

Методы предобработки текстовых данных для ускорения обучения языковых моделей

Сурков Максим Константинович

Научный руководитель: Ямщиков Иван Павлович

Санкт-Петербургская школа физико-математических и компьютерных наук
НИУ ВШЭ СПб

17 марта 2021 г.

Обработка естественного языка в реальной жизни

- социальные сети
- электронная почта
- службы доставки
- голосовые помощники
- переводчики
- чат боты



- ❶ классификация последовательностей
 - спам
 - грубая речь¹
- ❷ генерация выходной последовательности из исходной
 - машинный перевод
 - ответы на вопросы
- ❸ выделение информации из последовательностей
 - выделение именованных сущностей²

¹G. H. Paetzold et al., SemEval'19 Task 5: Hate Speech Identification with RNN.

²Vikas Yadav et al., SemEval'19 Task 12: Deep-Affix Named Entity Recognition of Geolocation Entities. ACL'19

Современные методы решения задач обработки естественного языка

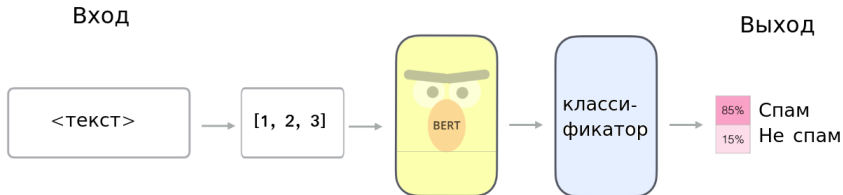
- ❶ Механизм внимания¹
- ❷ **BERT** (Google)²
- ❸ GPT-3 (OpenAI)³

¹Ashish Vaswani et al., Attention Is All You Need, 2017

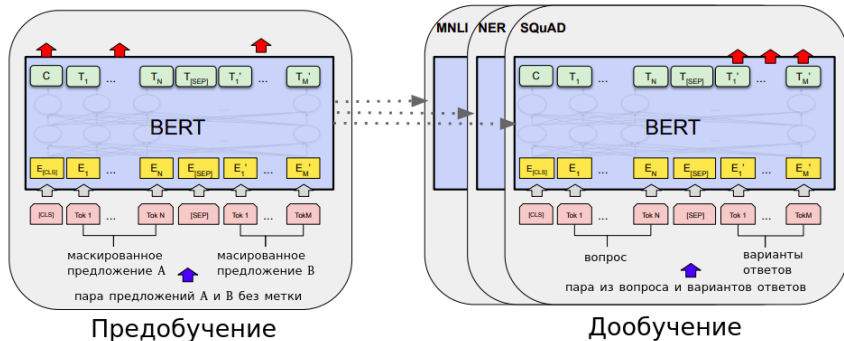
²Jacob Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019

³Tom B. Brown et al., Language Models are Few-Shot Learners, 2020

BERT. Использование



BERT. Обучение



BERT. Требуемые ресурсы

- количество параметров: 110M – 340M
- время на предобучение: от 2-4 дней до 1-2 недель¹
 - мировой рекорд: 47 минут на **1472** V100 GPU²
- время на дообучение: 1-2 дня
- размеры данных:

Набор данных	Размер
Wikipedia	3-600M
HND	600k-2M
s140	1.6M
IWSLT	200-230k
QQP	364k
MNLI	393k

¹При использовании 1x-4x GPU Nvidia Tesla V100 32Gb

²<https://developer.nvidia.com/blog/training-bert-with-gpus>

BERT. Существующие методы оптимизации

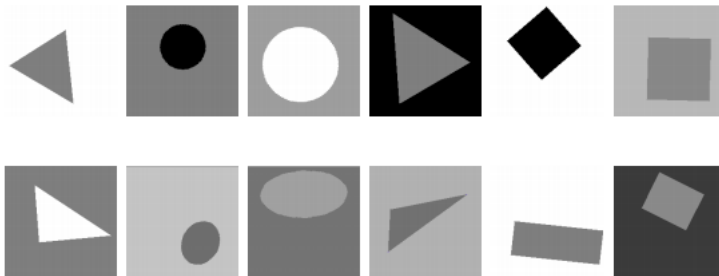
- квантизация¹
- дистилляция²
- прунинг³

