CS378: Natural Language Processing
Lecture 19: MT 2, Seq2seq Models

Greg Durrett

TEXAS
The University of Texas at Austin

Recap

Outline

- Armed with the idea of language models (P(**w**)) and Transformers (good models for this), we still need to actually put together an MT system

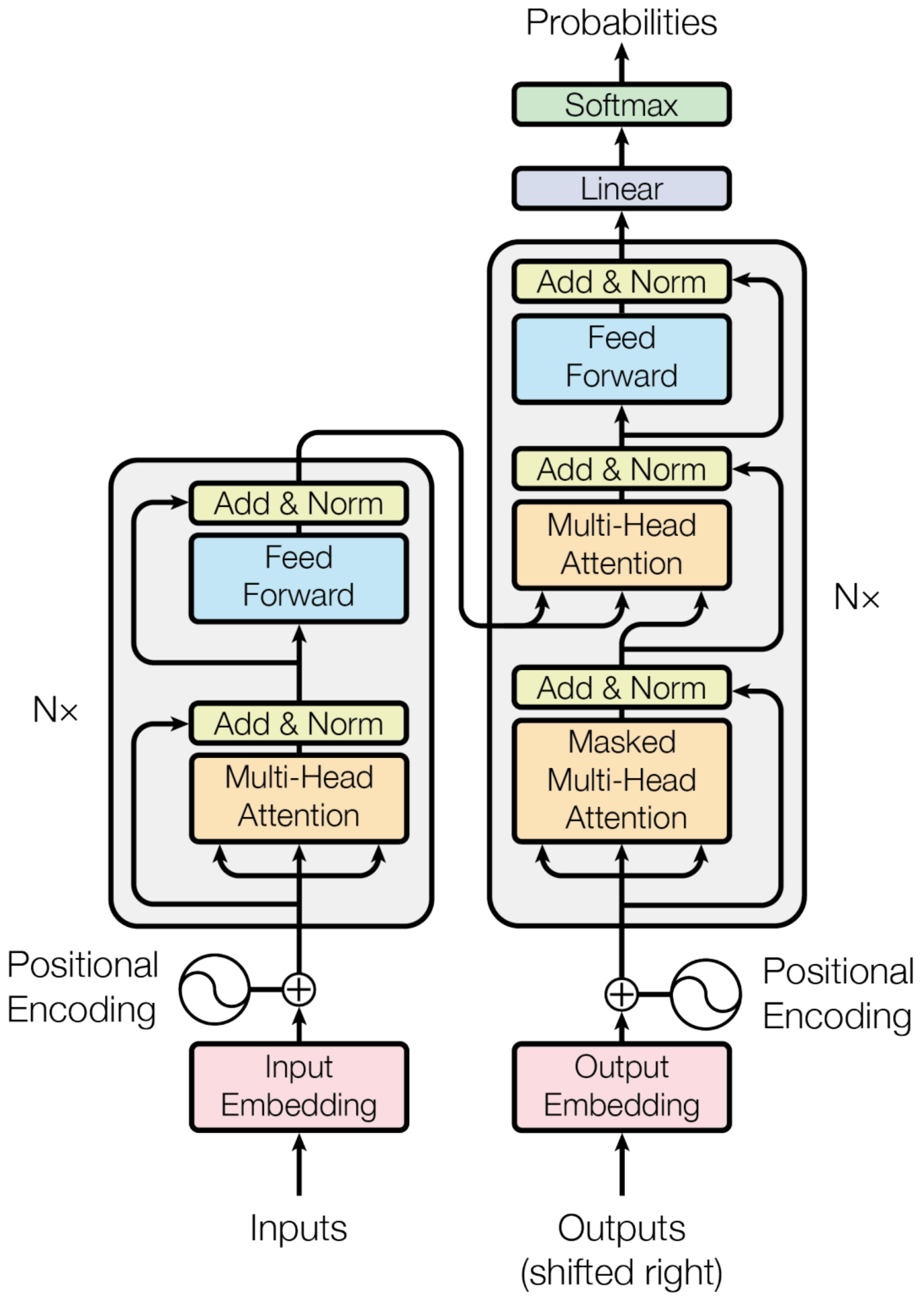
- Sequence-to-sequence (seq2seq) models: we define these as distributions P(**y**|**x**) and decide how to train and do inference. Training looks like LM training, inference is new.

