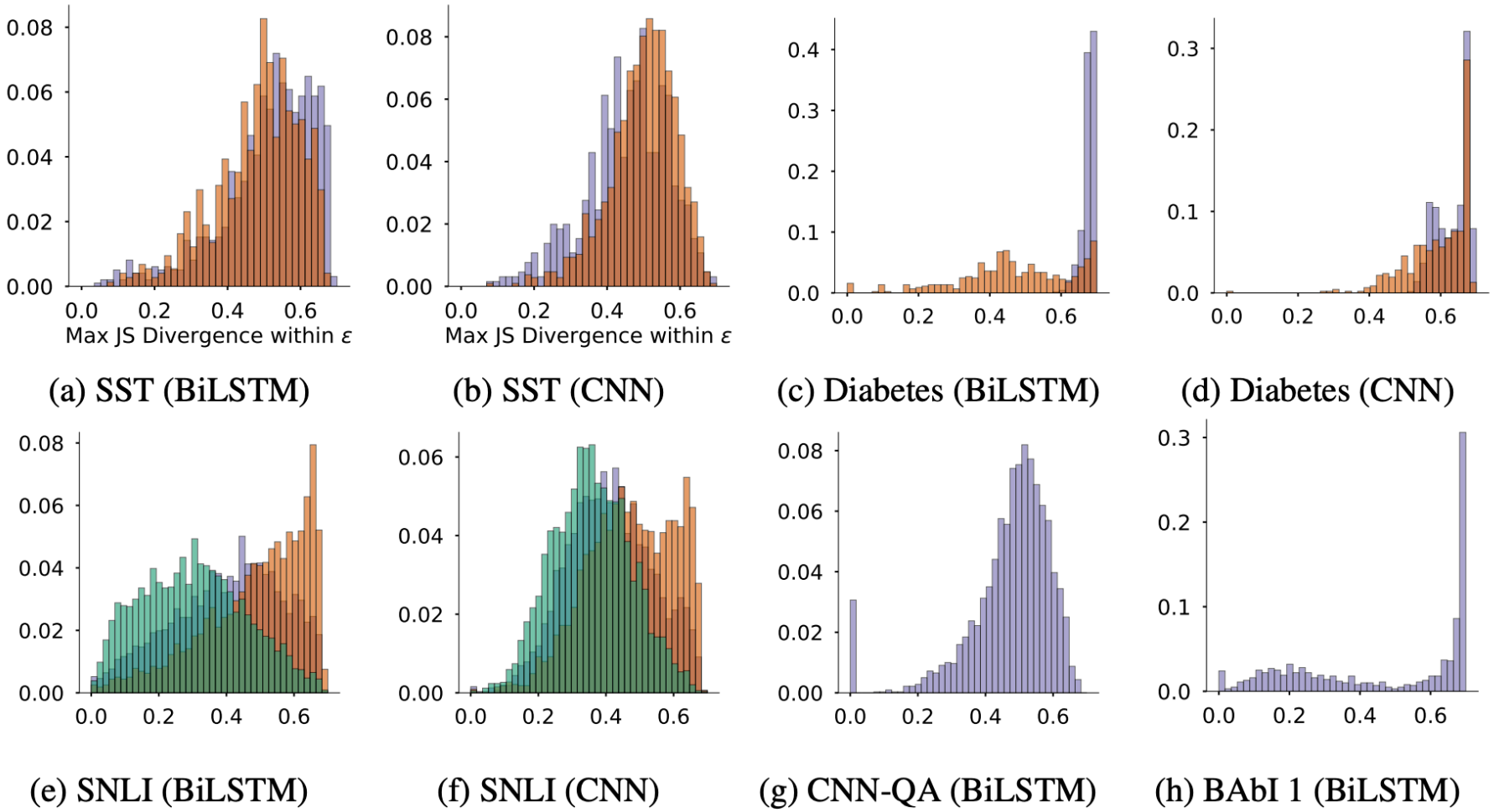


Figure 7: Histogram of **maximum adversarial JS Divergence** (**ϵ-max JSD**) between original and adversarial attentions over all instances. In all cases shown, $|\hat{y}^{adv} - \hat{y}| < \epsilon$. Encoders are specified in parantheses.

Is Attention Explanation?

Asks two questions:

1. To what extent do induced attention weights correlate with measures of feature importance (gradients and leave-one-out (LOO) methods)?

