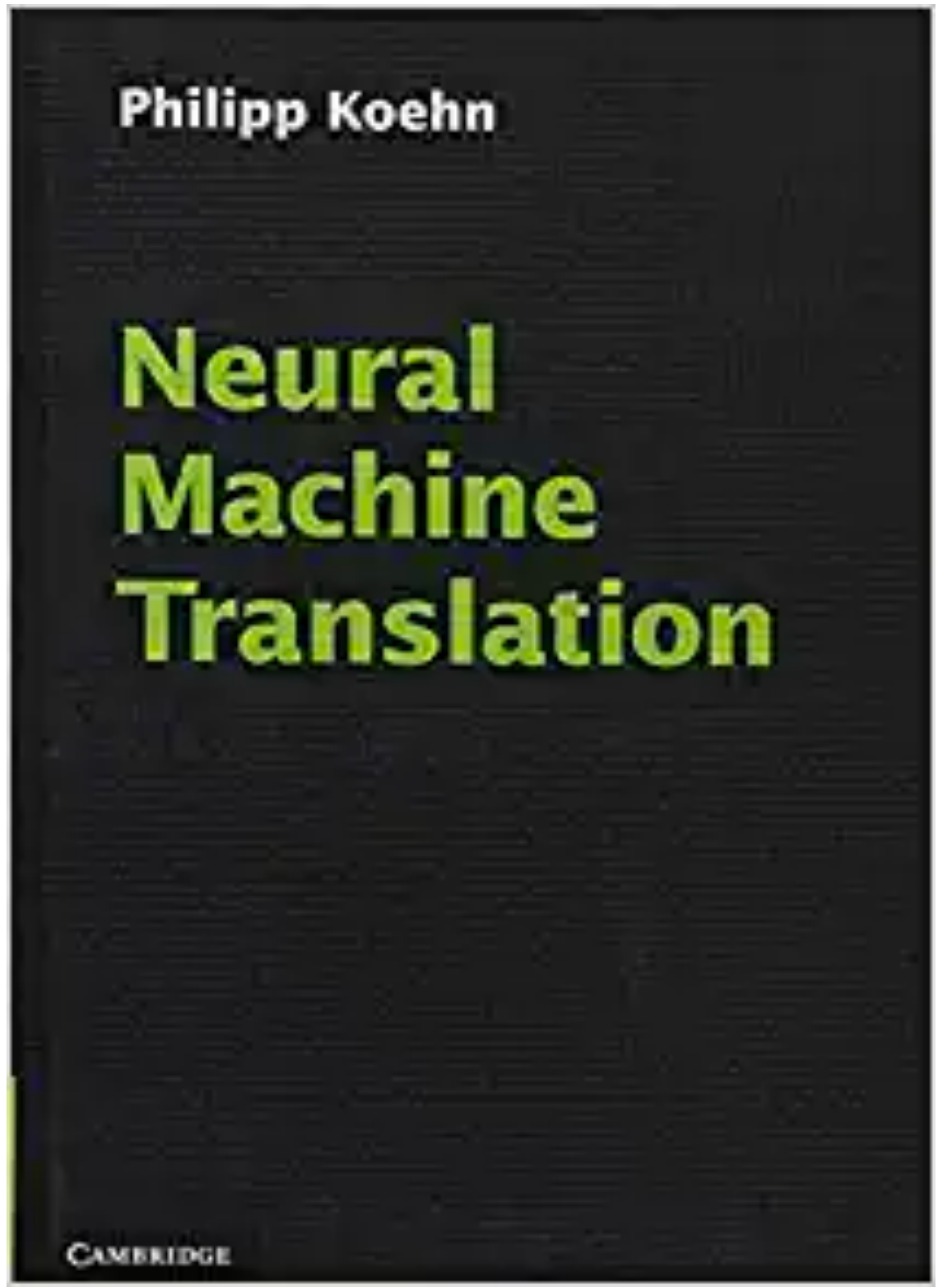
Encoder-Decoder (aka Seq2Seq)

Wei Xu

(many slides from Greg Durrett)

This Lecture

- Machine Translation
- Sequence-to-Sequence Model



- Reading — Eisenstein 18.3-18.5

MT Basics

MT Basics

People’s Daily, August 30, 2017

Trump Pope family watch a hundred years a year in the White House balcony

MT Ideally

- I have a friend => $\exists x$ `friend(x,self)` => J'ai un ami

  J'ai une amie

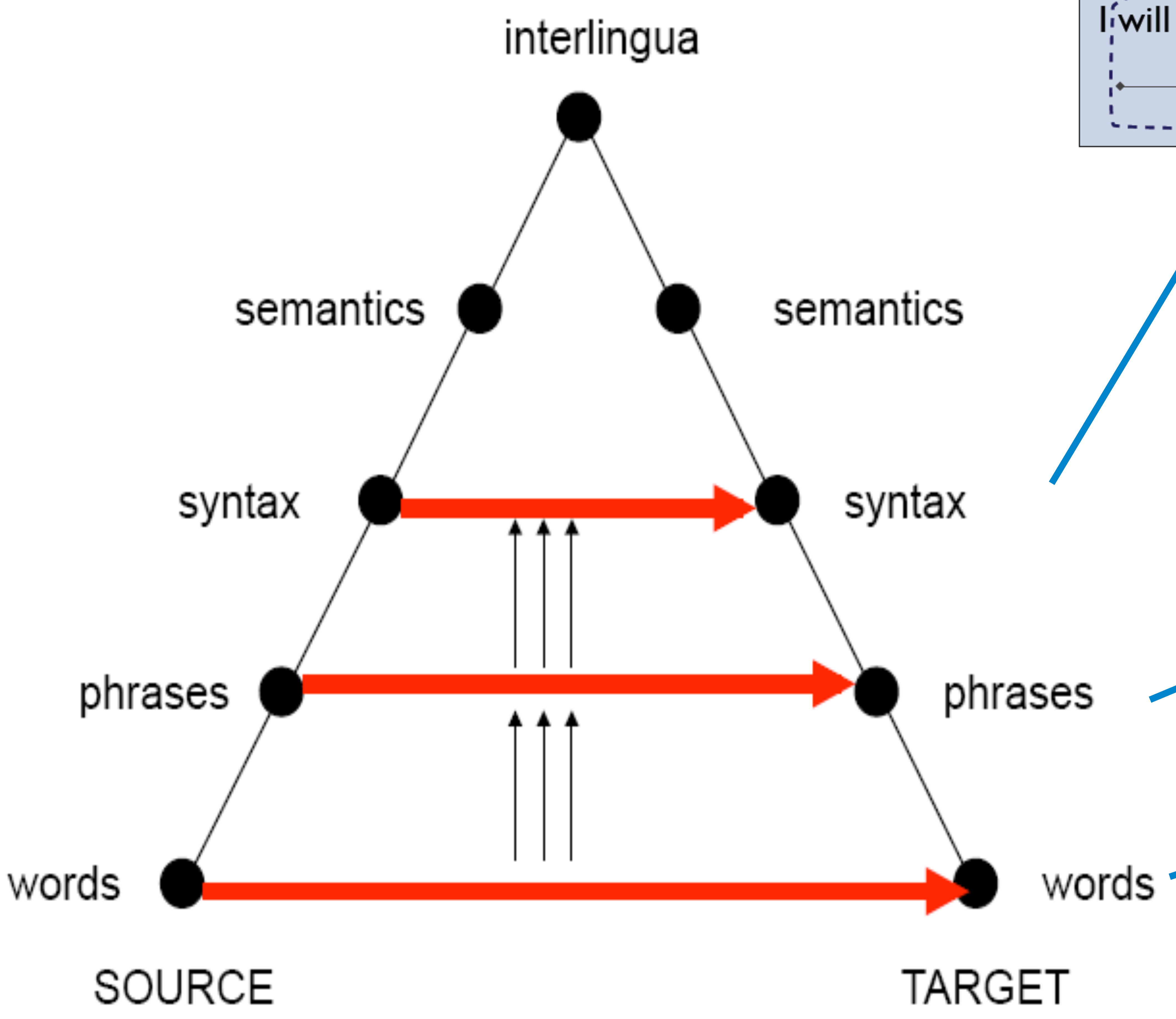
  - May need information you didn't think about in your representation
  - Hard for semantic representations to cover everything

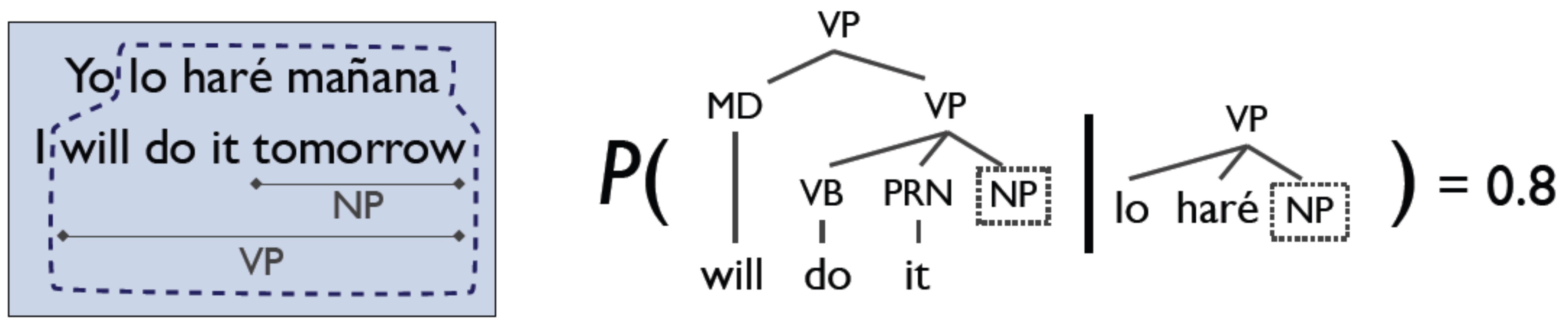
- Everyone has a friend => $\exists x \forall y$ `friend(x,y)` => Tout le monde a un ami

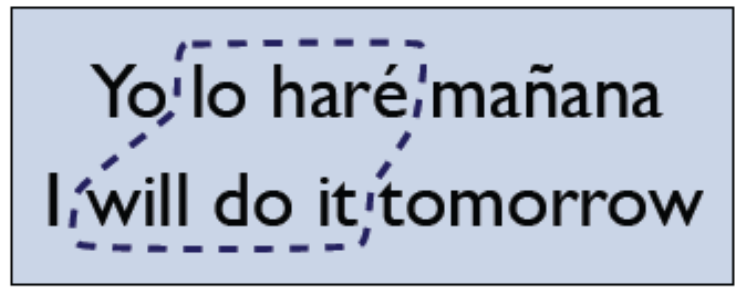
  $\forall x \exists y$ `friend(x,y)`

  - Can often get away without doing all disambiguation — same ambiguities may exist in both languages