- Subword tokenization: key practical implementation detail

Vaswani et al. (2017)

Seq2seq Models

Transformers: Complete Model

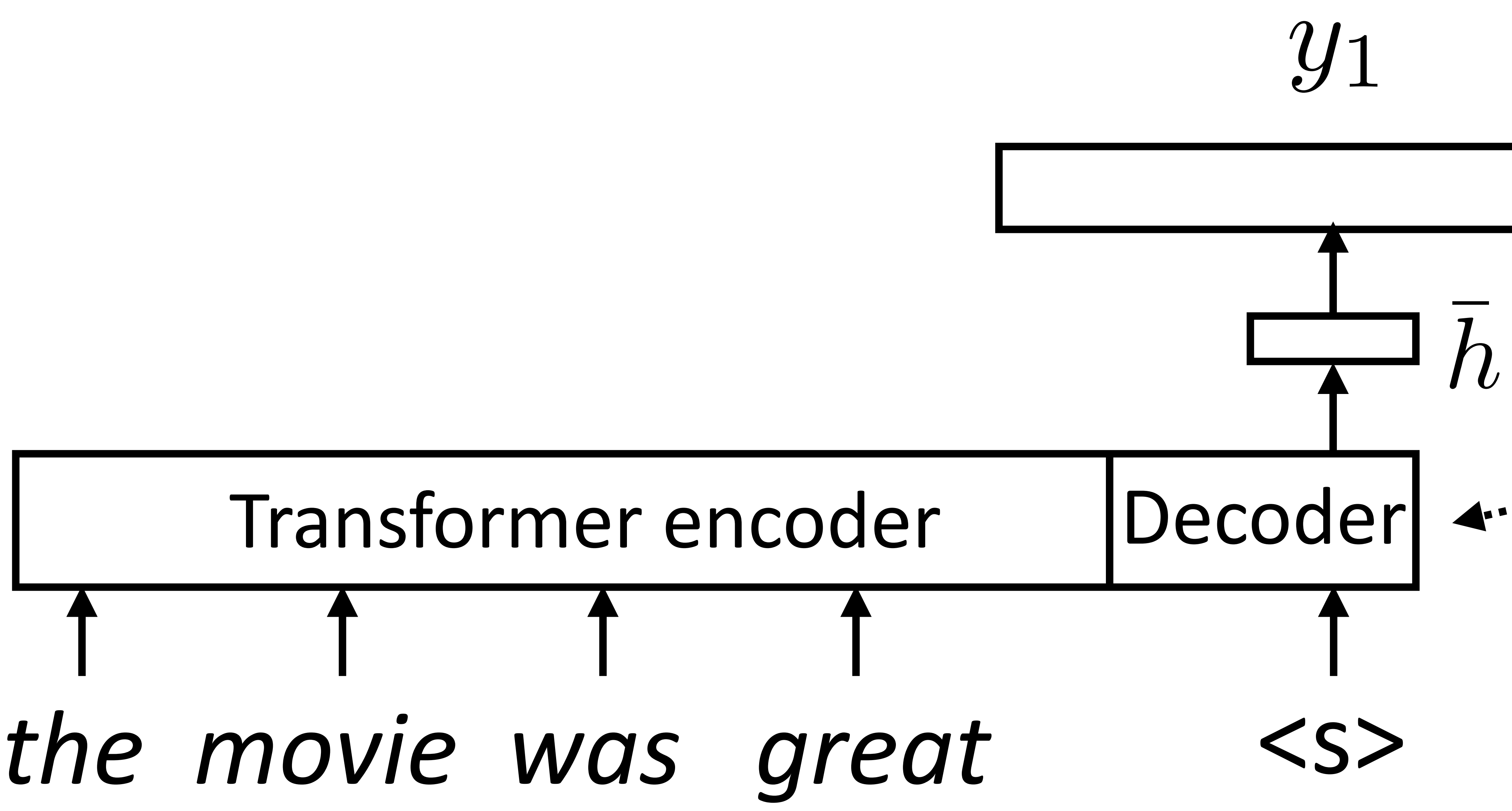


- Transformer encoder (A4) + decoder (looks back at encoder, but similar architecture)
- Decoder alternates attention over the output and attention over the input as well
- Decoder **consumes the previous generated tokens**. You need to run the whole decoder to predict token 1 of the output, then run the decoder again to predict token 2, etc.

