Экстрактивная суммаризация научных статей в области биомедицины

Никифоровская А. Б.

Научный руководитель: А. А. Шпильман

Научный консультант: О. Ю. Шпынов

Санкт-Петербургская школа физико-математических и компьютерных наук

НИУ ВШЭ – Санкт-Петербург

2020

Введение

Суммаризация — процесс получения резюме на основе данного текста.

- Экстрактивная использует только предложения из текста
- Абстрактивная использует новые фразы и предложения

PubTrends

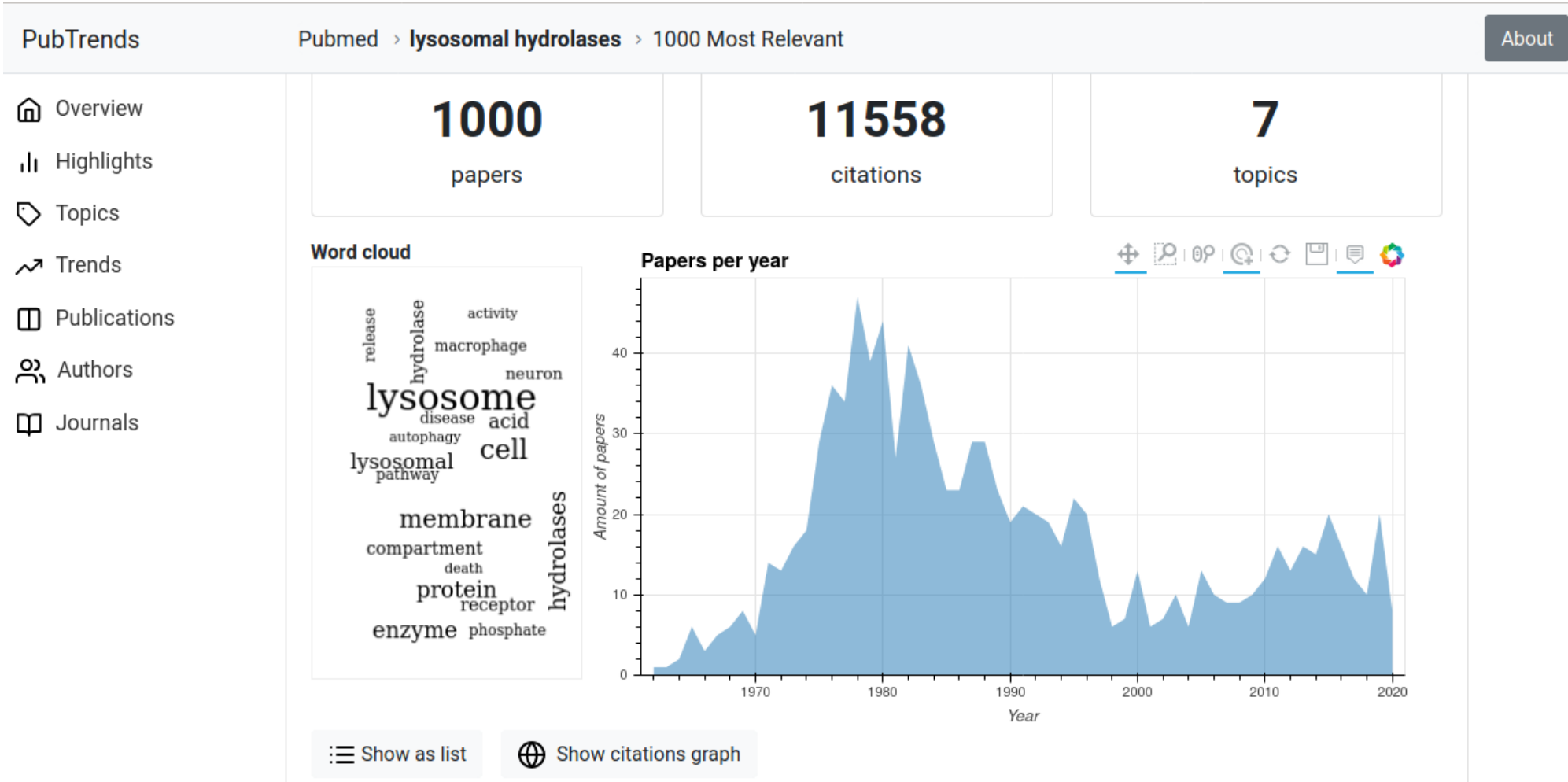
Веб-сервис, позволяющий анализировать статьи, соответствующие тому или иному запросу. [1]
Уже есть:

- Графы цитирования и коцитирования
- Ключевые слова тем
- Динамика цитирования

Не сокращают количество чтения, может не быть обзорный статей новой области.
Решение: генерация структуры литературного обзора.