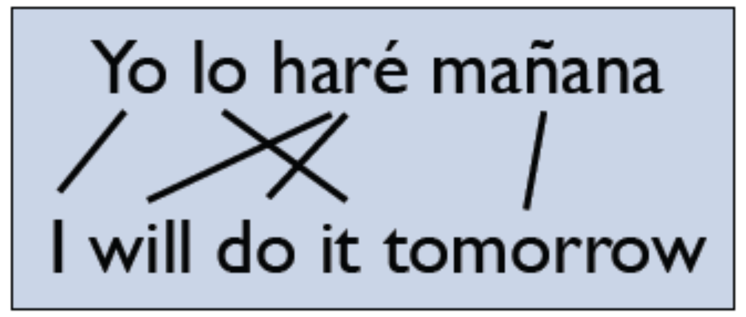
Levels of Transfer: Vauquois Triangle (1968)







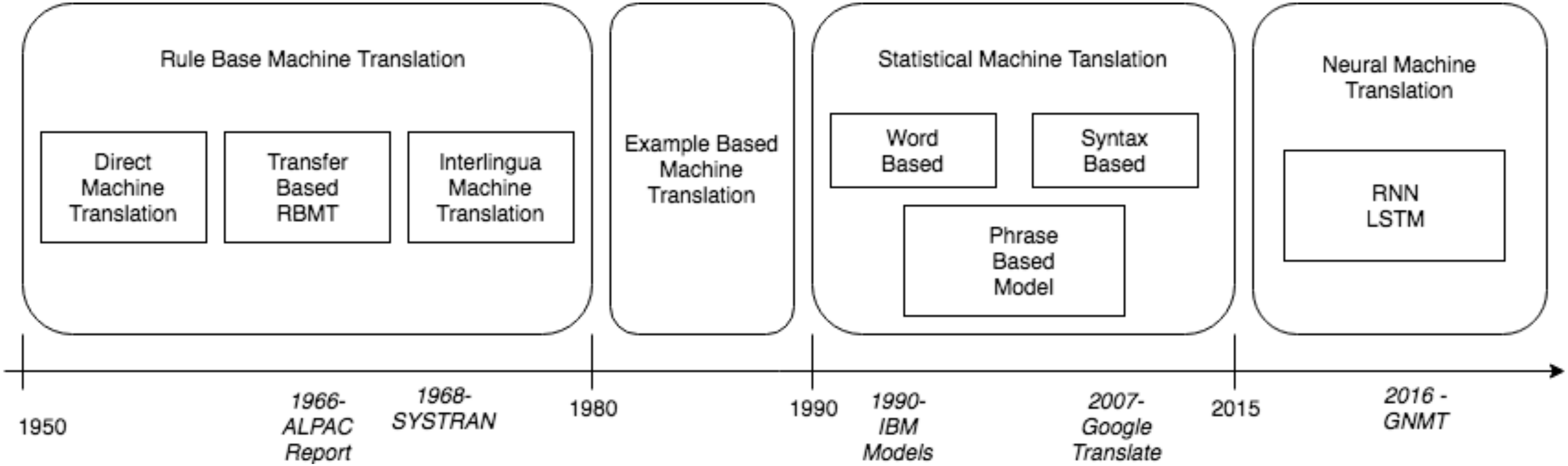
English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2



English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

Slide credit: Dan Klein

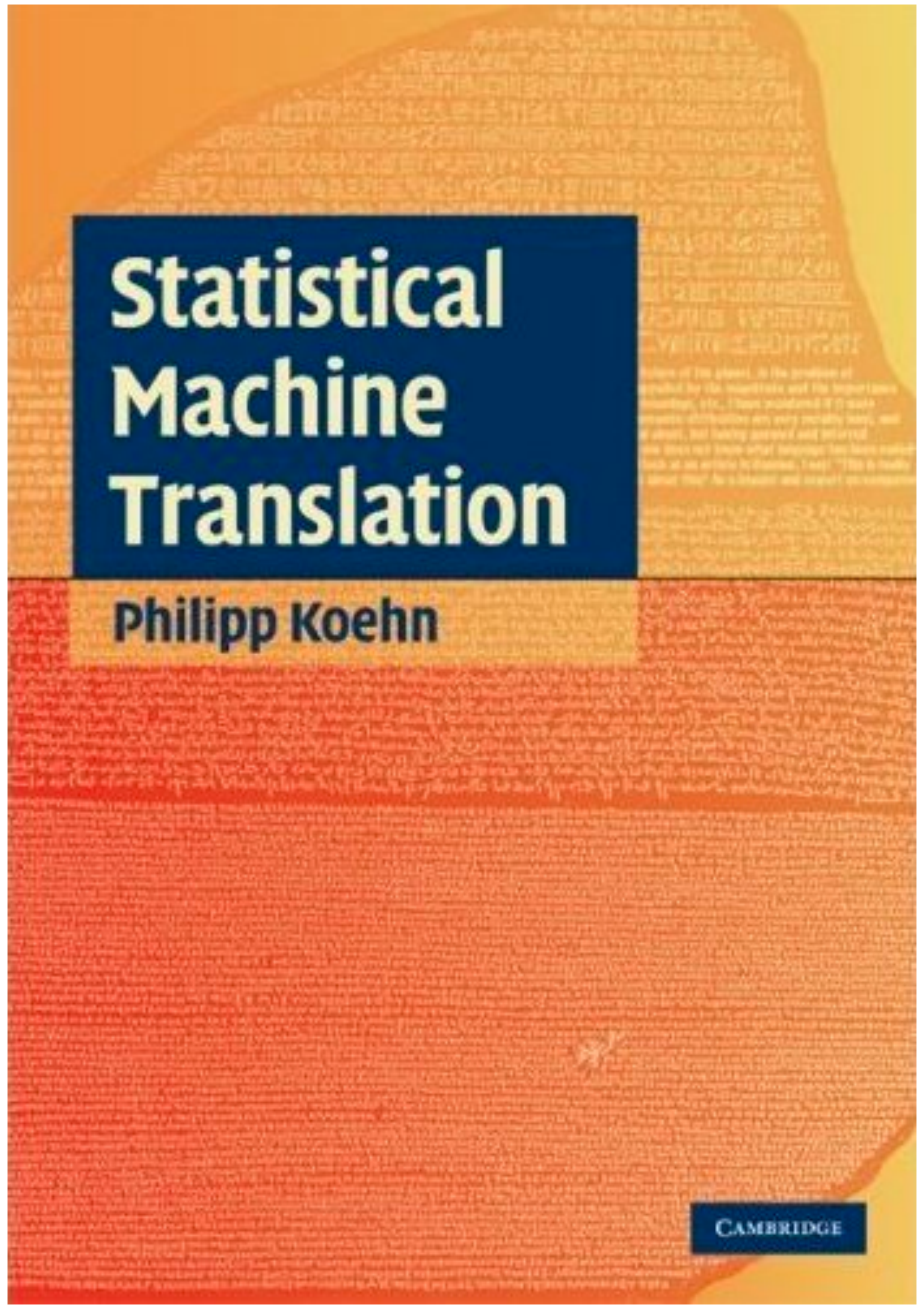
History of MT



Parallel Training Corpus

		facing with the swelling flow of through traffic zooming past their doors .		retahíla de inconvenientes que más y más gente tiene que soportar por el tráfico que pasa por delante_de sus casas , que aumenta a_diario .
5	#77501757	Weekend traffic bans and traffic **jams** are a curse to road transport .	#74765580	Las prohibiciones de conducir los fines de semana y los **embotellamientos** asolan el transporte por carretera .
6	#79500725	Some people also want to recoup the cost of traffic **jams** from those who get stuck in them , according to the ' polluter pays ' principle .	#76764676	Algunos son partidarios de que incluso los costes ocasionados por los **atascos** se carguen a el ciudadano que se encuentra atrapado en ellos , de conformidad con el principio de que " quien contamina paga " .
7	#79500765	I think this is an excellent principle and I would like to see it applied in full , but not to traffic **jams** .	#76764713	Me parece un principio acertado y estoy dispuesta a aplicarlo íntegramente , pero no sobre los **atascos** , ya_que éstos son un claro indicio de el fracaso de la política gubernamental en_materia_de infraestructuras .
8	#79500768	Traffic **jams** are indicative of failed government policy on the infrastructure front , which is why the government itself , certainly in the Netherlands , must be regarded as the polluter .	#76764747	Por eso es preciso subrayar que en estos casos quien contamina es el propio Gobierno , a el menos en los Países_Bajos .
9	#81309716	This would increase traffic **jams** , weaken road safety and increase costs .	#78586130	Esto aumentaría los **atascos** , mermaría la seguridad vial e incrementaría los costes .
10	#81997391	In the previous legislature , Parliament gave its opinion on the Commission ' s proposals on the simplification of vertical directives on sugar , honey , fruit juices , milk and **jams** .	#79281114	En efecto , durante la precedente legislatura , el Parlamento se manifestó sobre las propuestas de la Comisión relativas a la simplificación de directivas verticales sobre el azúcar , la miel , los **zumos** de frutas , la leche y las **confituras** .
11	#81998167	For **jams** , I personally reintroduced an amendment that was not accepted by the Committee on the Environment , Public Health and Consumer Policy , but which I hold to .	#79281936	Para las **confituras** , yo personalmente volví a introducir una enmienda que no fue aceptada por la Comisión_de_Medio_Ambiente , Salud_Pública y Política_de_el_Consumidor , pero que es importante para mí .
12	#81998209	It concerns not accepting the general use of a chemical flavouring in **jams** and marmalades , that is vanillin .	#79281966	Se trata de no aceptar la utilización generalizada de un aroma químico en las **confituras** y " marmalades " , a saber , la vainillina .
13	#82800065	This is highlighted particularly in towns where it is necessary to find ways of solving environmental problems and the difficulties caused by traffic **jams** .	#80085988	Esto se pone_de_relieve aún más en las ciudades , en las que hay que encontrar medios para eliminar los inconvenientes derivados de los problemas medioambientales y de la congestión de el tráfico .

