Pretraining

CSE 5525: Foundations of Speech and Natural Language Processing

https://shocheen.github.io/courses/cse-5525-spring-2026

THE OHIO STATE UNIVERSITY

Logistics

- Final Project Proposal: Due this Friday.
  - Will do another office hours tomorrow: 1-2.30pm in DL 581. Will announce on teams/canvas.

- Homework 3 will be released on Friday also.
  - Topic: finetuning / prompting.

Last Class Recap: Masked LMs

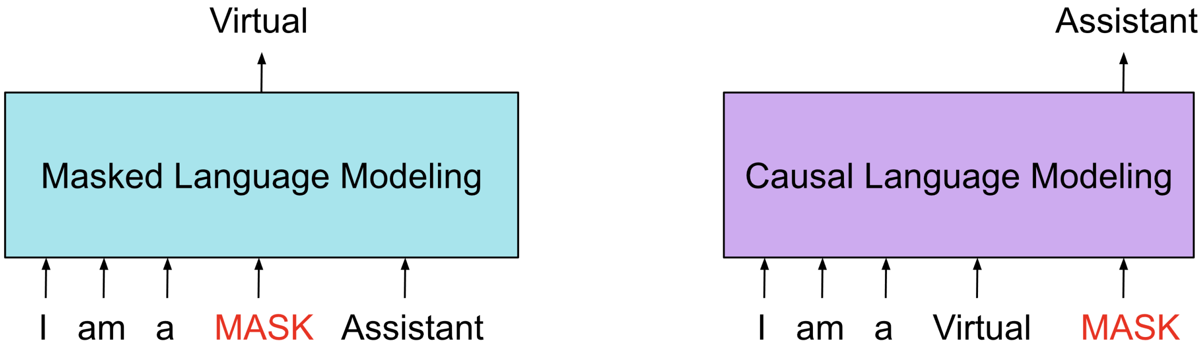


Image from https://www.holisticai.com/blog/from-transformer-architecture-to-prompt-engineering

Masked Language Modeling (MLM)

(text) Sylvester Stallone has made some **terrible** films in his lifetime, but this has got to be one of the **worst**. A totally **dull** story...

(masked text) Sylvester Stallone has made some **<mask>** films in his lifetime, but this has got to be one of the **<mask>**. A totally **<mask>** story...

ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER

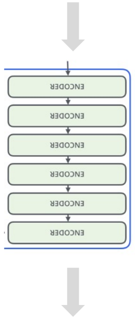
(output final layer) vector(Sylvester) ... vector(**<mask>**) vector(films) vector(in) vector(his) ... vector(the) vector(**<mask>**) ... vector(**<mask>**) vector(story)...

loss(predicted_word,“terrible”)

Masked Language Modeling (MLM)

(text) Sylvester Stallone has made some **terrible** films in his lifetime, but this has got to be one of the **worst**. A totally **dull** story...

(masked text) Sylvester Stallone has made some **<mask>** films in his lifetime, but this has got to be one of the **<mask>**. A totally **<mask>** story...



(output final layer) vector(Sylvester) ... vector(**<mask>**) vector(films) vector(in) vector(his) ... vector(the) vector(**<mask>**) ... vector(**<mask>**) vector(story)...

loss(predicted_word,“worst”)

loss(predicted_word,“dull”)

loss(predicted_word,“terrible”)

average these losses and do a gradient descent step

Masked LM: Only using Encoder transformer

- Encoders take a complete sequence as input (not just the prefix).

- Self-attention computes weighted sum over entire context (i.e., entire sequence)

- There is no generation problem, we just want representations
    - We will learn how to use them later on

BERT

What Do We Get?

- We can feed complete sentences to BERT
- For each token, we get a contextualized representation
  - Meaning: computed taking the other tokens in the sentence into acocunt
- While word2vec/glove vectors are forced to mix multiple senses, BERT can provide more instance-specific vectors

BERT

How Do We Use It?

- Widely supported by existing frameworks
  - E.g., Transformers library by Hugging Face
- We will soon see how to use it when working with annotated data
- Large BERT models quickly outperformed human performance on several NLP tasks
  - But what it meant beyond benchmarking was less clear
- Started an arms race towards bigger and bigger models, which quickly led to the LLMs of today

BERT

What It Is Not Great For?

- primarily used for discriminative tasks, Cannot generate text
    - Can manipulate to generate text but not the original purpose of this model.
    - Can train to generate text (more recently) – check out masked diffusion language models (MDLMs).

Finetuning a MLM-pretrained model

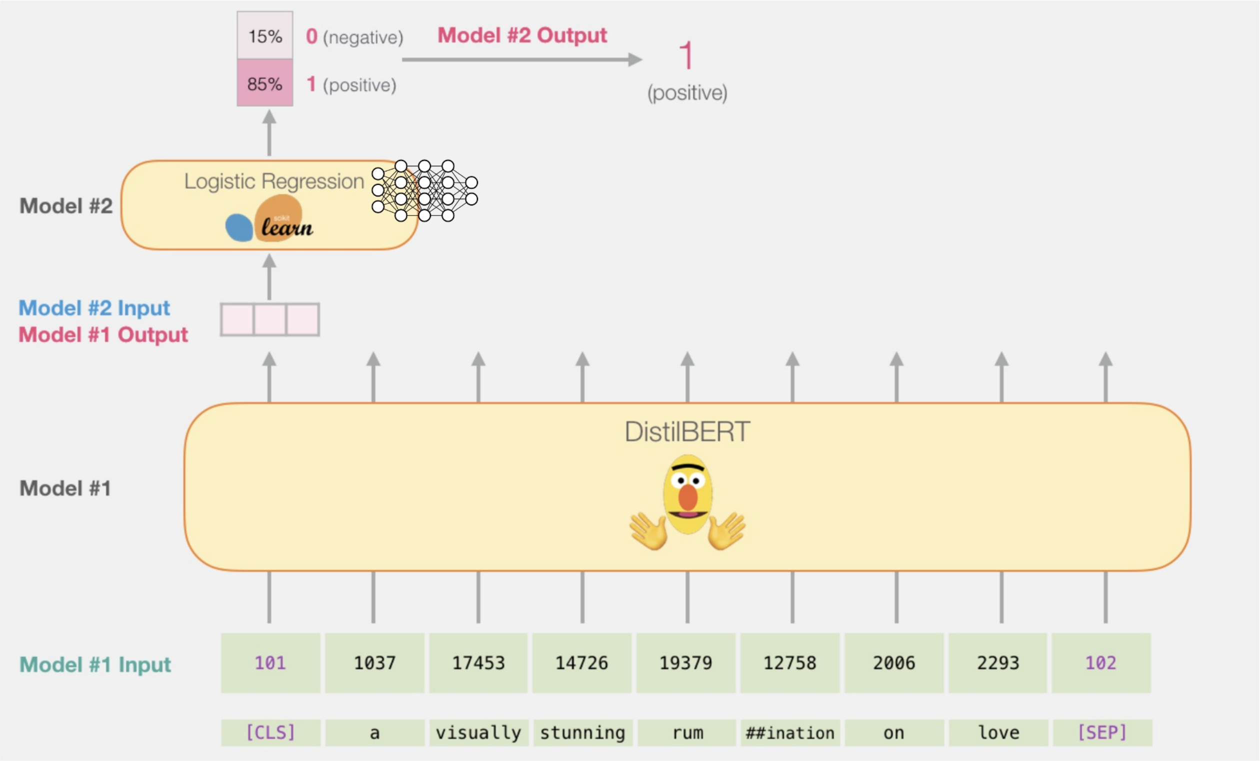
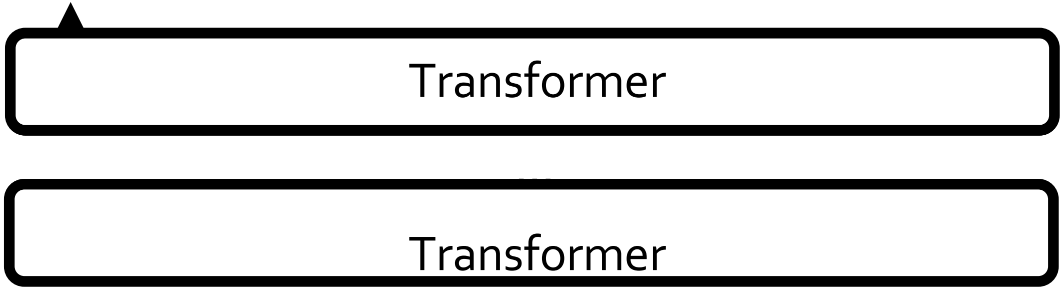


