Neural Machine Translation

CSCI 1460: Computational Linguistics

Lecture 9

Ellie Pavlick

Fall 2023

Topics

- SMT Followup: Putting it all together
- MT Evaluation
- Neural MT
  - Encoder-Decoder Models
  - Multilingual LMs and “Zero Shot” Cross-Lingual Transfer

Topics

- **SMT Followup: Putting it all together**
- MT Evaluation
- Neural MT
  - Encoder-Decoder Models
  - Multilingual LMs and “Zero Shot” Cross-Lingual Transfer

Full Phrase-Based MT System

$$\text{argmax}_{tgt \in TGT} P(src \mid tgt) P(tgt)$$

Full Phrase-Based MT System

$$\text{argmax}_{tgt \in TGT} P(src \mid tgt) P(tgt)$$

1. Translation Model

Full Phrase-Based MT System

$$\text{argmax}_{tgt \in TGT} P(src \mid tgt) P(tgt)$$

1. Translation Model
2. Language Model

Full Phrase-Based MT System

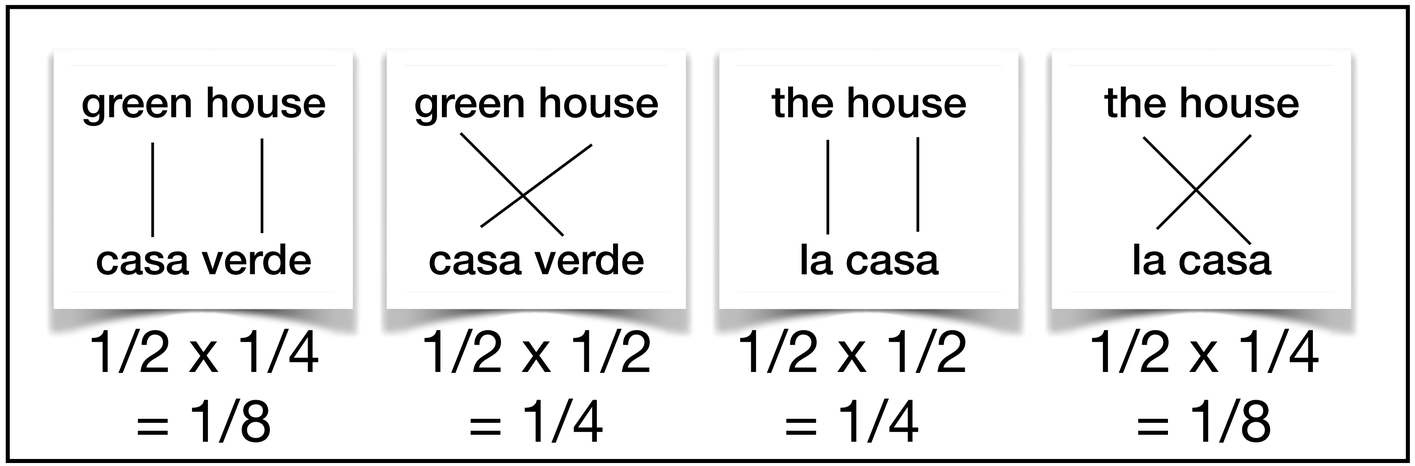
$$\text{argmax}_{tgt \in TGT} P(src \mid tgt) P(tgt)$$

1. Translation Model
2. Language Model
3. Decoder

Alignment: Produces a Translation Model

green house	casa verde
the house	la casa

a



t

		target		
		green	house	the
source	casa	1/2	1/2	1/2
	la	0	1/4	1/2
	verde	1/2	1/4	0

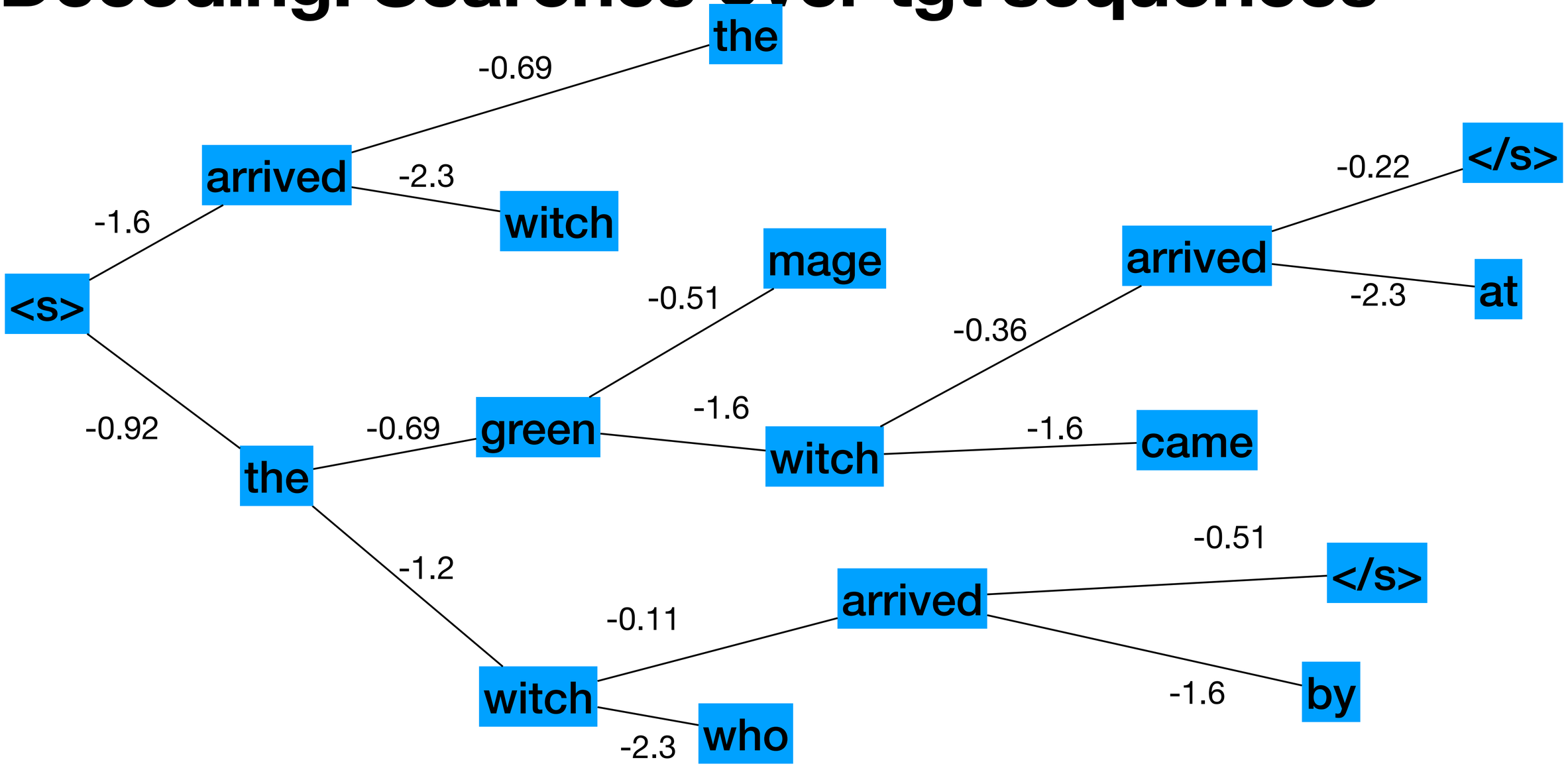
Lectures 9—12: Language Models

<s> can you tell me about any good cantonese restaurants close by </s>
<s> mid priced thai food is what i'm looking for </s>
<s> tell me about chez panisse </s>
<s> can you give me a listing of the kinds of food that are available </s>
<s> i'm looking for a good place to eat breakfast </s>
<s> when is caffe venezia open during the day </s>

$$P(w_0 \ldots w_n) \approx P(w_0 | \text{<s>}) \times P(w_1 | w_0) \times P(w_2 | w_1) \times \ldots \times P(w_n | w_{n-1})$$

P(tell me about caffe venezia) = P(tell|<s>) x P(me|tell) x P(about|me) x
P(caffe|about) x P(venezia|caffe) x P(</s>|venezia)

Decoding: Searches over tgt sequences



Full Phrase-Based MT System

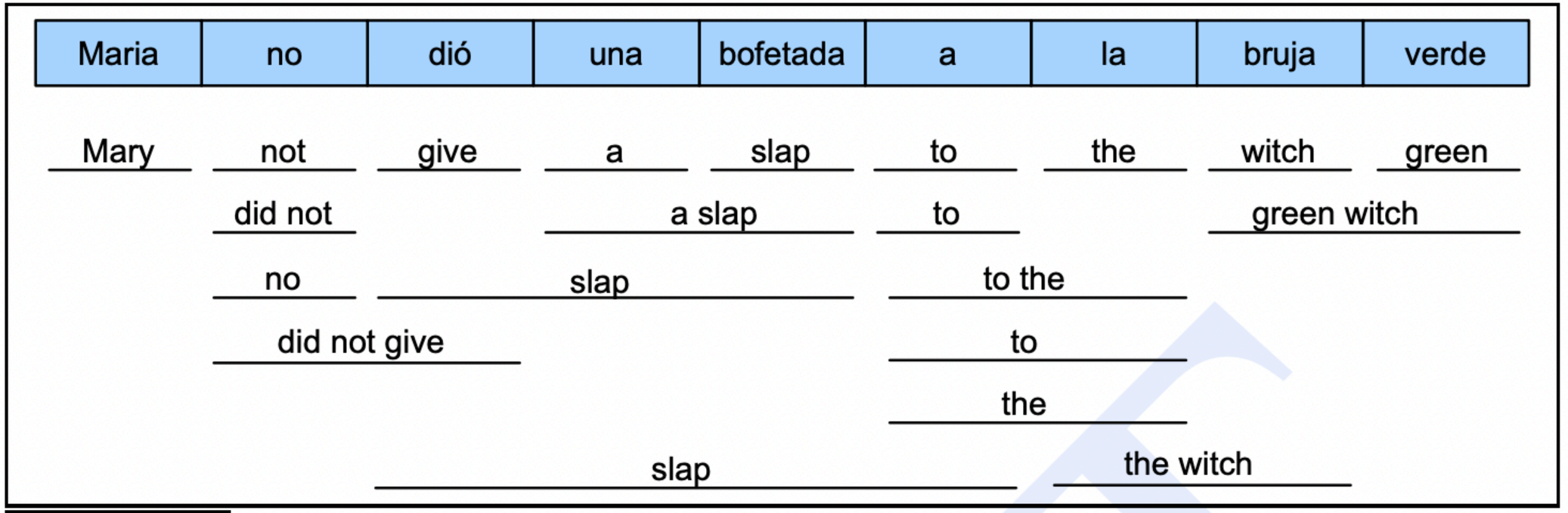


Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)

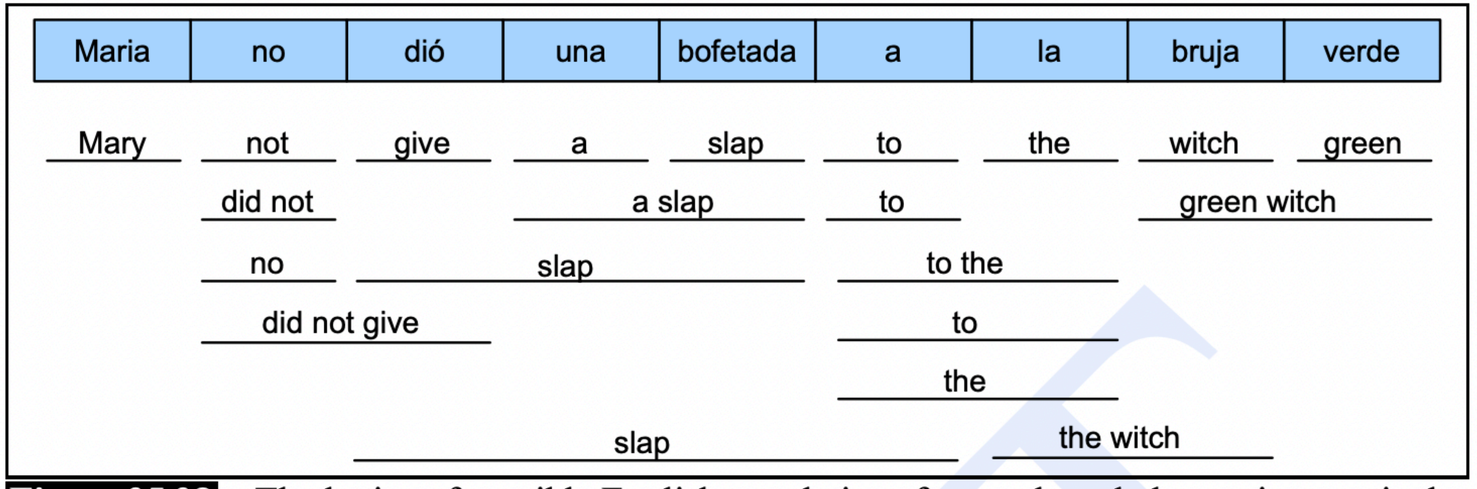
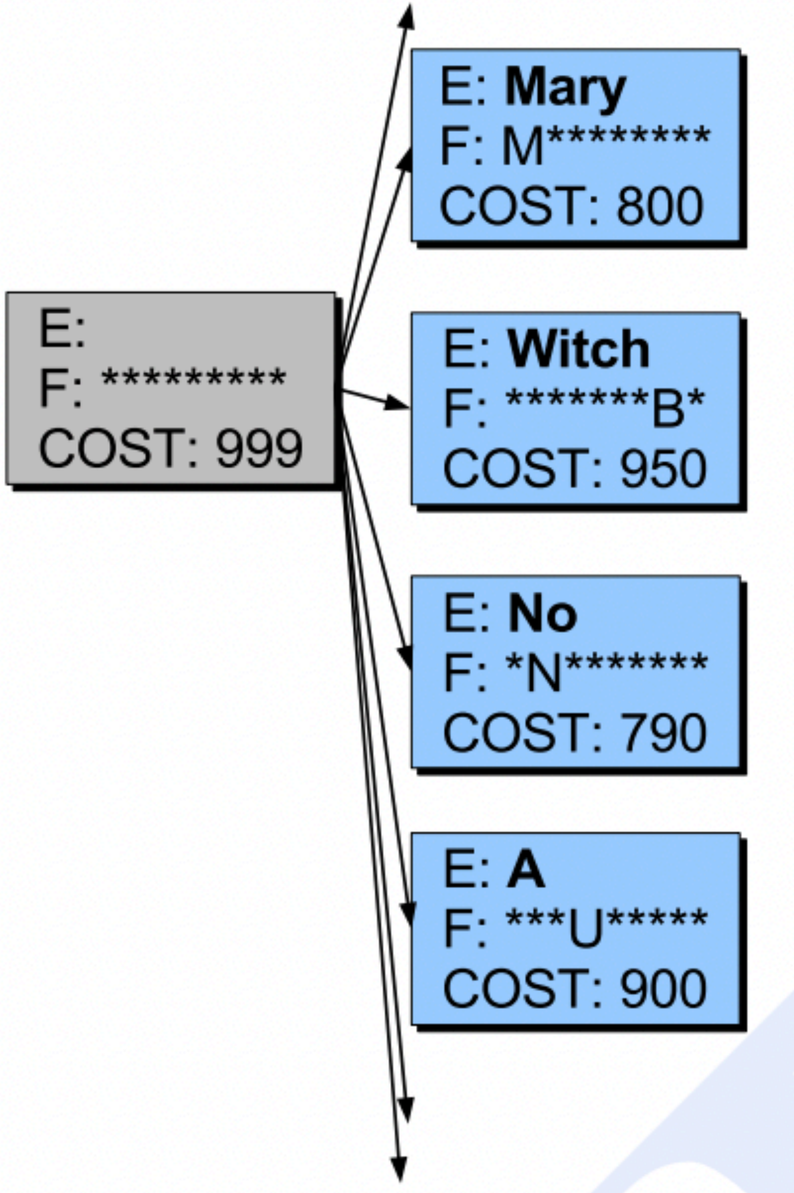


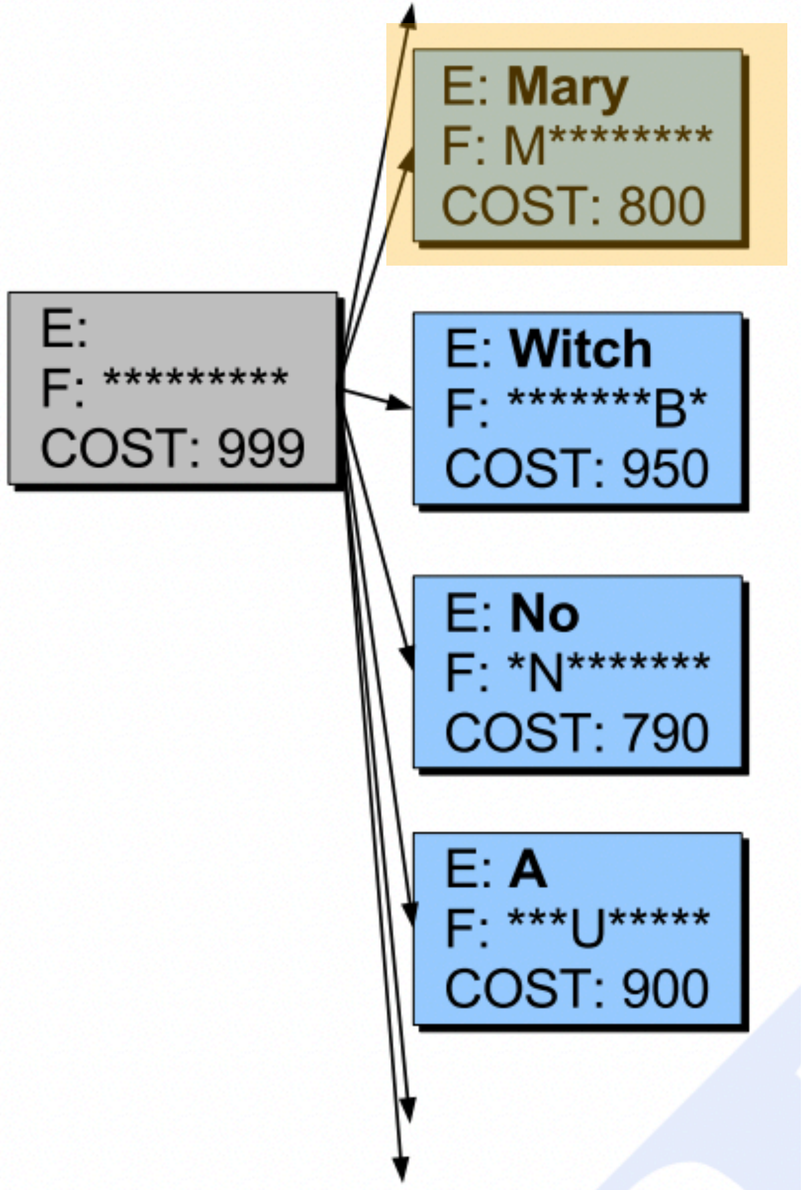
Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)



a) after expanding NULL

Maria	no	dió	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		to		green witch	
	no	slap			to the			
	did not give				to			
					the			
		slap				the witch		

Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)



a) after expanding NULL

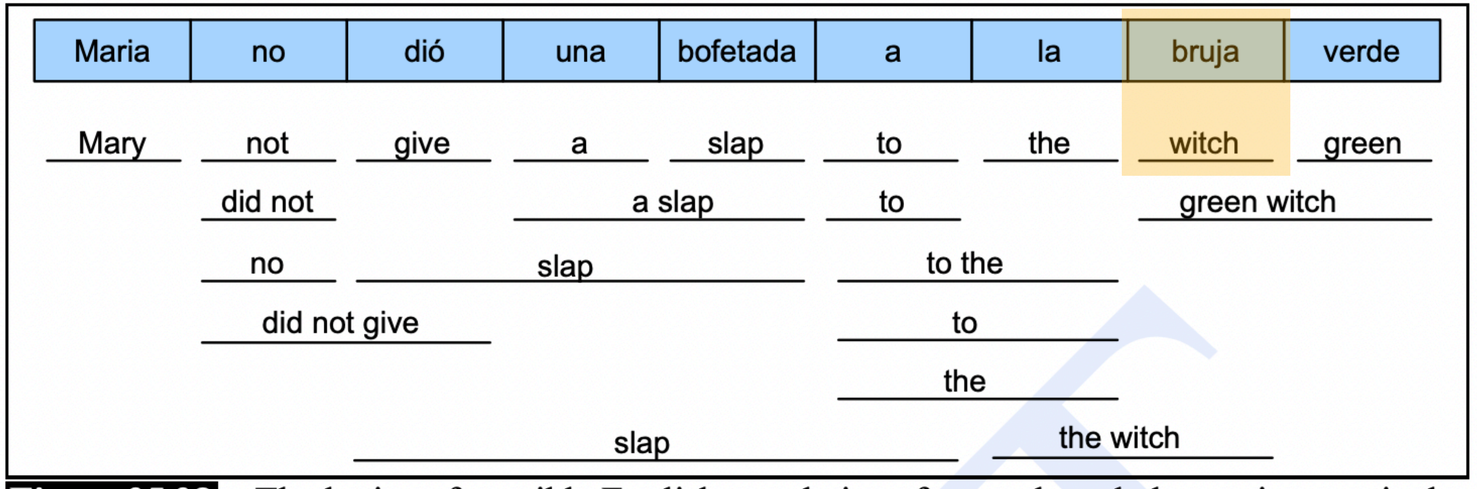
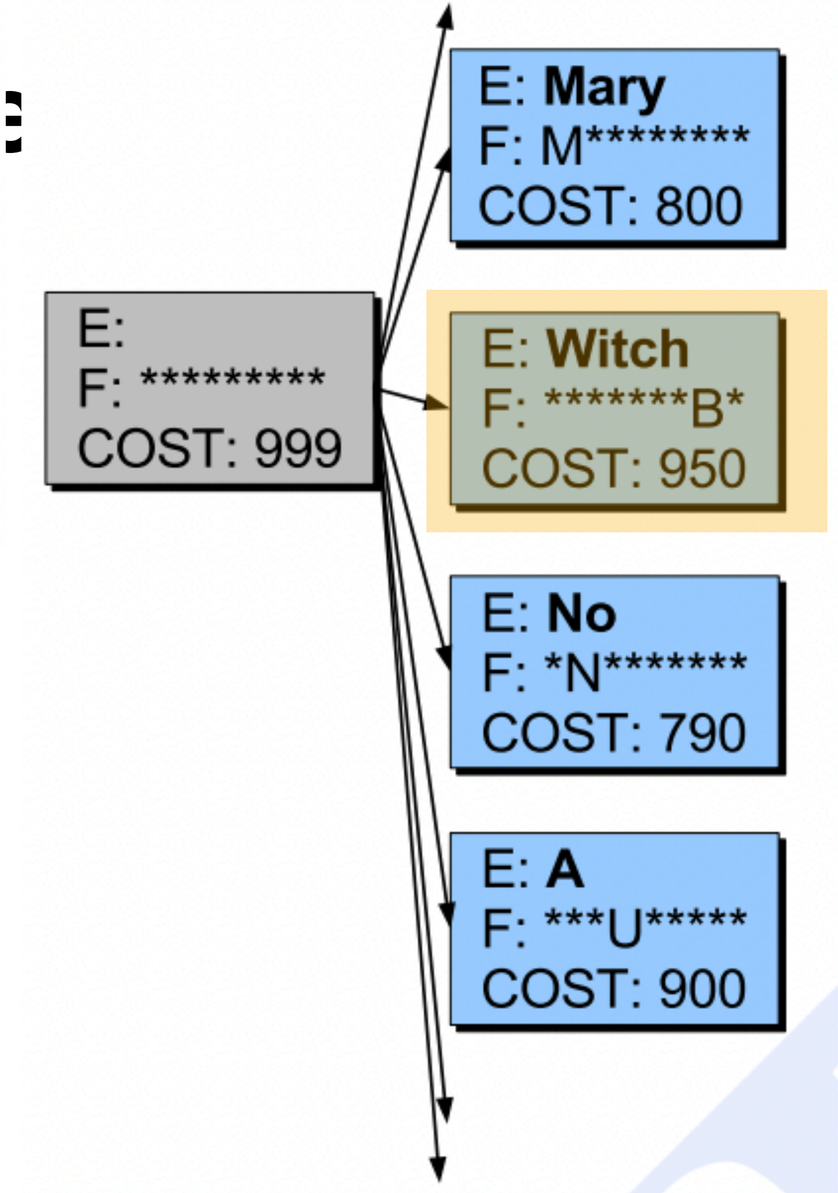


Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)



a) after expanding NULL

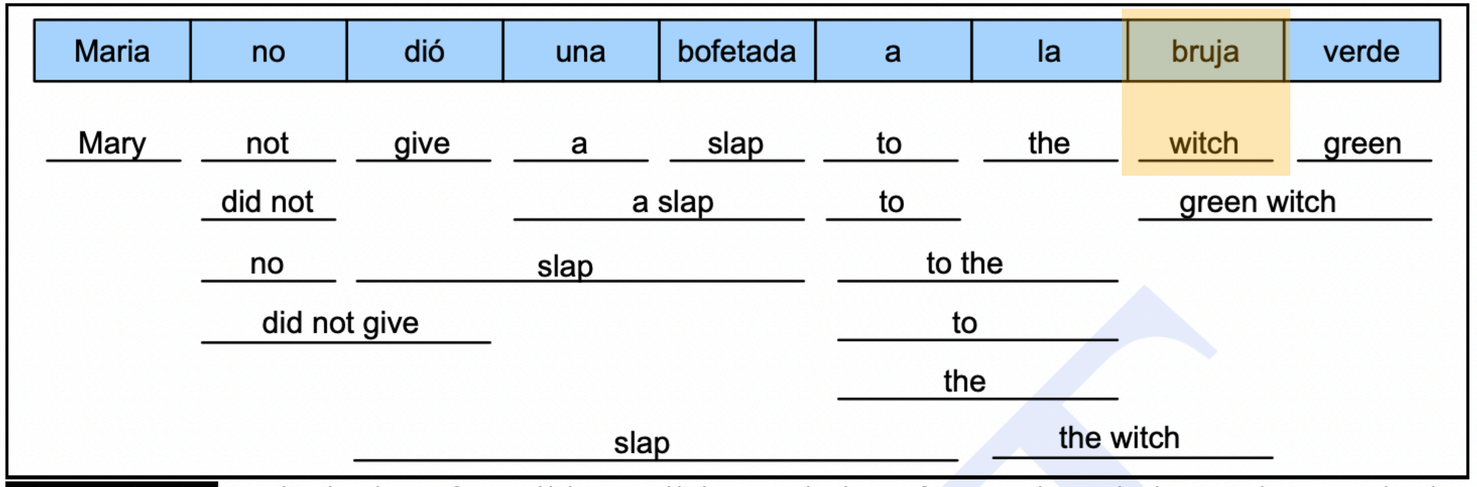
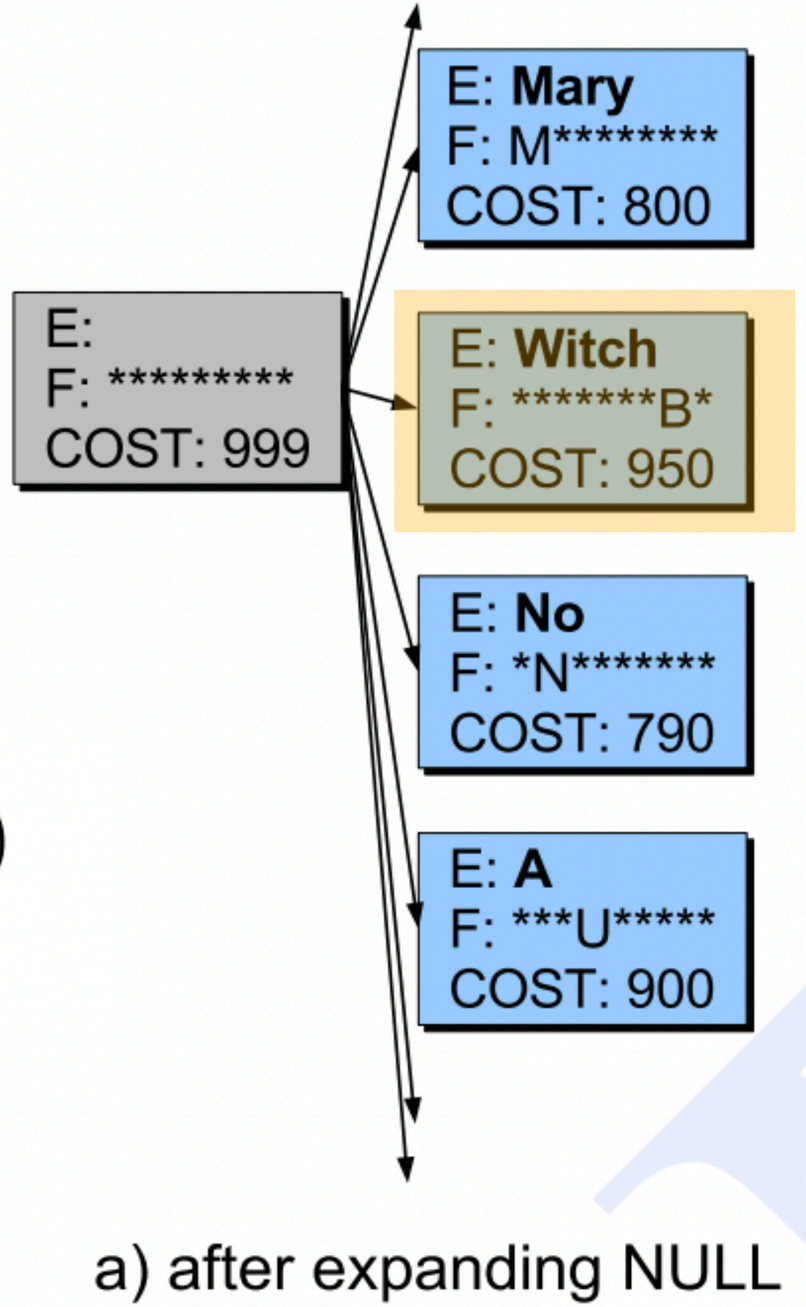


Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)



$$\text{cost}(E,F) = \prod_{i \in S} \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1}) P(E)$$

product over all positions in the (partial) sentence

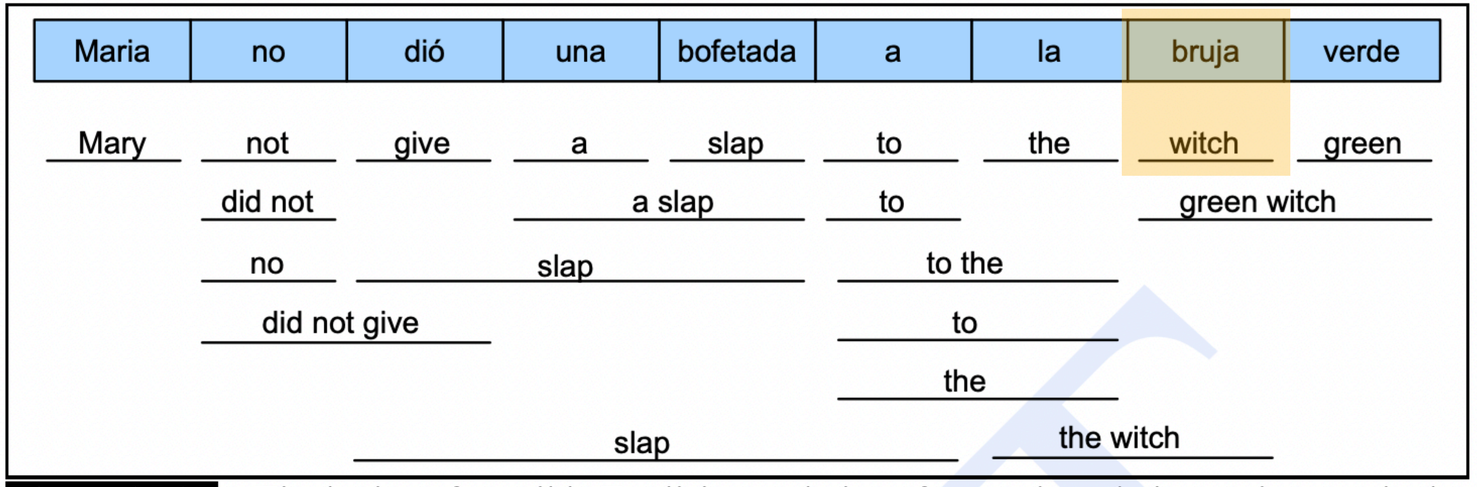
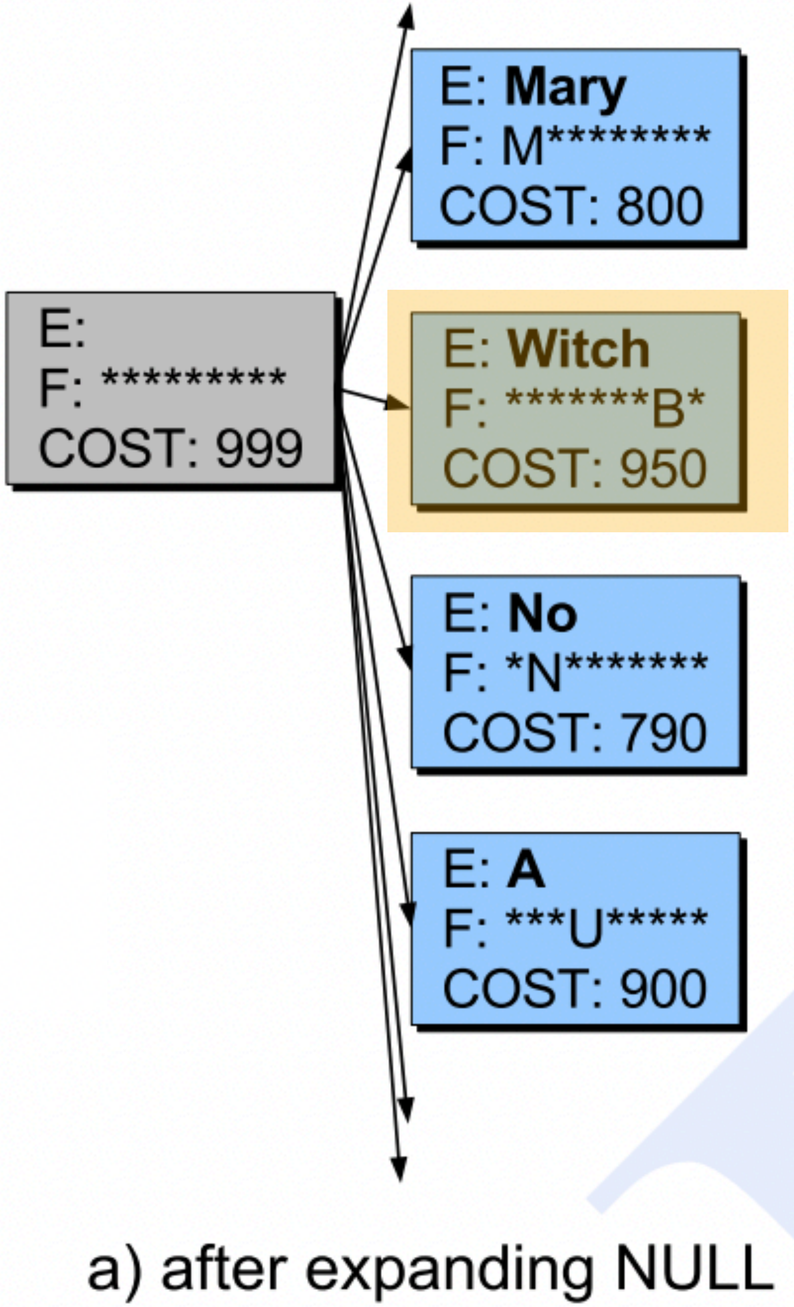


Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)



$$\text{cost}(E,F) = \prod_{i \in S} \phi(\overline{f}_i, \overline{e}_i) d(a_i - b_{i-1}) P(E)$$

translation probability
(e.g., from IBM alignment model)

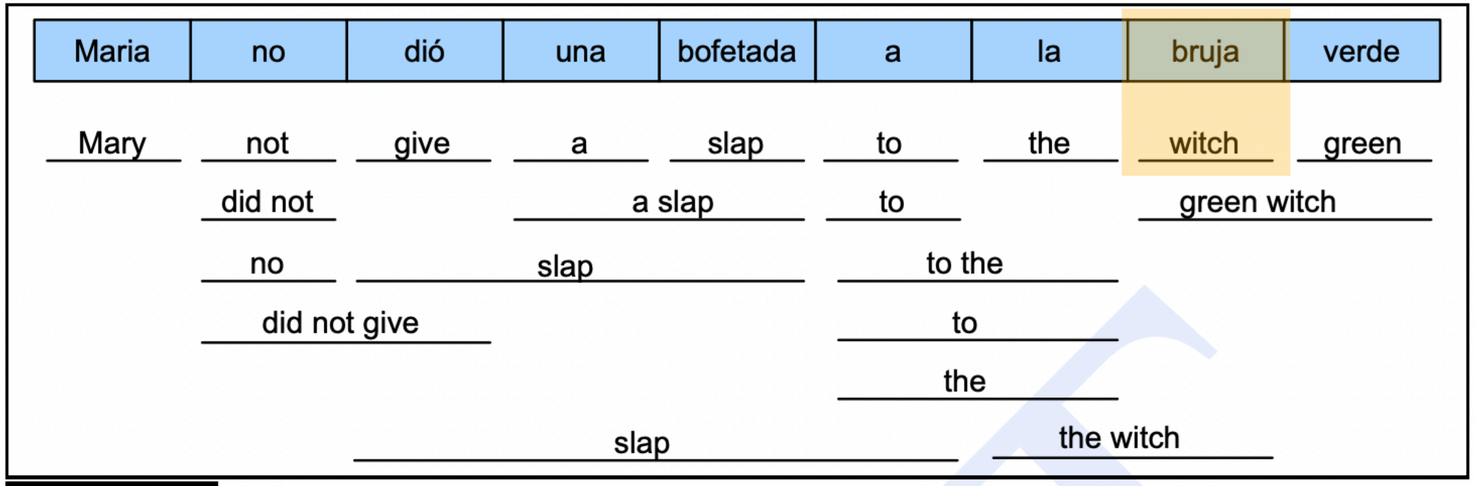
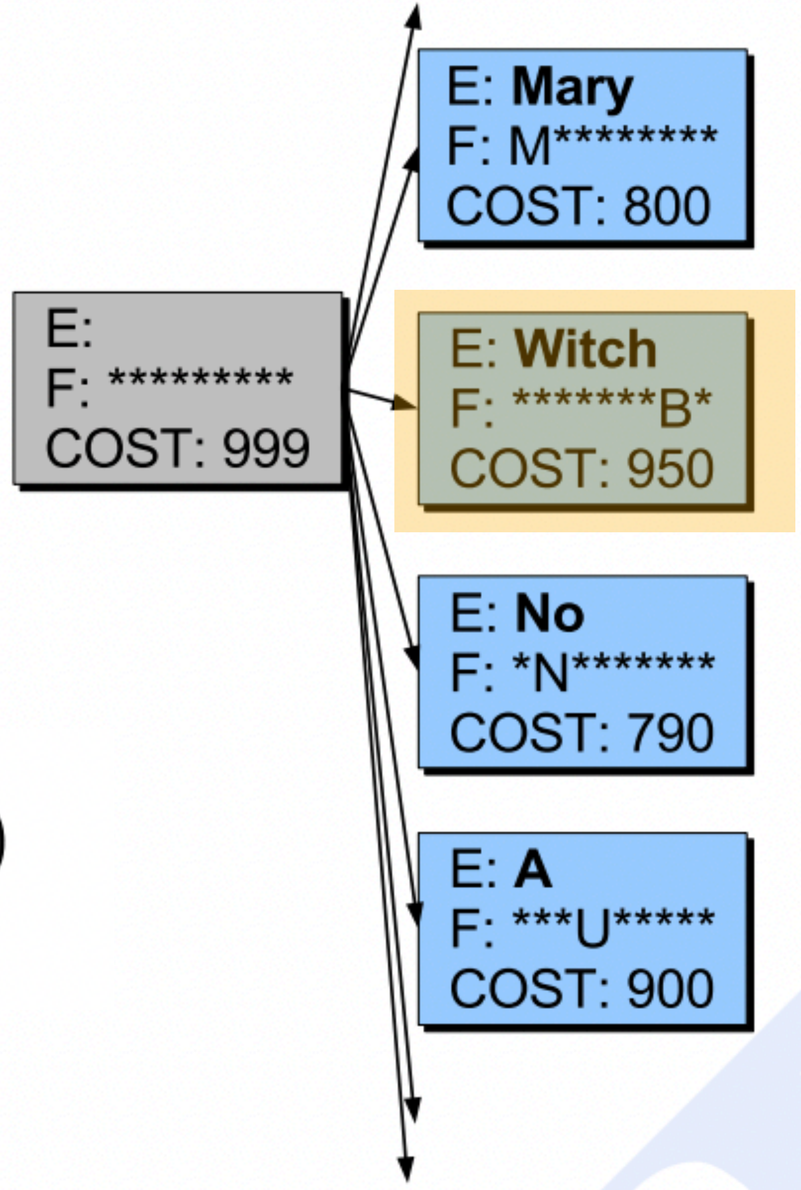


Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)



a) after expanding NULL

$$\text{cost}(E,F) = \prod_{i \in S} \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1}) P(E)$$

distortion probability

(deals with offsets, e.g., probability of translating the 8th word first)

