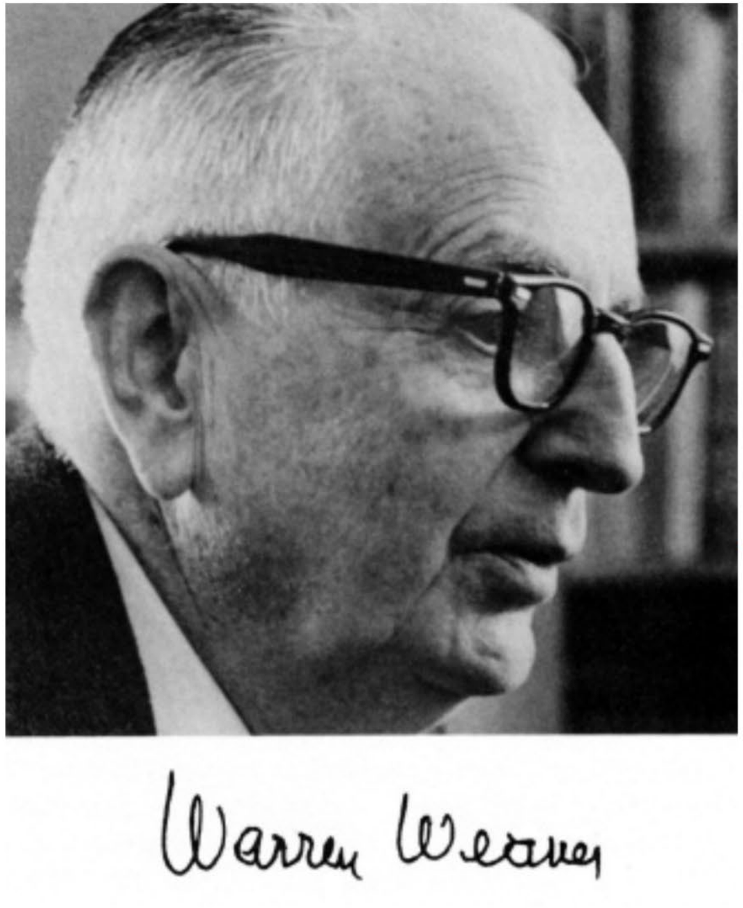
Machine Translation & "Foundational" Models

601.764

1/26/23



When I look at an article in Russian, I say:
”This is really written in English, but it has been coded in some strange symbols.
I will now proceed to decode”.

Warren Weaver,
American scientist (1894-1978)

Image courtesy of: Biographical Memoirs of the National Academy of Science, Vol. 57

Progress in MT

Warren Weaver's memo
1947

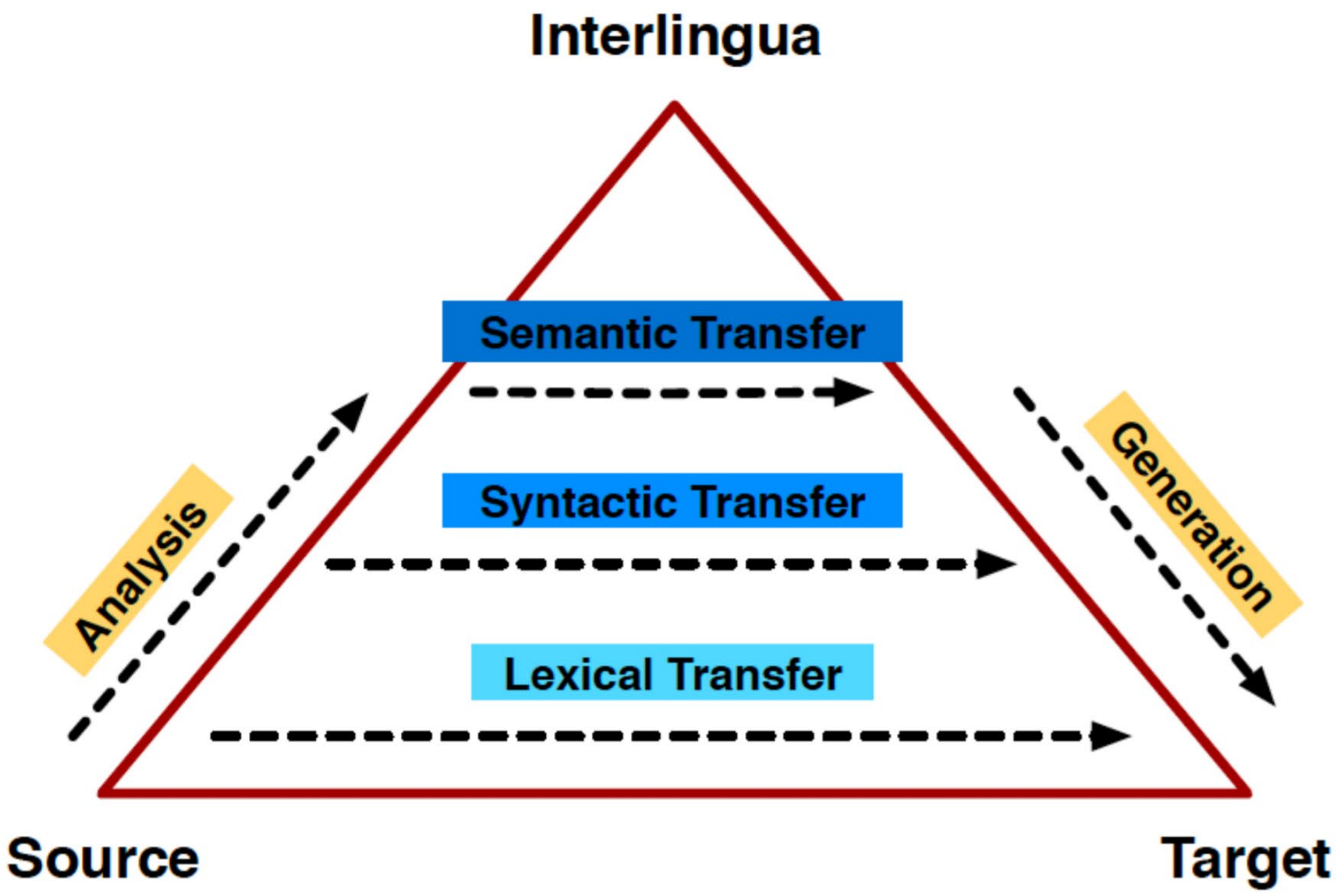
Founding of SYSTRAN. Development of Rule-based MT (RBMT)
1968

Seminal SMT paper from IBM
1993

DARPA TIDES, GALE, BOLT programs
Open-source of Moses toolkit
Development of Statistical MT (SMT)
Early 2000s

2011-2012: Early deep learning success in speech/vision
2015: Seminal NMT paper (RNN+attention)
2016: Google announces NMT in production
2017: New NMT architecture: Transformer
2010s-Present

Vauquois Triangle



Rule-Based Machine Translation

- Build Dictionaries
- Write Transformation Rules

```
"have" :=

if
    subject(animate)
    and object(owned-by-subject)
then
    translate to "kade...  aahe"
if
    subject(animate)
    and object(kinship-with-subject)
then
    translate to "laa...  aahe"
if
    subject(inanimate)
then
    translate to "madhye...  aahe"
```

Statistical Machine Translation

كيف حالك

How are you?

Statistical Machine Translation

كيف حالك →  → How are you?

Source sentence, f

Statistical Machine Translation

كيف حالك

Source sentence, f

How are you?

Target sentence, e

Statistical Machine Translation

$P(e \mid f)$

كيف حالك →  → How are you?

Source sentence, f

Target sentence, e

Statistical Machine Translation (SMT)

Data, Data, Data!

Statistical Machine Translation (SMT)

- Learn Dictionaries from Data

Statistical Machine Translation (SMT)

- Learn Dictionaries from Data   "farok" → "jjat"

Statistical Machine Translation (SMT)

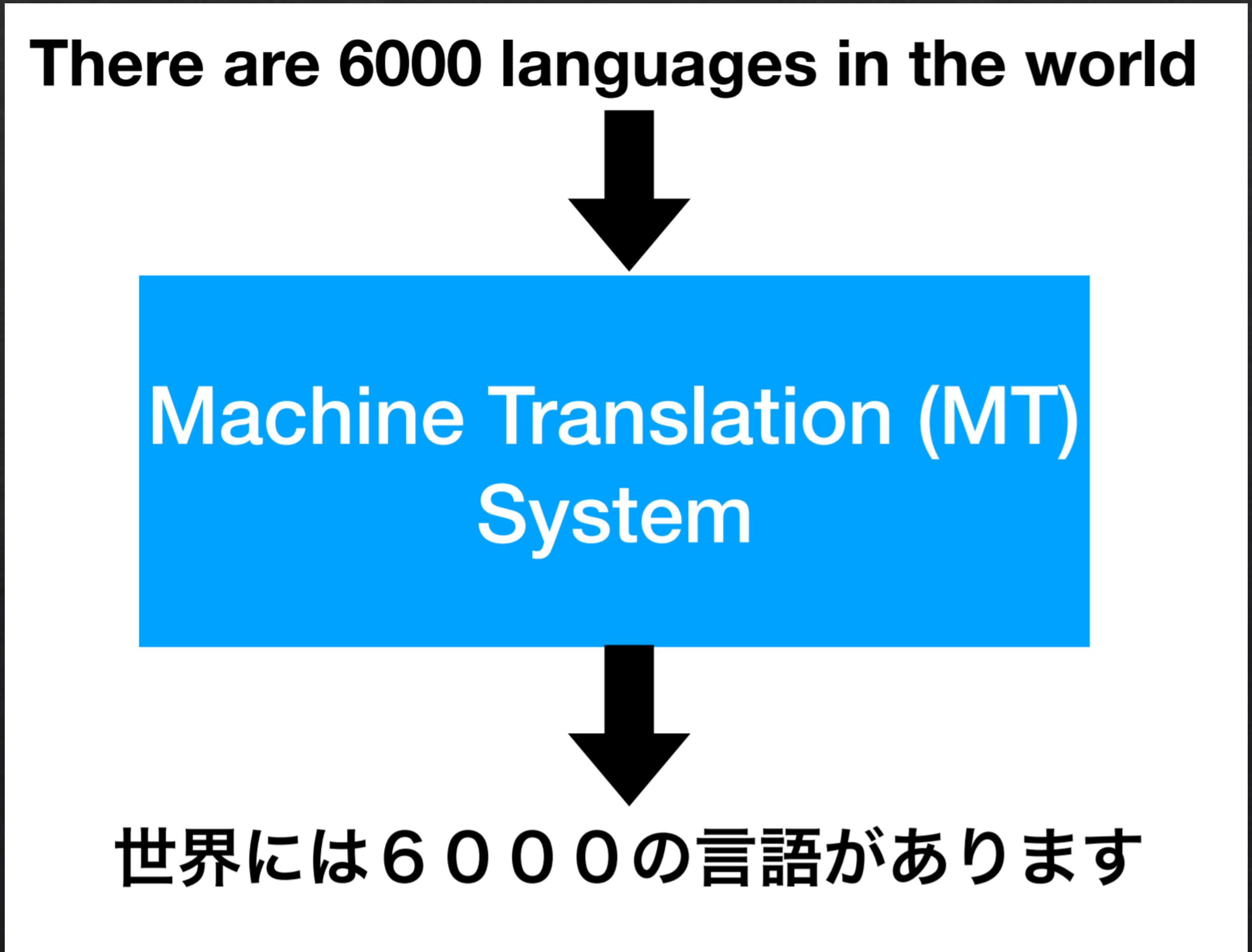
- Learn Dictionaries from Data "farok" → "jjat"
- Learn "Rules" from Data
- 1980 - 2015