Figure: Jay Alammar

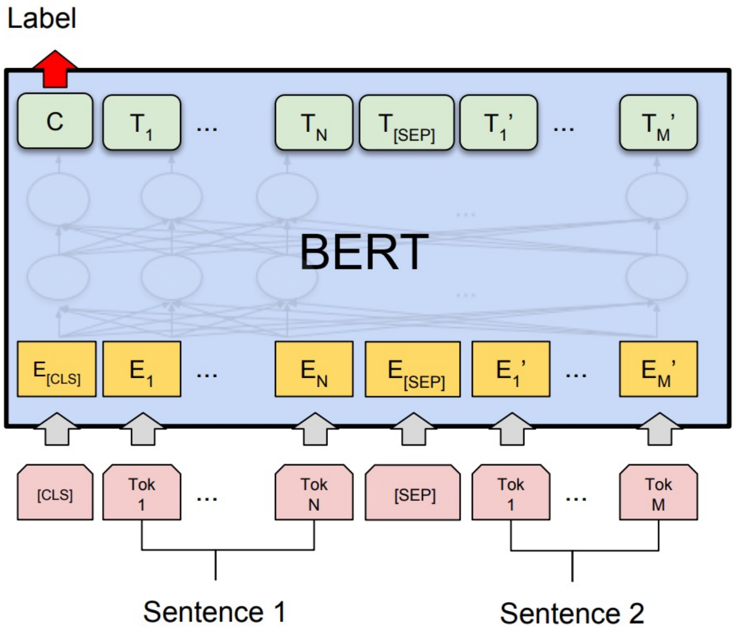
What can BERT do?

Text (Pair) Classification

Entails (first sentence implies second is true)



[CLS] A boy plays in the snow [SEP] A boy is outside



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

What can BERT do?

Question Answering (Span-based)

Q: What was Marie Curie the first female recipient of?

Passage: One of the most famous people born in Warsaw was Marie Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

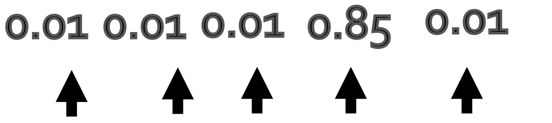
Answer = Nobel Prize

Question Answering with BERT

Q: What was Marie Curie the first female recipient of?

Passage: One of the most famous people born in Warsaw was Marie Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

- Predict answer as a pair of (start, end) indices given question q and passage p; compute a score for each word and softmax those

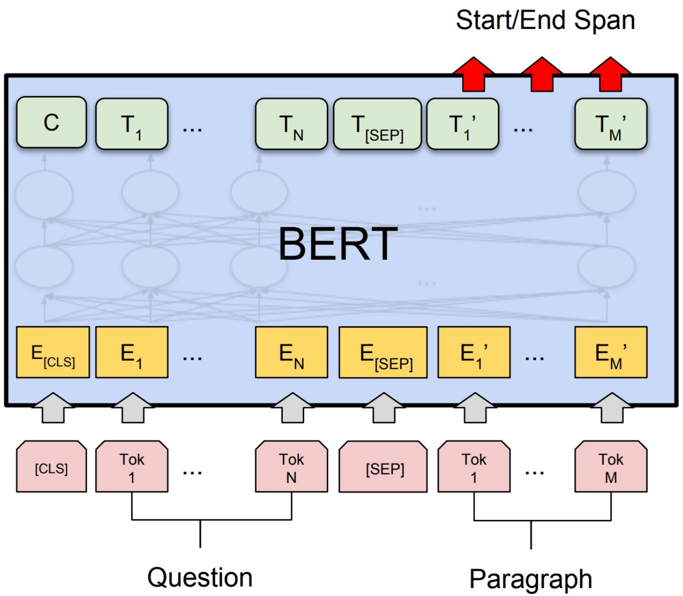


0.01 0.01 0.01 0.85 0.01

P(start | q, p) = recipient of the **Nobel Prize** .

P(end | q, p) = same computation but different params

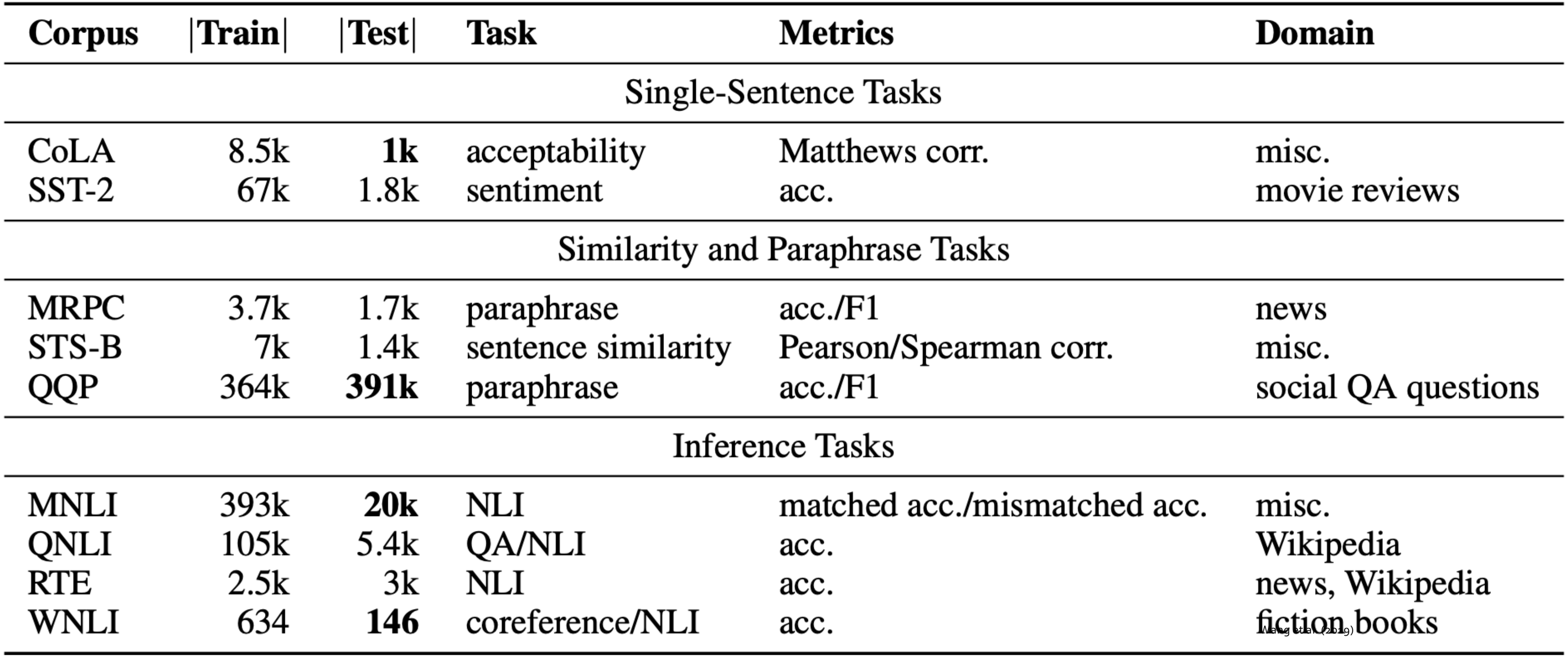
QA with BERT



What was Marie Curie the first female recipient of ? [SEP] One of the most famous people born in Warsaw was Marie ...

Devlin et al. (2019)

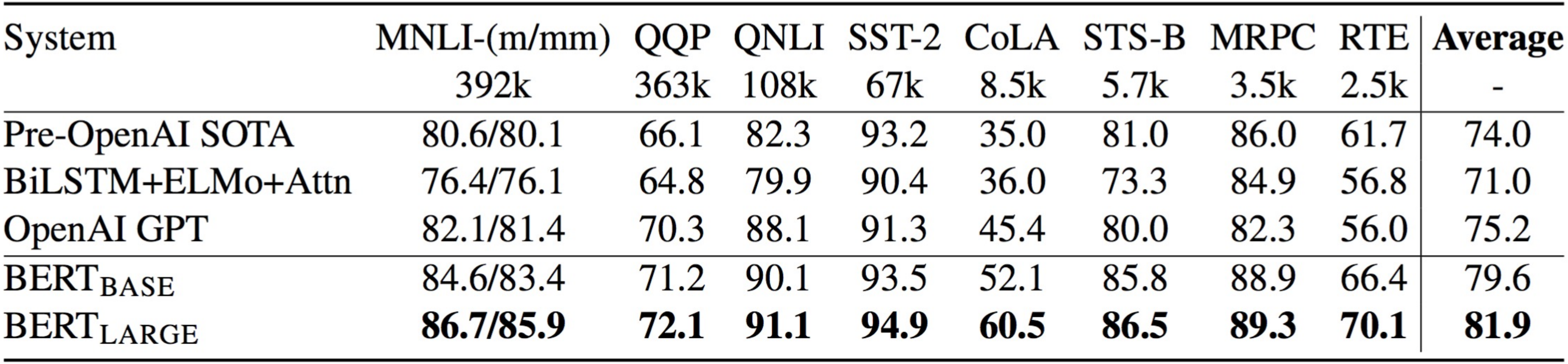
Evaluation: GLUE



Corpus	|Train|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Wang et al. (2019)

