

# Суммаризация текстов



# Суммаризация текстов

## Постановка задачи

- Сгенерировать краткое изложение исходного текста
- Важно сохранить не только смысл, но и важные факты, содержащиеся в тексте
- Пример: снippets новостей в сервисе Яндекс.Новости

### **NASA оценило угрозу экипажу из-за утечки воздуха на МКС**

Для поиска источника утечки на три дня задраят люки всех отсеков, специалисты за это время изучат уровни давления воздуха. Проверка экипажу станции угрожать не будет. Астронавт NASA Крис Кэссиди будет находиться в российском модуле "Звезда". [В источнике](#)

### **Смартфоны смогут определять степень опьянения владельца**

Американские ученые заявили, что с помощью смартфона можно определить, как сильно пьян его владелец. Об этом со ссылкой на Journal of Studies on Alcohol and Drugs пишет «ТАСС». Специалисты провели исследование среди 22 добровольцев от 21 до 43 лет. [В источнике](#)



# Суммаризация текстов

## Пример

Вот растение, которому в наш суматошный век истрёпанных нервов, изнурительных бессонниц и сдвинутой с места психики надо бы поставить красивый памятник: валериана, подобно матери, успокоит и усыпит, вернет так необходимое всем нам душевное равновесие.



© В. Солоухин



# Суммаризация текстов

Пример

Валериане следовало бы поставить памятник: она успокаивает, помогает заснуть, возвращает душевное равновесие.



# Суммаризация текстов

Формальная постановка задачи

Экстрактивная  
суммаризация

Абстрактивная  
суммаризация



# Суммаризация текстов

Формальная постановка задачи

Экстрактивная  
суммаризация



Абстрактивная  
суммаризация

Выбрать из текста важные  
словосочетания или  
предложения



# Суммаризация текстов

Формальная постановка задачи

Экстрактивная  
суммаризация



Выбрать из текста важные  
словосочетания или  
предложения

Абстрактивная  
суммаризация



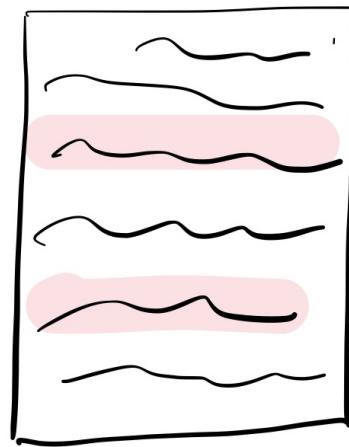
Сгенерировать новый  
согласованный текст,  
который будет содержать  
основные идеи исходного  
текста



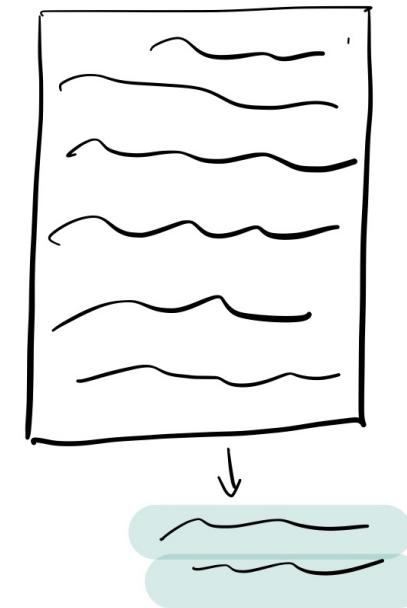
# Суммаризация текстов

Иными словами

Экстрактивная  
суммаризация



Абстрактивная  
суммаризация



# Систематизация методов

СУММАРИЗАЦИЯ



# Систематизация методов



# Систематизация методов



# Систематизация методов



# Систематизация методов



# **Постановка задачи. Экстрактивная суммаризация**



# Экстрактивная суммаризация

## Основные шаги

1. Разобьем текст на предложения и составим вектора предложений: можем использовать как модель мешка слов, так и векторные модели представления предложения

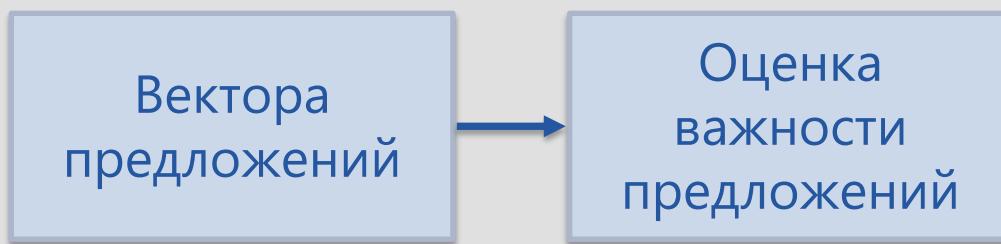
Вектора  
предложений



# Экстрактивная суммаризация

## Основные шаги

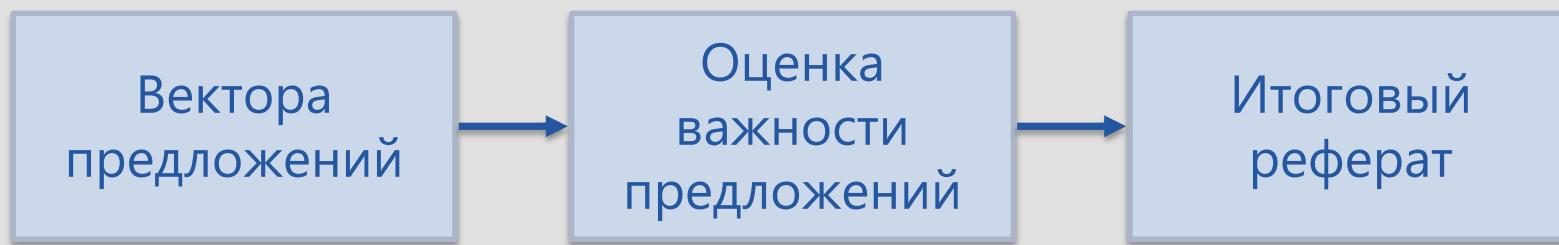
2. Найдем оценку важности каждого предложения – чем выше такая оценка, тем выше вероятность того, что предложение будет включено в итоговый реферат



# Экстрактивная суммаризация

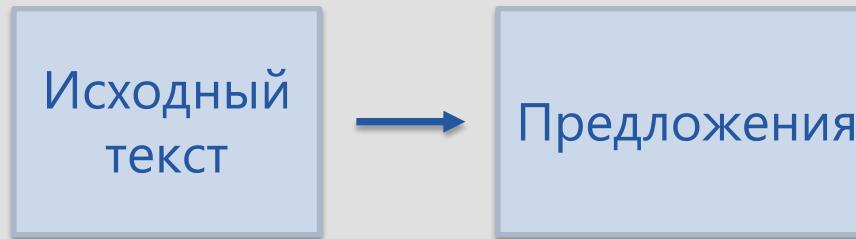
## Основные шаги

3. Выберем  $k$  наиболее важных предложений и составим из них итоговый реферат



# TextRank

## Базовый алгоритм экстрактивной суммаризации



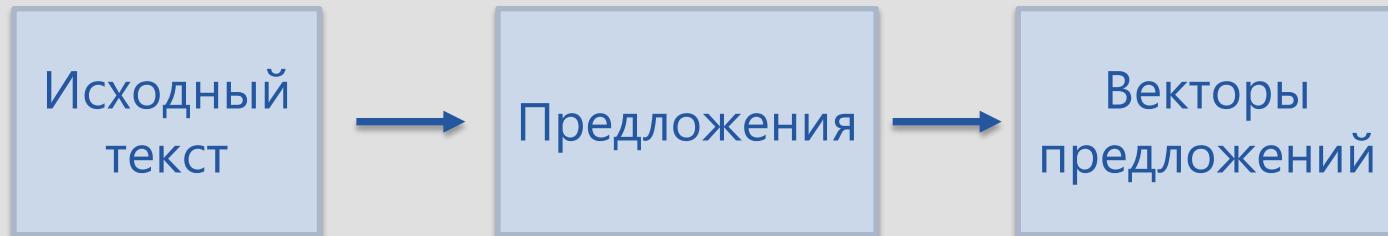
Разобъем исходный текст  
на предложения

- Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text, EMNLP



# TextRank

## Базовый алгоритм экстрактивной суммаризации



Составим вектора предложений:

- $tf - idf$  вектора
- Любая модель эмбеддингов предложений

► Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text, EMNLP



# TextRank

## Базовый алгоритм экстрактивной суммаризации



Оценим близость между  
предложениями по  
косинусной мере  
близости

► Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text, EMNLP



# TextRank

## Базовый алгоритм экстрактивной суммаризации



Составим граф:

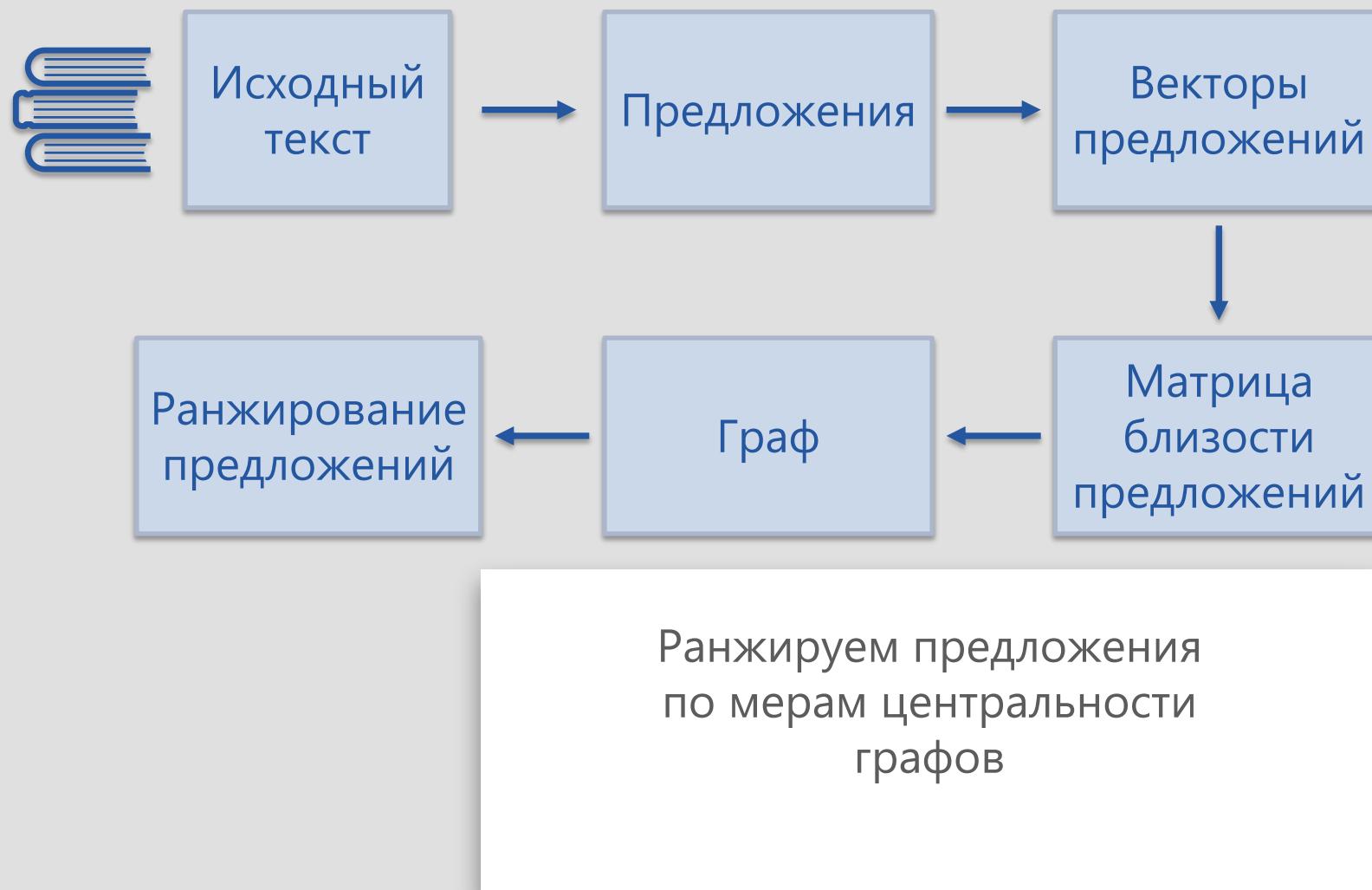
- Вершины - предложения
- Ребра соединяют близкие предложения

► Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text, EMNLP



# TextRank

## Базовый алгоритм экстрактивной суммаризации

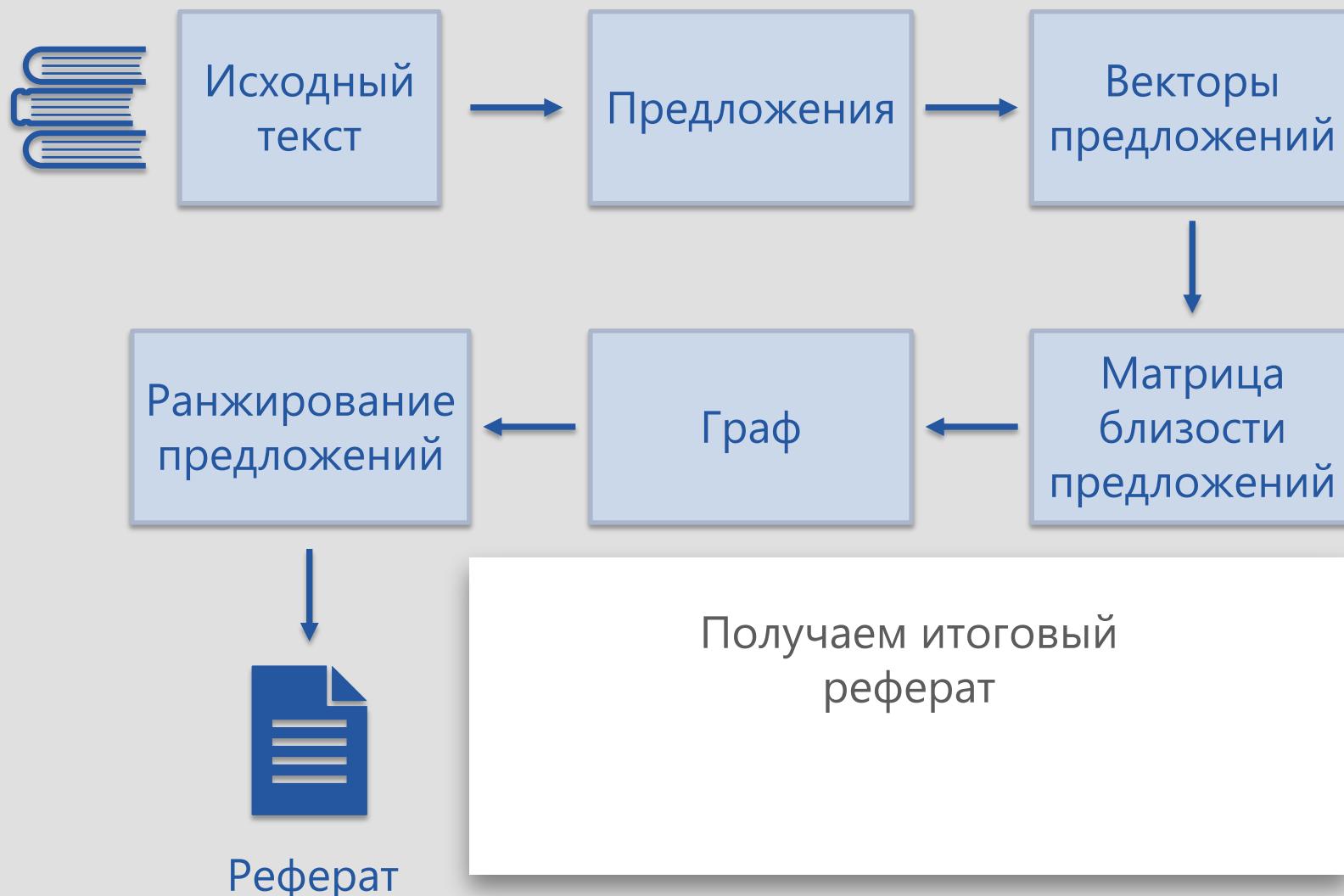


► Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text, EMNLP



# TextRank

## Базовый алгоритм экстрактивной суммаризации

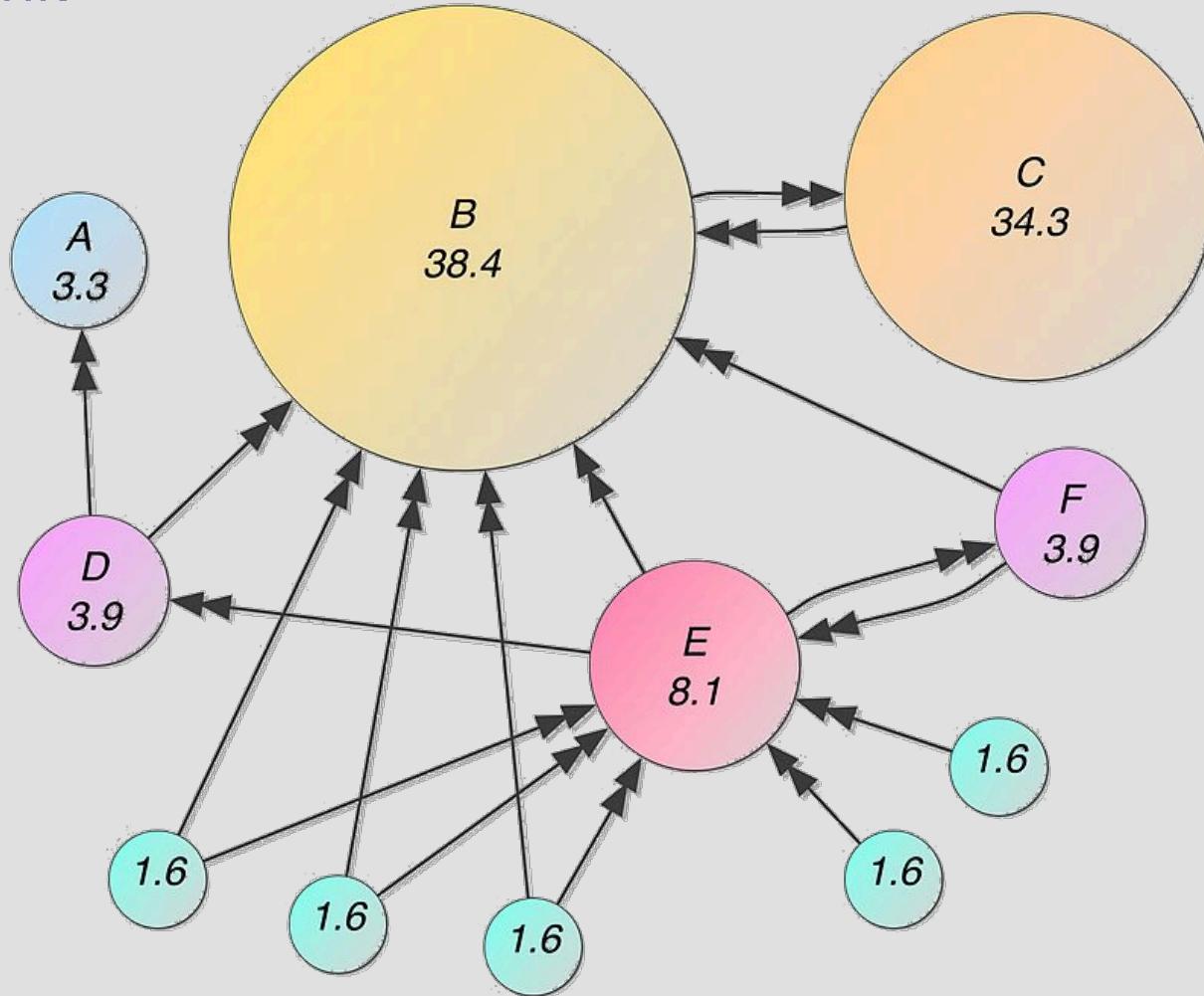


► Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text, EMNLP



# Меры центральности

## PageRank



► <https://en.wikipedia.org/wiki/PageRank>



# Меры центральности

## PageRank

$$PR(n_i) = \frac{1-d}{N} + d \sum_{n_j \in In(n_i)} \frac{PR(n_j)}{|Out(n_j)|}$$



