

Introduction to Neural Machine Translation (3/3)

Marine Carpuat

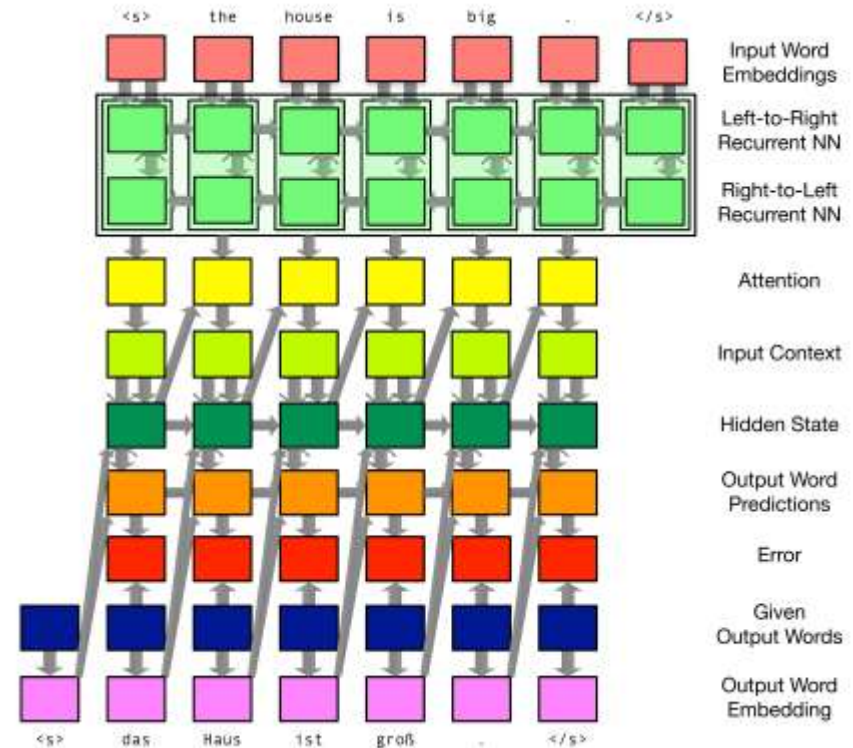
Computer Science

University of Maryland

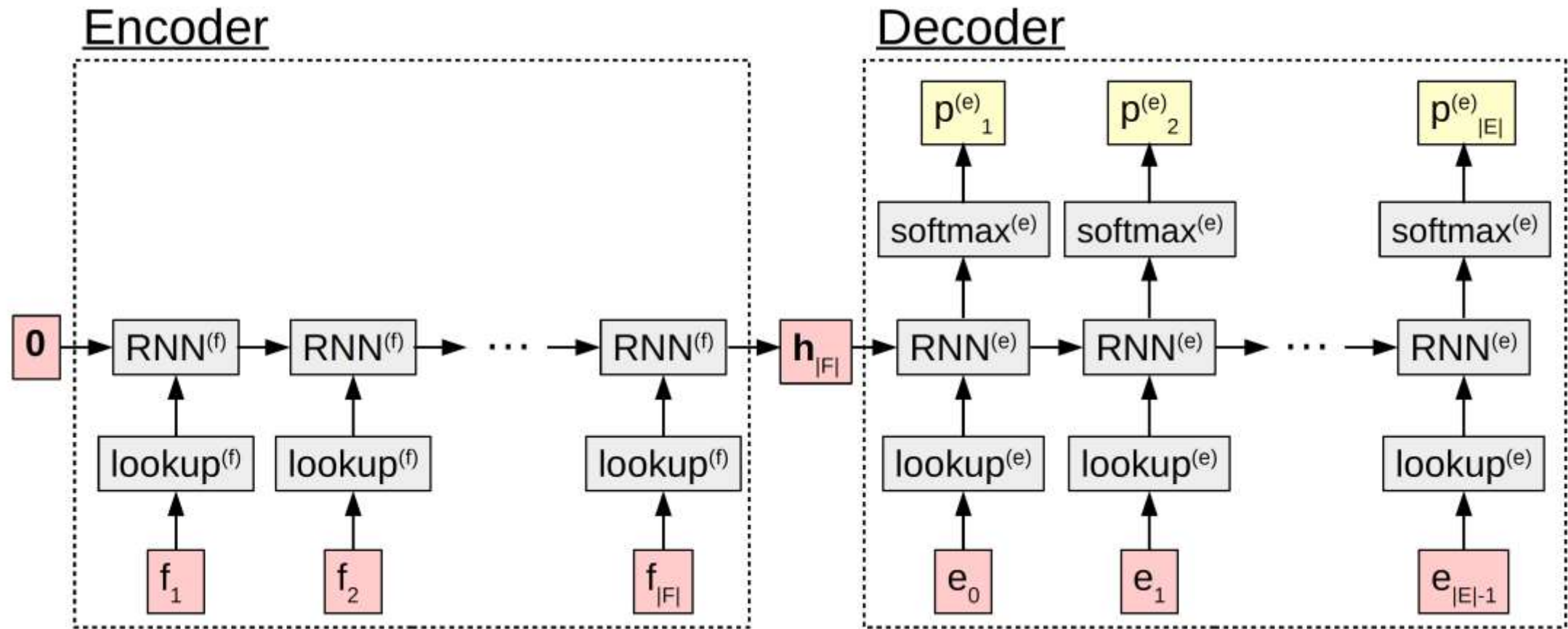


Roadmap

- Evaluating machine translation
- Introduction to neural networks
- Modeling sequences of words with neural language models
- Translating with encoder-decoder models
- Attention mechanism



An RNN Encoder-Decoder for $P(E|F)$



Generating Output (1): Ancestral Sampling

- Randomly generate words one by one
- Until end of sentence symbol
- Done!

```
while  $y_{j-1} \neq \text{"</s>"}$ :  
   $y_j \sim P(y_j \mid X, y_1, \dots, y_{j-1})$ 
```

Generating Output (2): Greedy search

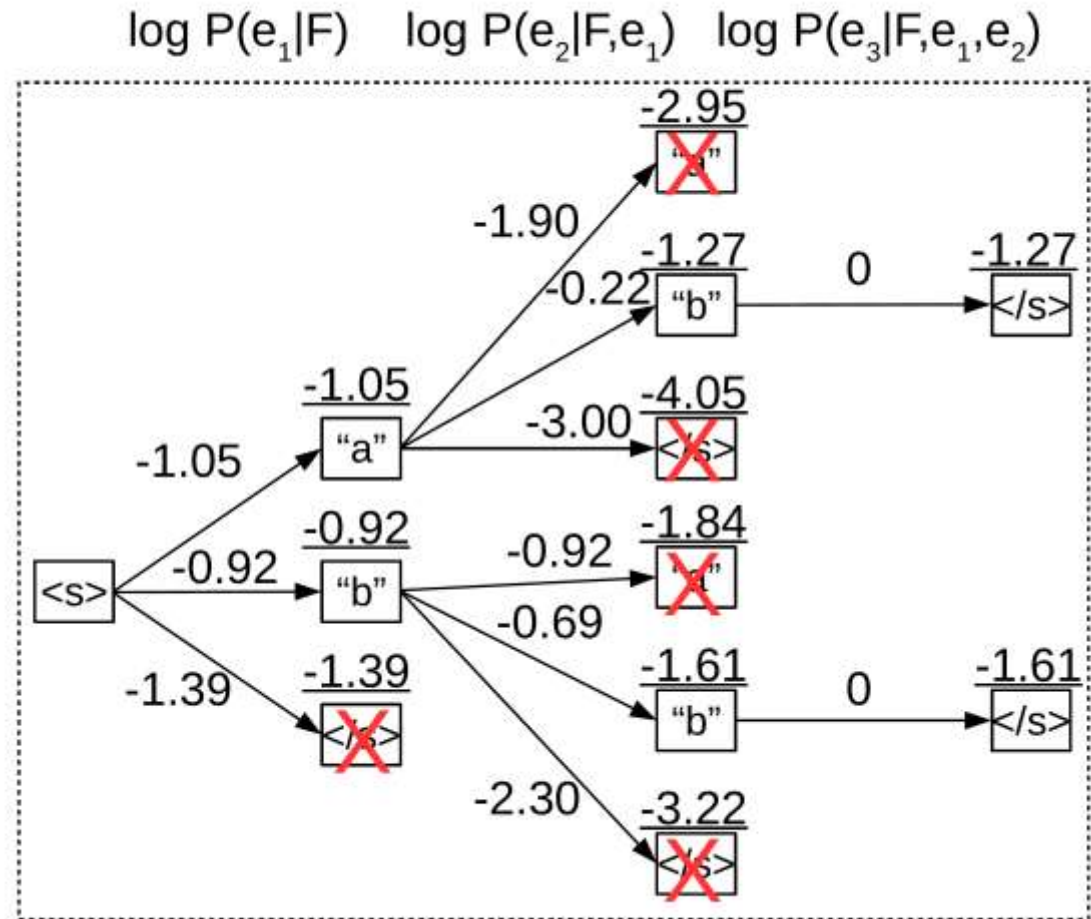
- One by one, pick single highest probability word

```
while  $y_{j-1} \neq \text{"</s>"}$ :  
   $y_j = \operatorname{argmax} P(y_j \mid X, y_1, \dots, y_{j-1})$ 
```