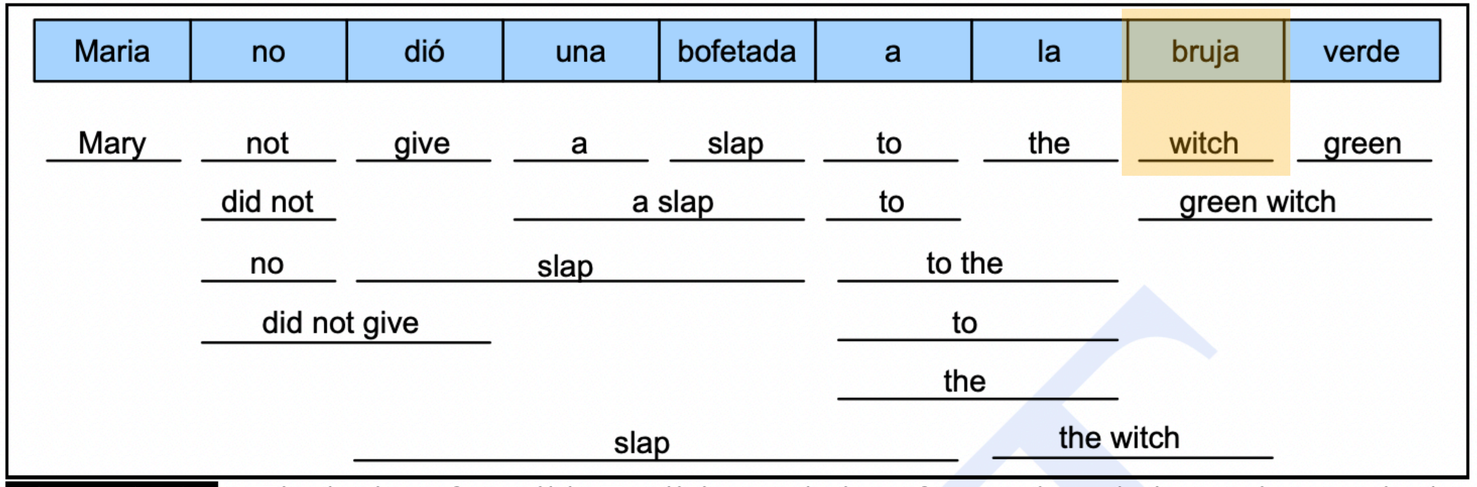
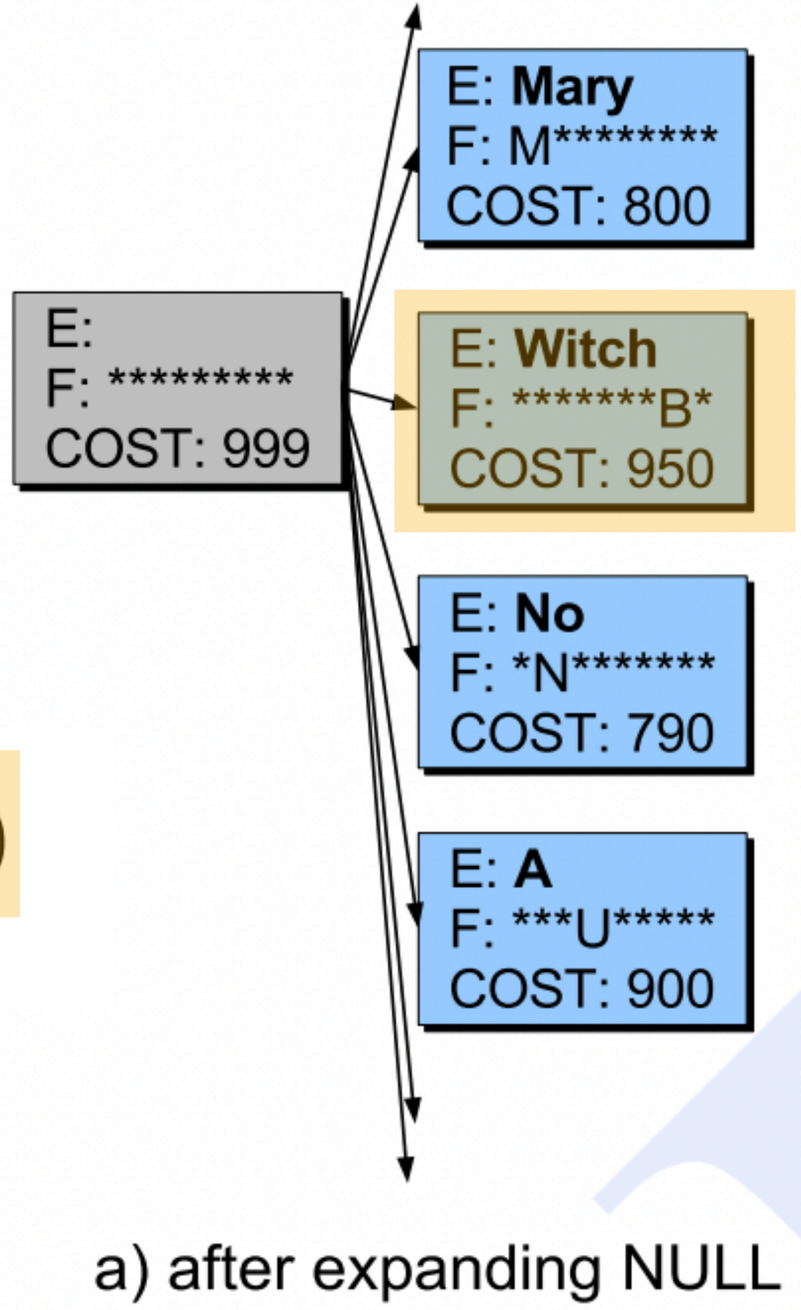


Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)

$$\text{cost}(E,F) = \prod_{i \in S} \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1}) P(E)$$

language model probability



a) after expanding NULL

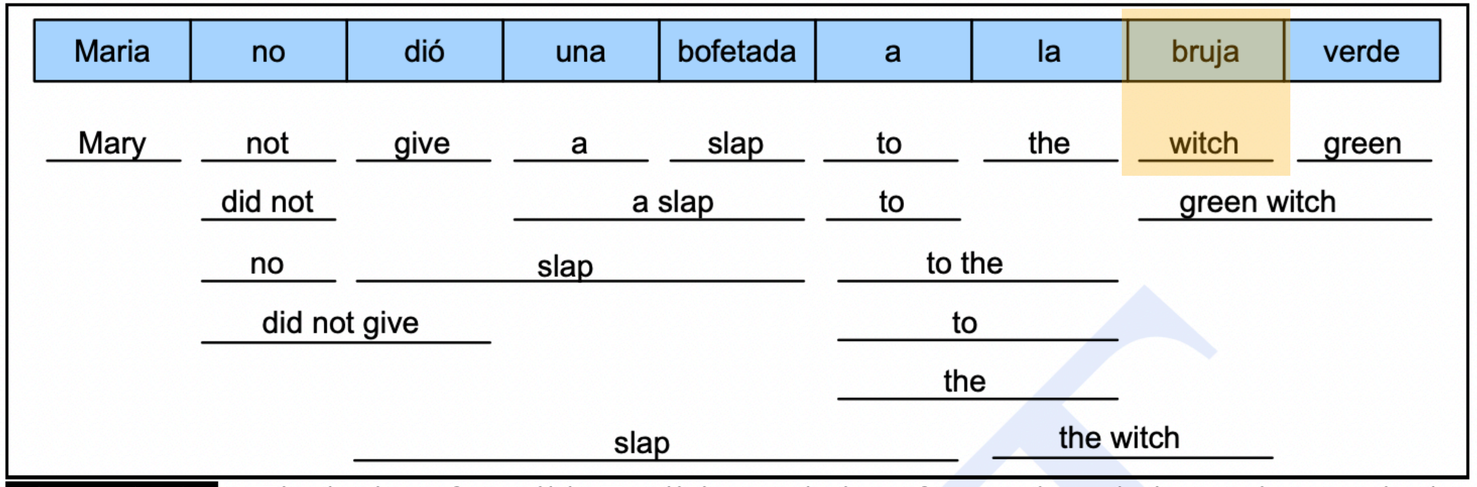
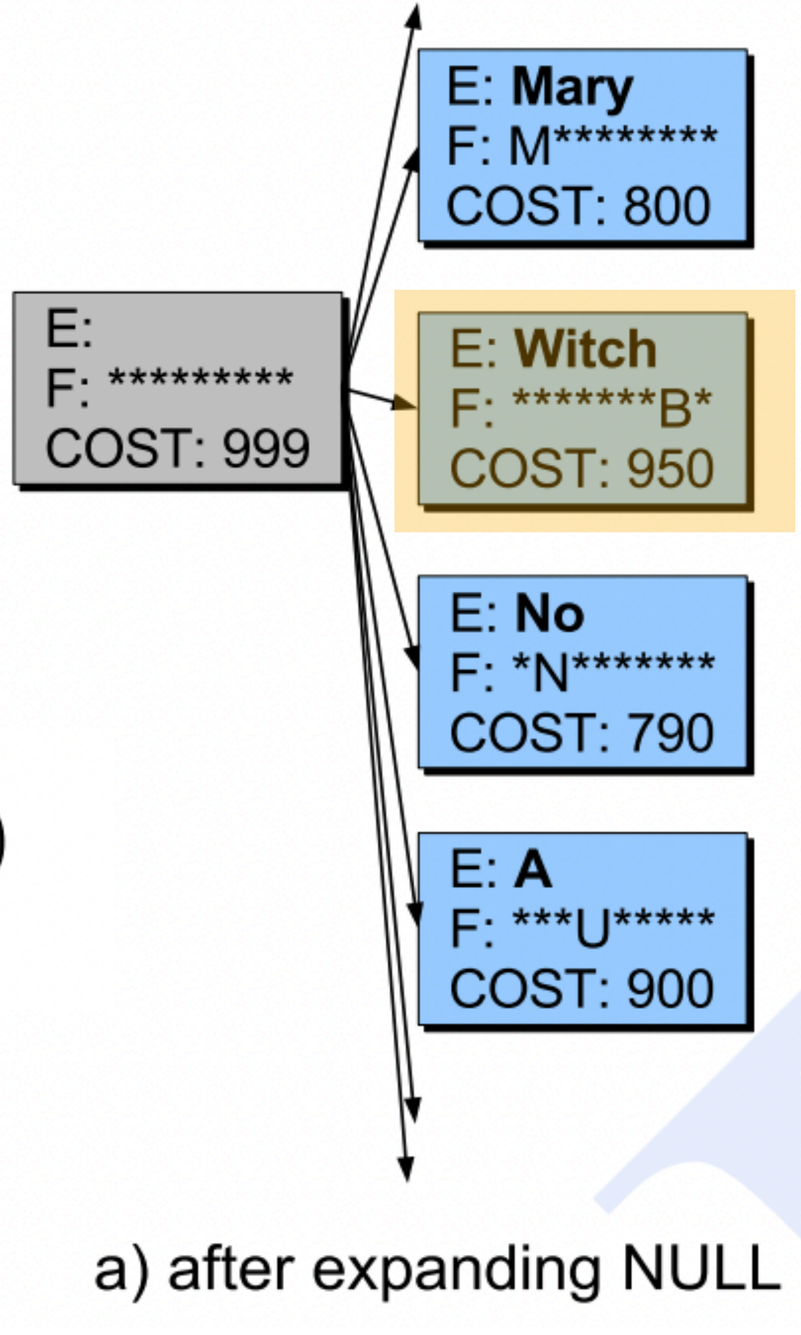


Figure 25.28 The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set. After Koehn (2003a)



$$\text{cost}(E,F) = \prod_{i \in S} \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1}) P(E)$$

(also combined with an estimated "future cost" of translating the remaining words.)

Topics

- SMT Followup: Putting it all together
- **MT Evaluation**
- Neural MT
  - Encoder-Decoder Models
  - Multilingual LMs and “Zero Shot” Cross-Lingual Transfer

MT Evaluation

- Human Evaluation:
  - Explicit ratings for fluency and faithfulness
  - Most reliable, but expensive to collect
  - Needs to be collected new for every system
  - Can’t be “hill climbed”

MT Evaluation

- Automatic Evaluations:
  - NLP likes automatic evals for standardization and optimization
  - Most popular metric for MT is **BLEU**

MT Evaluation

BLEU

- Assume we have an MT output (“candidate”) and are comparing against multiple human-generated “referent” translations
- Intuition: We should reward models for producing translations that contains lots of the same words/phrases as the references

MT Evaluation

BLEU

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party
Candidate 2	It is to insure the troops forever hearing the activity guidebook that party direct

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

MT Evaluation

BLEU

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party
Candidate 2	It is to insure the troops forever hearing the activity guidebook that party direct

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

MT Evaluation

BLEU

- BLEU: Weighted n-gram precision with a length penalty

$$Bleu = BP \times exp\left(\frac{1}{N}\sum_{n=1}^{N} logp_n\right)$$

MT Evaluation
BLEU

- BLEU: Weighted n-gram precision with a length penalty

$$Bleu = BP \times exp\left(\frac{1}{N}\sum_{n=1}^{N} logp_n\right)$$

weighted
ngram
precision

MT Evaluation
BLEU

- BLEU: Weighted n-gram precision with a length penalty

$$Bleu = BP \times exp\left(\frac{1}{N}\sum_{n=1}^{N} logp_n\right)$$

"brevity penalty"

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

unigram precision = 1

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

unigram precision = 2

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

unigram precision = 3

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

unigram precision = 4

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

unigram precision = 17

(every word except "obeys")

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

unigram precision = 17 / 18 = 0.94

(every word except "obeys")

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	party party party party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

unigram precision = 4 / 4 = 1.0

(every word except "obeys")

MT Evaluation

BLEU

weighted ngram precision

$$p_n = \frac{\sum_{c \in cand} \sum_{ngm \in c} count_{clip}(ngm)}{\sum_{c' \in cand} \sum_{ngm' \in c'} count(ngm')}$$

Candidate 1	party party party party

Reference 1	It is a guide to action that ensures that the military will forever heed the Party commands
Reference 2	It is the guiding prinicple which guarantees the military forces always being under the command of the Party
Reference 3	It is the practical guide for the army always to heed the directions of the party

"clipped" unigram precision = 4 / 4 = 1.0

(every word except "obeys")

MT Evaluation
BLEU

- BLEU: Weighted n-gram precision with a length penalty

$$Bleu = BP \times exp\left(\frac{1}{N}\sum_{n=1}^{N} logp_n\right)$$

harmonic mean of unigram, bigram, trigram, 4-gram precisions

MT Evaluation

BLEU

- BLEU: Weighted n-gram precision with a length penalty

$$Bleu = BP \times exp\left(\frac{1}{N}\sum_{n=1}^{N} logp_n\right)$$

brevity penalty

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

MT Evaluation

- Automatic Evaluations:
  - NLP likes automatic evals for standardization and optimization
  - Most popular metric for MT is **BLEU**
  - Newer variations exist: e.g., ML models like ESIM and BLEURT
  - Controversial—once systems are sufficiently good, metrics stop correlating with human judgments

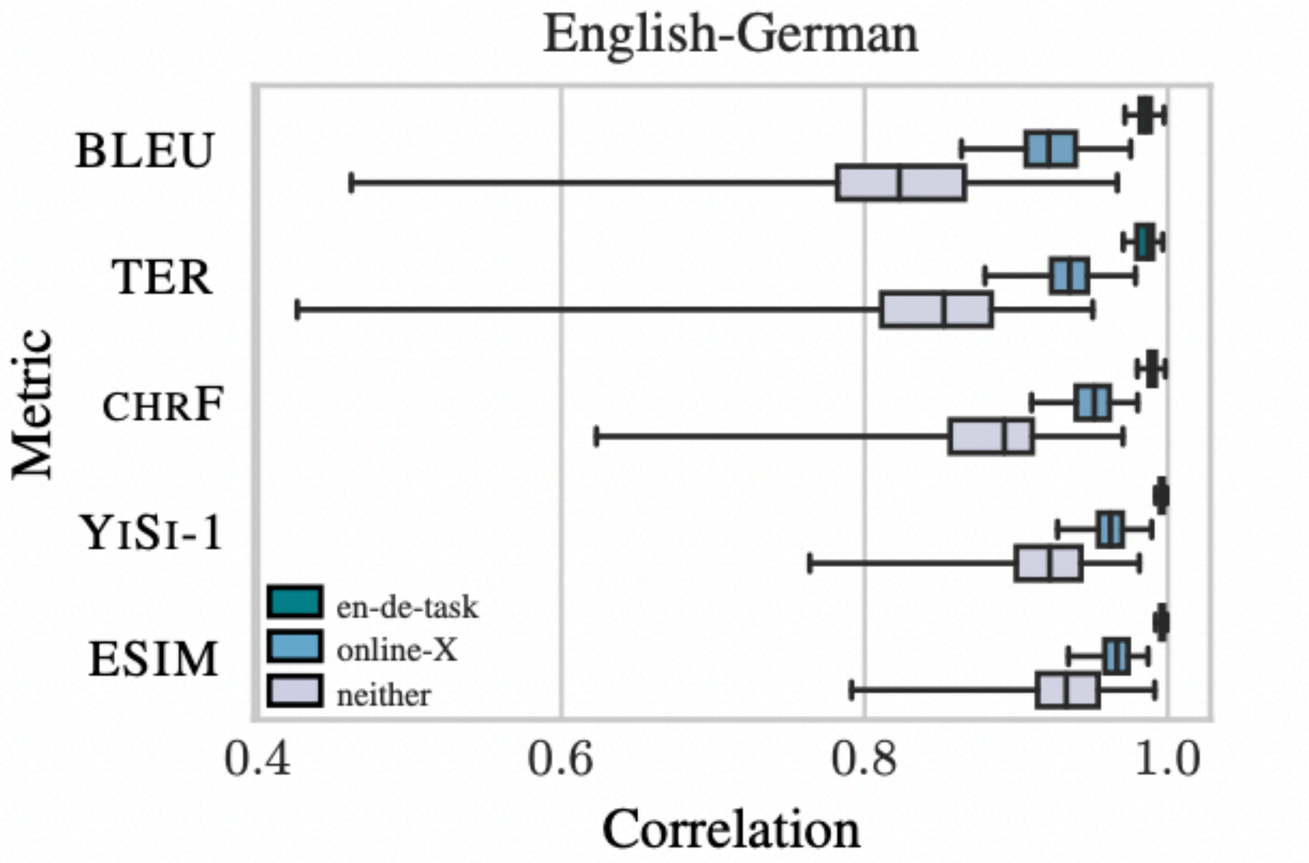
MT Evaluation

Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

Nitika Mathur **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia

nmathur@student.unimelb.edu.au {tbaldwin,tcohn}@unimelb.edu.au

MT Evaluation



Baseline metrics

- BLEU (Papineni et al., 2002b) is the precision of n-grams of the MT output compared to the reference, weighted by a brevity penalty to punish overly short translations. BLEU has high variance across different hyper-parameters and pre-processing strategies, in response to which sacreBLEU (Post, 2018) was introduced to create a standard implementation for all researchers to use; we use this version in our analysis.
- TER (Snover et al., 2006) measures the number of edits (insertions, deletions, shifts and substitutions) required to transform the MT output to the reference.
- CHRF (Popović, 2015) uses character n-grams instead of word n-grams to compare the MT output with the reference. This helps with matching morphological variants of words.

Best metrics across language pairs

- YISI-1 (Lo, 2019) computes the semantic similarity of phrases in the MT output with the reference, using contextual word embeddings (BERT: Devlin et al. (2019)).
- ESIM (Chen et al., 2017; Mathur et al., 2019) is a trained neural model that first computes sentence representations from BERT embeddings, then computes the similarity between the two strings. [3]

Source-based metric

- YISI-2 (Lo, 2019) is the same as YISI-1, except that it uses cross-lingual embeddings to compute the similarity of the MT output with the source.

Topics

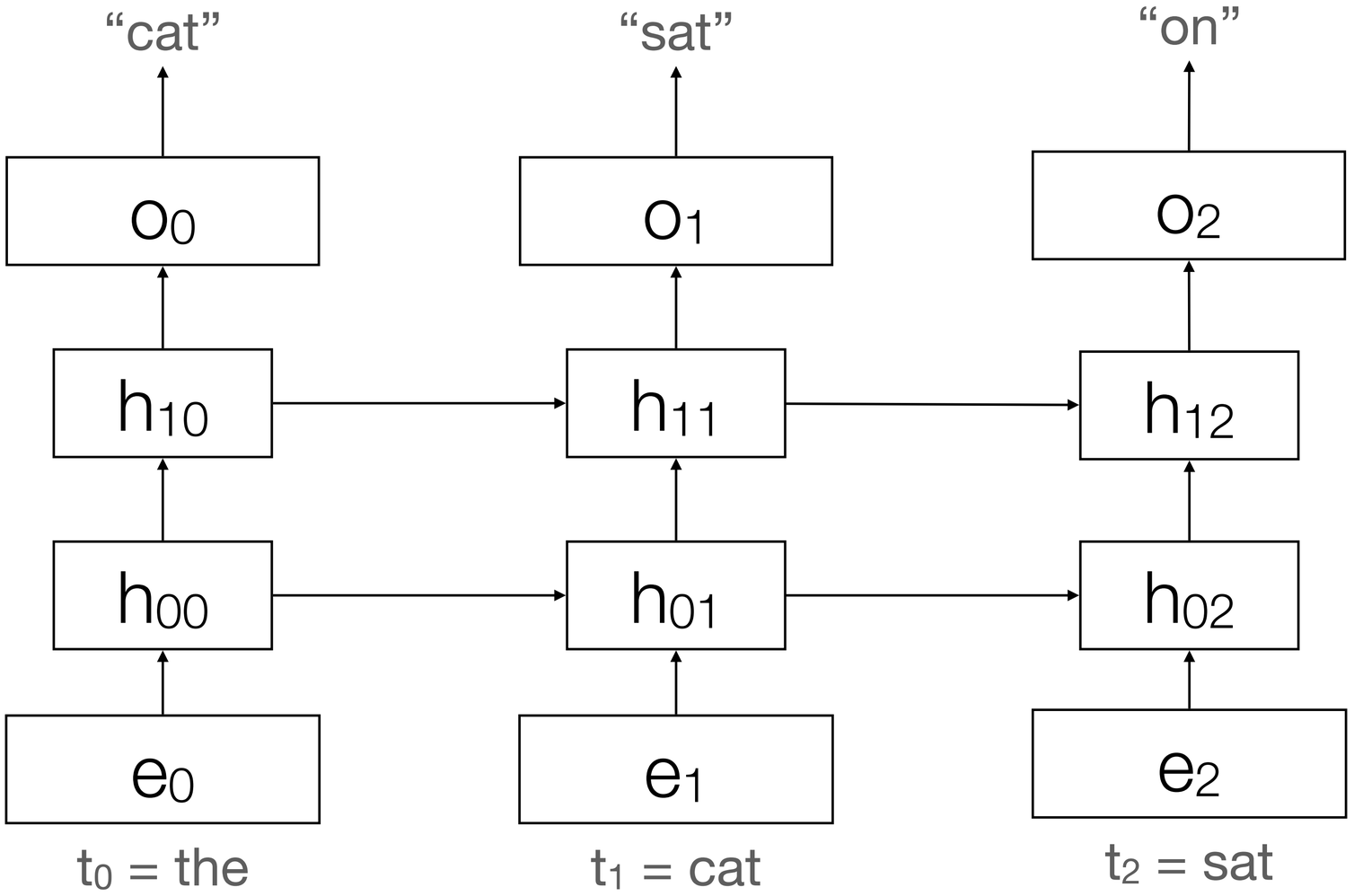
- SMT Followup: Putting it all together
- MT Evaluation
- **Neural MT**
  - **Encoder-Decoder Models**
  - Multilingual LMs and “Zero Shot” Cross-Lingual Transfer

