

Трансформерные модели. Часть 2.

Алексей Андреевич Сорокин

МГУ им. М. В. Ломоносова
весенний семестр 2022–2023 учебного года
Межфакультетский курс “Введение в компьютерную
лингвистику” 12 апреля, занятие 8

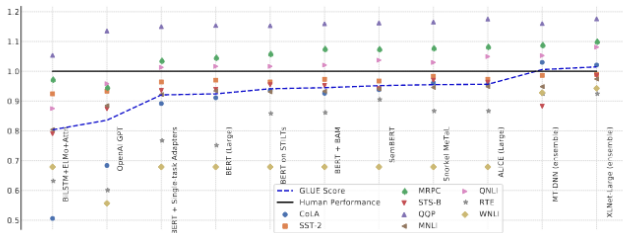
Набор GLUE

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

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

Набор GLUE

- Довольно быстро на GLUE было превышено качество, достигаемое человеком:



SuperGLUE

- Новый набор – SuperGLUE:

Table 1: The tasks included in SuperGLUE. *WSD* stands for word sense disambiguation, *NLI* is natural language inference, *coref.* is coreference resolution, and *QA* is question answering. For MultiRC, we list the number of total answers for 456/83/166 train/dev/test questions.

Corpus	Train	Dev	Test	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	F1 _a /EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

SuperGLUE

- Новый набор – SuperGLUE:

BoolQ	Passage: <i>Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.</i> Question: <i>is barq's root beer a pepsi product</i> Answer: No
CB	Text: <i>B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?</i> Hypothesis: <i>they are setting a trend</i> Entailment: Unknown
COPA	Premise: <i>My body cast a shadow over the grass.</i> Question: <i>What's the CAUSE for this?</i> Alternative 1: <i>The sun was rising.</i> Alternative 2: <i>The grass was cut.</i> Correct Alternative: 1

SuperGLUE

- Новый набор – SuperGLUE:

MultiRC

Paragraph: Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week

Question: Did Susan's sick friend recover? **Candidate answers:** Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)

ReCoRD

Paragraph: (CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electoral Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood

Query For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the <placeholder> presidency **Correct Entities:** US

SuperGLUE

- Новый набор – SuperGLUE:

RTE	Text: Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation. Hypothesis: Christopher Reeve had an accident. Entailment: False
WiC	Context 1: Room and <u>board</u> . Context 2: He nailed <u>boards</u> across the windows. Sense match: False
WSC	Text: Mark told <u>Pete</u> many lies about himself, which Pete included in his book. <u>He</u> should have been more truthful. Coreference: False

SuperGLUE

- BERT – базовое решение для SuperGLUE:
- Классификация пары предложений:
 - BoolQ, CB, RTE, WiC.
 - Логистическая регрессия на 2 класса для CLS-вектора.

SuperGLUE

- BERT – базовое решение для SuperGLUE:
- Классификация пары предложений:
 - BoolQ, CB, RTE, WiC.
 - Логистическая регрессия на 2 класса для CLS-вектора.
- Выбор варианта:
 - COPA, MultiRC, ReCoRD.
 - Для каждого варианта считается CLS-вектор:

$$h_i = \text{Encoder}([\text{Context}, \text{Variant}_i])[0]$$

SuperGLUE

- BERT – базовое решение для SuperGLUE:
- Классификация пары предложений:
 - BoolQ, CB, RTE, WiC.
 - Логистическая регрессия на 2 класса для CLS-вектора.
- Выбор варианта:
 - COPA, MultiRC, ReCoRD.
 - Для каждого варианта считается CLS-вектор:

$$h_i = \text{Encoder}([\text{Context}, \text{Variant}_i])[0]$$

- Эти представления пропускаются через дополнительную сеть:

$$s_i = \sigma(\langle w, h_i \rangle)$$

SuperGLUE

- BERT – базовое решение для SuperGLUE:
- Классификация пары предложений:
 - BoolQ, CB, RTE, WiC.
 - Логистическая регрессия на 2 класса для CLS-вектора.
- Выбор варианта:
 - COPA, MultiRC, ReCoRD.
 - Для каждого варианта считается CLS-вектор:

$$h_i = \text{Encoder}([\text{Context}, \text{Variant}_i])[0]$$

- Эти представления пропускаются через дополнительную сеть:

$$s_i = \sigma(\langle w, h_i \rangle)$$

- Далее классификатор выбирает самую большую s_i :

$$[p_1, \dots, p_K] = \text{softmax}([s_1, \dots, s_K])$$

- В случае нескольких правильных ответов – логистическая регрессия применяется к каждому h_i .