Attention weakly (and variably) correlates with gradients for many task – what is attended is not always what is most influential.

2. Would alternative attention weights necessarily yield different predictions?

Attention distribution can change drastically without changing the model output. (Some BIG caveats here.)

Attention is not explanation, 2019.

Is Attention Explanation?

Attention is not not Explanation

S Wiegreffe, Y Pinter - arXiv preprint arXiv:1908.04626, 2019 - arxiv.org

Attention mechanisms play a central role in NLP systems, especially within recurrent neural network (RNN) models. Recently, there has been increasing interest in whether or **not** the intermediate representations offered by these modules may be used to explain the …

☆ 99 Cited by 12 Related articles All 4 versions »

2. Would alternative attention weights necessarily yield different predictions?

Previous experiment modified attention *values* directly. Not the model. May be impossible to achieve the optimized attention.

Attention is not not explanation, 2020.

Is Attention Explanation?

Train a new model to:

- Mimic the output of the previous one
- Diverge in attention weights

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \text{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \lambda \text{ KL}(\boldsymbol{\alpha}_a^{(\mathbf{i})} \parallel \boldsymbol{\alpha}_b^{(\mathbf{i})}),$$

Previous paper's attention weights don't seem achievable in a real model.

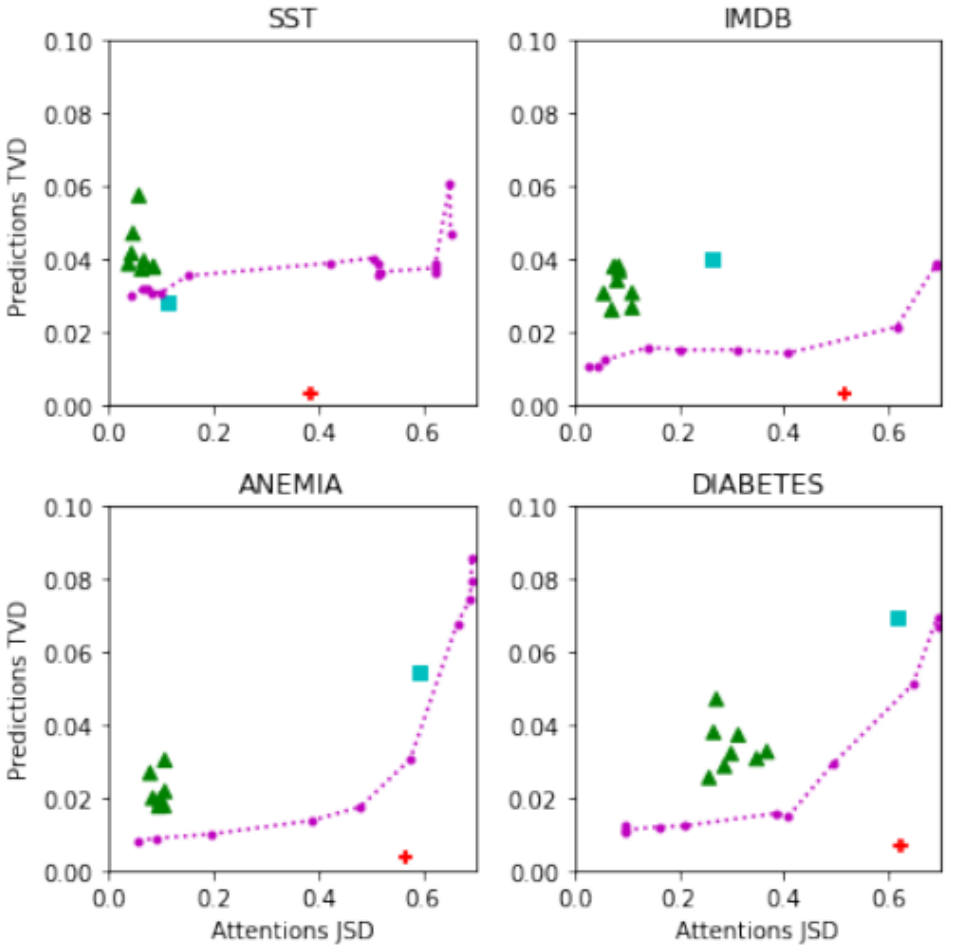
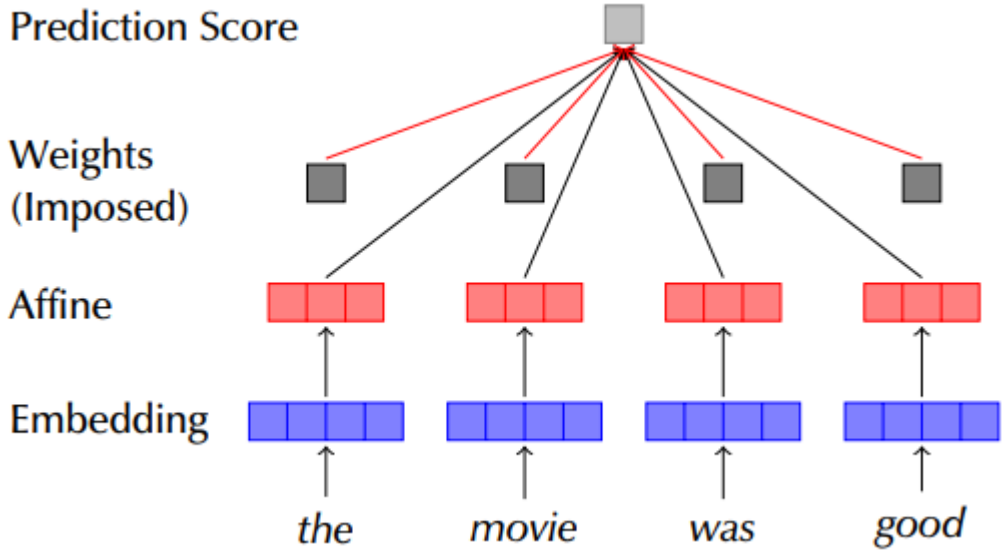