¹Sheng Shen et al., Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT, 2019

²Victor Sanh et al., DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020

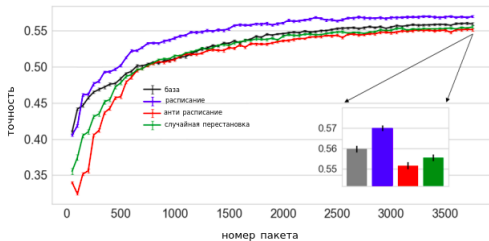
³Hassan Sajjad et al., Poor Man's BERT: Smaller and Faster Transformer Models, 2020

Обучение с расписанием. Начало



Y. Bengio et al., Curriculum learning, 2009

- компьютерное зрение¹



- обучение с подкреплением²

- глубокое обучение³

¹Guy Hach Cohen, Daphna Weinshall, On The Power of Curriculum Learning in Training Deep Networks, 2019

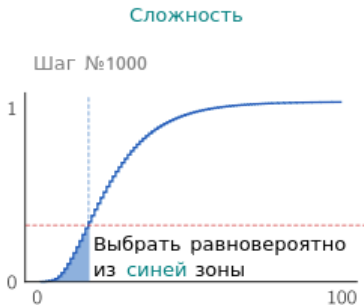
²Sanmit Narvekar et al., Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey, 2020

³Mermer et al., Scalable Curriculum Learning for Artificial Neural Networks, 2017

Обучение с расписанием в обработке языка

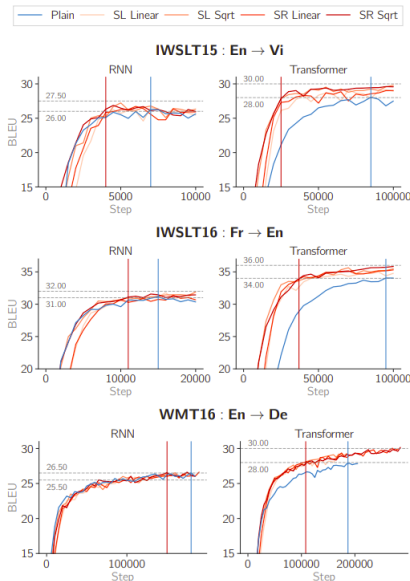
- Задача: машинный перевод
- Модель: BERT, LSTM
- Датасеты: IWSLT'15, IWSLT'16, WMT'16
- Алгоритм:

- 1 сортируем тексты по сложности (длина, логарифм вероятности правдоподобия)
- 2 в течение T шагов (рассмотрим шаг t)
 - считаем $c(t) \in [0, 1]$
 - строим батч из $c(t)$ **первых** текстов корпуса
 - шаг обучения



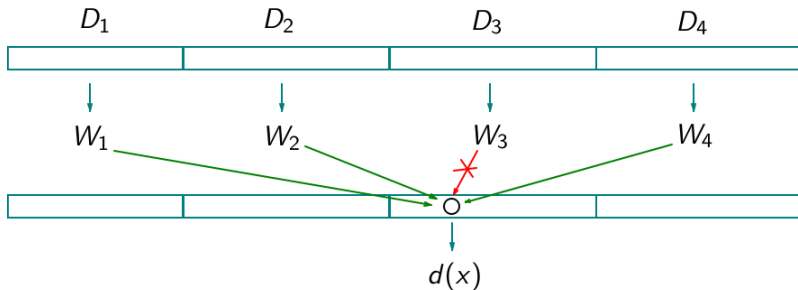
Е. А. Platanios et al., Competence-based Curriculum Learning for Neural Machine Translation, ACL'19

Обучение с расписанием в обработке языка



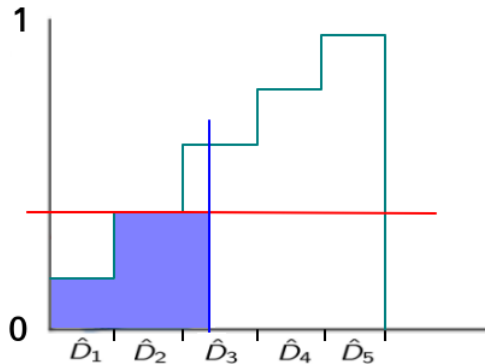
Обучение с расписанием в обработке языка