Phrase-based MT (very briefly)



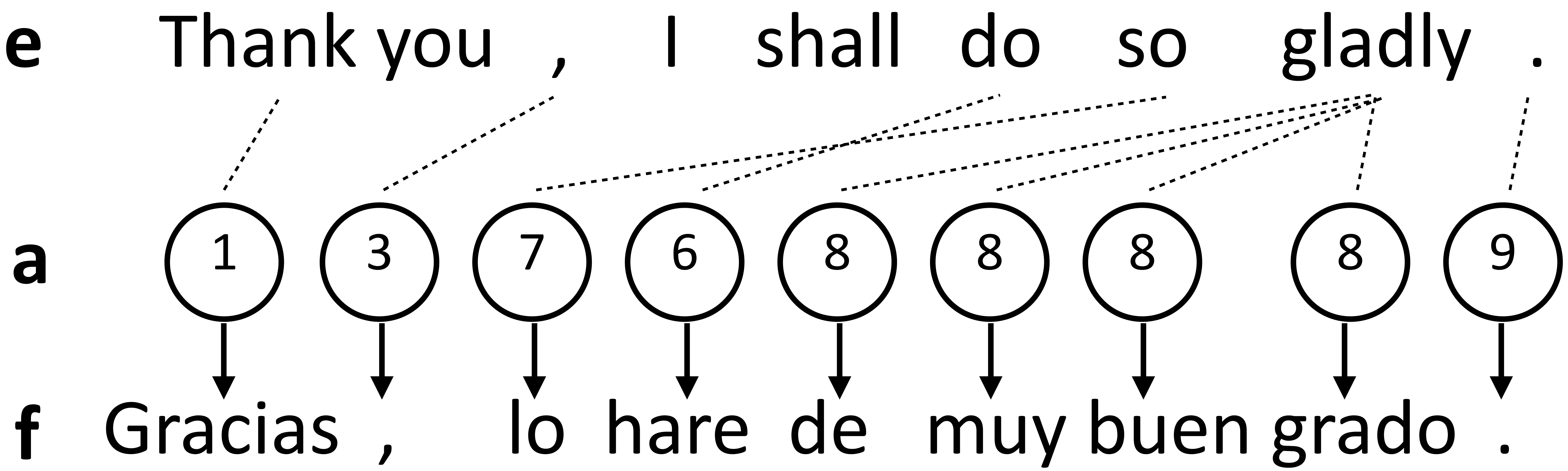
Phrase-Based MT

- Key idea: translation words better the bigger chunks you use
- Remember phrases from training data, translate piece-by-piece and stitch those pieces together to translate
  - How to identify phrases? Word alignment over source-target bitext
  - How to stitch together? Language model over target language
  - Decoder takes phrases and a language model and searches over possible translations
- NOT like standard discriminative models (take a bunch of translation pairs, learn a ton of parameters in an end-to-end way)

Word Alignment: IBM Model 1

- Each "Foreign" word is aligned to *at most* one English word

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{i=1}^{n} P(f_i|e_{a_i})P(a_i)$$



- Set P(a) uniformly (no prior over good alignments) = 1 / (#words in e + 1)
- $P(f_i|e_{a_i})$: word translation probability. Learn with EM (Eisenstein ch 18.2.2)

Brown et al. (1993)

Word Alignment

- Find contiguous sets of aligned words in the two languages that don't have alignments to other words

  de assister à la runion et ||| to attend the meeting and

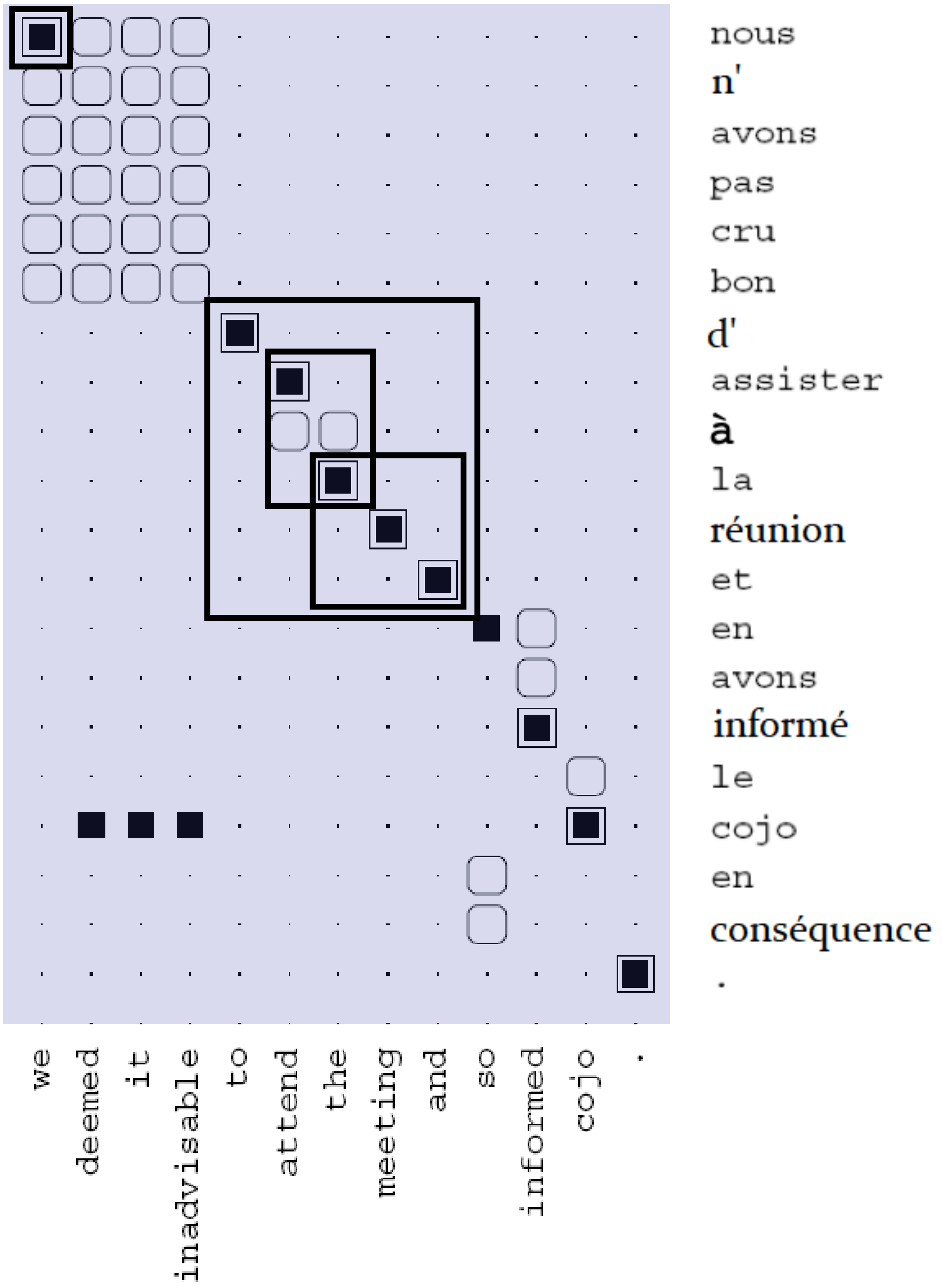
  assister à la runion ||| attend the meeting

  la runion and ||| the meeting and

  nous ||| we

  …

- Lots of phrases possible, count across all sentences and score by frequency

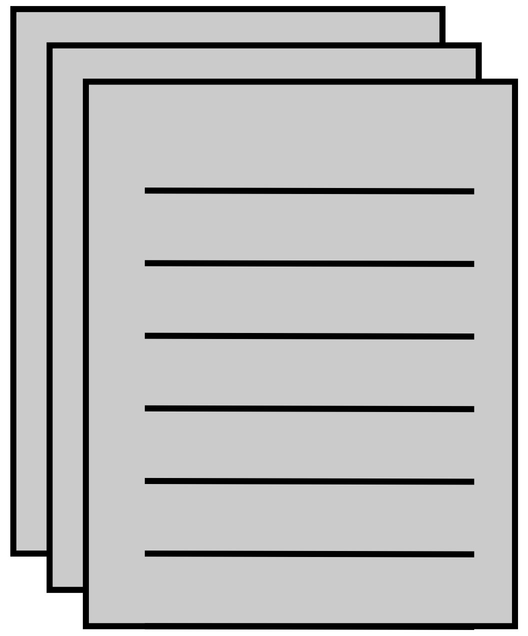


Phrase-Based MT

- Goal: translate from Foreign language to English

```
cat ||| chat ||| 0.9
the cat ||| le chat ||| 0.8
dog ||| chien ||| 0.8
house ||| maison ||| 0.6
my house ||| ma maison ||| 0.9
language ||| langue ||| 0.9
…
```

Phrase table $P(f|e)$



Unlabeled English data

Language model $P(e)$

$$P(e|f) \propto P(f|e)P(e)$$

Noisy channel model: combine scores from translation model + language model to translate foreign to English

"Translate faithfully but make fluent English"

MT Evaluation

Evaluating MT

- Fluency: does it sound good in the target language?
- Fidelity/adequacy: does it capture the meaning of the original?
- BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

		1-gram	2-gram	3-gram
hypothesis 1	I am exhausted	3/3	1/2	0/1
hypothesis 2	Tired is I	1/3	0/2	0/1
hypothesis 3	I I I	1/3	0/2	0/1
reference 1	I am tired			
reference 2	I am ready to sleep now and so exhausted			

Papineni et al. (2002)

Evaluating MT

- Fluency: does it sound good in the target language?
- Fidelity/adequacy: does it capture the meaning of the original?
- BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- Typically $N = 4$, $w_i = 1/4$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- r = length of reference
  c = length of system output

- Does this capture fluency and adequacy?

Papineni et al. (2002)