# Меры центральности

## PageRank

$$PR(n_i) = \frac{1-d}{N} + d \sum_{n_j \in In(n_i)} \frac{PR(n_j)}{|Out(n_j)|}$$

Важность вершины  $n_i$



# Меры центральности

## PageRank

Константа, обычно равна 0,85

$$PR(n_i) = \frac{1 - d}{N} + d \sum_{n_j \in In(n_i)} \frac{PR(n_j)}{|Out(n_j)|}$$

Число вершин в графе



# Меры центральности

## PageRank

Множество вершин, в которых есть ребро из  $n_j$

$$PR(n_i) = \frac{1 - d}{N} + d \sum_{n_j \in In(n_i)} \frac{PR(n_j)}{|Out(n_j)|}$$

Множество вершин, из которых есть ребро в  $n_i$



# Меры центральности

## PageRank

Значимость  
вершины  $n_j$

$$PR(n_i) = \frac{1 - d}{N} + d \sum_{n_j \in In(n_i)} \frac{PR(n_j)}{|Out(n_j)|}$$



# Экстрактивная суммаризация

## Заключение

- Рефераты, полученные с помощью экстрактивной суммаризации, корректны с точки зрения грамматики
- Простой алгоритм экстрактивной суммаризации TextRank не требует обучения, однако требует установления порогов на число предложений в реферате и на метрики важности и близости
- Применимы и другие алгоритмы ранжирования предложений, использующие, например, нейронные сети
- Однако, вся выразительная способность современных методов обработки текстов достигается только с помощью алгоритмов абстрактивной суммаризации



# Абстрактивная суммаризация



# Абстрактивная суммаризация

## Постановка задачи

- Вход: длинный текст
- Выход: сгенерированный автоматический короткий текст
- Основные модели: seq2seq модели
- Аналогия с машинным переводом: переводим длинный текст в короткий текст
- Нужны модели с обучением!



# Абстрактивная суммаризация

## Наборы данных, CNN/DailyMail

### Alessandra Ambrosio shows off her Latino tan and endless legs in edgy new fashion campaign

- Mother-of-two, 34, snapped up to front Dafiti's AW15 campaign
- Latin American e-tailer believe she embodies the style of the brand
- Recently named No. 8 on Forbes list of top-earning models

By BIANCA LONDON FOR MAILONLINE

PUBLISHED: 11:19 GMT, 24 April 2015 | UPDATED: 13:08 GMT, 24 April 2015



160  
shares

77  
[View comments](#)

Brazilian supermodel Alessandra Ambrosio goes back to her roots in an edgy new campaign shot in her home country.

The 34-year-old Victoria's Secret Angel shows off her Latino style and golden tan as she poses in a new campaign for online fashion retailer Dafiti, shot in São Paulo.

Dafiti, Latin America's largest online fashion retailer, has launched its own fashion collection, the Dafiti Collection, and signed Alessandra because they believe she embodies the style of the brand.

статьи

~ 766  
слов

рефераты

~ 53  
слова

обучающие  
данные

~ 300 тыс.

валидационные  
данные

~ 15 тыс.

тестовые  
данные

~ 10 тыс.

► Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. *Teaching machines to read and comprehend*. NeurIPS, 2015



# Абстрактивная суммаризация

## Наборы данных, WikiHow

### How to Help Save Rivers

#### Method 1 Reducing Your Water Usage

- 1 **Take quicker showers to conserve water.** One easy way to conserve water is to take shorter showers. Practice cutting your showers down to 10 minutes, then 7, then 5, every day.
- 2 **Wait for a full load of clothing before running a washing machine.** This will save water and electricity, so running a cycle for a couple of articles of clothing is better than running a cycle for one article.
- 3 **Turn off the water when you're not using it.** Avoid letting the water run while you brush your teeth or shave. Keep your hoses and faucets turned off as much as possible.

#### Method 2 Using River-Friendly Products

- 1 **Select biodegradable cleaning products.** Any chemicals you use in your home can end up in rivers. Choose natural soaps or create your own cleaning and disinfecting products from items like vinegar, lemon juice, and other natural products. These products have far less of an impact on the environment.
- 2 **Choose recycled products instead of new ones.** New products require more energy and resources to produce than recycled products. Reuse what you already own when possible. If you need to buy new products, look for ones made from recycled paper or other reused material.

статьи

~ 570  
слов

рефераты

~ 62  
слова

обучающие  
данные

~ 200  
тыс.

► Koupaee, M., & Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305.



# Абстрактивная суммаризация

## Меры качества, ROUGE

- Recall-Oriented Understudy for Gisting Evaluation
- $ROUGE - N, N \in [1,2,3,4]$
- $ROUGE - L$  (самая длинная подстрока)

$$ROUGE - N = \frac{\sum_{S \in sum} \sum_{n-gram \in S} Count_{match}(n-gram)}{\sum_{S \in sum} \sum_{n-gram \in S} Count(n-gram)}$$

► Lin, C. Y. Rouge: A package for automatic evaluation of summaries, ACL 2004



# Абстрактивная суммаризация

## Меры качества, ROUGE

$$ROUGE - N = \frac{\sum_{S \in Sum} \sum_{n-gram \in S} Count_{match}(n - gram)}{\sum_{S \in Sum} \sum_{n-gram \in S} Count(n - gram)}$$

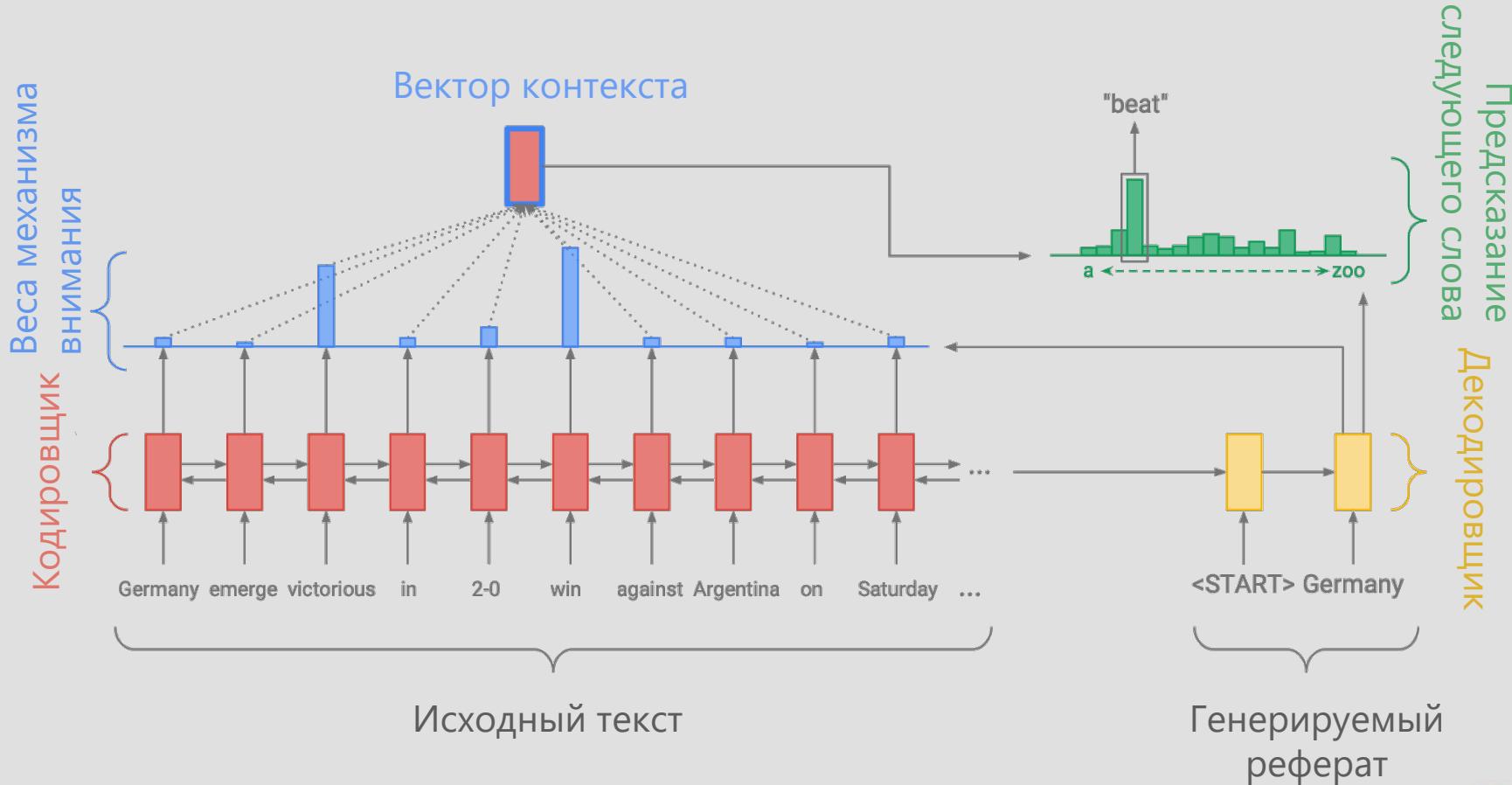
- *Sum* – множество предложений, реферат
- *Count<sub>match</sub>(n – gram)* – число совпавших *n* –грам с золотым стандартом
- *Count(n – gram)* – число *n* –грамм в золотом стандарте