[1] https://research.jetbrains.org/groups/biolabs/projects?project_id=56

PubTrends: интерфейс



Существующие подходы

- Использование структуры статьи для суммаризации [2]
- Использование предложений, в которых цитируется статья, для ее суммаризации [3]
- BERTSUM [4] – основан на модели BERT, лучшее в экстрактивной суммаризации.

---

[2] John M. Conroy et al. *Section mixture models for scientific document summarization.* 2017

[3] Cohan Arman et al. *Contextualizing citations for scientific summarization using word embeddings and domain knowledge.* 2017

[4] Liu Yang. *Fine-tune BERT for extractive summarization.* 2019.

Недостатки существующих решений

- У специализированных решений – невозможна работа в ограниченном количестве доступной информации про статьи
  - Часто имеются только аннотации статей, а не полный их текст
- BERTSUM не учитывает особенностей суммаризации статей
  - Построение текста
  - Длина текста

Цель и задачи

Цель: создать модуль для PubTrends, генерирующий обзоры статей.

Задачи:

- Собрать и обработать данные.
- Разработать и реализовать метод для суммаризации области.
- Протестировать модели автоматически.
- Реализовать удобный интерфейс для представления результатов суммаризации.
- Провести экспертную оценку.

Данные

PubMedCentral [5], Journal Article Tag Suite XML формат.
Постоянно обновляется.
Имеющаяся информация:

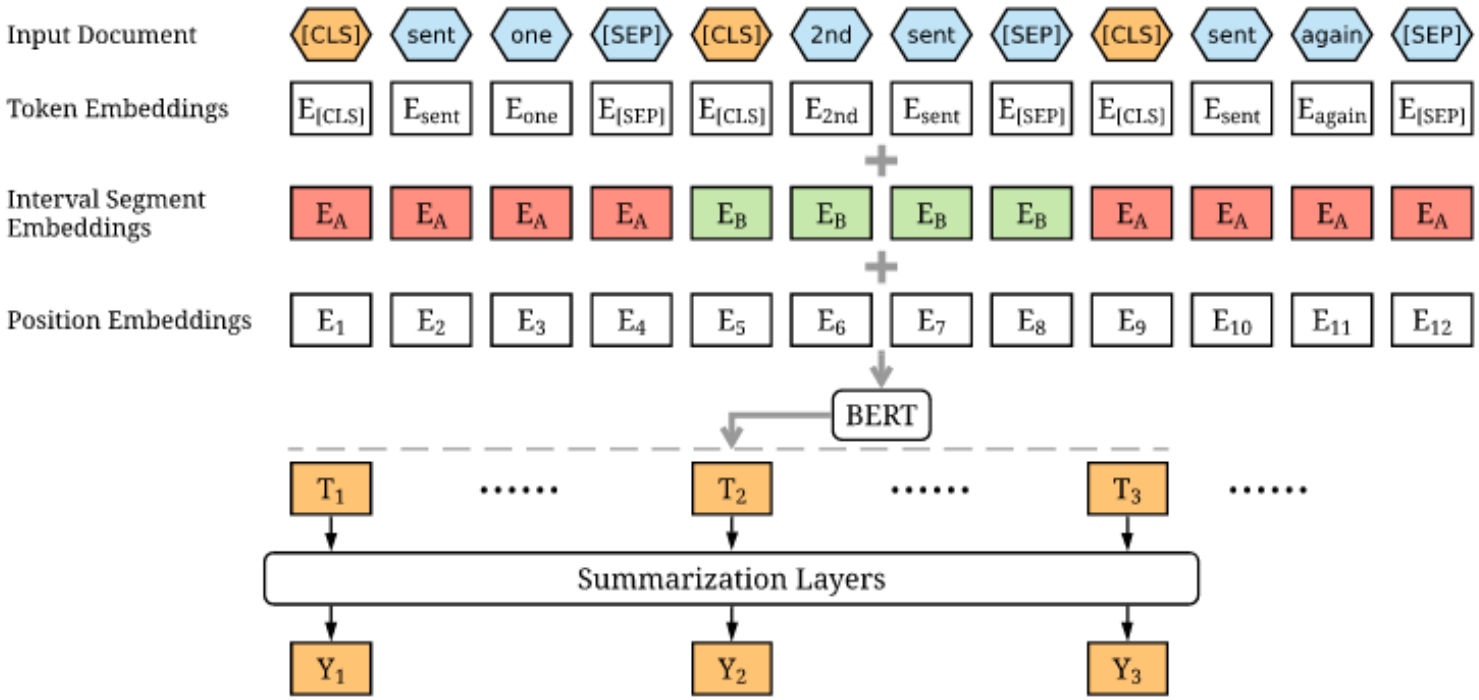
- PMID
- Название статьи
- Авторы
- Аннотация
- Основной текст с делением на разделы
- Ссылки, цитирование
- Подписи к таблицам и изображениям