https://github.com/mjpost/sacrebleu

Appraise - Human Evaluation Interface

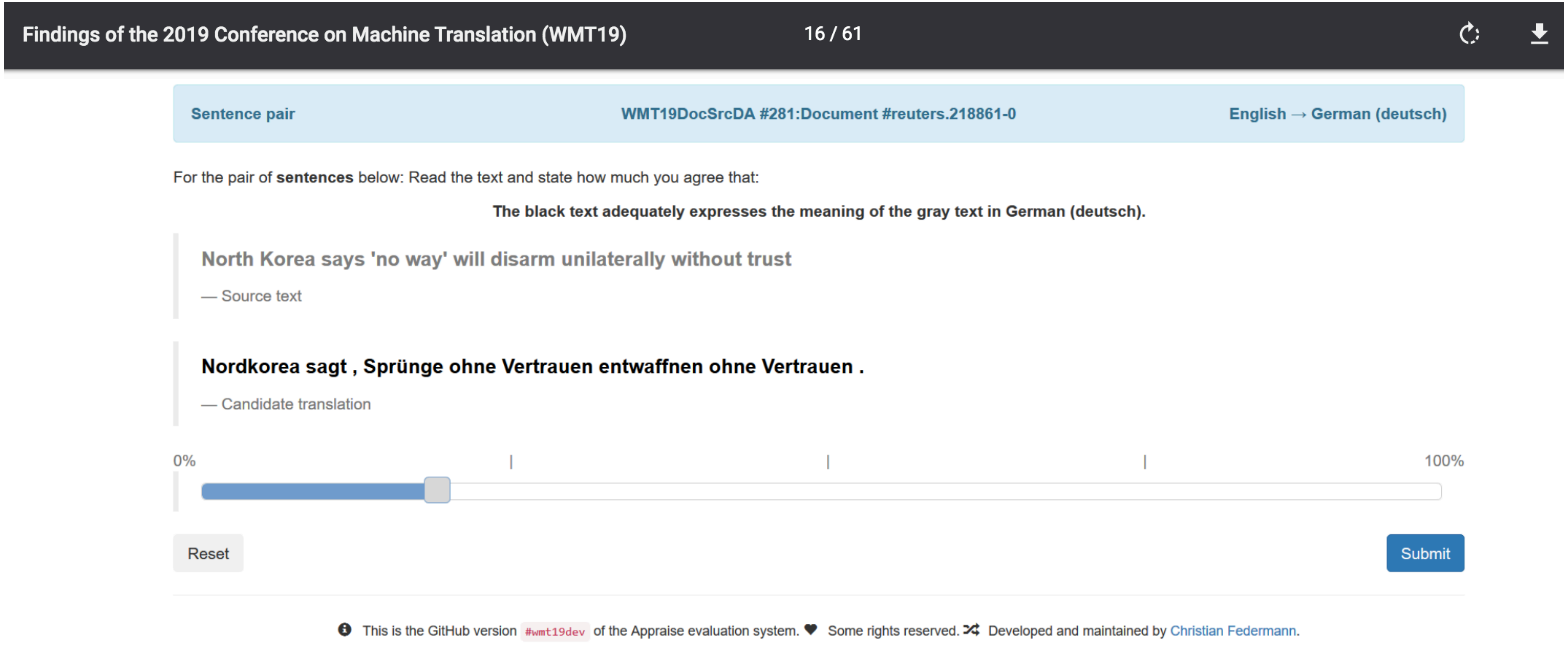
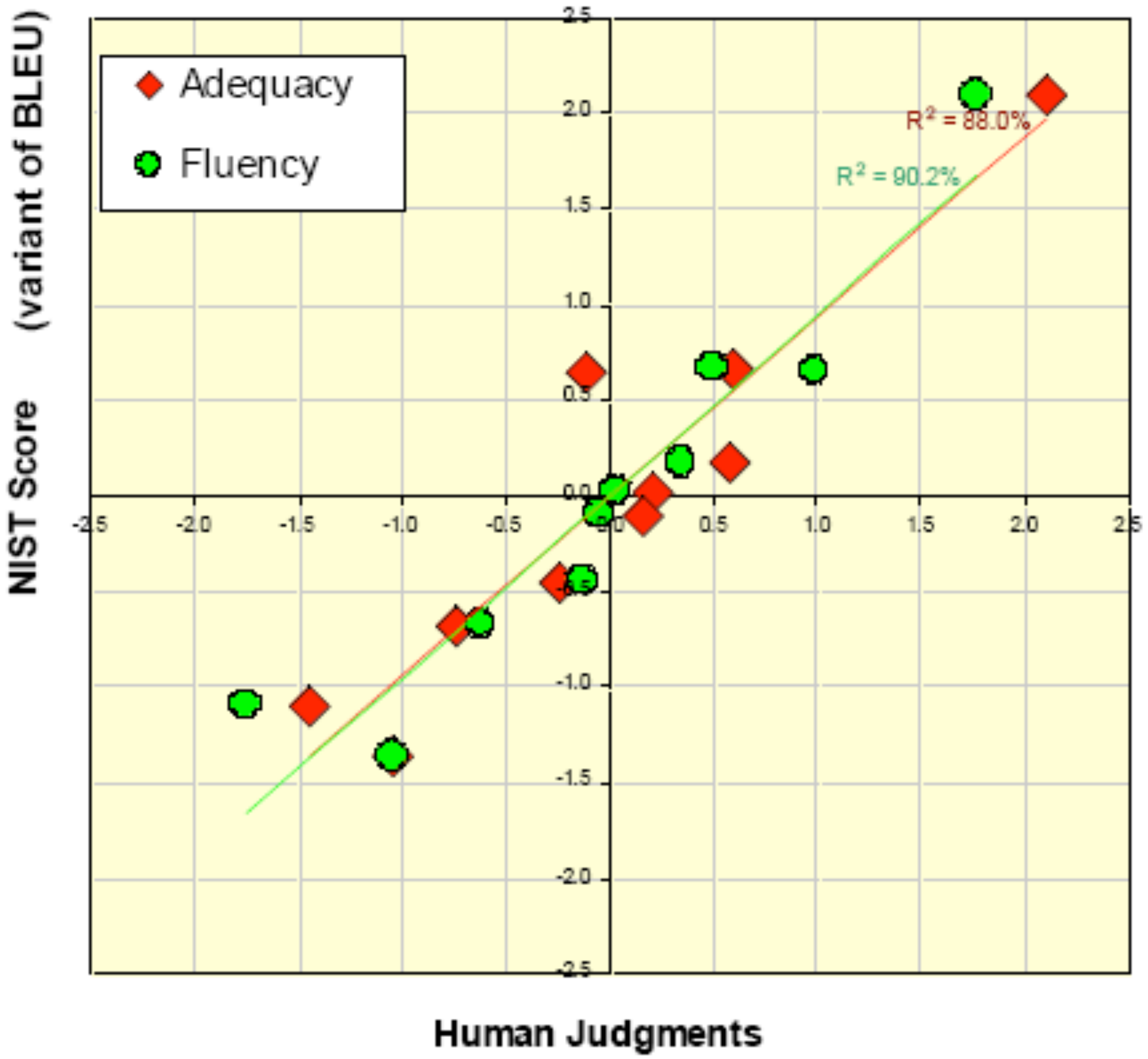


Figure 3: Screen shot of segment-rating portion of document-level direct assessment in the Appraise interface for an example English to German assessment from the human evaluation campaign. The annotator is presented with the machine translation output segment randomly selected from competing systems (anonymized) and is asked to rate the translation on a sliding scale.

Federmann (2010)

BLEU Score

- Better methods with human-in-the-loop
- HTER: human-assisted translation error rate
- If you're building real MT systems, you do user studies. In academia, you mostly use BLEU, COMET, etc.

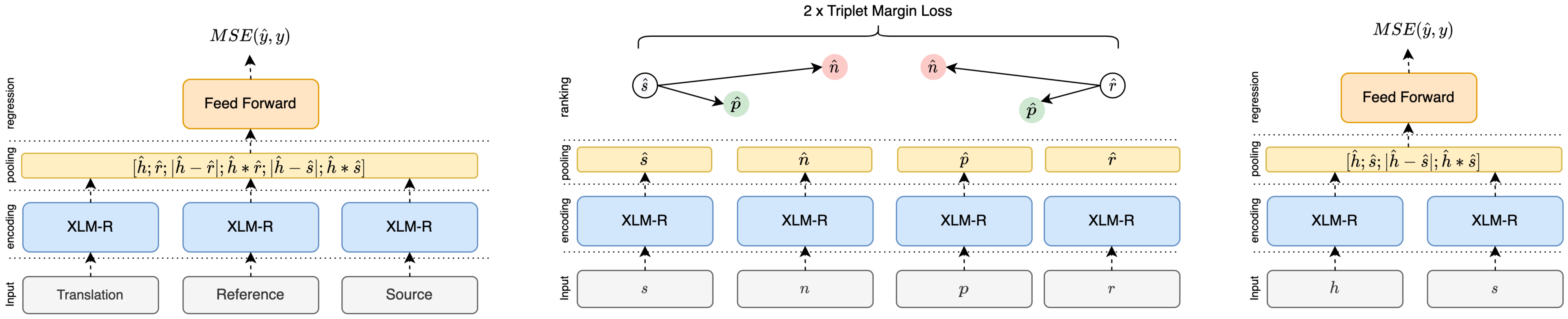


slide from G. Doddington (NIST)

Other MT Evaluation Metrics

- BLEU (2002): n-gram overlap
- METEOR (2005): also take into consideration of synonyms
- HTER (2009): human-assisted translation error rate
- BERTScore (2019): embedding-based
- BLEURT (2020) and COMET (2020): trained neural network model using human evaluation data
- and many more …

