Суммаризация sql-запросов

Алексей Заплатин

Январь 2023

Abstract

В данной работе будут дообучены четыре sequence-to-sequence модели для суммаризации sql-запросов. Две модели: mBart и mT5 будут резюмировать запросы на русском языке, а модели Bart и CodeT5 - на английском языке. Результаты работы доступны по ссылке: `https://github.com/alekz99/summarizationSQL`.

1 Introduction

Суммаризация – это задача создания краткой версии исходного документа, которая содержит его основную мысль. В целом, задача обработки программного кода становится популярной последние несколько лет и крупные компания проводят исследования в этой области, в частности Microsoft и модель CodeBert [Feng et al., 2020], и Salesforce и модель CodeT5 [Wang et al., 2021].

Часть задач, которые можно решать с помощью sequence-to-sequence моделей:

1. документирование программного кода;
2. суммаризация кода;
3. генерация комментариев к коду;
4. генерация кода;
5. проверка кода в среде разработки.

Большая часть исследований в этой области связана с английским языком. В данной работе будут использованы модели для суммаризации sql-запросов на английском и русском языках.

1.1 Team

Данная работа была выполнена студентом университета ИМТО, **Заплатиным Алексеем**.

2 Related Work

В данном разделе будут рассмотрены современные sequence-to-sequence модели, которые можно использовать для суммаризации кода.

Модель Bart – предобученная sequence-to-sequence модель, которая имеет стандартную архитектуру для машинного перевода на основе трансформеров, например, Bert - энкодер и GPT - декодер, [Lewis et al., 2019]. Bart обучается с помощью шумоподавления. Сначала текст искажается с помощью специальной функции "зашумления" данных, а модель, восстанавливая исходный текст, обучается. Способы зашумления данных:

- перемешивание предложений,
- маскирование токенов,
- циклический сдвиг текста,
- удаление токенов.

Модель mBart – расширенная модель Bart, которая предварительно обучена на крупномасштабных одноязычных корпусах на многих языках, включая русский язык [Liu et al., 2020].

Модель T5 – sequence-to-sequence модель, глобально аналогичная BART, разработка Google Research. T5 предобучается на задаче восстановления промежутков, то есть наборов подряд идущих токенов. При обучении промежутки исходного текста скрываются, и задача модели их сгенерировать. В отличие от BART, где текст генерируется целиком, T5 нужно сгенерировать только сами скрытые промежутки [Raffel et al., 2020].

Модель mT5 – многоязычный вариант T5, предварительно обученный на новом наборе данных на основе Common Crawl, охватывающем 101 язык [Xue et al., 2020].

Модель CodeT5 – T5 обученная на данных CodeSearchNet в многоязычной среде обучения(Ruby/JavaScript/Go/Python/Java/PHP) [Wang et al., 2021].

Модель CodeTrans – обученная T5 с использованием различных техник обучения: single-task learning, transfer learning, multi-task learning, and multi-task learning with fine-tuning [Elnaggar et al., 2021].

Для суммаризации на русском языке будут использованы модели mBart и mT5, так как только этим модели из вышеперечисленных поддерживают русский язык. Для суммаризации на английском языке будут использоваться модели Bart и CodeT5.

3 Model Description

Все выбранные модели имеют похожую структуру, которая состоит из кодировщика BERT и декодировщика GPT. Архитектура представлена на рис. 1.

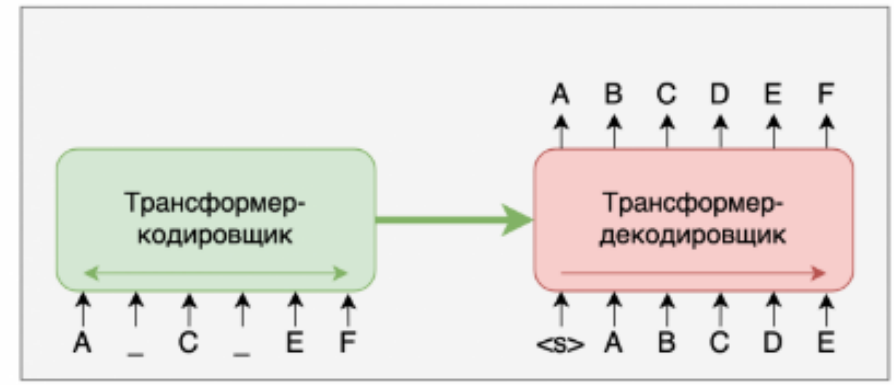

Figure 1: Архитектура seq-to-seq модели

Каждая модель будет дообучена на наборе данных, который будет предварительно предобработан и проанализирован.

4 Dataset

Для обучения модели использовался набор данных pauq[1]. Он состоит из 9876 объектов. Набор данных содержит следующие атрибуты:

1. id – первичный ключ,
2. db_id – первичный ключ базы данных,
3. source – источник объекта,
4. type – train/dev,
5. query – запрос на английском/русском языке,
6. question – вопрос к запросу на английском/русском языке,
7. sql – "разобранный" sql запрос на английском/русском языке,
8. question_toks – токены вопроса на английском/русском языке,
9. query_toks – токены запроса на английском/русском языке,
10. query_toks_no_values – токены запроса с заменой значений на английском/русском языке.