- Задача: классификация
- BERT
- Датасеты: SQuAD 2.0, NewsQA, GLUE
- Алгоритм: в течение T шагов



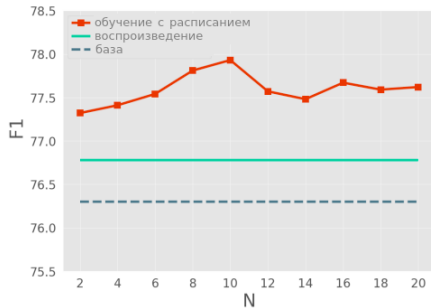
Benfeng Xu et al., Curriculum Learning for Natural Language Understanding, ACL'20

Обучение с расписанием в обработке языка



Обучение с расписанием в обработке языка

	MNLI-m	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	Avg
<i>results on dev</i>									
BERT Large	86.6	92.3	91.3	70.4	93.2	88.0	60.6	90.0	84.1
BERT Large*	86.6	92.5	91.5	74.4	93.8	91.7	63.5	90.2	85.5
BERT Large+CL	86.6	92.8	91.8	76.2	94.2	91.9	66.8	90.6	86.4
<i>results on test</i>									
BERT Large	86.7	91.1	89.3	70.1	94.9	89.3	60.5	87.6	83.7
BERT Large*	86.3	92.2	89.5	70.2	94.4	89.3	60.5	87.3	83.7
BERT Large+CL	86.7	92.5	89.5	70.7	94.6	89.6	61.5	87.8	84.1



Обучение с расписанием в обработке языка.

Направления для исследований

- Много важных задач обработки естественного языка с **большими** корпусами тренировочных данных
- Решаются с помощью **тяжелых** моделей, которые **долго** учатся
- Не исследованы метрики оценки сложности текста (длина - текущий предел)
- Эксперименты проведены только на определенных задачах
 - ACL'19 - только задача машинного перевода
 - ACL'20 - только задача классификации¹
- Не исследовано влияние обучения с расписанием на этапе предобучения

¹Не совсем честное обучение с расписанием; Не ускоряет; Требуется еще больших ресурсов

Цель: ускорить обучение языковой модели BERT с помощью обучения с расписанием за счет метрики оценки сложности текстовых данных на задачах предобучения, классификации и машинного перевода

Задачи:

- 1 Найти эффективные¹ метрики оценки сложности текста
- 2 Реализовать механизм подсчета найденных метрик на больших датасетах
- 3 Сравнить найденные метрики с существующими метриками оценки сложности текста
- 4 Исследовать влияние найденных метрик на скорость обучения языковой модели BERT

¹с точки зрения сокращения скорости обучения модели

- ❶ длина, вероятность правдоподобия¹
- ❷ информационный поиск
 - **tf-idf**
 - энтропия, семантическая сложность²
- ❸ средняя частота слова, самое редкое слово в предложении³
- ❹ число определенных частей речи⁴
- ❺ теория информации

¹E. A. Platanios et al., Competence-based Curriculum Learning for Neural Machine Translation, ACL'19

²Frans van der Sluis et al., Using Complexity Measures in Information Retrieval, 2010

³Xuan Zhang et al., An Empirical Exploration of Curriculum Learning for Neural Machine Translation, 2018

⁴Tom Kocmi, Ondrej Bojar, Curriculum Learning and Minibatch Bucketing in Neural Machine Translation, 2017

метрика	формула
Мультиинформация	$\sum_{v \in V} H_p(X_v) - H_p(X_V)$
Избыточная энтропия (ЕЕ)	$\left[\sum_{v \in V} H(X_{V \setminus \{v\}}) \right] - (n - 1)H(X_V)$
TSE	$\sum_{k=1}^{n-1} \frac{k}{n} C^{(k)}(X_V), \text{ где}$ $C^{(k)}(X_V) =$ $\frac{n}{k \binom{n}{k}} \sum_{A \subseteq V, A =k} H(X_A) - H(X_V)$
Переходная информация	:(

$$V = \{1, \dots, n\}$$

$$X_V = (X_1, \dots, X_n)$$

Nihat Ay et al., A **Unifying** Framework for Complexity Measures of Finite Systems,