- Problems
 - Often generates easy words first
 - Often prefers multiple common words to rare words

Generating Output (3): Beam Search

Idea: keep the k top hypotheses at each time step



Training

- Same as for RNN language modeling
- Training examples: pairs of sentences (E,F)
- Loss function
 - Negative log-likelihood of training data
 - Total loss for one example (sentence) = sum of loss at each time step (word)

Note that training loss differs from evaluation metric (BLEU)

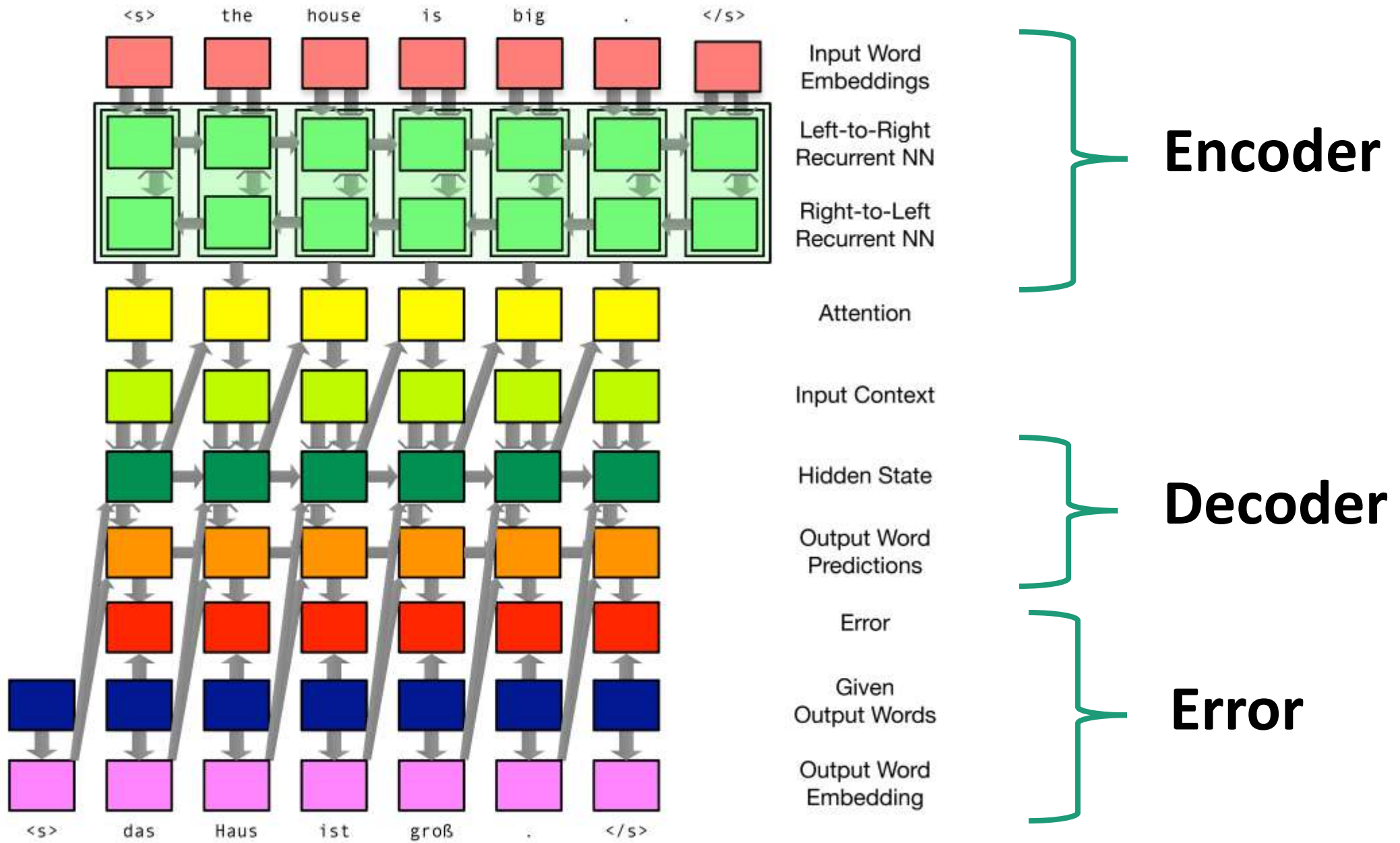
N-gram overlap between machine translation output and reference translation

Compute precision for n-grams of size 1 to 4

Add brevity penalty (for too short translations)

$$\text{BLEU} = \min \left(1, \frac{\text{output-length}}{\text{reference-length}} \right) \left(\prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}}$$

Typically computed over the entire corpus, not single sentences



Bidirectional encoder

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

$$\overleftarrow{h}_t^{(f)} = \begin{cases} \overleftarrow{\text{RNN}}^{(f)}(\mathbf{m}_t^{(f)}, \overleftarrow{h}_{t+1}^{(f)}) & t \leq |F|, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

