

Transfer Learning в NLP

Лекция 3

Seq2seq pre-training

План занятия

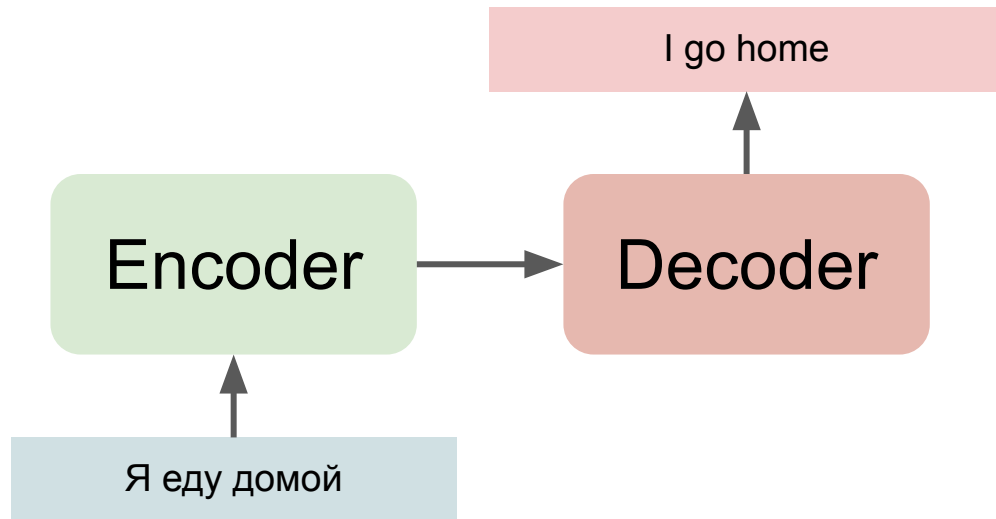
1. Напоминание о seq2seq
2. BART — ещё одна модель для self-supervised learning
3. T5 — сформулируем все задачи, как seq2seq

Напоминание

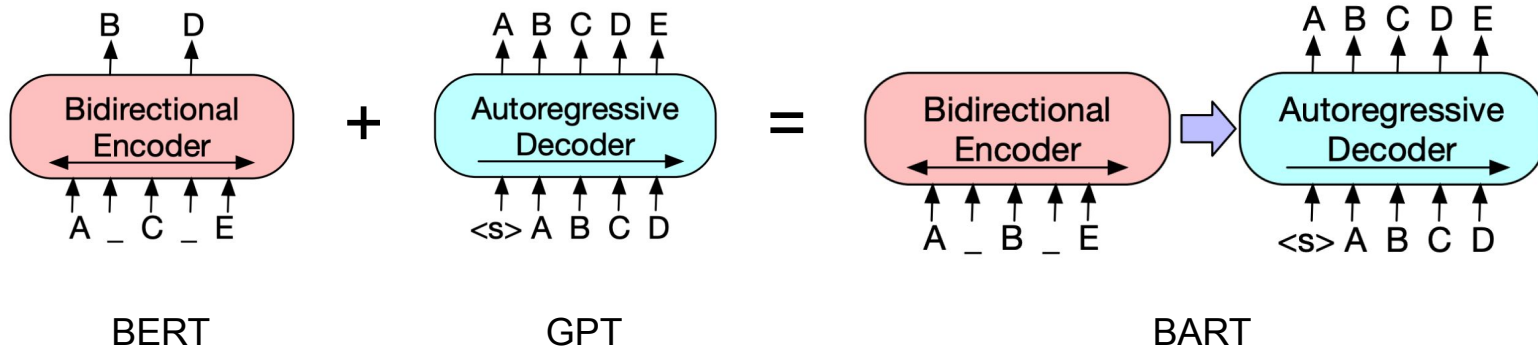
Seq2Seq — задача NLP, в которой нужно по входному тексту получить выходной текст.