Results



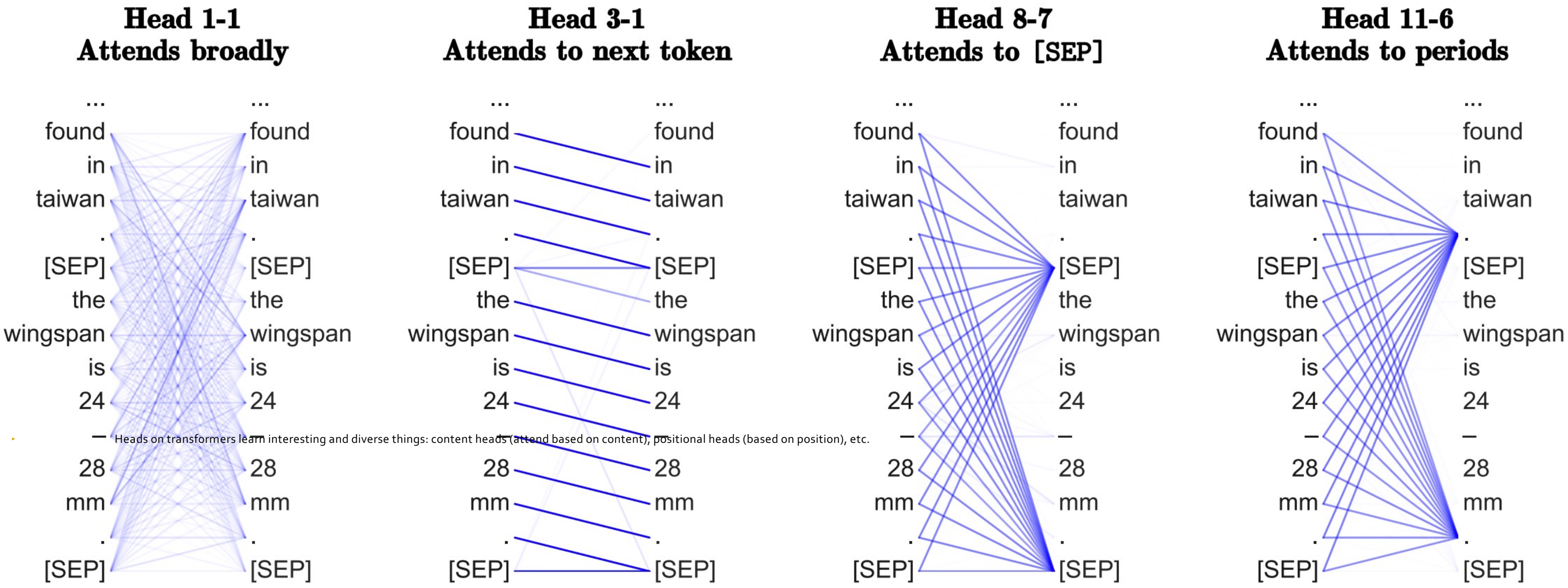
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT$_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT$_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

- Huge improvements over prior work

- Effective at “sentence pair” tasks: textual entailment (does sentence A imply sentence B), paraphrase detection

Devlin et al. (2018)

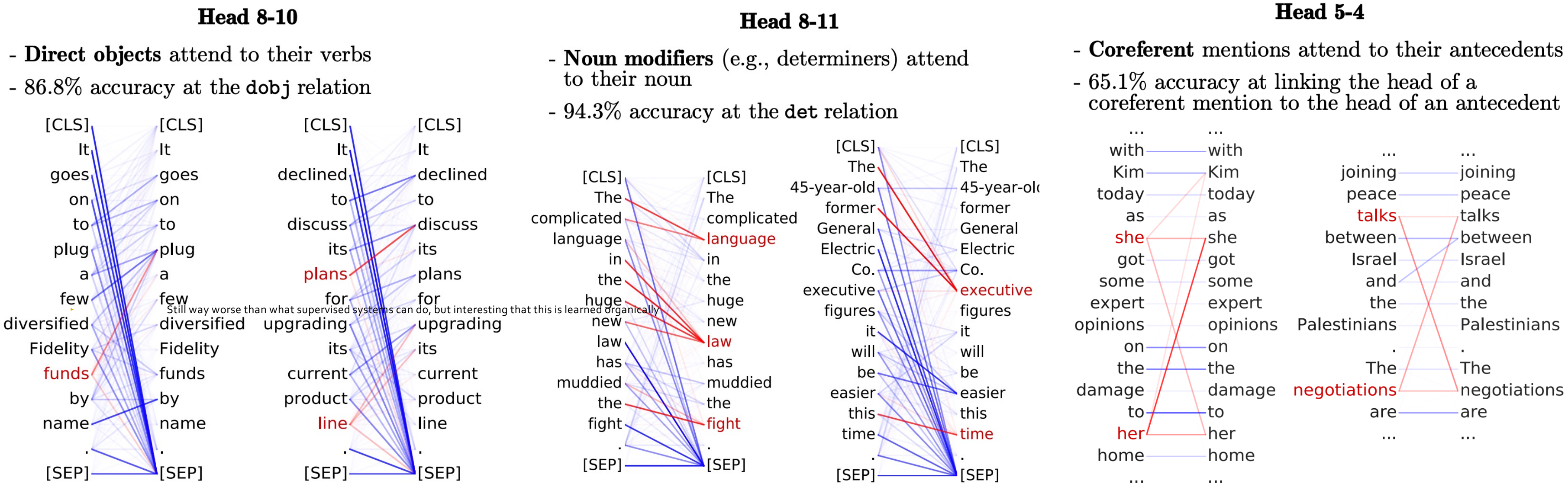
What does BERT learn?



- Heads on transformers learn interesting and diverse things: content heads (attend based on content), positional heads (based on position), etc.

Clark et al. (2019)

What does BERT learn?



Head 8-10

- **Direct objects** attend to their verbs
- 86.8% accuracy at the `dobj` relation

Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the `det` relation

Head 5-4

- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

[CLS] It goes on to plug a few diversified Fidelity funds by name . [SEP]
[CLS] It goes on to plug a few diversified Fidelity funds by name . [SEP]
[CLS] It declined to discuss its plans for upgrading its current product line . [SEP]
[CLS] It declined to discuss its plans for upgrading its current product line . [SEP]
[CLS] The complicated language in the huge new law has muddied the fight . [SEP]
[CLS] The complicated language in the huge new law has muddied the fight . [SEP]
[CLS] The 45-year-old former General Electric Co. executive figures it will be easier this time . [SEP]
[CLS] The 45-year-old former General Electric Co. executive figures it will be easier this time . [SEP]
... with Kim today as she got some expert opinions on the damage to her home ...
... with Kim today as she got some expert opinions on the damage to her home ...
... joining peace talks between Israel and the Palestinians . The negotiations are ...
... joining peace talks between Israel and the Palestinians . The negotiations are ...

Still way worse than what supervised systems can do, but interesting that this is learned organically

Clark et al. (2019)

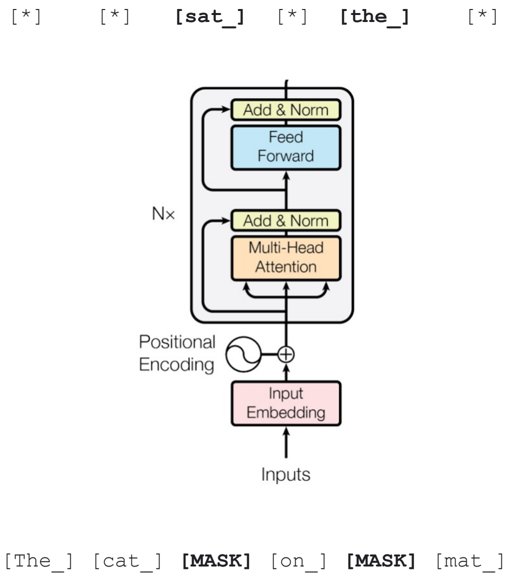
Masked LMs: Summary

Encoder-only transformer

- Masked language modeling (MLM), ~~next sentence prediction~~

⚓ These models are a good option if you want to solve a text classification problem for which you have thousands of labeled datapoints & you know how to train a model (which you all will know after this course)

⚓ Are not built for generation.



