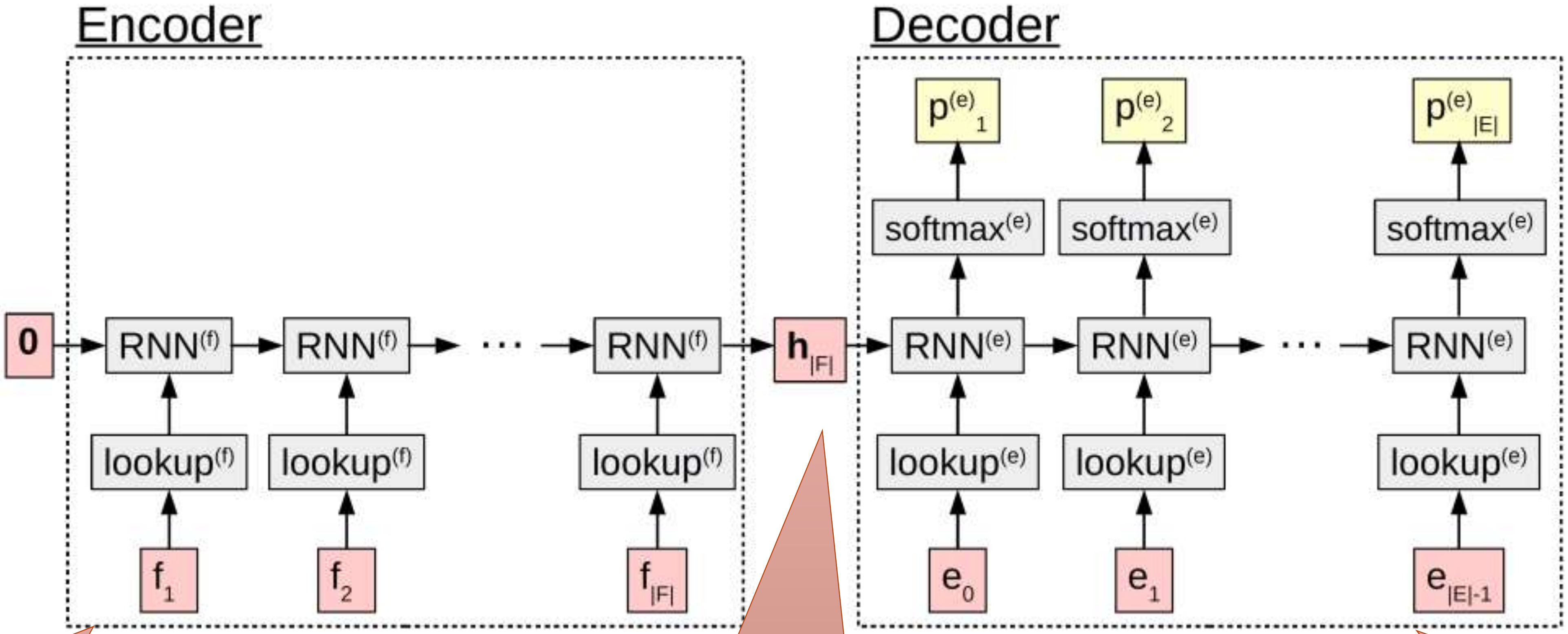
COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Attention models & Current topics in Neural MT

CMSC 470

Marine Carpuat

P(E|F) as an encoder-decoder model



The Encoder models the input/source entence F=(f1,….f|F|)

The decoder hidden state is initialized with the last hidden state of the encoder

The Decoder models the output/target sentence E=(e1,….e|E|).