Другие данные для тестирования

- SQuAD1.1 / SQuAD 2.0 – нахождение ответа в тексте.
- MNLI – проверка логического следования.
- SST-2 (Stanford Sentiment Treebank) – анализ тональности.
- RACE – понимание текста (выбор верного варианта ответа на вопрос).
- STS-B (Semantic Textual Similarity) – проверка схожести текстов.
- CNN/Daily Mail Corpus – автоматическое реферирование (для новостных текстов).
- WMT – машинный перевод, особенно часто
 - En-De, De-En, En-Fr, Fr-En.
 - En-Ro, Ro-En – для малоресурсного тестирования.

RussianSuperGLUE

DaNetQA

Passage: В период с 1969 по 1972 год по программе «Аполлон» было выполнено 6 полётов с посадкой на Луне.

Question: Был ли человек на луне?

Answer: Yes

PARus

Premise: Гости вечеринки прятались за диваном.

Question: Почему это произошло?

Alternative 1: Это была вечеринка-сюрприз.

Alternative 2: Это был день рождения.

Correct Alternative: 1

RussianSuperGLUE

RUSSE

Context 1: Бурные хорошие дорожки заглушили шаги. **Context 2:** Приятели решили выпить на дорожку в местном баре.

Sense match: False

TERRA

Text: Автор поста написал в комментарии, что произошла канализация.

Hypothesis: Автор поста написал про канализацию. **Entailment:** True

Бенчмарки для русского языка: SuperGLUE

- TERRa – проверка логического следования (аналог RTE),
- DaNetQA – ответ на да/нет-вопрос (аналог BoolQ),
- PaRUS – выбор правильного продолжения предложения (аналог COPA)

Бенчмарки для русского языка: SuperGLUE

- TERRa – проверка логического следования (аналог RTE),
- DaNetQA – ответ на да/нет-вопрос (аналог BoolQ),
- PaRUS – выбор правильного продолжения предложения (аналог COPA)
- Другие аналоги задач из SuperGLUE.
- Задачи не из SuperGLUE:
 - Распознавание именованных сущностей – Collection5,
 - Поиск ответа в тексте – SberQUAD
- Чаще всего качество данных хуже соответствующих английских задач.

Roberta

- Roberta – модификация BERT, отличающаяся:
 - Динамическим маскированием.
 - Отсутствием задачи проверки следования предложений друг за другом.
 - Большим размером батча при обучении.
 - Способом объединения предложений в батчи.

Roberta

- Roberta – модификация BERT, отличающаяся:
 - Динамическим маскированием.
 - Отсутствием задачи проверки следования предложений друг за другом.
 - Большим размером батча при обучении.
 - Способом объединения предложений в батчи.
- Также были изменены обучающий корпус и число шагов при обучении.
- Были изменены некоторые параметры оптимизатора и генерации обучающих данных.
- Словарь модели составлен на уровне байтов и расширен до 50000.

Модификации BERT

Roberta

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	94.6/89.4	90.2	96.4
BERT _{LARGE}						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
XLNet _{LARGE}						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

AIBERT

AIBERT – модификация BERT, отличающаяся:

- Общие параметры у всех слоёв Трансформера.

AIBERT

AIBERT – модификация BERT, отличающаяся:

- Общие параметры у всех слоёв Трансформера.
- Двухступенчатое вычисление эмбеддингов (позволяет уменьшить число параметров):

$$\begin{aligned}\tilde{e}_i &= V_1[0, \dots, 1, \dots, 0], \\ e_i &= V_2 \tilde{e}_i\end{aligned}$$

AIBERT

AIBERT – модификация BERT, отличающаяся:

- Общие параметры у всех слоёв Трансформера.
- Двухступенчатое вычисление эмбеддингов (позволяет уменьшить число параметров):

$$\begin{aligned}\tilde{e}_i &= V_1[0, \dots, 1, \dots, 0], \\ e_i &= V_2 \tilde{e}_i\end{aligned}$$

- Замена задачи проверки следования на распознавание порядка следования (в каком порядке идут два предложения в документе).

ALBERT: результаты

- Можно увеличить размер эмбедингов:

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

- Это приводит к росту результатов.

