

Attention + Neural MT

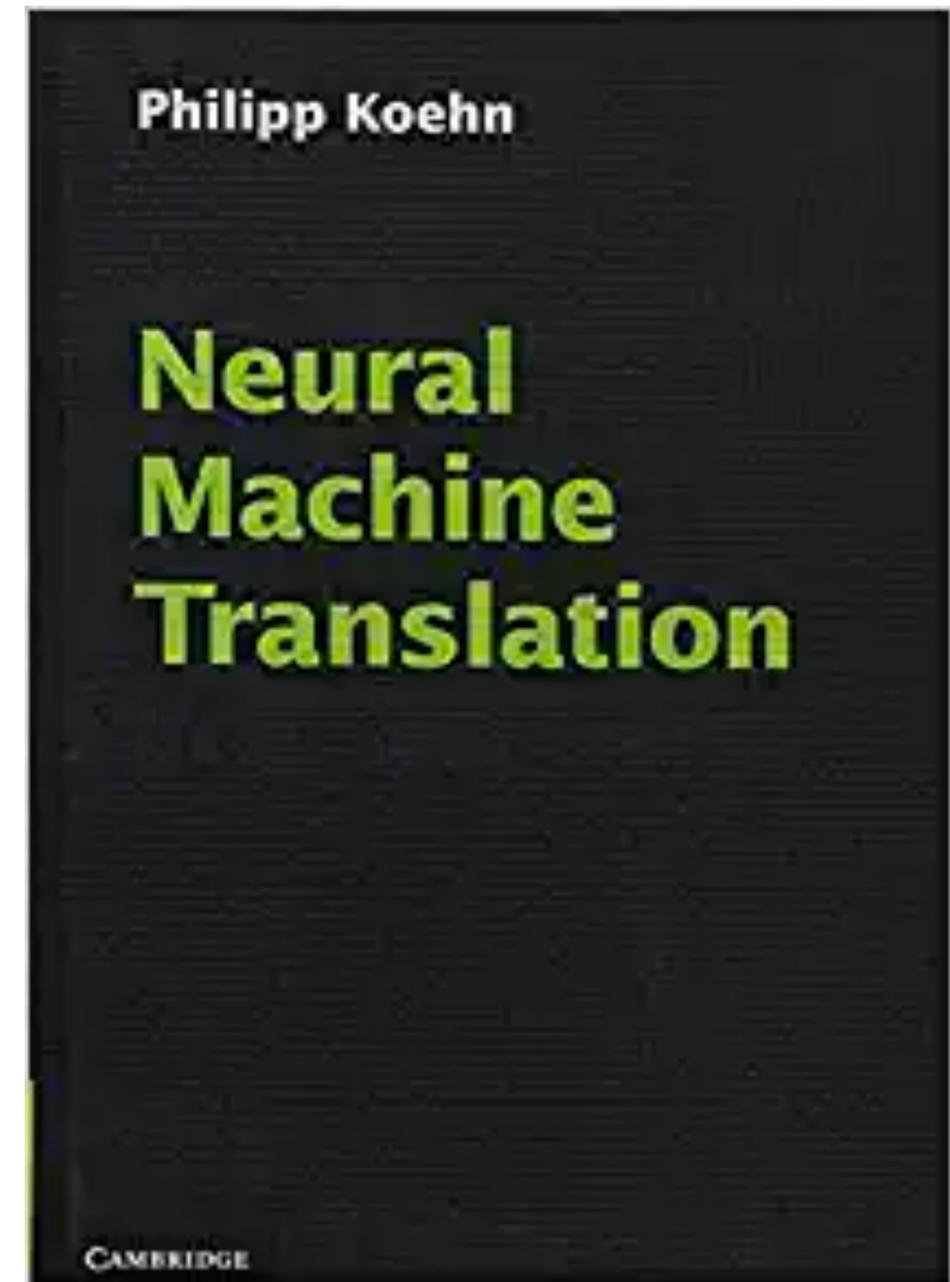
Wei Xu

(many slides from Greg Durrett)

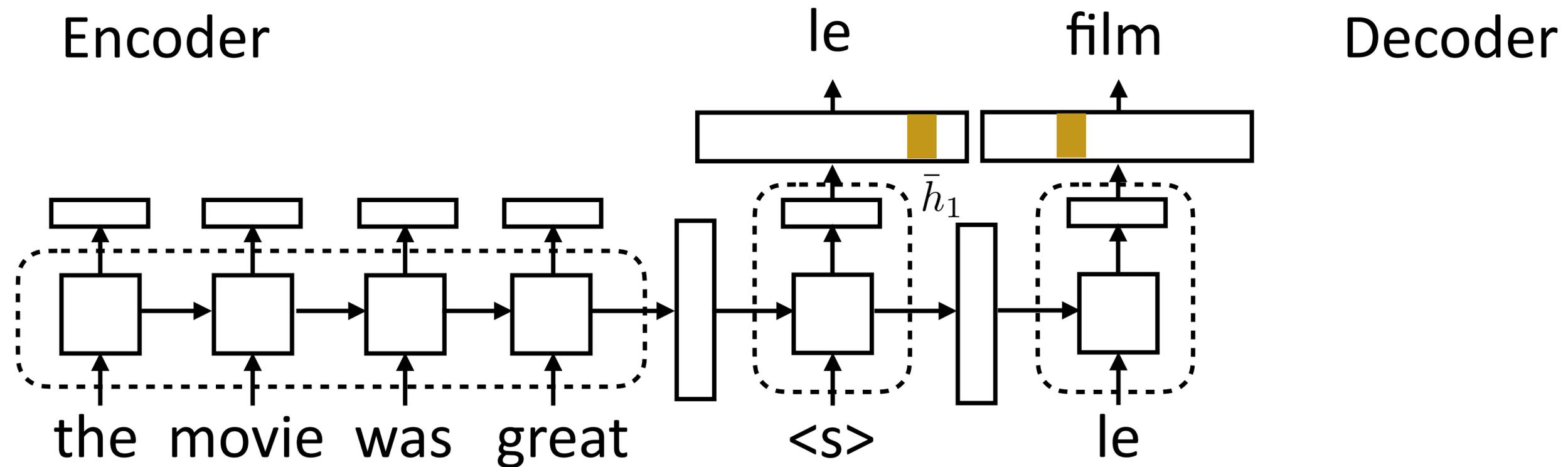
This Lecture

- ▶ Attention
- ▶ Copy/Pointer Network
- ▶ Neural Machine Translation

- ▶ Reading — Eisenstein 18.3-18.5



Recap: Seq2Seq Model



- ▶ Encoder: consumes sequence of tokens, produces a vector. Analogous to encoders for classification/tagging tasks $P(y_i | \mathbf{x}, y_1, \dots, y_{i-1}) = \text{softmax}(W \bar{h}_i)$
- ▶ Decoder: separate module, single cell. Takes two inputs: hidden state (vector h or tuple (h, c)) and previous token. Outputs token + new state

Results: Encoder-Decoder MT

- ▶ Kalchbrenner & Blunsom (2013), Bahdanau et al. (2014), Cho et al. (2014)
- ▶ Sutskever et al. (2014) paper: first major application of LSTMs to NLP
- ▶ Basic encoder-decoder with beam search

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	34.81

- ▶ SOTA = 37.0 then — not all that competitive...

Sutskever et al. (2014)

Attention

Problems with Seq2seq Models

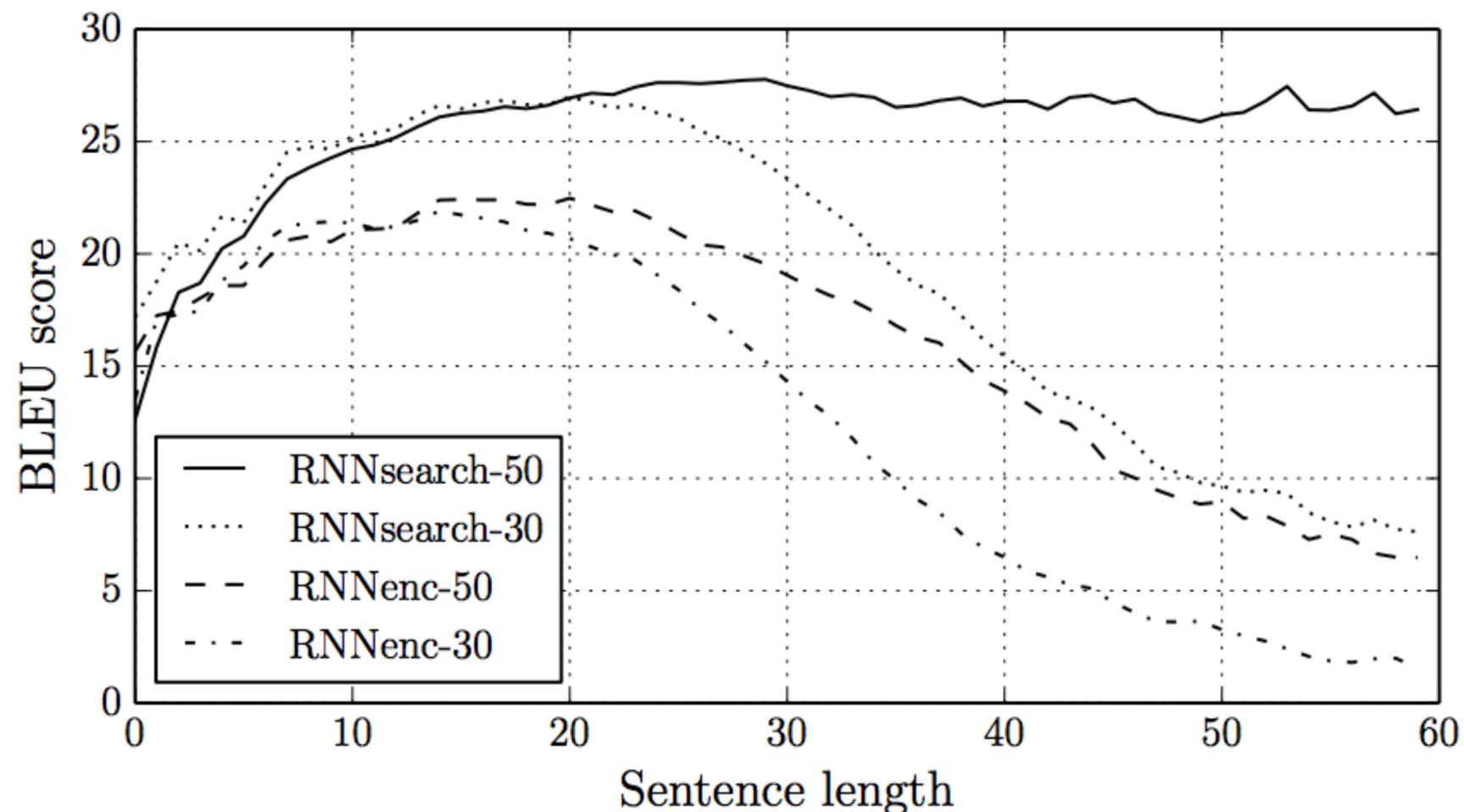
- ▶ Encoder-decoder models like to repeat themselves:

Un garçon joue dans la neige → *A boy plays in the snow **boy plays boy plays***

- ▶ Often a byproduct of training these models poorly. Input is forgotten by the LSTM so it gets stuck in a “loop” of generation the same output tokens again and again.
- ▶ Need some notion of input coverage or what input words we’ve translated

Problems with Seq2seq Models

- ▶ Bad at long sentences: 1) a fixed-size hidden representation doesn't scale; 2) LSTMs still have a hard time remembering for really long sentences

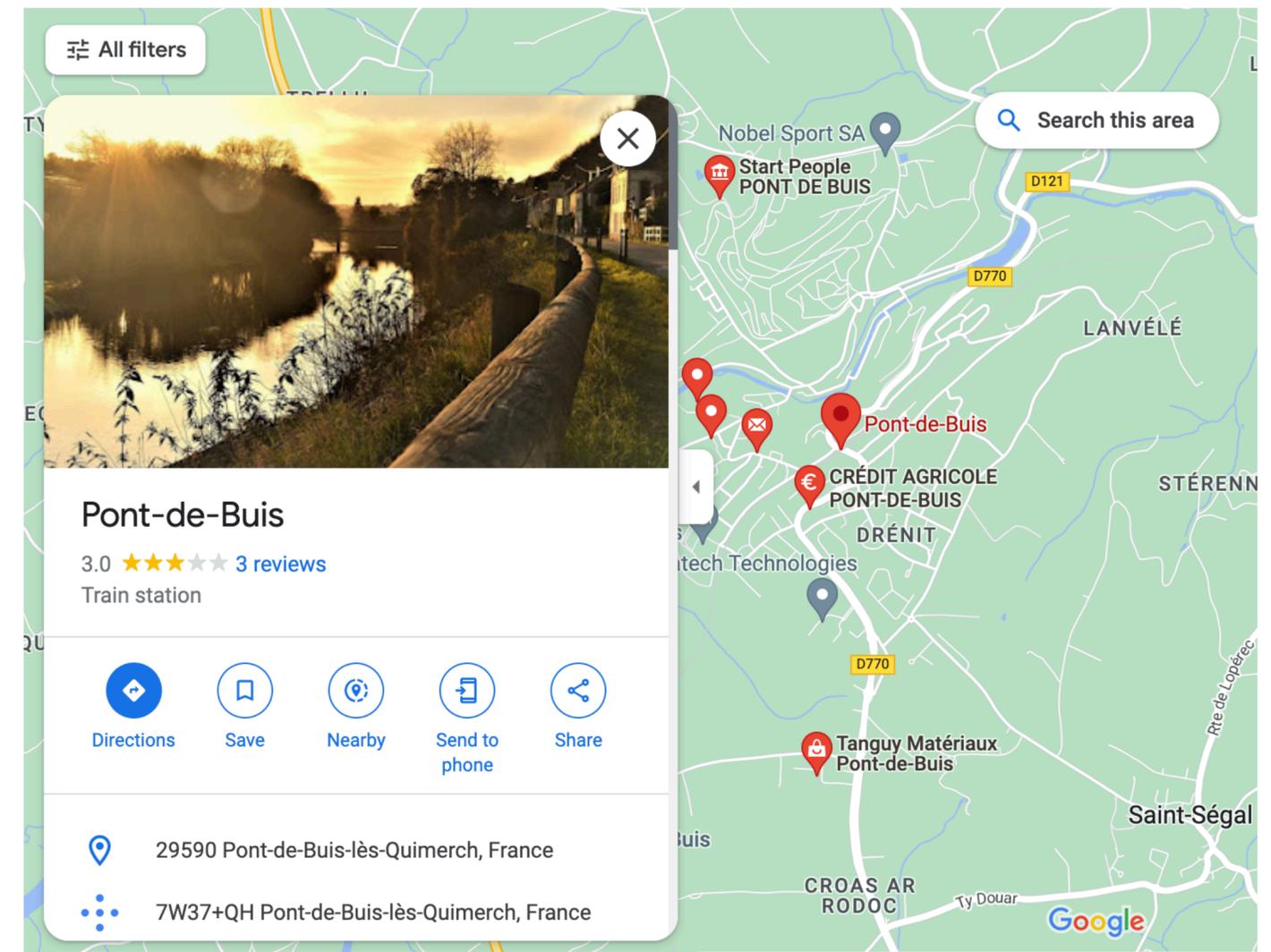


RNNenc: the model we've discussed so far

RNNsearch: uses attention

Rare/Unknown Words

The **ecotax portico** in **Pont-de-Buis**, around which a violent demonstration against the tax took place on Saturday, was taken down on Thursday morning.



Problems with Seq2seq Models

- ▶ Unknown words:

en: The ecotax portico in Pont-de-Buis, ... [truncated] ..., was taken down on Thursday morning