[*] [*] **[sat_]** [*] **[the_]** [*]

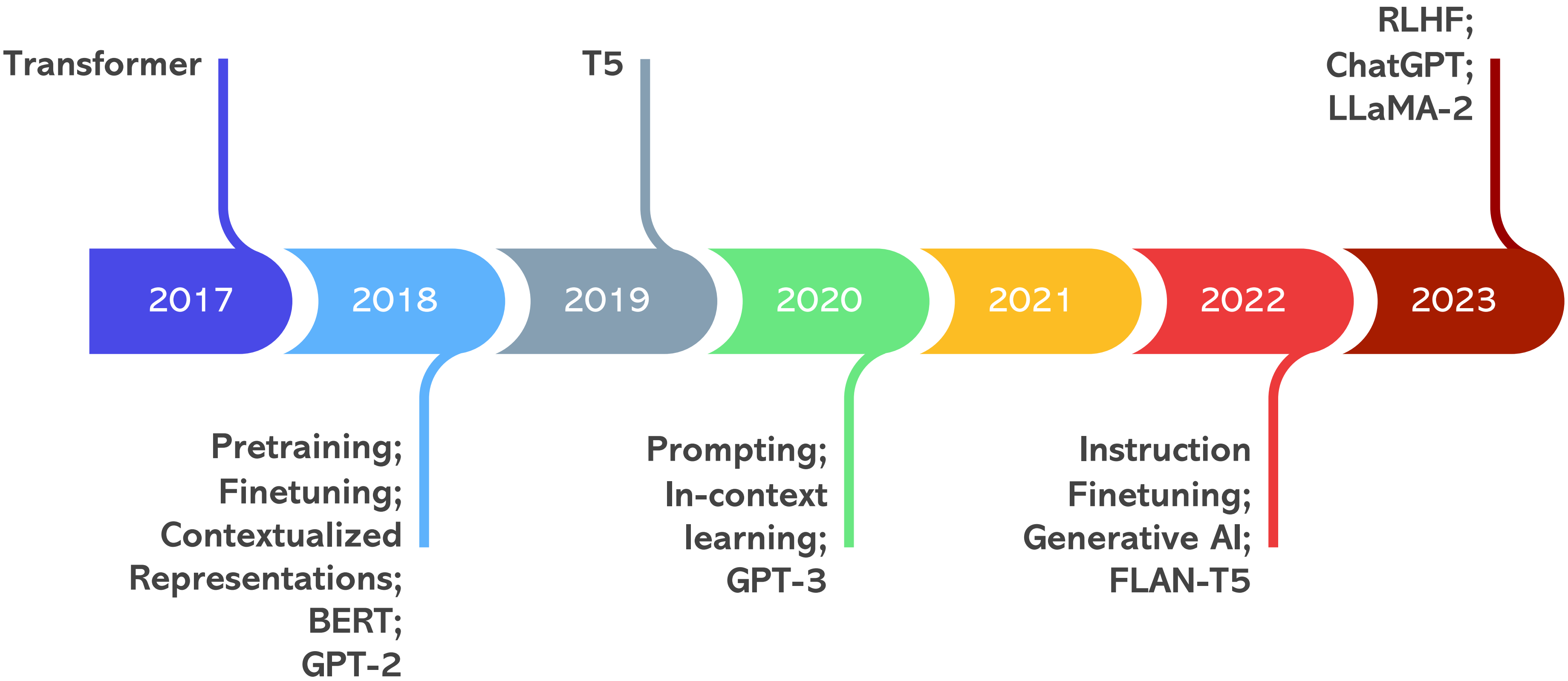
[The_] [cat_] **[MASK]** [on_] **[MASK]** [mat_]

Figure by: Lucas Beyer

Today

- How can we make BERT even better?

- We want pretraining benefits but also **generative capabilities.**



2017

Transformer

2018

Pretraining;
Finetuning;
Contextualized
Representations;
BERT;
GPT-2

2019

T5

2020

Prompting;
In-context
learning;
GPT-3

2021

2022

Instruction
Finetuning;
Generative AI;
FLAN-T5

2023

RLHF;
ChatGPT;
LLaMA-2

Why not just causal language modeling?

- GPT-1 (Generative pretrained transformers) came before BERT and BERT argued that you need bidirectional context to learn good representations.
    - This belief was maintained even with GPT2 but BERT was generally better than GPT1 and 2 on many tasks.
    - Spoiler alert: we do use primarily causal LMs for pretraining now.
- BERT achieved bidirectional context by learning using a “denoising objective”

Pretraining via denoising objectives

- What is denoising?
  - Add noise to your input, train a model to recover the original input from the noisy input
  - Goal: by learning to denoise, the model learns crucial details about the input.

- BERT uses masking a way to introduce noise.
  - Masked input a noisier version of original input.

- Lots of follow up works:
  - Can we built a generative model based on a denoising objective? → T5
  - Can we use other denoising objectives? → BART
  - What if used both causal and denoising objectives together → UL2

Today’s plan

- T5 (masked LM, encoder-decoder)

- BART (denoising LM, encoder-decoder)

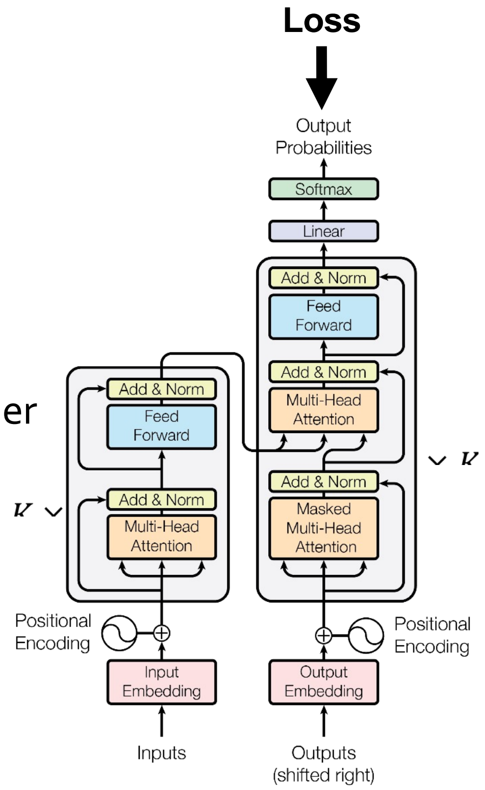
- UL2 (decoder only – mix of denoising + causal LM objectives)

- If time: How to decode from decoders (sampling algorithms).

Encoder-decoder

With Transformers

- The model is composed of two components
- Bidirectional **encoder** to process the input
- Autoregressive **decoder** to generate output
- Training is usually done with loss on the output
    - Propagates into the decoder and through it to the encoder

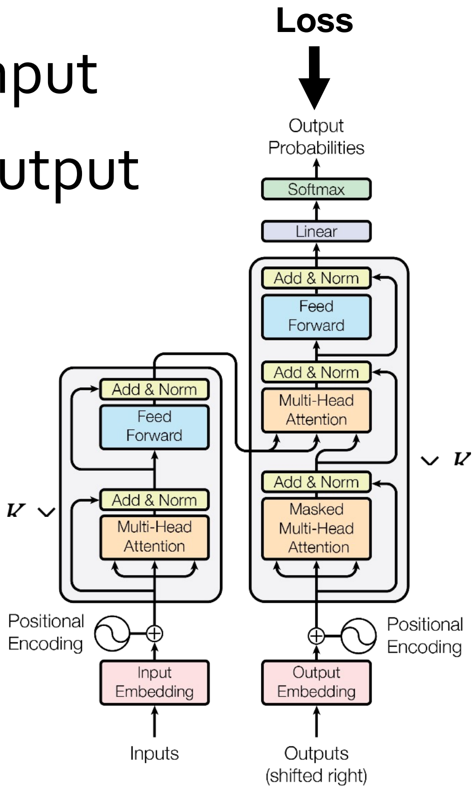


[Vaswani et al. 2017]

Encoder-decoder

With Transformers

- Bidirectional **encoder** to process the input
- Autoregressive **decoder** to generate output
- Why does this structure make sense?