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	94.1/88.3	88.1/85.1	88.0	95.2	82.3	88.7	0.3x

ALBERT: результаты

- Задача проверки порядка предложений оказывается более удачной для предобучения:

SP tasks	Intrinsic Tasks			Downstream Tasks					
	MLM	NSP	SOP	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg
None	54.9	52.4	53.3	88.6/81.5	78.1/75.3	81.5	89.9	61.7	79.0
NSP	54.5	90.5	52.0	88.4/81.5	77.2/74.6	81.6	91.1	62.3	79.2
SOP	54.0	78.9	86.5	89.3/82.3	80.0/77.1	82.0	90.3	64.0	80.1

- При этом исходный BERT не умеет её решать.

GPT-2

SYSTEM PROMPT
(HUMAN-WRITTEN)

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.

MODEL
COMPLETION
(MACHINE-
WRITTEN, 10
TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

GPT-2

- GPT-2 – однонаправленная языковая модель на основе Трансформера.
- Она обучалась только на задачу предсказания следующего слова:
 - 1,5 миллиарда параметров.
 - Обучение на 8 миллионах разнообразных интернет-страниц (порядка 40Гб текста).

GPT-2

- GPT-2 – однонаправленная языковая модель на основе Трансформера.
- Она обучалась только на задачу предсказания следующего слова:
 - 1,5 миллиарда параметров.
 - Обучение на 8 миллионах разнообразных интернет-страниц (порядка 40Гб текста).
- Основное применение – генерация продолжения текста по заголовку.
- Другое применение – few-shot learning: обучение на новую задачу по очень небольшому (10-100) количеству примеров.
- При этом задача должна быть сформулирована как языковое моделирование.

GPT-2: постановки задач

- Анализ тональности:

В этом ресторане отличная еда. sentiment = ?

- Переформулировка на языке GPT:

В этом ресторане отличная еда. Это X . – контекст

$p(X = \text{хорошо}) \geq p(X = \text{плохо}) \rightarrow \text{sentiment} = \text{positive}$

$p(X = \text{плохо}) \geq p(X = \text{хорошо}) \rightarrow \text{sentiment} = \text{negative}$

- Так можно переформулировать и другие задачи (ответ на вопросы, автоматическое реферирование, ...)
- Этот подход будет работать вообще без обучения.

Большие языковые модели

Модель	Авторы	Размер	Размер корпуса
LaMDA	Google	137B	168B
GPT-3	OpenAI	175B	300B
Jurassic	AI21	178B	300B
Gopher	DeepMind	280B	300B
MT-NLG	Microsoft+NVIDIA	530B	270B
Chinchilla	DeepMind	70B	1.4T
PaLM	Google	137B	168 B

Применение больших моделей

We use cookies to ensure that we give you the best experience on our website. If you continue to use this s

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many legs does a frog have?

A: A frog has four legs.

Q: Are there any animals with three legs?

A: No, there are no animals with three legs.

Q: Why don't animals have three legs?

A: Animals don't have three legs because they would fall over.

OpenAI GPT-3 Turing Test – Common Sense Questions

Применение больших моделей



Manuel Araoz @maraoz · Jul 20

This is my all-time favorite: rewrite emails in a calmer tone. This could save lives!! :)

- Original: "Stop fucking sending me emails. I hate it."
- Toned down: "Do not send me emails"
- Original: "I need to talk to you ASAP, call me"
- Toned down: "Kindly contact me when you read this, it's urgent."
- Original: "As you can read in my previous email, I already told you that we won't be able to make it."
- Toned down: "There might be a misunderstanding. Unfortunately, I don't think we'll be able to make it this time."
- Original: "HEY MARK! Pls send tomorrow's presentation ASAP. I need to work on it tonight. Thx."
- Toned down: "Mark, please send tomorrow's presentation as soon as possible. I need to work on it tonight."
- Original: "Your answer is a b.s. The phone is no good! Return the money! -Return the money! I'm not paying for this phone! -You are the worst company I've ever dealt with!"