Всего: 600 000 статей, 8 000 — обзорные

[5] https://www.ncbi.nlm.nih.gov/pmc/about/mscollection/

Метод: BERTSUM [4]

State-of-the-art



- Классификация: подходит предложение для резюме или нет
- Функция потерь: бинарная кросс-энтропия
- Подается весь имеющийся текст

[4] Liu Yang. *Fine-tune BERT for extractive summarization.* 2019.

Метод: модификация BERTSUM

Обзорные статьи хорошо резюмируют область.

Общая идея: использовать предложения в обзорных статьях, в которых цитируется данная статья, для обучения модели.

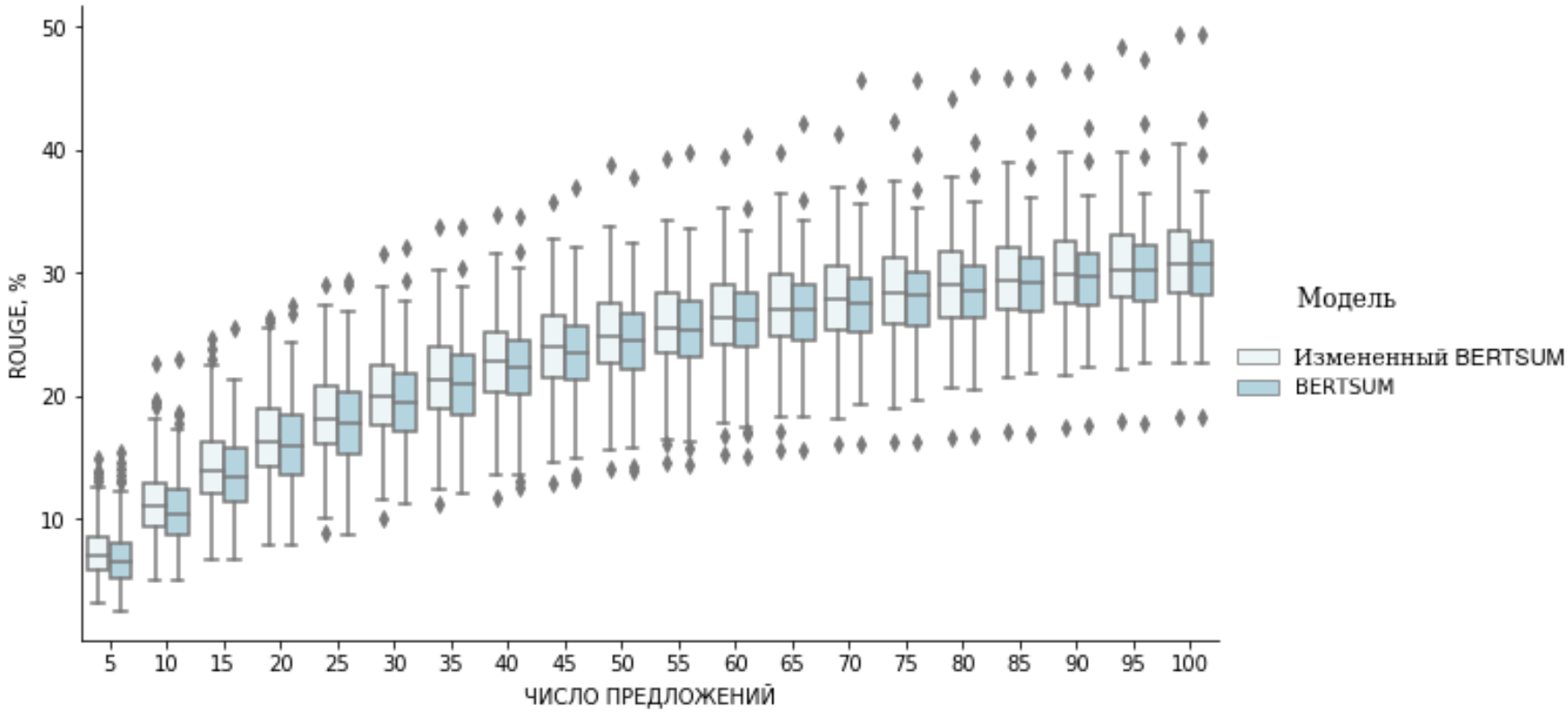
- Регрессия – приближаем метрику похожести ROUGE предложения на изначальную обзорную статью.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{обзорные статьи}\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{обзорные статьи}\}} \sum_{gram_n \in S} Count(gram_n)}$$