[Vaswani et al. 2017]

How to adapt Masked LM to an encoder-decoder setting

- Output is generated by decoder, and the loss is on the output

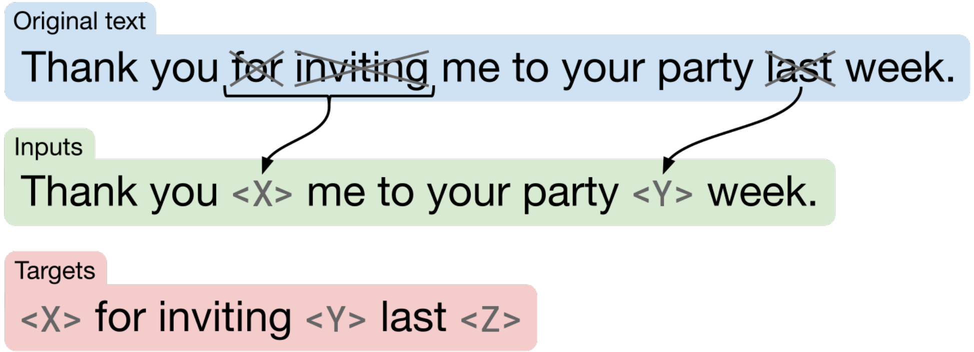
- Input is a sequence of tokens

[Lewis et al. 2019]

T5 (Text-to-Text Transfer Transformer)

Pretraining

- Pretraining is similar to the denoising objective of BERT:
  - Input: text with “masks” – but now spans removed (instead of just tokens)
  - Output: sequence of phrases to fill the gaps



Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

- Trained on the next token objective (only on the decoder; similar to a conditional LM)

[Raffel et al. 2019]

T5

What Do We Get?

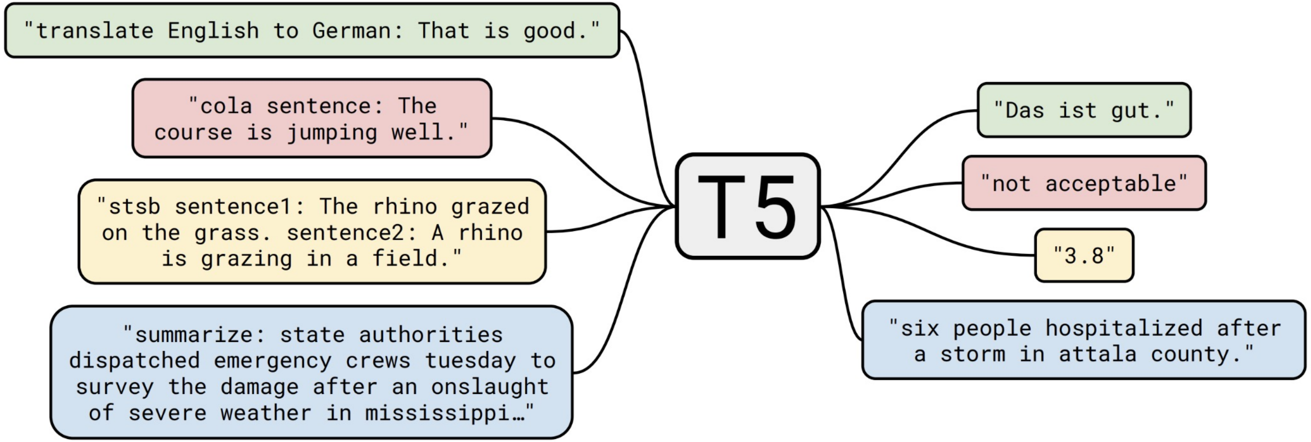
- BERT: a pre-trained encoder
- T5: pre-trained decoder and encoder

- [Lewis et al. 2019]

T5 (Text-to-Text Transfer Transformer)

Finetuning

- Frame any problem as a text-to-text problem.
- Initialize with pretrained T5 and finetune every task as a text to text generation task (no new parameters like classification heads required)

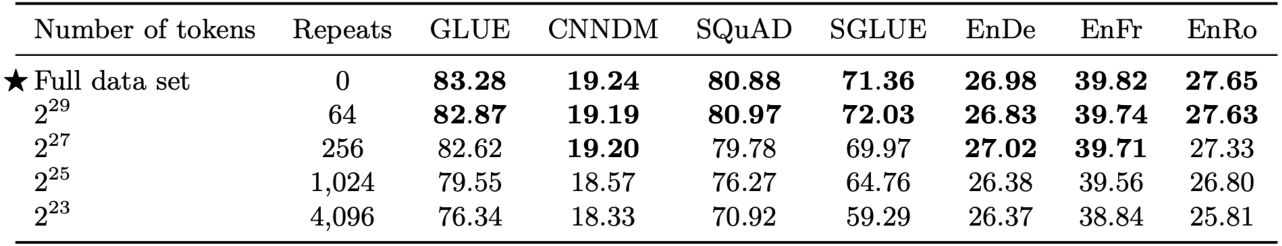

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

[Raffel et al. 2019]

T5 (Text-to-Text Transfer Transformer)

Results

- T5 was trained on one of the first very large corpora: 750GB of text, with pre-training using 2^{35} tokens
- First to show the impact of data scale
- It can solve both text classification and generation tasks.



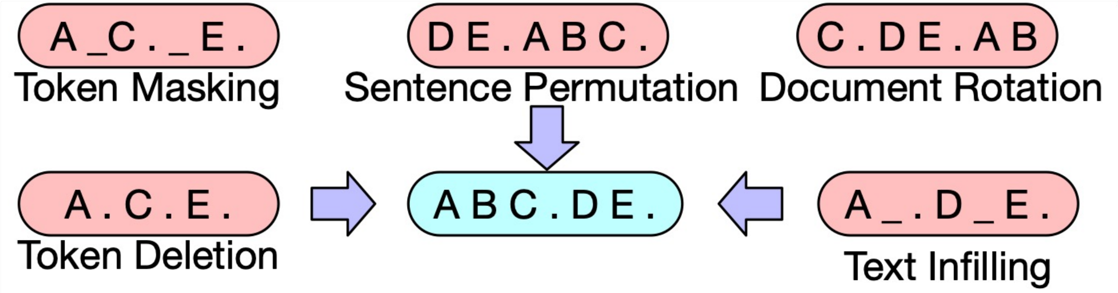
Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full data set	0	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**80.97**	**72.03**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	79.78	69.97	**27.02**	**39.71**	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

[Raffel et al. 2019]

A concurrent approach: BART

Bidirectional and Autoregressive Transformer

- Corrupt the input following five different recipes



- Try to recover the pre-corrupted input by generating it using the decoder
- Train on a lot of raw text data, just like with BERT and T5

[Lewis et al. 2019]

BART

How to Use?

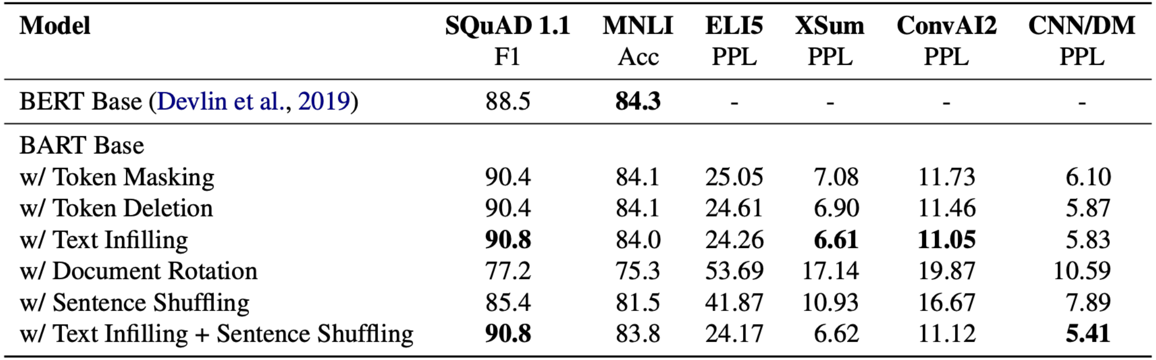
- Similar to BERT: fine-tune for the end task
- Add a classification head on just the encoder
- Or finetune like T5 on text-to-text tasks (no new parameters)
  - Some other heuristics are applied to make this feasible – check paper for details – not important in the context of current language models.

[Lewis et al. 2019]

BART

Performance

- Can do anything that BERT does
- But can also do generation tasks (e.g., summarization)



Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

- [Lewis et al. 2019]

BART and T5

Takeaways

- BART and T5 are very useful for all sorts of sequence-to-sequence tasks with language
  - T5 comes in different sizes
  - There are various customization (e.g., CodeT5)
- Extended the generalizations conclusions from BERT, and demonstrated the impact of data scale

Unified Language Modeling Paradigms

UL2

- So far, we have looked at multiple training objectives
  - Denoising
    - Masked token prediction –BERT
    - Masked span prediction – T5
    - Other denoising (shuffling, deleting, etc) – BART
  - Next token prediction (GPT)

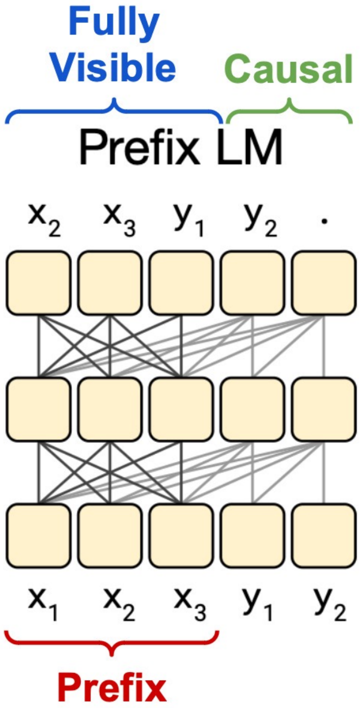
- Can we train a model that uses all of them?
  - Can we use encoder-only architecture?
  - Can we use encoder-decoder architecture?
  - Can we use decoder only architecture?