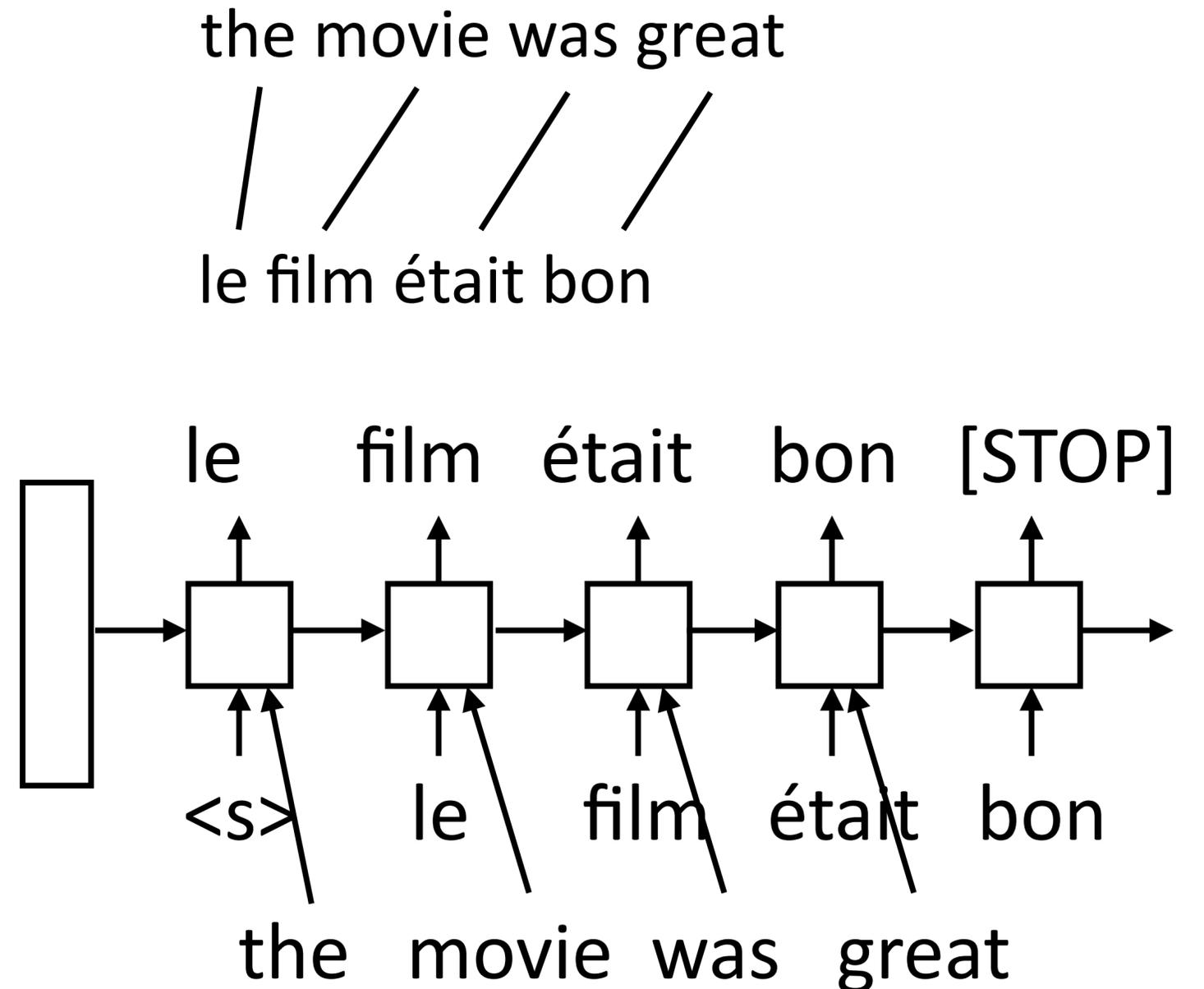
fr: Le portique écotaxe de Pont-de-Buis, ... [truncated] ..., a été démonté jeudi matin

nn: Le unk de unk à unk, ... [truncated] ..., a été pris le jeudi matin

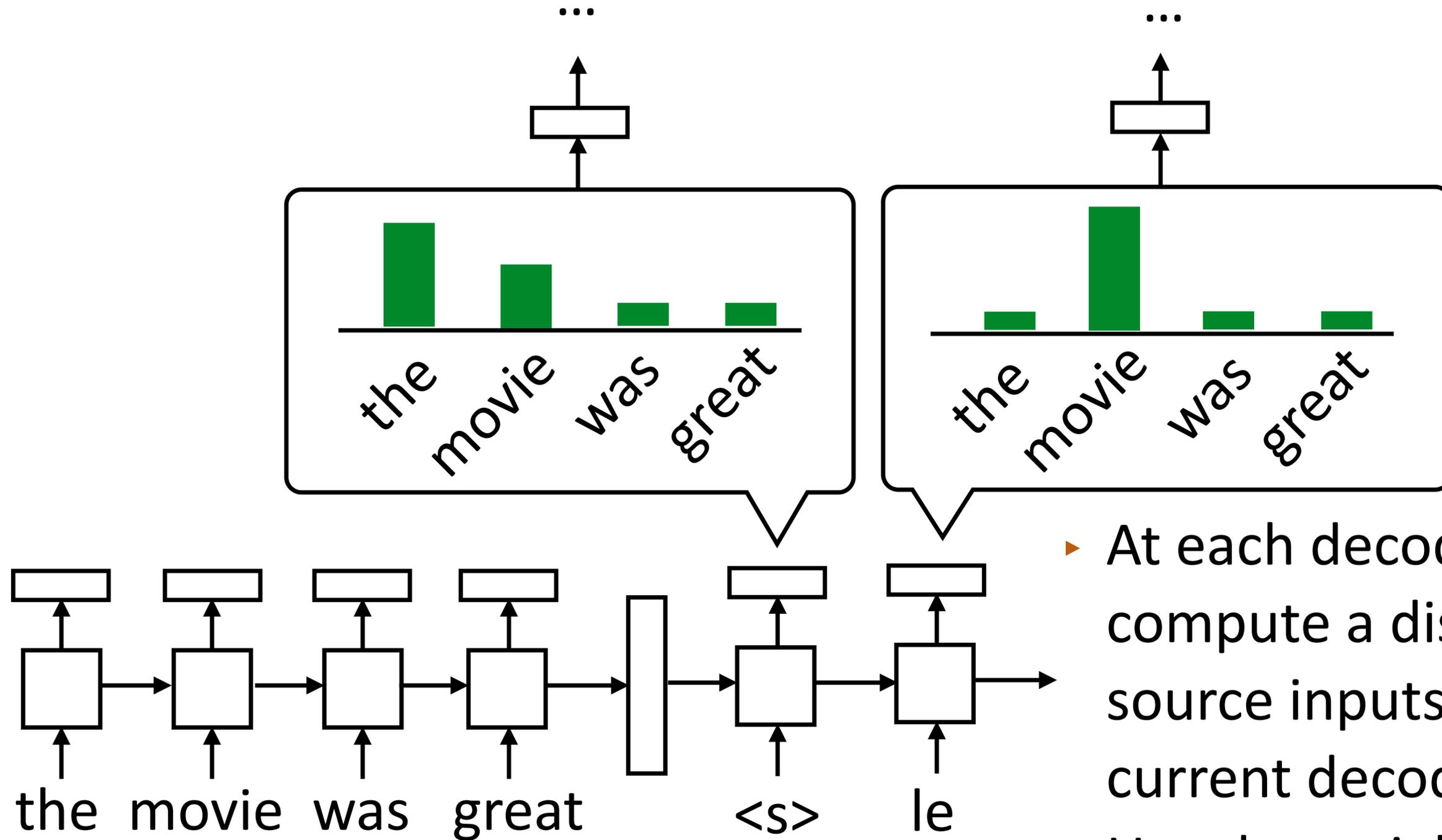
- ▶ Encoding these rare words into a vector space is really hard
- ▶ In fact, we don't want to encode them, we want a way of directly looking back at the input and copying them (Pont-de-Buis)

Aligned Inputs

- ▶ Suppose we knew the source and target would be word-by-word translated (recall the word alignment we talked about in phrase-based MT)
- ▶ Can look at the corresponding input word when translating — this could scale!
- ▶ Less burden on the hidden states
- ▶ How can we achieve this without hardcoding it?



Attention



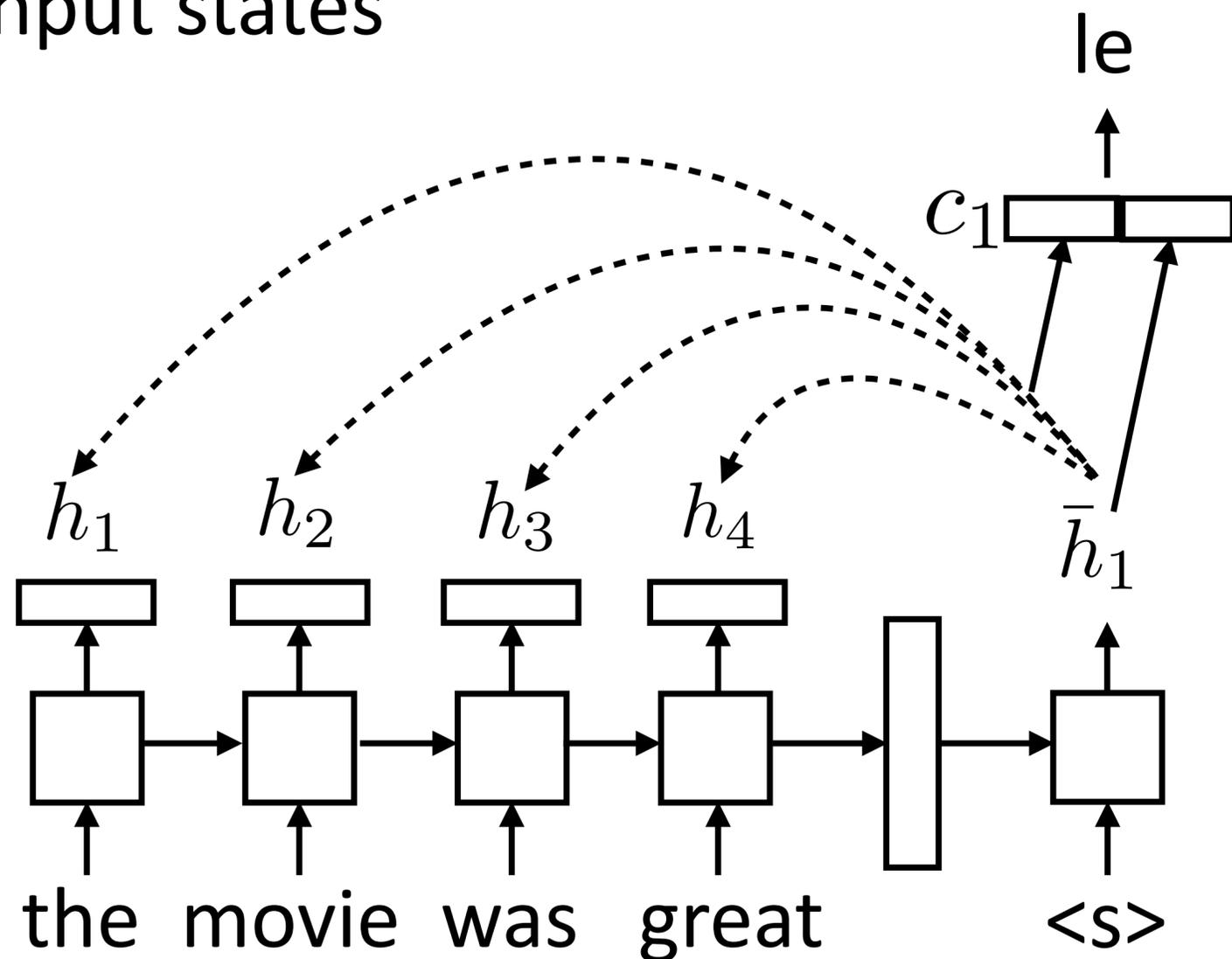
- ▶ At each decoder state, compute a distribution over source inputs based on current decoder state
- ▶ Use the weighted sum of input tokens to predict output

Attention

- ▶ For each decoder state, compute weighted sum of input states

- ▶ No attn: $P(y_i | \mathbf{x}, y_1, \dots, y_{i-1}) = \text{softmax}(W \bar{h}_i)$

$$P(y_i | \mathbf{x}, y_1, \dots, y_{i-1}) = \text{softmax}(W [c_i; \bar{h}_i])$$

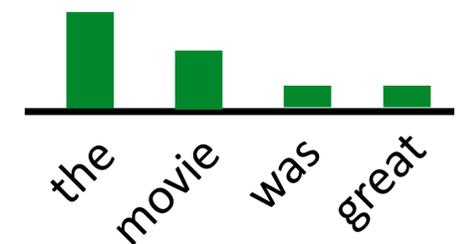


$$c_i = \sum_j \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})}$$

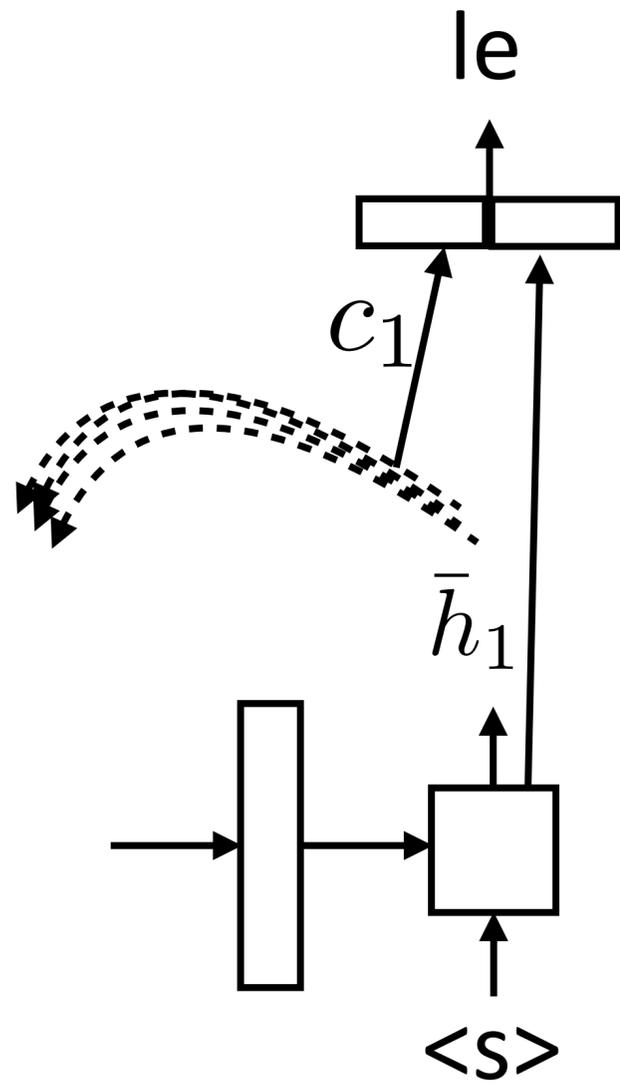
$$e_{ij} = f(\bar{h}_i, h_j)$$

- ▶ Weighted sum of input hidden states (vector)



- ▶ Some function f (e.g., dot product)

Attention



$$c_i = \sum_j \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})}$$

$$e_{ij} = f(\bar{h}_i, h_j)$$

$$f(\bar{h}_i, h_j) = \tanh(W[\bar{h}_i; h_j])$$

► Bahdanau+ (2014): additive

$$f(\bar{h}_i, h_j) = \bar{h}_i \cdot h_j$$

► Luong+ (2015): dot product

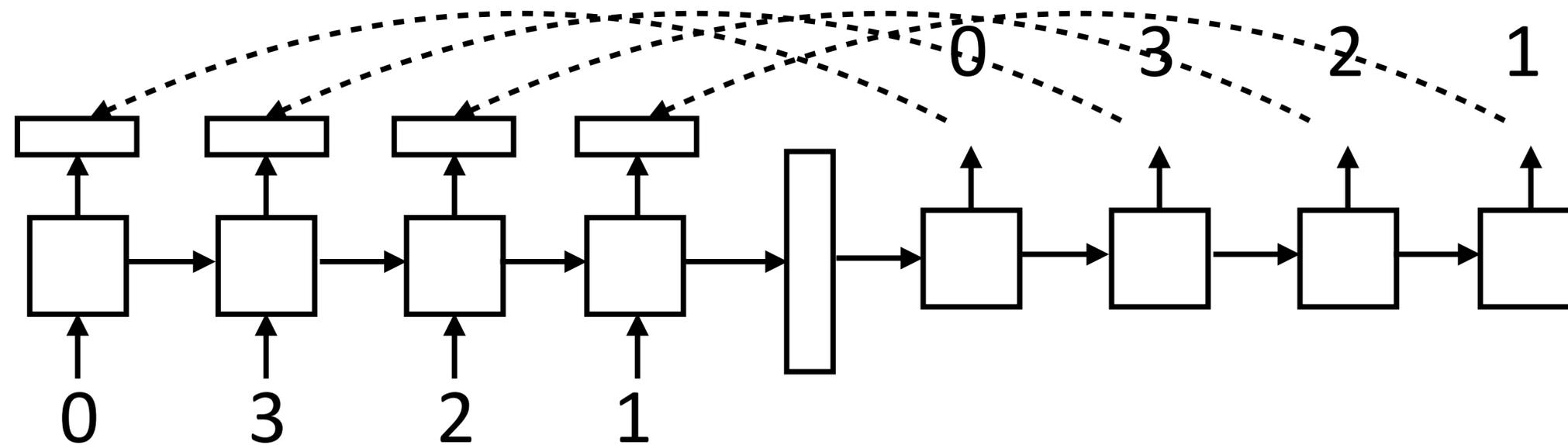
$$f(\bar{h}_i, h_j) = \bar{h}_i^\top W h_j$$

► Luong+ (2015): bilinear

► Note that this all uses outputs of hidden layers

What can attention do?

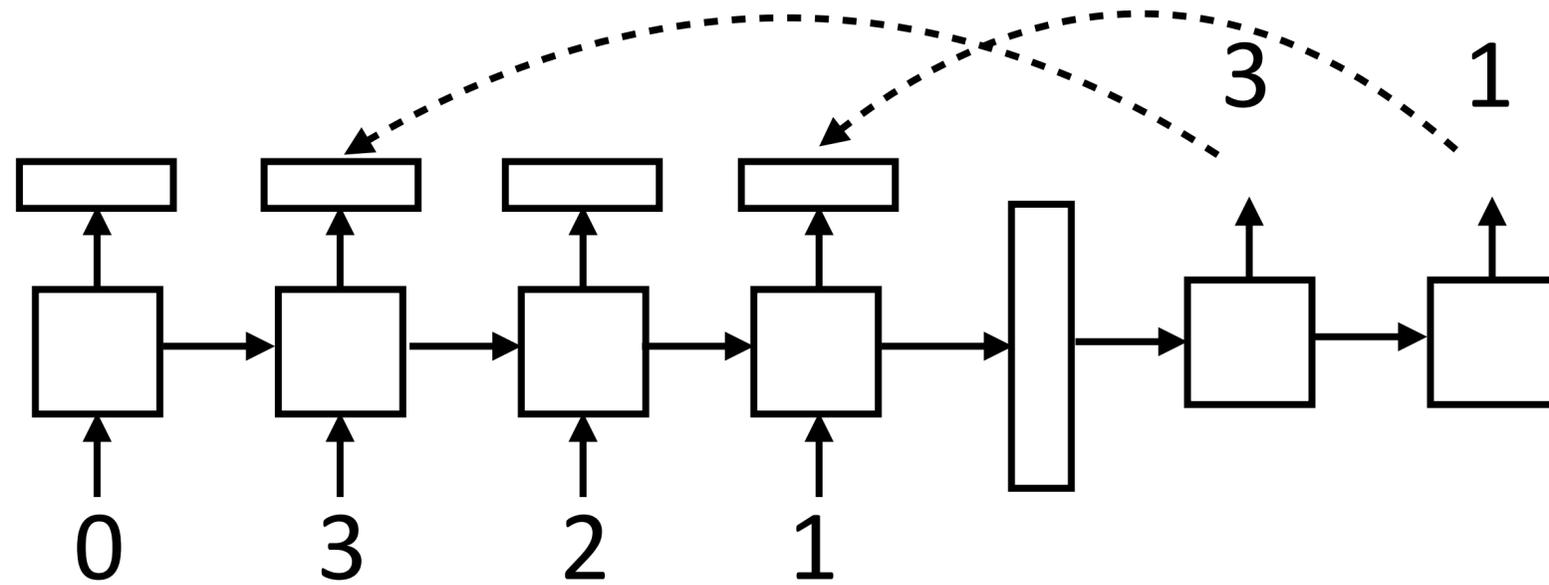
- ▶ Learning to copy — how might this work?



- ▶ LSTM can learn to count with the right weight matrix
- ▶ This is a kind of position-based addressing

What can attention do?