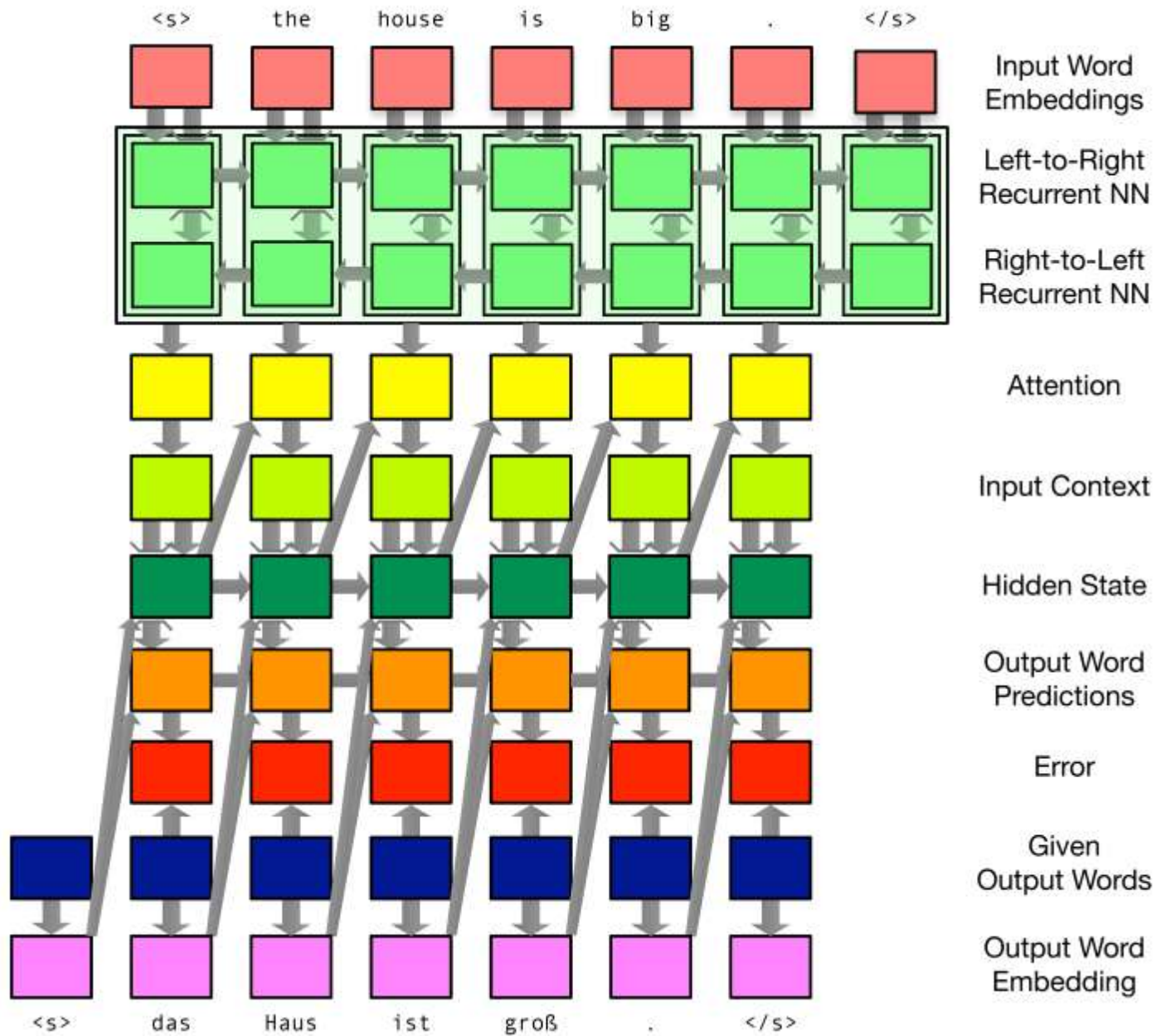
Motivation:

- Help bootstrap learning
- By shortening length of dependencies

$$\mathbf{h}_0^{(e)} = \tanh(W_{\vec{f}e} \vec{h}_{|F|} + W_{\overleftarrow{f}e} \overleftarrow{h}_1 + \mathbf{b}_e)$$

Motivation:

- Take 2 hidden vectors from source encoder
- Combine them into a vector of size required by decoder



Encoder

Attention

Decoder

Error

Attention Mechanism

- Some remaining issues with previous encoder-decoder model
 - Some long-distance dependencies remain a problem
 - A single vector represents the entire source sentence
 - No matter its length
- Solution: attention mechanism
 - This is what made neural machine translation work as well or better as earlier statistical models [Bahdanau et al. 2015]
 - 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

$$\begin{aligned}\vec{\mathbf{h}}_j^{(f)} &= \text{RNN}(\text{embed}(f_j), \vec{\mathbf{h}}_{j-1}^{(f)}) \\ \overleftarrow{\mathbf{h}}_j^{(f)} &= \text{RNN}(\text{embed}(f_j), \overleftarrow{\mathbf{h}}_{j+1}^{(f)}).\end{aligned}$$

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

- And concatenate them in a matrix

$$H^{(f)} = \text{concat_col}(\mathbf{h}_1^{(f)}, \dots, \mathbf{h}_{|F|}^{(f)}).$$

Attention model

Create a source context vector

$$\mathbf{c}_t = H^{(f)} \boldsymbol{\alpha}_t.$$

Context vector

Attention vector

- Attention vector:

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

Attention model

How to calculate attention scores

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

$$a_{t,j} = \text{attn_score}(\mathbf{h}_j^{(f)}, \mathbf{h}_t^{(e)}).$$

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

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

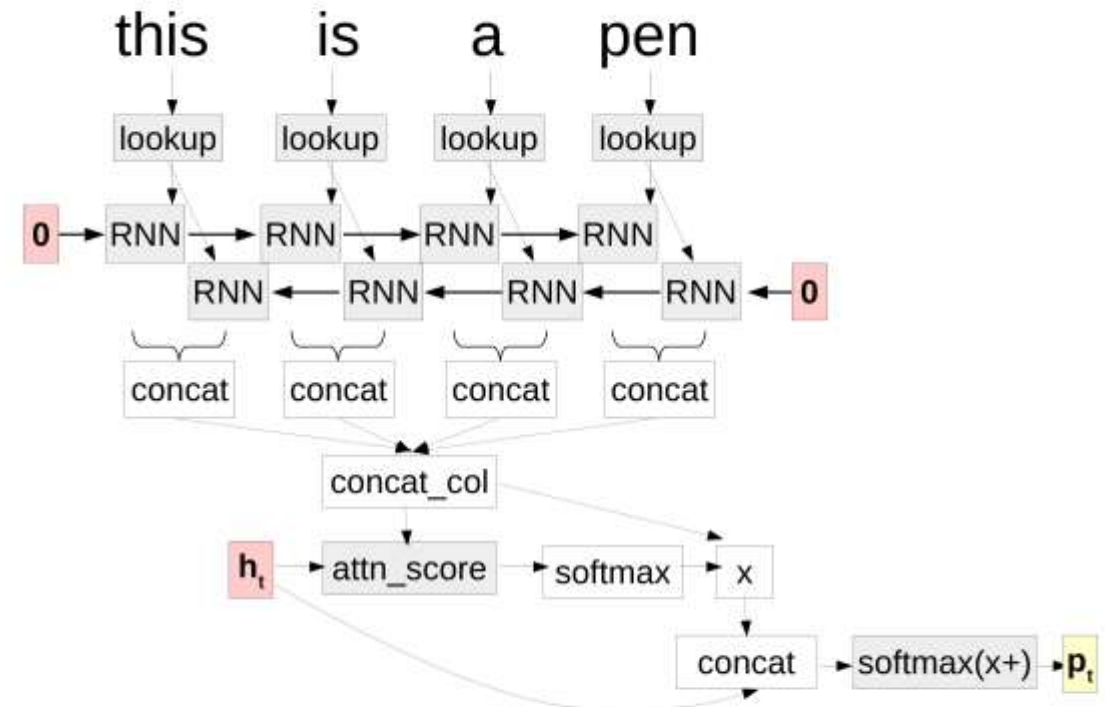


Figure 28: A computation graph for attention.

Attention model

Various ways of calculating attention score

- Dot product

$$\text{attn_score}(\mathbf{h}_j^{(f)}, \mathbf{h}_t^{(e)}) := \mathbf{h}_j^{(f)\top} \mathbf{h}_t^{(e)}.$$