COMET - Learnt Metric

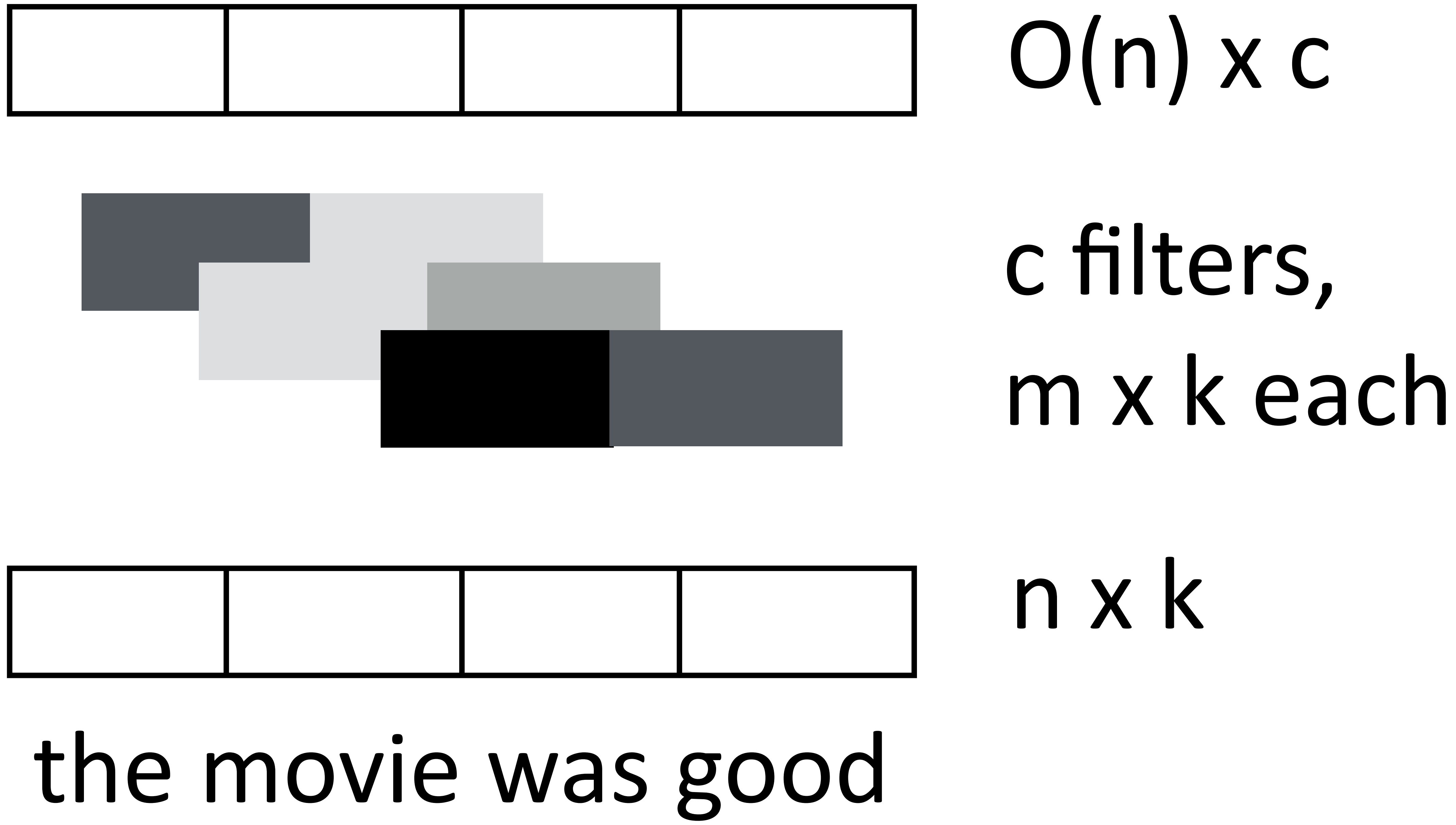


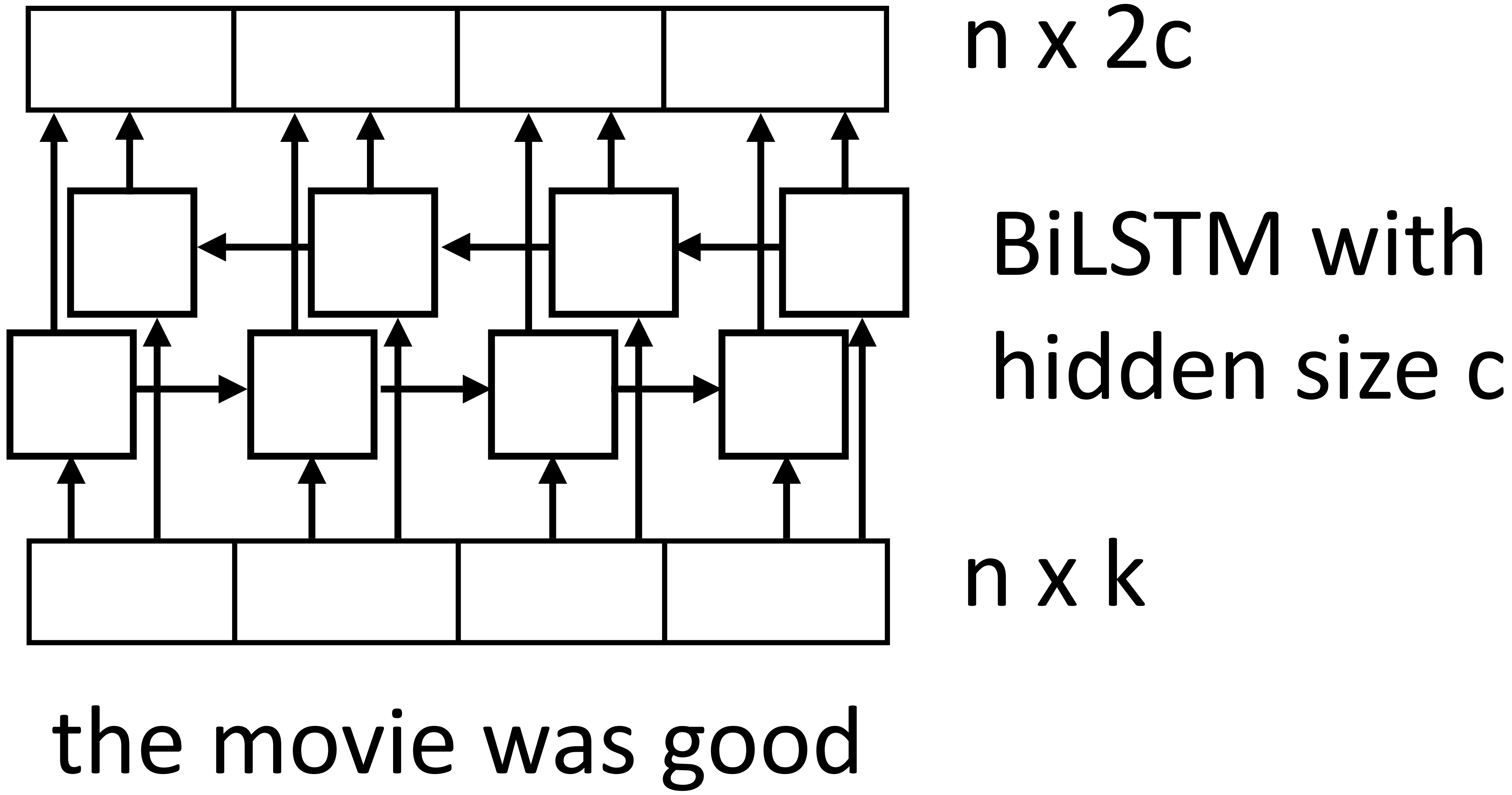
- Regression Metric (left): trained on a regression task using source, MT and reference; Ranking Metric (middle): optimize to encode good translations closer to the anchors (source, reference) while pushing bad translations away; Reference-less Metric (right): does not use the reference translation.

Rei et al. (2020)

Seq2Seq Models

Recall: CNNs vs. LSTMs

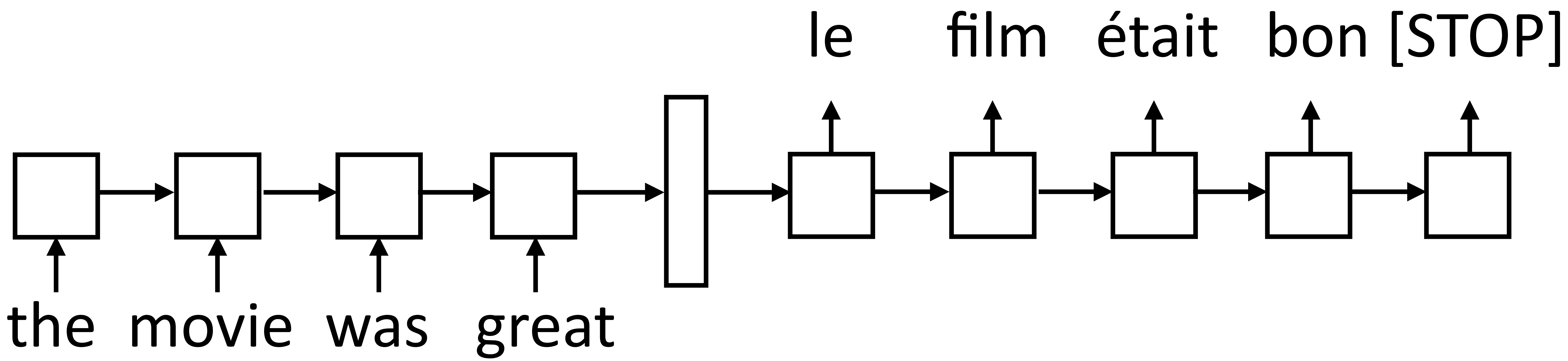




- Both LSTMs and convolutional layers transform the input using context
- LSTM: “globally” looks at the entire sentence (but local for many problems)
- CNN: local depending on filter width + number of layers

Encoder-Decoder

- Encode a sequence into a fixed-sized vector



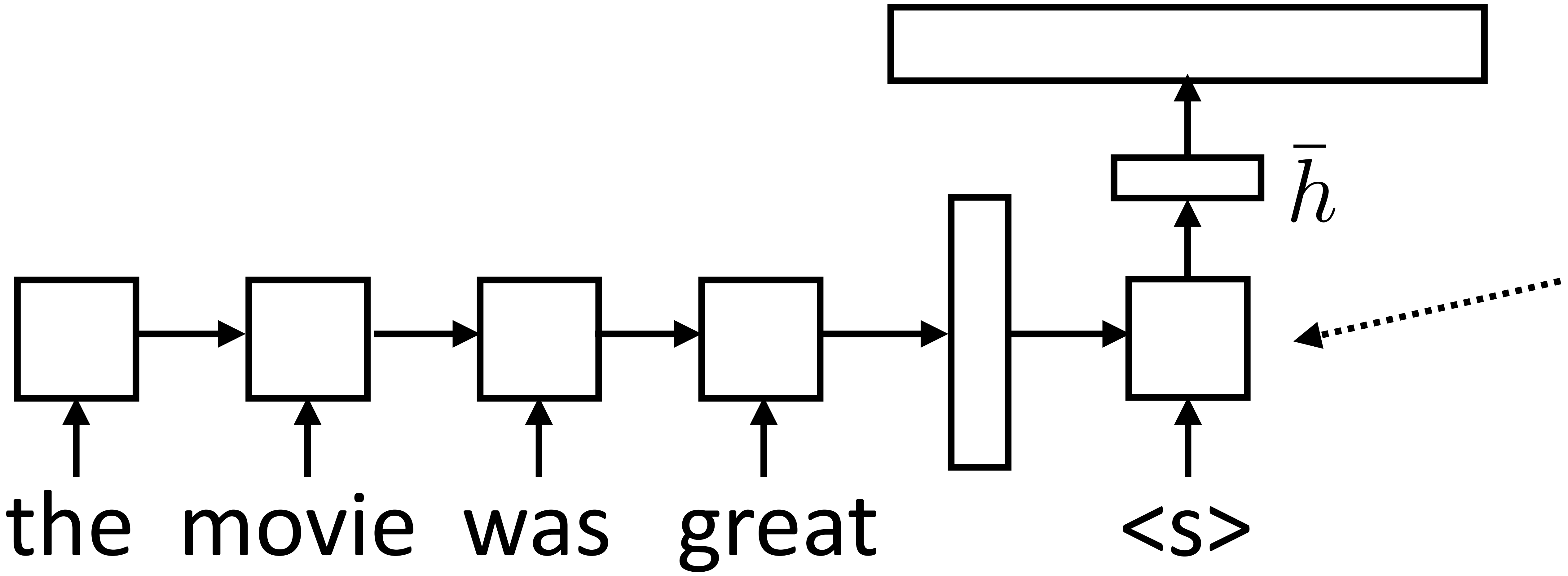
- Now use that vector to produce a series of tokens as output from a separate LSTM *decoder*
- Machine translation, NLG, summarization, dialog, and many other tasks (e.g., semantic parsing, syntactic parsing) can be done using this framework.

