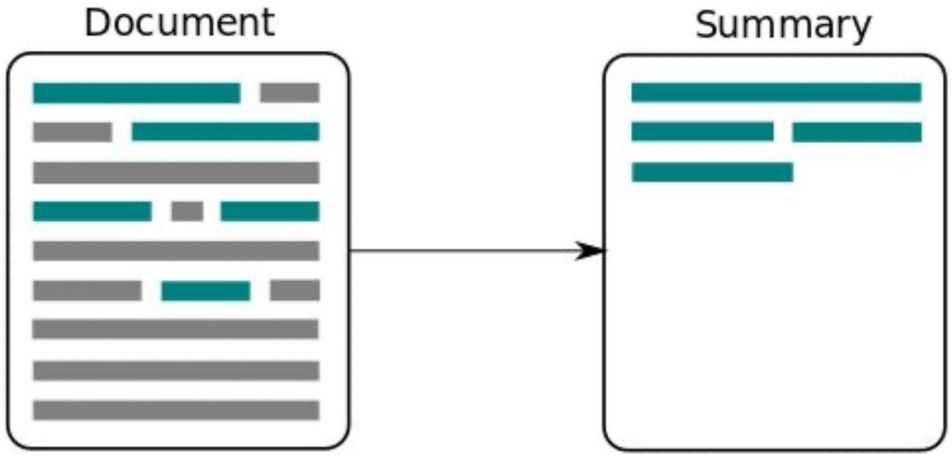
Суммаризация

Albina Akhmetgareeva
R&D NLP, SberDevices, AGI NLP

Задача суммаризации

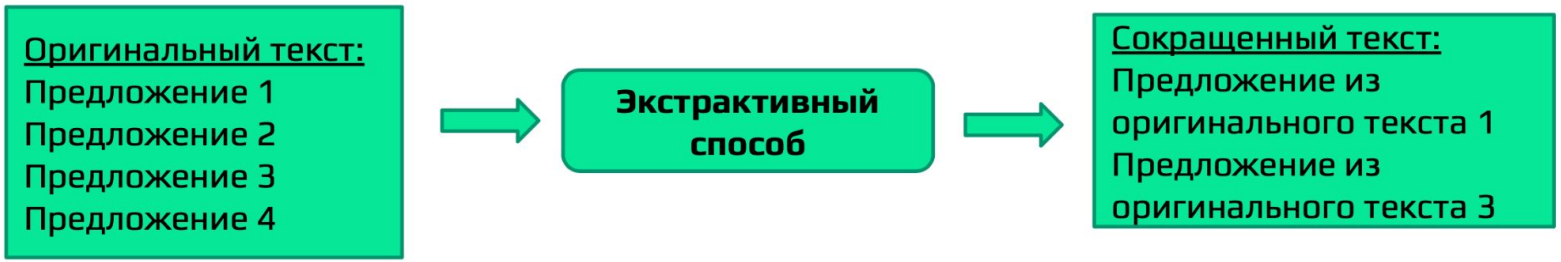
— это задача сокращения фрагмента текста до более короткой версии с одновременным сохранением ключевых информационных элементов и смысла содержания.



Задача суммаризации: пути решения

1. Абстрактивный

2. Экстрактивный



Прикладные задачи

- Создание заголовков новостей
- Реферирование скриптов диалогов
- Выделение знаний из диалогов чатбота
- Реферирование книг / статей (суммаризация длинных документов)
- Аннотирование документов (мультидок суммаризация)

Датасеты на русском

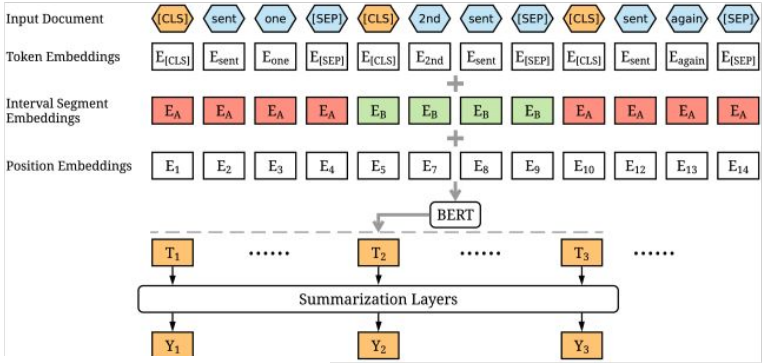
Датасеты	**Мощность**	**Саммари (целевая)**
MLSUM (CNN/Daily mail)	30k	заголовок
GusevGazeta	60k	абстракт
XLSum (ruBBC)	70k	абстракт
ria, lenta	600k	заголовок
wikihow	53k	первые предложения параграфа
yandexjobs	600	заголовок / подпункты
dialsum (перевод)	400	сокращенное описание диалога

