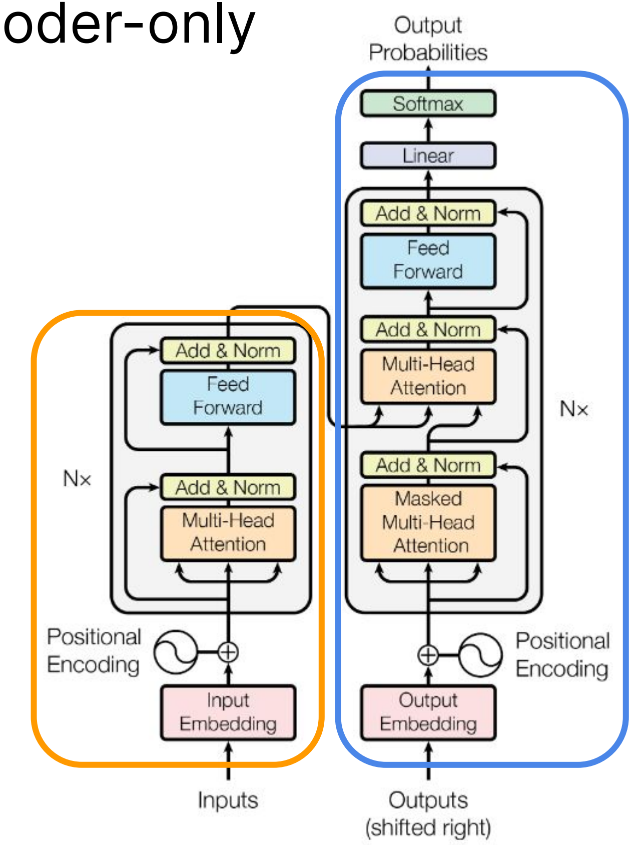
Encoder-only vs Decoder-only

Encoder



Decoder

Vaswani et al. (2017)

Encoder self-attention

	Never	gonna	give	you	up	,	never	gonna
Никогда	0,9	0,1						
тебя				1				
не	1							
подведу			0,5		0,5			
,						1		
никогда							0,9	0,1
<pad>								
<pad>								

Decoder **masked** self-attention

	Never	gonna	give	you	up	,	never	gonna
Никогда	0,9	0	0	0	0	0	0	0
тебя			0	0	0	0	0	0
не	1			0	0	0	0	0
подведу			0,5		0	0	0	0
,						0	0	0
никогда							0	0
<pad>								0
<pad>								

AIRI × Университет Сириус

Transfer Learning

Aydar Bulatov
bulatov@deeppavlov.ai

Vasily Konovalov
lecture materials

Feel free to open this lecture on your laptop

Telegram

Github of NLP Course

Feedback

Today:

01 Transfer Learning recap

02 Finetuning vs few-shot learning

03 Practice

When and how to transfer knowledge?

When do we need TL?

- Not enough data for target task

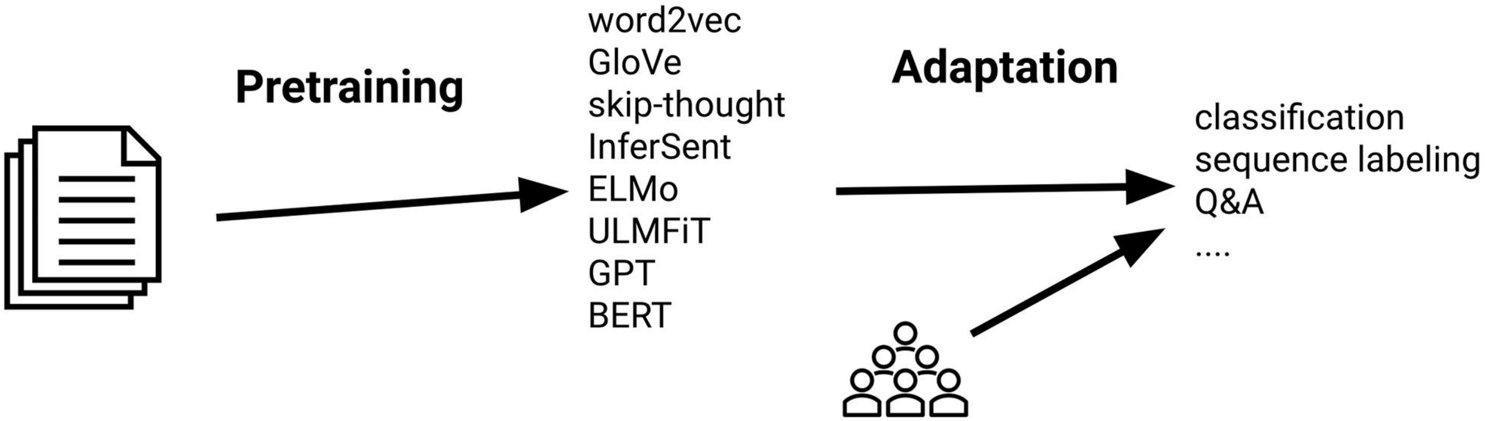
  like every real-world task ever

- Need a larger LM

  score ~ number of parameters, dataset size

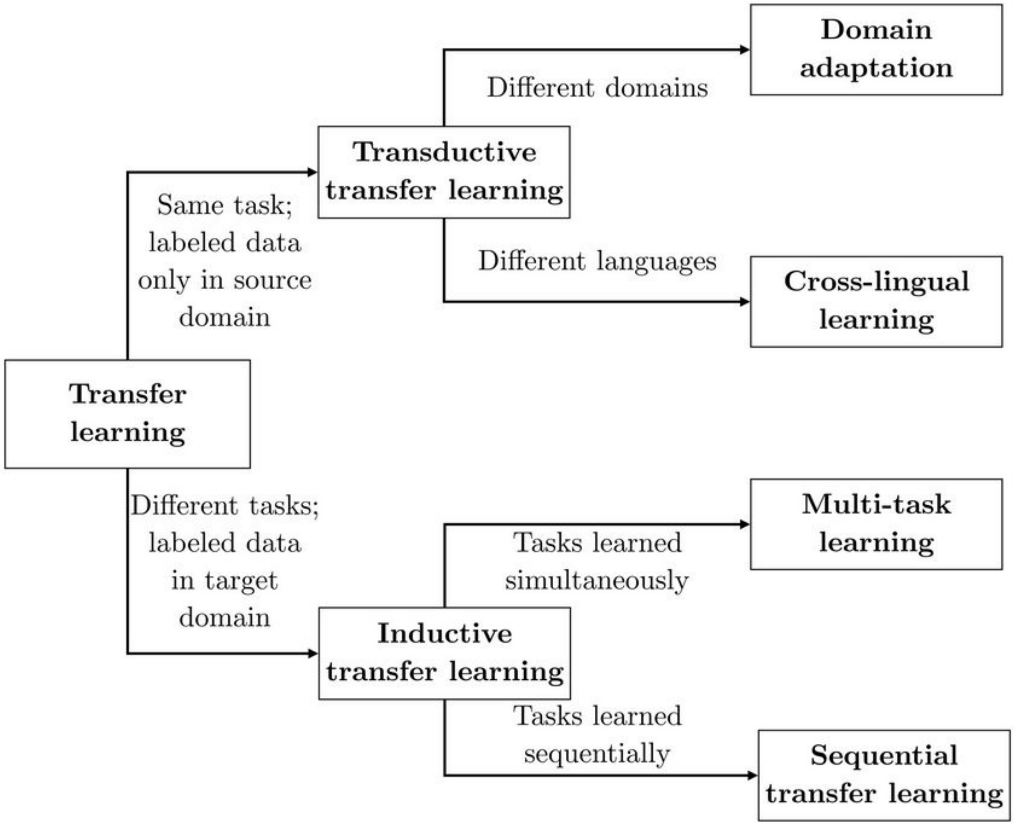
- Multiple training domains
- Multiple tasks at once