Bitexts

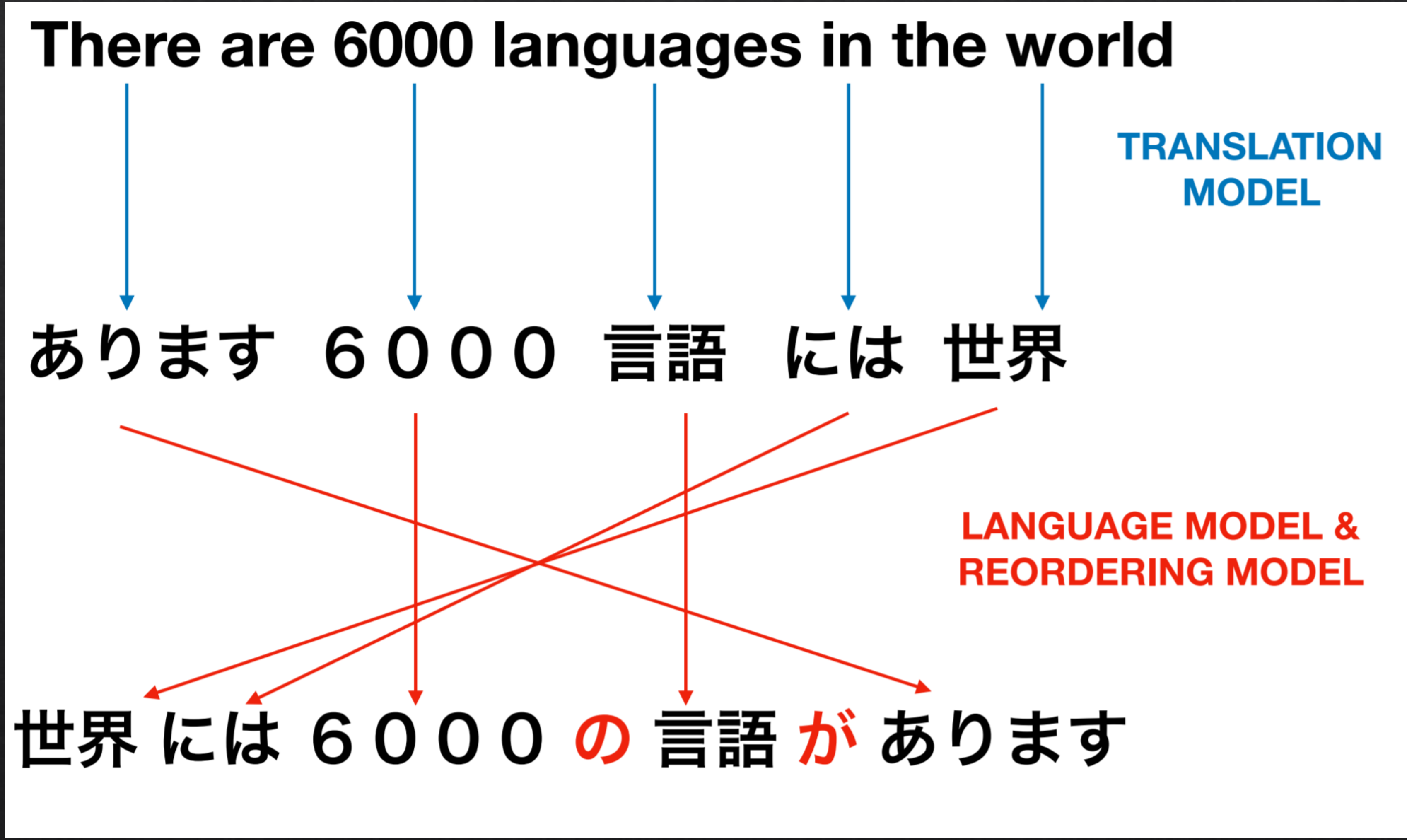
Johns hopkins university was founded
The mascot is a blue jay
Blue jays are the mascot
Lacrosse is a well-known sport

Đại học Johns hopkins được thành lậpLinh vật là một con chim giẻ cùi màu xanhGiẻ cùi xanh là linh vậtLacrosse là một môn thể thao nổi tiếng

Machine Translation (Abstraction)



Machine Translation (SMT) … simplified



SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	

SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	Input: Source Sentence

SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	

SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	Output: Target Sentence

SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	Output: Target Sentence
Automatically Learn from Bitext	

SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	Output: Target Sentence
Automatically Learn from Bitext	Automatically Learn from Bitext

SMT versus NMT

Statistical MT	**Neural MT**
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	Output: Target Sentence
Automatically Learn from Bitext	Automatically Learn from Bitext
Probabilistic Translation Model	

SMT versus NMT

Statistical MT	**Neural MT**
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	Output: Target Sentence
Automatically Learn from Bitext	Automatically Learn from Bitext
Probabilistic Translation Model	
Probabilistic Reordering Model	

SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	Output: Target Sentence
Automatically Learn from Bitext	Automatically Learn from Bitext
Probabilistic Translation Model	
Probabilistic Reordering Model	
Probabilistic Language Model	

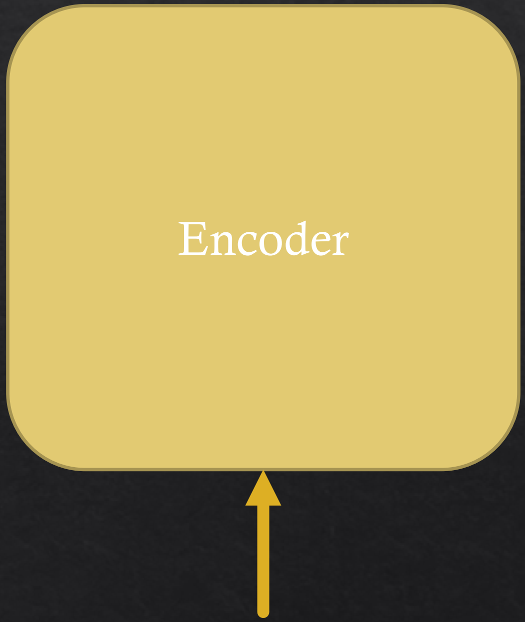
SMT versus NMT

Statistical MT	Neural MT
Input: Source Sentence	Input: Source Sentence
Output: Target Sentence	Output: Target Sentence
Automatically Learn from Bitext	Automatically Learn from Bitext
Probabilistic Translation Model	One Neural Model (Probabilistic)
Probabilistic Reordering Model	
Probabilistic Language Model	

Neural Machine Translation (NMT)

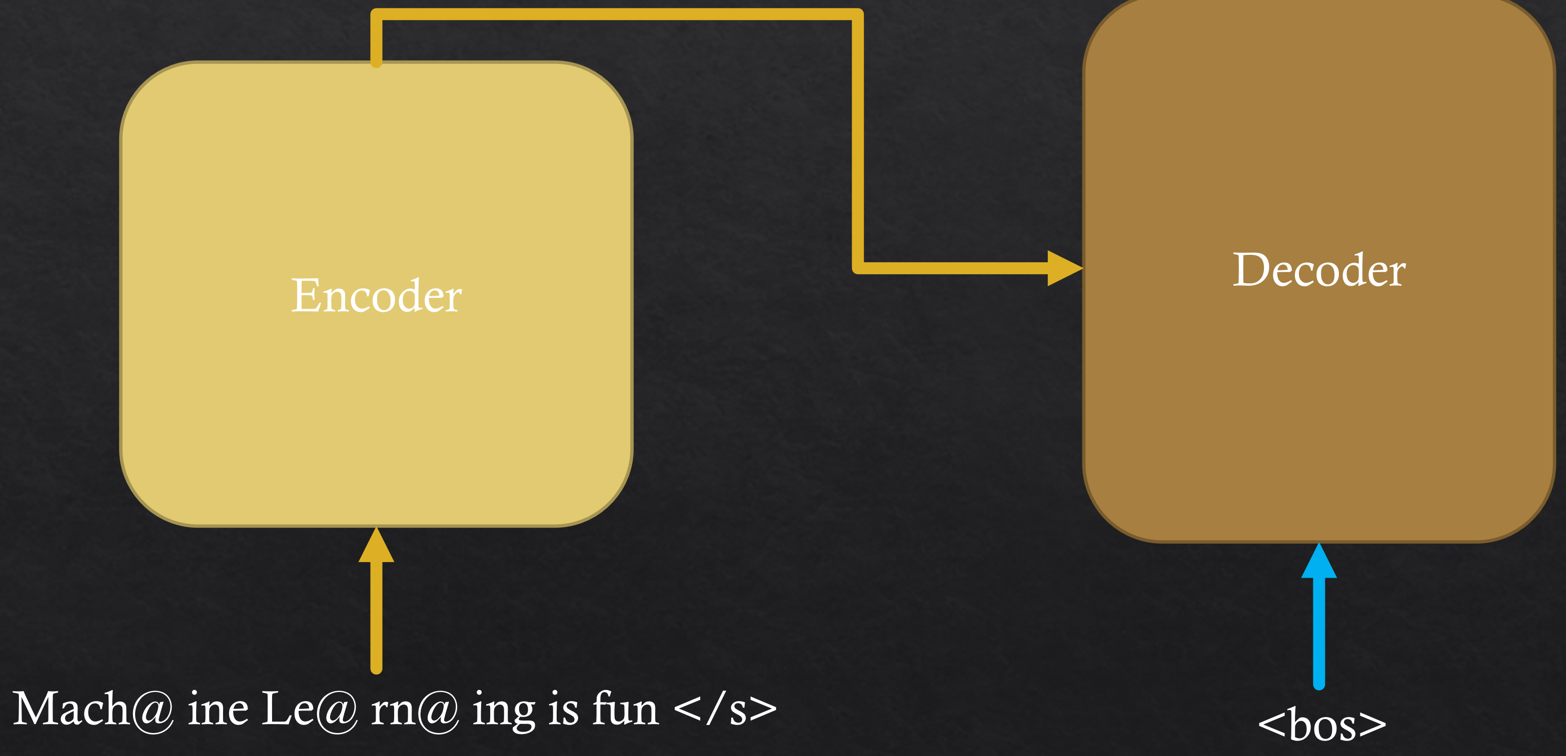
- Also a type of Statistical MT
- Represent words in high-dimensional, continuous, space
- $P(e|f)$

Neural Machine Translation (NMT)