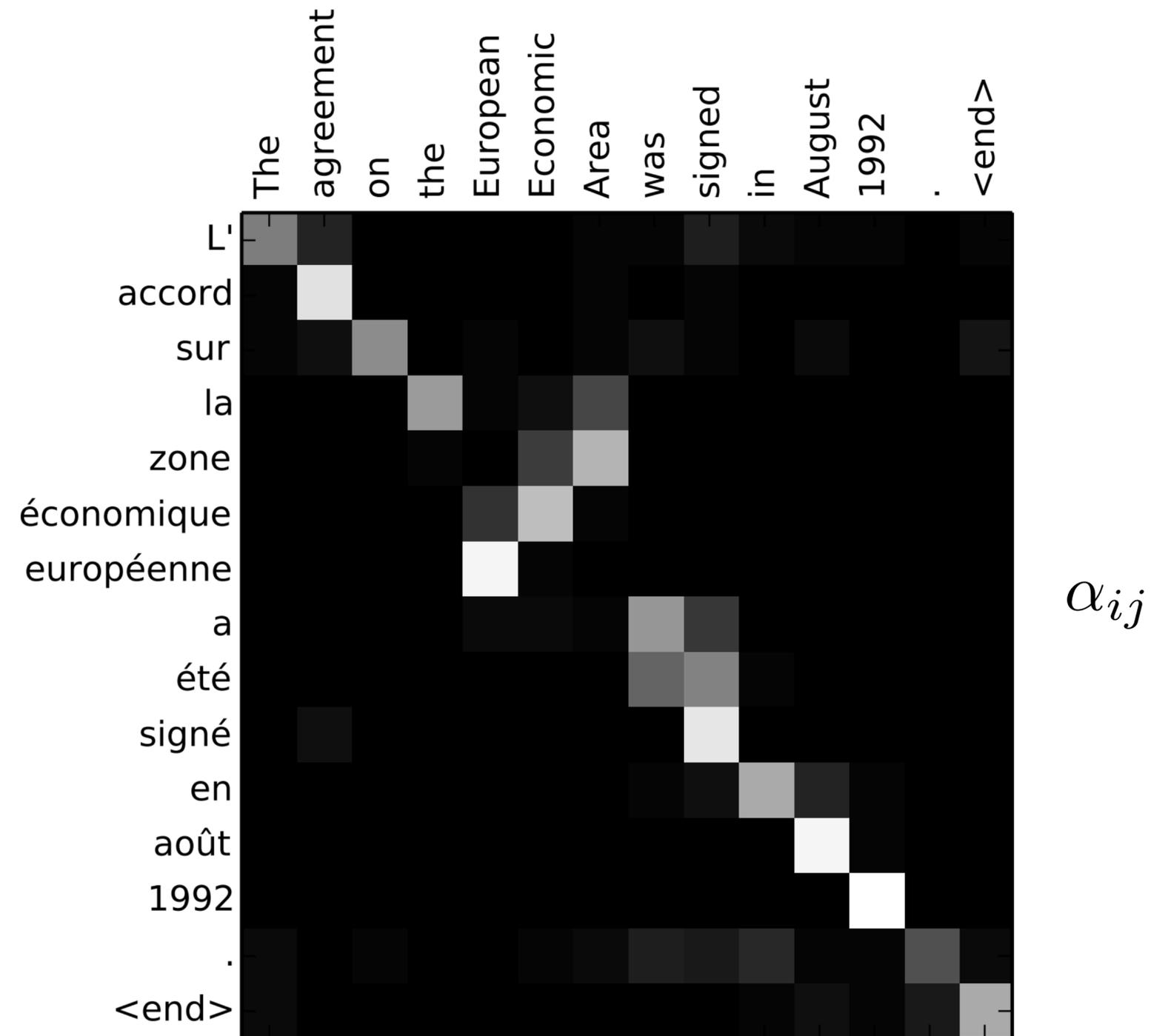
- ▶ Learning to subsample tokens



- ▶ Need to count (for ordering) and also determine which tokens are in/out
- ▶ Content-based addressing

Attention

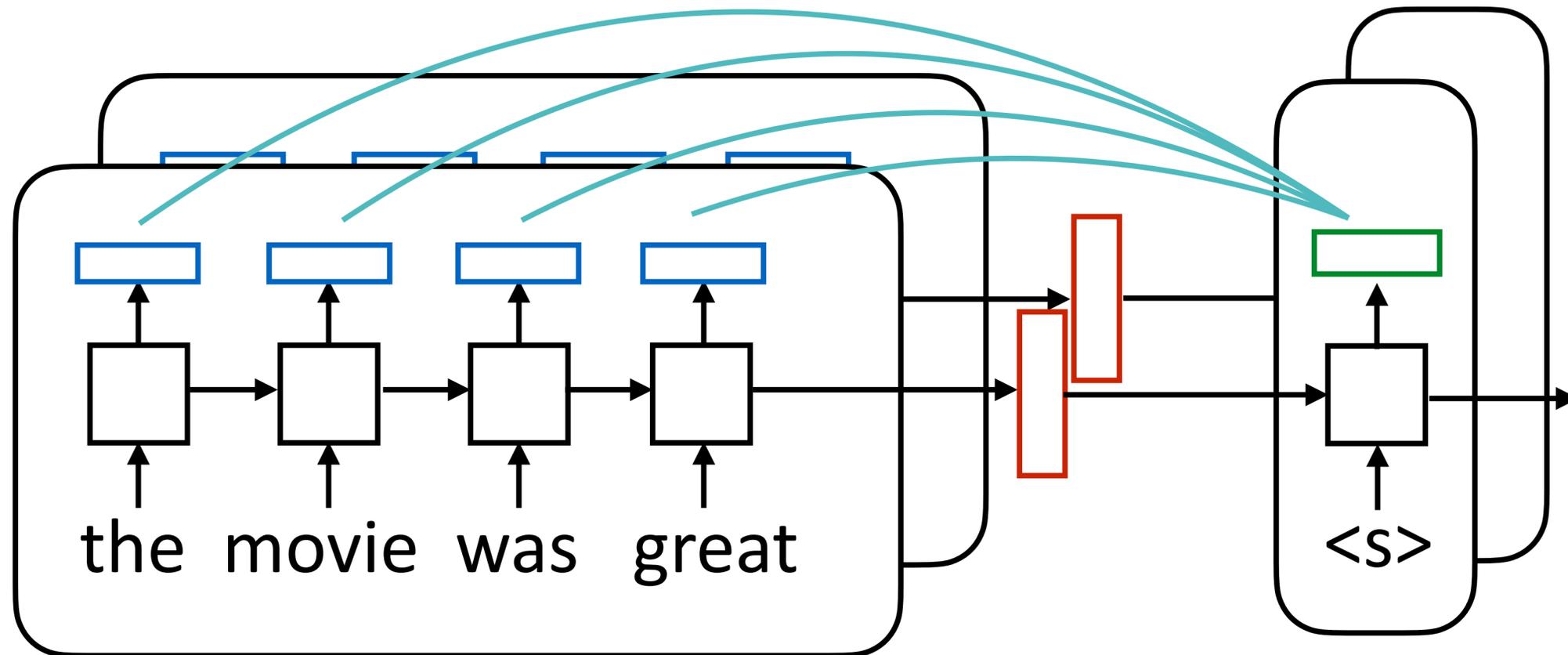
- ▶ Encoder hidden states capture contextual source word identity (“soft” word alignment)
- ▶ Decoder hidden states are now mostly responsible for selecting what to attend to
- ▶ Doesn’t take a complex hidden state to walk monotonically through a sentence and spit out word-by-word translations



Bahdanau et al. (2014)

Batching Attention

token outputs: batch size x sentence length x hidden size



hidden state: batch size x hidden size

$$e_{ij} = f(\bar{h}_i, h_j)$$
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})}$$

sentence outputs:
batch size x hidden size

attention scores = batch size x sentence length

c = batch size x hidden size

$$c_i = \sum_j \alpha_{ij} h_j$$

- ▶ Make sure tensors are the right size!

Luong et al. (2015)

Some MT Results

“Early” Neural MT

Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong Hieu Pham Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
{lmthang, hyhieu, manning}@stanford.edu

Abstract

An attentional mechanism has lately been used to improve neural machine translation (NMT) by selectively focusing on parts of the source sentence during translation. However, there has been little work exploring useful architectures for attention-based NMT. This paper examines two simple and effective classes of attentional mechanism: a *global* approach which always attends to all source words and a *local* one that only looks at a subset of source words at a time. We demonstrate the effectiveness of both approaches on the

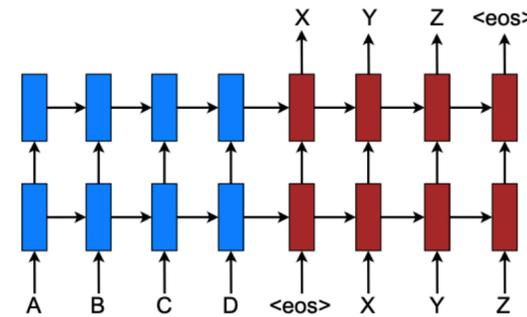


Figure 1: **Neural machine translation** – a stacking recurrent architecture for translating a source sequence A B C D into a target sequence X Y Z. Here, <eos> marks the end of a sentence.

ing plain SGD, (c) a simple learning rate schedule is employed – we start with a learning rate of 1; after 5 epochs, we begin to halve the learning rate every epoch, (d) our mini-batch size is 128, and (e) the normalized gradient is rescaled whenever its norm exceeds 5. Additionally, we also use dropout with probability 0.2 for our LSTMs as suggested by (Zaremba et al., 2015). For dropout models, we train for 12 epochs and start halving the learning rate after 8 epochs. For local attention models, we empirically set the window size $D = 10$.

Our code is implemented in MATLAB. When running on a single GPU device Tesla K40, we achieve a speed of 1K *target* words per second. It takes 7–10 days to completely train a model.

- ▶ TensorFlow first released in Nov 2015.
- ▶ PyTorch first released in 2016.

Luong et al. (2015)

MT Examples

src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and <i>Kerr</i> still love each other .
<i>best</i>	In an interview , however , Bloom said that he and <i>Kerr</i> still love .
base	However , in an interview , Bloom said that he and Tina were still <unk> .

- ▶ best = with attention, base = no attention
- ▶ NMT systems can hallucinate words, especially when not using attention
— phrase-based doesn't do this

MT Examples

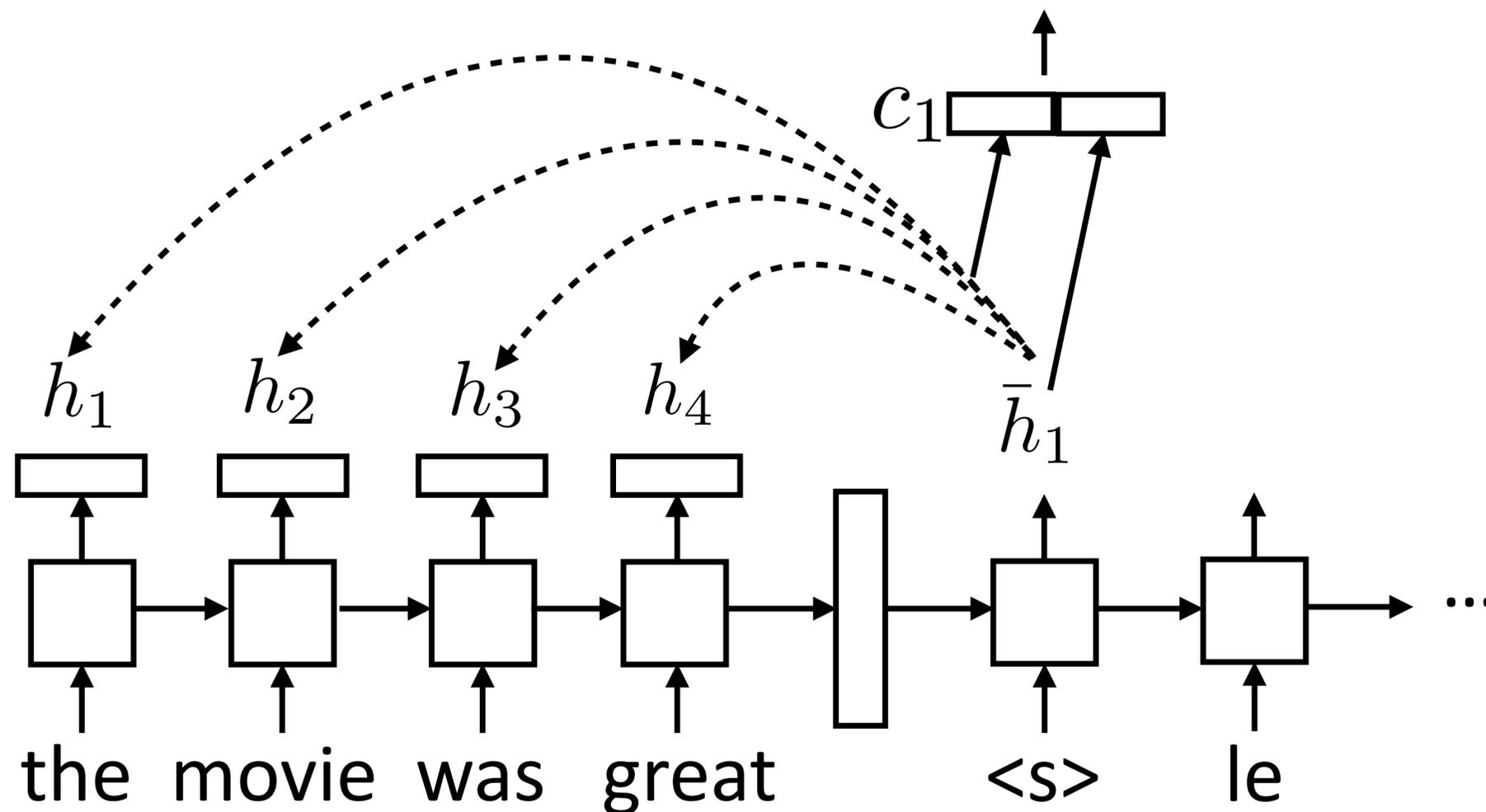
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The <i>austerity imposed by Berlin and the European Central Bank</i> , coupled with the straitjacket imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict <i>austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket</i> in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure imposed by the European Central Bank and the Federal Central Bank with the strict austerity imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

- ▶ best = with attention, base = no attention

Encoder-Decoder MT

- ▶ Better encoder-decoder with attention and handling of rare words

distribution over vocab + copying (more on this later)



Results: WMT English-French

- ▶ 12M sentence pairs

Classic phrase-based system: ~**33** BLEU, uses additional target-language data

Rerank with LSTMs: **36.5** BLEU (long line of work here; Devlin+ 2014)

Sutskever+ (2014) seq2seq single: **30.6** BLEU (input reversed)

Sutskever+ (2014) seq2seq ensemble: **34.8** BLEU

Luong+ (2015) seq2seq ensemble with attention and rare word handling:
37.5 BLEU

- ▶ But English-French is a really easy language pair and there's *tons* of data for it! Does this approach work for anything harder?

Results: WMT English-German

- ▶ 4.5M sentence pairs

Classic phrase-based system: **20.7** BLEU

Luong+ (2014) seq2seq: **14** BLEU

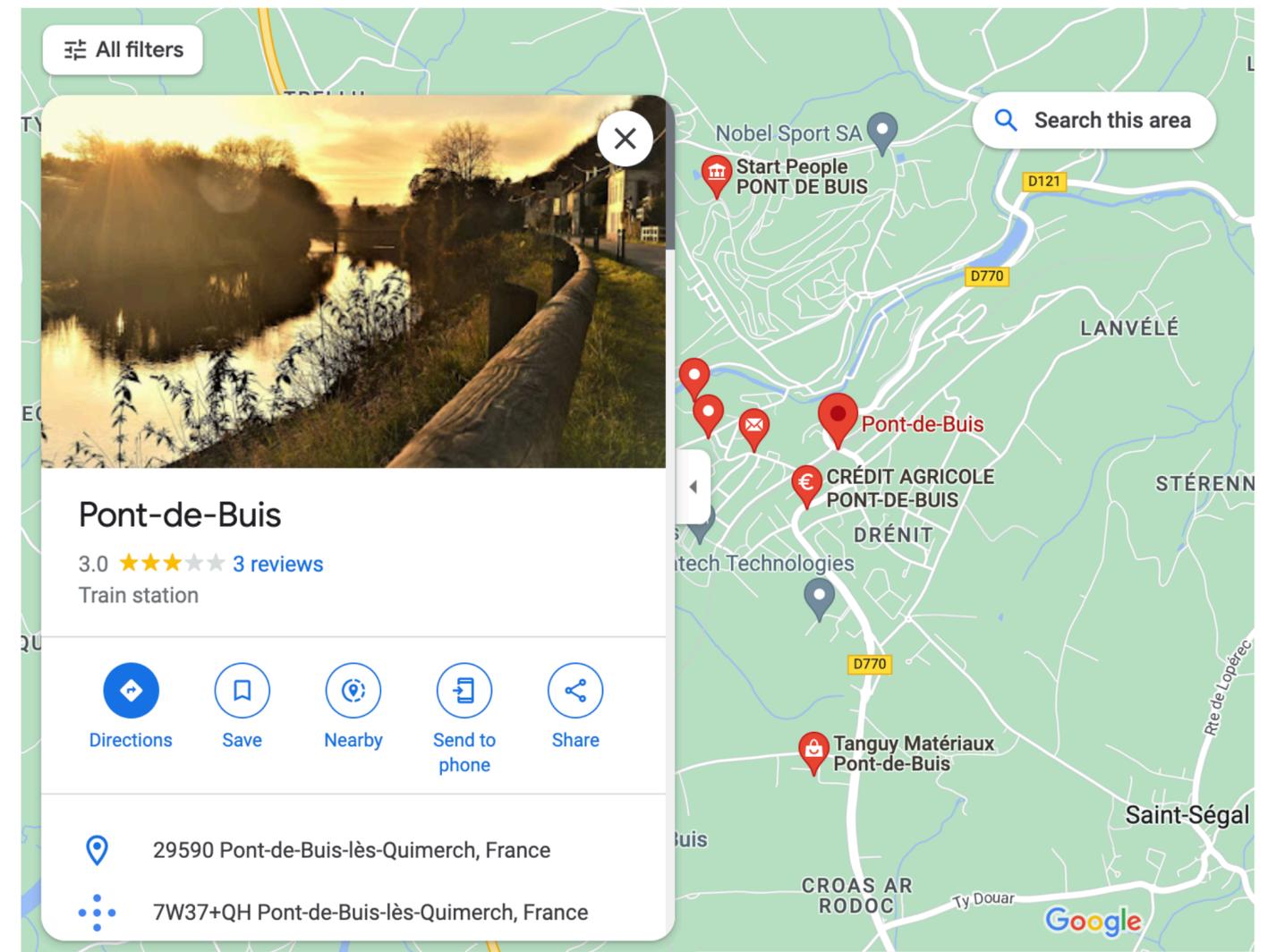
Luong+ (2015) seq2seq ensemble with rare word handling: **23.0** BLEU

- ▶ Not nearly as good in absolute BLEU, but not really comparable across languages
- ▶ French, Spanish = easiest
German, Czech = harder
Japanese, Russian = hard (grammatically different, lots of morphology...)