Transfer learning approach



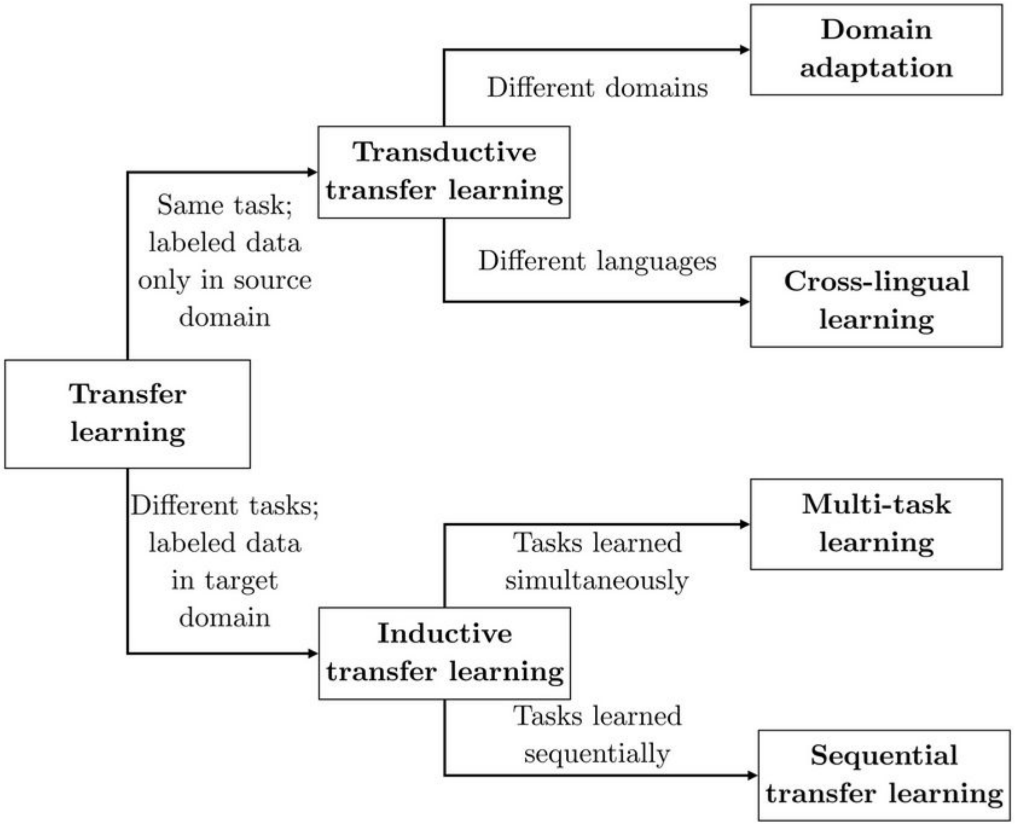
Ruder, 2019

Transfer learning taxonomy


Transfer learning
Same task; labeled data only in source domain
Transductive transfer learning
Different domains
Domain adaptation
Different languages
Cross-lingual learning
Different tasks; labeled data in target domain
Inductive transfer learning
Tasks learned simultaneously
Multi-task learning
Tasks learned sequentially
Sequential transfer learning

Ruder, 2019

Transfer learning taxonomy


Transfer learning
Same task; labeled data only in source domain
Transductive transfer learning
Different domains
Domain adaptation
Different languages
Cross-lingual learning
Different tasks; labeled data in target domain
Inductive transfer learning
Tasks learned simultaneously
Multi-task learning
Tasks learned sequentially
Sequential transfer learning

→ meta learning
→ lifelong learning

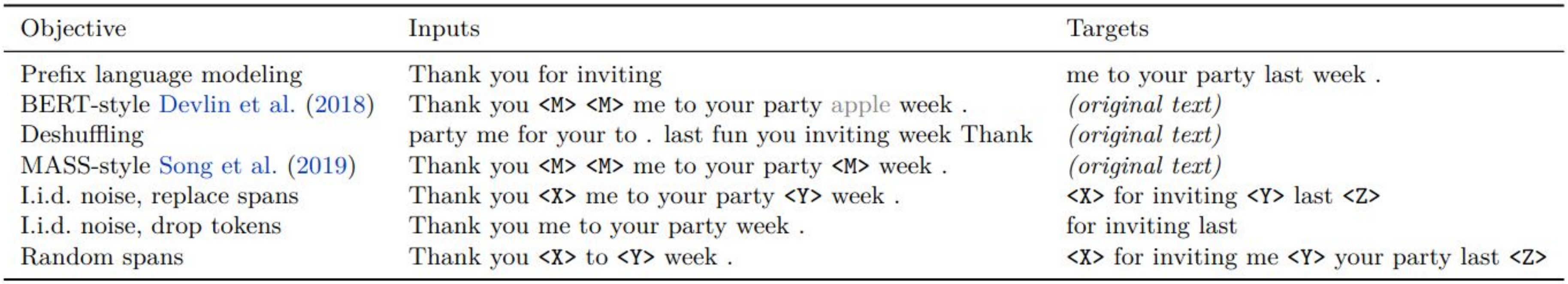
Ruder, 2019

Sequential transfer learning

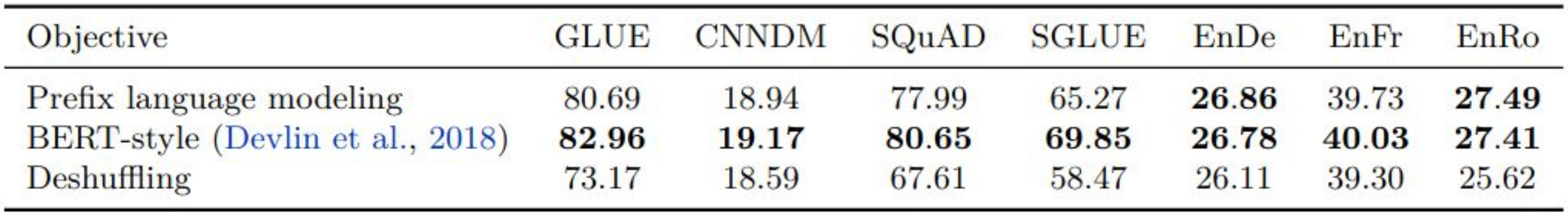
1) Pretraining
    a) no supervision
       LM, MLM, NSP, span corruption, deshuffling
    b) distant supervision
       relation extraction, sentiment analysis
    c) traditional supervision
       high-resource more general domain, paraphrase, nli, translation, ...

2) Fine-tuning
   your favourite downstream task

Pretraining objectives



Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>



Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Prefix language modeling	80.69	18.94	77.99	65.27	**26.86**	39.73	**27.49**
BERT-style (Devlin et al., 2018)	**82.96**	**19.17**	**80.65**	**69.85**	**26.78**	**40.03**	**27.41**
Deshuffling	73.17	18.59	67.61	58.47	26.11	39.30	25.62

Sequential transfer learning

Idea 1: from context-independent to contextualized representations

Word2Vec, GloVe → CoVe, ELMo

Idea 2: refuse from task-specific models

GPT: pretrain with LM loss → finetune with LM loss, task loss

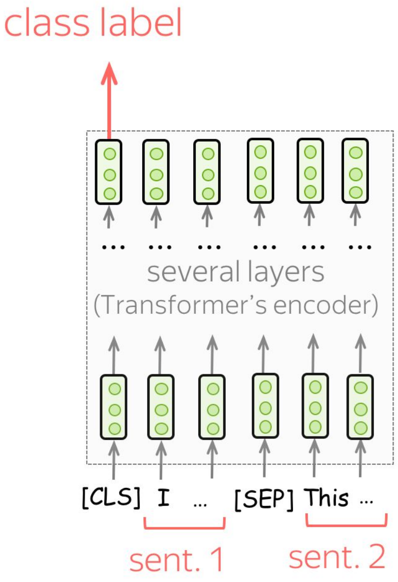
BERT: pretrain with MLM and NSP loss →

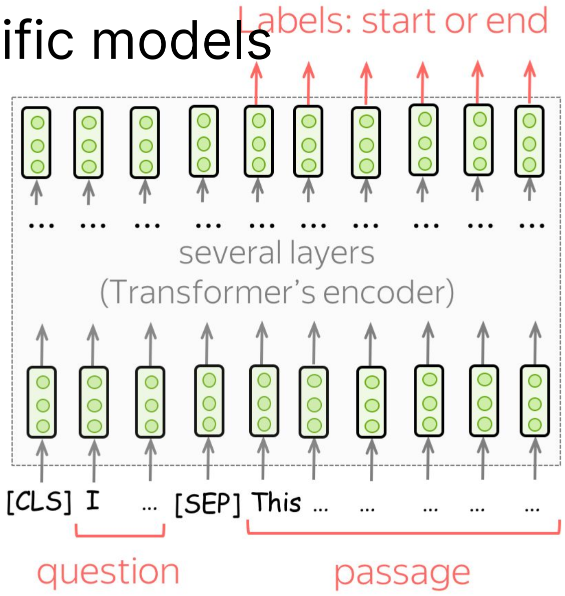
finetune with task loss from CLS token representations / passages