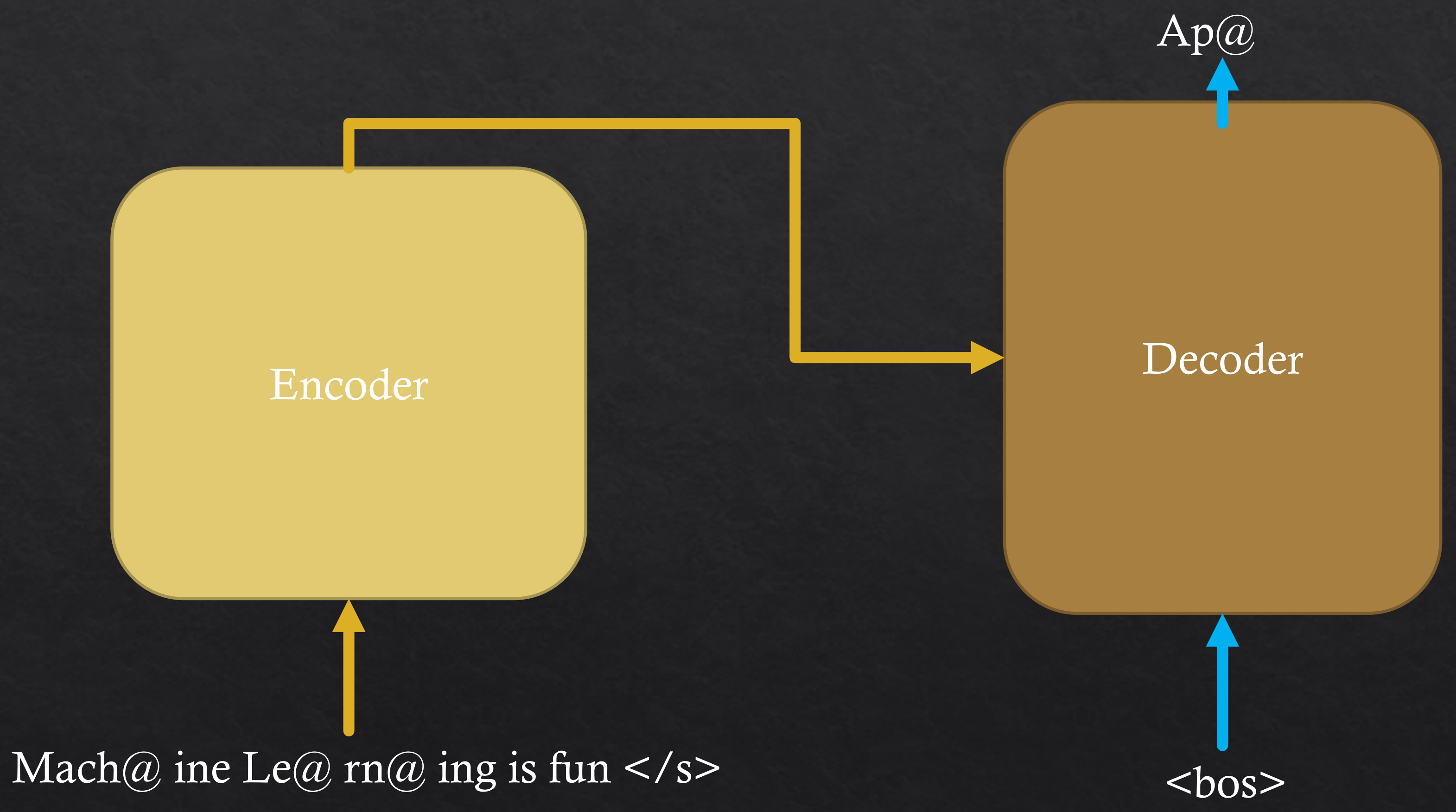


Mach@ ine Le@ rn@ ing is fun </s>

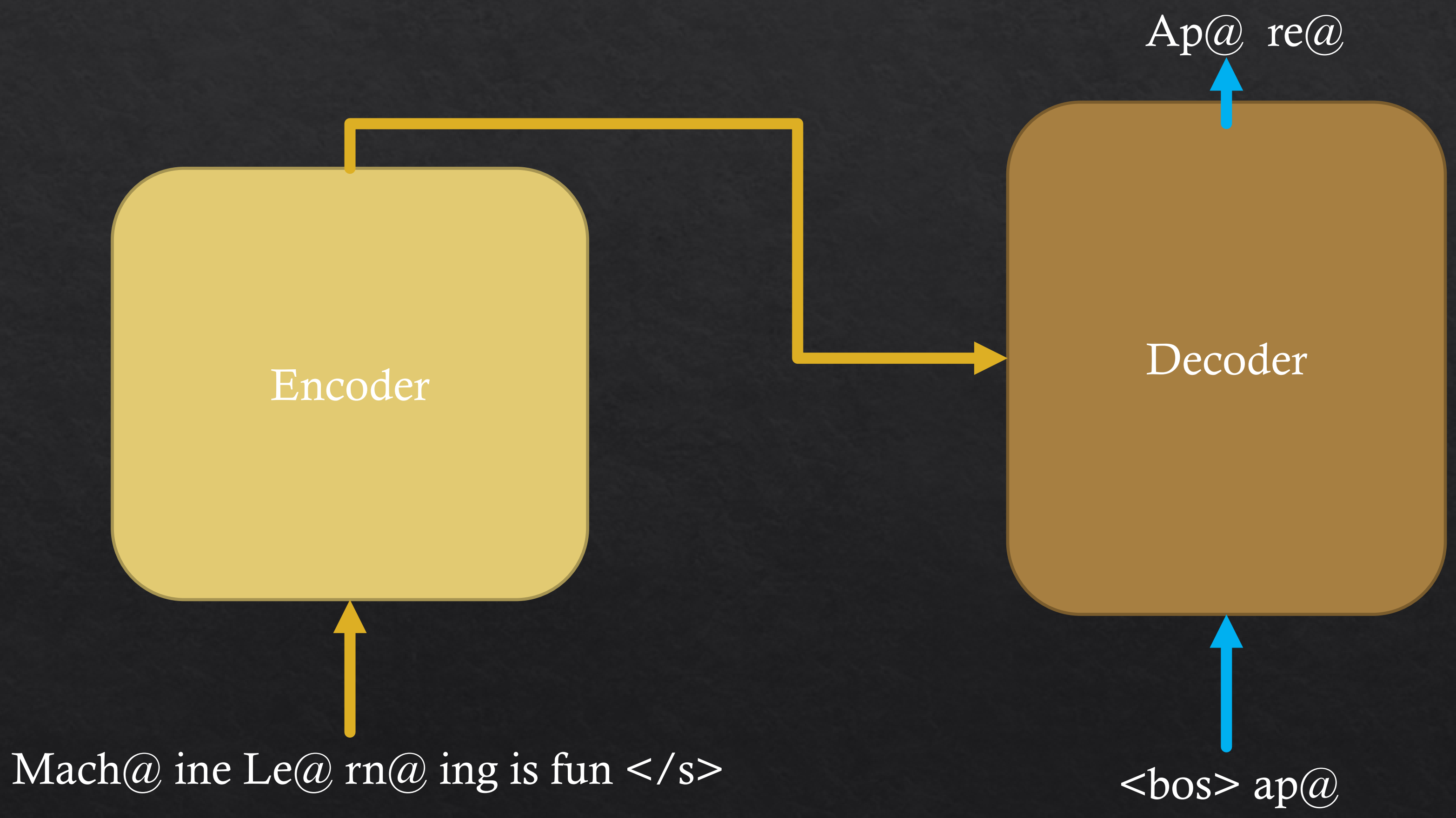
Neural Machine Translation (NMT)



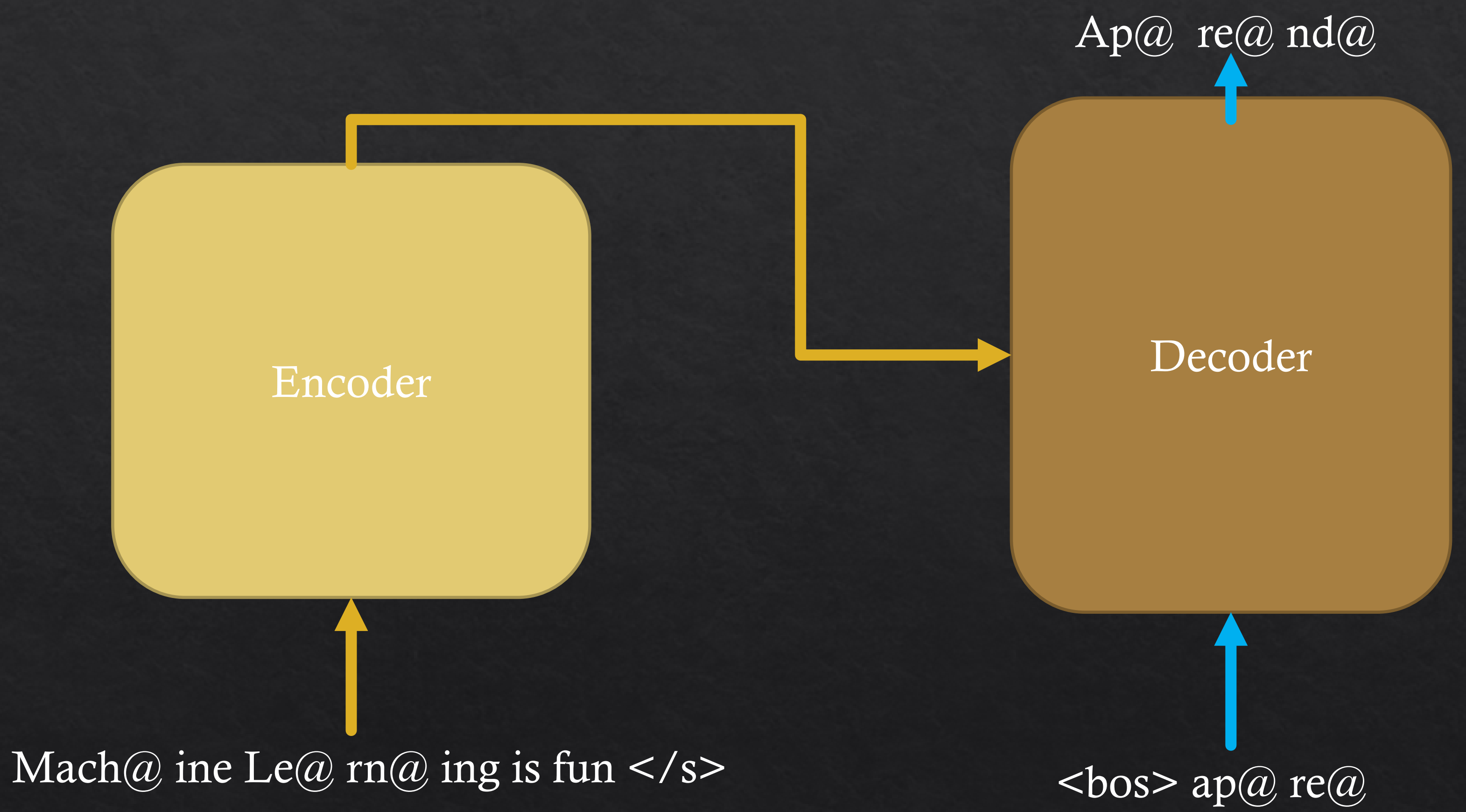
Neural Machine Translation (NMT)



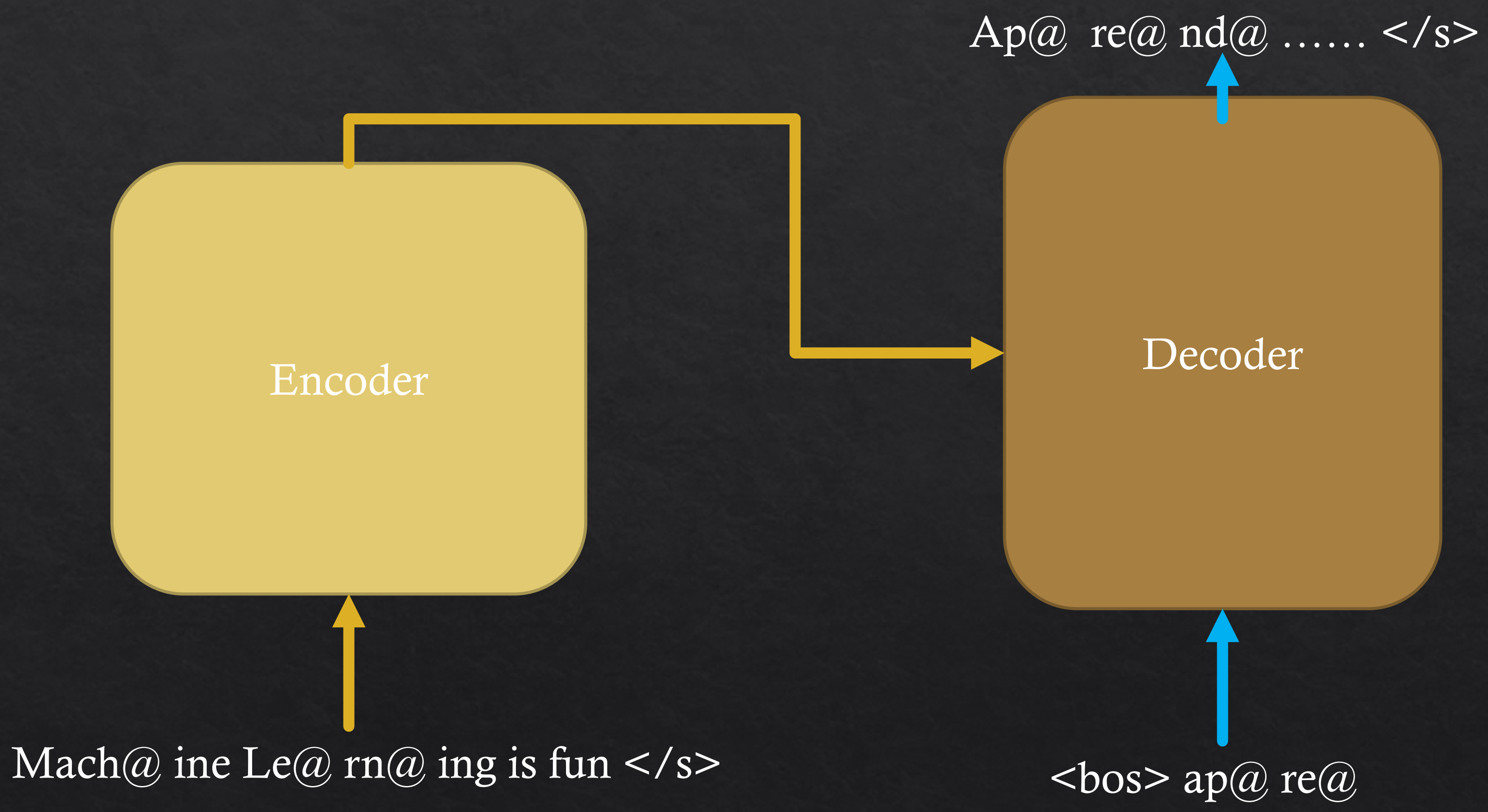
Neural Machine Translation (NMT)