2



18



72



OpenAI GPT-3 It can change tone of the sentence



Применение больших моделей

prompt	<p>Q: Tracy used a piece of wire 4 feet long to support tomato plants in the garden. The wire was cut into pieces 6 inches long. How many pieces did she obtain?</p> <p>A:</p>	prompt	<p>Q: Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?</p> <p>A:</p>
62b model	<p>She cut the wire into 6 inch pieces. This means she got $4 * 6 = 24$ pieces. The answer is 24. ✗</p>	62b model	<p>He travels from 1 to 4 PM. That means he travels for 3 hours. He travels back at 6 mph. That means he travels for $3 * 6 = 18$ miles. The answer is 18. ✗</p>
540b model	<p>The wire was 4 feet long. This means it was $4 * 12 = 48$ inches long. It was cut into pieces 6 inches long. This means she obtained $48 / 6 = 8$ pieces. The answer is 8. ✓</p>	540b model	<p>He travels at 10 mph from 1 to 4 PM. This means he travels 3 hours. 3 hours at 10 mph means he travels $3 * 10 = 30$ miles. He then travels back at 6 mph. This means he travels 6 miles per hour. He has to travel 30 miles, so it takes him $30 / 6 = 5$ hours. The answer is 5. ✓</p>

GPT-4

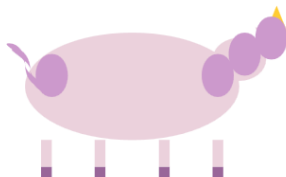
Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

```
Yes, I think I can,  
Though it might take a clever plan.  
I'll start by noting Euclid's proof,  
Which shows that primes aren't just aloof.  
Assume we have a finite list,  
Of primes, and that none have been missed.  
Multiply them all together,  
And add one, just to be clever.  
The result will be a number,  
That has no prime factors, oh wonder!  
But every number has a prime,  
That divides it, it's just a matter of time.  
So we've found a contradiction,  
And our finite list needs eviction.  
There must be infinitely many primes,  
And that's the end of my rhyming lines.s
```

Prompt: Draw a unicorn in TiKZ.

GPT-4: [Produces \LaTeX compiling to following picture.]



Большие языковые модели: комментарий

- Это примеры на середину 2022 года. С тех пор результаты языковых моделей стали несравнимо более впечатляющими.
- Сейчас языковые модели (GPT4) могут выполнять практически любое задание, сформулированное в виде инструкции:
 - *переведи текст,*
 - *ответь на вопрос,*
 - *напиши диалог в стиле древнегреческих философов*
 - ...
- Один из элементов – *instruction tuning* (обучение инструкциям).

Обучение инструкциям

- Многие задачи NLP можно переформулировать как следование инструкциям:
 - анализ тональности

Это отзыв о фильме, оцени его по пятибалльной шкале⟨source⟩
Оценка: ⟨target⟩

Обучение инструкциям

- Многие задачи NLP можно переформулировать как следование инструкциям:

- анализ тональности

Это отзыв о фильме, оцени его по пятибалльной шкале⟨source⟩
Оценка: ⟨target⟩

- Исправление грамматических ошибок

Это текст, который, возможно, содержит ошибки, исправь их⟨source⟩
Исправленный текст: ⟨target⟩

- Следующий этап: научить модель следовать произвольным инструкциям (не привязанным к конкретной задаче).

Обучение инструкциям

- Многие задачи NLP можно переформулировать как следование инструкциям:
- Следующий этап: научить модель следовать произвольным инструкциям (не привязанным к конкретной задаче).
- Этому можно обучить, сформулировав стандартные задачи в виде инструкций и обучая модель решать их все.

Обучение инструкциям

- Многие задачи NLP можно переформулировать как следование инструкциям:
- Следующий этап: научить модель следовать произвольным инструкциям (не привязанным к конкретной задаче).
- Этому можно обучить, сформулировав стандартные задачи в виде инструкций и обучая модель решать их все.
- При этом нужны не только типичные прикладные задачи, но и задачи "из жизни":
 - *сочини короткий рассказ для пятилетнего ребёнка*
 - *придумай рекламное объявление по карточке товара*
 - *тебе не продают в магазине алкоголь, напиши, что ты ответишь*

Мотивация

- Все описанные модели для предобучения (кроме GPT) представляют собой энкодеры.
- Они кодируют исходную последовательность в последовательность состояний той же длины.

Мотивация

- Все описанные модели для предобучения (кроме GPT) представляют собой энкодеры.
- Они кодируют исходную последовательность в последовательность состояний той же длины.
- Они не предобучаются на генерацию нового текста.
- Соответственно, не подходят для аналогичных задач:
 - Суммаризация.
 - Машинный перевод.
- GPT годится, но для обозримых размеров модели слишком плохо работает.
- Она училась продолжать текст, а не преобразовывать один текст в другой.

Задачи преобразования текста

- Сложнее всего задачи преобразования одного текста в другой:
 - Суммаризация.
 - Машинный перевод.
 - Моделирование диалога.