P(E|F) as an encoder-decoder model

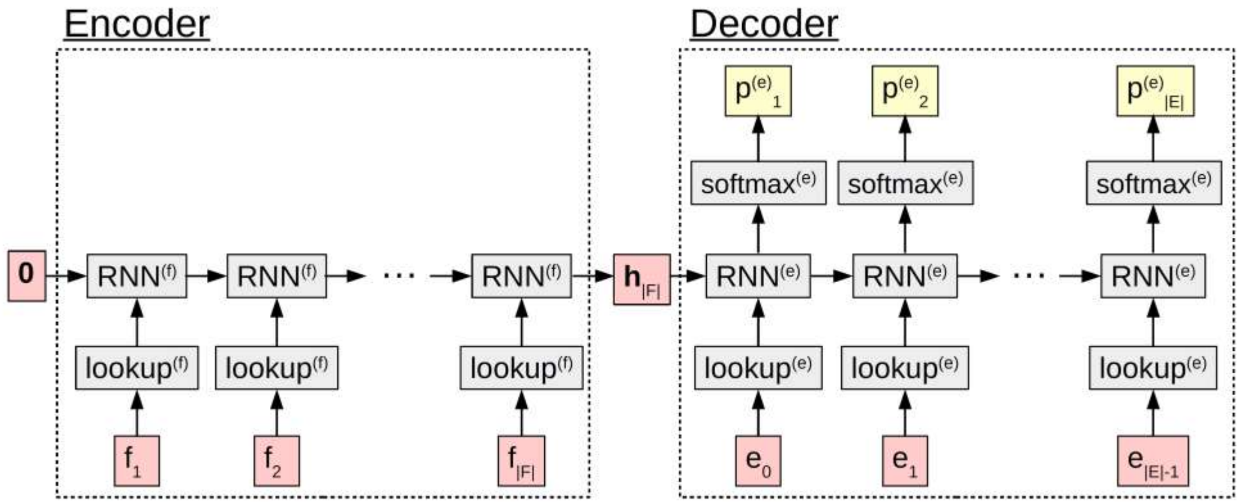
$$\boldsymbol{m}_t^{(f)} = M_{\cdot,f_t}^{(f)}$$

$$\boldsymbol{h}_t^{(f)} = \begin{cases} \mathrm{RNN}^{(f)}(\boldsymbol{m}_t^{(f)}, \boldsymbol{h}_{t-1}^{(f)}) & t \geq 1, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

$$\boldsymbol{m}_t^{(e)} = M_{\cdot,e_{t-1}}^{(e)}$$

$$\boldsymbol{h}_t^{(e)} = \begin{cases} \mathrm{RNN}^{(e)}(\boldsymbol{m}_t^{(e)}, \boldsymbol{h}_{t-1}^{(e)}) & t \geq 1, \\ \boldsymbol{h}_{|F|}^{(f)} & \text{otherwise.} \end{cases}$$

$$\boldsymbol{p}_t^{(e)} = \mathrm{softmax}(W_{hs}\boldsymbol{h}_t^{(e)} + b_s)$$



Problem with previous encoder-decoder model

- Long-distance dependencies remain a problem
- A single vector represents the entire source sentence
  - No matter its length
- Solution: attention mechanism
  - An example of incorporating inductive bias in model architecture

Attention model intuition

- Encode each word in source sentence into a vector
- When decoding, perform a linear combination of these vectors, weighted by "attention weights"
- Use this combination when predicting next word

[Bahdanau et al. 2015]

Attention model
Source word representations

- We can use representations from bidirectional RNN encoder

$$\overrightarrow{\boldsymbol{h}}_j^{(f)} = \mathrm{RNN}(\mathrm{embed}(f_j), \overrightarrow{\boldsymbol{h}}_{j-1}^{(f)})$$
$$\overleftarrow{\boldsymbol{h}}_j^{(f)} = \mathrm{RNN}(\mathrm{embed}(f_j), \overleftarrow{\boldsymbol{h}}_{j+1}^{(f)}).$$

$$\boldsymbol{h}_j^{(f)} = [\overleftarrow{\boldsymbol{h}}_j^{(f)}; \overrightarrow{\boldsymbol{h}}_j^{(f)}].$$