- Bilinear function

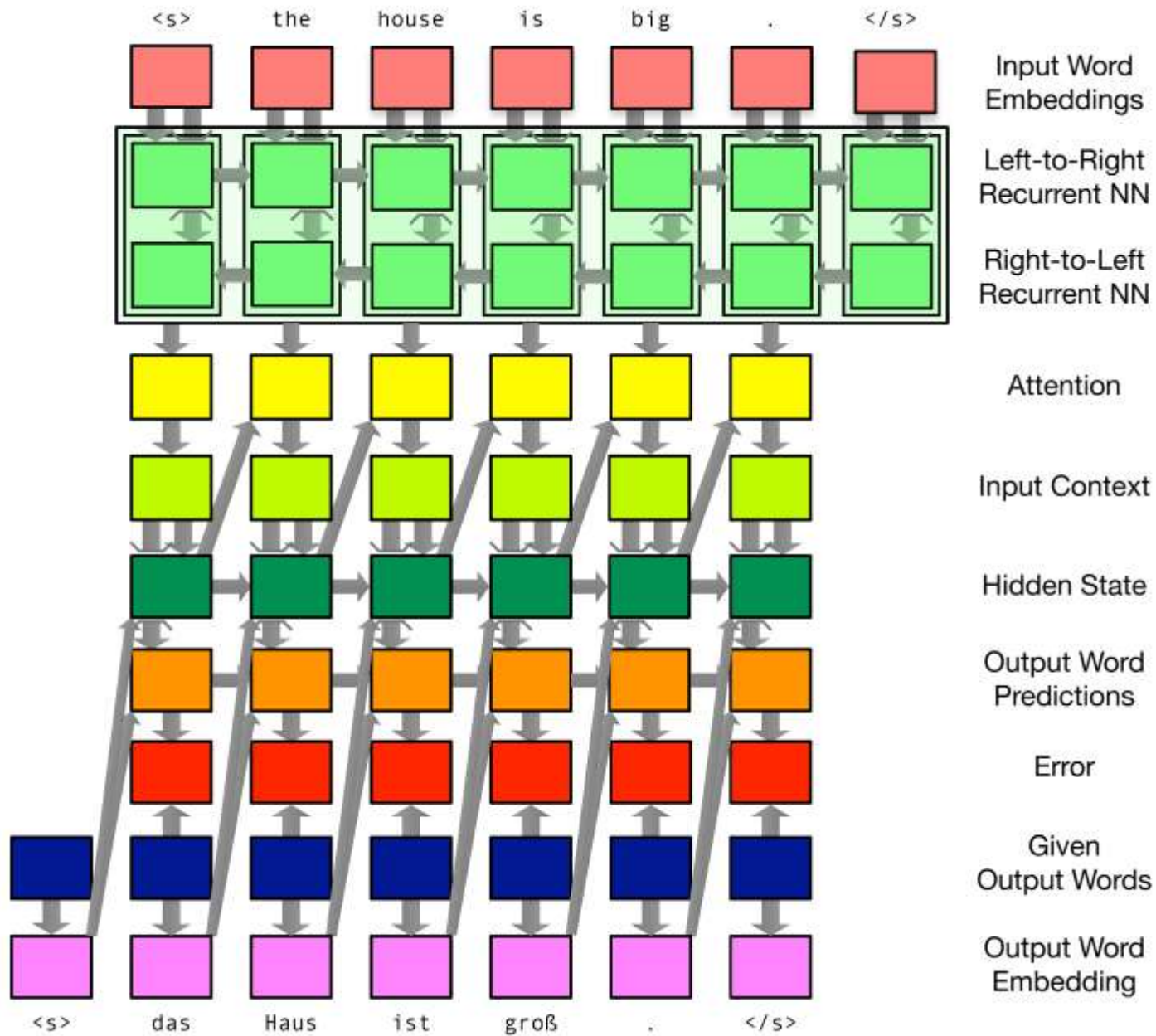
$$\text{attn_score}(\mathbf{h}_j^{(f)}, \mathbf{h}_t^{(e)}) := \mathbf{h}_j^{(f)\top} W_a \mathbf{h}_t^{(e)}.$$

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

$$\text{attn_score}(\mathbf{h}_t^{(e)}, \mathbf{h}_j^{(f)}) := \mathbf{w}_{a2}^\top \tanh(W_{a1}[\mathbf{h}_t^{(e)}; \mathbf{h}_j^{(f)}])$$

Advantages of attention

- Helps illustrate/interpret translation decisions
- Can help insert translations for OOV
 - By copying or look up in external dictionary
- Can incorporate linguistically motivated priors in model
- Can incorporate additional data



Encoder

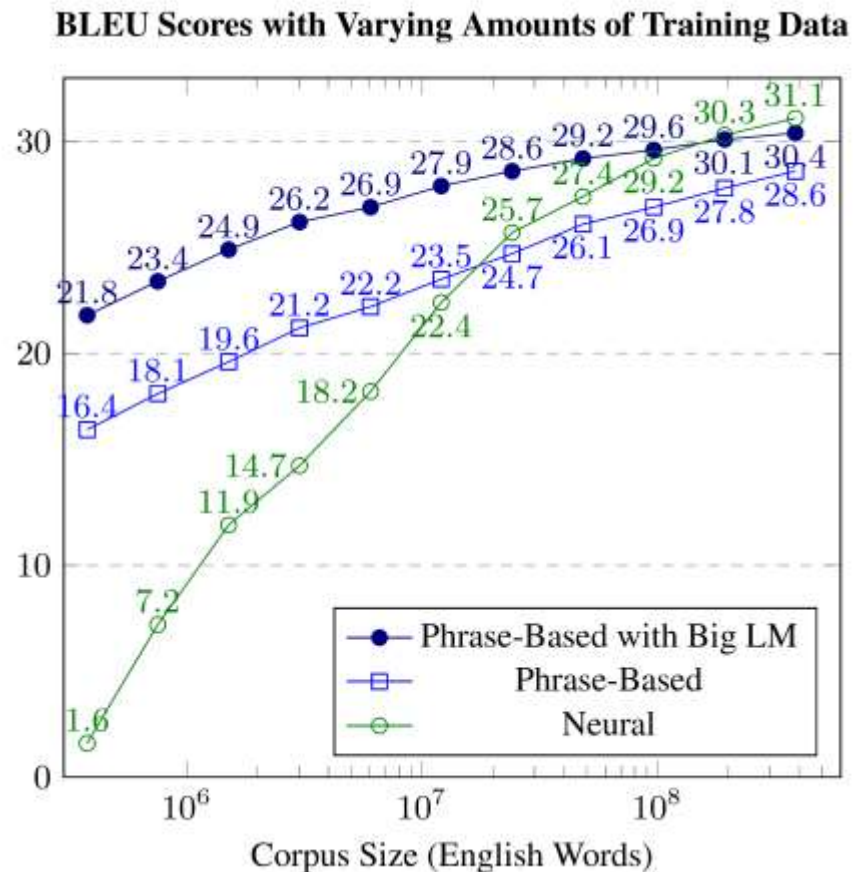
Attention

Decoder

Error

Where does neural MT break?