Encoder-Decoder Models

- AKA “sequence to sequence” or seq2seq
- Intuition: “Conditional” text generation/language modeling
- Like language modeling, but output is dependent on some input, in this case, the source sentence
- seq2seqs were originally designed for MT, but have since been applied to many other tasks
  - E.g., summarization, parsing, etc…

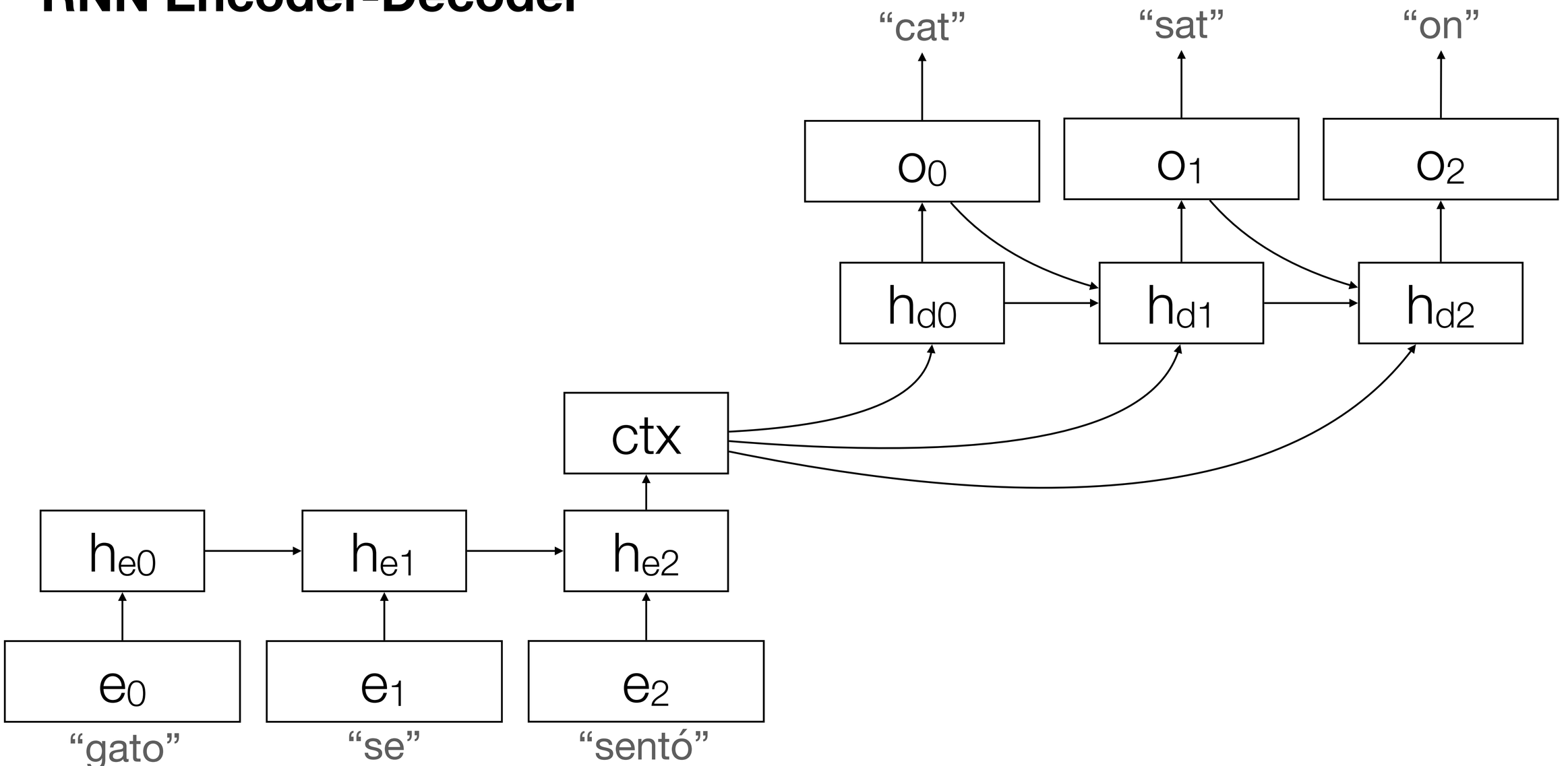
Encoder-Decoder Models

Basic RNN



Encoder-Decoder Models

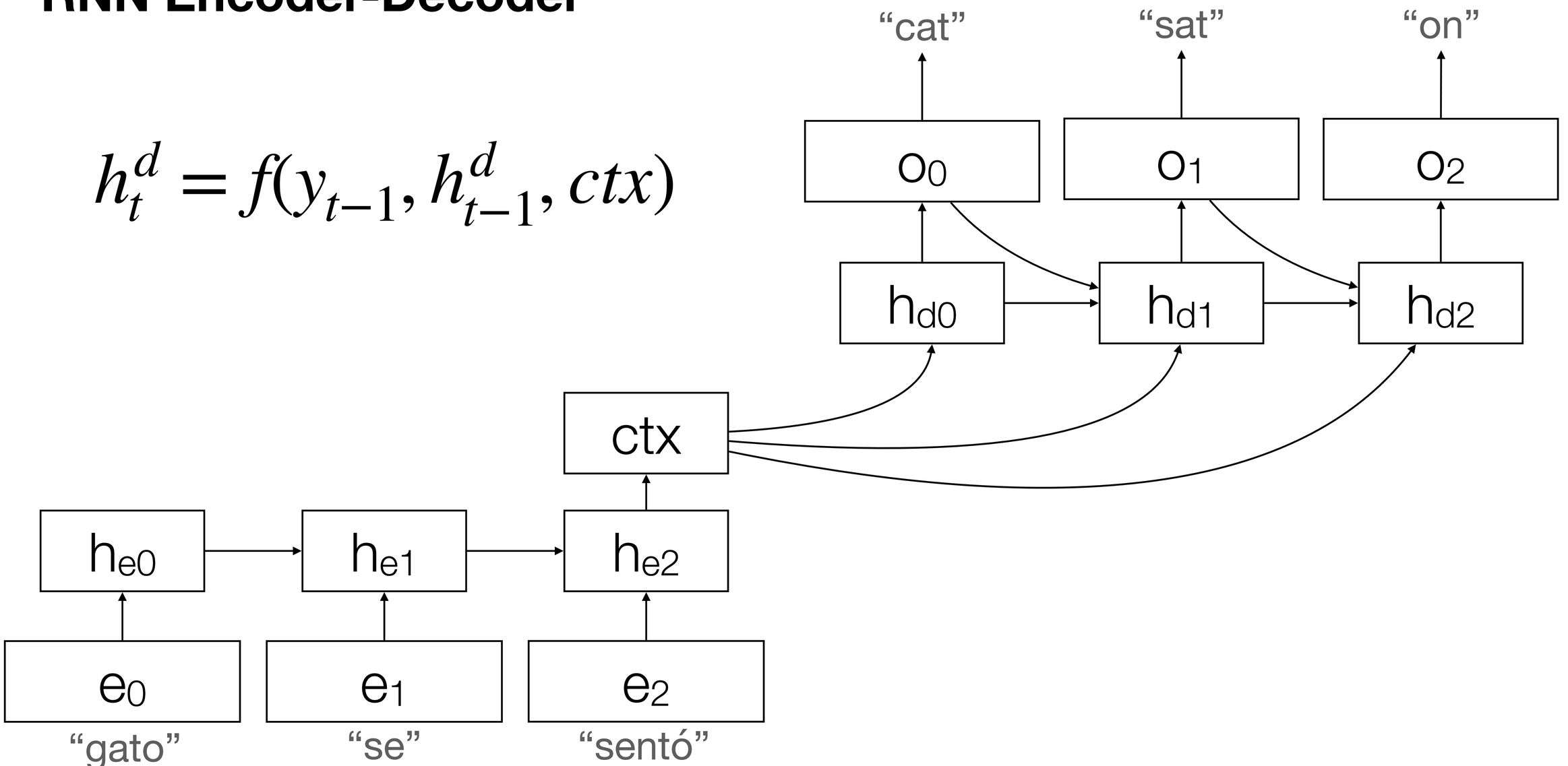
RNN Encoder-Decoder



Encoder-Decoder Models

RNN Encoder-Decoder

$$h_t^d = f(y_{t-1}, h_{t-1}^d, ctx)$$

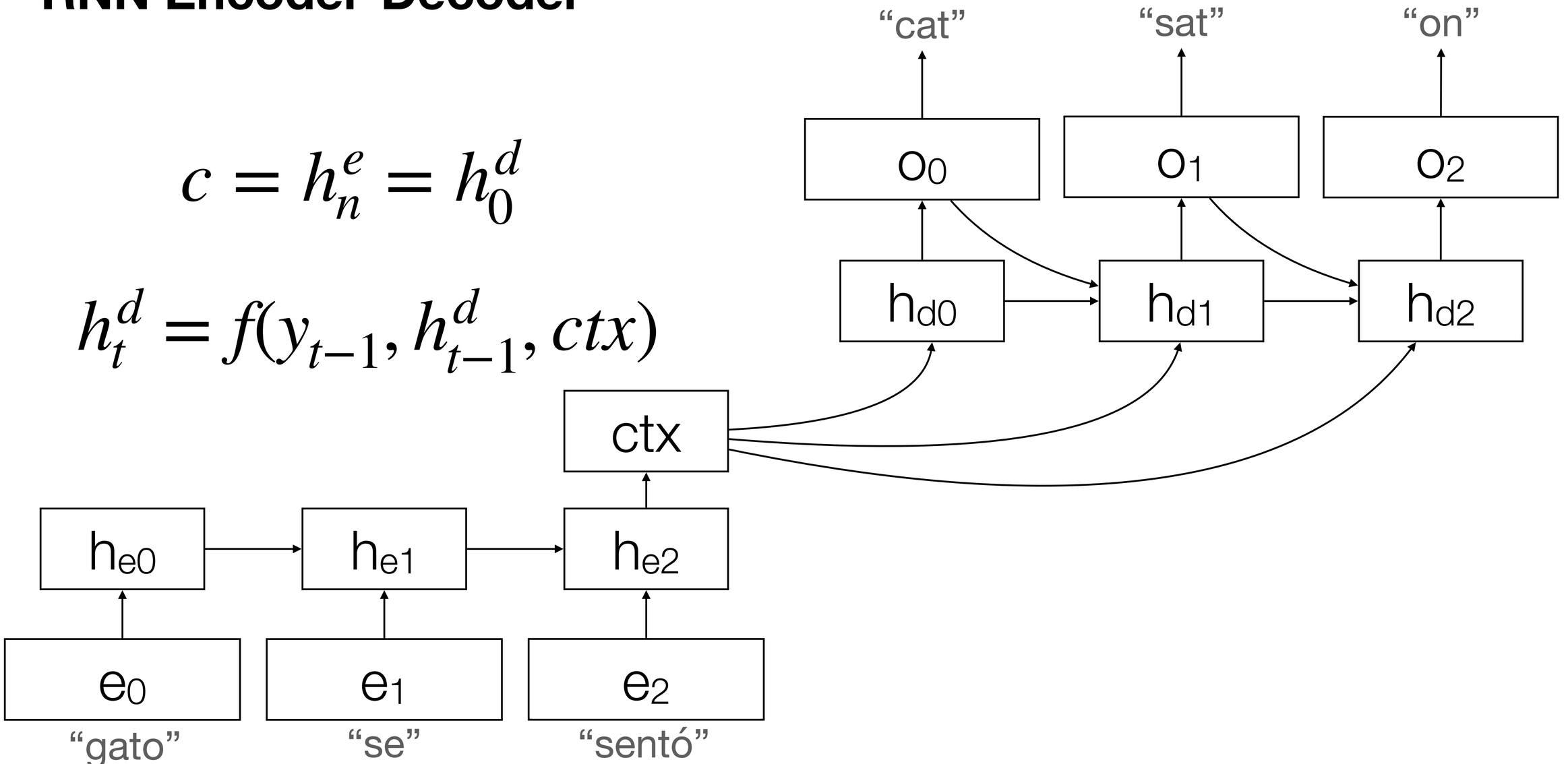


Encoder-Decoder Models

RNN Encoder-Decoder

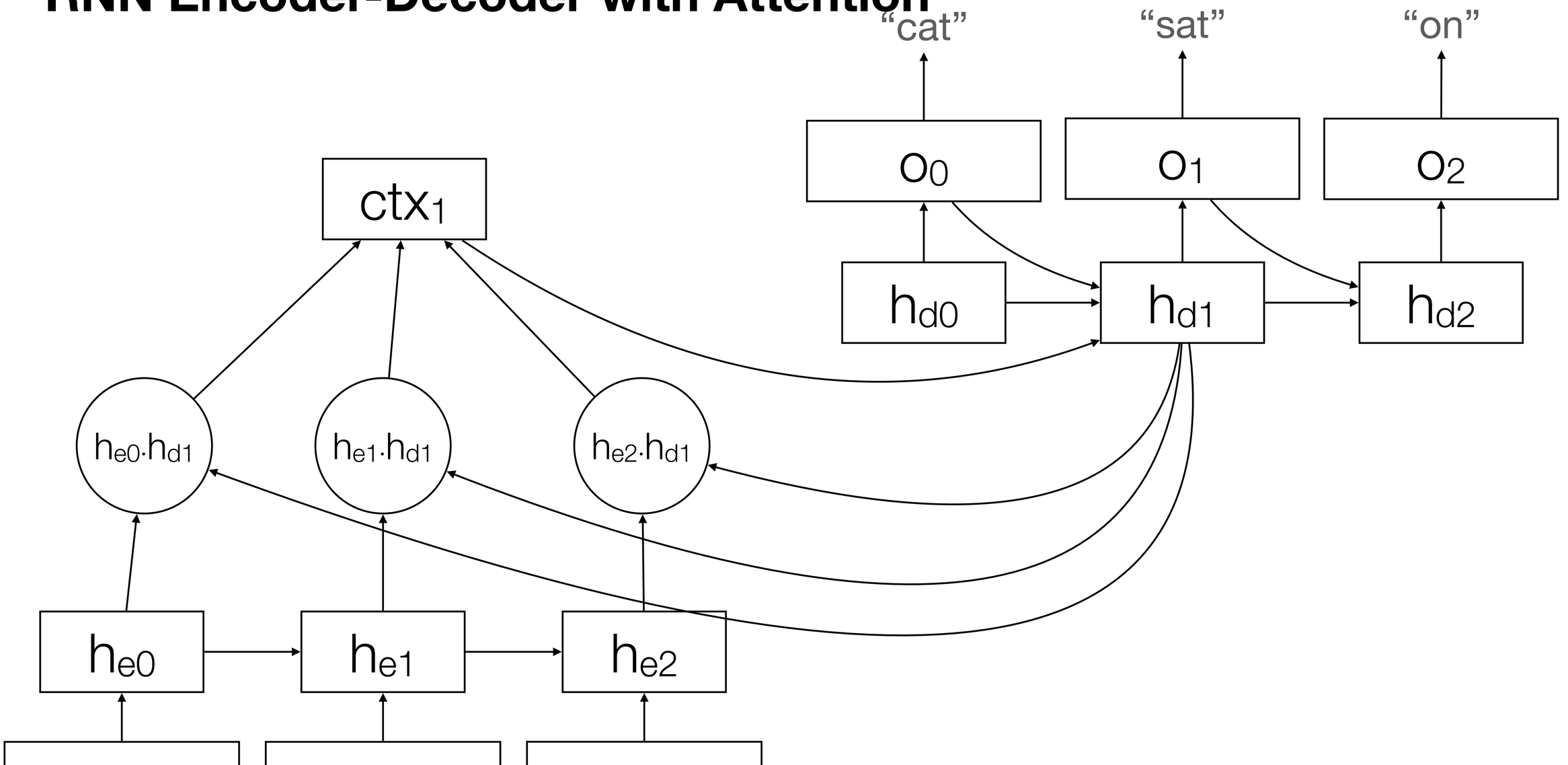
$$c = h_n^e = h_0^d$$

$$h_t^d = f(y_{t-1}, h_{t-1}^d, ctx)$$



Encoder-Decoder Models

RNN Encoder-Decoder with Attention



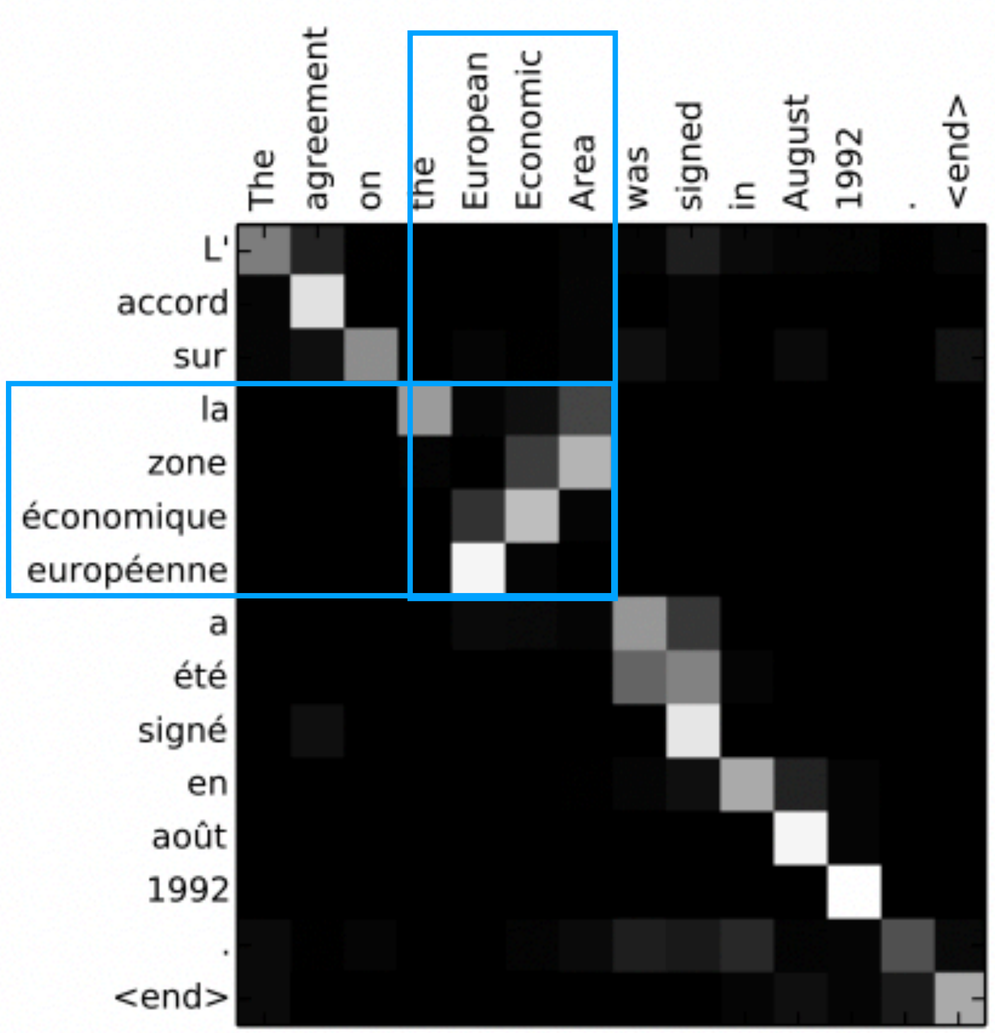
Encoder-Decoder Models

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

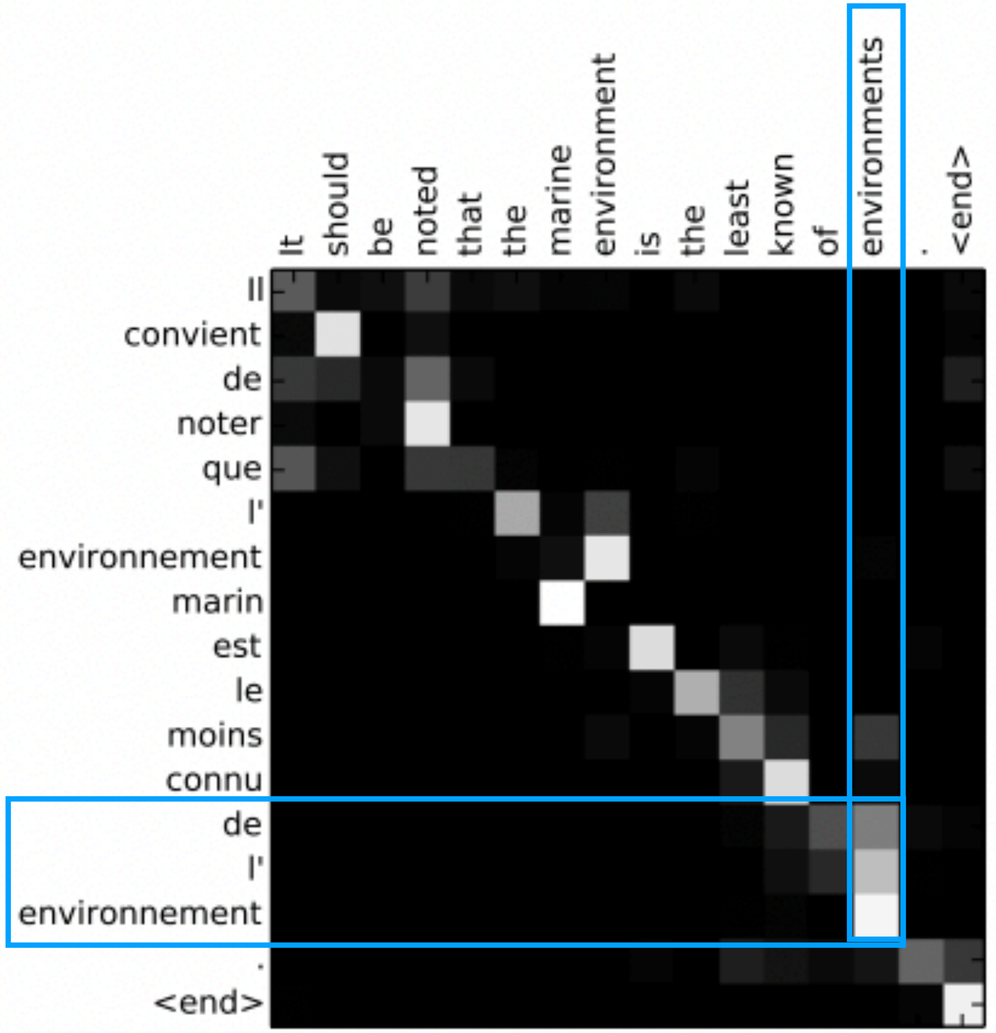
Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