► Lin, C. Y. Rouge: A package for automatic evaluation of summaries, ACL 2004



# Абстрактивная суммаризация

С помощью seq2seq моделей

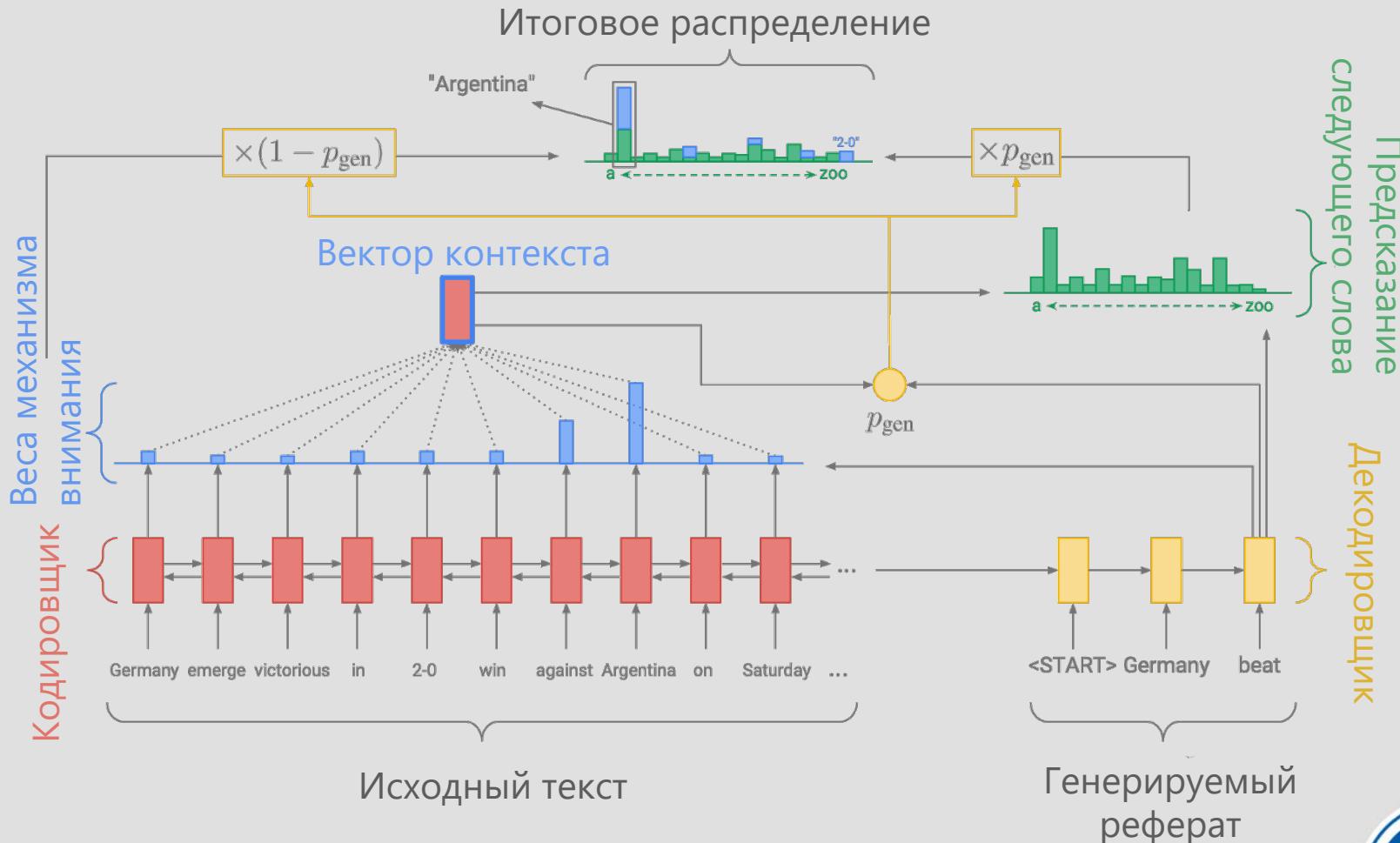


► See, A., Liu, P. J., & Manning, C. D. Get To The Point: Summarization with Pointer-Generator Networks. ACL 2017



# Абстрактивная суммаризация

С помощью seq2seq моделей с указателем

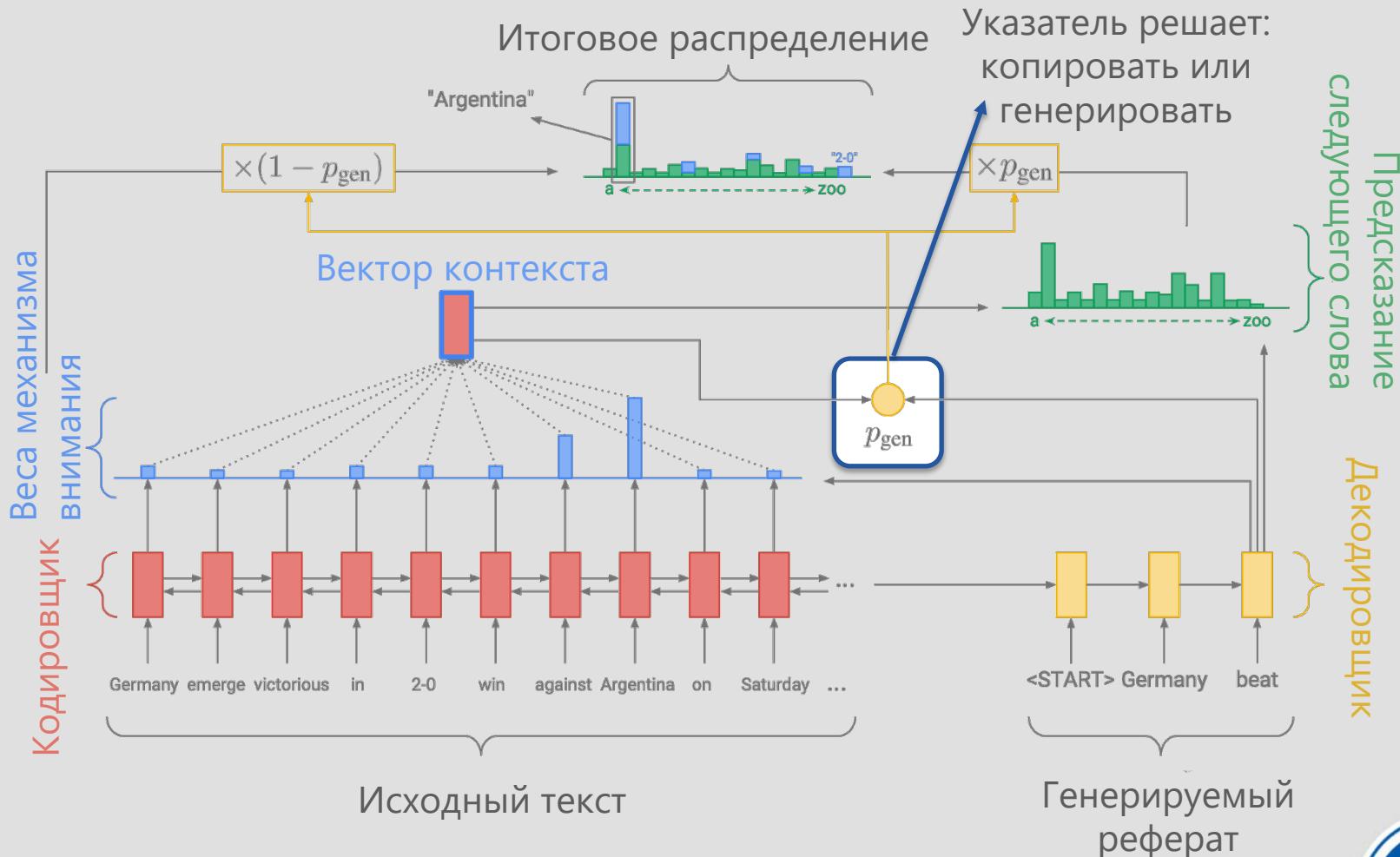


► See, A., Liu, P. J., & Manning, C. D. Get To The Point: Summarization with Pointer-Generator Networks. ACL 2017



# Абстрактивная суммаризация

С помощью seq2seq моделей с указателем



► See, A., Liu, P. J., & Manning, C. D. Get To The Point: Summarization with Pointer-Generator Networks. ACL 2017