Vaswani et al. (2017)

Seq2seq Model

- Generate next word conditioned on previous words (like a language model) **and** conditioned on the source
- W size is |vocab| x |hidden state|, softmax over entire vocabulary

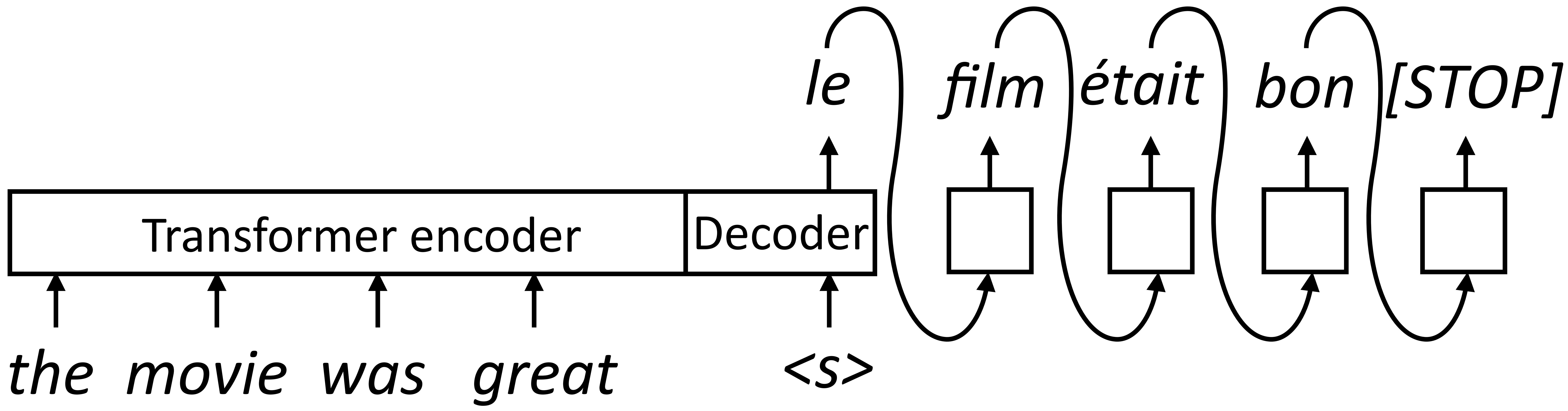


$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h})$$

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

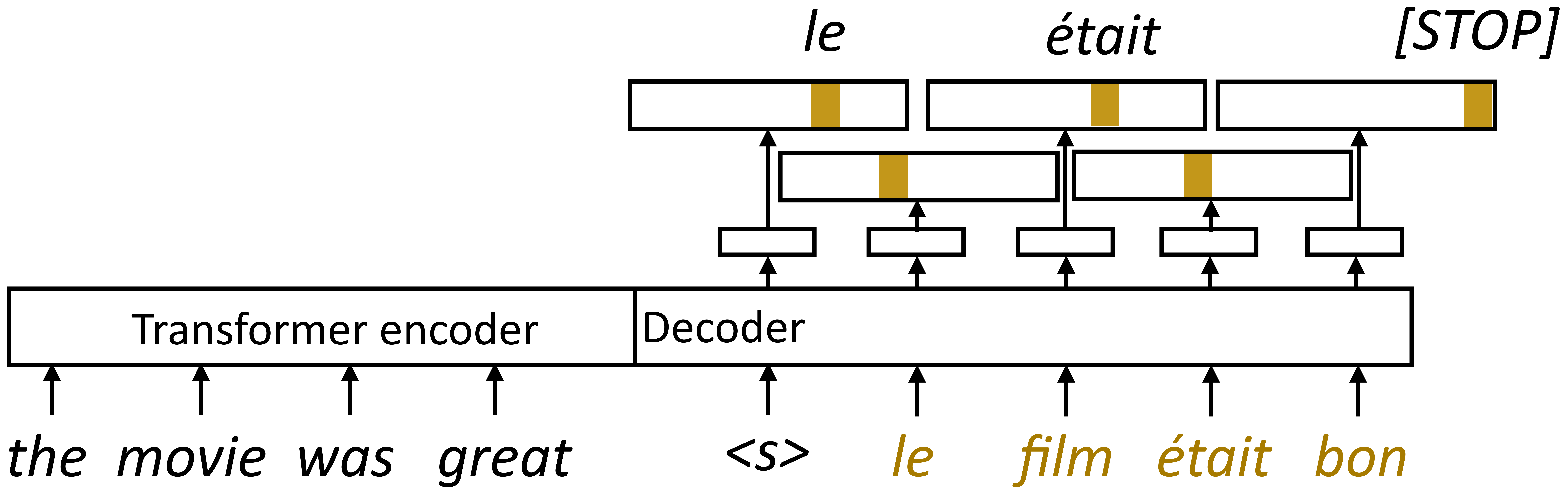
Decoder learns to both condition on **x** and generate plausible sequences **y** (if trained well/on enough data)

Inference (“Decoding”)



- During inference: need to compute the argmax over the word predictions and then run the next step of the decoder (which looks back at all previous encoder + decoder steps)
- Need to actually evaluate computation graph up to this point
- Decoder is advanced one state at a time until [STOP] is reached