Напоминание

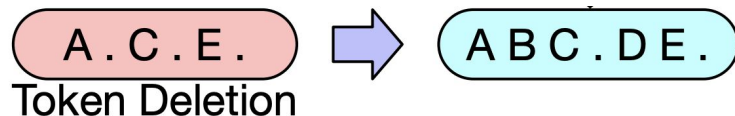
Seq2Seq — задача NLP, в которой нужно по входному тексту получить выходной текст.



BART — комбинация BERT и GPT



Гораздо большая свобода для формулирования задачи



Все задачи для тренировки BART

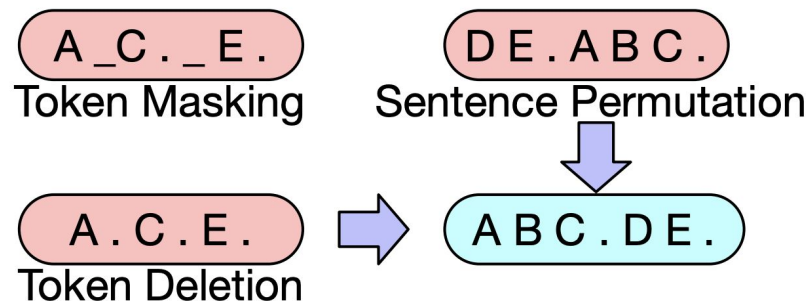
Гораздо большая свобода для формулирования задачи

A _ C . _ E .
Token Masking

A . C . E . ➡ A B C . D E .
Token Deletion

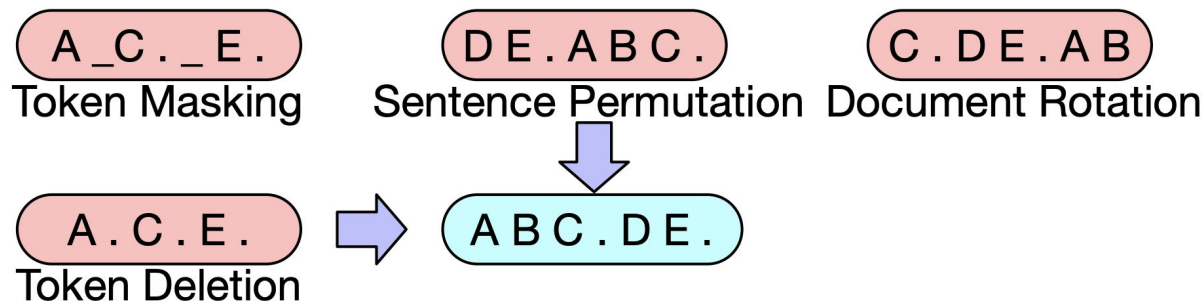
Все задачи для тренировки BART

Гораздо большая свобода для формулирования задачи



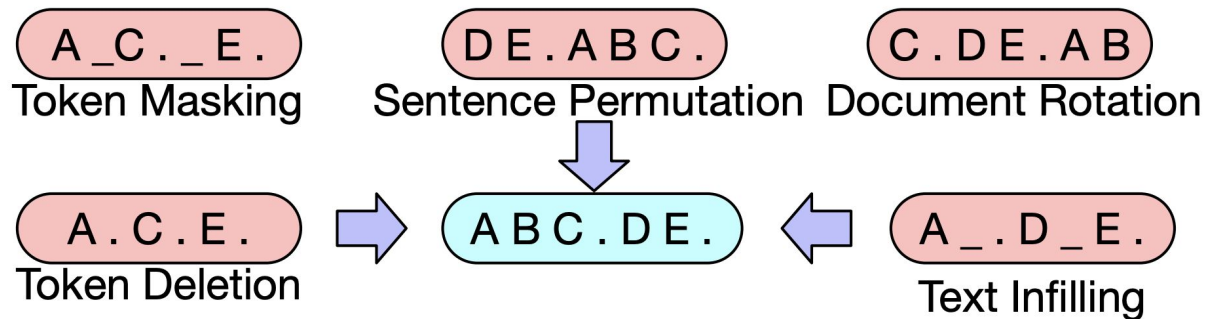
Все задачи для тренировки BART

Гораздо большая свобода для формулирования задачи



Все задачи для тренировки BART

Гораздо большая свобода для формулирования задачи



Все задачи для тренировки BART

Улучшили качество на seq2seq

1. Суммаризация

	CNN/DailyMail		
	R1	R2	RL
Lead-3	40.42	17.62	36.67
PTGEN (See et al., 2017)	36.44	15.66	33.42
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38
UniLM	43.33	20.21	40.51
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18
BART	44.16	21.28	40.90

