Фундаментальная информатика и информационные технологии

АВТОМАТИЧЕСКОЕ РЕФЕРИРОВАНИЕ ТЕКСТОВ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Выпускная квалификационная работа на степень бакалавра

Выполнил:
Северин Никита Николаевич

Научный руководитель:
старший преподаватель, к.ф.-м.н.
Юрушкин Михаил Викторович

Цель работы

Построение *экстрактивной* модели автоматического реферирования текстов русского языка для создания рефератов *смешанного* типа.

Задачи:

– сравнительный анализ методов решения задачи автоматического реферирования текстов

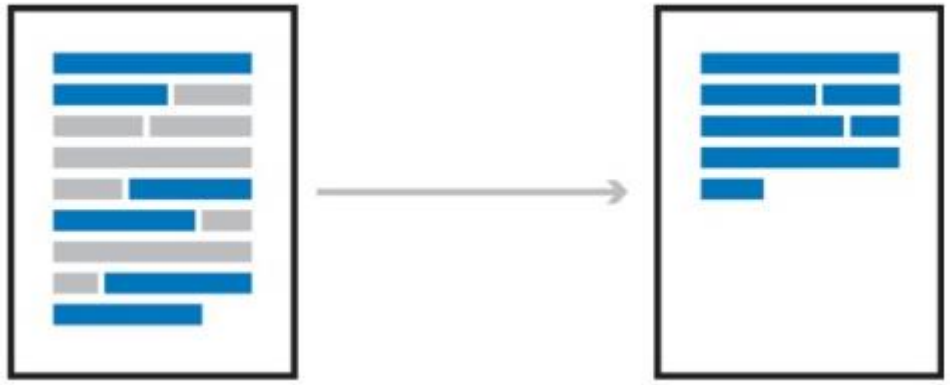
– анализ существующих подходов к извлечению признаков из текста

– разработка алгоритмов реферирования текстов на основе графовых методов и кластеризации предложений

– анализ качества работы полученных моделей

Обобщенный алгоритм экстрактивного реферирования

1) предобработка исходного текста
2) извлечение признаков из каждого предложения
3) выделение наиболее важных предложений



Методы извлечения признаков из текста

Модель мешка слов + TF-IDF

Модель BERT для русского языка (RuBERT)

Модель FastText

Модель Sentence-BERT для русского языка (RuSBERT)