1 Образование совместной случайной величины

$$T = (t_1, t_2, \dots, t_{i-1}, t_i, \dots, t_n)$$

$t_i \rightarrow \xi_{t_i}^i =: \mu_i$ – бинарная случайная величина



$$\xi = (\xi_{t_1}^1, \xi_{t_2}^2, \dots, \xi_{t_{i-1}}^{i-1}, \xi_{t_i}^i, \dots, \xi_{t_n}^n)$$

2 Вычисление энтропии

$$H(\mu) = \sum_{i=1}^n H(\mu_i | \mu_1, \mu_2, \dots, \mu_{i-1}) = \sum_{i=1}^n H(\mu_i | \mu_{i-L}, \dots, \mu_{i-1})$$

3 $L = 1$

$$H(\mu) = H(\mu_1) + H(\mu_2 | \mu_1) + \dots + H(\mu_i | \mu_{i-1}) + \dots + H(\mu_n | \mu_{n-1})$$

1 длина

2 tf-idf

$$\sum_{i=1}^n f(X_i) \log \frac{|D|}{|\{j : X_i \in X^{(j)}\}|}$$

- $x_i \rightarrow$ число текстов, в которых есть x_i

3 энтропия для вычисления EE, TSE

- длина \rightarrow число текстов с такой длиной
- $(i, x_i) \rightarrow$ число текстов, где $t_i = x_i$
- $(x_i) \rightarrow$ число текстов, где x_i является последним токеном
- $(i, x_{i-1}, x_i) \rightarrow$ число текстов, где на $(i-1)$ -й позиции стоит x_{i-1} , а на i -й позиции стоит x_i

4 EE, TSE - ?

$$EE(X) = \left[\sum_{v \in V} H(X_{V \setminus \{v\}}) \right] - (n-1)H(X_V) =$$

$$\left[\sum_{i=1}^n H(\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_n) \right] - (n-1)H(\mu)$$

- $\mathcal{O}(n^2)$
- $\mathcal{O}(n)$

$$\sum_{i=1}^n H(\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_n) =$$

$$= \sum_{i=1}^n H(\mu) - H(\mu_i | \mu_{i-1}) - H(\mu_{i+1} | \mu_i) + H(\mu_{i+1})$$

$$EE(X) = \sum_{i=2}^n H(\mu_i) - H(\mu_i | \mu_{i-1}) = \sum_{i=2}^n I(\mu_{i-1} : \mu_i)$$

$$\sum_{k=1}^{n-1} \frac{k}{n} C^{(k)}(X_V)$$
$$C^{(k)}(X_V) = \frac{n}{k \binom{n}{k}} \sum_{A \subseteq V, |A|=k} H(X_A) - H(X_V) =$$
$$= \frac{n}{k} \left[\frac{1}{\binom{n}{k}} \sum_{A \subseteq V, |A|=k} H(X_A) \right] - H(X_V)$$

$$\frac{1}{\binom{n}{k}} \sum_{A \subseteq V, |A|=k} H(X_A) = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq n} H(\mu_{i_1}, \mu_{i_2}, \dots, \mu_{i_k})$$

- ❶ $\mathcal{O}^*(2^n)$
- ❷ $\mathcal{O}(n^2)$ - динамическое программирование
- ❸ $\mathcal{O}(n)$

$$\sum_{i=1}^n A_i H(\mu_i) + \sum_{i=2}^n B_i H(\mu_i | \mu_{i-1})$$

$$A_i = \begin{cases} \binom{n-2}{k-1} / \binom{n}{k} = \frac{k(n-k)}{n(n-1)}, & i > 1 \\ \binom{n-1}{k-1} / \binom{n}{k} = \frac{k}{n}, & i = 1 \end{cases}$$