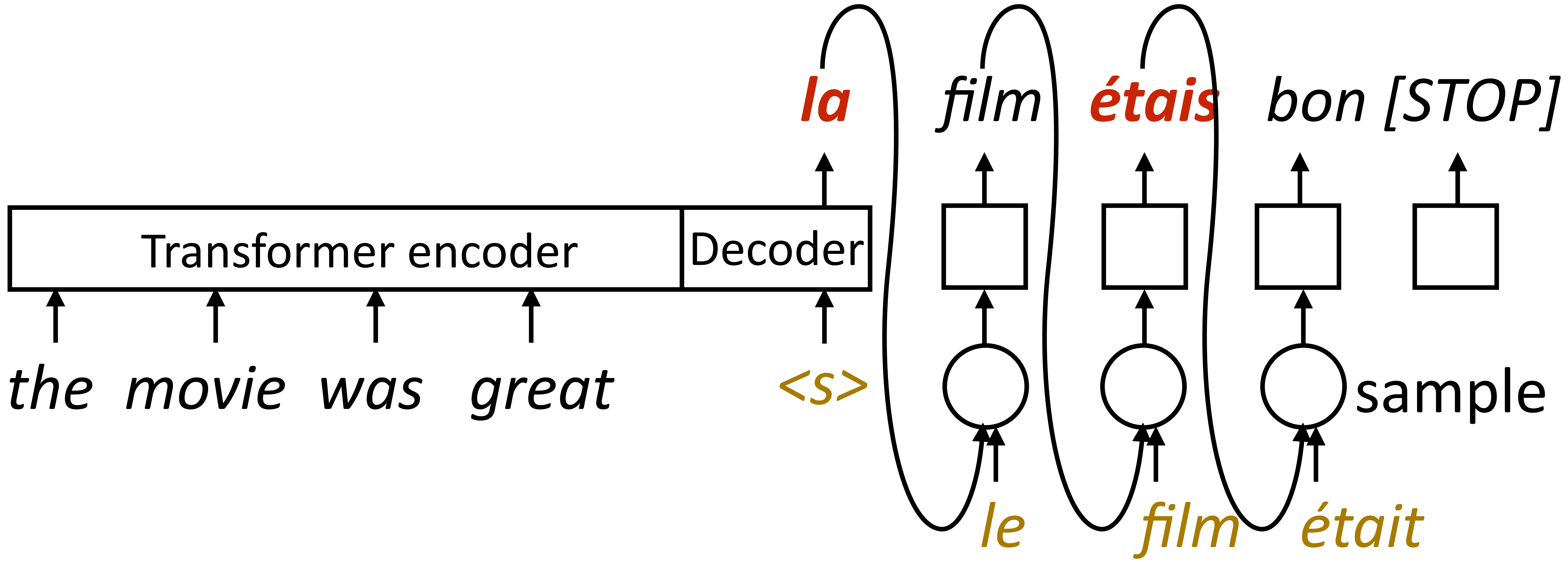
Training



- Objective: maximize $\sum_{(\mathbf{x},\mathbf{y})} \sum_{i=1}^{n} \log P(y_i^* | \mathbf{x}, y_1^*, \ldots, y_{i-1}^*)$
- One loss term for each target-sentence word, feed the correct word regardless of model's prediction (called "teacher forcing"). **Can train in "one go" like the language model, no need to run each step sequentially.**

Training: Scheduled Sampling

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



- Scheduled sampling: with probability p, take the gold as input, else take the model's prediction
- Starting with p = 1 (teacher forcing) and decaying it works best
- Not really used these days

Bengio et al. (2015)

Decoding Methods

Decoding Strategies

- LMs place a distribution $P(y_i \mid y_1, \ldots, y_{i-1})$
- seq2seq models place a distribution $P(y_i \mid \mathbf{x}, y_1, \ldots, y_{i-1})$
- Generation from both models looks similar; how do we do it?
  - Option 1: $\max_{y_i} P(y_i \mid y_1, \ldots, y_{i-1})$ — take greedily best option
  - Option 2: use beam search to find the sequence with the highest prob.
  - Option 3: sample from the model; draw y_i from that distribution
- Machine translation: use beam search. The top-scoring hypothesis is usually a great translation

Holtzman et al. (2019)

Decoding Strategies

- Story generation (this is with GPT-2):

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

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

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

- Beam search degenerates and starts repeating. If you see a fragment repeated 2-3x, it has very high probability to keep repeating

- Sampling is too noisy — introduces many grammatical errors

Holtzman et al. (2019)

Degeneration

- Beam search fails because the model is *locally normalized*
- Let's look at all the individual decisions that get made here