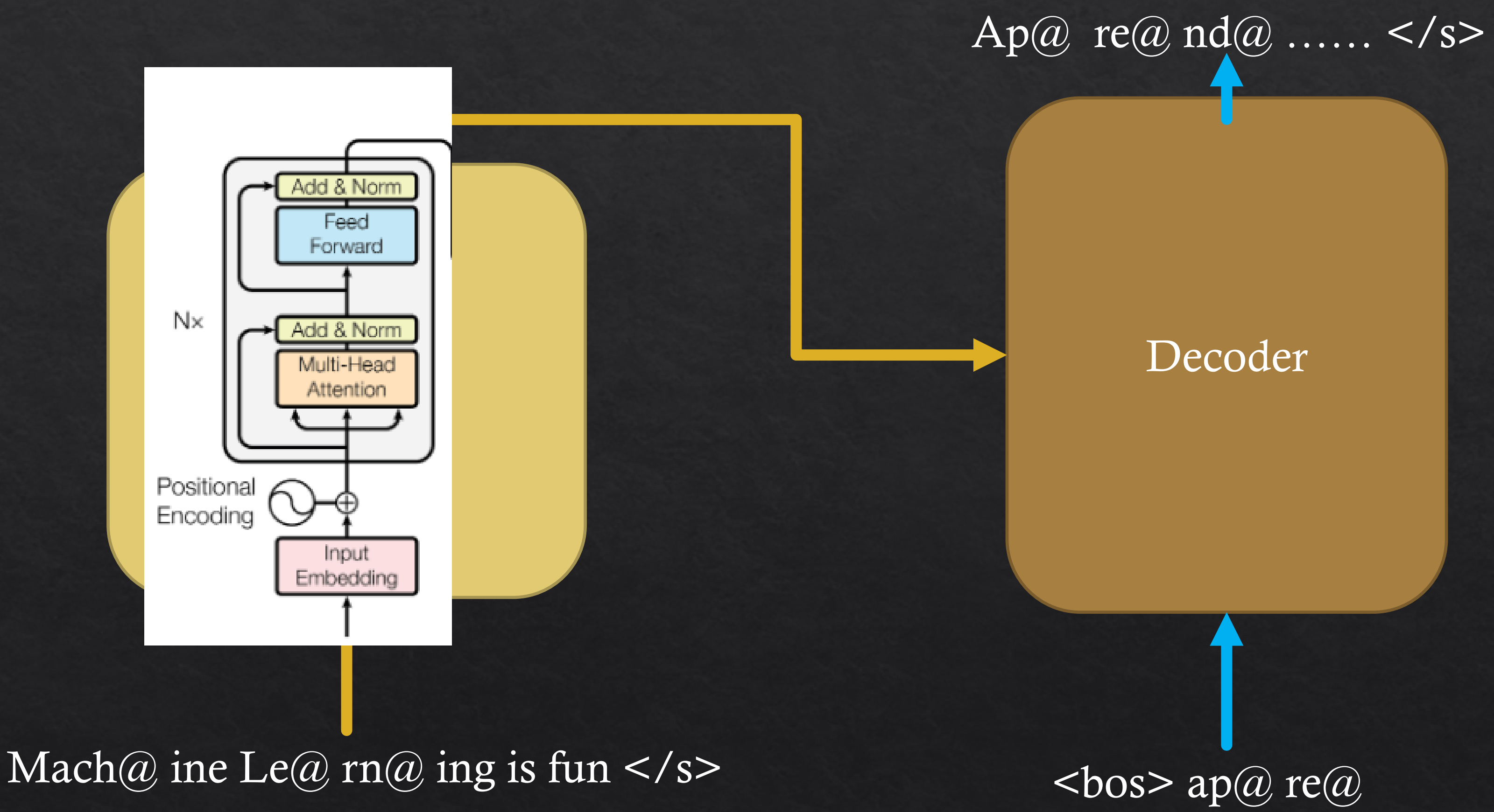
Neural Machine Translation (NMT)



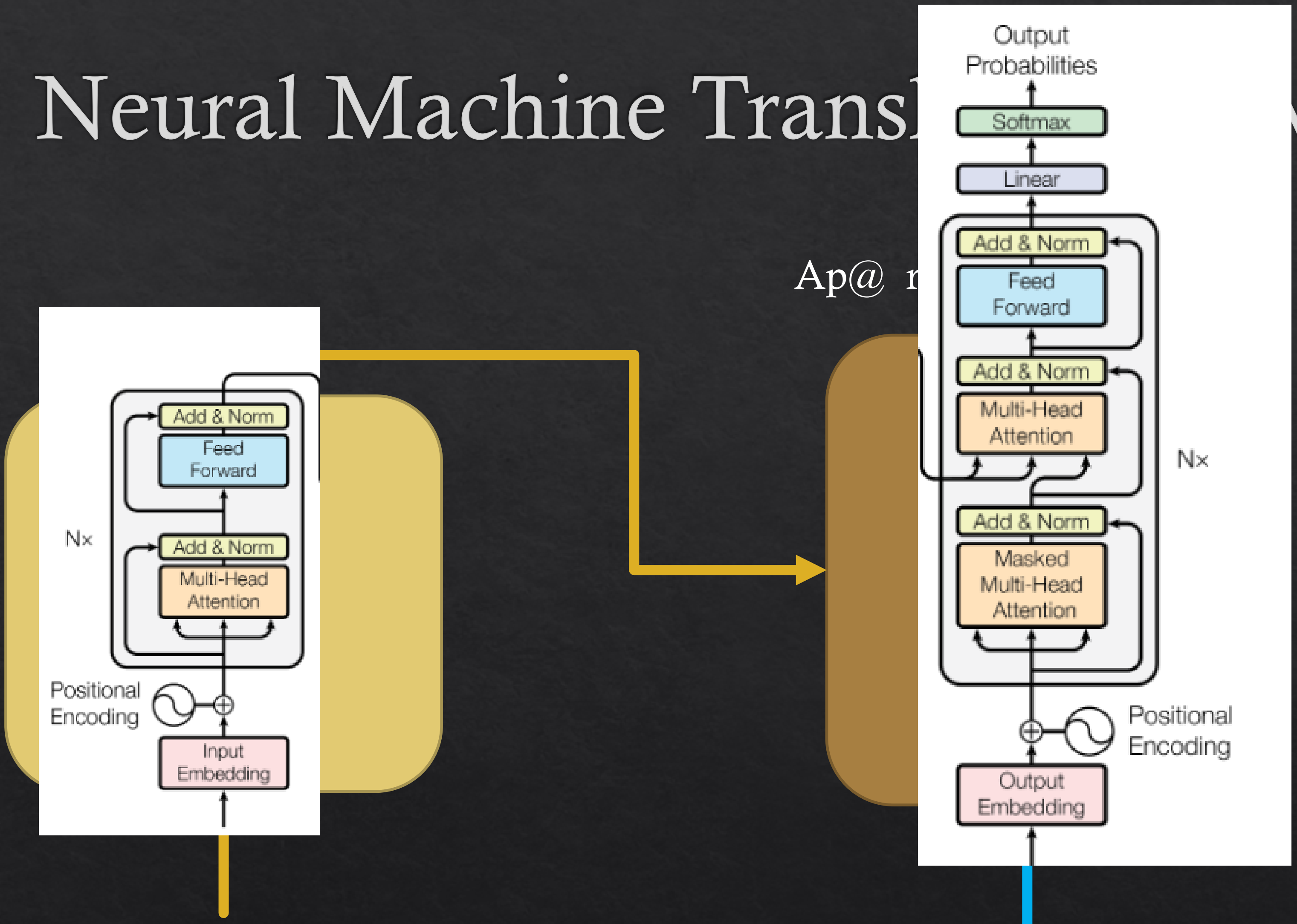
Neural Machine Translation (NMT)



Neural Machine Translation (NMT)



Neural Machine Transl MT)



Mach@ ine Le@ rn@ ing is fun </s>

<bos> ap@ re@

Vocabulary

One-Hot Vector

- Words correspond to index in vector
- *Fixed* size

One-Hot Vector

Dictionary

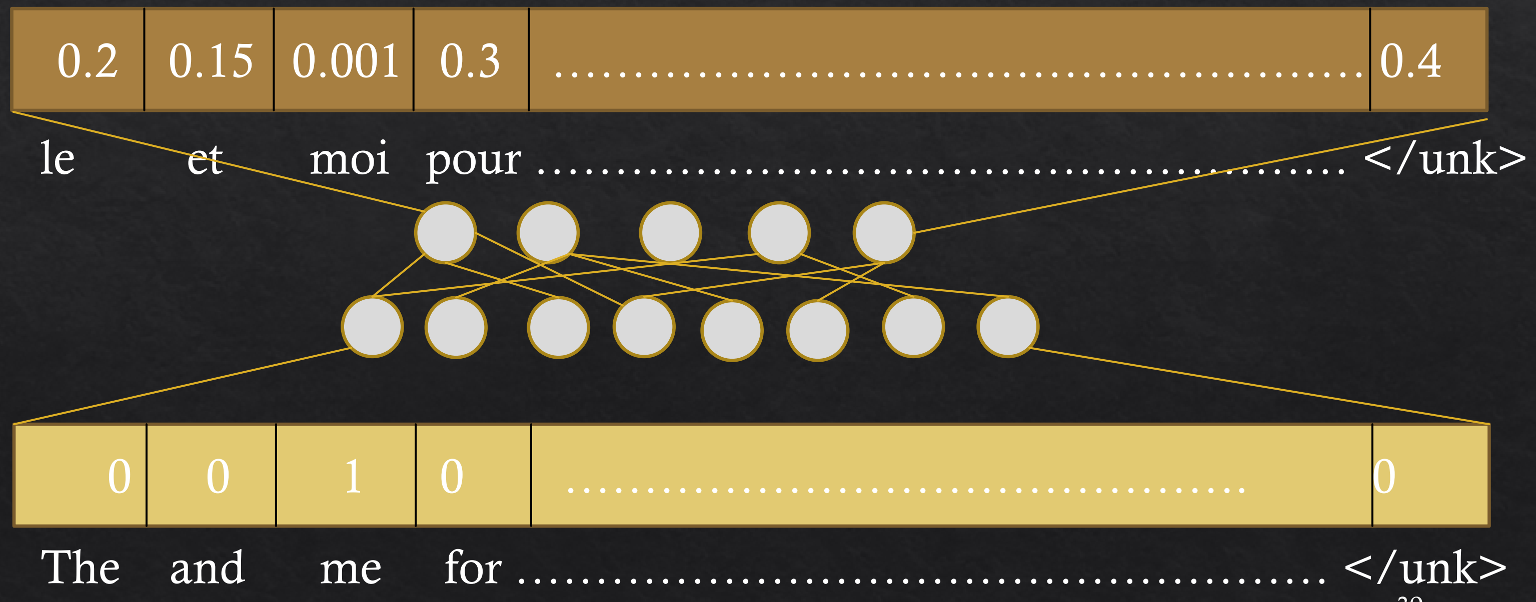
The	And	I	Dog	Johns	…….	Me	Cat

<u>Johns</u> Hopkins was founded in …..

0	0	0	0	1	……	0	0

Vocabulary Size

- Fixed Input & Output Vector Dimensions
- Out-of-vocabulary (OOVs)



Character Level

- No OOVs
- Very long sequences

Byte Pair Encoding

- Subword Unit
- Based on a compression algorithm
- Start small, repeatedly combine

peter piper picked a peck of pickled peppers

p@ e@ t@ e@ r p@ i@ p@ e@ r p@ i@ c@ k@ e@ d a p@ e@ c@ k o@ f p@ i@ c@ k@ l@ e@ d p@ e@ p@ p@ e@ r@ s

Vocabulary		Rules
p t i k a f s	e r c d o l	

p@ e@ t@ e@ r p@ i@ p@ e@ r p@ i@ c@ k@ e@ d a p@ e@ c@ k o@ f p@ i@ c@ k@ l@ e@ d p@ e@ p@ p@ e@ r@ s

Vocabulary		Rules
p	e	p@ e@ → pe@
t	r	
i	c	
k	d	
a	o	
f	l	
s	pe@	

pe@ t@ e@ r p@ i@ p@ e@ r p@ i@ c@ k@ e@ d a pe@ c@ k o@ f p@ i@ c@ k@ l@ e@ d pe@ p@ pe@ r@ s

Vocabulary		Rules
p t i k a f s pi@	e r c d o l pe@	p@ e@ → pe@ p@ i@ → pi@

pe@ t@ e@ r pi@ p@ e@ r pi@ c@ k@ e@ d a pe@ c@ k o@ f pi@ c@ k@ l@ e@ d pe@ p@ pe@ r@ s

Vocabulary		Rules
p	e	p@ e@ → pe@
t	r	p@ i@ → pi@
i	c	pi@ c@ → pic@
k	d	
a	o	
f	l	
s	pe@	
pi@	pic@	

pe@ t@ e@ r pi@ p@ e@ r pic@ k@ e@ d a pe@ c@ k o@ f pic@ k@ l@ e@ d pe@ p@ pe@ r@ s

Vocabulary		Rules
p	e	p@ e@ → pe@
t	r	p@ i@ → pi@
i	c	pi@ c@ → pic@
k	d	e@ d → ed
a	o	
f	l	
s	pe@	
pi@	pic@	

peter piper picked a peck of pickled peppers

Vocabulary		**Rules**
p	e	p@ e@ → pe@
t	r	p@ i@ → pi@
i	c	pi@ c@ → pic@
k	d	e@ d → ed
a	o	.
f	l	.
s	pe@	.
pi@	pic@	.
....		.
of		o@ f → of

How Good are our Translations?

- "Iyunivesithi yasekwa ngonyaka olandelayo ukusweleka kwakhe kwaye yanikezelwa kunyana wabo okuphela kwakhe."

- The university was founded in the year following his death and was dedicated to their only son.
- The university was following his death and was dedicated to their only son.
- Machine Translation always works perfectly.

MT Evaluation

- Human Evaluation (Expensive, but best)
- Automatic (Cheap, can be correlated)

BLEU Scores

- Modified *n*-gram precision
- BiLingual Evaluation Understudy

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus.

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around.