Структура 65.81% sql-запросов представлена на рис. 2.

	SQL structure	Amount of sample	Share of the tota
0	SELECT <ent> FROM <ent> WHERE <ent>	1328	13.70%
1	SELECT <ent> FROM <ent> JOIN <ent> WHERE <ent>	869	8.97%
2	SELECT <ent> FROM <ent>	809	8.35%
3	SELECT <ent> FROM <ent> JOIN <ent> JOIN <ent> ...	493	5.09%
4	SELECT <ent> , <ent> FROM <ent> GROUP BY <ent>	332	3.43%
5	SELECT <ent> , <ent> FROM <ent> JOIN <ent> WHE...	249	2.57%
6	SELECT <ent> FROM <ent> JOIN <ent> JOIN <ent> ...	237	2.45%
7	SELECT <ent> , <ent> FROM <ent> WHERE <ent>	235	2.42%
8	SELECT <ent> FROM <ent> GROUP BY <ent> ORDER B...	221	2.28%
9	SELECT <ent> , <ent> FROM <ent>	216	2.23%
10	SELECT <ent> FROM <ent> WHERE <ent> SELECT <en...	199	2.05%
11	SELECT <ent> FROM <ent> WHERE <ent> IN <ent> S...	190	1.96%
12	SELECT <ent> FROM <ent> GROUP BY <ent> HAVING ...	162	1.67%
13	SELECT <ent> , <ent> FROM <ent> JOIN <ent> GRO...	156	1.61%
14	SELECT <ent> FROM <ent> ORDER BY <ent> DESC LI...	148	1.53%
15	SELECT <ent> FROM <ent> JOIN <ent> GROUP BY <e...	147	1.52%
16	SELECT <ent> FROM <ent> JOIN <ent> GROUP BY <e...	104	1.07%
17	SELECT <ent> FROM <ent> WHERE <ent> INTERSECT ...	96	0.99%
18	SELECT <ent> , <ent> FROM <ent> JOIN <ent> GRO...	95	0.98%
19	SELECT <ent> FROM <ent> ORDER BY <ent>	91	0.94%

Figure 2: Структура 20 самых часто встречаемых sql-запросов

	Длина запроса en	**Длина запроса ru**	**Длина вопроса en**	**Длина вопроса ru**
count	8791	8791	8791	8791
mean	20.28	20.26	13.19	10.61
std	11.22	11.20	4.87	4.17
min	4	4	3	3
25%	11	11	10	8
50%	18	18	13	10
75%	27	27	16	13
max	86	86	45	36

Table 1: Статистики набора данных

Из набора данных были удалены объекты, где вопрос содержит меньше трех слов. Итоговая статистика набора данных представлена в табл. 1.

Для обучения были выбраны четыре атрибута набора данных: query_toks_en, question_en, query_toks_ru, question_ru. query_toks_en и query_toks_ru были объединены в строку, так как изначально это массивы с токенами.

Для обучения и оценки модели набор данных был разделен на три части. В табл. 2 представлено разделение.

	Train	**Valid**	**Test**
Количество объектов	7794	1001	1077

Table 2: Разделение набора данных

5 Experiments

5.1 Metrics

Для оценки полученных результатов будет использоваться метрика сглаженная BLEU-4, которая распространена для оценки задач суммаризации и машинного перевода.

BLEU использует измененную форму точности для сопоставления перевода кандидата с несколькими ссылочными переводами. Метрика изменяет простую точность, поскольку системы машинного перевода, как известно, генерируют больше слов, чем в справочном тексте.

5.2 Experiment Setup

Параметры обучения моделей представлены в табл. 3. Для обучения использовался бесплатный сервис Colab. Размер батча зависит от размера модели и ограничений сервиса по объему доступной оперативной памяти.

Был проведен ряд экспериментов, в ходе которых были выбраны лучшие модели. Из-за небольшого количества данных модели быстро переобучались, поэтому обучение происходило 1-2 эпохи.

6 Results

После обучения моделей были сгенерированы 1077 вопросов на тестовом наборе данных для каждой модели, которые доступны в репозитории. Значения метрики BLEU представлены в табл. 4.

По полученным результатам видно, что модель mBart лучше справляется с суммаризацией запросов на русском языке, чем mT5.

