Мультидокументная суммаризация: итоги-2024

Саша Коновалова, Алина Тиллабаева, Лиза Клыкова

Прошлый год: цели и результаты

- Датасет для мультидокументной суммаризации на русском языке (944 глав в 67 книгах)
- Сравнение алгоритмов TextRank, Hierarchical и предобученных многоязычных моделей mBART и mT5 на собранных данных

Этот год: цели

Новая метрика оценки качества суммаризации.

Зачем?

- Зависимость существующих метрик от длины саммари и наличия конкретных n-грамм
- Поощряют экстрактивные, а не абстрактивные саммари
- Низкая корреляция с человеческими оценками

Проблемы существующих метрик

Низкая корреляция с человеческими оценками (Wu et al., 2023):

Type	Method	CNN2022	SummEval	BBC2022	AVG
Overlap	ROUGE-1	0.466	0.431	0.469	0.461
	ROUGE-2	0.437	0.354	0.443	0.411
	ROUGE-L	0.422	0.322	0.436	0.393
	BLEU	0.475	0.372	0.502	0.450
	METEOR	0.514	0.473	0.561	0.516
Similarity	BERTScore	0.554	0.455	0.568	0.526
	MoverScore	0.456	0.385	0.442	0.428
LLM	GPT-D3	0.713	0.503	0.692	0.636
	DRPE	**0.816**	**0.683**	**0.784**	**0.761**

Существующие метрики

METEOR (Metric for Evaluation of Translation with Explicit ORdering):

- пересечение униграмм в исходном и сгенерированном текстах с учетом стемминга и т.д. (Lavie & Agarwal, 2007)

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation):

- отношение длины наибольшей общей последовательности (Longest Common Subsequence) к количеству униграмм (Lin & Och, 2004)

BERTScore:

- потокенная близость исходного и сгенерированного текста на основе контекстуальных эмбеддингов (Zhang et al., 2019)

Новое ТЗ

Цель: разработать метрику для оценки качества суммаризации

Проверка: значения полученной метрики коррелируют с экспертной (нашей) оценкой

Результат: код для подсчета метрики на Python 3; золотой стандарт

Новое ТЗ

Must have:

- метрика для оценки качества суммаризации на Python 3
- обоснование метрики и анализ результатов с опорой на human judgement

Should have:

- золотой стандарт: датасет саммари с экспертными оценками

Could have:

- python-библиотека для суммаризации и оценки качества

Золотой стандарт

- Метрика должна отражать некоторую реальность
- Реальность можно оценить с помощью human judgement
- Адекватность метрики = корреляция с экспертными оценками
- Наличие human judgement позволит сравнить и ранжировать существующие метрики