$$B_i = \frac{\binom{n-2}{k-2}}{\binom{n}{k}} = \frac{k(k-1)}{n(n-1)}$$

- наборы данных

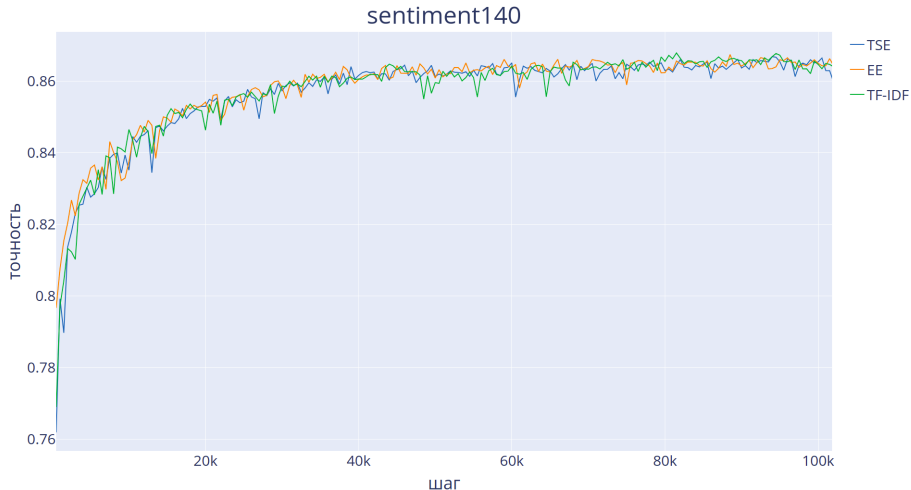
Набор данных	Размер
Hyperpartisan News Detection ¹	600k-2M
sentiment140	1.6M

- метрика качества модели – точность
- модель BERT-base
- метод сравнения метрик сложности текста
 - 1 фиксируем модель
 - 2 фиксируем датасет
 - 3 фиксируем семплер
 - 4 учим модели, используя сравниваемые метрики
 - 5 анализируем график обучения модели