Encoder-Decoder Models

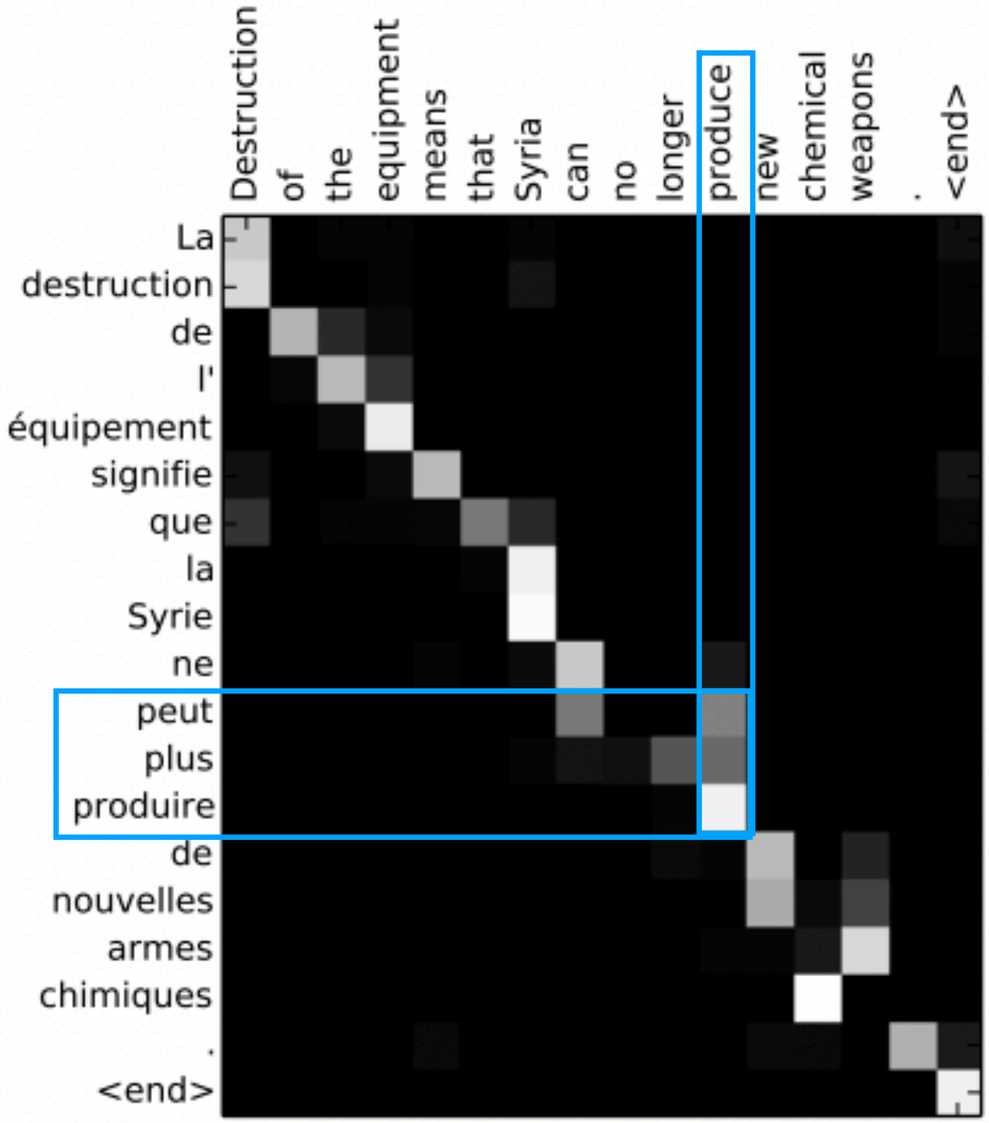


(a)

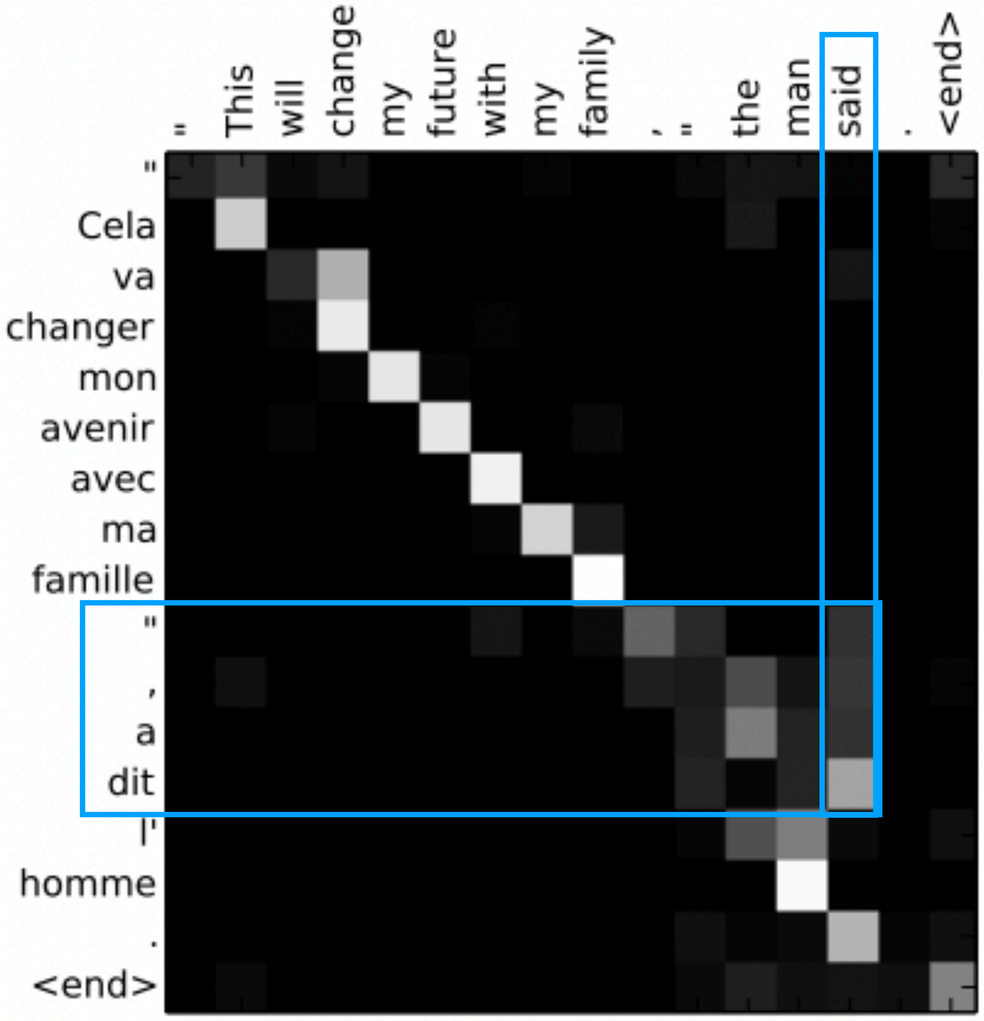


(b)

Encoder-Decoder Models



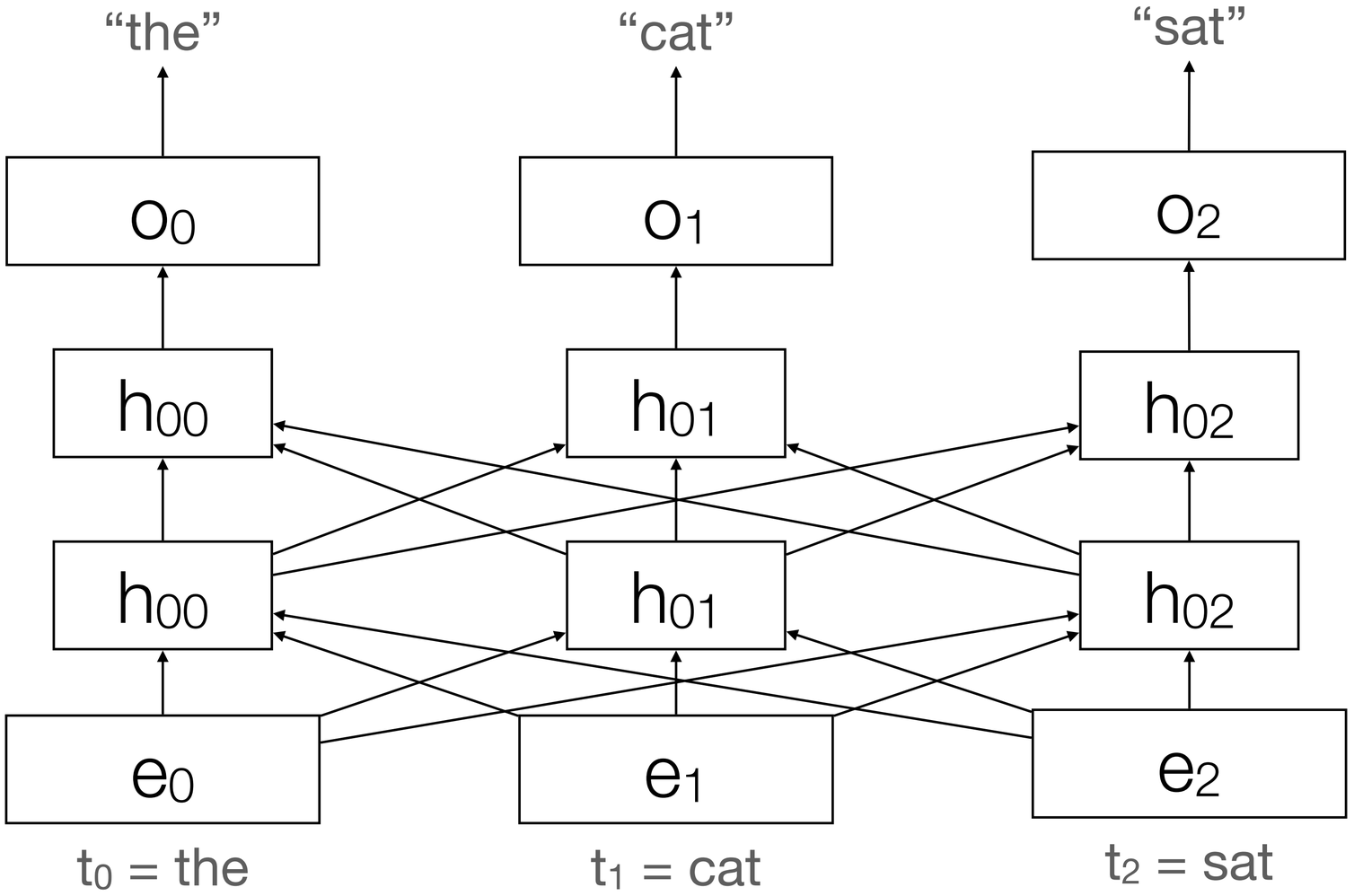
(c)



(d)

Encoder-Decoder Models

Transformer Encoder-Decoder



Encoder-Decoder Models

Recap: Transformer Blocks

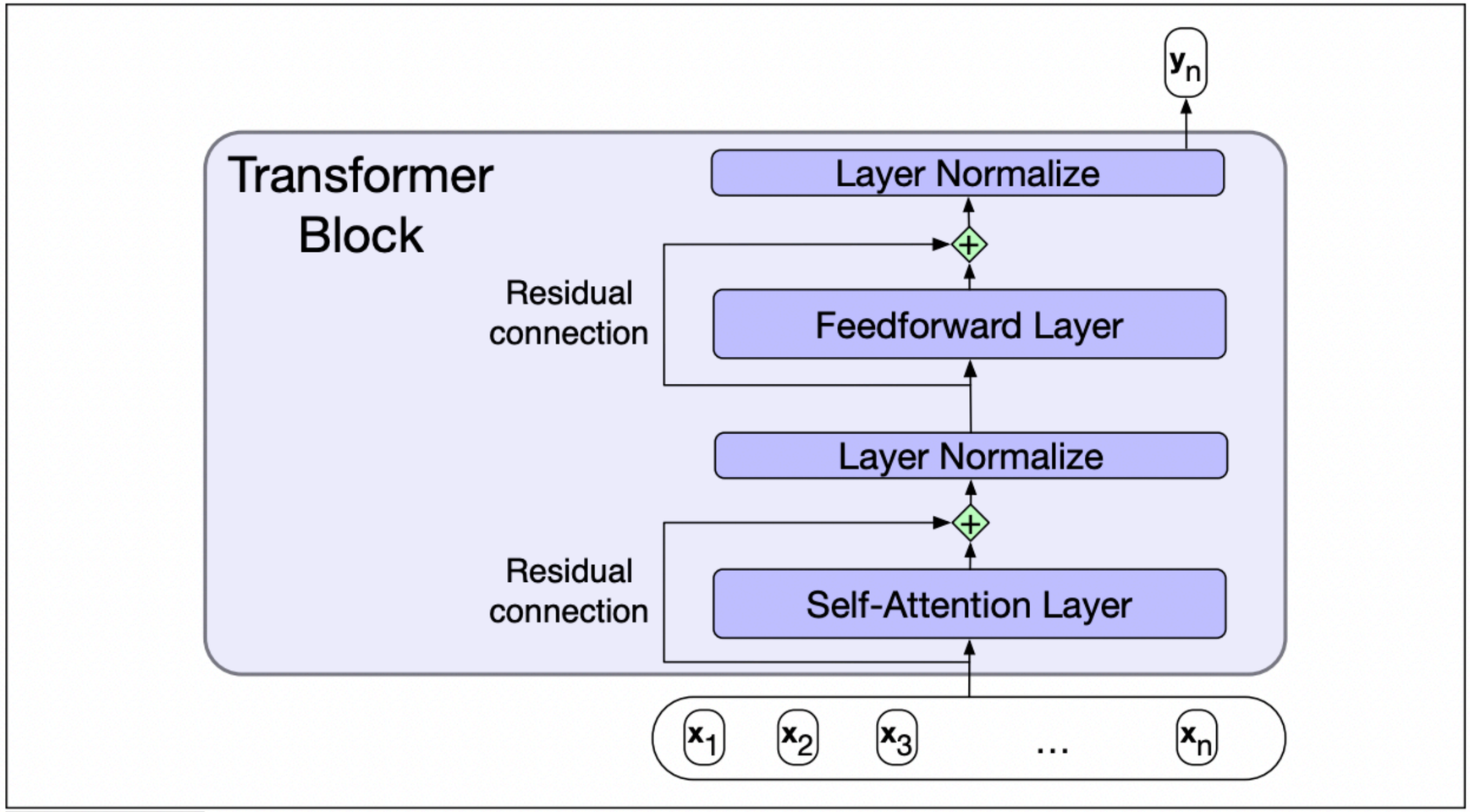
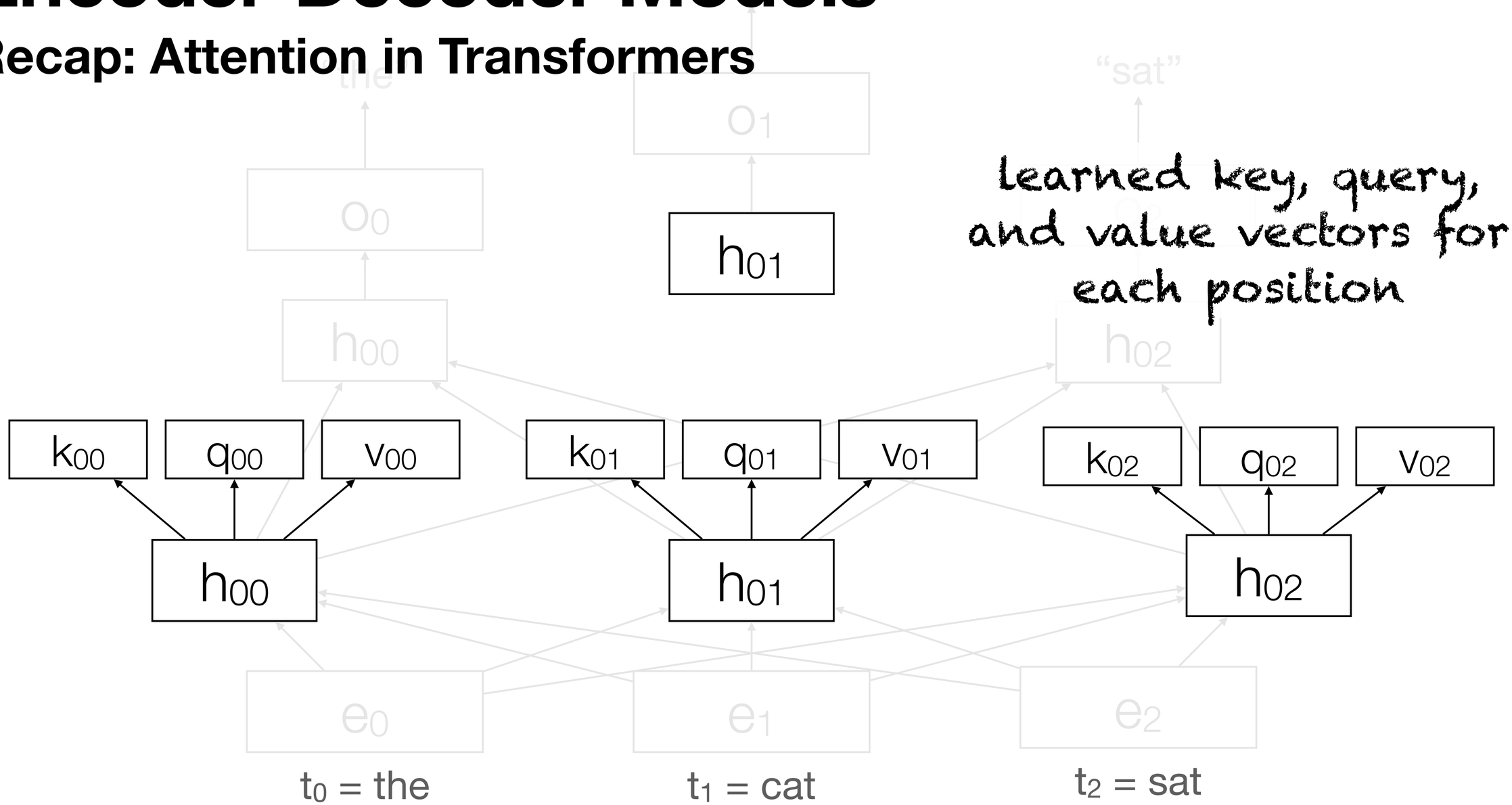


Figure 9.18 A transformer block showing all the layers.