Copy / Pointer Networks

Rare/Unknown Words

The **ecotax portico** in **Pont-de-Buis**, around which a violent demonstration against the tax took place on Saturday, was taken down on Thursday morning.



Unknown Words

en: The ecotax portico in Pont-de-Buis, ... [truncated] ..., was taken down on Thursday morning

fr: Le portique écotaxe de Pont-de-Buis, ... [truncated] ..., a été démonté jeudi matin

nn: Le unk de unk à unk, ... [truncated] ..., a été pris le jeudi matin

- ▶ Attention mechanism:

$$P(y_i | \mathbf{x}, y_1, \dots, y_{i-1}) = \text{softmax}(W[c_i; \bar{h}_i])$$

from attention from RNN
hidden state

- ▶ Problems: want to be able to copy named entities like Pont-de-Buis, but target word has to be in the vocabulary, attention + RNN need to generate good embedding to pick it.

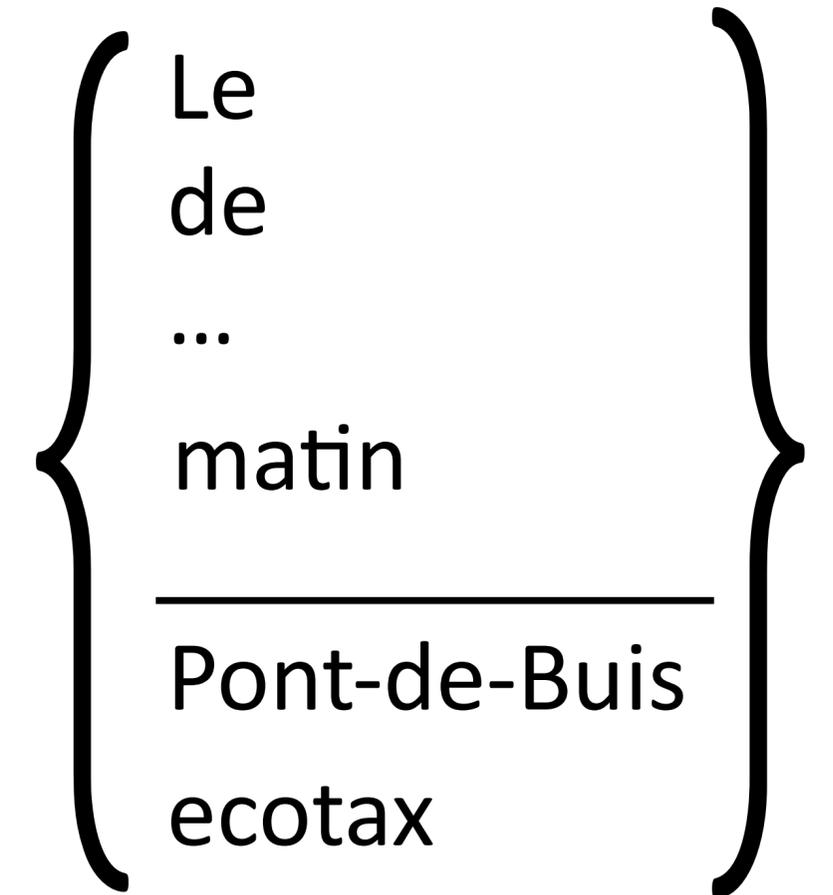
Jean et al. (2015), Luong et al. (2015)

Copying

en: The ecotax portico in Pont-de-Buis , ... [truncated] ..

fr: Le portique écotaxe de Pont-de-Buis , ... [truncated] .

nn: Le unk de unk à unk , ... [truncated] ..., a été pris

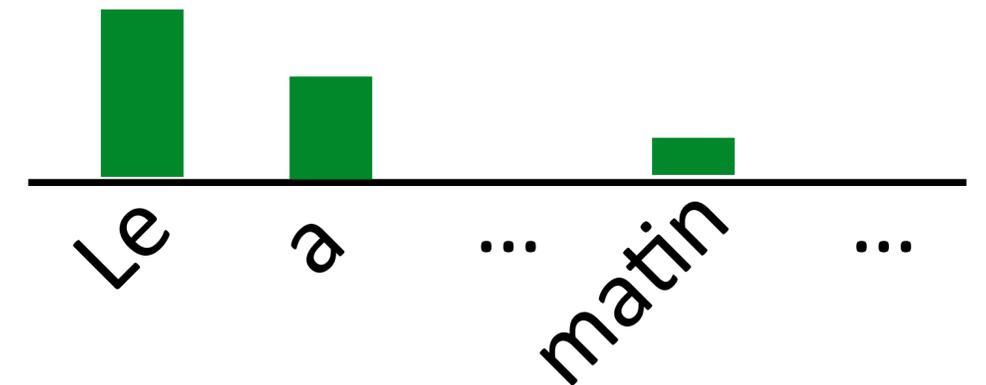


- ▶ Some words we want to copy may not be in the fixed output vocab (*Pont-de-Buis*)
- ▶ Solution: Vocabulary contains “normal” vocab as well as words in input.

Pointer Network

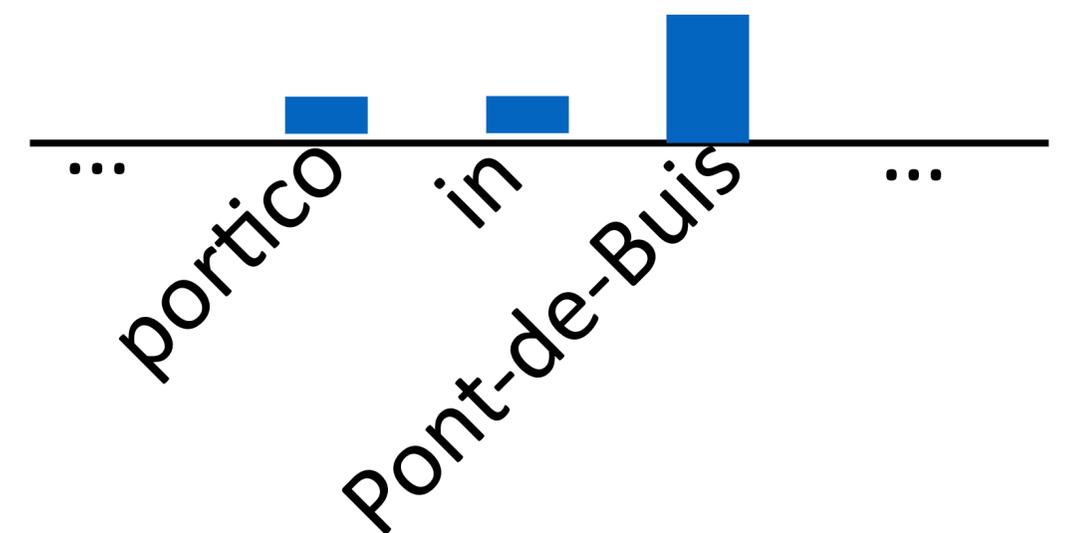
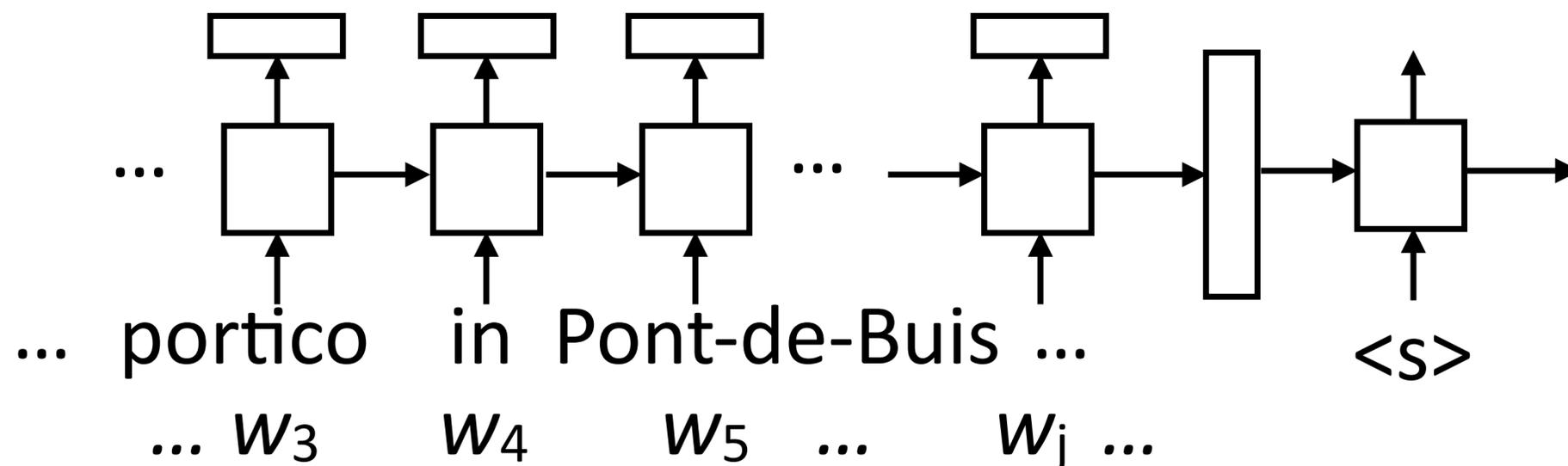
- Standard decoder with attention (P_{vocab}): softmax over vocabulary

$$P(y_i | \mathbf{x}, y_1, \dots, y_{i-1}) = \text{softmax}(W[c_i; \bar{h}_i])$$



- Pointer network (P_{pointer}): predict from *source* words, instead of target vocabulary

$$P_{\text{pointer}}(y_i | \mathbf{x}, y_1, \dots, y_{i-1}) \propto \begin{cases} h_j^\top V \bar{h}_i & \text{if } y_i = w_j \\ 0 & \text{otherwise} \end{cases}$$

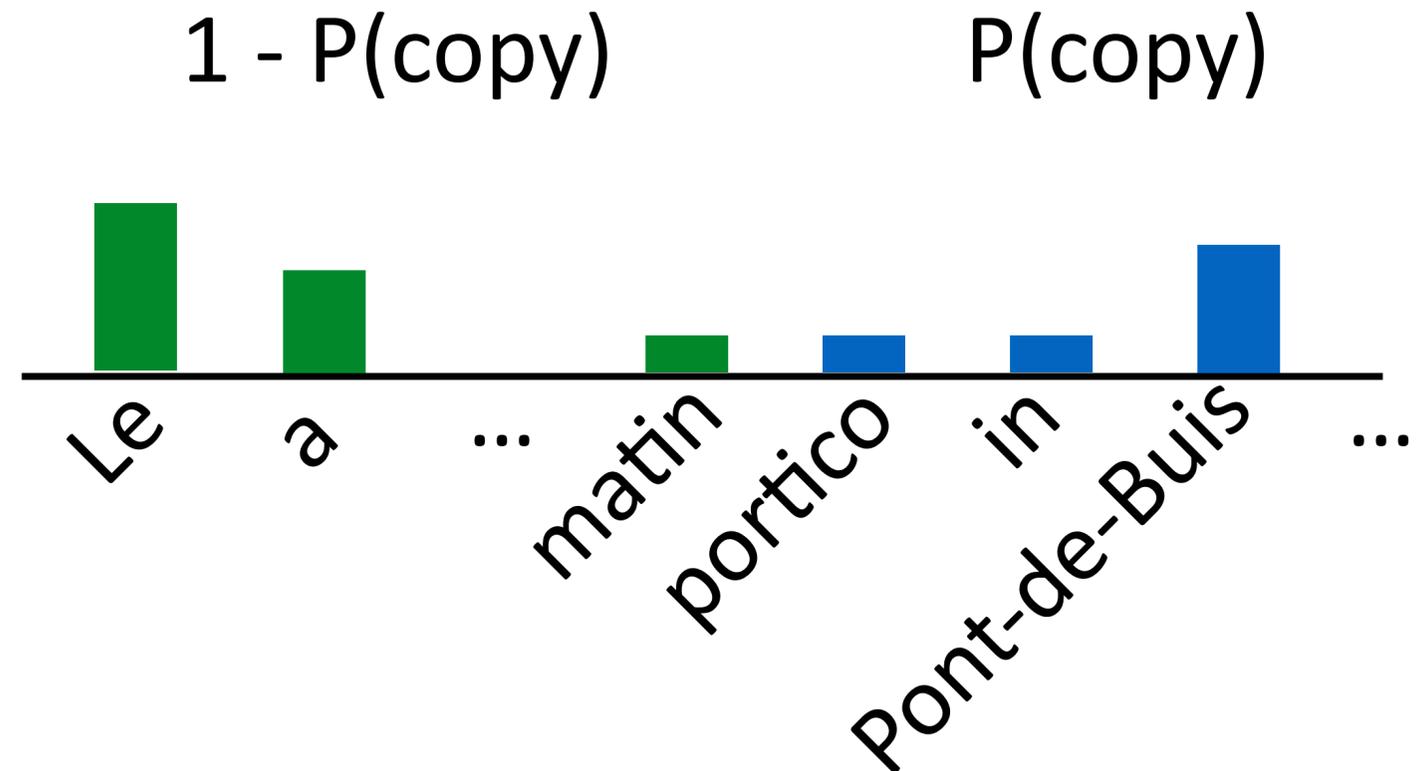


Pointer Generator Mixture Models

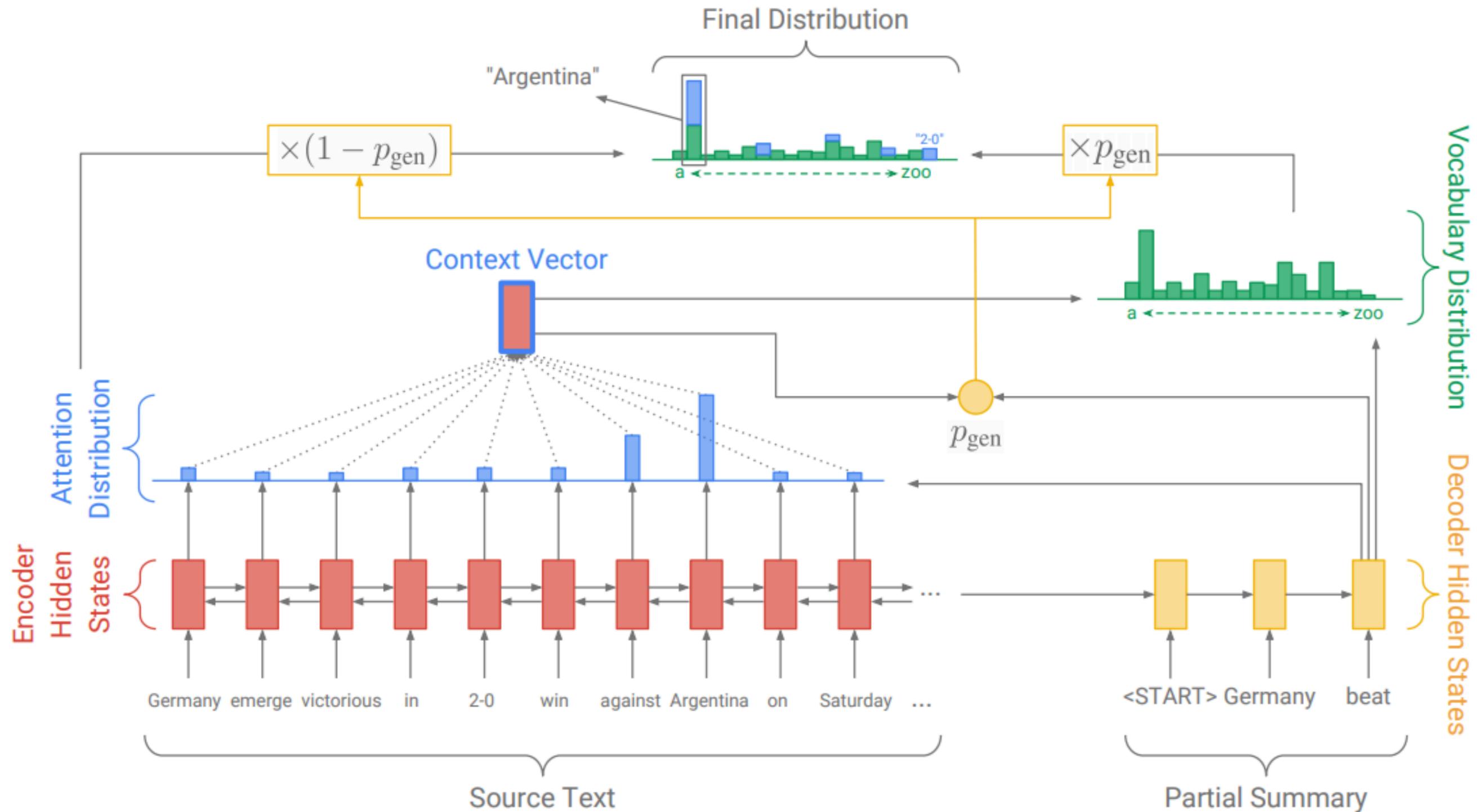
- ▶ Define the decoder model as a mixture model of P_{vocab} and P_{pointer}

$$P(y_i | \mathbf{x}, y_1, \dots, y_{i-1}) = P(\text{copy})P_{\text{pointer}} + (1 - P(\text{copy}))P_{\text{vocab}}$$

- ▶ Predict $P(\text{copy})$ based on decoder state, input, etc.
- ▶ Marginalize over copy variable during training and inference
- ▶ Model will be able to both generate and copy, flexibly adapt between the two