¹SemEval-2019 Task 4

Сравнение метрик

Без семплера



Сравнение метрик

семплер из ACL'19

sentiment140 + ACL'19



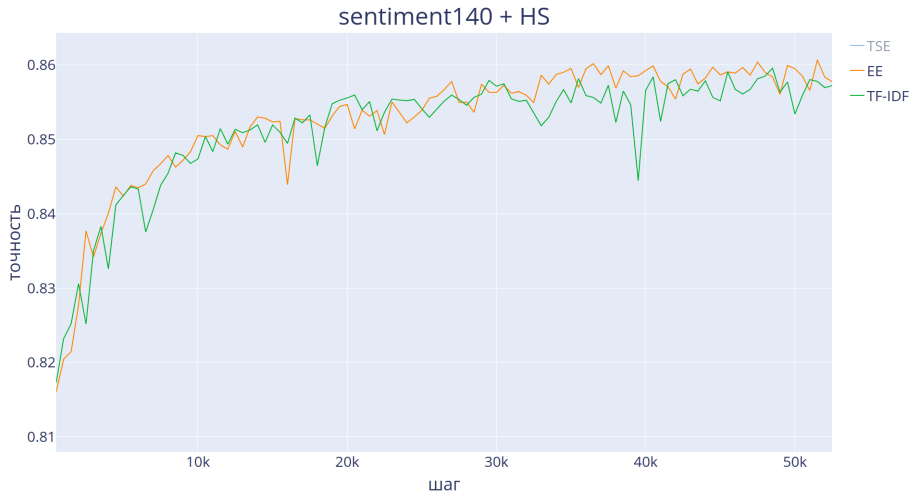
Сравнение метрик

семплер DB



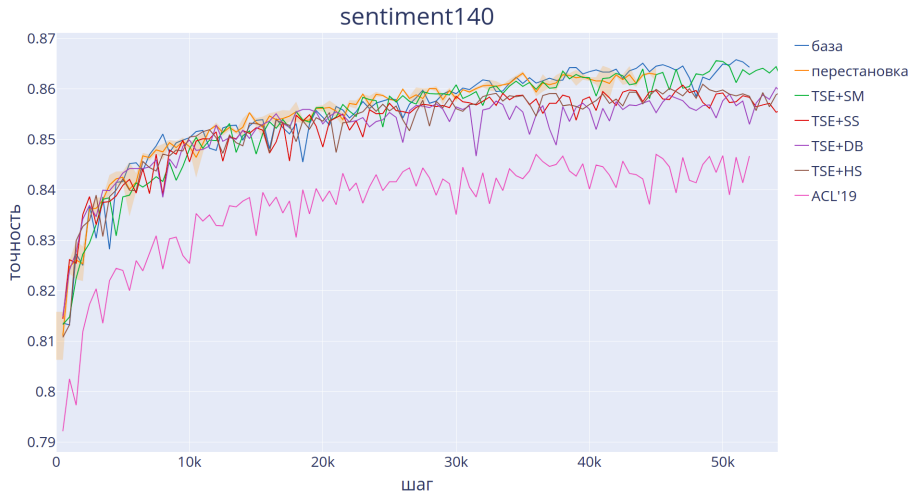
Сравнение метрик

гиперболический семплер



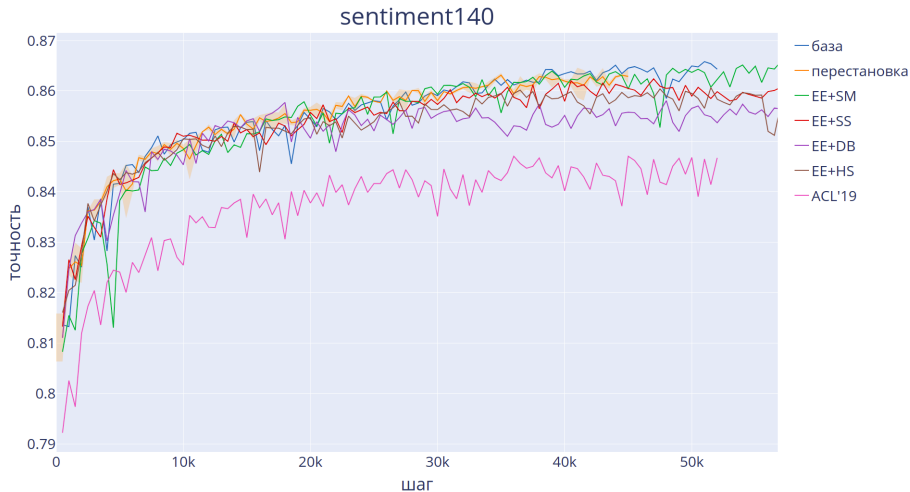
Влияние метрик на скорость обучения

TSE



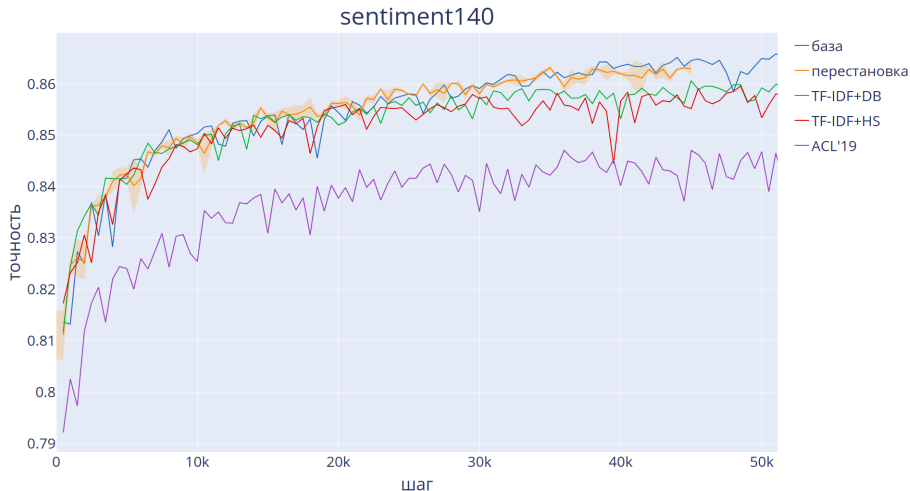
Влияние метрик на скорость обучения

EE



Влияние метрик на скорость обучения

TF-IDF



- ❶ Найдены метрики оценки сложности текста
 - метрики TSE и EE адаптированы под задачу обработки языка
 - $(TSE \approx EE) > TF-IDF > \text{длина}$
- ❷ Реализован механизм подсчета метрик на больших объемах данных
 - реализован механизм подсчета статистик для вычисления метрик
 - реализованы алгоритмы вычисления метрик
- ❸ Проведено сравнительное исследование метрик
 - задача классификации (sentiment140, HND)
 - показано, что влияние метрики зависит от семплера
 - показано ускорение обучения относительно существующих результатов¹
- ❹ Пока не удалось добиться существенного ускорения на задаче классификации относительно базового подхода

¹ACL'19

Дальнейший план работы

- исследовать отношения метрик и их влияние на скорость обучения на задаче машинного перевода
- попытаться обобщить подход вычисления TSE и EE на большие k
- попытаться добиться ускорения обучения на задаче классификации