- Функция потерь: среднеквадратичная ошибка.
- Текст подается перекрывающимися блоками меньшего размера

Автоматическое тестирование

Обзорная статья, все ссылки-статьи суммаризируются, выбирается топ предложений по оценке модели.
ROUGE – похожесть получаемых резюме на изначальные обзорные статьи.



Результаты суммаризации

Pubmed › **lysosomal hydrolases** › 1000 Most Relevant

Result review

This table shows the resulting review.

Show 10 entries Search:

#	Topic #	Sentence	PMID	Score
1	3	Altered mobility was also detected for the nonlysosomal enzyme adenosine deaminase-d. Deficient activities of other lysosomal enzymes were observed as previously reported.	848490	0.08617037
2	3	These results are consistent with the mucolipidosis defect(s) being associated with abnormal post-translatinal processing of multiple lysosomal enzymes and adenosine deaminase-d.	848490	0.0861625
3	2	In marked contrast biologically inactive substances such as latex particles or digestible substrates such as erythrocytes do not induce the selective release of acid hydrolases from macrophages.	181971	0.09417827
4	2	In vitro studies have shown that macrophages secrete a variety of products on exposure to different stimuli.	181971	0.09411629
5	3	The first line of evidence was obtained from analysis of inhibition of enzyme pinocytosis	266721	0.09982359

Экспертное тестирование

3 запроса, по каждому 20 наиболее важных статей на основе цитирования, из каждой аннотации по одному предложению.
Оценка от эксперта:

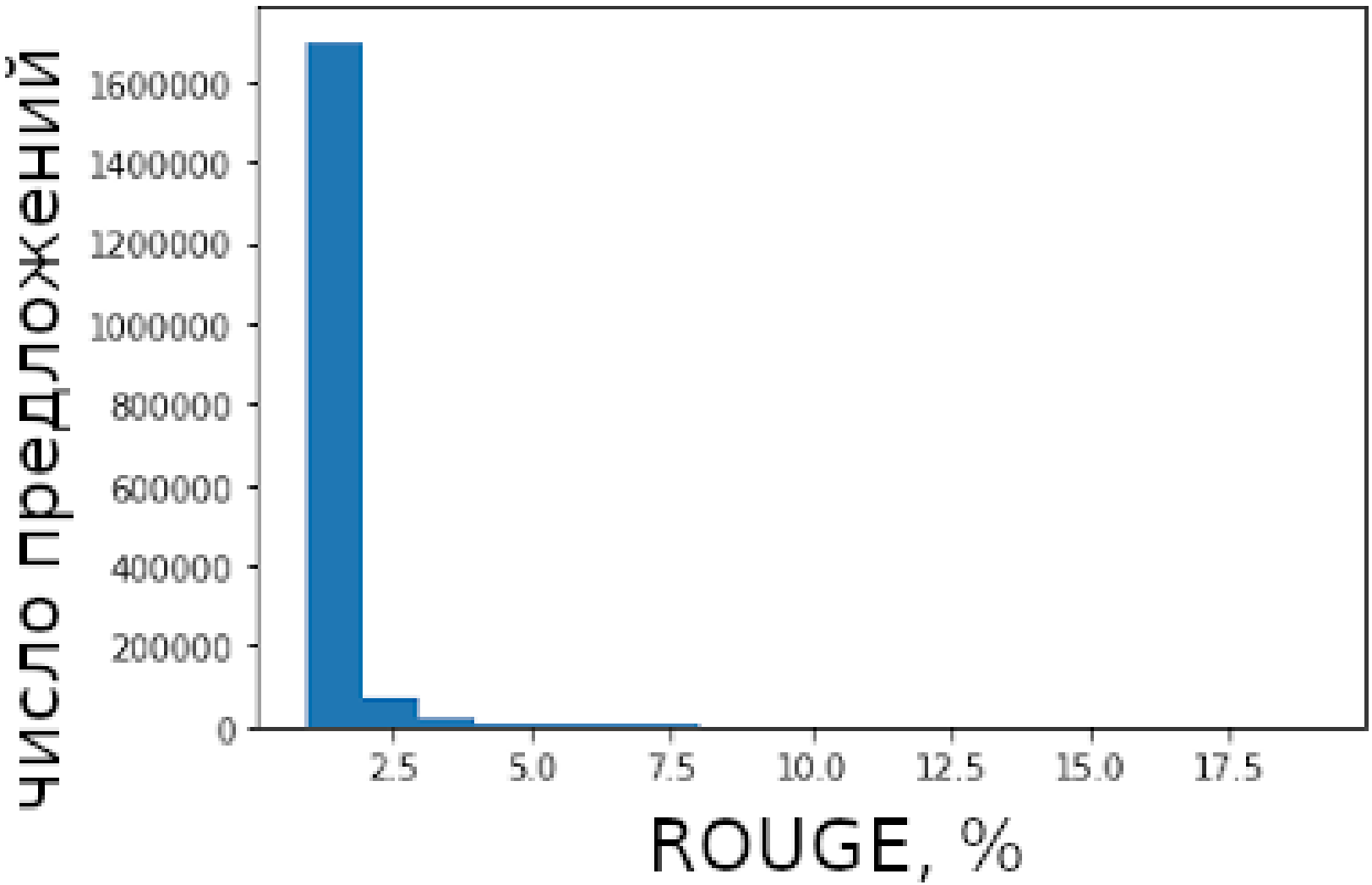
- не релевантно
- нейтрально
- полезно

Запрос №	Не релевантно	Нейтрально	Полезно
1	1	11	8
2	2	7	11
3	2	4	14
Среднее	**1.6**	**7.3**	**11**

Результаты

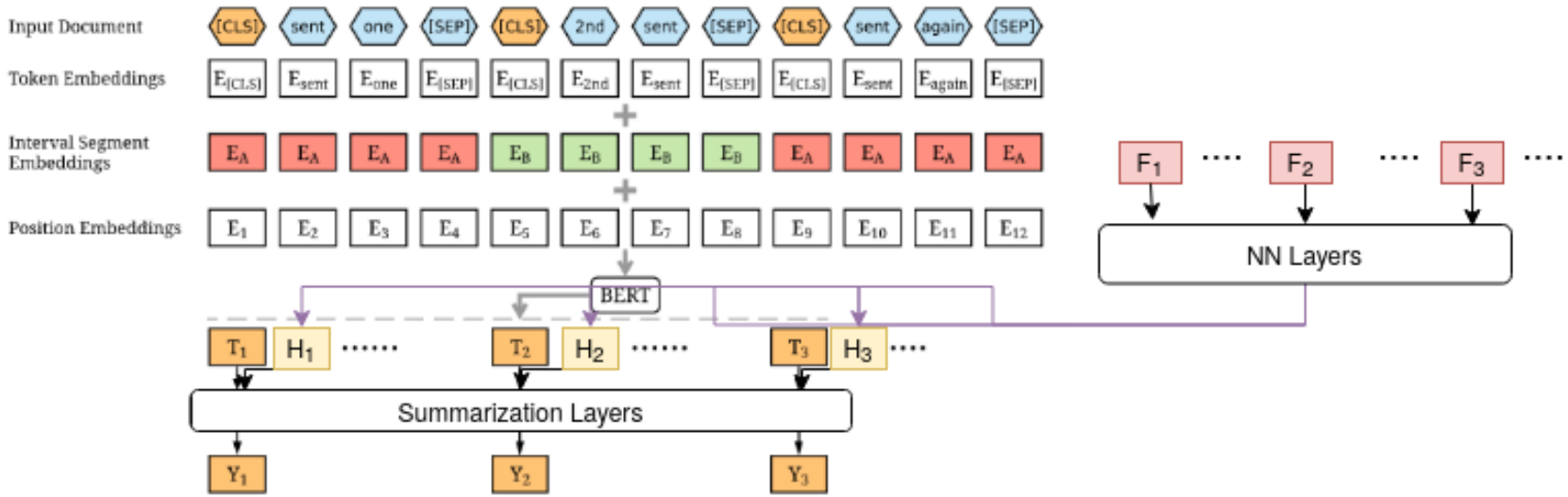
- Были собраны и организованы данные
- Разработана и реализована модель на основе лучшей в экстрактивной суммаризации с учетом особенностей научных статей
- Автоматическое тестирование и экспертная оценка показали
  - Преимущество разработанной модификации BERTSUM
  - Удовлетворенность результатами экспертом при условии наличия аннотаций, а не полных текстов
- Реализован модуль для веб-сервиса, визуализирующий результаты

Данные


число предложений
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
ROUGE, %

Модель с особенностями

К каждому предложению еще и вектор особенностей: похожесть на аннотацию, подписи.



Устройство сервиса