Sequential transfer learning

<u>Idea 1:</u> from context-independent to contextualized representations

<u>Idea</u> ask-specific models



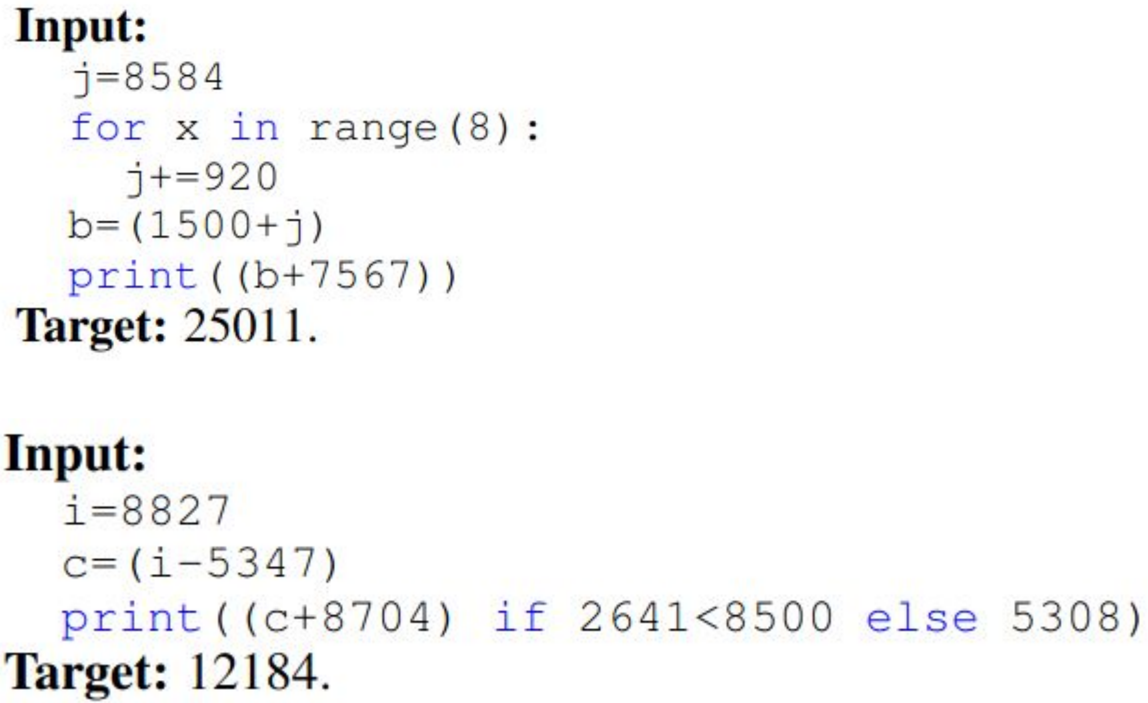


Curriculum learning

Task:
generate output given
symbol-level python code

parameters:
a = program length
b = program nesting
c = number of digits

model: LSTM



Input:

```
j=8584
for x in range(8):
  j+=920
b=(1500+j)
print((b+7567))
```

Target: 25011.

Input:

```
i=8827
c=(i-5347)
print((c+8704) if 2641<8500 else 5308)
```

Target: 12184.

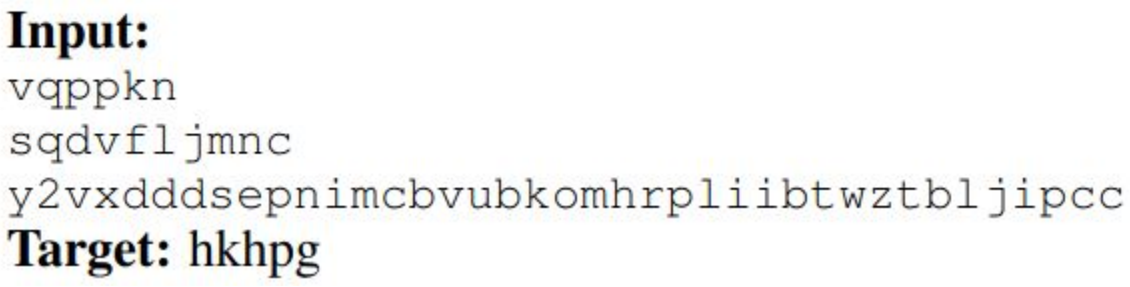
Learning to execute
Zaremba and Sutskever, 2015

Curriculum learning

Task:
generate output given
symbol-level python code

parameters:
a = program length
b = program nesting
c = number of digits

model: LSTM



Input:

```
vqppkn
sqdvfljmnc
y2vxdddsepnimcbvubkomhrpliibtwztbljipcc
```

Target: hkhpg

Learning to execute
Zaremba and Sutskever, 2015

Curriculum learning

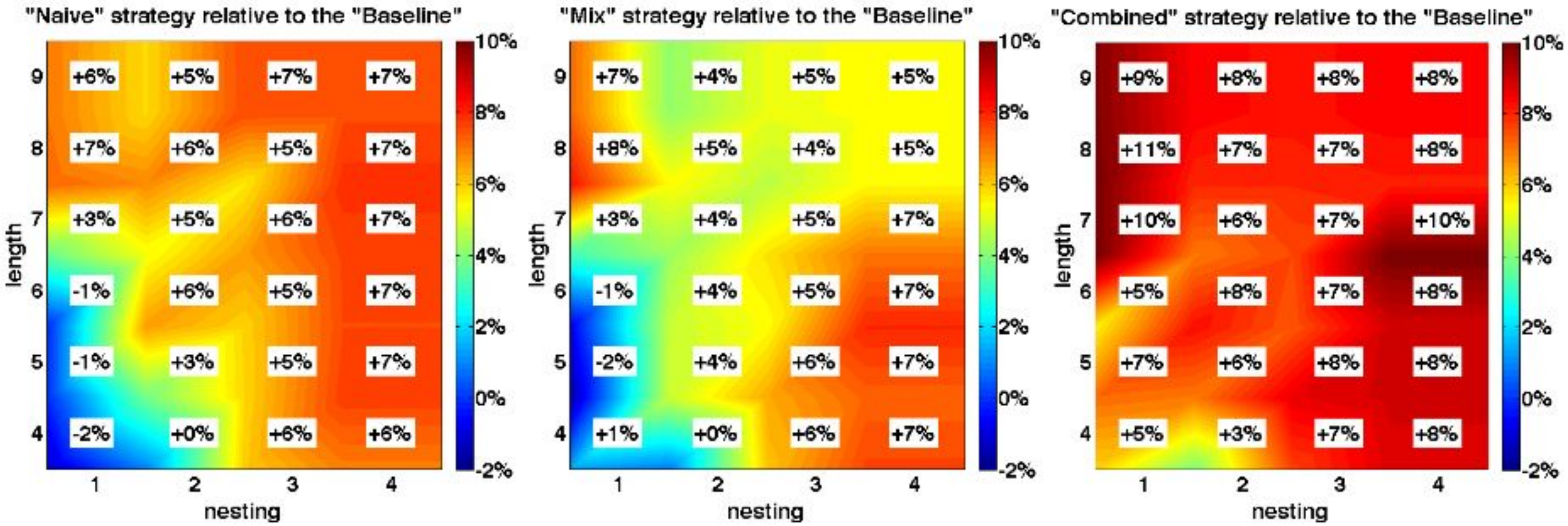
Approach 1: baseline, no curriculum
all training samples have length=a, nesting=b

Approach 2: naïve curriculum strategy
begin with length = 1, nesting = 1,
once on plateau, increase length by 1,
once length reaches a, reset to 1 and increase nesting by 1,
once length = a, nesting = b, stop.

Approach 3: mixed strategy
pick a random length from 1 to a, nesting from 1 to b,
generate training sample,
train, repeat.