# Seq2seq модели с указателем

- $P_{gen}$  – вероятность того, что следующее слово будет сгенерировано из словаря
- $P_{vocab}$  – распределение вероятностей слов из словаря
- $\alpha_i^t$  – веса механизма внимания
- Два набора вероятностей: генерируем или копируем следующее слово

$$P(w) = P_{gen}P_{vocab} + (1 - P_{gen}) \sum_{i:w_i=w} \alpha_i^t$$



# Seq2seq модели с указателем

- Два набора вероятностей: генерируем или копируем следующее слово

$$P(w) = P_{gen}P_{vocab} + (1 - P_{gen}) \sum_{i:w_i=w} \alpha_i^t$$

- Функция потерь:

$$loss = -\frac{1}{T} \sum_t \log P(w_t^*)$$



# Seq2seq модели с указателем

- $P_{gen}$  – вероятность того, что следующее слово будет сгенерировано из словаря

$$P_{gen} = \sigma(w_h^T h_t^* + w_h^T s_t + w_x^T x_t)$$

Вектор контекста



# Seq2seq модели с указателем

- $P_{gen}$  – вероятность того, что следующее слово будет сгенерировано из словаря

$$P_{gen} = \sigma(w_h^T h_t^* + w_h^T s_t + w_x^T x_t)$$

Текущее скрытое  
состояние  
декодировщика



# Seq2seq модели с указателем

- $P_{gen}$  – вероятность того, что следующее слово будет сгенерировано из словаря

$$P_{gen} = \sigma(w_h^T h_t^* + w_h^T s_t + w_x^T x_t)$$

Текущее слово в  
декодировщике



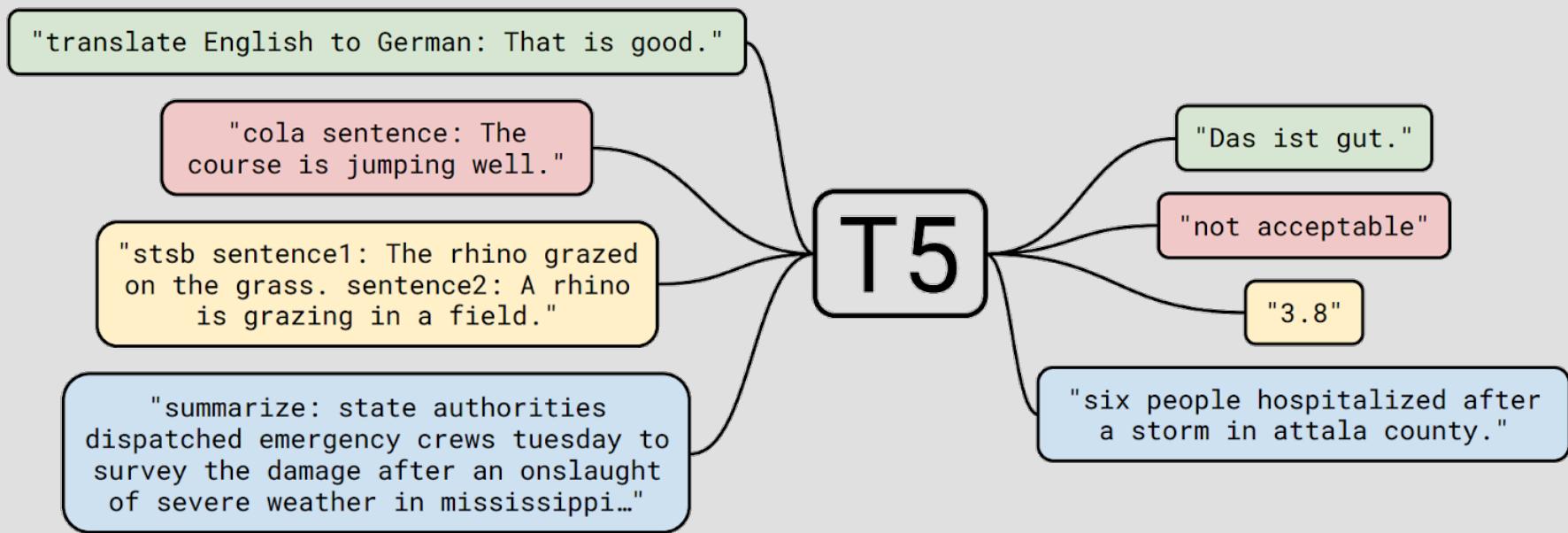
# Seq2seq модели для абстрактной суммаризации

- Проблема: качество измеряется с помощью ROUGE, а функция потерь ее не использует
- Более современные модели используют, конечно же, языковые модели на основе архитектуры Трансформер



# T5: Text-To-Text Transfer Transformer

- Большая seq2seq модель, умеющая решать одновременно несколько задач
- Унифицированный интерфейс: на вход подается управляющий код (summarize) и текст, на выходе – ответ



► Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.



# T5: Text-To-Text Transfer Transformer

## Предобучение



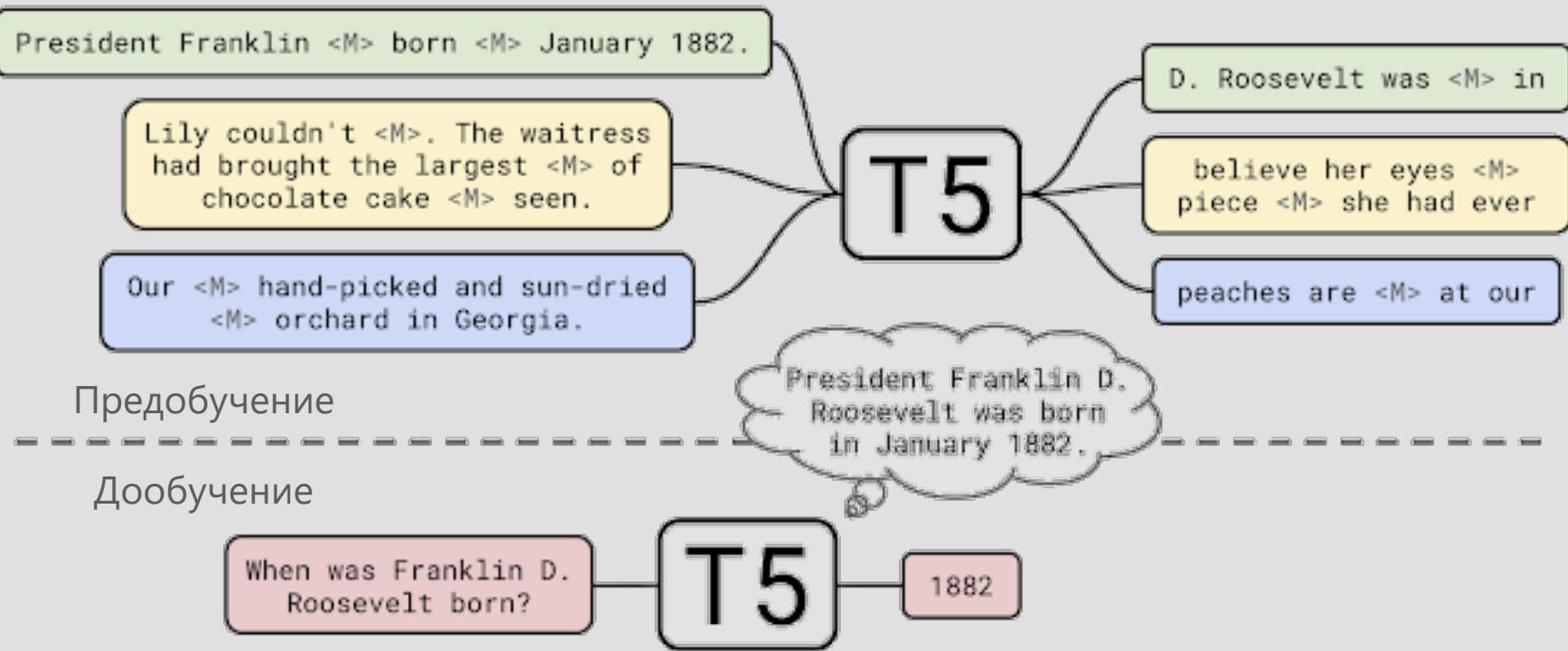
- Обучена на 750GB очищенных веб-текстов
- Использована архитектура кодировщик-декодировщик

► Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.



# T5: Text-To-Text Transfer Transformer

## Дообучение



- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.



# T5: Text-To-Text Transfer Transformer

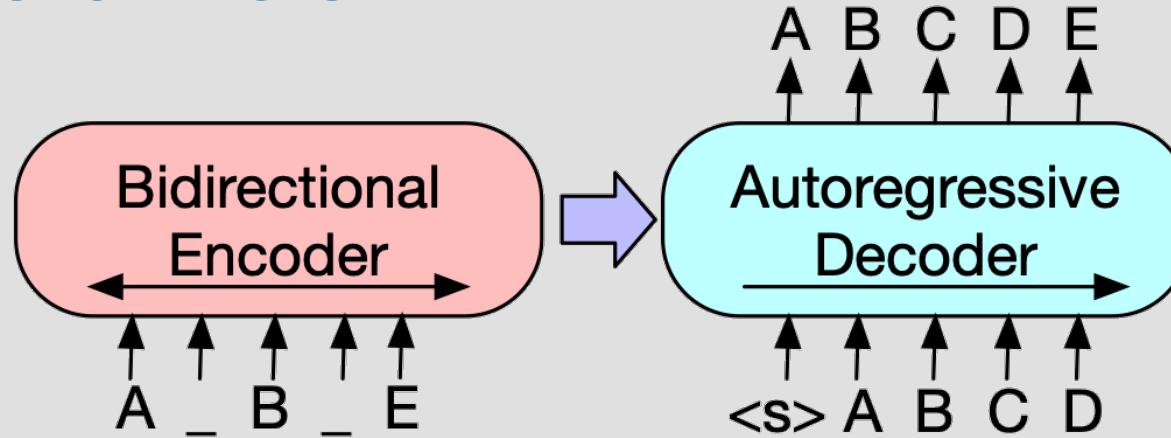
## Дообучение

- Дообучена решать несколько практических задач:
  - Поиск ответа на вопрос
  - Абстрактивная суммаризация
- Модель T5 показывает высокие результаты как на стандартных бенчмарках, GLUE и SuperGLUE, так и в практических задачах

► Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.