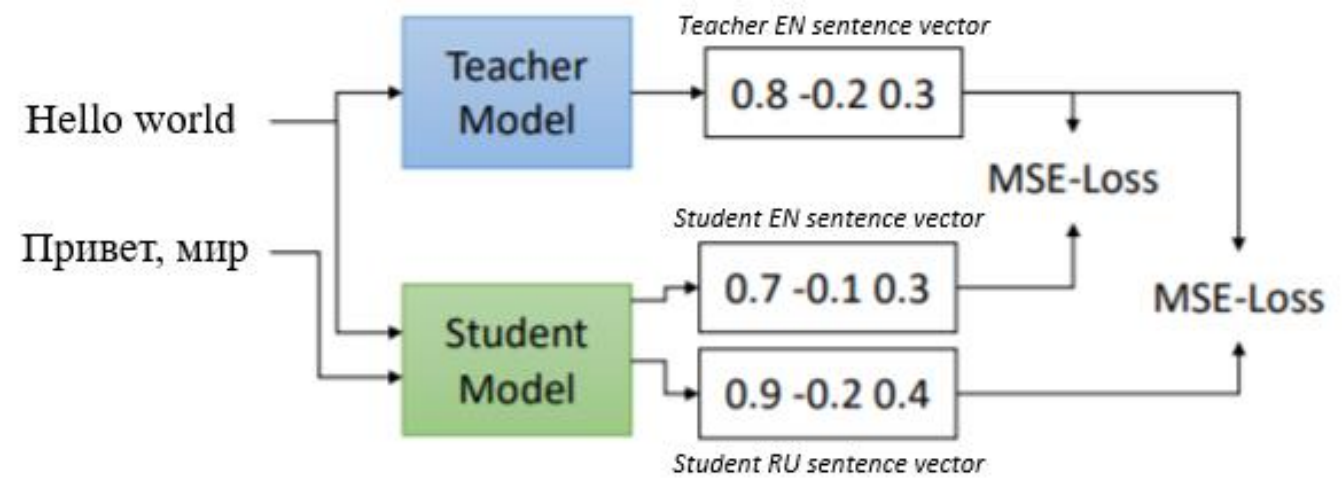
Модель CustomSBERT

Модель CustomSBERT

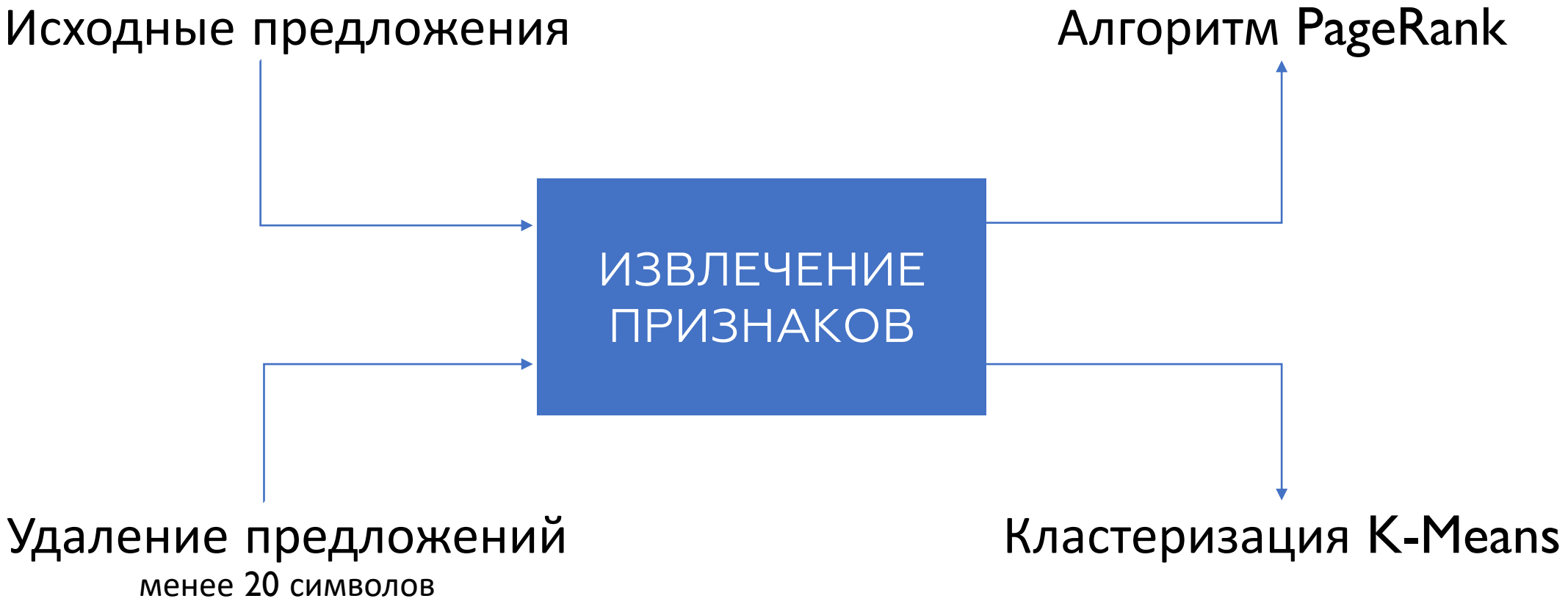
В основе: мультиязычная модель XLM-RoBERTa.

Модель-учитель: Sentence-BERT для английского языка (EnSBERT).

Материал для обучения: англо-русский параллельный корпус Яндекса.



Итоговые алгоритмы реферирования



Оценивание качества

Тестовый набор **4200** **научных** текстов:

- естественнонаучные, социально-экономические
- гуманитарные
- технические

ROUGE-1

совпадение униграмм

ROUGE-2

совпадение биграмм

ROUGE-L

Наибольшая подпоследовательность слов

Экспертное оценивание: критерии

Шкала оценивания: от 0 до 5

- отражение основных смысловых частей исходного текста
- логическая согласованность
- понятность