- And concatenate them in a matrix

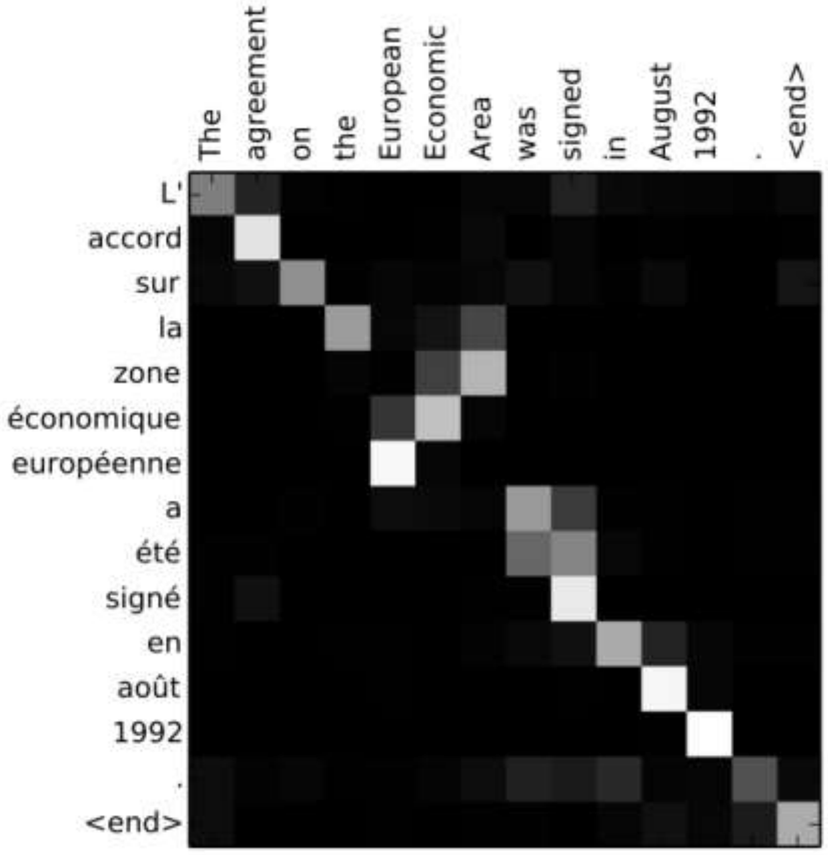
$$H^{(f)} = \mathrm{concat\_col}(\boldsymbol{h}_1^{(f)}, \dots, \boldsymbol{h}_{|F|}^{(f)}).$$

Attention model
Create a source context vector

- Attention vector:
  - Entries between 0 and 1
  - Interpreted as weight given to each source word when generating output at time step t

$$\underbrace{\boldsymbol{c}_t}_{\text{Context vector}} = H^{(f)} \underbrace{\boldsymbol{\alpha}_t}_{\text{Attention vector}}.$$

Attention model
Illustrating attention weights



Attention model
How to calculate attention scores

$$\boldsymbol{h}_t^{(e)} = \mathrm{enc}([\mathrm{embed}(e_{t-1}); \boldsymbol{c}_{t-1}], \boldsymbol{h}_{t-1}^{(e)}).$$

$$a_{t,j} = \mathrm{attn\_score}(\boldsymbol{h}_j^{(f)}, \boldsymbol{h}_t^{(e)}).$$

$$\boldsymbol{\alpha}_t = \mathrm{softmax}(\boldsymbol{a}_t).$$

$$\boldsymbol{p}_t^{(e)} = \mathrm{softmax}(W_{hs}[\boldsymbol{h}_t^{(e)}; \boldsymbol{c}_t] + b_s).$$