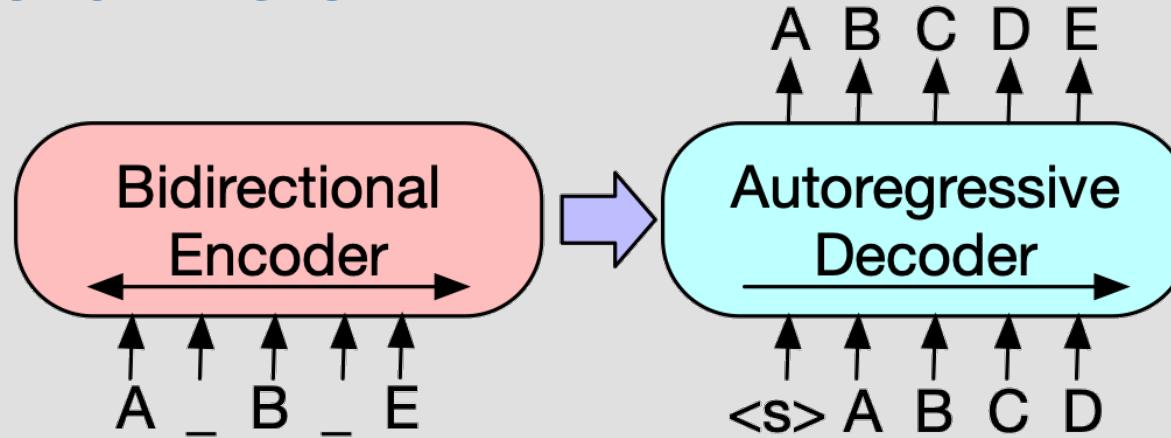
# BART: Bidirectional and Auto-Regressive Transformers



- Архитектура класса кодировщик-декодировщик
- 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*.



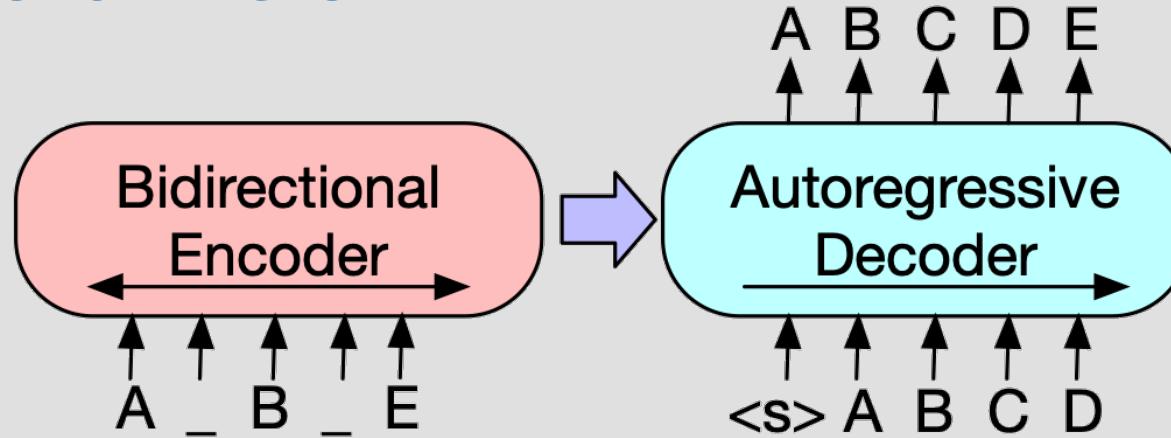
# BART: Bidirectional and Auto-Regressive Transformers



- Двунаправленный кодировщик (по аналогии с BERT) совмещен с авторегрессионным декодировщиком
- 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*.



# BART: Bidirectional and Auto-Regressive Transformers

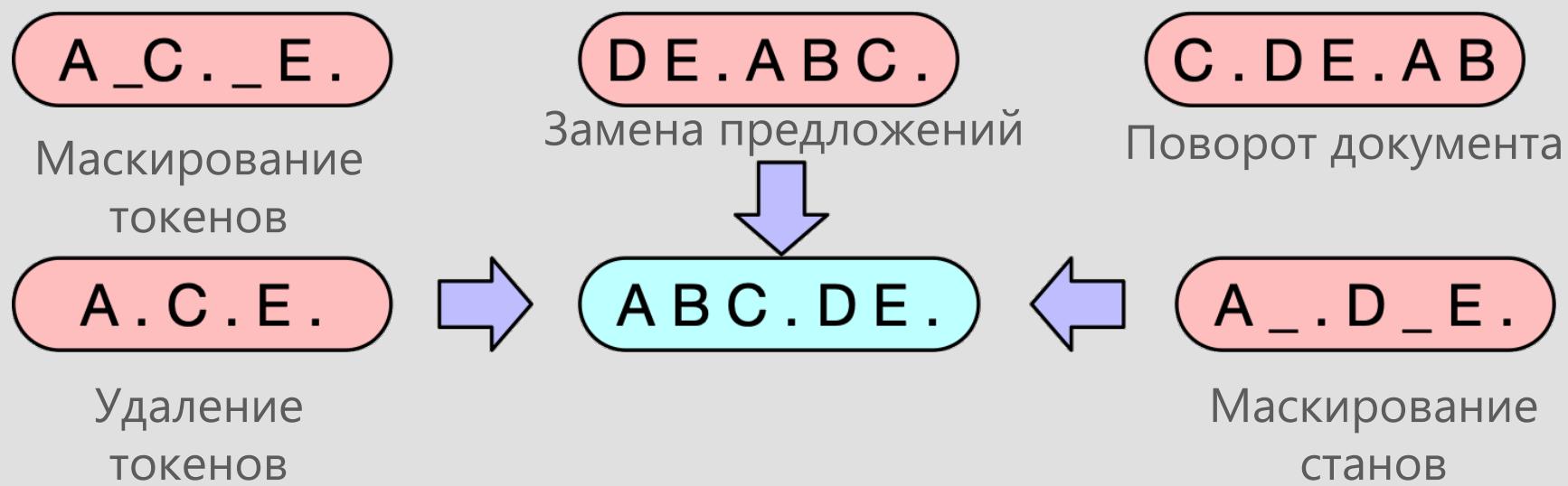


- Использует примерно сопоставимое с моделью BERT количество параметров, при этом применима в большем спектре задач
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.



# BART: Bidirectional and Auto-Regressive Transformers

По трансформированному входу требуется предсказать исходное предложение



- 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*.



# BART: Bidirectional and Auto-Regressive Transformers

- Модель BART показывает результаты, сопоставимые с RoBERTa, в задачах понимания текстов
- В задачах генерации текстов модель BART установила новые рекордные показатели
- Модель BART достаточно компактна и не сильно превосходит модель BERT по количеству параметровБолее современные модели используют, конечно же, языковые модели на основе архитектуры Трансформер

► 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*.



# Абстрактивная суммаризация

## Заключение

- Ранние подходы к абстрактной суммаризации использовали простые архитектуры класса кодировщик-декорвщик и показывали низкие результаты
- Метрики, используемые для оценки качества абстрактной суммаризации, можно оптимизировать с помощью обучения с подкреплением



# Абстрактивная суммаризация

## Заключение

- С появлением новых, более эффективных авторегрессионных языковых моделей, качество моделей абстрактивной суммаризации сильно выросло
- Основной сложностью в настоящее время остается подготовка качественных наборов данных и человеческая оценка качества



# **Упрощение текстов**



# Упрощение текстов

## Постановка задачи

- Упрощение на лексическом уровне [lexical simplification]: найти в тексте сложно слово и заменить на более простое
- Пример: плебисцит -> голосование
- Упрощение на уровне предложений [sentence simplification]: переписать предложение, заменить сложные грамматические конструкции на простые, убрать из предложения несущественные детали
- Пример: Итоги плебисцита, проведённого в России в 2020 г., имеют далеко идущие последствия -> Многое изменится в России после голосования в 2020 г.



# Упрощение текстов

## Социально-значимая задача

- Детям, неносителям языка и людям с когнитивными нарушениями сложно читать сложные тексты
- Адаптация текстов к разным уровням сложности — трудоёмкая работа
- Ее выполняют, например, разработчики методических материалов для изучения иностранных языков, или авторы школьных учебников
- Часть этой работы можно автоматизировать



# Упрощение текстов

## Формальная постановка задачи

- Упрощение текстов похоже на суммаризацию текстов или на машинный перевод — это ещё одна задача трансформации последовательности
- Архитектуры класса кодировщик-декодирование могут быть использованы для упрощения текстов
- Как всегда, основные сложности вызывает дефицит обучающих данных и необходимость использования нетривиальных метрик качества



# Упрощение текстов

## Формальная постановка задачи

- Упрощение текстов похоже на теггирование последовательности: для каждого слова надо определить, стоит ли его заменить, удалить, сохранить и т.д..
- Любая модель теггирования последовательности, включая рекуррентную нейронную сеть, может быть использована в такой постановке задачи
- Как всегда, основные сложности вызывает не только дефицит обучающих данных и необходимость использования нетривиальных метрик качества, но и необходимость в дополнительной разметке каждого слова



# Упрощение текстов

## Набор данных на основе Википедии

- Выровненный корпус пар предложений из англоязычной Википедии и упрощенной англоязычной Википедии

English  
Wikipedia

**Machine learning (ML)** is the study of computer algorithms that improve automatically through experience. It is seen as a subset of [artificial intelligence](#). Machine learning algorithms build a [mathematical model](#) based on sample data, known as "[training data](#)", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as [email filtering](#) and [computer vision](#), where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is closely related to [computational statistics](#), which focuses on making predictions using computers. The study of [mathematical optimization](#) delivers methods, theory and application domains to the field of machine learning. [Data mining](#) is a related field of study, focusing on [exploratory data analysis](#) through [unsupervised learning](#). In its application across business problems, machine learning is also referred to as [predictive analytics](#).



# Упрощение текстов

## Набор данных на основе Википедии

- Выровненный корпус пар предложений из англоязычной Википедии и упрощенной англоязычной Википедии

**Machine learning** gives [computers](#) the ability to [learn](#) without being explicitly programmed ([Arthur Samuel](#), 1959). It is a subfield of [computer science](#).