Задачи преобразования текста

- Сложнее всего задачи преобразования одного текста в другой:
 - Суммаризация.
 - Машинный перевод.
 - Моделирование диалога.
- Поэтому нужно предобучаться на задачу, требующую порождения одного текста по другому.
- Это должна быть задача восстановления “испорченного” или “неполного” текста (они не требуют разметки).

Постановки задач для предобучения

- Различные варианты задач.

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 .	<i>(original text)</i>
Deshuffling	party me for your to . last fun you inviting week Thank	<i>(original text)</i>
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	<i>(original text)</i>
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>

Постановки задач для предобучения

- Различные варианты задач.

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>

- Результаты.

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

- Вывод: нужно предобучаться на задачу заполнения пропусков.

Постановки задач для предобучения

- Различные варианты задач.

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)
L.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
L.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>

- Результаты.

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
BERT-style (Devlin et al., 2018)	82.96	19.17	80.65	69.85	26.78	40.03	27.41
MASS-style (Song et al., 2019)	82.32	19.16	80.10	69.28	26.79	39.89	27.55
★ Replace corrupted spans	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Drop corrupted tokens	84.44	19.31	80.52	68.67	27.07	39.76	27.82

Постановки задач для предобучения

- Различные варианты задач.

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)
L.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
L.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>

- Результаты.

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
BERT-style (Devlin et al., 2018)	82.96	19.17	80.65	69.85	26.78	40.03	27.41
MASS-style (Song et al., 2019)	82.32	19.16	80.10	69.28	26.79	39.89	27.55
★ Replace corrupted spans	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Drop corrupted tokens	84.44	19.31	80.52	68.67	27.07	39.76	27.82

- В T5 удаляются фрагменты текста произвольной длины (заменяются на маркеры).

Результаты: T5

- Результаты T5 для разных задач:

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4 ^a	69.2 ^b	97.1 ^a	93.6^b	91.5^b	92.7 ^b	92.3 ^b
T5-Small	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	90.3	71.6	97.5	92.8	90.4	93.1	92.8

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	74.8 ^c	90.7^b	91.3 ^a	91.0 ^a	99.2^a	89.2 ^a	91.8 ^a
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	75.1	90.6	92.2	91.9	96.9	92.8	94.5

Результаты: T5

- Результаты T5 для разных задач:

Model	SQuAD EM	SQuAD F1	SuperGLUE Average	BoolQ Accuracy	CB F1	CB Accuracy	COPA Accuracy
Previous best	90.1 ^a	95.5 ^a	84.6 ^d	87.1 ^d	90.5 ^d	95.2 ^d	90.6 ^d
T5-Small	79.10	87.24	63.3	76.4	56.9	81.6	46.0
T5-Base	85.44	92.08	76.2	81.4	86.2	94.0	71.2
T5-Large	86.66	93.79	82.3	85.4	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	89.9	90.3	94.4	92.0
T5-11B	91.26	96.22	88.9	91.2	93.9	96.8	94.8

Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCoRD Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4 ^d	52.5 ^d	90.6 ^d	90.0 ^d	88.2 ^d	69.9 ^d	89.0 ^d
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	88.1	63.3	94.1	93.4	92.5	76.9	93.8

T5: Выводы

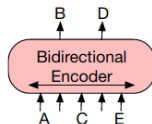
- Важно обучаться на большом количестве очищенных данных (Colossal Clean Crawled Corpus).

Data set	Size	GLUE	CNN3M	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	19.24	80.88	71.36	26.98	39.82	27.65
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	83.83	19.23	80.39	72.38	26.75	39.90	27.48
WebText-like	17GB	84.03	19.31	81.42	71.40	26.80	39.74	27.59
Wikipedia	16GB	81.85	19.31	81.29	68.01	26.94	39.69	27.67
Wikipedia + TBC	20GB	83.65	19.28	82.08	73.24	26.77	39.63	27.57

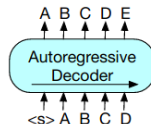
- Самая эффективная стратегия обучения – предобучение на большом корпусе + настройка под задачу. Многозадачность не помогает (на 2020г.).
- Впоследствии выяснилось, что многозадачности можно достичь с помощью шаблонов и обучения инструкциям.

BART: мотивация

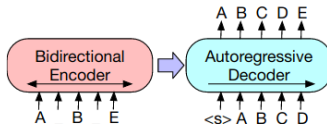
- Различные варианты постановки задачи предобучения:



(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.