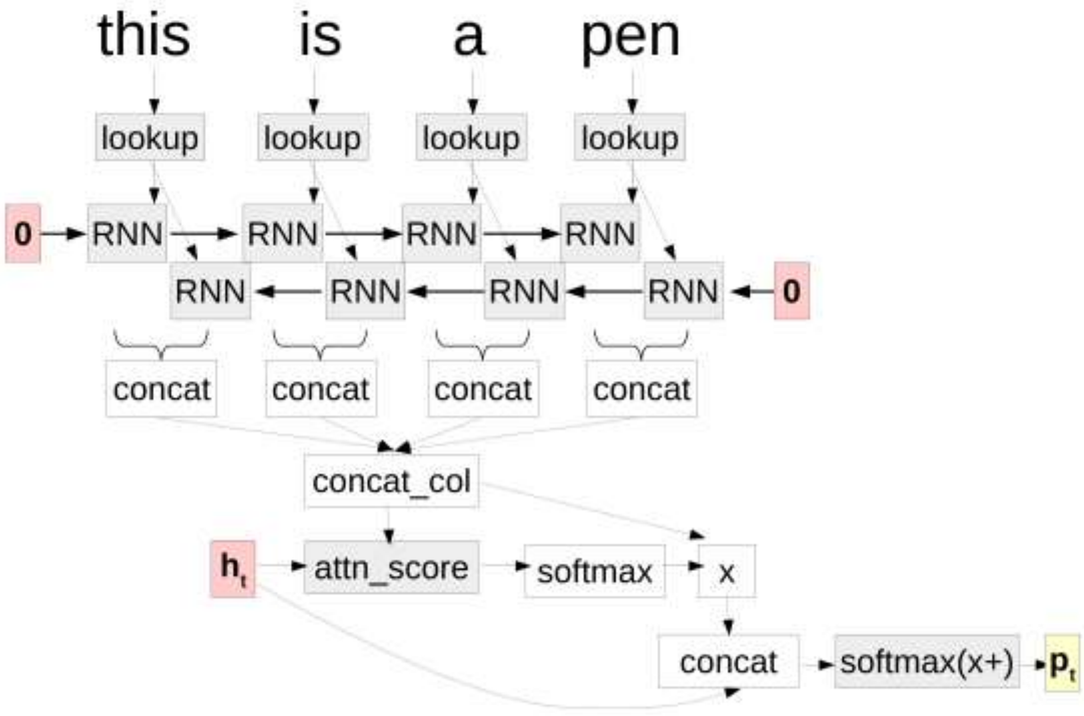


Figure 28: A computation graph for attention.

Attention model
Various ways of calculating attention score

- Dot product

$$\text{attn\_score}(\boldsymbol{h}_j^{(f)}, \boldsymbol{h}_t^{(e)}) := \boldsymbol{h}_j^{(f)\intercal} \boldsymbol{h}_t^{(e)}.$$

- Bilinear function

$$\text{attn\_score}(\boldsymbol{h}_j^{(f)}, \boldsymbol{h}_t^{(e)}) := \boldsymbol{h}_j^{(f)\intercal} W_a \boldsymbol{h}_t^{(e)}.$$

- Multi-layer perceptron (original formulation in Bahdanau et al.)

$$\text{attn\_score}(\boldsymbol{h}_t^{(e)}, \boldsymbol{h}_j^{(f)}) := \boldsymbol{w}_{a2}^{\intercal} \tanh(W_{a1}[\boldsymbol{h}_t^{(e)}; \boldsymbol{h}_j^{(f)}])$$

Advantages of attention

- Helps illustrate/interpret translation decisions

- Can help insert translations for out-of-vocabulary words
  - By copying or look up in external dictionary

- Can incorporate linguistically motivated priors in model

Attention extensions
Bidirectional constraints (Cohn et al. 2015)

- Intuition: attention should be similar in forward and backward translation directions

