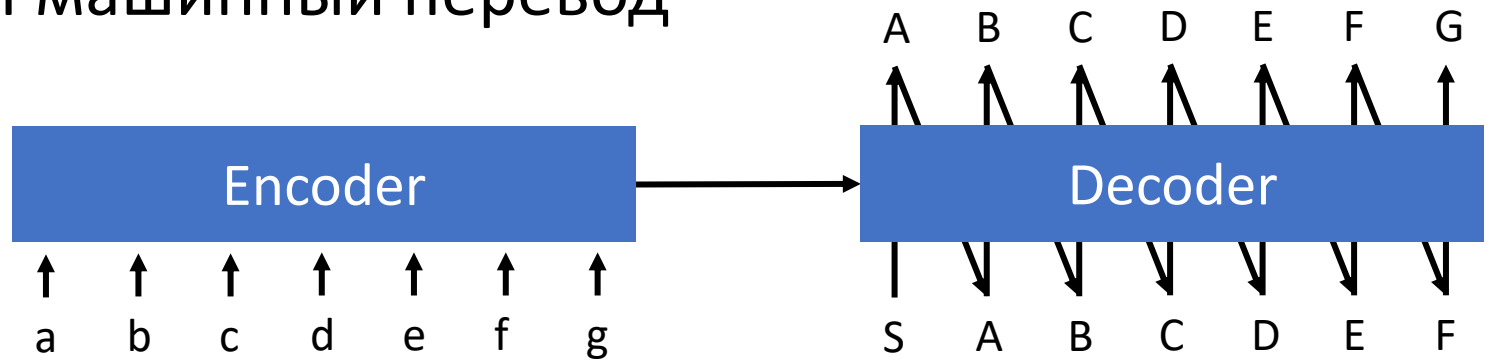
Mask-Predict: Parallel Decoding of Conditional Masked Language Models