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

P(Nacional | … Universidad) is high

P(Autónoma | … Universidad Nacional) is high

P(de | … Universidad Nacional Autónoma) is high

P(México | Universidad Nacional Autónoma de) is high

P(/ | … México) and P(Universidad | … México /) — these probabilities may be low. But those are just 2/6 words of the repeating fragment

- **Each word is likely given the previous words but the sequence is bad**

Holtzman et al. (2019)

Drawbacks of Sampling

- Sampling is "too random"

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by $\boxed{\text{town}}$ and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV

P(*y* | … they live in a remote desert uninterrupted by)

0.01 roads

0.01 towns

0.01 people

0.005 civilization

Good options, maybe accounting for 90% of the total probability mass. So a 90% chance of getting something good

…

0.0005 town

Long tail with 10% of the mass

Holtzman et al. (2019)

Nucleus Sampling

P(y | … they live in a remote desert uninterrupted by)

0.01 roads

0.01 towns ⟶ renormalize and sample

0.01 people

0.005 civilization

cut off after p% of mass

- Define a threshold p. Keep the most probable options account for p% of the probability mass (the *nucleus*), then sample among these.
- To implement: sort options by probability, truncate the list once the total exceeds p, then renormalize and sample from it

Holtzman et al. (2019)

Decoding Strategies

- LMs place a distribution $P(y_i \mid y_1, \ldots, y_{i-1})$
- seq2seq models place a distribution $P(y_i \mid \mathbf{x}, y_1, \ldots, y_{i-1})$
- How to generate sequences?
  - Option 1: $\max_{y_i} P(y_i \mid y_1, \ldots, y_{i-1})$ — take greedily best option
  - Option 2: use beam search to find the sequence with the highest prob.
  - ~~Option 3: sample from the model; draw y_i from that distribution~~
  - Option 4: nucleus sampling

Holtzman et al. (2019)

Subword Tokenization

Handling Rare Words

- Words are a difficult unit to work with: copying can be cumbersome, word vocabularies get very large
  - When you have 100,000+ words, the final matrix multiply and softmax start to dominate the computation
- Character-level models were explored extensively in 2016-2018 but simply don’t work well — becomes very expensive to represent sequences

Subword Tokenization

- Subword tokenization: wide range of schemes that use tokens that are **between characters and words** in terms of granularity

- These “word pieces” may be full words or parts of words

Input: _the **_eco tax** _port i co _in _Po nt - de - Bu is …

- _ indicates the word piece starting a word (can think of it as the space character).

Sennrich et al. (2016)

Subword Tokenization

- Subword tokenization: wide range of schemes that use tokens that are **between characters and words** in terms of granularity

- These "word pieces" may be full words or parts of words

Input: _the **_eco tax** _port i co _in _Po nt - de - Bu is …

Output: _le _port ique **_éco taxe** _de _Pont - de - Bui s

- Can achieve transliteration with this, subword structure makes some translations easier to achieve

Sennrich et al. (2016)

Byte Pair Encoding (BPE)

- Start with every individual byte (basically character) as its own symbol

```
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
```

- Count bigram character cooccurrences
- Merge the most frequent pair of adjacent characters

- Doing 8k merges => vocabulary of around 8000 word pieces. Includes many whole words

- Most SOTA NMT systems use this on both source + target

Sennrich et al. (2016)

Byte Pair Encoding (BPE)

(a)

Original:	furiously
BPE:	_fur | iously
Unigram LM:	_fur | ious | ly

(b)

Original:	tricycles
BPE:	_t | ric | y | cles
Unigram LM:	_tri | cycle | s

(c)

Original:	Completely preposterous suggestions
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 another technique based on a unigram language model: rather than greedily merge, find chunks which make the sequence look likely under a unigram LM

Bostrom and Durrett (2020)

Tokenization Today

- **All pre-trained** models use some kind of subword tokenization with a tuned vocabulary; usually between 50k and 250k pieces (larger number of pieces for multilingual models)

- As a result, classical word embeddings like GloVe **are not used**. All subword representations are randomly initialized and learned in the Transformer models

Neural MT

Results: WMT English-French

- 12M sentence pairs

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

PBMT + rerank w/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 BLEU scores aren't really comparable across languages
- French, Spanish = easiest
  German, Czech = harder
  Japanese, Russian = hard (grammatically different, lots of morphology…)

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 *Kerr* still love each other .
best	In an interview , however , Bloom said that he and *Kerr* 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

Luong et al. (2015)