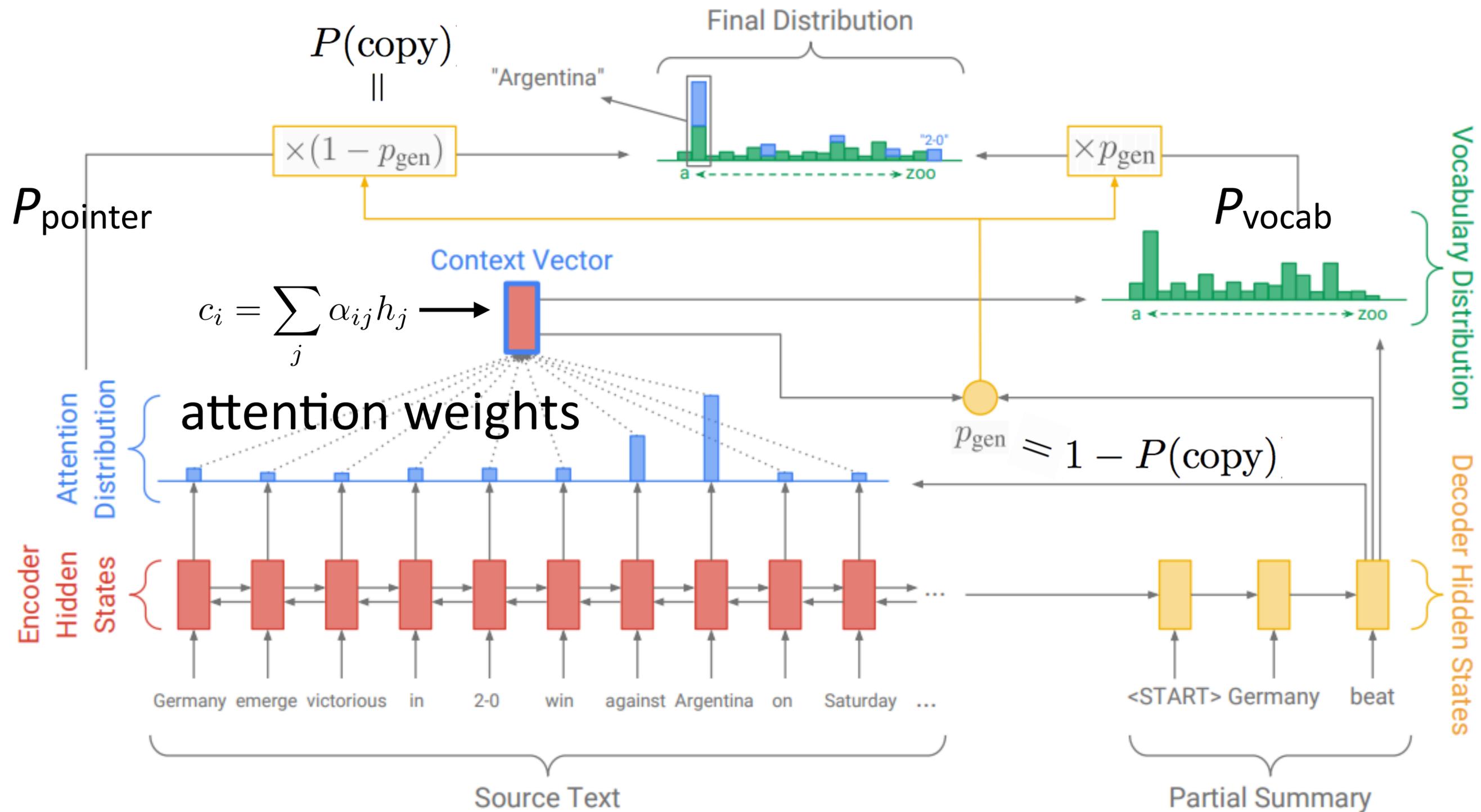


Copying in Summarization



See et al. (2017)

Copying in Summarization



See et al. (2017)

Copying in Summarization

	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	39.53	17.28	36.38	17.32	18.72
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

- ▶ maintain a coverage vector, which is the sum of attention distributions over all previous decoder timesteps

Copying in Summarization

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane amanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to **destabilize nigeria's economy**. UNK says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

Figure 1: Comparison of output of 3 abstractive summarization models on a news article. The baseline model makes **factual errors**, a **nonsensical sentence** and struggles with OOV words *muhammadu buhari*. The pointer-generator model is accurate but **repeats itself**. Coverage eliminates repetition. The final summary is composed from **several fragments**.

Tokenization

Recap: Problems with Seq2seq Models

- ▶ Unknown words:

en: The ecotax portico in Pont-de-Buis , ... [truncated] ... , was taken down on Thursday morning

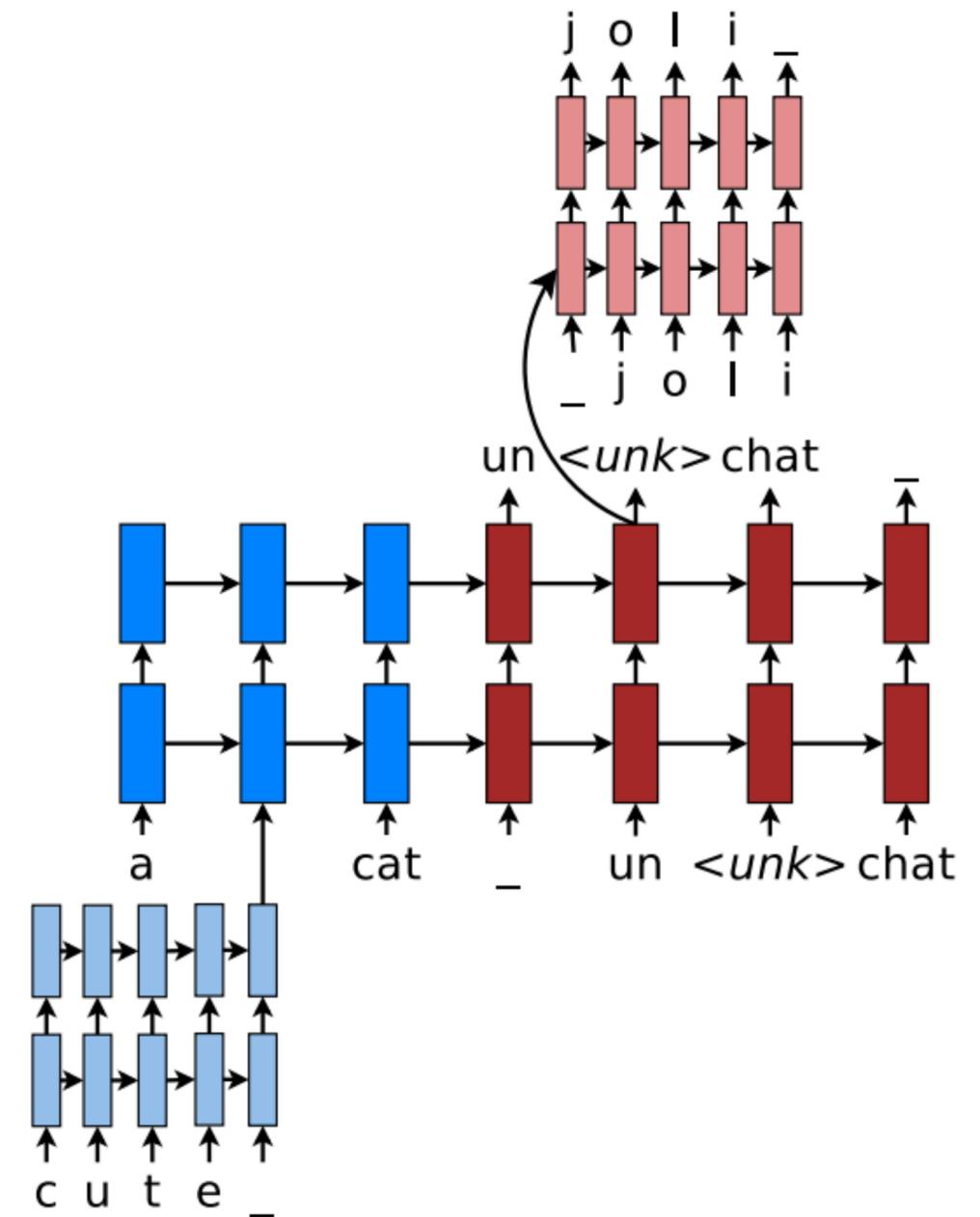
fr: Le portique écotaxe de Pont-de-Buis , ... [truncated] ... , a été démonté jeudi matin

nn: Le unk de unk à unk , ... [truncated] ... , a été pris le jeudi matin

- ▶ Encoding these rare words into a vector space is really hard

Character Models

- ▶ If we predict an *unk* token, generate the results from a character LSTM
- ▶ Can potentially transliterate new concepts, but architecture is more complicated and slower to train
- ▶ Models like this in part contributed to dynamic computation graph frameworks becoming popular



Character Models

initialization of parameters in $[-0.1, 0.1]$, (c) 6-epoch training with plain SGD and a simple learning rate schedule – start with a learning rate of 1.0; after 4 epochs, halve the learning rate every 0.5 epoch, (d) mini-batches are of size 128 and shuffled, (e) the gradient is rescaled whenever its norm exceeds 5, and (f) dropout is used with probability 0.2 according to (Pham et al., 2014). We now detail differences across the three architectures.

Word-based NMT – We constrain our source and target sequences to have a maximum length of 50 each; words that go past the boundary are ignored. The vocabularies are limited to the top $|V|$ most frequent words in both languages. Words not in these vocabularies are converted into $\langle unk \rangle$. After translating, we will perform dictionary⁵ lookup or identity copy for $\langle unk \rangle$ using the alignment information from the attention models. Such procedure is referred as the *unk replace* technique (Luong et al., 2015b; Jean et al., 2015a).

Character-based NMT – The source and target sequences at the character level are often about 5 times longer than their counterparts in the word-based models as we can infer from the statistics in Table 1. Due to memory constraint in GPUs, we

⁵Obtained from the alignment links produced by the Berkeley aligner (Liang et al., 2006) over the training corpus.

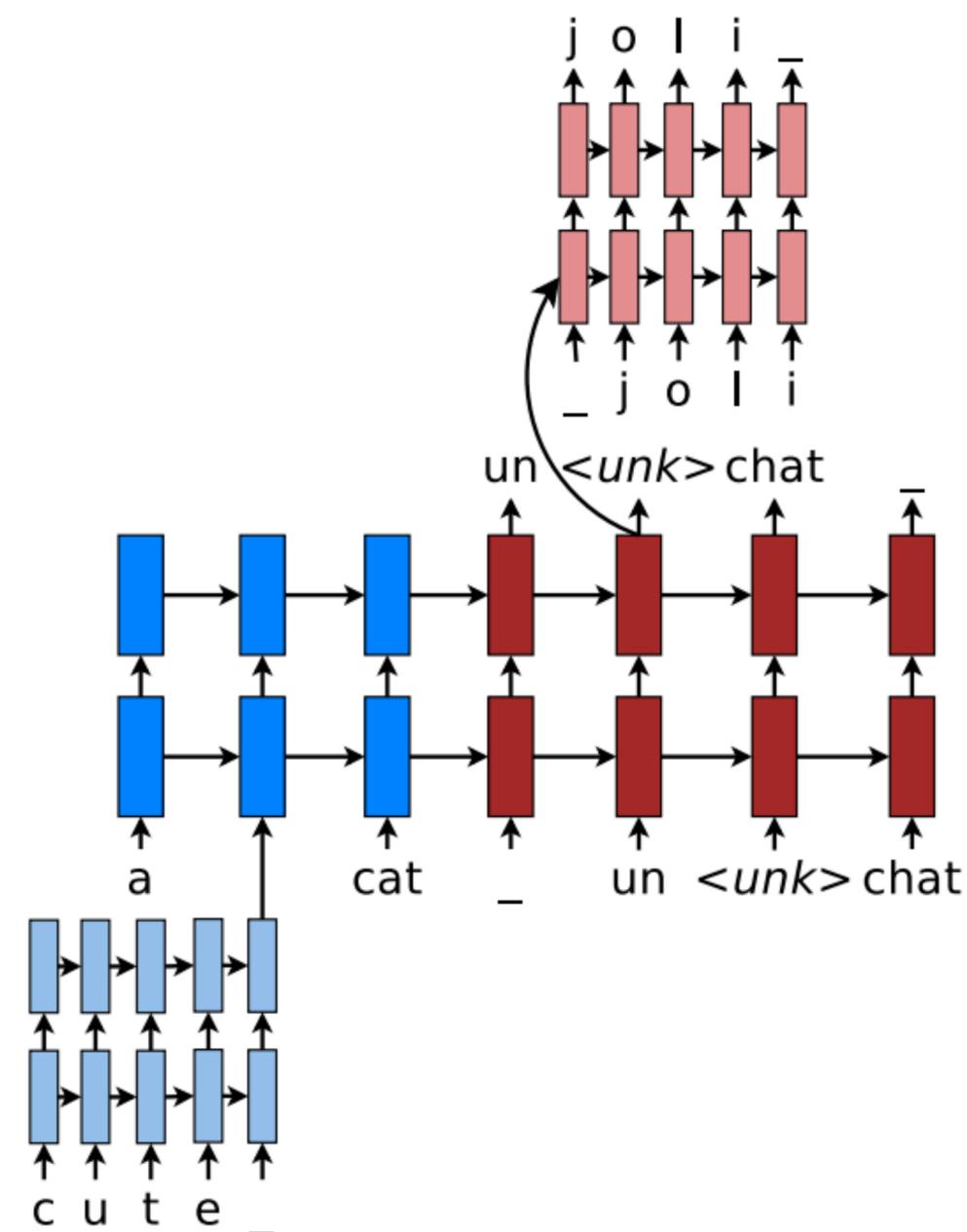
limit our source and target sequences to a maximum length of 150 each, i.e., we backpropagate through at most 300 timesteps from the decoder to the encoder. With smaller 512-dimensional models, we can afford to have longer sequences with up to 600-step backpropagation.

Hybrid NMT – The *word-level* component uses the same settings as the purely word-based NMT. For the *character-level* source and target components, we experiment with both shallow and deep 1024-dimensional models of 1 and 2 LSTM layers. We set the weight α in Eq. (5) for our character-level loss to 1.0.

Training Time – It takes about 3 weeks to train a word-based model with $|V| = 50K$ and about 3 months to train a character-based model. Training and testing for the hybrid models are about 10-20% slower than those of the word-based models with the same vocabulary size.

5.3 Results

We compare our models with several strong systems. These include the winning entry in WMT'15, which was trained on a much larger amount of data, 52.6M parallel and 393.0M mono-



Handling Rare Words

- ▶ Words are a difficult unit to work with: copying can be cumbersome, word vocabularies get very large
- ▶ Character-level models don't work well
- ▶ Solution: “word pieces” (which may be full words but may be subwords)

Input: *the* | ***eco tax*** | *port i co* | *in* | *Po nt - de - Bu is...*

Output: *le* | *port i que* | ***é co tax e*** | *de* | *Pont - de - Bui s*

- ▶ Can help with transliteration; capture shared linguistic characteristics between languages (e.g., transliteration, shared word root, etc.)

Wu et al. (2016)

Byte Pair Encoding (BPE)

- ▶ Start with every individual byte (basically character) as its own symbol
- ▶ Count bigram character cooccurrences
- ▶ Merge the most frequent pair of adjacent characters

Algorithm 1 Byte-pair encoding (Sennrich et al., 2016; Gage, 1994)

```
1: Input: set of strings  $D$ , target vocab size  $k$ 
2: procedure BPE( $D, k$ )
3:    $V \leftarrow$  all unique characters in  $D$ 
4:     (about 4,000 in English Wikipedia)
5:   while  $|V| < k$  do           ▷ Merge tokens
6:      $t_L, t_R \leftarrow$  Most frequent bigram in  $D$ 
7:      $t_{\text{NEW}} \leftarrow t_L + t_R$    ▷ Make new token
8:      $V \leftarrow V + [t_{\text{NEW}}]$ 
9:     Replace each occurrence of  $t_L, t_R$  in
10:       $D$  with  $t_{\text{NEW}}$ 
11:   end while
12:   return  $V$ 
13: end procedure
```

Byte Pair Encoding (BPE)

- ▶ Do this either over your vocabulary (original version) or over a large corpus (more common version)
- ▶ Final vocabulary size is often in 10k ~ 30k range for each language
- ▶ Most NMT systems use this on both source + target
- ▶ BPE tokenization takes the vocabulary V containing ordered merges and applies them to new text in the same order as they occurred during vocabulary construction

Word Pieces

- ▶ Alternatively, can learn word pieces based on unigram LM:

while voc size < target voc size:

- Build a language model over your corpus

- Merge pieces that lead to highest improvement in language model perplexity

- ▶ Result: way of segmenting input appropriate for translation
- ▶ SentencePiece library from Google: unigram LM & BPE
- ▶ Large pre-trained language models are all using this too!