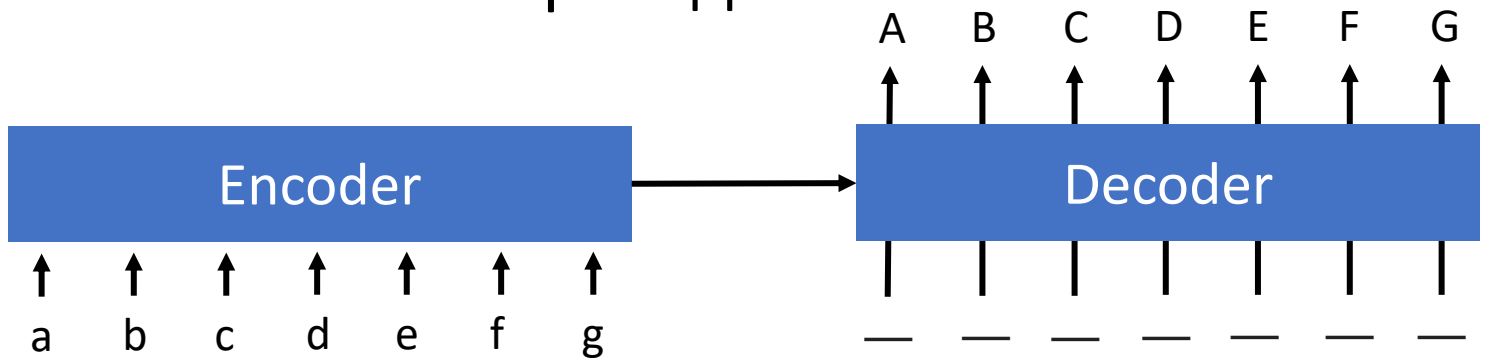
Илья Притуляк

Подходы к переводу текста

- Авторегрессионный машинный перевод

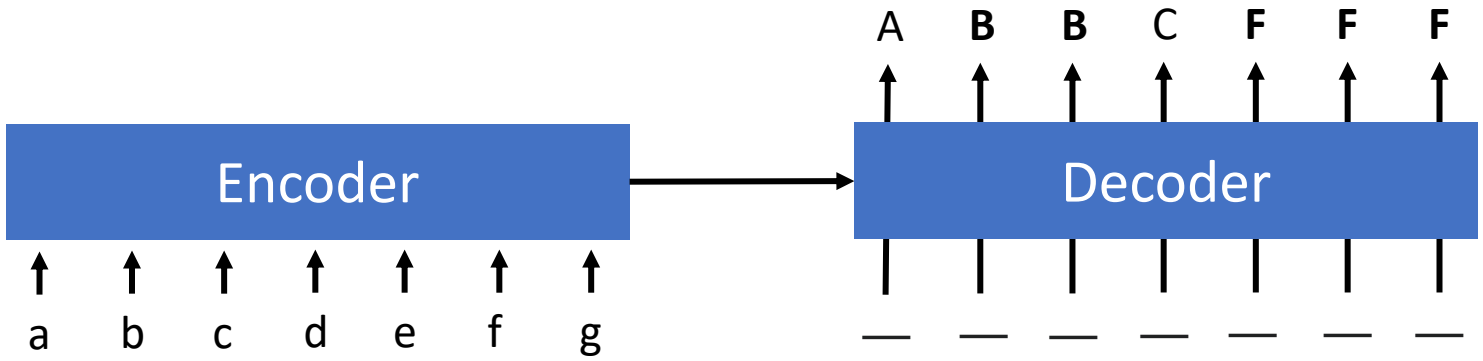


- Неавторегрессионный машинный перевод



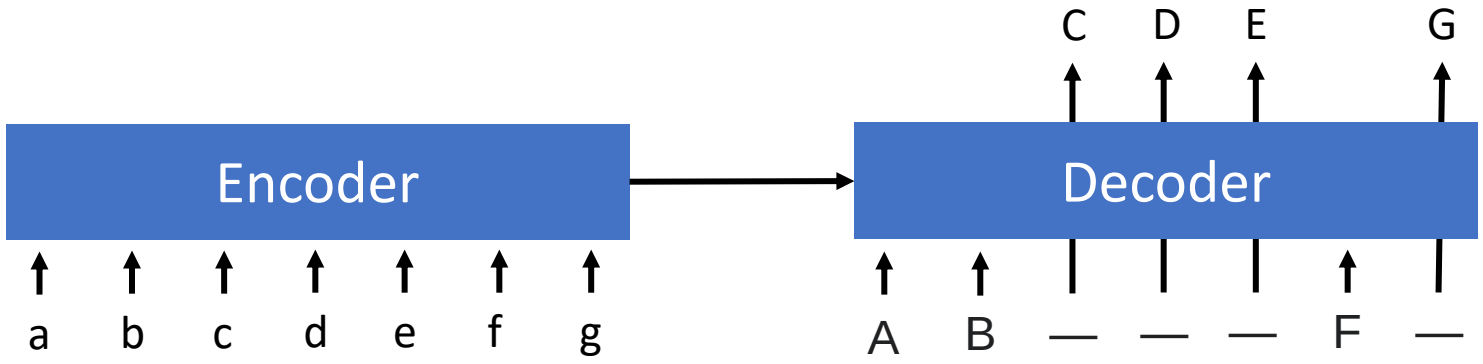
Проблема мультимодальности

- Хотим перевести с английского на русский "Thank you very much"
- Варианты переводов: "Спасибо тебе большое", "Большое тебе спасибо"
- Возможный перевод неавторегрессионной модели: "Спасибо тебе спасибо"