Neural MT only helps in high-resource settings



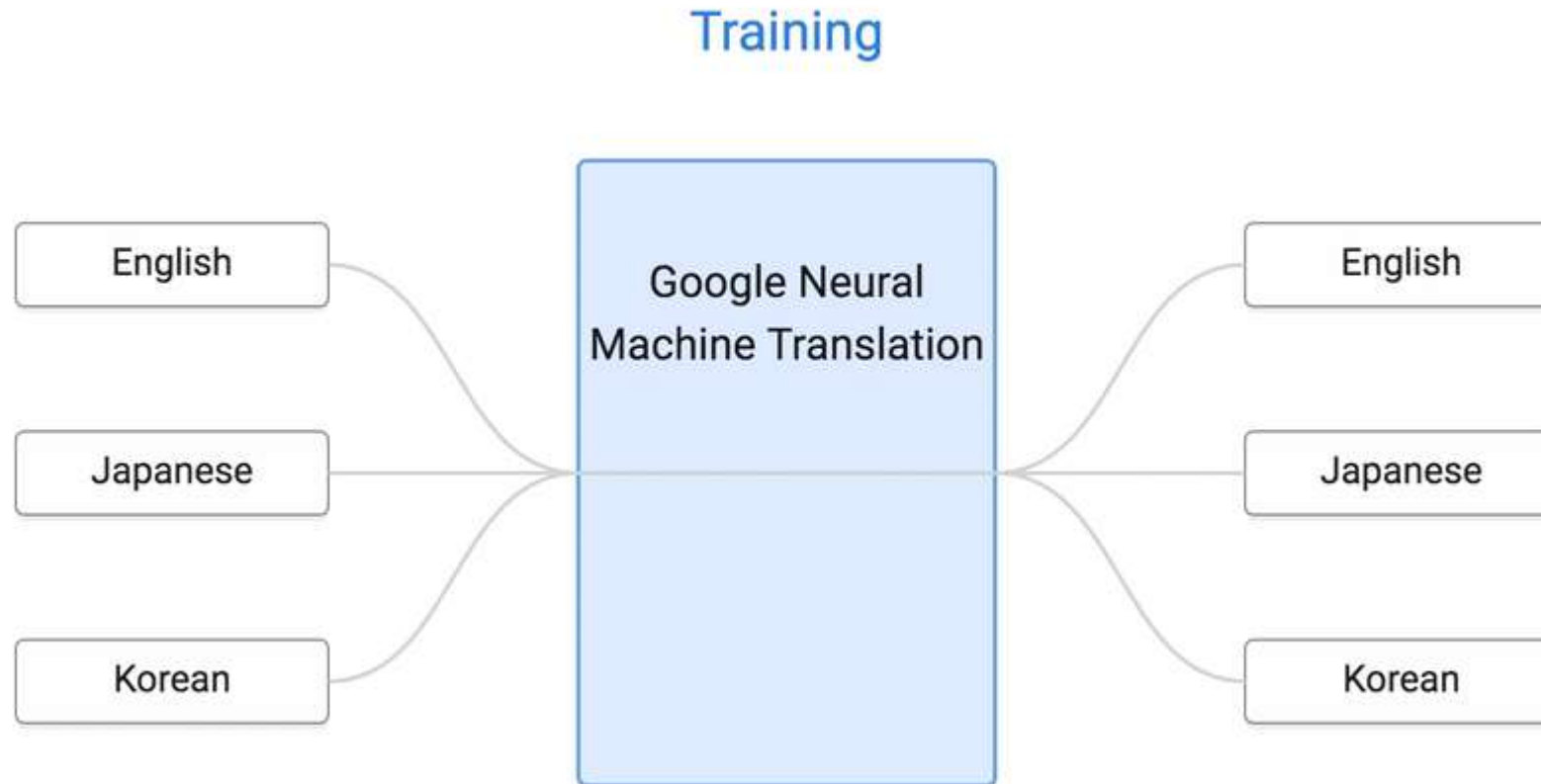
[Koehn & Knowles 2017]

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?`

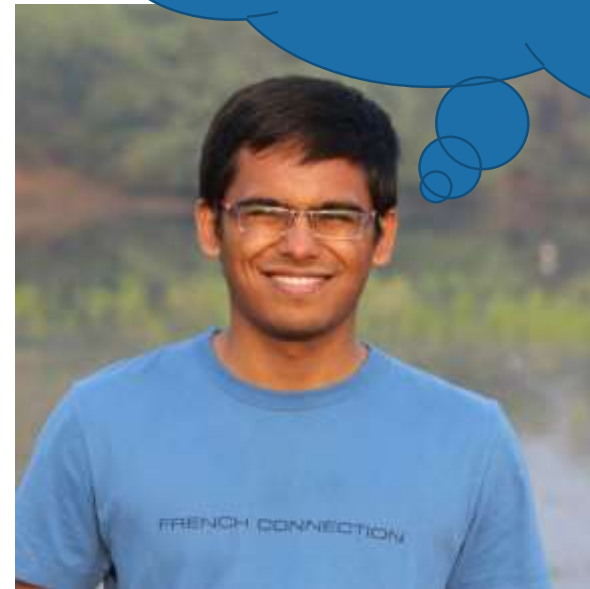
- Helps most for low-resources languages
- Enables zero-shot translation
- Can handle code-switched input

How can we make the most of parallel data?

Examples are not equally parallel

- Parallel segments are inevitably noisy
- Cross-lingual divergences abound
- “Traduttore, traditore”

Can we improve neural MT if we know which examples diverge in meaning?



Yogarshi Vyas

Divergent Translation Examples

someone wanted to cook bratwurst.

vous vouliez des saucisses grillées.

i don't know what i'm gonna do.

j'en sais rien.

- has the sake chilled? - no, it's fine.

- c'est assez chaud?

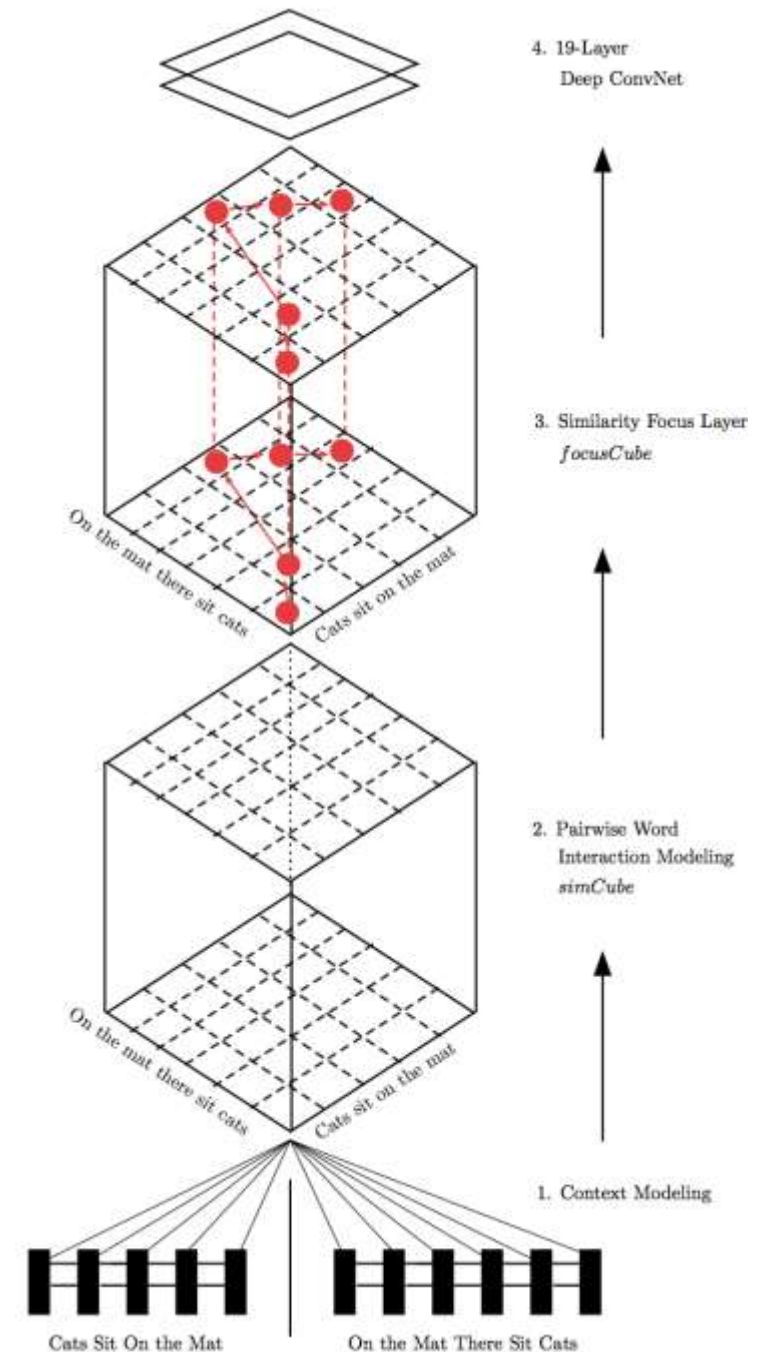
you help me with zander and i helped you with joe.

tu m'as aidé avec zander, je t'ai aidé avec joe.