Sutskever et al. (2014)

Model

- Generate next word conditioned on previous word as well as hidden state

- W size is |vocab| x |hidden state|, softmax over entire vocabulary



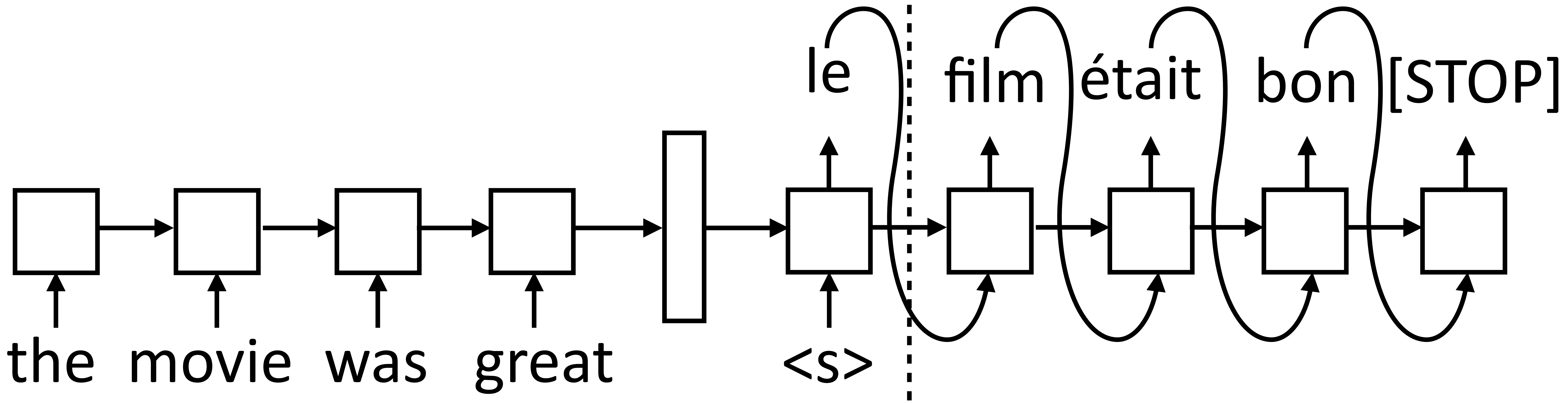
$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \text{softmax}(W\bar{h})$$

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1})$$

Decoder has separate parameters from encoder, so this can learn to be a language model (produce a plausible next word given current one)

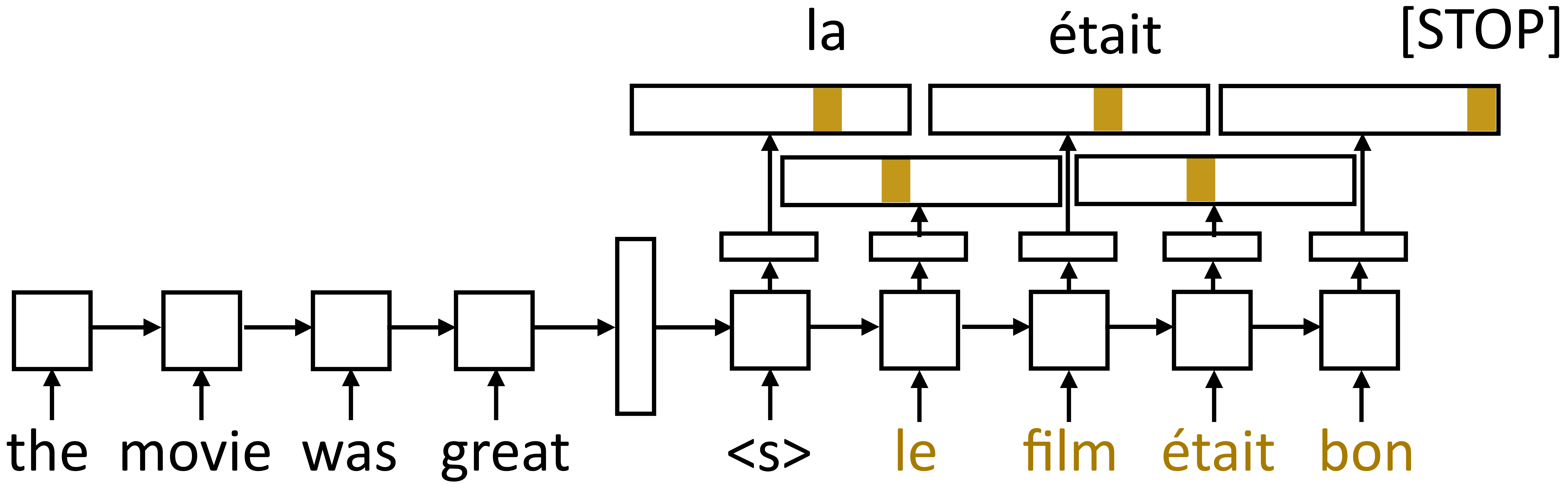
Inference

- Generate next word conditioned on previous word as well as hidden state



- During inference: need to compute the argmax over the word predictions and then feed that to the next RNN state
- Decoder is advanced one state at a time until [STOP] is reached

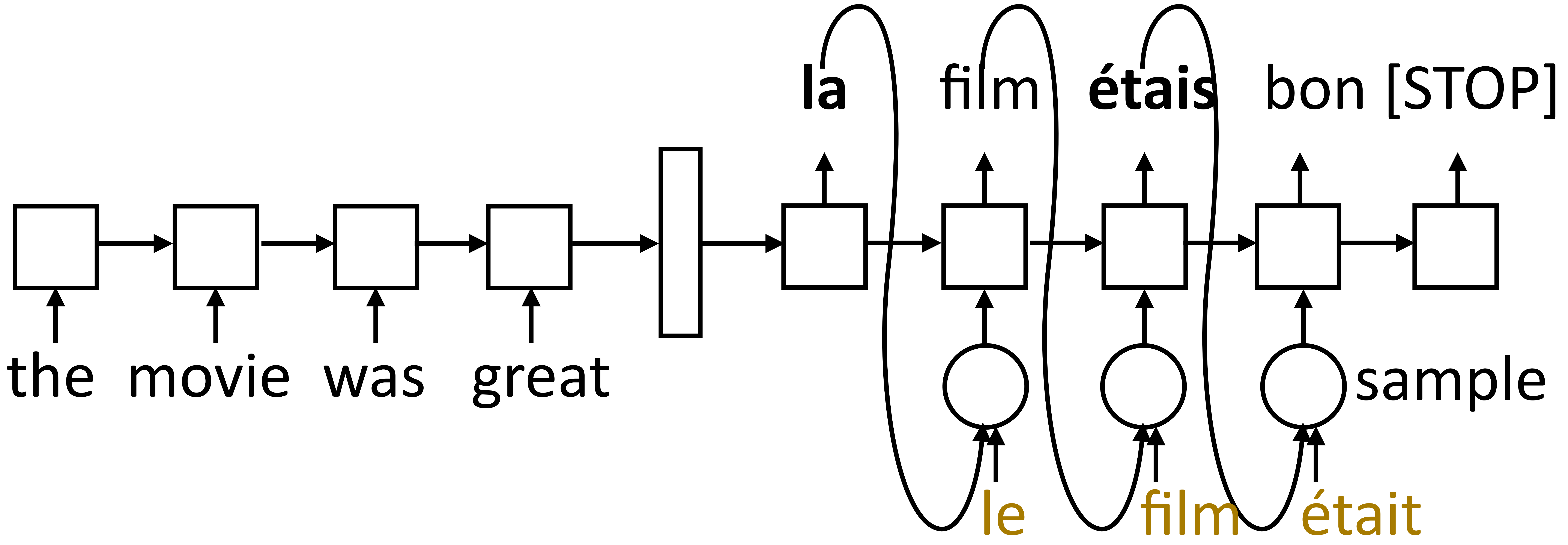
Training



- Objective: maximize $\sum_{(\mathbf{x},\mathbf{y})} \sum_{i=1}^{n} \log P(y_i^* | \mathbf{x}, y_1^*, \ldots, y_{i-1}^*)$
- One loss term for each target-sentence word, feed the correct word regardless of model's prediction (called "teacher forcing")

Training: Scheduled Sampling

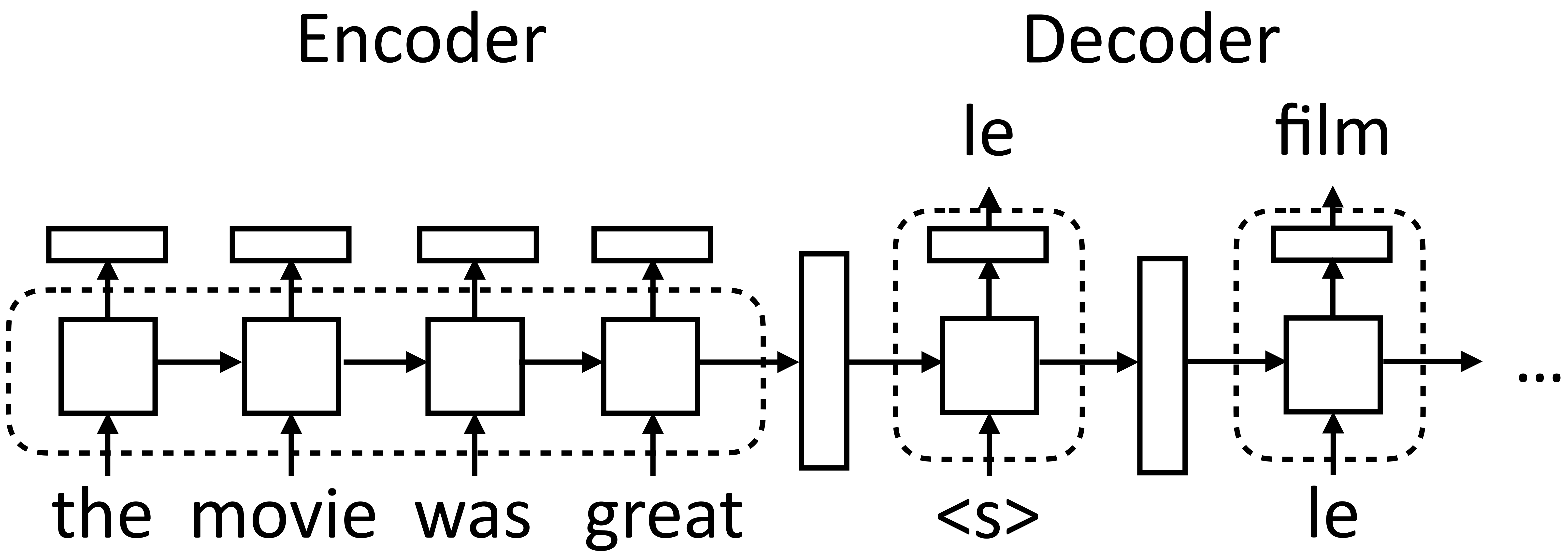
- Model needs to do the right thing even with its own predictions



- Scheduled sampling: with probability *p*, take the gold (human) translation as input, else take the model's prediction
- Starting with *p* = 1 and decaying it works best

Bengio et al. (2015)

Implementing seq2seq Models



- Encoder: consumes sequence of tokens, produces a vector. Analogous to encoders for classification/tagging tasks
- Decoder: separate module, single cell. Takes two inputs: hidden state (vector *h* or tuple (*h*, *c*)) and previous token. Outputs token + new state

Implementation Details

- Sentence lengths vary for both encoder and decoder:
  - Typically pad everything to the right length
- Encoder: Can be a CNN/LSTM/Transformer…
- Batching is a bit tricky:
  - encoder should use pack_padded_sequence to handle different lengths.
  - The decoder should pad everything to the same length and use a mask to only accumulate “valid” loss terms
  - Label vectors may look like [num timesteps x batch size x num labels]