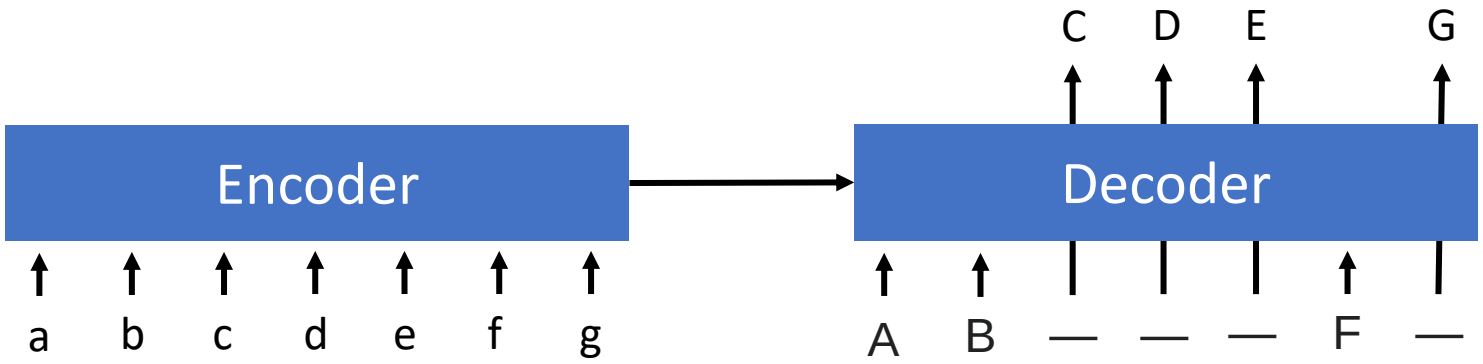
Архитектура и идея модели

- За основу взят обычный трансформер
- Self-attention в декодере теперь двунаправленный
- Хотим предсказать некоторое подмножество итоговых токенов на основе входного текста и остальных токенов



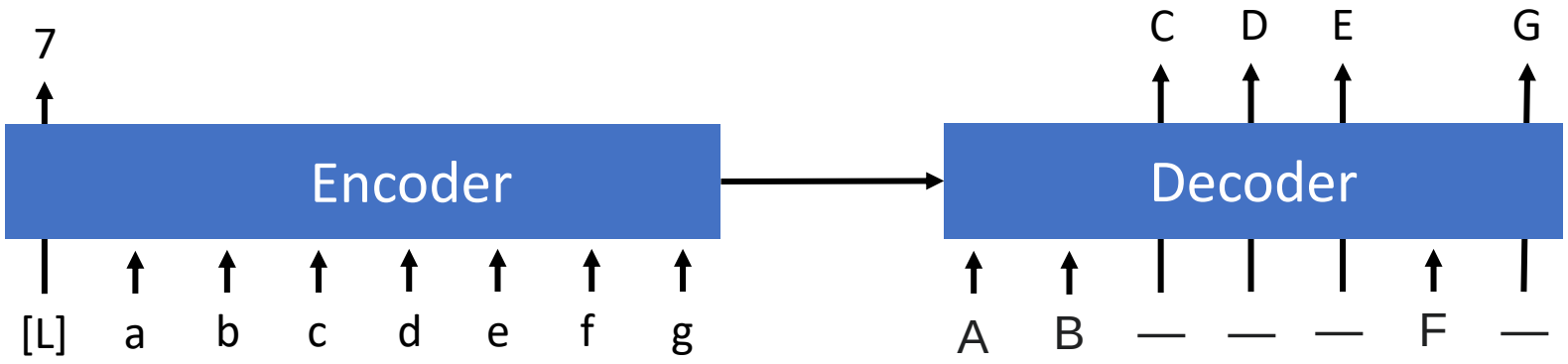
Обучение модели

- Маскируем k случайных токенов выходной последовательности, где k из $\mathrm{Uniform}(1, N)$ и предсказываем их
- Целевая функция: кросс-энтропия на маскируемых токенах