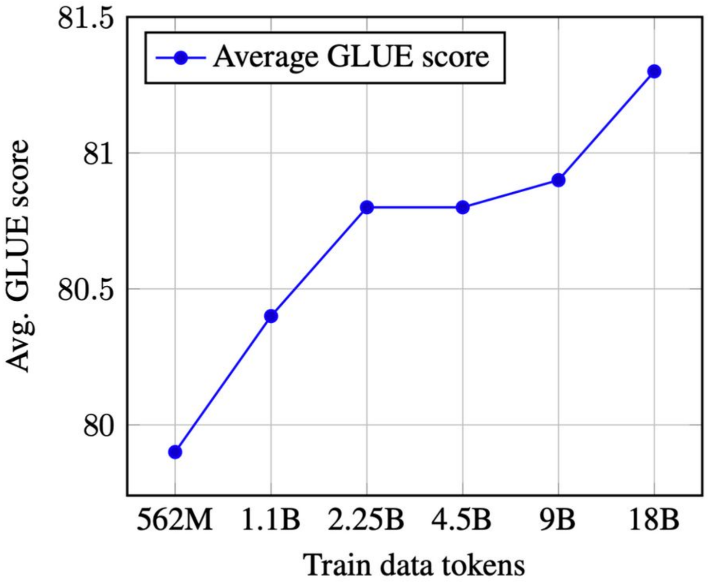
Approach 4: combined = mixed + naïve
samples from mixed and naive one after another

Curriculum learning



Relative prediction accuracy of the different strategies with respect to the baseline strategy.

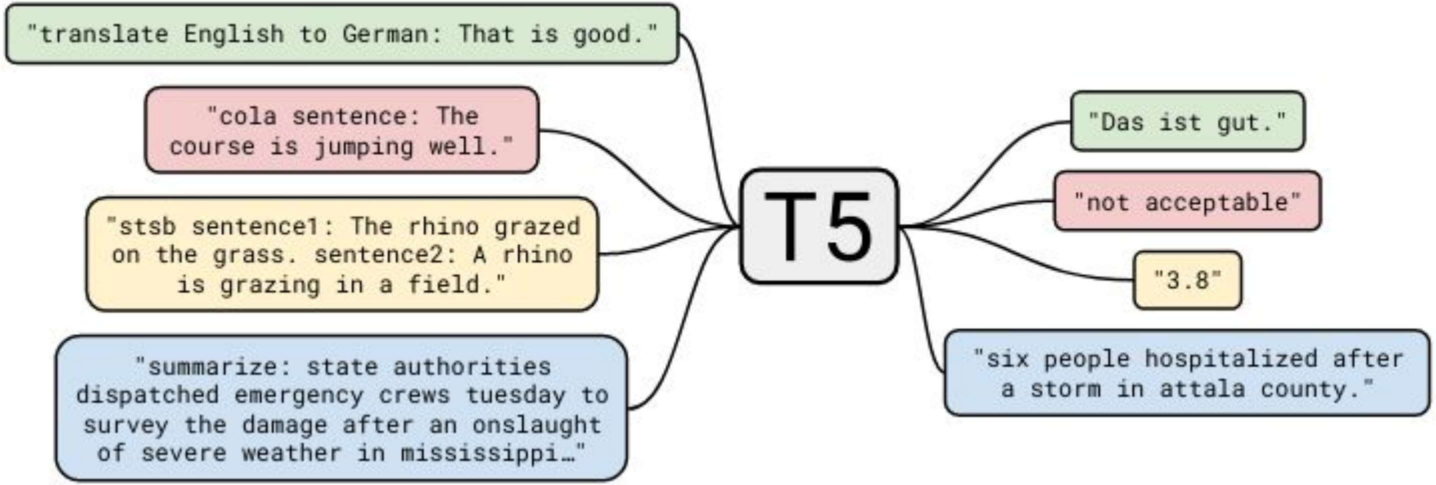
Multi-task learning



Sequential learning ignores knowledge from related tasks.
Want to improve generalization by using other tasks from domain.
Idea: use multi-task learning for language tasks

What model does that?

Multi-task learning



"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…"

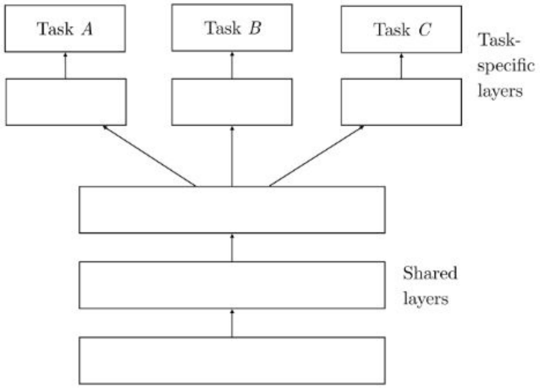
"Das ist gut."

"not acceptable"

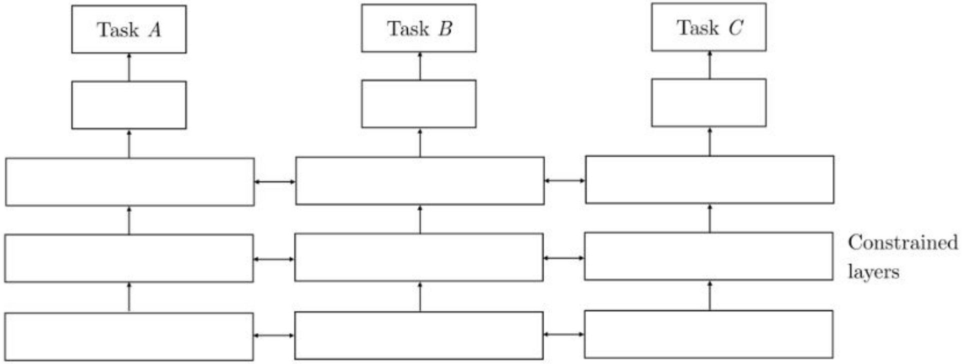
"3.8"

"six people hospitalized after a storm in attala county."

Multi-task learning



Hard parameter sharing



Soft parameter sharing
e.g. l2 distance, KL divergence

Why does it work?

Implicit data augmentation: MTL effectively increases the sample size that we are using for training our model

Attention focusing: If a task is very noisy or data is limited and high-dimensional, it can be difficult for a model to differentiate between relevant and irrelevant feature

Eavesdropping: Some features G are easy to learn for some task B, while being difficult to learn for another task A

Representation: bias MTL biases the model to prefer representations that other tasks also prefer.

Regularization: Finally, MTL acts as a regularizer by introducing an inductive bias.

Ruder, 2019

Cross-lingual learning

Idea: use multi-language datasets to create universal embeddings

word-level alignment

- mapping word representations
- pseudo-multi-lingual corpora
- joint methods

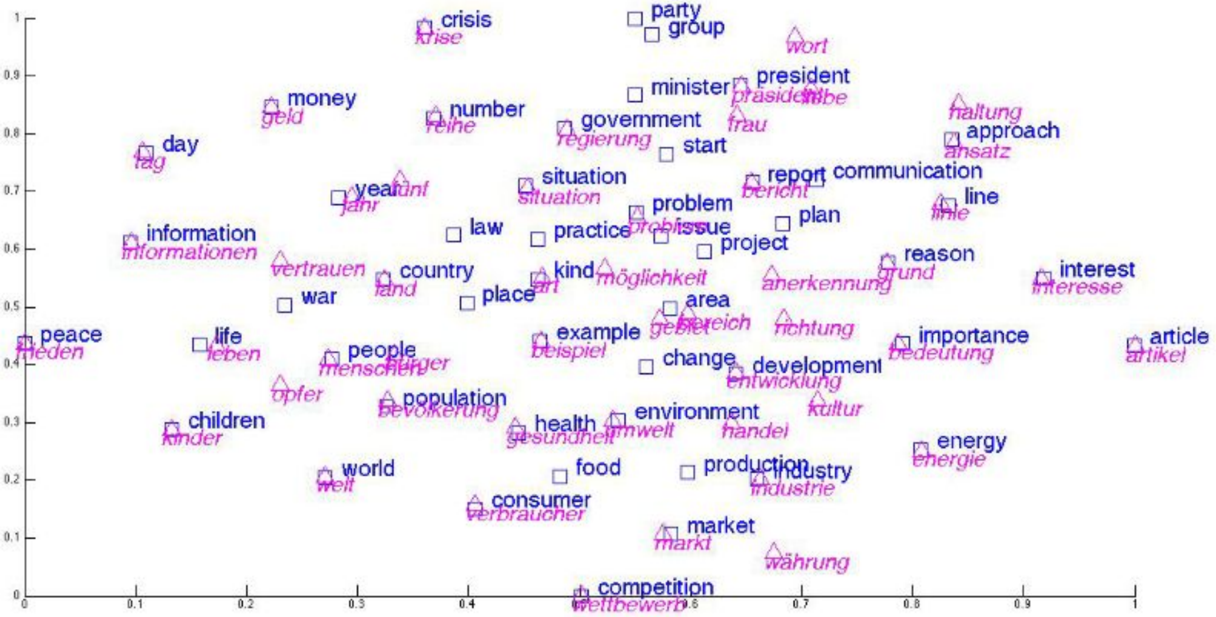
sentence-level alignment

- word-alignment based matrix factorization
- compositional sentence methods
- bilingual autoencoder
- bilingual skip-gram

document-level alignment

- pseudo-bilingual document aligned corpora
- concept-based methods
- extensions of sentence-aligned methods

Cross-lingual learning



Luong et al., 2015

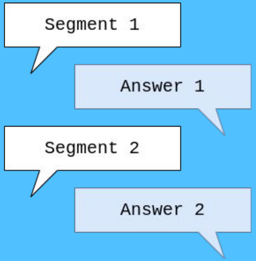
Domain adaptation

Problem:
In real world training and test data are not i.i.d.
e.g. different lexis

Ideas:
Ignore features that are not in the target domain
Bring training and test features into the common vector space

Solution
Multilingual models

What if the sequence is just too long?