Датасеты на английском

- CNN / Daily Mail (single document, many extractive)
- X-Sum (single doc, short summaries)
- Newsroom
- MultiNews (multi documents)
- DUC 2004 Task 1
- SAMSum Corpus
- Webis-TLDR-17 Corpus
- Gigaword
- BIGPATENT

Модели

Экстрактивные: **BertSum [1] (extractive)**

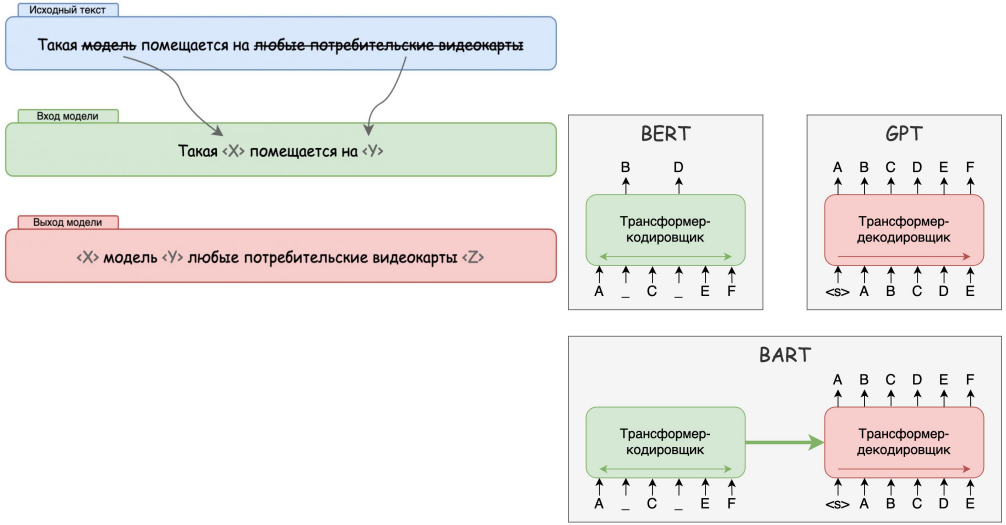


Абстрактивные:

BART [2], T5, Pegasus [3]

GPT-2, GPT-3 (decoders),

InstructGPT [4]



[1] https://arxiv.org/pdf/1908.08345.pdf
[2] https://arxiv.org/pdf/1910.13461.pdf
[3] https://arxiv.org/pdf/1912.08777.pdf
[4] https://arxiv.org/pdf/2203.02155.pdf
Картинки взяты из https://habr.com/ru/articles/596481/

Примеры работы абстрактивной суммаризации

Оригинал	Результат
“3 вещи, которые стоит знать о Платоне I. Его звали не Платон. Не всем известно, но философ был наследником аристократического семейства и получил от родителей имя Аристокл. Платон — это прозвище. Но под ним он стал популярен. II. Юный Платон баловался стихами. Но однажды, когда он нёс в театр свою новую трагедию, Платону попался Сократ. Разговор с Сократом изменил юного поэта: трагедию тот сжёг и ушёл из поэзии в философию. III. Наследие Платона — идеалистическая философская школа: согласно ей, есть мир идей, и он параллелен материальному миру, который мы все знаем. Прав ли был Платон? Философы до сей поры спорят об этом.”	“Три вещи, которые стоит знать о Платоне I. Платона звали не Платон. Его звали Аристокл. Платон — его прозвище. Но однажды он встретил Сократа, спел ему свою песню и ушёл в философию. II. После беседы с Сократом бросил писать стихи и ушёл в философию. III. Философия Платона - идеалистическая: согласно ей, существует мир идей, и он параллелен материальному миру. А вот прав ли был Платон? Философы до сей поры спорят.”
”Мы долго бродили с Касьяном по сечкам. Молодые отпрыски, ещё не успевшие вытянуться выше аршина, окружали своими тонкими, гладкими стебельками почерневшие, низкие пни; круглые губчатые наросты с серыми каймами, те самые наросты, из которых вываривают трут, лепились к этим пням; земляника пускала по ним свои розовые усики; грибы тут же тесно сидели семьями.”	“Касьян показал мне сечки. Вот они. Тихие, тенистые, с некрупными корнями. Под ними росли кусты земляники и грибов.”