Do we really need a separate encoder?

Prefix LM

- Share the parameters of the encoder and the decoder

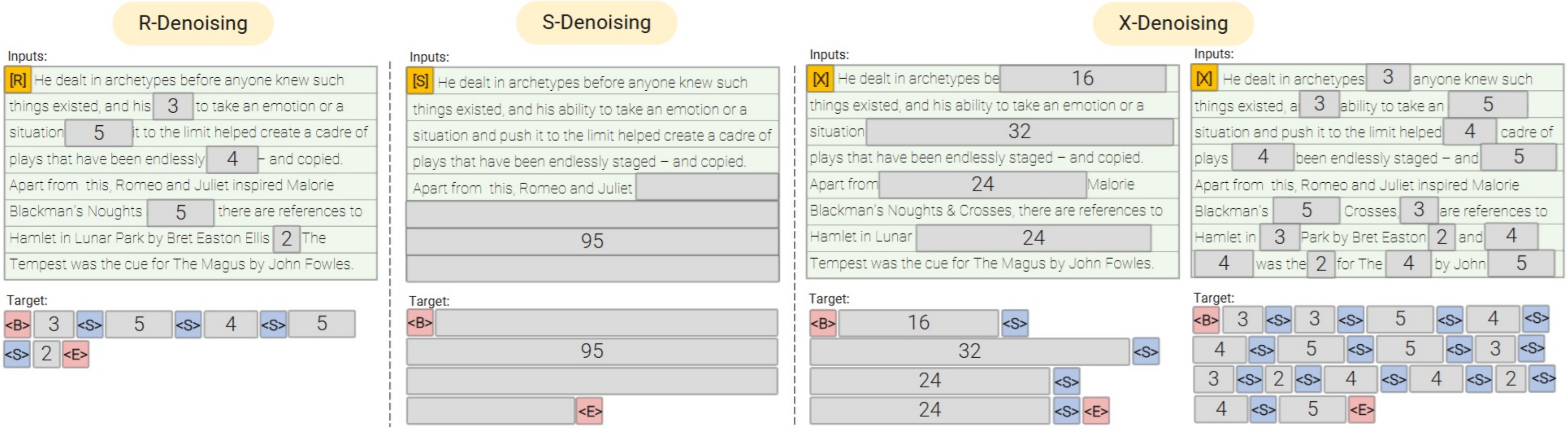
- Apply a bidirectional attention on the input and causal (masked) attention on the output.



Unified Language Modeling Paradigms

UL2

- Can use encoder-decoder or decoder-only prefix-LMs



R-Denoising

Inputs:

[R] He dealt in archetypes before anyone knew such things existed, and his 3 to take an emotion or a situation 5 it to the limit helped create a cadre of plays that have been endlessly 4 – and copied. Apart from this, Romeo and Juliet inspired Malorie Blackman's Noughts 5 there are references to Hamlet in Lunar Park by Bret Easton Ellis 2 The Tempest was the cue for The Magus by John Fowles.

Target:

<B> 3 <S> 5 <S> 4 <S> 5 <S> 2 <E>

S-Denoising

Inputs:

[S] He dealt in archetypes before anyone knew such things existed, and his ability to take an emotion or a situation and push it to the limit helped create a cadre of plays that have been endlessly staged – and copied. Apart from this, Romeo and Juliet 95

Target:

<B> 95 <E>

X-Denoising

Inputs:

[X] He dealt in archetypes be 16 things existed, and his ability to take an emotion or a situation 32 plays that have been endlessly staged – and copied. Apart from 24 Malorie Blackman's Noughts & Crosses, there are references to Hamlet in Lunar 24 Tempest was the cue for The Magus by John Fowles.

Target:

<B> 16 <S> 32 <S> 24 <S> 24 <S> <E>

Inputs:

[X] He dealt in archetypes 3 anyone knew such things existed, a 3 ability to take an 5 situation and push it to the limit helped 4 cadre of plays 4 been endlessly staged – and 5 Apart from this, Romeo and Juliet inspired Malorie Blackman's 5 Crosses, 3 are references to Hamlet in 3 Park by Bret Easton 2 and 4 4 was the 2 for The 4 by John 5

Target:

<B> 3 <S> 3 <S> 5 <S> 4 <S> 4 <S> 5 <S> 5 <S> 3 <S> 3 <S> 2 <S> 4 <S> 4 <S> 2 <S> 4 <S> 5 <E>

Figure 3: Mixture of denoisers for training UL2. Greyed out rectangles are masked tokens that are shifted to '*targets*' for prediction.

Summary of different pretraining objectives

- Denoising Objectives
- Next token prediction
- Combine them all.
- Which one should we use?
  - denoising objectives are great but pretty insufficient as a standalone objective -- less “loss exposure”, also a little contrived
  - Causal LMs – high loss exposure, natural to formulate – also enables other interesting phenomenon like “few-shot learning” – with enough data and scale, Causal LMs turned out to be just as good without needing bidirectional context.

What happened to BERT & T5? On Transformer Encoders, PrefixLM and Denoising Objectives — Yi Tay

Going forward – decoder only LMs

- We will (mostly) talk about decoder only LMs going forward as they are most commonly used now
    - Like GPTs, Claude, Gemini, Llama, Deepseek, Qwen, Kimi, GLM, OLMo, and many more.

What Can We Do with LMs?

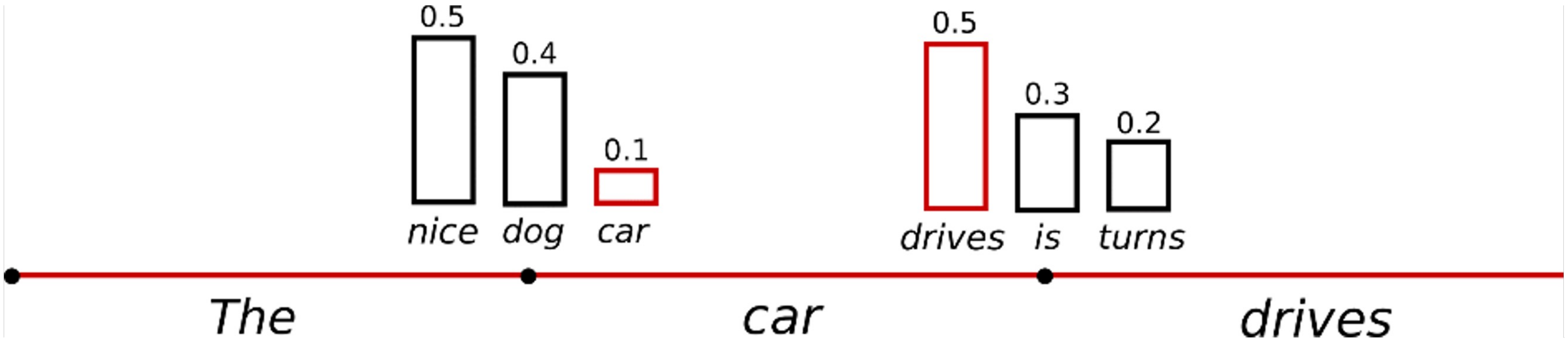
- Given a sequence $\overline{x}$ compute the probability of the sequence
  - $p(\overline{x}) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{i-1})$
- Given a prefix, generate a sequence autoregressively (i.e., generating one token at a time)
  - The prefix can be empty (sort of: always includes a start token)
  - This prefix is called a **prompt**

Decoding strategies

(Ancestral) Sampling

- Sampling:

$$x_i \sim p(x_i | x_1, \dots, x_{i-1}) \text{ until } x_i = \text{STOP}$$



This can often generate incoherent gibberish

Source: https://huggingface.co/blog/how-to-generate

Greedy Decoding

- Greedy (i.e., $\text{arg}max$):

$$x_i = \text{arg}\underset{x_1 \in \mathcal{V}}{max}\, p(x_i | x_1, \ldots, x_{i-1}) \text{ until } x_i = \text{STOP}$$