Обучение модели

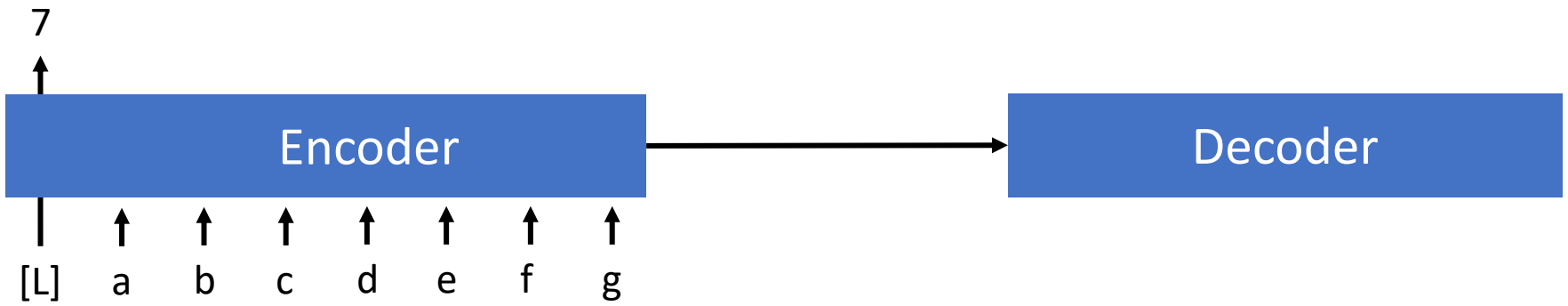
- Маскируем k случайных токенов выходной последовательности, где k из $\mathrm{Uniform}(1, N)$ и предсказываем их
- Целевая функция: кросс-энтропия на маскируемых токенах
- Длину текста предсказываем при помощи токена [LEN] в энкодере



Применение модели

Инициализация:

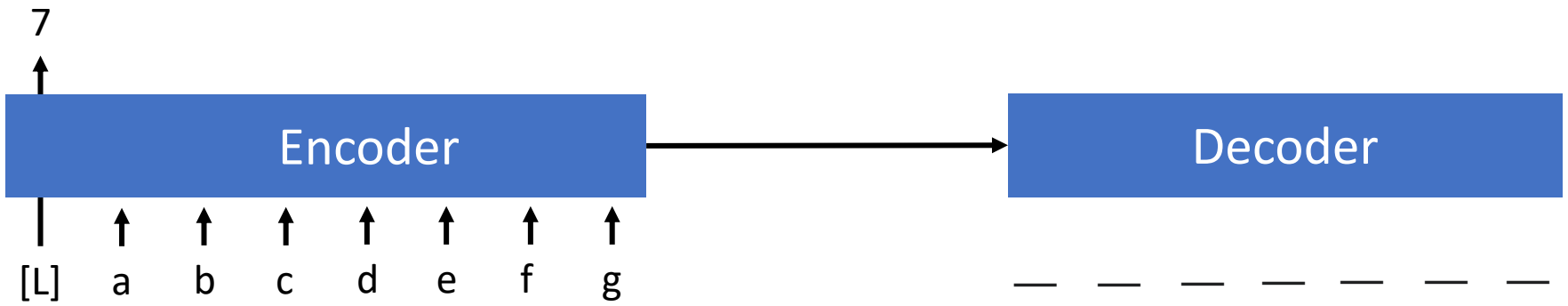
1. Предсказываем длину ответа N



Применение модели

Инициализация:

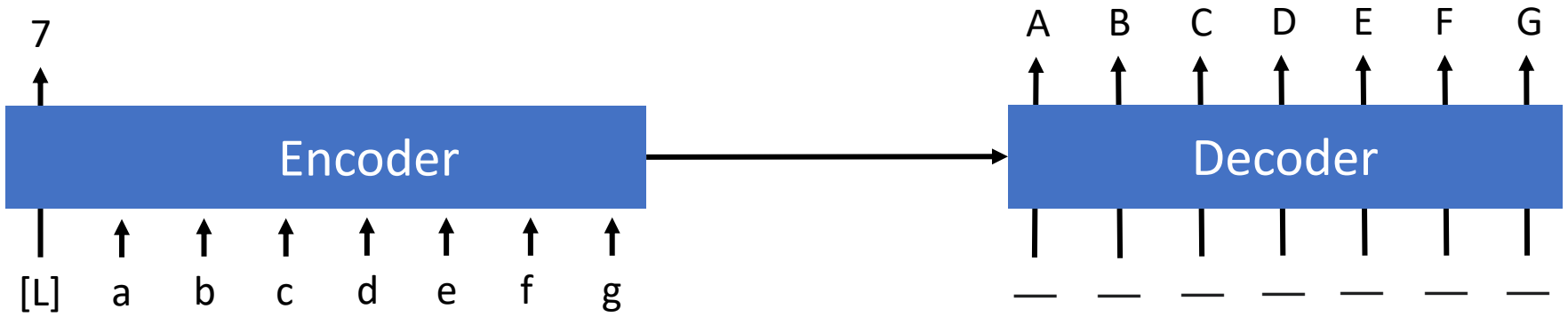
1. Предсказываем длину ответа N
2. Маскируем все токены