MT Examples

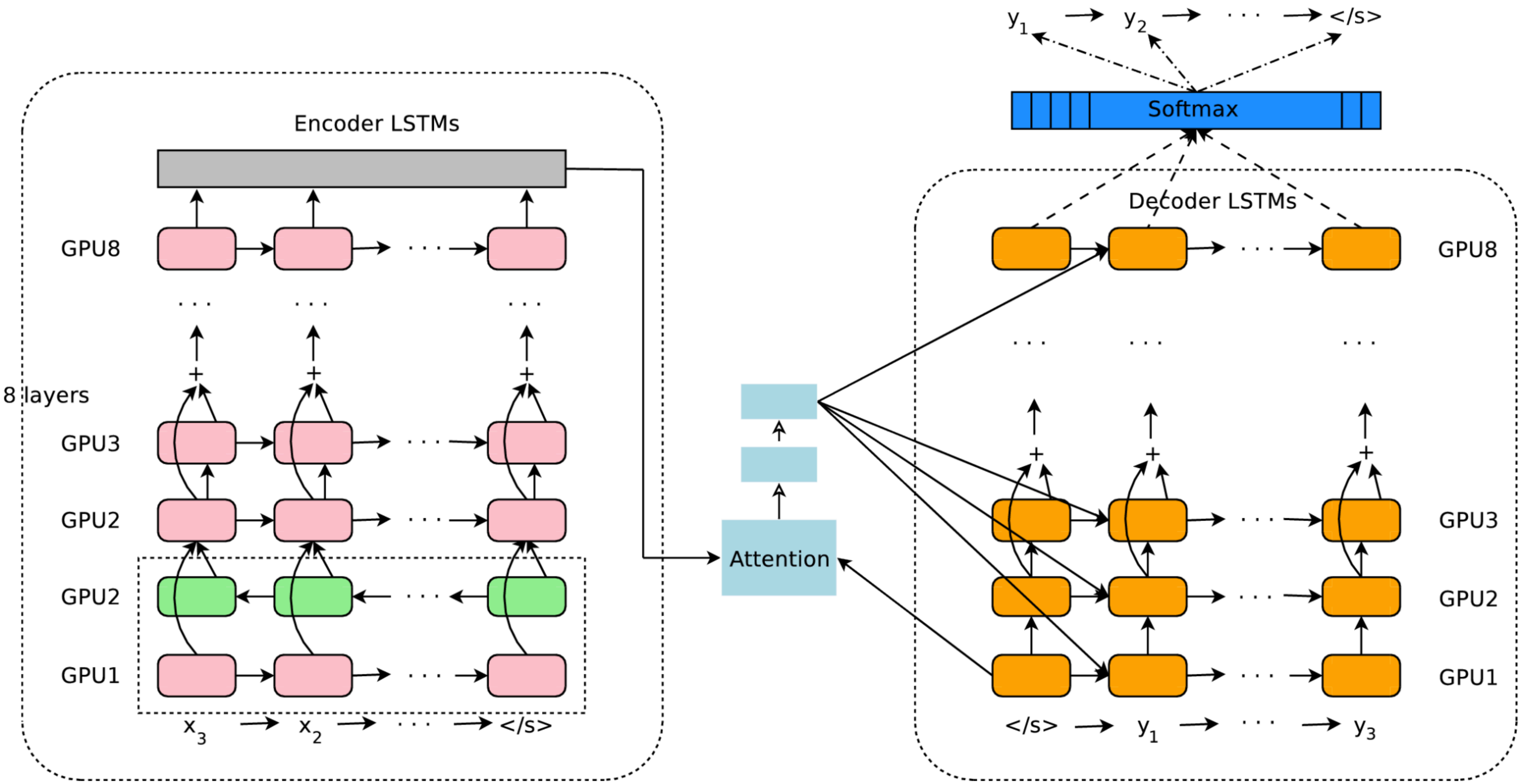
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The *austerity imposed by Berlin and the European Central Bank , 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 *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* 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

Luong et al. (2015)

Google NMT (2016)

Google's NMT System (2016)



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

Wu et al. (2016)

Google's NMT System (2016)