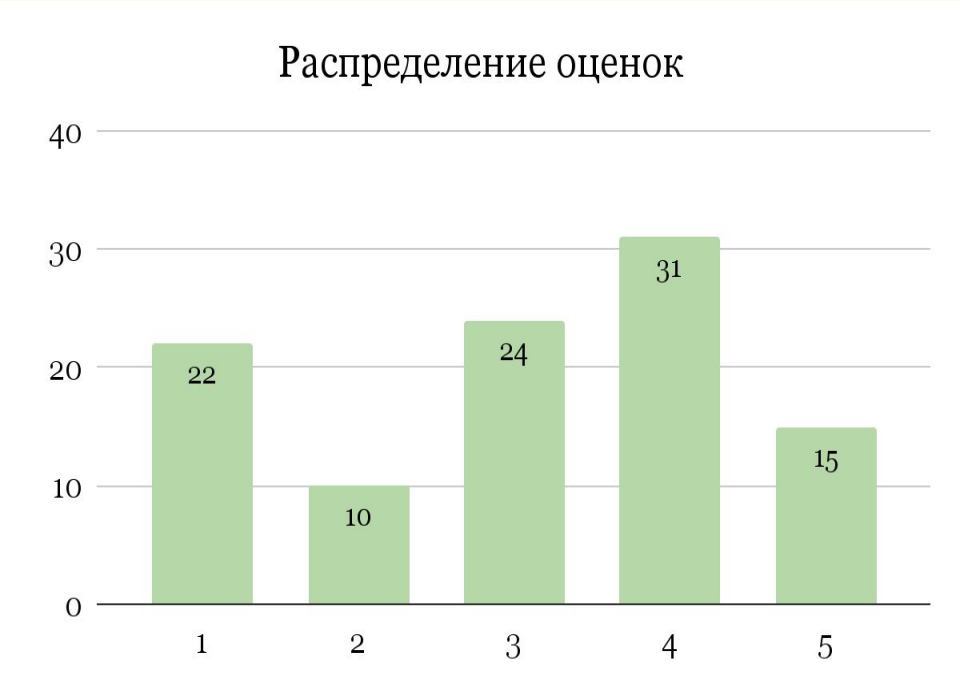
Золотой стандарт

- Оценки саммари 7 глав разных книг:
  - взяли разные по длине главы (0, 0.5 и 1 перцентиль), чтобы уменьшить bias
  - собрали разные варианты саммари, оценили от 1 до 5
  - для каждой главы выбрали одно лучшее саммари
- Используем оценки для сравнения метрик и анализа собственной

Золотой стандарт: результаты

Датасет

102 саммари (~15 на главу)



Влияние свойств саммари на оценку

Коэффициент корреляции Спирмена между экспертной оценкой и:

- Длиной саммари: 0.32
- Отношением длины саммари к длине главы: 0.54
- Кол-вом NE в саммари: 0.28
- Отношением кол-ва NE в саммари к кол-ву NE в главе: 0.57

Почему NER? Важно, чтобы саммари книг содержало ключевые имена собственные, локации и т.д.

Метрики vs. эксперты

Коэффициент корреляции Спирмена между экспертной оценкой и:

- METEOR: 0.56
- ROUGE-L: 0.59
- BERTScore: 0.51

→ результаты сравнимы с получаемыми путем подсчета длины / NE

book_title	chapter_title	human_score	is_best	comments	rouge-l
Гранатовый браслет	Глава 10. Встреча с Г. С. Ж.	3	ЛОЖЬ	слишком много деталей, написано цитатами	0,291
Преступление и наказание	Глава 2: Встреча с Мармеладовым	3	ЛОЖЬ	почти целиком цитаты, слишком длинное	0,258

Корреляция: анализ

Невысокая корреляция существующих метрик с экспертной оценкой – почему так?

- если метрика опирается на n-граммы, то она лучше подходит для extractive, чем для abstractive суммаризации
- наши саммари – скорее abstractive (написаны людьми)

Наша метрика: обоснование

- учитывает общепринятые метрики, показывающие хорошее качество на похожих задачах
- балансирует метрики для abstractive и extractive методов
- оптимизирована для текстов разной длины, так как использует не сами тексты, а соотношения и ключевые слова

Наша метрика: алгоритм

- Извлекаем ключевые слова
- Считаем BERTScore между ключевыми словами из текста и саммари
- Извлекаем именованные сущности и считаем пересечение
- Считаем ROUGE-L и делаем поправку на длину текстов

Наша метрика: алгоритм

$$Bert_{TextRank} + Bert_{Rake} \times \frac{(Ne_{rate} \times RougeL)}{Len_{rate}(Ne_{rate} + RougeL)}$$

- Bert_TextRank – BERTScore на графовых ключевых словах
- Bert_Rake – BERTScore на акцентированных ключевых словах
- Ne_rate – доля найденных именованных сущностей
- Rouge-L – ROUGE-L для наиболее длинной последовательности
- Len_rate – относительная длина в словах

Наша метрика: результаты

Коэффициент корреляции Спирмена между экспертной оценкой и:

- METEOR: 0.56
- ROUGE-L: 0.59
- BERTScore: 0.51
- **Нашей метрикой: 0.64**

Итоги

- Новый датасет (золотой стандарт для будущих задач)
- Сравнение существующих метрик с экспертной оценкой
- Анализ свойств саммари и их влияния на оценку (экспертную и существующие метрики)
- Новая метрика оценки качества суммаризации
- Код для подсчета и анализа метрик

Ссылки

Датасет саммари на Hugging Face

Код для подсчета метрик

Золотой стандарт

Гитхаб проекта

Гугл-папка

ТЗ

Литература

Lavie, A., & Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. https://aclanthology.org/W05-0909/

Lin, C., & Och, F.J. (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. *Annual Meeting of the Association for Computational Linguistics*.

Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D. (2023). Large Language Models are Diverse Role-Players for Summarization Evaluation. *Natural Language Processing and Chinese Computing*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1904.09675