Применение модели

Инициализация:

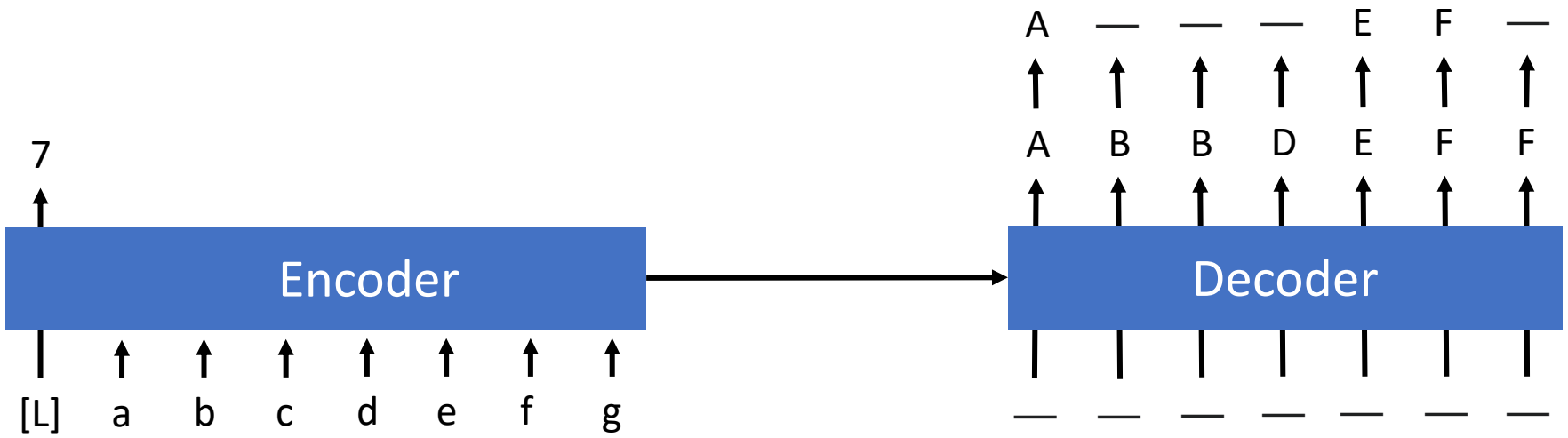
1. Предсказываем длину ответа N
2. Маскируем все токены
3. Предсказываем все токены с применением argmax



Применение модели

Итерация:

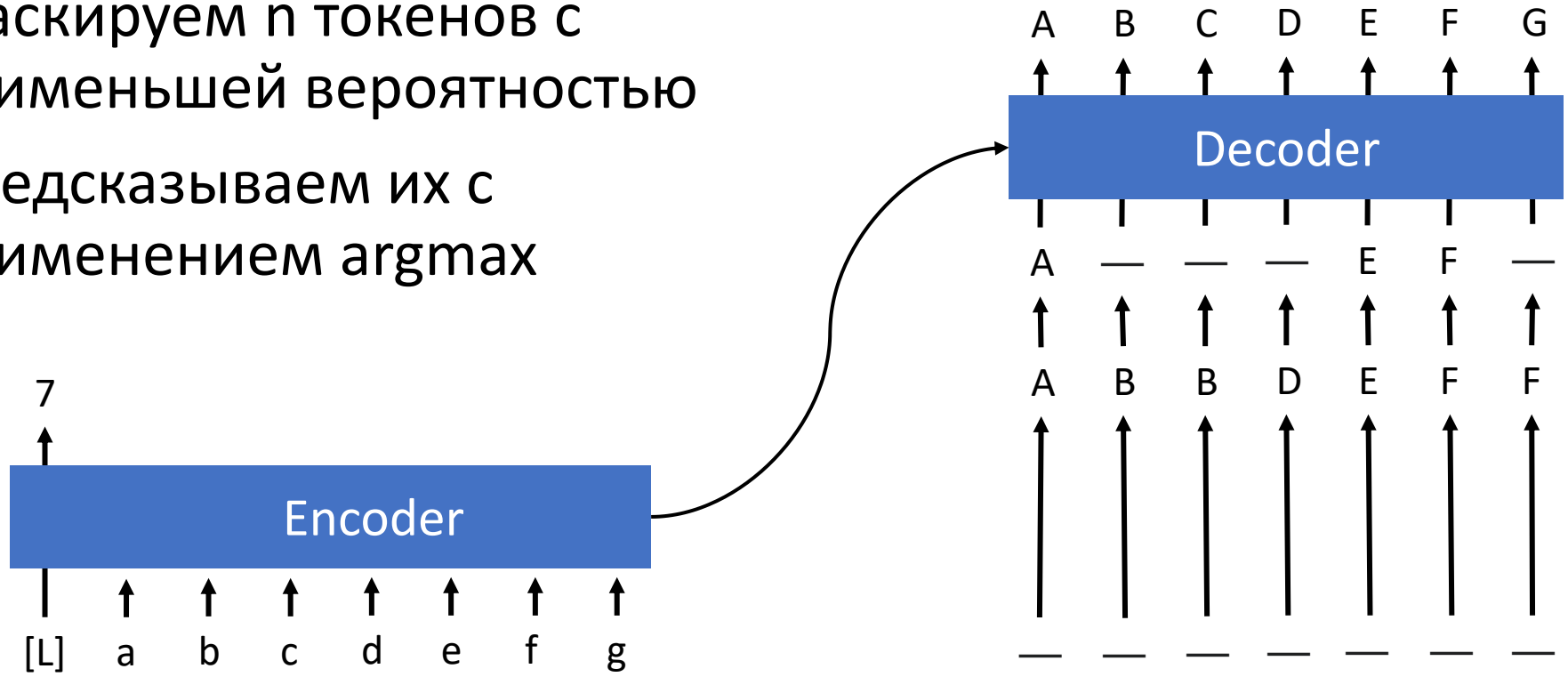
1. Маскируем n токенов с наименьшей вероятностью



Применение модели

Итерация:

1. Маскируем n токенов с наименьшей вероятностью
2. Предсказываем их с применением argmax



Применение модели

Сколько итераций?

- Константное: 1-10
- $\log N, \sqrt{N}, N$

Сколько маскировать?

$$n = \left(1 - \frac{current\ iteration}{total\ iterations}\right) N$$

Пример применения

src	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed on November 20th .

Улучшаем выбор длины

- Так как у нас есть распределение на длинах перевода, можем взять несколько наиболее вероятных из них
- Выбрать наилучшую можно по формуле

$$\frac{1}{N}\sum \log p_i$$

Постановка экспериментов

- Наборы данных: WMT’14 EN-DE (4.5M sentence pairs), WMT’16 EN-RO (610k pairs), WMT’17 EN-ZH (20M pairs) в обоих направлениях
- Токенизация при помощи BPE
- Основная метрика качества: BLEU
- Гиперпараметры модели в основном такие же, как и у трансформера

Постановка экспериментов

- Для обучения используются переводы большого трансформера
- Обучение происходит на батчах размера 128k токенов, 300k шагов
- После каждой эпохи качество замеряется на валидационной выборке
- Итоговая модель: усреднение пяти лучших состояний
- Beam size (авторегрессионная модель): 5
- Кандидатов длины (неавторегрессионная модель): 5