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

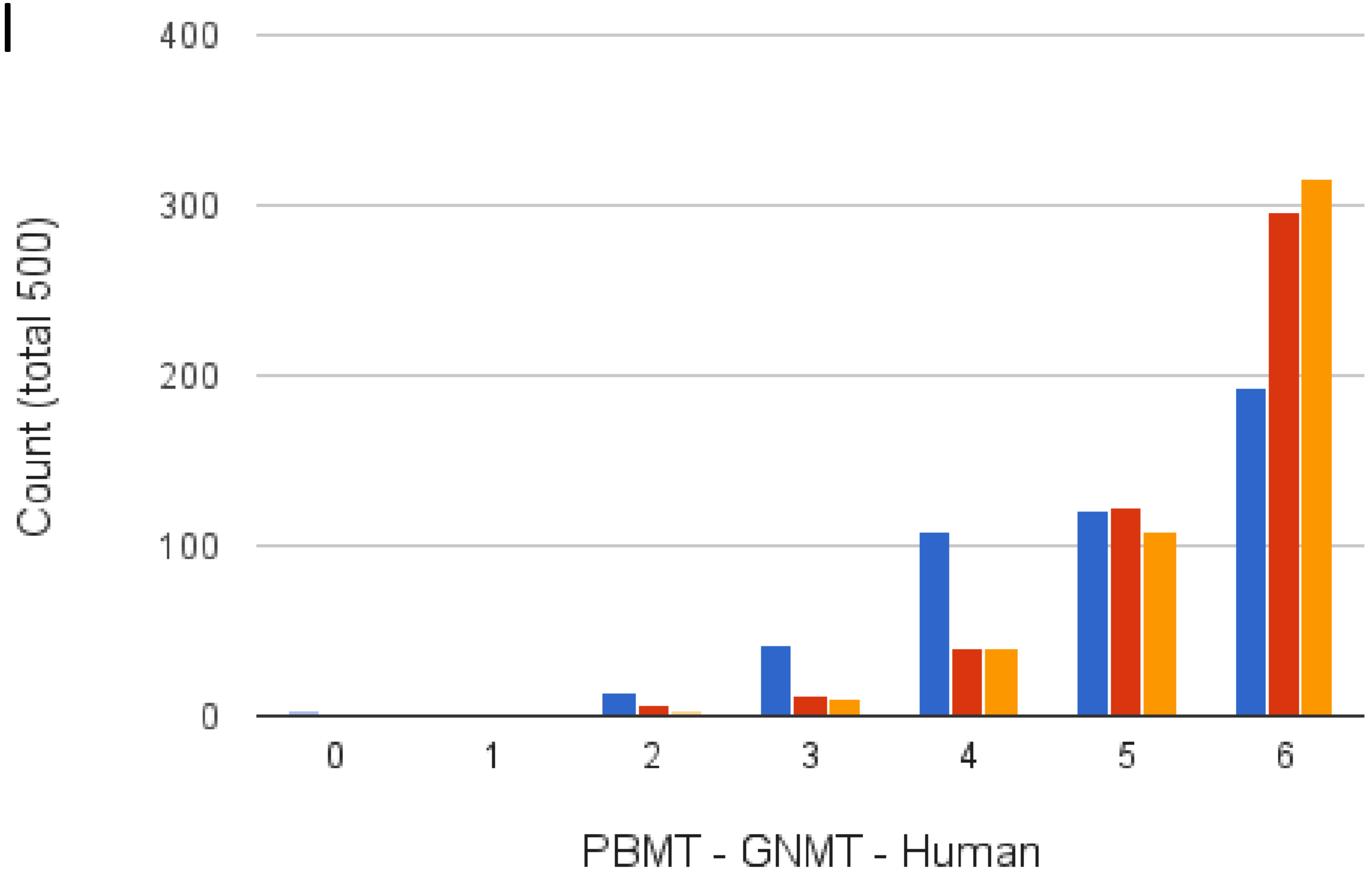
Luong+ (2015) seq2seq ensemble with rare word handling: 23.0 BLEU

Google's 32k word pieces: 24.2 BLEU

Wu et al. (2016)

Human Evaluation (En-Es)

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



Wu et al. (2016)

Transformer MT + Frontiers

Transformers

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**

- Big = 6 layers, 1000 dim for each token, 16 heads, base = 6 layers + other params halved

Vaswani et al. (2017)

Frontiers in MT: Small Data

ID	system	BLEU 100k	BLEU 3.2M
1	phrase-based SMT	15.87 ± 0.19	26.60 ± 0.00
2	NMT baseline	0.00 ± 0.00	25.70 ± 0.33
3	2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing)	7.20 ± 0.62	31.93 ± 0.05
4	3 + reduce BPE vocabulary (14k $\rightarrow$ 2k symbols)	12.10 ± 0.16	-
5	4 + reduce batch size (4k $\rightarrow$ 1k tokens)	12.40 ± 0.08	31.97 ± 0.26
6	5 + lexical model	13.03 ± 0.49	31.80 ± 0.22
7	5 + aggressive (word) dropout	15.87 ± 0.09	$\mathbf{33.60} \pm 0.14$
8	7 + other hyperparameter tuning (learning rate, model depth, label smoothing rate)	$\mathbf{16.57} \pm 0.26$	32.80 ± 0.08
9	8 + lexical model	16.10 ± 0.29	33.30 ± 0.08