Sennrich et al. (2016), Kudo (2018)

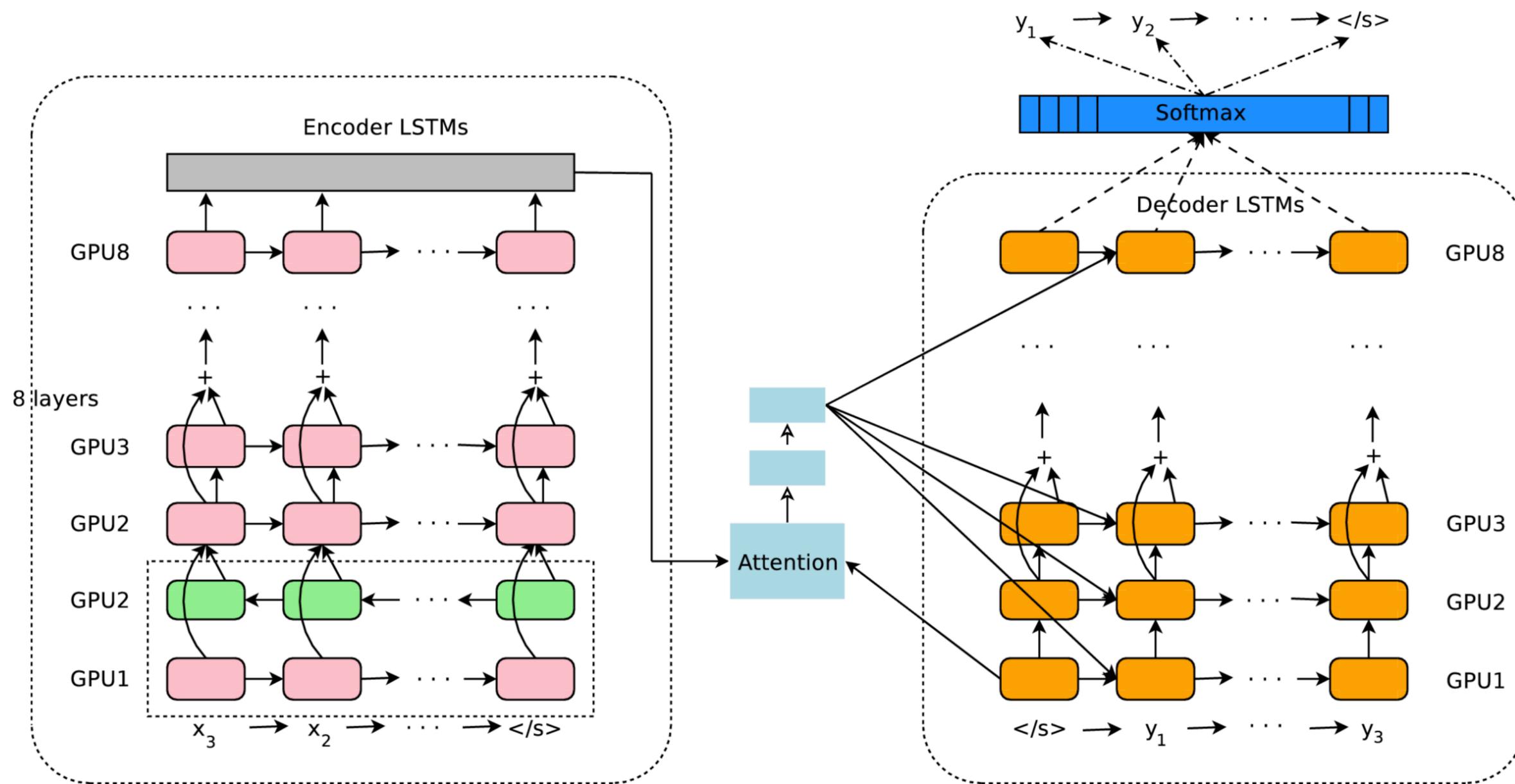
Comparison

	Original:	furiously		Original:	tricycles
(a)	BPE:	_fur iously	(b)	BPE:	_t ric y cles
	Unigram LM:	_fur ious ly		Unigram LM:	_tri cycle s
	Original:	Completely preposterous suggestions			
(c)	BPE:	_Comple t ely	_prep ost erous	_suggest ions	
	Unigram LM:	_Complete ly	_pre post er ous	_suggestion s	

- ▶ BPE produces less linguistically plausible units than word pieces (based on unigram LM)
- ▶ Some evidence that unigram LM works better in pre-trained Transformer models

Google NMT

Google's NMT System



- ▶ 8-layer LSTM encoder-decoder with attention, word piece vocabulary of 8k-32k

Google's NMT System

English-French:

Google's phrase-based system: 37.0 BLEU

Luong+ (2015) seq2seq ensemble with rare word handling: 37.5 BLEU

Google's 32k word pieces: **38.95** BLEU

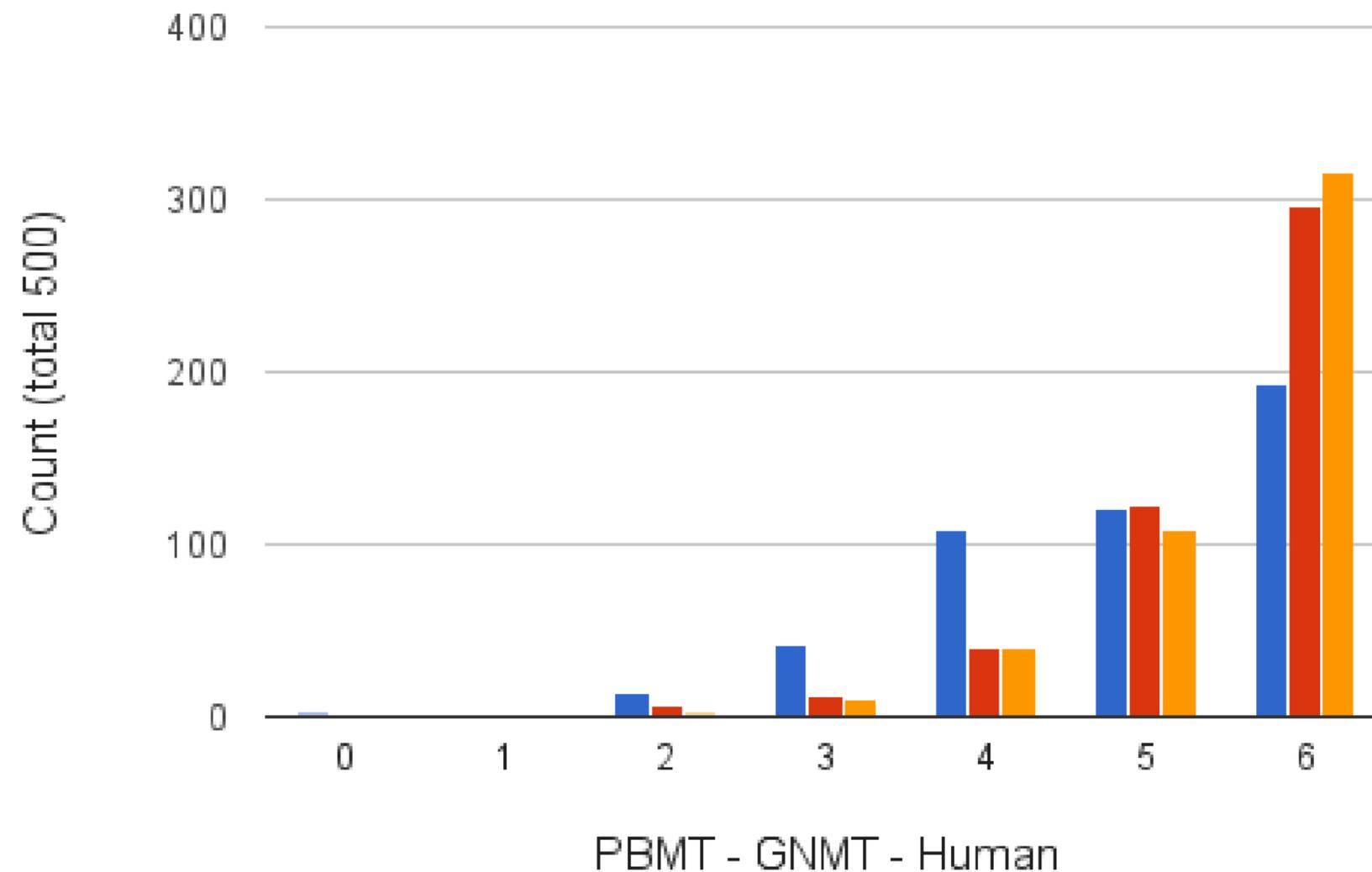
English-German:

Google's phrase-based system: 20.7 BLEU

Luong+ (2015) seq2seq ensemble with rare word handling: 23.0 BLEU

Google's 32k word pieces: **24.2** BLEU

Human Evaluation (En-Es)



- ▶ Similar to human-level performance *on English-Spanish*

Figure 6: Histogram of side-by-side scores on 500 sampled sentences from Wikipedia and news websites for a typical language pair, here English \rightarrow Spanish (PBMT blue, GNMT red, Human orange). It can be seen that there is a wide distribution in scores, even for the human translation when rated by other humans, which shows how ambiguous the task is. It is clear that GNMT is much more accurate than PBMT.

Google's NMT System

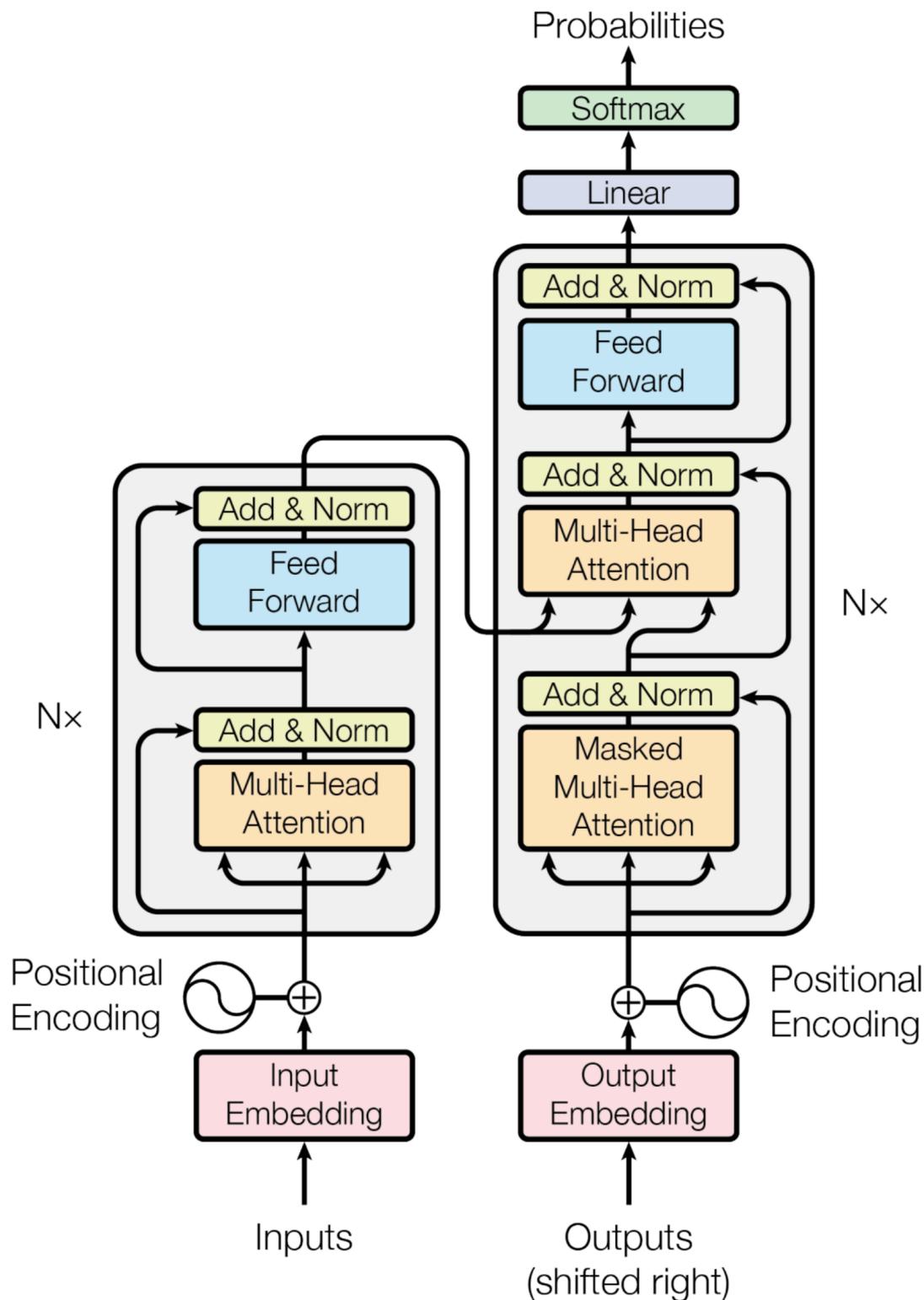
Source	She was spotted three days later by a dog walker trapped in the quarry	
PBMT	Elle a été repéré trois jours plus tard par un promeneur de chien piégé dans la carrière	6.0
GNMT	Elle a été repérée trois jours plus tard par un traîneau à chiens piégé dans la carrière.	2.0
Human	Elle a été repérée trois jours plus tard par une personne qui promenait son chien coincée dans la carrière	5.0

Gender is correct in GNMT
but not in PBMT

“sled” “walker”

The right-most column shows the human ratings on a scale of 0 (complete nonsense) to 6 (perfect translation)

Transformer (more later)



Model	BLEU	
	EN-DE	EN-FR
ByteNet [18]	23.75	
Deep-Att + PosUnk [39]		39.2
GNMT + RL [38]	24.6	39.92
ConvS2S [9]	25.16	40.46
MoE [32]	26.03	40.56
Deep-Att + PosUnk Ensemble [39]		40.4
GNMT + RL Ensemble [38]	26.30	41.16
ConvS2S Ensemble [9]	26.36	41.29
Transformer (base model)	27.3	38.1
Transformer (big)	28.4	41.8

Other MT research

Low-Resource MT

- ▶ Particular interest in deploying MT systems for languages with little or no parallel data
- ▶ BPE allows us to transfer models even without training on a specific language
- ▶ Pre-trained models can help further

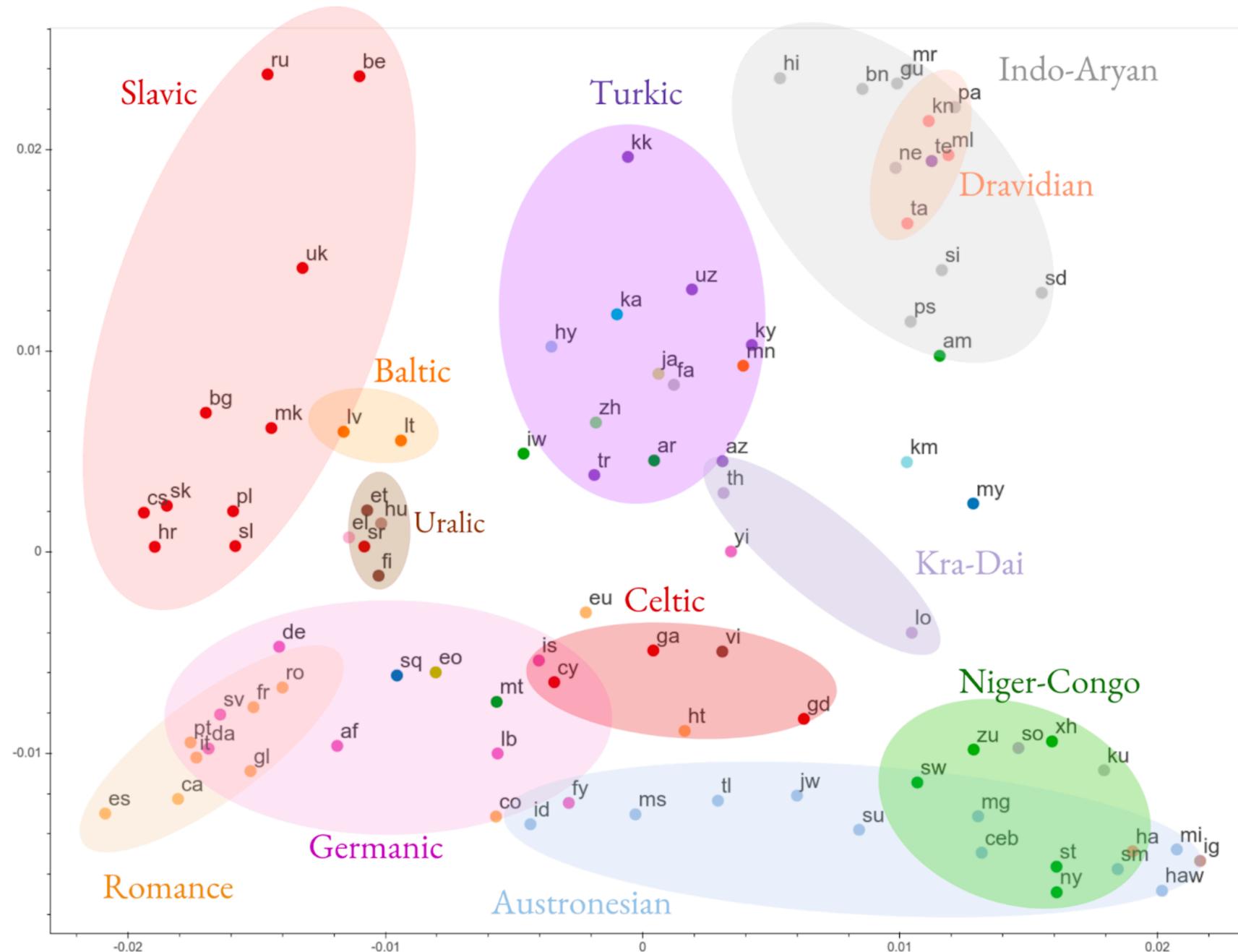
Burmese, Indonesian, Turkish
BLEU

Transfer	My→En	Id→En	Tr→En
baseline (no transfer)	4.0	20.6	19.0
transfer, train	17.8	27.4	20.3
transfer, train, reset emb, train	13.3	25.0	20.0
transfer, train, reset inner, train	3.6	18.0	19.1

Table 3: Investigating the model’s capability to restore its quality if we reset the parameters. We use En→De as the parent.