Figure 5: Averaged per-instance test set JSD and TVD from base model for each model variant. JSD is bounded at $\sim$ 0.693. ▲: random seed; ■: uniform weights; dotted line: our adversarial setup as λ is varied; +: adversarial setup from Jain and Wallace (2019).

Attention is not not explanation, 2020.

Another way to measure importance of attention:



Train a word-level model that uses the frozen attention weights from the original model.

Guide weights	Diab.	Anemia	SST	IMDb
UNIFORM	0.404	0.873	0.812	0.863
TRAINED MLP	0.699	0.920	0.817	0.888
BASE LSTM	**0.753**	0.931	**0.824**	**0.905**
ADVERSARY (4)	0.503	**0.932**	0.592	0.700

Uniform: uniform attention
Trained: train attention mechanism with MLP
Base LSTM: use attention generated by LSTM
Adversary: Use attention from previous paper

Gap between Uniform and Base LSTM suggests attention does capture some notion of token importance.

Attention is not not explanation, 2020.

Learning Objectives

Be able to answer:

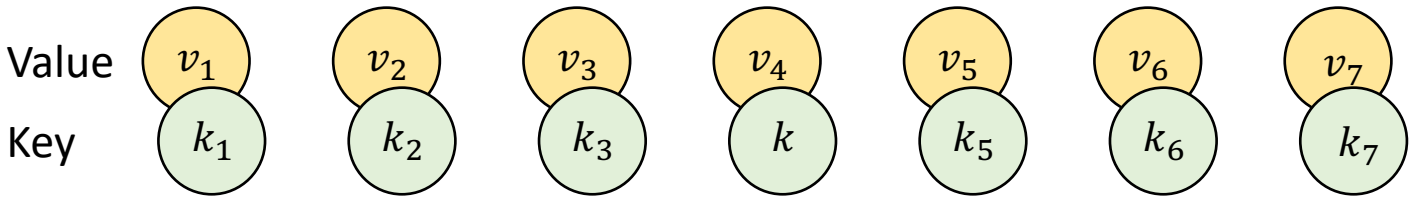
- ~~What is an attention mechanism?~~
- ~~Difference between soft and hard attention?~~

- ~~What are some examples of attention?~~

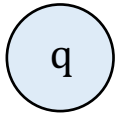
- ~~How does attention relate to interpretability?~~

Common Structure for Key-Query-Value Attention

Candidates



Query



Attention Function

$$\alpha_i = f(\mathrm{q}, \mathrm{k_i})$$

s.t. $\sum_i a_i = 1$

$\forall i, a_i \geq 0$

Attended Feature

$$c = \sum_i f(\mathrm{q}, \mathrm{k_i}) * \mathrm{c_i}$$

Next time

Next Time: *"If attention works so well, why do we need memory?"* –NLP 2017 - present