The idea came from work in [artificial intelligence](#). Machine learning explores the study and construction of [algorithms](#) which can [learn](#) and make predictions on [data](#). Such algorithms follow [programmed instructions](#), but can also make predictions or decisions based on data. They build a [model](#) from sample inputs. Machine learning is done where designing and programming explicit [algorithms](#) cannot be done. Examples include [spam filtering](#), detection of [network](#) intruders or malicious insiders working towards a data breach, [optical character recognition](#) (OCR), [search engines](#) and [computer vision](#).

Simple English  
Wikipedia



# Упрощение текстов

## Упрощенные новости, Newsela и Deutsche Welle



Spanish

570L

By National Geographic Society, adapted by Newsela staff

Published: 06/14/2019 Word Count: 421

Recommended for: Middle School - High School

Text Level: 3



Activities

*Ancient Greece was the world's first democracy.*



*The United States is also a democracy. Indeed, our form of government was partly based on Greece's. Below are some key ideas we borrowed from the Greeks.*

The United States became independent in 1776. Its founding fathers then set about creating a government for the new country. Their biggest model was ancient Greece's democracy. They borrowed many ideas from the Greeks.

Уровни сложности

► <https://newsela.com/>



# Упрощение текстов

## Упрощенные новости, Newsela и Deutsche Welle



Spanish

1040L ▾

Уровни  
сложности

By National Geographic Society, adapted by Newsela staff

Published: 06/14/2019 Word Count: 864

Recommended for: Middle School - High School

Text Level: 7



Activities

*The United States has a complex political system. The central element of this system is democracy, a form of government in which the ultimate power rests with the people. In the case of the United States, that power is exercised indirectly, through elected representatives. Although the United States has been a strong promoter of democracy, we did not invent it. The true pioneers of democracy were the ancient Greeks. As the article below shows, the men*



► <https://newsela.com/>



# Упрощение текстов

## Меры качества, BLEU, SARI, FKGL

- По аналогии с машинным переводом, можно использовать метрику BLEU. Однако, BLEU никак не учитывает структуру предложения и лексическую сложность
- SARI оценивает корректность трех операций: удаление, добавление и сохранение слова, по сравнению с золотым стандартом
- FKGL оценивает удобочитаемость предложения, но не учитывает грамматическую и семантическую корректность.



# Упрощение текстов

## Меры качества, SARI

$$SARI = \frac{1}{3} F_{\text{добавление}} + \frac{1}{3} F_{\text{удаление}} + \frac{1}{3} F_{\text{сохранение}}$$

Оценивается  $F$ -мера трех операций модели:

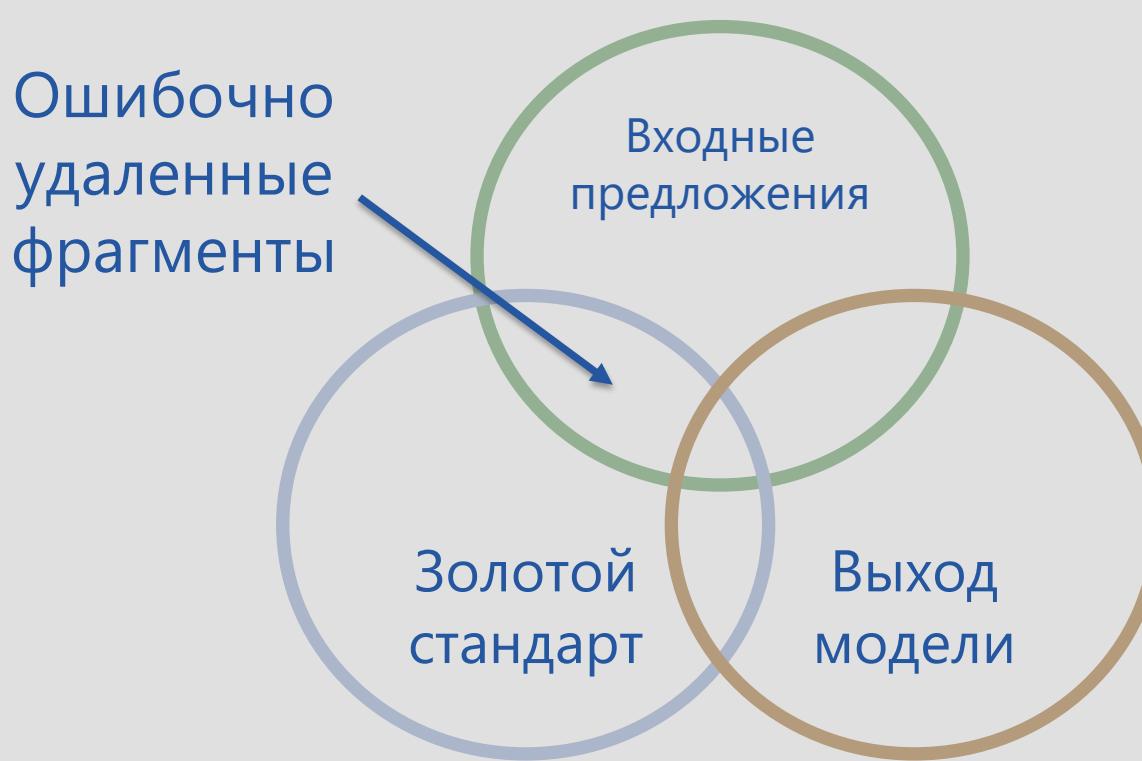
- Добавление: слова не было в входном предложении, но оно было добавлено и золотом стандарте
- Сохранение: слово было и в входном предложении, и в золотом стандарте
- Удаление: слово было во входном предложении и отсутствует в золотом стандарте

► Xu, Wei, Courtney Napolis, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing statistical machine translation for text simplification." TACL, 2016



# Упрощение текстов

## Меры качества, SARI

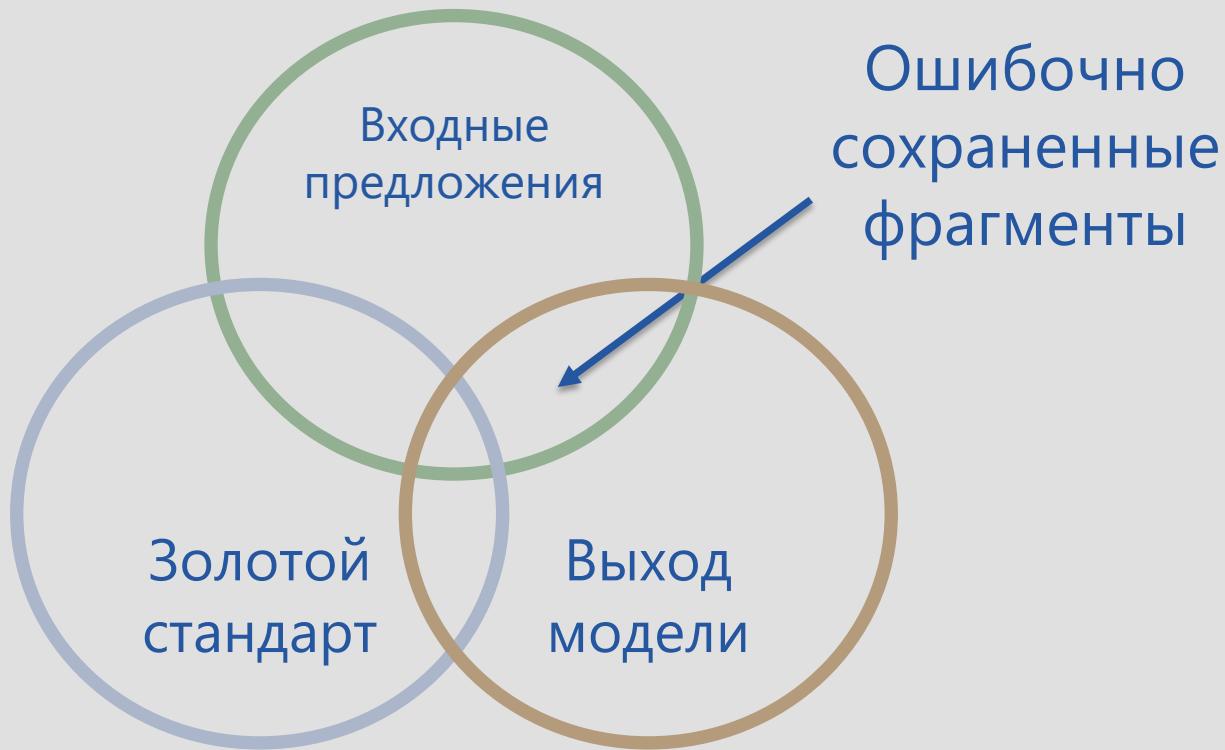


- Xu, Wei, Courtney Napolis, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing statistical machine translation for text simplification." TACL, 2016



# Упрощение текстов

## Меры качества, SARI



- Xu, Wei, Courtney Napolis, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing statistical machine translation for text simplification." TACL, 2016