Улучшили качество на seq2seq

1. Суммаризация
2. Машинный перевод

	CNN/DailyMail		
	R1	R2	RL
Lead-3	40.42	17.62	36.67
PTGEN (See et al., 2017)	36.44	15.66	33.42
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38
UniLM	43.33	20.21	40.51
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.22	38.76
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.22	38.76
BART	44.16	21.48	40.21
			RO-EN
			Baseline
			36.80
			Fixed BART
			36.29
			Tuned BART
			37.96

Улучшили качество на seq2seq

1. Суммаризация
2. Машинный перевод
3. Диалоговое моделирование

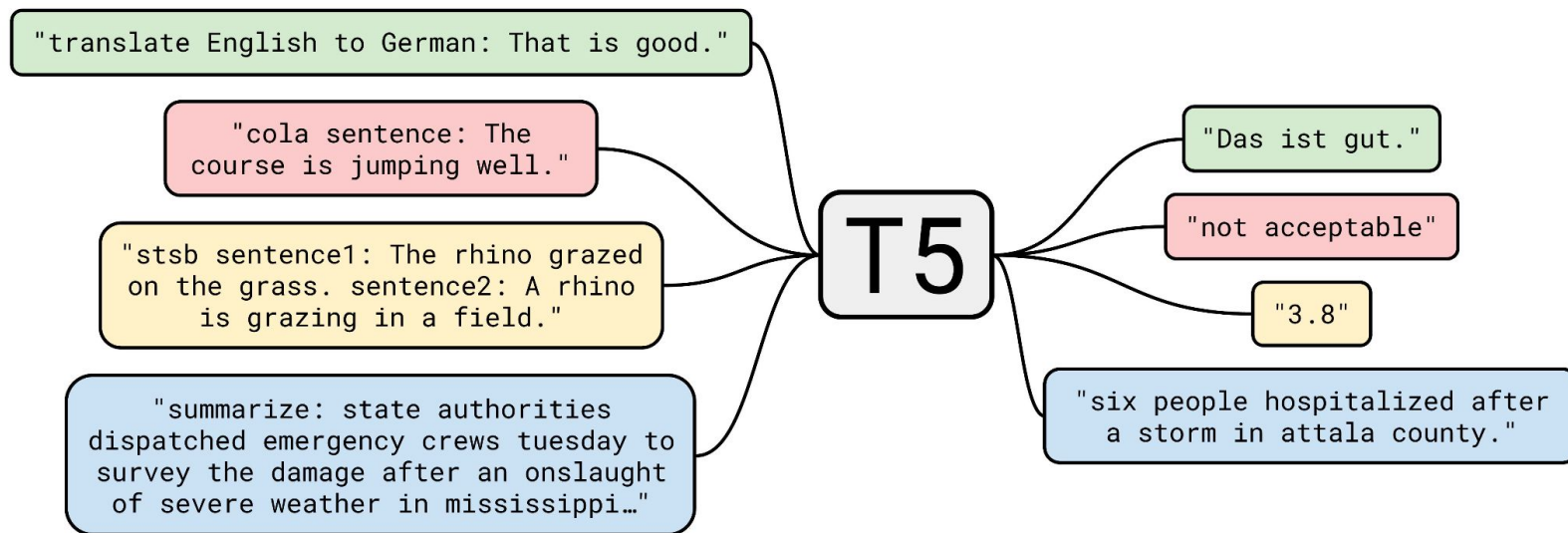
CNN/DailyMail			
	R1	R2	RL
Lead-3	40.42	17.62	36.67
PTGEN (See et al., 2017)	36.44	15.66	33.42
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38
UniLM	43.33	20.21	40.51
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.22	38.76
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.22	38.76
BART	44.16	21.48	40.76
RO-EN			
Baseline			36.80
Fixed BART			36.29
Tuned BART			37.96
ConvAI2			
	Valid F1	Valid PPL	
Seq2Seq + Attention	16.02	35.07	
Best System	19.09	17.51	
BART	20.72	11.85	