Implementation Details (cont')

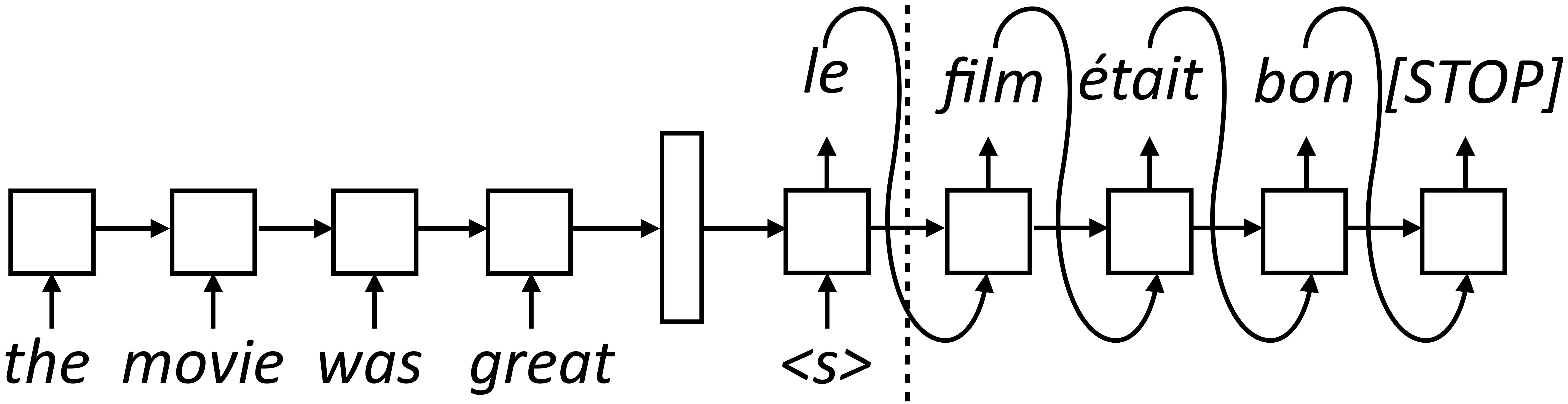
- Decoder: execute one step of computation at a time, so computation graph is formulated as taking one input + hidden state.
  - Test time: do this until you generate the [STOP] token
  - Training time: do this until you reach the gold stopping point
- Beam search: can help with lookahead. Finds the (approximate) highest scoring sequence:

$$\mathrm{argmax}_{\mathbf{y}} \prod_{i=1}^{n} P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1})$$

Decoding Strategies

Greedy Decoding

- Generate next word conditioned on previous word as well as hidden state



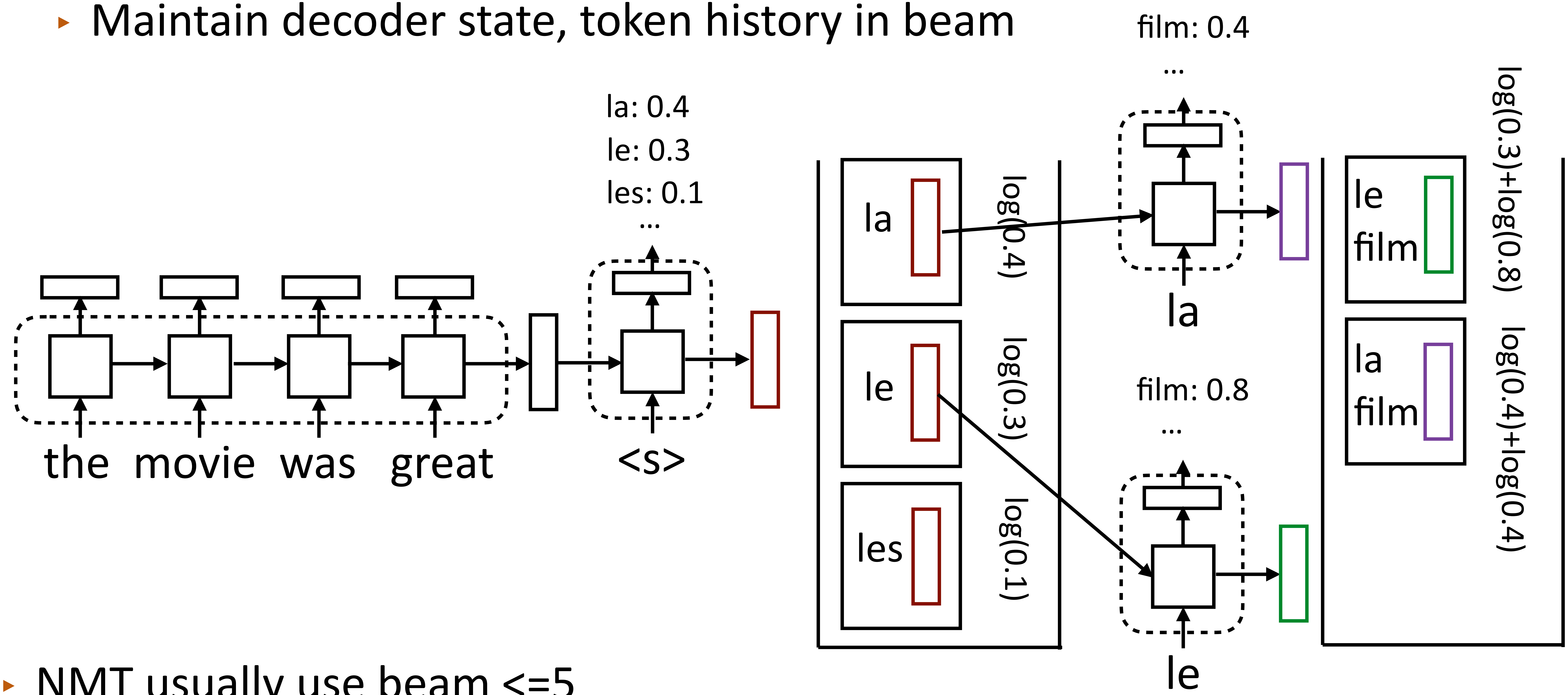
- During inference: need to compute the argmax over the word predictions and then feed that to the next RNN state. This is **greedy decoding**

$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h})$$

$$y_{\text{pred}} = \mathrm{argmax}_y P(y|\mathbf{x}, y_1, \ldots, y_{i-1})$$

Beam Search

- Maintain decoder state, token history in beam



- NMT usually use beam <=5
- Keep **both** *film* states! Hidden state vectors are different

Problems with Greedy Decoding

- Only returns one solution, and it may not be optimal
- Can address this with **beam search**, which usually works better…but even beam search may not find the correct answer! (max probability sequence)

Model	Beam-10	
	BLEU	#Search err.
LSTM*	28.6	58.4%
SliceNet*	28.8	46.0%
Transformer-Base	30.3	57.7%
Transformer-Big*	31.7	32.1%

A sentence is classified as search error if the decoder does not find the global best model score.

Stahlberg and Byrne (2019)

“Problems” with Beam Decoding

- For machine translation, the highest probability sequence is often the empty string, i.e.. a single </s> token!   (>50% of the time)

Search	BLEU	Ratio	#Search errors	#Empty
Greedy	29.3	1.02	73.6%	0.0%
Beam-10	30.3	1.00	57.7%	0.0%
Exact	2.1	0.06	0.0%	51.8%

- Beam search results in *fortuitous search errors* that avoid these bad solutions. NMT usually use beam <=5.
- Exact inference uses depth-first search, but cut off branches that fall below a lower bound.

Stahlberg and Byrne (2019)

Sampling

- Beam search may give many similar sequences, and these actually may be *too close* to the optimal. Can sample instead:

$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h})$$

$$y_{\mathrm{sampled}} \sim P(y|\mathbf{x}, y_1, \ldots, y_{i-1})$$

- Greedy solution can be uninteresting / vacuous for various reasons (so called text *degeneration)*. Sampling can help - especially for some text generation tasks.

Beam Search vs. Sampling

- These are samples from an unconditioned language model GPT-2 (not seq2seq model)

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

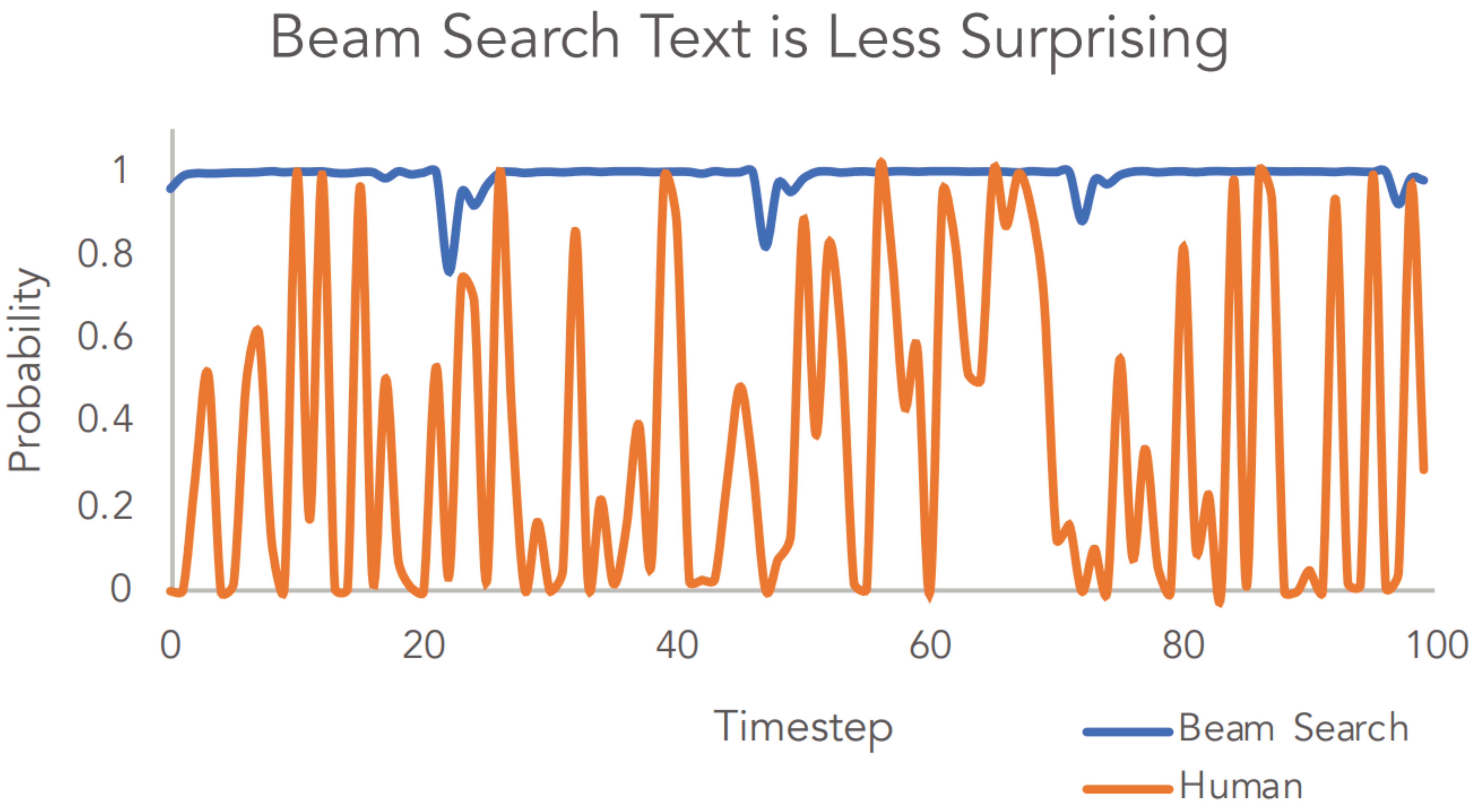
Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

- Sampling is better but sometimes draws too far from the tail of the distribution (relatively low prob. over thousands of candidate tokens).

