Министерство Науки и Высшего Образования Российской Федерации

Национальный Исследовательский Институт Высшая Школа Экономики

Факультет Компьютерных Наук

Школа Анализа Данных и Искусственного Интеллекта

РЕФЕРИРОВАНИЕ ТЕКСТА НА РУССКОМ ЯЗЫКЕ

Компьютерная лингвистика и анализ текстов

Студент *М.Д. Кирдин*

Преподаватель *Е.И. Большакова*

Москва, 2025г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Автоматическое реферирование текста является одной из основополагающих задач обработки естественного языка наряду с машинным переводом и распознаванием сущностей. Способы решения этой задачи делятся на две категории: извлекающие и генерирующие. Целью данной работы было провести сравнение этих подходов к решению задачи аннотирования текста. Извлекающие подходы были представлены алгоритмом *TextRank*, а генерирующие подходы — моделями с трасформерной архитектурой *FRED-T5-Summarize* (1.4 миллиарда параметров), а также *rut5-base* (246 миллионов параметров) с параметрами отрегулированными для решения задачи реферирования текстов на русском языке.

1 ХОД РАБОТЫ

Для сравнения двух различных подходов было решено использовать специализированный датасет, предложенный Ахметгареевой А. и др. [1]. Он состоит из 197 тыс. текстов в части предназначенной для обучения и 258 текстов проверенных вручную в части для тестов. Были взяты 40 первых примеров из тестового набора, аннотации, полученные в результате работы сравниваемых алгоритмов, были сохранены в виде текстовых файлов.

1.1 Реализация извлекающего алгоритма

Алгоритм *TextRank* является модификацией алгоритма *PageRank*, предложенного *Google* в 1998 году. В данной работе используется вариант данного алгоритма для извлечения предложений. Он основан на построении графа при помощи алгоритма *PageRank*, в котором вершинами являются предложения в тексте и извлечении n вершин с наибольшим значением внутренней метрики. В рамках данной работы был написан скрипт на языке *Python* с его реализацией, извлекающий ровно **3 лучших предложения**.

Для построения графа необходима матрица сходств предложений в реферируемом тексте. Она была получена как набор попарных косинусных расстояний между суммами эмбеддингов отдельных токенов. Эмбеддинги и токенизатор были взяты из библиотеки *SpaCy*.

1.2 Реализация генерирующих алгоритмов

Был написан скрипт на языке *Python*, который с помощью моделей *FRED-T5-Summarize* и *rut5-base* генерирует реферат текста. У последней, меньшей по размеру модели, параметры были отрегулированы(т.е. модель бы ла дообучена) на обучающей части выбранного датасета. Было рассмотрено несколько вариантов генерации ответов:

- с префиксами на английском и декодированием с помощью лучевого поиска;
- с префиксами на русском и декодированием с помощью лучевого поиска;
- с префиксами на английском и декодированием с помощью *promt lookup*;
- с префиксами на русском и декодированием с помощью *promt lookup*.

Со значениям параметров декодирующего слоя можно ознакомиться в таблице А.1. Такой выбор вариантов обусловлен тем, что некоторые модели (например, *rut5-base*) являются получены обрезанием многоязычных версий моделей, обученных преимущественно на данных на английском языке, вследствие чего модели могут генерировать лучшие результаты при использовании английских токенов.

РЕЗУЛЬТАТЫ

Выбранный извлекающий алгоритм не может производить новые данные, а также использует предложения полностью вследствие чего в некоторых случаях длинные предложения (например, с объемными перечислениями) извлекаются из текста неизменными, что отрицательно влияет на качество аннотации. Например:

Исходный текст	Аннотация
Мировой опыт свидетельствует, что для динамичного развития туризма необходимы следующие условия: *– стабильная социально-экономическая ситуация (в мире в целом, в отдельной стране и конкретном регионе);* *– отсутствие административно-чиновничьих барьеров при перемещениях через границы и в период гостевого пребывания;* *– притягательные рекреационные ресурсы (природно-климатические и культурно-исторические);* *– развитая инфраструктура туризма и квалифицированные кадры;* *– высокий уровень сервиса, обеспечение комфортного проживания, гостеприимство, культура и профессионализм персонала;* *– комфортабельный и безопасный транспорт, надежная связь;* *– свобода перемещения и гарантии прав путешествующих, обеспечение их безопасности;* *– высокая ответственность туристских организаций и их структурных подразделений за проведение конкретных туров;* *– положительный туристский имидж территории, высокая репутация обслуживающих туристов фирм и компаний.*	*Мировой опыт свидетельствует, что для динамичного развития туризма необходимы следующие условия:* *– стабильная социально-экономическая ситуация (в мире в целом, в отдельной стране и конкретном регионе);* *– отсутствие административно-чиновничьих барьеров при перемещениях через границы и в период гостевого пребывания;* *– притягательные рекреационные ресурсы (природно-климатические и культурно-исторические);* *– развитая инфраструктура туризма и квалифицированные кадры;* *– высокий уровень сервиса, обеспечение комфортного проживания, гостеприимство, культура и профессионализм персонала;* *– комфортабельный и безопасный транспорт, надежная связь;* *– свобода перемещения и гарантии прав путешествующих, обеспечение их безопасности;* *– высокая ответственность туристских организаций и их структурных подразделений за проведение конкретных туров;* *– положительный туристский имидж территории, высокая репутация обслуживающих туристов фирм и компаний.*