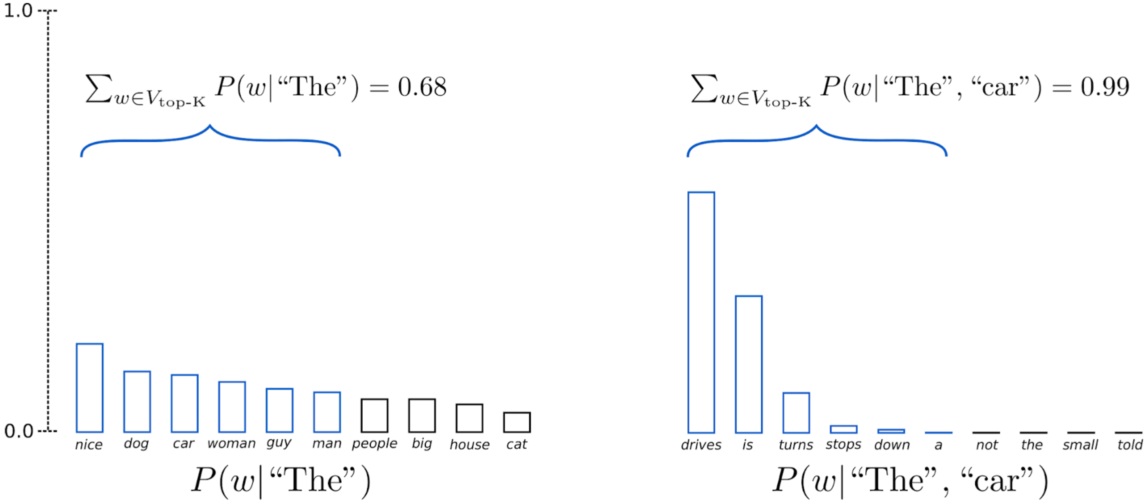
- How many different strings can we generate this way?

Top-k sampling [Fan et al., 2018]

Filter k most likely next tokens and redistributed the probability mass among only those k tokens, then sample from the new distribution

Problem: It doesn't dynamically adapt the number of words that are filtered from the next word probability distribution

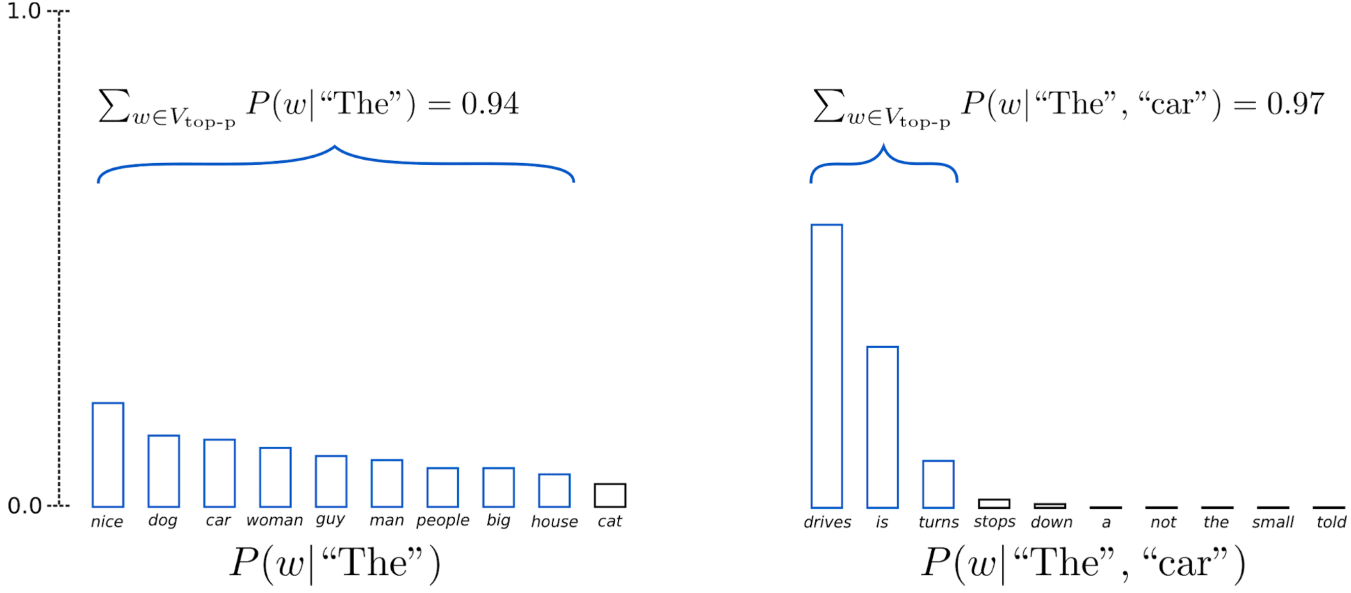
Reasonable candidates (left fig) are eliminated, and ill-fitted (right fig) are not



Source: https://huggingface.co/blog/how-to-generate

Top-p (nucleus) sampling [Holtzman et al., 2020]

Sample from the smallest possible set of tokens whose cumulative probability exceeds the probability p



Source: https://huggingface.co/blog/how-to-generate

Adjusting Distribution **Temperature**

- Let’s say we want something between sampling and greedy
  - Not fully deterministic
  - But to control how focused on the top of the distribution with high likelihood
- Add a temperature parameter to the softmax
  - Given z is the vector with logits, and $T \in \mathbb{R}$ in the temperature

$$p(y_i \mid x) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

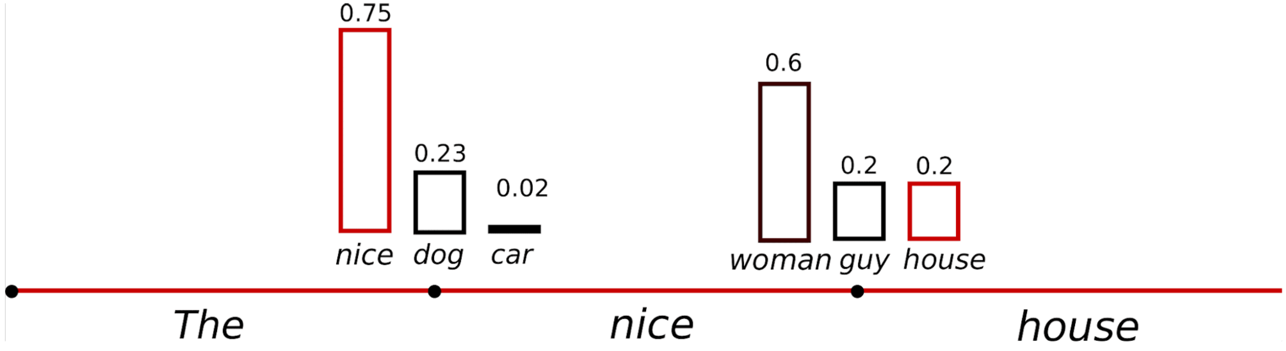
Higher T: softens probabilities.
Lower T: sharpens probabilities.

Adjusting Distribution **Temperature**

- Add a temperature parameter to the softmax
  - Given z is the vector with logits, and $T \in \mathbb{R}$ in the temperature

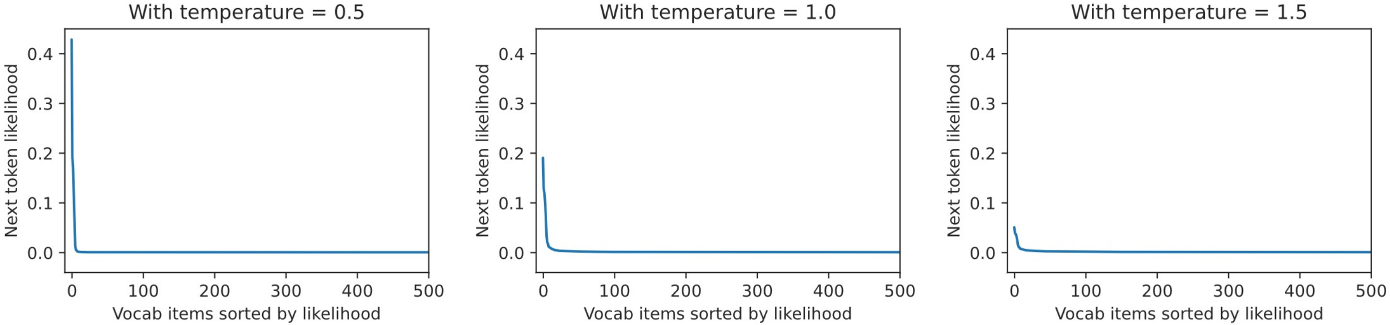
$$p(y_i \mid x) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Higher T: softens probabilities.
Lower T: sharpens probabilities.



Adjusting Distribution Temperature

- What happens with $T = 1$? $T = 0$ (or almost)? $T \in [0,1)$? $T > 1$?