- Synthetic small data setting: German -> English

Sennrich and Zhang (2019)

Frontiers in MT: Low-Resource

- Particular interest in deploying MT systems for languages with little or no parallel data

Burmese, Indonesian, Turkish

- BPE allows us to transfer models even without training on a specific language

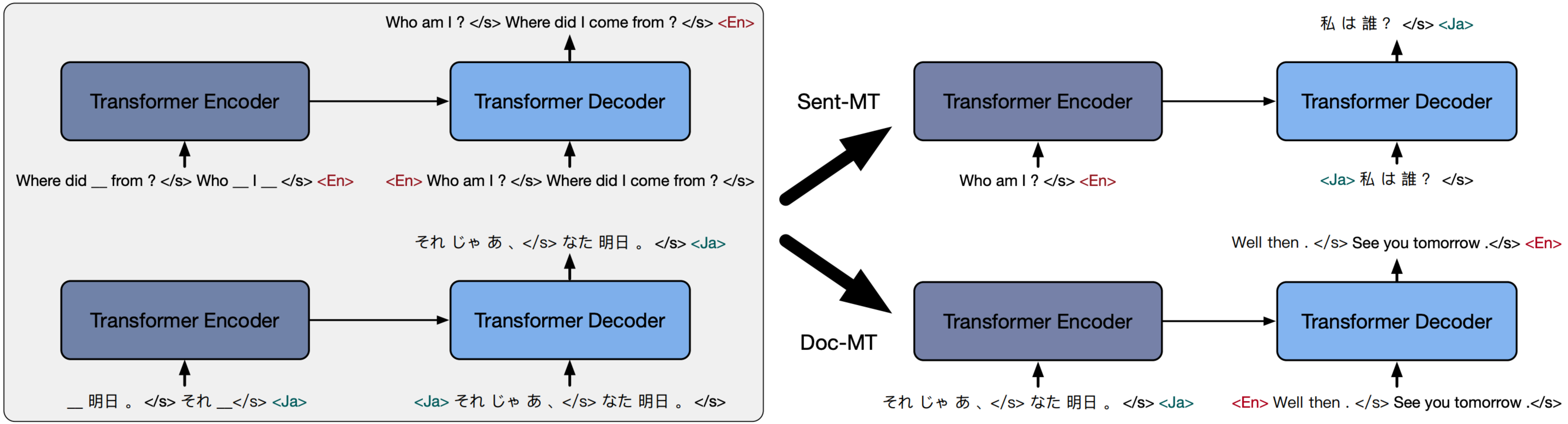
Transfer	BLEU My→En	Id→En	Tr→En
baseline (no transfer)	4.0	20.6	19.0
transfer, train	17.8	27.4	20.3
transfer, train, reset emb, train	13.3	25.0	20.0
transfer, train, reset inner, train	3.6	18.0	19.1

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

- Pre-trained models can help further

Aji et al. (2020)

Frontiers in MT: Multilingual Models



Multilingual Denoising Pre-Training (mBART)

Fine-tuning on Machine Translation

Yinhan Liu et al. (2020)

Frontiers in MT: Multilingual Models

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Languages	En-Nl		En-Ar		En-It		En-My		En-Ne		En-Ro	
Data Source	IWSLT17		IWSLT17		IWSLT17		WAT19		FLoRes		WMT16	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**

- Random = random initialization

Yinhan Liu et al. (2020)

Frontiers in MT: Multilingual Models

SOURCE Zh	针对政府的沉默态度,初级医生委员会执行委员会已于今日正式要求英国医学协会理事会召开特别会议批准旨在从九月初开始升级劳工行动的一项长期计划。
TARGET En	In response to the government's silence, JDC exec has today made a formal request for a special meeting of BMA Council to authorise a rolling programme of escalated industrial action beginning in early September.
mBART25 Ja-En	In response to the government's silence, the Council of Chief Medical Officers has formally requested today the Royal College of Physicians to hold a special meeting to approve a long-term workforce action that starts in September.
mBART25 Ko-En	In response to the government's silence, the Chief Medical Officers' Council is calling today for a special session at the Council of the British Medical Association, which is a long-term initiative to upgrade labor from September.
mBART25 Zh-En	In response to the government's silence, the Board of Primary Doctors has today formally asked the British Medical Association to hold a special meeting to approve a long-term plan that starts in the beginning of September.

Yinhan Liu et al. (2020)

Takeaways

- Transformers are state-of-the-art for machine translation
- They work really well on languages where we have a ton of data. When they don't: pre-training can help
- Next up: exploring pre-training in more detail (ELMo, BERT, GPT, etc.)