[1] https://github.com/ai-spiderweb/pauq

	mbart-large-cc25	mt5-base	bart-large	codet5-base-multi-sum
Язык запроса	en	en	en	en
Язык вопроса	ru	ru	en	en
Токен энкодера	Да	Да	Нет	Нет
Токен кодировщика	Да	Да	Нет	Нет
Макс. длина запроса	120	120	120	120
Макс. длина вопроса	80	80	80	80
Размер батча	16	10	24	24
Warmup step	Нет	Нет	100	100
Количество эпох обучения	1	2	1	1

Table 3: Параметры обучения моделей

Качество у обеих моделей невысокое. Модели Bart и CodeT5 показали высокое качество суммаризации на английском языке. В данной работе гипотеза о том, что обученная T5 на языках программирования (SQL не было в наборе данных) покажет лучшее качество, чем Bart, не подтвердилась. Полученное качество на английском языке сопоставимо с результатами работы [Elnaggar et al., 2021], где на наборе данных со Stack Overflow был получен результат 19.98.

	mbart-large-cc25	mt5-base	bart-large	codet5-base-multi-sum
BLEU	8.8664	5.0035	25.053	23.0288

Table 4: Значения сглаженной метрики BLEU-4

В табл. 5 представлены оригинальные и сгенерированные вопросы на русском языке, а в табл. 6 - на английском.

Оригинальный запрос	Оригинальный вопрос	mbart-large-cc25	mt5-base
select count (*) from singer ;	Сколько у нас певцов?	Какое общее количество исполнителей?	Найдите количество исполнителей, у которых есть песни.
select location , name from stadium where capacity between 5000 and 10000 ;	Каково расположение и названия всех станций вместимостью от 5000 до 10000?	Найдите местоположения и названия стадионов с вместимостью от 5000 до 10000 человек.	Найдите место и названия стадионов, в которых есть вместимость более 5000 и 10000.
select country from singer where age >40 intersect select country from singer where age <30 ;	Показать страны происхождения певца старше 40 лет и певца младше 30 лет.	Найдите страны, в которых есть исполнители старше 40 лет и исполнители младше 30 лет.	Каковы страны, в которых есть певцы, которые более 40 лет?
select count (*) from pets where weight >10 ;	Найдите количество домашних животных, вес которых превышает 10.	Найдите количество домашних животных весом более 10 кг.	Найдите количество pets, вес которых менее 10 кг.
select t1.fname from student as t1 join has_pet as t2 on t1.stuid = t2.stuid join pets as t3 on t3.petid = t2.petid where t3.pettype = "черепаха" intersect select t1.fname from student as t1 join has_pet as t2 on t1.stuid = t2.stuid join pets as t3 on t3.petid = t2.petid where t3.pettype = "попугай" ;	Как зовут учеников, у которых в качестве домашних животных есть и черепахи, и попугаи?	Найдите имена студентов, у которых есть домашние животные, которые могут быть как кровавыми, так и лошадиными.	Найдите имена студентов, у которых есть черепаха.

Table 5: Примеры оригинальных и сгенерированных вопросов на русском языке

Оригинальный запрос	Оригинальный вопрос	bart-large	codet5-base-multi-sum
select count (*) from singer ;	How many singers do we have?	How many singers are there?	How many singers are there?
select location , name from stadium where capacity between 5000 and 10000 ;	What are the locations and names of all stations with capacity between 5000 and 10000?	What are the locations and names of stadiums with capacity between 5000 and 10000?	What are the locations and names of all stadiums with capacity between 5000 and 10000?
select country from singer where age >40 intersect select country from singer where age <30 ;	Show countries where a singer above age 40 and a singer below 30 are from.	What are the countries that have both singers older than 40 and singers younger than 30?	What are the countries that have more than 40 and less than 30 singers?
select count (*) from pets where weight >10 ;	Find the number of pets whose weight is heavier than 10.	How many pets have a weight greater than 10?	How many pets have weight greater than 10?
select t1.fname from student as t1 join has_pet as t2 on t1.stuid = t2.stuid join pets as t3 on t3.petid = t2.petid where t3.pettype = "cat" intersect select t1.fname from student as t1 join has_pet as t2 on t1.stuid = t2.stuid join pets as t3 on t3.petid = t2.petid where t3.pettype = "dog" ;	What are the students' first names who have both cats and dogs as pets?	What are the first names of students who have both cats and dogs?	What are the first names of students who have cat or dog pet?

Table 6: Примеры оригинальных и сгенерированных вопросов на английском языке

7 Conclusion

В данной работе были рассмотрены существующие sequence-to-seqence модели. Были дообучены четыре модели для суммаризации sql-запросов

на русском и английском языках. Для дообучения моделей использовался набор данных pauq, который состоит из запросов на языке SQL и вопросов-пояснений на английском и русском языке.

Модель mBart показала лучшее качество на русском языке, но само качество модели невысокое. Модель Bart показала лучшее качество суммаризации запросов на английском языке, кроме того, полученное качество модели является высоким.

References

[Elnaggar et al., 2021] Elnaggar, A., Ding, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Severini, S., Matthes, F., and Rost, B. (2021). Codetrans: Towards cracking the language of silicon's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2104.02443*.

[Feng et al., 2020] Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., et al. (2020). Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155*.

[Lewis et al., 2019] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[Liu et al., 2020] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

[Raffel et al., 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

[Wang et al., 2021] Wang, Y., Wang, W., Joty, S., and Hoi, S. C. (2021). Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859*.

[Xue et al., 2020] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.