Предварительные выводы

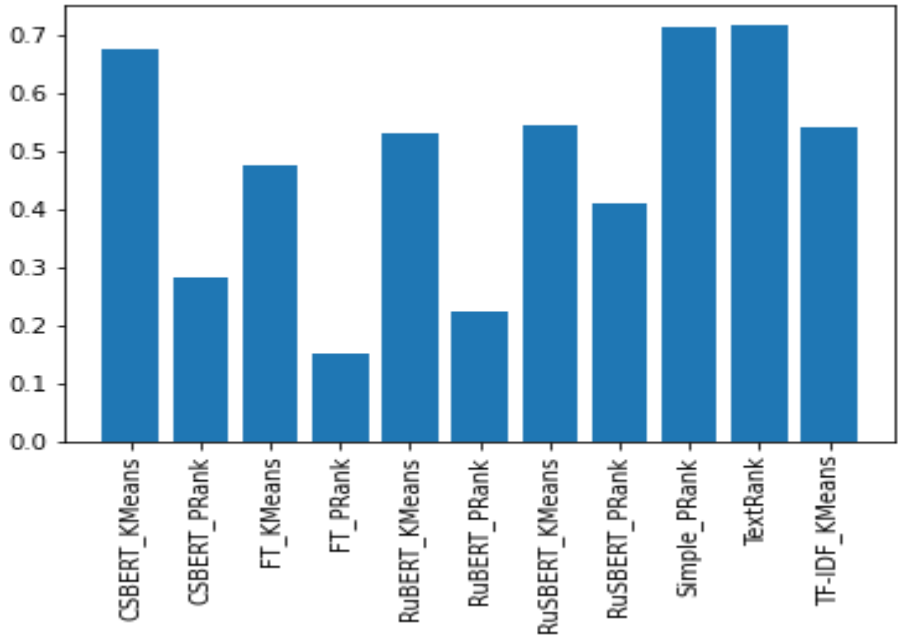
Лучшие алгоритмы:

графовый – **TextRank**

на основе
кластеризации –

CustomSBERT_KMeans

Нормированные экспертные оценки



“В то же время число абонентов огромно, поэтому не представляется возможным хранение индивидуального варианта меню для каждого пользователя. Группировка пользователей в **кластеры** *позволяет оптимизировать интерфейс для каждой конкретной группы, при этом избежав больших накладных расходов для хранения индивидуального дерева меню. В данной статье предлагается использовать информацию о контексте (или окружении) пользователя. Для каждой группы происходит выборка всех сессий, которые были открыты пользователями данного* **кластера**. *Где K – количество пользователей* **кластера**, `n_(j)` — *количество запросов, совершаемых пользователем в оптимизированном меню, `m_(j)` — количество запросов, совершаемых пользователем в первоначальном меню. Также был разработан метод, позволяющий адаптировать меню на основе данных, собранных в процессе работы пользователей с сервисом.*

(TextRank)

“Наиболее важными из них являются: точность предлагаемых изменений относительно реакции пользователя, степень изменения и **предсказуемости интерфейса**, *частота выполняемых адаптаций. Правильный выбор модели является крайне важным этапом разработки алгоритма, так как хранение лишней информации ресурсозатратно, а отсутствие какой-либо информации чревато потерей времени. Поскольку данная работа посвящена разработке метода автоматической адаптации* **USSD-меню**, *стоит немного рассказать о данной технологии. На вход* **нейронной сети** *подаются данные о* **кластеризуемом** *объекте. Этого должно хватить, поскольку пользователи открывшие данные контексты имеют схожие группы интересов, следовательно, имеют схожие контексты выполнения. Этот процесс выполняется для каждого* ***кластера*** *полученного на первом этапе (см. рис. 2). Каждой вершине присвоим вес `w_(i)` , определим это число следующим образом.*

(CustomSBERT_Kmeans)

Алгоритм «Mixed»

CustomSBERT_Kmeans

+ TextRank

“В данной статье предлагается использовать информацию о контексте (или окружении) пользователя.

Наиболее важными из них являются: точность предлагаемых изменений относительно реакции пользователя, степень изменения и предсказуемости **интерфейса**, *частота выполняемых адаптаций.*

Каждый алгоритм адаптации так или иначе использует математическую модель пользователя.

Правильный выбор модели является крайне важным этапом разработки алгоритма, так как хранение лишней информации ресурсозатратно, а отсутствие какой-либо информации чревато потерей времени.

Поскольку данная работа посвящена разработке метода автоматической адаптации **USSD-меню**, *стоит немного рассказать о данной технологии.*

На вход **нейронной сети** *подаются данные о* **кластеризуемом** *объекте.*

Этого должно хватить, поскольку пользователи открывшие данные контексты имеют схожие группы интересов, следовательно, имеют схожие контексты выполнения.

Этот процесс выполняется для каждого **кластера** *полученного на первом этапе (см. рис. 2).*

Каждой вершине присвоим вес `w_(i)` , определим это число следующим образом.

Также был разработан метод, позволяющий адаптировать **меню** *на основе данных, собранных в процессе работы пользователей с сервисом.*

(Mixed)

Результаты:

– разработана модель извлечения признаков из предложений на основе архитектуры Sentence-BERT для русского языка

– разработаны алгоритмы реферирования текстов на русском языке и проанализировано качество их работы

– выявлено, что алгоритмы TextRank и CustomSBERT_KMeans дополняют друг друга

– разработан алгоритм «Mixed», показывающий наилучшее качество реферирования текстов на русском языке

GitHub-репозиторий