# Упрощение текстов

## Меры качества, SARI

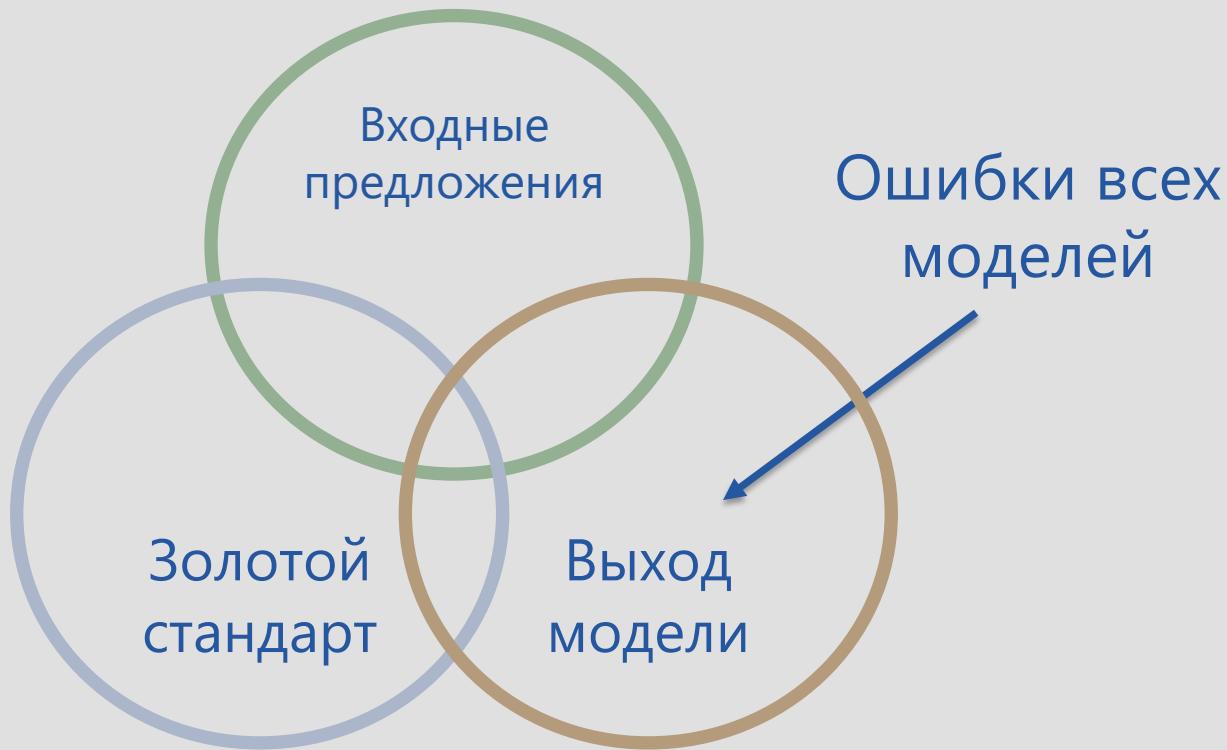


- Xu, Wei, Courtney Napolis, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing statistical machine translation for text simplification." TACL, 2016



# Упрощение текстов

Меры качества, SARI

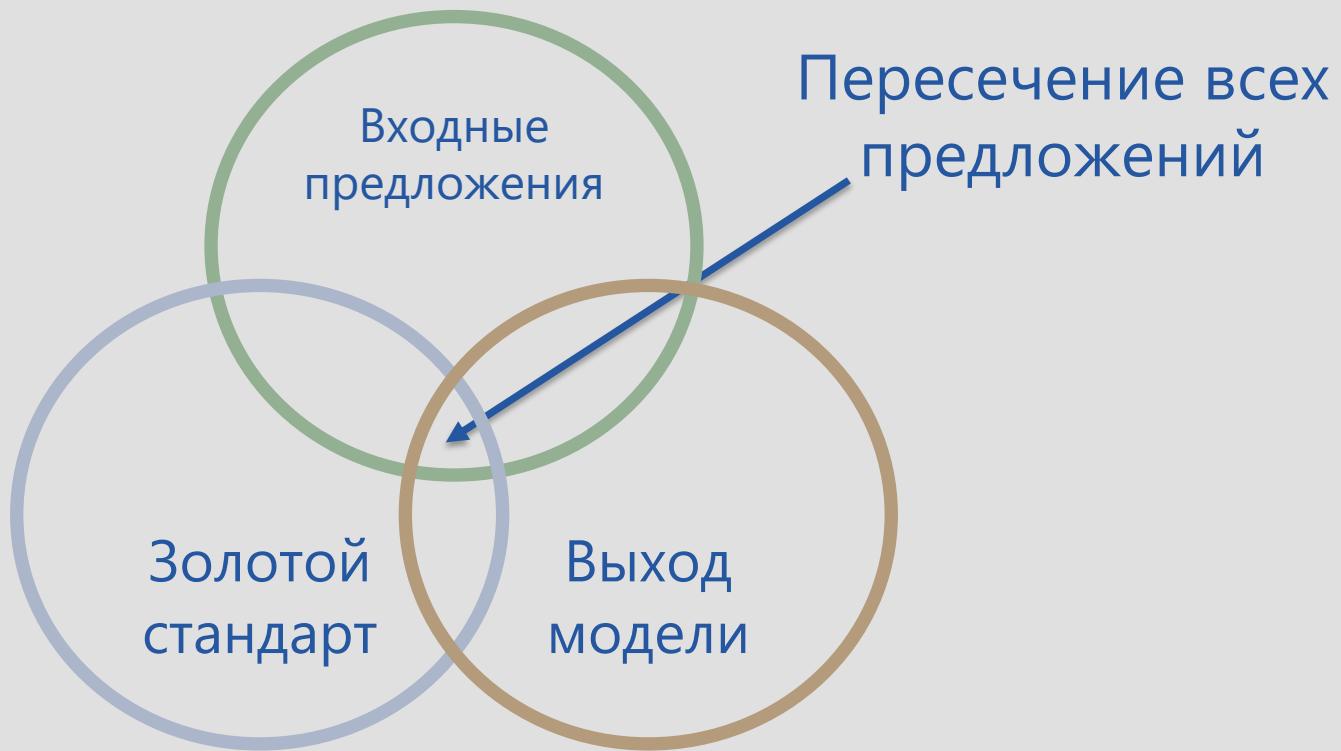


- Xu, Wei, Courtney Napolis, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing statistical machine translation for text simplification." TACL, 2016



# Упрощение текстов

## Меры качества, SARI



- Xu, Wei, Courtney Napolis, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing statistical machine translation for text simplification." TACL, 2016



# Упрощение текстов

Меры качества, SARI

$$SARI = \frac{1}{3} F_{\text{добавление}} + \frac{1}{3} F_{\text{удаление}} + \frac{1}{3} F_{\text{сохранение}}$$

$ope \in \{\text{добавление}, \text{удаление}, \text{сохранение}\}$

$$F_{ope}(n) = \frac{2 \times p_{ope}(n) \times r_{ope}(n)}{p_{ope}(n) + r_{ope}(n)}$$

$$F_{ope} = \frac{1}{k} \sum_{n=1,k} F_{ope}(n)$$

- Xu, Wei, Courtney Napolis, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. "Optimizing statistical machine translation for text simplification." TACL, 2016



# Упрощение текстов

## Меры качества, индекс Флеша

$$FKGL = 0.39 \left( \frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left( \frac{\text{total syllables}}{\text{total words}} \right) - 15.59$$

В системе образования США, индекс Флеша интерпретируется следующим образом: сколько лет нужно учиться, чтобы с лёгкостью прочитать данный текст



# Теггирование последовательностей

## Для упрощения предложения

- Множество операций: удаление (D), замена (R), перемещение (M), добавление (A)
- Автоматическая разметка пар сложное предложение – простое предложение
- Обучается двунаправленная LSTM сеть с несколькими слоями
- И получает превосходные результаты

► Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C. and Specia, L..  
*Learning how to simplify from explicit labeling of complex-simplified text pairs.* ICJNLP, 2017



# Теггирование последовательностей

Для упрощения предложения

Hershey <sup>D D D D D</sup> left no heirs when he died in 1945 , <sup>R</sup> giving most of his fortune to charity .

Hersey died in 1945 <sup>A</sup> and <sup>gave</sup> most of his fortune to charity .

<sup>R</sup> <sup>M</sup>  
[...] DeJongh remembered [...]

[...] remembered Amparo DeJongh [...]

- Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C. and Specia, L..  
*Learning how to simplify from explicit labeling of complex-simplified text pairs.* ICJNLP, 2017



# Seq2seq модели для упрощения текста

## ACCESS (AudienCe-CEntric Sentence Simplification)

- Обычная архитектура Трансформер дает умеренные результаты в задаче упрощения текстов
- Использование модели BART дает несущественный прирост в качестве

► *Martin, L., de la Clergerie, É.V., Sagot, B., Bordes, A., 2020, May.  
Controllable Sentence Simplification. LREC, 2020*



# Seq2seq модели для упрощения текста

## ACCESS (AudienCe-CEntric Sentence Simplification)

- Однако прорыв достигается за счет управляемых кодов – специальных префиксов, которые помогают управлять поведением модели:
  - NbChars – соотношение длины в символах исходного и упрощенного предложения
  - LevSim – редакционное расстояния, количество операций, которое нужно совершить, чтобы получить из исходного предложения упрощенное
  - WordRank – ранг слова в частотном словаре

► *Martin, L., de la Clergerie, É.V., Sagot, B., Bordes, A., 2020, May.  
Controllable Sentence Simplification. LREC, 2020*



# Seq2seq модели для упрощения текста

## ACCESS (AudienCe-CEntric Sentence Simplification)

- Прорыв достигается за счет управляющих кодов – специальных префиксов, которые помогают управлять поведением модели
- Обучающие данные составляются с использованием контролирующих кодов
  - Исходное предложение: <NbChars 0.3> <LevSim 0.4> He settled in London , devoting himself chiefly to practical teaching
  - Упрощенное предложение: He teaches in London
- На этапе тестирования управляющие коды не используются

► *Martin, L., de la Clergerie, É.V., Sagot, B., Bordes, A., 2020, May.  
Controllable Sentence Simplification. LREC, 2020*



# Упрощение текстов

## Заключение

- Задача упрощения текстов отчасти похожа на задачу суммаризации: сохраняя смысл предложения, требуется сформулировать его как можно проще
- Многие исследователи подчеркивают социальную значимость задачи упрощения текста
- Лидирующие подходы к решению задачи используют предобученные языковые модели и архитектуры машинного перевода
- Основные сложности связаны с дефицитом параллельных данных и сложностью субъективной оценки

► *Martin, L., de la Clergerie, É.V., Sagot, B., Bordes, A., 2020, May.  
Controllable Sentence Simplification. LREC, 2020*