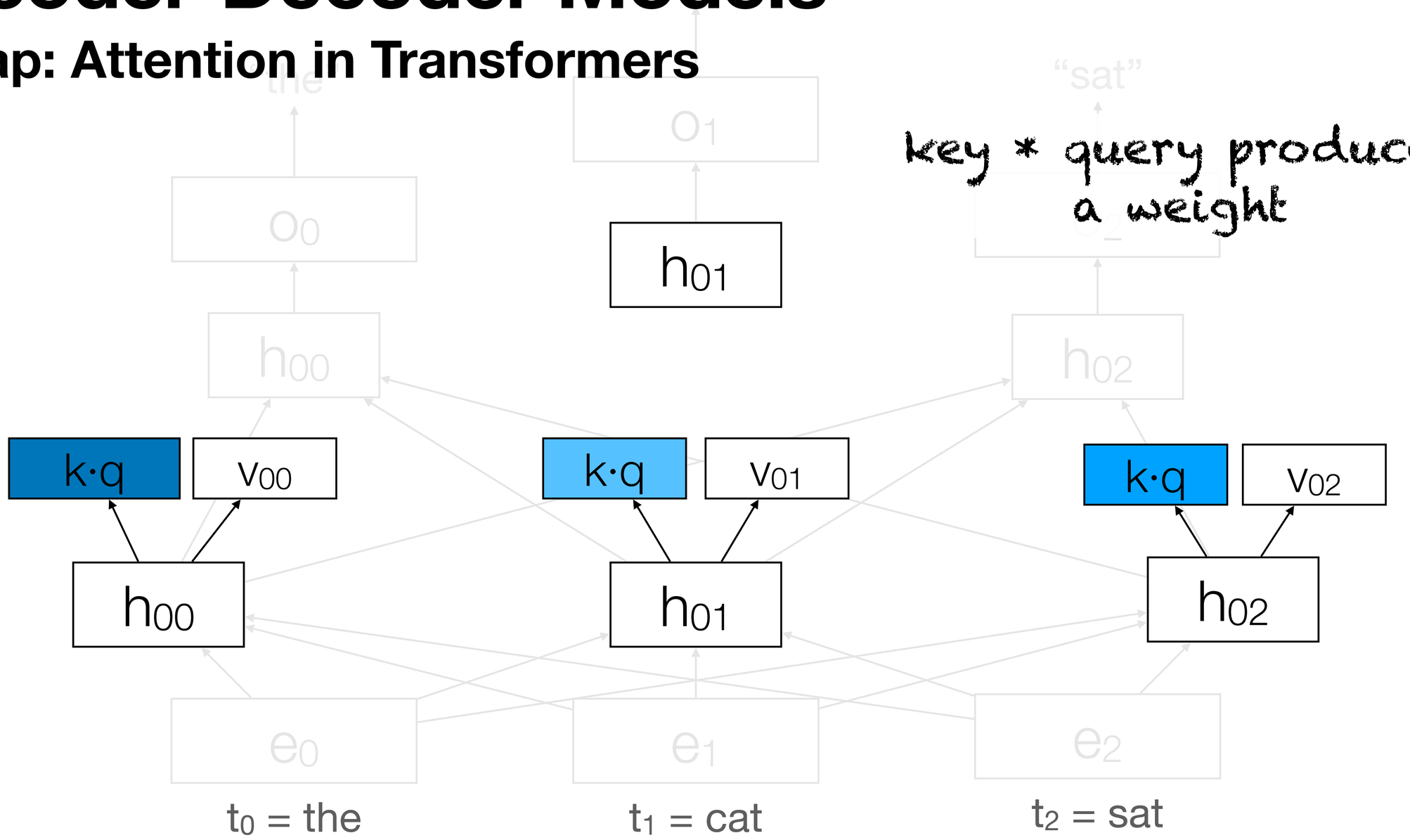
Encoder-Decoder Models

Recap: Attention in Transformers



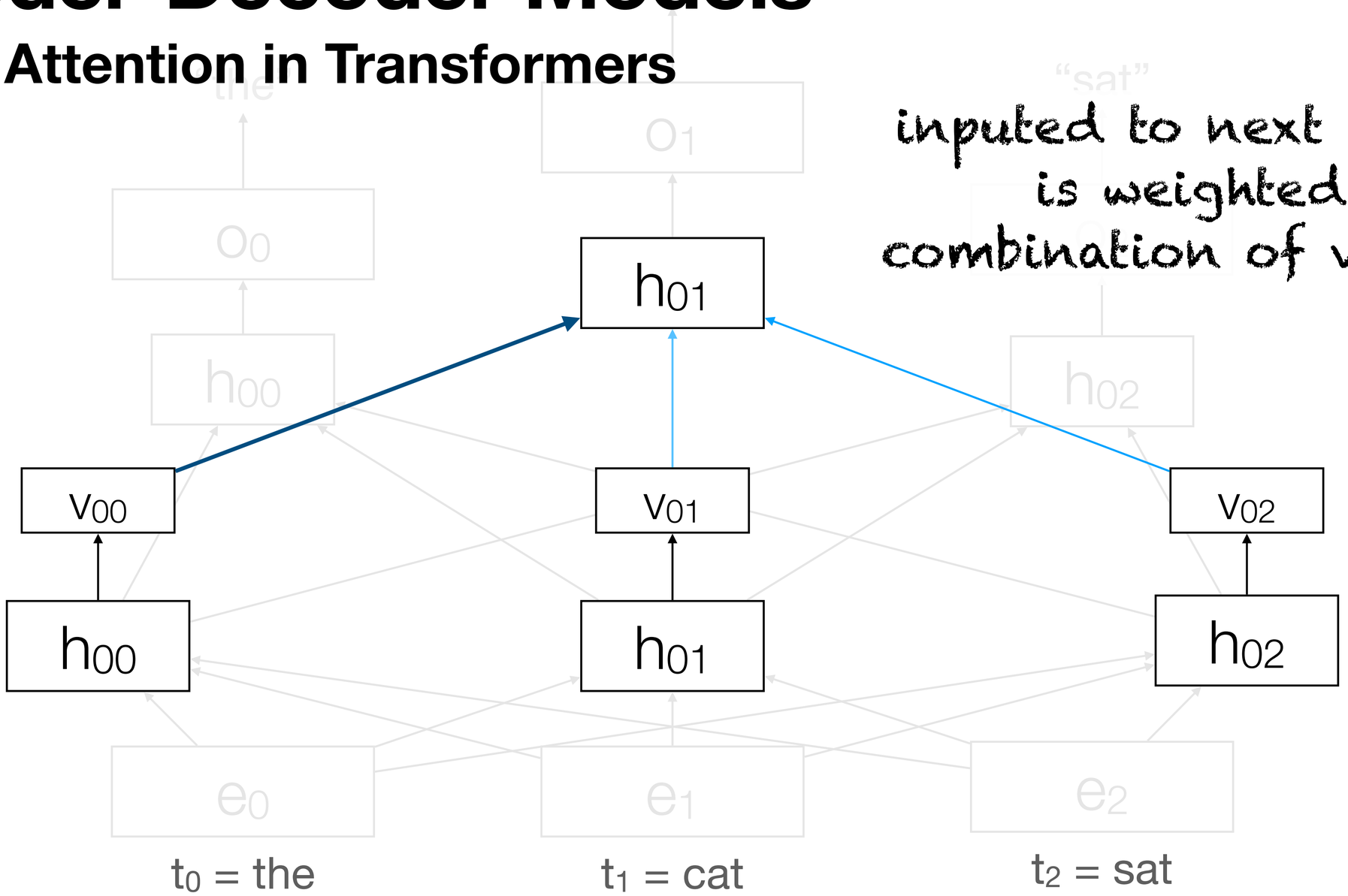
Encoder-Decoder Models

Recap: Attention in Transformers



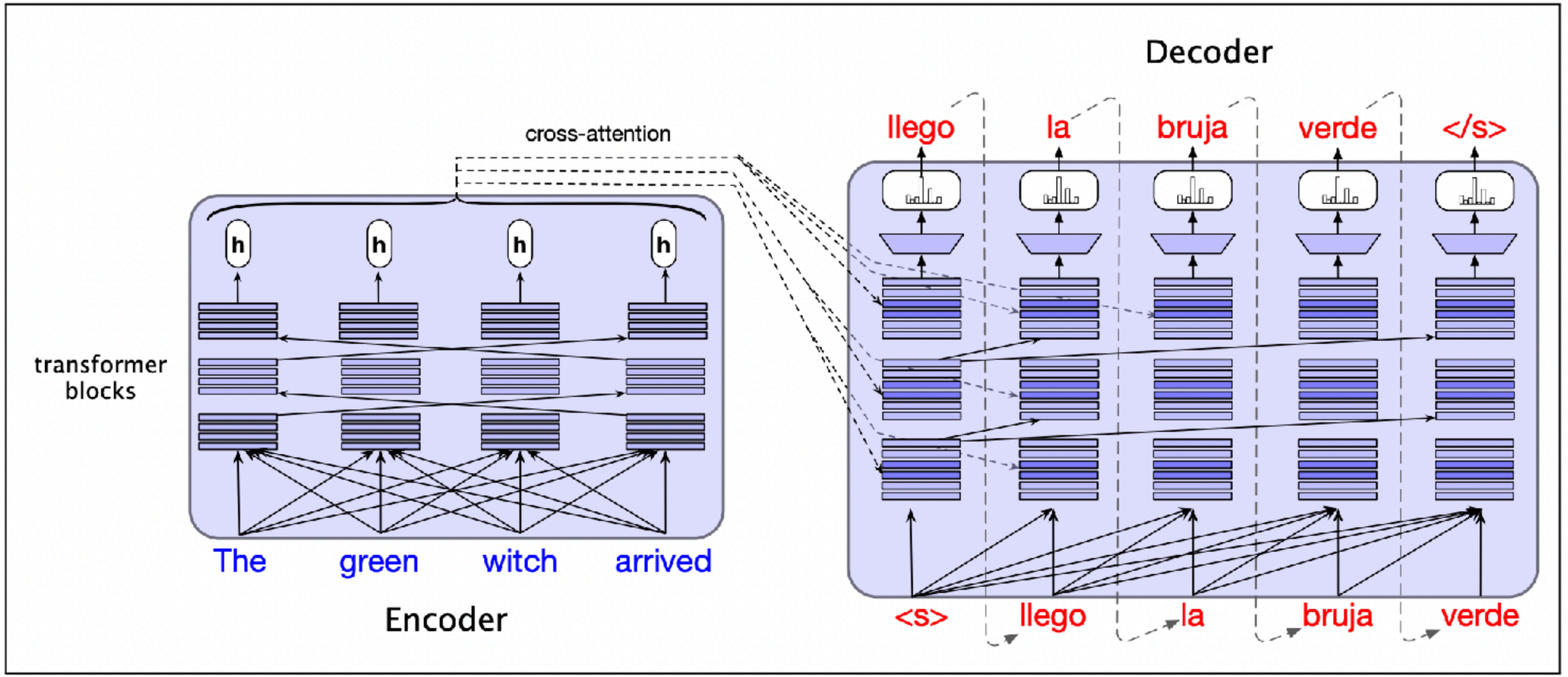
Encoder-Decoder Models

Recap: Attention in Transformers



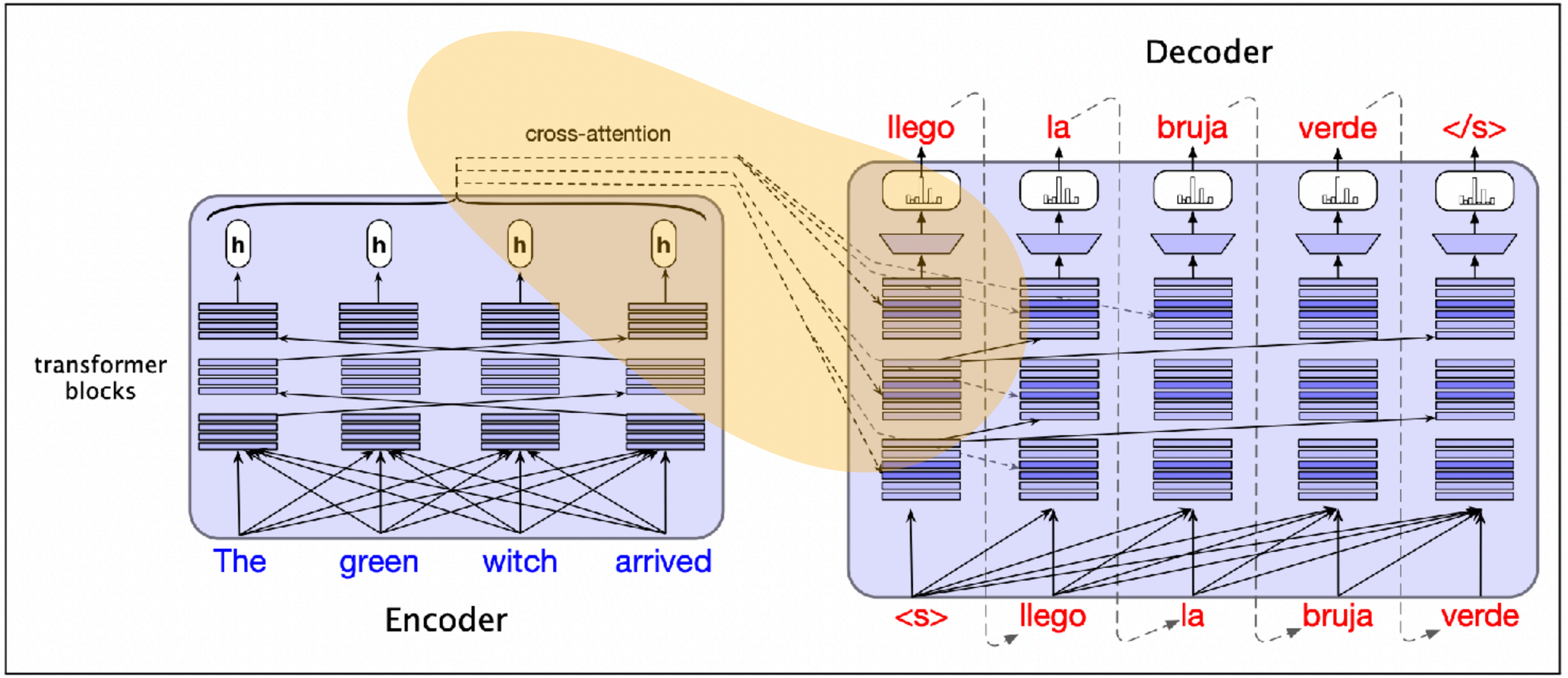
Encoder-Decoder Models

Transformer Encoder-Decoder



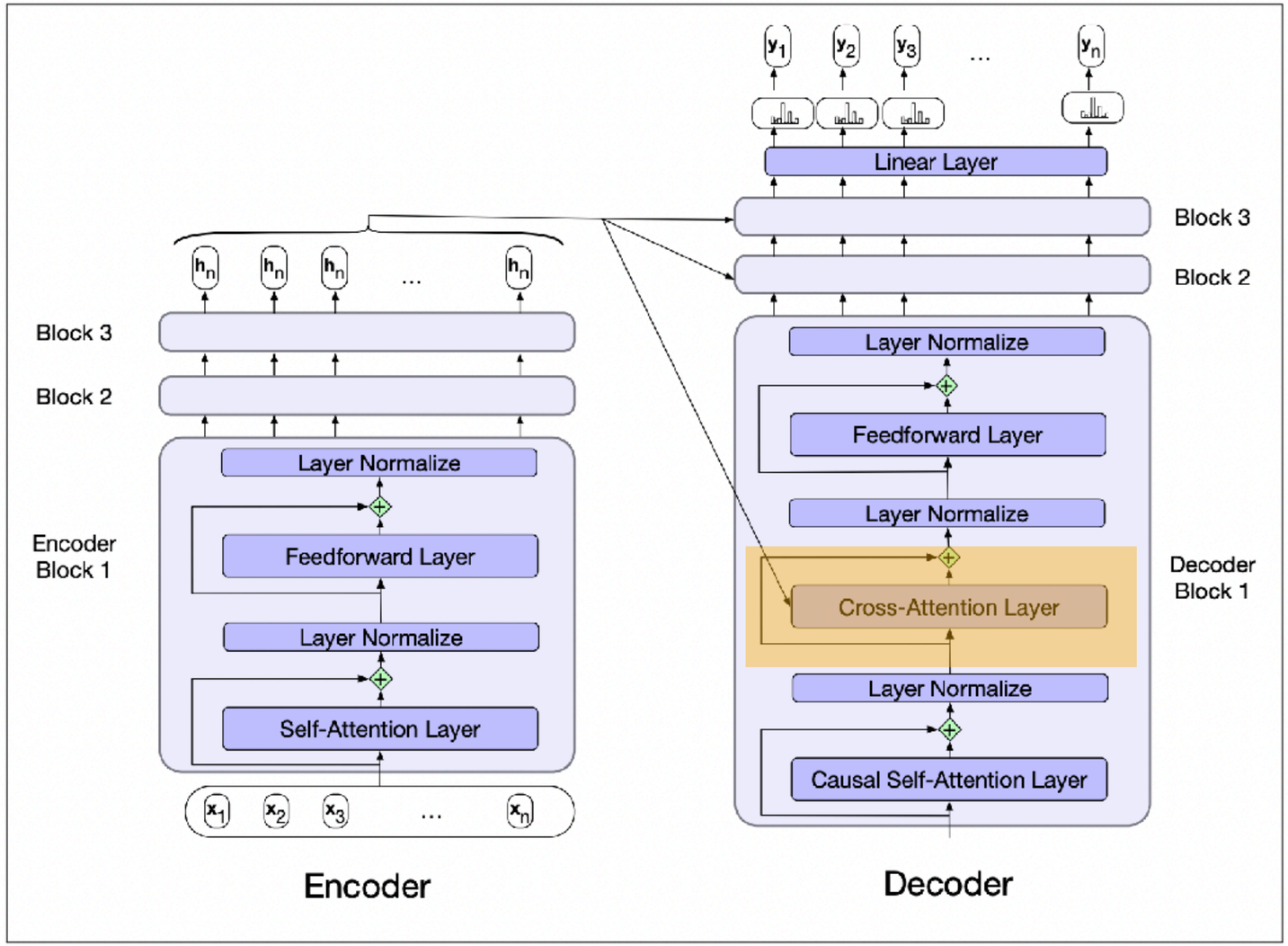
Encoder-Decoder Models

Transformer Encoder-Decoder



Encoder-Decoder Models

Transformer Encoder-Decoder

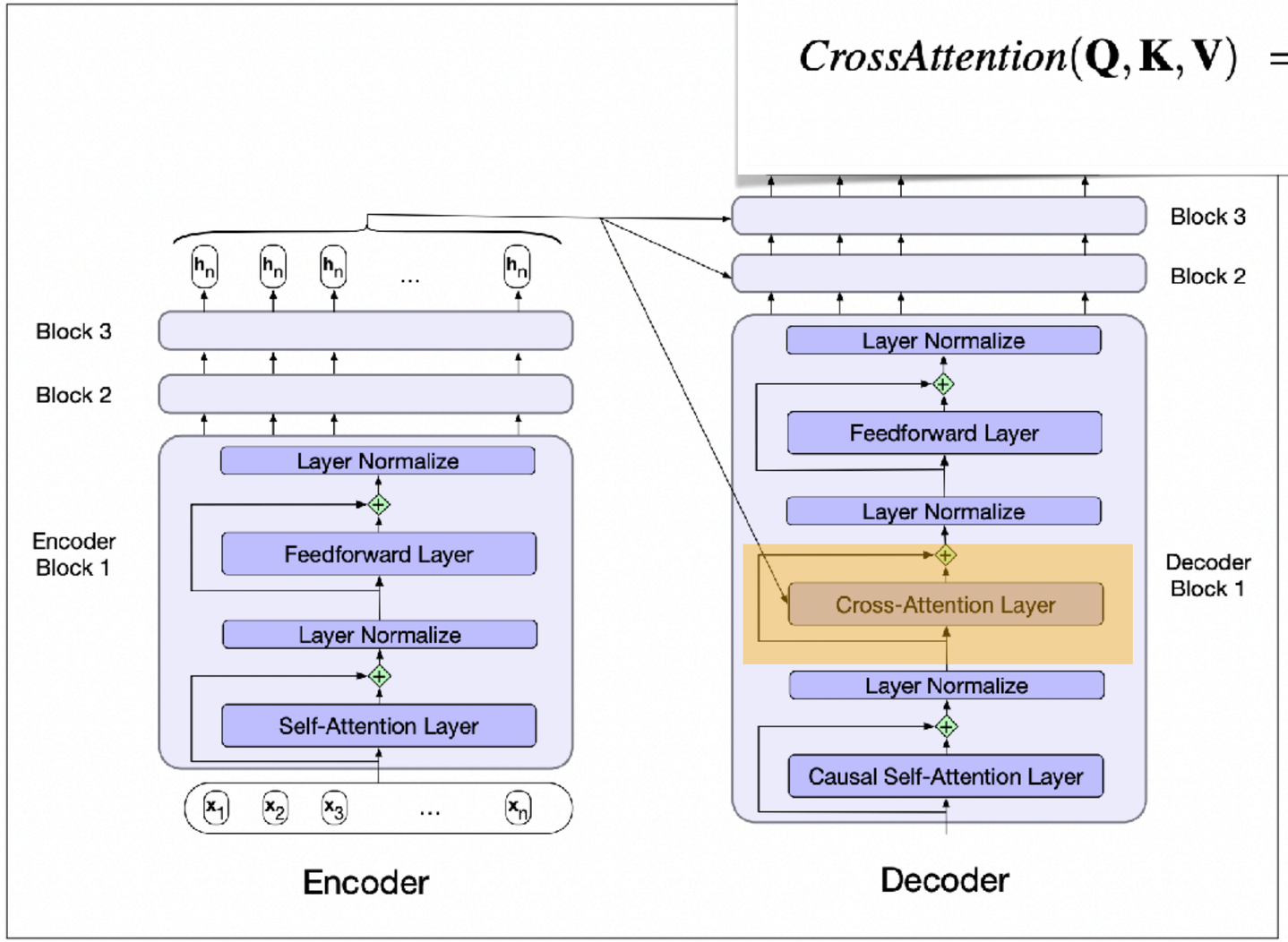


Encoder-Decoder Models

Transformer Encoder-Decoder

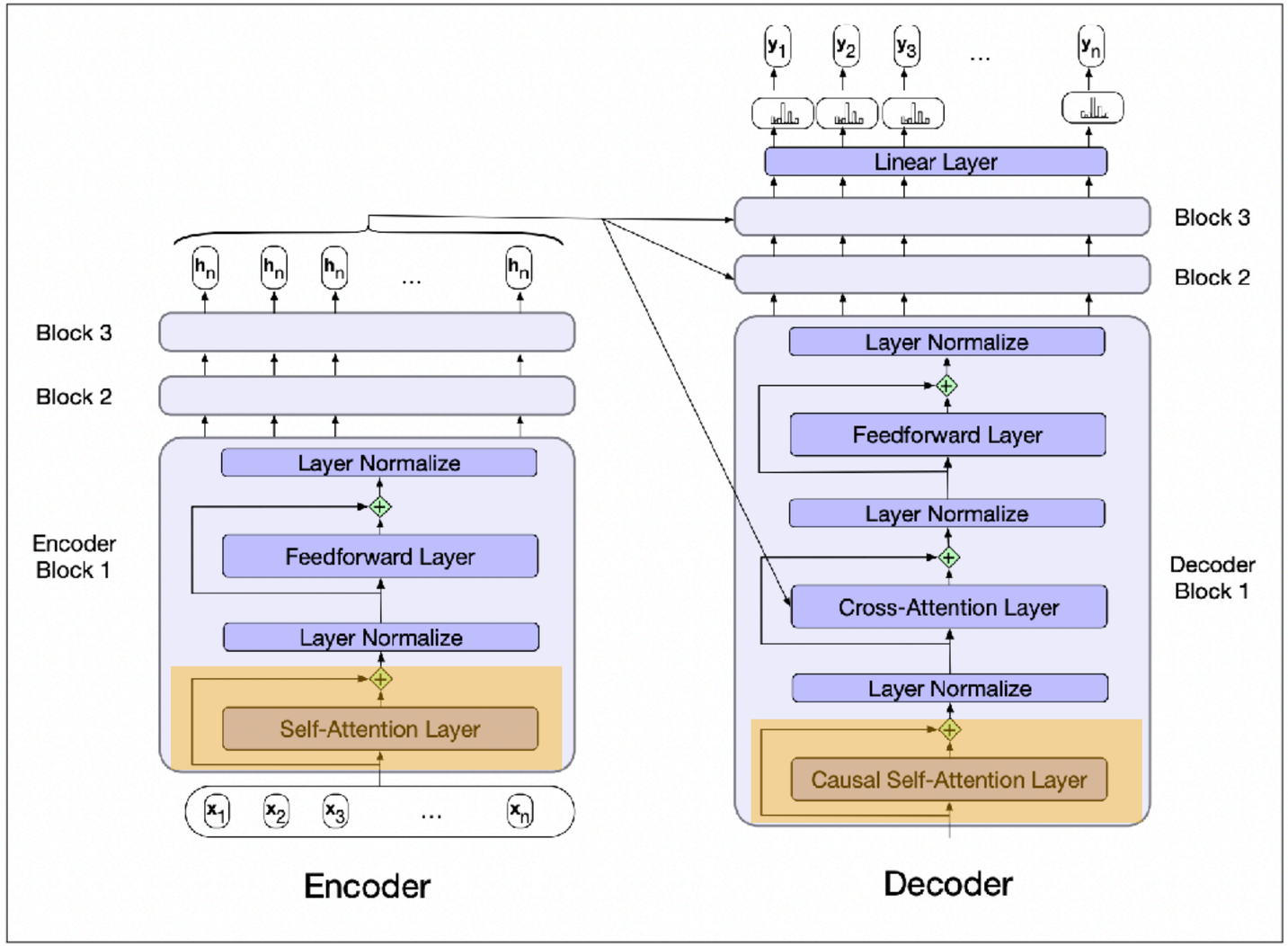
$$\mathbf{Q} = \mathbf{W}^{\mathbf{Q}}\mathbf{H}^{dec[i-1]};\ \mathbf{K} = \mathbf{W}^{\mathbf{K}}\mathbf{H}^{enc};\ \mathbf{V} = \mathbf{W}^{\mathbf{V}}\mathbf{H}^{enc}$$

$$CrossAttention(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d_k}}\right)\mathbf{V}$$



Encoder-Decoder Models

Transformer Encoder-Decoder



Encoders? Decoders? Encoder-Decoders?