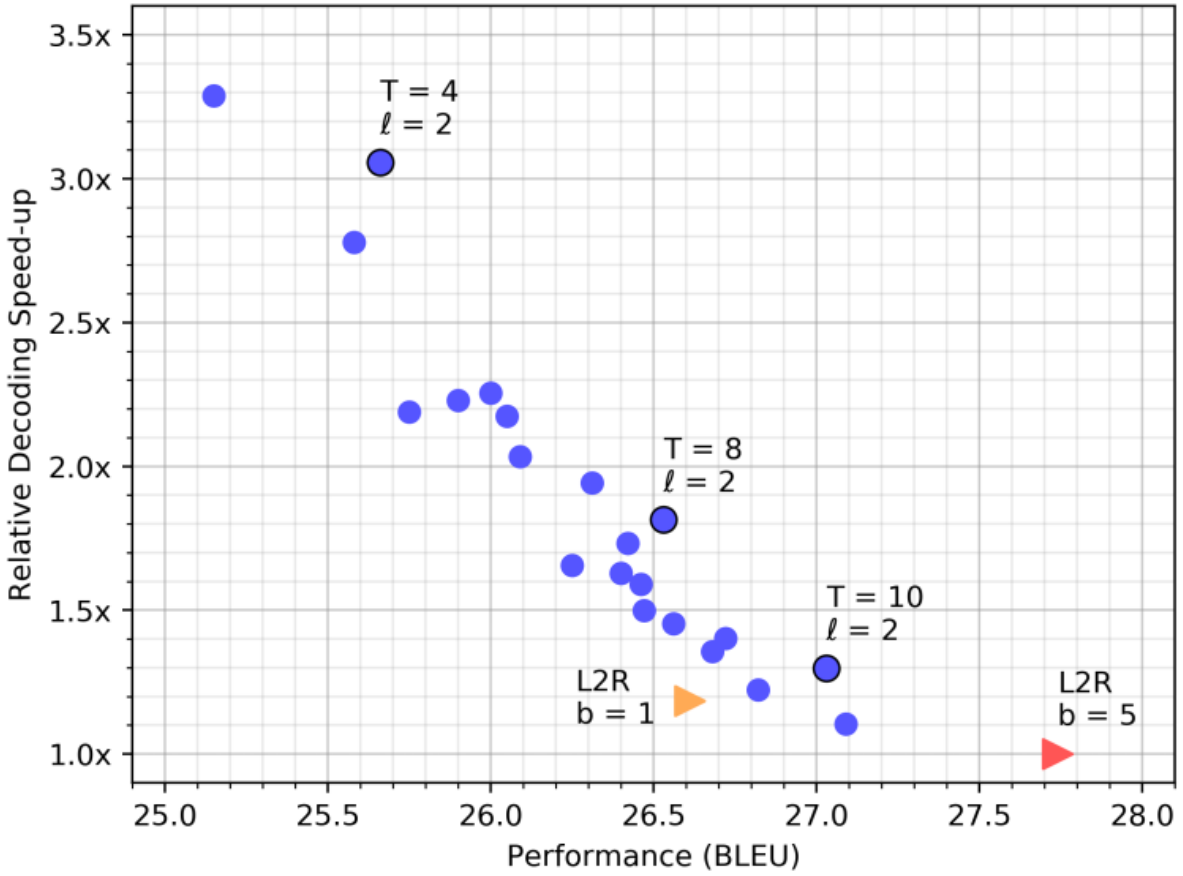
Качество перевода

Model	Dimensions (Model/Hidden)	Iterations	WMT'14 EN-DE	WMT'14 DE-EN	WMT'16 EN-RO	WMT'16 RO-EN
NAT w/ Fertility (Gu et al., 2018)	512/512	1	19.17	23.20	29.79	31.44
CTC Loss (Libovický and Helcl, 2018)	512/4096	1	17.68	19.80	19.93	24.71
Iterative Refinement (Lee et al., 2018)	512/512	1	13.91	16.77	24.45	25.73
	512/512	10	21.61	25.48	29.32	30.19
(Dynamic #Iterations)	512/512	?	21.54	25.43	29.66	30.30
Small CMLM with Mask-Predict	512/512	1	15.06	19.26	20.12	20.36
	512/512	4	**24.17**	**28.55**	**30.00**	30.43
	512/512	10	**25.51**	**29.47**	**31.65**	**32.27**
Base CMLM with Mask-Predict	512/2048	1	18.05	21.83	27.32	28.20
	512/2048	4	**25.94**	**29.90**	**32.53**	**33.23**
	512/2048	10	**27.03**	**30.53**	**33.08**	**33.31**
Base Transformer (Vaswani et al., 2017)	512/2048	N	27.30	— —	— —	— —
Base Transformer (Our Implementation)	512/2048	N	27.74	31.09	34.28	33.99
Base Transformer (+Distillation)	512/2048	N	27.86	31.07	— —	— —
Large Transformer (Vaswani et al., 2017)	1024/4096	N	28.40	— —	— —	— —
Large Transformer (Our Implementation)	1024/4096	N	28.60	31.71	— —	— —

Качество перевода

Model	Dimensions (Model/Hidden)	Iterations	WMT'17 EN-ZH	WMT'17 ZH-EN
Base CMLM with Mask-Predict	512/2048	1	24.23	13.64
	512/2048	4	32.63	21.90
	512/2048	10	33.19	23.21
Base Transformer (Our Implementation)	512/2048	N	34.31	23.74
Base Transformer (+Distillation)	512/2048	N	34.44	23.99