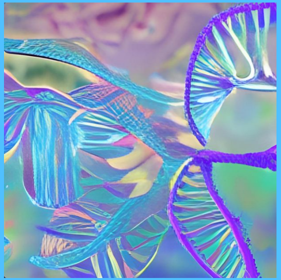

Segment 1

Answer 1

Segment 2

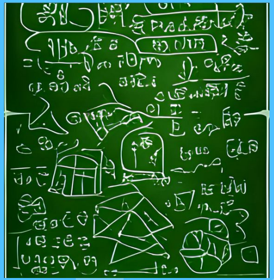
Answer 2

Chat-bots



DNA-related tasks

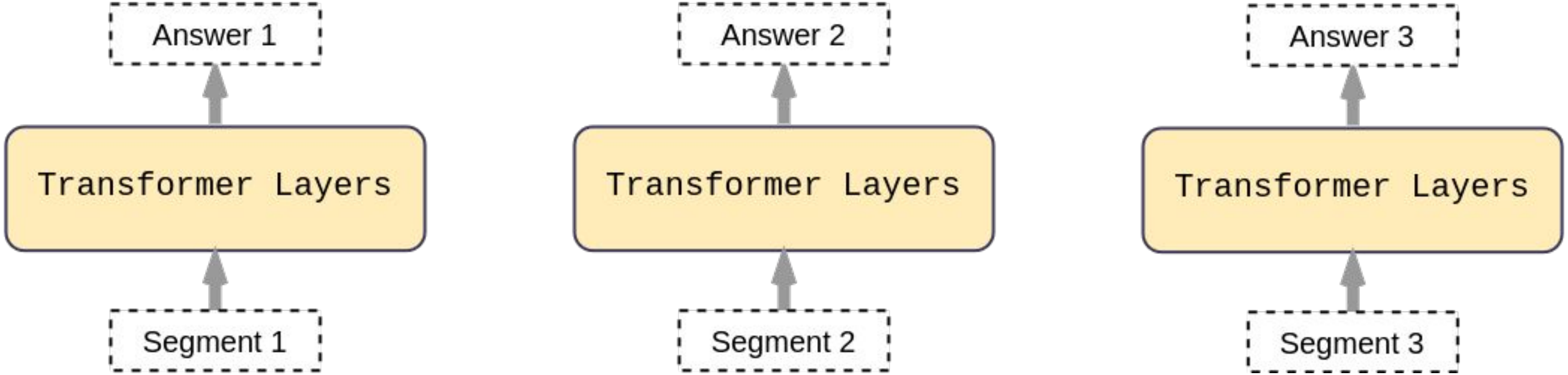
Long texts



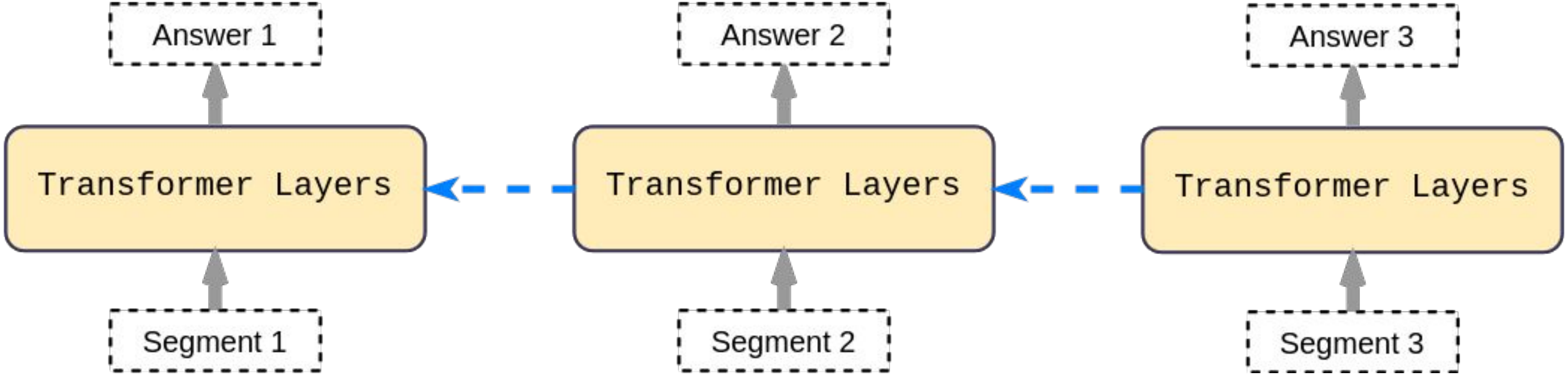
Step-by-step solutions

Data segmentation

How to remember data between segments?

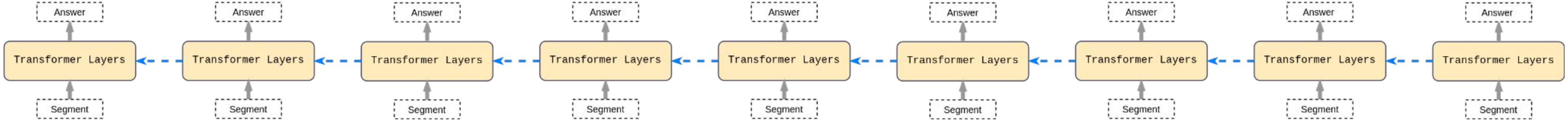


Data segmentation



Dai et al., 2019

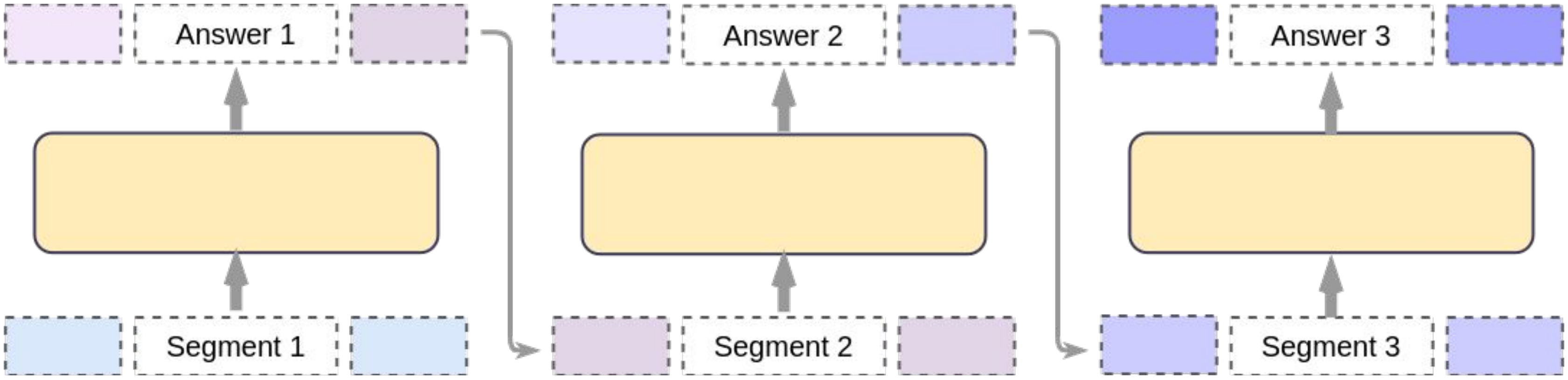
Data segmentation



What if the sequence is *really* long?