Massively Multilingual MT

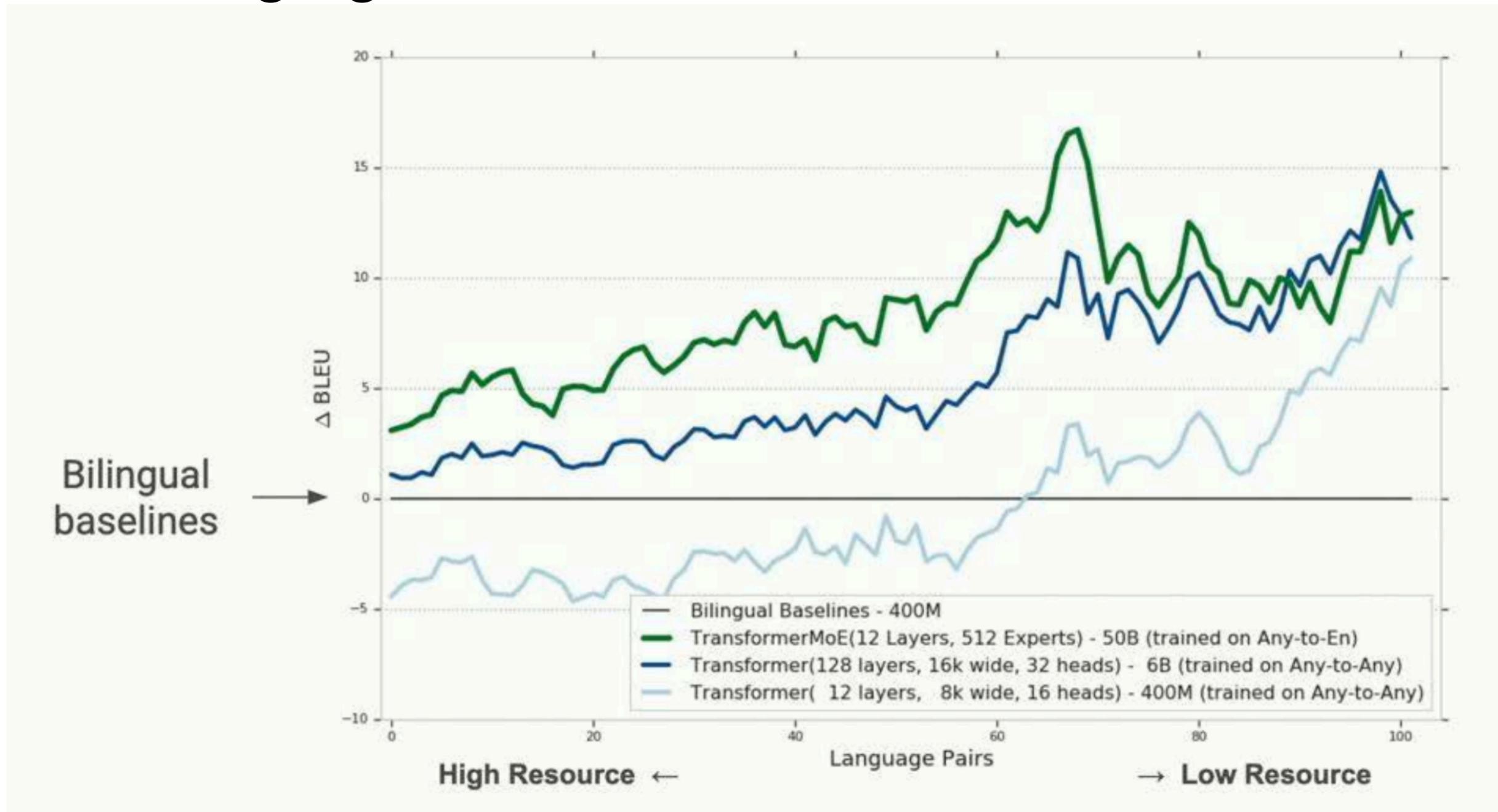
- For 103 languages



Arivazhagan et al. (2019), Kudugunta et al. (2019)

Massively Multilingual MT

- ▶ For 103 languages



Arivazhagan et al. (2019), Kudugunta et al. (2019)

Massively Multilingual MT

- ▶ For 200 languages (54B parameters)
 - ▶ Mixture of Expert (BOE) model. With more low-resource language pairs in the training data, the multilingual systems start to overfit.
 - ▶ Solutions: regularization, curriculum learning, self-supervised learning, and diversifying back-translation.

	eng_Latn-xx			xx-eng_Latn		
	MMTAfrica	M2M-100*	NLLB-200	MMTAfrica	M2M-100*	NLLB-200
hau_Latn	-/-	4.0/-	33.6/53.5	-/-	16.3/-	38.5/57.3
ibo_Latn	21.4/-	19.9/-	25.8/41.4	15.4/-	12.0/-	35.5/54.4
lug_Latn	-/-	7.6/-	16.8/39.8	-/-	7.7/-	27.4/46.7
luo_Latn	-/-	13.7/-	18.0/38.5	-/-	11.8/-	24.5/43.7
swh_Latn	40.1/-	27.1/-	37.9/58.6	28.4/-	25.8/-	48.1/66.1
wol_Latn	-/-	8.2/-	11.5/29.7	-/-	7.5/-	22.4/41.2
xho_Latn	27.1/-	-/-	29.5/48.6	21.7/-	-/-	41.9/59.9
yor_Latn	12.0/-	13.4/-	13.8/25.5	9.0/-	9.3/-	26.6/46.3
zul_Latn	-/-	19.2/-	36.3/53.3	-/-	19.2/-	43.4/61.5

Table 31: Comparison on FLORES-101 devtest on African Languages. We compare to two

Fan et al. (2022), NLLB Team (2022)

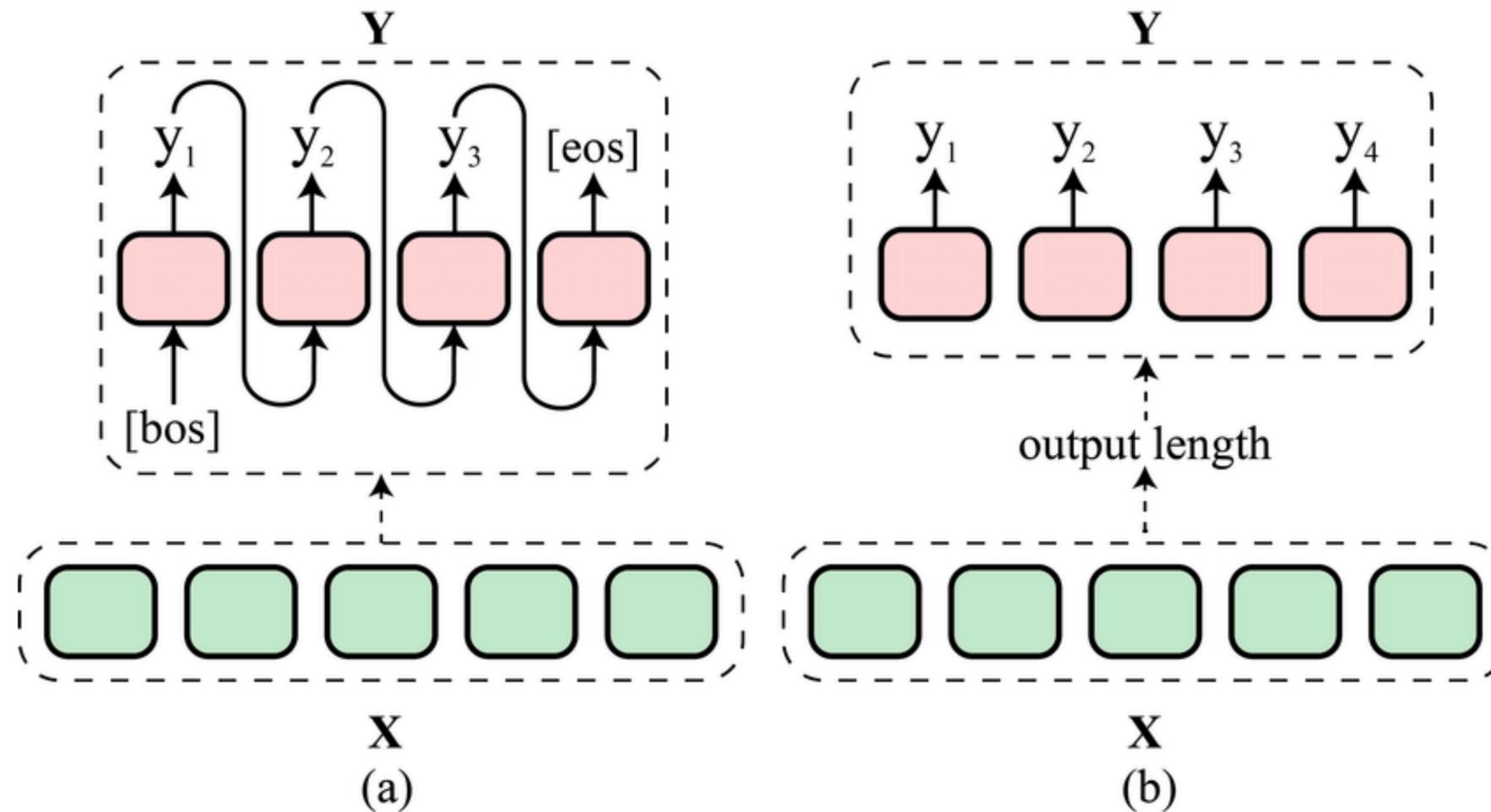
Unsupervised MT

Approach	Train/Val	Test	Loss
Supervised MT	L1-L2	L1-L2	$\mathcal{L}_{x \rightarrow y}^{MT} = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim (\mathcal{X}, \mathcal{Y})} [-\log p_{x \rightarrow y}(\mathbf{y} \mathbf{x})]$
Unsupervised MT	L1, L2	L1-L2	$\mathcal{L}_{x \leftrightarrow y}^{BT} = \mathbb{E}_{\mathbf{x} \sim \mathcal{X}} [-\log p_{y \rightarrow x}(\mathbf{x} g^*(\mathbf{x}))]$ $+ \mathbb{E}_{\mathbf{y} \sim \mathcal{Y}} [-\log p_{x \rightarrow y}(\mathbf{y} h^*(\mathbf{y}))]$ $g^*, h^* : \text{sentence predictors}$

- ▶ Common principles of unsupervised MT
 - ▶ Language models
 - ▶ (Iterative) Back-translation! - The goal of this model is to generate a source sentence for each target sentence in the monolingual corpus.

Lample et al. (2018)

Non-Autoregressive NMT



Gu et al. (2018), Ghazvininejad et al. (2019), Kasai et al. (2020)

WMT 2024

WMT 2024

[IMPORTANT DATES](#)

[OVERVIEW AND SHARED TASKS](#)

[PAPER SUBMISSION INFORMATION](#)

[Research Papers](#)

[System Papers](#)

[Presenting and poster format](#)

[ANNOUNCEMENTS](#)

[REGISTRATION AND VISA INFORMATION](#)

[ANTI-HARASSMENT POLICY](#)

[PAST CONFERENCES](#)

[ORGANIZERS](#)

[Invited Talk](#)

[Program Committee](#)

[Contact](#)

Conference

15-16 November, 2024

All deadlines are in AoE (Anywhere on Earth). Dates are specified with respect to EMNLP 2024.

OVERVIEW AND SHARED TASKS

This year's conference will feature the following shared tasks:

- a [general translation task \(former News task\)](#),
- a [translation into low-resource languages of Spain task](#),
- a [Low-Resource Indic Language Translation](#),
- a [Metrics Task](#),
- a [Chat Translation Task](#)
- a [Biomedical Translation Task](#)
- a [MT test suites sub-task](#)
- an [Open Language Data shared task](#)
- a [Quality Estimation Task](#)
- a [English-LowRes Multimodal Translation Task](#)
- a [Non-Repetitive Translation Task](#)
- a [Patent Task](#)
- a [Literary Translation Task](#)
- a [Patent Task](#)

WMT 2024

EMNLP 2024

NINTH CONFERENCE ON MACHINE TRANSLATION (WMT24)

November 15-16, 2024
Miami, Florida, USA

[\[HOME\]](#) [\[PROGRAM\]](#) [\[PAPERS\]](#) [\[AUTHORS\]](#)

TRANSLATION TASKS: [\[GENERAL MT \(NEWS\)\]](#) [\[LOW-RESOURCE LANGUAGES OF SPAIN\]](#) [\[INDIC MT\]](#) [\[CHAT TASK\]](#) [\[BIOMEDICAL\]](#) [\[MULTIINDIC22MT TASK\]](#)
[\[ENGLISH-TO-LOWRES MULTIMODAL MT TASK\]](#) [\[NON-REPETITIVE\]](#) [\[PATENT\]](#) [\[LITERARY\]](#)
EVALUATION TASKS: [\[METRICS TASK\]](#) [\[MT TEST SUITES\]](#) [\[QUALITY ESTIMATION\]](#)
OTHER TASKS: [\[OPEN LANGUAGE DATA INITIATIVE\]](#)

Subtask: MT Test Suites. "Help us break LLMs"

The “Test suites” sub-task will be again part of the WMT for the seventh consequent year.

What are the test suites?

Test suites are custom extensions to standard General Translation Task test sets constructed so that they can focus on particular aspects of the MT output. Test suites also evaluate these aspects in their custom way.

The test suite composition and its evaluation relies fully on the test suite provider — e.g. you. One can also call test suites an “unshared task”, where every participant is doing something different but related.

For our purpose, every test suite includes:

1. a source-side test set (i.e. a set of paragraphs in one language), and
2. an evaluation service (not necessarily automatic, not necessarily relying on any reference translations) for a particular target language.

As opposed to the standard evaluation process, which employs randomly selected test sets, test suites focus on particular translation phenomena or cases that consist of specific challenges to MT. Additionally, as opposed to having generic quality scores, test suites often produce separate fine-grained results for each case.

Since the usage of LLMs for translation is getting more popular, and we are expecting more LLMs participations in WMT this year, the theme of this year’s test suite sub-task is "Help us break LLMs", i.e. to reveal weaknesses and serious flaws of LLMs when translating, hidden within the overall high-quality generation.

Takeaways

- ▶ Can build MT systems with LSTM encoder-decoders, CNNs, or Transformers
- ▶ Word piece / byte pair models are really effective and easy to use
- ▶ State of the art systems are getting pretty good, but lots of challenges remain, especially for low-resource settings

Evaluating Robustness of Large Language Models with Neologisms (NeoBench)



Jonathan Zheng



Alan Ritter



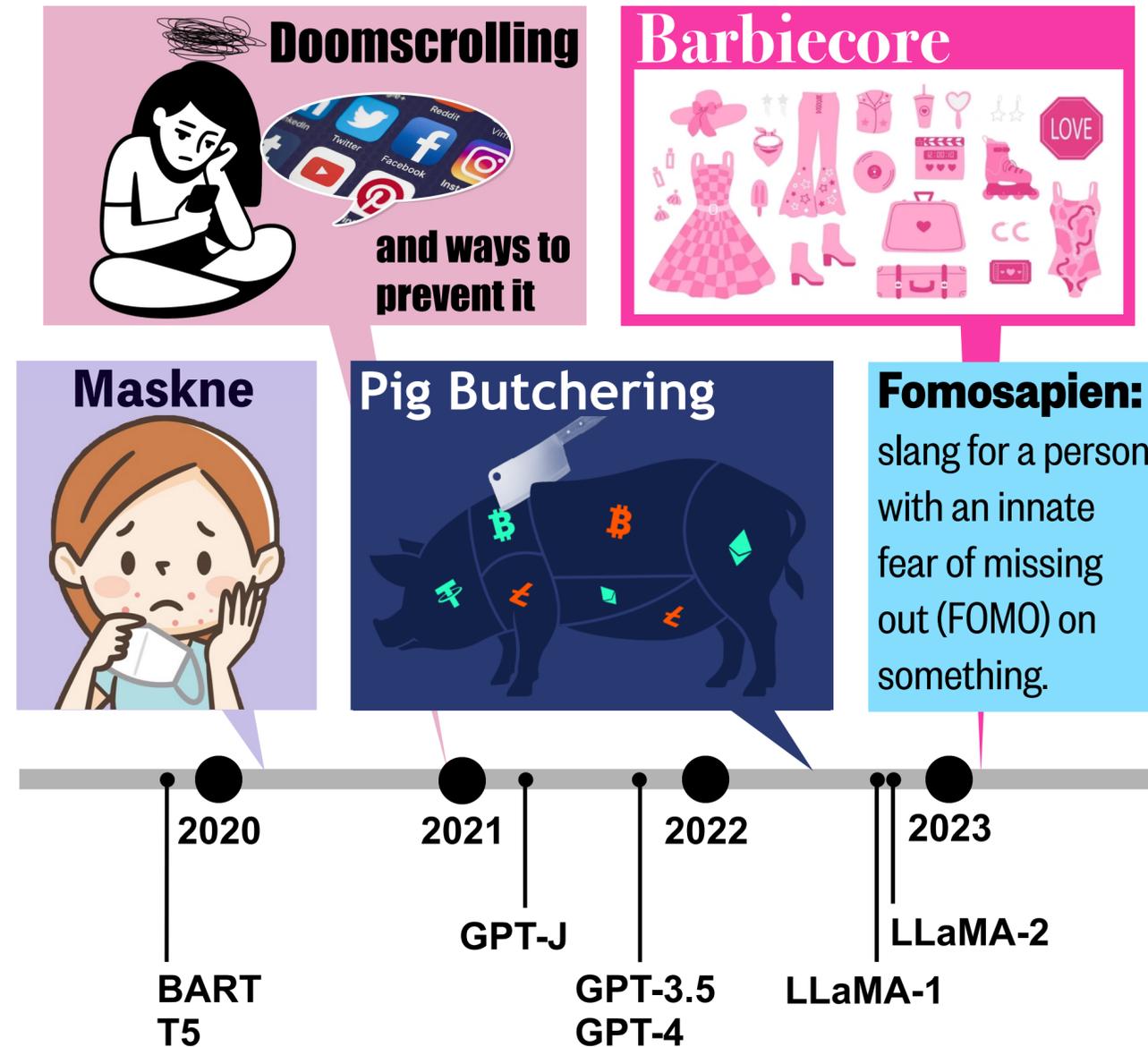
Wei Xu

A better technical solution for
marker-based label projection



NeoBench — evolving human languages

Data contamination, long-tail low-frequency words, tokenization, ...



We used 3 different methods to obtain 2,505 single- and multi-word neologisms.



NeoBench — human evaluation on translation

Models struggle to translate sentences that contains neologism (vs. non-neologism) word.