Large Transformer (Our Implementation)	1024/4096	N	35.01	24.65

Скорость перевода



Зачем нужно несколько итераций?

Ответ: чтобы решить проблему мультимодальности

Iterations	WMT'14 EN-DE		WMT'16 EN-RO	
	BLEU	Reps	BLEU	Reps
$T = 1$	18.05	16.72%	27.32	9.34%
$T = 2$	22.91	5.40%	31.08	2.82%
$T = 3$	24.99	2.03%	32.19	1.26%
$T = 4$	25.94	1.07%	32.53	0.87%
$T = 5$	26.30	0.72%	32.62	0.61%

Больше длина — больше итераций?

Ответ: да, но не критично

	$T = 4$	$T = 10$	$T = N$
$1 \leq N < 10$	21.8	22.4	22.4
$10 \leq N < 20$	24.6	25.9	26.0
$20 \leq N < 30$	24.9	26.7	27.1
$30 \leq N < 40$	24.9	26.7	27.6
$40 \leq N$	25.0	27.5	28.1

Больше кандидатов длины — лучше?

Ответ: да, но в меру

Length Candidates	WMT'14 EN-DE BLEU	WMT'14 EN-DE LP	WMT'16 EN-RO BLEU	WMT'16 EN-RO LP
$\ell = 1$	26.56	16.1%	32.75	13.8%
$\ell = 2$	27.03	30.6%	33.06	26.1%
$\ell = 3$	**27.09**	43.1%	**33.11**	39.6%
$\ell = 4$	**27.09**	53.1%	32.13	49.2%
$\ell = 5$	27.03	62.2%	33.08	57.5%
$\ell = 6$	26.91	69.5%	32.91	64.3%
$\ell = 7$	26.71	75.5%	32.75	70.4%
$\ell = 8$	26.59	80.3%	32.50	74.6%
$\ell = 9$	26.42	83.8%	32.09	78.3%
Gold	27.27	—	33.20	—

Необходима ли дистилляция?

Ответ: да

Iterations	WMT'14 EN-DE		WMT'16 EN-RO	
	Raw	Dist	Raw	Dist
$T = 1$	10.64	**18.05**	21.22	**27.32**
$T = 4$	22.25	**25.94**	31.40	**32.53**
$T = 10$	24.61	**27.03**	32.86	**33.08**

Список литературы

- https://arxiv.org/abs/1904.09324
- https://youtu.be/MpuJaNJIVs0