Улучшили качество на seq2seq

1. Суммаризация
2. Машинный перевод
3. Диалоговое моделирование
4. Question Answering

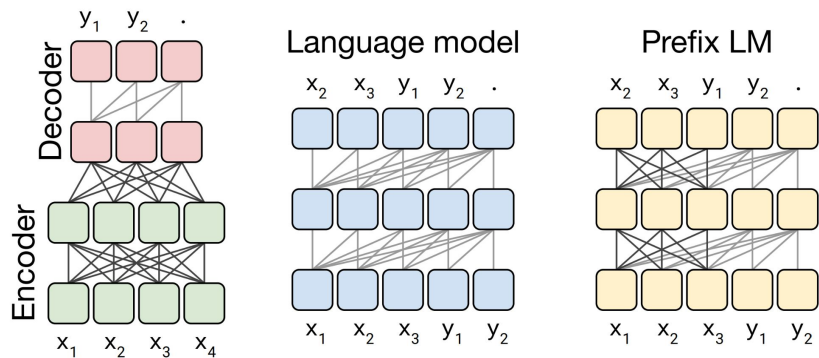
				CNN/DailyMail		
				R1	R2	RL
Lead-3				40.42	17.62	36.67
PTGEN (See et al., 2017)				36.44	15.66	33.42
PTGEN+COV (See et al., 2017)				39.53	17.28	36.38
UniLM				43.33	20.21	40.51
BERTSUMABS (Liu & Lapata, 2019)				41.72	19.22	38.76
BERTSUMEXTABS (Liu & Lapata, 2019)				42.13	19.22	38.76
BART				44.16	21.18	40.75
				RO-EN		
Best Extractive				23.5	3.1	17.5
Language Model				27.8	4.7	23.1
Seq2Seq				28.3	5.1	22.8
Seq2Seq Multitask				28.9	5.4	23.1
BART				30.6	6.2	24.3
				ConvAI2		
				Valid F1		Valid PPL
Seq2Seq + Attention				16.02		35.07
Best System				19.09		17.51
BART				20.72		11.85

T5 — Text-to-Text Transfer Transformer



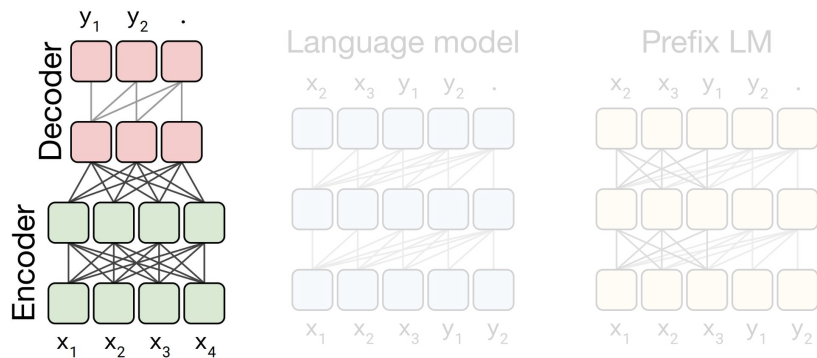
Обучим encoder-decoder трансформер восстанавливать целые куски текста