Однако в случае если текст достаточно объемный либо состоит из простых предложений качество реферирования значительно лучше:

Исходный текст	Аннотация
Тануки - традиционные японские звери-оборотни, символизирующие счастье и благополучие, обычно выглядящие как енотовидные собаки. Второй по популярности зверь-оборотень после кицунэ. В отличие от кицунэ, образ тануки практически лишен негативной окраски. Считается, что тануки — большие любители саке. Поэтому без его присутствия нельзя сделать хорошего сакэ. По этой же причине фигурки тануки, порой весьма большие, являются украшением многих питейных заведений. Они изображают тануки толстяком-добряком с заметным брюшком.	*Тануки - традиционные японские звери-оборотни, символизирующие счастье и благополучие, обычно выглядящие как енотовидные собаки. В отличие от кицунэ, образ тануки практически лишен негативной окраски. По этой же причине фигурки тануки, порой весьма большие, являются украшением многих питейных заведений.*

Качество аннотаций произведенных генерирующими алгоритмами было оценено с помощью метрики *ROUGE* относительно предложенных в датасете аннотаций текста. Согласно таблице 1, в целом лучевой поиск с русскими префиксами дает лучшие результаты при реферировании текста. При этом, модель *FRED-T5-Summarize* дает значительно лучшие результаты, чем модель *rut5-base*, что может быть обусловлено их разницей в размере.

Стоит заметить, что у обеих моделях во всех рассмотренных случаев со всеми рассмотренными метриками *precision* больше, чем *recall*. Это может быть объяснено тем, что сами сгенерированные аннотации короче предложенных в датасете аналогов.

Таблица 1. Значения метрик *ROUGE* для моделей *FRED-T5-Summarize* и *rut5-base* с отрегулированными параметрами

Метрика		*FRED-T5-Summarize*				*rut5-base*			
		ru, lookup	*en, lookup*	*ru, beam*	*en, beam*	*ru, lookup*	*en, lookup*	*ru, beam*	*en, beam*
rouge-1	*recall*	0.21	0.12	0.38	0.13	0.06	0.09	0.15	0.09
	precision	0.63	0.56	0.88	0.52	0.55	0.56	0.75	0.53
	f1-score	0.29	0.19	0.50	0.20	0.11	0.15	0.25	0.15
rouge-2	*recall*	0.11	0.06	0.32	0.07	0.03	0.04	0.09	0.04
	precision	0.37	0.32	0.79	0.29	0.29	0.26	0.54	0.27
	f1-score	0.15	0.09	0.43	0.10	0.05	0.05	0.15	0.07
rouge-l	*recall*	0.20	0.12	0.37	0.12	0.06	0.08	0.15	0.09
	precision	0.60	0.54	0.88	0.50	0.53	0.53	0.74	0.52
	f1-score	0.27	0.18	0.50	0.19	0.10	0.14	0.24	0.14

Рассмотрим некоторые рефераты, созданные рассматриваемыми моделями:

Исходная аннотация	Аннотация FRED-T5-Summarize	Аннотация rut5-base
Готский язык - это древний германский язык, использовавшийся готами с I по VIII век. Он известен благодаря письменным памятникам IV-VI веков, включая перевод Библии, сделанный епископом Вульфилой, который также создал готский алфавит. Готы взаимодействовали с Римской империей и ассимилировались с местным населением, что привело к утрате их языка. Славянские языки, включая русский, заимствовали некоторые слова из готского языка.	*Готский язык - это один из самых древних германских языков. Он был языком группы племен, которые обитали в I веке нашей эры по нижнему течению Вислы и Северном Причерноморье до VIII века на территории Восточной Римской империи (Византии). Славянские языки многое заимствовали оттуда: так русское слово «хлеб» — hlaifs является готским словом "hleaf".*	*В Европе готы изучали язык, который был основанным в первой половине нашей эры.*

Легко заметить, что генерируемые аннотации действительно короче тех, что предложены в самом датасете. Это может быть связано с ограниченным числом генерируемых токенов (512 штук). При этом также стоит обратить внимание на то, что, судя по аннотации *rut5-base*, генерирующие модели подвержены галлюцинациям.

Важно, что создание аннотаций с помощью генеративных алгоритмов занимало на порядок больше времени, чем с помощью извлекающего — 10 и 3 минуты на 40 текстов

для *FRED-T5-Summarize* и *rut5-base* соответственно и 2 минуты для алгоритма *PageRank*.

ЗАКЛЮЧЕНИЕ

Извлекающий алгормитм реферирования *TextRank* в отличие от генерирующих алгоритмов требует гораздо меньше вычислительных мощностей и не подвержен галлюцинациям, но при этом в силу того что он работает с целыми предложениями он лучше всего работает с текстами содержащими большое количество простых предложений. При этом генерирующие модели могут создавать новые тексты и обобщать информацию, что делает результирующие аннотации короче и содержательнее.

СПИСОК ИСТОЧНИКОВ

1. Akhmetgareeva A., Kuleshov I., Leschuk V., Abramov A., Fenogenova A., Towards Russian Summarization: can architecture solve data limitations problems? // https://sberlabs.com/publications?publication=1600 (2024).

ПРИЛОЖЕНИЕ А

Таблица А.1 Значения параметров декодирующего слоя

Параметр	Лучевой поиск	*Promt Lookup*
количество лучей	4	-
repetition_penalty	10.0	1.5
length_penalty	2.0	-
модель-помощник	-	Модель без дообучения
length_penalty	2.0	-
температура	-	0.4