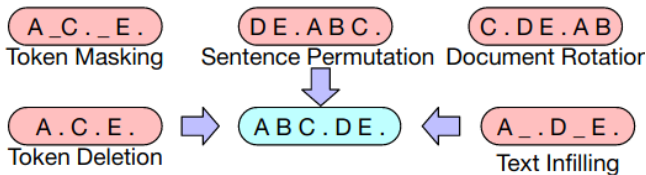


(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



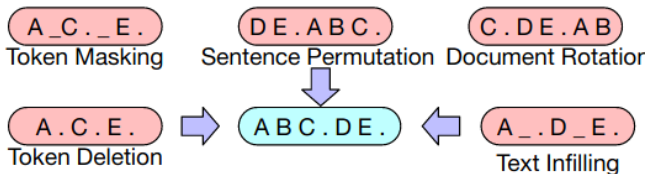
BART: постановка задачи

- BART объединяет все варианты восстановления текста:



BART: постановка задачи

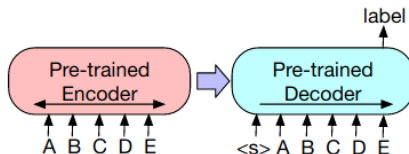
- BART объединяет все варианты восстановления текста:



- В итоге используется:
 - text infilling (заполнение пропусков, как в T5)
 - sentence permutation (восстановление порядка слов, как в XLNet).

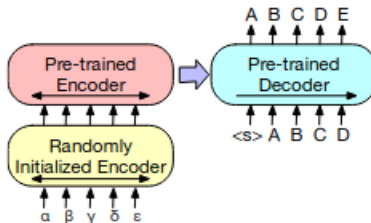
BART: постановка задачи

- При дообучении на классификацию один и тот же текст подаётся в энкодер и декодер:



BART: постановка задачи

- Для перевода добавляется ещё энкодер:



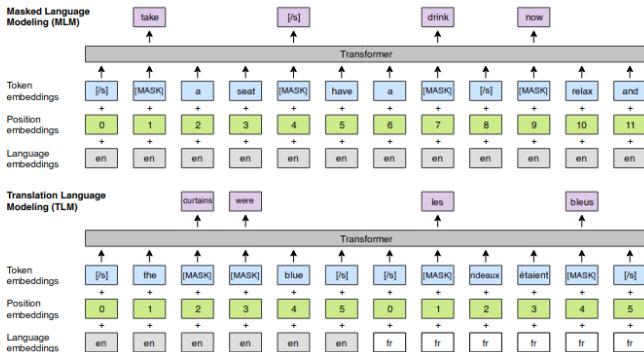
- Это нужно, чтобы адаптировать словарь модели под другой язык.

BART: применение

- Дообученный BART достигает передовых результатов на задачах порождения текста:
 - Автоматическое реферирование (суммаризация).
 - Порождение ответа на реплику в диалоге.
 - Малоресурсный машинный перевод

XLM

- XLM – это модификация мультязычного BERT, которая дополнительно обучается на языковое моделирование с переводом (Translation Language Modeling):



XLM: применение

- Инициализация для разных задач на языках, отличных от английского.
- Инициализация модели для машинного перевода.
- Языковое моделирование для малоресурсных языков.
- Побочный эффект – эмбединг перевода слова оказывается близок к исходному эмбедингу.

Мультиязычные модели

- XLM-R – модификация XLM, где вместо BERT взята Roberta.
- Наблюдения при обучении модели:
 - Чем больше языков, тем хуже качество в среднем.
 - Язык, отсутствующий в модели с малым числом языков, может иметь качество выше, чем когда он в ней появится.

Мультиязычные модели

- XLM-R – модификация XLM, где вместо BERT взята Roberta.
- Наблюдения при обучении модели:
 - Чем больше языков, тем хуже качество в среднем.
 - Язык, отсутствующий в модели с малым числом языков, может иметь качество выше, чем когда он в ней появится.
 - Мультиязычные модели выигрывают за счёт увеличения размеров словаря.
 - Обучаться надо с большим батчем (8тыс. примеров).
- После модификаций основной выигрыш на малоресурсных языках.

Мультиязычные модели

- mT5 – многоязычная версия T5 (107 языков).
 - Наибольшая по размеру модель (T5-XXL – 11 миллиардов параметров).
- мультиязычный BART (25 языков)
 - Позволяет достичь SOTA на машинном переводе, особенно для малоресурсных языков.
 - Хорошее предобучение для суммаризации.