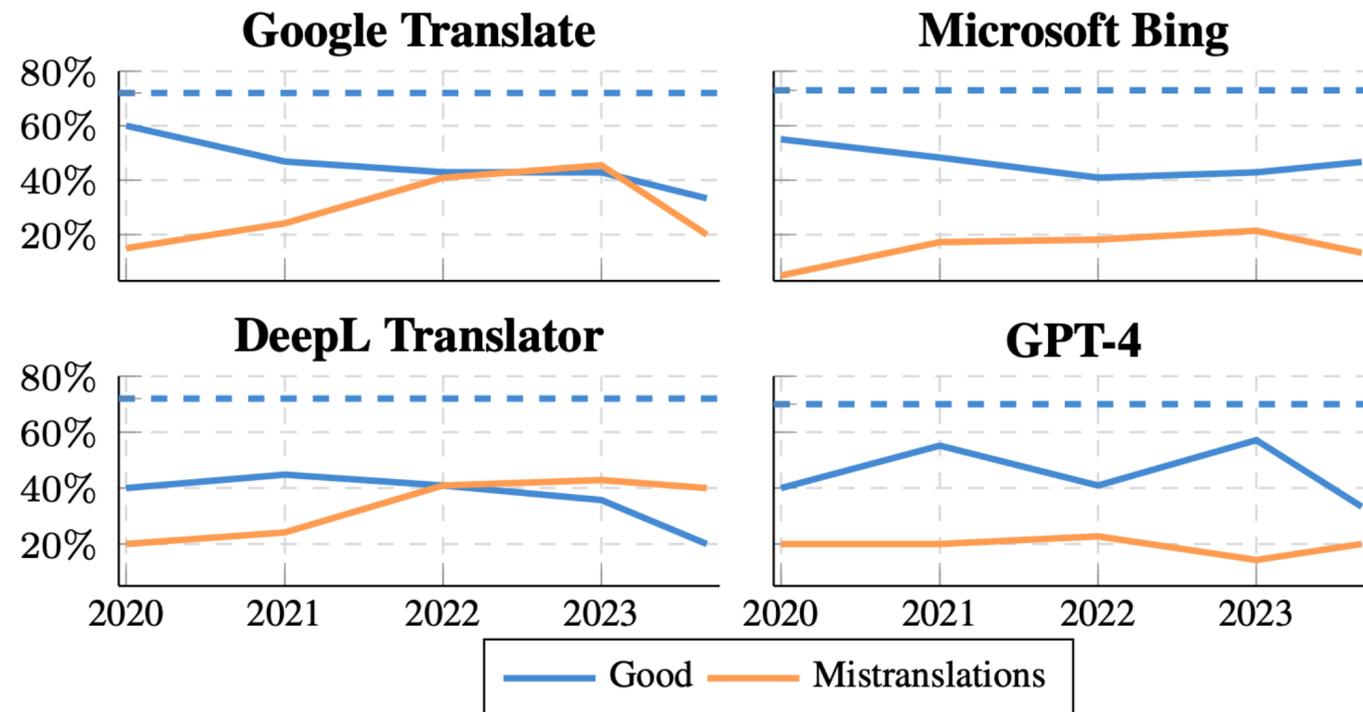


Figure 3: Percentage of good translations and mistranslations of neologism sentences over time. The dashed line represents the percentage of good translations achieved on non-neologism sentences.

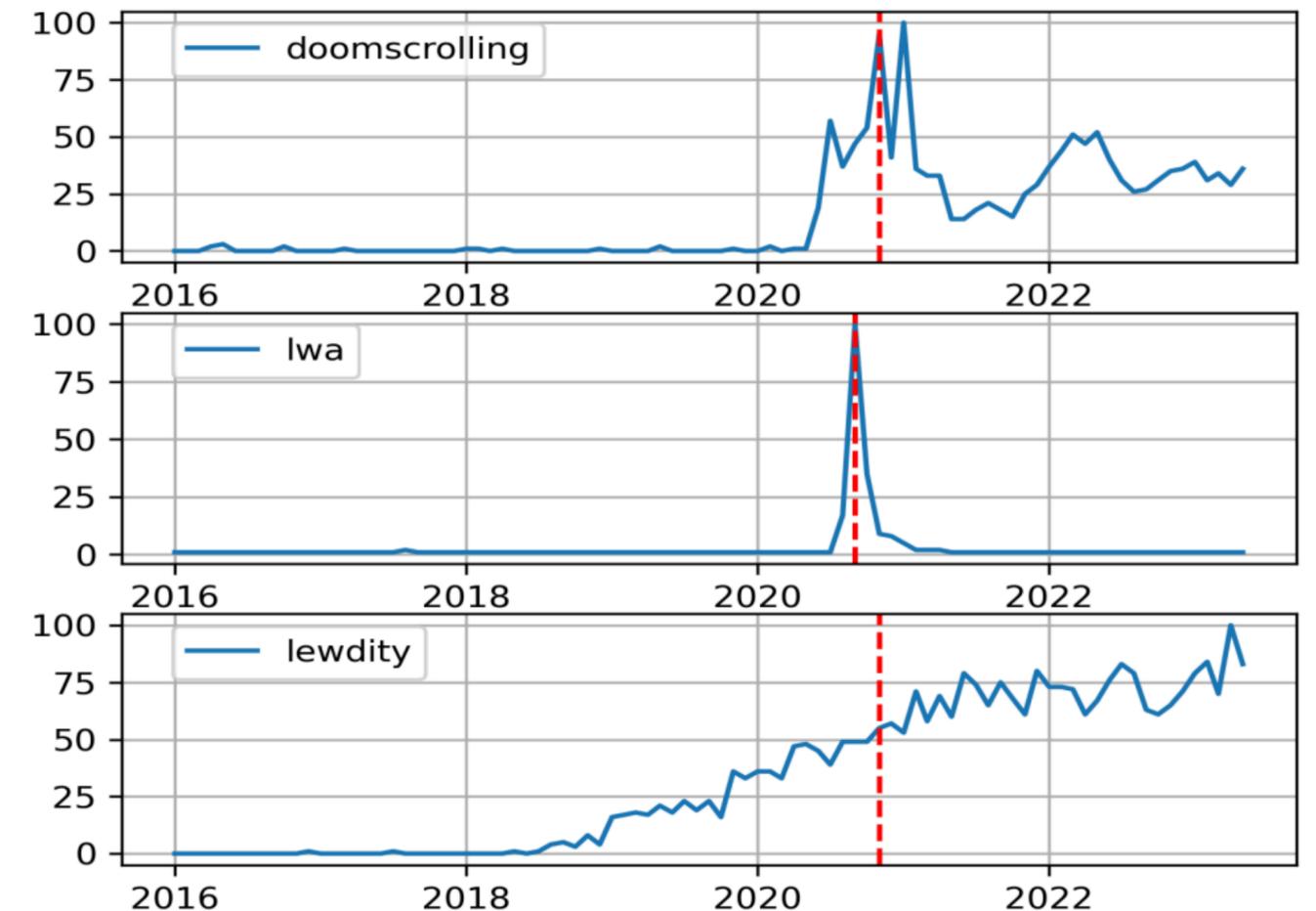


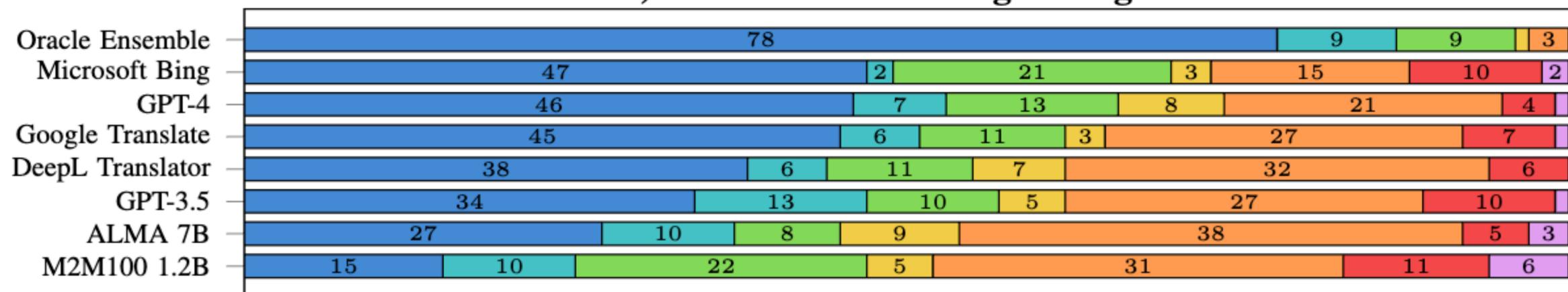
Figure 4: Example Google Trend lines measuring neologism prevalence. The dashed line estimates the date a neologism becomes popular while not yet conventional.



NeoBench — human evaluation on translation

Models struggle to translate sentences that contains neologism (vs. non-neologism) word.

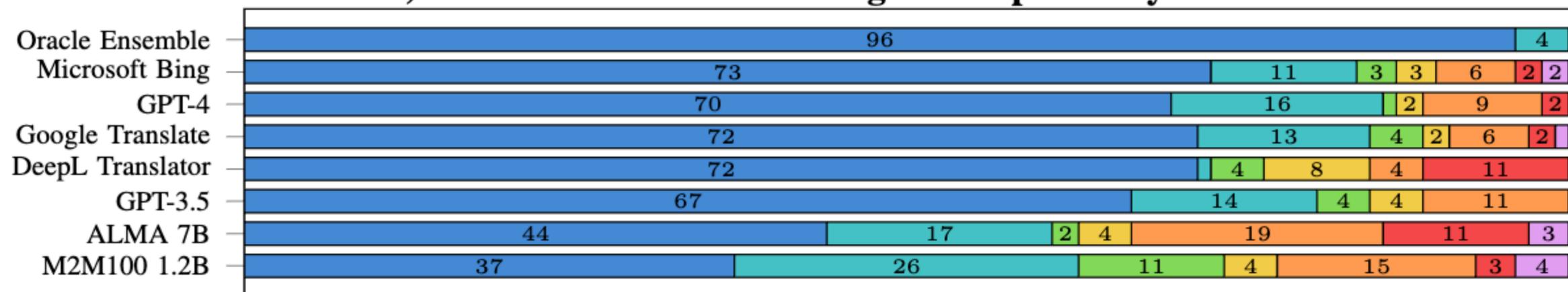
a) Sentences Containing Neologisms



Example:

Starting to think **doomscrolling** through the fall of civilization is having a negative effect on my mental health.

b) Same Sentences with Neologisms Replaced by Common Words



Example:

Starting to think **smoking** through the fall of civilization is having a negative effect on my mental health.



Thresh — Fine-grained Text Evaluation Tool

<https://github.com/davidheineman/thresh>

thresh.tools

A Unified, Customizable and Deployable Platform
for Fine-Grained Text Evaluation

Prompt (human-written):

France's former President Nicolas Sarkozy was found guilty of corruption on Monday and sentenced to three years in prison, a stunning fall from grace for a man who led his country and bestrode the world stage for five years.

Continuation:

Given the string of accusations against Mr Sarkozy and his hardline France, it could be argued that the 6/7 AIPAC scandal and the background of the president's class political career were alarming indeed. Some of Mr Sarkozy's closest aides had openly suggested that his remarks — in which he dismissed other American Democrats as "crazies" who used unsavoury language — sounded like he had been flirting with suicide. Nevertheless, - or perhaps because he did not believe in lost his job — he himself lost his job rather than resign from office in 2011 and became the president without having to resign. In 2008, however, he became a totally shut-door candidate for the presidential election.



NeoBench — perplexity, Cloze QA, definition

a) Definition Generation Output Examples	
Stablecoin	<p>Reference Definition: <i>A stablecoin is a type of cryptocurrency where the value of the digital asset is supposed to be pegged to a reference asset, which is either fiat money, exchange-traded commodities, or another cryptocurrency.</i></p> <p>Model Output (Correct): <i>Stablecoins are cryptocurrencies designed to maintain a stable value, typically by pegging their value to a specific asset or basket of assets, such as the US dollar, gold, or a combination of assets.</i></p>
Angel Shot	<p>Reference Definition: <i>An angel shot is a code to inform a bartender that a customer is not safe and needs assistance.</i></p> <p>Model Output (Incorrect): <i>An angel shot is a cocktail made with whiskey and cream, served in a shot glass.</i></p>
b) Machine Translation Output Examples	
Longcovid	<p>Input: Each reinfection increases the risk of longcovid, hospitalization, & death.</p> <p>Model Output (Correct): 每次再感染都会增加长新冠病毒、住院和死亡的风险。 <i>(Every reinfection increases the risk of long COVID, hospitalization, and death.)</i></p> <p>Human Translation: 每一次新冠感染都会提高出现后遗症、住院治疗，甚至死亡的风险。 <i>(Each COVID-19 infection increases the risk of developing sequelae, hospitalization, and even death.)</i></p>
Doomscrolling	<p>Input: Starting to think doomscrolling through the fall of civilization is having a negative effect on my mental health.</p> <p>Model Output (Incorrect): 开始认为在文明的衰落中滚动的厄运对我的心理健康产生了负面影响。 <i>(Start to think that the doom rolling in the decline of civilization is having a negative impact on my mental health.)</i></p> <p>Human Translation: 开始觉得，刷关于文明衰败的负能量新闻对我的心理健康产生了负面影响。 <i>(Starting to feel that scrolling through negative news about the decline of civilization is having a negative impact on my mental health.)</i></p>

Table 3: Example model definitions and translations for NEO-BENCH tasks. “Doomscrolling” is the act of spending an excessive amount of time reading negative news online. (English translations are shown for information only.)



NeoBench — automatic eval on translation

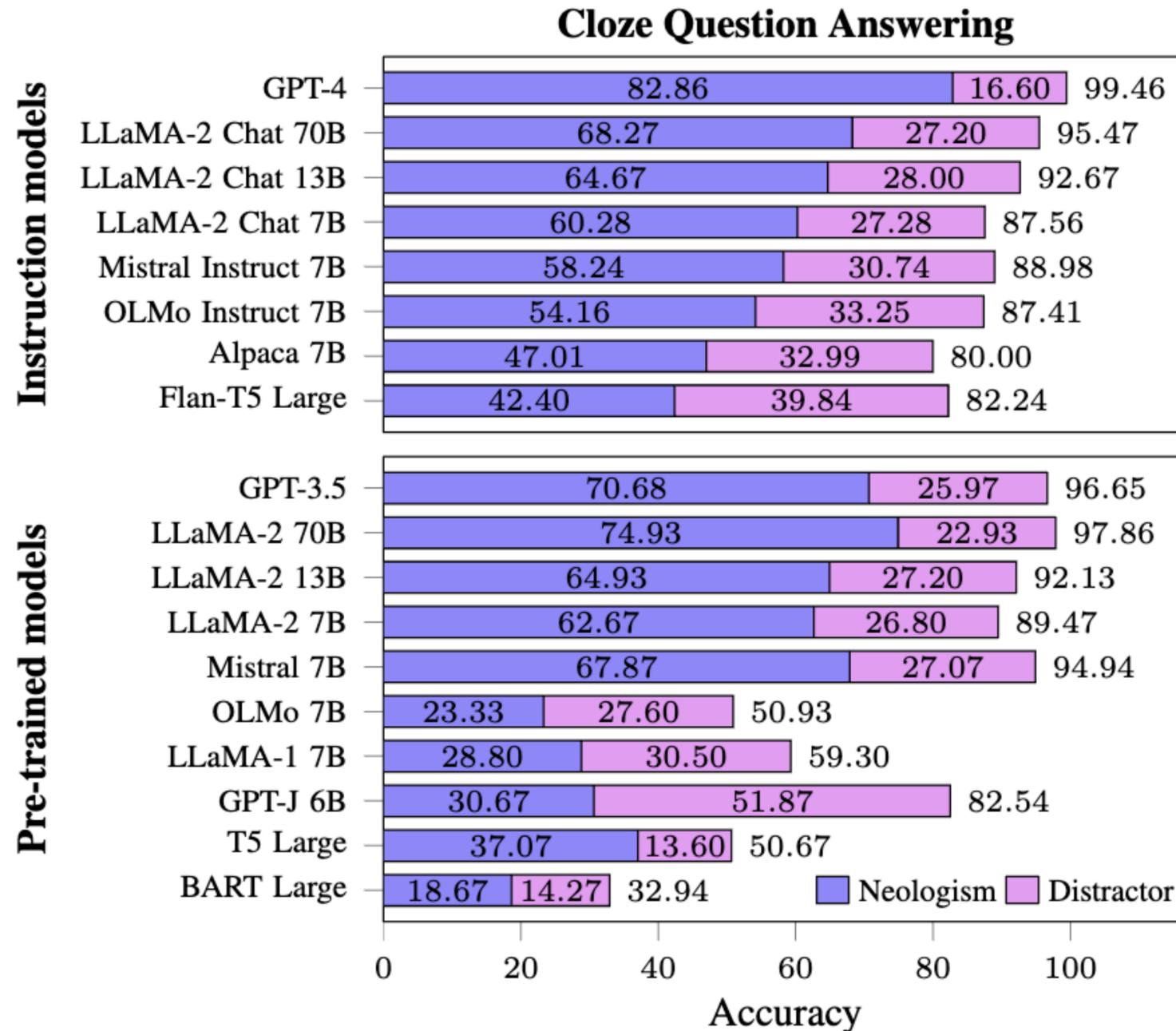
Automatic evaluation metrics do not show good system-level correlations with human evaluation.

Model (human rank)	Reference-Based Metrics				Reference-Free Metrics		
	BLEU \uparrow	COMET \uparrow	MX-23 _{XXL} \downarrow	MX-23 _{XL} \downarrow	COMET _{KIWI} \uparrow	MX-QE _{XXL} \downarrow	MX-QE _{XL} \downarrow
Bing Translator (1)	0.452 (2)	0.825 (5)	2.419 (6)	2.343 (6)	0.788 (5)	1.679 (5)	2.246 (5)
GPT-4 (2)	0.446 (3)	0.854 (1)	1.550 (1)	1.793 (1)	0.793 (3)	1.432 (3)	2.089 (3)
Google Translate (3)	0.507 (1)	0.853 (2)	1.825 (4)	1.945 (4)	0.800 (2)	1.429 (2)	1.940 (1)
DeepL Translator (4)	0.406 (4)	0.842 (3)	1.775 (3)	1.901 (3)	0.807 (1)	1.260 (1)	1.944 (2)
GPT-3.5 (5)	0.399 (5)	0.841 (4)	1.705 (2)	1.796 (2)	0.792 (4)	1.467 (4)	2.157 (4)
ALMA 7B (LLaMA-2) (6)	0.285 (7)	0.801 (6)	2.382 (5)	2.251 (5)	0.746 (6)	2.038 (6)	2.462 (6)
M2M100 1.2B (7)	0.337 (6)	0.776 (7)	3.454 (7)	3.142 (7)	0.745 (7)	2.821 (7)	2.976 (7)
Spearman's ρ	0.244	0.445	0.457	0.380	0.491	0.451	0.445



NeoBench — perplexity, Cloze QA, definition

Newer, larger LLMs work better; but, perplexity becomes worse after instruction tuning.



Neologism: *doomscrolling*

The silver lining of this website no longer functioning as an even vaguely reliable information source is that ___ has basically been completely undermined. It wouldn't even work now since everything is too geared to outrage clickbait and actual reporting has disappeared, so there is no point staying on the app.

- a) misinformation
- b) surfing
- c) **doomscrolling**
- d) lying
- e) gaming
- f) **anti-productivity (distractor)**

Table 2: Example passage in NEO-BENCH for multiple-choice Cloze Question Answering with correct neologism answers and partially correct distractor answers.



NeoBench — perplexity, Cloze QA, definition

Newer, larger LLMs work better; but, perplexity becomes worse after instruction tuning.