Plots from Daphne Ippolito / Chenyan Xiong, CMU LLMs course http://cmu-llms.org/

Decoding

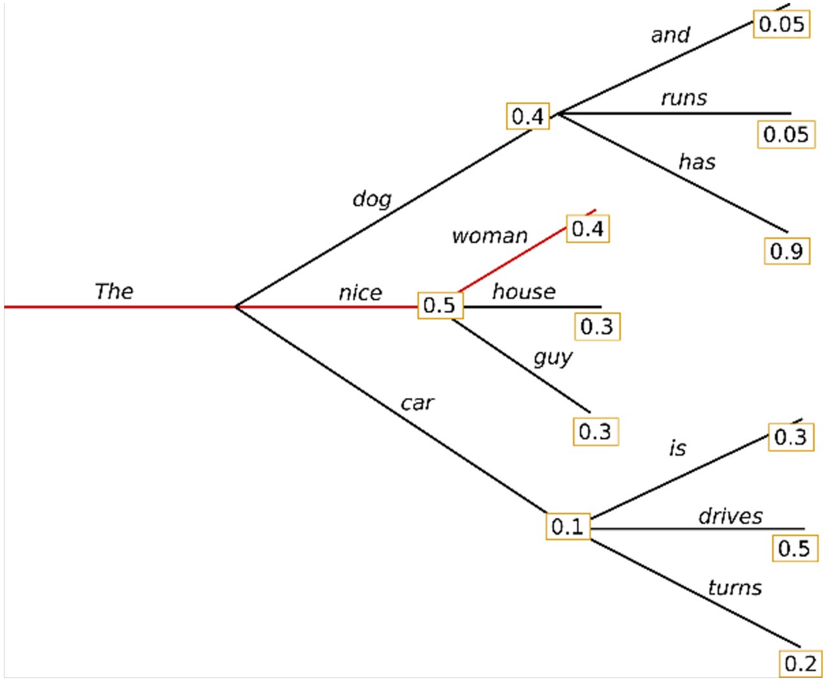
- Various decoding techniques: greedy, sampling, temperature-based, top-k, nucleus
- Most common: temperature-based
- Which are guaranteed to give you the **optimal** output? Will $\arg max$ give you the optimal output?

Decoding

- Various decoding techniques: greedy, sampling, temperature-based, top-k, nucleus
- Most common: temperature-based
- Which are guaranteed to give you the optimal output? Will $\mathrm{arg}max$ give you the optimal output?

	0,2	0,5	0,1
Output 1	I	love	Lucy
	0,2	0,1	0,99
Output 2	I	hate	Lucy

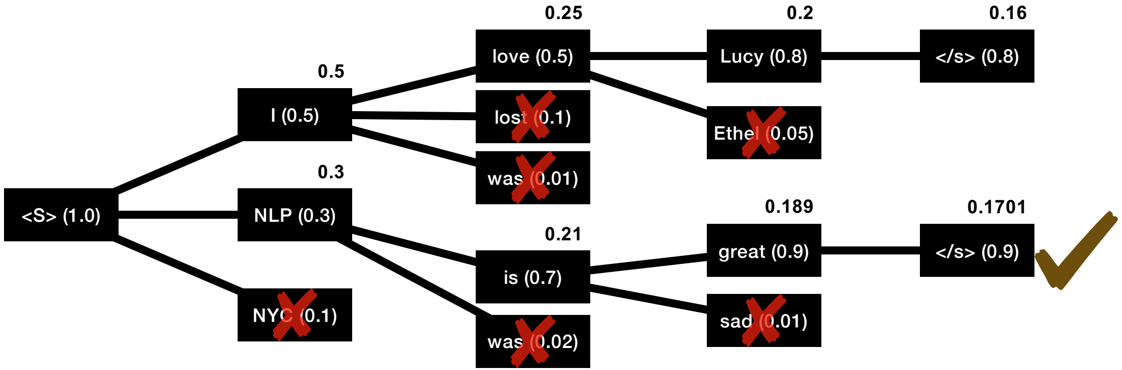
Greedy decoding/search

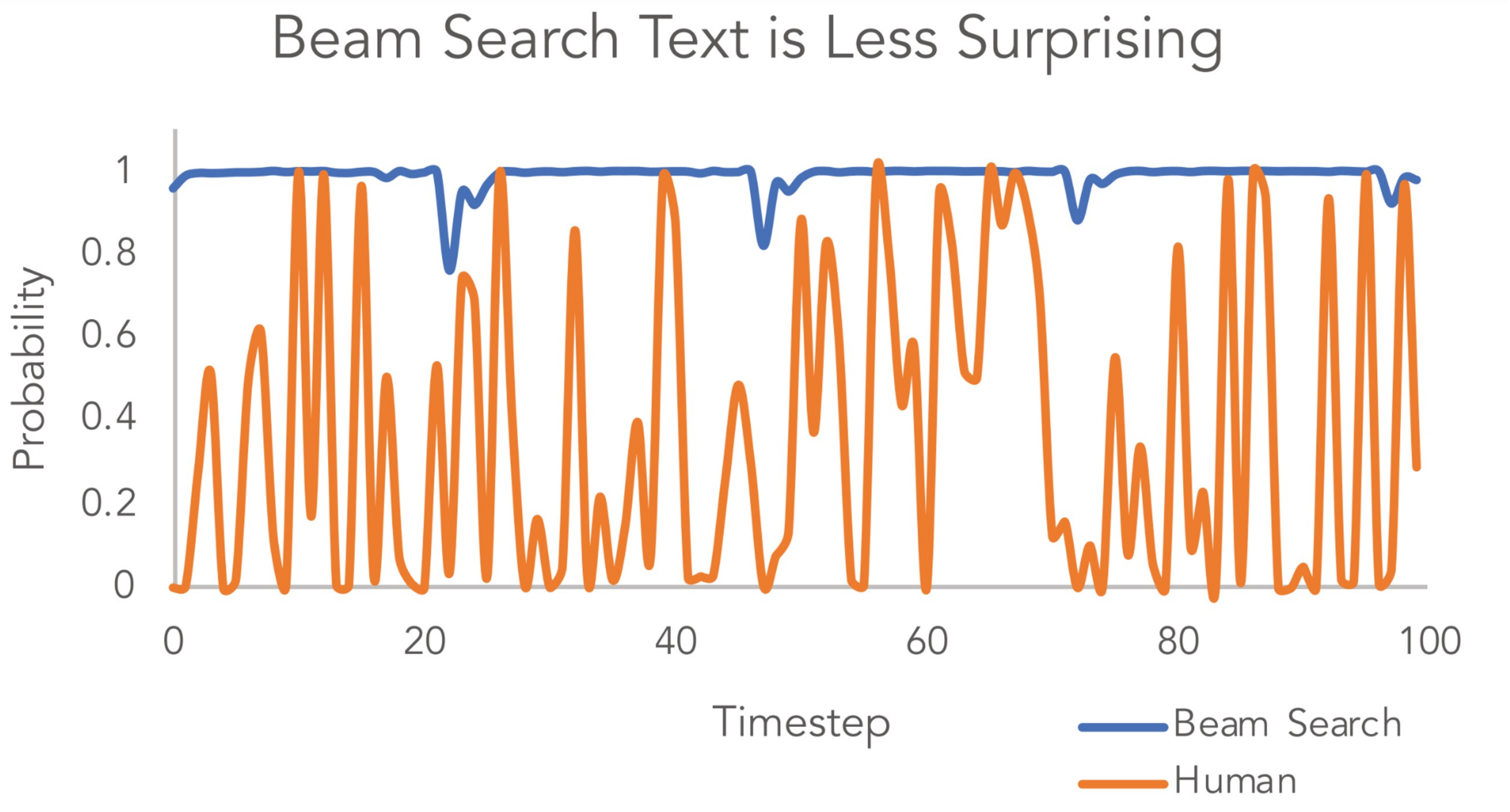


Source: https://huggingface.co/blog/how-to-generate

Beam Search

- Sampling techniques are not optimal
  - Following a single hypothesis is just not sufficient, but enumerating all is intractable
- **Beam search** is middle ground
  - Follow a set of hypothesis, always keeping the top ones
  - The size of the set B is a hyperparameter





Beam Search Text is Less Surprising

Source: [Holtzman et al., 2020]

Beam Search

- Sampling techniques are not optimal
  - Following a single hypothesis is just not sufficient, but enumerating all is intractable
- Beam search is middle ground
  - Follow a set of hypothesis, always keeping the top ones
  - The size of the set B is a hyperparameter
  - It's an approximation method
  - What happens with $B = 1$? $B = \infty$?
  - What is the cost of beam search compared to the sampling techniques we saw?
  - Can you combine sampling techniques with beam search?
  - Beam search is almost a relic of the past, was very common in tasks like translation, summarization, not used anymore.

https://aclanthology.org/2020.emnlp-main.170/