Pre-training этап



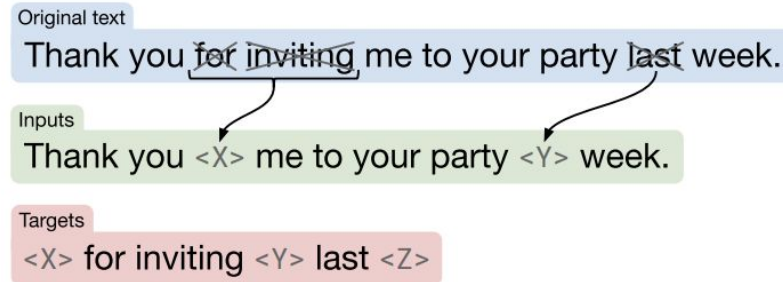
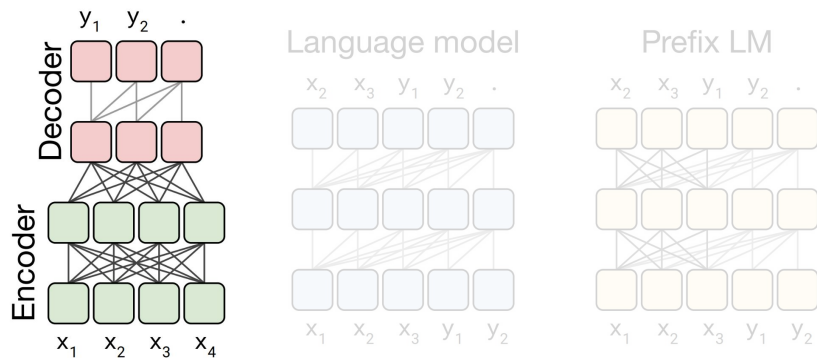
Обучим encoder-decoder трансформер восстанавливать целые куски текста

Pre-training этап



Обучим encoder-decoder трансформер восстанавливать целые куски текста

Pre-training этап



Влияние различных вариантов стратегий и подходов

1. Размеры и содержание датасетов
2. Количество шагов обучения
3. Сравнение с multitask learning
4. Маскирование vs Span Replacement vs Выбрасывание токенов
5. Shared параметры внутри encoder-decoder
6. ...

Получили SOTA результат почти везде

Обучили модель в нескольких вариациях:

- T5-small (60 млн.)
- T5-base (220 млн.)
- T5-large (770 млн.)
- T5-3B
- T5-11B

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

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

Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L
Previous best	33.8^e	43.8^e	38.5^f	43.47 ^g	20.30 ^g	40.63 ^g
T5-Small	26.7	36.0	26.8	41.12	19.56	38.35
T5-Base	30.9	41.2	28.0	42.05	20.34	39.40
T5-Large	32.0	41.5	28.1	42.50	20.68	39.75
T5-3B	31.8	42.6	28.2	42.72	21.02	39.94
T5-11B	32.1	43.4	28.1	43.52	21.55	40.69

<https://arxiv.org/pdf/1910.10683.pdf>

Получили SOTA результат почти везде

Обучили модель в нескольких вариациях:

- T5-small (60 млн.)
- T5-base (220 млн.)
- T5-large (770 млн.)
- T5-3B
- T5-11B

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 ^a	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 ^a	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

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

Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L
Previous best	33.8 ^e	43.8 ^e	38.5 ^f	43.47 ^g	20.30 ^g	40.63 ^g
T5-Small	26.7	36.0	26.8	41.12	19.56	38.35
T5-Base	30.9	41.2	28.0	42.05	20.34	39.40
T5-Large	32.0	41.5	28.1	42.50	20.68	39.75
T5-3B	31.8	42.6	28.2	42.72	21.02	39.94
T5-11B	32.1	43.4	28.1	43.52	21.55	40.69

<https://arxiv.org/pdf/1910.10683.pdf>

Итоги занятия

1. Вспомнили, что такое seq2seq-задачи, и привели несколько примеров.

Итоги занятия

1. Вспомнили, что такое seq2seq-задачи, и привели несколько примеров.
2. Познакомились с архитектурой BART, которая позволяет применять transfer learning для seq2seq-задач и узнали о способах зашумления текста в случае BART.

Итоги занятия

1. Вспомнили, что такое seq2seq-задачи, и привели несколько примеров.
2. Познакомились с архитектурой BART, которая позволяет применять transfer learning для seq2seq-задач и узнали о способах зашумления текста в случае BART.
3. Узнали, что такое T5 и каким способом обучалась эта модель