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16			

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16			

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16			

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15		

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	/14	

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	/13

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	/13

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	/13

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	/13

BLEU Scores

Gold: The Blue Jays are the mascot of Johns Hopkins University and can be seen around campus

Hyp: The Blue Jays are mascot of Johns Hopkins University University and can be seen around

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	4/13

BLEU Scores

$$BLEU = \min\left(1, e^{1-\frac{len(gold)}{len(hyp)}}\right) \sqrt[4]{\prod_{i=1}^{4} precision(i)}$$

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	4/13

BLEU Scores

$$BLEU = \underbrace{\min\left(1, e^{1-\frac{len(gold)}{len(hyp)}}\right)}_{\text{Brevity Penalty}} \sqrt[4]{\prod_{i=1}^{4} precision(i)}$$

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	3/13

BLEU Scores

$$BLEU = \min\left(1, e^{1-\frac{len(gold)}{len(hyp)}}\right) \sqrt[4]{\prod_{i=1}^{4} precision(i)}$$

$$BLEU = \min\left(1, e^{1-\frac{16}{15}}\right) \sqrt[4]{0.138}$$

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	3/13

BLEU Scores

$$BLEU = \min\left(1, e^{1-\frac{len(gold)}{len(hyp)}}\right) \sqrt[4]{\prod_{i=1}^{4} precision(i)}$$

$$BLEU = 0.94\sqrt[4]{0.138}$$

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	3/13

BLEU Scores

$$BLEU = \min\left(1, e^{1-\frac{len(gold)}{len(hyp)}}\right) \sqrt[4]{\prod_{i=1}^{4} precision(i)}$$

$$BLEU = 0.94 * 0.609$$

$$BLEU = 0.57$$

Unigrams	Bigrams	Trigrams	4-grams
14/16	12/15	9/14	3/13

BLEU Scores

- Calculate over *entire* test set (not one sentence)
- < 10 …. Pretty useless
- 10 – 20 … can get some meaning
- 20 – 30 … looks decent
- > 30 starts getting pretty good

BLEU Scores

- The large house
- A big mansion

Large Language Models
Foundational Models….

Language Modeling

- Create a model of language
- Frequently probabilistic/statistical
- Used for downstream tasks & predictions

Traditional Applications

- Autocorrect
- Translation
- Speech Recognition

2-gram

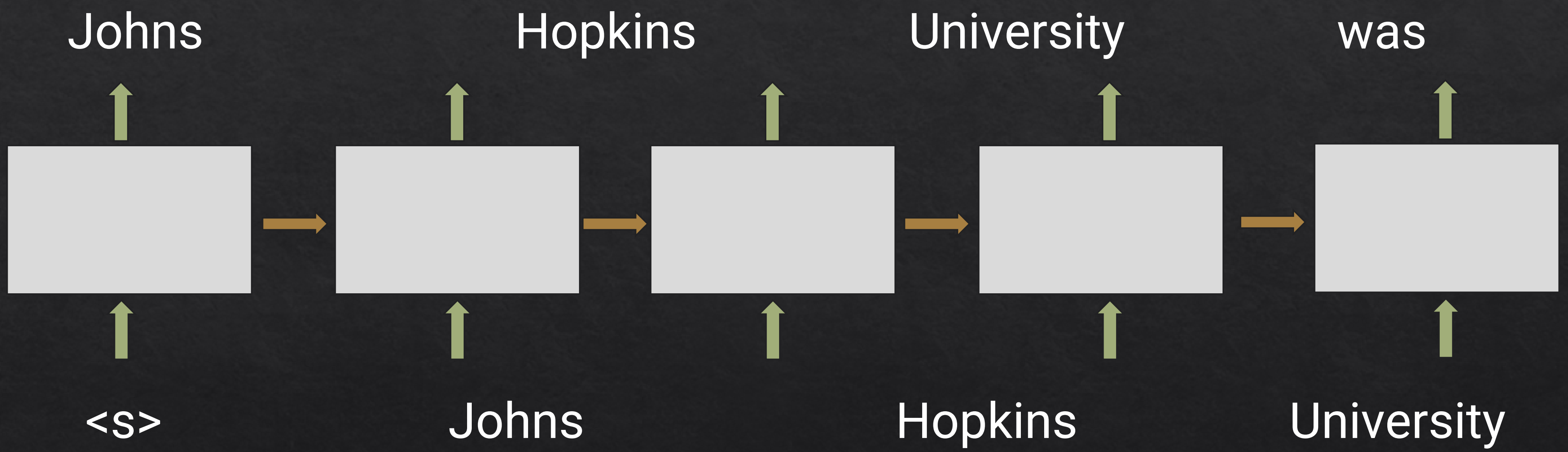
Johns ________

3-gram

Johns Hopkins __________

Backoff

RNN-LM



Masked Language Models

- No longer need to view everything left-to-right*
- Mask out random words in a sentence, not the sequence

BERT

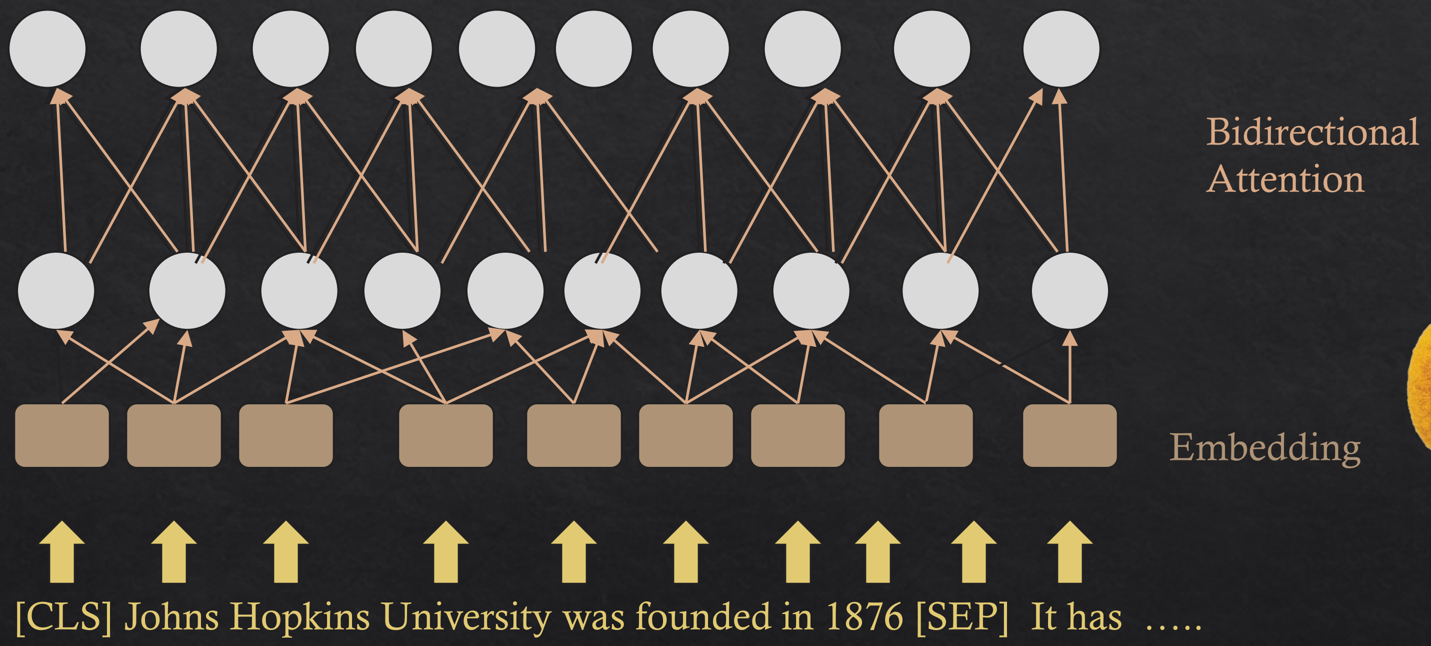
Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” 2018

BERT

[CLS] Johns Hopkins University was founded in 1876 [SEP] It has .....

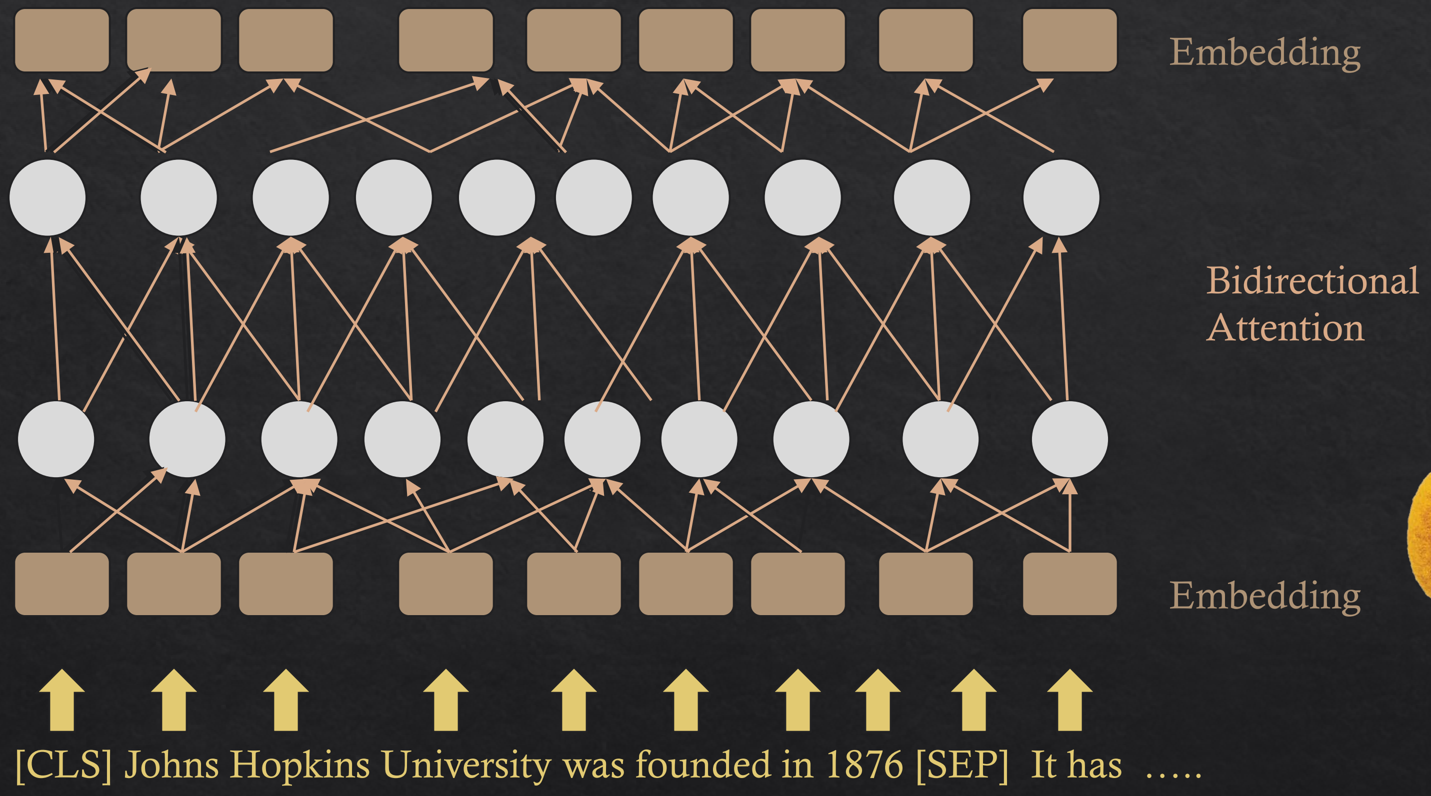
Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” 2018