Holtzman et al. (2019)

Beam Search vs. Sampling



Holtzman et al. (2019)

Decoding Strategies

- Greedy
- Beam search
- Sampling (e.g., top-k or Nucleus sampling)
  - Top-k: take the top k most likely words (k=5), sample from those
  - Nucleus: take the top p% (95%) of the distribution, sample from within that

Other Applications of Seq2Seq

Generation Tasks

- There are a range of seq2seq modeling tasks we will address
- For more constrained problems: greedy/beam decoding are usually best
- For less constrained problems: nucleus sampling introduces favorable variation in the output

Less constrained

More constrained

Unconditioned sampling/
e.g., story generation

Dialogue

Summarization
Data-to-text

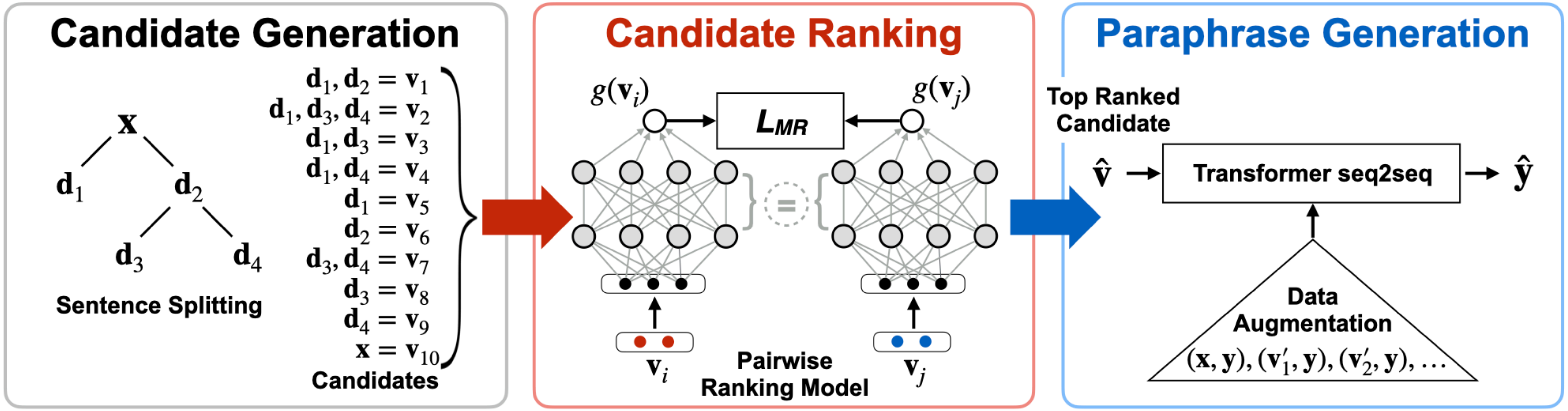
Translation

Text-to-text

Text-to-code

Text-to-Text Generation

- Text Simplification (with readability constraints)



Input sentece:

Since 2010, project researchers have uncovered documents in Portugal that have revealed who owned the ship

→ **seq2seq models** (RNN, Transformer) →

Generated Output:

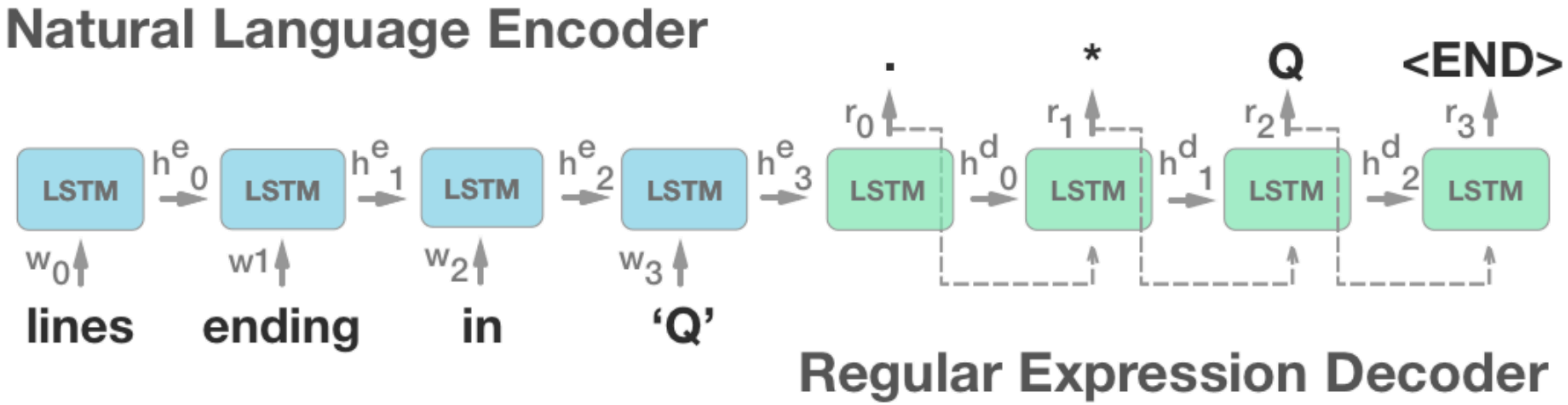
Scientists have found documents in Portugal.

They have also found out who owned the ship.

Mounica Maddela, Fernando Alva-Manchego, Wei Xu. "Controllable Text Simplification with Explicit Paraphrasing" (NAACL 2021)

Regex Prediction

- Seq2seq models can be used for many other tasks!
- Predict regex from text



- Problem: requires a lot of data: 10,000 examples needed to get ~60% accuracy on pretty simple regexes

Locascio et al. (2016)

Semantic Parsing as Translation

"what states border Texas"

$$\downarrow$$

$$\lambda\ x\ \mathtt{state(}\ x\ \mathtt{)}\ \wedge\ \mathtt{borders(}\ x\ \mathtt{,}\ \mathtt{e89}\ \mathtt{)}$$

- Write down a linearized form of the semantic parse, train seq2seq models to directly translate into this representation; using copy mechanism
- No need to have an explicit grammar, simplifies algorithms
- Might not produce well-formed logical forms, might require lots of data

Semantic Parsing/Lambda Calculus: https://www.youtube.com/watch?v=OocGXG-BY6k&t=200s

Jia and Liang (2016)

SQL Generation

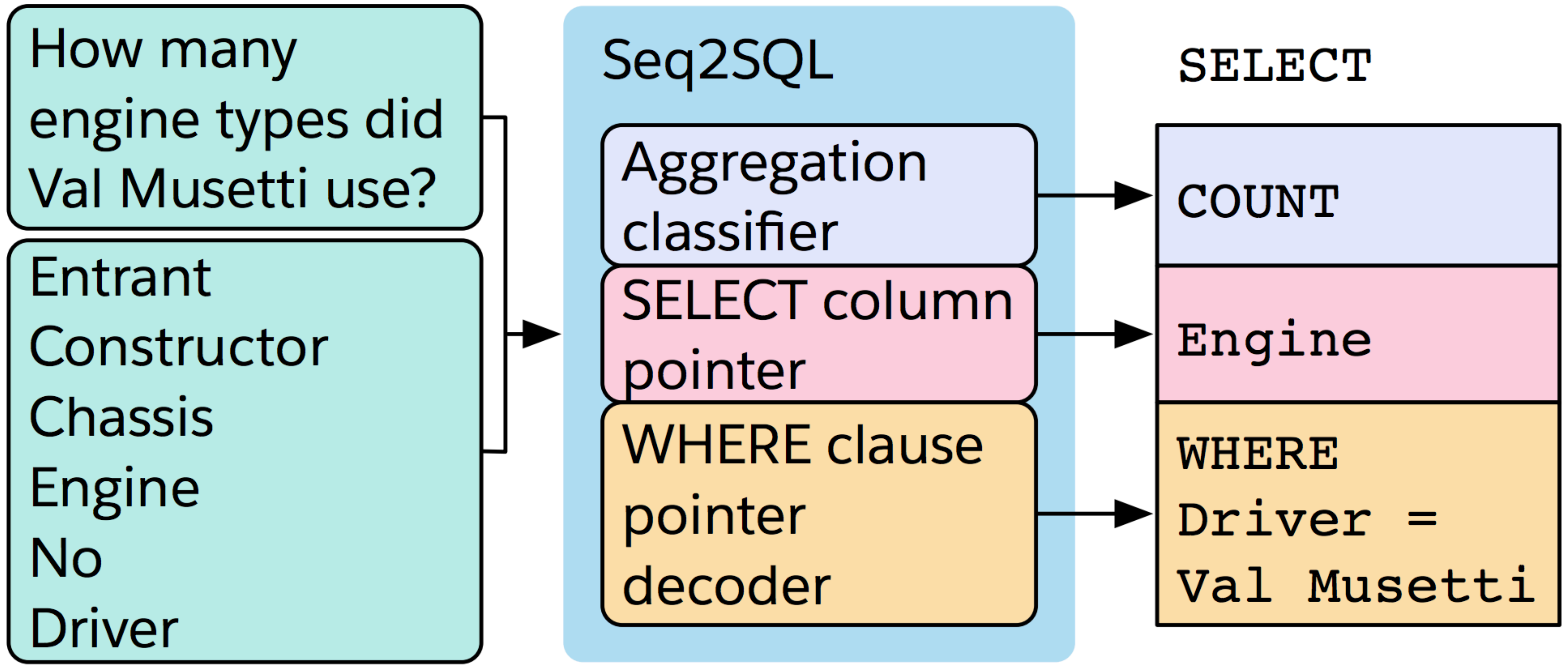
- Convert natural language description into a SQL query against some DB
- How to ensure that well-formed SQL is generated?
  - Three components
- How to capture column names + constants?
  - Pointer mechanisms

Question:

How many CFL teams are from York College?

SQL:

```
SELECT COUNT CFL Team FROM
CFLDraft WHERE College = “York”
```



Zhong et al. (2017)