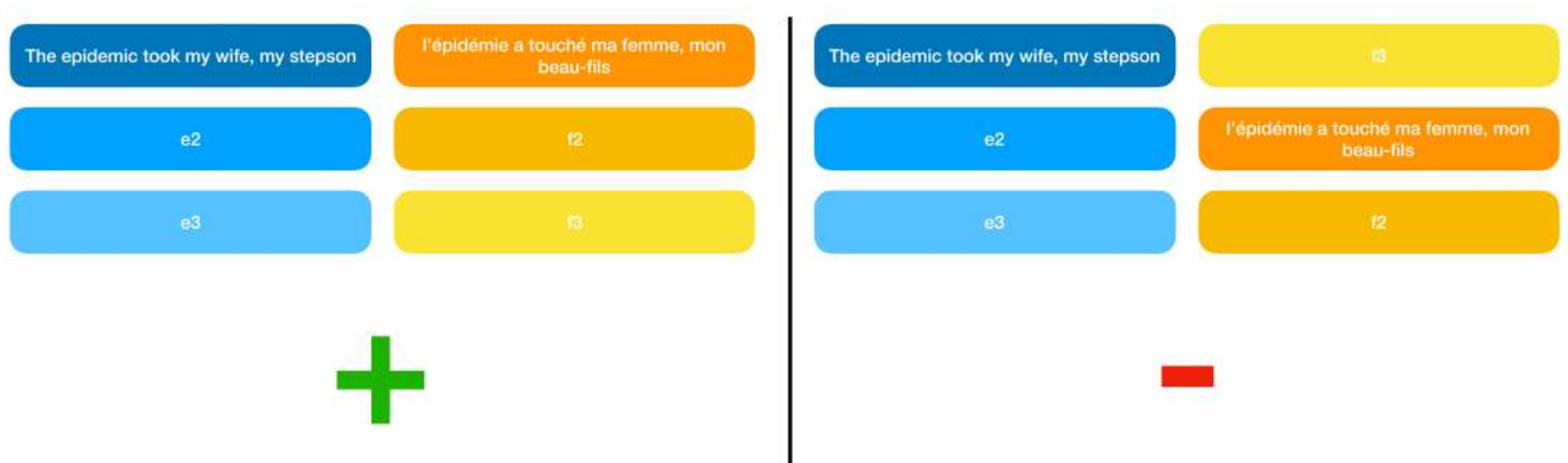
My fellow Americans, you chose to give me your trust, and I would like to express my deep gratitude.

Automatically detecting translation divergences with semantic similarity models

- Compare multiple sentence spans with deep convnet [He & Lin 2016]
- Initialize with bilingual word embeddings to enable cross-lingual comparisons

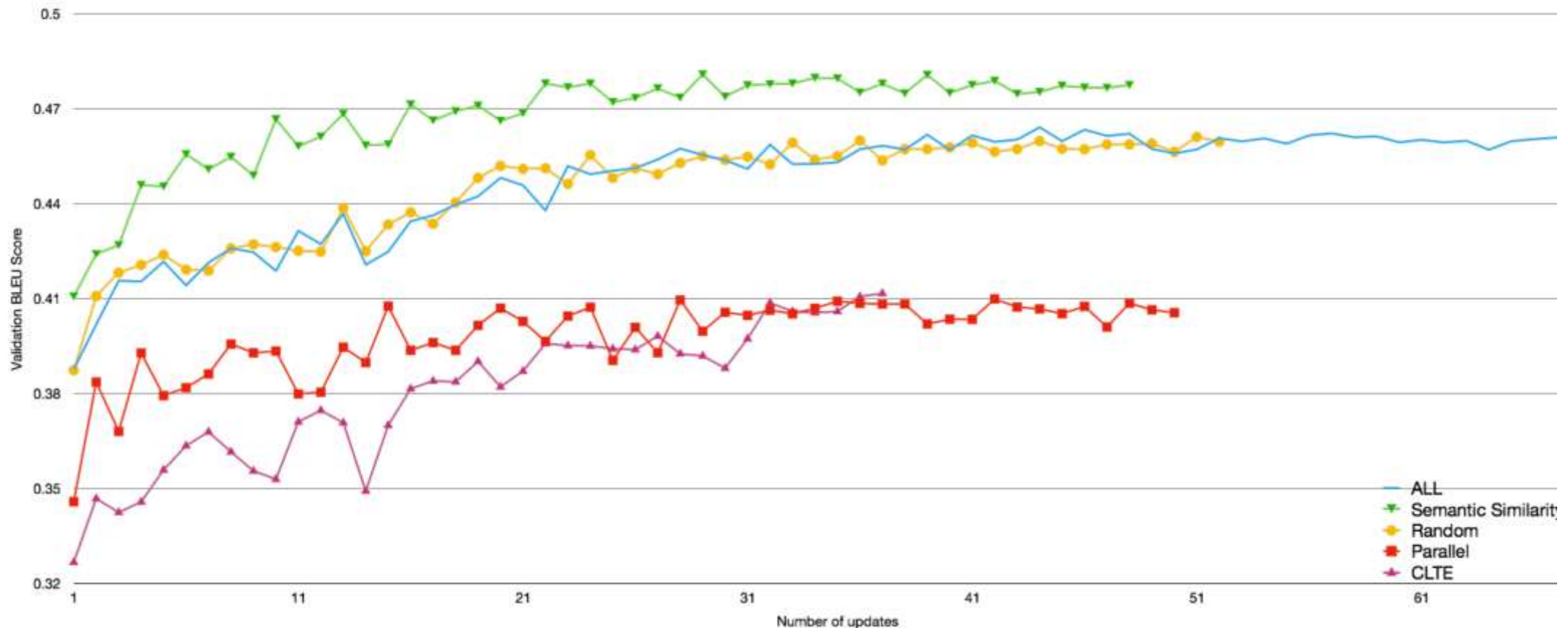


Creating synthetic supervision



Downsampling via cross-lingual semantic similarity helps NMT training

[Vyas, Niu & Carpuat, NAACL 2018]



In high resources settings, how
can we improve further?

Translate documents, not sentences!

In fairness, Miller did not attack **the statue** itself.

[...]

But he did attack **its meaning** [...]

HUMAN

Um fair zu bleiben, Miller griff nicht **die Statue** selbst an.

[...]

Aber er griff **deren Bedeutung** an [...]

MT

Fairerweise hat Miller **die Statue** nicht selbst angegriffen.

[...]

Aber er griff **seine Bedeutung** an [...]

Translate documents, not sentences!

Weidezaunprojekt ist elementar

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

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

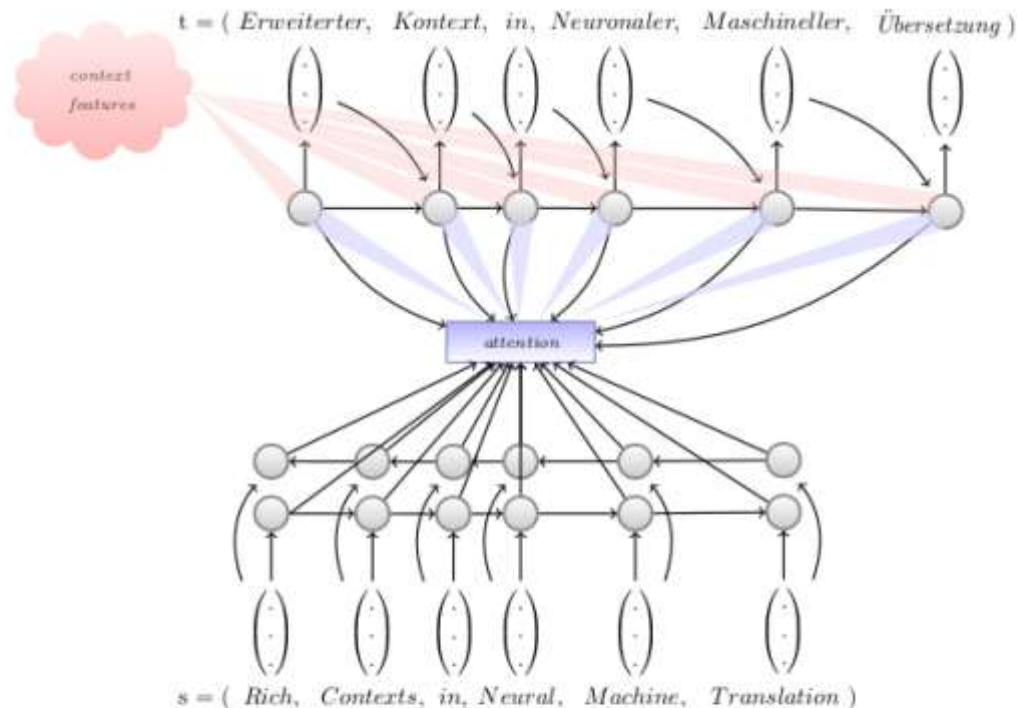
Translate documents, not sentences!