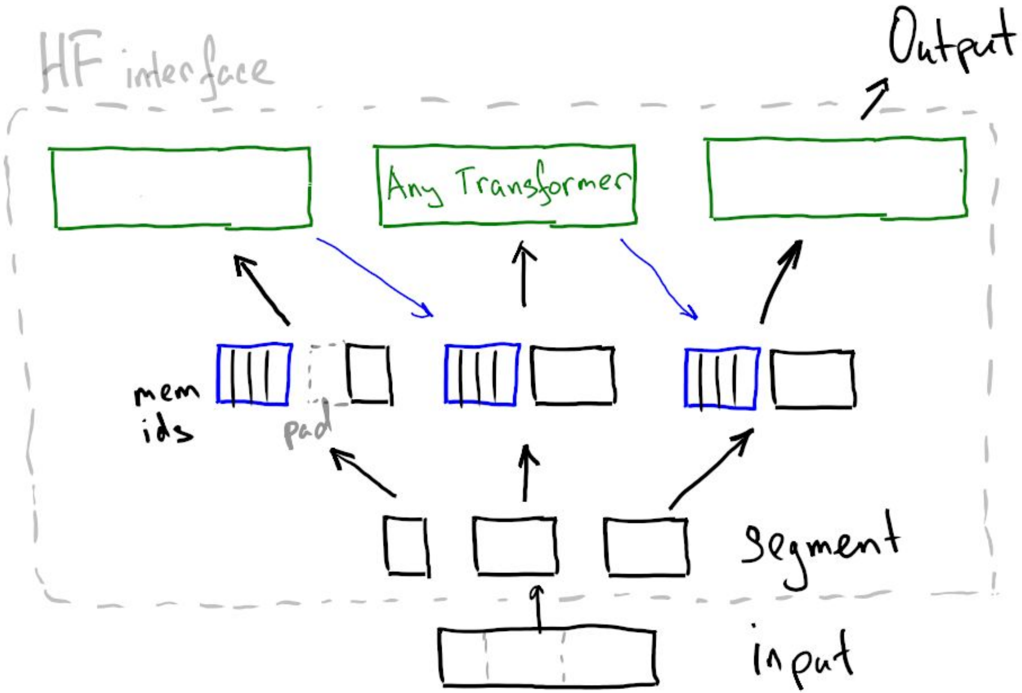
Recurrent Memory Transformer

Add special memory tokens
processed alongside sequence.

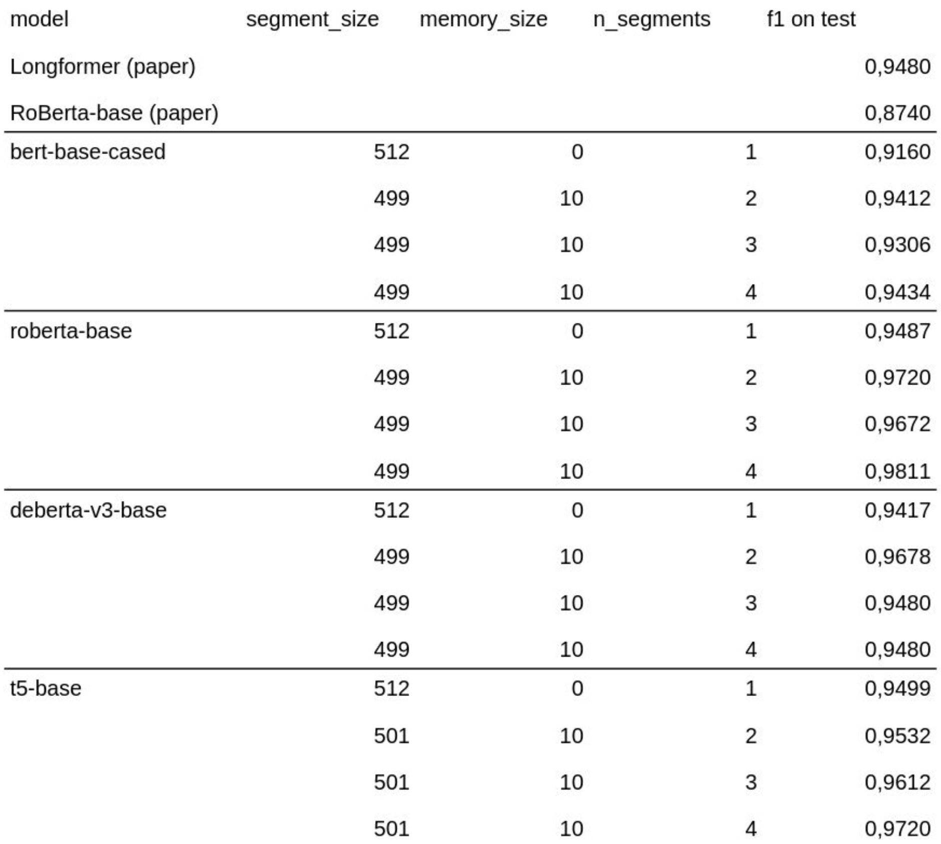


A. Bulatov, Y. Kuratov, M. Burtsev

Recurrent Memory Transformer



Hyperpartisan news detection



model	segment_size	memory_size	n_segments	f1 on test
Longformer (paper)				0,9480
RoBerta-base (paper)				0,8740
bert-base-cased	512	0	1	0,9160
	499	10	2	0,9412
	499	10	3	0,9306
	499	10	4	0,9434
roberta-base	512	0	1	0,9487
	499	10	2	0,9720
	499	10	3	0,9672
	499	10	4	0,9811
deberta-v3-base	512	0	1	0,9417
	499	10	2	0,9678
	499	10	3	0,9480
	499	10	4	0,9480
t5-base	512	0	1	0,9499
	501	10	2	0,9532
	501	10	3	0,9612
	501	10	4	0,9720

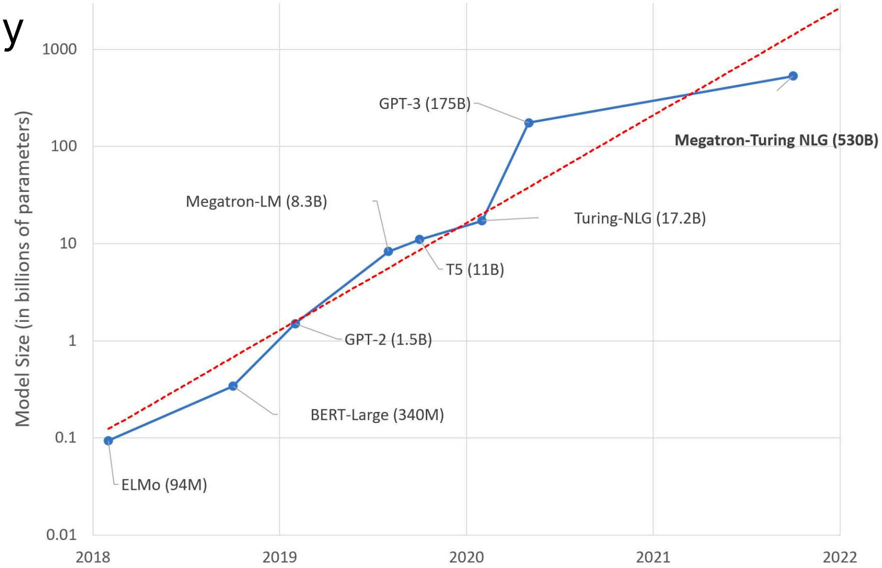
What if there is no data for finetuning?

GPT (Generative Pre-training Transformer)

GPT-1 - Improving Language Understanding by Generative Pre-Training
17M params, 12 layers, 12 heads

GPT-2 - Language Models are Unsupervised Multitask Learners
1.5B params, 48 layers,

GPT-3 - Language Models are **Few-Shot** Learners
175B params, 96 layers, 96 heads



GPT (Generative Pre-training Transformer)

N-shot learning

Suppose we have *very* limited training examples for the target task.