Примеры работы экстрактивной суммаризации

Оригинал	Результат
“3 вещи, которые стоит знать о Платоне I. Его звали не Платон. Не всем известно, но философ был наследником аристократического семейства и получил от родителей имя Аристокл. Платон — это прозвище. Но под ним он стал популярен. II. Юный Платон баловался стихами. Но однажды, когда он нёс в театр свою новую трагедию, Платону попался Сократ. Разговор с Сократом изменил юного поэта: трагедию тот сжёг и ушёл из поэзии в философию. III. Наследие Платона — идеалистическая философская школа: согласно ей, есть мир идей, и он параллелен материальному миру, который мы все знаем. Прав ли был Платон? Философы до сей поры спорят об этом.”	"3 вещи, которые стоит знать о Платоне I. Его звали не Платон. Разговор с Сократом изменил юного поэта: трагедию тот сжёг и ушёл из поэзии в философию. Наследие Платона - идеалистическая философская школа: согласно ей, есть мир идей, и он параллелен материальному миру, который мы все знаем."
”Мы долго бродили с Касьяном по сечкам. Молодые отпрыски, ещё не успевшие вытянуться выше аршина, окружали своими тонкими, гладкими стебельками почерневшие, низкие пни; круглые губчатые наросты с серыми каймами, те самые наросты, из которых вываривают трут, лепились к этим пням; земляника пускала по ним свои розовые усики; грибы тут же тесно сидели семьями.”	"Молодые отпрыски, ещё не успевшие вытянуться выше аршина, окружали своими тонкими, гладкими стебельками почерневшие, низкие пни ; круглые губчатые наросты с серыми каймами, те самые наросты, из которых вываривают трут, лепились к этим пням ; земляника пускала по ним свои розовые усики ; грибы тут же тесно сидели семьями."

Метрики

- **bleu** (use n-gram overlap)
- **rouge-n** (use n-gram overlap)
- **bertscore** (use cosine simularity on pre-trained contextual embeddings from BERT)
- **labsescore** (use cosine simularity on pre-trained contextual embeddings from LaBSE)
- **meteor** (harmonic mean of precision and recall on unigrams)
- **chrf** (F-score statistic for character n-gram matches)

Метрики

ROUGE-L: Статистика на основе самой длинной общей подпоследовательности (LCS).

Задача о самой длинной общей подпоследовательности естественным образом учитывает сходство структуры на уровне предложений и автоматически определяет самые длинные совместно встречающиеся в последовательности n-граммы.

ROUGE-N: Overlap of N-grams

ROUGE-1 refers to the overlap of ***unigram*** *(each word)* between the system and reference summaries; ROUGE-2 - of ***bigrams***.

- compute unique ngrams
- check overlap and length
- => F1 mesuare

Recall: $$\frac{|\text{ngrams}(ref) \ \& \ \text{ngrams}(hyp)|}{|\text{ngrams}(ref)|}$$

Precision: $$\frac{|\text{ngrams}(ref) \ \& \ \text{ngrams}(hyp)|}{|\text{ngrams}(hyp)|}$$

F1: $$2\frac{P * R}{R + P}$$

Сравнение методов на англоязычных данных

Model	ROUGE-1	ROUGE-2	ROUGE-L
BRIO (Liu et al., 2022)	49.07	25.59	40.40
PEGASUS (Zhang et al., 2019)	47.21	24.56	39.25
BART (Lewis et al., 2019)	45.14	22.27	37.25
BertSumExtAbs (Liu et al., 2019)	38.81	16.50	31.27

http://nlpprogress.com/english/summarization.html
https://arxiv.org/pdf/2203.16804.pdf