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

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]

Style Differences Matter for Translation

TO IMPROVE ACCURACY, FILL OUT THE OPTIONAL FIELDS BELOW

Is it more "Hey Dude" or "Dear Sir"?
Improve translation accuracy by telling us the tone of the content.

Informal

Translator

Informal
Friendly
Business
Formal
Other

Possible instructions
Voice
Links
Purpose & Audience

Casual, romantic, funny, serious etc.
To your website, screen shots or other docs.
This is going to my most important client etc.

Business from \$0.12 / word

Order total **\$520.80**

Estimated delivery **15 hours.**

I agree to the [Terms & Conditions](#) and [Quality Policy](#)
Updated on 03/16/2017

Payment method: Credit card PayPal

Pay & Confirm Order

View Full Quote

Can we control the
formality of machine
translation output?



Marianna
Martindale



Xing Niu

Ranking segments by formality in MT corpora

Formal delegates are kindly requested to bring their copies of documents to meetings . [UN]

in these centers , the children were fed , medically treated and rehabilitated on both a physical and mental level . [OpenSubs]

there can be no turning back the clock [UN]

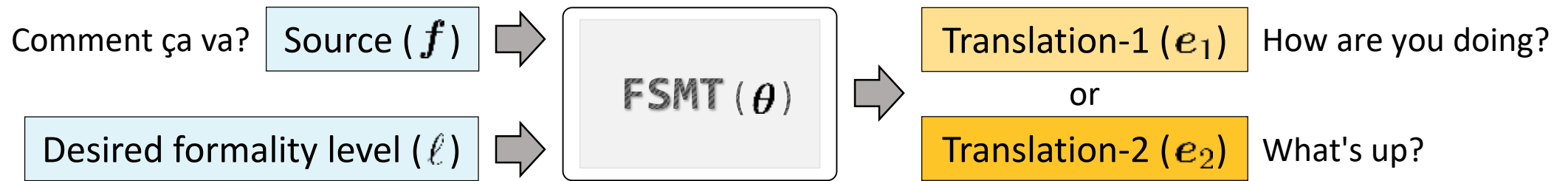
I just wanted to introduce myself [OpenSubs]

-yeah , bro , up top . [OpenSubs]

Informal

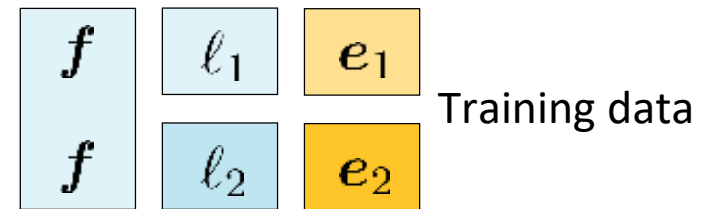


Formality-Sensitive Machine Translation (FSMT)

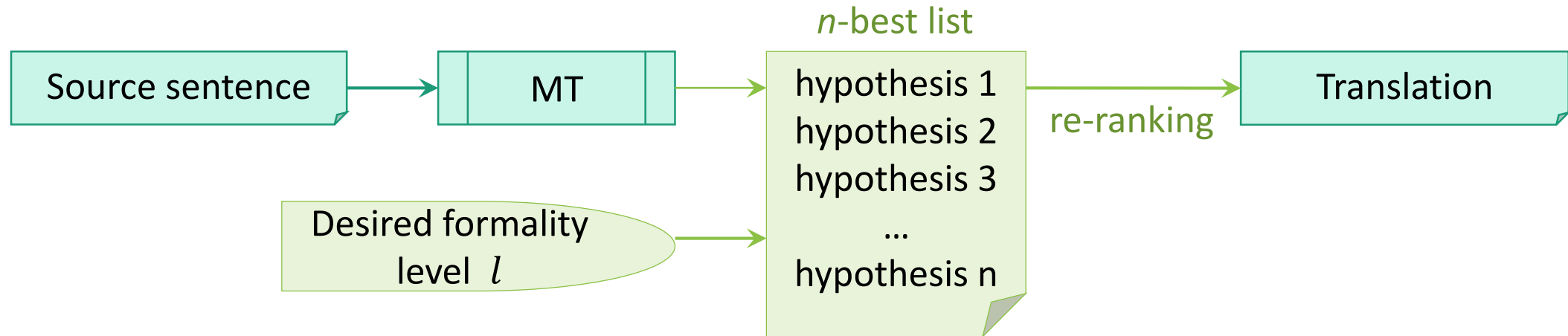


$$\hat{e} = \arg \max_e P(e|f, l; \theta)$$

$$\hat{\theta} = \arg \max_{\theta} \sum_{(f, l, e)} P(e|f, l; \theta)$$



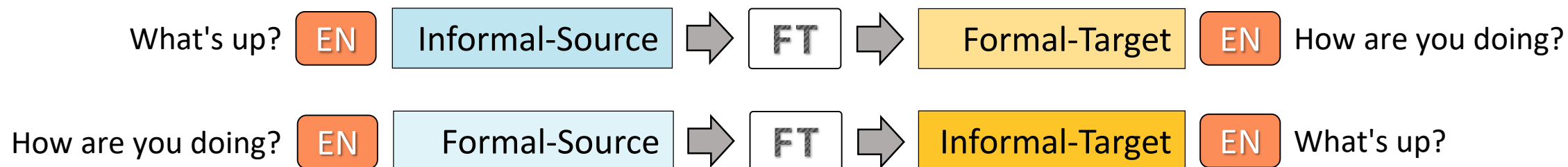
Formality-Sensitive MT: A re-ranking approach



- n -best list re-ranking with a new feature
 - $\Delta f(h, l) = | \text{Formality}(h) - l |$
 - Where $\text{Formality}(h)$ places sentences on $[-1,1]$ formality scale

[Niu, Martindale & Carpuat, EMNLP 2017]

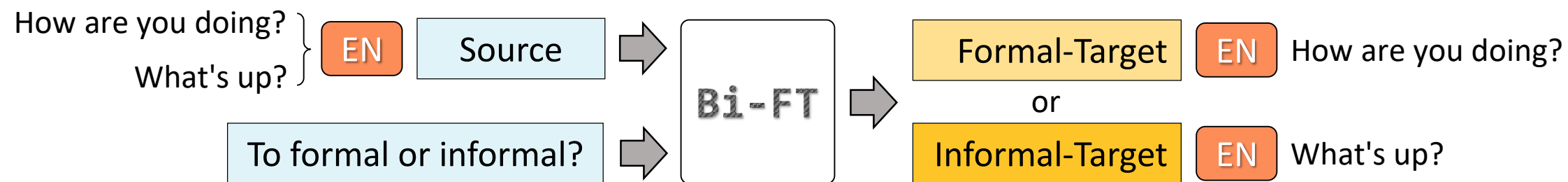
Formality Transfer (FT)