- The encoder-decoder model was first applied to MT and was one of the early big “successes” of neural language generation
- Many other models (namely, large LMs) are inspired by this architecture, but don’t “need” both the encoder and the decoder
  - Encoder-Decoder: Original Transformer Model (Vaswani et al, 2017)
  - Encoder-only: BERT and variants (ALBERT, DistilBERT, RoBERTa)
  - Decoder-only (i.e., auto-regressive): GPT

Topics

- SMT Followup: Putting it all together
- MT Evaluation
- **Neural MT**
  - Encoder-Decoder Models
  - **Multilingual LMs and “Zero Shot” Cross-Lingual Transfer**

Cross Lingual Transfer

- Goal: Train only on one language (English), but work in all languages
- Intuition: If the model learns a good representation of the languages under the hood (“interlingua”), it should be able to map the training it receives in one language to any other language

Multilingual Language Models

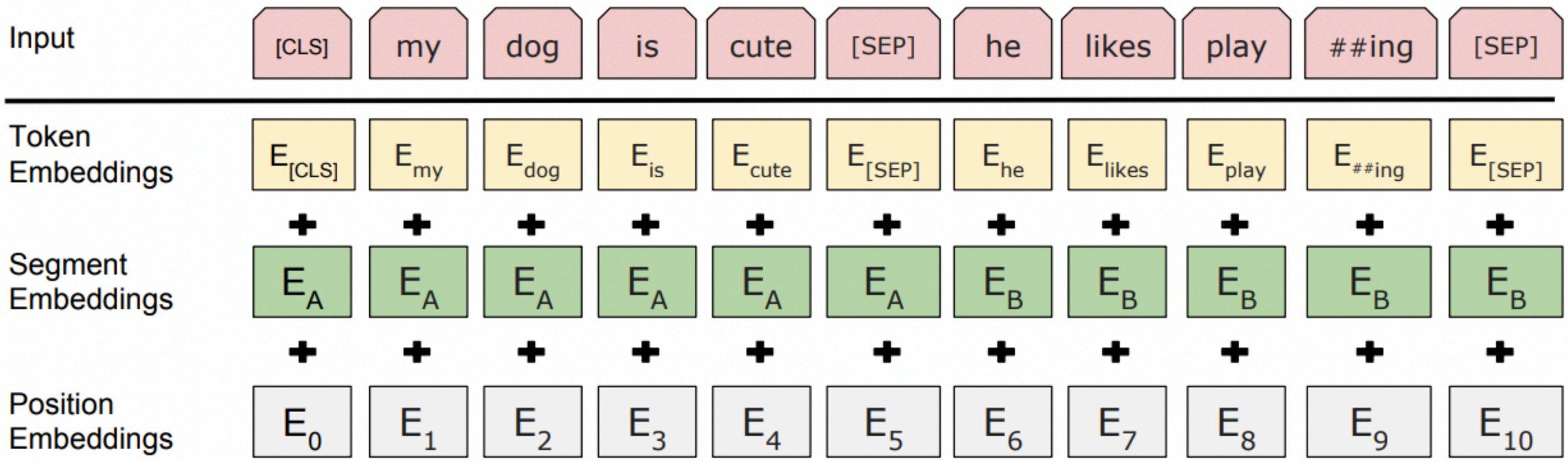
Cross-lingual Language Model Pretraining

Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau*
Facebook AI Research
Université Le Mans
aconneau@fb.com

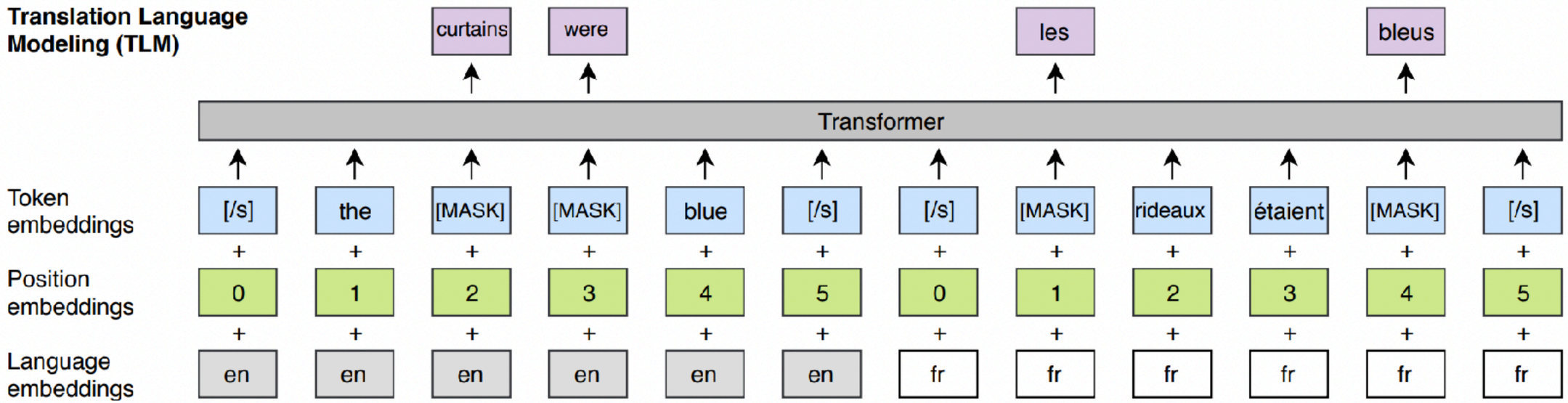
Multilingual Language Models

Recap: BERT



Multilingual Language Models

Multilingual MLM



Multilingual Language Models

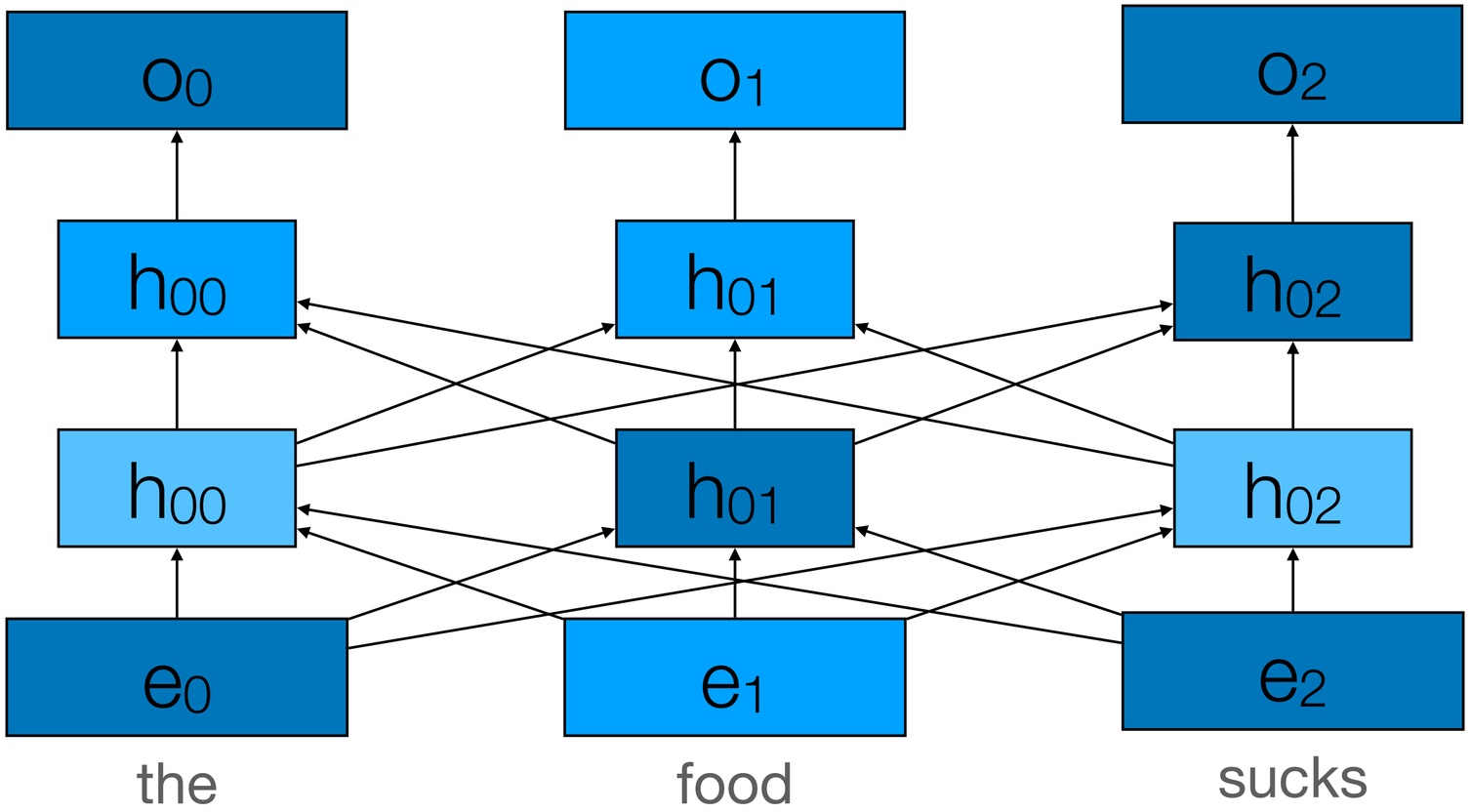
- Common multilingual models: mBERT, XLM-RoBERTa
- Just recently: mGPT, BLOOM
- Just monolingual data
- Shared wordpiece vocabulary
- No explicit mechanisms to encourage translation etc.

https://github.com/google-research/bert/blob/master/multilingual.md

Multilingual Language Models

Recap: Pretraining and Finetuning

Task: Sentiment Analysis
Input: “the food sucks”
Expected: Negative

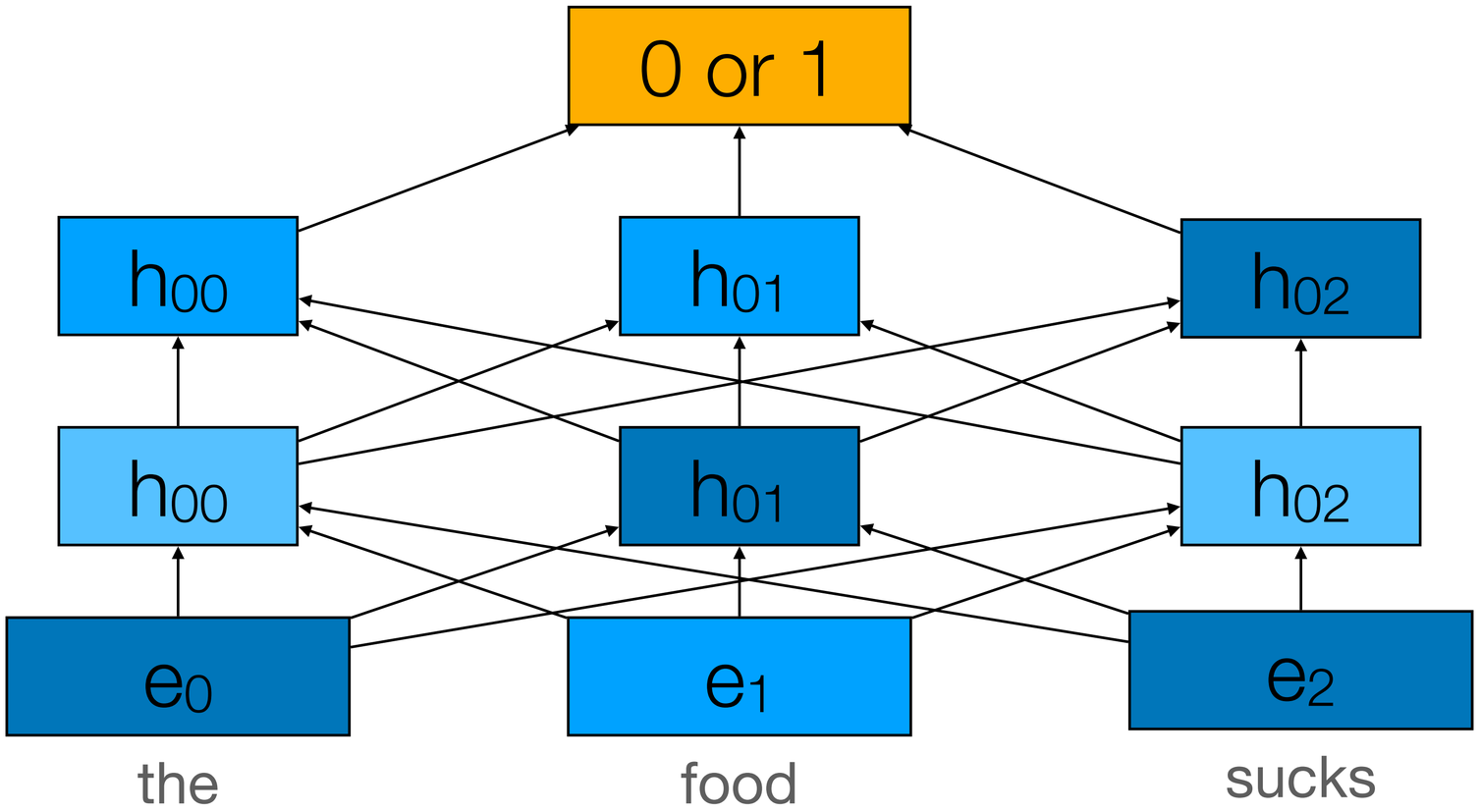


Multilingual Language Models

Recap: Pretraining and Finetuning

Task: Sentiment Analysis
Input: “the food sucks”
Expected: Negative

Adjust output layers to reflect target task

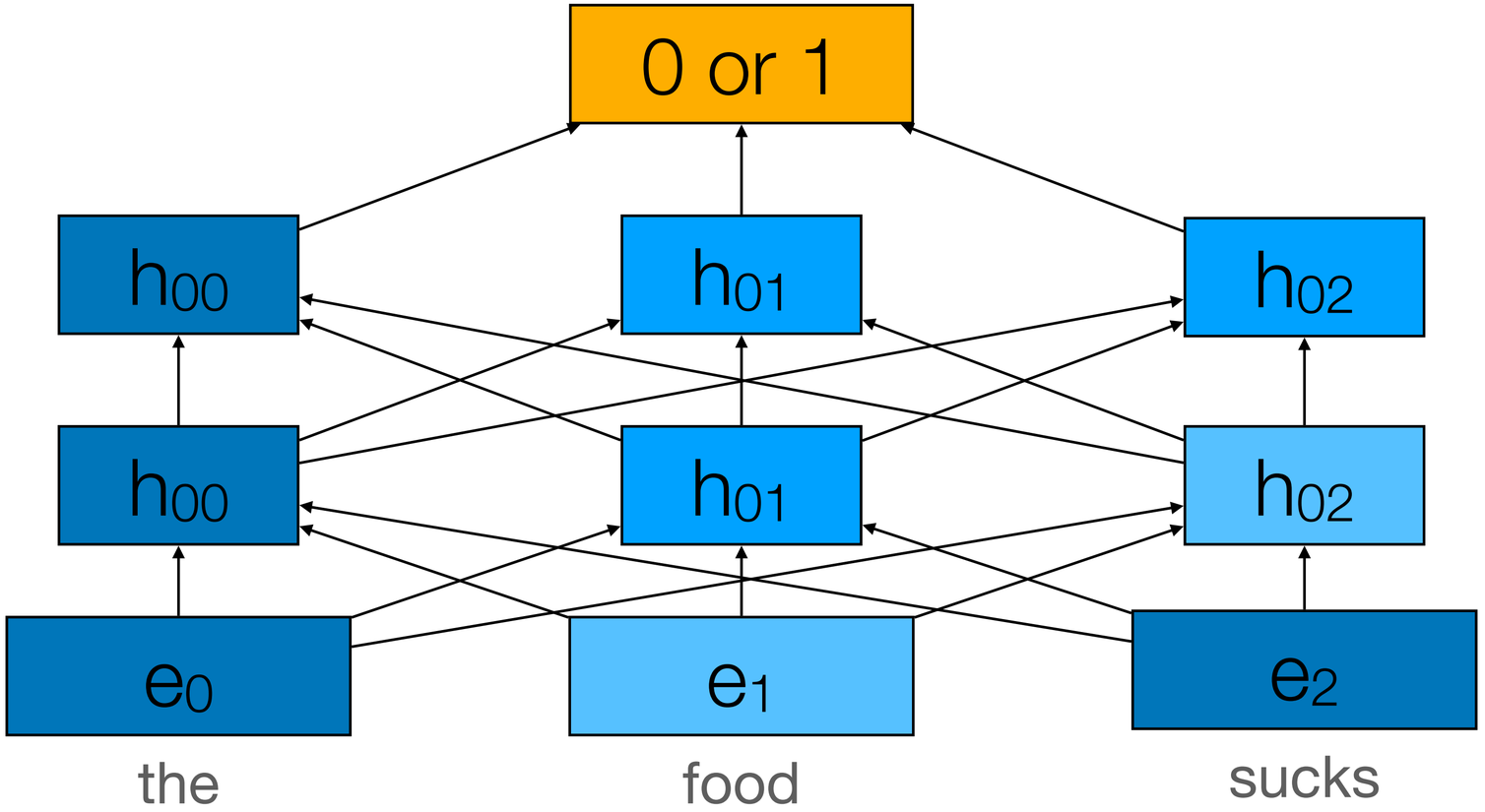


Multilingual Language Models

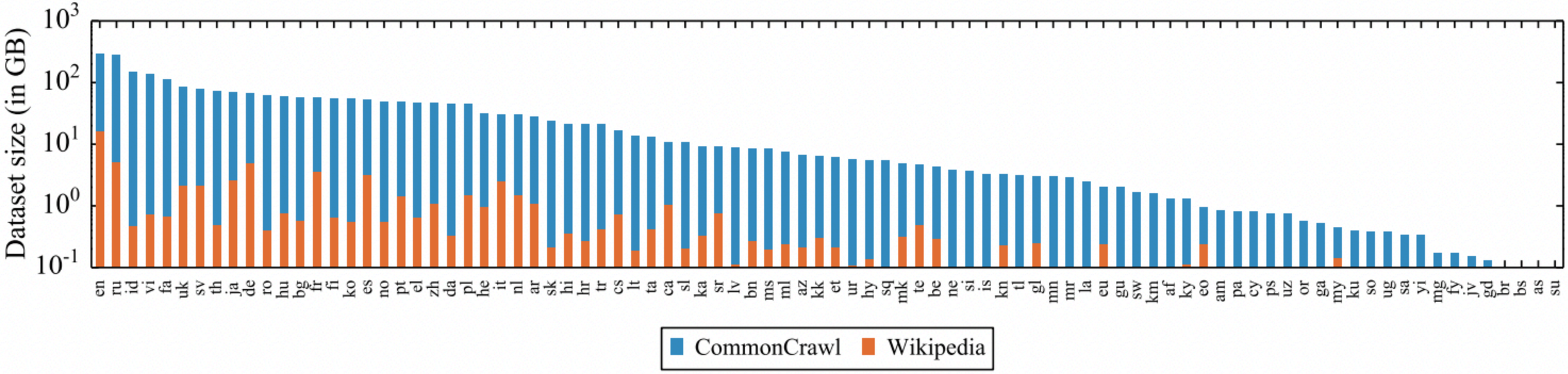
Recap: Pretraining and Finetuning

Task: Sentiment Analysis
Input: “the food sucks”
Expected: Negative

Keep training network, backproping through everything as needed.