BERT



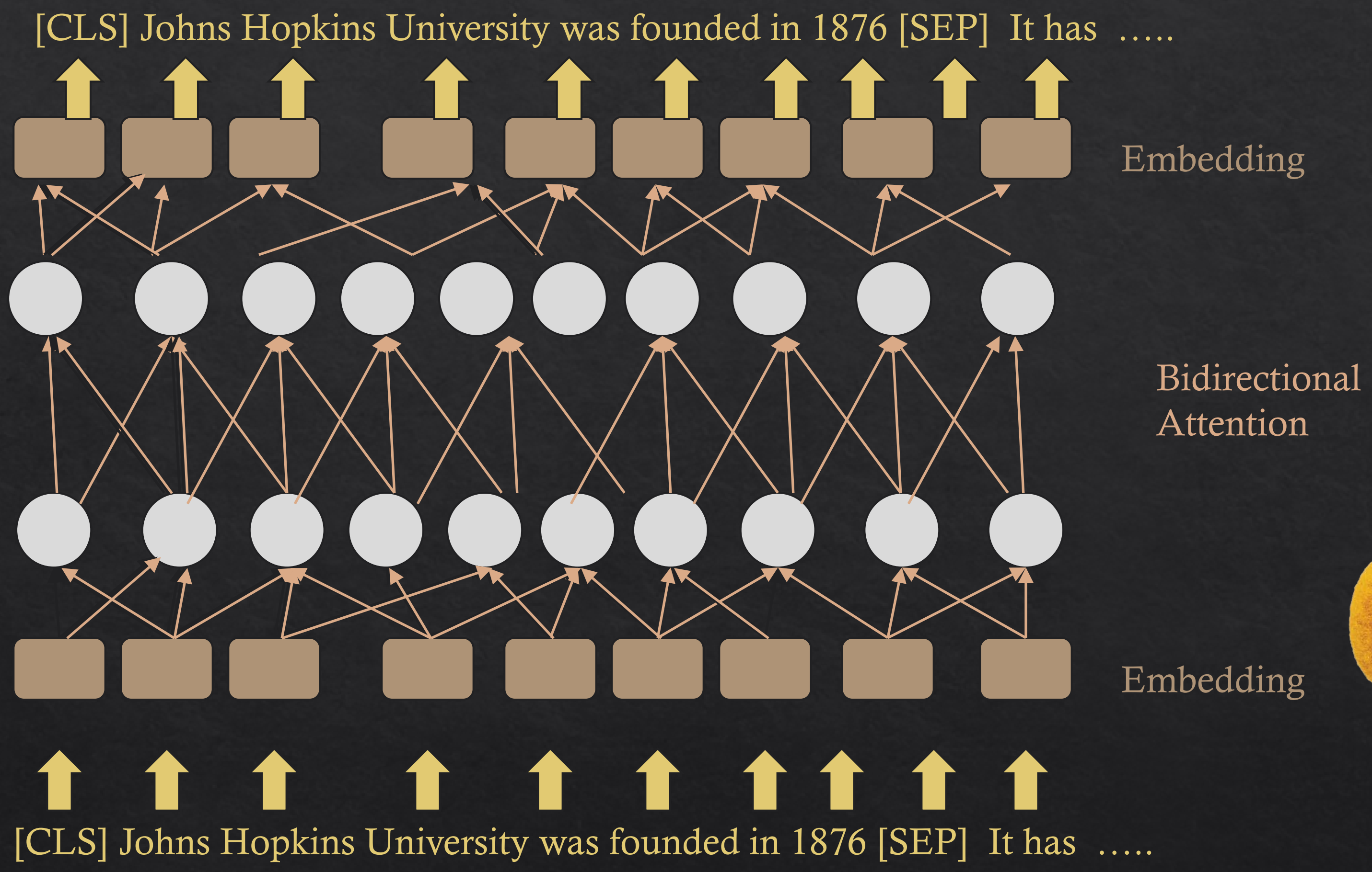
Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" 2018

BERT



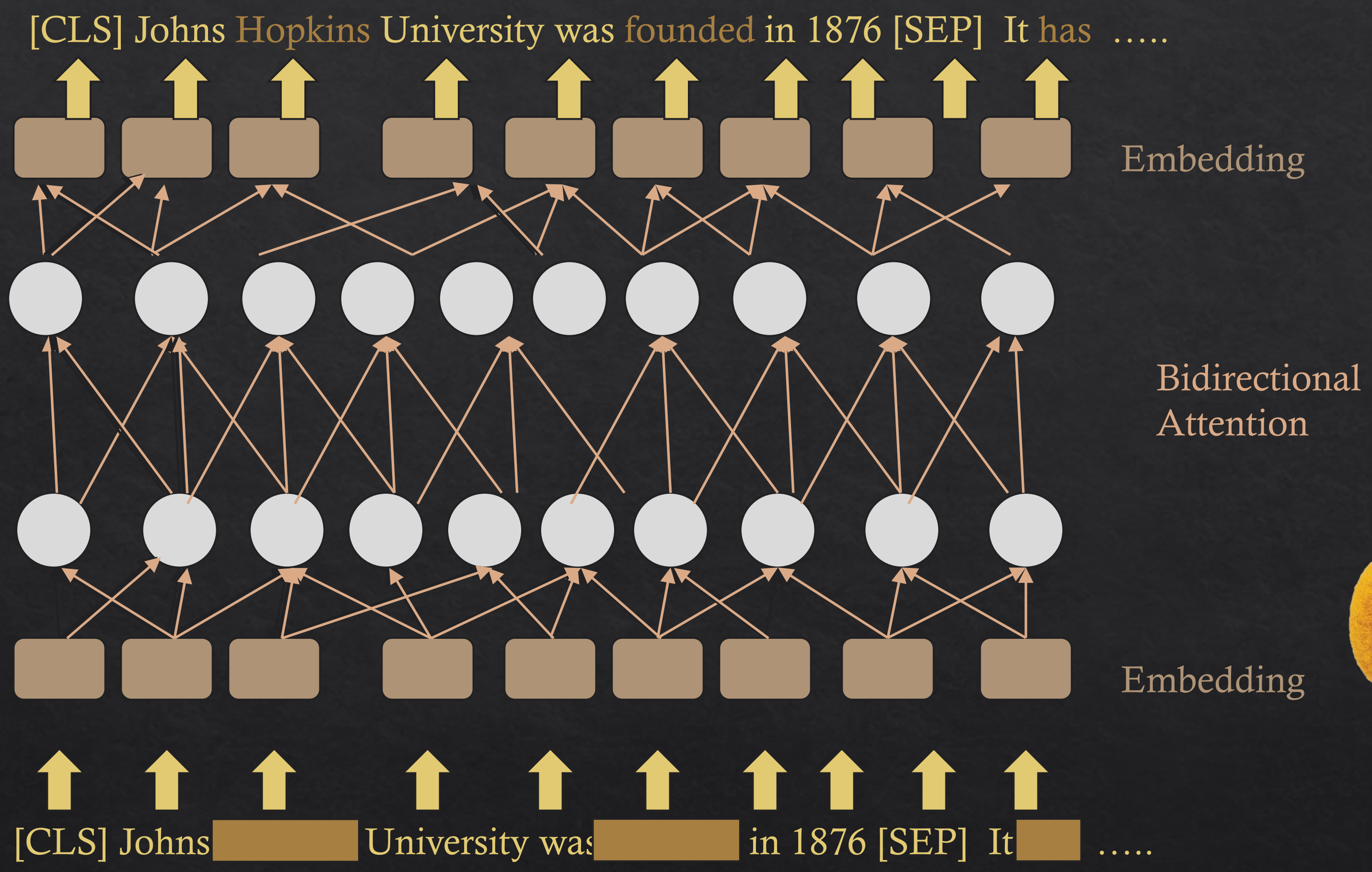
Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" 2018

BERT



Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" 2018

BERT



Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” 2018

BERT

- Masked Language Model
- Next Sentence Prediction

mBERT
104 langs

RoBERTa

- Robustly Optimized BERT Pretraining Approach
- BPE
- No Next Sentence Prediction
- Focus on Hyperparameters

Liu et al. 2019

XLM-R

- 100 Languages
- RoBERTa not BERT
- Not translation (unlike XLM)

Conneau et al. 2019

Curse of Multilinguality

- AFAIK, first mentioned in XLM-R Paper
- More languages hurt performance
- Beneficial for Low-Resource over High

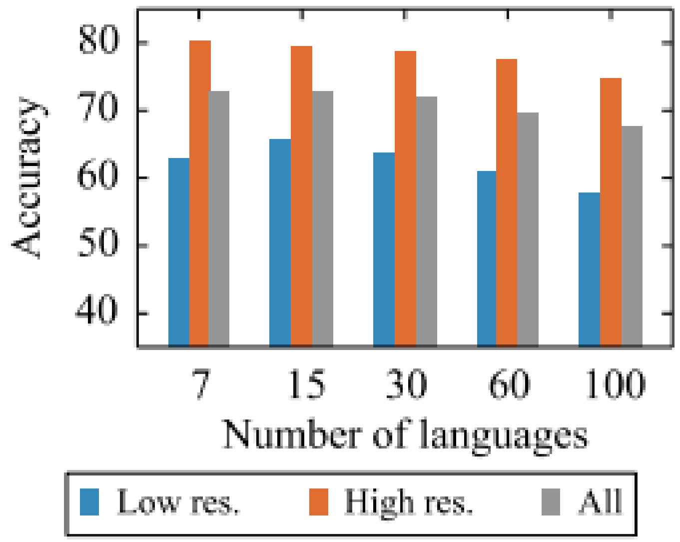


Figure 2: The transfer-interference trade-off: Low-resource languages benefit from scaling to more languages, until dilution (interference) kicks in and degrades overall performance.

Conneau et al. 2019

BiBERTs

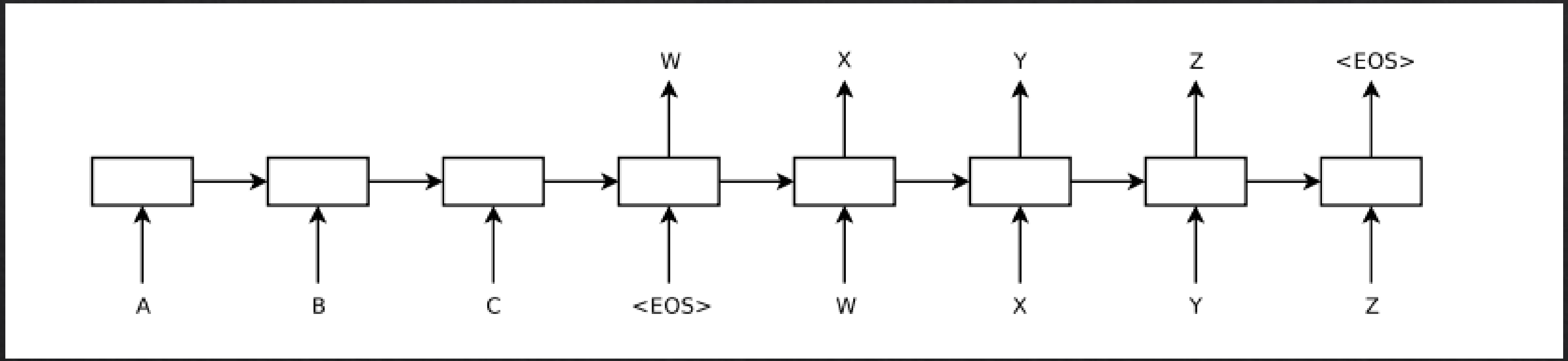
- 2 Languages
- Lan et al., 2020
- "An Empirical Study of Pre-trained Transformers for Arabic Information Extraction"
- Increased Performance on Cross-Lingual (not multilingual) tasks

Encoder	BLEU
Public	12.7
None	14.9
mBERT	15.7
GBv4	15.7
XLM-R	16.0
L64K	**16.2**
L128K	15.8

Table 2: BLEU scores of MT systems with different pre-trained encoders on English–Arabic IWSLT'17.

Yarmohammadi et al. 2021

Brief Detour…



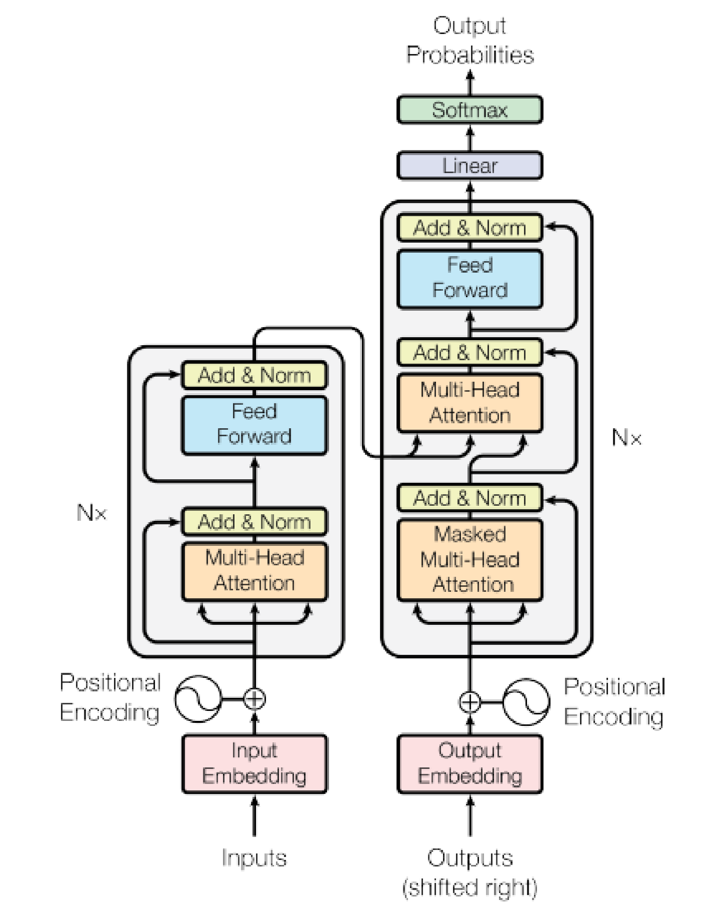


Figure 1: The Transformer - model architecture.