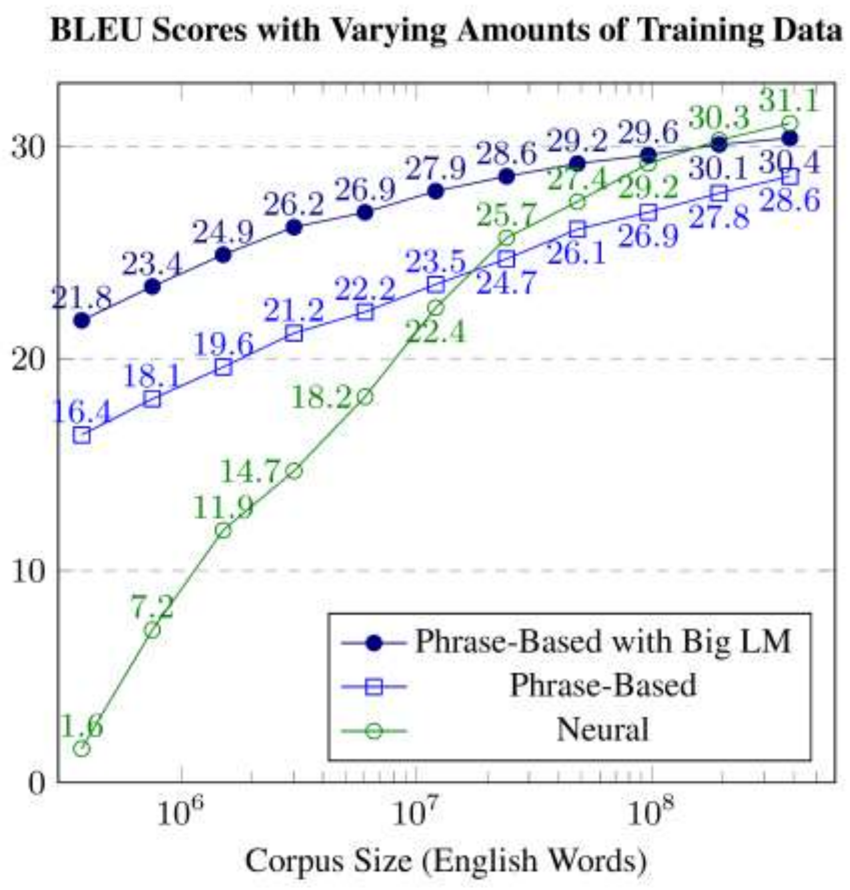
- Method: train so that we get a bonus based on the trace of matrix product for training in both directions

$$\mathrm{tr}(A_{X \to Y} A^{\intercal}_{Y \to X})$$

Attention extensions
An active area of research

- Attend to multiple sentences (Zoph et al. 2015)
- Attend to a sentence and an image (Huang et al. 2016)
- Incorporate bias from alignment models