A pre-trained model is a …

zero-shot learner, if it can solve the target task without any examples,
input: “The translation of “cheese” to french is ”

one-shot learner, if it can solve the target task if 1 example is included in input,
input: “The translation of “a bird” to french is “l’oiseau” .
The translation of “cheese” to french is ”

few-shot learner, if it can solve the target task if few examples are included in input.
input: “The translation of “a bird” to french is “l’oiseau” .
The translation of “a cat” to french is “le chat” .
…
The translation of “cheese” to french is ”

N-shot learning

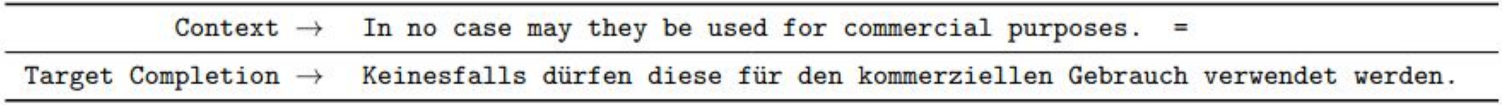
input: “The translation of “a bird” to french is “l’oiseau” .
The translation of “a cat” to french is “le chat” .
…
The translation of “cheese” to french is ”

What is the main difference from fine-tuning?

Zero-shot prompts from GPT-3

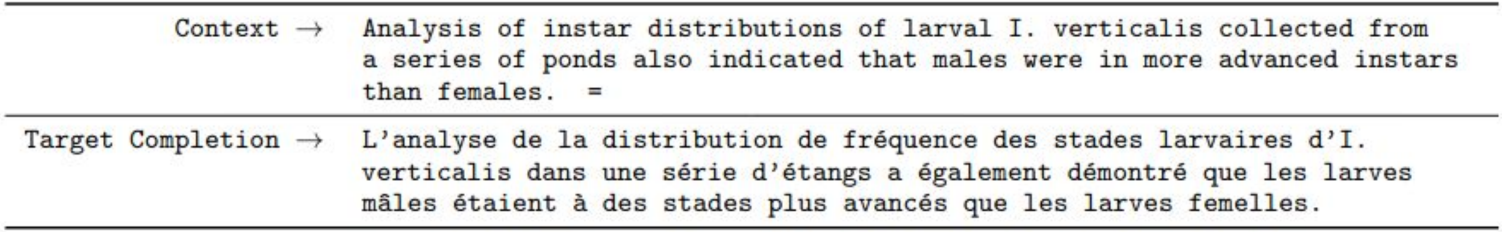
Prompt:
“Q: What is the {language} translation of {sentence} A: {translation}.”

Examples:



Context →	In no case may they be used for commercial purposes. =
Target Completion →	Keinesfalls dürfen diese für den kommerziellen Gebrauch verwendet werden.

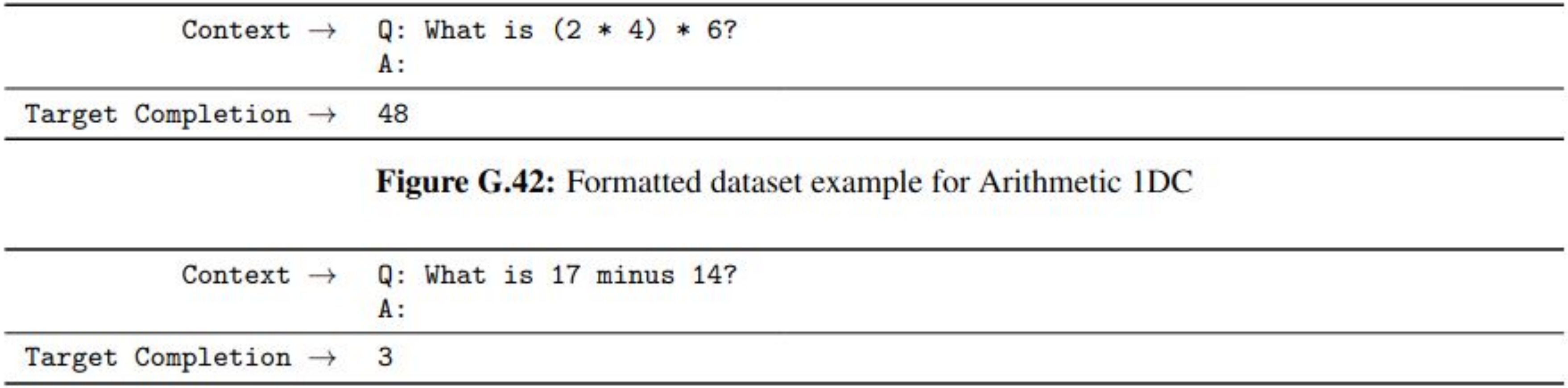
Figure G.37: Formatted dataset example for En→De



Context →	Analysis of instar distributions of larval I. verticalis collected from a series of ponds also indicated that males were in more advanced instars than females. =
Target Completion →	L’analyse de la distribution de fréquence des stades larvaires d’I. verticalis dans une série d’étangs a également démontré que les larves mâles étaient à des stades plus avancés que les larves femelles.

Figure G.38: Formatted dataset example for En→Fr

Zero-shot prompts from GPT-3

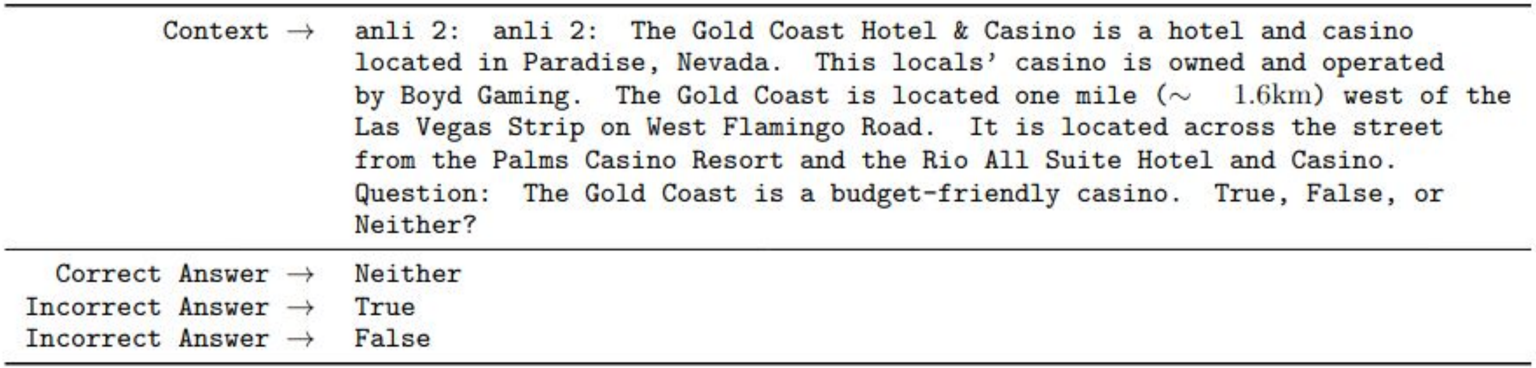


Context →	Q: What is (2 * 4) * 6? A:
Target Completion →	48

Figure G.42: Formatted dataset example for Arithmetic 1DC

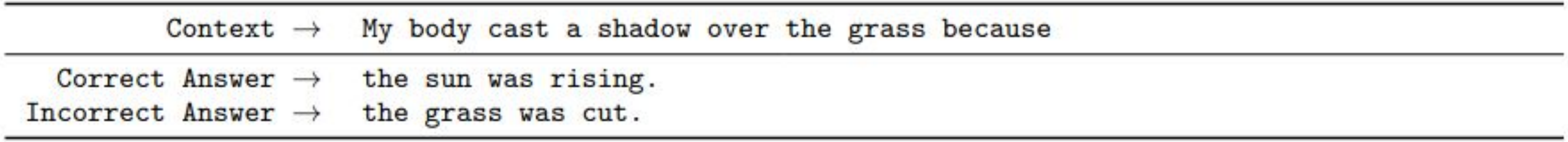
Context →	Q: What is 17 minus 14? A:
Target Completion →	3

Few-shot prompts from GPT-3



Context →	anli 2: anli 2: The Gold Coast Hotel & Casino is a hotel and casino located in Paradise, Nevada. This locals' casino is owned and operated by Boyd Gaming. The Gold Coast is located one mile (∼ 1.6km) west of the Las Vegas Strip on West Flamingo Road. It is located across the street from the Palms Casino Resort and the Rio All Suite Hotel and Casino. Question: The Gold Coast is a budget-friendly casino. True, False, or Neither?
Correct Answer →	Neither
Incorrect Answer →	True
Incorrect Answer →	False