Сравнение методов суммаризации на русском

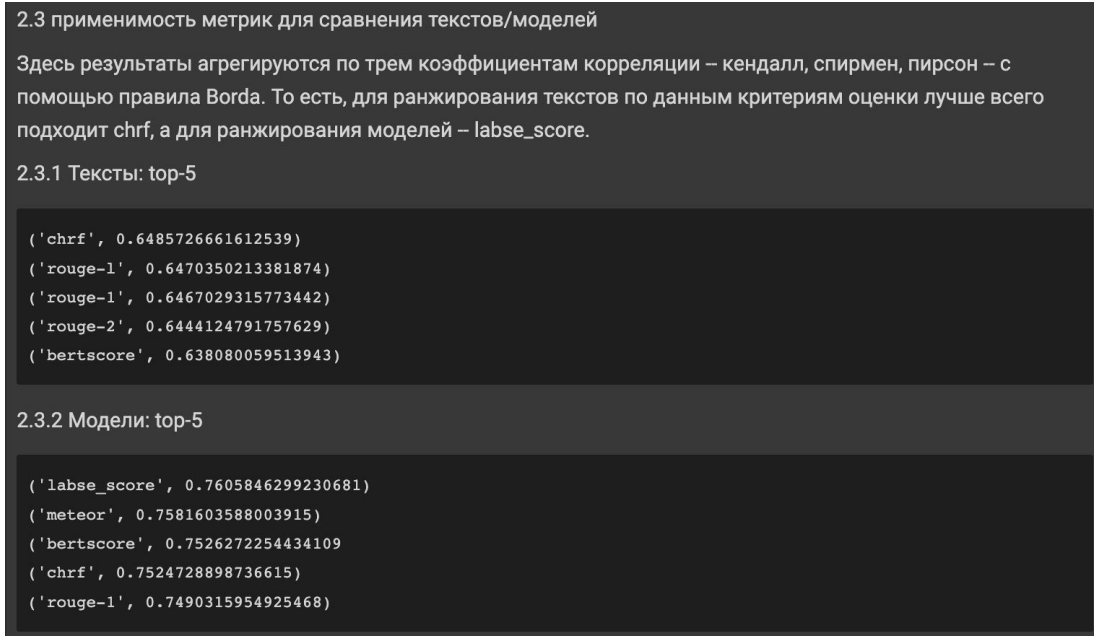
Модель \ метрика	bleu	rouge-1	rouge-2	rouge-l	bertscore	labse	meteor	chrf
Суммаризатор PRO (Sber)	7.124	0.245	0.077	0.229	0.731	0.778	0.289	38.871
Суммаризатор FREE (Sber)	2.003	0.163	0.03	0.151	0.694	0.651	0.155	24.175
csebuetnlp/mT5_multilingual_XLSum	0.188	0.069	0.007	0.064	0.622	0.415	0.039	8.508
UrukHan/t5-russian-summarization	0.847	0.094	0.012	0.086	0.612	0.467	0.067	11.3
cointegrated/rut5-base-absum	0.399	0.102	0.022	0.093	0.669	0.526	0.068	11.128
IlyaGusev/rut5_base_sum_gazeta	2.388	0.17	0.039	0.154	0.691	0.655	0.137	21.773
IlyaGusev/mbart_ru_sum_gazeta	2.605	0.171	0.04	0.153	0.689	0.633	0.14	21.823
gpt3turbo	4.884	0.216	0.056	0.203	0.725	0.754	0.219	33.213

https://300.ya.ru

Выводы по оценке и корреляциям с человеком

Корреляция между автоматическими метриками и критериями человеческой оценки:

- Нет ни одной автоматической метрики, которая бы коррелировала со всеми критериями человеческой оценки
- Surface-based метрики и Bertscore коррелируют с Originality, Relevance и искажением/наличием новых фактов. С grammar ничего не коррелирует.
- Метрики очень сильно коррелируют между собой. Labse меньше других



2.3 применимость метрик для сравнения текстов/моделей

Здесь результаты агрегируются по трем коэффициентам корреляции -- кендалл, спирмен, пирсон -- с помощью правила Borda. То есть, для ранжирования текстов по данным критериям оценки лучше всего подходит chrf, а для ранжирования моделей -- labse_score.

2.3.1 Тексты: top-5

```
('chrf', 0.6485726661612539)
('rouge-1', 0.6470350213381874)
('rouge-1', 0.6467029315773442)
('rouge-2', 0.6444124791757629)
('bertscore', 0.638080059513943)
```

2.3.2 Модели: top-5

```
('labse_score', 0.7605846299230681)
('meteor', 0.7581603588003915)
('bertscore', 0.7526272254434109
('chrf', 0.7524728898736615)
('rouge-1', 0.7490315954925468)
```

Кендалл, спирман, пирсон

Нерешенные проблемы

- мало данных
- модели галлюцинируют и искажают факты (нужны модули факт-чеккинга)
- задача сокращения длинных текстов без потери контекста
- оценка автоматическими метриками мало коррелирует с человеческой
- языковые модели (chatGPT [1]) дают эффективные саммари (измеряемые с помощью человеческой оценки), но эталонные сводки в стандартных наборах данных для суммирования (например, CNN/DM, XSUM) на самом деле хуже (при тех же человеческих оценках) [2]

[1] Training language models to follow instructions with human feedback https://arxiv.org/pdf/2203.02155.pdf
[2] HELM benchmark https://arxiv.org/pdf/2211.09110.pdf

Спасибо за внимание!

Вопросы?

Альбина Ахметгареева
ARAkhmetgareeva@sberbank.ru

SberDevices AGI NLP