Issue with Neural MT: it only works well in high-resource settings

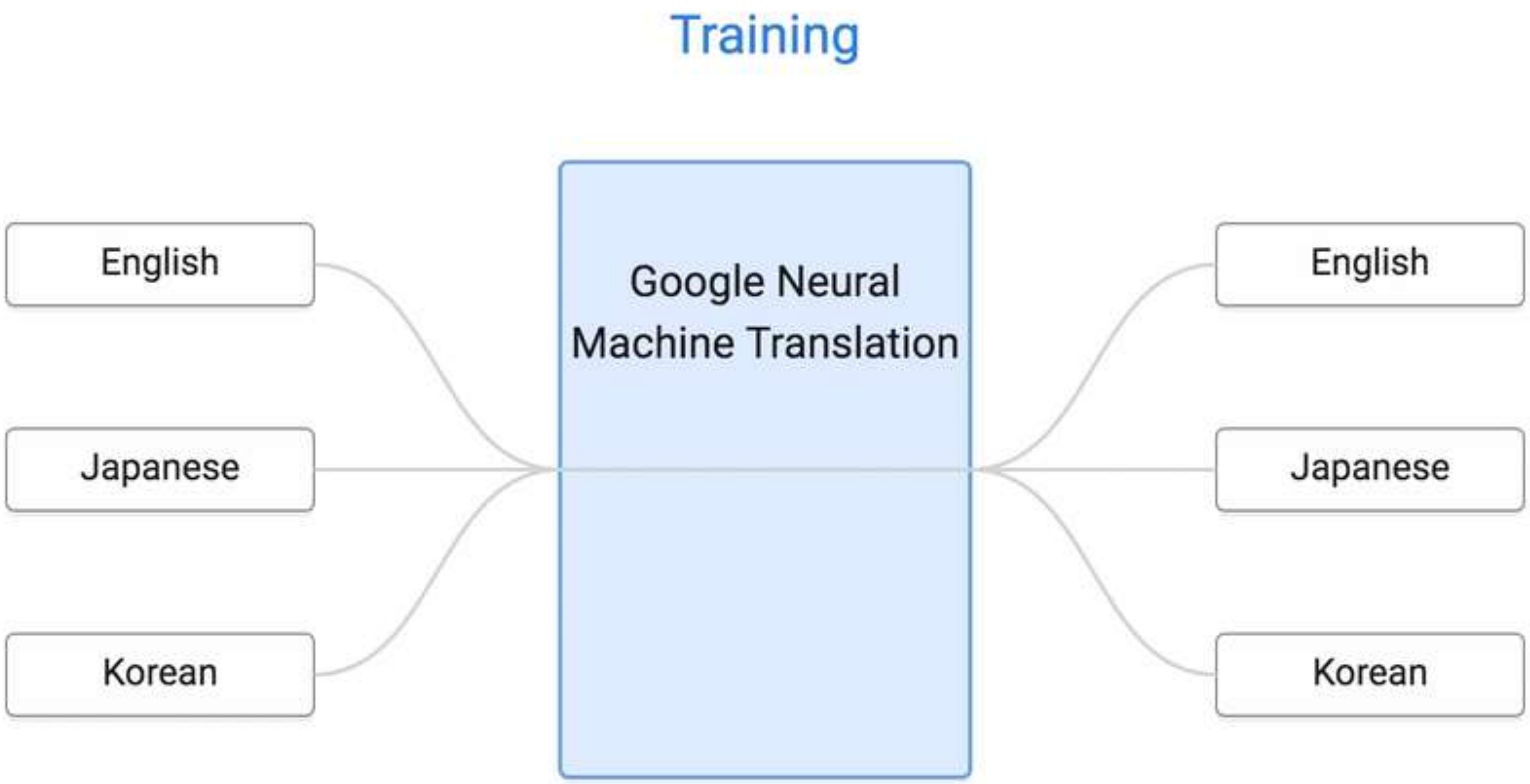


[Koehn & Knowles 2017]

Ongoing research

- Learn from other sources of supervision than pairs (E,F)
  - Monolingual text
  - Multiple languages
- Incorporate linguistic knowledge
  - As additional embeddings
  - As prior on network structure or parameters
  - To make better use of training data

The Google Multilingual NMT System
[Johnson et al. 2017]



The Google Multilingual NMT System
[Johnson et al. 2017]

- A simple idea
  - Train on sentence pairs in all languages
  - Add token to mark target language

```
<2es> Hello, how are you? -> Hola, ¿cómo estás?
```

- Helps most for low-resources languages
- Enables zero-shot translation
- Can handle code-switched input

Issue with NMT: Exposure Bias

- Mismatch between contexts seen at training and test time
  - During training, model produces outputs based on sequence prefix from reference translation
    - This is sometimes called **teacher forcing**
  - During decoding, the model produce outputs based on its own previous predictions
  - This is called the **exposure bias problem**

- Idea: expose model to its own predictions during training
  - Challenges: model predictions are very noisy, esp. during early training stages
  - Challenges:
  - Currently addressed using imitation learning and reinforcement learning algorithms

Issue with Neural MT: sentences are translated out-of-context

In fairness, Miller did not attack the statue itself.
[...]
But he did attack its meaning [...]

HUMAN	MT
Um fair zu bleiben, Miller griff nicht die Statue selbst an. [...] Aber er griff deren Bedeutung an [...]	Fairerweise hat Miller die Statue nicht selbst angegriffen. [...] Aber er griff seine Bedeutung an [...]

Example from Rico Sennrich (2018)

Issue with Neural MT: sentences are translated out-of-context

Weidezaunprojekt ist elementar

Das Fischerbacher Weidezaun-Projekt ist ein Erfolgsprojekt und wird im kommenden Jahr fortgesetzt.

HUMAN	MT
Pasture fence project is fundamental	Electric fence project is basic
The Fischerbach pasture fence project is a successful project and will be continued next year.	The Fischerbacher Weidezaun-Project is a success and will be continued in the coming year.