Multilingual Language Models



Multilingual Language Models

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

Multilingual Language Models

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

Multilingual Language Models

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	Wiki+CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
Lample and Conneau (2019)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
Lample and Conneau (2019)†	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Huang et al. (2019)	Wiki+MT	1	15	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
Lample and Conneau (2019)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R$_{\text{Base}}$	CC	1	100	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
XLM-R	CC	1	100	**89.1**	**85.1**	**86.6**	**85.7**	**85.3**	**85.9**	**83.5**	**83.2**	**83.1**	**83.7**	**81.5**	**83.7**	**81.6**	**78.0**	**78.1**	**83.6**

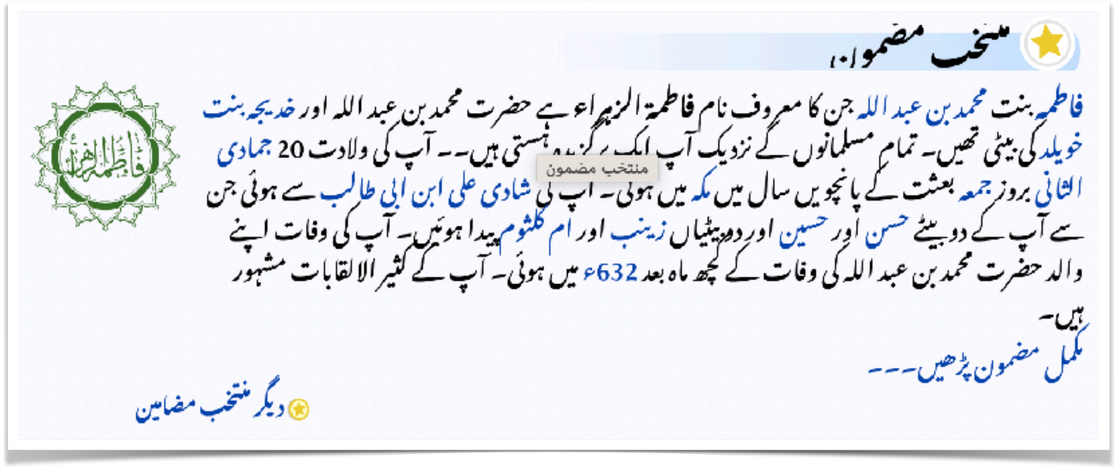
Cross Lingual Transfer

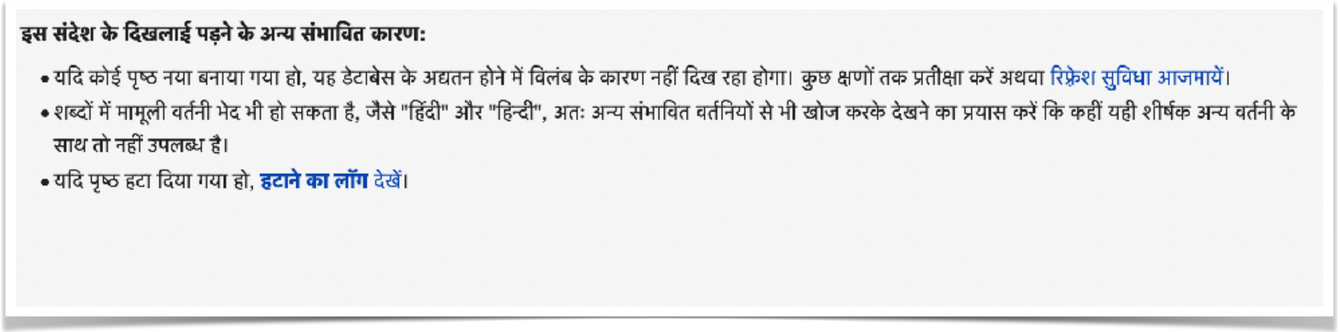
How multilingual is Multilingual BERT?

Telmo Pires* **Eva Schlinger** **Dan Garrette**
Google Research
{telmop,eschling,dhgarrette}@google.com

Cross Lingual Transfer

- What about transferring across languages with different scripts?
- I.e., zero overlap in vocabulary?
- E.g., train on Hindi (written in Devanagri), test on Urdu (written in Arabic)





Cross Lingual Transfer

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

Cross Lingual Transfer

Only ~6 points worse when trained on Urdu rather than Hindi

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

Cross Lingual Transfer

Larger performance drops for less similar language pairs

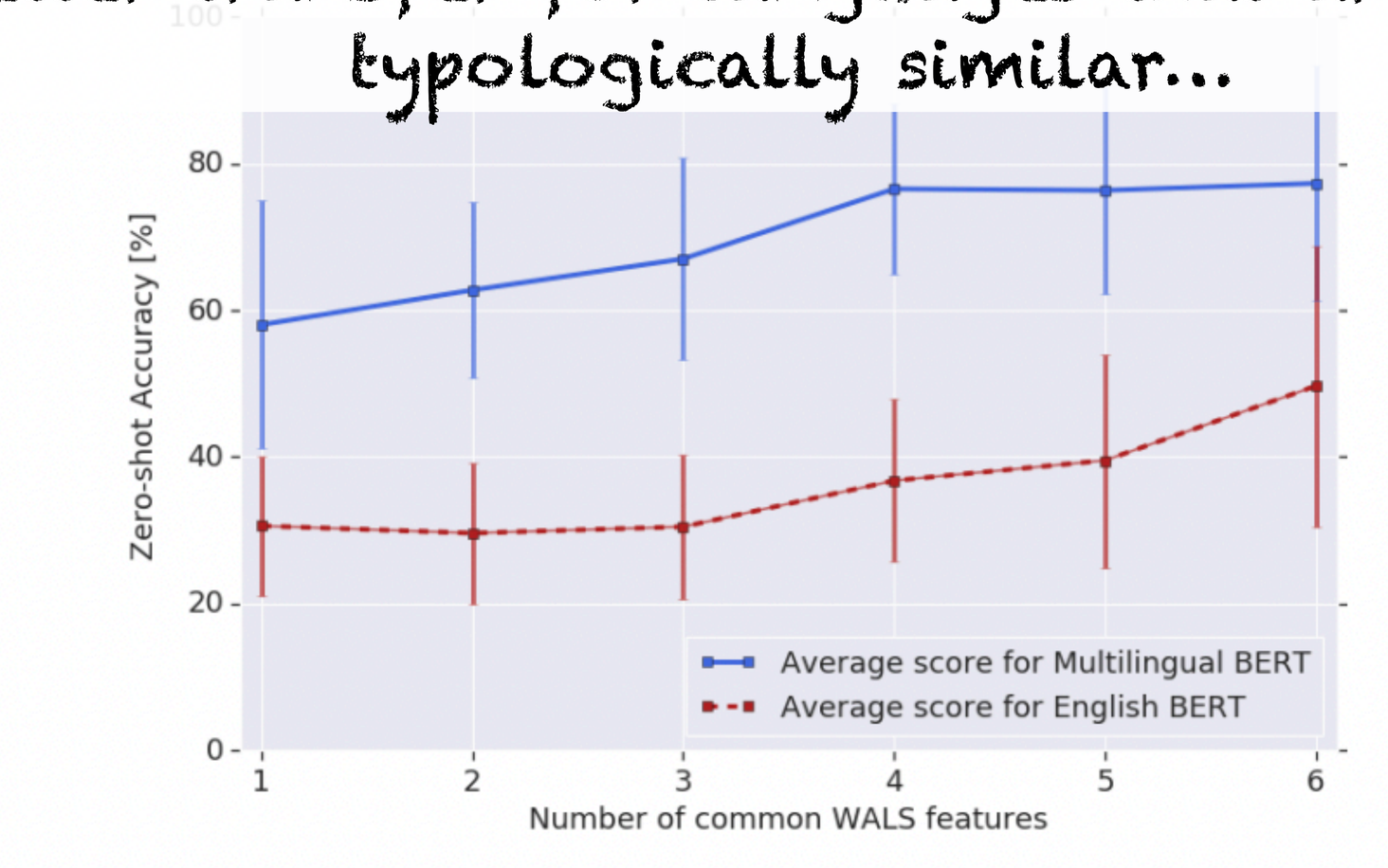
	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

Cross Lingual Transfer

Better transfer for languages that are more typologically similar...



Cross Lingual Transfer

...e.g., more syntactically similar

	SVO	SOV
SVO	**81.55**	66.52
SOV	63.98	**64.22**

(a) Subj./verb/obj. order.

	AN	NA
AN	**73.29**	70.94
NA	75.10	**79.64**

(b) Adjective/noun order.

Table 5: Macro-average POS accuracies when transferring between SVO/SOV languages or AN/NA languages. Row = fine-tuning, column = evaluation.

Cross Lingual Transfer

- Training on the language you care about is still better, in general
- Why would we do cross-lingual transfer then?
- A few reasons:
  - You might not have training data in the language you care about
  - You might not even have enough data to train an MT system for the language you care about
  - Cross-lingual models require only unlabeled, monolingual data, so are more feasible to train

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

BigScience Workshop

Major contributors

Teven Le Scao, Albert Villanova del Moral, Lucile Saulnier, Stas Bekman, Ellie Pavlick, Hugo Laurençon, Roman Castagné, Benoît Sagot, Jonathan Tow, Rachel Bawden, Niklas Muennighoff, Yacine Jernite, Alexandra Sasha Luccioni, Colin Raffel, François Yvon, Suzana Ilić, Margaret Mitchell, Albert Webson, Julien Launay, Victor Sanh, Thomas Wang, Stella Biderman, Daniel Hesslow, Iz Beltagy, Angelina McMillan-Major, Christopher Akiki, Olatunji Ruwase, Alexander M. Rush, Thomas Wolf, Pedro Ortiz Suarez, Samson Tan, Huu Nguyen

Dataset

Aaron Gokaslan, Aitor Soroa, Albert Villanova del Moral, Alexandra Sasha Luccioni, Alham Fikri Aji, Angelina McMillan-Major, Anna Rogers, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Akiki, Christopher Klamm, Colin Leong, Colin Raffel, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Ethan Kim, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hiếu Trần, Hugo Laurençon, Huu Nguyen, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Lucile Saulnier, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Margaret Mitchell, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Pawan Sasanka Ammanamanchi, Pedro Ortiz Suarez, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Samson Tan, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stella Biderman, Suhas Pai, Suzana Ilić, Sydney Zink, Teven Le Scao, Thomas Wang, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Yacine Jernite, Zaid Alyafeai, Zeerak Talat, Amit Alfassy, Adi Simhi, Efrat Levkovizh, Ariel Kreisberg Nitzav, Eyal Bar Natan, Stanislav Silberberg

Tokenization

Arun Raja, Benjamin Heinzerling, Benoît Sagot, Chenglei Si, Colin Raffel, Elizabeth Salesky, Lucile Saulnier, Manan Dey, Matthias Gallé, Pedro Ortiz Suarez, Roman Castagné, Sabrina J. Mielke, Samson Tan, Teven Le Scao, Thomas Wang, Wilson Y. Lee, Zaid Alyafeai

Prompt engineering

Abheesht Sharma, Albert Webson, Alexander M. Rush, Alham Fikri Aji, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Canwen Xu, Colin Raffel, Debajyoti Datta, Dragomir Radev, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jonathan Chang, Jos Rozen, Khalid Almubarak, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Manan Dey, Matteo Manica, Mike Tian-Jian Jiang, Nihal Nayak, Niklas Muennighoff, Rachel Bawden, Ryan Teehan, Samuel Albanie, Shanya Sharma, Sheng Shen, Srulik Ben David, Stella Biderman, Stephen H. Bach, Taewoon Kim, Tali Bers, Teven Le Scao, Thibault Fevry, Thomas Wang, Thomas Wolf, Trishala Neeraj, Urmish Thakker, Victor Sanh, Vikas Raunak, Xiangru Tang, Zaid Alyafeai, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh

Architecture and objective

Adam Roberts, Colin Raffel, Daniel Hesslow, Hady Elsahar, Hyung Won Chung, Iz Beltagy, Jaesung Tae, Jason Phang, Julien Launay, Lintang Sutawika, Lucile Saulnier, M Saiful Bari, Niklas Muennighoff, Ofir Press, Sheng Shen, Stas Bekman, Stella Biderman, Teven Le Scao, Thomas Wang, Victor Sanh, Zheng Xin Yong

Engineering

Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Niklas Muennighoff, Nouamane Tazi, Olatunji Ruwase, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stas Bekman, Stéphane Requena, Suraj Patil, Teven Le Scao, Thomas Wang, Tim Dettmers

Evaluation and interpretability

Ahmed Baruwa, Albert Webson, Alexandra Sasha Luccioni, Alham Fikri Aji, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Dragomir Radev, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Ellie Pavlick, François Yvon, Gerta Indra Winata, Hailey Schoelkopf, Jaesung Tae, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Khalid Almubarak, Lintang Sutawika, Manan Dey, Maraim Masoud, Margaret Mitchell, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Niklas Muennighoff, Oleg Serikov, Oskar van der Wal, Pawan Sasanka Ammanamanchi, Pierre Colombo, Rachel Bawden, Rui Zhang, Ruochen Zhang, Samson Tan, Sebastian Gehrmann, Shanya Sharma, Shayne Longpre, Stella Biderman, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Urmish Thakker, Verena Rieser, Vikas Raunak, Vitaly Protasov, Vladislav Mikhailov, Wilson Y. Lee, Yada Pruksachatkun, Zdeněk Kasner, Zeerak Talat, Zheng-Xin Yong, Shani Pais, Zachary Bamberger, Liam Hazan, Omer Antverg, Eli Bogdanov, Yonatan Belinkov

Broader impacts

Aaron Gokaslan, Alexandra Sasha Luccioni, Alham Fikri Aji, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Angelina McMillan-Major, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Chenghao Mou, Minh Chien Vu, Christopher Akiki, Danish Contractor, David Ifeoluwa Adelani, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Gérard Dupont, Giada Pistilli, Habib Rezanejad, Hessie Jones, Huu Nguyen, Ian Yu, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jaesung Tae, Jenny Chim, Jesse Passmore, Josh Seltzer, Julien Launay, Julio Bonis Sanz, Karen Fort, Khalid Almubarak, Livia Dutra, Long Phan, Mairon Samagaio, Manan Dey, Maraim Elbadri, Maraim Masoud, Margaret Mitchell, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Niklas Muennighoff, Nishant Subramani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Olivier Nguyen, Paulo Villegas, Pawan Sasanka Ammanamanchi, Priscilla Amuok, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Shanya Sharma, Shayne Longpre, Silas Wang, Somaieh Nikpoor, Sourav Roy, Stas Bekman, Stella Biderman, Suhas Pai, Suzana Ilić, Sylvain Viguier, Teven Le Scao, Thanh Le, Tobi Oyebade, Trieu Le, Tristan Thrush, Yacine Jernite, Yoyo Yang, Zach Nguyen, Zeerak Talat, Zheng Xin Yong

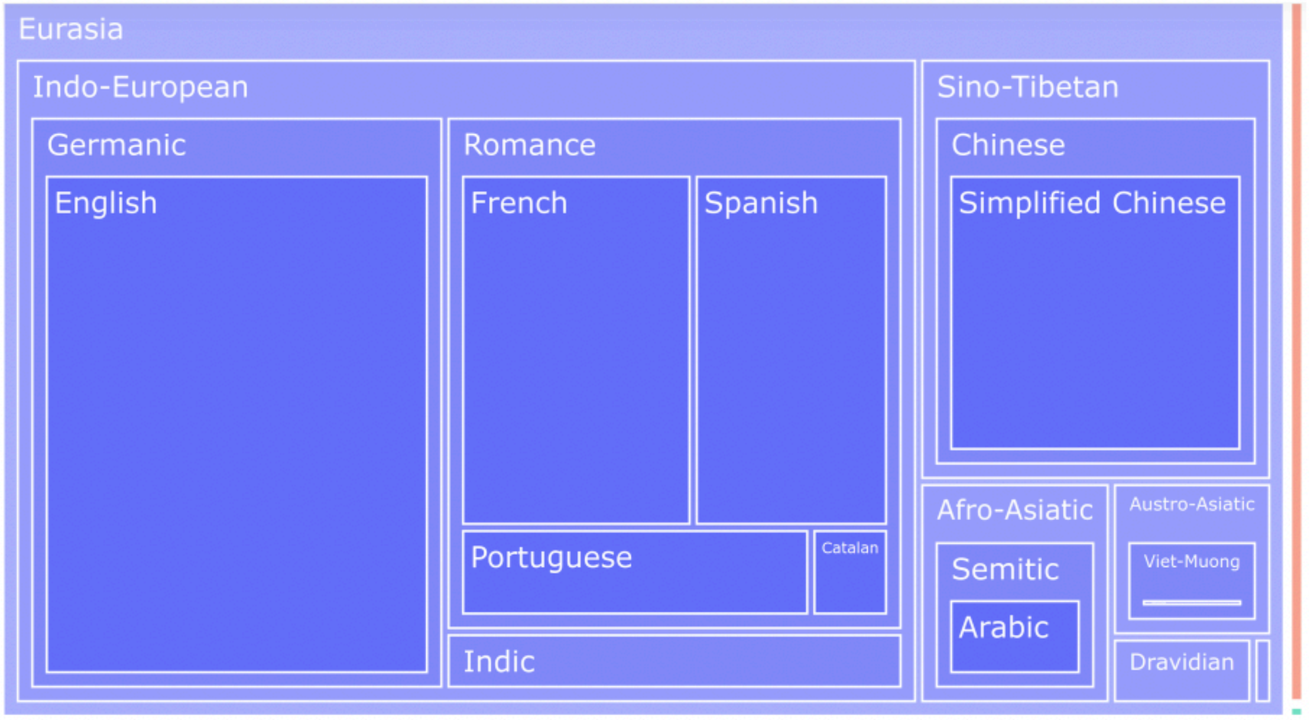
Applications

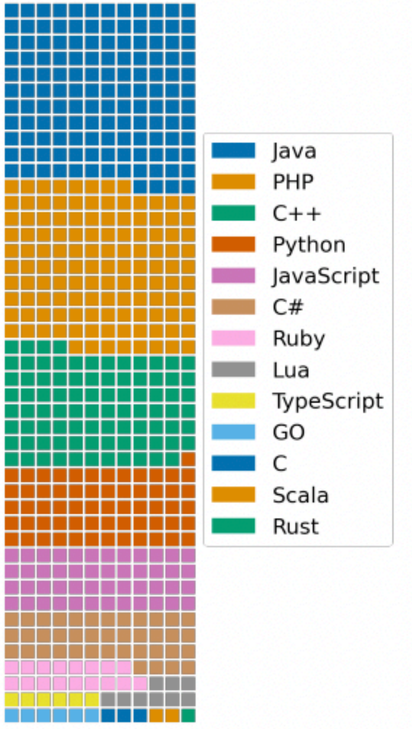
Abhinav Ramesh Kashyap, Albert Villanova del Moral, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Carlos Muñoz Ferrandis, Chenxi Zhou, Chirag Jain, Christopher Akiki, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Daniel van Strien, Danish Contractor, David Lansky, Debajyoti Datta, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Francesco De Toni, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jason Alan Fries, Javier de la Rosa, Jenny Chim, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Leon Weber, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minh Chien Vu, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shamik Bose, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Stella Biderman, Stephen H. Bach, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Trishala Neeraj, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye

Organization

Angela Fan, Christopher Akiki, Douwe Kiela, Giada Pistilli, Margot Mieskes, Mathilde Bras, Matthias Gallé, Suzana Ilić, Yacine Jernite, Younes Belkada, Thomas Wolf

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model





BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

Src↓	Trg→	eng	ben	hin	swh	yor
eng	M2M	–	23.04	28.15	29.65	2.17
	BLOOM	–	25.52	27.57	21.7	2.8
ben	M2M	22.86	–	21.76	14.88	0.54
	BLOOM	30.23	–	16.4	–	–
hin	M2M	27.89	21.77	–	16.8	0.61
	BLOOM	35.40	23.0	–	–	–
swh	M2M	30.43	16.43	19.19	–	1.29
	BLOOOM	37.9	–	–	–	1.43
yor	M2M	4.18	1.27	1.94	1.93	–
	BLOOM	3.8	–	–	0.84	–

(a) Low-resource languages

Src↓	Trg→	cat	spa	fre	por
cat	M2M	–	25.17	35.08	35.15
	BLOOM	–	29.12	34.89	36.11
spa	M2M	23.12	–	29.33	28.1
	BLOOM	31.82	–	24.48	28.0
glg	M2M	30.07	27.65	37.06	34.81
	BLOOM	38.21	27.24	36.21	34.59
fre	M2M	28.74	25.6	–	37.84
	BLOOM	38.13	27.40	–	39.60
por	M2M	30.68	25.88	40.17	–
	BLOOM	40.02	28.1	40.55	–

(b) Romance languages

All done!
Questions?