Sutskever et al. 2014
Vaswani et al. 2017

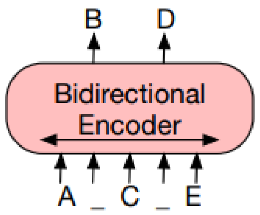
GPT-3

- Generative Pretrained Transformer
- 2048 Context Length
- 175 Billion Parameters
- DECODER

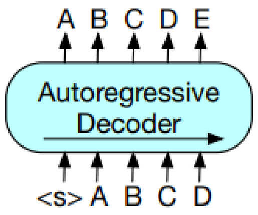
Brown et al. 2020

BART

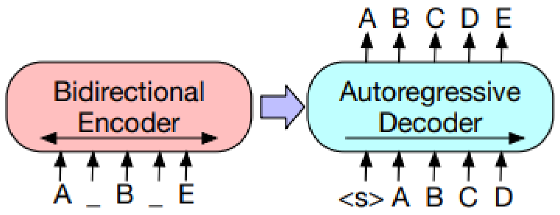
- Denoiser
- Encoder-Decoder
- Lewis et al. 2020



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure 1: A schematic comparison of BART with BERT (Devlin et al., 2019) and GPT (Radford et al., 2018).

BART

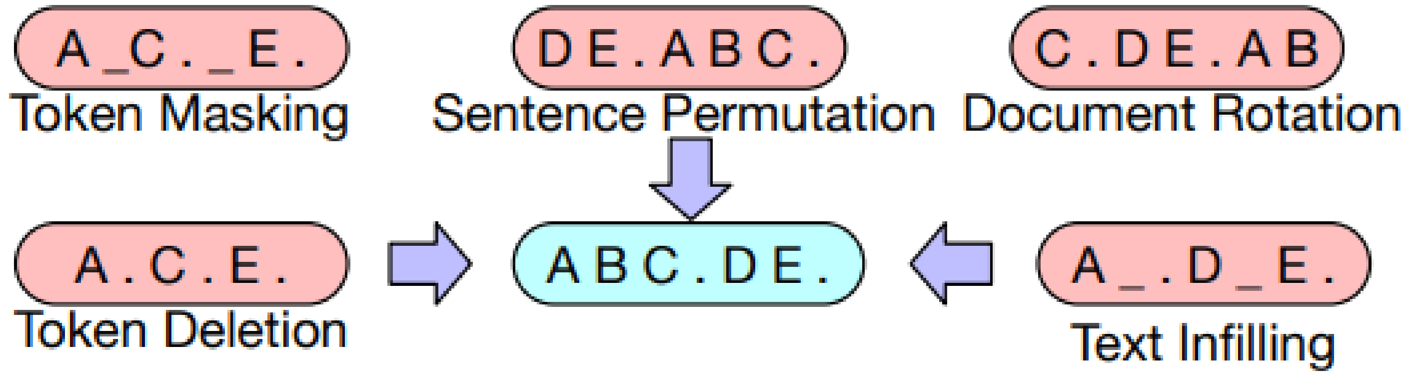


Figure 2: Transformations for noising the input that we experiment with. These transformations can be composed.

mBART

- Multilingual BART
- Liu et al. 2020
- 25, 50, 06 Languages?

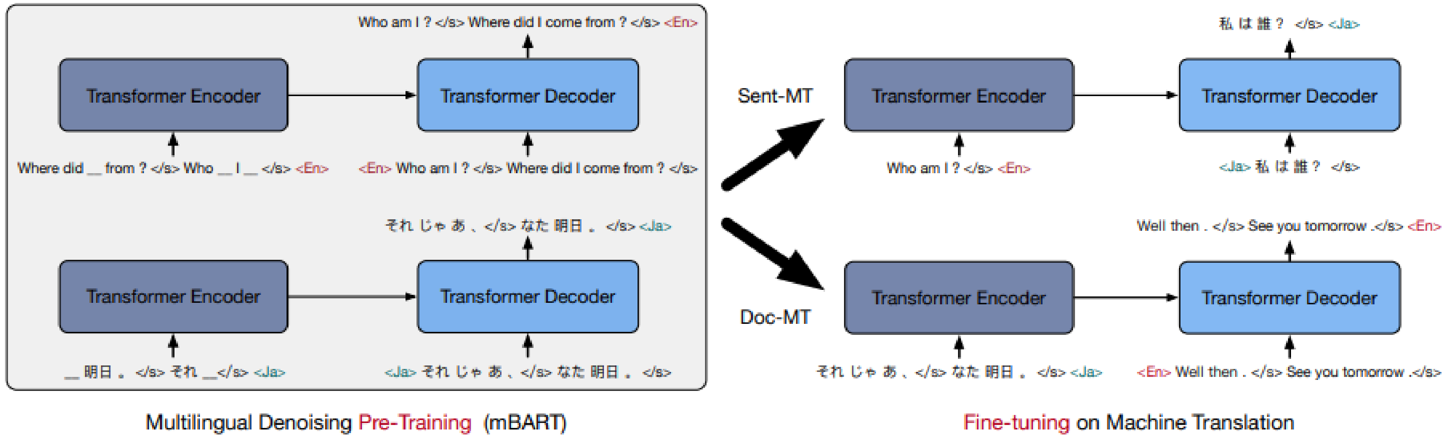
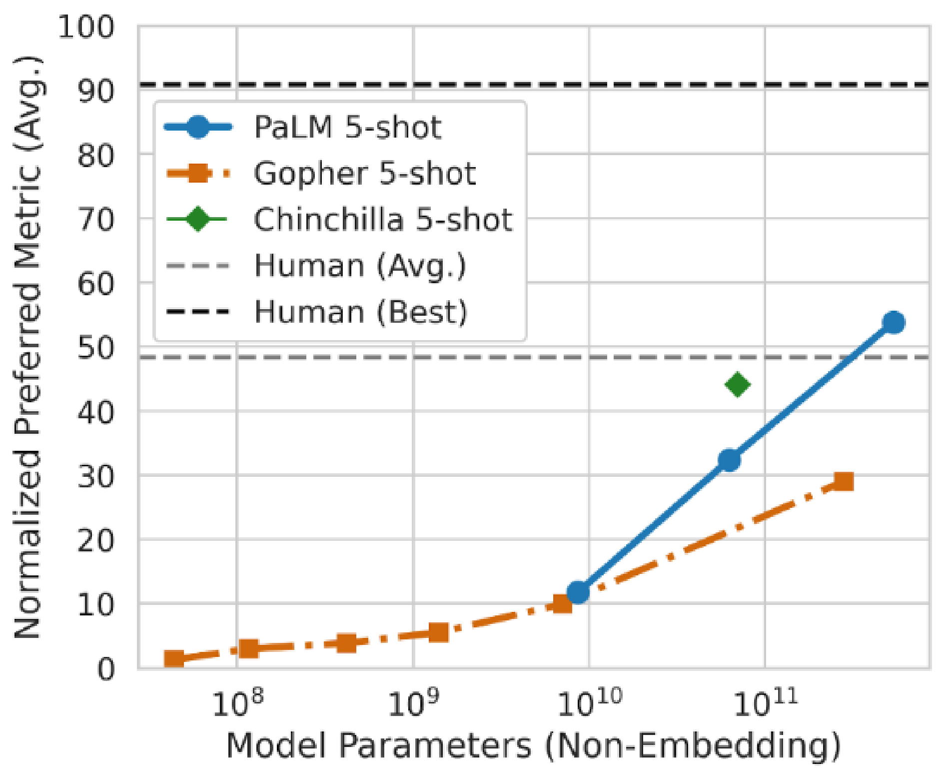


Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

Many More….

- Gopher 280 Billion
- Chinchilla 70 Billion
- LaMDA 137 Billion
- PaLM 540 Billion



Scaling behavior of PaLM on a subset of 58 BIG-bench tasks.

https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

T5

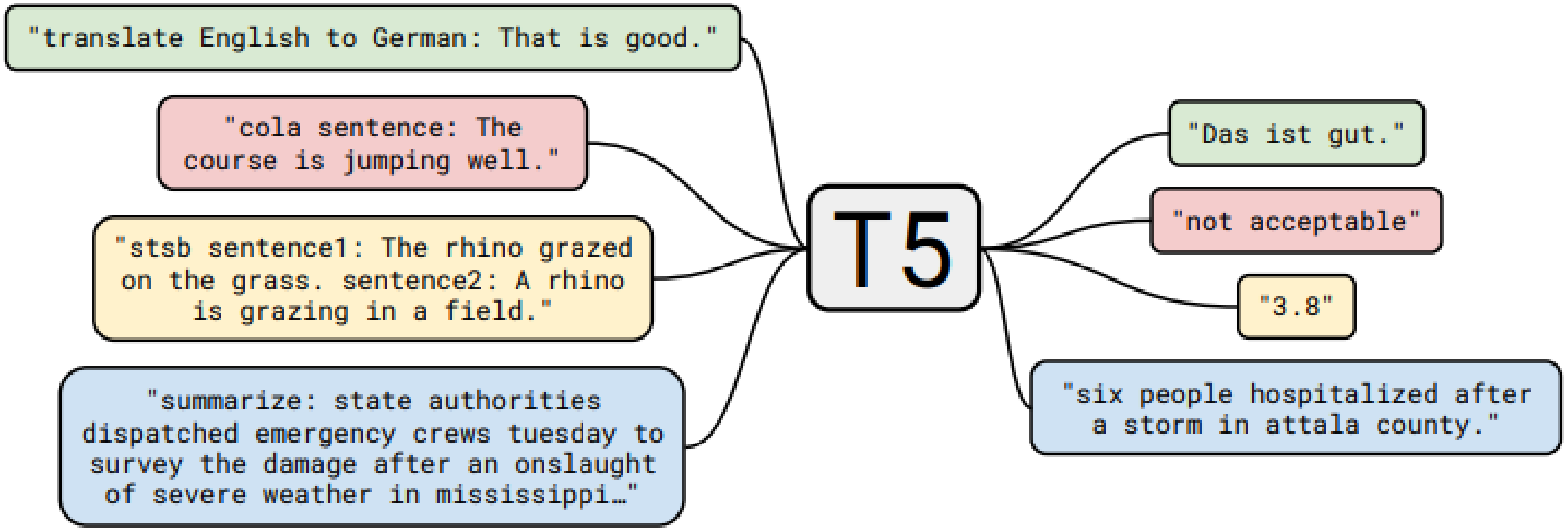


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

. 2020

mT5

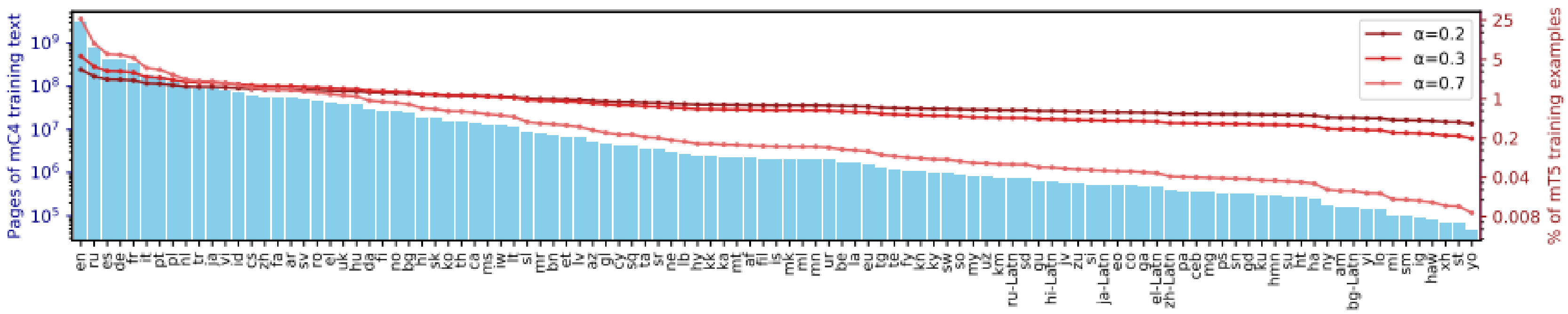


Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

Table 1: Comparison of mT5 to existing massively multilingual pre-trained language models. Multiple versions of XLM and mBERT exist; we refer here to the ones that cover the most languages. Note that XLM-R counts five Romanized variants as separate languages, while we ignore six Romanized variants in the mT5 language count.