Example from Rico Sennrich (2018)

Issue with Neural MT: sentences are translated out-of-context

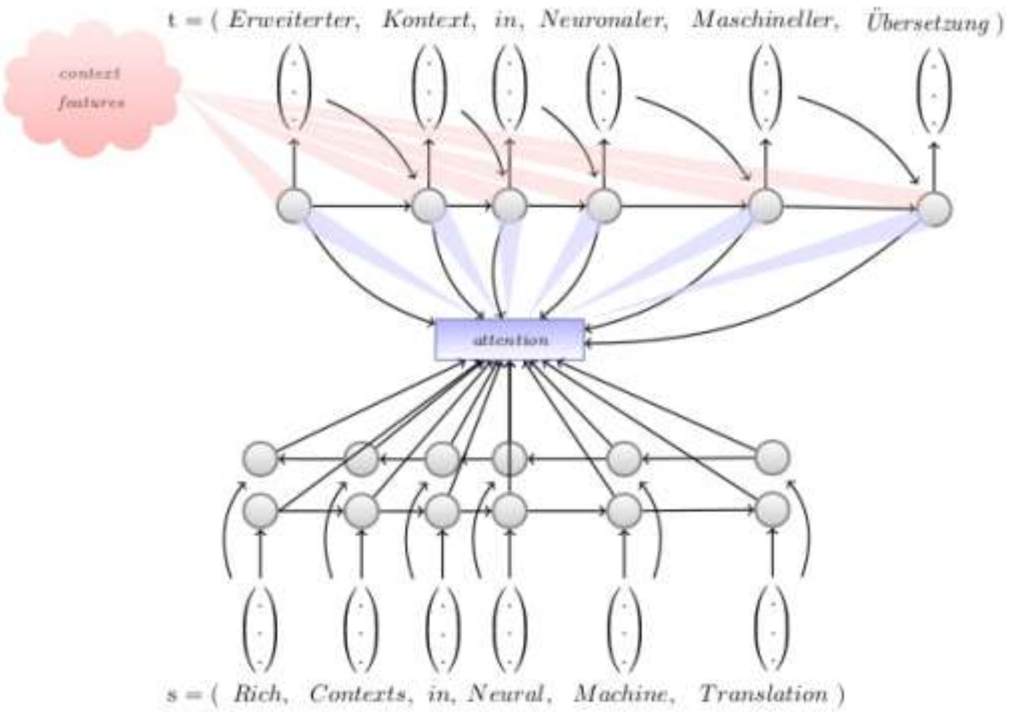
该款机器人使用语音合成、[...]

曾获得国际消费电子产品展（CES）[...]

HUMAN	MT
This robot uses speech synthesis, [...] with conversational [...] features.	Using speech synthesis [...] the robot has the functions of chatting conversation [...]
It has won two major CES awards [...]	Has won two awards at the International Consumer Electronics Exhibition (CES) [...]

Example from Rico Sennrich (2018)

Idea: Translate documents, not sentences!



contextual sentences as additional input

[Jean et al., 2017, Wang et al., 2017, Tiedemann and Scherrer, 2017, Bawden et al., 2018, Voita et al., 2018, Maruf and Haffari, 2018]

State-of-the-art neural MT models are very powerful, but still make many errors

https://www.youtube.com/watch?v=3-rfBsWmo0M

Beyond MT: Encoder-Decoder can be used as Conditioned Language Models to generate text Y according to some specification X

Input *X*	Output *Y* (**Text**)	Task
Structured Data	NL Description	NL Generation
English	Japanese	Translation
Document	Short Description	Summarization
Utterance	Response	Response Generation
Image	Text	Image Captioning
Speech	Transcript	Speech Recognition

Neural Machine Translation
What you should know

- How to formulate machine translation as a sequence-to-sequence transformation task
- How to model P(E|F) using RNN encoder-decoder models, with and without attention
- Algorithms for producing translations
  - Ancestral sampling, greedy search, beam search
- How to train models
  - loss functions, parameter update rules, batch vs online vs minibatch training
- Examples of weaknesses of neural MT models and how to address them
  - Bidirectional encoder, length bias, multilingual models
- Determine whether a NLP task should be addressed with neural sequence-to-sequence models