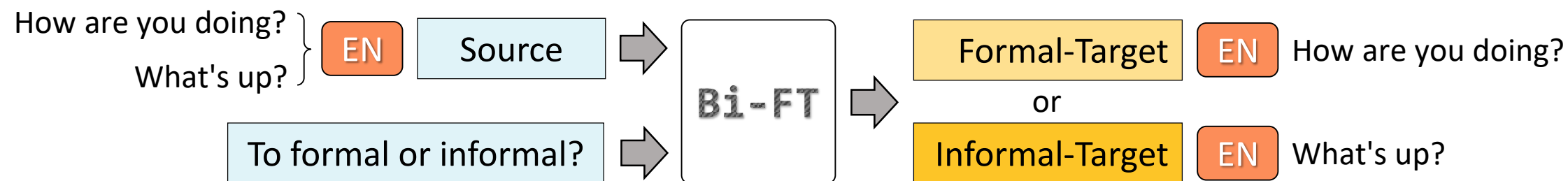
- Can be addressed with NMT models
- Given large parallel formal-informal corpus

[Rao and Tetreault, 2018]

Bi-Directional Formality Transfer (Bi-FT)



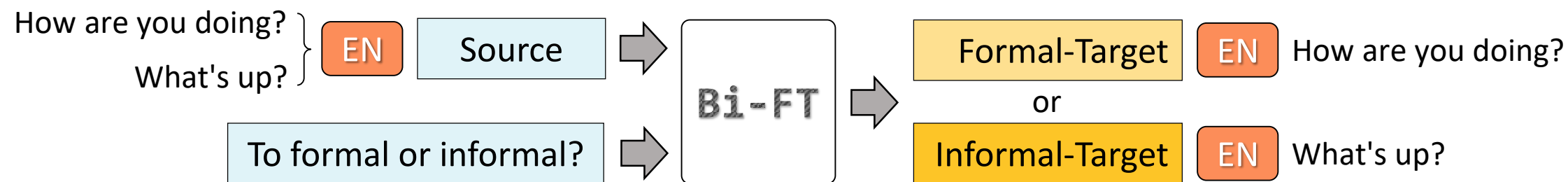
Bi-Directional Formality Transfer (Bi-FT)



- Both transfer directions share the same NMT model.
 - Implemented like multilingual NMT (Johnson et al., 2017)
 - Training data:

Informal-EN	Formal-EN
Formal-EN	Informal-EN

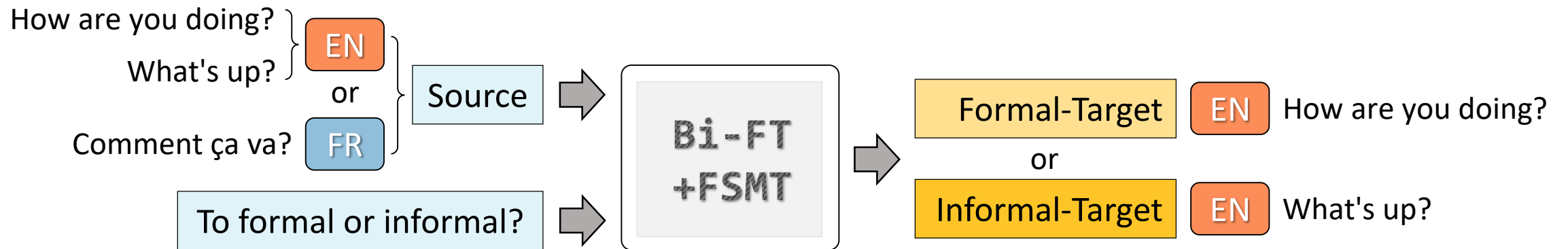
Bi-Directional Formality Transfer (Bi-FT)



- Both transfer directions share the same NMT model.
 - Implemented like multilingual NMT (Johnson et al., 2017)
 - Training data:

<F>	Informal-EN	Formal-EN
<I>	Formal-EN	Informal-EN

Formality Sensitive MT as Multitask Formality Transfer + MT





Results – FSMT (Human Evaluation)

Model (Informal vs. Formal)	Formality Diff Range = [-2,2]	Meaning Preservation Range = [0,3]
NMT-constraint	0.35	2.95
NMT MultiTask-random	0.32	2.90
PBMT-random	0.05	2.97

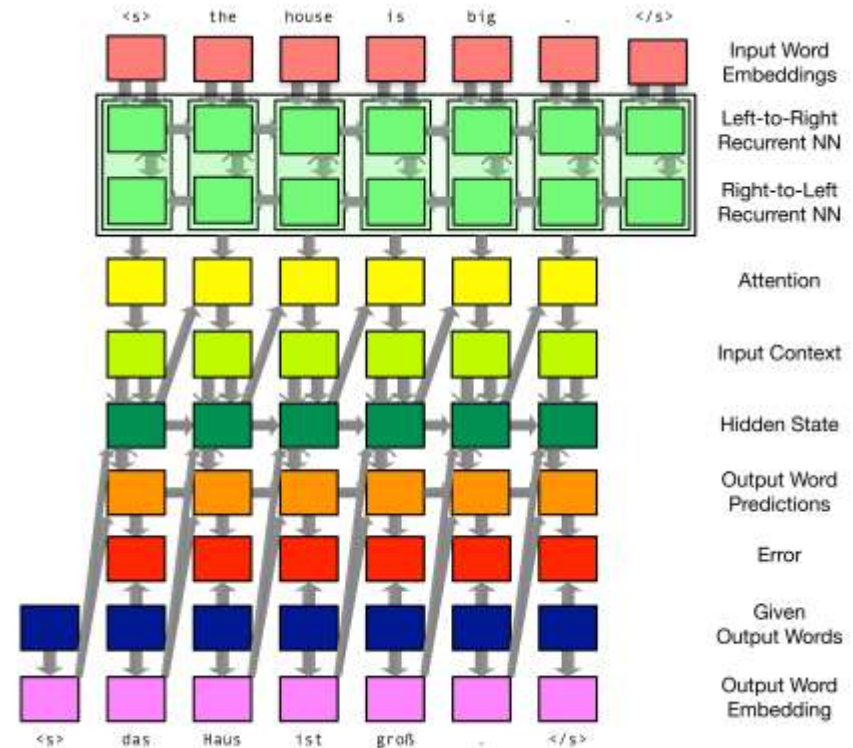
Human judgments on 300 translation pairs per system

Qualitative Analysis

# 1	Reference	Refrain from the commentary and respond to the question, Chief Toohey.
	MultiTask-random	You need to be quiet and answer the question, Chief Toohey.
Formal	NMT-constraint	Please refrain from any comment and answer the question, Chief Toohey.
	PBMT-random	Please refrain from comment and just answer the question, the Tooheys's boss.
	MultiTask-random	Shut up and answer the question, Chief Toohey.
Informal	NMT-constraint	Please refrain from comment and answer the question, chief Toohey.
	PBMT-random	Please refrain from comment and answer my question, Tooheys's boss.
# 2	Reference	Try to file any additional motions as soon as you can.
	 MultiTask-random	You should try to introduce the sharks as soon as you can.
Formal	NMT-constraint	Try to present additional requests as soon as you can.
	PBMT-random	Try to introduce any additional requests as soon as you can.
	 MultiTask-random	Try to introduce sharks as soon as you can.
Informal	NMT-constraint	Try to introduce extra requests as soon as you can.
	PBMT-random	Try to introduce any additional requests as soon as you can.

Summing up: Neural Machine Translation...

- Dramatically improved translation quality
- Implicitly captures useful generalizations about language
- But...
 - Models are opaque and use little linguistic insights
 - Performance drops in low-resource settings
 - We should do better/faster in high-resource settings



Beyond Translation: Neural Encoder-Decoders can be used for many other tasks

<u>Input X</u>	<u>Output Y (Text)</u>	<u>Task</u>
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

Resources

- [Neural Machine Translation and Sequence-to-Sequence Models: a Tutorial](#), by Graham Neubig
- [OpenNMT](#), an open-source neural machine translation system
- [Machine Translation Marathon](#)
- [Neural Network Methods for Natural Language Processing](#), by Yoav Goldberg

Introduction to Neural Machine Translation (3/3)

Marine Carpuat

Computer Science

University of Maryland