et al. 2020

ERNIE-M

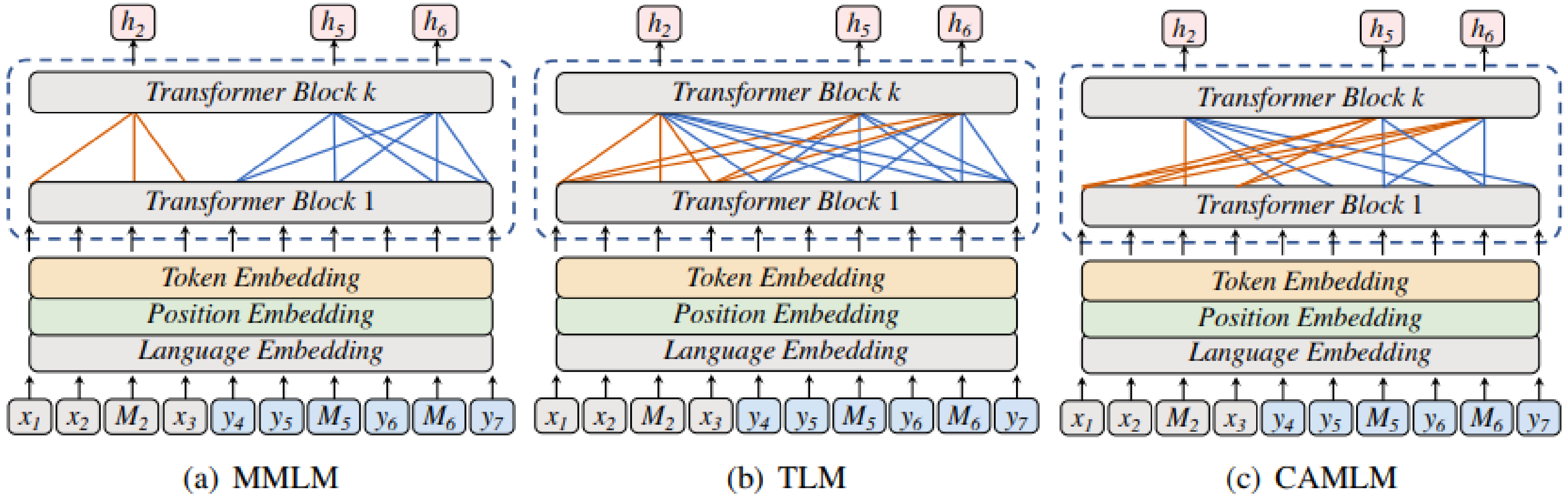


Figure 1: Overview of MMLM, TLM and CAMLM training. The input sentences in sub-figure (a) are monolingual sentences; x and y represent monolingual input sentences in different languages. The input sentences in sub-figures (b) and (c) are parallel sentences; x and y denote the source and target sentences of the parallel sentences, respectively. h indicates the token predicted by the model.

Baidu

Key Insight: Back-Translation

96 Languages

Ouyang et al. 2020

ERNIE-M

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	ERNIE-M	ERNIE Team	Baidu	Jan 1, 2021	80.9	87.9	75.6	72.3	91.9
2	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
3	Anonymous3	Anonymous3	Anonymous3	Jan 3, 2021	79.9	88.2	74.6	71.7	89.0
4	Polyglot	MLNLC	ByteDance	Nov 13, 2020	77.8	87.8	72.9	67.4	88.3
5	VECO	DAMO NLP Team	Alibaba	Sep 29, 2020	77.2	87.0	70.4	68.0	88.1
6	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
7	X-STILTs	Phang et al.	New York University	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
8	XLM-R (large)	XTREME Team	Alphabet, CMU	-	68.2	82.8	69.0	62.3	61.6
9	mBERT	XTREME Team	Alphabet, CMU	-	59.6	73.7	66.3	53.8	47.7

Multilingual Language Model Gains Research Attention

XTREME dataset (Hu et al. 2020)

http://research.baidu.com/Blog/index-view?id=151

SpanBERT

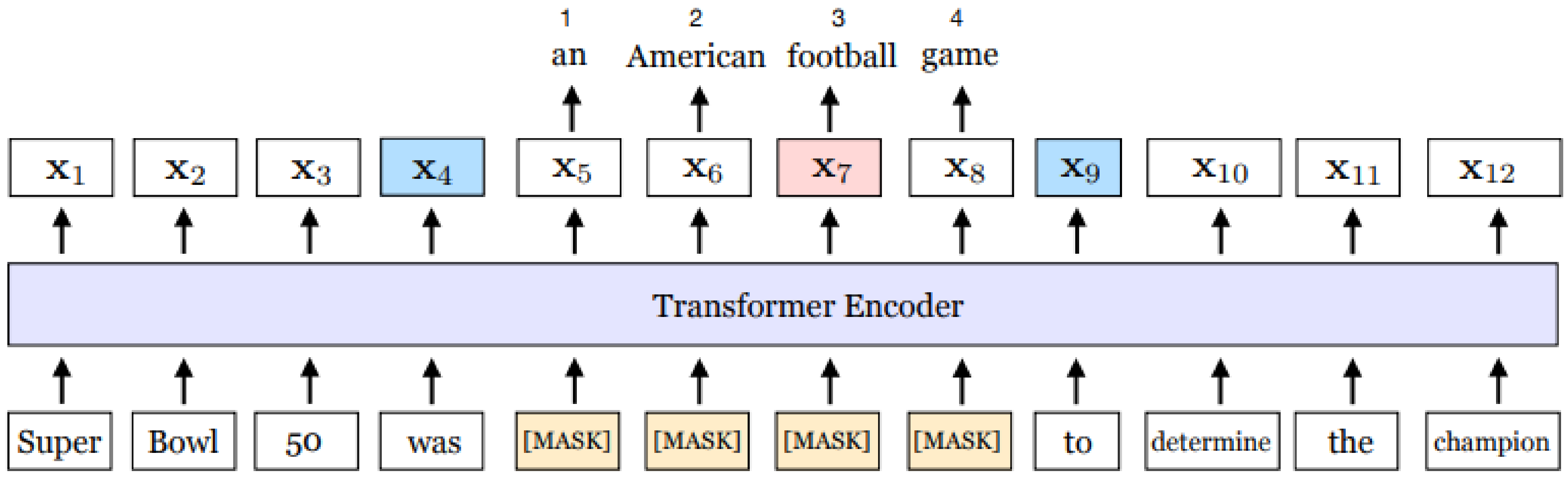


Figure 1: An illustration of SpanBERT training. The span *an American football game* is masked. The span boundary objective (SBO) uses the output representations of the boundary tokens, $\mathbf{x}_4$ and $\mathbf{x}_9$ (in blue), to predict each token in the masked span. The equation shows the MLM and SBO loss terms for predicting the token, *football* (in pink), which as marked by the position embedding $\mathbf{p}_3$, is the *third* token from x_4.

Joshi et al. 2019

SLAM

- Speech and LAnguage Modeling
- Bapna et al. 2021

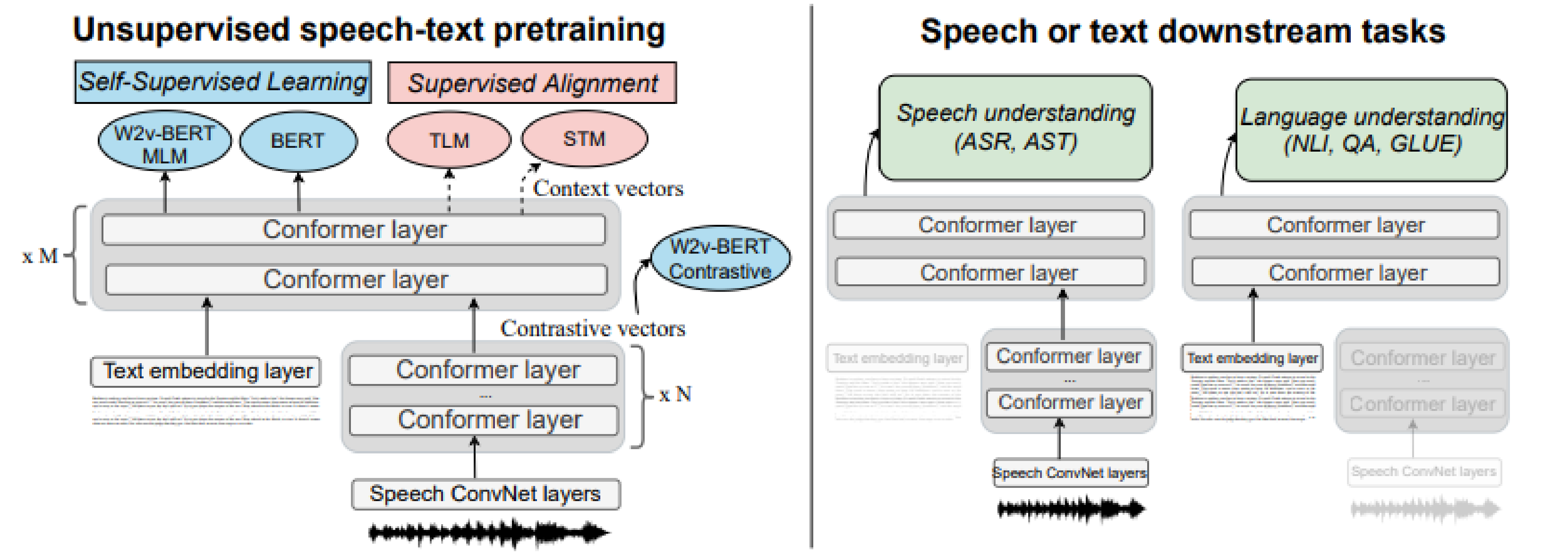


Figure 1: (Left) Our model consists of a text embedding layer and a speech-specific stack similar to w2v-BERT, the latter consisting of a ConvNet and a series of N Conformer layers. Both the text and speech output embeddings are fed to a series of N shared Conformer layers. Our unsupervised speech-text pre-training approach consists of self-supervised learning objectives (in blue), including w2v-BERT masked language modeling and contrastive losses, as well as the text BERT objective. This can be combined with supervised alignment losses (in red) which leverage speech-text annotated pairs. We leverage in particular the MLM variant of translation language modeling (TLM) and the ranking loss of speech-text matching (STM). **(Right)** Once pre-trained, the speech part of the shared architecture can be fine-tuned on speech understanding datasets like recognition or translation. The text part of the architecture can be fine-tuned on language understanding tasks.

mSLAM

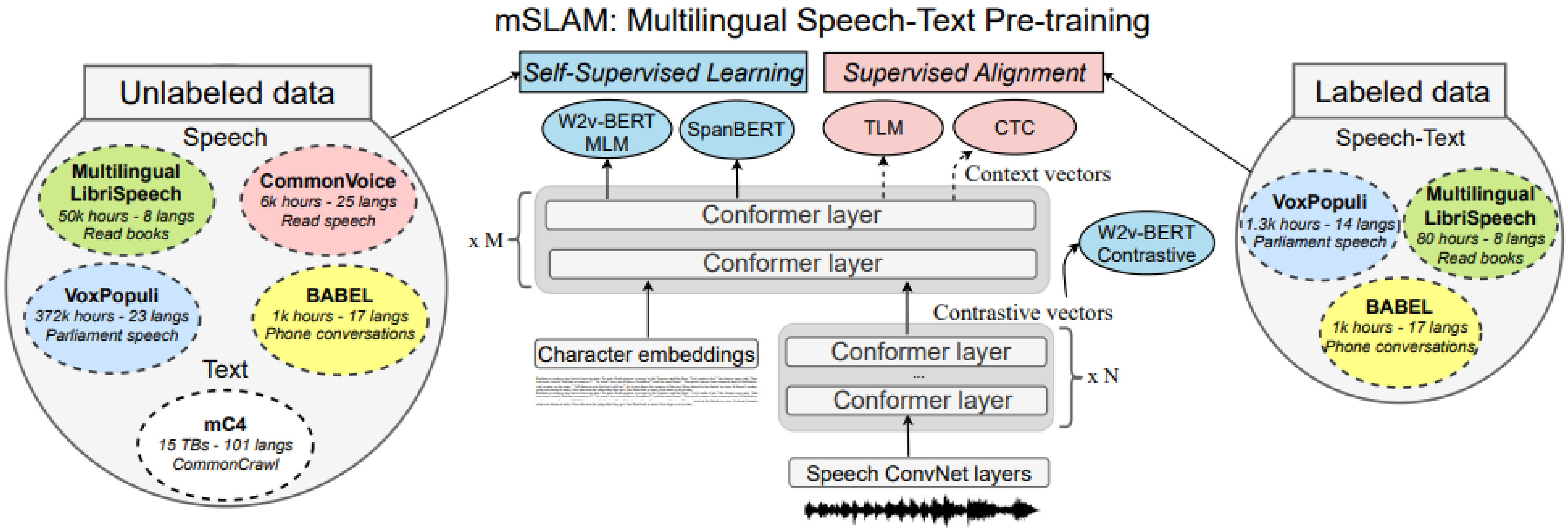


Figure 1: Multilingual Speech-Text Pretraining We pre-train a large multilingual speech-text Conformer on 429K hours of unannotated speech data in 51 languages, 15TBs of unannotated text data in 101 languages, as well as 2.3k hours of speech-text ASR data.

Bapna et al. 2021