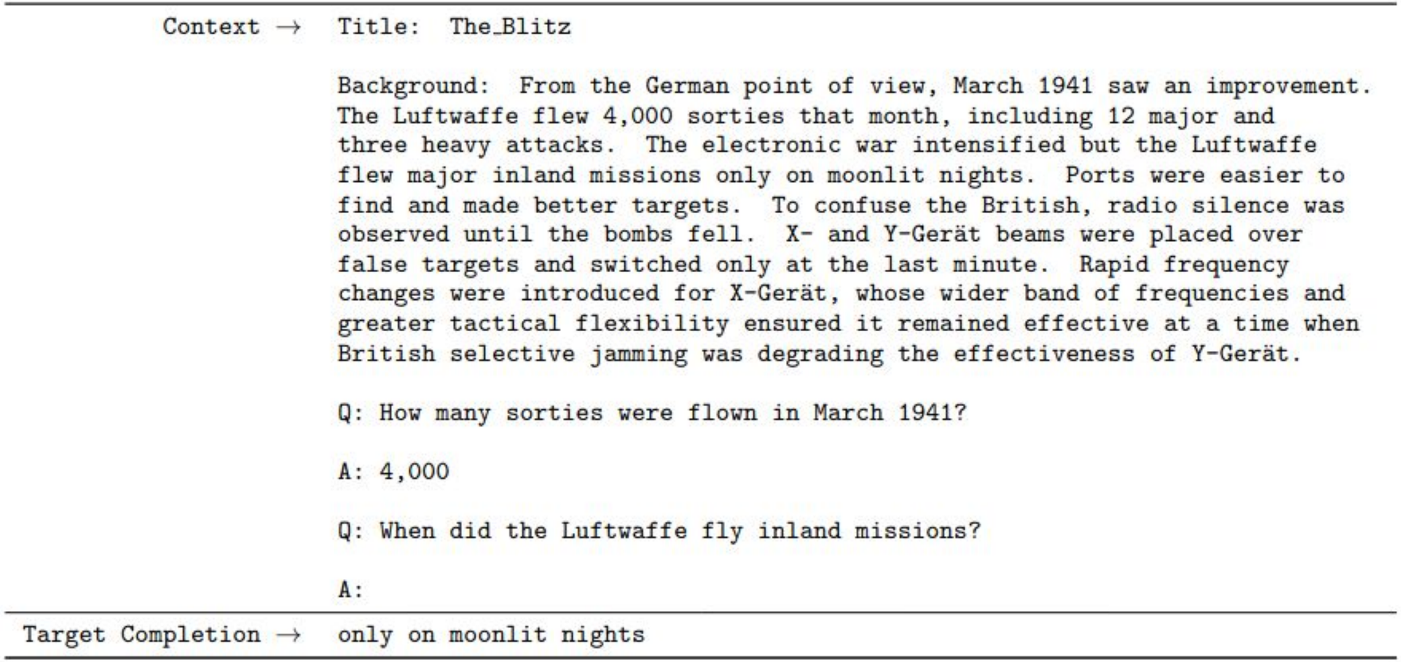
Figure G.2: Formatted dataset example for ANLI R2



Context →	My body cast a shadow over the grass because
Correct Answer →	the sun was rising.
Incorrect Answer →	the grass was cut.

Figure G.5: Formatted dataset example for COPA

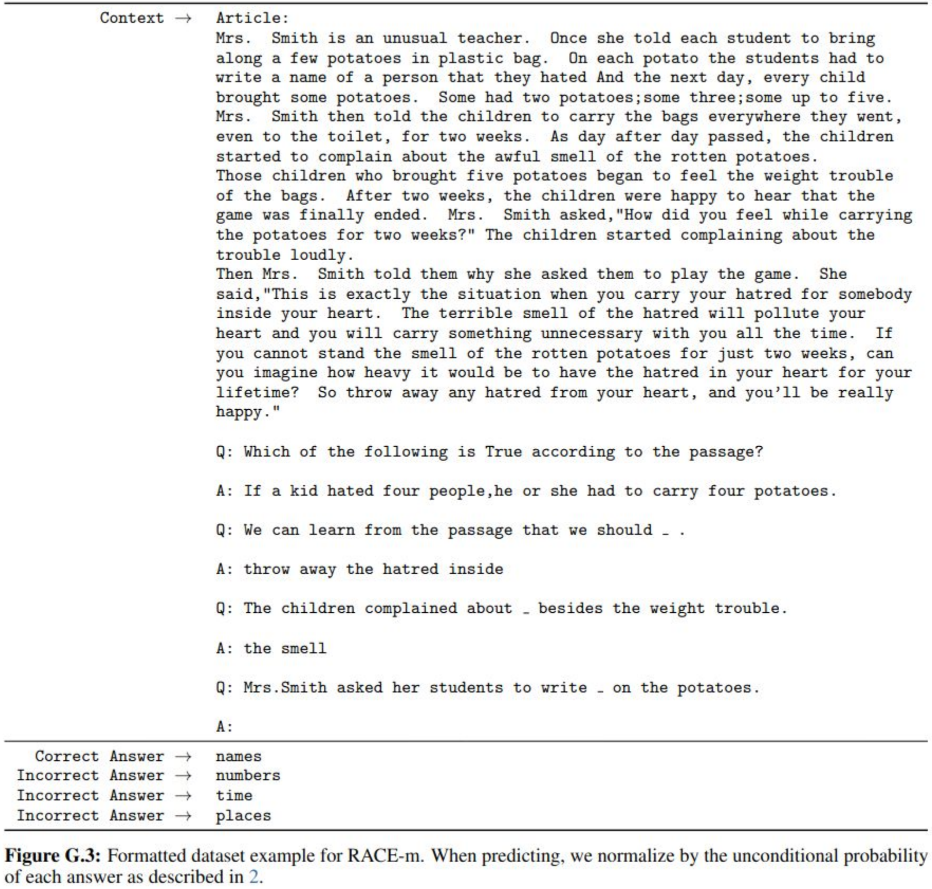
Few-shot prompts from GPT-3



Context →	Title: The_Blitz Background: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Gerät beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Gerät, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Gerät. Q: How many sorties were flown in March 1941? A: 4,000 Q: When did the Luftwaffe fly inland missions? A:
Target Completion →	only on moonlit nights

Figure G.28: Formatted dataset example for SQuADv2

Few-shot prompts from GPT-3



```
          Context → Article:
                    Mrs.  Smith is an unusual teacher.  Once she told each student to bring
                    along a few potatoes in plastic bag.  On each potato the students had to
                    write a name of a person that they hated And the next day, every child
                    brought some potatoes.  Some had two potatoes;some three;some up to five.
                    Mrs.  Smith then told the children to carry the bags everywhere they went,
                    even to the toilet, for two weeks.  As day after day passed, the children
                    started to complain about the awful smell of the rotten potatoes.
                    Those children who brought five potatoes began to feel the weight trouble
                    of the bags.  After two weeks, the children were happy to hear that the
                    game was finally ended.  Mrs.  Smith asked,"How did you feel while carrying
                    the potatoes for two weeks?" The children started complaining about the
                    trouble loudly.
                    Then Mrs.  Smith told them why she asked them to play the game.  She
                    said,"This is exactly the situation when you carry your hatred for somebody
                    inside your heart.  The terrible smell of the hatred will pollute your
                    heart and you will carry something unnecessary with you all the time.  If
                    you cannot stand the smell of the rotten potatoes for just two weeks, can
                    you imagine how heavy it would be to have the hatred in your heart for your
                    lifetime?  So throw away any hatred from your heart, and you'll be really
                    happy."

                    Q: Which of the following is True according to the passage?

                    A: If a kid hated four people,he or she had to carry four potatoes.

                    Q: We can learn from the passage that we should _ .

                    A: throw away the hatred inside

                    Q: The children complained about _ besides the weight trouble.

                    A: the smell

                    Q: Mrs.Smith asked her students to write _ on the potatoes.

                    A:
   Correct Answer → names
 Incorrect Answer → numbers
 Incorrect Answer → time
 Incorrect Answer → places
```

Figure G.3: Formatted dataset example for RACE-m. When predicting, we normalize by the unconditional probability of each answer as described in 2.

Next: seminar in colab

shorturl.at/CMQTU

AIRI

Artificial Intelligence Research Institute

airi.net

airi_research_institute

AIRI Institute

AIRI Institute

AIRI_inst

artificial-intelligence-research-institute

Telegram

Github of NLP Course

Feedback

AIRI × Университет Сириус